

# Development of EST database and transcriptome analysis in the leaves of *Brassica rapa* using a newly developed pipeline

Vignesh Dhandapani · Su Ryun Choi · Parameswari Paul · Yong-Kwon Kim · Nirala Ramchiary · Yoonkang Hur · Yong Pyo Lim

Received: 02 February 2012 / Accepted: 09 July 2012 / Published online: 01 December 2012  
© The Genetics Society of Korea and Springer 2012

## Abstract

*Brassica rapa* L. (AA,  $2n = 20$ ), an A genome diploid species of *Brassica* genus is of researchers interest recent days since enormous amount of data is available about the genome. Since EST analysis is a powerful tool in gene discovery we compared different existing methods and developed a new pipeline for EST computational analysis to analyze the available data. A total of 1,438 expressed sequence tags sizing from 83 to 2,023 base pairs were generated and subjected to various types of analysis. Cluster analysis of these ESTs identified 969 unique sequences called unigenes, with 162 contigs and 807 singlets. Similarity search produced 704 significant hits with  $E\text{-value} \geq 10^{-5}$ . The functions of the best hits were annotated by gene ontology (GO) analysis. Additionally, we classified 293 and 541 unigenes based on their functions, using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and protein structural domain databases, respectively. We also identified and categorized 171 microsatellites into di-, tri-, tetra-, and penta nucleotide repeats, and designed primers. Possible open reading frames (ORFs) were predicted for 960 unigenes, by the comparison with a primary protein sequence database. *In silico* mapping of partial unigenes were done in bacterial artificial chromosome (BAC) sequences, downloaded

from the *Brassica* genome project website. We determined 149 single nucleotide polymorphisms (SNPs) and 3 indels from the coding region of 27 unigenes of *B. rapa* and similar *Brassica napus* ESTs clusters. All the generated EST sequences were submitted to the GenBank EST database (dbEST) as accessions from CO749247 to CO750425.

**Keywords** *Brassica rapa*; Transcriptome; Expressed sequence tag (EST); Functional annotation

## Introduction

*Brassica rapa* a dicot annual herb, commonly known as turnip or turnip rape is widely cultivated as a leafy vegetable. The genus *Brassica* includes many species of which 6 species are cultivated mainly for producing vegetables, oilseeds, condiments, and fodder. *Brassica rapa* (AA,  $2n = 20$ ) is one among the 6 species of economically important, which include 2 other diploid species, *Brassica nigra* (BB,  $2n = 16$ ) and *Brassica oleracea* (CC,  $2n = 18$ ), and 3 amphidiploid species, *Brassica juncea* (AABB,  $2n = 36$ ), *Brassica napus* (AACC,  $2n = 38$ ), and *Brassica carinata* (BBCC,  $2n = 34$ ). Additionally, *B. rapa* is an “A” genome progenitor species of the 2 oilseed *Brassica* species, *B. napus* and *B. juncea*. The genetic relationships of *Brassica* species belonging to the U’s triangle are well documented (U, 1935). *B. oleracea* comprises the largest diversified form of vegetable crop that includes broccoli, cabbage, cauliflower, kale and Brussels sprouts. *B. rapa* Chinese cabbage (ssp *Pekinensis*) and Pakchoi (ssp *parakinensis*) are widely grown as leafy vegetables mainly in Asian countries such as China, Korea and Japan. These *Brassica* vegetables are an important source of dietary fiber, vitamin C, and anti-cancer compounds (Fahey and Talalay, 1995).

An expressed sequence tag or EST is a short sub-sequence of a cDNA sequence which may be used to identify the gene transcripts, and are instrumental in gene discovery and gene

V. Dhandapani · S. R. Choi · P. Paul · N. Ramchiary ·

Y. P. Lim (✉)

Korean Brassica Genome Resource Bank, Department of Horticulture Sciences, College of Agriculture and Life Sciences, Chungnam National University, Yuseong-gu, Daejeon-305764, Korea

e-mail: yplim@cnu.ac.kr

Y.-K. Kim

NH Seed Research & Development Center, National Agricultural Cooperative Federation, Sindu-ri 432-4, Gongdo-eup, Anseong-si, Gyeonggi-do, Korea

Y. Hur

Department of Biological Sciences, Chungnam National University, Yuseong-gu, Daejeon-305764, Korea

sequence determination (Adams et al, 1991). For EST sequencing, RNAs are reversely transcribed into double-stranded complementary DNA (cDNA) using reverse transcriptase enzyme, because the messenger RNA (mRNA) sequences are copies of expressed genes. These generated cDNAs are cloned to make libraries of transcribed genes of tissues, and the cDNA clones are randomly sequenced to obtain ESTs. The ESTs and cDNAs are the fundamental resources for transcriptome identification and analysis. Transcriptome analysis is one of the most effective means of gene identification, gene expression profiling, functional genome studies, single nucleotide polymorphism (SNP) characterization, proteome analysis, and construction of linkage maps.

Here, we developed a new pipeline to briefly analyze the generated *B. rapa* ESTs for transcriptome analysis, using highly efficient and error-less tools. Our study provides substantial information on transcriptome analysis of *B. rapa* leaves. The clustered high-quality ESTs were subjected to comparative analysis for mapping in genome sequence, similar gene identifications, putative functional annotations, structure predictions, and polymorphism studies, using a specially constructed pipeline with the appropriate tools for plant transcriptome analysis. From the annotation of ESTs, we identified many putative genes involved in leaf development and different environmental stresses. Additionally, by comparison of similar *B. napus* ESTs, we identified SNPs that play important roles in the coding regions of *B. rapa* and *B. napus*. Finally, we constructed a *Brassica* EST database with similar genes, their genome locations, annotated functions, SNPs, and microsatellites.

## Materials and Methods

### Plant materials and cDNA library construction

*Brassica rapa* L. ssp. *pekinensis*, inbred line Chiifu, was used as plant material. Total RNA was isolated from the young leaves of green house grown plants using TRIzol (Gibco-BRL, USA), and poly(A)<sup>+</sup> mRNA was extracted. The cDNA library was constructed using the ZAP-cDNA Synthesis and Cloning Kit (Stratagene, USA). First-strand cDNA was synthesized from poly(A)<sup>+</sup> mRNA (5 µg) with MMLV-RT (reverse transcriptase) at 37°C for 1 hour. A mixture of second-strand dNTP and DNA polymerase I was used in second-strand cDNA synthesis at 16°C for 2.5 hours. *Eco*RI and *Xho*I adaptors were added to phosphorylated cDNA ends. After size fractionation, cDNAs longer than 800 bp were cloned in Uni-ZAP XR vector and packaged with Gigapack III Gold (Stratagene, USA), according to the manufacturer's specifications.

### DNA isolation and sequencing

Each EST clone-containing bacteria were grown in 1.2 ml of YT culture medium, supplemented with ampicillin. The in-

sert DNA was isolated by a high throughput method using 96 well UNIFILTER (Whatman, USA). The cDNA inserts were amplified by PCR with T3 primers and purified using a Sephadex G-50 column (Sigma, Switzerland). Sequencing was performed using an ABI 3700 automatic sequencer (Perkin-Elmer Applied Biosystems).

### EST processing and clustering

The EST's were subjected to a newly developed pipeline (Fig. 1) for computational analysis, using highly efficient tools (Supplementary Table 1) for evaluation. As a preprocess, trimming and validation of ESTs were done, by cleaning the low-quality, low-complexity vectors using SeqTrim (Juan et al., 2010) software. The NCBI UniVec data set was utilized for the vector cleaning process. Sequences less than 100 bp were excluded from further analysis, to maintain the quality of the sequences. RepeatMasker (Smit et al., 2010) software was used for screening and removing the interspersed repeats and low-complexity ESTs. EST clustering was done by CAP3 (Huang and Madan, 1999) assembler, to produce high-fidelity consensus sequences, and to maintain a high level of sensitivity compared to other similar assemblers (Feng et al., 2000).

### Similarity search and translation of unigenes

To map the unigenes of the *B. rapa* genome, we downloaded the available bacterial artificial chromosome (BAC) sequences from the *B. rapa* genome sequencing project consortium (Kim et al., 2009), Multinational *Brassica* Genome Project (MBGP) website ([www.brassica.info](http://www.brassica.info)), and compared it using a BLAST standalone genome similarity search suite (James, 2002). In-home developed Perl script was used to screen the high-homologous regions with 80% query coverage and 90% identity cut-off in the genome sequence. We subsequently staged the homolog BAC sequences in the MBGP website, and mined the published physical maps (Kim et al., 2009; Park et al., 2005) to determine the location and chromosome.

We downloaded the recently updated *A. thaliana* TAIR 10<sup>th</sup> version gene sequences, available from [ftp://ftp.arabidopsis.org/home/tair/Sequences/blast\\_datasets/TAIR10\\_blastsets/](ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/) and compared these with the unigenes by using the megaBLAST algorithm to find homolog genes, because *A. thaliana* and *B. rapa* genomes are highly similar (Kim et al., 2009; Park et al., 2005). BLASTx (Christiam et al., 2009) algorithm was used to translate the EST sequences into 6 reading frames and highly homologous sequences were submitted to the OrfPredictor server (Min et al., 2005), to predict the most probable coding regions in all frames. Further the sub-cellular localization of ORF sequences was predicted by the TargetP online server (Olof et al., 2007).

### Functional characterization

Functional annotation of genes was done by using Blast2GO (B2G) comprehensive software (Ana and Stefan, 2008; Stefan