CrossMark

ORIGINAL ARTICLE

# Construction and characterization of a bacterial artificial chromosome library for *Camellia sinensis*

Yuling Tai[1,2] · Hao Wang[3] · Chaoling Wei[1] · Ling Su[2] · Min Li[1] · Liqun Wang[1] ·
Zhaozhao Dai[3] · Lingyun Shao[3] · Jiayue Jiang[1] · Meizhong Luo[3] · Hua Yang[1] ·
Weiwei Deng[1] · Bei Huang[1] · Xiaochun Wan[1] · Jun Sun[1,4]

**Abstract** Tea is a popular and natural non-alcoholic beverage, and is produced from fresh leaves of *Camellia sinensis*. Tea leaves contain many bioactive compounds that have significant health benefits. We constructed a high quality bacterial artificial chromosome (BAC) library by using the fresh petals of *C. sinensis* "Shuchazao" for genome sequencing and improvement of genomic assembly. BAC library is still a significant tool for studies of functional genomes and preservation of precious genetic resources. The BAC library contains 161,280 clones with an average insert size of 113 kb, which represents approximately 6.2-fold coverage of haploid genome equivalents of *C. sinensis*. We characterized 20 complete BAC clones and 738 BAC end sequences (BESs) ranging from 105 to 917 bp. In addition, we predicted cis-regulatory elements of *LAR* (leucoanthocyanidin reductase), *TCS* (caffeine synthase), and *TS* (theanine synthetase)

involved in tea characteristic metabolite synthesis and identified a larger number of light-responsive cis-acting elements in these three genes. Meanwhile, we analyzed alternative splicing of these three genes. Furthermore, 12 pairs of SSR primers were successfully amplified in tea plant DNA. The tea BAC library was a critical resource to accomplish de novo whole-genome sequencing, accelerate gene discovery and enhance molecular breeding of *C. sinensis*.

**Keywords** Tea plant · BAC library · BAC end sequencing · SSR · Alternative splicing

✉ Xiaochun Wan
xcwan@ahau.edu.cn

✉ Jun Sun
sunjun@ahau.edu.cn

1    State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, China

2    School of Life Science, Anhui Agricultural University, Hefei, China

3    College of Life Sciences and Technology, Huazhong Agricultural University, Wuhan, China

4    College of Horticulture, Anhui Agricultural University, Hefei, China

## Introduction

Tea is one of the most popular beverages consumed worldwide, and it is produced from the leaves of the plant *Camellia sinensis* (L.) O. Kuntze, which belongs to the family Theaceae. The leaf of *C. sinensis* is a unique organ usually used because it contains beneficial compounds. To date, nearly 4000 bioactive compounds (Mahmood et al. 2010), such as polyphenols (e.g., catechins), caffeine, and theanine (Liang et al. 2001; Mamati et al. 2006), have been identified in tea leaves. Tea has significant health benefits, such as autoxidation and anticancer activity (Sasazuki et al. 2012; Yen and Chen 1995) and it decreases high blood pressure (Hodgson et al. 2013), helps in preventing cardiovascular diseases (Hollman et al. 1999), and contributes to weight reduction (Auvichayapat et al. 2008).

Rapid progress in gene identification and isolation from *C. sinensis* has been made via RNA-Seq. For example, several studies have focused on the transcriptome of different tissues of *C. sinensis* (Shi et al. 2011a; Wu et al. 2013) and transcriptome profiles of *C. sinensis* during cold acclimation (Wang et al. 2013b). However, compared to many other crops

 Springer

**Fig. 1** *Camellia sinensis* flowers used for library construction



and model plant species, *C. sinensis* has a large (Taniguchi 2006; Wu et al. 2014) and high heterozygosity genome (Ma et al. 2010). Xia et al. (2017) reported the genome of tea tree (cultivar *Yunkang 10*) and offered valuable information to further research of tea tree. However, more publically available molecular resources are required for functional genomic research, genetic studies and breeding programs of *C. sinensis*.

BAC libraries are important resources for the construction of genetic and physical maps (Luo and Wing 2003), gene identification, map-based cloning, comparative genomics analysis of eukaryotic genomes, and marker

development (Shizuya et al. 1992). To date, hundreds of BAC libraries and physical maps have been constructed for model plants (Initiative 2000) and important fungi (Wang et al. 2013c), animals (Song et al. 2011), and crop plants such as rice (Ammiraju et al. 2006; Wang et al. 1995; Li 2005), wheat (Cenci et al. 2003; Janda et al. 2006; Nilmalgoda et al. 2003), maize (O'Sullivan et al. 2001; Tomkins et al. 2002; Wang et al. 2013a; Yim et al. 2002), soybean (Wu et al. 2004), tomato (Frary and Hamilton 2001), barley (Schulte et al. 2011), sorghum (Woo et al. 1994), and pearl millet (Allouis et al. 2001). These BAC libraries have been used for the development

**Fig. 2** Analysis of the insert sizes of the *Hind* III library. **a** DNA analysis of random BAC clones from the *Camellia sinensis* "Shuchazao" BAC library by using pulsed-field gel electrophoresis. *M* represents the midrange PFG marker. **b** Insert size distribution of the *Camellia sinensis* "Shuchazao" BAC library. In total, 430 random clones were analyzed
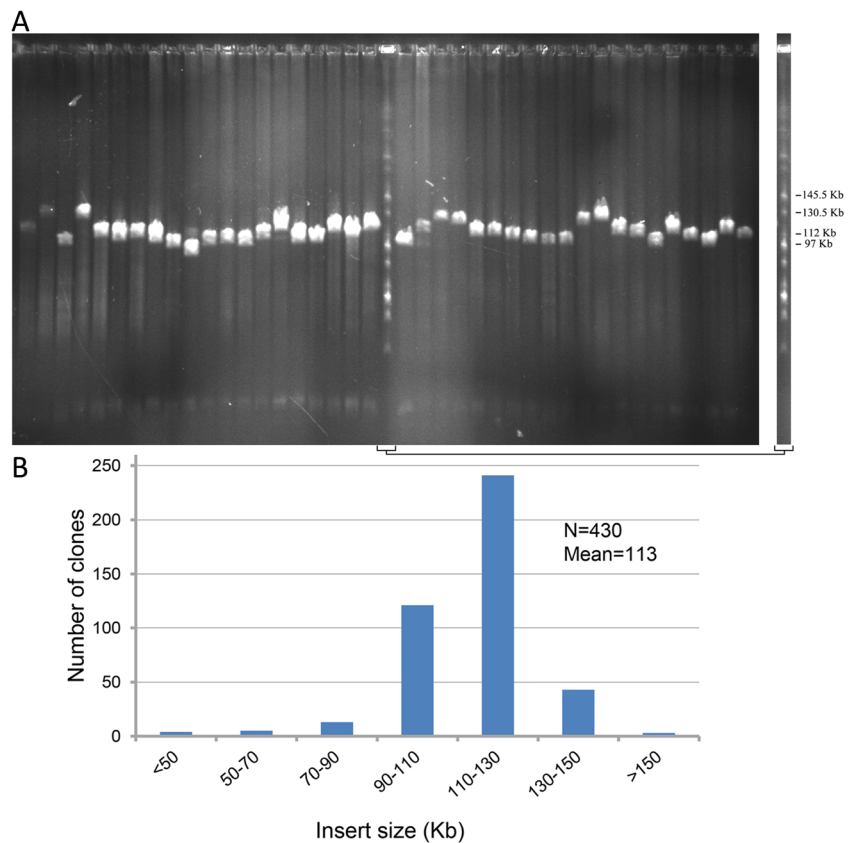
**Table 1** Characterization of 20 BACs

| BAC name | Raw reads number | Succeed reads number | Raw reads average length | Total BAC length | Coverage | GC% |
|---|---|---|---|---|---|---|
| Csi205B03 | 1536 | 1459 | 796 | 121161 | 9.59 | 38.91% |
| Csi205D12 | 1536 | 1504 | 884 | 110408 | 11.87 | 37.78% |
| Csi092H17 | 1536 | 1511 | 737 | 110950 | 10.04 | 36.32% |
| Csi044B12 | 1536 | 1361 | 740 | 83976 | 11.99 | 44.34% |
| Csi020o15 | 1536 | 1109 | 716 | 135319 | 5.87 | 34.57% |
| Csi106D7 | 1536 | 1475 | 784 | 112754 | 10.27 | 40.90% |
| Csi271P8 | 1152 | 1117 | 761 | 82780 | 10.27 | 33.82% |
| Csi037A20 | 2697 | 2398 | 707 | 131644 | 12.47 | 38.62% |
| Csi037L05 | 1938 | 1757 | 727 | 120359 | 9.26 | 37.66% |
| Csi037o01 | 1947 | 1786 | 693 | 116692 | 9.96 | 34.66% |
| Csi037o08 | 1843 | 1679 | 716 | 144,930 | 8.20 | 39.33% |
| Csi047o18 | 1928 | 1753 | 709 | 102542 | 11.55 | 37.99% |
| Csi274K14 | 2045 | 1786 | 699 | 125247 | 9.46 | 40.85% |
| Csi205A01 | 1404 | 1175 | 805 | 108791 | 8.28 | 40.02% |
| Csi205E16 | 1920 | 1801 | 806 | 112344 | 12.92 | 37.87% |
| Csi205H20 | 1980 | 1765 | 777 | 122164 | 10.49 | 39.44% |
| Csi205k12 | 1755 | 1556 | 778 | 109985 | 10.60 | 39.04% |
| Csi205k18 | 1520 | 1459 | 841 | 116287 | 10.55 | 38.34 |
| Csi205J21 | 2304 | 2135 | 823 | 115055 | 15.27 | 39.34% |
| Csi205E18 | 1920 | 1782 | 761 | 70165 | 19.33 | 34.84% |

of physical maps (Chen et al. 2002; Wei et al. 2009), BAC-to-BAC genome sequencing (Schnable et al. 2009; Wei et al. 2009), positional cloning of genes of agronomic importance, comparative analysis of genome structures (Ammiraju et al. 2008; Chen et al. 1997; Lin et al. 2012; Lu et al. 2009; Messing and Llaca 1998; Pan et al. 2014; Wang and Dooner 2012), and genome assemblies of whole-genome shotgun sequences (Yu et al. 2005). BAC end sequences (BESs) are genomic resources that enhance the value of BAC libraries by providing partial sequence information that can be used to understand genome content and architecture and develop genetic markers (David et al. 2008). BESs embedded in a

physical map can be used as anchors in genome comparisons to detect sequence assembly errors in the same source genome and large structural changes in phylogenetically close genomes (Zhu et al. 1997).

BAC libraries for *C. sinensis* will facilitate in DNA-marker analysis and functional gene cloning of this economically important species. Lin et al. (2011) constructed a BAC library for the cultivar "Chin-shin oolong." In this study, we constructed a high quality BAC library for *C. sinensis* "Shuchazao" using fresh petals. Three complete BAC clones and ends of 384 BAC clones selected randomly from the BAC library were sequenced and analyzed. Understanding the

**Fig. 3** BES length distribution; length was calculated after trimming the vector sequence
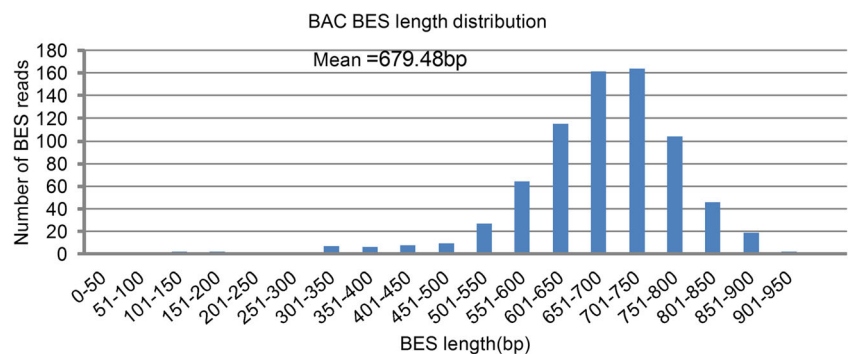
**Table 2** Repeat types within *C. sinensis* BAC sequences identified using Repeatmasker

| | Number of elements | Length occupied | Percentage of sequence |
|---|---|---|---|
| Retroelements | 266 | 279,641 bp | 12.40% |
|   SINEs | 2 | 177 bp | 0.01% |
|     Penelope | 0 | 0 bp | 0.00% |
|   LINEs | 18 | 10,754 bp | 0.48% |
|     CRE/SLACS | 0 | 0 bp | 0.00% |
|     L2/CR1/Rex | 0 | 0 bp | 0.00% |
|     R1/LOA/Jockey | 0 | 0 bp | 0.00% |
|     R2/R4/NeSL | 0 | 0 bp | 0.00% |
|     RTE/Bov-B | 7 | 5169 bp | 0.23% |
|     L1/CIN4 | 11 | 5585 bp | 0.25% |
| LTR elements | 246 | 268,710 bp | 11.91% |
|   BEL/Pao | 0 | 0 bp | 0.00% |
|   Ty1/Copia | 60 | 91,224 bp | 4.04% |
|   Gypsy/DIRS1 | 175 | 166,436 bp | 7.38% |
|   Retroviral | 0 | 0 bp | 0.00% |
| DNA transposons | 69 | 19,462 bp | 0.86% |
|   hobo-Activator | 14 | 6478 bp | 0.29% |
|   Tc1-IS630-Pogo | 0 | 0 bp | 0.00% |
|   En-Spm | 0 | 0 bp | 0.00% |
|   MuDR-IS905 | 0 | 0 bp | 0.00% |
|   PiggyBac | 0 | 0 bp | 0.00% |
|   Tourist/Harbinger | 13 | 2872 bp | 0.13% |
|   Other (mirage, P-element, Transib) | 0 | 0 bp | 0.00% |
|   Rolling-circles | 0 | 0 bp | 0.00% |
| Unclassified | 5 | 294 bp | 0.24% |
| Total interspersed repeats | | 299,397 bp | 13.28% |
| Small RNA | 2 | 141 bp | 0.01% |
| Satellites | 1 | 49 bp | 0.00% |
| Simple repeats | 15 | 2735 bp | 0.12% |
| Low complexity | 4 | 841 bp | 0.04% |

characteristics of the BAC library will facilitate in using it in functional genomic research of *C. sinensis*. The BAC library and BESs are valuable resources for assembly optimization of genome for *C. sinensis*.

## Materials and methods

### Plant materials

For construction of the BAC library, *C. sinensis* "Shuchazao" plants were grown at the De Chang fabrication base in Anhui, China. Fresh petals (Fig. 1) were collected from the tea plants and immediately frozen by submersion in liquid nitrogen, followed by short-term storage at −80 °C.

### Library construction

The nuclear DNA was extracted by s described by Luo (Luo and Wing 2003). Firstly, grind about 50 g of petals in liquid nitrogen with a mortar and a pestle, and polyvinylpyrrolidone (PVP40) was used for removal of phenolic compounds in the milling petal to reduce DNA degradation. Secondly, transfer the ground tissue into NIBM (NIB with 0.1% β-mercaptoethanol) for 15 min with frequent and gentle shaking. The precipitation was resuspended with NIBM after centrifugation at 2400 g for 15 min. Finally, the nuclear DNA was imbed with 1% low melting temperature agarose. Then, we conducted partial digestion tests by using different amounts of the restriction enzyme *Hind* III per reaction to determine optimal conditions. We observed that 1 U/μL per reaction and incubation at 37 °C for 15 min generated fragments with most being between 100 and 300 kb, and digested 8 plugs (high

**Table 3** Repeat types within *C. sinensis* BAC end sequences identified using Repeatmasker

|  | Number of elements | Length occupied | Percentage of sequence |
|---|---|---|---|
| Retroelements | 148 | 51,712 bp | 10.56% |
|   SINEs | 1 | 72 bp | 0.01% |
|     Penelope | 0 | 0 bp | 0.00% |
|   LINEs | 18 | 5242 bp | 1.07% |
|     CRE/SLACS | 0 | 0 bp | 0.00% |
|     L2/CR1/Rex | 0 | 0 bp | 0.00% |
|     R1/LOA/Jockey | 0 | 0 bp | 0.00% |
|     R2/R4/NeSL | 0 | 0 bp | 0.00% |
|     RTE/Bov-B | 4 | 1361 bp | 0.28% |
|     L1/CIN4 | 14 | 3881 bp | 0.79% |
|     LTR elements: | 129 | 46,398 bp | 9.47% |
|     BEL/Pao | 0 | 0 bp | 0.00% |
|     Ty1/Copia | 46 | 18,200 bp | 3.72% |
|     Gypsy/DIRS1 | 81 | 27,742 bp | 5.66% |
|     Retroviral | 0 | 0 bp | 0.00% |
|   DNA transposons | 13 | 1251 bp | 0.26% |
|     hobo-Activator | 3 | 196 bp | 0.04% |
|     Tc1-IS630-Pogo | 1 | 13 bp | 0.00% |
|     En-Spm | 0 | 0 bp | 0.00% |
|     MuDR-IS905 | 0 | 0 bp | 0.00% |
|     PiggyBac | 0 | 0 bp | 0.00% |
|     Tourist/Harbinger | 3 | 324 bp | 0.07% |
|     Other (Mirage, P-element, Transib) | 0 | 0 bp | 0.00% |
|     Rolling-circles | 0 | 0 bp | 0.00% |
| Unclassified | 2 | 1173 bp | 0.24% |
| Total interspersed repeats | | 54,136 bp | 11.05% |
| Small RNA | 1 | 72 bp | 0.01% |
| Satellites | 0 | 0 bp | 0.00% |
| Simple repeats | 227 | 9617 bp | 1.96% |
| Low complexity | 39 | 1971 bp | 0.40% |

molecular weight (HMW) DNA buried in low melting-point agarose) in this condition. After twice the size of selections, DNA fragments from 100 to 300 kb in size were eluted from the gel slices via electroelution and ligated into the pIndigoBAC536-s vector (Shi et al. 2011b). The ligated DNA was transformed into *Escherichia coli* strain DH10B (Invitrogen, USA) via electroporation.

**Insert sizing**

To evaluate the average insert size of the BAC library, we randomly selected 430 BAC clones from 384-well plates and isolated plasmid DNAs by using the alkaline lysis method (Sanmiguel and Bennetzen 1998). The DNAs were digested with I-*Sce*I, and separated by pulsed-field gel electrophoresis

**Table 4** SSR counts for BACs

| Motif | Counts | Average_length | Counts/mbp | GC content | Files/SSR |
|---|---|---|---|---|---|
| Mononucleotide | 16 | 18.56 | 45.27 | 0 | 0.19 |
| Dinucleotide | 36 | 21.14 | 101.86 | 0.24 | 0.08 |
| Trinucleotide | 13 | 28.85 | 36.78 | 0.15 | 0.23 |
| Tetranucleotide | 17 | 17.88 | 48.1 | 0.06 | 0.18 |
| Pentanucleotide | 23 | 17.74 | 65.08 | 0.26 | 0.13 |
| Hexanucleotide | 8 | 21.13 | 22.63 | 0.36 | 0.38 |

**Table 5**  SSR counts for BES

| Motif | Counts | Average_length | Counts/mbp | GC content | Files/SSR |
|---|---|---|---|---|---|
| Mononucleotide | 26 | 18.31 | 53.07 | 0.03 | 27.5 |
| Dinucleotide | 49 | 21.08 | 100.02 | 0.28 | 14.59 |
| Trinucleotide | 16 | 18.44 | 32.66 | 0.28 | 44.69 |
| Tetranucleotide | 24 | 18.96 | 48.99 | 0.1 | 29.79 |
| Pentanucleotide | 20 | 17 | 40.82 | 0.18 | 35.75 |
| Hexanucleotide | 7 | 23.43 | 14.29 | 0.33 | 102.14 |

(PFGE) on 1% agarose gels (Luo and Wing 2003). The insert size of each clone was estimated using the midrange PFG marker (New England BioLabs, USA) as the molecular weight standard.

### Shotgun library construction and sequencing of BAC clones

BAC clones were cultured in Luria broth medium containing 12.5 μg/mL chloramphenicol for 16 h, and then centrifuged to harvest the pellets. The QIAGEN large construct kit (QIAGEN Lot.12462, Hilden, Germany) was used to obtain high quality, genomic DNA-free BAC plasmids. The plasmid DNA was sheared to 3–5 kb fragments by using ultrasonic waves. Klenow fragments and T4 DNA polymerase were used for blunt-end repair. After

adding nucleotide A to 3′-end, DNA was ligated to the T-easy vector (Promega, Fitchburg, Wisconsin, USA). The DH5α competent cell (YEASTERN, ECOS Cat. No. FYE610-10VL, Taipei, Taiwan) was used for transformation. The transformed clones were maintained in 384-well plates for storage and sequencing.

Plasmids were extracted using AxyPrep Easy-96 (AXYGEN, AP-E96-P-24G, Corning, New York, USA) and dissolved in 30 μL of ddH₂O for Sanger sequencing. The sequencing reaction was performed according to the protocol of the Applied Biosystems, ABI BigDye terminator v 3.1 and v 1.1 cycle sequencing kits (Waltham, Massachusetts, USA). The following primers were used: T7 (5′-TAATACGACTCACTATAGGG-3′) and SP6 (5′-ATTTAGGTGACACTATAG-3′). Sequence data were collected using ABI 3730.
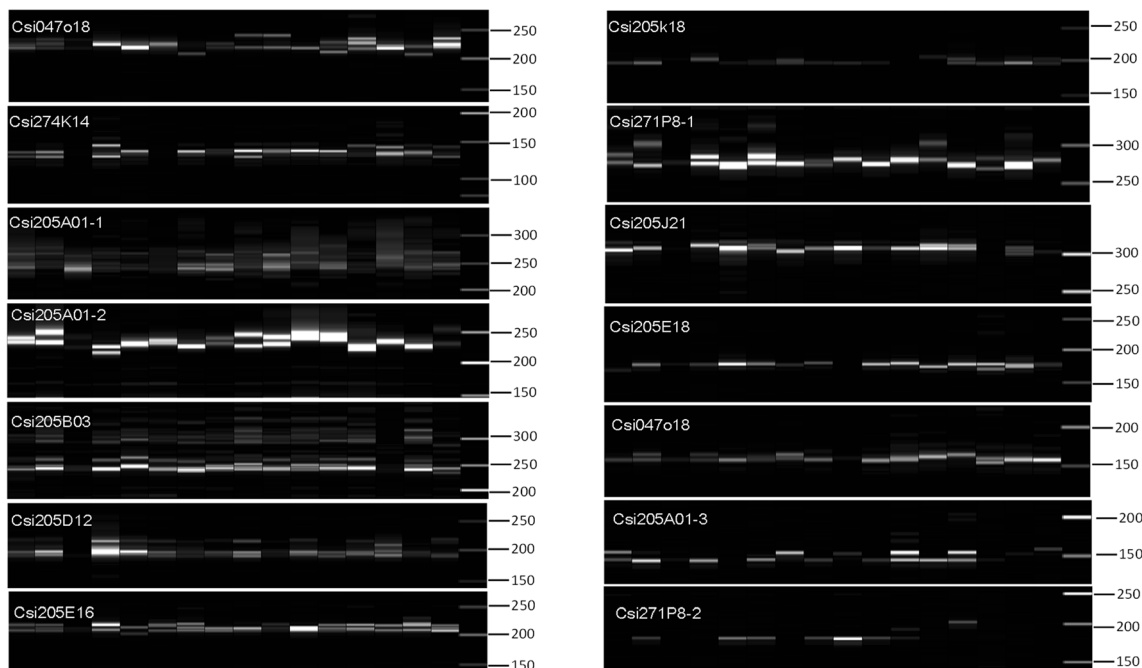


**Fig. 4**  The SSR patterns resulted from amplification with 14 pairs of SSR primers. 1 C.sinensis var. assamica cv. Zijuan. 2 C.sinensis cv. Anji-baicha, 3 C.sinensis cv. Anhui 1, 4 C.sinensis cv. Longjing-changye, 5 C.sinensis cv. Zhonghuang 1, 6 C.sinensis cv. Zhenong 113, 7 C.sinensis cv. Hongyan 12, 8 C.sinensis cv. Zhenong 117, 9 C.sinensis cv. Yingshuang, 10 C.sinensis cv. Fuzao 2, 11 C.sinensis cv. Fuding-dabaicha, 12 C.sinensis cv. Longjing 43, 13 C.sinensis cv. Shuchazao, 14 C.sinensis var. assamica cv. Yinghong 9, 15 C.sinensis cv. Tie-guanyin, 16 C.sinensis var. assamica cv. Yunkang 10

```
>LAR 1500nt
+ TTTTTTTTCC ATAACCAACA ATAAGAAGAT AATAATAATT TTTGAATTAA TATTAAGTAC AAAAAATTAA
- AAAAAAAAGG TATTGGTTGT TATTCTTCTA TTATTATTAA AAACTTAATT ATAATTCATG TTTTTTAATT

+ ATAAATTATT ACAAAAGTAT TTCACTTAAA ATTTTACAAT TAACTTTGAG GAAAATTAAA ACGAAAGTAT
- TATTTAATAA TGTTTTCATA AAGTGAATTT TAAAATGTTA ATTGAAACTC CTTTTAATTT TGCTTTCATA

+ TGCGTGTAAC TAGTGTATTA TGGGCAATGT AAAAAAATCA TATATTTAAG TGTTAAAATA AATTATATAA
- ACGCACATTG ATCACATAAT ACCCGTTACA TTTTTTTAGT ATATAAATTC ACAATTTTAT TTAATATATT

+ TGGTAAGATA AATCAAATTT TTACAAAATT TGAGTAAAAA TTTCATGTAA TAGTGATATC ACGTTATTGT
- ACCCATTCTAT TTAGTTTAAA AATGTTTTAA ACTCATTTTT AAAGTACATT ATCACTATAG TGCAATAACA

+ TGACATATGA TATTACCAAT CAATAGTAAA GCAACATATG GAAGGGTGAA CTCCACATGC AAATATAACT
- ACTGTATACT ATAATGGTTA GTTATCATTT CGTTGTATAC CTTCCCACTT GAGGTGTACG TTTATATTGA

+ AATACTCTAG TAATTTTTAC ATATATATAG TGAGATTAAT TATATTTGCA TATCAATACT AATGGATATT
- TTATGAGATAC ATTAAAAATG TATATATATC ACTCTAATTA ATATAAACGT ATAGTTATGA TTACCTATAA

+ ACTTTTATAT TGAGATTTTT TTCATAAAAT TTTCATACAT AAAAATATAA TTTGTTGTAG TTACTTCATT
- TGAAAATATA ACTCTAAAAA AAGTATTTTA AAAGTATGTA TTTTTATATT AAACAACATC AATGAAGTAA

+ TTTAATAAAA CAGTCACTAT TGTTTTCATA AATTTAAAAA TGCCCTTGAT ATTATATTCT AAATATTATT
- AAATTATTTT GTCAGTGATA ACAAAAGTAT TTAAATTTTT ACGGGAACTA TAATATAAGA TTTATAATAA

+ TTTGTTCAAA ATTTTATTTA TATACATATG TAAAAATAAA TAAATAAATA AATACCAGGG GGCCACATGA
- AAACAAGTTT TAAAATAAAT ATATGTATAC ATTTTTATTT ATTTATTTAT TTATGGTCCC CCGGTGTACT

+ CCCCTGGTCA CCGTGTGGCT CCGTCACTAG TTGCAACCGA GTACTCTTTG AATCATATTC CATTATATTT
- GGGGACCAGT GGCACACCGA GGCAGTGATC AACGTTGGCT CATGAGAAAC TTAGTATAAG GTAATATAAA

+ ATTAATTAAC TTCGGAAGTG GATTGATTTT TTAACTTTTT GACTATTTTA ACCGGGGGCT AAAAAGTTTT
- TAATTAATTG AAGCCTTCAC CTAACTAAAA AATTGAAAAA CTGATAAAAT TGGCCCCCGA TTTTTCAAAA

+ GTTTGGGAAG AGCTTTTTGA TTTTTTTGAC TTTTTTGTAAG AGAAAGAATG TAATGATTAT GGGTTAGGTG
- CAAACCCTTC TCGAAAAACT AAAAAAACTG AAAAACATTC TCTTTCTTAC ATTACTAATA CCCAATCCAC

+ TGATTTTTTT TTTTATTGAG CTAAAAAGCA TATTTTATCT TTTTGTAACT AAAATAACCT CAAATCCAAC
- ACTAAAAAAA AAAATAACTC GATTTTTCGT ATAAAATAGA AAAACATTGA TTTTATTGGA GTTTAGGTTG

+ CTAACCTTCT TCCTAAAGGA AGCCATATAT TGTTCTTAAA ATCCTAGATT TATTCAAGCG ACACAAAAAA
- GATTGGAAGA AGGATTTCCT TCGGTATATA ACAAGAATTT TAGGATCTAA ATAAGTTCGC TGTGTTTTTT

+ ATAATATCAT AAAATAAGCT TTTTTTTGTG AAAAAAACAA AAAAAGCATT TATGTCTACA CAAAATGTTA
- TATTATAGTA TTTTATTCGA AAAAAAACAC TTTTTTTGTT TTTTTCGTAA ATACAGATGT GTTTTACAAT

+ GGATTTTTTA TTATTCAAAA AAGAATTATA ATTGACCCCG TTGAAGTGAA AAATATTTTT TTTGAGAAAT
- CCTAAAAAAT AATAAGTTTT TTCTTAATAT TAACTGGGGC AACTTCACTT TTTATAAAAA AAACTCTTTA

+ TTCAAGTCGG GTCCCTTAAA TGTGGTAGGA ACCATTTAGT TATCCGAGCA CGATTTTCTC AACAATCAGT
- AAGTTCAGCC CAGGGAATTT ACACCATCCT TGGTAAATCA ATAGGCTCGT GCTAAAAGAG TTGTTAGTCA

+ CAATGCACAC TAGAAAAAAA TAAAAATAAA AATAAAATAA AATCAAAGGT ACACGTTAAA TACTTGAATT
- GTTACGTGTG ATCTTTTTTT ATTTTTATTT TTATTTTATT TTAGTTTCCA TGTGCAATTT ATGAACTTAA

+ AGATTATGAT ACACCAAACA ATGTAAAAAC ATTAAATAAA TAAGTTCTAC AAGAAAGTTG GAATTCCGTC
- TCTAATACTA TGTGGTTTGT TACATTTTTG TAATTTATTT ATTCAAGATG TTCTTTCAAC CTTAAGGCAG

+ CGTTTGAATT GAAGGGTAGC TGGAAATTAT GTGCTTCTAG TTATTTAATC TACCACCATC CTTGCCATCT
- GCAAACTTAA CTTCCCATCG ACCTTTAATA CACGAAGATC AATAAATTAG ATGGTGGTAG GAACGGTAGA

+ TCATACTCAT ATACACATAT ATATATATAT ATACACACAT ACACTTATCC AACGCGTGTCT CCAAATATCA
- AGTATGAGTA TATGTGTATA TATATATATA TATGTGTATA TGTGAATAGG TTGCGACAGA GGTTTATAGT

+ AGGACTAGAG GAAGGTACGT AACGAAAAG
- TCCTGATCTC CTTCCATGCA TTGCTTTTC
```

**Legend (color key):**
- ATCT-motif:part of a conserved DNA module involved in light responsiveness
- Box 4:part of a conserved DNA module involved in light responsiveness
- Box-W1:fungal elicitor responsive element
- CAAT-box:common cis-acting element in promoter and enhancer regions
- CCAAT-box:MYBHv1 binding site
- CGTCA-motif:cis-acting regulatory element involved in the MeJA-responsiveness
- G-Box:cis-acting regulatory element involved in light responsiveness
- G-box:cis-acting regulatory element involved in light responsiveness
- GT1-motif:light responsive element
- GTGGC-motif:part of a light responsive element
- HSE:cis-acting element involved in heat stress responsiveness
- CCAAT-box:MYBHv1 binding site
- MRE:MYB binding site involved in light responsiveness
- TC-rich repeats:cis-acting element involved in defense and stress responsiveness
- TCT-motif:part of a light responsive element
- TGACG-motif:cis-acting regulatory element involved in the MeJA-responsiveness
- chs-Unit 1 m1:part of a light responsive element
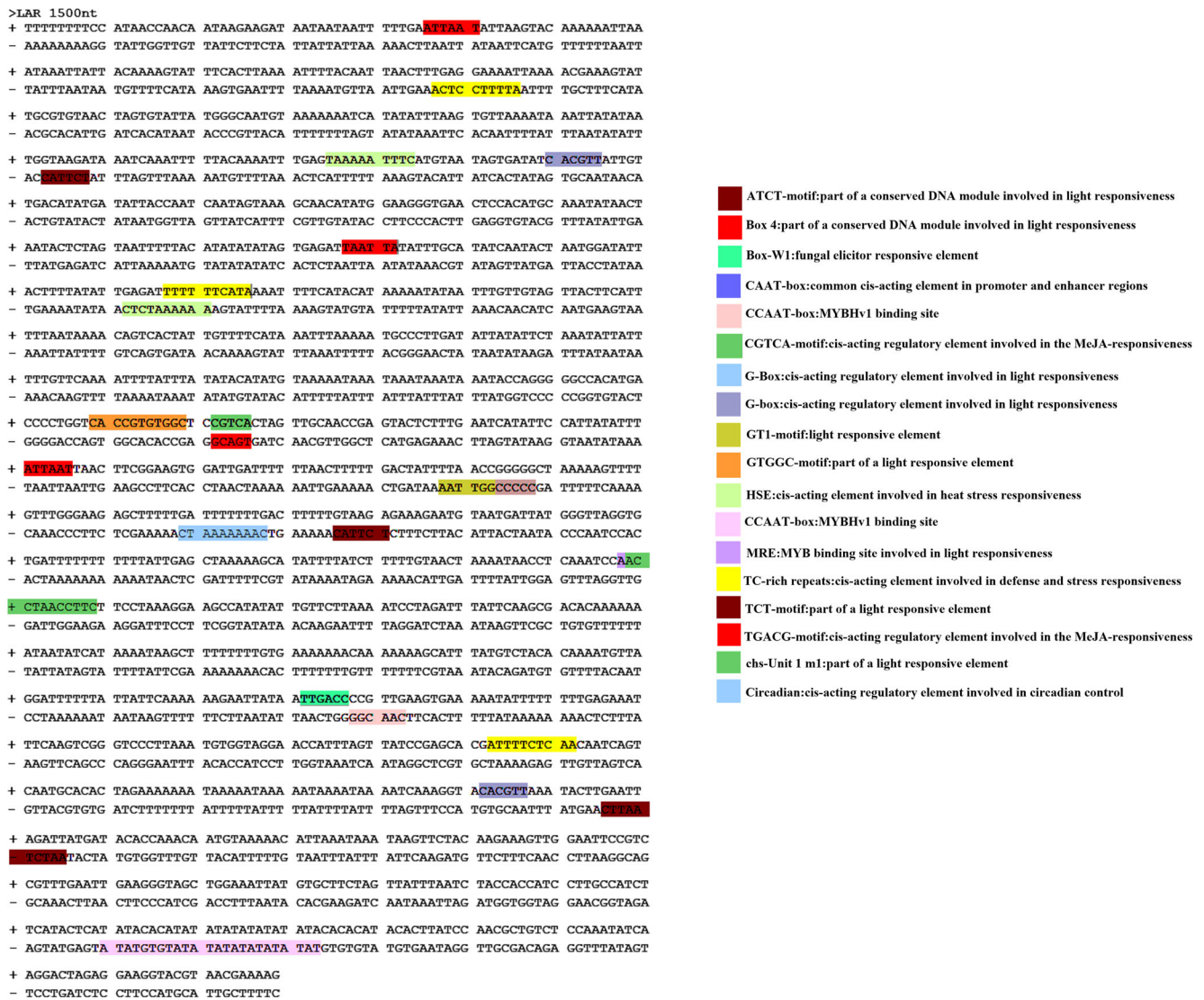- Circadian:cis-acting regulatory element involved in circadian control

**Fig. 5** Promoter prediction of *LAS*

## Shotgun sequence assembly

Raw sequence data were collected, and phred (Ewing and Green 1998) was used for base calling; cut-off value was set at 0.05 to get reliable sequences in the phd format. Phd2fasta was used to transfer the phd file to fasta and fasta quality files. The vector sequence was masked by crossmatch. The sequence was assembled using Phrap with default parameters. Gaps were closed using PCR, and BAC sequences were finished using Consed (Gordon 2003; Gordon et al. 1998).

## BAC end sequencing

The BAC plasmids for end sequencing were extracted using the standard alkaline lysis method and dissolved in 10 μL of ddH₂O for Sanger sequencing. The sequencing reaction was performed according to the protocol of ABI BigDye terminator v3.1 (https://tools.lifetechnologies.com/content/sfs/manuals/cms_081527.pdf). The following sequencing primers were used: pIndigoF, 5′-aacgacggccagtgaattg-3′; pIndigoR, 5′-gataacaatttcacacagg-3′. ABI 3730 was used to sequence BAC clones at both ends.

## Search for repetitive sequences in BAC sequences and BESs

BAC sequences and BESs were searched for repeat sequences by using RepeatMasker (Smit AFA, Hubley R & Green P. RepeatMasker Open-4.0.2013–2015 <http://www.repeatmasker.org>. Smit AFA, Hubley R & Green P. RepeatMasker Open-3.0.1996–2010 <http://www.repeatmasker.org>.) against the available repetitive sequence databases.
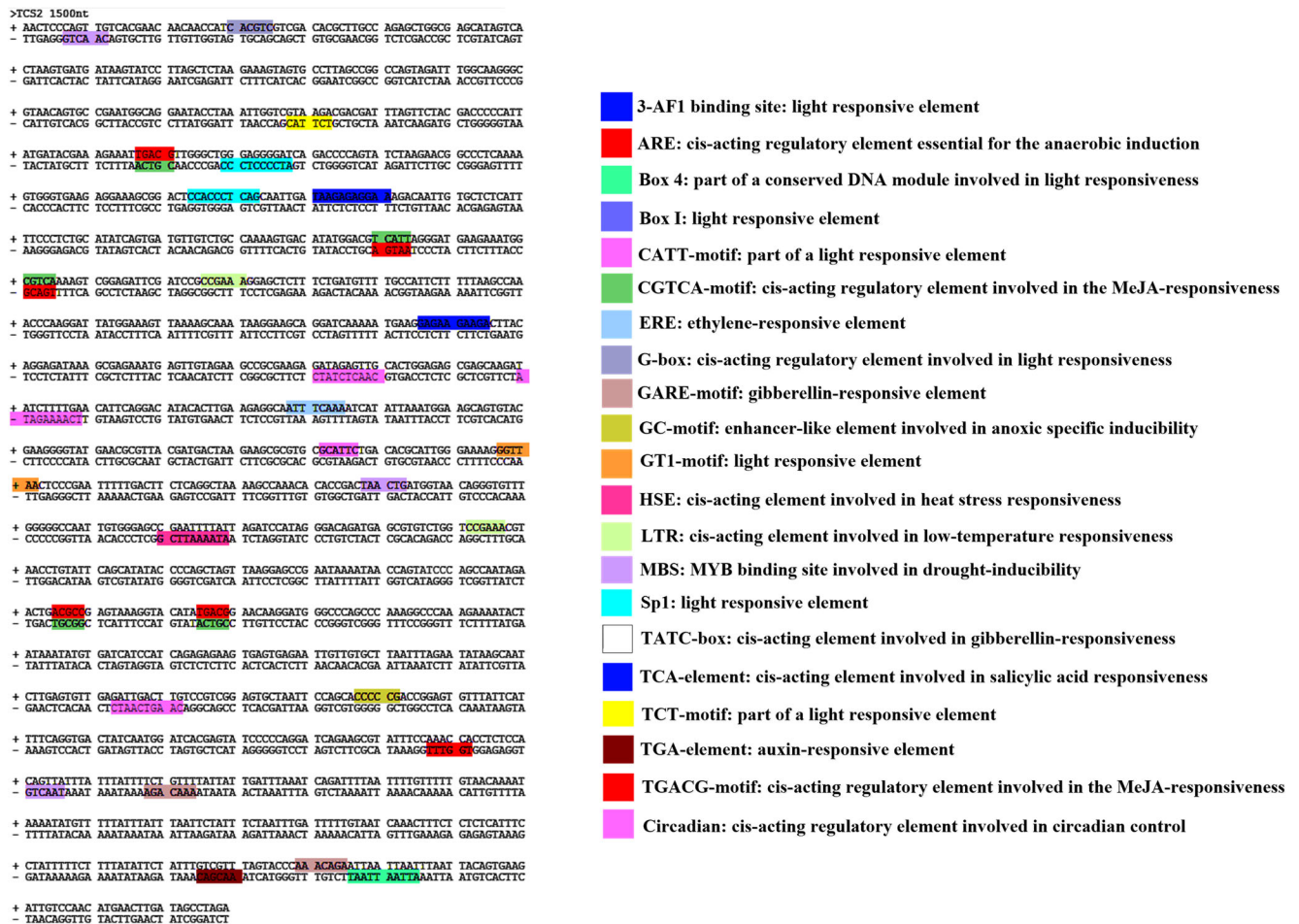
>TCS2 1500nt

```
+ AACTCCCAGT TGTCACGAAC AACAACCATC ACGTCGTCGA CACGCTTGCC AGAGCTGGCG AGCATAGTCA
- TTGAGGGTCA ACAGTGCTTG TTGTTGGTAG TGCAGCAGCT GTGCGAACGG TCTCGACCGC TCGTATCAGT

+ CTAAGTGATG ATAAGTATCC TTAGCTCTAA GAAAGTAGTG CCTTAGCCGG CCAGTAGATT TGGCAAGGGC
- GATTCACTAC TATTCATAGG AATCGAGATT CTTTCATCAC GGAATCGGCC GGTCATCTAA ACCGTTCCCG

+ GTAACAGTGC CGAATGGCAG GAATACCTAA ATTGGTCGTA AGACGACGAT TTAGTTCTAC GACCCCCATT
- CATTGTCACG GCTTACCGTC CTTATGGATT TAACCAGCAT TCTGCTGCTA AATCAAGATG CTGGGGGTAA

+ ATGATACGAA AGAAATTGAC GTTGGGCTGG GAGGGGATCA TCTAAGAACG GCCCTCAAAA
- TACTATGCTT TCTTTAACTG CAACCCGACC CTCCCCTAGT CTGGGGTCAT AGATTCTTGC CGGGAGTTTT

+ GTGGGTGAAG AGGAAAGCGG ACTCCACCCT CAGCAATTGA ATAAGAGAGA AAGACAATTG TGCTCTCATT
- CACCCACTTC TCCTTTCGCC TGAGGTGGGA GTCGTTAACT ATTCTCTCCT TTCTGTTAAC ACGAGAGTAA

+ TTCCCTCTGC ATATCAGTGA TGTTGTCTGC CAAAAGTGAC ATATGGACGT CATTAGGGAT GAAGAAATGA
- AAGGGAGACG TATAGTCACT ACAACAGACG GTTTTCACTG TATACCTGCA GTAATCCCTA CTTCTTTACT

+ CGTCAAAAGT CGGAGATTCG ATCCGCCGAA AGGAGCTCTT TCTGATGTTT TGCCATTCTT TTTAAGCCAA
- GCAGTTTTCA GCCTCTAAGC TAGGCGGCTT TCCTCGAGAA AGACTACAAA ACGGTAAGAA AAATTCGGTT

+ ACCCAAGGAT TATGGAAAGT TAAAAGCAAA TAAGGAAGCA GGATCAAAAA TGAAGGAGAA GAAGACTTAC
- TGGGTTCCTA ATACCTTTCA ATTTTCGTTT ATTCCTTCGT CCTAGTTTTT ACTTCCTCTT CTTCTGAATG

+ AGGAGATAAA GCGAGAAATG AGTTGTAGAA GCCGCGAAGA GATAGAGTTG CACTGGAGAG CGAGCAAGAT
- TCCTCTATTT CGCTCTTTAC TCAACATCTT CGGCGCTTCT CTATCTCAAC GTGACCTCTC GCTCGTTCTA

+ ATCTTTTGAA CATTCAGGAC ATACACTTGA AGAGGCAATT TCAAAATCAT ATTAAATGGA AGCAGTGTAC
- TAGAAAACTT GTAAGTCCTG TATGTGAACT TCTCCGTTAA AGTTTTAGTA TAATTTACCT TCGTCACATG

+ GAAGGGGTAT GAACGCGTTA CGATGACTAA GAAGCGCGTG GCCATTCTGA CACGCATTGG GAAAAGGGTT
- CTTCCCCATA CTTGCGCAAT GCTACTGATT CTTCGCGCAC GCGTAAGACT GTGCGTAACC CTTTTCCCAA

+ AACTCCCGAA TTTTTGACTT CTCAGGCTAA AAGCCAAACA CACCGACTAA CTGATGGTAA CAGGGTGTTT
- TTGAGGGCTT AAAAACTGAA GAGTCCGATT TTCGGTTTGT GTGGCTGATT GACTACCATT GTCCCACAAA

+ GGGGGCCAAT TGTGGGAGCC GAATTTTATT AGATCCATAG GGACAGATGA GCGTGTCTGG TCCGAAACGT
- CCCCCGGTTA ACACCCTCGG CTTAAAATAA TCTAGGTATC CCTGTCTACT CGCACAGACC AGGCTTGCA

+ AACCTGTATT CAGCATATAC CCCAGCTAGT TAAGGAGCCG AATAAAATAA CCAGTATCCC AGCCAATAGA
- TTGGACATAA GTCGTATATG GGGTCGATCA ATTCCTCGGC TTATTTTATT GGTCATAGGG TCGGTTATCT

+ ACTGACGCG AGTAAAGGTA CATATGACG AACAAGGATG GACCCGCCC AAAGGCCCAA AGAAAATACT
- TGACTGCGGC TCATTTCCAT GTATACTGCC TTGTTCCTAC CCGGGTCGGG TTTCCGGGTT TCTTTTATGA

+ ATAAATATGT GATCATCCAT CAGAGAGAAG TGAGTGAGAA TTGTTGTGCT TAATTTAGAA TATAAGCAAT
- TATTTATACA CTAGTAGGTA GTCTCTCTTC ACTCACTCTT AACAACACGA ATAAAATCTT ATATTCGTTA

+ CTTGAGTGTT GAGATTGACT TGTCCGTCGG AGTGCTAATT CCAGCACCCC CGACCGGAGT GTTTATTCAT
- GAACTCACAA CTCTAACTGA ACAGGCAGCC TCACGATTAA GGTCGTGGGG GCTGGCCTCA CAAATAAGTA

+ TTTCAGGTGA CTATCAATGG ATCACGAGTA TCCCCCAGGA TCAGAAGCGT ATTTCCAAAC CACCTCTCCA
- AAAGTCCACT GATAGTTACC TAGTGCTCAT AGGGGGTCCT AGTCTTCGCA TAAAGGTTG GTGGAGAGGT

+ CAGTTATTTA TTTATTTCTT GTTTTATTAT TGATTTAAAT CAGATTTTAA TTTTGTTTTT GTAACAAAAT
- GTCAATAAAT AAATAAAAGA CAAAATAATA ACTAAATTTA GTCTAAAATT ATCAAATTTTA CATTGTTTTA

+ AAAATATGTT TTTAATTTATT TAATTCTATT TCTAATTTGA TTTTTGTAAT CAAACTTTCT CTCTCATTTC
- TTTTATACAA AAATAAATAA ATTAAGATAA AGATTAAACT AAAAACATTA GTTTGAAAGA GAGAGTAAAG

+ CTAITTTTCT TTTATATTCT ATTTGTCGTT TAGTACCCAA ACAGAATTAA TTAATTTAAT TACAGTGAAG
- GATAAAAAGA AAATATAAGA TAAACAGCAA ATCATGGGTT TGTCTTAATT AATTAAATTA ATGTCACTTC

+ AITGTCCAAC ATGAACTTGA TAGCCTAGA
- TAACAGGTTG TACTTGAACT ATCGGATCT
```

| | |
|---|---|
| ■ (blue) | **3-AF1 binding site: light responsive element** |
| ■ (red) | **ARE: cis-acting regulatory element essential for the anaerobic induction** |
| ■ (green) | **Box 4: part of a conserved DNA module involved in light responsiveness** |
| ■ (blue-violet) | **Box I: light responsive element** |
| ■ (magenta) | **CATT-motif: part of a light responsive element** |
| ■ (green) | **CGTCA-motif: cis-acting regulatory element involved in the MeJA-responsiveness** |
| ■ (light blue) | **ERE: ethylene-responsive element** |
| ■ (gray-violet) | **G-box: cis-acting regulatory element involved in light responsiveness** |
| ■ (tan) | **GARE-motif: gibberellin-responsive element** |
| ■ (olive) | **GC-motif: enhancer-like element involved in anoxic specific inducibility** |
| ■ (orange) | **GT1-motif: light responsive element** |
| ■ (pink) | **HSE: cis-acting element involved in heat stress responsiveness** |
| ■ (light green) | **LTR: cis-acting element involved in low-temperature responsiveness** |
| ■ (lavender) | **MBS: MYB binding site involved in drought-inducibility** |
| ■ (cyan) | **Sp1: light responsive element** |
| □ (white) | **TATC-box: cis-acting element involved in gibberellin-responsiveness** |
| ■ (blue) | **TCA-element: cis-acting element involved in salicylic acid responsiveness** |
| ■ (yellow) | **TCT-motif: part of a light responsive element** |
| ■ (dark red) | **TGA-element: auxin-responsive element** |
| ■ (red) | **TGACG-motif: cis-acting regulatory element involved in the MeJA-responsiveness** |
| ■ (magenta) | **Circadian: cis-acting regulatory element involved in circadian control** |

**Fig. 6** Promoter prediction of *TCS*

## Simple sequence repeats (SSRs) analysis of BAC sequences and BESs

BESs were clustered using CAP3 (Huang and Madan 1999) to remove redundant sequences. Then BESs and complete BAC sequences were scanned for simple sequence repeats by using SciRoKo 3.4 (Kofler et al. 2007) in (perfect: total length) mode (minimum repeats 3, minimum total length 15, and no mismatch SSRs were considered). Motif length was set to 1–6; statistics were in the detailed mode.

A total of 16 cultivated tea plants were used in this study and the genomic DNA was extracted from fresh leaves following the manufacturer's protocol of CP Plant gDNA miniprep kit (Biomiga) with few modifications. Primer pairs were designed from BAC sequences using Primer 3 (Rozen and Skaletsky 1999) (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) software in batch mode. The target fragment size was set as 100–300 bp, the optimal annealing temperature 60 °C. The PCR fragments were detected by capillary electrophoresis with fragment analyzer 96 (Kuehnbaum et al. 2013).

## Gene annotation for BAC sequences and BESs

Fgene SH (http://linux1.softberry.com/) was used for gene prediction. Since there is no existing *Camellia sinensis* gene model, *Populus trichocarpa* was used as a replacement model for gene prediction. Blast2go was used for gene annotation: BLASTX (NCBI-NR database) was used with an E-value threshold of 1E-3; BLAST Top-hit species were clustered, and all functional domains were searched against the InterProScan database; then GO mapping and annotation were conducted.

## Cis-regulatory elements of promoters

We choose promoter sequences 1500 bp upstream of the start codon of *LAR*, *TCS*, and *TS*, and the cis-regulatory elements of three gen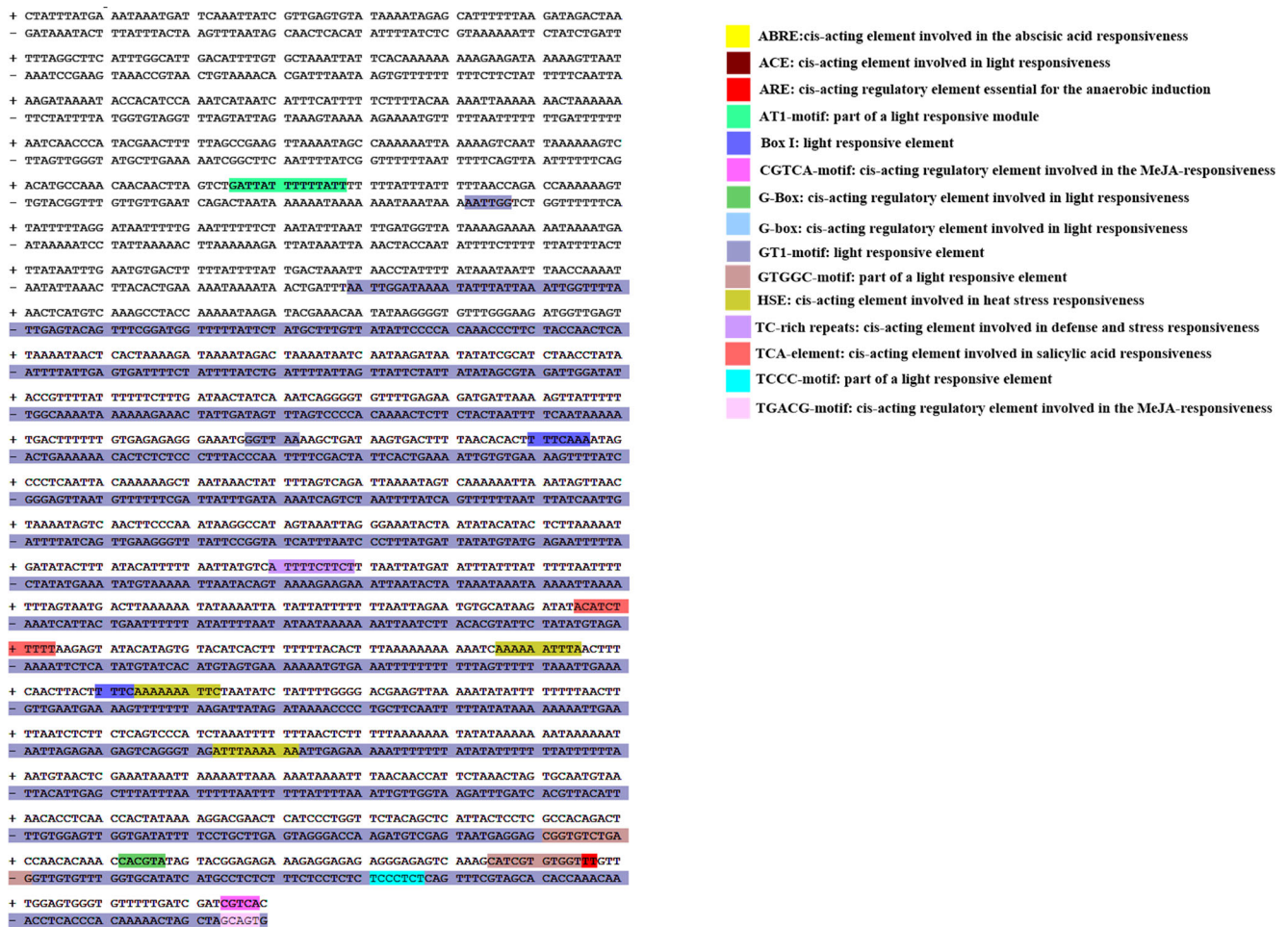es were predicted by promoter analysis software Plant CARE (http://bioinformatics.psb.ugent.be/webtools/plantcare /html/).

**Fig. 7** Promoter prediction of *TS*

## Analysis and validation of alternative splicing (AS)

The clean reads of tea transcriptomes were mapped to *LAR*, *TCS*, and *TS* BAC sequences by Tophat (Trapnell et al. 2009), then AS was analyzed by counting junction sites. The cds of three genes were mapped to the BAC sequences, and the introns of these genes were analyzed. Gene-specific primers were designed with Primer Premier 5 to confirm the predicted splicing events (Supplementary Table S1). PCR amplification was performed by KOD FX Neo (Toyobo), and PCR products were monitored by agarose gel electrophoresis and capillary electrophoresis.

## Results

### Construction of tea BAC library

The BAC library was constructed with *Hind* III in the pIndigoBAC536-s vector (Shi et al. 2011b). Fresh petals from *C. sinensis* "Shuchazao" were used for BAC library construction. Nuclei were washed with PVP40 to enhance DNA cloning ability because *C. sinensis* contains many polyphenols that may influence the construction of a large-insert DNA library. The *C. sinensis* BAC library contains 161,280 clones arrayed in 420 microtiter (384 wells) plates. To determine the insert sizes of the library clones, we analyzed a random sample of 430 clones from the library. The clones were digested with I-*Sce*I, and evaluated on 1% agarose CHEF (CHEF-DRIII apparatus; Bio-Rad, Hercules, CA) gels (Fig. 2a). Result showed that the clones had insert sizes ranging from 45 to 200 kb, with an average insert size of 113 kb (Fig. 2b). Only 22 clones (5.11%) had insert sizes less than 90 kb, with 9 clones having inserts smaller than 70 kb. About 287 clones (67%) had insert fragments larger than 110 kb, some with inserts larger than 140 kb. All 430 clones had inserts, and no empty clone was found. On the basis of the insert sizes of the 430 clones, the average insert size of the BAC library was estimated to be approximately 113 kb; thus, this BAC library would represent 6.2-fold coverage of the *C. sinensis* genome (based on a genome size of 2940 Mb) (Huang et al. 2013). These results indicate that the tea BAC library has a very high quality.
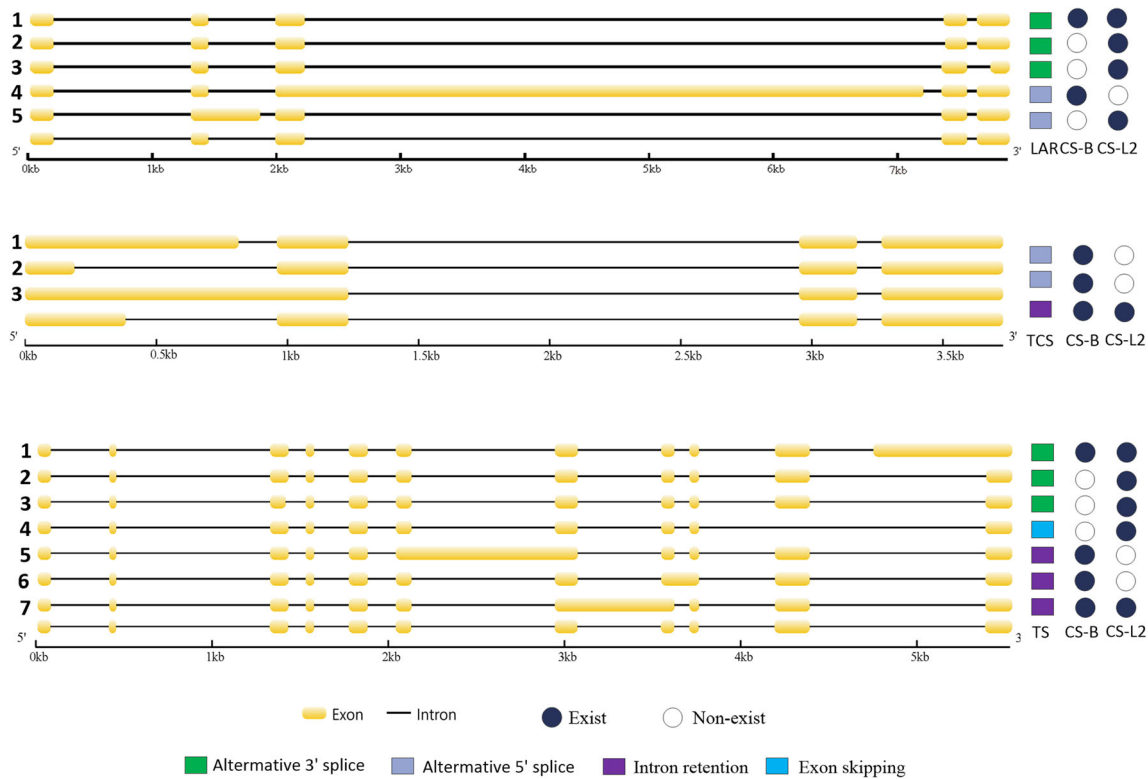
Fig. 8  AS prediction of *LAR TCS* and *TS* by tea transcriptomic reads

## Statistics of BAC sequences and BESs

To further evaluate the quality of the tea BAC library and make a general genome survey for tea, we sequenced 20 complete BAC clones and both ends of 384 BAC clones randomly selected from the BAC library using Sanger sequencing method. PFGE results showed that they all contained the DNA inserts (around 110 kb). Details of the three BAC sequences are presented in Table 1. The GC content is 38.12% and the average BAC insert length is around 113 kb. We successfully obtained 738 raw BES reads, with a success rate of 96.1%. The total length of these BESs was 501,459 bp, and average read length (after trimming the vector sequence) was 679.48 bp (Fig. 3); GC content was 39.55%. Majority of the sequences had lengths between 500 and 800 bp (86.04%).

## Assessment of contamination of mitochondrial and chloroplast DNA

For plant BAC libraries, mitochondrial and especially chloroplast DNA contaminations are usually a serious problem. To assess the contamination rate by mitochondrial and chloroplast DNA in the tea BAC library, all BESs were searched using the BLASTN algorithm against the chloroplast and mitochondrial genome sequence databases. The BLAST results showed no significant matches between BESs and mitochondrial and chloroplast sequences. All BESs were also searched

to check for the bacterial DNA contamination rate and no any significant matches were found either. These results further demonstrated that the tea BAC library has a very high quality and is a valuable resource for genome sequencing, gene cloning and other purposes in tea genome studies.

## Analysis of repeat sequences

Repeat sequences in the 20 BACs and 715 BESs (nonredundancy) were searched using RepeatMasker against the available repetitive sequence databases and found to be 13.44 and 13.42%, respectively. Long terminal repeat (LTR) retrotransposons were the most abundant class of transposable elements, constituting 11.91 and 9.47% of the BAC sequences and BESs, respectively. LTR retrotransposons are among the most abundant constituents of eukaryotic genomes (Havecker et al. 2004). Few lines and sines were found in these sequences (Tables 2 and 3).

## Analysis of SSRs and SSR primer development

SSRs are effective molecular markers in genetic and genomic analysis. Tables 4 and 5 listed all SSRs in the BACs and BESs. In total, 255 SSRs were found in 20 BACs and BESs (901 and 142, respectively). Among 6 SSR motifs, mononucleotide were the most frequent (372), and then were the dinucleotide (366) in 20 BACs. (Supplementary Fig. S1). The average

**Fig. 9** Capillary electrophoresis
(**a**) and agarose gel (**b**) of AS
amplified from *LAR TCS* and *TS*
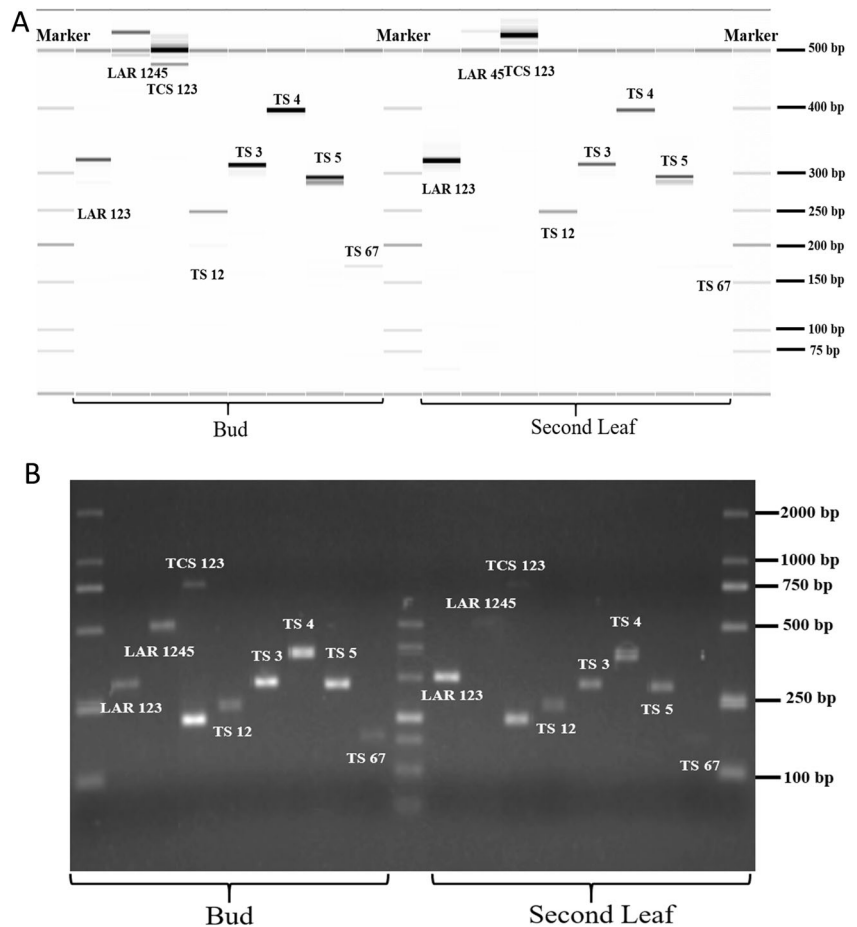


length of SSRs in BACs and BESs were 20.74 and 19.5 bp, respectively. For BESs, 142 (19.9%) BES files contained SSRs of which 112 files contained 1 SSR, 26 files contained 2 SSRs, 3 files contained 3 SSRs and only 1 file contained 4 SSRs (Supplementary Table S2 and S3). Based on this analysis, 14 pairs of SSR primers successfully amplified tea plant DNA (Supplementary Table S4), and most primer pairs amplified one or two fragments, meanwhile some primers amplified more than five fragments such as Csi205A01-1 and Csi205B03. (Fig. 4). (Supplementary Fig. S2). SSR markers have very useful for breeding of tea plants, and these results will be considerable value for the construction of a genetic map for tea plants as well as marker-assisted selection.

## Gene prediction and annotation

Genes in the three BACs (*LAR*, *TCS*, and *TS*) were predicted. These three genes were the key enzyme genes in characteristic metabolic pathways of tea plants. The results showed that there are 30 predicted gene models in each BAC clone. The gene span was estimated to be 3.93 kb per gene (Supplementary Table S5). BLAST top hits analysis using the predicted gene sequences in the BACs (Supplementary Figs. S3a) and the whole BAC end sequences as queries

(Supplementary Fig. S3b) showed an absolute predominance to *Vitis vinifera*. These observations were consistent with the results of previous studies, in which unigenes of tea transcriptomes were reported to have the highest homology to genes of *V. vinifera* (Wu et al. 2014; Zhang et al. 2015). In GO classification, "metabolic and cellular process" was the most common functional sub-category in the category "biology process" for both BESs and BACs. For BESs, the sub-category "catalytic function" exceeded "binding function" under the category "molecular function." For the category "cell component," "cell function" was more common than the others (Supplementary Fig. S4).

## Promoters analysis of *LAR*, *TCS*, and *TS*

Some important elements that make up core promoter such as TATA box and CAAT box were identified in *LAR*, *TCS*, and *TS*. Additionally, a larger number of light-responsive cis-acting elements such as ATCT-motif, Box 4, G-Box, GT1, GTGGC-motif, MRE, TCT-motif, chs-Unit 1 m 1, 3-AF1 binding site, Box I, CATT-motif, TCCC-motif (Figs. 5, 6, and 7) were identified in *LAR, TCS,* and *TS*. According to previous reports, the biosynthesis of catechins (Wang et al. 2012; Jin et al. 2016), caffeine (Chen.

2010; Liu et al. 2013a), and theanine (Matsuura and Kakuda 1990) were greatly influenced by light. Additionally, many regulatory elements involved in binding site of MeJA-responsiveness and ethylene-responsive were also identified. There were also several cis-regulatory elements such as ARE, TC-rich repeats, GC-box, HSE, LTR, MBS, Box-W1, and TCA-element, which served as a putative binding sites for specific TFs in response to biotic and abiotic stresses.

### Alternative splicing (AS) analysis of *LAR*, *TCS*, and *TS*

The process of AS can be spliced in different patterns to produce different mRNA structurally and functionally. AS may be one of mechanisms that accounts for the greater macromolecular and cellular complexity in higher living organisms. Every conceivable pattern of alternative splicing is found in nature, and the main types of alternative splicing including intron retention, alternative 3′ splice sites, alternative 5′ splice sites and exon skipping. We obtained four alternative 5′ splice sites and two alternative 3′ splice sites in LAR gene. Moreover, there are two alternative 5′ splice sites and one intron retention in TCS gene. In addition, more AS types in *TS*, we found three alternative 3′ splice sites, three intron retention and one exon skipping (Fig. 8). To experimentally confirm the accuracy of the identified AS analysis, primers were designed for PCR amplification. The results showed that size of the fragments and the bands on the agarose gel and capillary electrophoresis were mostly consistent with the alternative splicing isoforms (Fig. 9).

### Discussion

Most studies have focused on investigating metabolic genes (Liu et al. 2012; Wang et al. 2014b) and the transcriptome (Jia et al. 2015; Zhang et al. 2015) of *C. sinensis*. In this study, we constructed a high quality, publicly available BAC library for *C. sinensis*. The average insert size of *C. sinensis* BAC clones was 113 kb, representing approximately 6.2-fold overage of the *C. sinensis* genome (2940 Mb) (Huang et al. 2013).

For BAC libraries, contaminations of chloroplast and mitochondrial DNA origins are serious problems and the rates are usually used as important indicators for the quality of plant BAC libraries. We sequenced the ends of 384 random clones and did not find any matches to chloroplast and mitochondrial DNA sequences. The rates are lower than those reported for other plant BAC libraries (Anistoroaei et al. 2011; Cao et al. 2014; Feng et al. 2015; Liu et al. 2013b; Xia et al. 2014). The low rate of chloroplast DNA contamination may be attributable to the use of the petal for BAC library construction. The petals of *C. sinensis* contain few chloroplasts and are thus a good material for BAC library construction. Our sequence data set of BACs and BESs from this library also provided a

glimpse into the sequence composition of the tea genome. Plant genomes usually contain large proportions of repeat sequences. For example, the rice genome contains >35% (Li 2005; Wang et al. 2014a) and the maize genome contains >80% of repeat sequences (Schnable et al. 2009). From the limited tea BAC sequences and BESs, we only found 18.21 and 13.42%, respectively, that match the known repeat sequences. This may mean that the tea genome contains a large amount of unknown or genome-specific repeat sequences. A high quality de novo whole-genome sequence of tea is required to reveal the structure and composition of the genome. The tea BAC library provides an important foundation for whole-genome sequencing assembly and annotation during de novo sequencing of the tea genome.

AS, the process, can generate multiple transcripts from a single coding sequence, moreover, AS of mRNA allows structurally and functionally distinct mRNA and protein variants to be produced. It has recently been proposed as a mechanism by which higher-order diversity is generated. It also provides evolutionary flexibility. We found that there were many regulatory elements involved in binding site of light in *LAR*, *TCS*, and *TS*. This indicated that light can be most important factors involved in regulation of core secondary metabolite genes through regulatory region.

### Conclusions

A high quality BAC library for *C. sinensis* with an average insert size of 113 kb and over 6.2-fold coverage of the genome was constructed. This publicly available resource will provide a useful platform for genetic and genomic research of *C. sinensis*, such as gene identification, functional genomic research, physical mapping, gene isolation and regulation, as well as complex analysis of targeted genomic regions. The AS and promoter analysis of *LAR*, *TCS*, and *TS* will provide important reference for functional genomic research.

**Authors' contributions**  Conceived and designed the experiment: WC and SJ; analyzing data: WL, WH, TL, JY, LZ, YH, and DW; experiment: SL, DZ, SY, HB, WQ, and LM. TL and WH contributed to writing the text.

**Compliance with ethical standards**

**Competing interests**  The authors declare that they have no competing interests.

# References

Allouis S, Qi X, Lindup S, Gale M, Devos K (2001) Construction of a BAC library of pearl millet, *Pennisetum glaucum*. Theor Appl Genet 102:1200–1205

Ammiraju JS, Meizhong L, Goicoechea JL, Wenming W, Dave K, Christopher M et al (2006) 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 Res 16:140–147

Ammiraju JS, Lu F, Abhijit S, Yeisoo Y, Song X, Jiang N et al (2008) Dynamic evolution of Oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. Plant Cell 20: 3191–3209

Anistoroaei R, ten Hallers B, Nefedov M, Christensen K, de Jong P (2011) Construction of an American mink Bacterial Artificial Chromosome (BAC) library and sequencing candidate genes important for the fur industry. BMC Genomics 12:354

Auvichayapat P, Montira P, Oratai T, Narong A (2008) Effectiveness of green tea on weight reduction in obese Thais: a randomized, controlled trial. Physiol Behav 93:486–491

Cao W, Fu B, Wu K, Li N, Zhou Y, Gao Z et al (2014) Construction and characterization of three wheat bacterial artificial chromosome libraries. Int J Mol Sci 15:21896–21912

Cenci A, Chantret N, Kong X, Gu Y, Andersonet OD, Fahima T, Distelfeld A, Dubcovsky J (2003) Construction and characterization of a half million clone BAC library of durum wheat (*Triticum turgidum* ssp. durum). Theor Appl Genet 107:931–939

Chen, P. (2010). Shading effects and influences on photosynthesis and quality component of tea in tea plantation. Hunan Agricultural University

Chen M, SanMiguel P, De Oliveira A, Woo S-S, Zhang H, Wing R, Bennetzen J (1997) Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. Proc Natl Acad Sci 94: 3431–3435

Chen M, Gernot P, Barbazuk B, Goicoechea J, Barbara B, Fang G et al (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

David P, Sévignac M, Thareau V, Catillon Y, Kami J, Gepts P, Thierry L, Vale'rie G (2008) BAC end sequences corresponding to the B4 resistance gene cluster in common bean: a resource for markers and synteny analyses. Mol Gen Genomics 280:521–533

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phredII Error probabilities. Genome Res 8:186–194

Feng C, Bluhm BH, Correll JC (2015) Construction of a spinach bacterial artificial chromosome (BAC) library as a resource for gene identification and marker development. Plant Molecular. Biology Reporter: 1–10

Frary A, Hamilton CM (2001) Efficiency and stability of high molecular weight DNA transformation: an analysis in tomato. Transgenic Res 10:121–132

Gordon D (2003) Viewing and editing assembled sequences using Consed. Current protocols in bioinformatics 11.12(11-11.12):43

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225–225

Hodgson JM, Croft KD, Woodman RJ, Puddey IB, Fuchs D, Draijer R, Lukoshkova E, Head GA (2013) Black tea lowers the rate of blood pressure variation: a randomized controlled trial. Am J Clin Nutr 97: 943–950

Hollman PC, Feskens EJ, Katan MB (1999) Tea flavonols in cardiovascular disease and cancer epidemiology. Proc Soc Exp Biol Med 220: 198–202

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Huang H, Tong Y, Zhang Q-J, Gao L-Z (2013) Genome size variation among and within camellia species by using flow cytometric analysis. PLoS One 8:e64981

Initiative AG (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796

Janda J, Šafář J, Kubaláková M, Jet B, Kovářová P, Suchánková P et al (2006) Advanced resources for plant genomics: a BAC library specific for the short arm of wheat chromosome 1B. Plant J 47:977–986

Jia X, Deng Y, Sun X, Liang L, Ye X (2015) Characterization of the global transcriptome using Illumina sequencing and novel microsatellite marker information in seashore paspalum. Genes Genomics 37:77–86

Jin Q, Chen Z, Sun W, Lin F, Xue Z, Huang Y, Tang X (2016) Cloning and Bioinformatical analysis of anthocyanin synthase gene and its promoter in *Camellia sinensis*. J Tea Sci 36:219–228

Kofler R, Schlötterer C, Lelley T (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23: 1683–1685

Kuehnbaum NL, Kormendi A, Britz-McKibbin P (2013) Multisegment injection-capillary electrophoresis-mass spectrometry: a high-throughput platform for metabolomics with high data fidelity. Anal Chem 85:10664–10669

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

Liang Y, Ma W, Lu J, Wu Y (2001) Comparison of chemical compositions of *Ilex latifolia Thumb* and *Camellia sinensis* L. Food Chem 75:339–343

Lin J, Kudrna D, Wing RA (2011) Construction, characterization, and preliminary BAC-end sequence analysis of a bacterial artificial chromosome library of the tea plant (*Camellia sinensis*) J Biomed Biotechnol 2011:476723

Lin H, Xia P, Wing RA, Zhang Q, Luo M (2012) Dynamic intra-japonica subspecies variation and resource application. Mol Plant 5:218–230

Liu Y, Gao L, Liu L, Yang Q, Lu Z, Nie Z, Wang Y, Xia T (2012) Purification and characterization of a novel galloyltransferase involved in catechin galloylation in the tea plant (*Camellia sinensis*). J Biol Chem 287:44406–44417

Liu J, Yuan D, Si H, Pang X, Tang X, Yang J (2013a) Effects of shading on ingredients of tea shoots in different seasons. Southwest China J Agric Sci 26:115–118

Liu C, Guo Y, Lu T, Wu H, Na R, Li X, Guan W, Ma Y (2013b) Construction and preliminary characterization analysis of Wuzhishan miniature pig bacterial artificial chromosome library with approximately 8-fold genome equivalent coverage. Biomed Res Int

Lu F, Ammiraju JS, Sanyal A, Zhang S, Song R, Chen J, Li G, Sui Y, Song X et al (2009) Comparative sequence analysis of MONOCULM1-orthologous regions in 14 Oryza genomes. Proc Natl Acad Sci 106:2071–2076

Luo M, Wing RA (2003) An improved method for plant BAC library construction. In: Plant functional genomics. Springer, pp 3–19

Ma J, Zhou Y, Ma C, Yao M, Jin J, Wang X, Chen L (2010) Identification and characterization of 74 novel polymorphic EST-SSR markers in the tea plant, *Camellia sinensis* (Theaceae). Am J Bot 97:e153–e156

Mahmood T, Akhtar N, Khan BA (2010) The morphology, characteristics, and medicinal properties of *Camellia sinensis*' tea. J Med Plants Res 4:2028–2033

Mamati GE, Liang Y, Lu J (2006) Expression of basic genes involved in tea polyphenol synthesis in relation to accumulation of catechins and total tea polyphenols. J Sci Food Agric 86:459–464

Matsuura T, Kakuda T (1990) Effects of precursor, temperature, and illumination on Theanine accumulation in tea callus. Agric Biol Chem 37:2033–2051

Messing J, Llaca V (1998) Importance of anchor genomes for any plant genome project. Proc Natl Acad Sci 95:2017–2020

Nilmalgoda SD, Cloutier S, Walichnowski AZ (2003) Construction and characterization of a bacterial artificial chromosome (BAC) library of hexaploid wheat (*Triticum aestivum* L.) and validation of genome coverage using locus-specific primers. Genome 46:870–878

O'Sullivan DM, Ripoll P, Rodgers M, Edwards K (2001) A maize bacterial artificial chromosome (BAC) library from the European flint inbred line F2. Theor Appl Genet 103:425–432

Pan Y, Deng Y, Lin H, Kudrna DA, Wing RA, Li L, Zhang Q, Luo M (2014) Comparative BAC-based physical mapping of Oryza sativa ssp. indica var. 93–11 and evaluation of the two rice reference sequence assemblies. Plant J 77:795-805

Rozen S, Skaletsky H (1999) Primer3 on the WWW for general users and for biologist programmers. Bioinforma Methods Protocol:365–386

Sanmiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 82:37–44

Sasazuki S, Tamakoshi A, Matsuo K, Ito H, Wakai K, Nagata C, Mizoue T, Tanaka K, Tsuji I, Inoue M, Tsugane S (2012) Green tea consumption and gastric cancer risk: an evaluation based on a systematic review of epidemiologic evidence among the Japanese population. Jpn J Clin Oncol 42:335–346

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schulte D, Ariyadasa R, Shi B, Fleury D, Saski C, Atkins M, Pieter D, Wu C, Andreas G, Peter L, Nils S (2011) BAC library resources for map-based cloning and physical map construction in barley (*Hordeum vulgare* L.) Bmc Genomics 12:247

Shi C, Yang H, Wei C, Yu O, Zhang JC, Sun J, Li Y, Chen Q, Xia T, Wan X (2011a) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics 12:131

Shi X, Zeng H, Xue Y, Luo M (2011b) A pair of new BAC and BIBAC vectors that facilitate BAC/BIBAC library construction and intact large genomic DNA insert exchange. Plant Methods 7:33

Shizuya H, Birren B, Kim U-J, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci 89:8794–8797

Song X, Goicoechea J, Ammiraju J, Luo M, He R, Lin J et al (2011) The 19 genomes of Drosophila: a BAC library resource for genus-wide and genome-scale comparative evolutionary research. Genetics 187: 1023–1030

Taniguchi F, Hirai, N. and Yamaguchi, S, (2006) Estimation of the genome size of tea (Camellia sinensis), camellia (C. japonica), and their interspecific hybrids by flow cytometry. Journal of the Remote Sensing Society of Japan (Japan):1-7

Tomkins J, Davis G, Main D, Yim Y, Duru N, Musket T, Goicoechea J, Frisch D, Coe E, Wing R (2002) Construction and characterization of a deep-coverage bacterial artificial chromosome library for maize. Crop Sci 42:928–933

Trapnell C, Pachter L, Salzberg S (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111

Wang Q, Dooner HK (2012) Dynamic evolution of bz orthologous regions in the Andropogoneae and other grasses. Plant J 72:212–221

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

Wang Y, Gao L, Shan Y, Liu Y, Tian Y, Xia T (2012) Influence of shade on flavonoid biosynthesis in tea ( Camellia sinensis (L.) O. Kuntze). Sci Hortic 141:7–16

Wang C, Shi X, Liu L, Li H, Ammiraju J, Kudrna D, Xiong W, Wang H et al (2013a) Genomic resources for gene discovery, functional genome annotation, and evolutionary studies of maize and its close relatives. Genetics 195:723–737

Wang X, Zhao Q, Ma C, Zhang Z, Cao H, Kong Y (2013b) Global transcriptome profiles of *Camellia sinensis* during cold acclimation. BMC Genomics 14:415

Wang X, Liu Q, Wang H, Luo C-X, Wang G, Luo M (2013c) A BAC based physical map and genome survey of the rice false smut fungus *Villosiclava virens*. BMC Genomics 14:883

Wang X, Kudrna D, Pan Y, Wang H, Liu L, Lin H, Zhang J, Song X et al (2014a) Global genomic diversity of *Oryza sativa* varieties revealed by comparative physical mapping. Genetics 196:937–949

Wang Y, Xu Y, Gao L, Yu O, Wang X, He X, Jiang X, Liu Y, Xia T (2014b) Functional analysis of Flavonoid 3′ 5′-hydroxylase from ea plant (*Camellia sinensis*): critical role in the accumulation of catechins. BMC Plant Biol 14:347

Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K et al (2009) The physical and genetic framework of the maize B73 genome. PLoS Genet 5:e1000715

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

Wu C, Nimmakayala P, Santos F, Springman R, Scheuring C, Meksem K, Lightfoot D, Zhang H (2004) Construction and characterization of a soybean bacterial artificial chromosome library and use of multiple complementary libraries for genome physical mapping. Theor Appl Genet 109:1041–1050

Wu H, Chen D, Li J, Yu B, Qiao X, Huang H, He Y (2013) De novo characterization of leaf transcriptome using 454 sequencing and development of EST-SSR markers in tea (*Camellia sinensis*). Plant Mol Biol Report 31:524–538

Wu Z, Li X, Liu Z, Xu Z, Zhuang J (2014) De novo assembly and transcriptome characterization: novel insights into catechins biosynthesis in *Camellia sinensis*. BMC Plant Biol 14:277

Xia Z, Wu H, Watanabe S, Harada K (2014) Construction and targeted retrieval of specific clone from a non-gridded soybean bacterial artificial chromosome library. Anal Biochem 444:38–40

Xia E, Zhang H, Sheng J, Li K, Zhang Q, Kim C, et al (2017) The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. Molecular Plant

Yen G, Chen H (1995) Antioxidant activity of various tea extracts in relation to their antimutagenicity. J Agric Food Chem 43:27–32

Yim Y, Davis G, Duru N, Musket T, Linton E, Messing J, McMullen M, Soderlund C, Polacco M, Gardiner J, Coe E (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

Yu J, Wang J, Lin W, Li S, Li H, Zhou J et al (2005) The genomes of *Oryza sativa*: a history of duplications. PLoS Biol 3:e38

Zhang H, Xia E, Huang H, Jiang J, Liu B, Gao L (2015) De novo transcriptome assembly of the wild relative of tea tree (*Camellia taliensis*) and comparative analysis with tea transcriptome identified putative genes associated with tea quality and stress response. BMC Genomics 16:298

Zhu H, Choi S, Johnston AK, Wing RA, Dean RA (1997) A large-insert (130 kbp) bacterial artificial chromosome library of the rice blast fungus *Magnaporthe grisea*: genome analysis, contig assembly, and gene cloning. Fungal Genet Biol 21:337–347

**Data achieving statement**

The raw data of tea plant BAC are now are available in the NCBI SRA under project accession number PRJNA385558. The experiment accession numbers were SRX2982355. The BAC sequences of tea plant will be submitted to the European Nucleotide Archive database (ENA; http://www.ebi.ac.uk/ena) under project number accession PRJEB21650 if the manuscript is accepted for publication in the tree genetics and genomes prior to publication.

The Illumina RNA-seq data used in this study has been deposited in the NCBI SRA (http://trace.ncbi.nlm.nih.gov/Traces/sra) with accessions SRR1928149.