GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Survey of simple sequence repeats in woodland strawberry (*Fragaria vesca*)

L. Guan[1], J.F. Huang[1], G.Q. Feng[1], X.W. Wang[1], Y. Wang[1], B.Y. Chen[1] and Y.S. Qiao[1,2]

[1]College of Horticulture, Nanjing Agricultural University, Nanjing, China
[2]Engineering Research Center of Horticultural Crop Germplasm Enhancement and Utilization, Ministry of Education of the People's Republic of China, Nanjing, China

Corresponding author: Y.S. Qiao
E-mail: qiaoyushan@njau.edu.cn

Genet. Mol. Res. 12 (3): 2637-2651 (2013)
Received August 30, 2012
Accepted March 16, 2013
Published July 30, 2013
DOI http://dx.doi.org/10.4238/2013.July.30.3

ABSTRACT. The use of simple sequence repeats (SSRs), or microsatellites, as genetic markers has become popular due to their abundance and variation in length among individuals. In this study, we investigated linkage groups (LGs) in the woodland strawberry (*Fragaria vesca*) and demonstrated variation in the abundances, densities, and relative densities of mononucleotide, dinucleotide, and trinucleotide repeats. Mononucleotide, dinucleotide, and trinucleotide repeats were more common than longer repeats in all LGs examined. Perfect SSRs were the predominant SSR type found and their abundance was extremely stable among LGs and chloroplasts. Abundances of mononucleotide, dinucleotide, and trinucleotide repeats were positively correlated with LG size, whereas those of tetranucleotide and hexanucleotide SSRs were not. Generally, in each LG, the abundance, relative abundance, relative density, and the proportion of each unique SSR all declined rapidly as the repeated unit increased. Furthermore, the lengths and frequencies of SSRs varied among different LGs.

Key words: Simple sequence repeats; Chloroplast; *Fragaria vesca*; Genome; Molecular marker

INTRODUCTION

Cultivated strawberry (*Fragaria* x *ananassa* Duch.) originated 250 years ago and is among the youngest of all crop species (Darrow, 1966). Genomically, *F.* x *ananassa* is also one of the most complex crops, harboring 8 sets of chromosomes (2n = 8x = 56), which are derived from as many as 4 different diploid ancestors. Paradoxically, the small and basic (X = 7) genome size of the woodland strawberry, about 240 Mb, makes it particularly suitable for genomic research (Shulaev et al., 2011). Based on this genetic advantage, an international consortium selected the *F. vesca* (2n = 2x = 14) sequence as a genomic reference for the family Rosaceae (Shulaev et al., 2008). Moreover, *F. vesca* may also provide other advantages as a reference genomic system for Rosaceae research, including its relatively short generation time compared to other perennials, easy vegetative propagation, and its small herbaceous structure relative to woody plants such as apple or peach.

Simple sequence repeats (SSRs), also known as microsatellites, comprise tandemly repeated genetic loci of 1 to 6 bp. SSRs are highly abundant and exhibit extensive levels of polymorphism in both eukaryotic and prokaryotic genomes. They are found in both protein-coding and in noncoding regions of genomes, but are more abundant in noncoding regions than in exons (Tóth et al., 2000; Li et al., 2011). Compared with an average mutation rate of approximately 10^{-9} single nucleotide substitutions per generation for eukaryotic genomes, SSR locus mutation rates are high, ranging from 10^{-3} to 10^{-6} per generation (Schug et al., 1997; Xu et al., 2000). This relatively high mutation rate is induced by the slipped strand mispairing of DNA polymerase and unequal crossing over events, and results in an increase of one or more repeat motifs (Levinson and Gutman, 1987; Richards and Sutherland, 1992). Consequently, the lengths of SSRs are increased throughout the genome. Furthermore, the mutation rate of SSRs generally increases as the length of the repeated unit grows (Wierdl et al., 1997). It has also been suggested that SSRs undergo a life cycle; they are born, they grow, and then ultimately die. The entire life cycle of an SSR may span tens, or even hundreds of millions of years (Messier et al., 1996). Fixation of *de novo* generated SSRs is determined by the interplay among the type of repeat, the specific genomic position, and the genetic or biochemical background of the cell (Tóth et al., 2000). Furthermore, different chromosomes in the same species tend to have similar SSR types but different SSR densities (Kruglyak et al., 2000). SSR motifs, abundances, and mutation rates vary among species and exhibit a wide range of genetic properties (Cruz et al., 2005).

According to a set of precise classification rules, SSRs are divided into 3 categories: perfect SSRs (without interruptions in the runs of the repeated sequence), imperfect SSRs (one or more interruptions in the run of the repeated sequence), and compound SSRs (adjacent tandem simple repeats separated by one or more base pairs). The polymorphic information content values obtained from perfect SSRs are more informative than those obtained from imperfect or compound SSRs (Weber, 1990). In addition, SSRs are also categorized into 2 classes, based on the length of repeats: Class I $\geq$ 20 bp and Class II = 12-19 bp. Class I repeats are highly polymorphic, but Class II repeats tend to be less variable (Temnykh et al., 2001). Class II SSRs were most commonly found in a study of 5 *Eucalyptus* species (Rabello et al., 2005). In addition, microsatellite primers are normally conserved within species and may even be transferable among different taxa (Yin et al., 2004; Castillo et al., 2008; Guan et al., 2012). Their high transferability, ubiquity, and char-

acterization of repeatability and co-dominance have made microsatellites one of the most powerful genetic markers for use in genetic linkage map construction, genetic diversity detection, gene mapping, and molecular-assisted breeding (Tautz et al., 1986; Plaschke et al., 1995; Robinson et al., 2004).

The availability of complete genome sequences for many organisms has made it possible to conduct genome-wide analyses of SSRs. To date, genome information of horticultural crops has been used to analyze the regulation of SSR distribution in the genome (Cai et al., 2009; Guan et al., 2011), and SSR markers developed from specific crop species are successfully applied in genetic analyses of related species (Guan et al., 2011, 2012). In this study, we screened the whole nuclear and chloroplast genomes of *F. vesca* in order to determine the distribution of SSRs, SSR sequence types, the longest SSR, and the relationship between SSR abundance and the size of its corresponding linkage group (LG). Since accumulating evidence has suggested that SSRs function in regulating gene expression (Künzler et al., 1995; Moxon and Wills, 1999), this study will provide baseline information that will increase insight into the evolutionary history of *F. vesca*.

MATERIAL AND METHODS

Genome and chloroplast data were obtained from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov) (Table 1). A lower number of sequences were analyzed than are estimated from the genome size (X = 7) of *F. vesca* (~240 Mb), likely because some regions, such as telomeres and centromeres, have been omitted in these LGs.

Table 1. Data collection of linkage groups and chloroplast in *Fragaria vesca*.

No. of research object	Public web site	Sequence analyzed (Mb)
LG 1	http://www.ncbi.nlm.nih.gov/nuccore/CM001053.1	22.68
LG 2	http://www.ncbi.nlm.nih.gov/nuccore/CM001054.1	33.31
LG 3	http://www.ncbi.nlm.nih.gov/nuccore/CM001055.1	27.88
LG 4	http://www.ncbi.nlm.nih.gov/nuccore/CM001056.1	23.29
LG 5	http://www.ncbi.nlm.nih.gov/nuccore/CM001057.1	29.33
LG 6	http://www.ncbi.nlm.nih.gov/nuccore/CM001058.1	38.22
LG 7	http://www.ncbi.nlm.nih.gov/nuccore/CM001059.1	23.40
Genome		198.11
Chloroplast	http://www.ncbi.nlm.nih.gov/nuccore/NC_015206.1	156 (kb)

Querying for SSR was supported by the Perl script search module MISA (http://pgrc.ipk-gatersleben.de/misa), which allows for the identification of perfect and compound SSRs (Varshney et al., 2002). Perfect SSRs were defined as sequences of 10 or more mononucleotide repeats, 6 or more dinucleotide repeats, and 5 or more tri-, tetra-, penta-, or hexanucleotide repeats. Compound SSRs were considered when repeats in same sequence were separated by a maximum of 100 bp. Repeat number and genome location were recorded and reported in an output file. The data were processed and tabulated using Microsoft Excel 2003. All microsatellite data were normalized as the number of SSRs per Mb in order to effectively compare sequences among all LGs (relative abundance). The estimated relative density of repeats (bp/Mb) for each LG was calculated as the total SSR length (bp) in each

analyzed sequence (Mb). Differences among occurrences of repeats in exons, introns, and intergenic regions were not considered.

RESULTS

We analyzed perfect SSRs of the nuclear (198.11 Mb) and chloroplast (155.69 kb) genomes in *F. vesca*, and calculated abundance, relative abundance (Table 2), density, relative density (Figures 1 and 2), and the proportion of microsatellites of different lengths (Figure 3). The results showed that the total number of SSRs was 80,350, which consisted of 61,495 (76.53%) perfect SSRs. We observed that the highest and lowest abundance of SSRs were found in LG 6 (12,121) and LG 1 (6859), respectively. The percentage of perfect SSRs was very stable among all LGs (the largest proportion was 77.57% in LG 4, and the smallest proportion was 75.85% in LG 7) (Table 2). In contrast, there were only 46 SSRs in total in the *F. vesca* chloroplast, 37 of which were perfect SSRs (80.43%). Based on the relative abundance measures, the abundances of mononucleotide, dinucleotide and trinucleotide SSRs were found to be positively correlated with the size of their corresponding LGs (correlation coefficients were all >0.95; Figure 4), but the SSRs of tetranucleotides to hexanucleotides were not (data not shown). In addition, the relative abundances in the nuclear genome and in the chloroplast were 310.41 and 237.64 repeats/Mb, respectively. Finally, the highest and the lowest abundances were observed in LG 5 (320.53) and LG 1 (302.43), respectively.

SSR density, the average length (kb) of sequences analyzed from a single SSR locus, were 3.22 kb/repeat and 4.21 kb/repeat in the nuclear and chloroplast genomes, respectively. The highest density was observed in LG 5 (3.12 kb/repeat) and the lowest density was in LG 1 (3.31 kb/repeat). We also found that the density distribution of mononucleotide, dinucleotide, and trinucleotide repeats was stable among all LGs, but varied in tetranucleotide and hexanucleotide repeats (Figure 1). These results were consistent with a survey of SSR density distribution in the apple genome (*Malus* x *domestica* Borkh) (Guan et al., 2011). The relative density of perfect SSRs was 4940 bp/Mb and 2858 bp/Mb in the nuclear and chloroplast genome, respectively, with the highest relative abundance found in LG 5 (5153 bp/Mb), and the lowest found in LG 7 (4805 bp/Mb) (Figure 2). The relative abundance of all LGs decreased as the number of repeats increased (Figure 2). The relative density of SSRs is lower in *F. vesca* compared to that of the human genome (3150 Mb) or in genomes of other mammals and higher eukaryotic organisms, but is higher than relative densities of unicellular organisms and other lower eukaryotes (Figure 5). A scan for the longest and the most frequent type of SSRs in the *F. vesca* genome (Table 3 and Table S1) revealed that the dinucleotide repeat AC in LG 1 [(AC)102] was the longest (Table 3), and the mononucleotide motif A/T [in LG 6, (A/T)8495] was the most frequent (Table S2). The same type of repeats showed similar proportions in each LG (Figure 3). The maximum proportion of mononucleotide repeats of more than 10 units was found in LG 7 (42.06%). In contrast, no hexanucleotide repeats of more than 10 units were found anywhere (Figure 3). Only 5 types of SSR were identified in the chloroplast (Figure 3): 5 units of trinucleotide repeats (2.70% of total SSRs), 6 units of dinucleotide repeats (2.70%), more than 10 units of dinucleotide repeats (2.70%), 10 units of mononucleotide repeats (56.76%), and more than 10 units of mononucleotide repeats (35.14%).

Table 2. Abundance and relative abundance of perfect SSRs in *Fragaria vesca*.

No. of research object	GenBank accession	LG size (Mb)	Total number of SSRs (Percent of perfect SSRs)	Repeat type						
				Mono	Di	Tri	Tetra	Penta	Hexa	Total
LG 1	CM001053.1	22.68	8,992 (76.28%)	4023/177.38*	2102/92.68	652/28.75	60/2.65	11/0.49	11/0.49	6859/302.43
LG 2	CM001054.1	33.31	13,778 (75.86%)	6258/187.87	3176/95.35	897/26.93	82/2.46	17/0.51	22/0.66	10,452/313.78
LG 3	CM001055.1	27.88	11,117 (76.44%)	5247/188.20	2401/86.12	777/27.87	34/1.22	20/0.72	19/0.68	8498/304.81
LG 4	CM001056.1	23.29	9,119 (77.57%)	4222/181.28	2127/91.33	647/27.78	55/2.36	17/0.73	6/0.26	7074/303.74
LG 5	CM001057.1	29.33	12,138 (77.45%)	5696/194.20	2761/94.14	841/28.67	64/2.18	19/0.65	20/0.68	9401/320.53
LG 6	CM001058.1	38.22	15,859 (76.43%)	7437/194.58	3507/91.76	1053/27.55	70/1.83	31/0.81	23/0.60	12,121/317.14
LG 7	CM001059.1	23.4	9,347 (75.85%)	4384/187.35	2034/86.92	593/25.34	36/1.54	18/0.77	25/1.07	7090/302.99
Genome		198.11	80,350 (76.53%)	37,267/188.11	18,108/91.40	5460/27.56	401/2.02	133/0.67	126/0.64	61,495/310.41
Chloroplast	NC_015206.1	155.69 kb	46 (80.43%)	34/218.37	2/12.85	1/6.42	-	-	-	37/237.64

*SSR abundance/relative abundance. Relative abundance was defined as the total number of SSRs per Mb of the sequence analyzed. LG = linkage groups.

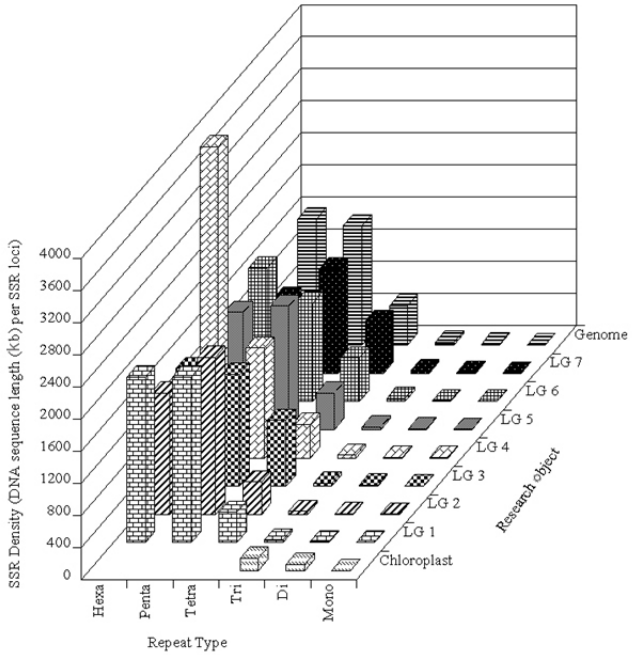

Figure 1. Simple sequence repeat (SSR) density across the whole genome and chloroplast in *Fragaria vesca*. Density was defined as the average length (kb) of DNA sequence analyzed from single SSR loci.

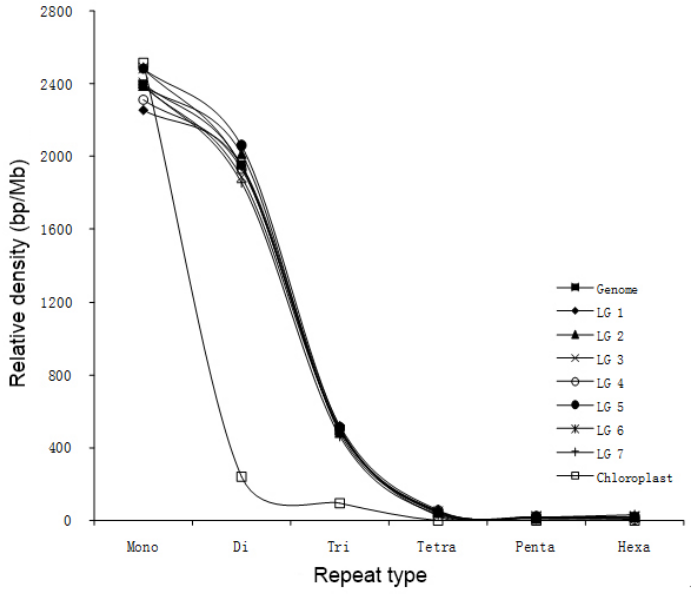

Figure 2. Relative density of simple sequence repeats (SSRs) across the whole genome and chloroplast in *Fragaria vesca*. Relative density was defined as the total sequence length (bp) contributed by each SSR per Mb of DNA sequence analyzed.

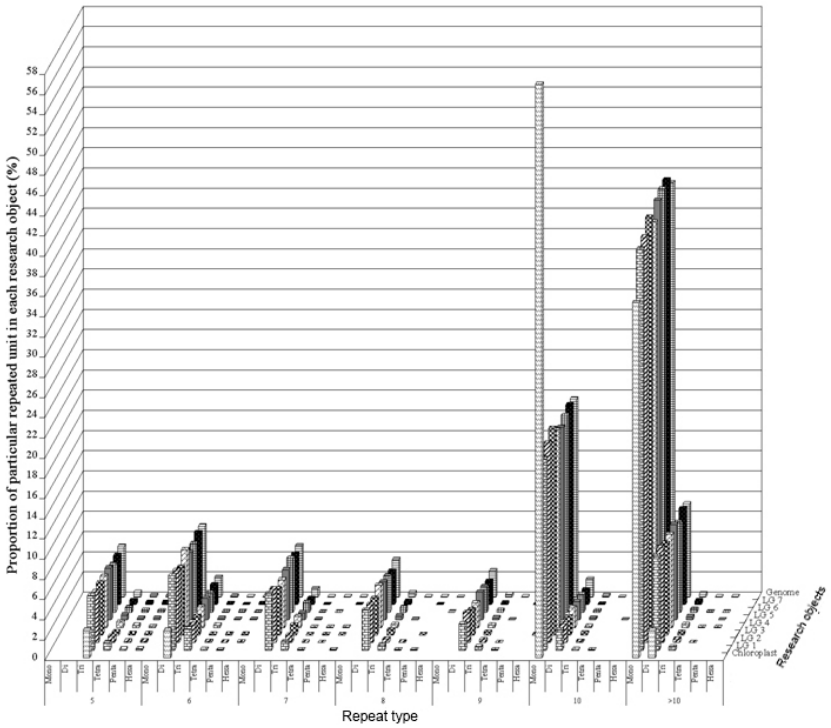

Figure 3. Proportion of particular repeated units according to their length in *Fragaria vesca*.

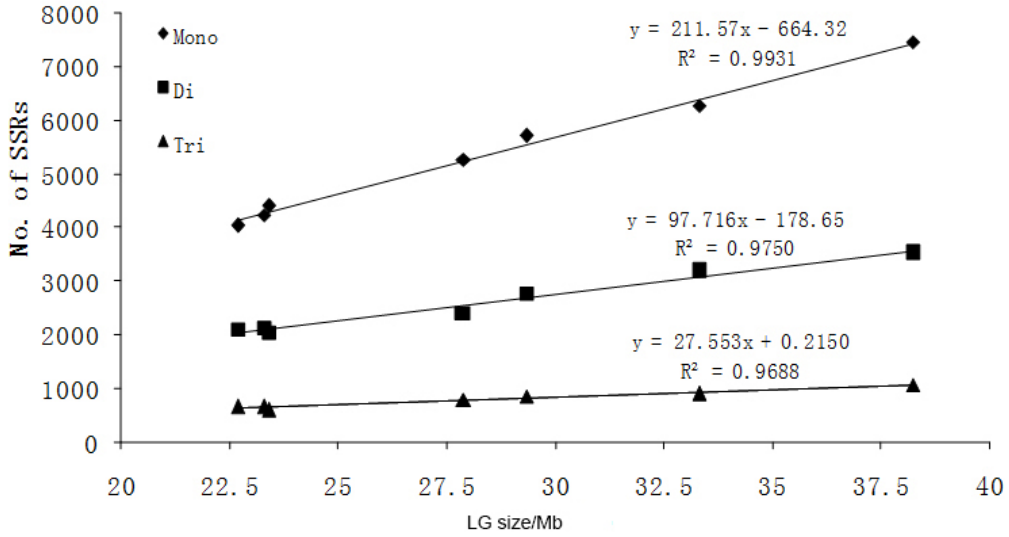

Figure 4. Correlation analysis between the linkage group (LG) size and its abundance of mononucleotide, dinucleotide and trinucleotide repeats.

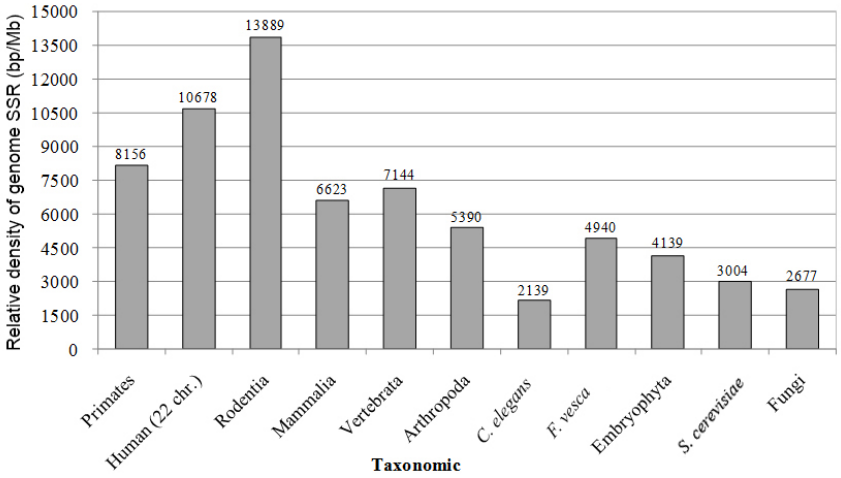

Figure 5. Relative density of perfect simple sequence repeats (SSRs) in different eukaryotic genome. Data on relative density of SSRs in other eukaryotic genome were collected from Tóth et al. (2000).

Table 3. The longest perfect SSRs of genome and chloroplast in *Fragaria vesca*.

No. of research object	Repeat type					
	Mono	Di	Tri	Tetra	Penta	Hexa
LG 1	(A)36	(AC)102	(ATT)36	(TACG)10	(AAAAT)8	(CTCAGG)7
	(T)35	(TC)71	(AGA)19	(AAAG)9	(TCAAA)6	(ATAATC)5
	(A)34	(TC)64	(ATA)18	(AAAT)8	(TCAAA)6	(ATAATC)5
	(A)34	(GA)61	(AAT)16	(CTCA)7	(TCTCT)5	(ATAATC)5
LG 2	(A)40	(CT)72	(AAT)22	(TATG)13	(CGTTA)7	(AAAGCA)6
	(A)37	(CT)64	(AAT)19	(AAAC)9	(GTCCC)6	(AAAGCA)6
	(C)34	(CT)63	(ATT)17	(CTTT)8	(GTCCC)6	(AAAGCA)6
	(C)34	(AG)59	(AAG)16	(CTTT)8	(TCCCG)5	(AAAGCA)6
LG 3	(A)39	(CT)93	(ATT)22	(AAAG)7	(GTAAC)8	(TCATCT)9
	(T)36	(TG)77	(AAT)19	(AAGA)6	(TCCCG)7	(AAATCC)7
	(T)36	(AG)72	(AAG)16	(AAGA)6	(TCCCG)7	(TTGATT)6
	(A)35	(CT)64	(AAG)16	(AAGA)6	(TCCCG)7	(TTGATT)6
LG 4	(T)38	(GA)64	(TAA)18	(TCTT)12	(ACGGG)13	(AGTGCT)7
	(A)36	(CT)62	(TAA)18	(TTTC)7	(GAAAA)7	(CCCTCT)5
	(A)36	(AG)58	(TAA)18	(TTTC)7	(TTTTA)5	(CCCTCT)5
	(A)36	(CT)56	(ATT)17	(AAAG)6	(TTTTA)5	(CCCTCT)5
LG 5	(T)42	(AG)63	(TAA)23	(CTTT)9	(ACGGG)11	(AAACAA)6
	(T)38	(CT)60	(ATA)21	(CAAA)8	(GTAAC)10	(AAACAA)6
	(T)37	(CT)60	(GAA)19	(CAAA)8	(TTTCC)6	(AAACAA)6
	(A)36	(CT)60	(CTT)18	(AGAA)7	(TTTCC)6	(AAACAA)6
LG 6	(A)44	(AG)74	(TTC)26	(TACG)35	(TCCCG)11	(CCGTCC)8
	(A)42	(CT)63	(CAT)22	(TACG)18	(AAAAT)7	(GCAAGT)6
	(A)42	(CT)63	(TCT)18	(ATAC)7	(TTTTC)6	(GCAAGT)6
	(T)40	(CT)60	(AGA)17	(ATAC)7	(TTTTC)6	(GCAAGT)6
LG 7	(A)46	(GA)70	(CTT)34	(TTTC)7	(AGAAA)9	(ACGGAC)7
	(T)41	(GA)67	(CTT)34	(TTAA)6	(TACGT)8	(CTTGAA)6
	(A)34	(GA)62	(TCT)26	(TTAA)6	(TTTAT)5	(CTTGAA)6
	(A)34	(CT)61	(TCT)23	(TTAA)6	(TTTAT)5	(CTTGAA)6
Chloroplast	(A)29	(TA)13	(TTA)5	-	-	-
	(A)22	(TA)6	-	-	-	-
	(T)16	-	-	-	-	-
	(T)13	-	-	-	-	-

Mononucleotide repeats

Mononucleotide repeats were the most frequently found SSR type in all LGs and in the chloroplast (Table S1). Our study revealed an overrepresentation of A/T (4523 in total) relative to C/G repeats (Table S2), and the proportions of A/T repeat motifs in the contents of each LG were all higher than 90%, with the maximum and the minimum proportions located in LG 7 (96.47%) and LG 4 (95.90%), respectively. The perfect mononucleotide repeat abundance was 37,267 (60.60%), with LG 6 (7437) and LG 5 (4023) showing the highest and lowest abundances, respectively. The relative abundance of mononucleotide repeats in the whole genome was 188.11 repeats/Mb, with the highest relative abundance observed in LG 6 (194.58 repeats/Mb), and the lowest relative abundance in LG 1 (177.38 repeats/Mb). The density and relative density of mononucleotide repeats were 5.32 kb/repeat and 2401 bp/Mb, respectively. The maximum density and relative density were observed in LG 6 (5.14 kb/repeats) and LG 5 (2488 bp/Mb), respectively, and minimum values were both in LG 1, with 5.64 kb/repeat and 2259 bp/Mb, respectively. The maximum and minimum overall proportions of mononucleotide repeats were 42.06% (>10 units in LG 7) and 18.94% (10 units in LG 1), respectively (Figure 3). The longest stretches of mononucleotide repeats were of type A, and located in LG 7 [(A)46]. Overall, it was observed that A/T repeats were much longer and more frequent than C/G repeats (Table 3 and Table S1).

Dinucleotide repeats

As molecular markers, dinucleotide repeats are generally more useful than other SSRs and are particularly valuable due to their high mutation rates. We demonstrated that AT/TA repeats are predominant in the *F. vesca* genome, with a systematic decrease observed in the frequencies of CT/TC, AG/GA, GT/TG, AC/CA, and CG/GC repeats (Table S3). This might be due to the high content of A/T repeats observed in individual LGs (Table S1). The abundance of perfect dinucleotide repeats was 18,108 (29.45% of all SSRs). LG 6 had the highest (3507), and LG 7 (2034) the lowest, abundance of dinucleotide repeats. However, the highest relative abundance of dinucleotide repeats was observed in LG 2 (95.35 repeats/Mb) and the lowest was observed in LG 3 (86.12 repeats/Mb), with an average value of 91.40 repeats/Mb among all LGs. The density distribution of dinucleotide repeats among LGs was stable, with an average of 10.94 kb/repeat and the highest and lowest densities were observed in LG 2 (10.49 kb/repeat) and LG 3 (11.61 kb/repeat), respectively. The relative density of dinucleotide repeats was also similar among LGs, with an average of 1956 bp/Mb. The maximum proportion of dinucleotide repeats was 9.48 (>10 units in LG 7), and the minimum proportion was 1.48% (10 units in LG 7) (Figure 3). LG 5 (2059 bp/Mb) and LG 7 (1859 bp/Mb) had the highest and the lowest relative densities, respectively. The longest dinucleotide repeat was found to be (AC)102 in LG 1 (Table 3), and the most frequent types of dinucleotide repeats were extremely similar among different LGs (Table S1).

Trinucleotide repeats

In general, the abundance of perfect trinucleotide repeats was 5460, representing 8.86% of all SSRs. LG 6 contained the highest number (1053) of trinucleotide repeats, where-

as LG 7 showed the lowest content (593). The relative abundance of trinucleotide repeats was 27.56 repeats/Mb, with the highest value observed in LG 1 (28.75 repeats/Mb), and the lowest in LG 7 (25.34 repeats/Mb). The density and relative density distributions were stable among all LGs and their averages were 36.28 kb/repeat and 499 bp/Mb, respectively. The highest density (34.79 kb/repeat) and relative density (522 bp/Mb) values were both found in LG 1, whereas the lowest density (39.46 Kb/repeat) and relative density (462 bp/Mb) values were both found in LG 7. The maximum and minimum proportion of trinucleotide repeats was 5.38% (5 units in LG 1) and 0.07% (10 units in LG 1), respectively (Figure 3). We also concluded that the longest trinucleotide repeats were found in LG 1 [(AAT)36], which was also rich in A/T repeats (Table 3). A-containing or T-containing trinucleotide repeats were also the most frequently observed SSR motifs, with the highest frequency trinucleotide repeats being AAG/CTT, with an abundance of 2423 (Table S1), representing 44.38% of the total.

Tetranucleotide repeats

The abundance, relative abundance, density, and relative density of tetranucleotide repeats were found to be much lower than those of mononucleotide, dinucleotide or trinucleotide repeats, and their values fluctuated widely among the different LGs. The highest and the lowest abundances were observed in LG 2 (82) and LG 3 (34), respectively. The average number of tetranucleotide repeats in the whole genome was 401 (0.65% of all perfect repeats in the genome). However, the highest and the lowest relative abundances were found in LG 1 (2.65 repeats/Mb) and LG 3 (1.33 repeats/Mb), respectively, with an average of 2.02 repeats/Mb. The average density of tetranucleotide repeats in the genome was 494.04 kb/repeat, with the highest (378 kb/repeat) and the lowest (820 kb/repeat) densities observed in LG 1 and LG 3, respectively. The highest and lowest relative densities of tetranucleotide repeats were 56 bp/Mb in LG 1 and 26 bp/Mb in LG 3, respectively. The maximum proportion of tetranucleotide repeats observed was 0.75% (5 units in LG 1), whereas in LG 3, LG 4, LG 5, and LG 6, there were no tetranucleotide repeats of more than 10 units (Figure 3). The longest tetranucleotide repeats were A and T rich, and the longest type of tetranucleotide repeat was (TACG)35 in LG 6 (Table 3). The most frequent tetranucleotide repeats were all low in abundance and the most common tetranucleotide repeat was found to be AAAT/ATTT, with a total of 179 (Table S1).

Pentanucleotide and hexanucleotide repeats

As expected, the abundance, relative abundance, density, and relative density of pentanucleotide and hexanucleotide repeats were all lower than those of shorter repeated units, and the average values of these 2 repeated types were similar among all LGs. Average values of pentanucleotide and hexanucleotide repeats were 133 and 126, respectively, relative abundances were 0.67 repeats/Mb and 0.64 repeats/Mb, respectively, densities were 14,890 kb/repeat and 1572 kb/repeat, respectively, and relative densities were 19 bp/Mb and 21 bp/Mb, respectively. We found that the highest values of abundance (31), relative abundance (0.81 repeats/Mb), density (1233 kb/repeats), and relative density (22 bp/Mb) of pentanucleotide repeats were all in LG 6, whereas the lowest values of abundance (11), relative abundance (0.49 repeats/Mb), density (2062 kb/repeats), and relative density (13 bp/Mb) were all in LG 1. With respect to hexanucleotide repeats, the highest and the lowest averages were found in

LG 7 and LG 4, respectively. The highest and the lowest values of abundance, relative abundance, density, and relative density were, respectively: 25 and 6, 1.07, and 0.26 repeats/Mb; 936 and 3882 kb/repeat; and 35 and 8 bp/Mb. The maximum proportion of pentanucleotide repeats observed overall was 0.21% (5 units in LG 7) and hexanucleotide repeats of 8 units or more were seldom observed (Figure 3). There was no particular pattern in the composition of the longest pentanucleotide repeats, with the longest type being (ACGGG)13 in LG 4. The composition of the longest hexanucleotide repeats was similar among all LGs, with the longest being (TCATCT)9 in LG 3 (Table 3). Based on these results, pentanucleotide and hexanucleotide repeats appear to be rare and complicated in the *F. vesca* genome. Due to their low application as SSR markers, their frequency patterns are not shown here.

DISCUSSION

Relationship between SSR abundance and genome size

Generally, because eukaryotes have larger genomes, SSRs in the genome sequence might be relatively diluted. Some studies have shown that the number and diversity of SSRs in the genome are highly correlated with increasing genome size. That is, mutation of SSR types results in increased microsatellite diversity (Guo, 2004). In addition, many other researchers have demonstrated that the distribution of specific repeats is particular to different genome regions (such as introns, exons, and intergenic regions). For example, in all vertebrates, the microsatellite distribution of introns and intergenic regions was quite similar, but the abundance of CCG triplets varied across regions; introns do not contain this type of repeat, whereas it is relatively abundant in intergenic regions, and is one of the most highly abundant repeats in exons (Tóth et al., 2000). Cai et al. (2009) also demonstrated that the distribution of the same repeats varies among different regions of genes. They further provided a new method for applying SSR data across different species. However, others argue that whole-genome analyses of SSRs would be more informative to study their distributions or locations (Li et al., 2007). In this study, we examined SSRs at the genome-wide scale in order to provide useful information for the development of polymorphic SSR markers in *F. vesca*.

Previous studies have demonstrated the abundance of SSRs in plant genomes to be 5 times lower than that of mammals (Lagercrantz et al., 1993). Furthermore, comparison of SSR abundances among different species could be informative about evolutionary history and relationships. Although the abundance of SSRs does not strictly increase with the evolutionary status of a species, there is nonetheless a positive correlation between SSR abundance and genome size. In analyses of SSRs among several species, ranging from fungi to mammals, 2 other general trends emerged. First, SSR abundance was higher in species of higher evolutionary positions, and second, SSR abundances were more similar among closely related species across several taxa, from rodents to humans (Guo, 2004).

The most frequent and the longest perfect SSRs

Tóth et al. (2000) found that the distribution of each class of SSRs (from mononucleotide to hexanucleotide repeats) was variable among humans and 10 other eukaryotic species. For example, (GT)n was found to be the most frequent repeat in mammals, whereas in plants

and vertebrates, the dominant repeats were (AT)n and (CT)n, respectively (Lagercrantz et al., 1993; Paxton et al., 1996). Our results revealed that mononucleotide repeats were the most frequent SSRs in the *F. vesca* genome (such as A/T, representing 52.92% of total SSRs), followed by dinucleotide repeats (AT/TA, 14.43%), which is in stark contrast to results from previous studies (Cardle et al., 2000; Varshney et al., 2002; Rabello et al., 2005). This notable difference with previous results might be due to the fact that previous genome information was not complete, or that molecular hybridizations were the only methods previously available. Such limitations have made comprehensive statistical analyses of mononucleotide repeats in the whole genome difficult, and the potential bias introduced by errors in distinguishing exons from intergenic regions could not be ignored (Tóth et al., 2000). In the present study, comparisons across different LGs indicated large variations in the distribution of the longest repeats from mononucleotides to hexanucleotides (Table 3). This phenomenon might imply that the same repeat type of SSR in different LGs might have different abilities in driving evolution. For example, different abundances of the same type of trinucleotide repeats across different LGs would lead to quite different expressions of its encoded protein. This variation might also be a consequence of various protein functions or different selection pressures during a species' evolutionary history, or a result of codon bias across LGs. In brief, this non-random distribution may be a result of mutability differences or repair efficiency bias in the mismatch repair system, which might lead to the overexpression of SSRs (Harr and Schlötterer, 2000).

In this study, dinucleotide and trinucleotide repeats tended to be longer than the other types of SSRs (Table S1). The motif (AT/TA)n was found to be one of the most frequent repeats, whereas the longest repeats were observed in the AC/TC/AG groups. This observation was similar to results of a previous study that compared abundances among long SSRs in introns and intergenic regions of higher taxa, and found that AC, AG, and AT dinucleotide repeats showed striking dominance (Tóth et al., 2000). Overall, we confirmed that when developing polymorphic primers for genetic analysis, repeats containing AC/TC/AG groups should be used preferentially, and the palindrome formation, especially induced by AT/TA, should be avoided.

Relationship between evolution and repeated unit length

The relative lack of longer repeated units observed in the *F. vesca* genome could possibly be explained due to their downward mutation bias and short lifespan (Harr and Schlötterer, 2000). However, the high content of shorter repeated unit lengths generally indicates that a species is of a relatively higher evolutionary level. Similarly, according to the high number of shorter repeats in the apple genome, a previous analysis has confirmed that its genome has a long evolutionary history, or represents a high evolution level (Guan et al., 2011).

Our results confirmed that the SSRs' length per Mb was negatively correlated with increases in the repeated unit length (Figure 2). This was also consistent with results of previous studies that investigated the relationship between SSR abundance and repeated unit length (Samadi et al., 1998; Tóth et al., 2000). Using bioinformatic analysis, a systematic exploration of the relationship between the repeated unit length and mutation frequency in humans and 26 other eukaryotic species revealed that shorter repeated units are more frequent than longer repeated units, and also occur more frequently in highly recombinant genome regions. This was particularly evident with respect to the abundance of dinucleotide and trinucleotide

repeated units, although, the opposite trend was observed for marine and terrestrial organisms (Guo, 2004). This may indicate that SSRs have played an important role in genome evolution, and the processes responsible for SSR generation and fixation may have undergone alterations during evolution (Tóth et al., 2000).

Since the discovery of the polymorphic nature of repeats, SSRs have become one of the most important molecular markers for genetic studies. SSR abundance and repeated unit length might be good indicators of repeat mutations. The length distribution of all SSRs indicated that the abundance of repeats decreased rapidly with increases in the repeated unit length (Figure 2). Results from a screen of the rice genome suggested that it contained approximately 5700-10,000 microsatellites, with the relative frequency of different repeats decreasing with increasing size of the repeated unit (McCouch et al., 1997). SSR length has also been shown to be relatively longer in higher organisms; lower eukaryotes, like fungi, have shorter SSRs relative to those of higher organisms (Figure 5). In addition, a study comparing different fungal species demonstrated that larger genomes have longer SSRs than smaller genomes (Karaoglu et al., 2005). This might be due to higher mutation rates and less stable structures in longer repeats. As a consequence, longer repeats have drawn more attention with respect to their application in genetic research (Lim et al., 2004; Karaoglu et al., 2005). Here, 2 different patterns were observed with respect to the general relationship between the proportion of SSR lengths and the number of repeated units. First, there were well-defined decay curves when moving from trinucleotide to hexanucleotide repeats, and second, this trend was not observed in either mononucleotide or dinucleotide repeats, which is in concordance with the results of Varshney et al. (2002) and Rabello et al. (2005).

In conclusion, our study of SSRs in the completely sequenced *F. vesca* nuclear and chloroplast genomes contributes a small step towards a better understanding of the nature of these important sequences. The data provided here about SSR composition, length and distribution could be practical for choosing optimal repeat motifs for SSR isolation in the *F. vesca* genome, as well as for other relevant crops. The analyzed SSR data should also facilitate further research about genome organization.

ACKNOWLEDGMENTS

Research supported by the Jiangsu Agriculture Science and Technology Innovation Fund (JASTIF), CX (12) 2014.

Supplementary material

REFERENCES

Cai B, Li CH, Yao QH, Zhou J, et al. (2009). Analysis of SSRs in grape genome and development of SSR database. *J. Nanjing Agric. Univ.* 32: 28-32.

Cardle L, Ramsay L, Milbourne D, Macaulay M, et al. (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156: 847-854.

Castillo A, Budak H, Varshney RK, Dorado G, et al. (2008). Transferability and polymorphism of barley EST-SSR markers used for phylogenetic analysis in *Hordeum chilense*. *BMC Plant Biol.* 8: 97.

Cruz F, Pérez M and Presa P (2005). Distribution and abundance of microsatellites in the genome of bivalves. *Gene* 346: 241-247.

Darrow GM (1966). The Strawberry: History, Breeding and Physiology. Holt, Rinehart and Winston, New York, USA.

Guan L, Zhang Z, Wang X, Xue H, et al. (2011). Evaluation and application of the SSR loci in apple genome. *China Agr. Sci*. 44: 4415-4428.

Guan L, Huang J, Liu J, Gao Z, et al. (2012). Development of polymorphic SSR primers for pear based on apple genome. *Acta Bot. Boreal.-Occident. Sin.* 32: 48-53.

Guo WJ (2004). Primary Research on the Microsatellite Distribution and Function in Genomes and the Relevant Computational Methodology. In: Crop Genetics and Breeding Sichuan Agricultural University, Sichuan, China.

Harr B and Schlötterer C (2000). Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* 155: 1213-1220.

Karaoglu H, Lee CM and Meyer W (2005). Survey of simple sequence repeats in completed fungal genomes. *Mol. Biol. Evol.* 22: 639-649.

Kruglyak S, Durrett R, Schug MD and Aquadro CF (2000). Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. *Mol. Biol. Evol.* 17: 1210-1219.

Künzler P, Matsuo K and Schaffner W (1995). Pathological, physiological, and evolutionary aspects of short unstable DNA repeats in the human genome. *Biol. Chem. Hoppe Seyler* 376: 201-211.

Lagercrantz U, Ellegren H and Andersson L (1993). The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucl. Acids Res.* 21: 1111-1115.

Levinson G and Gutman GA (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4: 203-221.

Li S, Yin T, Wang M and Tuskan GA (2011). Characterization of microsatellites in the coding regions of the *Populus* genome. *Mol. Breed.* 27: 59.

Li W, Chen HG, Li W, Zhang AX, et al. (2007). Analysis of simple sequence repeats in genomes of *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *Yi Chuan* 29: 1154-1160.

Lim S, Notley-McRobb L, Lim M and Carter DA (2004). A comparison of the nature and abundance of microsatellites in 14 fungal genomes. *Fungal Genet. Biol.* 41: 1025-1036.

McCouch SR, Chen X, Panaud O, Temnykh S, et al. (1997). Microsatellite marker development, mapping and applications in rice genetics and breeding. *Plant Mol. Biol.* 35: 89-99.

Messier W, Li SH and Stewart CB (1996). The birth of microsatellites. *Nature* 381: 483.

Moxon ER and Wills C (1999). DNA microsatellites: agents of evolution? *Sci. Am.* 280: 94-99.

Paxton RJ, Thoren MPA, Tengö J and Estoup A, et al. (1996). Mating structure and nestmate relatedness in a communal bee, *Andrena jacobi* (Hymenoptera, Andrenidae), using microsatellite. *Mol. Ecol.* 5: 511-519.

Plaschke J, Ganal MW and Röder MS (1995). Detection of genetic diversity in closely related bread wheat using microsatellite markers. *Theor. Appl. Genet.* 91: 1001-1007.

Rabello E, Souza AN, Saito D and Tsai SM (2005). *In silico* characterization of microsatellites in *Eucalyptus* spp.: Abundance, length variation and transposon associations. *Genet. Mol. Biol.* 28: 582-588.

Richards RI and Sutherland GR (1992). Dynamic mutations: a new class of mutations causing human disease. *Cell* 70: 709-712.

Robinson AJ, Love CG, Batley J and Barker G, et al. (2004). Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics* 20: 1475-1476.

Samadi S, Artiguebielle E, Estoup A, Pointier JP, et al. (1998). Density and variability of dinucleotide microsatellites in the parthenogenetic polyploid snail *Melanoides tuberculata*. *Mol. Ecol.* 7: 1233-1236.

Schug MD, Mackay TF and Aquadro CF (1997). Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nat. Genet.* 15: 99-102.

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43: 409-116.

Shulaev V, Korban SS, Sosinski B, Abbott AG, et al. (2008). Multiple models for Rosaceae genomics. *Plant Physiol.* 147: 985-1003.

Tautz D, Trick M and Dover GA (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322: 652-656.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, et al. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* 11: 1441-1452.

Tóth G, Gaspari Z and Jurka J (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10: 967-981.

Varshney RK, Thiel T, Stein N, Langridge P, et al. (2002). *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol. Biol. Lett.* 7: 537-546.

Weber JL (1990). Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* 7: 524-530.

Wierdl M, Dominska M and Petes TD (1997). Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 146: 769-779.

Xu X, Peng M and Fang Z (2000). The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* 24: 396-399.

Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, et al. (2004). Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. *Theor. Appl. Genet.* 109: 451-463.