Beverage Plant
Research

ARTICLE

https://doi.org/10.48130/BPR-2021-0005
Beverage Plant Research **2021**, 1: 5

TeaPGDB: Tea Plant Genome Database

Xiaogang Lei[1], Ya Wang[1], Yuhan Zhou[1], Yongzhong Chen[2], Hongyuan Chen[2], Zhongwei Zou[3], Lin Zhou[4], Yuanchun Ma[1], Fei Chen[1*], and Wanping Fang[1*]

[1] *College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China*
[2] *The High-performance Computing Platform of Bioinformatics Center, Nanjing Agricultural University, Nanjing 210095, China*
[3] *Department of Plant Science, University of Manitoba, Winnipeg, R3T2N2, Canada*
[4] *Forestry and Pomology Research Institute, Shanghai Academy of Agricultural Sciences, Shanghai 201403, China*
* Corresponding author, E-mail: feichen@njau.edu.cn; fangwp@njau.edu.cn

Abstract

As the most widely consumed beverage in the world, tea has various nutritional, economic, and global cultural values. With the development of the third-generation sequencing technology, several genome sequences of tea plants have been published. These genomic data have pivotal information that is of benefit to tea plant breeders and biologists in advancing tea plant improvement and the final quality of tea products. We hereby present the integrative online database, Tea Plant Genome Database (TeaPGDB; http://eplant.njau.edu.cn/tea), which incorporates the published genome sequences of tea plants. The current release of TeaPGDB hosts published tea plant genome data with various online tools, including JBrowse, gene search, SSR search, BLAST. TeaPGDB also contains a download server, which provides access for the download of genome-related data and rich annotation files. TeaPGDB is committed to collecting, integrating, and annotating published tea plant genome data, providing data support for research on tea plant heredity, evolution, breeding for resistance, plant improvement, and facilitating the characterization of important traits or flavor related genes in the community. Compared with other tea plant databases, this database not only contains more complete genome data and gene annotation information, but also has a user-friendly interface for researchers in the field.

Citation: Lei X, Wang Y, Zhou Y, Chen Y, Chen H, et al. 2021. TeaPGDB: Tea Plant Genome Database. *Beverage Plant Research* 1: 5 https://doi.org/10.48130/BPR-2021-0005

INTRODUCTION

Tea is one of the three most popular non-alcoholic beverages in the world, with important economic, health and cultural value[1]. Tea plants not only have both a long history and wide range of cultivation. So far, tea plants have been planted in more than 60 countries of the world[2]. In the past few decades, theanine, caffeine, tea polyphenols, mineral elements and other substances that contribute to human health and tea quality, have been focused on and studied by tea researchers[3−7]. Before the release of the *Camellia sinensis* var. *sinensis* (CSS) 'Shuchazao' genome, studies on the molecular genetics and breeding of tea plants progressed slowly due to the lack of genomic datasets[1]. The self-incompatibility of tea plants are another reason why tea plant breeding and genetics progressed slowly[8]. Different from other crops, tea plants are self-incompatible species, and the rate of hybrid breeding is relatively low. These factors lead to a lack of high-generation segregating populations and a lack of sufficient offspring, which is not conducive to the construction of genetic maps. Genetic mapping is the basis of molecular biology and is essential for the study of genetics and genomics, such as quantitative trait mapping, molecular marker-assisted breeding and comparative genome research[9,10]. In addition, the basic biological characteristics has been narrowed by limited knowledge of tea plant phylogenetic biology and functional genomics[8]. In general, there are lots of research factors that hinder tea plants genetics and breeding, but almost all of them are related to tea plant genome research.

The rapid development of plant genomics has accelerated and advanced molecular biological characterization of important factors in plant vegetative and reproductive growth, tolerance to stress, and secondary metabolites, which benefits horticultural crop genetics, evolution, and improvement. In particular, over the past decade, with the development and improvement of third-generation sequencing and assembly technologies, several important horticultural crop genomes were sequenced and resequenced such as *Isatis indigotic*[11], willow (*Salix suchowensis*)[12], asparagus fern (*Asparagus setaceus*)[13], wild hemp (*Cannabis sativa*)[14], *Cladopus chinensis*[15], sweet cherry (*Prunus avium* L.)[16], *Kandelia obovata*[17], star fruit (*Averrhoa carambola* L.)[18], and apple (*Malus domestica*)[19], generating a large amount of important genomic data. To facilitate the use and mining of genomic data, corresponding genomic databases have been established such as *citrus*[20] strawberry (*Fragaria* × *ananassa*)[21], kiwifruit (*Actinidia* spp.)[22], tomato (*Lycopersicon esculentum Miller*)[23], pineapple (*Ananas comosus* L.)[24], and cabbage (*Brassica oleracea* var. *capitata* L.)[25]. Because of the large genome size, high heterozygosity, and complexity of tea plants, the deciphering of the tea plant genome has become a significant barrier in tea science research. Similar to the genome sequencing of other horticultural crops, eight genomes of six tea plant species have been quickly deciphered, among which *C. sinensis* var. *sinensis* 'Shuchazao' has been sequenced and assembled three times[1,26,27]. Since the publication of the genome of the *C. sinensis* var. *assamica* (CSA) 'Yunkang 10' in 2017, the genomes of *C. sinensis* var. *sinensis* (CSS) 'Biyun', CSS 'Longjing 43', the wild tea plant DASZ, CSS 'Shuchazao' and 'Huangdan' have been published successively[1,26−32]. Although these genomes have greatly promoted tea science research, it is difficult to obtain detailed information such as the sequence, chromosome location, function, and gene annotation when people look at particular functional genes related to fertility, in response to stress, resistance, or aroma formation. In addition, users need to introduce many external tools to perform gene or protein sequence alignments and query detailed information of genes. The whole process is very cumbersome, complicated and time-consuming. Although tea plant has published eight genomes in different cultivars, it is still difficult to select the proper genome sequence as a reference in research activities since the assembly quality of each genome is hard to evaluate. In attempts to understand the unknown functions of thousands of proteins and coding sequences in the individual genome, we need to perform alignments using several different databases to make inferences[33]. In this study, we developed a tea plant genome database (TeaPGDB), which integrates all the currently published tea plant genomes. The new integrated database fully annotated the tea plant genome (organellar genome contained), including the gene localization on the chromosome, gene families identification, Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation, Gene Ontology (GO) annotation, signal peptide prediction, etc. We also provide several useful online tools, such as gene search, BLAST, JBrowse, etc. By using a gene search engine, tea researchers can easily figure out the chromosomal location of the target gene, gene family identification, GO annotations and KEGG annotations and other detailed information. In addition, users can also use online tools to align gene or protein sequences with published tea plant genomes, where we can view the relevant information of the gene in JBrowse based on the chromosomes. Unknown gene sequences are aligned and searched in published data to find homologous genes, which is important to advance the molecular and biological studies in tea plants.

RESULTS

Species and tea flavor gene pages

TeaPGDB has several unique pages including species and tea flavor gene pages. We introduce and collect data on tea species with released genomes. These pages contain the results of BUSCO analysis of all tea plant genomes, including genes involved in the synthesis of tea plant quality components (catechins, caffeine, theanine). TeaPGDB facilitates tea researchers in understanding the current status of tea plant genome research and determine the most suitable reference genome. TeaPGDB can also promote the study of tea quality components.

Data

The TeaPGDB database contains all released genome data, and all gene and protein sequence data annotated. Compared with TPIA or other databases, the annotated genomes in this database are more comprehensive and informative. These annotation data can benefit research in tea biochemical composition, resistance, and breeding, which is of great significance for promoting tea plant performance and improvement. Compared with TPIA, TeaPGDB contains more complete SSR data, and it has SSR data for all released genomes.

Online retrieval and analysis tools

TeaPGDB provides several online retrieval and analysis tools to facilitate data analysis and information retrieval. Gene and SSR search is an information retrieval tool. Users can obtain detailed gene annotation information and SSR information by entering the correct number. JBrowse is a genome visualization tool that can display genome sequence information and gene annotation information to users, and users can also upload data for customization. BLAST online analysis tool provides an online sequence comparison function, you can select tea plant data or other species data in the database for sequence comparison to find homologous sequences. Compared with TPIA, this database has a more advanced blast online analysis tool, and its reference database contains the protein, genome, and coding sequence files of the released genomes.

Download and Community modules

Download and Community modules are convenient for users to download original and annotated data, and provide user feedback to facilitate developers in order to optimize the database. It also provides access to tea-related and genome-related databases, which is conducive to the interaction and collaboration between databases.

MATERIALS AND METHODS

Dataset

We obtained the project number of the genomic data in National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov), National Genomics Data Centre (NGDC, https://bigd.big.ac.cn), TPIA, Github or figshare (https://figshare.com) where the genomic data is located from the published genomic article, and then downloaded the published genomic data from the corresponding website. Genome sequences of CSA 'Yunkang 10', CSS 'Shuchazao' (AHAU released it), CSS 'Biyun', CSS 'Shuchazao' (TRI released it), wild tea plant (DASZ), CSS 'Huangdan' were derived from TPIA (http://tpia.teaplant.org/download.html), TPIA (http://tpia.teaplant.org/download.html), NGDC (https://ngdc.cncb.ac.cn/gwh/Assembly/8796/show),Github (https://github.com/JiedanChen/TeaGenomeData), figshare (https://figshare.com/articles/journal_contribution/Assembly_and_annotation_of_DASZ_genome/12560462/1), NGDC (https://bigd.big.ac.cn/bioproject/browse/PRJCA003382), respectively[1,26−30,32]. We will annotate, upload and visualize these data using the particular online tools such as gene search, download and JBrowse etc. (Table 1).

We collected 17 groups of published tea plant organellar genomes from NCBI. These genomic data can be obtained through the GenBank project. The chloroplast genome of *C. sinensis* var. *assamica*[34], *C. sinensis* 'Anhua'[35], *C. sinensis* var. *pubilimba*[36], *C. sinensis* var. *sinensis* 'Tieguanyin'[37], *C. sinensis*

Table 1. The dataset integrated in TeaPGDB.

Data type	Counts	Data size
Tea plant cultivars	6	–
Nuclear genomes[1]	7	21.5 Gb
Organellar genomes[2]	17	3.8 Mb
Coding sequences	271,430	352 Mb
Proteins	271,430	150 Mb
Gene family identification	4,460	25 M
GO term	71,996	59 M
KEGG	90,893	1.2 G
Signal peptide	19,156	13 M
SSR	2,363,168	152 M

[1] *C. sinensis* var. *sinensis* 'Shuchazao' has two nuclear genome sequences.
[2] *C. sinensis* var. *assamica* 'Yunkang 10' has two mitochondrion genome sequences.

'Longjing 43'[38], *C. sinensis* 'Wuyi Narcissus'[39], *C. sinensis* 'Dahongpao'[40], *C. sinensis* 'Baijiguan', *C. sinensis* 'Rougui', *C. sinensis* 'Shuijingui', *C. sinensis* var. *sinensis* 'Baiye 1'[41], *C. sinensis* dehungensis[36], *C. sinensis* 'sangmok'[42], *C. sinensis* 'Tieluohan'. The mitochondrion genome of *C. sinensis assamica* mitochondrion genome[43], *C. sinensis* var. *assamica* 'Yunkang 10' chromosome 1[44], *C. sinensis* var. *assamica* 'Yunkang 10' chromosome 2[44], *C. sinensis* var. *assamica* 'Yunkang 10' plastid genome[45], are respectively available in the NCBI database under project numbers MH394407.1, MH042531.1, KJ806280.1, MW148820.1, KF562708, MT612435.1, MT773374.1, MT773373.1, MT773375.1, MT773376.1, MN086819.1, KJ806279.1, LC488797.1, MT773377.1, NC_043914.1, MK574876.1, MK574877.1, MH019307.1.

Sequence annotation tools

InterProScan (version 5.1.2)[46] is used to analyze the protein domain functions of tea plant genes. Gene ontology annotation was conducted by searching against the databases including PANTHER,SMART, PRINTS, Gene3D and Pfam etc[47]. Hmmer (version 3.1b2)[48] is used to identify gene families using of the hmmscan program in hmmer against the Pfam database. Program Kofam_scan (version 1.2.0) in kofamkoala software is employed to annotate proteins[49]. Signalp (version 5.0)[50] is adopted to predict the signal peptide of the protein sequences. In this study, all the programs were carried out in local mode and with full default parameters. In addition, we use R (version 3.6.1)[51] language to draw gene structure diagrams, including exons, mRNA 3′ UTR and 5′UTR etc. The gene structure diagram mainly illustrates the two-dimensional position relationship of genes and the localization on the chromosome. The blastn program in the ncbi-blast-2.11.0+ compares the tea plant genome sequences with the nr database to obtain the homologous sequences in the nr database.

Database development pipeline

We implemented TeaPGDB using four major tools, those include Apache httpd (version 2.2)[52], HTML5[53], PHP (version 7.4)[54] and MySQL (version 8.0.23)[55] (Fig. 1). The relevant data sets are restored in Linux platform with MySQL database. We use html to build the basic framework of the webpage, and connect the database and the webpage through apache, php and mysql, and then realize the process of users querying gene-related information. TeaPGDB is an interactive and user-friendly website. Its landing page is generated by HTML, CSS[56], JavaScript[57], jQuery[58], Bootstrap etc (Fig. 1b). The gene and SSR search interface are composed of PHP, MySQL, Apache and web building tools. The sequenceserver (version1.0.14)[59] page is designed based on blast+ (version 2.9.0+)[60]. We also implemented JBrowse[61] as a genome browser for gene model visualization. MySQL, HTML and PHP were used to build tea plant flavor gene pages, and BUSCO software (version 5.1.2)[62] was introduced to evaluate the assembly quality of tea plant genome and protein files. Finally, incorporating HTML, CSS and JavaScript, the download page provides users with access and data download capability for their use. (Fig. 1c).

UNIONIZATION OF THE TEAPGDB

In 2021, TeaPGDB contains the genomes of six published tea plant species, generating a total of seven sets of genome data. Among them, CSS 'Shuchazao' contains two genomes with different assembly levels. We have collected and sorted the most important genes related to the quality of tea plants such as catechins, theanine, and caffeine. Each tea plant species contains the relevant genes in tabular format, which is designed for investigation of the related gene functions involved in the determination of tea quality. Compared with other tea plant databases such as Tea Plant Information Archive (TPIA)[63], Tea Metabolome Database (TMDB)[64] and gene co-expression network database (TeaCoN)[65], TeaPGDB includes more complete genome data and gene annotation content. TeaPGDB also has a species page, which mainly highlights the tea plant cultivars of released genomes (organellar genomes contained), i.e., genome general information, and genome assembly quality assessment. It also contains online analysis and search tools such as BLAST, JBrowse, gene search and SSR search, etc. Therefore, we have the ability to analyze and retrieve the required information easily through designated platforms. TeaPGDB contains a download page and a Community module. Users can download genomic data and detailed annotation data. Under the Community module, users can not only understand the research progress of the tea plant genome, but also conveniently access other relevant databases. Through the server, users can also give their feedback and suggestions to potentially further improve the website.

The portal's homepage

A clear and concise TeaPGDB website homepage has been built (Fig. 2). Currently, the website contains five main parts: a navigation bar (Fig. 2a), species gallery, toolkit, a brief introduction, and other panels. The left panel of the navigation is the species gallery (Fig. 2b) and the right are the toolkits (Fig. 2c). The central part is a brief induction of TeaPGDB (Fig. 2d). In order to meet the needs of scientists with different expertise, we briefly describe recent progress in the field of tea plant and tea plant genome research. It mainly introduces tea plant-related information through the middle module of the homepage, including tea plant research, botanical classification, tea plant genomes and the goal of the establishment of the website. The bottom of the homepage is used for other related information, news, citations (Fig. 2e) and a small plugin showing global visitors (Fig. 2f).

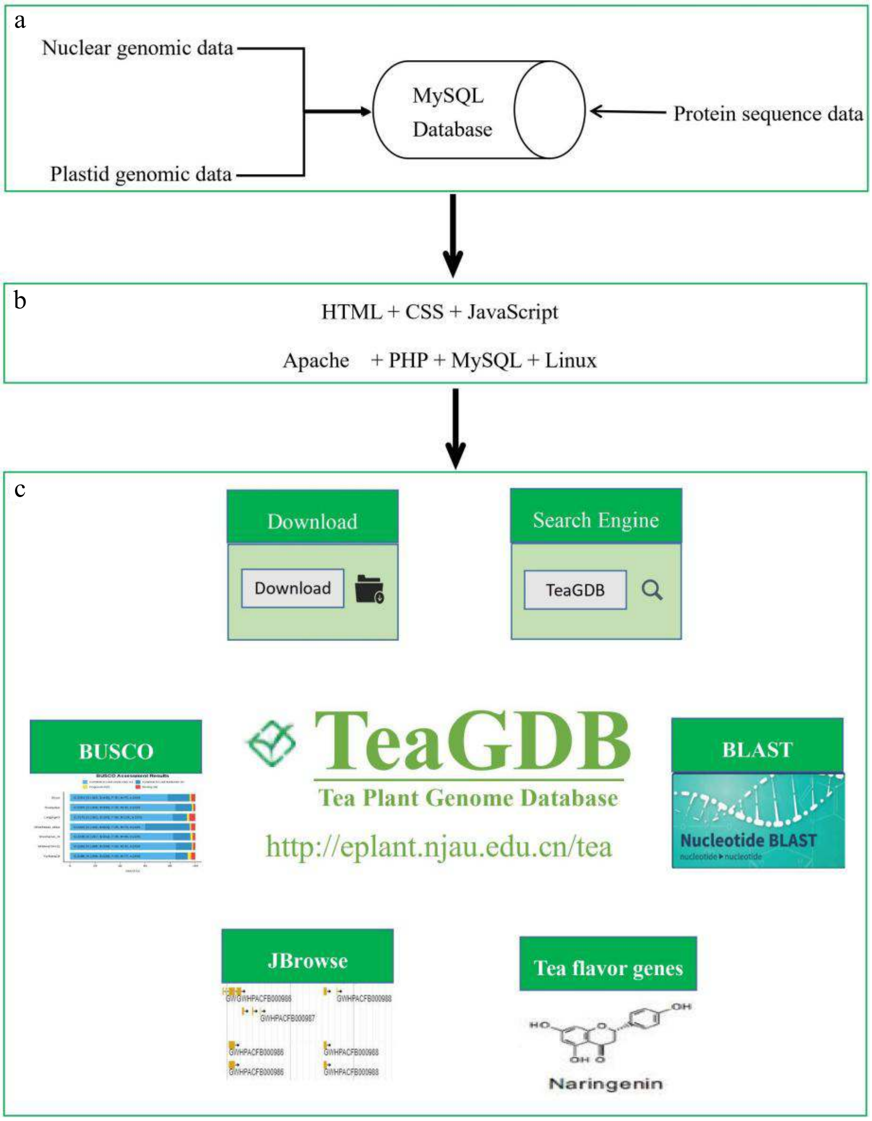


Fig. 1 Overview of the TeaPGDB server. (a) Genomic data, (b) middleware on the server, (c) application layer for public users.

Species page

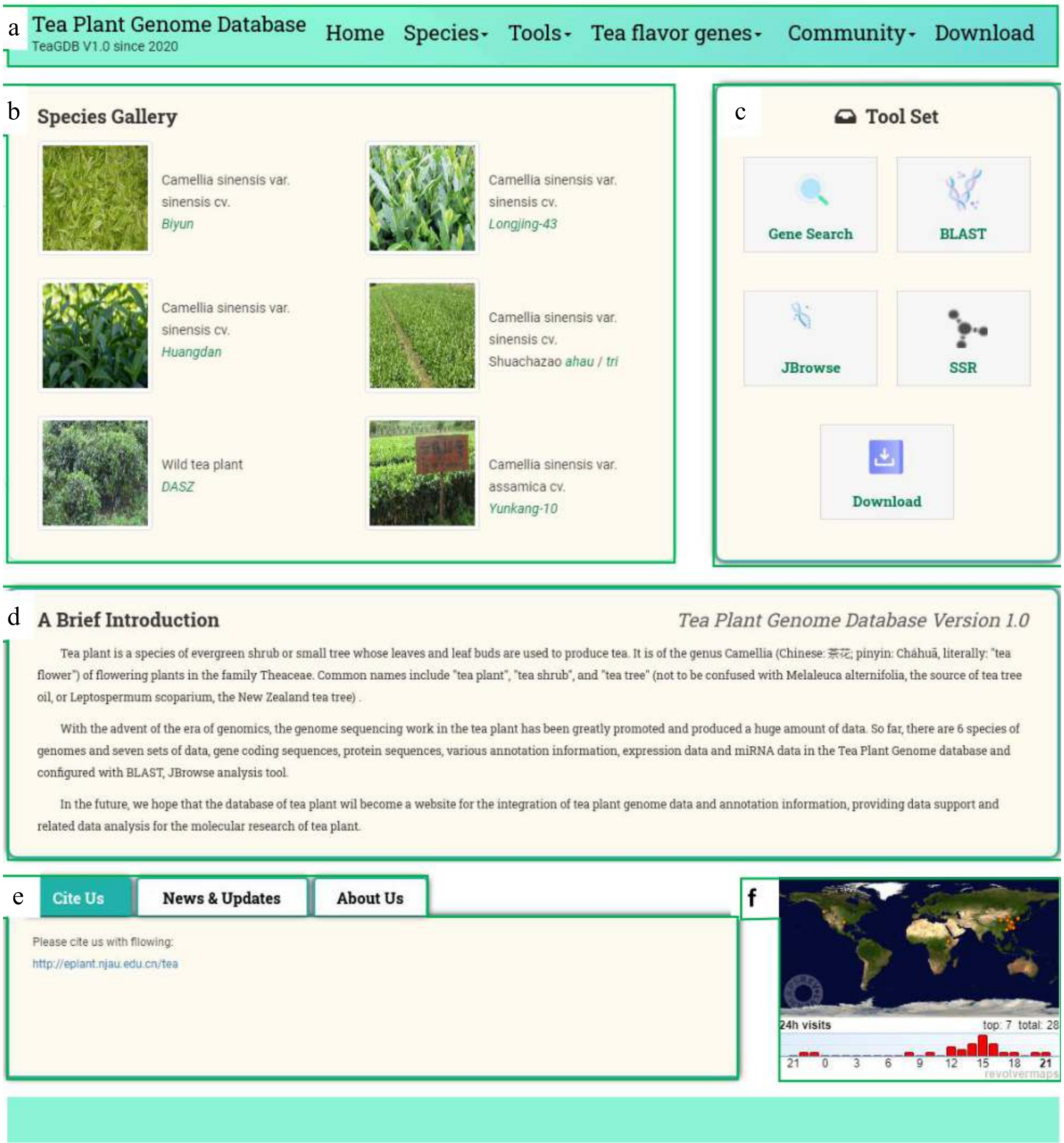
The introduction pages of tea plant varieties are used to give a brief description of each tea plant where the genome has been released. These pages mainly include introduction to tea plant varieties and the analysis of the tea plant genome. BUSCO (Version 5.1.2)[62] is used to evaluate the assembly quality of the genome and proteome, and finally the results are displayed in a tablealongside the other six tea plant genomes.

Tea flavor genes

The tea flavor genes page includes genes related to the biosynthesis of catechins, caffeine and theanine (Fig. 3). The genes of CSS 'Shuchazao' involved in quality components and their synthesis pathway have been identified in previous studies[1,31], while the related genes from several other varieties are aligned by sequenceserver. The known genes are compared with the genome databases of other unknown tea plants, using parameters --evalue = 10^{-5}. Thus, the relevant genes could be underlined, and users can then check the molecular and biological function of these genes in TeaPGDB. After confirming that all the gene functions are consistence with the involved biological processes, these genes are written and summarized in the tea flavor genes' table. The tea flavor genes are categorized into three groups according to their function in certain compound biosynthesis (Fig. 3a). The gene in each pathway can be linked to the gene annotation result page, which is convenient for researchers to view and derive detailed information of the gene (Fig. 3b).

Gene search

The gene search interface is connected to the MySQL database through PHP, Apache and HTML to realize the search function of gene annotations. When the user inputs the ID of the gene, a complete set of annotation results for the gene can be displayed via the website (Fig. 4). For each gene analysed, the user can select a different tea plant species through the drop-down menu, and the sample ID of the gene will appear in the search box. After clearing the gene ID through the 'clean all' button, the user can input the


a
Tea Plant Genome Database
TeaGDB V1.0 since 2020
Home
Species
Tools
Tea flavor genes
Community
Download
b
Species Gallery
Camellia sinensis var. sinensis cv. Biyun
Camellia sinensis var. sinensis cv. Longjing-43
Camellia sinensis var. sinensis cv. Huangdan
Camellia sinensis var. sinensis cv. Shuachazao ahau / tri
Wild tea plant DASZ
Camellia sinensis var. assamica cv. Yunkang-10
c
Tool Set
Gene Search
BLAST
JBrowse
SSR
Download
d
e
Cite Us
News & Updates
About Us
f
24h visits
top: 7 total: 28
21 0 3 6 9 12 15 18 21
revolvermaps

A Brief Introduction

Tea Plant Genome Database Version 1.0

Tea plant is a species of evergreen shrub or small tree whose leaves and leaf buds are used to produce tea. It is of the genus Camellia (Chinese: 茶花; pinyin: Cháhuā, literally: "tea flower") of flowering plants in the family Theaceae. Common names include "tea plant", "tea shrub", and "tea tree" (not to be confused with Melaleuca alternifolia, the source of tea tree oil, or Leptospermum scoparium, the New Zealand tea tree) .

With the advent of the era of genomics, the genome sequencing work in the tea plant has been greatly promoted and produced a huge amount of data. So far, there are 6 species of genomes and seven sets of data, gene coding sequences, protein sequences, various annotation information, expression data and miRNA data in the Tea Plant Genome database and configured with BLAST, JBrowse analysis tool.

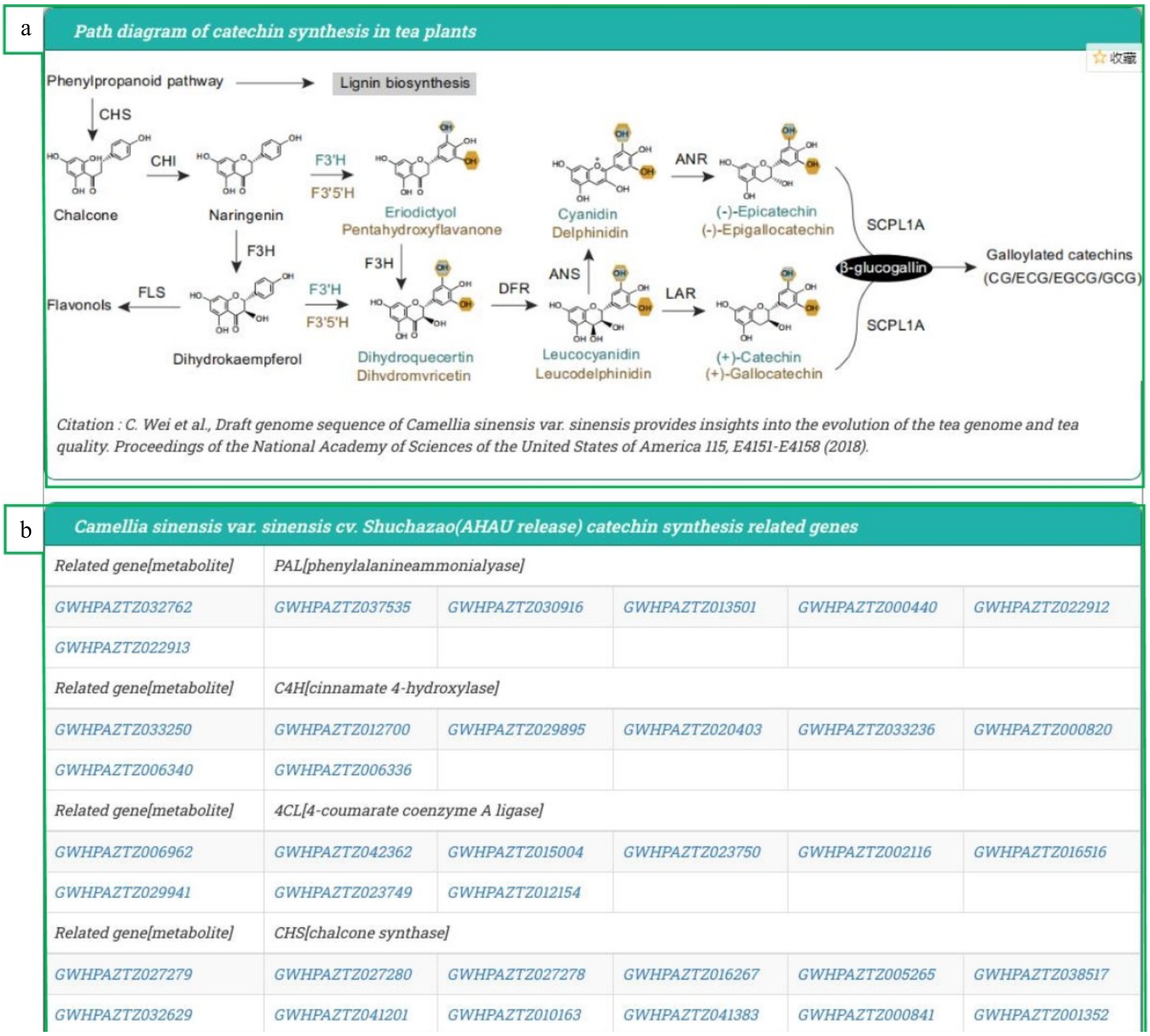
In the future, we hope that the database of tea plant wil become a website for the integration of tea plant genome data and annotation information, providing data support and related data analysis for the molecular research of tea plant.

Please cite us with flowing:
http://eplant.njau.edu.cn/tea

Fig. 2 Homepage of Tea Plant Genome Database (TeaPGDB). The homepage consists of five main parts: (a) species gallery, (b) carousel diagram, (c) toolkit, (d) a brief introduction, (e) news, citation and introduction information, (f) user access map.

correct gene ID based on the format of the sample ID (Fig. 4a).

The results page of the search mainly includes Data Related Information, DNA Coding & Protein Sequence, Gene Identification, KEGG Orthology, Gene Family Identification, Gene Attributes, Gene Ontology Classification, Signal Peptide, PFAM hits, Interpro hits, Gene Structure and Blastn Searches (Fig. 4b). Among them, DNA Coding & Protein Sequence will show the option to download the sequence file. Data Related Information includes the genome source and research institutions. Gene Identification explains gene products and gene names. In addition, users can find out the possible functions and products of this gene based on gene identification. KEGG Orthology shows KEGG annotation results that are mainly displayed by the ko number linked to the KEGG website. Through the ko number, users can figure out the metabolic pathways in which the gene inquired is involved. Gene family identification is predicted by the conserved domains of proteins and it shows the gene family to which the gene belongs. Gene Attributes indicate the position and total length of the chromosome where the gene is located, and the length of the predicted protein. Gene Ontology is visualized in tabular form, mainly including inferred functions, alignment databases, alignment positions and GO numbers. Through the GO annotation results, the function of the gene and its involvement in biological processes could be shown. The Signalp Peptide result navigates the cleavage site of the protein sequence and the presence or absence of a signal peptide. PFAM hits contain the position and annotation results of the sequence compared with the database. Interproscan is used to annotate protein sequence files to get Interpro hits including predicted products, the start and end positions of the sequence aligned with the databases and annotations. The gene structure diagram mainly contains the position of the



Citation : C. Wei et al., Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. Proceedings of the National Academy of Sciences of the United States of America 115, E4151-E4158 (2018).

Camellia sinensis var. sinensis cv. Shuchazao(AHAU release) catechin synthesis related genes

Related gene[metabolite]	*PAL[phenylalanineammonialyase]*				
GWHPAZTZ032762	*GWHPAZTZ037535*	*GWHPAZTZ030916*	*GWHPAZTZ013501*	*GWHPAZTZ000440*	*GWHPAZTZ022912*
GWHPAZTZ022913					
Related gene[metabolite]	*C4H[cinnamate 4-hydroxylase]*				
GWHPAZTZ033250	*GWHPAZTZ012700*	*GWHPAZTZ029895*	*GWHPAZTZ020403*	*GWHPAZTZ033236*	*GWHPAZTZ000820*
GWHPAZTZ006340	*GWHPAZTZ006336*				
Related gene[metabolite]	*4CL[4-coumarate coenzyme A ligase]*				
GWHPAZTZ006962	*GWHPAZTZ042362*	*GWHPAZTZ015004*	*GWHPAZTZ023750*	*GWHPAZTZ002116*	*GWHPAZTZ016516*
GWHPAZTZ029941	*GWHPAZTZ023749*	*GWHPAZTZ012154*			
Related gene[metabolite]	*CHS[chalcone synthase]*				
GWHPAZTZ027279	*GWHPAZTZ027280*	*GWHPAZTZ027278*	*GWHPAZTZ016267*	*GWHPAZTZ005265*	*GWHPAZTZ038517*
GWHPAZTZ032629	*GWHPAZTZ041201*	*GWHPAZTZ010163*	*GWHPAZTZ041383*	*GWHPAZTZ000841*	*GWHPAZTZ001352*

Fig. 3 The pathways and genes responsible for tea flavor in TeaPGDB. Tea plant quality components catechins, caffeine and theanine synthesis genes. (a) Compound synthesis route, (b) related genes.

gene and the relative position of mRNA, exon, 3'URT, 5'UTR. Blastn Searches contain the identity, sequence length of the alignment and E-value. If a new tea plant genome is published, we will continue to incorporate and annotate it to enrich the content of the gene search page.

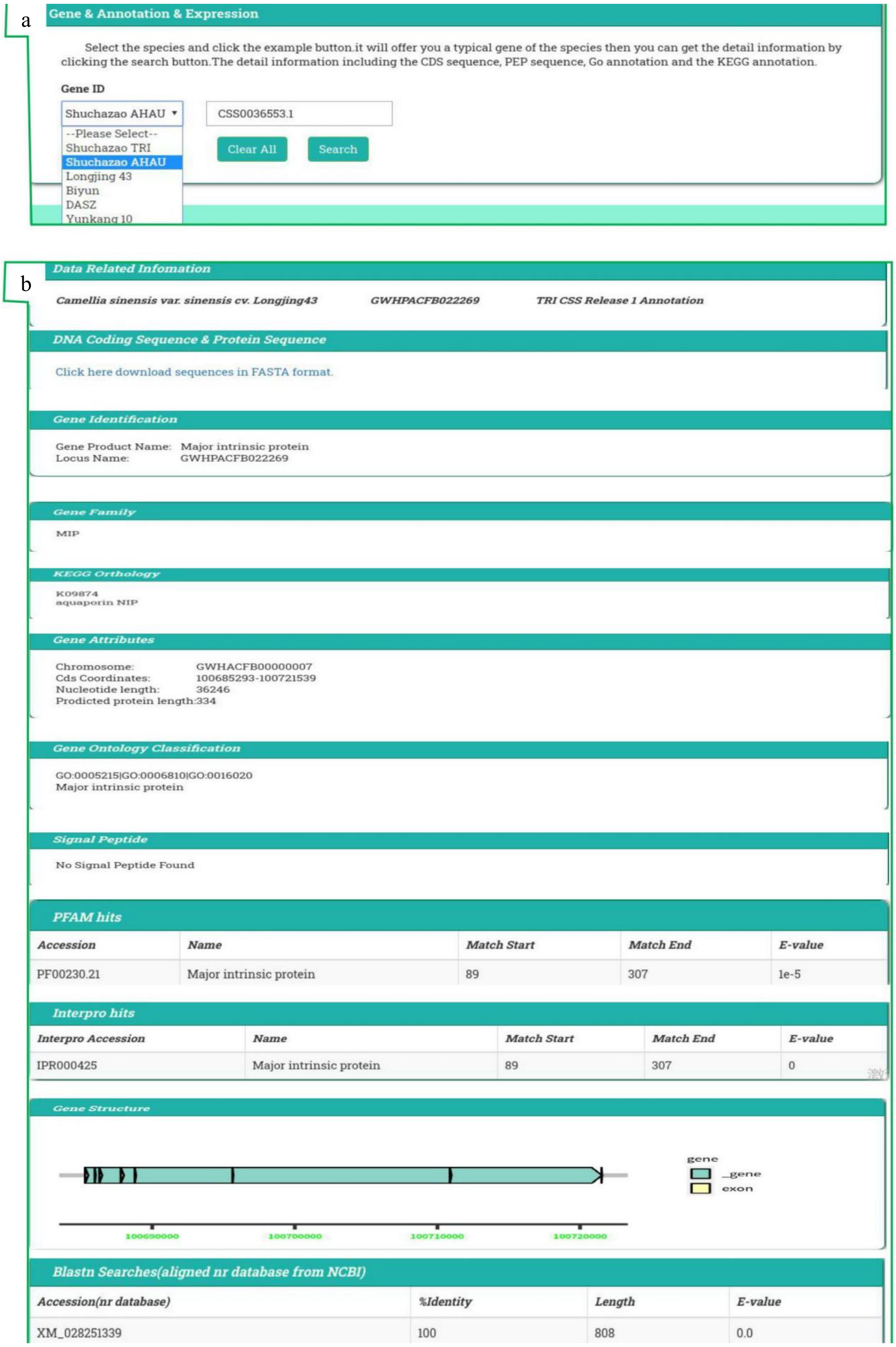
Genomic views through JBrowse

One of the key objectives of TeaPGDB is to collect, organize and annotate genomic data, to enable researchers in the field globally to obtain an intuitive understanding of tea plant genomic data and the functions, structure, composition and other information contained in the genes. In order to reach this goal, we established JBrowse, an interactive genome browser, which is widely used to navigate large-scale high-throughput sequencing data under the framework of the genome[66]. JBrowse is a highly flexible and customized genome browser (Fig. 5). Users can upload their own sequence data sets for visualization and comparison with the data sets in TeaPGDB[63].

Currently JBrowse is divided into two parts, the first part is the nuclear genome, and the second is the Organellar genome (Fig. 5a). The Organellar genome data contains mitochondria and chloroplast genomes, which are downloaded from NCBI. The JBrowse of the tea plant genome contains annotation information such as genome, CDS, SSR etc (Fig. 5b). Users could get detailed information, such as location, length, type and id, by clicking the bar on JBrowse (Fig. 5c). We have capacity to add more novel and publicly available data and analysis into JBrowse. Moreover, if the tea plant genome is continuously published, we will update and enrich the content of JBrowse.

BLAST tool

The webpage online Blast tool performs online comparison and analysis of gene sequences or protein sequences through Sequencesever (version 1.0.14) (Fig. 6). Users could use the BLAST online tool for sequence alignment by inputting a gene sequence or importing a sequence file in the input box



Gene & Annotation & Expression

Select the species and click the example button.it will offer you a typical gene of the species then you can get the detail information by clicking the search button.The detail information including the CDS sequence, PEP sequence, Go annotation and the KEGG annotation.

Gene ID

Shuchazao AHAU ▾ CSS0036553.1

Data Related Infomation

Camellia sinensis var. sinensis cv. Longjing43 *GWHPACFB022269* *TRI CSS Release 1 Annotation*

DNA Coding Sequence & Protein Sequence

Click here download sequences in FASTA format.

Gene Identification

Gene Product Name: Major intrinsic protein
Locus Name: GWHPACFB022269

Gene Family

MIP

KEGG Orthology

K09874
aquaporin NIP

Gene Attributes

Chromosome: GWHACFB00000007
Cds Coordinates: 100685293-100721539
Nucleotide length: 36246
Prodicted protein length:334

Gene Ontology Classification

GO:0005215|GO:0006810|GO:0016020
Major intrinsic protein

Signal Peptide

No Signal Peptide Found

PFAM hits

Accession	*Name*	*Match Start*	*Match End*	*E-value*
PF00230.21	Major intrinsic protein	89	307	1e-5

Interpro hits

Interpro Accession	*Name*	*Match Start*	*Match End*	*E-value*
IPR000425	Major intrinsic protein	89	307	0

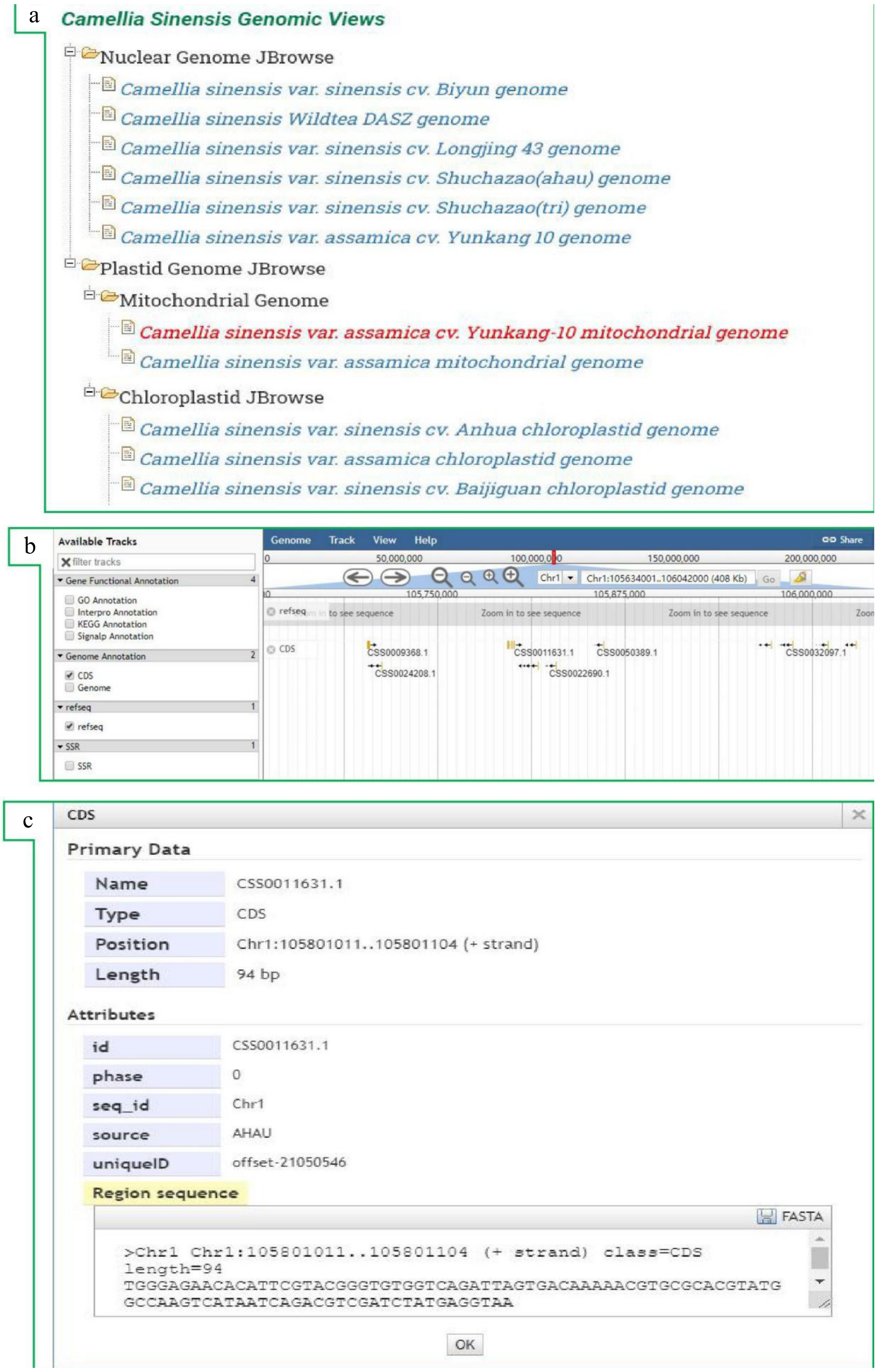
Gene Structure

Blastn Searches(aligned nr database from NCBI)

Accession(nr database)	*%Identity*	*Length*	*E-value*
XM_028251339	100	808	0.0

Fig. 4 An example of gene search and the returned results. Using Gene ID to search for gene annotation results. (a) Search box and drop-down options, (b) search results page containing detailed annotation information.

(Fig. 6a), and selecting the databases for comparison (Fig. 6c). These databases are all established based on published tea plant genomes, cds and protein sequences. They also include cds and protein sequence files from other plants such as strawberry, grape and Arabidopsis as optional databases (Fig. 6b). The option box can be used to enter parameters,



Camellia Sinensis Genomic Views

- Nuclear Genome JBrowse
  - *Camellia sinensis var. sinensis cv. Biyun genome*
  - *Camellia sinensis Wildtea DASZ genome*
  - *Camellia sinensis var. sinensis cv. Longjing 43 genome*
  - *Camellia sinensis var. sinensis cv. Shuchazao(ahau) genome*
  - *Camellia sinensis var. sinensis cv. Shuchazao(tri) genome*
  - *Camellia sinensis var. assamica cv. Yunkang 10 genome*
- Plastid Genome JBrowse
  - Mitochondrial Genome
    - *Camellia sinensis var. assamica cv. Yunkang-10 mitochondrial genome*
    - *Camellia sinensis var. assamica mitochondrial genome*
  - Chloroplastid JBrowse
    - *Camellia sinensis var. sinensis cv. Anhua chloroplastid genome*
    - *Camellia sinensis var. assamica chloroplastid genome*
    - *Camellia sinensis var. sinensis cv. Baijiguan chloroplastid genome*

CDS

Primary Data

Name	CSS0011631.1
Type	CDS
Position	Chr1:105801011..105801104 (+ strand)
Length	94 bp

Attributes

id	CSS0011631.1
phase	0
seq_id	Chr1
source	AHAU
uniqueID	offset-21050546

Region sequence

FASTA

```
>Chr1 Chr1:105801011..105801104 (+ strand) class=CDS
length=94
TGGGAGAACACATTCGTACGGGTGTGGTCAGATTAGTGACAAAAACGTGCGCACGTATG
GCCAAGTCATAATCAGACGTCGATCTATGAGGTAA
```

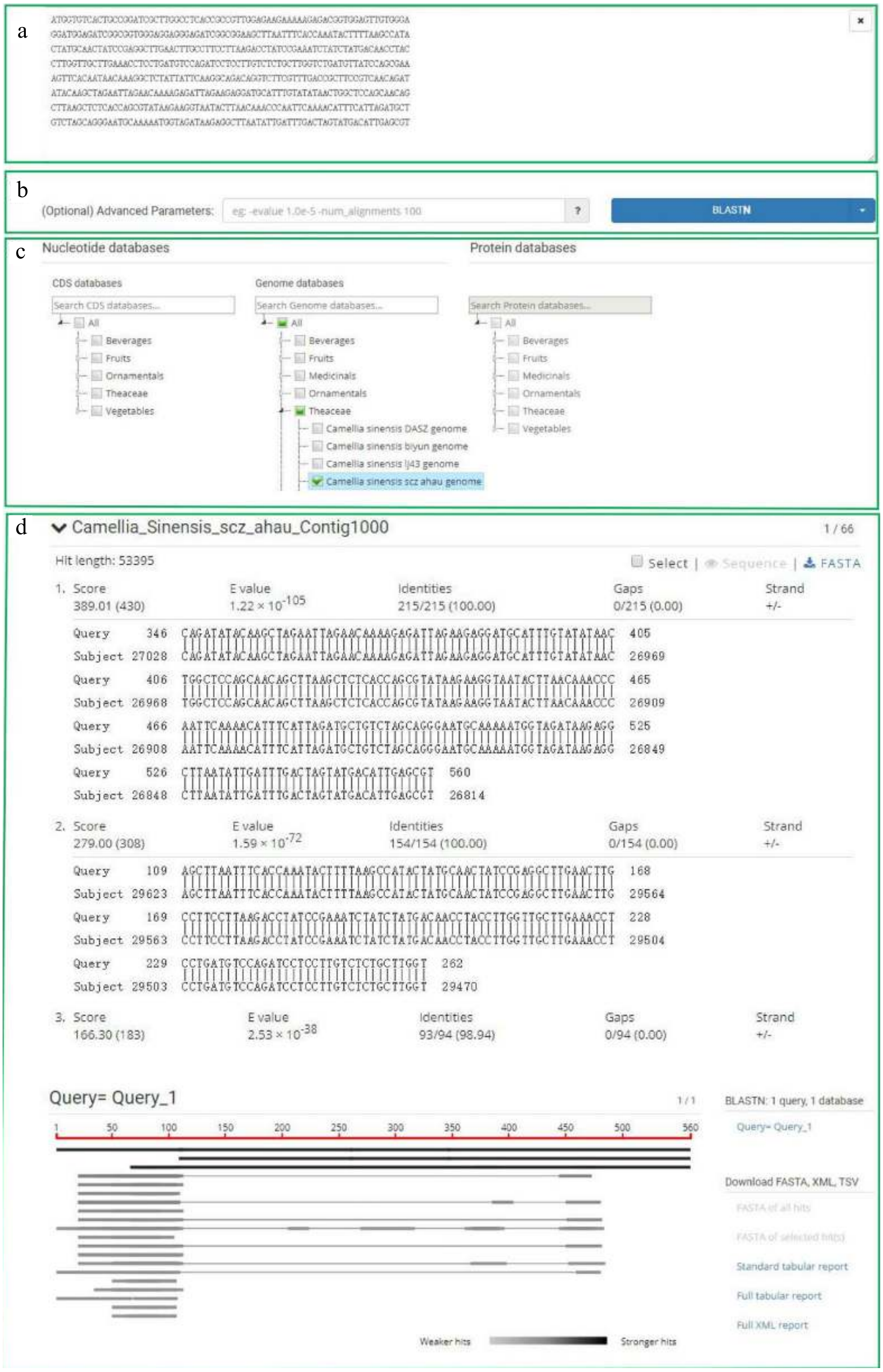
OK

Fig. 5 JBrowse tools. Users can visualize the location of the query gene in the genome with the help of JBrowse. (a) JBrowse Genome Data Selection Page, (b) visual interface of genetic data on the chromosome, (c) details of the genetic data.

such as -e-value, -num_alignments, etc. The BLAST result contains the alignment result including the queried sequence and referred sequence from the database, which indicates homologous genes or a desired gene (Fig. 6d). Again, we will upload new genome, cds, and protein sequence files to update BLAST.

SSR search

The open-source tool Krait[67] was used for the mining of SSRs from different data sources. Perfect SSRs from genomic and chloroplast genome sequences were identified for five different categories such as di- to hexa-nucleotide with a minimum repeat motif length of ≥ 18 bp. These include di-



a

```
ATGGTGTCACTGCCGGATCGCTTGGCCTCACCGCCGTTGGAGAAGAAAAAGAGACGGTGGAGTTGTGGGA
GGATGGAGATCGGCGGTGGGAGGAGGGAGATCGGCGGAAGCTTAATTTCACCAAATACTTTTAAGCCATA
CTATGCAACTATCCGAGGCTTGAACTTGCCTTCCTTAAGACCTATCCGAAATCTATCTATGACAACCTAC
CTTGGTTGCTTGAAACCTCCTGATGTCCAGATCCTCCTTGTCTCTGCTTGGTCTGATGTTATCCAGCGAA
AGTTCACAATAACAAAGGCTCTATTATTCAAGGCAGACAGGTCTTCGTTTGACCGCTTCCGTCAACAGAT
ATACAAGCTAGAATTAGAACAAAAGAGATTAGAAGAGGATGCATTTGTATATAACTGGCTCCAGCAACAG
CTTAAGCTCTCACCAGCGTATAAGAAGGTAATACTTAACAAACCCAATTCAAAACATTTCATTAGATGCT
GTCTAGCAGGGAATGCAAAAATGGTAGATAAGAGGCTTAATATTGATTTGACTAGTATGACATTGAGCGT
```

b

(Optional) Advanced Parameters: eg: -evalue 1.0e-5 -num_alignments 100

c

Nucleotide databases

CDS databases

Search CDS databases...

- □ All
  - □ Beverages
  - □ Fruits
  - □ Ornamentals
  - □ Theaceae
  - □ Vegetables

Genome databases

Search Genome databases...

- ■ All
  - □ Beverages
  - □ Fruits
  - □ Medicinals
  - □ Ornamentals
  - ■ Theaceae
    - □ Camellia sinensis DASZ genome
    - □ Camellia sinensis biyun genome
    - □ Camellia sinensis lj43 genome
    - ☑ Camellia sinensis scz ahau genome

Protein databases

Search Protein databases...

- □ All
  - □ Beverages
  - □ Fruits
  - □ Medicinals
  - □ Ornamentals
  - □ Theaceae
  - □ Vegetables

d

Camellia_Sinensis_scz_ahau_Contig1000 1 / 66

Hit length: 53395

```
1. Score           E value                Identities          Gaps           Strand
   389.01 (430)    1.22 × 10^-105         215/215 (100.00)    0/215 (0.00)   +/-

   Query      346  CAGATATACAAGCTAGAATTAGAACAAAAGAGATTAGAAGAGGATGCATTTGTATATAAC  405
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Subject  27028  CAGATATACAAGCTAGAATTAGAACAAAAGAGATTAGAAGAGGATGCATTTGTATATAAC  26969
   Query      406  TGGCTCCAGCAACAGCTTAAGCTCTCACCAGCGTATAAGAAGGTAATACTTAACAAACCC  465
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Subject  26968  TGGCTCCAGCAACAGCTTAAGCTCTCACCAGCGTATAAGAAGGTAATACTTAACAAACCC  26909
   Query      466  AATTCAAAACATTTCATTAGATGCTGTCTAGCAGGGAATGCAAAAATGGTAGATAAGAGG  525
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Subject  26908  AATTCAAAACATTTCATTAGATGCTGTCTAGCAGGGAATGCAAAAATGGTAGATAAGAGG  26849
   Query      526  CTTAATATTGATTTGACTAGTATGACATTGAGCGT  560
                   |||||||||||||||||||||||||||||||||||
   Subject  26848  CTTAATATTGATTTGACTAGTATGACATTGAGCGT  26814

2. Score           E value                Identities          Gaps           Strand
   279.00 (308)    1.59 × 10^-72          154/154 (100.00)    0/154 (0.00)   +/-

   Query      109  AGCTTAATTTCACCAAATACTTTTAAGCCATACTATGCAACTATCCGAGGCTTGAACTTG  168
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Subject  29623  AGCTTAATTTCACCAAATACTTTTAAGCCATACTATGCAACTATCCGAGGCTTGAACTTG  29564
   Query      169  CCTTCCTTAAGACCTATCCGAAATCTATCTATGACAACCTACCTTGGTTGCTTGAAACCT  228
                   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   Subject  29563  CCTTCCTTAAGACCTATCCGAAATCTATCTATGACAACCTACCTTGGTTGCTTGAAACCT  29504
   Query      229  CCTGATGTCCAGATCCTCCTTGTCTCTGCTTGGT  262
                   ||||||||||||||||||||||||||||||||||
   Subject  29503  CCTGATGTCCAGATCCTCCTTGTCTCTGCTTGGT  29470

3. Score           E value                Identities          Gaps           Strand
   166.30 (183)    2.53 × 10^-38          93/94 (98.94)       0/94 (0.00)    +/-
```

Query= Query_1 1 / 1

BLASTN: 1 query, 1 database

Fig. 6 The integrated BLAST tool. Users could search tea plant homologous genes by using the BLAST function. (a) Sequence input box, (b) optional parameters for sequence alignment, (c) sequence alignment database, (d) sequence alignment results.

nucleotide repeats ≥ 10 bp, tri-nucleotide repeats ≥ 6 bp, tetra-nucleotide repeats ≥ 5 bp, penta-nucleotide repeats ≥4 bp, and hexa-nucleotide repeats of ≥ 4 bp[68]. This tool was built to help users massively detect polymorphic SSRs (PolySSR) in tea plants, which is an important and widely used genetic marker for tea plant population genetics, marker assistant selection, and molecular breeding[63]. Users can enter the base sequence in the input box to search for the specific SSR contained in a single tea species.

Compared with TeaMiD, TPGDB contains all published tea plant genome data and most of the data is chromosome-scale, while TeaMiD only contains three tea plant genomes: CSS 'Shuchazao', CSA 'Yunkang 10' and *C. assamica* L (O). Kuntze 'TV-1'[69], and all data is scaffold-scale. For SSRs, the number and quality of SSR sequences identified by the genome are positively correlated with the assembly quality of the genome. For example, the number of di-, tri-, tetra-, penta-, and hexanucleotide simple sequence repeats in the chromosome-scale genome data CSS 'Shuchazao' (2020 released) is significantly more than that in the scaffold-scale genome data CSS 'Shuchazao' (2018 released) (Table 2). Nowadays, high-quality assembled genome data of tea plants are continuously released, and the reference data used by researchers is therefore the latest released genome data. Therefore, we collect all the published high-quality genomic data and identify the SSRs of these data, so that researchers can efficiently collect and query SSRs. In order to speed up the research process of tea plant breeding for resistance, we will continue to optimize the SSR search module and add new functions during database updates.

Table 2. SSRs integrated in TPGDB and TeaMiD.

Database	Count	Species	Assembly scale	Di	Tri	Tetra	Penta	Hexa
TPGDB	7	CSS 'Biyun'	Chromosome	200 688	106 538	31 272	32 222	22 695
		CSS 'Shuchazao'[1]	Chromosome	202 217	108 561	32 822	33 629	23 712
		CSS 'Shuchazao'[2]	Chromosome	163 949	82 721	28 356	31 275	21 408
		CSS 'Huangdan'	Chromosome	190 146	106 242	31 604	32 562	23 956
		CSS 'Longjing 43'	Chromosome	195 729	109 900	31 620	32 195	21 882
		CSA 'Yunkang 10'	Scaffold	118 758	49 111	17 059	22 058	15 878
		Wild tea plant[3]	Chromosome	190 468	104 621	30 014	31 930	22 488
TeaMiD	3	CSS 'Shuchazao'[4]	Scaffold	118 777	21 352	17 096	5 183	4 585
		CSA 'Yunkang 10'	Scaffold	163 982	33 223	28 426	7 289	6 091
		CA 'TV-1'[5]	Scaffold	138 689	25 392	18 829	5 720	5 281

[1] *C. sinensis* var. *sinensis* 'Shuchazao' was released by AHAU in 2020.
[2] *C. sinensis* var. *sinensis* 'Shuchazao' was released by TRI in 2020.
[3] Wild tea plant (DASZ) was released by HZAU in 2020.
[4] *C. sinensis* var. *sinensis* 'Shuchazao' was released by AHAU in 2018.
[5] *C. assamica* L (O). Kuntze 'TV-1' was released by IARI.

Downloads

The Download page mainly includes the cds and protein sequence files of tea plant genomes and chloroplast genomes. It includes the annotation data files obtained through software such as hmmer, signalp, interproscan, kofamkoala etc. All these data are available freely to the research community.

Community modules

For the Community module, we collected the references on tea plant genome research. Users can click on the article title link and link to the download page. If users encounter problems on the website or have suggestions, they can contact us through the 'Contact us' page. If users want to learn more about the tea plant genome, co-expression, other plant genome databases, and other plant genetic research-related institutions and platforms, they can click directly on the 'more databases' page to obtain related database websites, which facilitate users in obtaining further resources for genomic research.

CONCLUSION

With the rapid development of sequencing technology and bioinformatics software, more and more tea plant genomes have been released. How to efficiently collect, organize, and annotate genetic data to mine the enormous biological value of genomic data will become more important. We conducted information mining on genomic data and established TeaPGDB. The goal of TeaPGDB is not only a genomic information database, but is also a user-friendly portal that currently provides simple genetic data analysis tools. It will be useful and important to add and update more advanced genetic data analysis tools based on bioinformatics in the future.

We also hope to continuously update and improve TeaPGDB. In the future, we will develop and establish more gene online analysis tools to facilitate tea researchers in conducting online analysis. In summary, this new database incorporates published tea plant genomes, multiple analysis tools, new features for tea plant genomic data analysis, gene function characterization, SSR marker selection, flavor related gene mapping, and updates, which is easily accessed and can potentially benefit the tea plant research community.

ACKNOWLEDGMENTS

Wanping Fang, Fei Chen acknowledge the support by the high-performance computing platform of bioinformatics center, Nanjing Agricultural University. Wanping Fang acknowledge this research supported by the National Key R & D Program of China (2020YFE0202900), The National Natural Science Foundation of China (31972460, 31870680), the earmarked fund for China Agriculture Research System (CARS-19), the Key Research and Development Program of Jiangsu Province (BE2019379), Jiangsu Agriculture Science and Technology Innovation Fund (CX(20)2004), Innovation and Extension Projects of Forestry Science and Technology in Jiangsu Province (LYKJ-Changzhou[2020]03), Changzhou

Science and Technology Support Program (Agriculture CE20202003), and Chuzhou Science and Technology Support Program (2020ZN009).

Conflict of interest

The authors declare that they have no conflict of interest.

Dates

Received 25 May 2021; Accepted 25 June 2021; Published online 30 June 2021

REFERENCES

1. Xia E, Zhang H, Sheng J, Li K, Zhang Q, et al. 2017. The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Molecular Plant* 10:866−77
2. Chen Y, Yu M, Xu J, Chen X, Shi J. 2009. Differentiation of eight tea (*Camellia sinensis*) cultivars in China by elemental fingerprint of their leaves. *Journal of the Science of Food and Agriculture* 89:2350−55
3. Juneja LR, Chu DC, Okubo T, Nagato Y, Yokogoshi H. 1999. L-theanine - a unique amino acid of green tea and its relaxation effect in humans. *Trends in Food Science & Technology* 10:199−204
4. Mukhtar H, Ahmad N. 2000. Tea polyphenols: prevention of cancer and optimizing health. *The American Journal of Clinical Nutrition* 71:1698S−1702S
5. Koshiishi C, Kato A, Yama S, Crozier A, Ashihara H. 2001. A new caffeine biosynthetic pathway in tea leaves: utilisation of adenosine released from the *S*-adenosyl-L-methionine cycle. *FEBS Letters* 499:50−54
6. Zhang Z, Li Y, Qi L, Wan X. 2006. Antifungal activities of major tea leaf volatile constituents toward *Colletorichum camelliae Massea*. *Journal of Agricultural and Food Chemistry* 54:3936−40
7. Zhang S, Xuan H, Zhang L, Fu S, Wang Y, et al. 2017. TBC2health: a database of experimentally validated health-beneficial effects of tea bioactive compounds. *Briefings in Bioinformatics* 18:830−36
8. Xia E H, Tong W, Wu Q, Wei S, Zhao J, et al. 2020. Tea plant genomics: achievements, challenges and perspectives. *Horticulture Research* 7:7
9. Ott J, Wang J, Leal SM. 2015. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics* 16:275−84
10. Zhang C, Wang L, Wei K, Wu L, Li H, et al. 2016. Transcriptome analysis reveals self-incompatibility in the tea plant (*Camellia sinensis*) might be under gametophytic control. *BMC Genomics* 17:359
11. Kang M, Wu H, Yang Q, Huang L, Hu Q, et al. 2020. A chromosome-scale genome assembly of *Isatis indigotica*, an important medicinal plant used in traditional Chinese medicine. *Horticulture Research* 7:18
12. Wei S, Yang Y, Yin T. 2020. The chromosome-scale assembly of the willow genome provides insight into Salicaceae genome evolution. *Horticulture Research* 7:45
13. Li S, Wang J, Dong R, Zhu H, Lan L, et al. 2020. Chromosome-level genome assembly, annotation and evolutionary analysis of the ornamental plant *Asparagus setaceus*. *Horticulture Research* 7:48
14. Gao S, Wang B, Xie S, Xu X, Zhang J, et al. 2020. A high-quality reference genome of wild *Cannabis sativa*. *Horticulture Research* 7:73
15. Xue T, Zheng X, Chen D, Liang L, Chen N, et al. 2020. A high-quality genome provides insights into the new taxonomic status and genomic characteristics of *Cladopus chinensis* (Podostemaceae). *Horticulture Research* 7:46
16. Xanthopoulou A, Manioudaki M, Bazakos C, Kissoudis C, Farsakoglou AM, et al. 2020. Whole genome re-sequencing of sweet cherry (*Prunus avium* L.) yields insights into genomic diversity of a fruit species. *Horticulture Research* 7:60
17. Hu M, Sun W, Tsai WC, Xiang S, Lai X, et al. 2020. Chromosome-scale assembly of the *Kandelia obovata* genome. *Horticulture Research* 7:60
18. Fan Y, Sahu SK, Yang T, Mu W, Wei J, et al. 2020. Dissecting the genome of star fruit (*Averrhoa carambola* L.). *Horticulture Research* 7:94
19. Peace CP, Bianco L, Troggio M, van de Weg E, Howard NP, et al. 2019. Apple whole genome sequences: recent advances and new prospects. *Horticulture Research* 6:59
20. Li Q, Qi J, Qin X, Dou W, Lei T, et al. 2020. CitGVD: a comprehensive database of citrus genomic variations. *Horticulture Research* 7:12
21. Liu T, Li M, Liu Z, Ai X, Li Y. 2021. Reannotation of the cultivated strawberry genome and establishment of a strawberry genome database. *Horticulture research* 8(1):41
22. Yue J, Liu J, Tang W, Wu Y, Tang X, et al. 2020. Kiwifruit Genome Database (KGD): a comprehensive resource for kiwifruit genomics. *Horticulture Research* 7:117
23. Yano K, Aoki K, Shibata D. 2007. Genomic Databases for Tomato. *Plant Biotechnology* 24:17−25
24. Xu H, Yu Q, Shi Y, Hua X, Tang H, et al. 2018. PGD: Pineapple Genomics Database. *Horticulture Research* 5:66
25. Kim C, Park D, Seol Y, Yoon U, Lee G, et al. 2012. An online database for genome information of agricultural plants. *Bioinformation* 8:1059−61
26. Chen J, Zheng C, Ma J, Jiang C, Ercisli S, et al. 2020. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Horticulture Research* 7:63
27. Xia E, Tong W, Hou Y, An Y, Chen L, et al. 2020. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Molecular Plant* 13:1013−26
28. Wang X, Feng H, Chang Y, Ma C, Wang L, et al. 2020. Population sequencing enhances understanding of tea plant evolution. *Nature Communications* 11:4447
29. Zhang Q, Li W, Li K, Nan H, Shi C, et al. 2020. The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. *Molecular Plant* 13:935−38
30. Zhang W, Zhang Y, Qiu H, Guo Y, Wan H, et al. 2020. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nature Communications* 11:3719
31. Wei C, Yang H, Wang S, Zhao J, Liu C, et al. 2018. Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. *PNAS* 115:E4151−E4158
32. Wang P, Yu J, Jin S, Chen S, Yue C, et al. 2021. Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. *Horticulture research* 8:107
33. Guo W, Chen J, Li J, Huang J, Wang Z, et al. 2020. Portal of Juglandaceae: A comprehensive platform for Juglandaceae study. *Horticulture Research* 7:35
34. Zeng C, Hollingsworth PM, Yang J, He Z S, Zhang Z R, et al. 2018. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Methods* 14:43

35. Dong M, Liu S, Xu Z, Hu Z, Ku W, et al. 2018. The complete chloroplast genome of an economic plant, *Camellia sinensis* cultivar Anhua, China. *Mitochondrial DNA Part B - Resources* 3:558−59

36. Huang H, Shi C, Liu Y, Mao S, Gao L. 2014. Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evolutionary Biology* 14:151

37. Chen S, Li R, Ma Y, Lei S, Ming R, et al. 2021. The complete chloroplast genome sequence of *Camellia sinensis* var. *sinensis* cultivar Tieguanyin (Theaceae). *Mitochondrial DNA Part B-Resources* 6:395−96

38. Ye X, Zhao Z, Xu Y, Xu H, Sun L. 2015. The Phylogenetic Analysis of *Camellia sinensis* cv. Longjing 43. *Journal of the Korean Tea Society, special* 21:63−66

39. Li L, Hu Y, He M, Zhang B, Wu W, et al. 2021. Comparative chloroplast genomes: insights into the evolution of the chloroplast genome of *Camellia sinensis* and the phylogeny of Camellia. *BMC genomics* 22:138

40. Li L, Hu Y, Wu L, Chen R, Luo S. 2021. The complete chloroplast genome sequence of *Camellia sinensis* cv. *Dahongpao*: a most famous variety of Wuyi tea (Synonym: *Thea bohea* L.). *Mitochondrial DNA Part B - Resources* 6:3−5

41. Hao W, Wang S, Yao M, Ma J, Xu Y, et al. 2019. The complete chloroplast genome of an albino tea, *Camellia sinensis* cultivar 'Baiye 1'. *Mitochondrial DNA Part B - Resources* 4:3143−44

42. Lee DJ, Kim CK, Lee TH, Lee SJ, Moon DG, et al. 2020. The complete chloroplast genome sequence of economical standard tea plant, *Camellia sinensis* L. cultivar Sangmok, in Korea. *Mitochondrial DNA Part B-Resources* 5:2841−42

43. Rawal HC, Kumar PM, Bera B, Singh NK, Mondal TK. 2020. Decoding and analysis of organelle genomes of Indian tea (*Camellia assamica*) for phylogenetic confirmation. *Genomics* 112:659−68

44. Zhang F, Li W, Gao C, Zhang D, Gao L. 2019. Deciphering tea tree chloroplast and mitochondrial genomes of *Camellia sinensis* var. *assamica*. *Scientific Data* 6:209

45. Jia X, Zhang W, Fernie AR, Wen W. 2019. *Camellia sinensis* (Tea). *Trends in Genetics* 37:201−2

46. Finn R D, Attwood T K, Babbitt P C, Bateman A, Bork P, et al. 2017. InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research* 45:D190−D199

47. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research* 49:D412−D419

48. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, et al. 2018. HMMER web server: 2018 update. *Nucleic Acids Research* 46:W200−W204

49. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, et al. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36:2251−52

50. Armenteros JJA, Tsirigos KD, Sønderby, CK, Petersen T N, Winther O, et al. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* 37:420−23

51. Ginestet C. 2011. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society Series a - Statistics in Society* 174:245−46

52. Gong X, Zhang D. 2011. Research on Web Server Based on Red5, Tomcat and Apache. *Advanced Materials Research* 282−283:721−25

53. Hoy MB. 2011. HTML5: A new standard for the web. *Medical Reference Services Quarterly* 30:50−55

54. Prokhorenko V, Choo KKR, Ashman H. 2016. Intent-Based Extensible Real-Time PHP Supervision Framework. *IEEE Transactions on Information Forensics and Security* 11:2215−26

55. Di Giacomo M. 2005. MySQL: Lessons learned on a digital library. *IEEE Software* 22:10−13

56. Korpela J. 1998. Lurching toward Babel: HTML, CSS and XML. *Computer* 31:103−4

57. Wei S, Xhakaj F, Ryder BG. 2016. Empirical study of the dynamic behavior of JavaScript objects. *Software-Practice and Experience* 46:867−89

58. Lee S-U, Moon I-Y. 2011. A study of user interaction using jQuery in Web Application. *Journal of Advanced Navigation Technology* 15(4):626−31

59. Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, et al. 2019. Sequenceserver: a modern graphical user interface for custom BLAST databases. *Molecular Biology and Evolution* 36:2922−24

60. Wang H, Ooi BC, Tan KL, Ong TH, Zhou L. 2003. BLAST++: BLASTing queries in batches. *Bioinformatics* 19:2323−24

61. Westesson O, Skinner M, Holmes I. 2013. Visualizing next-generation sequencing data with JBrowse. *Briefings in Bioinformatics* 14:172−77

62. Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In Gene prediction. methods in molecular biology, ed. Kollmar M. vol 1962. New York: Humana. pp. 227−45. https://doi.org/10.1007/978-1-4939-9173-0_14

63. Xia EH, Li FD, Tong W, Li PH, Wu Q, et al. 2019. Tea Plant Information Archive: a comprehensive genomics and bioinformatics platform for tea plant. *Plant Biotechnology Journal* 17:1938−53

64. Yue Y, Chu G, Liu X, Tang X, Wang W, et al. 2014. TMDB: A literature-curated database for small molecular compounds found from tea. *BMC Plant Biology* 14:243

65. Zhang R, Ma Y, Hu X, Chen Y, He X, et al. 2020. TeaCoN: a database of gene co-expression network for tea plant (*Camellia sinensis*). *BMC Genomics* 21:461

66. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, et al. 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* 17:66

67. Du L, Zhang C, Liu Q, Zhang X, Yue B. 2018. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics* 34:681−83

68. Dubey H, Rawal HC, Rohilla M, Lama U, Kumar PM et al. 2020. TeaMiD: a comprehensive database of simple sequence repeat markers of tea. *Database* 2020:baaa013

69. Mondal TK, Rawal HC, Bera B, Kumar PM, Choubey M, et al. 2019. Draft genome sequence of a popular Indian tea genotype TV-1 [*Camellia assamica L.* (*O*). Kunze]. *bioRxiv* Preprint