Molecular cloning of two genes encoding cinnamate 4-hydroxylase (C4H) from oilseed rape (*Brassica napus*)

CHEN Anhe, CHAI Yourong, LI Jiana, CHEN Li

College of Agronomy and Biotechnology, Southwest University, 216 Tiansheng Road, Beibei, Chongqing, 400716, P R China. Emails: chaiyourong@tom.com ; ljn1950@swu.edu.cn

Abstract

Cinnamate 4-hydroxylase (C4H) is a key enzyme of phenylpropanoid pathway, which synthesizes numerous secondary metabolites participating in plant development and stress adaption. Two *C4H* isoforms, the 2192-bp *BnC4H-1* and 2108-bp *BnC4H-2*, were cloned from oilseed rape (*Brassica napus*). They both have two introns and a 1518-bp open reading frame encoding a 505-amino-acid polypeptide. BnC4H-1 is 57.73 kDa with an isoelectric point of 9.11, while 57.75 kDa and 9.13 for BnC4H-2. They share only 80.6% identities on nucleotide level but 96.6% identities and 98.4% positives on protein level. Showing high homologies to *Arabidopsis thaliana* C4H, they possess a conserved p450 domain and all P450-featured motifs, and are identical to typical C4Hs at substrate-recognition sites and active site residues. They are most probably associated with endoplasmic reticulum by one or both of the N- and C-terminal transmembrane helices. Phosphorylation may be a necessary post-translational modification. Their secondary structures are dominated by alpha helices and random coils. Most helices locate in the central region, while extended strands mainly distribute before and after this region. Southern blot indicated about 9 or more *C4H* paralogs in *B. napus*. In hypocotyl, cotyledon, stem, flower, bud, young- and middle-stage seed, they are co-dominantly expressed. In root and old seed, *BnC4H-2* is dominant over *BnC4H-1*, with a reverse trend in leaf and pericarp. Paralogous *C4H* numbers in Brassicaceae genomes and possible roles of conserved motifs in 5′ UTR and the 2nd intron are discussed.

Key words: cinnamate 4-hydroxylase (C4H); cloning; expression; oilseed rape (*Brassica napus*)

Introduction

Phenylpropanoid pathway produces a large number of biologically important secondary metabolites through ligin, flovonoid and other branch pathways. Lignins play fundamental roles in mechanical support, solute conductance and disease resistance. Flavonoids attract pollinators and protect plants from UV irradiation, however it is in favor of intrusion s by fungi and animals. Other phenylpropanoids such as salicylic acid act as signaling molecules. Manipulation of phenylpropanoid pathway metabolites has long been a hotspot worldwide (Dixon et al., 1996).

Cinnamate 4-hydroxylase (C4H, EC 1.14.13.11) catalyzes the hydroxylation of trans-cinnamic acid to form 4-hydroxycinnamate and is the second key enzyme of common phenylpropanoid pathway (Russell, 1971). It belongs to CYP73A subfamily of cytochrome P450-dependent monooxygenase superfamily, and plays a pivotal role at the interface between cytosolic phenylpropanoid pathway and membrane-localized electron-transfer reactions (Chapple, 1998). 54 *C4H* genes have been isolated from plants, but, except *Arabidopsis thaliana C4H*, no full-length *C4H* gene has been cloned from the agronomically important family Brassicaceae.

In oilseed rape (*Brassica napus* L.), many efforts were focused on the genetic improvement for phenylpropanoid-related traits .. Improvement of disease resistance needs quicker and enhanced cell wall lignification in response to pathogen invasion. Genetic engineering of lignin pathway flux, monolignol ratio and lignin composition provide a promising strategy to cope with these problems (Anterola and Lewis, 2002). Lacking of yellow-seeded genotypes together with instability of yellow seed phenotype have largely retarded breeding procedures. (Heneen and Brismar, 2001). The most typical feature of yellow seed trait is the reduction of lignin and flavonoid pigments in the seed coat. Study on *B. napus C4H* gene will help to dissect the mechanism of yellow seed trait formation and lay the base for transgenic creation of stable yellow-seeded *B. napus*. Here we report the cloning and molecular characterization of two isoforms, *BnC4H-1* and *BnC4H-2*, of *C4H* gene family from *B. napus*.

Materials and Methods

Plant materials and nucleic acid isolation The black-seeded line 5B of *B. napus* was used in this study. Samples were collected from the root hypocotyl, cotyledon, stem, leaf, bud, flower, silique pericarp, and seeds of 10, 20 and 30 d after flowering (DAF) to isolate the total RNA using a CTAB method .RNA samples were digested with RNase-free DNase I. Total genomic DNA was isolated from fresh leaves using a CTAB-based method.

3′ and 5′ cDNA end amplification 5-μg mixture of total RNA from various organs was used to generate first strand total cDNA using GeneRacer Kit (Invitrogen, USA). Forward primers FC4H3-1 (5′-TGATGATGTACAACAACATGTTCCG-3′) and FC4H3-2 (5′-CCTCACATGAACCTCCATGATGC-3′) were synthesized. FC4H3-1 was paired with GeneRacer 3′-Primer to carry out the primary amplification of 3′ RACE in a standard 50-μl *Taq* PCR system containing 0.5 μl total cDNA as template annealed at 50°C. One μl of 50-fold diluted PCR product was used as template for 3′-nested PCR using primer

FC4H3-2 and GeneRacer 3′-Nested Primer with an anneal temperature of 55°C. In 5′ RACE, antisense primer RC4H5-1 (5′-GCATCATGGAGGTTCATGTGAGG-3′) was paired with GeneRacer 5′-Primer, while RC4H5-2 (5′-CGGAACATGTTGTTGTACATCATCA-3′) with GeneRacer 5′-Nested Primer, for primary and nested PCRs respectively with parameters the same as corresponding 3′ RACE primary PCRs. PCR products were subcloned into pMD18-T and sequenced using primers M13F/M13R.

Full-length cDNA and genomic sequence amplification Sense primers FBNC4-2 (5′-AGCAGCTCCTTCTGCT TTCTC-3′) and FBNC4-3 (5′-TCAGCAGCTCCTTCTGCTTTC-3′), and antisense primers RBNC4-1 (5′-CAAAACAGTGGGAACCAATAGTTATTG-3′) and RBNC4-7 (5′-CCGAAGAAACAACACATTGAATA TTCAAC-3′), were designed and combined into 4 primer pairs. 0.5-μl total first strand cDNA and 0.5 μg total genomic DNA were used for amplification of full-length cDNAs and genomic sequences respectively, annealed at 62°C. Subcloning and sequencing were performed as described above.

Southern blotting 45-μg total genomic DNA was fully digested with *Dra*I, *Eco*RI, *Eco*RV and *Hin*dIII respectively, separated, and transferred to nylon membrane. Primer pair FBNC4-2/RC4H5-2 was used to amplify a 656-bp conserved BnC4H-1 cDNA fragment. Digoxigenin-11-dUTP labeling, hybridization at 42°C for 16 h, washing, and immunological detection were performed (DIG kits, Roche, Germany).

RT-PCR detection Oligo(dT)$_{20}$-directed reverse transcription of 5-μg total RNA of each sample was performed using SuperScript III First-Strand Synthesis SuperMix (Invitrogen, USA). Primers FBNC4-2 and RBNC4-2N (5′-TTTGGTGAGGTTCGGGGAG-3′) were used to isoform-specifically amplify a 357-bp region of *BnC4H-1*, while FBNC4-1 and RBNC4-1N (5′-CCTTTCGTGGCCGAATCAAG-3′) for specific amplification of a 516-bp region of *BnC4H-2*. 0.5-μg first strand cDNA of each sample was taken as template in a 50-μl standard *Taq* PCR with 30 cycles annealed at 62°C. A 542-bp *ACT2* gene fragment was used as internal control annealed at 55°C.

Results

Sequence cloning of BnC4H-1 *and* BnC4H-2 Nested PCR of the 3′ RACE resulted in a band of about 600 bp. Sequenced clones showed two different 3′ cDNA ends, 496 bp and 524 bp in net length. 5′ RACE nested PCR yielded a specific band of about 700 bp. Sequenced clones resulted in 2 different 5′ cDNA ends, 656 bp and 658 bp. The 4 primer pairs all yielded specific bands of about 1750 bp in full-length cDNA amplifications, but the bands of FBNC4-2/FBNC4-1 and FBNC4-3/FBNC4-1 were a little longer than those of FBNC4-2/RBNC4-7 and FBNC4-3/FBNC4-7. Sequencing results of the 2 longer bands were practically identical to each other and the same case with the 2 shorter bands. This indicated that only 2 full-length cDNAs, denoted *BnC4H-1* and *BnC4H-2* here, were obtained. Alignment indicated that the right primer pairs for *BnC4H-1* and *BnC4H-2* were FBNC4-2/RBNC4-1 and FBNC4-3/RBNC4-7 respectively. So they were used to amplify genomic sequences of *BnC4H-1* and *BnC4H-2* and showed bands of about 2200 bp and 2100 bp respectively. Comparison between these two genomic sequences found no differentiations in all the coding regions.

Molecular characterization of nucleotide sequences of BnC4H-1 *and* BnC4H-2 *Basic parameters* The genomic sequence and full-length cDNA of *BnC4H-1* are 2192 bp and 1742 bp respectively. At 879-949 bp and 1084-1462 bp, 2 introns are contained. Its cDNA has a 93-bp 5′ UTR and a 131-bp 3′ UTR, between which is a 1518-bp ORF. The genomic sequence and full-length cDNA of *BnC4H-2* were 2108 bp and 1716 bp respectively. The 2 introns are found between 881-945 bp and 1080-1406 bp. The 5′ UTR, ORF and 3′ UTR of *BnC4H-2* are 95 bp, 1518 bp and 103 bp, respectively. The G+C contents of their ORFs are 50.59% and 49.28%, while their non-coding regions have G+C contents of less than 40%.

Homologies and parental-species origin NCBI blastn indicated that the coding regions of *BnC4H-1* and *BnC4H-2* showed high identities to known *Brassica C4H* tags and *AtC4H* (U71080 and NM_128601). They also showed moderate identities to many non-cruciferous *C4H*/*CYP73A* genes. The 466-bp *C4H-BO-1* from *B. oleracea* (AF230674) demostrated only 1 bp of difference to *BnC4H-1*, while all other fragments including those from *B. napus* show identities of less than 94%. These suggest that *BnC4H-1* is a novel *B. napus C4H* gene, which has no tag in the Genbank database and is undoubtedly transmitted from the parental species *B. oleracea*. On the other hand, the 314-bp *C4H-BN-7* from *B. napus* (AF230673) was practically identical to *BnC4H-2*.

BnC4H-1 shows 80.6% and 87.4% identities to *BnC4H-2* on genomic and cDNA levels respectively. Their 5′ UTRs are completely identical to each other and their coding regions are of high uniform (90.6%), while their introns and 3′ UTRs are of quite low identities (59.7% to 47.5%). *BnC4H-1* shows 75.9% and 83.1% identities to *AtC4H* on genomic and cDNA levels respectively. The identities of their ORF, 5′ UTR, 3′ UTR, intron 1 and intron 2 are 87.1%, 59.1%, 58.4%, 63.5% and 45.1% respectively. *BnC4H-2* shows 77.3% and 83.1% identities to *AtC4H* on genomic and cDNA levels, respectively. The identities of their ORF, 5′ UTR, 3′ UTR, intron 1 and intron 2 are 87.0%, 60.0%, 53.3%, 56.5% and 50.8% respectively.

Possible cis-elements The 3′ UTR of *BnC4H-2* contains a canonical polyadenylation signal $A_{2017}ATAAA_{2022}$, but none was detected in *BnC4H-1*. According to the new definition of CpG island (Takai and Jones, 2002), a 532-bp CpG island was predicted in *BnC4H-1* at A_{83}-A_{620}. No CpG island was found in *BnC4H-2*. The 5′ UTRs of *BnC4H-1* and *BnC4H-2* are completely identical to each other in the corresponding 93-bp region. Though these two 5′ UTRs are of low similarities to the 5′ UTR of *AtC4H*, a 19-bp highly conserved region AGCAGCTCCTTCTGCTTTC was identified at the beginning of the 5′ UTR of all the 3 genes. Most regions of the 2 introns of the 3 genes are not conserved, but a highly conserved region was detected at the beginning of the 2nd intron, especially a 16-bp sequence CTTGTAGGATACGTAA corresponding to

1112-1127 bp of *BnC4H-1*. The 1100-1115 bp of *BnC4H-2* is completely identical in the 3 genes. The sequences have been submitted to GenBank under accession numbers from DQ485129to DQ485132.

Conservation and structural features of the deduced BnC4H-1 and BnC4H-2 proteins

The ORFs of *BnC4H-1* and *BnC4H-2* both encode a polypeptide of 505 aa. BnC4H-1 possesses a Mw of 57.73 kDa and a pI of 9.11, and 57.75 kDa and 9.13 for BnC4H-2.

Homologies BnC4H-1 show 96.6% identities and 98.4% positives to BnC4H-2. SUPERFAMILY alignment revealed that BnC4H-1 and BnC4H-2 both belong to the cytochrome P450 family. NCBI blastp indicated that BnC4H-1 and BnC4H-2 show very wide similarities to C4Hs from other plants. On whole molecule scale, BnC4H-1 and BnC4H-2 shows identities/positives of 95.8%/98.0% and 95.4%/97.8% to AtC4H (AAB58355) respectively. They also show similarities to non-C4H P450s such as F3′H, F5H, F3′5′H, C3′H, etc.

Conserved domains/motifs and active site residues NCBI Conserved Domain search detected two conserved cytochrome P450 domains dominating most part of BnC4H-1 and BnC4H-2: pfam00067 (p450) and COG2124 (CypX). They have nearly overlapping locations: pfam00067 at F_{43}-G_{48} and COG2124 at Q_{48}-G_{475} in BnC4H-1, and pfam00067 at F_{43}-V_{499} and COG2124 at Q_{48}-G_{475} in BnC4H-2. BnC4H-1 and BnC4H-2 have all the P450-featured motifs, such as the haem-iron binding domain P_{439}FGVGRRSCPG$_{449}$, the T-containing binding pocket motif A_{306}AIETT$_{311}$, the E_{363}-R_{366}-R_{420} triad, and the hinge motif P_{35}PGP(M/I)PIP$_{41}$ (Chapple, 1998). Residues for enzymatic active sites of C4H might involve I_{109}, K_{113}, V_{118}, F_{220}, E_{301}, N_{302}, I_{303}, V_{305}, A_{306}, T_{310}, R_{366}, R_{368}, A_{370}, I_{371}, P_{372}, L_{374}, V_{375}, P_{376}, H_{377}, K_{484}, F_{488}, and L_{490} etc. They distribute in five substrate recognition sites (SRS) signature motifs of C4H/CYP73A5: SRS1 (S_{100}RTRNVVFDIFTGKGQDMVFTVY$_{122}$), SRS2 (L_{216}AQSFEYNY$_{224}$), SRS4 (I_{299}VENINVAAIETTLWS$_{314}$), SRS5 (R_{368}MAIPLLVPH$_{377}$) and SRS6 (K_{484}GGQFSLHI$_{492}$), respectively (Schoch et al., 2003). BnC4H-1 and BnC4H-2 have all the five SRS motifs with 100% identities. These results indicated that BnC4H-1 and BnC4H-2 are orthologous proteins of AtC4H (CYP73A5) and are most probably catalytically functional.

Possible post-translational modifications NetPhos 2.0 predicted 19 potential phosphorylation sites in BnC4H-1 and 21 in BnC4H-2. Phosphorylation may be a prerequisite for normal functioning of them. NetNGlyc 1.0 and PROSITE predicted BnC4H-1 and BnC4H-2 to have a potential N-glycosylation site at position 85 (**N**LTK).

Signal peptide/anchor and subcellular localization SignalP 3.0 predicted that BnC4H-1 has a probability of 0.408 to have a signal peptide and a probability of 0.584 to have a signal anchor, whereas 0.438 and 0.553 for BnC4H-2. Predotar, Softberry-ProtComp 6.0 and WoLFPSORT definitely predicted BnC4H-1 and BnC4H-2 to be ER-membrane bound, like other C4Hs. Both TMpred and SOSUI predicted 2 strong transmembrane helices at both terminal regions of BnC4H-1 and BnC4H-2, and the positions and sequences are completely identical between the 2 proteins.

Secondary and tertiary structures Predicted by SOPMA, the secondary structures of BnC4H-1 and BnC4H-2 are mainly composed of alpha helices (48.71% for both) and random coils (33.47% and 34.65%), while extended strands (12.67% for both) and beta turns (5.15% and 3.96%) also contribute. H_{12}-R_{366} of them is dominated by alpha helices connected by random coils, with a 62-aa huge alpha helix at L_{202}-Q_{263} in BnC4H-1. In BnC4H-2 this huge helix is cut into two major helices by some random coils, but other 2 helices (H_{125}-F_{138} and K_{141}-K_{163}) found in BnC4H-1 have merged into a large helix (H_{125}-K_{163}) in BnC4H-2. The N-terminal large helix of them covers the predicted signal peptide/anchor and the N-terminal transmembrane helix. Extended strands mainly disperse at two regions: one is the ~100-residue C-terminal region, and another is the ~130-residue region between the N-terminal helix and the central helices. SWISS-MODEL predicted tertiary structures of BnC4H-1 and BnC4H-2 are very similar to the reported P450 crystal structure of a globular protein (Rupasinghe et al., 2003). The haem is located in the center and is surrounded by several large alpha helices. The C4H-signature motifs SRS1, SRS2 and SRS4 distribute in the helices 4, 8 and 10, respectively. The haem-iron binding motif P_{439}FGVGRRSCPG$_{449}$ locates in the SRS6.

Southern blot detection of C4H *homologues in the genome of* B. napus In Southern blot, *Dra*I, *Eco*RI, *Eco*R*V* and *Hin*dIII digestions resulted in 5, 6, 8 and 9 hybridization bands respectively. A few bands are quite weak, but they can still be identified as specific hybridization bands. All the 4 enzymes have no cutting site in genome sequence of *BnC4H-1* and *BnC4H-2*, so it is suggested that the *B. napus* genome may contain about 9 or more *C4H* members and some members have the same digestion maps for *Dra*I and *Eco*RV, respectively.

Transcription levels of BnC4H-1 *and* BnC4H-2 *in various organs of* B. napus RT-PCR results indicated that *BnC4H-1* and *BnC4H-2* have similar expression patterns in view of organ specificity, but differences are still obvious. The transcription of *BnC4H-1* can be distinctly detected in all analyzed organs except in 30 DAF seed. Its expression in the hypocotyl, stem, cotyledon, leaf, bud, flower and silique pericarp shows no great difference, but the expression in the root and seed is distinctly lower. Expression of *BnC4H-2* can be detected in all the 11 organs analyzed. Its expression in the hypocotyl, stem, root, cotyledon, bud and flower is obviously higher than in the leaf, seed of all stages and silique pericarp, with the lowest still in the 30 DAF seed.

Discussion

How many C4H *genes in* Brassica *and Brassicaceae?* *A. thaliana* genome has experienced a shrinking process but the genus *Brassica* has highly replicated genomes. Basic 'diploid' *Brassica* species are likely derived from hexaploid ancestry, and it is expected that in *B. napus* there might exist about 6 genes orthologous to each gene from *A. thaliana* (Lysak et al., 2005). In a consensus genetic marker (ACGM) analysis, a pair of *AtC4H*-based conserved primers amplified 7, 4 and 3 *C4H* fragments from *B. napus*, *B. oleracea* and *B. rapa* respectively, and these were considered reliable to represent the total *C4H*

genes in the respective genomes (Fourmann et al., 2002). But our Southern blot result indicated that there might be as many as 9 or more *C4H* genes in *B. napus*. Overestimation of copy numbers may be caused by enzyme cutting within hybridized region in some unknown genes, but there is a strong factor leading to underestimation of copy numbers, i.e. identical digestion maps for highly homologous genes especially in the amphidiploid *B. napus* whose two parental species and *Brassica*-duplicated genes are quite near in evolution.

Two other important facts also favor the high copy number assumption. First, identities of *BnC4H-1* to *C4H-BO-1* indicated that a *B. napus* gene was basically unchanged from its donor gene from a parent species, a fact also proved by Fourmann et al. (2002). But in their research none of the 7 *BnC4H* tags was assigned to an explicit parental locus, and *vise versa* for the 7 parental *C4H* tags. It strongly suggests that, at least a part of, the 7 *BnC4H* tags have no receptor-donor relationship with the 7 parental-species *C4H* tags. This is to say that *C4H* gene numbers in *B. napus* should be more than 7 and also more than 4 and 3 in *B. oleracea* and *B. rapa* respectively. At least *BnC4H-1* is the 8th *C4H* gene in *B. napus* succeeding the 7 tags. Second, *C4H-BO-4* forms an almost triangle relationships with *AtC4H* and other known *C4H* tags/genes. On protein level, surprisingly, C4H-BO-4 is more divergent from other *Brassica* C4Hs than AtC4H. Perhaps after diverging with *Arabidopsis*, hexaploidization of *Brassica* ancestor resulted in most of the known *Brassica C4H* genes/tags, but *C4H-BO-4* might be resulted from another duplication event prior to the triplication event.

In the crucial common phenylpropanoid pathway, in sharp contrast to four *PAL* and four *4CL* genes, *A. thaliana* enigmatically contains only one *C4H* gene. Most plants contain a small family of *C4H* genes. Studies suggest that prior to the separation of monocots and dicots, or even earlier, the *C4H* gene has duplicated. Quite divergent Class 1 and Class 2 of *C4H* genes have been identified (Betz et al., 2001). The separation of a dicot *M. crystallinum* C4H from all other dicot, monocot, even gymnosperm C4Hs, suggests the duplication of the *C4H* gene prior to the divergence of gymnosperm and angiosperm species. Hence, C4H in ancestral species of dicot families may be encoded by more than one *C4H* genes. The evolution route of C4H may resemble those of PAL and 4CL. That is to say, C4H is encoded basically by multiple genes in higher plants, and the monogenic status of some plants is resulted from gene loss.

Possible important cis elements deserve further study As the non-coding regions are basically of low conservation, a 19-bp region in the right beginning of the 5′ UTR and a 16-bp sequence at the 5′ of the 2nd intron are highly conserved. The transcription initiation sites of the 2 *BnC4H* genes are quite near that of the *AtC4H* (Bell-Lelong et al., 1997), and conservation of certain proximal structures might be a determinative factor. Possible role of the 16-bp region may be involved in regulating transcription or transcript processing. Furthermore, *BnC4H-1* and *BnC4H-2* differ from each other in CpG island and polyadenylation signal. Whether these differences have any relation to their different tissue specificities deserves to be clarified.

Conservation of protein structure and differentiation of tissue specificity of C4H *genes expression* As it has been mentioned, similarities among *BnC4H-1*, *BnC4H-2* and *AtC4H* are significantly higher and higher as aligned from on genomic, cDNA, ORF to on protein levels. This implies that in Brassicaceae the *C4H* gene family is highly conserved on protein level both orthologously and paralogously. At conserved motifs and active site residues, BnC4H-1 is identical to BnC4H-2 and they show little difference with AtC4H.

However, expression patterns of *C4H* genes even within one species seem to be more differentiated. Though the wide-expression features of *BnC4H-1* and *BnC4H-2* resemble those of *AtC4H* (Bell-Lelong et al., 1997), obvious differences also exist. Co-domination, domination, and complementation all exist between *BnC4H-1* and *BnC4H-2* in tissue specificity. This functional divergence may be an evolution strategy to allocate the "redundant" family members especially in an amphidiploid species like *B. napus*. In certain tissues, isoforms of common phenylpropanoid pathway enzymes might be combined with certain branch pathway enzymes to form pathway-specific enzyme complex (Winkel-Shirley, 1999). The features of tissue specificity of the two isoforms observed here favor this assumption. Maybe *BnC4H-1* and *BnC4H-2* have completed functional differentiation in certain tissues. Other explanations for the differed tissue specificities still exist, such as mutations of cis-elements and preference of advantageous isoform(s) in certain organs.

References

Omitted.