

## Research Article

# Construction, Characterization, and Preliminary BAC-End Sequence Analysis of a Bacterial Artificial Chromosome Library of the Tea Plant (*Camellia sinensis*)

Jinke Lin,<sup>1,2</sup> Dave Kudrna,<sup>1</sup> and Rod A. Wing<sup>1,3</sup>

<sup>1</sup> School of Plant Sciences, Arizona Genomics Institute, The University of Arizona, Tucson AZ 85721, USA

<sup>2</sup> Department of Tea Sciences, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China

<sup>3</sup> BIO5 Institute, University of Arizona, Tucson AZ 85721, USA

Correspondence should be addressed to Rod A. Wing, rwing@ag.arizona.edu

Received 29 July 2010; Accepted 28 October 2010

Academic Editor: Yong Lim

Copyright © 2011 Jinke Lin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We describe the construction and characterization of a publicly available BAC library for the tea plant, *Camellia sinensis*. Using modified methods, the library was constructed with the aim of developing public molecular resources to advance tea plant genomics research. The library consists of a total of 401,280 clones with an average insert size of 135 kb, providing an approximate coverage of 13.5 haploid genome equivalents. No empty vector clones were observed in a random sampling of 576 BAC clones. Further analysis of 182 BAC-end sequences from randomly selected clones revealed a GC content of 40.35% and low chloroplast and mitochondrial contamination. Repetitive sequence analyses indicated that LTR retrotransposons were the most predominant sequence class (86.93%–87.24%), followed by DNA retrotransposons (11.16%–11.69%). Additionally, we found 25 simple sequence repeats (SSRs) that could potentially be used as genetic markers.

## 1. Introduction

Since 1994, bacterial artificial chromosome (BAC) libraries have become an invaluable resource tool to initiate genomics research in the areas of genome sequencing, physical mapping, positional cloning, complex analysis of targeted genomic regions, and analysis of gene structure and function [1–8]. More recently, physical map tiles of BACs have been shown to be suitable for next generation sequencing of chromosomal regions or whole genomes [9].

BAC-end sequencing (BES) is a powerful tool that enhances the value of BAC libraries as a genomic resource by providing partial sequence information that can be used to understand genome content and architecture and develop genetic markers [10, 11]. Physical maps constructed from fingerprinted BAC clones, together with associated BAC-end sequence information, can be used to: construct BAC fingerprint-/BES- based physical maps [1, 12] which can be aligned to reference genome sequences; sequence genomes by “walking” from one clone to the next [13]; anchor whole

genome shotgun sequence data; integrate genetic linkage maps with physical maps [14].

Tea (*Camellia sinensis*) is an important world beverage [15]. About 3.9 million tons of tea is produced (<http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor>) and consumed yearly with an estimated world value of \$9.04 billion (\$2.39 per kg; <http://www.fao.org/docrep/meeting/019/K8336E.pdf>). Throughout history, it has become well known that tea is one of the most beneficial beverages to drink due to its health attributes. Tea polyphenols (catechins) and antioxidants, saponin, polysaccharides, L-theanine, pigments, and tea water extracts are the most important components that contribute to tea's medicinal qualities [16–19].

There are different varieties or types of tea, that is, green tea, black tea, oolong tea, white tea, and so on [20], (see also <http://jingtea.com/tea-knowledge/tea-varieties>). Oolong tea is a special variety of tea and has been studied for its effects on diabetes, eczema, allergies, bacterial infections, dental caries (cavities), obesity, cardiovascular disease, cancer, and

its antioxidant properties [21–26]. It has been shown that consumption of oolong tea stimulates both energy expenditure and fat oxidation in normal weight men [21]. Oolong tea is semifermented which occurs when the leaves are gently rolled during processing [20], and this gives oolong tea a unique appearance and flavor. Chin-shin oolong is one of the most widely cultivated oolong tea varieties worldwide [27].

It has been suggested that functional genomic research should be a major emphasis of tea genetics and breeding in the future [28]. Although rapid progress of gene identification and isolation from tea plants has been made in the past several years [29], the study of the tea plant genome lags far behind other crop species due to the lack of good genomic research tools. This is most likely due to the difficulties of preparing high-quality tea DNA and due to the distinctness of tea plant from other taxa (i.e., perennial nature, high inbreeding depression, unavailability of distinct mutants of different biotic and abiotic stress, and large genome size of 4Gb [30]). One key genomic tool that is completely lacking for tea is the availability of a high-quality, deep-coverage, BAC library. In this paper, we report the construction of a high-quality, publicly available BAC library of tea plant from the variety Chin-shin oolong. We generated and analyzed a limited data set of BAC-end sequences from this library which provided an early glimpse into the sequence composition of the tea genome.

## 2. Materials and Methods

**2.1. Germplasm and Plant Tissue Processing.** Sixty 6-year-old plants, derived from a clonally propagated single mother plant of a *Camellia sinensis* cultivar Chin-shin oolong, were selected as the plant germplasm source. Healthy young shoot tips and the uppermost two leaves were collected (similar to harvesting top-quality tea), then washed quickly to remove debris, and immediately frozen by submersion in liquid nitrogen followed by short-term storage at  $-80^{\circ}\text{C}$ . All plant growth and tissue selection was kindly provided by Dr. Francis Zee (USDA, Hilo, Hawaii).

**2.2. Preparation of High Molecular Weight (HMW) Tea Plant DNA in Agarose Plugs.** Tea genetic resources have lagged behind other important plants due to the complex and difficult nature of adequate extraction of quality nuclear DNA. However, a detailed manuscript describing the experiments that led to the successful isolation of high-quality, high molecular weight, nuclear DNA, suitable for BAC library construction, was recently made available [31]. The method for tea plant BAC library construction utilized the standard Arizona Genomics Institute (AGI) protocol [32], the method of Luo and Wing [3], and Ammiraju et al. [2]. Modifications necessary to achieve successful construction of a tea plant BAC library are described as follows.

Twenty grams of frozen tissue was homogenized and transferred to a flask containing 200 mL of prechilled extraction buffer (10 mM Tris-HCL, pH 8.0, 10 mM EDTA, pH 8.0, 100 mM KCl, 0.5 M sucrose, 4 mM spermidine, 1 mM spermine, 0.10% w/v L-ascorbic acid, 2.00% w/v PVP-40,

0.13% w/v sodium diethyldithiocarbamate trihydrate; PVP-40 was only about 50 percent dissolved) and  $400\ \mu\text{l}$   $\beta$ -mercaptoethanol. The homogenate was filtered into a flask that contained 200 mL of the same above prechilled extraction buffer and  $400\ \mu\text{l}$   $\beta$ -mercaptoethanol. The homogenate was filtered a second time into a fresh flask that contained 45 mL of prechilled extraction buffer with 10.00% v/v Triton X-100. The mixture was centrifuged for 15 min at 3250 rpm at  $4^{\circ}\text{C}$ . The resulting pellet was washed several times with the same buffer and resuspended in 1 mL of prechilled extraction buffer, incubated in a  $45^{\circ}\text{C}$  water bath for 5 minutes, and gently mixed with one-third volume of 1.00% low melting temperature agarose (in extraction buffer) at  $45^{\circ}\text{C}$ . The mixture was allowed to solidify after transferring to plug molds. Twenty-four plugs were transferred into a 50 mL-Falcon tube, containing 40 mL of standard proteinase K solution (1.00% w/v N-lauroylsarcosine (sodium salt), 0.1 mg/mL proteinase K, dissolved in 0.5 M EDTA, pH 9.4), and incubated in a hybridization oven at  $50^{\circ}\text{C}$  with a gentle rotation for 24 h. After the plugs were rewashed with fresh proteinase K solution for an additional 24 h, they were rewashed in 2 additional solutions. First, two times with 40 mL  $\text{T}_{10}\text{E}_{10}$  containing 1 mM PMSF (phenylmethanesulfonyl fluoride) and then twice with 40 mL TE, each time for about 1 h at room temperature with gentle shaking. The plugs were stored in 70.00% ethanol at  $-20^{\circ}\text{C}$  (for long-term storage).

**2.3. Restriction Digestion of HMW DNA and Isolation of Size-Selected Fragments.** Two and a half DNA plugs were used to establish optimal *Hind*III partial digestion conditions. Formal partial digestion, using 5 DNA plugs, was performed using 5U of *Hind*III restriction enzyme added to each sample (per half plug). Digested samples were loaded to 1.00% agarose gel and subjected to pulsed-field gel electrophoresis (PFGE). DNA was visualized, and agarose fragments, containing specific DNA sizes, were cut from the gel slabs. A second- and third-PFGE run of the fragments was performed to further purify the DNA and remove small DNA fragments. After finishing the third size selection, the gel fractions containing different sized fragments (B2, B1, A2) were recovered and stored at  $-20^{\circ}\text{C}$  in 70.00% ethanol (for long-term storage).

**2.4. Ligation of Sized DNA Fragments.** For each size-selected fraction, the high molecular weight genomic DNA was electroeluted from the agarose at  $4^{\circ}\text{C}$ . Pipet tips (with cutoff tips) were used when manipulating high molecular weight genomic DNA to avoid mechanical shearing. The DNA concentrations were estimated, and 120 ng–200 ng DNA was used to ligate to the linearized (*Hind*III site) and dephosphorylated vector pIndigoBAC536 *Swa*I, commonly known as pAGIBAC1 [2]. Separately, ligations were mixed well by tapping and then incubated in a water bath at  $16^{\circ}\text{C}$  for 19 hours. Ligation samples were transferred into 0.1 M glucose/1.00% agarose cones to desalt for 1.5 h on ice. Ligations were transferred into fresh microcentrifuge tubes and stored at  $4^{\circ}\text{C}$  until transformation tests were completed.

TABLE 1: Composition of the *Camellia sinensis* tea BAC library, CSBCBa.

	Size selection range (kb)	Number of clones	Average insert size (kb)	Genome equivalent coverage	Insert size range (kb)	Proportion of the whole library size (%)
Total Library	190–350	401,280	135	13.5	8–316	100%
Sublibrary 1 (Ligation B2)	300–350	40,320	140	1.4	8–316	10.05
Sublibrary 2 (Ligation B1)	240–300	160,896	140	5.6	32–185	40.10
Sublibrary 3 (Ligation A2)	190–240	200,064	130	6.5	19–180	49.85

2.5. *Transformation Testing, Library Name, and BAC Clone End Sequencing (BES)*. Ligation samples were tested to determine transformation efficiency and cloned insert quality. The methods for test transformations, BAC DNA isolations, BAC insert size analyses, bulk transformations, colony arrays, and library characterizations followed exactly the AGI protocols (<http://www.genome.arizona.edu/information/publications/wingpub/construct.pdf> and [http://www2.genome.arizona.edu/pdfs/publications/meizhong\\_metmb.pdf](http://www2.genome.arizona.edu/pdfs/publications/meizhong_metmb.pdf)). From each HMW fraction (B2, B1, A2), 2.3  $\mu$ l of each ligation was used to transform 20  $\mu$ l of DH10B T1 phage-resistant *E. coli* cells (Invitrogen) by electroporation using the Cell Porator and Voltage Booster electroporation system (Life Technologies). Electroporation was performed on ice at 327 DC V with fast charge rate at a low resistance (4 k $\Omega$ ) and a capacitance of 330  $\mu$ F. The cells were transferred into 3 mL of SOC media, and incubated at 37°C for 1 h with shaking at 250 rpm, followed by the addition of an equal volume of sterile glycerol and gentle shaking for 3 minutes. These mixtures were immediately frozen by submersion into liquid nitrogen followed by long-term storage at –80°C.

To evaluate these transformation tests, 300  $\mu$ l of each (containing cells, SOC, and glycerol) were spread on Petri dishes (containing LB–X-gal–IPTG agar with 12.5  $\mu$ g/mL of chloramphenicol, 80  $\mu$ g/mL X-gal, and 100  $\mu$ g/mL IPTG) and incubated at 37°C overnight. Five-hundred-seventy-six white recombinants (positive for insert), 192 from each of the three sublibraries (B2, B1, A2), were randomly selected and grown overnight at 37°C in 1.2 mL LB broth (with chloramphenicol 12.5  $\mu$ g/mL) with shaking at 220 rpm. BAC DNAs were isolated and digested with *NotI* to release the BAC insert. Digestions were separated by PFGE at 6V/cm, switch time from 5 to 15 s, angle 120°, and run for 16 h followed by staining, destaining, and visualization.

Transformed *E. coli* from the B2 ligation, selected to contain the largest insert sizes, were prepared for array to 384-well microtiter dishes. Five mL of the mixture from the B2 ligation were spread on LB–X-gal–IPTG–CM agar Q-trays (22.5  $\times$  22.5 cm) and incubated at 37°C overnight. The white (insert positive recombinant) clones were robotically picked (using a Genetix Q-Bot; Genetix Ltd.) into forty-eight 384-well microtiter plates containing freezing media and stored at –80°C. Hybridization screening filters were printed from library copies to facilitate future experiments.

To further evaluate this sub library (B2 ligation), 576 BACs were randomly selected from these forty-eight plates and grown overnight at 37°C with shaking at 220 rpm in 1.2 mL LB supplemented with chloramphenicol (12.5  $\mu$ g/mL). BAC DNAs were isolated and digested with *NotI*. The digested clones were separated by PFGE and analyzed. Additionally, 192 BAC clones were subjected to BAC clone end sequencing (BES) using the method of Kim et al. [12], and the resultant sequences were analyzed for chloroplast and mitochondrial genome contaminations and for repeat content using BLASTN searches with settings that required 98% identity over lengths of at least 51 bp.

The successful *C. sinensis* BAC library, composed of *E. coli* transformants from each of the three ligations, was named CSBCBa.

### 3. Results

3.1. *Construction and Characterization of the Tea BAC Library*. *Camellia sinensis* cultivar Chin-shin oolong was chosen to construct the BAC library since it is one of the most widely cultivated oolong tea varieties. To avoid contamination with small, trapped DNA fragments and improve the size and uniformity of the inserts, the high molecular weight (HMW) genomic DNA was partially digested with *HindIII* and three separate size fractionations were collected. Following ligations into the *HindIII* site of the pAGIBAC1 vector, the three size fractionations were transformed, and the new tea BAC library, consisting of 401,280 BAC clones, was named CSBCBa; see Table 1. The overall average insert size of the total library was calculated to be 135 kb with inserts ranging from 8 to 314 kb, and the total library coverage was estimated to equal 13.54 haploid genome equivalents based upon mathematical calculations that utilize a genome size of 4,000 Mb. Sub library 1 (Ligation B2) was composed of 40,320 clones with an average insert size 140 kb; Sub library 2 (Ligation B1) was composed of 160,896 clones with an average insert size 140 kb; Sub library 3 (Ligation A2) was composed of 200,064 clones with an average insert size 130 kb. Five-hundred-seventy-six BACs, 192 from each of the three sublibraries, were randomly selected and analyzed by *NotI* restriction enzyme digestion in order to evaluate the library (Figure 1). Based on this sample size, no empty vector clones were visualized. Over 74.40% of the BAC

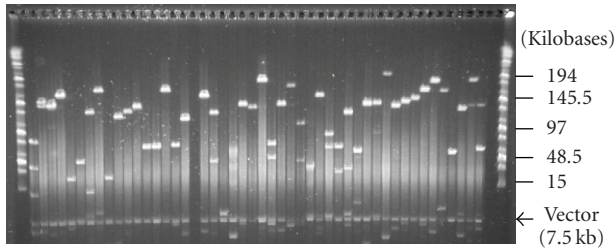


FIGURE 1: Insert DNA analysis of random BAC clones from the *Camellia sinensis* *Hind*III BAC library, CSBCBa, by pulsed-field gel electrophoresis. DNA samples were digested with *Not*I and separated by pulsed-field gel electrophoresis with ramp pulse time of 5–15 s at 6 V/cm at 14°C in 0.5× TBE buffer for 16 h. Markers used are midrange I (outside lanes). The 7.5-kb common band corresponds to the cloning vector pIndigoBAC536 *Swal* (PAGIBAC1).

clones were shown to carry DNA inserts greater than 130 kb, with a reasonable fraction (22.30%) carrying inserts larger than 170 kb (Figure 2). BAC clones from sub library 1, Ligation B2, were spread on agar dishes, clones (18,432) were robotically picked and arrayed to 48- 384well plates, and hybridization screening filters were printed to facilitate future experiments.

**3.2. Analysis of Repetitive Sequences from Pilot BAC-End Sequence of *C. sinensis*.** To examine the quality of the library and to obtain a preliminary view of the major repetitive element content of the *C. sinensis* genome, unidirectional end sequencing was performed on 192 random BAC clones which produced 182 reads with an average length of 581 high-quality bases; the BESs are available at the URL: [http://www2.genome.arizona.edu/files/camellia\\_BES.txt](http://www2.genome.arizona.edu/files/camellia_BES.txt). Of the 182 random BESs, the total number of nucleotides was 111,695 bp and the GC level was 40.35% (Table 2).

Likewise, the 182 random BAC-end sequences were evaluated using BLAST searches and queries to the transposable element databases of *Arabidopsis* (*Arabidopsis thaliana*) at Genetic Information Research Institute and of rice (*Oryza sativa*) at the Arizona Genomics Institute. This preliminary analysis of repetitive sequences from pilot BAC-end sequences indicated that 97 to 98 percent of the BESs contained sequences related to transposable elements. LTR retrotransposons were the predominant class of repeat elements in *C. sinensis* (*Arabidopsis* 86.93% and rice 87.24%; Table 2). The second major class of transposable elements belongs to DNA retrotransposon (*Arabidopsis* 11.16% and rice 11.69%). LINE and non-LTR elements were also observed though at much lower amounts.

BLASTN searches against the chloroplast and mitochondrial genomes of *Arabidopsis* (*Arabidopsis thaliana*) and rice (*Oryza sativa*) using the 182 random BESs revealed no significant homology found, thus indicating that the BAC library contained very low levels of chloroplast or mitochondrial DNA contamination. Visualization of the inserts from the

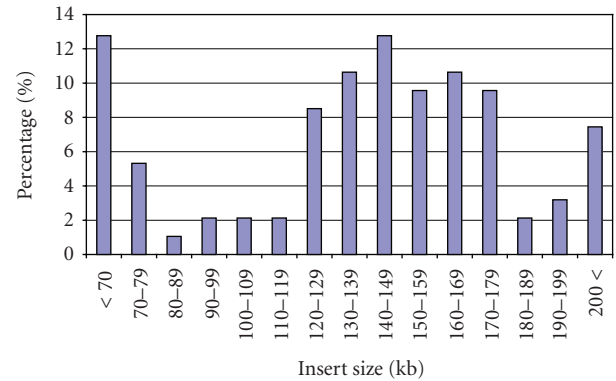


FIGURE 2: Insert size distribution of *Camellia sinensis* BAC clones from the CSBCBa library. The average insert size of the BAC library was calculated to be 135 kb from an analyzed random sample size of 576 BAC clones.

random 576 BACs from pulsed-field electrophoresis also indicated no patterns typical of organelle contamination (data not shown). Further repeat analysis of the 182 random BESs revealed 25 different microsatellite repeats (Table 3). The 25 simple sequence repeats (SSRs) were found from 17 different BAC clones (from the forward and reverse reads), thus suggesting that eight BACs represented genomic regions that contained high repeat content. Interestingly, one BES read, > \_camellia\_g1-12-3-07\_D02\_g1, contained two different dimer, and one pentamer, SSRs.

## 4. Discussion

Generally, the world's agriculture and food systems come from the tremendous biological diversity encompassed in over 200 independently (and perhaps convergent) domesticated species of Angiosperms [33]. Paradoxically, much of the potential genomic research with these taxa has been centered on a few model species, or taxonomic families, that represent only a very small amount of this diversity. In the past decade, prominent plant genome initiatives have resulted in genomic resources of enormous magnitude. These include molecular/genetic maps, transcript databases, large-insert libraries, five complete genome sequences (rice [14], *Arabidopsis* [34], poplar [35], grape [36], and maize [37]), and a suite of several others scheduled for completion. The ultimate goals of these projects are to continue to discover new ways to meet future world agricultural and food needs while simultaneously providing an understanding of functional systems biology. However, the expected dramatic advances in theoretical and applied plant biology for other taxa, following these innovations, have been critically slow due to two major obstacles: lack of adequate genomic tools and resources to efficiently and effectively transfer existing genomic knowledge to other economically and diverse agriculturally important species and lack of representation of novel genes and regulatory networks that underlie key traits of agriculture and ecology in the sequenced model species. Despite the significant economic impact of tea and other similar commodities (i.e., coffee, cocoa), a lack of

TABLE 2: Analysis of repetitive elements from preliminary BAC-end sequence (BES) of the *Camellia sinensis* tea BAC library, CSBCBa.

No. of BES	No. of BES nucleotides (bp)	BES GC contents (%)	Compared to	Repeat coverage (bp/%)	LTR repeat coverage (bp/%)	DNA repeat coverage (bp/%)	LINE repeat coverage (bp/%)	Non-LTR retrotransposon repeat coverage (bp/%)
182	111,695	40.35	Rice	5,251/4.70	4,581/87.24	614/11.69	56/1.06	0/0
			Arabidopsis	5,457/4.88	4,744/86.93	609/11.16	0/0	104/1.90

TABLE 3: Identification of simple sequence repeats (SSRs) from preliminary BAC-end sequence (BES) of the *Camellia sinensis* tea BAC library, CSBCBa.

Sequence filename	Type	Motif	No. of repeats	SSR start	SSR end
>_camellia_g1-12-3-07_G11_g1	di	ac	6	370	381
>_camellia_g1-12-3-07_A01_g1	di	ag	7	173	186
>_camellia_g1-12-3-07_C10_g1	di	ag	8	133	148
>_camellia_g1-12-3-07_D02_g1	di	ag	6	461	472
>_camellia_g1-12-3-07_H02_g1	di	ag	5	585	594
>_camellia_b1-12-3-07_G05_b1	di	at	5	306	315
>_camellia_g1-12-3-07_C10_g1	di	at	6	121	132
>_camellia_g1-12-3-07_D02_g1	di	at	11	194	215
>_camellia_g1-12-3-07_D08_g1	di	at	5	388	397
>_camellia_b1-12-3-07_A03_b1	di	ct	13	358	383
>_camellia_b1-12-3-07_D05_b1	di	ct	8	1	16
>_camellia_g1-12-3-07_A06_g1	di	ct	7	208	221
>_camellia_g1-12-3-07_B11_g1	di	ct	7	310	323
>_camellia_g1-12-3-07_G11_g1	di	ct	5	287	296
>_camellia_g1-12-3-07_B06_g1	di	ga	18	516	551
>_camellia_g1-12-3-07_C10_g1	di	ga	18	586	621
>_camellia_g1-12-3-07_D02_g1	di	ga	8	538	553
>_camellia_b1-12-3-07_C09_b1	di	ta	6	124	135
>_camellia_b1-12-3-07_F10_b1	di	ta	8	29	44
>_camellia_b1-12-3-07_G12_b1	di	ta	11	437	458
>_camellia_g1-12-3-07_D03_g1	di	ta	5	555	564
>_camellia_b1-12-3-07_A03_b1	di	tc	8	314	329
>_camellia_g1-12-3-07_B06_g1	di	tg	6	505	516
>_camellia_b1-12-3-07_H09_b1	tri	aat	5	197	211
>_camellia_g1-12-3-07_D02_g1	penta	gagat	9	556	600

adequate public genomic resources, especially access to large-insert libraries, has contributed to a lack of advanced genetic knowledge useable for modern breeding and improvement.

Recent published tea plant works involving molecular tools have shown an AFLP and RAPD marker-based linkage map [38], identified cDNAs involved in secondary metabolism [39], described sequence analysis from 4320 tea ESTs [40], and used intersimple sequence repeats to analyze genetic variability of somaclonal embryo-derived tea plants [41]. A recent review described the use of molecular resources for tea cultivar classification, the identification of parentage of Mulberry scale resistance, and more advanced linkage maps with SSR markers [42]. While these and other reports provide important insight to the understanding of

the tea plant, the availability and utilization of BAC resources are absent.

Here, we report the construction and public availability of a high-quality, deep-coverage, large-insert, bacterial artificial chromosome (BAC) library of cultivated tea plant (*Camellia sinensis*) variety Chin-shin oolong. This BAC library resource is publicly available in the form of whole libraries, filters, and individual clones, through the AGI BAC/EST Resource Center (<http://www.genome.arizona.edu/orders>), and we expect it to be extensively used worldwide for the analysis of genome evolution and organization, positional cloning, and eventual gap closure of a *C. sinensis* reference sequence.

This tea BAC library was made possible after significant methodological improvements were accomplished in the

preparation of purified high molecular weight (HMW) DNA [31] and subsequent partial enzymatic digestions and ligations. We showed that, despite expected difficulties in obtaining HMW DNA of reliable quality from tissues known to contain compounds deleterious to established extraction techniques, our results yielded ligations that produced a BAC library with low organellar contamination, such as chloroplast and mitochondria, yet with high transformation efficiencies and large-insert sizes. Though the genome size of tea is quite large, 4 Gb, the described BAC library has an average insert size of 135 kb and provides over 13x genome coverage to allow for adequate utility. It was divided into three sublibraries: sub library 1 (Ligation B2) composed of 40,320 clones with average insert sizes of 140 kb, sub library 2 (Ligation B1) composed of 160,896 clones with average insert sizes of 140 kb, and sub library 3 (Ligation A2) composed of 200,064 clones with average insert sizes of 130 kb.

Fresh tea leaves are rich in both volatile secondary compounds such as tea polyphenols and carbohydrate matrices such as tea polysaccharides. It has been shown that tea leaves contain 20–33% dry weight of polyphenols [43], and polysaccharides have been shown to be present at 7.02% dry weight in oolong tea [44]. These types of compounds contribute negatively [32] toward the handling of high molecular weight (HMW) DNA necessary for construction of large-insert genomic clone libraries. Tea polyphenols must be prevented from interacting with the nuclear DNA, and tea polysaccharides must be prevented from trapping nuclei in the process of tissue homogenization. In this study, to prepare high-quality HMW DNA, the frozen tissue was homogenized with a buffer (composed of combining ingredients from two methods but requiring the PVP-40 to be partially undissolved) in double volumes followed by additional filtration. Double volumes of buffer were found to dilute the polyphenols and polysaccharides, as evidenced by the absence of sticky nuclear pellets, thus allowing for increased removal during the centrifugation steps. The combining of the nuclei with the low melting temperature agarose at a lower ratio of 1 : 3 (instead of 1 : 1) was required to concentrate the nuclei but was performed in the presence of the undissolved PVP-40. To lower the organellar DNA contamination and to further eliminate tea polysaccharides and tea polyphenols in the process of preparing tea plant HMW DNA plugs, the use of additional washing steps were performed [45].

Removing small restriction fragments is vital for construction of a high-quality BAC library [46, 47]. Tea plant genomic plugs prepared by the method used in this paper contained abundant HMW DNA and produced satisfactory restriction fragments when digested with *Hind*III. To avoid contamination with small-trapped DNA fragments and improve the size and uniformity of the inserts, we performed three separate size selections. Usually two size selections are sufficient, but visual examination of the DNA fragments after the second size selection revealed that an additional selection was required (data not shown). A detailed manuscript describing the experiments for the appropriate isolation of high-quality, high molecular weight DNA that led to the successful construction of this tea plant BAC library (tea

plant nuclei isolation, buffer compositions, HMW DNA, large-insert ligations, etc.) was recently published [31], and this method will also yield sufficient quality tea DNA for other purposes, such as next generation sequencing (NGS).

Pulsed-field gel electrophoresis analysis of 576 random BAC clone plasmids showed that the majority of *C. sinensis* cloned DNA inserts were present as single *Not*I fragments. This indicated that the *C. sinensis* genome apparently contained few *Not*I sites, a feature commonly observed with the genomes of other plant dicot species and contrary to the results obtained with monocot species [48].

Long terminal repeats (LTRs), a component of genome repeat analysis [49–51], are sequences of DNA that are repeated hundreds or thousands of times in the genome. They are often found in retrotransposons and flanking functional genes [51, 52]. LTR retrotransposons constitute a significant portion of most eukaryote genomes and in plants have been suggested to be causative for the dramatic differences of genome sizes (and polyploidization) and for disruptions of genome organization and structure [2, 51, 53, 54]. While insertions of DNA have been attributed to amplifications of retrotransposons [51, 54–57], deletions have been suggested to involve all DNA sequence class types and may be the result of homologous recombination and/or illegitimate recombination [56]. Recent analysis of BES from different *Oryza* species [2] found good correlation with flow cytometric genome sizing and repeat content. Our preliminary analysis of tea repetitive sequences from pilot BAC-end sequences indicated that over 98 percent of the tea genome could be repetitive. We found that LTR retrotransposons are the predominant class of repeat elements in *C. sinensis* followed by DNA retroelements. These results support the correlation of large genome size with proliferation of repeat content since the tea genome is quite large, 4,000 Mb, which is 1.6X maize and 0.8X barley whose genomes are approximately 80–95 percent repetitive. Therefore, it is not surprising that the majority of tea BESs contained sequences highly enriched with transposable elements.

## 5. Conclusion

A high-quality, deep-coverage, *Hind*III BAC library of tea plant (*Camellia sinensis*) has been constructed. The library, named CSBCBa, is publicly available from the Arizona Genomics Institute Resource Center (<http://www.genome.arizona.edu/orders/>). The average insert size of the library was 135 kb, it contained very low organellar contamination, and it provides 13.54x genome equivalent coverage from a total of 401,280 clones. Analysis of BAC clone end sequences revealed that the repetitive fraction of the tea genome was highly enriched for LTRs and DNA TEs. This public resource will provide a useful platform for genomics research, such as genome sequencing, DNA fingerprinting and physical mapping, gene identification, isolation and regulation, as well as complex analysis of targeted genomic regions. Research of genomic information and gene expression patterns of tea plant have advanced slowly probably owing to the lack of an available high-quality BAC library as one key genomic tool.

Since this BAC library has been made available to the public, we expect that the most advanced work with DNA tools, which has been initiated by other researchers, will move forward for deepening genomics research of tea plant. The specific genes of tea plant, such as the genes involved in the oxidative pathways of catechins by polyphenol oxidase [20], which are involved in the pathways of the powerful antioxidant epigallocatechin-3-gallate (EGCG), may be revealed by use of this resource. Previous work on tea plant genetics, such as linkage mapping, genetic integrity of somaclonal variants [38, 41, 58, 59], and tea plant functional genes and biopathways [39, 40, 60, 61], may be integrated and/or expanded with this library to help unravel the tea plant genome.

## Acknowledgments

The authors thank Dr. Francis T. Zee (Research Leader, Supervisory Research Horticulturist at USDA/ARS, Hilo, Hawaii) for helpful discussions and collecting and providing tea plant tissues for this research. Additional thanks, to the staff of AGI for robotic and barcode handling of the clones, DNA template extractions, BES and data pipeline.

## References

- [1] R. A. Wing, J. S. S. Ammiraju, M. Luo et al., "The *Oryza* map alignment project: the golden path to unlocking the genetic potential of wild rice species," *Plant Molecular Biology*, vol. 59, no. 1, pp. 53–62, 2005.
- [2] J. S. S. Ammiraju, M. Luo, J. L. Goicoechea et al., "The *Oryza* bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus *Oryza*," *Genome Research*, vol. 16, no. 1, pp. 140–147, 2006.
- [3] M. Luo and R. A. Wing, "An improved method for plant BAC library construction," in *Plant Functional Genomics: Methods and Protocols*, E. Grotewold, Ed., pp. 3–19, Humana Press, Totowa, NJ, USA, 2003.
- [4] S. Choi and R. A. Wing, "The construction of bacterial artificial chromosome (BAC) libraries," in *Plant Molecular Biology Manual*, S. Gelvin and R. Schilperoort, Eds., pp. 1–28, Kluwer Academic Publishers, Norwell, Mass, USA, 2nd edition, 2000.
- [5] N. Bayou, R. M'rad, A. Belhaj et al., "De novo balanced translocation t (7;16) (p22.1; p11.2) associated with autistic disorder," *Journal of Biomedicine and Biotechnology*, vol. 2008, no. 1, Article ID 231904, 5 pages, 2008.
- [6] B. Crawford, A. A. Hussain, and N. M. Jideama, "Evidence of a genomic biomarker in normal human epithelial mammary cell line, MCF-10A, that is absent in the human breast cancer cell line, MCF-7," *Journal of Biomedicine and Biotechnology*, vol. 2006, Article ID 43181, 5 pages, 2006.
- [7] H. Shizuya, B. Birren, U.-J. Kim et al., "Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 18, pp. 8794–8797, 1992.
- [8] J. M. Breen, T. Wicker, X. Kong et al., "A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block," *BMC Plant Biology*, vol. 10, no. 1, article 98, 2010.
- [9] S. Rounsley, P. Marri, Y. Yu et al., "De novo next generation sequencing of plant genomes," *Rice*, vol. 2, pp. 35–43, 2009.
- [10] P. David, M. Sévignac, V. Thareau et al., "BAC end sequences corresponding to the B4 resistance gene cluster in common bean: a resource for markers and synteny analyses," *Molecular Genetics and Genomics*, vol. 280, no. 6, pp. 521–533, 2008.
- [11] Y. Han and S. S. Korban, "An overview of the apple genome through BAC end sequence analysis," *Plant Molecular Biology*, vol. 67, no. 6, pp. 581–588, 2008.
- [12] H. R. Kim, B. Hurwitz, Y. Yu et al., "Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus *Oryza*," *Genome Biology*, vol. 9, no. 2, article R45, 2008.
- [13] The Rice Chromosome 3 Sequencing Consortium, "Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species," *Genome Research*, vol. 15, pp. 1–10, 2005.
- [14] International Rice Genome Sequencing Project, "The map-based sequence of the rice genome," *Nature*, vol. 436, no. 7052, pp. 793–800, 2005.
- [15] Z. M. Chen, "Prospect on tea industry in the year of 2000," *Journal of Tea Science*, vol. 14, no. 2, pp. 81–88, 1994.
- [16] N. Khan and H. Mukhtar, "Tea polyphenols for health promotion," *Life Sciences*, vol. 81, no. 7, pp. 519–533, 2007.
- [17] X.-D. Jia, C. Han, and J.-S. Chen, "Tea pigments induce cell-cycle arrest and apoptosis in HepG2 cells," *World Journal of Gastroenterology*, vol. 11, no. 34, pp. 5273–5276, 2005.
- [18] M. Monobe, K. Ema, F. Kato, and M. Maeda-Yamamoto, "Immunostimulating activity of a crude polysaccharide derived from green tea (*Camellia sinensis*) extract," *Journal of Agricultural and Food Chemistry*, vol. 56, no. 4, pp. 1423–1427, 2008.
- [19] J. F. Bukowski and S. S. Percival, "L-theanine intervention enhances human  $\gamma\delta$  T lymphocyte function," *Nutrition Reviews*, vol. 66, no. 2, pp. 96–102, 2008.
- [20] S. M. Karori, F. N. Wachira, J. K. Wanyoko, and R. M. Ngunjiri, "Antioxidant capacity of different types of tea products," *African Journal of Biotechnology*, vol. 6, no. 19, pp. 2287–2296, 2007.
- [21] W. Rumpel, J. Seale, B. Clevidence et al., "Oolong tea increases metabolic rate and fat oxidation in men," *Journal of Nutrition*, vol. 131, no. 11, pp. 2848–2852, 2001.
- [22] M. Sano, M. Suzuki, T. Miyase, K. Yoshino, and M. Maeda-Yamamoto, "Novel antiallergic catechin derivatives isolated from oolong tea," *Journal of Agricultural and Food Chemistry*, vol. 47, no. 5, pp. 1906–1910, 1999.
- [23] S. Sato, A. Adachi, Y. Sasaki, and M. Ghazizadeh, "Oolong tea extract as a substitute for uranyl acetate in staining of ultrathin sections," *Journal of Microscopy*, vol. 229, no. 1, pp. 17–20, 2008.
- [24] Y. Kuroda and Y. Hara, "Antimutagenic and anticarcinogenic activity of tea polyphenols," *Mutation Research*, vol. 436, no. 1, pp. 69–97, 1999.
- [25] L. K. Han, T. Takaku, J. Li et al., "Antiobesity action of oolong tea," *International Journal of Obesity and Related Metabolic Disorders*, vol. 23, pp. 98–105, 1999.
- [26] H. Kurihara, H. Fukami, Y. Toyoda et al., "Inhibitory effect of oolong tea on the oxidative state of low density lipoprotein (LDL)," *Biological and Pharmaceutical Bulletin*, vol. 26, no. 5, pp. 739–742, 2003.
- [27] S. B. Zhou and X. D. Chen, "Tea in Taiwan," *Chinese Tea*, vol. 2006, no. 2, pp. 34–35, 2006.

- [28] L. Chen, Z.-X. Zhou, and Y.-J. Yang, "Genetic improvement and breeding of tea plant (*Camellia sinensis*) in China: from individual selection to hybridization and molecular breeding," *Euphytica*, vol. 154, no. 1-2, pp. 239–248, 2007.
- [29] C. Ma and L. Chen, "Research progress on isolation and cloning of functional genes in tea plants," *Molecular Plant Breeding*, vol. 4, no. 3, supplement, pp. 16–22, 2006.
- [30] J. Tanaka, F. Taniguchi, N. Hirai, and S. Yamaguchi, "Estimation of the genome size of tea (*Camellia sinensis*), *Camellia* (*C. japonica*), and their interspecific hybrids by flow cytometry," *Tea Research Report*, vol. 101, pp. 1–7, 2006.
- [31] J. K. Lin, D. Kudrna, and R. A. Wing, "High Molecular Weight (HMW) genomic DNA preparation from Tea plant (*Camellia sinensis*) for BAC library construction," *Journal of Agricultural Science and Technology*, vol. 3, no. 1, pp. 1–10, 2009.
- [32] D. Peterson, J. Tomkins, D. Frisch, R. Wing, and A. Paterson, "Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide," *Journal of Agricultural Genomics*, vol. 5, pp. 1–100, 2000.
- [33] A. H. Paterson, "Leafing through the genomes of our major crop plants: strategies for capturing unique information," *Nature Reviews Genetics*, vol. 7, no. 3, pp. 174–184, 2006.
- [34] Arabidopsis Genome Initiative, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*," *Nature*, vol. 410, no. 6826, p. 299, 2001.
- [35] A. M. Brunner, V. B. Busov, and S. H. Strauss, "Poplar genome sequence: functional genomics in an ecologically dominant plant species," *Trends in Plant Science*, vol. 9, no. 1, pp. 49–56, 2004.
- [36] O. Jaillon, J.-M. Aury, B. Noel et al., "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla," *Nature*, vol. 449, no. 7161, pp. 463–467, 2007.
- [37] P. S. Schnable, D. Ware, R. S. Fulton et al., "The B73 maize genome: complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, 2009.
- [38] C. A. Hackett, F. N. Wachira, S. Paul, W. Powell, and R. Waugh, "Construction of a genetic linkage map for *Camellia sinensis* (tea)," *Heredity*, vol. 85, no. 4, pp. 346–355, 2000.
- [39] J.-S. Park, J.-B. Kim, B.-S. Hahn et al., "EST analysis of genes involved in secondary metabolism in *Camellia sinensis* (tea), using suppression subtractive hybridization," *Plant Science*, vol. 166, no. 4, pp. 953–961, 2004.
- [40] L. Chen, L.-P. Zhao, and Q.-K. Gao, "Generation and analysis of expressed sequence tags from the tender shoots cDNA library of tea plant (*Camellia sinensis*)," *Plant Science*, vol. 168, no. 2, pp. 359–363, 2005.
- [41] J. Thomas, D. Vijayan, S. D. Joshi, S. Joseph Lopez, and R. Raj Kumar, "Genetic integrity of somaclonal variants in tea (*Camellia sinensis* (L.) O Kuntze) as revealed by inter simple sequence repeats," *Journal of Biotechnology*, vol. 123, no. 2, pp. 149–154, 2006.
- [42] J. Tanaka and F. Taniguchi, "Tea," in *Genome Mapping and Molecular Breeding in Plants*, C. Kole, Ed., vol. 6 of *Technical Crops*, pp. 119–125, 2007.
- [43] X. C. Wan, *Tea Biochemistry*, China Agriculture Press, Beijing, China, 3rd edition, 2003.
- [44] B. Q. Fu, M. Y. Xie, X. P. Nie et al., "Method simplified in assaying tea polysaccharide," *Food Science*, vol. 22, no. 11, pp. 69–73, 2001.
- [45] B. Yüksel and A. H. Paterson, "Construction and characterization of a peanut *HindIII* BAC library," *Theoretical and Applied Genetics*, vol. 111, no. 4, pp. 630–639, 2005.
- [46] C. J. Zeng, H. J. Pan, S. B. Gong, J. Q. Yu, Q. H. Wan, and S. G. Fang, "Giant panda BAC library construction and assembly of a 650-kb contig spanning major histocompatibility complex class II region," *BMC Genomics*, vol. 8, article 315, 2007.
- [47] B. Chalhoub, H. Belcram, and M. Caboche, "Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size," *Plant Biotechnology Journal*, vol. 2, no. 3, pp. 181–188, 2004.
- [48] S. Noir, S. Patheyron, M.-C. Combes, P. Lashermes, and B. Chalhoub, "Construction and characterisation of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.)," *Theoretical and Applied Genetics*, vol. 109, no. 1, pp. 225–230, 2004.
- [49] E. R. Havecker, X. Gao, and D. F. Voytas, "The diversity of LTR retrotransposons," *Genome Biology*, vol. 5, no. 6, article 225, 2004.
- [50] L. Gao, E. M. McCarthy, E. W. Ganko, and J. F. McDonald, "Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequences," *BMC Genomics*, vol. 5, article 18, 2004.
- [51] A. Zuccolo, A. Sebastian, J. Talag et al., "Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*," *BMC Evolutionary Biology*, vol. 7, article 152, 2007.
- [52] T. Sasaki and B. Burr, "International Rice Genome Sequencing Project: the effort to completely sequence the rice genome," *Current Opinion in Plant Biology*, vol. 3, no. 2, pp. 138–141, 2000.
- [53] J. S. S. Ammiraju, A. Zuccolo, Y. Yu et al., "Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus *Oryza*," *Plant Journal*, vol. 52, no. 2, pp. 342–351, 2007.
- [54] B. L. Hurwitz, D. Kudrna, Y. Yu et al., "Rice structural variation: a comparative analysis of structural variation between rice and three of its closest relatives in the genus *Oryza*," *Plant Journal*, vol. 63, no. 6, pp. 990–1003, 2010.
- [55] K. M. Devos, J. K. M. Brown, and J. L. Bennetzen, "Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*," *Genome Research*, vol. 12, no. 7, pp. 1075–1079, 2002.
- [56] J. Ma and J. L. Bennetzen, "Rapid recent growth and divergence of rice nuclear genomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 34, pp. 12404–12410, 2004.
- [57] J. Ma, K. M. Devos, and J. L. Bennetzen, "Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice," *Genome Research*, vol. 14, no. 5, pp. 860–869, 2004.
- [58] S. S. Kaundun and S. Matsumoto, "Identification of processed Japanese green tea based on polymorphisms generated by STS-RFLP analysis," *Journal of Agricultural and Food Chemistry*, vol. 51, no. 7, pp. 1765–1770, 2003.
- [59] J. Huang, J. Li, Y. Huang et al., "Construction of AFLP molecular markers linkage map in tea plant," *Journal of Tea Science*, vol. 25, no. 1, pp. 7–15, 2005.
- [60] J. Q. Jin, H. R. Cui, X. C. Gong, W. Y. Chen, and Y. Xin, "Studies on tea plants (*Camellia sinensis*) germplasms using EST-SSR marker," *Hereditas*, vol. 29, no. 1, pp. 103–108, 2007.
- [61] P. A. N. Punyasiri, I. S. B. Abeysinghe, V. Kumar et al., "Flavonoid biosynthesis in the tea plant *Camellia sinensis*: properties of enzymes of the prominent epicatechin and catechin pathways," *Archives of Biochemistry and Biophysics*, vol. 431, no. 1, pp. 22–30, 2004.