Plant Mol Biol (2008) 67:581–588
DOI 10.1007/s11103-008-9321-9

An overview of the apple genome through BAC end sequence analysis

Yuepeng Han · Schuyler S. Korban

Received: 6 December 2007 / Accepted: 14 March 2008 / Published online: 3 June 2008

Abstract The apple, *Malus* × *domestica* Borkh., is one of the most important fruit trees grown worldwide. A bacterial artificial chromosome (BAC)-based physical map of the apple genome has been recently constructed. Based on this physical map, a total of ~2,100 clones from different contigs (overlapping BAC clones) have been selected and sequenced at both ends, generating 3,744 high-quality BAC end sequences (BESs) including 1,717 BAC end pairs. Approximately 8.5% of BESs contain simple sequence repeats (SSRs), most of which are AT/TA dimer repeats. Potential transposable elements are identified in ~21% of BESs, and most of these elements are retrotransposons. About 11% of BESs have homology to the *Arabidopsis* protein database. The matched proteins cover a broad range of categories. The average GC content of the predicted coding regions of BESs is 42.4%; while, that of the whole BESs is 39%. A small number of BES pairs were mapped to neighboring chromosome regions of *A. thaliana* and *Populus trichocarpa*; whereas, no pairs are mapped to the *Oryza sativa* genome. The apple has a higher degree of synteny with the closely related *Populus* than with the distantly related *Arabidopsis*. BAC end sequencing can be used to anchor a small proportion of the apple genome to the *Populus* and possibly to the *Arabidopsis* genomes.

Keywords Apple · Bacterial artificial chromosome · BAC end sequences · Simple sequence repeats · Synteny

Y. Han · S. S. Korban (✉)
Department of Natural Resources and Environmental Sciences, University of Illinois, Urbana, IL 61801, USA
e-mail: korban@uiuc.edu

Abbreviations

BAC	Bacterial artificial chromosome
BES	BAC end sequence
SSR	Simple sequence repeat
TE	Transposable element
EST	Expressed seqeuence tag

Introduction

The domesticated apple, *Malus* × *domestica* Borkh., is a member of the Rosaceae family. The family consists of over 100 genera and 3,000 species, most of which are perennial trees, shrubs, and herbs (Tatum et al. 2005). The apple is self-incompatible and highly heterozygous diploid with a base chromosome number of 17. Although the apple is a diploid ($2n = 34$), it has an allopolyploid origin (Chevreau et al. 1985). The apple is not only a major economic fruit crop grown world-wide, but also serves as an important model species for functional genomics research of woody perennial angiosperms due to its relative small genome size of 750 Mb/haploid (Tatum et al. 2005).

Bacterial artificial chromosome (BAC) libraries have been extensively used in genomics research due to their large DNA inserts, high cloning efficiency, and stable maintenance of foreign DNA. In plants, BAC libraries have been constructed for a variety of species such as *Arabidopsis* (Choi et al. 1995), rice (Wang et al. 1995), maize (Yim et al. 2002), sorghum (Woo et al. 1994), soybean (Shoemaker et al. 1996; Salimath and Bhattacharyya 1999; Tomkins et al. 1999; Meksem et al. 2000), papaya (Ming et al. 2001), and apple (Vinatzer et al. 1998; Xu et al. 2001). These libraries have made invaluable contributions to plant genomic studies including map-based or positional cloning of genes, genome-wide physical map construction

(Mozo et al. 1999; Klein et al. 2000; Chen et al. 2002; Xu and Korban 2002; Shultz et al. 2006; Han et al. 2007), genome sequencing (The *Arabidopsis* Genome Initiative 2000; International Rice Genome Sequencing Project 2005), and comparative genomics (O'Neill and Bancroft 2000; Ilic et al. 2003).

Earlier, BAC end sequencing has been proposed as a viable and efficient strategy for genome sequencing projects (Venter et al. 1996). Since then, it has become an important component of genomics research efforts as BESs are very useful in genome assembly and chromosome walking. For example, BESs can serve as sequence tag connectors (STCs) for selecting minimum overlapping clones targeted for genome sequencing (Mahairas et al. 1999). BAC end sequence (BES) pairs combined with BAC-fingerprinted contigs can serve as a primary scaffold for whole-genome shotgun sequence assembly. BESs are useful for generating comparative physical maps (Larkin et al. 2003; Shultz et al. 2007b). Moreover, BESs are valuable resources for the development of genetic markers such as BAC-end sequence-based microsatellite markers (Shultz et al. 2007a). In addition, analysis of BES data can provide an overview of the sequence composition, including gene density and presence of potential transposable elements (TEs) as well as microsatellites, of an unsequenced genome (Lai et al. 2006).

Recently, we have developed a genome-wide BAC physical map of the apple (*M.* × *domestica*) (Han et al. 2007). In order to develop genetic markers to integrate the physical and genetic maps, a total of ~2,100 BAC clones, selected from 1,767 different contigs, were sequenced at both ends, and resulting in 3,744 BESs. These BAC clones were selected from different contigs, thus suggesting they were randomly distributed across the apple genome. Hence, BESs derived from these BAC clones provided a unique opportunity to gain insights into the organization of the apple genome. Here, we report on the analysis of 3,744 BESs, and focus our attention primarily on microsatellite content, repeat element composition, GC content, protein-coding regions, and comparative mapping of BAC-end sequence pairs to other sequenced plant genomes. These BESs will serve as useful resources for genetic marker development, integration of physical and genetic maps, and whole genome sequencing of the apple.

Materials and methods

Source of BAC clones and BAC end sequencing

Two complementary BAC libraries (*Bam*HI and *Hin*dIII) from apple cv. GoldRush were used. The BAC vectors for *Bam*HI and *Hin*dIII libraries were pBeloBAC11 and pIndigoBAC-5, respectively. BAC clones, picked from 384-well microplates, were inoculated in 96-deep well plates containing 1.5 ml of 2× LB medium plus 12.5 µl/ml chloramphenicol. Plates were incubated at 37°C with continuous shaking at 325 rpm for 20–24 h. BAC DNA was then isolated using a modified alkaline lysis method. BAC end sequencing was performed using an ABI Big Dye Terminator v3.1 (ABI, CA, USA), and analyzed on an ABI 3730x1 instrument. Base-calling and sequence trimming were performed with PHRED software (Ewing and Green 1998) using the default parameters. The output of sequence data was converted into a FASTA format, and vector sequences were masked. Terminal vector sequences were then trimmed, and BESs shorter than 100 bp were discarded.

Identification of simple sequence repeats

Five classes of simple sequence repeats (SSRs), including mono-, di-, tri-, tetra-, and penta-nucleotide tandem repeats, were scanned for all trimmed BESs larger than 100 bp in size. SSRs recorded for the final dataset included monomers with at least 20 repeats and dimers to pentamers with at least 15 bp in length.

Analysis of repetitive sequences

BESs were compared with The Institute for Genomic Research (TIGR) plant repeat databases (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/) using BLAST at a cut-off value of 10^{-5}. Repetitive sequences were annotated according to the best match in the repeat database, and classified based on TIGR codes for plant repetitive sequences (http://www.tigr.org/tdb/e2k1/plant.repeats/repeat.code.shtml).

Annotation

To identify protein-coding regions, BESs with no homology to the repeat sequence database were compared with the protein database of *Arabidopsis thaliana* (ftp://ftp.arabidopsis.org/home/tair/Proteins/) using BLASTX at a cut-off value of 10^{-6}. Those BESs significantly matched to the *Arabidopsis* protein database were annotated based on the original *A. thaliana* protein database annotation.

Comparative genome mapping

All pairs of BESs were compared with whole genome sequences of *Arabdopsis*, rice (*Oryza sativa*) and poplar (*Populus trichocarpa*) using TBLASTX at a cut-off value of 10^{-6}. Whole genome sequence databases of *A. thaliana*, rice, and poplar were downloaded from The National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/genomes/static/euk.html). If a pair of BESs had

significant hits, separated by at least 10 kb and not more than 300 kb in the target genome, the tiled BAC was considered to be potentially colinear with the target genome (Lai et al. 2006).

Results

BAC end sequencing

A total of 2,112 BAC clones from cv. GoldRush were sequenced at both ends. Of these BAC clones, 62.2% and 37.8% were from *Bam*HI and *Hin*dIII libraries, respectively. Following trimming and vector sequence removal, 3,744 high-quality BESs were generated (Table 1). Of these BESs, 1,717 were paired end reads. The size of these BESs ranged from 100 to 910 bp with an average of 636 bp, thus corresponding to a total length of ~2.4 Mb. The G + C content of these BESs ranged from 11% to 66% with an average of 39%.

Simple sequence repeats

A total of 320 SSRs or microsatellites were discovered within the BESs, and these contained a variety of repeat types (Table 2). Di-nucleotide repeats were the most abundant, accounting for 48.1% of all SSRs, followed by penta- and mono-nucleotide repeats which accounted for 19.4% and 17.2%, respectively. Both tetra- and tri-nucleotide repeats occurred relatively rarely and accounted for 6.9% and 8.4%, respectively, of all SSRs. Of the di-nucleotide repeats, AT/TA was the most abundant, accounting for 56.5% of all di-nucleotide repeats; while, AG/CT, TC/GA, GT/AC, and TG/CA repeats accounted for 18.8%, 11.0%, 9.1%, and 4.5%, respectively. No GC/CG repeats were found in this study. Moreover, length distribution of all SSRs indicated that the frequency of repeats decreased with repeat length (Fig. 1). Among the monomer repeats, A/T was predominant, while G/C occurred very rarely (Table 2). Thus, AT/TA dimer repeats were the most abundant SSRs in apple BESs. In addition, 19 pairs of SSRs were clustered within the same BESs. Of the paired SSRs, 4 tetramers were clustered with both dimers and trimers, and 6 pentamers were clustered with both dimers and trimers.

Table 1 Statistical information and composition of apple BAC end sequences (BESs)

Total number of BESs	3,744
No. of paired BESs	1717
No. of non-paired BESs	310
Total length (bp)	2,380,428
Minimum length (bp)	100
Average length (bp)	636
Maximum length (bp)	910
Sequence composition	
Potential transposable elements (%)	20.9
Simple sequence repeats (%)	6.5
Protein coding regions (%)	10.9
Unknown genomic sequences (%)	61.7

Table 2 Distribution of simple sequence repeats in apple BESs

Repeat	Type	No.
Monomer	A/T	54
	G/C	1
Dimer	AT/TA	87
	AG/CT	29
	GA/TC	17
	AC/GT	14
	CA/TG	7
Trimer	AAT/ATA	5
	ATT/TTA/TAT	4
	AAG/GAA	5
	CTT/TTC	4
	ACA/CAA	3
	GAT/GAC/CCT/TTG/GCG	6
Tetramer	AAAT/AATA/ATAA	3
	ATTA/TAAT/TTAT/TTTA	5
	Other	14
Pentamer	AAAAT/AATAA	7
	ATATA/TAATA	2
	TTTAT/TTTTA/TATTT	9
	ATATT/TATAT/TTATA/TTTAA	4
	Other	40
Total		320

Transposable elements

A total of 3,744 BESs were compared with the plant repeat database revealing that 786 (20.9%) BESs were homologous to TEs (Table 3). Of these potential TEs, class I transposons or retrotransposons represented the most abundant repeats, accounting for 88.2% of TEs. Whereas, class II transposons and miniature inverted repeat TEs were relatively rare, and accounting for 10.9% and 0.9%, respectively. Among the retrotransposons identified in BESs, the total number of long terminal repeat (LTR) retrotransposons, including *Ty1-copia* and *Ty3-gypsy*, were 2.8 times higher than those of non-LTR retrotransposons, such as LINE and SINE (Table 3). In addition, more than half of the retrotransposons (54.1%) and most of the transposons (70.9%) could not be clearly assigned to a specific type (Table 3).

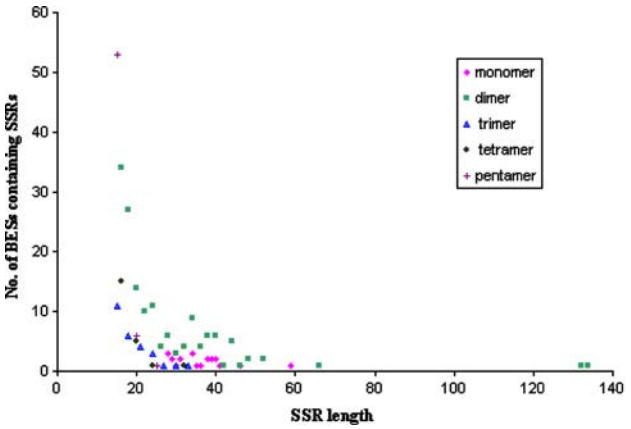

Fig. 1 Length distribution of different types of SSRs identified within apple BESs

Protein coding regions

A total of 2,958 BESs with no homology to the plant repeat database was compared with the *Arabidopsis* nucleolar protein database. Of the total BESs, 323 (8.6%) were homologous to *Arabidopsis* proteins at an E-value of $<1e^{-19}$. Functional annotation of putative gene products was then carried out using the Gene Ontology assignment of the *Arabidopsis* proteome (http://www.arabidopsis.org/tools/bulk/go/index.jsp). The predicted genes covered a broad range of functional categories, such as cellular components, metabolism, signal transduction, and response to stress (Fig. 2). With 8.6% of BESs having homologous sequences to the *Arabidopsis* protein database, this suggested that the total coding region of the apple genome was approximately 64.5 Mb, based on an estimated genome size of 750 Mb (Tatum et al. 2005). Given the assumption of an average gene length of 2 kb, similar to that of *Arabidopsis* (The *Arabidopsis* Genome Initiative 2000), the total gene content of the apple was estimated to be ~32,250. Moreover, the average GC content of the predicted coding regions of BESs was 43%.

Comparative mapping of apple BAC ends to other plant genomes

In order to gain insight into the syntenic relationships between apple and other plant species, apple BESs were BLASTed against whole genome sequences of three sequenced plants, including *A. thaliana*, poplar (*P. trichocarpa*), and rice (*O. sativa*). If paired BAC ends mapped to the target genome with a span of 10 kb to 300 kb along with proper orientation, then they were deemed potentially colinear with the target genome. A total of 894 BESs, including 107 BAC end pairs, had significant hits to the *Arabidopsis* genome. Amino acid identities of these hits ranged from 23% to 96% with an average of 49.1%. Of 107 BES pairs, 28 had the top BLAST hit to the same *Arabidopsis* chromosome and three were mapped to the *Arabidopsis* genome with a span of 69–300 kb (Table 4). Similarly, when apple BESs were compared with the *Populus* genome, 1,110 BESs, including 154 BAC end pairs, had significant matches. Amino acid identities of these matches ranged from 20% to 97% with an average of 53.3%. Among 154 BAC end pairs, 15 had the top match to the same *Populus* chromosome and eight were mapped to the *Populus* genome with a span of 12–65 kb (Table 4). Moreover, BESs of the eudicot apple were also BLASTed against the genome of the monocot model plant rice. The results revealed that a total of 907 BESs, including 106 BAC end pairs, had significant hits to the rice genome. The amino acid identities of these hits ranged from 20% to 96% with an average of 48.6%. Of 106 BES pairs, 12 had the top hit to the same chromosome. However, no pairs of apple BESs were mapped to the same rice chromosome separated by more than 10 kb and

Table 3 Summary of potential transposon contents in apple BESs

Class	Transposon type	Number of BES
Retrotransposon	*Ty1-copia*	190
	Ty3-gypsy	62
	LINE	65
	SINE	1
	Unclassified	375
Transposon	Ac/Ds	4
	CACTA, En/Spm	14
	Mutator (MULE)	3
	ping/pong/SNOOPY	4
	Unclassified	61
Miniature inverted repeat transposable element (MITE)	MITE-adh, type D	5
	Micron	2
	Total	786

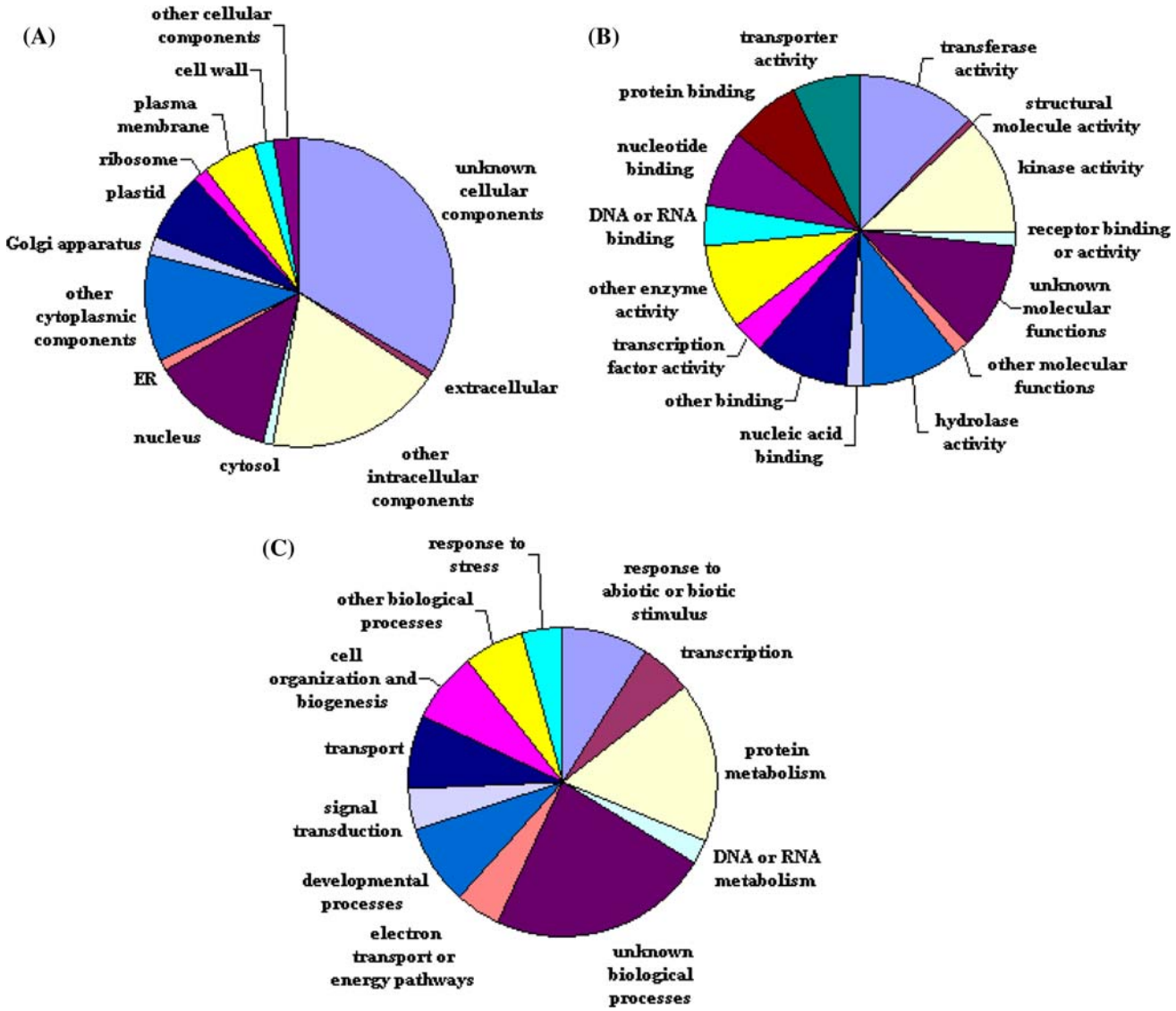

Fig. 2 Gene ontology annotation of apple BESs. (**a**) Cellular component; (**b**) Molecular function; (**c**) Biological process

less than 500 kb. This suggested that the colinearity relationship between apple and rice has heavily eroded since the divergence of eudicots from monocots.

Discussion

Analysis of BESs is an efficient approach for developing an understanding of sequence content and complexity of an unsequenced genome (Lai et al. 2006; Cheung and Town 2007). This approach relies on sequencing ends of BAC clones randomly selected from BAC libraries. In this study, we took advantage of the genome-wide BAC-based physical map of the apple, and collected a set of BAC clones. Analysis of BESs from the BAC set has provided an early glance at the apple genome before the whole genome sequence becomes available. The results presented herein indicate that the apple genome contains a large number of potential TEs and microsatellites, and it has a higher degree of colinearity with the *Populus* genome than with the *Arabidopsis* genome.

Genomic GC content is one of the most important features of a genome. Genomes with a low GC content are expected to have shorter exons than those with a high GC content (Xia et al. 2003). Based on comparisons of apple BESs with the *Arabidopsis* protein database, the average GC content of coding regions of the apple genome is ~43%, which is similar to that of the *Arabidopsis* genome (~42.7%; The *Arabidopsis* Genome Initiative 2000). Moreover, *Arabidopsis* and apple genomes represent sister clades within the dicot subclass Rosidae. Therefore, it is reasonable to assume that the average gene length of the

Table 4 Comparative mapping of paired apple BAC ends to other plant genomes

Plant species	Paired apple BAC ends	Chromosomal location	Coordinates (bp)	Span (bp)
Arabidopsis thaliana	KB01003X1C08r/KB01003X1C08f	1	5102253–4799732	302,521
	KB01009X1B11f/KB01009X1B10r	1	5166190–5061215	104,975
	KB01014X1D04r/KB01014X1D04f	2	1772675–1703340	69,335
Populus trichocarpa	KB01008X1A11r/KB01008X1A11f	VIII	226194–255183	28,989
	KB01012X1A06r/KB01012X1A06f	XVIII	1496344–1448318	48,026
	KB01013X1A08r/KB01013X1A08f	XI	1560563–1495884	64,679
	KB01013X1D08r/KB01013X1D08f	XIV	2875992–2836479	39,513
	KB01019X1A12r/KB01019X1A12f	I	38014–13132	24,882
	KB01018X1G01f/KB01018X1G01r	II	2693817–2711121	17,304
	KB01020X1C01f/KB01020X1C01r	VI	4640410–4628146	12,264
	KB01154X1D07f/KB01154X1D07r	VI	2427766–2441077	13,311

apple is similar to that of *Arabidopsis*. Based on this assumption, the total number of apples genes is predicted to be ~32,250, which is rather consistent with results obtained from analysis of our apple EST database (182,241 5′ and 3′ reads) indicating that the total gene content of apple is ~29,000 (unpublished data).

Plant genomes contain a variety of TEs such as transposons, retrotransposons, and miniature inverted-repeat TEs (MITEs). The most abundant TEs in plant genomes are retrotransposons and MITEs (Feschotte et al. 2002). In this study, TEs are identified in ~21% of apple BESs. Of these TEs, 88.2% belong to retrotransposons, thus suggesting that the apple genome consists of abundant retrotransposons. The ratio of *Ty3-gypsy* to *Ty1-copia* retrotransposons in apple BESs is 1:3, and it is different from those reported for the *Arabidopsis* (1:1; The *Arabidopsis* Genome Initiative 2000) and rice (2:1; International Rice Genome Sequencing Project 2005) genomes. Moreover, ~11.6% of BESs contain unclassified TEs (Table 3), suggesting that novel repeats constitute a significant portion of the apple genome. On the other hand, MITEs have been reported and are present in high copy numbers in the apple genome (Han and Korban 2007). However, few MITEs have been identified in apple BESs. Similarly, few MITEs have been found in papaya BESs (Lai et al. 2006). The detection of MITEs in BESs may be significantly biased by either the restriction enzyme used to generate the BAC library or the secondary structures of MITEs influencing BAC end sequencing.

SSRs constitute a special class of tandemly repeated DNA. SSRs have several advantages over other molecular markers, including high polymorphism due to the high mutation rate affecting the number of repeat units, abundance in whole eukaryotic genomes, and co-dominant inheritance (Tóth et al. 2000; Katti et al. 2001). SSRs have been extensively used for genome mapping in plants such as rice (Coburn et al. 2002; McCouch et al. 2002), maize (Sharopova et al. 2002), wheat (La Rota et al. 2005; Gao et al. 2004), and papaya (Eustice et al. 2007). BESs are useful resources for the development of SSR markers, and BAC-end sequence-based SSRs have been successfully used to develop genetic maps in cotton (Frelichowski et al. 2006) and soybean (Shultz et al. 2007a). In this study, analysis of apple BESs has revealed that 6.5% BESs contain SSRs. This suggests that the development of BES-based SSRs is a potentially feasible approach for either constructing or saturating the genetic map for apple. Moreover, the most abundant SSRs identified in apple BESs are A/T monomer and AT/TA dimer repeats. This is in agreement with previous findings indicating that AT-rich SSRs are predominant in *Arabidopsis* (Tamanna and Khan 2005), soybean (Shultz et al. 2007a), and papaya (Lai et al. 2006). In addition, most of the SSRs identified in apple BESs are 20−40 bp in length, and very few SSRs are larger than 50 bp in length (Table 1). The length distribution of apple BES-based SSRs is consistent with a previous finding that the frequency of repeats decreases exponentially with repeat length (Katti et al. 2001).

SSR analysis has been reported for expressed sequence tags (ESTs) from apple (Newcomb et al. 2006). Here, we further compare the composition of BES-based SSRs with that of EST-derived SSRs in apple. AT and AG repeats are the most abundant of di-nucleotide repeats in both BES-based and EST-derived SSRs. Both BESs and ESTs have few GC repeats. However, the frequencies of different types of repeats are different between BES-based SSRs and EST-derived SSRs. For example, AT and AG repeats account for ~57% and 18.8% of di-nucleotide repeats identified in BESs, respectively; while, AT and AG repeats constitute 7.6% and 88% of di-nucleotide repeats derived from ESTs, respectively (Newcomb et al. 2006). The frequency of di-nucleotide repeats is higher than that of

tri-nucleotide repeats for BES-based SSRs; whereas, the frequency of di- and tri-nucleotide repeats in EST-derived SSRs is comparable (Newcomb et al. 2006). These inconsistencies may be attributed to the fact that the composition and frequency of SSRs are different between genomic DNA and coding region sequences. Moreover, it is worth mentioning that the minimum length used to define SSRs is different between BES-based SSRs and EST-derived SSRs. The minimum size of BES-based SSRs is 15 bp; while, it is 12 bp for EST-derived SSRs. The differences in the minimum length of SSRs may also contribute to observed inconsistencies of SSR distribution between BESs and ESTs.

Comparative genetic mapping studies have revealed colinear chromosome segments among closely related species such as Poaceae (Devos and Gale 2000), Solanaceae (Tanksley et al. 1992), and Brassicaceae (O'Neill and Bancroft 2000). However, analysis of colinear chromosome segments is not well suited for distantly related species (Paterson et al. 1996). Recently, with the completion of whole genome sequences of model plants such as *Arabidopsis* and rice, an alternative analysis approach, microsynteny, has been developed to investigate colinearity among distantly related species. In this study, the extent of colinearity between apple and each of the three sequenced plant species, the eudicots *Populus* and *Arabidopsis* along with the monocot rice, has been determined by mapping apple BAC end pairs to the model plant genomes. A total of 154, 107, and 106 apple BES pairs have been identified to be homologous to *Populus*, *Arabidopsis*, and rice genomes, respectively. Among these BESs pairs, 8 (5.2%), 3 (2.8%), and 0 BES pairs have been mapped to *Populus*, *Arabidopsis*, and rice genomes, respectively, with a span of 10 to 300 kb. The apple and *Populus* represent two sister orders within the Eurosids I clade; whereas, *Arabidopsis* is a member of the order Brassicales within the Eurosids II clade. Thus, results presented in this study indicate that the apple has a higher degree of synteny with the closely related *Populus* than with the distantly related *Arabidopsis*. Therefore, in the future, comparative genetic mapping can be carried out between apple and poplar genomes using a microsynteny approach. Moreover, 28 BES pairs of apple map to the same chromosomes of *Arabidopsis*. Among those, 25 map to the same chromosome regions with a span of either <10 kb or more than 300 kb. This finding suggests that the degeneration of microsynteny between apple and *Arabidopsis* may be due to extensive rearrangements of the *Arabidopsis* genome (Blanc et al. 2000).

Acknowledgements This project was supported by the USDA Cooperative State Research, Education and Extension Service—National Research Initiative—Plant Genome Program grant No. 2005-35300-15538

References

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell 12:1093–1102

Chen M, Presting G, Barbazuk WG, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu J, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo M, Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA, Wing RA (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

Cheung F, Town CD (2007) A BAC end view of the *Musa acuminata* genome. BMC Plant Biol 7:29

Chevreau E, Lespinasse Y, Gallet M (1985) Inheritance of pollen enzymes and polyploid origin of apple (*Malus* × *domestica* Borkh). Theor Appl Genet 71:268–277

Choi S, Creelman RA, Mullet JE, Wing RA (1995) Construction and characterization of a bacterial artificial chromosome library from *Arabidopsis thaliana*. Weed World 2:17–20

Coburn J, Temnykh S, Paul E, McCouch SR (2002) Design and application of microsatellite marker panels for semi-automated genotyping of rice (*Oryza sativa* L.). Crop Sci 42:2092–2099

Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. Plant Cell 12:636–646

Eustice M, Yu Q, Lai CW, Hou S, Thimmapuram J, Liu L, Alam M, Moore PH, Presting GG, Ming R (2007) Development and application of microsatellite markers for genomic analysis of papaya. Tree Genet Genomes doi: 10.1007/s11295-007-0112-2

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: Where genetics meets genomics. Nat Rev Genet 3:329–341

Frelichowski JE, Palmer MB, Main D, Tomkins JP, Cantrell RG, Stelly DM, Yu J, Kohel RJ, Ulloa M (2006) Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends. Mol Gen Genomics 275:479–491

Gao LF, Jing RL, Huo NX, Li Y, Li XP, Zhou RH, Chang XP, Tang JF, Ma ZY, Jia JZ (2004) One hundred and one new microsatellite loci derived from ESTs (EST-SSRs) in bread wheat. Theor Appl Genet 108:1392–1400

Han Y, Korban SS (2007) *Spring*: a novel family of miniature inverted-repeat transposable elements is associated with genes in apple. Genomics 90:195–200

Han Y, Gasic K, Marron B, Beever JE, Korban SS (2007) A BAC-based physical map of the apple genome. Genomics 89:630–637

Ilic K, San Miguel PJ, Bennetzen JL (2003) A complex history of rearrangements in an orthologous region of the maize, sorghum, and rice genomes. Proc Natl Acad Sci USA 100:12265–12270

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

Klein PE, Klein RR, Cartinhour SW, Ulanch PE, Dong J, Obert JA, Morishige DT, Schlueter SD, Childs KL, Ale M, Mullet JE (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res 10:789–807

La Rota M, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice wheat and barley. BMC Genomics 6:23–35

Lai CWJ, Yu Q, Hou S, Skelton RL, Jones MR, Lewis KLT, Murray J, Eustice M, Guan P, Agbayani R, Moore PH, Ming R, Presting GG (2006) Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome. Mol Genet Genomics 276:1617–4615

Larkin DM, Everts-van der Wind A, Rebeiz M, Schweitzer PA, Bachman S, Green C, Wright CL, Campos EJ, Benson LD, Edwards J, Liu L, Osoegawa K, Womack JE, de Jong PJ, Lewin HA (2003) A cattle-human comparative map built with cattle BAC-ends and human genome sequence. Genome Res 13:1966–1972

Mahairas GG, Wallace JC, Smith K, Swartzell S, Holzman T, Keller A, Shaker R, Furlong J, Young J, Zhao S, Adams MD, Hood L (1999) Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. Proc Natl Acad Sci USA 96:9739–9744

McCouch SR, Teytelman L, Xu Y, Lobos KB, Clare K, Walton M, Fu B, Maghirang R, Li Z, Xing Y, Zhang Q, Kono I, Yano M, Fjellstrom R, DeClerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res 9:199–207

Meksem K, Zobrist K, Hyten D, Quangzhou T, Zhang H, Lightfoot DA (2000) Two large-insert soybean genomic libraries constructed in a binary vector: applications in chromosome walking and genome wide physical mapping. Theor Appl Genet 101:747–755

Ming R, Moore PH, Zee F, Abbey CA, Ma H, Paterson AH (2001) Construction and characterization of a papaya BAC library as a foundation for molecular dissection of a tree-fruit genome. Theor Appl Genet 102:892–899

Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloska S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S et al (1999) A complete BAC-based physical map of the *Arabidopsis thaliana* genome. Nat Genet 22:271–275

Newcomb RE, Crowhurst RN, Gleave AP, Rikkerink EHA, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ, Laing WA, McArtney S, Nain B, Ross GS, Snowden KC, Souleyre EJF, Walton EF, Yauk YK (2006) Analysis of expressed sequence tags from apple. Plant Physiol 141:147–166

O'Neill CM, Bancroft I (2000) Comparative physical mapping of segments of the genome of *Brassica oleracea* var. *alboglabra* that are homeologous to sequenced regions of chromosomes 4 and 5 of *Arabidopsis thaliana*. Plant J 23:233–243

Paterson AH, Lan T, Reischmann KP, Chang C, Lin Y, Liu S, Burow MD, Kowalski SP, Katsar CS, DelMonte TA, Feldmann KA, Schertz KF, Wendel JF (1996) Toward a unified genetic map of higher plants, transcending the monocot–dicot divergence. Nat Genet 14:380–382

Salimath SS, Bhattacharyya MK (1999) Generation of a soybean BAC library, and identification of DNA sequences tightly linked to the *Rps1-k* disease resistance gene. Theor Appl Genet 98:712–720

Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, Bergstrom D, Houchins K, Melia-Hancock S, Musket T, Duru N, Polacco M, Edwards K, Ruff T, Register JC, Brouwer C, Thompson R, Velasco R, Chin E, Lee M, Woodman-Clikeman W, Long MJ, Liscum E, Cone K, Davis G, Coe EH Jr (2002) Development and mapping of SSR markers for maize. Plant Mol Biol 48:463–481

Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, Kochert GA, Boerma HR (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics 144:329–338

Shultz JL, Kurunam D, Shopinski K, Iqbal MJ, Kazi S, Zobrist K, Bashir R, Yaegashi S, Lavu N, Afzal AJ, Yesudas CR, Kassem MA, Wu C, Zhang HB, Town CD, Meksem K, Lightfoot DA (2006) The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of *Glycine max*. Nucleic Acids Res 34:D758–D765

Shultz JL, Kazi S, Bashir R, Afzal JA, Lightfoot DA (2007a) The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor Appl Genet 114:1081–1090

Shultz JL, Ali S, Ballard L, Lightfoot DA (2007b) Development of a physical map of the soybean pathogen *Fusarium virguliforme* based on synteny with *Fusarium graminearum* genomic DNA. BMC Genomics 8:262–268

Tamanna A, Khan AU (2005) Mapping and analysis of simple sequence repeats in the *Arabidopsis thaliana* genome. Bioinformation 1:64–68

Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB (1992) High density molecular linkage maps of the tomato and potato genomes. Genetics 132:1141–1160

Tatum T, Stepanovic S, Biradar DP, Rayburn AL, Korban SS (2005) Variation in nuclear DNA content in *Malus* species and cultivated apples. Genome 48:924–930

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

Tomkins JP, Mahalingam R, Smith H, Goicoechea JL, Knap HT, Wing RA (1999) A bacterial artificial chromosome library for soybean PI 437654 and identification of clones associated with cyst nematode resistance. Plant Mol Biol 41:25–32

Tóth G, Gáspári Z, Jurka J (2000) Microastellites in different eukaryotic genomes: survey and analysis. Genome Res 10: 967–981

Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381:364–366

Vinatzer BA, Zhang H-B, Sansavini S (1998) Construction and characterization of a bacterial artificial chromosome library of apple. Theor Appl Genet 97:1183–1190

Wang GL, Holsten TE, Song WY, Wang HP, Ronald PC (1995) Construction of a rice bacterial artificial chromosome library and identification of clones linked to the *Xa-21* disease resistance locus. Plant J 7:525–533

Woo SS, Jiang J, Gill BS, Paterson AH, Wing RA (1994) Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res 22:4922–4931

Xia X, Xie Z, Li W (2003) Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J Mol Evol 56:362–370

Xu M, Korban SS (2002) A cluster of four receptor-like genes resides in the *Vf* locus that confers resistance to apple scab disease. Genetics 162:1995–2006

Xu M, Song J, Cheng Z, Jiang J, Korban SS (2001) A bacterial artificial chromosome (BAC) library of *Malus floribunda* 821 and contig construction for positional cloning of the apple scab resistance gene *Vf*. Genome 44:1104–1113

Yim YS, Davis GL, Duru NA, Musket TA, Linton EW, Messing JW, McMullen MD, Soderlund CA, Polacco ML, Gardiner JM, Coe EH Jr (2002) Characterization of three maize bacterial artificial chromosome libraries toward anchoring of the physical map to the genetic map using high-density bacterial artificial chromosome filter hybridization. Plant Physiol 130:1686–1696