Botanical Studies (2010) 51: 7-16.

Construction of a full-length enriched cDNA library and analysis of 3111 ESTs from roots of *Bupleurum chinense* DC.

Chun SUI, Jian-He WEI*, Shi-Lin CHEN, Huai-Qiong CHEN, L.M. DONG, and Cheng-Min YANG

Institute of Medicinal Plant Development (IMPLAD), Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100193, P. R. China

(Received June 24, 2008; Accepted June 18, 2009)

ABSTRACT. *Radix Bupleuri* (Chaihu), sourced from the dried roots of *Bupleurum* species, is a traditional Chinese medicine with anti-inflammatory, anti-pyretic, and anti-hepatotoxic efficacy. It is widely used in China, Japan, Korea, and other countries in south Asia. A full-length enriched cDNA library derived from the roots of *B. chinense* DC. was constructed for the first time by the SMART technique in this study to initiate the functional genomic research of this important medicinal plant. The titre of the library was 1.1×10^6. From the library, randomly selected 3902 clones were 5′ single-pass sequenced, among which 3111 high quality ESTs were generated and 1650 uniESTs were identified with 377 contigs and 1273 singleton ESTs. The estimated average cDNA insert size was 1.1 kb, and the fullness ratio was ca. 51.5%. BlastX analysis of all uniESTs resulted in 949 (57.5%) homology to previously identified genes, 680 (41.2%) matched to unknown, unnamed, or hypothetical protein genes, and 21 clones had no hit. Gene ontology (GO) annotation of uniESTs showed that approximately 1002, 957, and 861 were assigned molecular function, biological process, and cellular component GO terms, respectively. KEGG pathway analysis indicated that 307 uniESTs may be involved in 31 metabolic pathways, in which at least five uniESTs were contained, by comparing each with *Arabidopsis* metabolic pathways. Through SSR searching within 1650 uniESTs, 86 potentially useful SSR loci were identified in 82 uniESTs. The library and EST data provide a platform to study the molecular mechanisms of various physiological phenomena of *Bupleurum*. The set of SSR loci would potentially be useful molecular markers for the germplasm identification, genetic diversity analysis, and gene mapping of *Bupleurum*.

Keywords: *Bupleurum chinense* DC.; Full-length enriched cDNA library; ESTs; Switching mechanism at 5' end of RNA transcript (SMART); SSR.

INTRODUCTION

Radix Bupleuri (Chaihu), sourced from the dried roots of *Bupleurum* species (*Umbelliferae* family), is an important Traditional Chinese Medicine (TCM) which has anti-inflammatory, anti-pyretic, and anti-hepatotoxic properties (Yen et al., 2005; Pan, 2006). It had been used in ancient China for about 2, 000 years when it was recorded in "*Shen Nong's Materia Medica*," the first monographic work of its kind in China. According to the "*Pharmacopoeia of the People's Republic of China*," *B. chinense* DC. is one of the two official *Radix Bupleuri* source species (the other being *B. scorzonerifolium* Willd.) widely distributed in China as a wild and cultivated species. It has also been introduced to other countries. Crude drug, decoction pieces, and extracts of *Radix Bupleuri* are annually exported from mainland China to Japan, Korea, and Southeast Asian countries. However, *B. falcatum* L., restricted to Japan and Korea, and *B. kaoi* Liu, C. Y. Chao & Chuang, endemic to Taiwan, have also been pharmacologically studied and used as the source species for *Radix Bupleuri* in their native countries, respectively. Most research about *Bupleurum* has focused on its classification (Urgamal et al., 2007), authentication (Yang et al., 2007), cultivation and breeding (Wei et al., 2003), physiology (Aoyagi et al., 2001), identification of medicinal compounds (Liu et al., 2002), and pharmacology (Pan, 2006). Pharmacologically active components like saikosaponins, volatile oils, and polysaccharides have been found in *Radix Bupleuri*. Saikosaponins, a class of triterpenoid saponin, make up its major active component, and these are used as a quality control standard for medicinal *Bupleurum*. Little molecular information about the

*Corresponding author: E-mail: wjianh@263.net; jhwei@implad.ac.cn; Tel: +86-10-6281-8841; Fax: +86-10-6281-8841.

secondary metabolism of these species is available. A few laboratories have recently launched preliminary studies to research the biosynthetic pathways of saikosaponins (Kim et al., 2006; Chen et al., 2007).

A full-length enriched cDNA library, with higher level full-length cDNA sequences (usually 30-70% of the total sequenced clones) than that of EST or cDNA common library, has been constructed and annotated largely from model organisms like *Arabidopsis*, rice, maize, *Drosophila*, and mice to analyze the functions of genes (Jia et al., 2006). As to medicinal plants, EST or common cDNA libraries were constructed from plant species such as *Panax ginseng* (Jung et al., 2003; Kim et al., 2006) and *Crocus sativus* (D'Agostino et al., 2007). For the *Umbelliferae* family, the first cDNA library was constructed using carrot somatic embryos (Lin et al., 1996). Subsequently, seven cDNA or EST libraries have been generated according to dbEST (Vilaine et al., 2003; Kwon et al., 2004; Divol et al., 2005; Park and Park, 2006). Most of these cDNA libraries were subtractive or differentially displayed using materials with special treatment, and the numbers of sequenced ESTs or cDNAs from these cDNA libraries were small. For plants in the genus *Bupleurum*, only one subtractive EST library was constructed using the adventitious roots of *B. kaoi* (Chen et al., 2007).

In this study, we constructed a full-length enriched cDNA library of *B. chinense* root, sequenced and bioinformatically analyzed more than three thousand ESTs. The cDNA library and these sequence data will help to isolate some metabolic functional genes and also will increase useful bioinformatics references for research on other umbelliferous plants. We also identified some interesting potential SSR markers detected in our ESTs. This is the first report to analyze a full-length enriched cDNA library using medicinal roots of traditionally cultivated *Bupleurum*.

MATERIALS AND METHODS

Plant materials

Bupleurum chinense cv. Zhongchai No. 1, which is a mass-selected cultivar of *B. chinense*, was used to construct a full-length enriched cDNA library. The roots of one-year old plants were harvested when flowering, frozen immediately in liquid nitrogen, and then stored at -80°C until RNA was extracted.

RNA extraction and full-length enriched cDNA library construction

Total RNA was extracted using TRIzol reagent (GIBCO BRL) according to the manufacturer's guideline. From approximately 6 g of frozen tissue, 728 μg total RNA was obtained. About 8.84 μg mRNA was then isolated using an Oligotex mRNA Kit (QIAGEN). The full-length cDNA library was constructed from approximately 3 μg mRNA by the SMART technique (Wellenreuther et al., 2004). After first-strand cDNA synthesis, long distance PCR (LD-PCR) and proteinase K digestion, PCR products were digested with restriction enzyme *Sfi*I to generate directional cloning ends. The *Sfi*I-digested double-strand cDNA was then size fractioned by a CHROMA SPIN-400 Column, and six fractions of cDNA fragments (ranging from 500 bp to 4 kb) were pooled. Fractioned cDNA was cloned into *Sfi*I-digested reconstructive pBluescript II SK vector (sequences between *Eco*RI and *Not*I sites replaced by *Sfi*I A and *Sfi*I B adaptors for directional insertion) and transformed into *Escherichia coli* DH-5α competent cells.

Evaluation of full-length enriched cDNA library and sequence analysis of ESTs

One microlitre ligation products were transformed to evaluate the recombination rate and capacity of a full-length enriched cDNA library. Plasmid DNA from a total of 136 clones was extracted, *Sfi*I-digested, and analyzed on 1% agarose gel. Fragment size was estimated using AlphaImager 2200. Three thousand, nine hundred and two clones were randomly selected and 5' single-pass sequencing was performed using an ABI 3730 Sequencer. The Phred/Phrap/Consed software package was used to call bases, remove vector sequences and low quality readings, and assemble uniESTs. Inserts with lengths >=450 bp and N<5 were defined as valid sequences. All sequences were compared to genes in the non-redundant protein database (nr) and nucleic acid databases (nt) of NCBI using BlastX and BlastN searches. Sequences with identity>90% over 100 bp were clustered as single uniESTs. To evaluate the fullness ratio of the library, clones corresponding to the identified genes were aligned to determine whether they contained a 5' UTR and a putative ATG translation initiation codon. When a sequence contained a putative translation initiation codon, it was defined as a full-length cDNA. Assembled uniESTs were annotated via a GO (Gene Ontology) term using a GoPipe standalone package (Chen et al., 2005) and further compared with the *Arabidopsis* metabolism pathway defined by KEGG (http://www.kegg.com) after BlastX of uniESTs and *Arabidopsis* protein sequences with a threshold E-value 1E-10 and an overall identity of 50%. Additionally, uniESTs assembled in the library were compared with all nucleotide squences of *Bupleurum* obtained from GenBank via a BlastN algorithm.

SSR searching

All uniESTs obtained in sequence assembly were analyzed for SSR using SSRHunter 1.3 (Li and Wan, 2005). To ensure the accuracy, only SSR loci within a 600 bp range of uniEST sequences were counted and for later primer design, at least 50 bp nucleotides were present before and after SSR loci. If two or more SSRs were within one uniEST, and the distance between them was less than 50 bp, they were identified as one compound SSR locus. Dinucleotides that were equal or repeated more than 9 times and tri-nucleotides repeated over 5 times were selected.

RESULTS

Overall features of *B. chinense* root full-length enriched cDNA library

One microlitre of transformed ligation mixture yielded 1113 recombinant clones, and the titre of the full-length enriched cDNA library was about 1.1×10^6 clones. To evaluate the size and distribution of insert cDNA clones, a total of 136 clones were randomly selected and digested by *Sfi*I, allowing us to obtain average cDNA insert sizes and the cDNA length distribution profiles. Most insert lengths ranged from 0.5 to 2 kb. The estimated average insert size was 1.1 kb. The size distribution of insert DNA clones is shown in Figure 1.

A total of 3902 recombinant clones were randomly selected and sequenced from the 5' end. After eliminating low-quality sequences and contaminated clones, 3111 valid sequences in all were obtained. Clustering and assembly of these ESTs resulted in a total of 1650 uniESTs with 377 contigs and 1273 singleton ESTs. BlastX and BlastN analysis showed that of 1650 uniESTs, 949 (57.5%) were homologous to previously identified genes, 680 (41.2%) matched to unknown, unnamed, or hypothetical protein genes, and 21 clones had no hit. Clones corresponding to the identified genes were tested to determine whether they contained a 5' UTR and a putative ATG translation initiation codon. Among these sequences matched to a known gene, 489 (ca. 51.5%) clones were predicted to contain a putative ATG translation initiation codon, and among them 159 single sequences were deposited in the dbEST division of GenBank (Accession numbers: FG341847-FG342005). General features of the library and sequencing statistics are listed in Table 1.

Gene Ontology annotation and metabolism pathway analysis

All uniESTs were assigned the Gene Ontology (GO) terms using the sequence comparison results of BlastX. Of the uniESTs, 1002 (60.7%), 957 (58.0%), and 861 (52.2%) were assigned molecular function, biological process, and cellular component GO terms, respectively. One uniEST did not exclusively belong to a single GO term. Figure 2 illustrates the GO term distribution of all uniESTs. The first two largest categories in three GO terms were binding and catalytic activity, physiological process and cellular process, cell and intracellular. These results which relate to the storage function of the root were consistent with similar studies on *P. ginseng* root (Jung et al., 2003). Comparison with the KEGG *Arabidopsis* pathway showed that 307 uniESTs represented 31 metabolic pathways on the prerequisite that each pathway contained at least 5 uniESTs (Table 2). The full-length enriched cDNA library we constructed is thus useful in identifying functional genes of *B. chinense*. BlastN analysis of our uniESTs with *Bupleurum* nucleotide sequences obtained from GenBank indicated that only three were homologous and the hits were AM409304 (cinnamic acid 4-hydroxylase), AM409290 (short-chain dehydrogenase/reductase), and AM409292 (omega-6 fatty acid desaturase). Therefore, nearly all uniESTs that we sequenced may represent novel transcripts from *Bupleurum*.

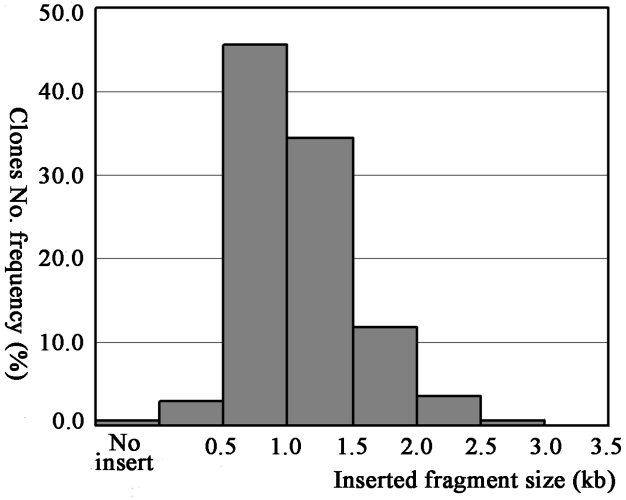

Figure 1. Size distribution of cDNAs in the library. Fragment sizes were determined by *sfi*I digestion of 136 random selected clones.

Table 1. General features of root of *B. chinense* full-length enriched cDNA library and sequencing statistics.

Titre (pfu)	1.1×10^6
Average cDNA insert size	1.1 kb
Total clones sequenced	3902
Sequences passed qualiy check	3111 (79.7%)
Clustered (contigs)	377
Unassembled (singletons)	1273
Unigenes (contigs+ singletons)	1650
Observed redundancy[a]	88.5%
No hit to nr (BLASTX)	22
EST matches with E-value in BLASTX $\geq 1\times10^{-14}$	366 (11.8%)
EST matches with E-value in BLASTX $< 1\times10^{-14}$	2724 (88.2%)

[a]Oberved redundancy: (EST# after quality check-Unigene #)/ Unigene # (Lindqvist et al., 2006).

To visualize the transcript abundance of *Bupleurum* root, the contigs assembled from the library were analyzed. Of these, the sixteen most abundant transcripts (EST number equal to or larger than 18) observed in this library are listed in Table 3. The contig with the largest number of ESTs was identified to be dehydrin protein. Contigs with the second and third largest ESTs were a pathogenesis-related protein-like protein 1 and a putative

stress-responsive protein, respectively. Moreover, two proteins belonging to the LEA family and another stress-related protein ranked among the sixteen contigs with the most ESTs.

SSR discovery

A total of 86 potentially useful SSR loci were identified in 82 uniESTs (four uniESTs have two loci each), comprising ca. 4.97% and 2.64% of the total uniESTs and total EST sequences, respectively. Trinucleotide repeats were the most abundant (48.8%), followed by di- (46.5%) and hexanucleotide repeats (1.2%). Also there were two loci with di- and trinucleotide compound repeats and one locus with di- and pentanucleotide compound repeats (Table 4).

DISCUSSION

Radix Bupleuri is an important and commonly used TCM in mainland China and Taiwan, Japan, Korea, and Southeast Asian countries, but molecular biology about the source species is poorly researched. Chen et al. (2007) reported the construction of PCR-select cDNA subtraction libraries and transcriptional changes in MeJA-induced adventitious roots of *B. kaoi*. They obtained a total of 834 ESTs representing 532 uniESTs. Kim et al. (2006) cloned core sequences of five isoprenoid pathway genes by homology-based RT-PCR and determined the correlation of transcripts of these genes with saikosaponin accumulation in *B. falcatum*. In our study, a set of 3111 ESTs representing 1650 uniESTs acquired from *B. chinense* would expand the genic sequence pool of the genus *Bupleurum*. Approximately 51.5% full-length cDNAs of those sequences which match to a known gene would provide a robust approach to identifying different functional genes in agriculture and pharmacology. In addition, compared to other crops, umbelliferous crops receive little research attention, especially in molecular biology (Rubatzky et al., 1999). Before we submitted our ESTs (May of 2008), only 8226 nucleotide sequences (5863 Nucleotides + 2363 ESTs) were registered, and only five genera—*Apium* (2224 ESTs), *Daucus* (640 Nucleotides + 38 ESTs), *Eryngium* (477 Nucleotides), *Cymopterus* (233 Nucleotides), and *Petroselinum* (214 Nucleotides)—had more than 200 nucleotide sequences according to GenBank. Therefore the sequence information we obtained from *B. chinense* could also provide a useful basis for researching other important medicinal plants and vegetables in the *Umbelliferae* family, e.g. *Angelica sinensis* (Oliv.) Diels, *A. dahurica* Maxim., *Saposhnikovia divaricata* (Turcz.) Schischk., *Pimpinella anisum* L., *Carum carvi* L., *Daucus carota* L. and *Apium graveolens* L.

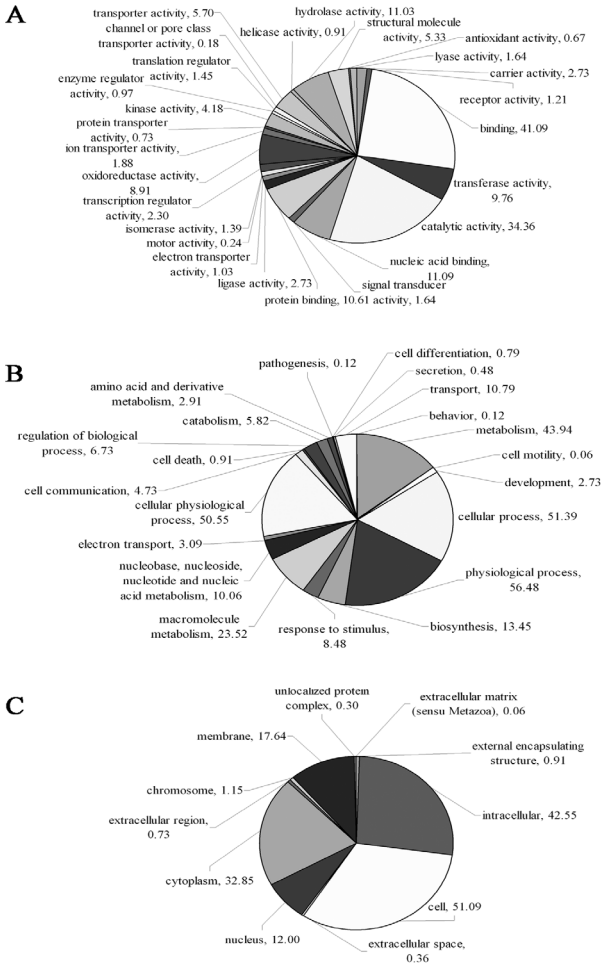

Figure 2. Gene Ontology annotation of 1650 uniESTs. A, B and C indicate molecular function, biological process, and cellular component, respectively.

Of sixteen contigs with the most abundant ESTs, nine had a (putative) relationship with stress responses, including biotic stresses (contig 25, 110) (Mosolov and Valueva, 2005) and abiotic stresses like drought, cold, or salt (contig 23, 115, 50, 41, 310) (Chang and Zhu, 2002; Lopez et al., 2004; Yakubov et al., 2005; Goyal et al., 2005). This may suggest that the natural drought environment experienced in November in Beijing induced the anti-stress responses that *Bupleurum* plants exhibit when flowering. Contig 45 with the fourth most abundant EST numbers showed a high homology with *Broad bean wilt virus* 2 (BBWV2) or *Patchouli mild mosaic virus* (PatMMV), which both belong to the genus *Fabavirus* and share a high level of similarity (Qi et al., 2000). BlastX and BlastN implied that the nucleoside and amino acid identities were 92-95% and 89-100% (different from each EST clone), respectively. The homology length was 510-797 bp and 142-265 amino acids, and the homology position was the RNA-dependent RNA polymerase (RdRp) encoded fragment of the virus RNA1. In addition, another contig (4 ESTs) and one singleton, which had homology with a different fragment of RNA1 of BBWV2 or PatMMV, were in the library. From these data, we concluded primarily that *Bupleurum* plants used for constructing the cDNA library were infected by a virus strain or isolate belonging to the genus *Fabavirus* in the family *Comoviridae*. According to Plant Virus Online

Table 2. A list of 31 pathways bearing at least five *B. chinense* homologous uniESTs by BlastX with protein sequences of *Arabidopsis*.

KEGG identifier	Pathway	Unigene Nos.	KEGG identifier	Pathway	Unigene Nos.
ath03010	Ribosome	58	ath03060	Protein export	7
ath00010	Glycolysis	17	ath00051	Fructose and mannose metabolism	6
ath00190	Oxidative phosphorylation	17	ath00071	Fatty acid metabolism	6
ath04120	Ubiquitin mediated proteolysis	17	ath00100	Biosynthesis of steroids	6
ath00710	Carbon fixation	13	ath00271	Methionine metabolism	6
ath03050	Proteasome	12	ath00450	Selenoamino acid metabolism	6
ath00230	Purine metabolism	11	ath00632	Benzoate degradation via CoA ligation	6
ath01040	biosynthesis of unsaturated fatty acids	11	ath00220	Urea cycle and metabolism of amino groups	5
ath00350	Tyrosine metabolism	10	ath00251	Glutamate metabolism	5
ath00620	Pyruvate metabolism	10	ath00360	Phenylalanine metabolism	5
ath00680	Methane metabolism	10	ath00480	Glutathione metabolism	5
ath00500	Starch and sucrose metabolism	9	ath00562	Inositol phosphate metabolism	5
ath04070	Phosphatidylinositol signaling system	8	ath00940	Phenylpropanoid biosynthesis	5
ath00020	Citrate cycle (TCA cycle)	7	ath00960	Alkaloid biosynthesis II	5
ath00030	Pentose phosphate pathway	7	ath04130	SNARE interactions in vesicular transport	5
ath00380	Tryptophan metabolism	7			

Table 3. BlastX results of 16 contigs with most redundant ESTs in the *B. chinense* full-length enriched cDNA library.

Contig	No. of ESTs	Percentage of total	NCBI BlastX			
			Accession No.	Species	Gene name	e-value
contig23	147	4.73%	BAD86644	*Daucus carota*	Dehydrin protein	8e-63
contig25	119	3.83%	BAD04841	*Daucus carota*	Pathogenesis-related protein-like protein 1	1e-116
contig13	65	2.09%	AAT01418	Tamarix androssowii	Putative stress-responsive protein	3e-93
contig45	48	1.54%	BAB83045	*Broad bean wilt virus* 2	210 kDa protein precursor	1e-144
contig110	45	1.45%	AAU81597	*Petunia* x *hybrida*	Cysteine proteinase inhibitor	1e-140
contig115	30	0.96%	CAA33406	*Brassica napus*	Late embryogenesis abundant protein 76	1e-119
contig182	24	0.77%	CAO65275	*Medicago truncatula*	Eukaryotic/archaeal ribosomal protein S3	9e-26
contig310	22	0.71%	AAC62510	*Pimpinella brachycarpa*	Metallothionein-1-like protein	0.0
contig6	20	0.64%	AAK83601	*Arabidopsis thaliana*	Glyceraldehyde-3-phosphate dehydrogenase	1e-119
contig50	20	0.64%	AAA61564	*Gossypium hirsutum*	Desiccation protectant protein Lea14 homolog	1e-114
contig53	20	0.64%	EAY98759	*Oryza sativa*	Hypothetical protein OsI_019992	1e-126
contig109	20	0.64%	AAD51854	*Vitis vinifera*	Stress related protein	8e-76
contig41	19	0.61%	ABB29477	*Panax ginseng*	Tonoplast intrinsic protein	1e-150
contig158	19	0.61%	AAR20771	*Arabidopsis thaliana*	GRAM domain-containing protein / ABA-responsive protein-related	1e-135
contig10	18	0.58%	AAY87906	*Sesamum indicum*	Caleosin B	1e-88
contig97	18	0.58%	ABL67651	*Citrus* cv. Shiranuhi	Putative auxin-repressed/dormancy-associated protein	1e-73
Total	654	21.02%				

Table 4. A list of 86 SSR loci identified in 82 uniESTs with repeat motifs and BlastX results shown.

Unigene ID	Repeat	NCBI BlastX	
		Species	Gene name
Bc-0002	$(TA)_6TT(TA)_5$	*Petroselinum crispum*	Common plant regulatory factor 7
Bc-0013	$(AAG)_8$	*Tamarix androssowii*	Putative stress-responsive protein
Bc-0017	$(TTC)_5$	*Vitis vinifera*	Hypothetical protein
Bc-0023	$(TA)_6(CA)_5$	*Daucus carota*	Dehydrin protein
Bc-0024	$(AT)_{10}$	*Glycine max*	SCOF-1
Bc-0025	$(TA)_{10}$	*Daucus carota*	Pathogenesis-related protein-like protein 1
Bc-0029	$(TA)_{10}$	*Ricinus communis*	Cyclophilin
Bc-0085	$(TCC)_5$	*Vitis vinifera*	CBF4 transcription factor
Bc-0113	$(GAT)_6$	*Solanum tuberosum*	Histone deacetylase 2a-like
Bc-0115	$(AT)_{12}$	*Brassica napus*	Late embryogenesis abundant protein 76 (LEA 76)
Bc-0119	$(GT)_5(AT)_5$	*Arabidopsis thaliana*	Unknown protein
Bc-0125	$(TC)_5(TA)_6+(AT)_{11}$	*Medicago truncatula*	Hypothetical protein MtrDRAFT_AC174144g16v1
Bc-0137	$(AT)_{10}$		No hit
Bc-0140	$(AT)_{10}$	*Arabidopsis thaliana*	Late embryogenesis abundant protein-like
Bc-0150	$(TG)_{11}$	*Craterostigma plantagineum*	Group 4 LEA protein
Bc-0199	$(AT)_{10}$	*Oryza sativa*	Hypothetical protein OsI_031007
Bc-0211	$(AAT)_6$	*Capsicum annuum*	WRKY-type transcription factor
Bc-0233	$(TA)_9+(AAC)_7$	*Arabidopsis thaliana*	Rcd1-like cell differentiation protein, putative
Bc-0254	$(TTA)_5$	*Lycopersicon esculentum*	TAGL12 transcription factor
Bc-0259	$(GT)_{11}$	*Solanum tuberosum*	Cinnamoyl CoA reductase
Bc-0274	$(AT)_9$	*Mesembryanthemum crystallinum*	Major latex protein homolog
Bc-0285	$(CA)_6CT(CA)_6$	*Vitis vinifera*	Hypothetical protein
Bc-0309	$(ACA)_5$	*Ricinus communis*	Cyclophilin
Bc-0310	$(TTA)_7$	*Pimpinella brachycarpa*	Metallothionein-1-like protein
Bc-0341	$(GCT)_5$	*Arabidopsis thaliana*	ARF GAP-like zinc finger-containing protein ZIGA3
Bc-0356	$(ACT)_5$	*Homo sapiens*	hCG1999844
Bc-0413	$(TA)_9$	*Solanum tuberosum*	snakin2
Bc-0418	$(GGT)_3+(TGG)_5$	*Rattus norvegicus*	PREDICTED: similar to keratin associated protein 10-7
Bc-0448	$(AT)_9$	*Arabidopsis thaliana*	Phosphatidylethanolamine-binding family protein
Bc-0461	$(AGA)_5$	*Vitis vinifera*	Hypothetical protein
Bc-0466	$(CA)_{13}$	*Vitis vinifera*	Putative ripening-related protein
Bc-0473	$(AT)_9$	*Trichomonas vaginalis* G3	Hypothetical protein TVAG_343280
Bc-0478	$(TA)_{10}$	*Vitis vinifera*	Hypothetical protein
Bc-0507	$(GGA)_8$	*Lycopersicon esculentum*	Mature anther-specific protein LAT61
Bc-0515	$(CCG)_5$	*Glycine max*	Aspartate aminotransferase glyoxysomal isozyme AAT1 precursor
Bc-0557	$(AT)_9(TATAT)_4$	*Oryza sativa*	Os01g0633400 putative YZ1
Bc-0573	$(CCG)_5$	*Arabidopsis thaliana*	RNA recognition motif (RRM)-containing protein
Bc-0603	$(TC)_9$	*Arabidopsis thaliana*	UBC22 (ubiquitin-conjugating enzyme 18); ubiquitin-protein ligase
Bc-0635	$(TAT)_6$	*Ostreococcus tauri*	Unnamed protein product
Bc-0659	$(AGA)_5$	*Glycine max*	Histone deacetylase HDT1 (Histone deacetylase 2a)
Bc-0700	$(GAA)_2GAC(GAA)_5$ $(AAG)_2(GAA)_4$	*Arabidopsis thaliana*	Unknown protein
Bc-0776	$(GCA)_5$	*Drosophila melanogaster*	Flap endonuclease 1 CG8648-PA
Bc-0794	$(AT)_{11}$	*Arabidopsis thaliana*	Hypothetical protein

Table 4. (Continued)

Unigene ID	Repeat	NCBI BlastX	
		Species	Gene name
Bc-0821	$(TCA)_5(TC)_8$	*Olea europaea*	Cytochome b5
Bc-0824	$(AAG)_5$	*Solanum tuberosum*	Eukaryotic translation initiation factor 2 beta subunit-like
Bc-0850	$(AT)_9$	*Lycopersicon esculentum*	Lecithine cholesterol acyltransferase-like protein
Bc-0861	$(AAG)_5$	*Oryza sativa*	Hypothetical protein
Bc-0861	$(ATC)_5$	*Oryza sativa*	Hypothetical protein
Bc-0915	$(AAT)_{11}$	*Ipomoea batatas*	ATP synthase delta' chain, mitochondrial precursor
Bc-0949	$(GGA)_6$	*Arabidopsis thaliana*	Hypothetical protein
Bc-0960	$(GAA)_5$		No hit
Bc-0973	$(AC)_9$	*Vitis vinifera*	Hypothetical protein
Bc-0995	$(TC)_7(TA)_8$	*Oryza sativa*	Os10g0533900 unknown protein
Bc-1029	$(AAG)_6$	*Lycopersicon esculentum*	Bax inhibitor
Bc-1045	$(AT)_9$	*Arabidopsis thaliana*	Got1-like family protein
Bc-1063	$(AT)_9$	*Oryza sativa*	Os02g0693200 DnaJ protein-like
Bc-1088	$(AAG)_5+(AAG)_6$	*Arabidopsis thaliana*	Transducin family protein / WD-40 repeat family protein
Bc-1120	$(TTC)_5$	*Arabidopsis thaliana*	Unknown protein
Bc-1126	$(TGT)_5$	*Vitis vinifera*	Hypothetical protein
Bc-1150	$(TTC)_5$	*Vitis vinifera*	Unnamed protein product
Bc-1169	$(GT)_5+(AC)_5$	*Vitis vinifera*	Unnamed protein product
Bc-1186	$(AC)_{10}$	*Vitis vinifera*	Unnamed protein product
Bc-1192	$(TA)_9$	*Homo sapiens*	PREDICTED: hypothetical protein
Bc-1192	$(TG)_6(TA)_8$	*Homo sapiens*	PREDICTED: hypothetical protein
Bc-1249	$(TTC)_6$	*Oryza sativa*	Os03g0119000 ABR017Cp, putative, expressed
Bc-1269	$(AT)_{13}$	*Arachis hypogaea*	Subtilisin-like protease
Bc-1275	$(TA)_{10}$	*Arabidopsis thaliana*	ATRPABC24.3 ('Arabidopsis thaliana RNA polymerase I, II and III 24.3 kDa subunit'); DNA binding / DNA-directed RNA polymerase
Bc-1329	$(TA)_{10}$		No hit
Bc-1335	$(CA)_{11}(TA)_7$	*Vitis vinifera*	Hypothetical protein
Bc-1377	$(GT)_{10}(GA)_{10}$	*Arabidopsis thaliana*	Subtilase family protein
Bc-1397	$(TA)_9$	*Vitis vinifera*	Hypothetical protein
Bc-1419	$(AGA)_7$	*Oryza sativa*	Hypothetical protein
Bc-1423	$(AAC)_5$	*Vitis vinifera*	Hypothetical protein
Bc-1461	$(GAA)_5$		No hit
Bc-1470	$(CTA)_7$	*Glycine max*	Dehydration responsive element binding protein
Bc-1503	$(TA)_{10}$	*Cucumis sativus*	3-ketoacyl-CoA thiolase; acetyl-CoA acyltransferase
Bc-1511	$(AAGTAG)_5$	*Vitis vinifera*	Hypothetical protein
Bc-1511	$(GCA)_8A(ATG)_5$	*Vitis vinifera*	Hypothetical protein
Bc-1542	$(TGA)_6$	*Vitis vinifera*	Hypothetical protein
Bc-1542	$(AAG)_5$	*Vitis vinifera*	Hypothetical protein
Bc-1543	$(GGT)_5$	*Vitis vinifera*	Hypothetical protein
Bc-1561	$(CAA)_5$	*Solanum tuberosum*	Oligouridylate binding protein-like protein
Bc-1562	$(AT)_6+(AT)_6$	*Prunus persica*	Major latex-like protein
Bc-1563	$(CTG)_5$	*Medicago truncatula*	Pumilio/Puf RNA-binding
Bc-1590	$(GTC)_5$	*Arabidopsis thaliana*	Unknown protein
Bc-1630	$(TG)_9$	*Medicago truncatula*	Galactose mutarotase-like

(http://image.fs.uidaho.edu/vide/descr115.htm), one of the BBWV synonyms was Parsley virus 3, which infects parsley asymptomatically, but until now, no Fabavirus has been reported to infect *Bupleurum* plants. Virus inoculation, virion isolation, and complete sequence analysis will be needed to verify the virus infection of our experimental plant material.

Saikosaponins, the main pharmacologically active component of *Bupleurum*, are synthesized via the isoprenoid pathway by cyclization of 2, 3-oxidosqulene to produce oleanane (β-amyrin) or a dammarane triterpenoid skeleton, and the triterpenoid backbone is modified by P450 and glycosyltransferases (Haralampidis et al., 2002; Chen et al., 2007). In the report of Chen et al. (2007), ESTs of a β-amyrin synthase gene (*β-AS*), a UDP-glucosyltransferase gene, and two *P450* genes were identified, and the transcripts of these genes were all upregulated in the adventitious roots of *B. kaoi* induced by MeJA. Of 3111 ESTs sequenced in our study, one 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase gene (*HMGR*), five glycosyltransferase genes (*GT*), and five cytochrome P450 genes (*P450*) were included. *GT* and *P450* comprise superfamilies genes in plants. They are seldom cloned by the regular PCR-based method, and their function is difficult to determine. To screen the large-scale ESTs of induced or developmental EST or cDNA libraries and subsequently verify their function is an effective measure (Osbourn, 2003; Goossens et al., 2003). Whether these genes play a role in saikosaponin biosynthesis is of interest. Although previous investigation showed roots of flowering *B. chinense* plants had the highest content of saikosaponin (Yang et al., 2006), no more cDNAs involved in saikosaponin biosynthesis were identified in the sequenced cDNA pool. Consistent with that, the ginseng leaf cDNA library contained five ESTs of ginsenoside-synthesizing enzymes out of 2,896 ESTs (0.17%), and the ginseng root cDNA library featured two out of 3,808 ESTs (0.05%) (Kim et al., 2006). This phenomenon implies that the transcript abundances of genes involved in saikosaponin biosynthesis are naturally rather low or the plant developmental stage during which roots of plants have the highest content of saikosaponin may not be the time that saikosaponin is rapidly biosynthesized. Additionally, it is possible that the transcripts of some related genes degrade and regenerate rapidly. Some reports such as research by Kim et al. (2006) have shown that the gene expression responsible for the biosynthesis of secondary metabolic products is upregulated dramatically when metabolites are increased artificially. The behavior of plants when cultivated remains to be characterized.

SSR (also referred to as microsatellite), with remarkable attributes such as codominant inheritance and ease of detection, has become preferred over other molecular markers like RAPD, ISSR, and AFLP. With the growing amount of EST data, EST-SSRs have been identified by some plant EST or cDNA libraries (Lindqvist et al., 2006; Chen et al., 2006). Mining SSRs from EST data has proven to be a time and cost saving method (Ceresini et al., 2005). During our SSR search in the *B. chinense* root full-length enriched cDNA library, only loci for which there were at least 50 bp for potential flanking primers before and after the repeat site were accounted for. Therefore, future work will need the design of primers for these 86 SSR loci. This could lead to the development of a set of SSR markers. These putative expressed SSR markers plus those genomic SSR markers we have developed for *B. chinense* (Sui et al., 2009) would be extremely useful for analyzing genetic diversity, germplasm identification, and genetic map construction for the *Bupleurum* genus plant.

Acknowledgements. This research was supported by the National Key Project of Scientific and Technical Supporting Programs funded by the Ministry of Science & Technology of China (No. 2006BAI09B01), by the Special Funds in Basic Scientific Research for Non-Profit Research Institutes financed by the Ministry of Finance, People's Republic of China (No. YZ-1-10), and by the Special Funds in Scientific and Technological Research financed by the State Administration of Traditional Chinese Medicine (No. 2004ZX06-3).

LITERATURE CITED

Aoyagi, H., Y. Kobayashi, K. Yamada, M. Yokoyama, K. Kusakari, and H. Tanaka. 2001. Efficient production of saikosaponins in *Bupleurum falcatum* root fragments combined with signal transducers. Appl. Microbiol. Biotechnol. **57:** 482-488.

Ceresini, P.C., C.L.S. P. Silva, R.F. Missio, E.C. Souza, C.N. Fischer, I.R. Guillherme, I. Gregorio, E.H.T. da Silva, R.M.B. Cicarelli, M.T.A. da Silva, J.F. Garcia, G.A. Avelar, L.R.P. Neto, A.R. Marçon, M.B. Junior, and D.C. Marini. 2005. Satellyptus: Analysis and database of microsatellites from ESTs of *Eucalyptus*. Genet. Mol. Biol. **28:** 589-600.

Chang, T.J. and Z. Zhu. 2002. Study advances of plant metallothionein: expression characteristics and functions of plant MT gene. Biotechnology Bulletin. **5:** 1-5, 92.

Chen, C.X., P. Zhou, Y.A. Choi, S. Huang, and F.G. Gmitter Jr. 2006. Mining and characterizing microsatellites from citrus ESTs. Theor. Appl. Genet. **112:** 1248-1257.

Chen, L.R., Y.J. Chen, C.Y. Lee, and T.Y. Lin. 2007. MeJA-induced transcriptional changes in adventitious roots of *Bupleurum kaoi*. Plant Sci. **173:** 12-24.

Chen, Z.Z., C.H. Xue, S. Zhu, F.F. Zhou, X.F.B. Ling, G.P. Liu, and L.B. Chen. 2005. GoPipe: Streamlined gene ontology annotation for batch anonymous sequences with statistics. Prog. Biochem. Biophys. **32:** 187-191.

D'Agostino, N., D. Pizzichini, M.L. Chiusano, and G. Giuliano. 2007. An EST database from saffron stigmas. BMC Plant Biol. **7:** 53.

Divol, F., F. Vilaine, S. Thibivilliers, J. Amselem, J.C. Palauqui, C. Kusiak, and S. Dinant. 2005. Systemic response to aphid infestation by *Myzus persicae* in the phloem of *Apium graveloens*. Plant Mol. Biol. **57:** 517-540.

Goossens, A., S.T. Hakkinen, I. Laakso, T. Seppanen-Laakso, S. Biondi, V. De Sutter, F. Lammertyn, A.M. Nuutila, H. Soderlund, M. Zabeau, D. Inze, and K.M. Oksman-Caldentey. 2003. A functional genomics approach toward the understanding of secondary metabolism in plant cells. Proc. Natl. Acad. Sci. USA **100:** 8595-8600.

Goyal, K., L.J. Walton, and A. Tunnacliffe. 2005. LEA proteins prevent protein aggregation due to water stress. Biochem. J. **388:** 151-157.

Haralampidis, K., M. Trojanowska, and A.E. Osbourn. 2002. Biosynthesis of triterpenoid saponins in plants. Adv. Biochem. Eng. Biotechnol. **75:** 31-49.

Jia, J.P., J.J. Fu, J. Zheng, X. Zhou, J.L. Huai, J.H. Wang, M. Wang, Y. Zhang, X.P. Chen, J.P. Zhang, J.F. Zhao, Z. Su, Y.P. Lv, and G.Y. Wang. 2006. Annotation and expression profile analysis of 2073 full-length cDNAs from stress induced maize (*Zea mays* L.) seedlings. Plant J. **48:** 710-727.

Jung, J.D., H.W. Park, Y. Hahn, C.G. Hug, D.S. In, H.J. Chung, J.R. Liu, and D.W. Choi. 2003. Discovery of genes for ginsenoside biosynthesis by analysis of ginseng expressed sequence tags. Plant Cell Rep. **22:** 224-230.

Kim, M.K., B. S. Lee, J. G. In, H. Sun, J. H. Yoon, and D. C. Yang. 2006. Comparative analysis of expressed sequence tags (ESTs) of ginseng leaf. Plant Cell Rep. **25:** 599-606.

Kwon, S. J., S. W. Hong, N. S. Kim, and J. C. Kim. 2004. Isolation of callus-specific mRNAs from differentiating embryogenic somatic calli of *Pimpinella brachycarpa* by cDNA-AFLP. Mol. Cell. **17:** 39-44.

Li, Q. and J.M. Wan. 2005. SSRHunter: Development of a local searching software for SSR sites. Hereditas **27:** 808-810.

Lin, X.Y., G.J. Hwang, and J.L. Zimmerman. 1996. Isolation and characterization of a diverse set of genes from carrot somatic embryos. Plant Physiol. **112:** 1365-1374.

Lindqvist, C., A.C. Scheen, M.J. Yoo, P. Grey, D. Oppenheimer, J. Leebens-Mack, D. Soltis, P. Soltis, and V. Albert. 2006. An expressed sequence tag (EST) library from developing fruits of a Hawaiian endemic mint (*Stenogyne rugosa*, Lamiaceae): characterization and microsatellite markers. BMC Plant Biol. **6:** 16-30.

Liu, Q.X., L. Tan, Y.J. Bai, H. Liang, and Y.Y. Zhao. 2002. A survey of the studies on saponins from *Bupleurum* in past 10 years. China J. Chin. Mat. Med. **27:** 7-11, 45.

Lopez, F., A. Bousser, I. Sissoeff, J. Hoarau, and A. Mahe. 2004. Characterization in maize of ZmTIP2-3, a root-specific tonoplast intrinsic protein exhibiting aquaporin activity. J. Exp. Bot. **55:** 539-541.

Mosolov, V.V. and T.A. Valueva. 2005. Proteinase inhibitors and their function in plants: a review. Appl. Biochem. Microbiol. **41:** 261-282.

Osbourn, A.E. 2003. Molecules of interest. Saponins in cereals. Phytochemistry **62:** 1-4.

Pan, S.L. (ed.) 2006. *Bupleurum* Species: Scientific Evaluation and Clinical Applications. Taylor & Francis Group, 272 pp.

Park, J.S. and S.G. Park. 2006. Identification of differentially expressed genes involved in spine formation on seed of *Daucus carota* L. (Carrot), using annealing control primer system. J. Plant Biol. **49:** 133-140.

Qi, Y.Z., X.P. Zhou, Z.Y. Xue, and D.B. Li. 2000. Complete sequence of *Broad bean wilt virus* China isolate RNA2 and its polyprotein digestion site. Prog Nat. Sci. **10:** 805-811.

Rubatzky, V.E., C.F. Quiros, and P.W. Simon (eds.). 1999. Carrots and Related Vegetable Umbelliferae (Crop Production Science in Horticulture). CABI Publishing, 304 pp.

Sui, C., J.H. Wei, S.L. Chen, H.Q. Chen, and C.M. Yang. 2009. Development of genomic SSR and potential EST-SSR markers in *Bupleurum chinense* DC. Afr. J. Biotechnol. **8:** 6233-6240.

Urgamal, M., C. Sanchir, and M.L. Zhang. 2007. Classification and distribution of *Bupleurum* L. (*Umbelliferae* Juss.) in Mongolia. Bull. Bot. Res. **27:** 20-24.

Vilaine, F., J.C. Palauqui, J. Amselem, C. Kusiak, R. Lemoine, and S. Dinant. 2003. Towards deciphering phloem: a transcriptome analysis of the phloem of *Apium graveolens*. Plant J. **36:** 67-81.

Wei, J.H., H.Z. Cheng, K.T. Li, W.L. Ding, Z.X. Xu, and Q.L. Chu. 2003. Study on organogenesis and dry substance accumulation of *Bupleurum chinense* DC. J. Chin. Med. Mat. **26:** 469-471.

Wellenreuther, R., I. Schupp, A. Poustka, and S. Wiemann. 2004. SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones. BMC Genomics **5:** 36.

Yakubov, B., O. Barazani, A. Shachack, L. J. Rowland, O. Shoseyov, and A. Golan-Goldhirsh. 2005. Cloning and expression of a dehydrin-like protein from *Pistacia vera* L. Trees-Struct. Funct. **19:** 224-230.

Yang, C.M, J.H. Wei, H.Z. Cheng, S.L. Chen, F.J. Ma, and Z.W. Huang. 2006. Study on the content undulation of saikosaponin in *Bupleurum chinense* DC. J. Chin. Med. Mat. **29:** 316-318.

Yang, Z.Y., Z. Chao, K.K. Huo, H. Xie, Z.P. Tian, and S.L. Pan. 2007. ITS sequence analysis used for molecular identification of the *Bupleurum* species from northwestern China. Phytomedicine **14:** 416-422.

Yen, M.H., T.C. Weng, S.Y. Liu, C.Y Chai, and C.C. Lin. 2005. The hepatoprotective effect of *Bupleurum kaoi*, an endemic plant to Taiwan, against dimethylnitrosamine-induced hepatic fibrosis in rats. Biol. Pharm. Bull. **28:** 442-448.

柴胡全長富集 cDNA 文庫構建及其 3111 個 EST 序列測定分析

隋 春　魏建和　陳士林　陳懷瓊　董樂萌　楊成民

中國醫學科學院及北京協和醫學院藥用植物研究所

中藥柴胡，爲柴胡屬植物的乾燥根，具有抗炎、退熱和保肝等功效。在中國、日本、韓國及南亞其他國家和地區廣泛應用。本文利用 SMART 技術構建了藥用柴胡屬植物北柴胡 (*Bupleurum chinense* DC.) 根的全長富集 cDNA 文庫，以啓動藥用柴胡功能基因組研究。文庫滴度爲 1.1×10^6。從文庫中隨機選擇了 3902 個克隆進行 5' 端單反應測序，獲得了 3111 個高品質 EST 序列，包括 377 個序列重疊群 (contigs) 和 1273 個單一序列 (singletons) 共 1650 個獨立 EST (uniESTs)。文庫插入片段平均長度約 1.1 kb，全長比率約 51.5%。BlastX 分析結果表明 949 (57.5%) 個獨立 EST 和已鑒定基因同源，680 (41.2%) 個獨立 EST 與 GenBank 中未知、未命名或推測的蛋白基因同源，21 個獨立 EST 沒有找到同源基因。獨立 EST 的 GO (Gene Ontology) 注釋顯示 1002 、957 和 861 個獨立 EST 分別歸屬於細胞功能、生物過程和細胞組分三大類。與 KEGG 中擬南芥代謝途徑進行了比較分析，結果表明，307 個獨立 EST 可能參與 31 個代謝途徑（每個代謝途徑至少包含 5 個獨立 EST）。1650 個獨立 EST 中 82 個含有總計 86 個微衛星位點 (SSR)。本研究首次構建的柴胡根全長富集 cDNA 文庫及測序分析的 EST 克隆，爲研究這一中藥材的各種生理現象的分子基礎提供了有效平臺。從文庫資料中挖掘的具有潛在標記價值的 SSR 位點將爲柴胡種質鑒定，遺傳多樣性分析及基因定位提供有效分子標記。

關鍵詞 ： 北柴胡（*Bupleurum chinense* DC.）； 全長富集 cDNA 文庫 ； 表達序列標籤（ESTs）；SMART（Switching mechanism at 5' end of RNA transcript）技術 ；SSR。