Biochemical Genetics (2019) 57:607–622
https://doi.org/10.1007/s10528-019-09912-6

ORIGINAL ARTICLE

De Novo Assembled Transcriptome Analysis and Identification of Genic SSR Markers in Red-Flowered Strawberry

Yan Ding[1] · Li Xue[1] · Rui-xue Guo[1] · Gang-jun Luo[1] · Yu-tong Song[1] · Jia-jun Lei[1]

Received: 30 June 2018 / Accepted: 19 February 2019 / Published online: 1 March 2019

Abstract

Red-flowered strawberry is a new ornamental flower derived from intergeneric hybridization (*Fragaria*×*Potentilla*). To date, few molecular markers have been reported for this plant. RNA sequencing provides a relatively fast and low-cost approach for large-scale detection of simple sequence repeats (SSRs). In the present study, we profiled the transcriptome of red-flowered strawberry by Illumina HiSeq 2500 to identify SSRs related to petal color. Based on 2 million clean reads of red and white flowers from red-flowered strawberry hybrids, we assembled 91,835 unigenes with an average length of 717 bp. After functional annotation and prediction, there were 47,058 unigenes; of these, 26,861 had a gene ontology annotation, with 14,264 SSR loci. Mononucleotide SSRs were the predominant repeat type (47.20%, $n=6724$), followed by di- (32.50%, $n=4641$), tri- (19.10%, $n=2729$), tetra- (0.90%, $n=132$), hexa- (0.2%, $n=21$), and penta- (0.10%, $n=16$) nucleotide repeats. The most frequent di-, tri-, and tetra-nucleotide repeats were AG/CT, AAG/CTT, and AAAG/CTTT, respectively. PCR amplification with 105 SSR primer pairs yielded four bands specific to red flowers, namely UgRFsr57622, UgRFsr94149, UgRFsr40142, and UgRFsr54608; corresponding 4 trait-specific markers were found to co-segregate with white and red flower color in hybrid population, demonstrating that the genic SSR marker is useful to discriminate between white and red flowers in strawberry. Markers to discriminate flower color in red-flowered strawberry will be useful for early selection of progeny and for breeding management.

Keywords Red-flowered strawberry · Transcriptome sequencing · Simple sequence repeat (SSR) · Flower color · Marker-assisted selection

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10528-019-09912-6) contains supplementary material, which is available to authorized users.

✉ Jia-jun Lei
jiajunleisy@163.com

[1] College of Horticulture, Shenyang Agricultural University, Shenyang 110866, China

Introduction

All cultivars and wild species of the genus *Fragaria* are white-flowered. Red-flowered strawberry, an intergeneric hybrid obtained several decades ago by crossing *F. ananassa* ($2n = 8x = 56$) and *Potentilla palustris* ($2n = 6 \times = 42$) has great market potential and economic value owing to its beautiful flower and edible fruit (Ellis 1962; Mabberley 2002; Bentvelsen and Bouw 2006).

Although molecular markers have been identified in many fruit trees, only sequence-characterized amplified regions derived from random amplified polymorphic DNA (RAPD) have been reported in red-flowered strawberry (Yan et al. 2006). Molecular markers are useful for screening offspring with target traits at seed and seedling stages, which can improve the reliability of early selection and shorten the breeding period (Bematzky et al. 1992). DNA-based molecular markers include restriction fragment length polymorphism (RFLP), random amplified polymorphism (RAPD), simple sequence repeat (SSR), amplified fragment length polymorphisms (AFLP), single nucleotide polymorphisms (SNPs), and cleaved amplified polymorphic sequences (CAPS), among others. SSR markers are widely used owing to their high degree of polymorphism and reproducibility (Miah et al. 2013; Cuadrado et al. 2008). The genetic map of *Citrus* has been constructed based on 67 de novo identified SSRs and 129 expressed sequence tag SSR markers (Chen et al. 2006). However, the traditional approach of SSR marker development is labor-intensive, costly, inefficient, and cannot meet the research needs for complex processes (Röder et al. 1998; Ujino et al. 1999; Uzunova and Ecke 1999).

Next-generation sequencing technologies allow the rapid and efficient development of SSR markers on a large scale from transcriptome data (Zalapa et al. 2012; Wang et al. 2014a). RNA sequencing (RNA-seq) technology based on next-generation high-throughput Illumina/Solexa sequencing can provide general information on the transcriptome with high speed and accuracy and at low cost. RNA-seq technology supports de novo sequencing for non-model species for which a reference sequence is lacking, and can be used to identify unknown and rare transcripts. It has been applied to many plants including black raspberry (Hyun et al. 2014), *Rosa chinensis* cv. 'Pallida' (Yan et al. 2014), apple (Bai et al. 2014), strawberry (Pillet et al. 2015), sweet cherry (Wei et al. 2015b), and sesame (Zhang et al. 2012), among others. Illumina-based RNA-seq has also facilitated the identification of 57 polymorphic SSR loci in lily (*Lilium* Oriental hybrid 'Sorbonne') (Du et al. 2015), and a new orange trait-specific SSR marker was developed by transcriptome analysis in orange head Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*) (Zhang et al. 2015).

In this study, we developed a comprehensive set of genic SSRs based on Illumina RNA-HiSeq 2500 sequencing combined with de novo annotation, and identified unitranscript sequences associated with a red flower color. A total of 14,264 genic SSRs from 12,077 unigenes were generated in the de novo transcriptome, and four trait-specific genic SSR markers that co-segregated with the red flower trait were identified that can be useful for red-flowered strawberry selection.

Materials and Methods

Plant Materials

Plant materials used in this experiment were obtained from the Strawberry Germplasm Repository of Shenyang Agricultural University in China. 'Pink Princess', 'Pink Beauty', and 'Pretty Beauty' were three red-flowered strawberry cultivars released by Shenyang Agricultural University with pink, pink, and dark pink flowers, respectively (Xue et al. 2014). While 'Honeoye' was a famous strawberry cultivar released by New York State Agricultural Experiment Station, Geneva, USA in 1979. In this paper, four hybridizations, Pink Princess × Pink Beauty, Pink Princess × Honeoye, Pretty Beauty × Honeoye, and Pink Princess × Pretty Beauty were carried out in 2012. The six white-flowered hybrids (W) and six red-flowered hybrids (R) from the cross of Pink Princess × Pink Beauty were used for transcriptome analysis and molecular marker development, respectively. The samples of 'W' and 'R' were used to construct four libraries, which were named as PH_R1, PH_R2, PH_W1 and PH_W2, respectively (Fig. 1). Petals were frozen in liquid nitrogen and stored at − 80° until RNA extraction; Hybrids of other crosses (Pink Princess × Honeoye, Pretty Beauty × Honeoye, and Pink Princess × Pretty Beauty) were used for molecular marker identification. Their fresh young leaves were collected and rapidly frozen in liquid nitrogen and stored at − 80°until DNA extraction.

RNA Extraction, cDNA Library Construction, and Sequencing

Total RNA was extracted using the modified cetyl trimethylammonium bromide (CTAB) method (Shan et al. 2008). RNA quality was analyzed by 1.0% agarose gel electrophoresis, and concentration and purity (optical density OD260/280 and OD260/230) were evaluated with a NanoDrop 2000c spectrophotometer (Thermo

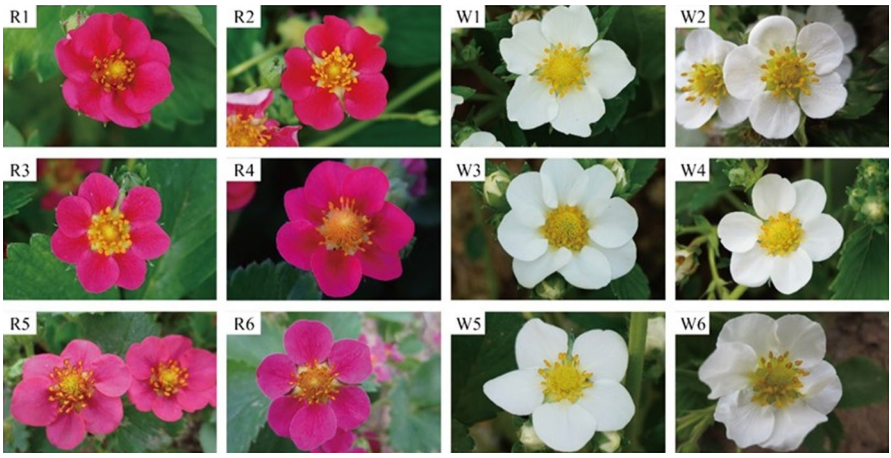

Fig. 1 Flowers of red- and white-flowered strawberry used in deep sequencing. Both R1-6 and W1-6 were selected from the cross of Pink Princess × Pink Beauty (Color figure online)

Fisher Scientific, Waltham, MA, USA). The RNA integrity was assessed with an Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA, USA).

A total amount of 3 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) following manufacturer's recommendations and index codes were added to attribute sequences to each sample. The library quality was assessed on the Agilent Bioanalyzer 2100 system. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. After clustering, the libraries were sequenced using an Illumina HiSeq 2500 platform and 125 bp/150 bp paired-end reads were generated. Four cDNA libraries, PH_R1, PH_R2, PH_W1, and PH_W2 were constructed with two biological replicates by Beijing Novogene Bioinformatics Technology Co., Ltd (Beijing, China).

De Novo Transcriptome Assembly and Functional Annotation

The raw data in FASTQ format were deposited in the National Center for Biotechnology Information (NCBI) database (accession no. SRP136399), and the reads were processed with in-house Perl scripts. Clean data were obtained from raw data by removing reads containing adapter or poly-*N* sequences as well as low-quality reads. Clean reads were assembled using Trinity software (http://trinitytyrnaseq.sourceforge.net) with the following parameters: K-mer = 25 and group pair distance = 300. The Trinity short read assembly program with default settings was used to reassemble the clean and high-quality transcriptome sequence data (Grabherr et al. 2011). The longest transcript of each gene was considered as a unigene, and a unigene general library of four samples was obtained. These transcripts were used for further bioinformatics analyses. The assembled sequences were functionally annotated based on NCBI protein (Nr) and nucleotide (Nt), Swiss-Prot, Protein Family (PFAM), Gene Ontology (GO), EuKaryotic Orthologous Groups (KOG), and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases using Basic Local Alignment Search Tool (BLAST)x (E-value $\leq 1e^{-5}$).

Identification of Novel Genic SSRs and Primer Design and Synthesis

Using the MISA microsatellite program (Thiel et al. 2003), we searched the screened unigenes (574 kb) from the transcriptome of red-flowered strawberry for microsatellite sites. The number of repetitions used to select SSRs was at least ten times for single nucleotides, six times for dinucleotides, and five times for tri-, tetra-, penta-, and hexanucleotides. Primers were synthesized by Shanghai GENEWIZ Biotechnology Co. (Shanghai, China) and are shown in Table S1.

DNA Isolation, PCR Profiling, and Screening

DNA was isolated from young leaves of red-flowered strawberry using the modified CTAB method (Liu et al. 2003). DNA quantity and quality were evaluated with a NanoDrop 2000c spectrophotometer. The DNA was normalized to a concentration of 20 ng/l and then stored at − 20° until use. Purified DNA was used for primary screening of all the primer pairs by polyacrylamide gel electrophoresis (PAGE).

PCR amplification was performed on a 96-well thermal cycler (Veriti 96-Well Thermal Cycler). The volume of the reaction was 20 µL. The PCR program was as follows: 1 min at 94°; 35 cycles of 1 min at 94°, 2 min at the appropriate annealing temperature, and 2 min at 72°; and 7 min at 72°. All primers were initially screened by PAGE. PCR products were separated on an 8% non-denaturing polyacrylamide gel at 160 V for 2–2.5 h and visualized by rapid silver staining (Wei et al. 2012).

Detection of Genic SSR Primers

Primer specificity was verified by 3% (w/v) agarose gel electrophoresis. The PCR product band of the expected size was recovered and directly cloned into the pGEM-T vector (Tiangen Biotech, Beijing, China). The sequencing results were aligned to the original sequence with BLAST.

Results

Red-Flowered Strawberry Transcriptome Sequence Assembly

High-throughput sequencing was performed on the Illumina HiSeq 2500 platform and the original image data were converted to raw reads by base calling. The number of raw reads of PH_R1, PH_R2, PH_W1, and PH_W2 were 62,225,658, 57,080,440, 53,740,688, and 51,381,864, respectively. After removing the impurities and adapters, 58,914,640, 54,309,114, 50,732,796, and 48,856,812 clean reads were obtained with a total base number of 26.58 G. The total nucleotide number of each sample was greater than 6 G. The Q20 ratio of each sample was greater than 95%, and GC content was approximately 46% (Table 1). Thus, the sequencing data met the quality requirement for subsequent analyses; transcript sequences assembled with Trinity

Table 1 Overview of sequencing and assembly of transcriptome

Samples	Raw reads	Clean reads	Clean bases	Error (%)	Q20 (%)	Q30 (%)	GC content (%)
PH_R1	62,225,658	58,914,640	7.36 G	0.04	95.57	91.36	46.11
PH_R2	57,080,440	54,309,114	6.78 G	0.04	95.91	91.94	46.21
PH_W1	53,740,688	50,732,796	6.34 G	0.04	95.87	91.87	46.03
PH_W2	51,381,864	48,856,812	6.10 G	0.04	95.86	91.84	45.89

Note Q20 and Q30 percentages are proportion of nucleotides with quality value larger than 20 and 30, respectively; GC percentage is proportion of guanine and cytosine nucleotides among total nucleotides

Table 2 Statistics of the assembled transcripts and unigenes

Variety	Total number	Mean length (bp)	N50 (bp)	500–1 Kbp	1K–2Kbp	> 2Kbp	Total nucleotides
Transcripts	139,292	909	353	29,636	24,591	16,526	126,608,737
Unigenes	91,835	717	282	17,680	10,330	7122	65,851,703

Table 3 Unigene function annotation

Annotation database	No. of unigenes	Percentage (%)
Annotated in NR	38,288	41.69
Annotated in NT	36,092	39.3
Annotated in KEGG	12,832	13.97
Annotated in SwissProt	26,032	28.34
Annotated in PFAM	26,247	28.58
Annotated in GO	26,861	29.24
Annotated in KOG	14,710	16.01
Annotated in all databases	6797	7.4
Annotated in at least one database	47,058	51.24
Total unigenes	91,835	100

were used as reference sequences. Based on the sequence information, the 139,292 transcripts were obtained with the mean length of 909 bp and N50 length of 353 bp, and 91,385 unigenes were obtained with the mean length of 717 bp and N50 length of 282 bp. There were 17,452 unigenes with length > 1kbp (Table 2). The assembly completeness was assessed by the BUSCO tool with the result of 90%, which indicated the high quality of transcriptome.

Functional Annotation of Unigenes

The unigenes were searched against the Nr, Nt, Swissprot, PFAM, KOG, KEGG, and GO databases. Moreover, 51.24% (47,058) of assembled unigenes had homologs in at least one of the databases. Based on the Nr database, 38,288 unigenes were annotated, and 25,410 unigenes had homologs in *F. vesca* with the highest frequency of 66.5%. A total of 47,058 (51.24%) red-flowered strawberry unigenes have been annotated in at least one of the above-mentioned databases (Table 3). Of these, 26,032 unigenes were annotated to the Swissprot database (28.34%). A total of 6797 unigenes were annotated to functional genes in all seven databases (7.4%).

The species distribution showed that the species with the most annotations in the Nr database was *F. vesca*, followed by *Acyrthsiphon pisum*, *Prunus mume*, *Prunus persica*, and *Zootermopsis nevadensis*; the number of unigenes for these species annotated by BLASTx alignment were 25,410 (66.5%), 3668 (9.6%), 955 (2.5%), 917 (2.4%), and 764 (2.0%), respectively (Fig. 2).

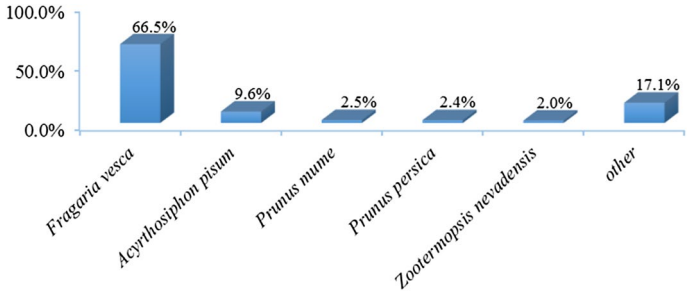

Fig. 2 The species-based distribution of the top BLASTx hits for each assembled red-flowered strawberry unigenes search against non-redundant protein database (Nr)

GO Annotation of Assembled Transcripts

A total of 26,861 (29.24%) unigenes were mapped to GO terms in the cellular component, molecular function, and biological process categories (Fig. 3) and were further classified into 46 terms. In the cellular component category, cell part (31.27%), cell (31.27%), organelle (20.83%), and macromolecular complex (20.61%) were the four most enriched GO terms (Table S2).

In the biological process category, cellular process (55.56%), metabolic process (54.43%), and single-organism process (41.56%) were the three most highly enriched terms, followed by regulation of biological regulation (18.00%), biological process (16.66%), localization (15.95%), and response to stimulus (11.83%). In the molecular function category, binding (56.84%) and catalytic activity (45.77%) were the two most highly enriched GO terms, followed by transporter activity (7.03%) and structural molecule activity (3.48%). This suggests that the identified unigenes are involved in growth, development, metabolism, and apoptosis in red-flowered strawberry.

Characteristics of Genic SSRs in the Red-Flowered Strawberry Transcriptome

SSRs were abundant in the assembled unigene dataset; 12,077 unigenes with 14,264 SSR loci were identified from 139,292 transcripts. Among them, 1823 contained more than one SSR locus, and 758 SSR loci existed in complex form. One SSR locus was found every 10.70 kb of unigene sequence. Among all repeat types, SSR length ranged from 10 to 228 bp, with an average length of 16.58 bp. The most abundant repeat units were single-base repeat sequences—i.e., mononucleotides (47.2%), followed by di- (32.5%), tri- (19.1%), tetra- (0.9%), hex- (0.2%), and penta- (0.1%) nucleotides (Fig. 4a). The repeated unit number of SSR loci ranged from 5 to 42. The number of mononucleotide repeated unit ranges from 9 to 23. Of them, 10 and 11 repeated units were the most dominant, with the SSR number of 3127 (21.92%) and 1367 (9.30%), respectively. Most of the di-nucleotide had 5 to 11 repeat units, the majority of which had 6 and 7 repeat units, with the SSR number

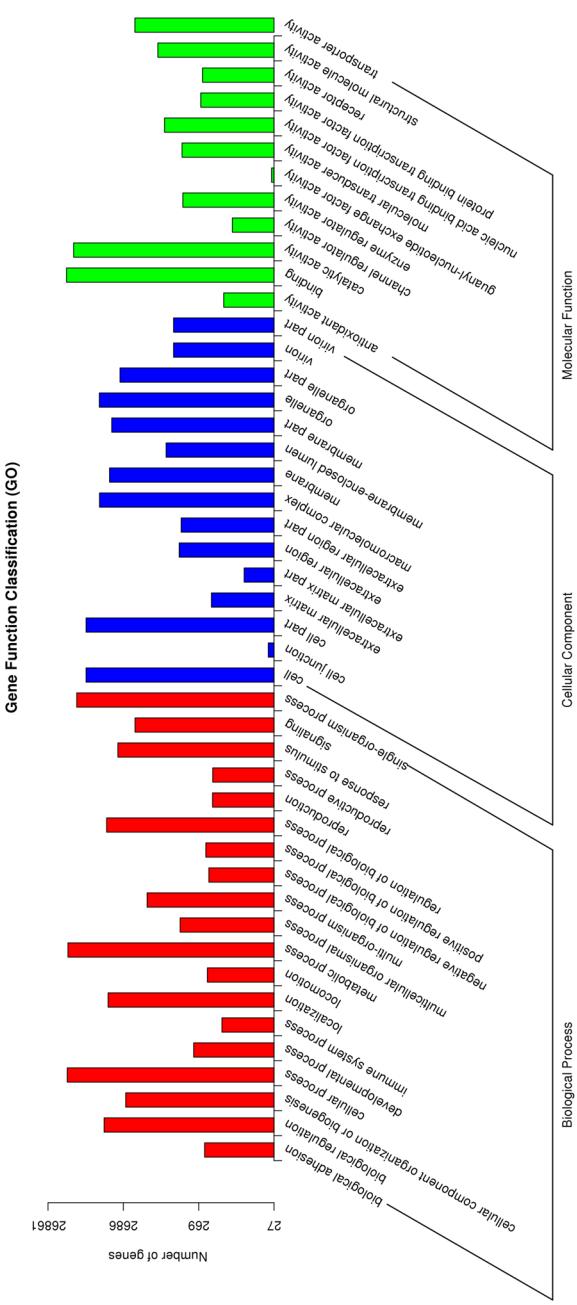

Fig. 3 Gene Ontology (GO) classification of unigene sequences containing SSR loci from red-flowered strawberry. The results are summarized in three main categories: cellular component, molecular function and biological process (Color figure online)

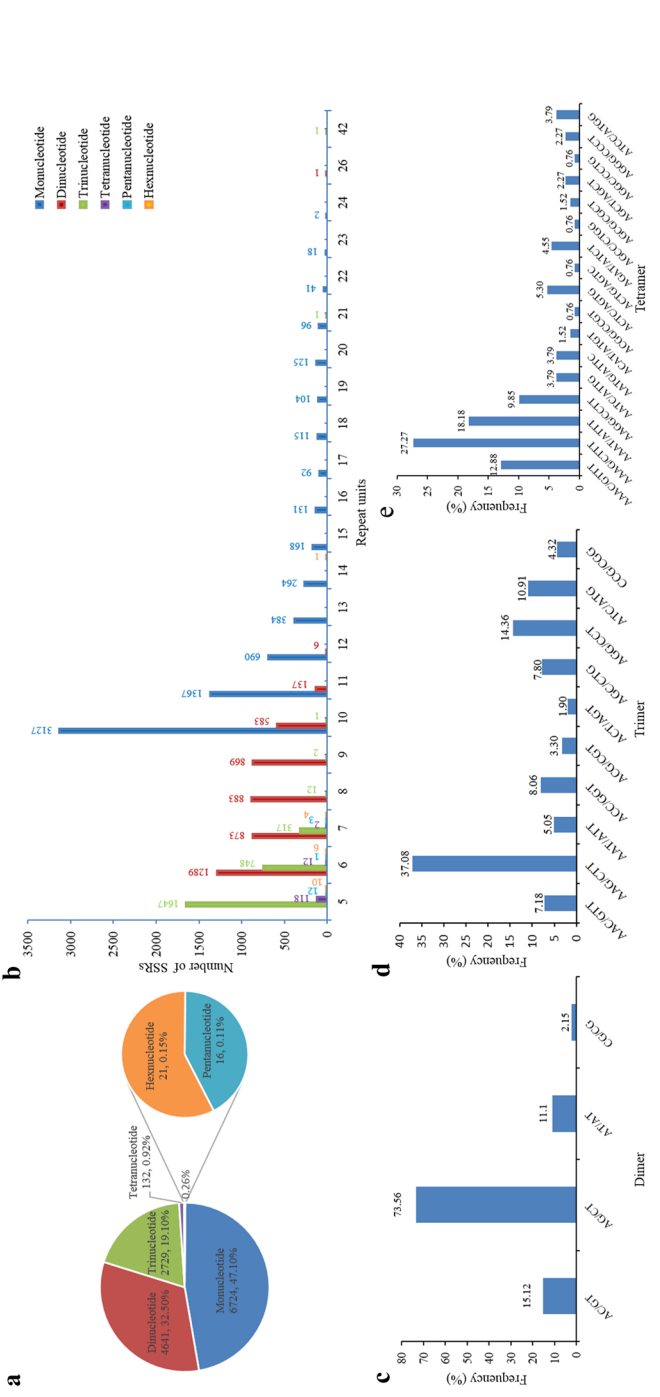

Fig. 4 Characterization of SSRs in red-flowered strawberry transcriptome. **a** Distribution of different SSR repeat motif types, **b** Number of different repeat motifs, **c–e** Fequency distribution of major SSRs based on main motif sequence

of 1289 (9.04%) and 873 (6.12%), respectively. Most of the tri-nucleotide had 5 to 7 repeat units, and the 5 and 6 repeat units were the most dominant, with the SSR number of 1647 (11.55%) and 748 (6.12%), respectively. The any repeat units of the remaining nucleotide were relatively low (< 1.2%) (Fig. 4b). Within identified SSRs, 176 motif sequence types were searched and the most common motifs were the tetra-nucleotide AAAG (4.54%), tri-nucleotide GAA (8.79%), and di-nucleotide GA (21.33%).

The 31 most common motifs and frequencies of individual SSR units were shown in Fig. 4c–e. The three most abundant motifs of the di-nucleotide repeat units were AG/CT, AC/GT and AT/AT with the frequencies of 52.13, 10.72 and 7.86%, respectively. Of the tri-nucleotide repeat units, AAG/CTT, AGG/CCT, ATC/ATG, ACC/GGT and AGC/CTG were the most abundant with the frequencies of 15.45, 5.99, 4.55, 3.36, and 3.25%, respectively. The four most abundant motifs of the tetra-nucleotide repeat units were AAAG/CTTT, AAAT/ATTT, AAAC/GTTT and AAGG/CCTT with the frequencies of 27.27, 18.18, 12.88 and 9.85%, respectively. However, other motifs shown in Fig. 4c accounted for less than 3% of total SSRs, and the GC motif was not detected.

Development and Detection of Trait-Specific Markers

A total of 12,077 unigenes containing SSRs from red strawberry transcriptome data were used for primer design; of these, 7621 (63.10%) sequences containing SSRs were used to develop SSR primers. Based on the transcriptome annotation, we first chose 33 SSR loci in the sequence of structure gene and transcription factor involved in anthocyanin biosynthesis pathway. Then, we randomly chose 12 SSR loci from each of mono-, di-, tri-, hex-, tetra- and penta-nucleotides, respectively (Table S2). A total of 105 SSR primers were synthesized and initial screening by polyacrylamide gel electrophoresis was performed for PCR amplicons with genomic DNA from red- and white-flowered parents and their hybrid progenies (Fig. S1).

The 105 SSR primer pairs were selected for polymorphism amplification of 176 red- and white-flowered hybrids from crosses of Pink Princess × Pink Beauty, Pink Princess × Honeoye, Pretty × Honeoye, and Pink Princess × Pretty Beauty. The 94 SSR primer pairs amplified 184 clear bands with an effective amplification rate of 90.48%. The 55 SSR primer pairs were polymorphic and 79 polymorphic bands were amplified, yielding a polymorphism rate of 42.93%. PCR products were verified by polyacrylamide gel electrophoresis (Fig. S1a–d). The 4 SSR primer pairs, UgRFsr57622, UgRFsr94149, UgRFsr40142 and UgRFsr54608, were associated with the red flower trait of red-flowered strawberry, among which only UgRFsr57622 was polymorphic (Fig. S1a).

An analysis of the representative UgRFsr57622 primer pair by 3% (w/v) agarose gel electrophoresis showed that only one band was amplified from the white-flowered hybrids used for transcriptome sequencing from Pink Princess × Pink Beauty and the white-flowered hybrids used for molecular marker identification from Pink Princess × Honeoye, Pretty × Honeoye, and Pink Princess × Pretty Beauty; in addition, a second band was amplified in all red-flowered hybrids

from transcriptome-sequenced (Pink Princess × Pink Beauty) and hybrid progeny (Pink Princess × Honeoye, Pretty × Honeoye, and Pink Princess × Pretty Beauty) (Fig. 5a). A single band was amplified from all red flowers with UgRFsr94149 and UgRFsr40142 primer pairs (Fig. 5b, c), with a secondary band amplified by the UgRFsr54608 primer pair (Fig. 5d). There were no amplicons in white-flowered progenies using these three primer sets (Fig. 5b–d).

The specific fragments were recovered and sequenced after cloning. The fact that the sequences matched the original sequence indicated that the trait markers were highly specific. All four trait-specific markers were readily identified as specific to red- and white-flowered genotypes by specific fragment marker analysis on polyacrylamide.

Discussion

High-throughput sequencing technology can reveal transcript sequences and allows the analysis of transcriptomes for species for which there is no available genome information. The obtained transcriptome data were searched against known protein databases (Nr, Swiss-Prot, KOG, and KEGG). Many non-model plants have been sequenced using this method including *Salvia miltiorrhiza*, buckwheat, bayberry, black raspberry, lily, and others (Hua et al. 2011; Logacheva et al. 2011; Feng

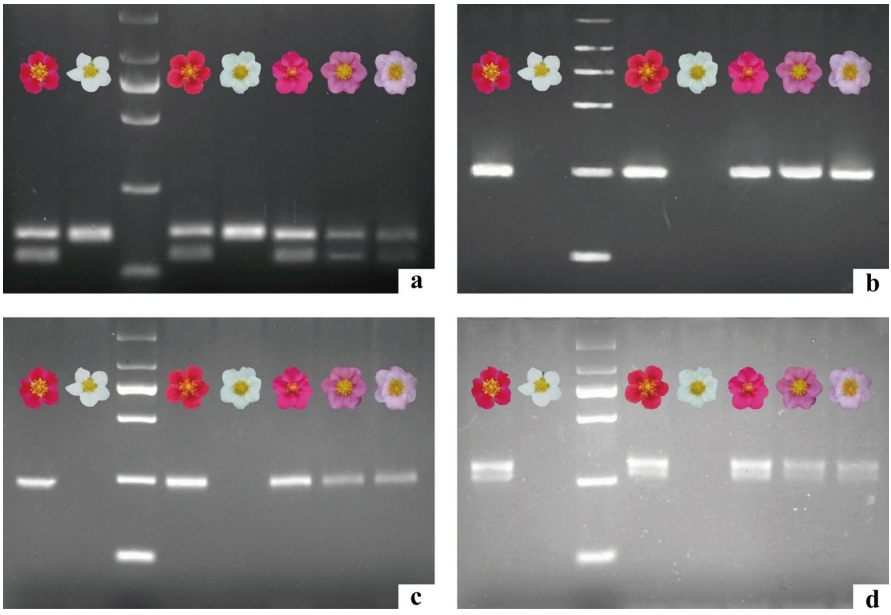

Fig. 5 Specific primers were used to amplify the gel electrophoresis pattern. **a–d** Were gel electrophoresis amplified by primers UgRFsr57622, UgRFsr94149, UgRFsr40142 and UgRFsr54608, respectively. 1 and 2 was from the cross of Princess × Pink Beauty used for RNA- seq. 3–7 was randomly selected from Princess × Honeoye, Pretty × Honeoye, Pink Princess × Pretty Beauty used for molecular marker identification (Color figure online)

et al. 2012; Hyun et al. 2012; Du et al. 2015). In this study, red- and white-flowered hybrids of Pink Princess × Pink Beauty were sequenced on an Illumina HiSeq 2500 platform. In this study, a total of 91,835 unigenes were obtained by transcriptome sequencing and 51.24% (47,058) unigenes have been annotated in at least one of the above-mentioned databases (Table 3). On the one hand, the relatively low number of genes annotated in this work could be due to the fact that these unigenes were new genes of red-flowered strawberry or that the genetic resources in the current database were limited; On the other hand, another probable reason was due to the high ploidy (8*x*) of red-flowered strawberry, such as 6*x* chrysanthemum (Hong et al. 2015; Xu et al., 2013) and 4*x* red-fleshed kiwifruit (Li et al. 2018). Xu et al. (2018) used 8*x* sugarcane (*Saccharum officinarum* L.) to obtain pokkah boeng disease and drought stress tolerance genotype, and obtained 93,115 unigenes by RNA-seq with lower annotation unigenes (46.74%).

SSRs have been screened based on transcriptome sequencing in many in angiosperm species including castor bean seeds, rubber tree, and sesame (Qiu et al. 2010; Li et al. 2012; Zhang et al. 2012). Transcriptome-derived SSR markers have been detected near or within functional gene squences (Li et al. 2002; Morgante et al. 2002). SSR primers based on the transcriptome can identify high-quality flanking regions (Silva et al. 2013). In this study, about 10.24% of red-flowered strawberry transcriptome sequences contained SSR loci, which was consistent with the frequency range (2.65–16.82%) reported in other dicotyledonous species (Zheng et al. 2013; Wang et al. 2014b; Chen et al. 2015). The characteristics of SSR nucleotides differ in various plants, and in barley, maize, rice, sorghum, wheat, gossypium, peanut among others, trinucleotides have the highest repetition frequency (Kantety et al. 2002; Wang et al. 2006; Liang et al. 2009). However, in our study the most abundant repeat sequence type was dinucleotides, which was consistent with findings in sesame, some Rosaceae plants, and lily (Zhang et al. 2013; Jung et al. 2005; Du et al. 2015). Among di-nucleotide repeat sequences, the AC/GT motif was the most common in red-flowered strawberry, which is similar to what has been reported for lily, petunia, and calla lily (Du et al. 2015; Wei et al. 2015a; Wei et al. 2016). Although the functional significance of most SSRs in plant transcriptome regions is unclear, a high frequency of purine-pyrimidine motifs homologous to AG/CT in 50 untranslated regions was shown to be associated with acid metabolism and nucleic acid regulation (Martienssen and Colot 2001; Scaglione et al. 2009; Wöhrmann and Weising 2011). Furthermore, $(GC)_n$ repeats are exceedingly rare in most eukaryotic genomes (Feng et al. 2009; Biswas et al. 2012).

Correct selection and separation of traits of interest is critical for plant breeding. Molecular marker-assisted breeding combines modern molecular biology and traditional genetic breeding, whereby the detection, location, and tracking of a single or multiple genes are linked to a target trait. Selected SNPs have been used to genotype populations of wild and cultivated soybeans (Guo et al. 2018). SNPs among control samples and proton beam-induced mutations have previously been detected in soybean with a genotyping-by-sequencing approach (Kim et al. 2018). In a study of gentian flowers of different colors, four allelic variations were associated with a loss of the anthocyanin biosynthesis regulatory gene *Gentiana triflora MYB3* and 4-bp deletion was existent in the second exon of the anthocyanidin synthase (ANS) gene,

which was separately determined based on three PCR-based molecular markers and a CAPS marker that successfully distinguished between white and blue gentian flowers (Nakatsuka et al. 2012). The *F3′5′H* gene controlling anthocyanin biosynthesis in petunias was also identified with PCR-based molecular markers (Matsubara et al. 2006).

At present, a large number of molecular markers have been studied in octoploid-cultivated strawberries, most of which are PCR-based. SSR markers were used to distinguish between Florida strawberry genotypes (Brunings et al. 2010). A high-resolution SSR map of *F. virginiana* was developed to analyze 30 linkage groups in each of the male and female maps (Spigler et al. 2010), and molecular markers have been used to select the flavor volatiles mesifurane and γ-decalactone in the fruit of cultivated strawberry (Cruzrus et al. 2017). Red-flowered strawberry have a long juvenile period, and the selection of flower color requires a longer period of time. Thus, techniques for early identification of flower color using DNA markers are extremely useful for breeding of red-flowered strawberry. In this study, four trait-specific markers were obtained from SSRs identified by RNA-seq; All four trait-specific markers can be useful to discriminate flower color in red-flowered strawberry-breeding programs for early selection of progeny and for breeding management.

Acknowledgements This work was supported by Natural Science Fund of China (No. 31701964).

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Bai Y, Dougherty L, Xu K (2014) Towards an improved apple reference transcriptome using RNA-seq. Mol Genet Genom 289(3):427–438. https://doi.org/10.1007/s00438-014-0819-3

Bematzky R, Mulcahy DL, Tuskan GA (1992) Marker-aided selection in a backcross breeding program for resistance to chestnut blight in the American chestnut. Can J For Res 22(22):1031–1035. https://doi.org/10.1139/x92-137

Bentvelsen G, Bouw B (2006) Breeding ornamental strawberries. Acta Hortic 708:455–458. https://doi.org/10.17660/ActaHortic.2006.708.80

Biswas MK, Chai L, Mayer C, Xu Q, Guo W, Deng X (2012) Exploiting bac-end sequences for the mining, characterization and utility of new short sequences repeat (SSR) markers in *Citrus*. Mol Biol Rep 39(5):5373–5386. https://doi.org/10.1007/s11033-011-1338-5

Brunings AM, Moyer C, Peres N, Folta KM (2010) Implementation of simple sequence repeat markers to genotype Florida strawberry varieties. Euphytica 173(1):63–75. https://doi.org/10.1007/s10681-009-0112-4

Chen C, Zhou P, Choi YA, Huang S, Gmitter FG Jr (2006) Mining and characterizing microsatellites from citrus ESTs. Theor Appl Genet 112(7):1248–1257. https://doi.org/10.1007/s00122-006-0226-1

Chen H, Liu L, Wang L, Wang S, Somta P, Cheng X (2015) Development and validation of EST-SSR markers from the transcriptome of adzuki bean (*Vigna angularis*). PLoS ONE 10:e0131939. https://doi.org/10.1371/journal.pone.0131939

Cruzrus E, Sesmero R, Sánchezsevilla JF, Ulrich D, Amaya I (2017) Validation of a PCR test to predict the presence of flavor volatiles mesifurane and γ-decalactone in fruits of cultivated strawberry (*Fragaria*×*ananassa*). Mol Breed 37(10):131. https://doi.org/10.1007/s11032-017-0732-7

Cuadrado A, Cardoso M, Jouve N (2008) Physical organisation of simple sequence repeats (SSRs) in Triticeae: structural, functional and evolutionary implications. Cytogenet Genome Res 120(3–4):210–219. https://doi.org/10.1159/000121069

Du F, Wu Y, Zhang L, Li XW, Zhao XY, Wang WH, Gao ZS, Xia YP (2015) De novo assembled transcriptome analysis and SSR marker development of a mixture of six tissues from *Lilium* Oriental hybrid 'Sorbonne'. Plant Mol Biol Rep 33(2):281–293. https://doi.org/10.1007/s11105-014-0746-9

Ellis JR (1962) *Fragaria-Potentilla* intergeneric hybridization and evolution in *Fragaria*. Proc Linn Sac Land 173:99–106. https://doi.org/10.1111/j.1095-8312.1962.tb01300.x

Feng SP, Li WG, Huang HS, Wang JY, Wu YT (2009) Development, characterization and cross- species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). Mol Breed 23(1):85–97. https://doi.org/10.1007/s11032-008-9216-0

Feng C, Chen M, Xu C, Bai L, Yin X, Li X, Allan AC, Ferguson IB, Chen KS (2012) Transcriptomic analysis of Chinese bayberry (*Myrica rubra*) fruit development and ripening using RNA-Seq. BMC Genom 13(1):19. https://doi.org/10.1186/1471-2164-13-19

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652. https://doi.org/10.1038/nbt.1883

Guo Y, Su B, Tang J, Zhou F, Qiu LJ (2018) Gene-based SNP identification and validation in soybean using next-generation transcriptome sequencing. Mol Genet Genom 293(3):623–633. https://doi.org/10.1007/s00438-017-1410-5

Hong Y, Tang XJ, Huang H, Zhang Y, Dai SL (2015) Transcriptomic analyses reveal species-specific light-induced anthocyanin biosynthesis in chrysanthemum. BMC Genom 16(1):202

Hua W, Zhang Y, Song J, Zhao L, Wang Z (2011) De novo transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. Genomics 98(4):272–279. https://doi.org/10.1016/j.ygeno.2011.03.012

Hyun TK, Rim Y, Jang HJ, Kim CH, Park J, Kumar R, Lee SY, Kim BC, Bhak J, Nguyen-Quoc B, Kim SW, Lee S, Kim JY (2012) De novo transcriptome sequencing of *Momordica cochinchinensis* to identify genes involved in the carotenoid biosynthesis. Plant Mol Biol 79(4–5):413–427. https://doi.org/10.1007/s11103-012-9919-9

Hyun TK, Lee S, Rim Y, Kumar R, Han X, Lee SY, Lee CH, Kim JY (2014) De-novo RNA sequencing and metabolite profiling to identify genes involved in anthocyanin biosynthesis in Korean black raspberry (*Rubus coreanus* Miquel). PLoS ONE 9(2):e88292. https://doi.org/10.1371/journal.pone.0088292

Jung S, Abbott A, Jesudurai C, Tomkins J, Main D (2005) Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs. Funct Integr Genom 5(3):136–143. https://doi.org/10.1007/s10142-005-0139-0

Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48(5–6):501–510. https://doi.org/10.1023/A:1014875206165

Kim WJ, Ryu J, Im J, Kim SH, Kang SY, Lee JH, Jo SH, Ha BK (2018) Molecular characterization of proton beam-induced mutations in soybean using genotyping-by-sequencing. Mol Genet Genom. https://doi.org/10.1007/s00438-018-1448-z

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11:2453–2465. https://doi.org/10.1046/j.1365-294X.2002.01643.x

Li D, Deng Z, Qin B, Liu X, Men Z (2012) De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). BMC Genom 5(18):192. https://doi.org/10.1186/1471-2164-13-192

Li Y, Fang J, Qi X, Lin M, Zhong Y, Sun L (2018) A key structural gene, *AaLDOX*, is involved in anthocyanin biosynthesis in all red-fleshed kiwifruit (*Actinidia arguta*) based on transcriptome analysis. Gene 30(648):31–41. https://doi.org/10.1016/j.gene.2018.01.022

Liang X, Chen X, Hong Y, Liu H, Zhou G, Li S, Guo B (2009) Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species. BMC Plant Biol 9(1):1–9. https://doi.org/10.1186/1471-2229-9-35

Liu L, Guo W, Zhu X, Zhang T (2003) Inheritance and fine mapping of fertility-restoration for cytoplasmic male sterility in *Gossypium hirsutum* L. Theor Appl Genet 106:461–469. https://doi.org/10.1007/s00122-002-1084-0

Logacheva MD, Kasianov AS, Vinogrado DV, Samigullin TH, Gelfand MS, Makeev VJ, Penin AA (2011) De novo sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). BMC Genom 12(1):30. https://doi.org/10.1186/1471-2164-12-30

Mabberley DJ (2002) *Potentilla* and *Fragaria* (Rosaceae) reunited. Telopea 9(4):793–801. https://doi.org/10.7751/telopea20024018

Martienssen RA, Colot V (2001) DNA methylation and epigenetic inheritance in plants and filamentous fungi. Science 293(5532):1070–1074. https://doi.org/10.1126/science.293.5532.1070

Matsubara K, Chen S, Lee JX, Kodama H, Kokubun H, Watanabe H, Ando T (2006) PCR-based markers for the genotype identification of flavonoid-3′,5′-hydroxylase genes governing floral anthocyanin biosynthesis in commercial petunias. Breed Sci 56(4):389–397. https://doi.org/10.1270/jsbbs.56.389

Miah G, Rafii MY, Ismail MR, Puteh AB, Rahim HA, Islam KhN, Latif MA (2013) A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. Int J Mol Sci 14:22499–22528. https://doi.org/10.3390/ijms141122499

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30(2):194–200. https://doi.org/10.1038/ng822

Nakatsuka T, Saito M, Sato-Ushiku Y, Yamada E, Nakasato T, Hoshi N, Fujiwara K (2012) Development of DNA markers that discriminate between white- and blue-flowers in Japanese gentian plants. Euphytica 184(4):335–344. https://doi.org/10.1007/s10681-011-0534-7

Pillet J, Yu HW, Chambers AH, Whitaker VM, Folta KM (2015) Identification of candidate flavonoid pathway genes using transcriptome correlation network analysis in ripe strawberry (*Fragaria*×*ananassa*) fruits. J Exp Bot 66(15):4455–4467. https://doi.org/10.1093/jxb/erv205

Qiu L, Yang C, Tian B, Yang JB, Liu A (2010) Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). BMC Plant Biol 10(1):1–10. https://doi.org/10.1186/1471-2229-10-278

Röder MS, Korzun V, Wendehake K, Plaschke J, Tixier MH, Leroy P, Ganal MW (1998) A microsatellite map of wheat. Genetics 149(4):2007–2023

Scaglione D, Acuadro A, Portis E, Taylor CA, Lanteri S, Knapp SJ (2009) Ontology and diversity of transcript-associated microsatellites mined from a globe artichoke EST database. BMC Genom 10(1):1–17. https://doi.org/10.1186/1471-2164-10-454

Shan LL, Li X, Wang P, Cai C, Zhang B, Sun CD, Zhang WS, Xu CJ, Ferguson I, Chen KS (2008) Characterization of cDNAs associated with lignification and their expression profiles in loquat fruit with different lignin accumulation. Planta 227(6):1243–1254. https://doi.org/10.1007/s00425-008-0696-2

Silva PI, Martins AM, Gouvea EG, Pessoa-Filho M, Ferreira ME (2013) Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genom 14(1):1–9. https://doi.org/10.1186/1471-2164-14-17

Spigler RB, Lewers KS, Johnson AL, Ashman TL (2010) Comparative mapping reveals autosomal origin of sex chromosome in octoploid *Fragaria virginiana*. J Hered 101(2):S107–S117. https://doi.org/10.1093/jhered/esq001

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106(3):411–422

Ujino T, Kawahara T, Tsumura Y, Nagamitsu T, Yoshimaru H, Ratnam W (1999) Development and polymorphism of simple sequence repeat DNA markers for *Shorea curtisii*, and other dipterocarpaceae species. Heredity 81(4):422–428. https://doi.org/10.1038/sj.hdy.6884230

Uzunova MI, Ecke W (1999) Abundance, polymorphism and genetic mapping of microsatellites in oilseed rape (*Brassica napus* L.). Plant Breed 118(4):323–326. https://doi.org/10.1139/g09-084

Wang CB, Guo WZ, Cai CP, Zhang TZ (2006) Characterization, development and exploitation of EST derived microsatellites in *Gossypium raimondii* Ulbrich. Chin Sci Bull 51(5):557–561. https://doi.org/10.1007/s11434-006-0557-y

Wang BH, Zhu P, Yuan YL, Wang CB, Yu CM, Zhang HH, Zhu XY, Wang W, Yao CB, Zhuang ZM, Li P (2014a) Development of EST-SSR markers related to salt tolerance and their application in

genetic diversity and evolution analysis in *Gossypium*. Genet Mol Res 13(2):3732–3746. https://doi.org/10.4238/2014.May.13.1

Wang S, Zhang Z, Jiang NH, Zhang GH, Sha BC, Yang SC, Chen JW (2014b) SSR information in transcriptome of *Pinellia ternata*. J Chin Med Mater 37(9):1566–1569. https://doi.org/10.13863/j.issn1001-4454.2014.09.015

Wei ZZ, Luo LB, Zhang HL, Xiong M, Wang X, Zhou D (2012) Identification and characterization of 43 novel polymorphic EST-SSR markers for arum lily, *Zantedeschia aethiopica* (Araceae). Am J Bot 99(12):493–497. https://doi.org/10.3732/ajb.1200228

Wei C, Tao X, Li M, He B, Yan L, Tan X, Zhang Y (2015a) De novo transcriptome assembly of *Ipomoea nil* using Illumina sequencing for gene discovery and SSR marker identification. Mol Genet Genom 290(5):1873–1884. https://doi.org/10.1007/s00438-015-1034-6

Wei H, Chen X, Zong X, Shu H, Gao D, Liu Q (2015b) Comparative transcriptome analysis of genes involved in anthocyanin biosynthesis in the red and yellow fruits of sweet cherry (*Prunus avium* L.). PloS ONE 10(3):e0121164. https://doi.org/10.1371/journal.pone.0121164

Wei Z, Sun Z, Cui B, Zhang Q, Xiong M, Wang X, Zhou D (2016) Transcriptome analysis of colored calla lily (*Zantedeschia rehmannii* Engl.) by Illumina sequencing: de novo assembly, annotation and EST-SSR marker development. Peer J 4(9):e2378. https://doi.org/10.7717/peerj.2378

Wöhrmann T, Weising K (2011) In silico mining for simple sequence repeat loci in a pineapple expressed sequence tag database and cross-species amplification of EST–SSR markers across Bromeliaceae. Theor Appl Genet 123(4):635–647. https://doi.org/10.1007/s00122-011-1613-9

Xu Y, Gao S, Yang Y, Huang M, Cheng L, Wei Q, Fei Z, Gao J, Hong B (2013) Transcriptome sequencing and whole genome expression profiling of chrysanthemum under dehydration stress. BMC Genom 14(1):662. https://doi.org/10.1186/1471-2164-14-662

Xu S, Wang J, Shang H, Huang Y, Yao W, Chen B, Zhang M (2018) Transcriptomic characterization and potential marker development of contrasting sugarcane cultivars. Sci Rep 8(1):1683. https://doi.org/10.1038/s41598-018-19832-x

Xue L, Lei JJ, Dai HP, Deng MQ (2014) Two new red-flowered strawberry cultivars 'Pink Beauty' and 'Pretty Beauty'. Acta Hortic 1049:231–234

Yan Y, Ma HX, Lei JJ, Yu GH (2006) Conversion of RAPD marker linked to red-flowered Gene of red-flowered strawberry to SCAR marker. Mol Plant Breed 4(5):690–694 **(in Chinese)**

Yan H, Zhang H, Chen M, Jian H, Baudino S, Caissard JC, Bendahmane M, Li S, Zhang T, Zhou N, Qiu X, Wang Q, Tang K (2014) Transcriptome and gene expression analysis during flower blooming in *Rosa chinensis* 'Pallida'. Gene 544(2):96–103. https://doi.org/10.1016/j.gene.2014.02.008

Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99:193–208. https://doi.org/10.3732/ajb.1100394

Zhang HY, Wei LB, Miao HM, Zhang TD, Wang CY (2012) Development and validation of Cgenic-SSR markers in sesame by RNA-seq. BMC Genom 13(1):316. https://doi.org/10.1186/1471-2164-13-316

Zhang J, Li H, Zhang M, Hui M, Wang Q, Li L, Zhang L (2013) Fine mapping and identification of candidate *Br-or* gene controlling orange head of Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). Mol Breed 32(4):799–805. https://doi.org/10.1007/s11032-013-9907-z

Zhang J, Yuan H, Fei Z, Pogson BJ, Zhang L, Li L (2015) Molecular characterization and transcriptome analysis of orange head Chinese cabbage (*Brassica rapa* L. ssp. *pekinensis*). Planta 241(6):1381–1394. https://doi.org/10.1007/s00425-015-2262-z

Zheng XF, Pan C, Diao Y, You YN, Yang CZ, Hu ZL (2013) Development of microsatellite markers by transcriptome sequencing in two species of *Amorphophallus* (Araceae). BMC Genom 14:490. https://doi.org/10.1186/1471-2164-14-490