Single base substitution causing the fragrant phenotype and development of a type-specific marker in aromatic coconut (*Cocos nucifera*)

A. Vongvanrungruang[1], C. Mongkolsiriwatana[2,3], T. Boonkaew[1], O. Sawatdichaikul[4], K. Srikulnath[1,5,6] and S. Peyachoknagul[1,5,7]

[1]Department of Genetics, Faculty of Science, Kasetsart University, Chatuchak, Bangkok, Thailand
[2]Division of Genetics, Department of Science, Faculty of Liberal Arts and Science, Kasetsart University, Kamphaeng Saen Campus, Nakhon Pathom, Thailand
[3]Research Unit of Genetic Technology and Applications, Department of Science, Faculty of Liberal Arts and Science, Kasetsart University, Kamphaeng Saen Campus, Nakhon Pathom, Thailand
[4]Functional Food Unit, Division of Nutrition and Health, Institute of Food Research and Product Development, Kasetsart University, Chatuchak, Bangkok, Thailand
[5]Center for Advanced Studies in Tropical Natural Resources, National Research University-Kasetsart University, Chatuchak, Bangkok, Thailand
[6]Laboratory of Animal Cytogenetics and Comparative Genomics, Department of Genetics, Faculty of Science, Kasetsart University, Chatuchak, Bangkok, Thailand
[7]Department of Biology, Faculty of Science, Naresuan University, Phitsanulok, Thailand

Corresponding author: S. Peyachoknagul
E-mail: fscisrp@ku.ac.th

Genet. Mol. Res. 15 (3): gmr.15038748
Received April 29, 2016
Accepted June 8, 2016
Published September 19, 2016
DOI http://dx.doi.org/10.4238/gmr.15038748

Copyright © 2016 The Authors. This is an open-access article distributed under the terms of the Creative Commons Attribution ShareAlike (CC BY-SA) 4.0 License.

ABSTRACT. The fragrance gene, betaine aldehyde dehydrogenase 2 (*Badh2*), has been well studied in many plant species. The objectives of this study were to clone *Badh2* and compare the sequences between aromatic and non-aromatic coconuts. The complete coding region was cloned from cDNA of both aromatic and non-aromatic coconuts. The nucleotide sequences were highly homologous to *Badh2* genes of other plants. *Badh2* consisted of a 1512-bp open reading frame encoding 503 amino acids. A single nucleotide difference between aromatic and non-aromatic coconuts resulted in the conversion of alanine (non-aromatic) to proline (aromatic) at position 442, which was the substrate binding site of BADH2. The ring side chain of proline could destabilize the structure leading to a non-functional enzyme. *Badh2* genomic DNA was cloned from exon 1 to 4, and from exon 5 to 15 from the two coconut types, except for intron 4 that was very long. The intron sequences of the two coconut groups were highly homologous. No differences in *Badh2* expression were found among the tissues of aromatic coconut or between aromatic and non-aromatic coconuts. The amino acid sequences of BADH2 from coconut and other plants were compared and the genetic relationship was analyzed using MEGA 7.0. The phylogenetic tree reconstructed by the Bayesian information criterion consisted of two distinct groups of monocots and dicots. Among the monocots, coconut (*Cocos nucifera*) and oil palm (*Elaeis guineensis*) were the most closely related species. A marker for coconut differentiation was developed from one-base substitution site and could be successfully used.

Key words: Coconut; Fragrance gene; Differentiation marker; *Badh2* gene

INTRODUCTION

The coconut tree (*Cocos nucifera* Linn.) is a member of the palm family Arecaceae and grows throughout tropical countries. Coconuts are generally classified as tall or dwarf varieties. They are found in all areas of Thailand, but are mainly grown commercially in the central region of the country. Aromatic coconut is derived from a mutation in the dwarf-type variety 'Musi' (Tangchatuporn, 2000). The juice of fresh young aromatic coconuts is sweet and has a pandan leaf-like aroma (Tangchatuporn, 2000). Aromatic coconut juice is very popular among foreigners and local people. The fragrant smell of coconut juice is similar to that of pandan leaf (*Pandanus amaryllifolius* Roxb.), aromatic rice (*Oryza sativa* L.), and the flower of *Vallaris glabra* (L.) Kuntze (Wongpornchai et al., 2003), and is mainly derived from the aromatic compound 2-acetyl-1-pyrroline (2AP). This chemical was first discovered in cooked rice by Buttery et al. (1982) and is also found in volatile oil from pandan leaf (Buttery et al., 1983). In the aromatic rice varieties Khao Dawk Mali 105 and Basmati, the quantity of 2AP is about 100-times higher than in non-aromatic varieties (Grosch and Schieberle, 1997), making these two varieties very popular and valuable on the world market.

The pathway of 2AP biosynthesis in *P. amaryllifolius*, *V. glabra*, coconut, and in other plants has not been reported. However, in rice, biosynthesis of this aromatic compound results from the metabolism of 4-aminobutyraldehyde (AB-ald) to 1-pyrroline, which is then converted to 2AP due to the lack of the functional enzyme, betaine aldehyde dehydrogenase homologue 2 (BADH2) (Yoshihashi et al., 2002; Vanavichit et al., 2005).

BADH proteins contain an NAD-dependent aldehyde dehydrogenase domain and a cysteine active site, which are highly conserved and necessary for the functional activity of these enzymes. In rice, BADH1 and BADH2 differ in an aldehyde dehydrogenase cysteine active site and are encoded by two different genes, *Badh1* and *Badh2*, respectively. *Badh1* is located on chromosome 4 and is involved in stress tolerance, whereas *Badh2* is located on chromosome 8 and is involved in 2AP accumulation (Nakamura et al., 2001; Yoshihashi et al., 2004; Bradbury et al., 2005). Aromatic and non-aromatic rice carry different alleles of *Badh2*, with the recessive allele, *badh2*, found in aromatic rice varieties. Mutation of *Badh2* results from an 8-bp deletion and 3 single nucleotide polymorphisms (SNPs) in exon 7, an insertion of 7 bp in exon 8, or a deletion of 7 bp from exon 2. These mutations generate a premature stop codon and a non-functional truncated protein, resulting in the subsequent accumulation of 2AP. The functional wild-type BADH2 enzyme catalyzes the conversion of AB-ald to 4-aminobutyric acid (GABA), which reduces the accumulation of 2AP (Bradbury et al., 2005; Amarawathi et al., 2008; Shi et al., 2008). Transformation of aromatic rice with the wild-type *Badh2* significantly reduced the amount of 2AP (Bradbury et al., 2008; Chen et al., 2008). Mutation of *Badh2* was also reported to generate mutants with aromatic flavor in soybean (Arikit et al., 2011; Juwattanasomran et al., 2011), sorghum (Yundaeng et al., 2013), and cucumber (Yundaeng et al., 2015).

In the present study, *Badh2* was cloned from aromatic and non-aromatic coconut. The sequence and expression of *Badh2* were analyzed in order to identify the allele responsible for the fragrant phenotype.

MATERIAL AND METHODS

Plant materials and DNA extraction

Aromatic and non-aromatic coconut leaves were collected from the experimental field of Chumphon Horticultural Research Center, Chumphon Province, Thailand. Root and fruit pericarp samples were collected from aromatic coconuts for RNA extraction. Total genomic DNA was extracted from leaves as described by Agrawal et al. (1992), and kept at -80°C until required. The quality and concentration of DNA were measured by 1% agarose gel electrophoresis and spectrophotometric analysis.

RNA extraction and reverse transcription

Total RNA was extracted from 4-6 samples of coconut roots, leaves, and fruit pericarps using the cetyltrimethylammonium bromide method (Yu and Goh, 2000), and lithium chloride precipitation as described by Stiekema et al. (1988). DNA was removed from total RNA samples using RQ1 RNase-Free DNase (Promega, Madison, WI, USA). DNaseI-treated RNA (5 μg) was added to 1X reaction buffer, 5 μM oligo$(dT)_{18}$ primer, 20 U RiboLock RNase inhibitor, 1 mM dNTP mix, and 200 U RevertAid M-MuLV reverse transcriptase enzyme

and reverse transcribed using Thermo Scientific RevertAid First-Strand cDNA Synthesis Kit (Thermo Scientific, USA), following the manufacturer instructions. The obtained cDNA was used as a template for PCR amplification.

Cloning of the full-length cDNA of *Badh2*

CnBadh2-cds1 primers (Table 1) were designed based on nine nucleotide sequences of *Badh2* from four plant species deposited in the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov). Genomic DNA (75 ng) from aromatic or non-aromatic coconut samples was added to 50 µL 1X PS GXL buffer, 0.2 mM dNTPs, 0.25 µM each of forward and reverse primers, and 0.375 U GXL DNA polymerase (TaKaRa, Japan), and PCR was performed with an initial denaturation at 94°C for 5 min, followed by 29 cycles at 98°C for 10 s, 55-60°C for 15 s, and 68°C for 90 s, and a final extension of 68°C for 8 min. The PCR products were detected by electrophoresis on 1% agarose gel. The DNA fragments were subsequently extracted from the ethidium bromide-stained gel, cloned using pGEM-T Easy Vector System I (Promega), and transformed into DH5α competent *Escherichia coli*. Nucleotide sequences of the DNA fragments were determined by Macrogen Co., Ltd. (South Korea). The new primers were designed from the obtained sequences in order to clone full-length cDNA by semi-nested PCR.

Table 1. Names and sequences of the primers used to clone full-length cDNA.

Primer name	Primer sequence (5'-3')	Ta (°C)	Size range (bp)
CnBadh2-cds1	F: GATGAAGCDGCATGGGACAT R: ATCTTCTCATACTGTCCTTCACTAACA	55	717
CnBadh2-cds2	F: AGCATGTCGGGCGCGATCCCT R: ACTGCTTCAGGCACATAACATGTCTAC	55	1,512
Badh2-3N1	GATGAAGCDGCATGGGACAT	55	
Badh2-3N2	GATGTGGATATTGAGAAAGCTGTTG	55	
Badh2-5N1	CACCCAAAGAGAGTCCACTCAAC	55	
Badh2-5N2	ACCCGGTCACTATGTTTAGGACACC	55	
Oligo(dT)$_{15}$VN	TTTTTTTTTTTTTTTVNNNN		
AD	NTGCGASWGANAWGAA		

To amplify the 5'- and 3'-regions of *Badh2*, new Badh2-3N1, Badh2-3N2, Badh2-5N1, and Badh2-5N2 (Table 1) primers were designed from the partial *Badh2* gene sequences obtained from the previously isolated fragment, and used together with the arbitrary degenerate primer (AD) (Liu and Whittier, 1995) or Oligo(dT)$_{15}$VN primer. The positions of the primers are shown in Figure 1. The 5'- and 3'-regions of *Badh2* were amplified by semi-nested PCR using the same cycling condition previously described. The PCR products were detected by electrophoresis on 1% agarose gel. The DNA fragments were subsequently cloned into pGEM-T easy vector (Promega) and sequenced. The nucleotide sequences were assembled using the CAP3 program (http://doua.prabi.fr/software/cap3) and primers for the complete coding region, named CnBadh2-cds2, (Table 1) were designed. Full-length cDNAs were cloned from aromatic and non-aromatic coconuts. Their sequences were compared against the NCBI database using the BLASTx and BLASTn programs (http://blast.ncbi.nlm.nih.gov/Blast.cgi). All nucleotide sequences were determined from three to five DNA fragments amplified with corresponding primers. The sequences from both coconut types were compared

using CLUSTAL O (1.2.1) multiple-sequence alignment (http://www.ebi.ac.uk/Tools/msa/clustalo/). The complete sequences of *Badh2* from aromatic and non-aromatic coconuts were submitted to the DNA data bank of Japan (DDNJ) under accession numbers LC119054 and LC119055, respectively.

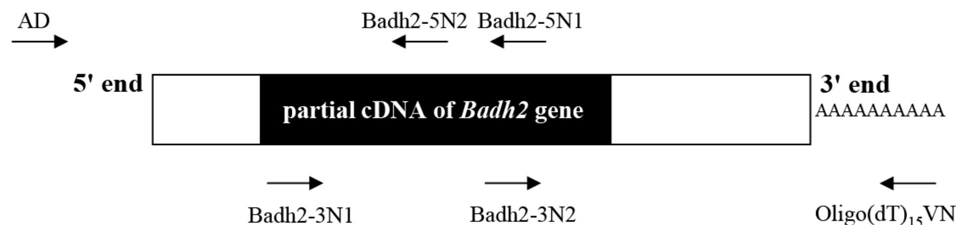

Figure 1. Position and orientation of the primers used to clone the 5'- and 3'-region of *Badh2*.

Homology modeling and protein structural analysis of BADH2

The encoded protein sequences of alleles *A* and *B* of coconut BADH2 (accession numbers LC119055 and LC119054, respectively), were used to identify potential protein templates via protein-protein BLAST (BLASTp). According to the BLAST results, the structure of the related protein, aminoaldehyde dehydrogenase 1a from *Zea mays* (zmamadh1a), PDB 4I8P chain A, was selected as the protein template. Selected template coordinates were then retrieved from the Protein Data Bank (www.pdb.org). Homology models of alleles *A* and *B* of coconut BADH2 were generated in SWISSMODEL WORKSPACE (Biasini et al., 2014). The proposed models for BADH2 were then verified by PROCHECK (Laskowski et al., 1993) by submitting the coordinates to the Structure Analysis and Verification Server version 4, hosted by UCLA (http://services.mbi.ucla.edu/SAVES/). Illustrations of the structural proteins were generated by Discovery Studio 3.5 Visualizer (Accelrys Inc., San Diego, CA, USA).

Sequence analysis and phylogenetic tree reconstruction

Nucleotide sequence comparisons were performed against NCBI using the BLASTx and the BLASTn programs (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Then, multiple-sequence alignment was performed using CLUSTAL O (1.2.1) (McWilliam et al., 2013; Kumar et al., 2016) to compare the putative amino acid sequences with known sequences from other *Badh2* genes deposited in NCBI. All protein sequences were aligned using the default parameters of Molecular Evolutionary Genetics Analysis version 7.0 (MEGA 7) for larger datasets (Kumar et al., 2016). The amino acid substitution model and parameters were calculated by Aminosan (Tanabe, 2011), based on the Bayesian information criterion. The phylogenetic tree was reconstructed in MrBayes 3.2 (Ronquist et al., 2012). A total of 10,000 trees were constructed from 1,000,000 generations by sampling every 100 generations. The first 2500 trees (25%) were discarded as burn-in.

Cloning of *Badh2* from genomic DNA

The genomic DNA sequence of the plant *Badh2* was retrieved from GenBank (http://www.ncbi.nlm.nih.gov/) and compared to the cDNA sequence of coconut *Badh2* to determine

the position of all exons and introns. Five primer pairs, named CnBadh2-G1, CnBadh2-G2, CnBadh2-G3, CnBadh2-G4, and CnBadh2-G5 (Table 2), were designed to amplify *Badh2* from genomic DNA of aromatic and non-aromatic coconuts. The PCR products were cloned and sequenced. The new primers were designed as appropriate to clone the entire gene by primer walking. The contiguous sequences of aromatic coconut were submitted to DDBJ under accession numbers LC125458 and LC125459, while the contiguous sequences of non-aromatic coconut were submitted under the accession numbers LC125460 and LC125461.

Table 2. Names and sequences of the primers used to clone genomic DNA.

Primer name	Primer sequence (5'-3')	Ta (°C)	Size range (bp)
CnBadh2-G1	F: AGCATGTCGGGCGCGATCCCT R: CCAAGGCGTCTGCAAGATCTGC	55	1056
CnBadh2-G2	F: GCTGCTGGTTGTACAGCTG R: CCACCAAGTTCCAATGAAACAGG	55	664
CnBadh2-G3	F: ATGACTGCTGCTGCACAAAC R: GCTGAACCAGACAACTCAGAC	55	936
CnBadh2-G4	F: GACAAATGGTCAAATATGCAGTGC R: TCACTGCTTCAGGCACATAAC	55	3013
CnBadh2-G5	F: GAGCTGGCTAAGCTAGAGACG R: TGCAGCAGCCATAATCC	55	10,000

Gene expression analysis by semi-quantitative RT-PCR

PCR primers, CnBadh2 (Table 3), were designed from the exon region of *Badh2*. Gene expression was determined by semi-quantitative PCR using 50 ng first-strand cDNA from the root, leaf, and fruit pericarp as a template. Other PCR components and cycling conditions were as previously described; however, the number of cycles varied, with 26, 28, 30, and 32 cycles used for the semi-quantification of gene expression in each sample. To control for RNA quantity and quality, primers used to amplify part of the actin gene were added to the reaction mixture as a duplex PCR (Thanananta et al., 2006). The PCR products were detected by electrophoresis on 1.5% agarose gel.

Table 3. Names and sequences of the primers used for semi-quantitative RT-PCR.

Primer name	Primer sequence (5'-3')	Ta (°C)	Size range (bp)
CnBadh2	F: AGCGCTGCAACCGTGTATC R: CTACAGCTTGGAGGGAGGTG	55	222
Actin	F: AAACGCGTATGGCTGATGCTGAGGATATCC R: AATCCAGCACGATACCAGTTGTACG	55	480

Development of DNA markers to differentiate aromatic and non-aromatic coconuts

Nucleotide sequences of *Badh2* from aromatic and non-aromatic coconuts were aligned using the default parameter of CLUSTAL O (1.2.1) multiple-sequence alignment. Two primer pairs (Table 4) were designed based on the differences identified in the sequences. The first primer pair, CnBadh2-m1, was used to amplify the DNA of aromatic coconut, while the second pair, CnBadh2-m2, was specific for non-aromatic coconut.

Table 4. Names and sequences of the primers used to differentiate coconuts.

Primer name	Primer sequence (5'-3')	Ta (°C)	Size range (bp)
CnBadh2-m1	F: TGCTCTGCAGGAAATTCAAC R: CTACAGCTTGGAGGGAGGTG	65	290
CnBadh2-m2	F: ACCTAGGTTTATGCATGCAC R: GCAGTTTATCCATACAATTCCAGC	67	553

RESULTS

Cloning of full-length *Badh2* cDNA

RNA from coconut roots was reverse transcribed to generate first-strand cDNA and then used for PCR amplification. First, the 717-bp fragment of *Badh2* was amplified using the CnBadh2-cds1 primer pair (Table 1). To amplify the 5'- and 3'-regions of the gene, new primers were subsequently designed based on the obtained gene sequence and used together with the AD primer (Liu and Whittier, 1995) or Oligo(dT)$_{15}$VN primer. Sequences and position of the primers are shown in Table 1 and Figure 1, respectively. The 3'-fragment was amplified by semi-nested PCR using Badh2-3N1 and Oligo(dT)$_{15}$VN primers in the first round, and Badh2-3N2 and Oligo(dT)$_{15}$VN primers in the second round. The 800-bp fragment PCR product was overlapped with the first fragment and extended to the 3'-end of the gene. Primers specific for the 5'-end of the gene were developed from the obtained sequence and used with AD primers. The first PCR was performed using Badh2-5N1 and AD primers. The product from the first round of PCR was used for the second PCR with the Badh2-5N2 and AD primers. The PCR product from the 700-bp fragment was overlapped with the first fragment and extended to the 5'-end of the gene. The sequences of the three fragments were then assembled into a complete *Badh2* gene, and the primer pair CnBadh2-cds2 was used to amplify the complete gene. *Badh2* was cloned from both aromatic and non-aromatic coconuts, and its sequences were compared by CLUSTAL O (1.2.1) multiple-sequence alignment. *CnBadh2* in both types of coconut consisted of 15 exons with a 1512-bp open reading frame encoding 503 amino acids. Three alleles of *Badh2* were identified from coconuts, named *A*, *B*, and *C*, where allele *A* was specific to a non-aromatic variety. There were two single-base substitutions in the coding sequence (CDS) of *CnBadh2* at positions 1307 and 1324 in exons 13 and 14, respectively. Guanine (G) was mutated to cytosine (C) in exon 14 resulting in an amino acid change from alanine (GCT; A) in non-aromatic coconuts to proline (CCT; P) in aromatic coconuts. Additionally, variation within non-aromatic coconuts was found in the third allele, in which base C in exon 13 was replaced by base T, resulting in an amino acid change from alanine (GCA; A) to valine (GTA; V) (**Figure S1**).

Homology modeling and structural analysis of the BADH2 protein

Amino acid sequence alignment of three BADH2 proteins revealed high similarity, with two amino acid differences at positions 436 and 442, and only one amino acid difference between non-aromatic and aromatic coconuts at position 442 (allele *A* and *B*). At this position, alanine was present in the non-aromatic coconut, and the wild-type allele changed to proline in the aromatic coconut. Amino acid sequences of the coconut BADH2 protein showed close similarity to other monocot species (Figure 2). Protein bioinformatics revealed that the primary sequence of the *A* allele shared 99.7% sequence identity with that of the *B* allele, with only one different residue identified, A_A^{442}/P_B^{442} (Figure 3).

```
Zea mays                  STETGKRIMTSAAQMVKPVSLELGGKSPLIVFDDIGDIDKAVEWTMFGIFANAGQVCSAT  300
Hordeum vulgare           SYATGQKIMVAAAPTVKPVTLELGGKSPIVVFDDV-DIDKAVEWTLFGCFWTNGQICSAT  298
Leymus chinensis          SYATGQKIMVAAAPTVKAVTLELGGKSPIVVFDDV-DIDKAVEWTLFGCFWTNGQICSAT  297
Zoysia tenuifolia         SSATGQKIMTAAATMVKPVSLELGGKSPIVVFDDV-DIDKAVEWTLFGCFWTNGQICSAT  298
Sorghum bicolor           SFETGKKIMAAAAPMVKPVTLELGGKSPIVVFDDV-DIDKAVEWTLFGCFWTNGQICSAT  299
Oryza sativa Japonica     SYETGKKIMASAAPMVKPVSLELGGKSPIVVFDDV-DVEKAVEWTLFGCFWTNGQICSAT  297
Oryza sativa Indica       SYETGKKIMASAAPMVKPVSLELGGKSPIVVFDDV-DVEKAVEWTLFGCFWTNGQICSAT  297
Musa acuminata            STETGKKIMTSAAQMVKPVSLELGGKSPIIVFEDV-EVEKAVEWTLFGCFWTNGQICSAT  299
Elaeis guineensis         STETGRRIMTAAAQTIKPVSLELGGKSPIIVFEDV-DIEKAVEWSLFGCFWTNGQICSAT  297
Cocos nucifera A          STETGRRIMTAAAQTIKPVSLELGGKSPIIVFEDV-DIEKAVEWTLFGCFWTNGQICSAT  297
Cocos nucifera B          STETGRRIMTAAAQTIKPVSLELGGKSPIIVFEDV-DIEKAVEWTLFGCFWTNGQICSAT  297

Zea mays                  SRLLLHEKIAKKFLDRLVAWAKNIKVSDPLEEGCRLGSVISEGQYEKIKKFISTAR----  356
Hordeum vulgare           SRLLIHKNIAKEFVDRMVAWSKNIKVSDPLEEGCRLGPVVSEGQYEKIKKFVANAK----  354
Leymus chinensis          SRLLIHKNIAKEFVDRMVAWSKNIKVSDPLEEGCRLGPVVSEGQYEKIKKFISNAK----  353
Zoysia tenuifolia         SRLLIHKKIAKEFIERMVAWAKNIKVSDPLEEGCRLGPVVSEGQYEKIKKFVANAK----  354
Sorghum bicolor           SRLLIHTKIAKEFNERMVAWAKNIKVSDPLEEGCRLGPVVSEGQYEKIKKFISNAK----  355
Oryza sativa Japonica     SRLILHKKIAKEFQERMVAWAKNIKVSDPLEEGCMLGPVVSEGQYEKIKQFVSTAK----  352
Oryza sativa Indica       SRLILHKKIAKEFQERMVAWAKNIKVSDPLEEGCRLGPVVSEGQYEKIKQFVSTAK----  352
Musa acuminata            SRLLLHETIAEEFLGKLVTWAKNIKVSDPLEEGCRLGPVVSEGQYEKINKFISTAK----  355
Elaeis guineensis         SRLLVHESISKEFMERLVAWAKNIKISNPLEEGCRLGPVVSEGQYEKIKKFISTAK----  353
Cocos nucifera A          SRLLVHESISKEFMERLVAWAKNIKISNPLEEGCRLGPVVGEGQYEKIKKFISTAK----  353
Cocos nucifera B          SRLLVHESISKEFMERLVAWAKNIKISNPLEEGCRLGPVVGEGQYEKIKKFISTAK----  353

Zea mays                  ----AGAVISNDQERCERISKALHSGIIWINCSQPCFVQAPWGGNKRSGFGRELGEWGLD  480
Hordeum vulgare           ----AGAVISGDRERCQRLAEEIEAGCIWVNCSQPCFCQAPWGGNKRSGFGRELGEGGID  478
Leymus chinensis          ----AGAVISGDRERCQRLAEEIDAGCIWVNCSQPCFCQAPWGGNKRSGFGRELGEGGID  477
Zoysia tenuifolia         ----AGAVLSADRERCQRVAEEIDAGAIWVNCSQPCFCQAPWGGNKRSGFGRELGDGGLD  478
Sorghum bicolor           ----AGAVISGDRERCQRLSEEIDAGCIWVNCSQPCFCQAPWGGNKRSGFGRELGEGGID  479
Oryza sativa Japonica     ----AGAVLSGDRERCQRLTEEIDAGIIWVNCSQPCFCQAPWGGNKRSGFGRELGEGGID  477
Oryza sativa Indica       ----AGAVLSGDRERCQRLTEEIDAGIIWVNCSQPCFCQAPWGGNKRSGFGRELGEGGID  477
Musa acuminata            ----AGAVISKDTERCQRISEMIQAGVIWVNCSQPCFCQAPWGGTKRSGFGRELGEWGLD  479
Elaeis guineensis         ----AGAVISKDPERCNRISEEIQAGIVWVNCSQPCFCQAPWGGNKRSGFGRELGEWGLD  477
Cocos nucifera A          ----AGAVISKDPERCNRASEEIQAGIVWINCSQPCFCQAPWGGNKRSGFGRELGEWGLD  477
Cocos nucifera B          ----AGAVISKDPERCNRASEEIQPGIVWINCSQPCFCQAPWGGNKRSGFGRELGEWGLD  477
```

Figure 2. Alignment of deduced amino acid sequences of the BADH2 protein from coconuts, showing close similarity of BADH2 sequences among monocot species. Gray areas represent the conserved domain required for functional activity of the gene. Amino acid differences are shown in red.

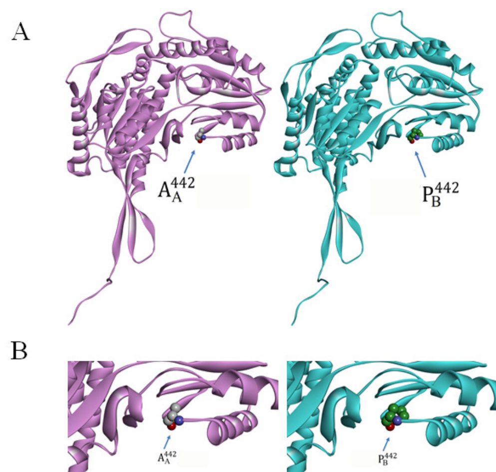

Figure 3. Tertiary structure of BADH2. **A.** Pink represents allele *A* and cyan represents allele *B*. **B.** Amino acid substitution site is enlarged.

Sequence analysis and phylogenetic tree reconstruction

BADH2 belongs to the AMADH superfamily of enzymes. To evaluate the evolutionary relationship of the coconut *Badh2* with those of various plant species, BADH2 protein sequences from 22 different plant species [*Arabidopsis thaliana* (AAK44148.1), *Brassica napus* (NP 001303116.1), *B. rapa* (XP 009150209.1), *Capsella rubella* (XP 006290943.1), *Camelina sativa* (XP 010503438.1), *Cocos nucifera* (LC119055), *Cucumis sativus* (AJF20760.1), *Elaeis guineensis* (XP 010913436.1), *Glycine max* (ADN03185.1), *Hordeum vulgare* (BAB62846.1),

Leymus chinensis (BAD86758.1), *Malus domestica* (AKC00600.1), *Musa acuminata* (XP 009417704.1), *Nicotiana tabacum* (AII99844.1), *O. sativa Indica* Group (ACF06149.1), *O. sativa Japonica* Group (ABI84118.1), *Solanum lycopersicum* (ACI43573.1), *Sorghum bicolor* (AGZ15751.1), *Tarenaya hassleriana* (XP 010550673.1), *Vitis vinifera* (XP 002283690.1), *Z. mays* (ACS74868.1), and *Zoysia tenuifolia* (BAD34949.1)] were obtained from GenBank. The phylogenetic tree was reconstructed based on the deduced amino acid sequences of their predicted proteins using MrBayes 3.2. BADH2 of vascular plants was separated into two groups containing dicot and monocot species. Coconut BADH2 showed a close relationship with other monocots, and the closest relationship was found with the oil palm *E. guineensis* as a monophyletic branch (Figure 4).

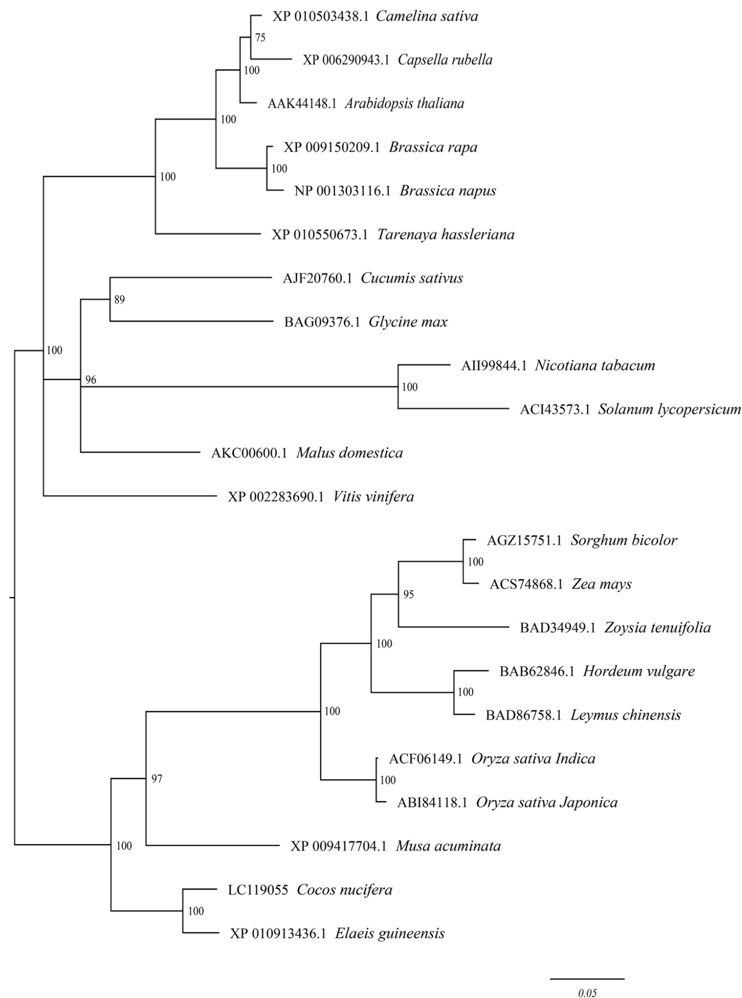

Figure 4. Bayesian phylogram constructed using MEGA 7.0 program clarifying the phylogenetic relationship of BADH2 proteins from coconut and other plants. The amino acid substitution model and parameters were calculated by Aminosan based on the Bayesian information criterion. The bar below the tree indicates the distance value, and numbers on each branch indicate bootstrap values.

Cloning of *Badh2* from genomic DNA

One-base pair difference was identified between the two alleles from aromatic and non-aromatic coconuts, and their genomic DNA sequence was subsequently determined. Five primer pairs successfully amplified the DNA of *Badh2*. In both coconut types, primers CnBadh2-G1 could amplify a 1056-bp fragment ranging from exon 1 to part of exon 4. Primers CnBadh2-G2 could amplify a 664-bp fragment ranging from part of exon 5 to part of exon 8. Primers CnBadh2-G3 could amplify a 936-bp fragment ranging from part of exon 7 to part of exon 9. Primers CnBadh2-G4 could amplify a 3013-bp fragment ranging from part of exon 9 to exon 15. After a long PCR amplification, a product of about 10,000 bp from exon 4, intron 4, and exon 5 was obtained from both coconut types using the CnBadh2-G5 primer. The end sequences of the product were confirmed to be part of exon 4 and 5. However, the full sequence of intron 4 was not determined, as this was not necessary. The obtained sequences could be assembled into two contigs of 1150 and 4059 bp in the aromatic coconut, and 1151 and 4059 bp in the non-aromatic coconut. The sequences were submitted to DDBJ under the accession numbers LC125458 and LC125459 for aromatic coconut, and LC125460 and LC125461 for non-aromatic coconut, respectively. The first fragment consisted of exons 1, 2, 3, 4, introns 1, 2, 3, and the 5'-part of intron 4, and the second fragment ranged from the 3'-part of intron 4 to exon 15. The structure of *Badh2* comprised 15 exons and 14 introns. The size of each region is shown in **Table S1**. The size and sequence of the exons and introns were highly similar in both coconut types. A few base substitutions and indels were identified in introns 2, 10, and 12, and a base substitution was found in exon 14 (Figure 5).

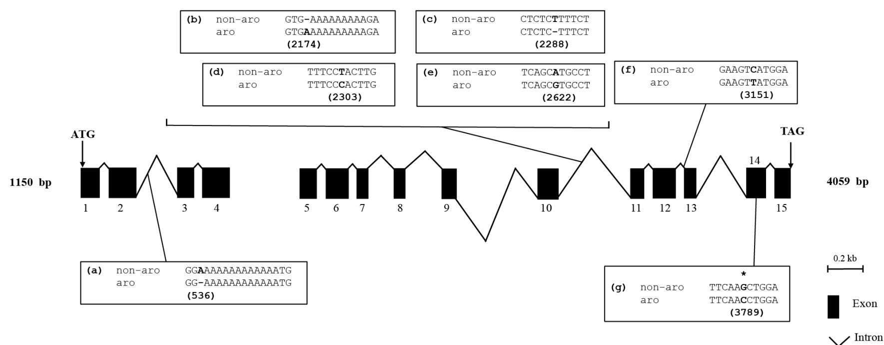

Figure 5. Structure of *CnBadh2* showing various base substitutions among aromatic and non-aromatic coconuts. Start codon (ATG), 15 exons (solid boxes), 14 introns (lines) and stop codon (TAG) are illustrated. Single nucleotide polymorphisms (SNPs) and indels are shown in bold. **a.-f.** Site of variable bases in introns. **g.** The SNP in exon 14 causing an amino acid change in aromatic coconut is indicated by an asterisk.

Expression of *Badh2* in coconut

Gene expression in the two coconut types was determined using semi-quantitative RT-PCR. Two samples of cDNA each from the root, leaf, and fruit pericarp of aromatic coconuts, and two leaf samples from non-aromatic coconuts were amplified for 26, 28, 30, and 32 cycles using CnBadh2 primers for *Badh2*. Duplex PCR was performed with the addition of the Actin primer (Juntawong et al., 2014) to amplify the actin gene in the same reaction. The

reaction was performed for 28 cycles and no differences were found among all samples. The intensity of bands at about 480 bp for the actin gene and 222 bp for *Badh2* were consistent between duplicated tissue samples from aromatic coconut, and between leaf samples of the two coconut types (Figure 6). No differences in the level of *Badh2* mRNA expression were observed. However, differences in function may be associated with altered protein structure between the two coconut types.

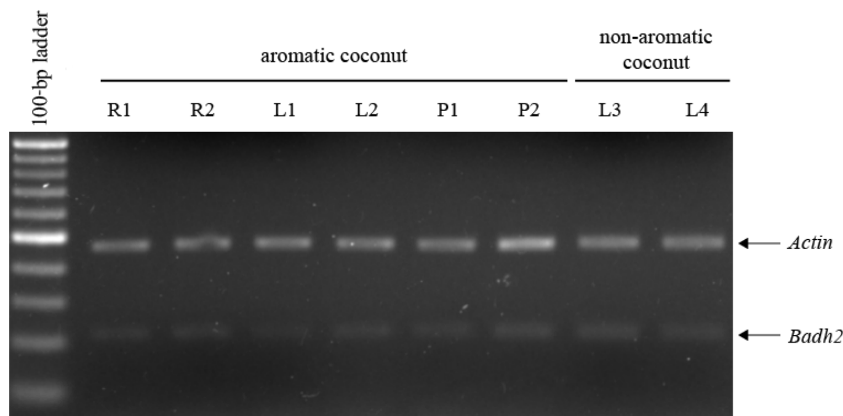

Figure 6. DNA band generated by semi-quantitative RT-PCR of *CnBadh2* from roots (R1 and R2), leaves (L1 and L2), fruit pericarps (P1 and P2) of aromatic coconut, and leaves of non-aromatic coconut (L3 and L4) compared to the actin gene.

Development of DNA markers to differentiate aromatic and non-aromatic coconuts

Two pairs of primers were developed to specifically amplify the DNA of aromatic or non-aromatic coconuts. The CnBadh2-m1 primers could amplify the 290-bp band from aromatic coconut and CnBadh2-m2 primers could amplify the 553-bp band in non-aromatic coconut, respectively (Figure 7).

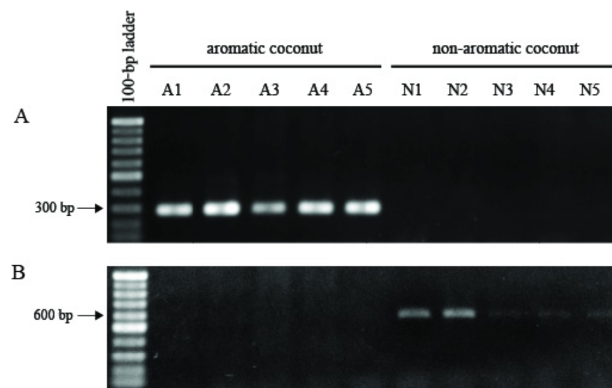

Figure 7. A. A 290-bp DNA band amplified using CnBadh2-m1 primer in aromatic coconut (A1-A5) and **B.** a 553-bp DNA band amplified using CnBadh2-m2 primer in non-aromatic coconut (N1-N5).

DISCUSSION

In many plant species, aromatic characteristics are derived from mutation of *Badh2*. The deletion of 8 bp and three SNPs in exon 7, insertion of 7 bp in exon 8, or deletion of 7 bp in exon 2 of rice *Badh2* (Bradbury et al., 2005; Amarawathi et al., 2008; Shi et al., 2008), deletion of 1444 bp of the sorghum *Badh2* (Yundaeng et al., 2013), or an SNP in the soybean *Badh2* (Juwattanasomran et al., 2011) result in the fragrant phenotype. These mutations produce a truncated or non-functional protein, which leads to the accumulation of the fragrant compound 2AP. In aromatic coconut, a base substitution was identified, which resulted in an amino acid change from alanine to proline at position 442 near the C-terminal of the protein. This might result in the fragrant phenotype in coconut, because this amino acid change was located within the substrate binding domain of the enzyme (Chen et al., 2008; Wongpanya et al., 2011). Proline contains a side chain with a ring structure that may affect the overall structure of this domain of the enzyme (Figure 3), resulting in instability and a non-functional enzyme. A single base substitution causing one amino acid change in the highly conserved motif also leads to the development of fragrance in soybean (Juwattanasomran et al., 2011) and cucumber (Yundaeng et al., 2015). In contrast, no mutation was found in *Badh2* of some fragrant rice varieties, suggesting that the mutation might occur within the promoter region, or the intron of the gene (Shao et al., 2013).

The genomic structure of the coconut *Badh2* consists of 15 exons and 14 introns, similar to the structure in rice (Bradbury et al., 2005). However, one intron (intron 4) is very long, at about 10,000 bp, which is longer than *Badh2* of other plants (Bradbury et al., 2005; Juwattanasomran et al., 2011). This suggests that an insertion or duplication event occurred in this intron, which does not affect the transcription of the gene. Furthermore, mRNA transcription was not affected by base substitution and indels in other introns, because no differences were observed in the size and amount of mRNA (cDNA) in aromatic and non-aromatic coconuts.

The phylogenetic tree reconstructed from the deduced amino acid sequence consisted of two groups containing monocot and dicot species (Figure 4). BADH2 proteins of coconut were grouped in the same cluster as those from other monocots; however, they were more closely related to oil palm (*E. guineensis*) BADH2 than to other cereals. When BADH1 and BADH2 were analyzed, the two isoforms of BADH in dicot species were clustered in the same branch, whereas those of monocots were separated into two subgroups of BADH1 and BADH2 (data not shown), indicating that duplication of the ancestral genes occurred earlier in the monocot species. The mutation of genes associated with the development of the fragrant phenotype then occurred randomly at distinctive sites in each plant (Bradbury et al., 2005; Juwattanasomran et al., 2011; Yundaeng et al., 2015).

No differences in the mRNA expression of *Badh2* were observed between the two coconut types, and no differences were found in the tissues of aromatic coconut. However, the BADH2 protein from aromatic coconut contained proline at the substrate binding domain of the enzyme instead of alanine, which might be responsible for the break-down of structure within this region (Figure 3). This base substitution was used to develop a type-specific DNA marker that could successfully differentiate aromatic from non-aromatic coconuts.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the Kasetsart University Research and Development Institute (KURDI) and the Graduate Study Research Scholarship for International Publications, Graduate School, Kasetsart University, Thailand. We also thank Chumphon Horticultural Research Center, Chumphon Province, Thailand, for providing the coconut samples.

REFERENCES

Agrawal GK, Pandey RN and Agrawal VP (1992). Isolation of DNA from *Choerospondias axillaris* leaves. *Biotech. Biodiv. Lett.* 2: 19-24.

Amarawathi Y, Singh R, Singh AK, Singh VP, et al. (2008). Mapping of quantitative trait loci for basmati quality traits in rice (*Oryza sativa* L.). *Mol. Breed.* 21: 49-65. http://dx.doi.org/10.1007/s11032-007-9108-8

Arikit S, Yoshihashi T, Wanchana S, Uyen TT, et al. (2011). 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: 75-87. http://dx.doi.org/10.1111/j.1467-7652.2010.00533.x

Biasini M, Bienert S, Waterhouse A, Arnold K, et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42: W252-8. http://dx.doi.org/10.1093/nar/gku340

Bradbury LMT, Fitzgerald TL, Henry RJ, Jin Q, et al. (2005). The gene for fragrance in rice. *Plant Biotechnol. J.* 3: 363-370. http://dx.doi.org/10.1111/j.1467-7652.2005.00131.x

Bradbury LMT, Gillies SA, Brushett DJ, Waters DLE, et al. (2008). Inactivation of an aminoaldehyde dehydrogenase is responsible for fragrance in rice. *Plant Mol. Biol.* 68: 439-449. http://dx.doi.org/10.1007/s11103-008-9381-x

Buttery RG, Ling LC and Juliano BO (1982). 2-Acetyl-1-pyrroline: An important aroma component of cooked rice. *Chem. Ind.* 958-959.

Buttery RG, Juliano BO and Ling LC (1983). Identification of rice aroma compound 2-acetyl-1-pyrroline in pandan leaves. *Chem. Ind.* 478-479.

Chen S, Yang Y, Shi W, Ji Q, et al. (2008). *Badh2*, encoding betaine aldehyde dehydrogenase, inhibits the biosynthesis of 2-acetyl-1-pyrroline, a major component in rice fragrance. *Plant Cell* 20: 1850-1861. http://dx.doi.org/10.1105/tpc.108.058917

Grosch W and Schieberle P (1997). Flavor of cereal products-a review. *Cereal Chem.* 74: 91-97. http://dx.doi.org/10.1094/CCHEM.1997.74.2.91

Juntawong P, Sirikhachornkit A, Pimjan R, Sonthirod C, et al. (2014). Elucidation of the molecular responses to waterlogging in Jatropha roots by transcriptome profiling. *Front. Plant Sci.* 5: 658. http://dx.doi.org/10.3389/fpls.2014.00658

Juwattanasomran R, Somta P, Chankaew S, Shimizu T, et al. (2011). A SNP in *GmBADH2* gene associates with fragrance in vegetable soybean variety "Kaori" and SNAP marker development for the fragrance. *Theor. Appl. Genet.* 122: 533-541. http://dx.doi.org/10.1007/s00122-010-1467-6

Kumar S, Stecher G and Tamura K (2016). Mega7: Molecular evolutionary genetics analysis version 7.0. *Mol. Biol. Evol.* 33: 1870-1874. http://dx.doi.org/10.1093/molbev/msw054

Laskowski RA, MacArthur MW, Moss DS and Thornton JM (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26: 283-291. http://dx.doi.org/10.1107/S0021889892009944

Liu YG and Whittier RF (1995). Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. *Genomics* 25: 674-681. http://dx.doi.org/10.1016/0888-7543(95)80010-J

McWilliam H, Li W, Uludag M, Squizzato S, et al. (2013). Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* 41: W597-600. http://dx.doi.org/10.1093/nar/gkt376

Nakamura T, Nomura M, Mori H, Jagendorf AT, et al. (2001). An isozyme of betaine aldehyde dehydrogenase in barley. *Plant Cell Physiol.* 42: 1088-1092. http://dx.doi.org/10.1093/pcp/pce136

Ronquist F, Teslenko M, van der Mark P, Ayres DL, et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61: 539-542. http://dx.doi.org/10.1093/sysbio/sys029

Shao G, Tang S, Chen M, Wei X, et al. (2013). Haplotype variation at *Badh2*, the gene determining fragrance in rice. *Genomics* 101: 157-162. http://dx.doi.org/10.1016/j.ygeno.2012.11.010

Shi W, Yang Y, Chen S and Xu M (2008). Discovery of a new fragrance allele and the development of functional markers for the breeding of fragrant rice varieties. *Mol. Breed.* 22: 185-192. http://dx.doi.org/10.1007/s11032-008-9165-7

Stiekema WJ, Heidekamp F, Dirkse WG, van Beckum J, et al. (1988). Molecular cloning and analysis of four potato tuber mRNAs. *Plant Mol. Biol.* 11: 255-269. http://dx.doi.org/10.1007/BF00027383

Tanabe AS (2011). Kakusan4 and Aminosan: two programs for comparing nonpartitioned, proportional and separate models for combined molecular phylogenetic analyses of multilocus sequence data. *Mol. Ecol. Resour.* 11: 914-921. http://dx.doi.org/10.1111/j.1755-0998.2011.03021.x

Tangchatuporn S (2000). Aromatic coconut. Chonburi College of Agriculture and Technology, Department of Vocational Education, MOE, Bangkok.

Thanananta T, Pongtongkam P, Thongpan L, Kaveeta A, et al. (2006). Effect of short day photoperiod on DNA methylation and expression of a gene in rice KDML105. *Afr. J. Biotechnol.* 5: 1375-1382.

Vanavichit A, Yoshihashi T, Wanchana S, Areekit S, et al. (2005). Cloning of *Os2AP*, the aromatic gene controlling the biosynthetic switch of 2-acetyl-1-pyrroline and gamma aminobutyric acid (GABA) in rice. Proceedings of the 5th International Rice Genetics Symposium. International Rice Research Institute, Philippines, 44.

Wongpanya R, Boonyalai N, Thammachuchourat N, Horata N, et al. (2011). Biochemical and enzymatic study of rice BADH wild-type and mutants: an insight into fragrance in rice. *Protein J.* 30: 529-538. http://dx.doi.org/10.1007/s10930-011-9358-5

Wongpornchai S, Sriseadka T and Choonvisase S (2003). Identification and quantitation of the rice aroma compound, 2-acetyl-1-pyrroline, in bread flowers (*Vallaris glabra* Ktze). *J. Agric. Food Chem.* 51: 457-462. http://dx.doi.org/10.1021/jf025856x

Yoshihashi T, Huong NTT and Inatomi H (2002). Precursors of 2-acetyl-1-pyrroline, a potent flavor compound of an aromatic rice variety. *J. Agric. Food Chem.* 50: 2001-2004. http://dx.doi.org/10.1021/jf011268s

Yoshihashi T, Nguyen TTH and Kabaki N (2004). Area dependency of 2-acetyl-1-pyrroline content in an aromatic rice variety, Khao Dawk Mali 105. *Jpn. Agric. Res. Q.* 38: 105-109. http://dx.doi.org/10.6090/jarq.38.105

Yu H and Goh CJ (2000). Identification and characterization of three orchid MADS-box genes of the AP1/AGL9 subfamily during floral transition. *Plant Physiol.* 123: 1325-1336. http://dx.doi.org/10.1104/pp.123.4.1325

Yundaeng C, Somta P, Tangphatsornruang S, Wongpornchai S, et al. (2013). Gene discovery and functional marker development for fragrance in sorghum (*Sorghum bicolor* (L.) Moench). *Theor. Appl. Genet.* 126: 2897-2906. http://dx.doi.org/10.1007/s00122-013-2180-z

Yundaeng C, Somta P, Tangphatsornruang S, Chankaew S, et al. (2015). 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: 1881-1892. http://dx.doi.org/10.1007/s00122-015-2554-5

Supplementary material

Figure S1. Alignment of the amino acid sequence of betaine aldehyde dehydrogenase 2 (BADH2) protein encoded by *CnBadh2* from aromatic (**A**) and non-aromatic (**B** and **C**) coconuts. The positions of amino acid differences are shown in bold and underlined.

Table S1. Regions of Badh2 gene components.