PLOS ONE

RESEARCH ARTICLE

Development of EST-SSR markers in flowering Chinese cabbage (*Brassica campestris* L. ssp. *chinensis* var. *utilis* Tsen et Lee) based on *de novo* transcriptomic assemblies

Jingfang Chen[1☯], Ronghua Li[1☯], Yanshi Xia[1], Guihua Bai[2], Peiguo Guo[1]*, Zhiliang Wang[1], Hua Zhang[3], Kadambot H. M. Siddique[4]

1 International Crop Research Center for Stress Resistance, College of Life Sciences, Guangzhou University, Guangzhou, China, **2** Hard Winter Wheat Genetics Research Unit, United States Department of Agriculture–Agricultural Research Service, Manhattan, Kansas, United States of America, **3** Guangzhou Academy of Agricultural Sciences, Guangzhou, China, **4** The UWA Institute of Agriculture and School of Agriculture & Environment, The University of Western Australia, Perth WA, Australia

☯ These authors contributed equally to this work.
* guopg@gzhu.edu.cn

Citation: Chen J, Li R, Xia Y, Bai G, Guo P, Wang Z, et al. (2017) Development of EST-SSR markers in flowering Chinese cabbage (*Brassica campestris* L. ssp. *chinensis* var. *utilis* Tsen et Lee) based on *de novo* transcriptomic assemblies. PLoS ONE 12(9): e0184736. https://doi.org/10.1371/journal.pone.0184736

Editor: Xiang Jia Min, Youngstown State University, UNITED STATES

Received: May 28, 2017

Accepted: August 30, 2017

Published: September 13, 2017

Data Availability Statement: The raw data were deposited into the NCBI SRA database under the accession number SRP114731. Reviewer / collaborator link to metadata: ftp://ftp-trace.ncbi.nlm.nih.gov/sra/review/SRP114731_20170818_131603_2f29acc3347fd6f6ff001c1337a94fd2 and https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP114731. The assembled transcripts data including SSR data were deposited into the NCBI TSA database; this Transcriptome Shotgun Assembly project has been deposited at DDBJ/

Abstract

Flowering Chinese cabbage is one of the most important vegetable crops in southern China. Genetic improvement of various agronomic traits in this crop is underway to meet high market demand in the region, but the progress is hampered by limited number of molecular markers available in this crop. This study aimed to develop EST-SSR markers from transcriptome sequences generated by next-generation sequencing. RNA-seq of eight cabbage samples identified 48,975 unigenes. Of these unigenes, 23,267 were annotated in 56 gene ontology (GO) categories, 6,033 were mapped to 131 KEGG pathways, and 7,825 were assigned to clusters of orthologous groups (COGs). From the unigenes, 8,165 EST-SSR loci were identified and 98.57% of them were 1–3 nucleotide repeats with 14.32%, 41.08% and 43.17% of mono-, di- and tri-nucleotide repeats, respectively. Fifty-eight types of motifs were identified with A/T, AG/CT, AT/AT, AC/GT, AAG/CTT and AGG/CCT the most abundant. The lengths of repeated nucleotide sequences in all SSR loci ranged from 12 to 60 bp, with most (88.51%) under 20 bp. Among 170 primer pairs were randomly selected from a total of 4,912 SSR primers we designed, 48 yielded unambiguously polymorphic bands with high reproducibility. Cluster analysis using 48 SSRs classified 34 flowering Chinese cabbage cultivars into three groups. A large number of EST-SSR markers identified in this study will facilitate marker-assisted selection in the breeding programs of flowering Chinese cabbage.

Introduction

Molecular markers are widely used in plant science for various applications. Due to quick advancements in molecular biology, the marker technologies rapidly evolved from restriction fragment length polymorphisms (RFLP) to simple sequence repeats (SSR) and single

EMBL/GenBank under the accession GFUS00000000. The version described in this paper is the first version, GFUS01000000.

Funding: This work was supported by the Guangdong Natural Science Foundation of China (http://www.gdstc.gov.cn) through grant 2015A030313500 to PG, by the Provincial Key International Cooperative Research Platform and the Major Scientific Research Project of Guangdong Higher Education (http://www.gdhed.edu.cn) through grant 2015KGJHZ015 to PG, and by the Science and Technology Plan of Guangdong of China (http://www.gdstc.gov.cn) through grant 2016B020201001 to RL. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

nucleotide polymorphisms (SNP). SSR markers are tandem repeats of 1–6 bp of nucleotides that are widespread in whole genomes of different species. SSR can be divided into genomic SSR and expressed-sequence-tag-based SSR (EST-SSR). SSR markers have many advantages over other markers including a high level of polymorphism, co-dominant inheritance, and high specificity and repeatability [1, 2]. They are widely used to analyze genetic diversity and biological evolution, construct genetic linkage maps, and map quantitative trait loci (QTLs) and for marker-assisted breeding [3]. SSR markers have been developed in many crop species and applied to genetic improvement of these species, including rice bean, maize, and sesame [4–6].

Flowering Chinese cabbage (*Brassica campestris* L. ssp. *chinensis* var. *utilis* Tsen et Lee) is an important vegetable crop in southern China. It is a popular vegetable crop for its rich nutritional value and favorable taste. Although flowering Chinese cabbage research has recently focused more on agronomic traits, physiological properties and genetic diversity [7], transcriptome analysis and development of molecular markers for breeding are still in its infancy [8]. Exploring new SSR markers in flowering Chinese cabbage will facilitate genetic and molecular studies of this species. The conventional method for SSR development involves constructing a genome or EST library, sequencing all clones in the library, and identifying SSRs. SSR primers are then designed according to the bilateral sequences of the identified SSR loci [2]. However, this method is labor intensive and produces relatively a few SSRs [6]. Recent developments in genome sequencing and *de novo* assembly provide opportunities to simultaneously develop a large number of SSR markers. RNA-sequencing (RNA-seq), using next generation sequencing technology, makes it possible to develop a large number of SSR markers in transcribed regions of a genome. Using RNA-seq to develop SSRs has advantages of relatively cost effective and direct use of expressed gene sequences [6]. Development of SSR markers with genetic information using RNA-seq and their application in diversity analysis have been reported in many species including Chinese cabbage, radish, carnation and moth orchid [9–12], but not in flowering Chinese cabbage. Although both flowering Chinese cabbage and Chinese cabbage belong to the *Brassica rapa* genome, their morphological traits and planting areas differ significantly. Breeders and researchers of flowering Chinese cabbage consider Chinese cabbage as an alien species in breeding program. The available SSR markers from *Brassica rapa* including Chinese cabbage can be transferred to flowering Chinese cabbage, but the efficiency of transferability is less than 37.6% [8]. In addition, potential SSRs identified in Chinese cabbage by Ding et al. [9] were not further detected and utilized. Therefore, developing SSR markers using RNA-seq data is needed in flowering Chinese cabbage.

As an important abiotic stress for flowering Chinese cabbage, heat stress has serious impacts on its quality and production. Based on experiments investigating genes responsive to heat stress at the seedling stage in flowering Chinese cabbage by using RNA-seq, transcriptome data of four genotypes under control and heat stress conditions were collected, and 8,165 SSR loci were identified. We designed primers for 4,912 SSRs, randomly selected 170 SSRs for validation in four flowering Chinese cabbage genotypes, and analyzed 48 polymorphic SSRs in a diversity panel of 34 cultivars.

Material and methods

Plant materials

Four flowering Chinese cabbage genotypes, Sijiu-19, Youlv 501, 3T-6 and Liuye 50, provided by Guangzhou Academy of Agricultural Sciences, Guangzhou, China, were sown in pots and grown in a growth chamber with temperatures set at 28/22˚C for 14/10 h (day/night). Each genotype had six pots with 3 seedlings per pot. At five-leaf stage, three pots per genotype were

moved to a different growth chamber for 6 h heat stress treatment at 38˚C/29˚C (14/10 h). After heat stress, the first fully expanded leaves of selected plants were harvested, together with the plants at the same stage in the control groups, snap frozen in liquid nitrogen, and stored at –80˚C for RNA extraction. Young leaves of all the four genotypes were also harvested for DNA isolation.

DNA and RNA isolation

Genomic DNA was extracted using a modified CTAB method [13] and total RNA was isolated using a Trizol RNA extraction kit (Invitrogen, USA) [14]. The concentration and quality of extracted RNA and DNA were evaluated using a microplate spectrophotometer (BioTek Company, USA) and 1% agarose gel electrophoresis, respectively. DNA was diluted to 10 ng·μl^{-1} for SSR analysis.

Sequence analysis

The total RNA with equal quantity was pooled from each sample, the mRNA was purified from the pooled RNA using Oligo dT attached magnetic beads, fragmented, and used to synthesize the cDNA. Double-stranded cDNA was purified and end-repaired, then single nucleotide adenine and the sequence adaptors were added to the cDNA fragments for ligation. The target fragments were PCR amplified for construction of cDNA libraries. The cDNA libraries were sequenced using an Illumina HiSeq2000 platform (Annoroad Gene Company, Beijing, China).

Raw data was filtered by removing the adaptors, low quality reads (Phred quality score $\leq$ 19) and reads with unknown bases. De novo assembly of full-length transcripts was performed using Trinity as decribed by Grabherr et al. [15] with three steps: 1) contigs were assembled by Inchworm software; 2) Bruijn graphs were built using Chrysalis software; 3) alternatively spliced transcripts were obtained and paralogous transcripts were separated through Butterfly software.

Functional annotation of unigenes

To annotate unigenes, putative protein sequences were predicted from all unigenes using TransDecoder tools (http://transdecoder.sourceforge.net/). The functions of putative protein sequences were annotated using Trinotate tools (http://trinotate.github.io/) against NCBI non-redundant (Nr) (http://www.ncbi.nlm.nih.gov/genbank/), Swiss-Prot (http://www.uniprot.org/), and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (http://www.genome.jp/kegg/pathway.html) databases, using the E value $\leq 1e^{-5}$. GO (Gene Ontology) functional classifications (http://www.geneontology.org/) including molecular function, biological process, and cellular component clusters were obtained by using Blast2GO [16]. Furthermore, all unigenes were aligned to the COG (clusters of orthologous groups) (http://www.ncbi.nlm.nih.gov/COG) to predict their possible functions.

SSR primer design and SSR marker screening

SSRs from unigenes were identified using the MIcroSAtellite identification tool (MISA, http://pgrc.ipk-gatersleben.de/misa/) [17]. Repeat sequence motifs included one to six nucleotides with repetitions of 12, 6, 5, 5, 5 and 5.

SSR primers were designed based on the bilateral sequence of SSR using BatchPrimer3 v1.0 software at http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi. Lengths of the primers ranged from 18–25 bp with most around 20 bp, and melting temperatures (Tm)

ranged from 55–60˚C. PCR products ranged from 100–300 bp. The percentages of GC content in the designed primers ranged from 25–65% with most around 50%. Other parameters were set as default. All primers were synthesized by the Sangon Biotech Company (Shanghai, China).

Validation and application of SSR markers

Four genotypes were used to validate designed SSR primers. The PCR reactions were carried out in a 10μl volume containing 20 ng of template DNA, 1×PCR buffer, 2.0 mM of $MgCl_2$, 0.25 μM of each forward and reverse primer, 0.25 mM of dNTPs, 0.75 u of *Taq* DNA polymerase. PCR amplification was performed using an ABI9700 PCR cycler (Applied Biosystems, USA) using the following program: 94˚C pre-denaturation for 5 min, 35 cycles of 94˚C denaturation for 45 s, varied annealing temperatures ranging from 55–60˚C with different primers for 45 s, 72˚C extension for 1 min, and a final step at 72˚C for 10 min. The PCR products were visualized in a 6% polyacrylamide gel after silver staining [18].

Thirty-four flowering Chinese cabbage cultivars (S1 Table) were genotyped using 48 validated SSRs. Polymorphic information content (PIC) was calculated using PIC_Calc 0.6 software [19], and observed heterozygosities (*Ho*) and expected heterozygosities (*He*) were obtained by using the POPGENE 1.32 software [20]. Genetic similarity of these accessions was analyzed using NTsys 2.10e software [21]. Based on the data of genetic similarity coefficients, a dendrogram tree was constructed using the unweighted pair-groups from the UPGMA module. The reliability of dendrogram was tested through bootstrap analysis based on 1000 trials using Phylip v3.69 [22].

Results

Sequencing data and *de novo* assembly

A total of 2.95×10^8 clean reads with 36.9 Gb sequence were generated from eight samples. The sequence data were deposited into the NCBI SRA database under the accession number SRP114731 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP114731). A total of 48,975 unigenes with 38.2 Mb sequence were obtained with an average length of 779 bp per unigene. N50 and N90 were 1,317 and 301 bp, respectively, which represent the sequence length of the smallest contig in the set that contains the fewest (largest) contig whose combined length produces at least 50% or 90% of the assembly, respectively [23]. Near half of the unigenes (22,677 or 46.30%) are 201–400 bp, 6,699 unigenes (13.68%) are 401 to 600 bp, and 19,599 unigenes (40.02%) are over 600 bp with 12,733 of them over 1,000 bp (Fig 1). The number of unigenes increased with the decrease in the sequence lengths of unigenes. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GFUS00000000. The version described in this paper is the first version, GFUS01000000.

Unigene annotation

Of the 48,975 unigenes, 37,494 (76.56%) and 24,524 (50.07%) showed significant similarity to proteins in the Nr and Swiss-Prot databases, respectively, and 24,406 (49.83%) were found in both databases. However, 6,864 (14.02%) couldn't be located in any existing databases (S2 Table).

Ground on Nr annotation, 23,267 (47.51%) unigenes showed functions in biological process, cellular component and molecular function, which can be further classified into 56 functional groups in Gene Ontology (GO) (S1 Fig). The seven major GO groups were cell (14,977, 64.37%), cell part (14,977, 64.37%), binding (11,222, 48.23%), cellular process (10,747,

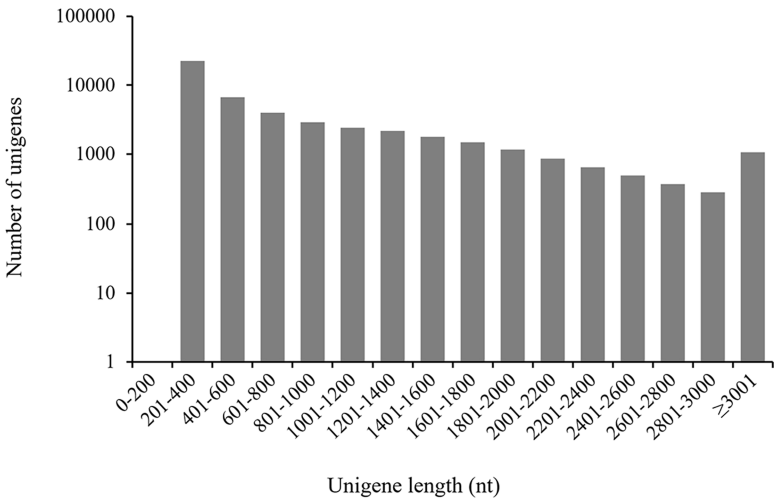

Fig 1. Length distribution of assembled unigenes from the transcriptome of flowering Chinese cabbage.

https://doi.org/10.1371/journal.pone.0184736.g001

46.19%), organelle (9,952, 42.77%), single-organism process (8,455, 36.34%) and metabolic process (7,762, 33.36%). The smallest groups, with only one unigene assigned, included biological phase, detoxification, virion part and synapse part.

In addition, functional prediction and classification of all unigenes were performed against the COG database (Fig 2). Of the 48,975 unigenes, 7,825 (15.98%) were classified into 24 molecular functional families. Among them the five predominant families were 'general function prediction only' (1,427, 18.24%), 'post-translational modification, protein turnover, chaperones' (838, 10.71%), 'replication, recombination and repair' (773, 9.88%), 'transcription' (748, 9.56%) and 'translation, and ribosomal structure and biogenesis' (688, 8.79%). The

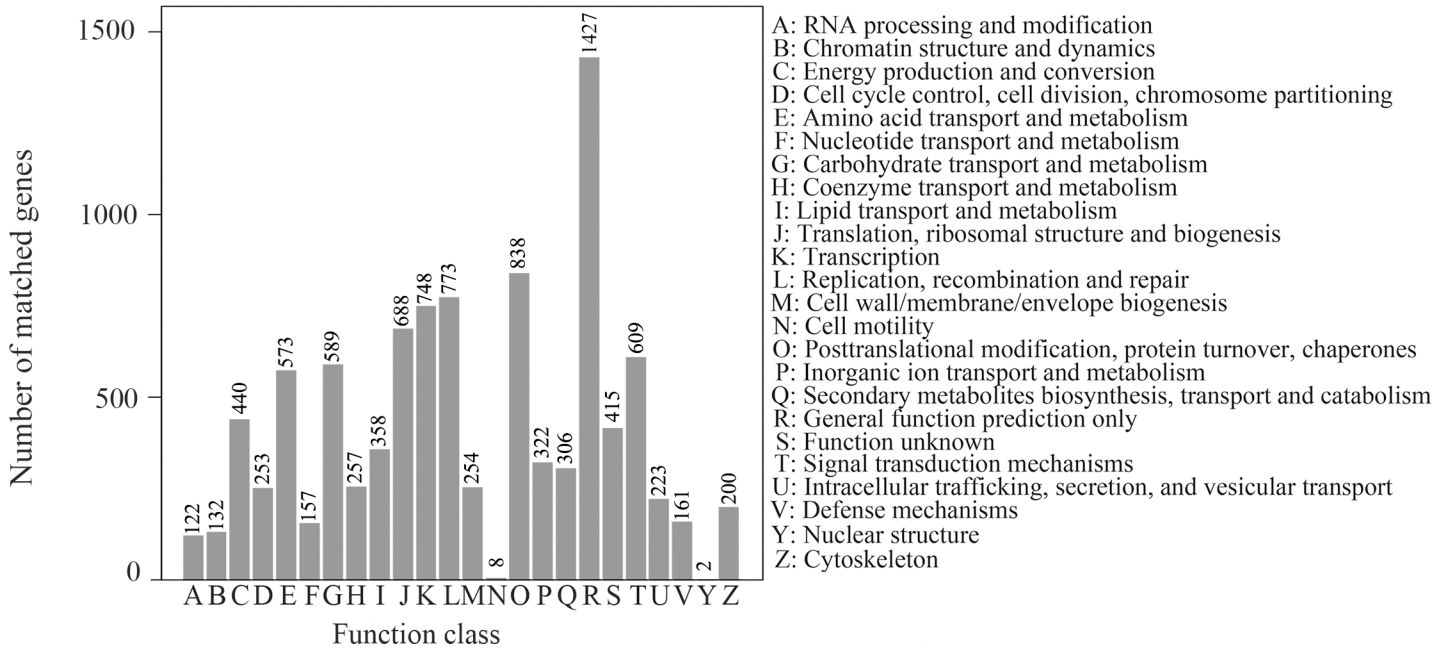

Fig 2. Classification of the assembled unigenes by clusters of orthologous groups (COG) analysis.

https://doi.org/10.1371/journal.pone.0184736.g002

smallest families, with only eight and two unigenes assigned, were 'cell motility' and 'nuclear structure', respectively. The proportions of other families ranged from 1.6% to 7.8%.

The identified 48,975 unigenes were further analyzed in the KEGG database, and 6,033 (12.32%) of them had significant matches in 5 categories including metabolism (3,013, 49.94%), genetic information processing (2,141, 35.49%), environmental information processing (538, 8.92%), cellular processes (499, 8.27%) and organismal systems (359, 5.95%). These categories related to 131 KEGG pathways with 1 to 402 unigenes that were involved in an individual pathway (S3 Table). Of these KEGG pathways, plant hormone signal transduction (402, 6.66%) had the highest unigene hits, followed by biosynthesis of amino acids (367, 6.08%), carbon metabolism (357, 5.92%) and ribosome (306, 5.07%). All other pathways had fewer than 300 unigenes involved (Table 1).

Frequency and distribution of SSR loci in flowering Chinese cabbage transcriptome

Screening the 48,975 unigenes, 8,165 SSR loci (distribution frequency of 16.68%) were discovered from 6,778 unigenes (S4 Table). Among these unigenes, 1,149 contained more than 1 SSR locus, and 413 SSRs presented in compound formation. In an average, every 4.67 kb sequences have an SSR locus in the flowering Chinese cabbage transcriptome.

Among the six types of SSR loci identified, trinucleotide repeats were the most frequent (7.2%), followed by dinucleotide (6.85%), whereas tetra, penta and hexanucleotide repeats were the least frequent (Table 2). The mean distances of the mono to hexa SSR loci types across the 38.17 Mb of unigene sequences were negatively correlated with their distribution frequency. Hexa and mononucleotide repeats had the longest (3,816.93 kb) and the shortest (10.83 kb) mean distances, respectively. The discovered SSR repeat sequence length per locus was 15.47 bp, with hexa SSRs being the longest (37.2 bp) and mononucleotide the shortest (13.93bp).

Table 1. The 20 KEGG pathways that have the highest unigene numbers.

Number	Name of pathway	No. of unigenes	Pathway ID
1	Plant hormone signal transduction	402 (6.66%)	ko04075
2	Biosynthesis of amino acids	367 (6.08%)	ko01230
3	Carbon metabolism	357 (5.92%)	ko01200
4	Ribosome	306(5.07%)	ko03010
5	Plant-pathogen interaction	296 (4.91%)	ko04626
6	Protein processing in endoplasmic reticulum	278 (4.61%)	ko04141
7	Spliceosome	251 (4.16%)	ko03040
8	Starch and sucrose metabolism	247 (4.09%)	ko00500
9	Ubiquitin mediated proteolysis	240 (3.98%)	ko04120
10	Purine metabolism	209 (3.46%)	ko00230
11	Endocytosis	206 (3.41%)	ko04144
12	RNA degradation	186 (3.08%)	ko03018
13	Amino sugar and nucleotide sugar metabolism	184 (3.05%)	ko00520
14	RNA transport	179 (2.97%)	ko03013
15	mRNA surveillance pathway	167 (2.77%)	ko03015
16	Oxidative phosphorylation	167 (2.77%)	ko00190
17	Pyrimidine metabolism	167 (2.77%)	ko00240
18	Phenylpropanoid biosynthesis	163 (2.70%)	ko00940
19	Cysteine and methionine metabolism	158 (2.62%)	ko00270
20	Peroxisome	151 (2.50%)	ko04146

https://doi.org/10.1371/journal.pone.0184736.t001

Table 2. Distribution of SSR loci among the transcriptome of flowering Chinese cabbage.

SSR type	Number	Distribution frequency (%)	Average distance (kb)	Mean length (bp)	Percentage (%)	Number of motifs
Mono-	1,169	2.39	32.65	13.93	14.32	2
Di-	3,354	6.85	11.38	14.72	41.08	4
Tri-	3,525	7.2	10.83	16.43	43.17	10
Tetra-	92	0.19	414.88	21.13	1.13	20
Penta-	15	0.03	2,544.62	26.33	0.18	12
Hexa-	10	0.02	3,816.93	37.2	0.12	10
Total	8,165	16.68	6,831.29	129.74	100	58

https://doi.org/10.1371/journal.pone.0184736.t002

Among the 8,165 SSR loci identified, the percentage of trinucleotide repeats was the highest (43.17%), followed by dinucleotide repeats (41.08%), mononucleotide repeats (14.32%), tetranucleotide repeats (1.13%), pentanucleotide repeats (0.18%) and hexanucleotide repeats (0.12%) (Table 2). The six SSR types consisted of 58 motifs of repeat sequences (S5 Table) including 20 tetranucleotide, 12 pentanucleotide, 10 trinucleotide, 10 hexanucleotide, 4 dinucleotide and two mononucleotide repeats. The most frequent motif type in the SSR loci was AG/CT (32.09%), followed by AAG/CTT (14.78%) and A/T (14.12%) (Fig 3). The remaining motifs were all less than 10%.

SSR validation

The Fig 4 showed the length distribution of 8,165 SSR loci identified from the transcriptome of flowering Chinese cabbage. These SSR loci could be divided into three groups based on the length of repeat motifs. First group, which was the most abundant and comprised 5,265 SSR loci (64.48%), had repeat motif lengths $\leq$ 15 bp; second group with 1,962 (24.03%) SSR loci had repeat motif lengths between 16 and 19 bp; and third group, which was the least with 938 (11.49%) SSR loci, had repeat motif lengths $\geq$ 20 bp.

A total of 4,912 primer pairs were designed using BatchPrimer3 software, and 170 (S6 Table) were randomly selected from these loci with repeat motif lengths $\geq$ 20 bp for PCR using the four flowering Chinese cabbage accessions (Fig 5).

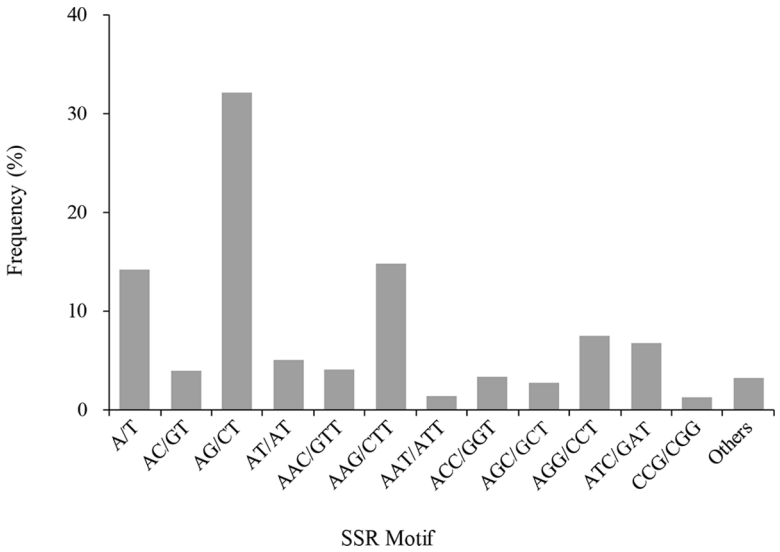

Fig 3. Distribution of SSR motifs in the flowering Chinese cabbage transcriptome.

https://doi.org/10.1371/journal.pone.0184736.g003

Ninety-six of the 170 tested primer pairs produced clear PCR products, with an amplification efficiency rate of 56.47%. Of these, 41 produced PCR products of the expected fragment size, 22 pairs were shorter and 33 pairs were longer than expected sizes. Forty-eight of the 96 primer pairs that produced PCR showed polymorphism among the four flowering Chinese cabbage accessions (S7 Table). Ten randomly selected PCR products were sequenced, and results showed that all of them contained the corresponding SSR loci.

Diversity analysis using the validated SSRs

Thirty-four accessions of flowering Chinese cabbage were genotyped using 48 polymorphic SSRs. A total of 213 alleles were scored, and the number of alleles for SSR loci ranged from 2 to 9 with an average of 4.44. The PIC value of these 48 SSRs ranged from 0.19 to 0.85, with an average value of 0.56.

Genetic similarities of the 34 accessions ranged from 0.46 to 0.90 with an average of 0.62. The largest genetic similarity coefficient (0.90) was observed between "2011–15" and "18264" (Fig 6); whereas the least similarity (0.46) occurred between "No.12" and "802–1". The 34 accessions could be divided into three clusters with 10, 14 and 10 accessions in cluster 1, 2 and 3, respectively. Cluster 3 was supported by the highest bootstrap value of 85%.

Discussion

Transcriptome assembly and functional annotation

De novo assembly of short reads of transcriptome without using a reference genome has been widely applied in sequence analysis of non-model organisms recently [24, 25]. In the present study, RNA-seq was performed in flowering Chinese cabbage, a non-model vegetable crop with limited transcriptome data in public databases. We used N50 value and mean length of unigenes as key indicators to evaluate the quality of the assembly [24] and obtained 1,317 bp

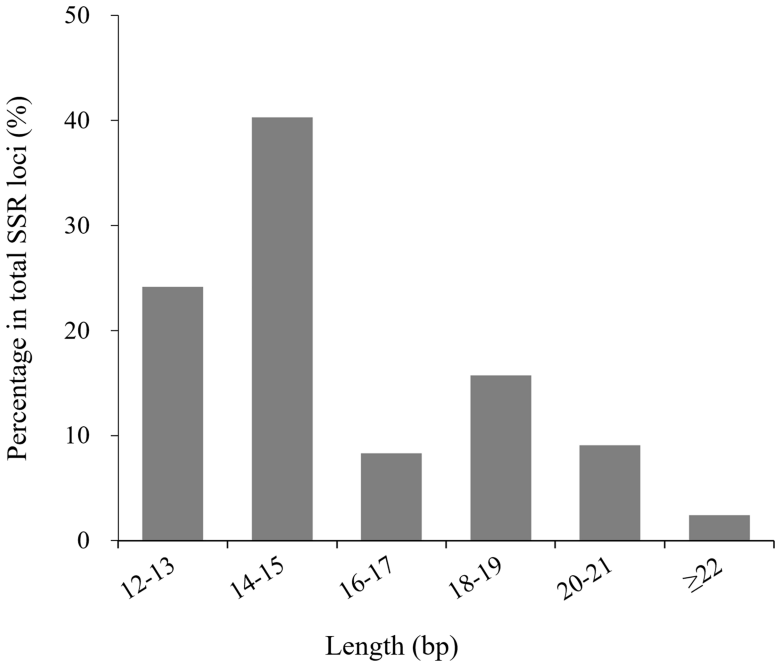

Fig 4. Distribution of the repeat sequence lengths of SSR loci among the transcriptome of flowering Chinese cabbage.

https://doi.org/10.1371/journal.pone.0184736.g004

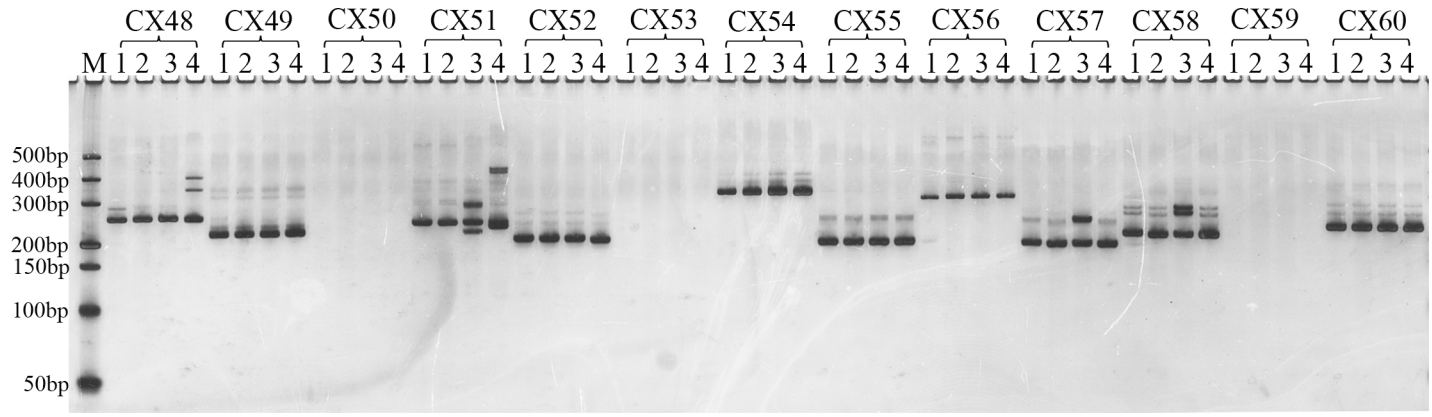

Fig 5. Partial image of the PCR amplification products of 170 selected SSRs in four flowering Chinese cabbage genotypes. Lane M is DNA size marker; lanes 1–4 refer to the four accessions Sijiu-19 caixin, Youlv 50, 3T6 and Liuye 50, respectively; CX48 to CX60 represent SSR primers.

https://doi.org/10.1371/journal.pone.0184736.g005

and 779 bp for these indicators, respectively, in the flowering Chinese cabbage. These values were higher than those in the previous reports for *Lilium Oriental* (N50 = 988 bp and mean length = 673 bp), *Houttuynia cordata* (N50 = 1,051 bp and mean length = 679 bp), and red clover (N50 = 933 bp and mean length = 622 bp) [24, 26, 27], suggesting that our *de novo* transcriptome assembly in flowering Chinese cabbage was effective and accurate.

Of the 48,975 assembled unigenes in flowering Chinese cabbage, 19,599 (40.02%) were longer than 600 bp and 12,733 (26.0%) were longer than 1,000 bp. A total of 37,494 (76.56%) were annotated in Nr database, which was more than those matched for alfalfa (71.8%), *Houttuynia cordata* (62.52%), Tibetan *Sophora moorcroftiana* (47.12%) and mung bean (53.0%) [24, 28, 29, 30]. Unigenes with longer sequences and a higher annotation rate were the indicators of high quality assembly [31, 32] that effectively captured a large portion of the transcriptome.

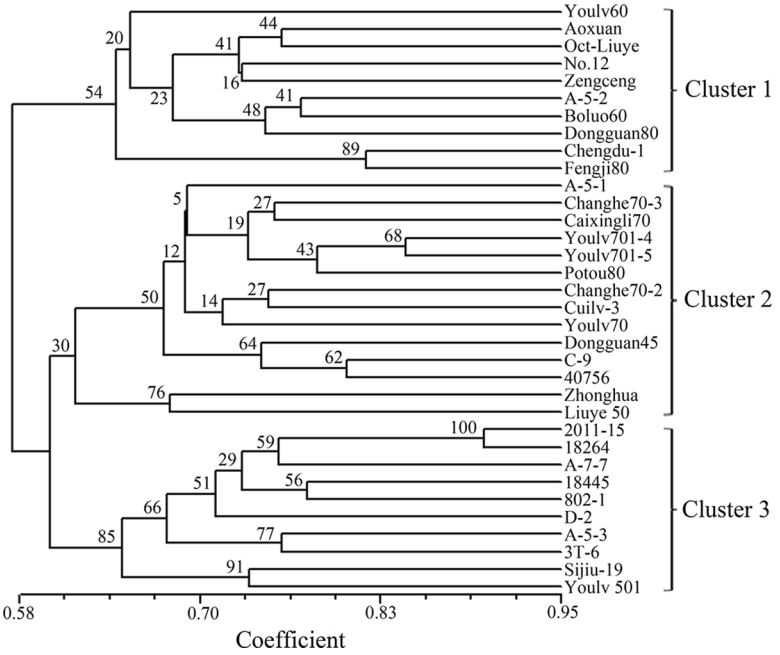

Fig 6. UPGMA dendrogram depicting genetic diversity estimated among 34 flowering Chinese cabbage accessions using 48 SSR markers. The numbers in the dendrogram are bootstrap values.

https://doi.org/10.1371/journal.pone.0184736.g006

SSR loci in the transcriptome of flowering Chinese cabbage

Flowering Chinese cabbage is a special *Brassica* vegetable cultivated mainly in southern China. Unlike other *Brassica* vegetables such as Chinese cabbage, only a few usable SSR markers have been reported in flowering Chinese cabbage. Rapid development of next-generation sequencing technology makes it possible for massive development SSR markers using transcriptome sequencing data [10, 33, 34].

In this study, we generated 48,975 unigenes from 295 million reads by *de novo* assembly. We detected 8,165 SSR loci that represent 16.68% of the generated unigenes. This frequency is higher than a previous report of radish (16.32%) [10], and lower than those in Chinese cabbage (20.16%) and radish (23.79%) [9, 35].

The mean SSR density in flowering Chinese cabbage was 213.9 SSR/Mb, which is higher than the previously reported for radish (202.84 SSR/Mb) [10], and lower than that in Chinese cabbage (255.43 SSR/Mb) [9]. The differences in distribution frequencies and densities of SSR loci may be due to the differences in genomes of different species, the quantity and length of unigenes in the transcriptome data, and the analytic tools used and SSR loci screening conditions [10, 36].

Of the six types of repeat sequence units in the current study, trinucleotide repeats had the highest frequency, followed by dinucleotides and mononucleotides, which agrees with cotton, Chinese cabbage, non-heading Chinese cabbage, and radish [9, 35, 37, 38]. Of the 58 SSR motifs identified in the SSR loci, A/T, AG/CT and AAG/CTT were the richest motifs in mono-, di-, and tri-nucleotide repeat sequence units, respectively, and also the three most abundant motifs in SSR, which agrees with those in Chinese cabbage, non-heading Chinese cabbage and radish [9, 35, 38].

Validation of EST-SSRs in flowering Chinese cabbage

The levels of polymorphic SSR are reported to vary with the lengths of repeat motifs of SSRs, and the longer the repeat motifs, the higher the possibility of polymorphisms [39, 40]. Therefore we assume that SSRs with repeat motif length $\geq$ 16 bp are more likely to have polymorphisms. Of 676 designed primers with the repeat motif length $\geq$ 16 bp, 170 were randomly selected for validation in four flowering Chinese cabbage genotypes. Ninety-six (56.47%) primer pairs produced unambiguous amplicons. The successful amplification rate is higher than those reported in rice bean (21.3%) and cotton (47.7%) [4, 37], but lower than those reported in Chinese cabbage (79.2%) and radish (83.5%) [9, 10]. The possible reasons for low amplification rate might because of at least one primer across a splice site, or amplicons with a large intron or chimeric cDNA contigs [41].

Among the 170 randomly tested primer pairs, 48 (28.24%) were polymorphic in the four flowering Chinese cabbage genotypes, which was slightly lower than those in mung bean (33%) with 31 accessions [30], peanut (40.63%) with six accessions [42], and Chinese cabbage (70.8%) with 24 accessions [9]. However, the polymorphism rate might be affected by numbers of accessions and their genetic backgrounds. For example, Zhang et al. [6] determined the level of polymorphisms for the EST-SSRs developed from sesame transcriptomic data, and found polymorphisms in 32 (11.59%) of the 300 EST-SSRs tested in a panel of 24 cultivated accessions, but the numbers of polymorphic EST-SSRs reached 167 (60.51%) when a wild accession was added to the panel. Thus, the polymorphic level of SSRs in the current study might be higher if the accession number was increased or more diversified germplasm lines were added.

Genetic diversity in the 34 accessions of flowering Chinese cabbage

All 48 polymorphic SSRs were used to genotype 34 cultivated accessions. A total of 213 allelic variations were detected with an average of 4.44 alleles per SSR. The PIC values of these SSRs ranged from 0.19 to 0.85 with an average of 0.56, which is higher than that reported for mung

bean (0.34) [30], but lower than those reported for *Houttuynia cordata* (0.72) and *Hemarthria* (0.71) [24, 43]. A UPGMA dendrogram based on genetic similarity coefficients separated 34 accessions into three distinct clusters, which did not match with their origin or maturity category, but agrees with the previous reports [7, 44] that the genetic basis of the origin and maturity categories were mixed in these cultivated accessions. This may be resulted from a narrow genetic background or limited breeding materials/resources used by breeders as observed in sesame [6]. Therefore, more diverse accessions and alien species should be added to breeding programs to expand the genetic backgrounds of breeding parents of flowering Chinese cabbage. In addition, 6 of the 34 accessions used in the current study (Aoxuan, No.12, Sijiu-19, Youlv 501, Oct-Liuye and Youlv60) were part of a genetic diversity panel used by Guo et al. [7]. They used nine microsatellite-anchored fragment length polymorphism (MFLP) markers to cluster 32 accessions into two groups, with Aoxuan, No.12, Sijiu-19 and Youlv 501 in one group, and Oct-Liuye and Youlv60 in the other. In the current study, Aoxuan, No.12, Oct-Liuye and Youlv60 were clustered in the same group, while Sijiu-19 and Youlv 501 were clustered in another group, indicating that different molecular markers might influence estimates of genetic similarity in flowering Chinese cabbage, which agrees with a previous report [45].

In conclusion, our sequence analysis of transcriptome identified 8,165 EST-SSR loci with 98.57% of 1–3 nucleotide repeats. We designed 4,912 SSR primer pairs and identified 48 polymorphic amplicons in four genotypes from 170 randomly selected primer combinations. Genetic diversity analysis using the 48 markers classified 34 cultivated accessions of flowering Chinese cabbage into three groups. The newly developed EST-SSR markers will be a valuable resource for genetic diversity analysis, QTL mapping and marker-assisted breeding in flowering Chinese cabbage.

Supporting information

S1 Fig. Gene ontology classification of the assembled unigenes.
(TIF)

S1 Table. Information on the 34 cultivated accessions of flowering Chinese cabbage used for diversity analysis.
(DOC)

S2 Table. Annotation of all the unigenes.
(XLS)

S3 Table. KEGG pathways for 6033 unigenes.
(DOC)

S4 Table. Summary of the SSR investigation.
(DOC)

S5 Table. Fifty-eight motifs of six types of SSRs and their repeat numbers in the transcriptome of flowering Chinese cabbage.
(DOC)

S6 Table. 170 SSR primer pairs for testing availability in four flowering Chinese cabbage genotypes.
(DOC)

S7 Table. Characterization of 48 polymorphic EST-SSRs in the 34 accessions of flowering Chinese cabbage.
(DOC)

Author Contributions

Conceptualization: Ronghua Li, Peiguo Guo.

Data curation: Jingfang Chen, Ronghua Li, Yanshi Xia, Peiguo Guo.

Formal analysis: Jingfang Chen, Ronghua Li, Peiguo Guo.

Funding acquisition: Ronghua Li, Peiguo Guo.

Investigation: Jingfang Chen, Ronghua Li, Yanshi Xia, Peiguo Guo, Zhiliang Wang, Hua Zhang.

Methodology: Yanshi Xia, Peiguo Guo.

Project administration: Ronghua Li, Yanshi Xia, Peiguo Guo.

Resources: Jingfang Chen, Ronghua Li, Yanshi Xia, Zhiliang Wang.

Software: Jingfang Chen, Ronghua Li.

Supervision: Ronghua Li, Yanshi Xia, Peiguo Guo.

Validation: Jingfang Chen, Ronghua Li, Zhiliang Wang.

Visualization: Jingfang Chen, Ronghua Li, Guihua Bai, Peiguo Guo, Kadambot H. M. Siddique.

Writing – original draft: Jingfang Chen, Guihua Bai, Peiguo Guo, Kadambot H. M. Siddique.

Writing – review & editing: Jingfang Chen, Ronghua Li, Guihua Bai, Peiguo Guo, Kadambot H. M. Siddique.

References

1. Morgante M, Olivieri A. PCR-amplified microsatellites as markers in plant genetics. Plant Journal. 1993; 13(1): 175–182.
2. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Research. 2000; 10(7): 967–981. PMID: 10899146
3. Tuler A, Carrijo T, Nóia L, Ferreira A, Peixoto A, Silva F. SSR markers: a tool for species identification in *Psidium* (*Myrtaceae*). Molecular Biology Report. 2015; 42(11):1501–1513.
4. Chen H, Chen X, Tian J, Yang Y, Liu Z, Hao X, et al. Development of gene-based SSR markers in rice bean (*Vigna umbellata* L.) based on transcriptome data. PLoS ONE. 2016; 11(3): e0151040. https://doi.org/10.1371/journal.pone.0151040 PMID: 26950544
5. Galvão K, Ramos HC, Santos P, Entringer G, Vettorazzi J, Pereira M. Functional molecular markers (EST-SSR) in the full-sib reciprocal recurrent selection program of maize (*Zea mays* L.). Genetics and Molecular Research. 2015; 14(3): 7344–7355. https://doi.org/10.4238/2015.July.3.10 PMID: 26214413
6. Zhang H, Wei L, Miao H, Zhang T, Wang C. Development and validation of genic-SSR markers in sesame by RNA-seq. BMC Genomics. 2012; 13: 316. https://doi.org/10.1186/1471-2164-13-316 PMID: 22800194
7. Guo P, Xu L, Xia Y, Huang H, Zhang H, Zheng Y, et al. Genetic diversity analysis for germplasm of flowering Chinese cabbage by using fluorescent microsatellite-anchored fragment length polymorphism. Acta Horticulturae Sinica. 2015; 42(2): 350–360.
8. Chen J, Li R, Wang Z, Xia Y, Guo P, Siddique K. Transferability and application of *Brassica* SSR markers in flowering Chinese cabbage. Northern Horticulture. 2016; 15:86–91.
9. Ding Q, Li J, Wang F, Zhang Y, Li H, Zhang J, et al. Characterization and development of EST-SSRs by deep transcriptome sequencing in Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). International Journal of Genomics. 2015; 2015: 473028. https://doi.org/10.1155/2015/473028 PMID: 26504770
10. Zhai L, Xu L, Wang Y, Cheng H, Chen Y, Gong Y, et al. Novel and useful genic-SSR markers from de novo transcriptome sequencing of radish (*Raphanus sativus* L.). Molecular Breeding. 2014; 33(3):611–624.

11. Yagi M, Yamamoto T, Isobe S, Hirakawa H, Tabata S, Tanase K, et al. Construction of a reference genetic linkage map for carnation (*Dianthus caryophyllus* L.). BMC Genomics. 2013; 14:734. https://doi.org/10.1186/1471-2164-14-734 PMID: 24160306
12. Tsai C, Shih H, Wang H, Lin Y, Chang C, Chiang Y, et al. RNA-seq SSRs of moth orchid and screening for molecular markers across genus *Phalaenopsis* (*Orchidaceae*). PLoS ONE. 2015; 10(11): e0141761. https://doi.org/10.1371/journal.pone.0141761 PMID: 26523377
13. Li R, Xia Y, Liu S, Sun L, Guo P, Miao S, et al. CTAB-improved method of DNA extraction in plant. Research and Exploration in Laboratory. 2009; 28(9): 14–16.
14. Guo P, Baum M, Grando S, Ceccarelli S, Bai G, Li R, et al. Differentially expressed genes between drought-tolerant and drought-sensitive barley genotypes in response to drought stress during the reproductive stage. Journal of Experimental Botany. 2009; 60(12): 3531–3544. https://doi.org/10.1093/jxb/erp194 PMID: 19561048
15. Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011; 29(7), 644–652. https://doi.org/10.1038/nbt.1883 PMID: 21572440
16. Conesa A, Götz S, Garcíagómez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005; 21(18): 3674–3676. https://doi.org/10.1093/bioinformatics/bti610 PMID: 16081474
17. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theoretical and Applied Genetics. 2003; 106: 411–422. https://doi.org/10.1007/s00122-002-1031-0 PMID: 12589540
18. Liu W, Li R, Ayalew H, Xia Y, Bai G, Yan G, et al. Development of a simple and effective silver staining protocol for detection of DNA fragments. Electrophoresis. 2017; 38(8), 1175–1178. https://doi.org/10.1002/elps.201700009 PMID: 28145034
19. Nagy S, Poczai P, Cernák I, Gorji A, Hegedűs G, Taller J. PICcalc: an online program to calculate polymorphic information content for molecular genetic studies. Biochemical Genetics. 2012; 50(9–10), 670–672. https://doi.org/10.1007/s10528-012-9509-1 PMID: 22573137
20. Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S. PopGen: populationbased recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genetics. 2006; 9(55–61)
21. Rohlf F. Multivariate Analysis System, Version 2.10e. Applied Biostatistics. 2000; Inc, New York.
22. Felsenstein J. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 1989; 5:164–166.
23. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010; 95(6): 315–327. https://doi.org/10.1016/j.ygeno.2010.03.001 PMID: 20211242
24. Wei L, Li S, Liu S, He A, Wang D, Wang J, et al. Transcriptome analysis of *Houttuynia cordata* Thunb. by Illumina paired-end RNA sequencing and SSR marker discovery. PLoS ONE. 2014; 9(1): e84105. https://doi.org/10.1371/journal.pone.0084105 PMID: 24392108
25. Russell J, Hackett C, Hedley P, Liu H, Milne L, Bayer M, et al. The use of genotyping by sequencing in blackcurrant (*Ribes nigrum*): developing high-resolution linkage maps in species without reference genome sequences. Molecular Breeding. 2014; 33(4): 835–849.
26. Du F, Wu Y, Zhang L, Li X, Zhao X, Wang W, et al. De novo assembled transcriptome analysis and SSR marker development of a mixture of six tissues from *Lilium* Oriental hybrid 'Sorbonne'. Plant Molecular Biology Reporter. 2015; 33(2): 281–293.
27. Yates S, Swain M, Hegarty M, Chernukin I, Lowe M, Allison G, et al. *De novo* assembly of red clover transcriptome based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. BMC Genomics. 2014; 15(1): 453.
28. Liu Z, Chen T, Ma L, Zhao Z, Zhao P, Nan Z, et al. Global transcriptome sequencing using the Illumina platform and the development of EST-SSR markers in autotetraploid alfalfa. PLoS ONE. 2013; 8(12): e83549. https://doi.org/10.1371/journal.pone.0083549 PMID: 24349529
29. Li H, Yao W, Fu Y, Li S, Guo Q. *De novo* assembly and discovery of genes that are involved in drought tolerance in Tibetan *Sophora moorcroftiana*. PLoS ONE. 2015; 10(1): e111054. https://doi.org/10.1371/journal.pone.0111054 PMID: 25559297
30. Chen H, Wang L, Wang S, Liu C, Blair M, Cheng X. Transcriptome sequencing of mung bean (*Vigna radiata* L.) genes and the identification of EST-SSR markers. PLoS ONE. 2015; 10(4): e0120273. https://doi.org/10.1371/journal.pone.0120273 PMID: 25830701
31. Blavet N, Charif D, Oger-Desfeux C, Marais GA, Widmer A. Comparative high-throughput transcriptome sequencing and development of SiESTa, the *Silene* EST annotation database. BMC Genomics. 2011; 12:376. https://doi.org/10.1186/1471-2164-12-376 PMID: 21791039

32. Tian D, Pan X, Yu Y, Wang W, Zhang F, Ge Y, et al. De novo characterization of the *Anthurium* transcriptome and analysis of its digital gene expression under cold stress. BMC Genomics. 2013; 14(1): 827.

33. Dutta S, Kumawat G, Singh B, Gupta D, Singh S, Dogra V, et al. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. BMC Plant Biology. 2011; 11:17. https://doi.org/10.1186/1471-2229-11-17 PMID: 21251263

34. Wei W, Qi X, Wang L, Zhang Y, Hua W, Li D, et al. Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. BMC Genomics. 2011; 12:451. https://doi.org/10.1186/1471-2164-12-451 PMID: 21929789

35. Wang S, Wang X, He Q, Liu X, Xu W, Li L, et al. Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. Plant Cell Reports. 2012; 31(8): 1437–1447. https://doi.org/10.1007/s00299-012-1259-3 PMID: 22476438

36. Biswas M, Chai L, Mayer C, Xu Q, Guo W, Deng X. Exploiting BAC-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in Citrus. Molecular Biology Report. 2012; 39: 5373–5386.

37. Zhang X, Ye Z, Wang T, Xiong H, Yuan X, Zhang Z, et al. Characterization of the global transcriptome for cotton (*Gossypium hirsutum* L.) anther and development of SSR marker. Gene. 2014; 551(2): 206–213. https://doi.org/10.1016/j.gene.2014.08.058 PMID: 25178523

38. Song X, Ge T, Li Y, Hou X. Genome-wide identification of SSR and SNP markers from the non-heading Chinese cabbage for comparative genomic analyses. BMC Genomics. 2015; 216: 328.

39. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Research. 2001; 11(8): 1441–1452. https://doi.org/10.1101/gr.184001 PMID: 11483586

40. Singh H, Deshmukh RK, Singh A, Singh AK, Gaikwad K, Sharma TR, et al. Highly variable SSR markers suitable for rice genotyping using agarose gels. Molecular Breeding. 2010; 25(2): 359–364.

41. Varshney RK, Grosse I, Hahnel U, Siefken R, Prasad M, Stein N, et al. Genetic mapping and BAC assignment of EST-derived SSR markers shows nonuniform distribution of genes in the barley genome. Theoretical and Applied Genetics. 2006; 113: 239–250. https://doi.org/10.1007/s00122-006-0289-z PMID: 16791690

42. Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, et al. De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea* L.). BMC Genomics. 2012; 13(1): 90.

43. Huang X, Yan H, Zhang X, Zhang J, Frazier T, Huang D, et al. *De novo* transcriptome analysis and molecular marker development of two *Hemarthria* species. Frontiers in Plant Science. 2016; 7: 496. https://doi.org/10.3389/fpls.2016.00496 PMID: 27148320

44. Qiao Y, Huang H, Li G, Zhang H, Li Z, Zheng Y, et al. Genetic diversity of Chinese flowering cabbage (*Brassica campestris* L.) and relatives based on morphological phenotypes and SRAP. Molecular Plant Breeding. 2012; 10: 1369–1375.

45. Pejic I, Ajmone-Marsan P, Morgante M, Kozumplick V, Castiglioni P, Taramino G, et al. Comparative analysis of genetic similarity among maize inbred lines detected by RFLPs, RAPDs, SSRs, and AFLPs. Theoretical and Applied Genetics. 1998; 97: 1248–1255.