Theor Appl Genet (2012) 124:87–96
DOI 10.1007/s00122-011-1689-2

ORIGINAL PAPER

Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (*Vaccinium macrocarpon* Ait.)

H. Zhu · D. Senalik · B. H. McCown · E. L. Zeldin · J. Speers · J. Hyman · N. Bassil · K. Hummer · P. W. Simon · J. E. Zalapa

Received: 20 April 2011 / Accepted: 17 August 2011 / Published online: 9 September 2011

Abstract The American cranberry (*Vaccinium macrocarpon* Ait.) is a major commercial fruit crop in North America, but limited genetic resources have been developed for the species. Furthermore, the paucity of codominant DNA markers has hampered the advance of genetic research in cranberry and the Ericaceae family in general. Therefore, we used Roche 454 sequencing technology to perform low-coverage whole genome shotgun sequencing of the cranberry cultivar 'HyRed'. After de novo assembly, the obtained sequence covered 266.3 Mb of the estimated 540–590 Mb in cranberry genome. A total of 107,244 SSR loci were detected with an overall density across the genome of 403 SSR/Mb. The AG repeat was the most frequent motif in cranberry accounting for 35% of all SSRs and together with AAG and AAAT accounted for 46% of all loci discovered. To validate the SSR loci, we designed 96 primer-pairs using contig sequence data containing perfect SSR repeats, and studied the genetic diversity of 25 cranberry genotypes. We identified 48 polymorphic SSR loci with 2–15 alleles per locus for a total of 323 alleles in the 25 cranberry genotypes. Genetic clustering by principal coordinates and genetic structure analyzes confirmed the heterogeneous nature of cranberries. The parentage composition of several hybrid cultivars was evident from the structure analyzes. Whole genome shotgun 454 sequencing was a cost-effective and efficient way to identify numerous SSR repeats in the cranberry sequence for marker development.

Communicated by A. Schulman.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1689-2) contains supplementary material, which is available to authorized users.

H. Zhu · B. H. McCown · E. L. Zeldin
Department of Horticulture,
University of Wisconsin, Madison, WI, USA

D. Senalik · P. W. Simon · J. E. Zalapa (✉)
Vegetable Crops Research Unit, Department of Horticulture,
University of Wisconsin, USDA-ARS, Madison, WI, USA
e-mail: jezalapa@wisc.edu; Juan.Zalapa@ars.usda.gov

J. Speers · J. Hyman
Biotechnology Center, DNA Sequencing Facility,
University of Wisconsin, Madison, WI, USA

N. Bassil · K. Hummer
National Clonal Germplasm Repository,
USDA-ARS, Corvallis, OR, USA

Introduction

In the recent years, advances in molecular biology have allowed the development of high-throughput DNA sequencing technology. Large-scale sequencing studies resulting in complete genomic sequences have been performed in plants ranging from the 125 Mb *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000) to the 2.5 Gb *Zea mays* (Schnable et al. 2009). One application of this new technology in plants is the possibility of rapid and cost-effective discovery of unlimited numbers of molecular markers (Abdelkrim et al. 2009; Allentoft et al. 2009; Santana et al. 2009; Tangphatsornruang et al. 2009; Csencsics et al. 2010; Saarinen and Austin 2010; Perry and Rowe 2011). In particular, Roche 454 pyrosequencing can generate ~500 Mb of sequence data per run, with read lengths reaching 400–500 bp (Perry and Rowe 2011). This capability increases the likelihood of isolating sequences containing usable simple sequence repeats (SSRs) for genetic studies. SSRs are tandem, repetitive DNA sequences with a basic motif of less than six base pairs.

SSR markers are particularly useful in genetic studies because they are easy to use, highly reproducible, multiallelic, codominant, relatively abundant, and widely distributed in most plant genomes (Zalapa et al. 2008).

The North American cranberry (*Vaccinium macrocarpon* Ait.; $2n = 2x = 24$) is a diploid, perennial, woody plant in the Ericaceae family. The genome size of cranberry has been estimated to be 540–590 Mb, about five times the size of *Arabidopsis thaliana* (Costich et al. 1993). Cranberries are highly prized for their tart flavor and nutritional and medicinal attributes (Eck 1990). In the recent years, the demand for cranberry products has increased due to the presence of abundant quantities of antioxidants and other phytochemicals in the cranberry fruit (e.g., anthocyanins, flavonols, and proanthocyanidins; Wang and Stretch 2001; Kalt 2002). Although a fair amount of effort has been devoted to the development and testing of clonal cultivars (McCown and Zeldin 2003), little information is available regarding genetic aspects of the species. Random amplified polymorphic DNA (RAPD) and sequenced characterized amplified region (SCAR) markers have been used to assess the genetic diversity in wild cranberry populations (Stewart and Excoffier 1996; Debnath 2007) and for DNA fingerprinting of clonal cultivars (Novy et al. 1994, 1996; Novy and Vorsa 1995; Polashock and Vorsa 2002). More recently, SSR markers developed in blueberry have been tested and cross amplified in cranberry (Rowland et al. 2003; Boches et al. 2005; Bassil et al. 2009).

The primary motivation for the development of molecular markers in cranberry is their potential use in marker-assisted selection (MAS) during breeding. A prerequisite for MAS is the development of abundant genome-wide molecular markers. Next-generation sequencing technologies offer the possibility of generating large numbers of SSR markers in cranberry faster and at a lower cost compared to library-based methods (Abdelkrim et al. 2009; Allentoft et al. 2009; Tangphatsornruang et al. 2009; Santana et al. 2009; Cavagnaro et al. 2010; Csencsics et al. 2010; Saarinen and Austin 2010; Perry and Rowe 2011). The objectives of this study were to (1) use Roche 454 pyrosequencing technology to conduct a large-scale genome-wide characterization of SSR loci; (2) test the amplification of a subset of primer-pairs and search for polymorphic loci; and (3) assess the usefulness of SSR loci for cultivar characterization. This study is the first large-scale characterization attempt of SSR markers in cranberry. We expect the derived SSR loci will be immediately useful for cultivated germplasm characterization and DNA fingerprinting, genetic diversity and population structure of wild populations, and linkage map construction and QTL analysis and other studies in the Ericaceae family.

Materials and methods

Plant materials and DNA isolation

To isolate SSR loci, a single micropropagated cranberry 'HyRed' individual (McCown and Zeldin 2003) was used for shotgun sequencing of the cranberry genomic DNA using the Roche 454 GS FLX Titanium platform sequencing technology (Roche 454 Life Sciences, Branford, CT, USA). 'HyRed' is an early, high fruit color hybrid cultivar developed from crosses between widely grown cultivars 'Stevens' and a derivative of 'Ben Lear' (Online Resource 1). We also examined the genetic diversity of 25 cranberry cultivars/accessions obtained from the US Department of Agriculture–Agricultural Research Service (USDA-ARS) National Clonal Germplasm Repository (NCGR; Corvallis, OR, USA) and from collections preserved at the University of Wisconsin-Madison and Rutgers University (Dana 1983; Online Resource 1). Young leaves from single individuals (a single upright) were collected and freeze-dried for 72 h. using a BenchTop lyophilizer (Virtis Inc., Gardiner, NY, USA). DNA was extracted using a DNeasy kit (Qiagen, Valencia, CA, USA), and concentrations were measured in a Turner Quantech Fluorometer (Barnstead, Dubuque, IA, USA).

Library preparation

A single library was prepared from 'HyRed' genomic DNA. A total of 350 ng of double stranded genomic DNA at a final concentration of 3.5 ng/μl was randomly fragmented by sonication to an average size of 600 bp using the Bioruptor XL (Diagenode, Denville, NJ, USA). Library fragment end repair and double stranded DNA adaptor ligation were performed using the GS FLX Titanium Rapid Library Preparation Kit (Roche Applied Science, Indianapolis, IN, USA). Library fragments were size selected using AMPure XP SPRI beads (Agencourt Bioscience, Beverly, MA, USA) as indicated in the GS FLX Titanium Rapid Library Preparation Method Manual. The library was quantified in a 96-well black plate using the Synergy 2 Multi-Mode Microplate Reader (Biotek, Winooski, VT, USA) at an excitation wavelength of 485 ± 20 nm and an emission wavelength of 528 ± 20 nm.

Emulsion PCR and pyrosequencing

The 'HyRed' cranberry genomic DNA library was clonally amplified via emulsion polymerase chain reaction (emPCR) using the GS FLX Titanium Lib-L LV emPCR Kit (Roche 454 Life Sciences, Branford, CT, USA) following the manufacturer's recommendations. The amount of library to be used in emPCR amplification was previously determined

using the titration method described in the GS FLX Titanium emPCR Method Manual—Lib-L SV (Roche 454 Life Sciences, Branford, CT, USA). DNA library capture was performed using a ratio of one molecule of library per DNA capture bead. After amplification, reactions were collected, emulsions were broken, and beads containing clonally amplified DNA were enriched according to the manufacturer's protocol. Enriched DNA beads were counted using the CASY Model DT cell counter (Roche 454 Life Sciences, Branford, CT, USA). Enriched DNA beads were deposited into the wells of a GS FLX Titanium Pico Titer Plate fitted with a 2-region gasket according to the manufacturer's recommendations (Roche 454 Life Sciences, Branford, CT, USA). Image analysis and signal processing were performed using GS Run Browser v2.3 (Roche 454 Life Sciences, Branford, CT, USA).

Genomic SSR isolation

From a single run, 1.67 million reads were obtained for a total 621 Mb with 372 bp average read length and 399 bp median read length. In order to detect the differences in assembly between de novo sequencing software, we independently assembled the raw reads using gsAssembler (Newbler version 2.3; Roche 454 Life Sciences, Branford, CT, USA) and CLC Genomics Workbench (version 4.6.1; CLC Bio, Aarhus, Denmark). We randomly selected 100 assembled sequences from gsAssembler and searched for equivalent contigs produced using CLC Genomics Workbench. All sampled contigs from gsAssembler had equivalent CLC Genomic Workbench contigs with at least 50% of overlapping bases and 95% identity (data not presented). Therefore, gsAssembler shotgun sequence assembly of the cranberry 'HyRed' was selected to create FASTA and FASTQ files for all contigs and unassembled singletons. For comparison purposes, the complete plastid and mitochondrial genome of carrot (*Daucus carota*; Ruhlman et al. 2006) and tobacco (*Nicotiana tabacum*; Sugiyama et al. 2005) sequences were downloaded from the National Center for Biotechnology (NCBI) database (GSS Section). Homology of 'HyRed' cranberry contigs and singletons to carrot plastid and tobacco mitochondria was performed using MUMmer3.0 (Kurtz et al. 2004). All plastid and mitochondrial sequence hits were removed from the analysis, creating a putative nuclear-only 454 sequence data set.

Detection of SSRs and primer design

Using the nuclear-only 454 sequence, a large-scale, genome-wide SSR search was performed in both contigs and singletons using MIcroSAtellite identification tool (MISA; Thiel et al. 2003). Initially, we considered both perfect and compound repeats of basic motifs ranging from 2- to 6-bp of SSRs with a minimum length of 12 bp and repeat lengths of di-6, tri-4, tetra-3, penta-3, and hexa-3. Oligonucleotide primers were designed from the SSR flanking sequences using Primer3 with the following parameters: 100–300 bp in length (optimum 250 bp) and minimum, optimum, and maximum values for primer length (bp): 18–22–25; T_m (°C): 52–55–58; GC content (%): 40–50–60. SSR locator (da Maia et al. 2008) was used to confirm the results obtained with MISA.

PCR and SSR amplification

Cranberry SSR forward primers were appended at the 5′end with the M13 sequence (5′-CACGACGTTGT AAAACGAC-3′) to allow indirect labeling of reactions. Reverse primers were appended with the sequence GTTTCTT (PIG) at the 5′ end to promote nontemplated (A) addition and to facilitate subsequent genotyping (Brownstein et al. 1996). The M13 universal primer was labeled with carboxyfluorescein (FAM) fluorescent tag. Polymerase chain reaction (PCR) was performed in 8 μL total volume using 3.5 μL, 1× JumpStart REDTaq ReadyMix (Sigma, St. Louis, MO, USA); 2 μL, 5 ng/μL genomic DNA; 1.25 μL of H_2O; 0.5 μL, 5 μM M13-FAM primer; 0.5 μL, 5 μM reverse/0.5 μM forward primer; 0.125 μL, 5 M betaine (Sigma, St. Louis, MO, USA); and 0.125 μL, 50 mg/ml BSA (CHIMERx, Milwaukee, WI, USA). Thermocycling conditions consisted of an initial melting step (94°C for 3 min), followed by 30 cycles of 94°C for 15 s, 55°C for 90 s, and 72°C for 2 min, and a final elongation step (72°C for 20 min), followed by an indefinite soak at 4°C. PCR products (2 μL) were combined with 15 μL Hi-Di formamide (Applied Biosystems, Foster City, CA, USA) and 0.5 μL of carboxy-X-rhodamine (ROX) standard (GeneFlo-625 ROX; CHIMERx, Milwaukee, WI, USA). SSR allele genotyping was performed using an ABI 3730 fluorescent sequencer (POP-6 and a 50-cm array; Applied Biosystems, Foster City, CA, USA). Alleles were scored using GeneMarker Software version 1.5 (SoftGenetics, State College, PA, USA).

SSR data analysis

We examined the genetic diversity of 25 cranberry genotypes using SSR loci (Online Resource 1). GeneAlEx 6.4 (Peakall and Smouse 2006) was used to calculate the observed (Na) and effective (Ne) number of alleles, Shannon's information index (*I*), and levels of observed (Ho) and expected (He) heterozygosity. We conducted a principal coordinates analysis (PCoA) based on genetic distances estimated between pairs of individuals as computed by GeneAlEx 6.4 (Smouse and Peakall 1999). We also used the program STRUCTURE 2.2 to calculate the degree of

ancestry (q) of each cranberry individual based on the multilocus data (Pritchard et al. 2000). We selected the option of correlated allele frequencies, a burn-in period of 50,000 steps, and 100,000 MCMC replicates; each run was replicated 10 times to ensure consistency of results.

Results

Assembly of the cranberry genomic sequence

A total of 1,539,643 (620 Mb) cleaned reads (i.e., adaptors and low-quality reads removed) of cranberry sequence were obtained from the single run of Roche 454 pyrosequencing (Table 1). The average read length was 372 bases, and the total sequence obtained covered an estimated 266.3 Mb, about half of the cranberry genome (540–590 Mb). After plastid and mitochondrial hits were removed from the analysis, a total of 1,345,703 putative *V. macrocarpon* reads were identified of which 650,427 reads were assembled into 188,792 contigs (μ length = 383 bp, covering an estimated 72.3 Mb), and 506,484 reads (μ length = 336 bp) remained as singletons for a total of 695,275 consensus genomic sequences (Table 1). The vast majority of contigs, 178,026 or 94% of all contigs, were assembled from <10 reads (26% from 2 reads, 28% from 3 reads, 16% from 4 reads, and 24% from 5–10 reads; data not presented).

Table 1 Number of cranberry (*Vaccinium macrocarpon* Ait.) 454 sequences before and after assembly

Item	Total number
Reads generated	1,670,163
Cleaned reads (620,183,270 bp; μ length = 372 bp)	1,539,634
Genomic reads	1,345,703
Contigs sequence (72,261,495 bp; μ length = 383 bp)	188,792
Singleton sequence (170,275,913 bp; μ length = 336 bp)	506,484

Discovery of cranberry SSR loci

We identified SSRs in the contig and singleton cranberry sequence data using MISA with a minimum ≥3 repeat units and/or 12 bp (Table 2). A total of 90,018 (13%) sequences out of the 695,276 consensus genomic sequences (18,069 contigs and 71,949 singletons) contained SSRs (Table 2). Ninety-four percent of the 90,018 sequences containing SSRs were ≥100 bp (Table 2), and a total of 14,221 of the sequences (16%, 2,576 contigs and 11,645 singletons; data not presented) contained more than one SSR locus.

A total of 107,244 potentially amplifiable SSR loci (21,305 loci in contigs and 85,939 loci in singletons; Table 3) were identified in the 90,018 sequences (Table 2). On average, there was one SSR locus for every 2.5 kb of the assembled 266.3 Mb of the cranberry genome. The total length of all SSR sequence was estimated at about 2.13% of the covered cranberry genome. Dinucleotide motifs were the most common SSR motif accounting for 48% of the total SSR loci discovered, and hexanucleotide motifs were the least represented repeat type with a 4% of the total SSR loci discovered (Table 3).

Frequencies and distributions of SSR types in cranberry

We examined the SSR motif distribution with regard to repeat number (Fig. 1). For all five repeat types, SSR frequency decreased sharply as the repeat number increased, and the longer SSR motifs showed more evidence of this change (Fig. 1). As a consequence, the mean repeat number in dinucleotides (10.9) was more than twice the number of trinucleotides and three times the number of tetra-, penta- and hexanucleotides (3.3, 3.2 and 3.2, respectively) (Table 3). The analysis of individual SSR types in cranberry genomic sequence revealed that some motifs were more prevalent than others in each class (Online Resource 2). For example, while the AG dinucleotide motif was dramatically overrepresented, the CG motif was the least frequent dinucleotide (Online Resource 2). In fact, the AG repeat was the most frequent motif in the cranberry genome, accounting for 35% of the total SSR loci

Table 2 Characterization of simple sequence repeats (SSRs) in genomic sequences of cranberry (*Vaccinium macrocarpon* Ait.) generated by 454 sequencing

Item	50–100 bp	101–200 bp	201–300 bp	301–400 bp	>400 bp	Total
Contig sequence	28,996	34,963	24,146	16,787	83,764	188,656
Singleton sequence	30,322	56,329	89,916	140,828	189,089	506,484
Sequence containing SSR	5,332	12,106	15,520	20,127	36,933	90,018
Designed SSR primers	335	11,056	15,500	20,123	36,932	83,946
SSR primers in contigs	13	1,552	1,807	1,609	12,491	17,472
SSR primers in singletons	322	9,504	13,693	18,514	24,441	66,474

Table 3 Repeat types in SSRs from cranberry (*Vaccinium macrocarpon* Ait.) identified by 454 sequencing

Motif length	Number of loci identified	Frequency (%)	Mean repeat number	Number of loci primers designed	Percentage SSRs suitable for primer design (%)
Di	51,315	48	10.9	44,998	88
Tri	19,831	18	4.9	19,510	98
Tetra	20,244	19	3.3	19,727	97
Penta	11,398	11	3.2	11,160	98
Hexa	4,456	4	3.2	4,359	98
Total	107,244	100	–	99,754	93

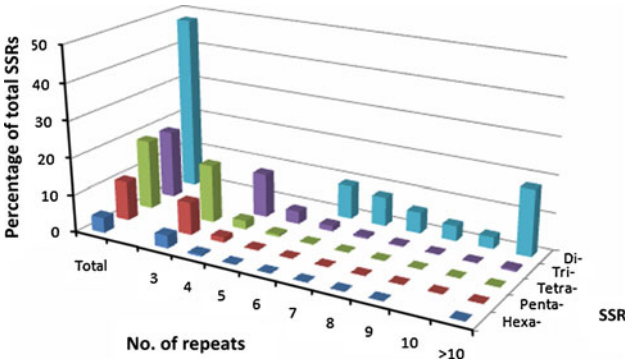

Fig. 1 Frequencies (%) of repeat types with repeat numbers in SSRs from cranberry (*Vaccinium macrocarpon* Ait.) identified by 454 sequencing

discovered (Online Resource 2). Similarly, the AAG trinucleotide motif was the most common (5% of loci discovered), and CCG was the rarest. Motifs AAAT and AAAAT were the most abundant tetra-, and pentanucleotide repeats types, accounting for 8% of the total SSR loci discovered (Online Resource 2). Analysis of the frequencies of SSR motifs revealed that AT-rich motifs were the most abundant motifs, except in dinucleotides.

Cranberry SSR primers design

The 90,018 SSR-containing sequences were screened for suitable flanking sites for PCR primer design using MISA. A total of 83,946 sequences contained suitable (93%; 17,472 contigs and 66,464 singletons) flanking sites for SSR primer design. Primers were successfully designed for a total 99,754 SSR loci within the 83,946 sequences: di-, tri-, and tetranucleotide motifs represented 84% of the primers designed (Table 3). Trinucleotides had the greatest percentage (98%) of SSR sequences suitable for primer design followed by dinucleotides with 88%.

To confirm and select a set of SSR loci for validation, we reran the SSR motif search in the 188,792 contigs (Table 1) using SSR locator (da Maia et al. 2008). This time, our goal was to isolate only perfect SSRs within the contig sequences. A multistep approach using a search for 24, 12 and 6 bp motifs allowed us to purge short repeat lengths and contigs with compound SSRs. Based on the contig sequence search results generated using SSR locator, we designed 96 primer-pairs to amplify perfect SSRs of different lengths and motif types and used them to amplify the DNA from 25 cranberry genotypes (Online Resource 1).

Validation of cranberry SSR loci

Ninety primer-pairs (94%) produced amplification products in the 25 cranberry samples included in our study. While 23 primer-pairs (24%) showed monomorphic allelic patterns, 19 primers (20%) produced allelic patterns not consistent with single locus segregation in a diploid species. The remaining 48 primers (50%) produced a maximum of two fragments of the expected size lengths per individual (Online Resource 3), and were subsequently considered single polymorphic loci. The sequences of such polymorphic SSR loci were submitted to GenBank. Since the cranberry plant material used herein (Online Resource 1) cannot be considered a natural population under random mating, we could not perform Hardy–Weinberg equilibrium (HWE) tests for each locus, and/or linkage disequilibrium test between loci. However, as precedent for future studies, the 48 polymorphic primer-pairs were used to describe cranberry SSR locus and germplasm genetic diversity (Table 4).

Genetic diversity characteristics of cranberry SSR loci

We detected 323 alleles in the 25 individual plants using the 48 loci (Table 4). The number of different alleles (Na) for each primer-pair ranged from 2 to 15, with an average of 4.8 alleles per locus. The effective number of alleles (Ne) ranged from 1.1 to 7.7, with an average of 2.8 per locus. Shannon's information index (*I*) for each primer-pair ranged from 0.2 to 2.3, with an average of 1.1. Observed heterozygosity (Ho) and expected heterozygosity (He) for each primer pair ranged from 0.04 to 0.96 and 0.08 to 0.87, respectively. Primer vm55441 had the highest diversity (Table 4).

Table 4 Genetic diversity characteristics of 25 individual cranberry (*Vaccinium macrocarpon* Ait.) genotypes based on 48 SSR loci (323 alleles) developed from 454 sequencing

Primers	N	Na	Ne	I	Ho	He
vm54133	25	5	2.6	1.2	0.64	0.62
vm54153	24	3	1.3	0.5	0.21	0.25
vm54428	24	2	1.1	0.2	0.13	0.12
vm55288	25	2	2.0	0.7	0.36	0.49
vm55441	25	15	7.7	2.3	0.96	0.87
vm59107	25	3	1.1	0.3	0.04	0.11
vm68798	25	6	2.1	1.1	0.52	0.53
vm69485	25	5	2.2	1.1	0.44	0.55
vm72062	25	5	4.1	1.5	0.64	0.76
vm78806	24	10	6.6	2.0	0.54	0.85
vm79687	21	2	1.6	0.5	0.10	0.36
vm83024	24	5	2.5	1.1	0.58	0.60
vm89040	25	6	3.2	1.4	0.56	0.69
vm01649	25	4	2.8	1.2	0.56	0.64
vm04084	25	7	4.1	1.6	0.72	0.76
vm04249	25	2	1.1	0.2	0.08	0.08
vm05418	25	2	2.0	0.7	0.32	0.50
vm07778	25	2	1.5	0.5	0.28	0.34
vm09532	15	4	1.9	0.9	0.33	0.48
vm09703	25	2	1.6	0.6	0.44	0.39
vm10462	25	2	2.0	0.7	0.24	0.49
vm12486	25	2	1.2	0.3	0.20	0.18
vm13742	24	2	1.4	0.5	0.21	0.31
vm13780	25	6	2.5	1.1	0.48	0.61
vm13884	25	4	2.9	1.1	0.56	0.66
vm21169	20	5	3.2	1.3	0.30	0.69
vm23232	25	3	1.9	0.7	0.40	0.49
vm25796	25	10	2.9	1.6	0.60	0.65
vm26877	25	8	5.2	1.8	0.60	0.81
vm27120	25	5	2.6	1.2	0.56	0.61
vm28527	25	8	3.6	1.6	0.72	0.73
vm31502	25	2	1.9	0.7	0.72	0.46
vm31701	25	7	5.6	1.8	0.72	0.82
vm32279	25	2	1.1	0.2	0.08	0.08
vm33770	25	4	2.7	1.1	0.68	0.63
vm34671	25	6	3.2	1.3	0.76	0.69
vm38401	25	9	5.0	1.8	0.72	0.80
vm39030	24	9	6.6	2.0	0.79	0.85
vm40600	25	9	3.5	1.6	0.80	0.71
vm45176	25	4	1.8	0.8	0.36	0.44
vm46208	25	4	3.9	1.4	0.72	0.75
vm48827	23	3	1.4	0.6	0.17	0.30
vm50753	25	2	1.3	0.4	0.28	0.24
vm51409	25	2	2.0	0.7	0.64	0.49
vm51985	24	7	4.4	1.6	0.75	0.77
vm52204	25	3	1.2	0.3	0.16	0.15
vm52682	24	6	4.7	1.6	0.79	0.79
vm53000	25	6	3.9	1.5	0.72	0.74
Mean	24.4	4.8	2.8	1.1	0.48	0.54

N number of genotypes examined, *Na* number of different alleles, *Ne* number of effective alleles, *I* Shannon's information index, *Ho* Observed Heterozygosity, *He* Expected Heterozygosity

Discrimination of cranberry genotypes using mined SSR loci

Based on the 48 SSR loci data PCoA clearly separated all of the cranberry individuals in our study, except for 'Ben Lear' and 'Potter's Favorite', which were genetically identical at the 323 alleles (Fig. 2a). The first principal coordinate accounted for 27.6% of the genetic variance, and the second coordinate accounted for 20.1% of the variance. We conducted structure analyzes of the 25 cranberry samples. The most likely true value of *K* identified by STRUCTURE was $K = 3$ (Online Resource 4), but we are presenting $K = 6$ since it more clearly represents some of the known relationships among several of the cultivars used in this study. Each of the 25 cranberry individuals was assigned membership probability values to the six genetic clusters based on their multilocus genotypes (Fig. 2b). Several of the samples possessed a major probability of membership to one of the six clusters, e.g., group I: 'Early Black', 'Searles', PI:618231, 'GH1', and PI:618030; group II: 'McFarlin' and 'Stevens'; group III: PI:618025 and 'Beckwith'; group V: 'Potter's Favorite' and 'Ben Lear'; and group VI: 'Howes', 'Prolific', and 'Crowley'. The rest of the cranberry cultivars/accessions showed admixed genetic patterns.

Discussion

Simple sequence repeats have traditionally been isolated de novo by constructing genomic libraries enriched for a few targeted SSR motifs and using recombinant DNA technologies for the isolation and sequencing of clones containing SSRs (Zalapa et al. 2008). However, such library-based approaches have proven time consuming and costly, rely on low-throughput sequencing, and can isolate only the targeted enriched SSR motif types. The present work demonstrates the use of emerging shotgun sequencing technologies to rapidly and cost-effectively generate large numbers of genome-wide sequences containing virtually any SSR motif type in a genetically understudied crop

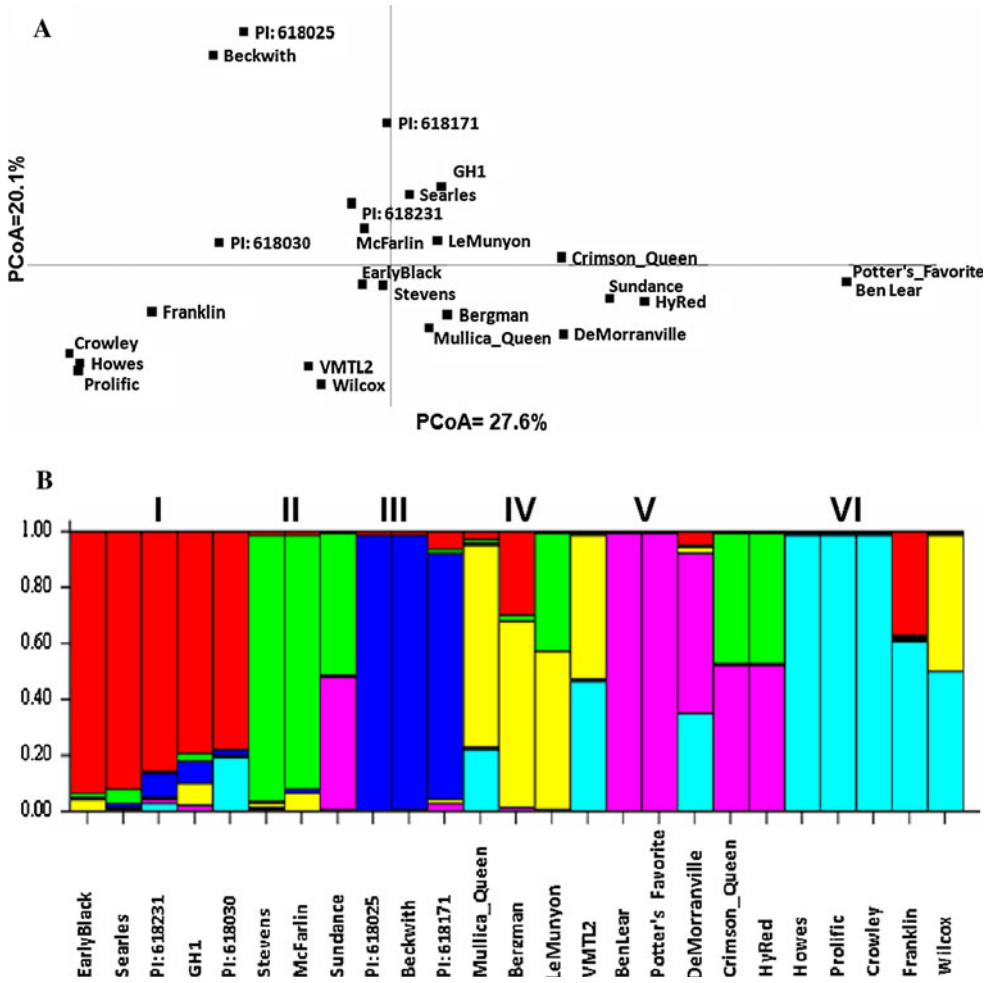

Fig. 2 Principal coordinate (**a**) and STRUCTURE (**b**) analyzes of 25 cranberry (*Vaccinium macrocarpon* Ait.) genotypes based on 48 SSR loci developed from 454 sequence data. In **b**, each *grid* represents an individual and each *color* a population ($K = 6$; Roman numerals). (color in online)

species such as cranberry, a commercially important US dominated crop.

Discovery and mining of genomic SSR loci using 454 pyrosequencing technology has had successful applications in several plant species (Tangphatsornruang et al. 2009; Cavagnaro et al. 2010; Csencsics et al. 2010). Therefore, we used Roche 454 GS FLX Titanium platform sequencing technology to generate 1,345,703 of high quality cranberry genomic reads (~620 Mb) from the single run at a cost of about US $14 per Mb. Reads were assembled de novo into 695,275 sequences (188,792 contigs and 506,484 singletons) of which 13% contained at least one SSR of 12 bp in size and 3 repeat units (Tables 1, 2). The relatively large size of the sequences (μ length = 372 bp) obtained using 454 sequencing made it computationally easy to assemble and mine the sequence data for SSRs, especially using CLC Genomics workbench (version 4.6.1) and SSR locator (da Maia et al. 2008). We identified one SSR approximately every 2.5 kb of the 266.3 Mb of the covered cranberry genome for a total of 107,244 SSR loci (Table 3). The total length of cranberry SSRs (di- to hexanucleotides) contributed about 2.13% to 266.3 Mb of the assembled sequence, which translated to an overall density across the genome of ~403 SSR/Mb. The observed SSR frequencies and densities in cranberry (Fig. 1; Online Resource 2) are consistent with those reported in other plant genomes (Cavagnaro et al. 2010). Our data is also consistent with the observation in many plant species that whole genome SSR frequency is inversely related to genome size and to the proportion of repetitive DNA, indicating that most SSRs reside in regions pre-dating the recent genome expansion (Morgante et al. 2002). Low coverage 454 pyrosequencing was a powerful method to generate potentially unlimited numbers of SSR markers for genetics studies in cranberry and the Ericaceae family.

Frequency analyzes of various nucleotide repeats in cranberry revealed that dinucleotide repeats were the most abundant SSRs followed by tetra-, tri, penta- and hexanucleotide repeats (Fig. 1; Online Resource 2). These findings are in agreement with reports of highly abundant dinucleotide repeats in the genomic DNA of many plant species (Wang et al. 1994; Tangphatsornruang et al. 2009; Cavagnaro et al. 2010; Victoria et al. 2011). Overall, AT-rich motifs were the most abundant SSRs in the cranberry

genome, except in dinucleotides (Online Resource 2). Also, similar to several understudied plant species (Yu et al. 2009), AG motifs were the most frequent SSR motifs, accounting for 35% of the total loci discovered, and together with AAG and AAAT accounted for 46% of all SSR loci in cranberry (Online Resource 2). The least frequent dinucleotide motif was GC accounting for <1% of the dinucleotides, and other GC-rich repeat SSR motifs were also rare (Online Resource 2). This result agrees with other studies indicating that genomic SSRs with GC-rich repeats are rare in many plant species (Wang et al. 1994; Tangphatsornruang et al. 2009; Cavagnaro et al. 2010). Since the isolation of SSR motifs using conventional library-based approaches is limited by cost, 454 pyrosequencing provides an inexpensive and efficient alternative for the isolation of numerous genome-wide SSR loci in cranberry, in which motif choice can have an effect on the SSR variability detected.

Sequence technology differences (Illumina vs. 454 pyrosequencing), variations in algorithms (e.g., MISA vs. SSR locator), search parameters (e.g., perfect vs. imperfect SSR), and differences in the type (e.g., genomic vs. transcribed) and sizes (i.e., genome coverage) of data sets used for SSR marker isolation can influence their detection and development (Cavagnaro et al. 2010). In contrast with Illumina that averages 50 bp read lengths, our 454 sequencing data reached an average of 382 bp in length. The longer read lengths increased our chances of successfully designing primers while making it possible to identify long SSR repeats comparable to the sizes obtained using conventional library–based approaches (Zalapa et al. 2008). In fact, 68% of the SSR-containing sequences for which primers were successfully designed were >300 bp and 86% were larger than 200 bp in length (Table 2). Since perfect (vs. compound) SSRs (Buschiazzo and Gemmell 2006) as well as longer SSR repeats (Kelkar et al. 2008) are known to exhibit greater allelic variability, we conducted an extensive search for perfect SSRs using MISA and SSR locator to isolate longer repeats and true perfect SSRs. After isolating suitable SSR sequences, we designed and tested 96 primer-pairs using 25 cranberry genotypes. Polymorphic patterns typical of diploid segregation (i.e., a maximum of two peaks per individual) were observed using 48 of the primers (50%; Table 4) that included varied repeat motifs (Online Resource 3). Moreover, all the primers were designed from sequences assembled from <20 reads (Online Resource 3), indicating that they were generated by chance from nonrepetitive DNA and likely represent single SSR loci (Lander and Waterman 1988). 454 pyrosequencing generated large amounts of genomic sequence of ideal length to develop high quality SSR markers for genetic analysis in cranberry and comparative genomics in the Ericaceae family.

The paucity of DNA markers has hampered the advance of genetic research in cranberry and the Ericaceae family. Next-generation sequencing technologies provide a viable method to produce abundant genome-wide loci for diversity, linkage mapping, marker-assisted breeding, and other studies. In the present study, we used 48 cranberry SSR loci developed using 454 sequence data to study the relationship of 25 cranberry genotypes. Our results (Table 4) confirmed the highly heterogeneous nature of cranberry (Stewart and Excoffier 1996; Debnath 2007). Recently, several cranberry cultivars were differentiated using transferrable blueberry SSR loci (Bassil et al. 2009). Our genetic clustering by PCoA analysis also differentiated the clonally propagated cranberry cultivars tested, except for 'Potter's Favorite' and 'Ben Lear' (Fig. 2a). The exact origins of 'Potter's Favorite' and 'Ben Lear' are not clear, but it is known from records from the USDA-ARS NCGR collection that both cultivars originated in Wisconsin. While 'Potter's Favorite' was collected in 1895 and is listed as a chance seedling, 'Ben Lear' is a 1900 introduction selected from the wild (USDA-ARS NCGR). Thus, it is possible that 'Potter's Favorite' and 'Ben Lear' are not unique, but additional analyzes including more samples of both cultivars will be needed to confirm their genetic identity. Thus, one of the applications of these cranberry SSR markers is DNA fingerprinting, which will likely results in the identification of genetic redundancy in cranberry germplasm and/or error during the propagation of clonal cranberry cultivars (Novy et al. 1994, 1996; Novy and Vorsa 1995; Polashock and Vorsa 2002). Our SSR marker data also point to the possibility of identifying and tracking hybrid backgrounds in cranberry breeding programs as well as identifying the existence of heterotic gene pools for the prioritizing of controlled crosses in cranberry. In this regard, parentage composition of several hybrid cultivars was clearly evident in structure analyzes, especially among newer cranberry cultivars (released after 2003), which are less likely to be contaminated genetically (Fig. 2b). For example, 'HyRed', 'Sundance', and 'Crimson Queen' proved to be first generation hybrids of similar 'Stevens' × 'Ben Lear' crosses generated in different breeding programs in Wisconsin and New Jersey (Fig. 2b; Online Resource 1). Similarly, New Jersey cultivars 'DeMoranville' and 'Mullica Queen' showed parental contributions in general agreement with their progenitor's genetic background, 'Franklin' × 'Ben Lear' and ('Howes' × 'Searles') × 'LeMunyon', respectively (Fig. 2b; Online Resource 1). Further, the development of genome-wide SSR loci and their application in the characterization of each cultivar's unique genetic fingerprint and testing of inter- and intra-cultivar diversity will allow us to better assess and classify the genetic differences and purity among and within clonal cranberry cultivars.

In conclusion, whole genome shotgun 454 sequencing was a cost-effective and efficient way to identify numerous SSR repeats in the cranberry sequence for marker development. Another potential significant outcome from our study is the use of these sequences for gene identification in comparative whole genome sequence and transcriptome analyzes with other members of the Ericaceae family.

Acknowledgments The authors thank PS100, Eric Wiesman, Lisa Wasko, Beth Workmaster, Rebecca Harbut, Shawn Steffan, Jim Polashock, Nick Vorsa, and Rod Serres for their help with different aspects of this work. This research was supported by USDA-ARS (Project # 3655-21220-001-00) funding provided to J.E.Z.

References

Abdelkrim J, Robertson BC, Stanton JAL, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46:185–192

Allentoft ME, Schuster SC, Holdaway RN, Hale ML, McLay E, Oskam C, Gilbert MTP, Spencer P, Willerslev E, Bunce M (2009) Identification of microsatellites from an extinct moa species using high throughput (454) sequence data. Biotechniques 46:195–200

Bassil N, Oda A, Hummer KE (2009) Blueberry microsatellite markers identify cranberry cultivars. Acta Hort 810:181–187

Boches PS, Bassil NV, Rowland LJ (2005) Microsatellite markers for *Vaccinium* from EST and genomic libraries. Mol Ecol Notes 5:657–660

Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by TAQ DNA polymerase: primer modifications that facilitate genotyping. Biotechniques 20:1004–1010

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28:1040–1050

Cavagnaro PF, Senalik DA, Yang L, Simon PW, Harkins TT, Kodira CD, Huang S, Weng Y (2010) Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). BMC Genomics 11:569–586

Costich DE, Ortiz R, Meagher TR, Bruederle LP, Vorsa N (1993) Determination of ploidy level and nuclear DNA content in blueberry by flow cytometry. Theor Appl Genet 86:1001–1006

Csencsics D, Brodbeck S, Holderegger R (2010) Cost-effective, species-specific microsatellite development for the endangered dwarf bulrush (*Typha minima*) using next-generation sequencing technology. J Hered 101:789–793

Dana MN (1983) Cranberry cultivar list. Frt Var J 37:88–95

da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI, de Oliveira AC (2008) SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. Int J Plant Genomics 41:2696

Debnath SC (2007) An assessment of the genetic diversity within a collection of wild cranberry (*Vaccinium macrocarpon* Ait) clones with RAPD-PCR. Genet Resour Crop Ev 54:509–517

Eck P (1990) The American cranberry. Rutgers University Press, New Brunswick

Kalt W (2002) Health functional phytochemicals of fruits. Hortic Rev 27:269–315

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18:30–38

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Open source MUMmer 30 is described in: versatile and open software for comparing large genomes. Genome Biol 5:R12

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2:231–239

McCown BH, Zeldin EL (2003) 'HyRed' and early, high fruit color cranberry hybrid. HortScience 38:304–305

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Novy RG, Vorsa N (1995) Identification of intracultivar genetic heterogeneity in cranberry using silver-stained RAPDs. HortScience 30:600–604

Novy RG, Kobak C, Goffreda J, Vorsa N (1994) RAPDs identify varietal misclassification and regional divergence in cranberry (*Vaccinium macrocarpon* Ait.). Theor Appl Genet 88:1004–1010

Novy RG, Vorsa N, Patten K (1996) Identifying genotypic heterogeneity in McFarlin' cranberry: a randomly-amplified polymorphic DNA (RAPD) and phenotypic analysis. J Am Soc Hort Sci 2:210–215

Peakall R, Smouse PE (2006) GenAlEx 6: genetic analysis in excel population genetic software for teaching and research. Mol Ecol Notes 6:288–295

Perry JC, Rowe L (2011) Rapid microsatellite development for water striders by next-generation sequencing. J Hered 102:125–129

Polashock JJ, Vorsa N (2002) Development of SCAR markers for DNA fingerprinting and germplasm analysis of American cranberry. J Am Soc Hortic Sci 127:677–684

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Rowland LJ, Dhanaraj AL, Polashock JJ, Arora R (2003) Utility of blueberry-derived EST-PCR primers in related Ericaceae species. HortScience 38:1428–1432

Ruhlman T, Lee SB, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H (2006) Complete plastid genome sequence of *Daucus carota*: implications for biotechnology and phylogeny of angiosperms. BMC Genomics 7:222–234

Saarinen EV, Austin JD (2010) When technology meets conservation: increased microsatellite marker production using 454 genome sequencing on the endangered Okaloosa darter (*Etheostoma okaloosae*). J Hered 101:784–788

Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni OX, Hammond GNA, Wingfield MJ, Wingfield BD (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques 46:217–223

Schnable P, Ware D, Fulton R, Stein J, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves T, Minx P, Reily A, Courtney L, Kruchowskz S, Tomlinson C et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82:561–573

Stewart CN, Excoffier L (1996) Assessing population genetic structure and variability with RAPD data: application to *Vaccinium macrocarpon* (American cranberry). J Evol Biol 9:153–171

Sugiyama Y, Watase Y, Nagase M, Makita N, Yagura S, Hirai A, Sugiura M (2005) The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Genet Genom 272:603–615

Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, Seehalak W, Sommanas W, Tragoonrung S, Srinives P (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean [*Vigna radiata* (L) Wilczek]. BMC Plant Biol 9:137–148

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

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 106:411–422

Victoria FC, da Maia LC, de Oliveira AC (2011) In silico comparative analysis of SSR markers in plants. BMC Plant Biol 11:15

Wang SY, Stretch AW (2001) Antioxidant capacity in cranberry is influenced by cultivar and storage temperature. J Agric Food Chem 49:969–974

Wang Z, Weber JL, Zhong G, Tanksley SD (1994) Survey of plant short tandem DNA repeats. Theor Appl Genet 88:1–6

Yu JW, Dixit A, Ma KH, Chung JW, Park YJ (2009) A study on relative abundance, composition and length variation of microsatellites in eighteen underutilized crop species. Genet Resour Crop Evol 56:237–246

Zalapa JE, Brunet J, Guries RP (2008) Isolation and characterization of microsatellite markers for red elm (*Ulmus rubra* Muhl) and cross-species amplification with Siberian elm (*Ulmus pumila* L). Mol Ecol Resour 8:109–112