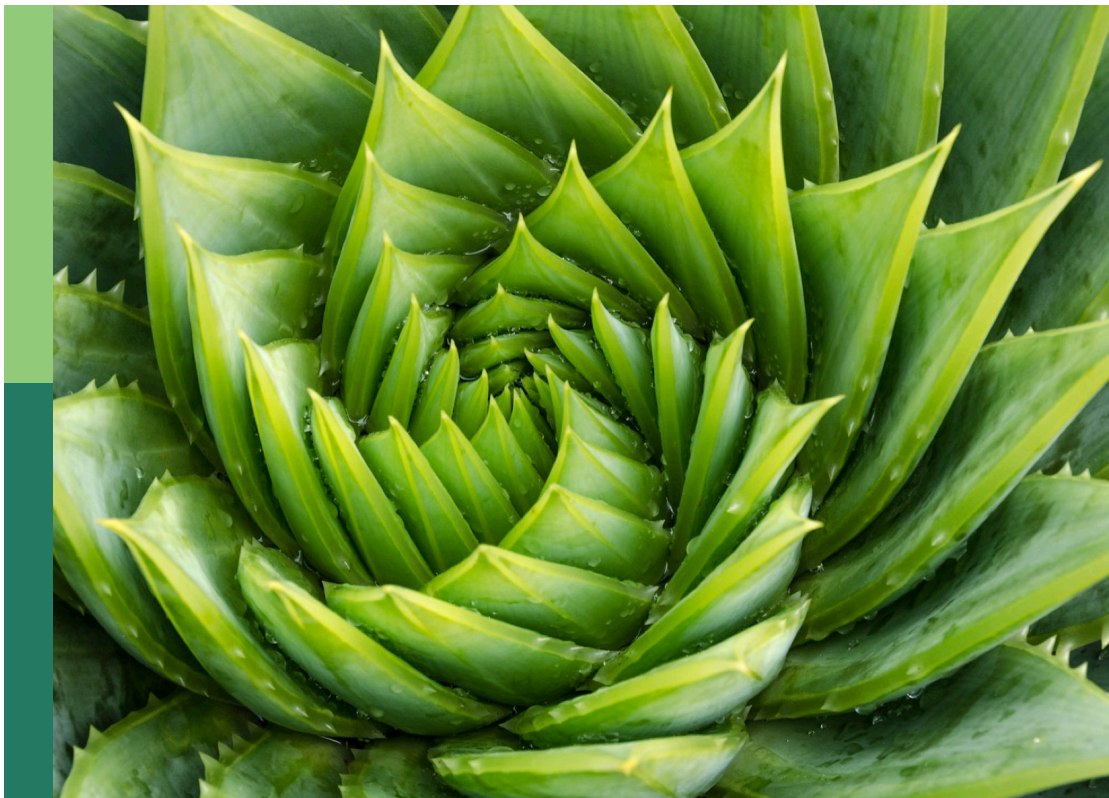
frontiers | Research Topics

Approaches and applications in plant genome assembly and sequence analysis

Edited by
Weihua Pan, Ergude Bao, Surya Saha and Jianyu Zhou

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-83252-012-3
DOI 10.3389/978-2-83252-012-3

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Approaches and applications in plant genome assembly and sequence analysis

Topic editors

Weihua Pan — Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, China

Ergude Bao — Beijing Jiaotong University, China

Surya Saha — Boyce Thompson Institute (BTI), United States

Jianyu Zhou — Nankai University, China

Citation

Pan, W., Bao, E., Saha, S., Zhou, J., eds. (2023). *Approaches and applications in plant genome assembly and sequence analysis*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-83252-012-3

Table of contents

frontiers in Plant Science

TECHNOLOGY AND CODE
published: 03 January 2022
doi: 10.3389/fpls.2021.813036

SLDMS: A Tool for Calculating the Overlapping Regions of Sequences

*Yu Chen[1], DongLiang You[1], TianJiao Zhang[1] and GuoHua Wang[1,2]**

[1] College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2] State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin, China

In the field of genome assembly, contig assembly is one of the most important parts. Contig assembly requires the processing of overlapping regions of a large number of DNA sequences and this calculation usually takes a lot of time. The time consumption of contig assembly algorithms is an important indicator to evaluate the degree of algorithm superiority. Existing methods for processing overlapping regions of sequences consume too much in terms of running time. Therefore, we propose a method SLDMS for processing sequence overlapping regions based on suffix array and monotonic stack, which can effectively improve the efficiency of sequence overlapping regions processing. The running time of the SLDMS is much less than that of Canu and Flye in dealing with the sequence overlap interval and in some data with most sequencing errors occur at both the ends of the sequencing data, the running time of the SLDMS is only about one-tenth of the other two methods.

Keywords: algorithm, sequence analysis, genome assembly, contig assembly, overlapping regions, application

OPEN ACCESS

Edited by:
Ergude Bao,
Beijing Jiaotong University, China

Reviewed by:
Bin Liu,
Beijing Institute of Technology, China
Dong-Jun Yu,
Nanjing University of Science and Technology, China
Hongmin Cai,
South China University of Technology, China

***Correspondence:**
GuoHua Wang
ghwang@nefu.edu.cn

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *11 November 2021*
Accepted: *29 November 2021*
Published: *03 January 2022*

Citation:
Chen Y, You D, Zhang T and Wang G (2022) SLDMS: A Tool for Calculating the Overlapping Regions of Sequences. Front. Plant Sci. 12:813036. doi: 10.3389/fpls.2021.813036

INTRODUCTION

Due to the limitations of existing gene sequencing technology, we cannot directly obtain the entire gene sequence, but can only use existing sequencing methods to sequence the genes of the species to be tested to generate sequence fragments and then further genome assembly to restore the original genes. The genome assembly problem is also one of the most important and difficult problems in bioinformatics today.

The two algorithms commonly used in genome assembly are the overlap-layout-consensus (OLC) (Li, 2012) algorithm and the de-bruijn-graph (DBG) (Li, 2012) algorithm, which use different methods to convert the assembly problem into a graph-theoretic related problem. By creating an edge-weighted graph of the sequencing data, the resulting edge-weighted graph is processed to find relevant pathway information in the graph for use in downstream genome assembly work. All the algorithms derive the optimal path from the edge-weighted graph to obtain the initial contig (Huang, 1992).

Most applications for genome assembly are based on one of the algorithms such as the Canu (Koren et al., 2017), which chooses to use the MHAP (Koren et al., 2017) algorithm to detect the overlap in noisy sequences to obtain the overlapping regions between sequences. Additionally, the Flye (Lin et al., 2016) software uses the ABruijn (Lin et al., 2016) algorithm to combine the OLC and DBG algorithms, generates its own unique A-bruijn-graph (ABG) graph, and obtains the overlapping regions of the nodes in the graph and some other assembly software can also complete the same work. These software are usually more time-consuming in the sequence alignment process. For example, Flye takes longer to find overlapping regions of sequences and Canu is slower to correct sequencing data, etc.

This article presents a new software for overlapping regions calculation called the SLDMS, a tool that uses gene sequencing data as input and supports the fastq and fasta formats. It can calculate an output overlapping regions information between sequencing data and write it to a file, so that other applications can use it. Compared to other genome assembly software that calculates overlapping regions, our monotonic stack and suffix array-based design approach is more efficient and provides richer pathway information for downstream genome assembly software to use as a reference. At the same time, the SLDMS can be easily integrated into the genomic analysis process.

METHODS

The overall workflow of the SLDMS (**Figure 1**) includes four steps: (i) data preprocessing; (ii) building a suffix array; (iii) selecting the software version and establishing the relevant data structure; and (iv) traversing the suffix array and output the results of overlapping regions.

Data Structure

The SLDMS needs three arrays when obtaining overlapping regions information, These three arrays are the suffix array (SA) array (Manber and Myers, 1993) longest common prefix (LCP) array (Fischer, 2010), and Document array (DA) array (Muthukrishnan, 2002). First, we briefly introduce these three arrays. The SA array is the suffix array and SA(i) represents the starting position of the suffix whose string rank is i in the original string. The LCP array is the longest common prefix array and LCP(i) represents the longest common prefix of the suffixes represented by SA(i) and SA(i-1). The DA array is a document array. DA(i) represents the number of strings in the input data to which the suffix ranked i belongs to. This array can be obtained in the process of obtaining SA and LCP.

The meaning of the elements stored in the three arrays is given in **Figure 2**. The SA array: The string above the array is the suffix represented by each item in the array and the value stored in the array is the starting position of the suffix it represents in the original string. LCP array: The string above the array is the suffix represented by each item in the array and the value stored in the array is the length of the longest common prefix between the suffix it represents and the suffix ranked one place ahead of it; to calculate this length, we ignore the ending symbols of the suffix. The DA array: The string above the array is the suffix represented by each item in the array and the value stored in the array is the source of the suffix. For example, DA(i) = 20, "babbc" belongs to the 20th input sequence.

Algorithm Principle

We define read as a piece of data in the sequencing data and its representation in the computer is a string. Before finding the overlapping regions information of the sequencing data, first consider the case of finding the overlapping regions information for two reads. Suppose the two reads are str1 and str2; if the tail of str1 and the head of str2 overlap and the overlap starts at position i, then the suffix suf [suf = str1(i:)] of str1 must be the same as some prefix of str2. In other words, the overlap part is the prefix of str2 (otherwise, str2 is the substring of str1 and the splicing is equivalent to discarding str2, so there is no need to splice str1 with str2 in this case), so the longest overlap part of str1 and str2 must be a suffix belonging to str1 that is ranked before str2 in the dictionary order. The overlapping regions information can be obtained by sorting all the suffixes of str1 with str2 and processing the suffixes ranked before str2 (**Figure 3**). As shown in **Figure 3**, the set of suffixes in **Figure 3** does not show suffixes belonging to str2 because read cannot be connected to itself, so it will make a judgment on the belonging of suffixes and ignore these suffixes belonging to itself when calculating the candidate answers of str2.

Extend the case of two reads overlapping to a set of reads. All the suffixes of the reads are sorted, the best overlapping regions information of each read must exist in some suffix ranked before it, and the read to which this suffix belongs is the maximum possible adjacent node of the current read after building the graph.

Since reading itself is also a suffix of reads, when traversing the set of suffixes, if we encounter a certain read itself, we are able to guarantee that the suffix with its optimal overlapping regions information must have been traversed. Then, the problem is transformed into that the longest common prefix, which is calculated for all the suffixes ranked before it and the current string and when the length of the common prefix is equal to the length of this suffix, we consider this suffix as a candidate answer and select the best or top-K optimal as the final answer among all the candidate answers.

In the design of the algorithm, we choose two advanced data structures, suffix array, and monotone stack, because the information stored in the suffix array is the suffix of the string sorted in dictionary order, which corresponds to the suffix set in **Figure 3**. The reason for choosing the monotonic stack is that the monotonic stack can work with the LCP array to filter the set of suffixes and remove those suffixes that are not likely to be the answer and improve the computation speed in this way. When we get a certain candidate answer with length x, the rest of the candidates with length greater than x must not match exactly with the subsequent reads; this is because their common prefixes have a maximum length of x, so these candidates should be removed. The monotonic stack exists to remove this part of information.

IMPLEMENTATION

Constructing the SA Array, LCP Array, and DA Array

Since the method of constructing the SA and LCP arrays in a single string is quite mature, the SLDMS stitches all the reads in the data into a single string, splitting the sequence with American Standard Code for Information Interchange (ASCII) code 1, and ending the stitching with ASCII code 0. This allows the read set to be treated as a string and we call this stitched together extra-long string the "original string." For this part, we use gsufsort (Louza et al., 2020) software based

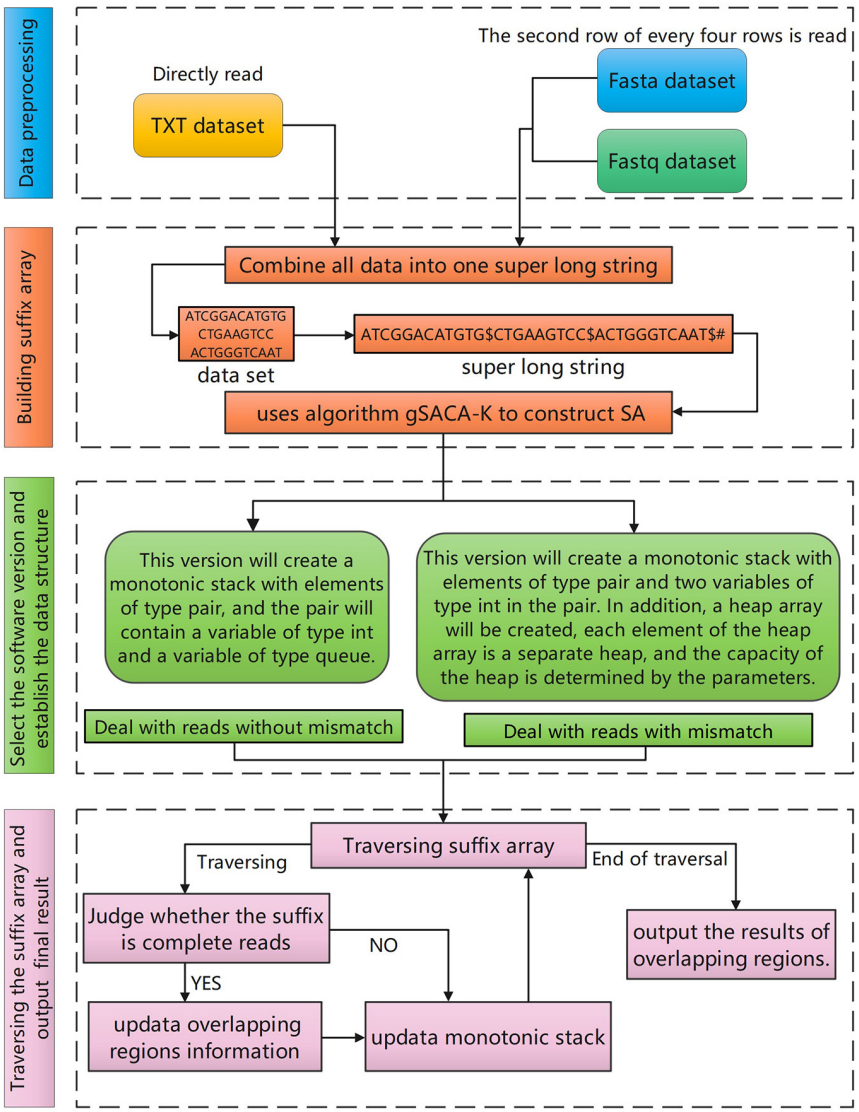


FIGURE 1 | The overall workflow of the SLDMS.

on the gSACA-K (Louza et al., 2017) algorithm to obtain the three arrays.

Maintaining the Monotonic Stack

Assuming that the number of reads is *n*, after obtaining the information of the SA array and LCP array, the first n + 1 items of the SA array are traversed. These n + 1 items are the positions of the interval $ between strings and the string terminator # in the original string. Therefore, we can obtain the start and end positions of each read in the original string and record them in the Fi array and Se array. For example, the start position of the xth read is Fi(x) = SA(X − 1) + 1 and the end position is Se(x) = SA(x). According to the start position and end position of each read, its length was also calculated as LEN(x) = [SE(x) − Fi(x) + 1].

After obtaining the above information, the matching process of suffixes and reads can be optimized by maintaining a monotonic stack. For different input data, different strategies are adopted and the SLDMS was designed in two versions. The first version considers the data to be completely correct and can directly perform the overlapping regions calculation. The accuracy of the result depends on the input dataset and if the dataset is completely correct, the result is also completely correct. Therefore, the input data required to use this version should be either corrected high-accuracy data or raw high-accuracy data such as the PacBio-HiFi (Hon et al., 2020) dataset and the Sanger dataset. The second version allows some differences between reads when performing overlapping regions calculations. Overlapping regions information can be obtained for data with some errors, the accuracy of the information fluctuates depending on the characteristics of the errors in the

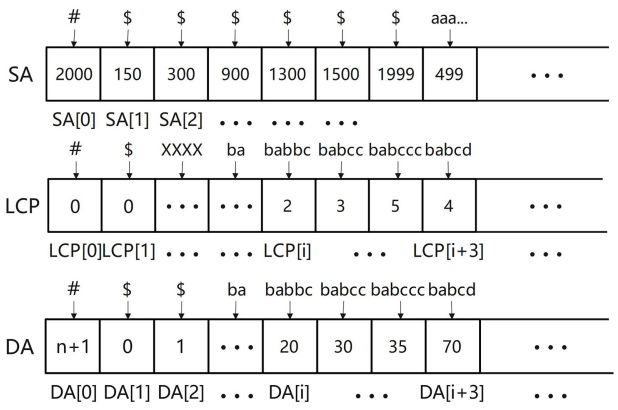


FIGURE 2 | The elements stored in three arrays.

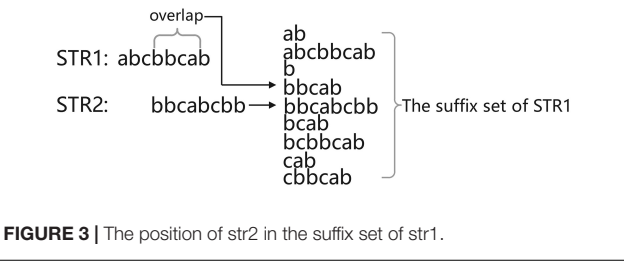


FIGURE 3 | The position of str2 in the suffix set of str1.

data, and the accuracy of the results of the runs varies from one dataset to another. After experiments, it is found that the accuracy of the run results is greatly improved when the error part in the reads is gathered at both the ends. Therefore, for the second version, if the errors in the dataset are completely random, it is recommended to correct the complete data first before using the first version or correct the data center part first before using the second version of the software.

For the part of data error correction, we suggest that the third-generation sequencing data PacBio can be used for self-error correction (Hon et al., 2020) or the second-generation sequencing Illumina data can be used for error correction of the third-generation sequencing data PacBio (Mahmoud et al., 2017) such as PBCR (Koren et al., 2012) in the famous Celera Assembler (Schatz, 2006; Denisov et al., 2008) software and LoRDEC (Leena and Eric, 2014) error correction tool. For the datasets with some regularity of data errors (the errors of the sequencing appear on both the sides of the reads), the second version of this software can be chosen directly.

Deal With Reads Without Mismatch

In this version, since the data can maintain a high accuracy rate, it is enough to directly obtain the overlapping regions information of the reads and the problem of error correction of the reads is not involved. The algorithm idea is as follows.

Build a monotonic stack. The stack is implemented by array simulation to facilitate the acquisition of data in the stack. The element type stored in the stack is a structure similar to a pair designed by us. Its first element is of the int type, which is used to represent the length of the suffix stored in the current element. Its second element is a rolling array, which is used to store the DA information of the suffix that meets the first condition. The length of the array can be set artificially; for example, the length of the array is set to *n*, i.e., to store the top n best answers for each read. In this way, we can obtain more overlapping information between reads. The struct design of the monotone stack and stack elements is found in **Supplementary Figure 1**, where vector is the structure of the stack, pair is the element stored in stack, and queue is the main part of storing information in the pair, which is realized by a rolling array (**Supplementary Figure 2**).

Maintain a monotonous stack (**Figure 4A**). Let one suffix ranked Y be str and for all the suffixes ranked before Y, assuming their rank is X, their longest common prefix with str, i.e., the length of LCP, must be equal to min[LCP(X + 1:Y)]. According to this property, when traversing the SA and LCP arrays, each time a new LCP(i) is traversed and the elements of the stack whose first item is larger than LCP(i) can be taken off the stack because for these elements and the following suffix the LCP cannot be greater than LCP(i), so these suffixes become useless information and can be cleaned up. Start to operate the elements in the monotonic stack from the top of the stack; if the first element at the top of the stack is larger than LCP(x), just get this top element out of the stack directly and loop this operation until it is impossible to get out of the stack. After clearing, check the pair at the top of the stack, whether its first element is equal to the length of the string corresponding to the current SA(i) [len = se(DA(i) – SA(i))], if it is equal, put DA(i) into the second scrolling array of the pair and update the array. If it is not equal, create a new pair element whose first is equal to len and whose initial values in the second element are set as follows: head = 0, tail = 1, have = 1, size = k, and data(0) = DA(i). After the element is created, this element is put on the stack. Since all the elements in the stack whose first is greater than len have been removed before entering the stack, each element entering the stack must be the largest element in the stack, so the monotonicity of the stack can be guaranteed, which is exactly the reason for using a monotonic stack.

Get overlapping regions information (**Figure 4B**). In the process of maintaining the monotonic stack, if the current suffix is an ordinary suffix, just follow the normal process of maintaining the monotonic stack and if the current suffix is a complete read [the value of SA(i) is equal to fi(DA(i))], then we should add the process of information acquisition to the normal maintenance process; at this time, you can maintain the monotonic stack information to obtain the required overlapping regions information. The way to obtain the required overlapping regions information is very simple; first of all, we must first check the top of the stack elements to ensure that the top of the stack elements are not expired (if the elements are expired, they can be taken out of the stack). Then, read the data from the top of the stack and read the second item of each element; these are the reads that are most likely to overlap with the current read and output the overlapping regions information to the result file for use in building the edge weight graph. By default, the first ten possible results are provided, the longest overlapping regions read is usually selected, and the specific read chosen as the path

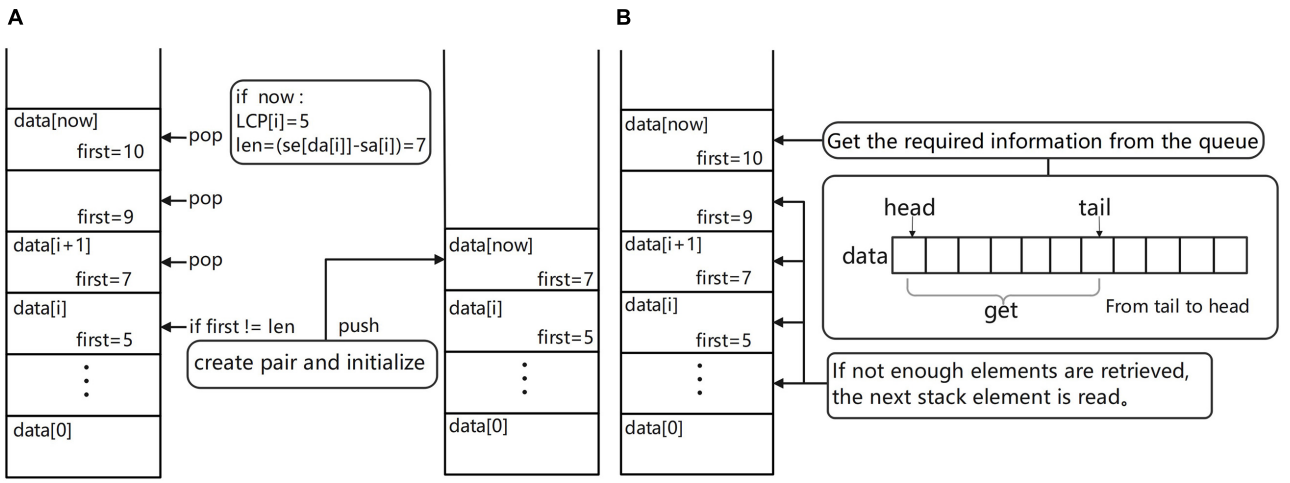


FIGURE 4 | (A) Schematic diagram of the monotonic stack maintenance process. **(B)** Diagram of the process of obtaining overlapping regions information.

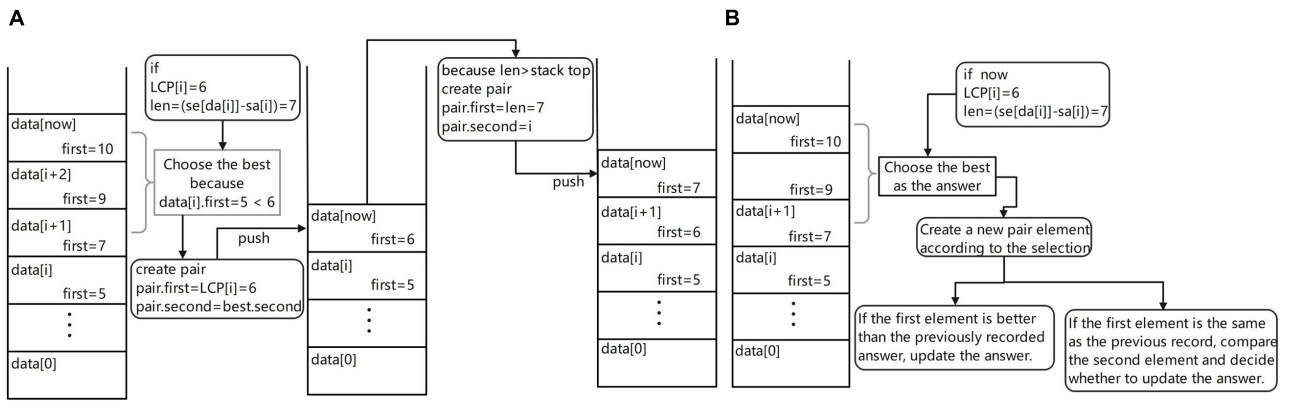


FIGURE 5 | (A) Maintenance process diagram of monotone stack and **(B)** Overlap regions information acquisition flow diagram.

in the edge weight graph can be freely chosen according to the subsequent software requirements.

Deal With Reads With Mismatch

The first-generation sequencing data are high-accuracy data, but they are no longer in mainstream use because of their expensive sequencing price. The second-generation sequencing data are short-read data with high accuracy and are more suitable for use with DBG assembly software based on K-mer counting (Wang et al., 2020) such as the SOAPdenovo (Xie et al., 2014) software. The third-generation sequencing data are long-read data with a high error rate and there would be a high error rate if the sequencing data was matched exactly, so this version allows for slight differences in sequence during the matching process. This version is an alternative solution to cope with the situation in which the data cannot be completely corrected because of the long correction time or high correction cost of the dataset.

In this version, the error part of the input data should appear at both the ends of the data as far as possible or the data center part has been corrected to ensure that there will be no error in the middle part of each read. The closer the error location is to both the ends, the better the overlapping regions information will be. In this version, a parameter K will be entered, which determines the maximum allowable cutting size of the sequence head and tail when the algorithm is matching. If K is set speculatively without knowing much about the dataset, there may be some error in the result obtained from a single run due to the parameters. But, the speed of fault-tolerant matching is very fast, we can get the final result by inputting different fault-tolerant parameters and running this version many times. We can also get the result at one time by accurately setting K on the basis of knowing the error distribution of dataset. The optimal value of K is set to ensure that the error data at both the ends can be excised on the basis of as small as possible. The algorithm idea is as follows.

Build a monotone stack. The monotone stack in this version chooses to use another structure similar to a pair. Its first element stores X characters in all the previous reads within the fault tolerance range that match the current suffix. This first element represents X. The second element is no longer a rolling array, but represents the sequence number of the suffix with length X in the

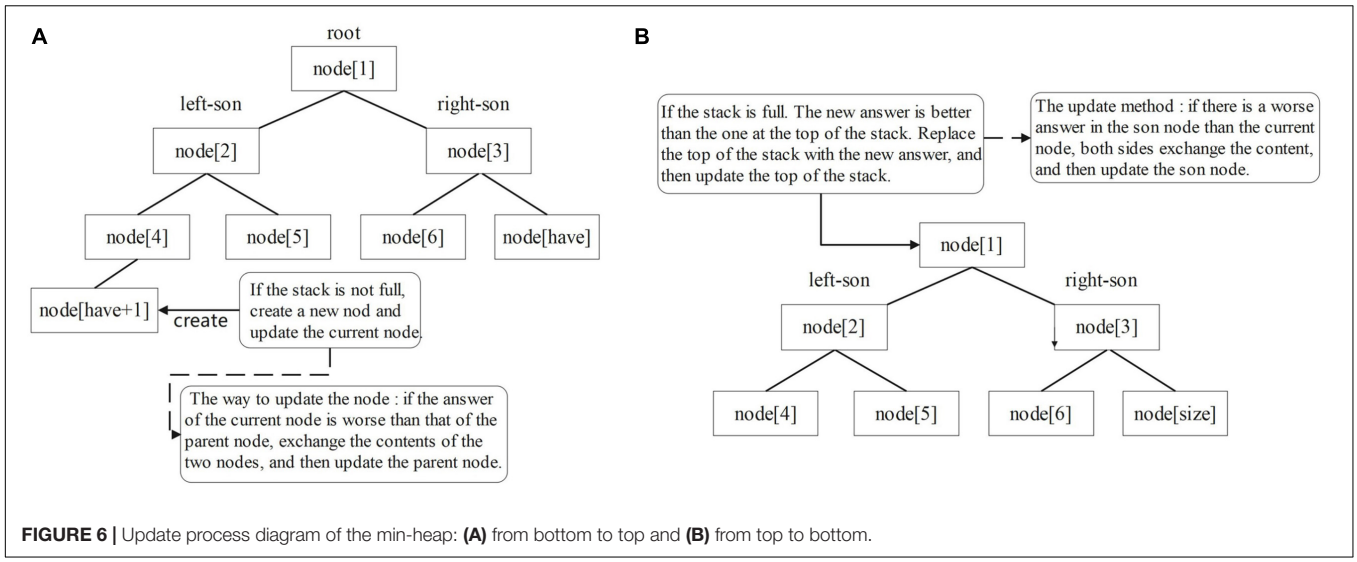


FIGURE 6 | Update process diagram of the min-heap: **(A)** from bottom to top and **(B)** from top to bottom.

suffix array within the fault tolerance range. The final result is obtained by maintaining and optimizing all the second elements in the monotone stack whose first elements are greater than LCP(i). The concept diagram of the monotone stack is shown in **Supplementary Figure 3**.

Maintain the monotonous stack (**Figure 5A**). Different from the previous version, if the first element at the top of the stack is greater than LCP(x), it does not directly take the element at the top of the stack out of the stack, but takes out all the stack elements whose first element is greater than or equal to LCP(x) and selects the best element after comparison to access back to the top of the stack. The element that meets the condition [len(STR) – LCP(x)] is the best and its first item is assigned to LCP(x) and its second item to the second of the optimal element. This process is equivalent to allowing an excision operation at the end of the read, where the wrong part is excised and matched again and the length of the allowed excision is set by the user of the software. After updating the stack, use the current suffix information to create a new pair element. If the length is larger than the top of the stack, put it on the stack. If the length is the same, replace the top of the stack. Of course, to prevent self-loops in the graph constructed from the last obtained overlapping regions information, the stack update is performed by ignoring the suffix of the read itself.

Get overlapping regions information (**Figure 5B**). When traversing the suffix array under the previous version, the result is obtained only when the current suffix is a complete read. In this version, the result is obtained when the first character of the current suffix is the first K characters of the original string, but of course K is determinable and this operation is equivalent to allowing the head of the sequencing data to be cut and the allowed cut length is K. Each cut method is tried, so that each read is compared several times and stores the maximum possible result. Therefore, two auxiliary arrays are needed for multiple result comparisons, the ANS array and the LEN array, where the ANS array stores the ordinal number of its result in the SA array and the LEN array stores the matched lengths. In the process of maintaining the monotonic stack, if the current suffix is the first K suffixes of the string to which it belongs, it is compared with the top element of the stack and the better result is stored. The final ANS array is obtained after several maintenance sessions. After traversing the SA array under this version, the resulting DA[ANS(i)] is the optimal overlapping regions information to be obtained.

Maintain overlapping regions information. To facilitate subsequent assembly software, the SLDMS provide more overlapping regions information for subsequent software. In this version, a top-k overlap suffix set is also provided for each sequence to facilitate subsequent work on genome assembly and parameters are required to set k before running the software. The data structure used to maintain this set of suffixes is the min heap and the top of the heap stores the Kth good overlapping regions (**Supplementary Figure 4**). The reason for choosing this data structure is that the heap can efficiently maintain the largest or smallest value in the heap, so a min-heap of capacity K is created and the top of the heap is the worst quality of the candidate answer and when a new candidate answer is encountered, it only needs to be compared with the top of the heap, which facilitates the update of the answer.

Overlapping regions information maintenance method: create a min-heap for each read to maintain the top-k suffix set, the data in the heap store the position of the corresponding suffix in the SA array, and Len stores the matching length between the suffix and the current reads. Node(1) corresponds to the top of the heap. The larger the size of the heap, the more information is obtained and the longer the corresponding program takes to run. To obtain the required information, the top element of the stack is compared with the data in the min-heap, in addition to maintaining the optimal value of the ans array, when traversing the suffix array and encountering a suffix that can obtain the answer [i-fi(da(i)) $\leq$ k]. If the amount of data in the min-heap is less than k, put the top element of the stack directly into the min-heap and update the heap from the bottom up (**Figure 6A**). This is because the capacity of the heap is K. There is still space in

the heap to store the candidate answers, so the candidate answers can be put directly into the heap and the heap can be updated. Otherwise, compare the top element of the stack with the top element of the heap. If the top element of the heap is better, do not update the elements in the heap; otherwise, use the top element of the stack to replace the top element of the heap and update the heap from top to bottom (**Figure 6B**). This is because there is no space in the heap to store extra candidate answers, so we must choose between the K answers in the heap and the current candidate answers and delete the worst quality answer. Using this method, software users can obtain more overlapping regions information, which makes the edge weight graph based on this overlap information more high-quality information to facilitate subsequent software processing.

Output the Final Overlapping Regions Result

The SLDMS software builds an edge-weight graph with read as the point and overlapping regions information as the edge based on the overlap information after obtaining the overlapping regions information, which contains the overlap position information of the two reads in addition to the length of the overlap for use in obtaining the initial contig. Each path in the graph is stitched into a longer read according to the overlapping regions information, which is the initial contig and the SLDMS software stores all the information in this edge weight graph into a file for the next step of obtaining the contig.

When processing reads without mismatch, the SLDMS only counts the overlapping regions information when a complete read is encountered, a complete read is only encountered once, and the result is not updated again in the subsequent maintenance process, so the output to the file is in the order of the encountered reads, which is a way to update the data processing and the result output synchronously. When processing reads with mismatch, the SLDMS collects data for a read several times in the process of maintaining the monotonic stack, so the stored result information may be updated by subsequent maintenance and the program design idea of separating data processing and result output is adopted. To ensure the accuracy of the results, the maintained results are output in the order of the input data after the program has processed all the data.

Although the output of the two strategies is different, logically they both fill in an array of final results and the i-th element of this array stores the answer of the i-th read. The two strategies differ only in the order of filling in the array, one is filling in the array in order and the other is filling in the array in disorder, but the final goal is to fill in the array completely.

Accuracy Analysis

To prove the universality of the SLDMS software, we write a program to generate the simulated gene sequence randomly, traverse the generated gene sequence many times, and randomly take a substring for simulated gene sequencing. In the first version, the substring is completely correct and in the second version, random errors are generated at both the ends of the substring and used for the SLDMS software input. At the same time, we record the start and end positions of each read and store them in the check file for the final accuracy test.

We consider that the result of each read is correct when it refers to its adjacent read in the gene sequence. After using the SLDMS software to run these input data, the output file and check file are combined to prove the accuracy. According to the results in the output file, we determined whether there was a common part in the interval of the two reads in the check file. If there is a common part, it means that the two reads should be assembled together. We think this is the correct result. We write a program to do this work. First, the program reads the output file and the check file; find each pair of reads and the corresponding overlapping regions information in the output file (if the overlap length is less than 100, it is regarded as invalid data and discarded directly) and then find the corresponding interval in the check file and check the interval. If there is a common part in the two intervals, we can find the corresponding interval in the check file, which is considered the correct result. After calculation, the accuracy of the two versions of the SLDMS is above 99.99%.

The SLDMS software itself and the program code used in the above accuracy proof process are stored on the GitHub website at https://github.com/Dongliang-You/sldms.

RESULTS

The SLDMS and Flye and Canu software were tested on 6 PacBio-HiFi datasets of different sizes and sequenced species and 32 simulated datasets (16 ultrahigh-accuracy datasets and 16 datasets with errors at both ends of the read) on a desktop computer with an Intel Core (TM) i7-9700 CPU (3.00 GHz 8-core processor), 32 GB RAM, and 477 GB hard disk. Due to the limited hardware conditions of the test environment, the oversized dataset was cut where the descriptions of the PacBio-HiFi dataset are shown in **Table 1**, which are the datasets downloaded from the official National Center for Biotechnology Information (NCBI) website. The Sequence Read Runs (SRR) in the description represents the data record of the dataset on the website and the specific data information can be viewed at the official website of the NCBI according to the data record information, which is located at https://www.ncbi.nlm.nih.gov/. The data information of the simulated dataset is shown in **Supplementary Tables 1, 2**.

The timing in the experiment starts when the sequencing data are read and ends when the contig is ready to be acquired. This means that it is necessary to be able to build the edge-weight graph from the overlapping regions information to obtain the contig.

For the SLDMS versions that do not allow for mismatching, six PacBio-HiFi datasets with 16 high-accuracy simulated datasets were chosen for testing and comparison with both the Canu and Flye software. When running the Flye software, the genomeSize parameter is the best estimate according to the uasge.md documentation of the Flye software, which is approximately 1% of the file size. The min-overlap parameter is set to 1,000 and the rest of the parameters are the default parameters. When running the Canu software, the genomeSize parameter is the best guess value entered according to the usage documentation,

TABLE 1 | The PacBio-HiFi dataset used in the experiment and its description.

Dataset	Size(MB)	Description
Z. mays	579	WGS of *Zea mays* "B73" using PacBio HiFi Sequencing(SRR11606869)
*E_coli*_K12	1,914	WGS of *E. coli* K12 with PacBio HiFi DNA sheared on Megaruptor to 20 kb(SRR10971019)
F. × *ananassa*_part1	1,131	WGS of *Fragaria* × *ananassa* Royal Royce using PacBio HiFi Sequencing (SRR11606867)
F. × *ananassa*_part2	2,190	WGS of *Fragaria* × *ananassa* "Royal Royce" using PacBio HiFi Sequencing(SRR11606867)
*M. musculus*_part1	1,567	WGS of *Mus musculus* "C57/BL6J" using PacBio HiFi Sequencing(SRR11606870)
*M. musculus*_part2	2,760	WGS of *Mus musculus* "C57/BL6J" using PacBio HiFi Sequencing(SRR11606870)

which is the same as the genomeSize parameter of the Flye software and the rest of the parameters are default values without any restrictions.

The time required for different software programs to run the PacBio-HiFi datasets to find the overlapping regions is given in **Figure 7** and **Supplementary Table 3**. The SLDMS software ran faster than Flye on all the datasets, faster than Canu on most datasets, and only slightly slower than the Canu software on the *M. musculus*_part1 dataset due to the nature of the algorithm of the Canu software, which makes it potentially efficient at running certain datasets. This result suggests that the SLDMS software runs more efficiently than Canu and Flye for sequence alignment on most PacBio-HiFi datasets.

The time required for different software programs to run ultrahigh-accuracy simulation datasets to find overlapping regions is given in **Figures 8**, **9** and **Supplementary Table 4**. As shown in the two line graphs in the first row of **Figure 8**, these are the results of running the ultrahigh accuracy simulation dataset with different average lengths for the same data volume of 5,000, 10,000, 15,000, and 20,000 reads. As shown in the two line graphs in the second row of **Figure 8**, these are the results of running the ultrahigh-accuracy simulation dataset with different data volumes for the same average data lengths of 5,000, 10,000, 15,000, and 20,000. As seen from the graphs in all the runs, the SLDMS software has a shorter run time than the other two, which suggests that in most cases, it is a good choice to use the SLDMS software to find the overlapping regions information.

For the SLDMS versions that allow for mismatching, there are no real data available that match the conditions for this version to run, so it was only possible to test this version using simulated data. Each read in the simulated data was divided into three parts in order, with the first and third parts being 80% accurate, the second part being 100% accurate, and the second part being at least half the length of the read. The Canu and Flye software selected PacBio-Raw for the data type during the tests and the rest of the parameter settings were the same as the previous version.

The time required for different software programs to run the simulation datasets with errors to find overlapping regions is given in **Figure 9**, in which row 1 is an experiment with the average length of the data as the variable and row 2 is an experiment with the amount of data as the variable. As seen from the graph, when running these datasets, both when running datasets with different average lengths for the same amount of data and when running datasets with different amounts of data for the same average length, the SLDMS runs much faster than the other two. This shows that using the SLDMS to run this dataset with errors only at both the ends of the data is much better and takes less time than using the other two software tools.

In the experiment, we tested the performance of the different software programs on different datasets. The SLDMS software

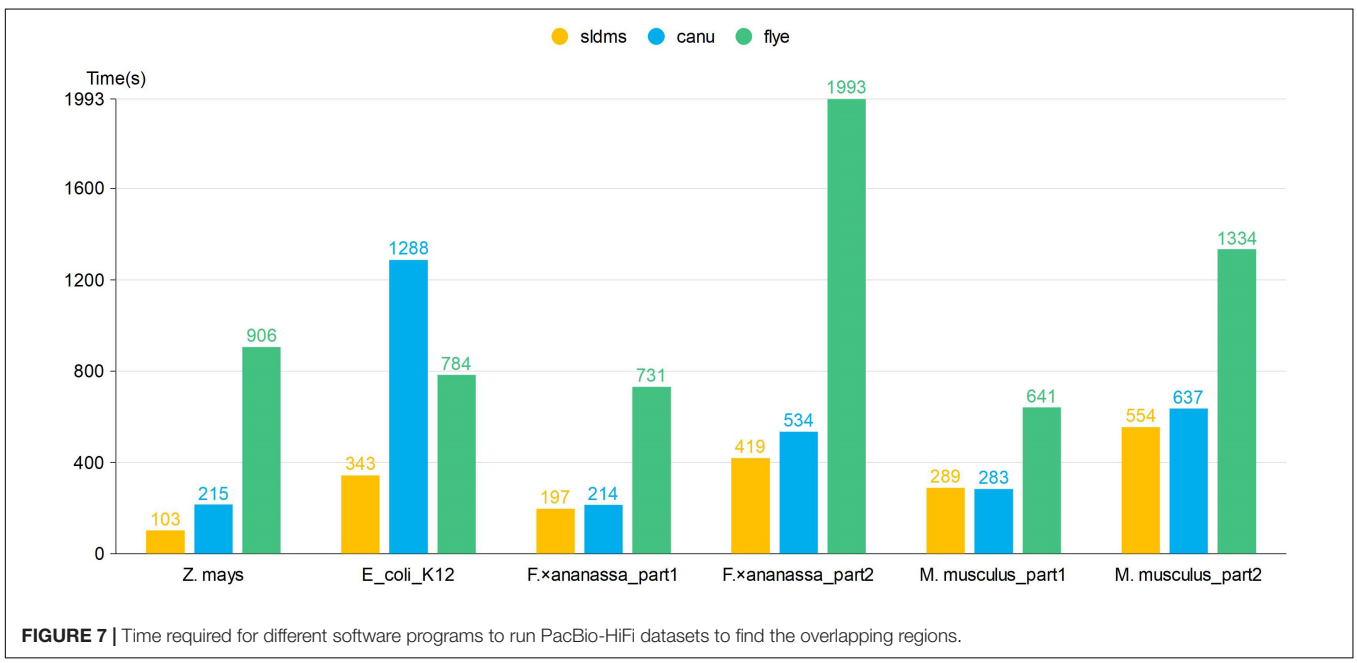


FIGURE 7 | Time required for different software programs to run PacBio-HiFi datasets to find the overlapping regions.

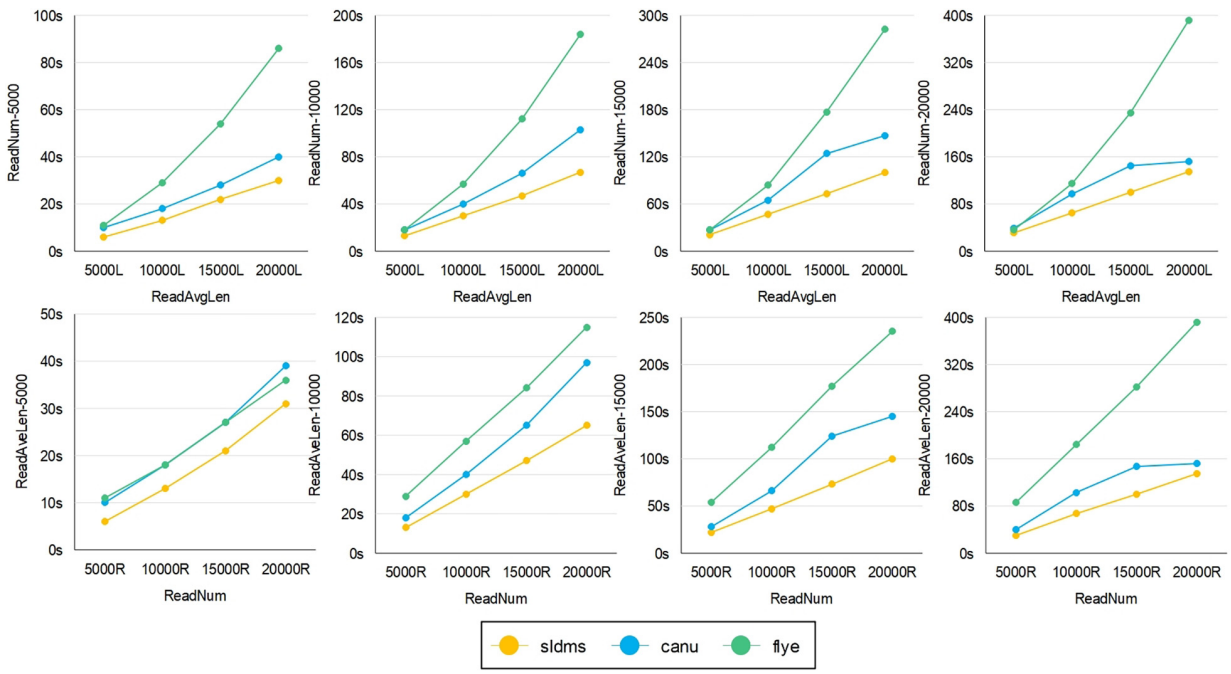


FIGURE 8 | Time required for different software programs to run ultrahigh-accuracy simulation datasets to find overlapping regions.

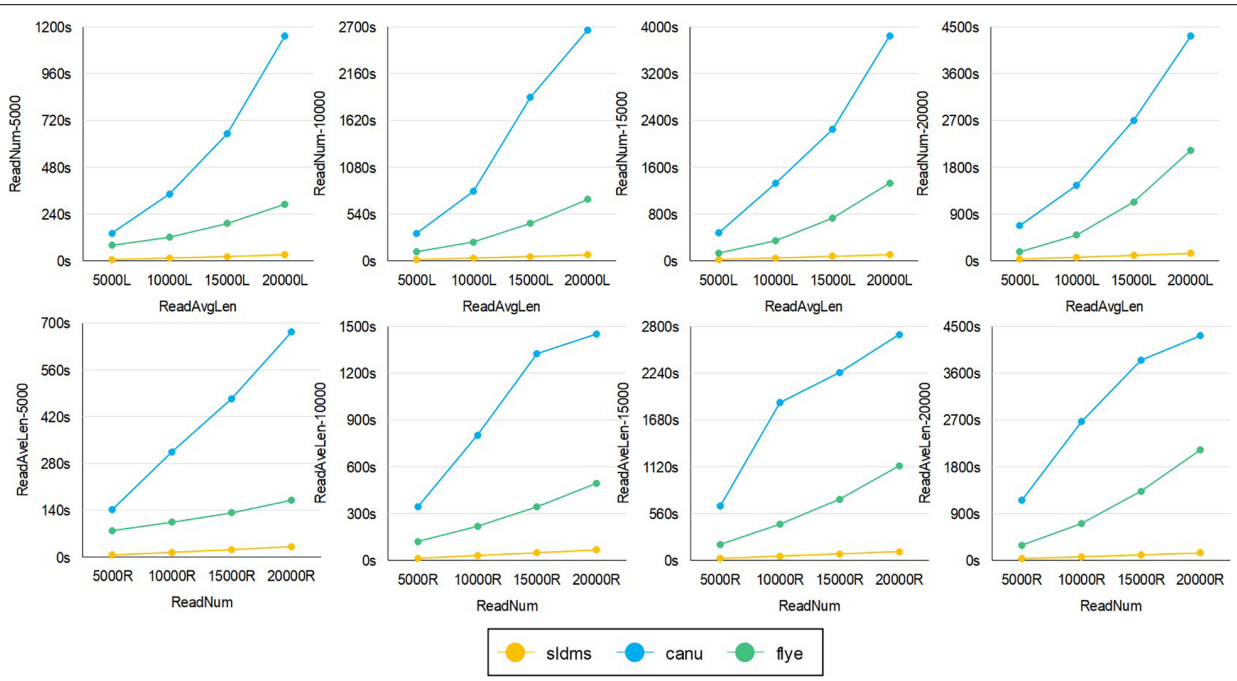


FIGURE 9 | The time required for different software to run simulation datasets with errors to find overlapping regions.

was stable and efficient in obtaining overlapping regions information for the various test data. The running time of the SLDMS is only related to the size of the input dataset and does not fluctuate depending on differences in the accuracy of the data. This means that the SLDMS is suitable for processing a wide variety of data without the worry that the SLDMS will take a particularly long time to process a particular type of data.

The SLDMS software uses the gsufsort algorithm to calculate the three arrays of SA, LCP, and DA information, which takes a significant amount of time. If new methods are developed to

obtain this information more quickly, the SLDMS software will be more efficient.

DISCUSSION

The main contribution of the proposed method SLDMS for extracting information about the overlapping regions between sequences based on suffix arrays and monotonic stacks is to substantially improve the time efficiency of calculating the overlapping regions. Obtaining overlapping regions information is useful in many bioinformatics applications. As the price of sequencing technology decreases and genome sequencing technology develops, it becomes easier to obtain sequencing data with a wide range of characteristics and higher accuracy. When assembling these sequencing data, it is essential to efficiently extract overlapping regions information between sequences to provide a more favorable environment for subsequent genome assembly work.

The Flye software uses the ABruijn algorithm, which combines the OLC and DBG algorithms to generate its own unique ABG graph, obtains the overlapping regions information of nodes in the graph, and then processes the ABG graph to obtain contigs. In this process, it takes considerable time to process the graph, so we can see from the experimental results that the Flye software takes about twice as long to run as the SLDMS on almost all the datasets.

The Canu software uses the MHAP algorithm to aggregate reads with the same k-mer for error correction and pruning and then obtains the overlapping regions information. When dealing with high-accuracy data without error correction, the SLDMS is around 20% faster than Canu. With respect to error correction, the Canu software is very slow, regardless of the error characteristics of the data, indicating that the Canu software does not take full advantage of the error characteristics of the data. The SLDMS does a very good job in this respect and in some data with most sequencing errors occur at both the ends of the sequencing data, the SLDMS runs at many times the speed of Canu.

The SLDMS obtains the three arrays of SA, LCP, and DA by processing the input data and quickly finds the overlapping regions information in the input data with the help of these three arrays and the monotonic stack. The calculation speed of the overlapping regions information is improved. The experimental results show that compared with the other two kinds of software, the SLDMS have faster speed in the calculation of the overlapping regions and with the help of the SA array containing all the suffixes, it also has the ability of data fault tolerance by cutting the suffixes. This shows that the SLDMS is very efficient as an approach based on suffix arrays and monotonic stacks and that suffix arrays are still the ideal data structure for solving the problem of calculating overlapping regions of gene sequences.

ABOUT THE SLDMS

The SLDMS is an open source software tool developed in C and can only be run on Linux systems. The link to the project is (https://github.com/Dongliang-You/sldms). Permission from the author is required before use for non-academic purposes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://bioinfor.nefu.edu.cn/chenyu/sldms_web/, sldms.

AUTHOR CONTRIBUTIONS

YC: conceptualization. DY: software. GW: writing – original draft. TZ: writing – review and editing. All the authors conducted experiments and read and agreed to the published version of the manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (61771165, 62072095, and 62172087), the National Key R&D Program of China (2021YFC2100100), the Fundamental Research Funds for the Central Universities (2572021BH01), and the Innovation Project of State Key Laboratory of Tree Genetics and Breeding (Northeast Forestry University) (2019A04).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.813036/full#supplementary-material

REFERENCES

Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S., et al. (2008). Consensus generation and variant detection by Celera Assembler. *Bioinformatics* 2008:btn074. doi: 10.1093/bioinformatics/btn074

Fischer, J. (2010). Wee LCP. *Inf. Process Lett.* 110, 317–320. doi: 10.1016/j.ipl.2010.02.010

Hon, T., Mars, K., Young, G., Tsai, Y. C., and Rank, D. R. (2020). Highly accurate long-read hifi sequencing data for five complex genomes. *Sci. Data* 7:077180. doi: 10.1101/2020.05.04.077180

Huang, X. (1992). A contig assembly program based on sensitive detection of fragment overlaps. *Genomics* 14, 18–25. doi: 10.1016/S0888-7543(05)80277-0

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., et al. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 693–700. doi: 10.1038/nbt.2280

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Leena, S., and Eric, R. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* 30, 3506–3514. doi: 10.1093/bioinformatics/btu538

Li, Z. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief. Funct. Genom.* 11, 25–37. doi: 10.1093/bfgp/elr035

Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., Chaisson, M., and Pevzner, P. A. (2016). Assembly of long error-prone reads using de Bruijn graphs. *Pro. Natl. Acad. Sci. U S A.* 113:E8396. doi: 10.1073/pnas.1604560113

Louza, F. A., Gog, S., and Telles, G. P. (2017). Inducing enhanced suffix arrays for string collections. *Theor. Comput. Sci.* 678, 22–39. doi: 10.1016/j.tcs.2017.03.039

Louza, F. A., Telles, G. P., Gog, S., Prezza, N., and Rosone, G. (2020). gsufsort: constructing suffix arrays, LCP arrays and BWTs for string collections. *Algorithms Mol. Biol.* 15:18. doi: 10.1186/s13015-020-00177-y

Mahmoud, M., Zywicki, M., Twardowski, T., and Karlowski, W. M. (2017). Efficiency of pacbio long read correction by 2nd generation illumina sequencing. *Genomics* 2017:S0888754317301660. doi: 10.1016/j.ygeno.2017.12.011

Manber, U., and Myers, G. (1993). Suffix Arrays: A New Method for On-Line String Searches. *SIAM J. Comput.* 22, 935–948. doi: 10.1137/0222058

Muthukrishnan, S. (2002). Efficient algorithms for document retrieval problems. *Proc. SODA* 2002, 657–666.

Schatz, M. (2006). *Celera Assembler Celera Assembler Overview*. Honolulu, HI: University Of Hawaii.

Wang, J., Chen, S., Dong, L., and Wang, G. (2020). Chtkc: a robust and efficient k-mer counting algorithm based on a lock-free chaining hash table. *Brief. Bioinformat.* 22:bbaa063. doi: 10.1093/bib/bbaa063

Xie, Y., Wu, G., Tang, J., Luo, R., Jordan, P., Liu, S., et al. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 12:1660. doi: 10.1093/bioinformatics/btu077

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/fpls.2021.830042

Efficient Generation of RNA Secondary Structure Prediction Algorithm Under PAR Framework

Haihe Shi and Xiaoqian Jing*

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China

Prediction of RNA secondary structure is an important part of bioinformatics genomics research. Mastering RNA secondary structure can help us to better analyze protein synthesis, cell differentiation, metabolism, and genetic processes and thus reveal the genetic laws of organisms. Comparative sequence analysis, support vector machine, centroid method, and other algorithms in RNA secondary structure prediction algorithms often use dynamic programming algorithm to predict RNA secondary structure because of their huge time and space consumption and complex data structure. In this article, the domain of RNA secondary structure prediction algorithm based on dynamic programming (DP-SSP) is analyzed in depth, and the domain features are modeled. According to the generative programming method, the DP-SSP algorithm components are interactively designed. With the support of PAR platform, the DP-SSP algorithm component library is formally realized. Finally, the concrete algorithm is generated through component assembly, which improves the efficiency and reliability of algorithm development.

Keywords: algorithm component library, feature modeling, PAR platform, RNA secondary structure prediction algorithm, generative programming

OPEN ACCESS

Edited by:
Weihua Pan, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences (CAAS), China

Reviewed by:
Mingzhi Liao, Northwest A&F University, China
Jingjing Liu, Nanjing University of Aeronautics and Astronautics, China

****Correspondence:***
Haihe Shi
haiheshi@jxnu.edu.cn

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *06 December 2021*
Accepted: *16 December 2021*
Published: *21 January 2022*

Citation:
Shi H and Jing X (2022) Efficient Generation of RNA Secondary Structure Prediction Algorithm Under PAR Framework. Front. Plant Sci. 12:830042. doi: 10.3389/fpls.2021.830042

INTRODUCTION

RNA is one of the important macromolecules in organisms and plays an important role in protein synthesis, cell differentiation, metabolism, and genetic process. Especially in HIV and other viruses, genetic information is carried directly by RNA rather than DNA (Jiang et al., 2002). To better analyze the role of RNA molecules in the life process, it is necessary to understand the molecular structure of RNA. The molecular structure of RNA can be divided into three levels (Peter and Rdf, 2000; Yang, 2013), namely, primary structure, secondary structure, and tertiary structure. Primary structure refers to a sequence composed of four bases (A, U, C, and G) of RNA. The secondary structure is a two-dimensional planar structure formed by pairing partial bases on the basis of the primary structure, and tertiary structure is a three-dimensional structure formed by folding on the basis of secondary structure (Yu, 2009; Huang et al., 2014). It has been proved that RNA tertiary structure plays a decisive role in RNA function, but the prediction of RNA tertiary structure largely depends on the prediction of secondary structure (Shuaimin, 2019; Bowen, 2021; Zhaokui and Yuanchao, 2021). Therefore, RNA secondary structure prediction is an important and hot issue in the field of RNA research. Since 1980s, RNA secondary structure prediction algorithms have emerged one after another, which can be roughly divided into two categories: one is comparative

sequence analysis method, such as Stochastic Context-free Grammar (SCFG) model (Dowell and Eddy, 2004) and Covariance Model (CM) (Eddy and Durbin, 1994); and the other is the prediction method based on dynamic programming. Typical examples include the maximum base pairs algorithm proposed by Nussinov of Weizmann Institute of Science (Nussinov et al., 1978); the minimum free energy algorithm designed by Zuker of Division of Biological Sciences, National Research Council of Canada (Zuker and Stiegler, 1981); the partition function algorithm established by McCaskill of Max-Planck lnstitut fur Biophysikalische Chemie (McCaskill, 1990); and the helix-based prediction algorithm studied by the team of Harbin Institute of Technology (Xia, 2008). Because the former requires a large number of homologous RNA sequences in advance to predict, the time and space complexity of the algorithm is particularly high, and long sequences cannot be analyzed well, people often use dynamic programming model to predict RNA secondary structure.

Most of the existing RNA secondary structure prediction algorithms based on dynamic programming focus on the optimization of specific steps of specific algorithms, such as accelerating the execution speed of algorithms through parallelization technology, and the optimization results will have different effects on different sequences. In addition, the complexity of the RNA secondary structure prediction problem and the diversity of algorithm design strategies make the reliability of the algorithm development difficult to guarantee and the development cost high, which is not convenient for researchers to study.

In this article, the domain of RNA secondary structure prediction algorithm based on dynamic programming (DP-SSP) is regarded as a specific domain. Through in-depth analysis of the DP-SSP domain, the commonness and differences of the domain are extracted, and the generic algorithm component library in the DP-SSP domain is designed by combining domain engineering, feature modeling, formal method PAR, and other related technologies. Then, the abstract generic programming language Apla is used to formalize the implementation. Finally, using the program conversion system of PAR platform, the components of the component library are manually assembled according to the configuration knowledge and generate a specific algorithm, thereby improving the development efficiency of the RNA secondary structure prediction algorithm and ensuring the reliability of the algorithm development.

The section "Materials and Methods" introduces related theories and methods of domain engineering, generative programming (GP), formal method PAR, and so on. The section "Domain Analysis and Abstraction of RNA Secondary Structure Prediction Algorithm" analyzes the domain of RNA secondary structure prediction algorithm domain, establishes the domain feature model of DP-SSP, and implements it by using the generic abstract programming language Apla in the formal method PAR, finally establishing a high abstract component library based on abstract data type (ADT). The section "Results" shows the process of developing Zuker algorithm based on the component library and gives the experimental results of the algorithm and the comparison with other algorithms. Finally, in the section "Discussion," the full text is summarized and prospected.

MATERIALS AND METHODS

PAR Framework

PAR framework (Jinyun, 1993, 1998; Xue, 1997, 2015; Shi and Xue, 2009; Wang and Xue, 2009) includes a practical formal method and corresponding support platform. PAR platform includes requirements design language SNL, algorithmic modeling language Radl, abstract programming language Apla, and a series of conversion rules and automatic conversion tools. PAR focuses on the design and implementation of algorithms, supports most of the current mainstream algorithm design technologies, includes a new development strategy of loop invariant, and implements the distributed transaction processing system and relational database mechanism. By using PAR method to develop algorithms, we can have a deeper understanding of the algorithm and avoid the difficulty of selecting the design method. The agile generic mechanism is one of the important features of PAR. Regardless of the data type, data value, calculation operation, or user-defined ADT, it can be a generic parameter. Apla can directly use ADTs and abstract processes programming. It not only has the advantages of concise mathematical language but also has the characteristics of expressing unambiguity. Due to its high abstraction, Apla is very suitable for describing abstract algorithm programs. The following describes the implementation mechanism and constraint mechanism of Apla generics:

(1) Apla includes the concepts of type variables, type domains, operation variables, operation domains, ADT variables, and ADT domains. It uses sometype, someaction, and someadt to represent type domain, operation domain, and ADT domain, respectively, and implements the parametric operation of type, function, program, and custom ADT. In the instantiation process, actual parameters that meet the relevant attribute conditions can be passed in to implement different program units.
(2) Generic constraints describe the types and composition of generic parameters in detail. The implementation of generic constraint mechanism can greatly improve the reliability of generic programming, which is one of the necessary conditions for the real implementation of generic programming. PAR platform implements relevant generic constraints on generic parameters such as basic data type, ADT type and subroutine type, and proposes corresponding constraint description, matching, and detection mechanisms, which are still being improved.

In addition, PAR platform also supports the transformation of Apla into executable high-level programming languages such as C++, Java, C#, and Delphi, which has a good support for the development of components. PAR has established two ways to formalize the way to develop programs, and its platform architecture is shown in **Figure 1**. The first is that for quantitative problems, the PAR method can convert the SNL requirement model into the Radl specification model, then into the Radl

algorithm model, and further into the Apla abstract program model, and finally into a high-level language program that can be run directly. The second way is that for nonquantitative problems, we can manually design the Apla program directly through the SNL requirement model, supplemented by the corresponding formal proof, and then convert the Apla program into an executable program.

Domain Engineering

Domain engineering (Neighbors, 1989; Li et al., 1999; Hu and Wei, 2008) is the basic process of software reuse, and its purpose is to acquire and use reusable resources in a specific domain to develop high-quality software efficiently and at low cost. Domain engineering mainly analyzes, designs, and implements the domain. Domain analysis includes a series of activities such as system scope definition, domain requirement definition, and related terminology analysis, and finally, the results are reflected in the domain model. The domain design completed the architecture design of the system family in the domain, identified the corresponding functions and related constraints, and made plans for the subsequent implementation process. Domain implementation uses appropriate technology to complete the development of reusable resources such as architecture and components. These three stages adopt the idea of gradual refinement in practical application and modify and improve the completed results at any time according to changes in requirements.

Domain analysis is the basis of domain engineering. The generated domain model affects the quality of subsequent work. It usually adopts a combination of top-down and bottom-up analysis to repeat domain analysis activities. Top-down analysis takes into account the needs of future systems in the domain, while the bottom-up analysis mainly considers the existing systems and the reusable resources accumulated by previous development. After years of efforts by researchers, many methods have been used in domain analysis, such as organization domain modeling (ODM), object-oriented analysis (OOA), and feature-oriented domain model (FODM) (Wartrk, 1999; Chastek et al., 2001). To carry out domain analysis activities efficiently, Zhang and Mei (2003) put forward a feature modeling method FODM, which focuses on the characteristics of services, functions, and behaviors in the domain and discusses the presentation form of a feature model and its detailed modeling process.

Generative Programming

Generative programming (Czarnecki et al., 2000; Fan and Zhang, 2005) is a new type of software paradigm, which accords with the idea of software reuse. It uses components and makes software products in an automated way, which is of great significance to solve the "software crisis." There are two steps to implement GP. The first is to change the current software development mode into the development of software system families and develop generators to automatically assemble components. GP is an example of domain engineering application, which needs to make full use of existing domain knowledge to complete component development in corresponding domain. Finally, generator is used to develop new software in the field by means of component assembly, without the need to follow the steps of software engineering to start programming from scratch.

The purpose of GP is to realize the production automation of components and applications, and the key part of GP is to establish domain models for system families. The generative domain model consists of three parts: problem space, solution space, and corresponding configuration knowledge. The problem space mainly includes the needs of application programmers and users for the system, and the requirements are generally described by the concepts and characteristics of the program; the solution space includes the relevant components that can solve the demand problem and the combination mode of each component, and it is required to achieve the maximum composability, and the redundancy between the combinations must be minimized, and the highest reusability of the components must be achieved as much as possible. The configuration knowledge is the mapping relationship between the problem space and the solution space, which avoids the occurrence of illegal feature combination and sets the default parameters and rules of features. Configuration knowledge is the mapping relationship between the problem space and the solution space, which avoids the illegal feature combination and sets the default parameters and rules of features.

Concepts Related to RNA Secondary Structure Prediction

RNA sequence: RNA sequence refers to the primary structure $S = S_1S_2S_3 \ldots S_n$ of RNA, where $S_i \in (G, C, U, A)$, $1 \leq i \leq n$.

Base pair: If $(S_i, S_j) \in \{GC, AU, UG\}$, $1 \leq i < j \leq n$, then (S_i, S_j) constitutes a base pair.

RNA secondary structure: A set of base pairs.

RNA secondary structure prediction: Input an RNA sequence, predict the secondary structure through some algorithm, and follow the following rules in the prediction process:

(1) A base cannot be paired with two or more bases at the same time. That is, there are base pairs (S_i, S_j) and (S_k, S_l). If $i = k$, then $j = l$.
(2) If $i < g < j < h$ or $g < i < h < j$, (S_i, S_j) and (S_g, S_h) cannot appear in the secondary structure, that is, pseudoknot cannot appear in the secondary structure.
(3) If there are base pairs (S_i, S_j), then $|j - i| \geq 4$, that is, the length of hairpin loop structure should be ≥ 4.

Hairpin loop: A structure consisting of one base pair (S_i, S_j) and all unpaired bases closed by it.

Stem: A structure composed of two adjacent base pairs (S_i, S_j), (S_{i+1}, S_{j-1}).

Bulge loop: It is composed of two base pairs (S_i, S_j) and (S_k, Sl), and the two base pairs are adjacent at one end and not adjacent at the other end ($k = i + 1, k < l < j - 1$ or $l = j - 1, i + 1 < k < l$).

Interior loop: It is composed of two base pairs (S_i, S_j) and (S_k, S_l), and the two base pairs are not adjacent at both ends $(i + 1 < k < l < j = 1)$.

Multibranched loop: A structure closed by three or more base pairs.

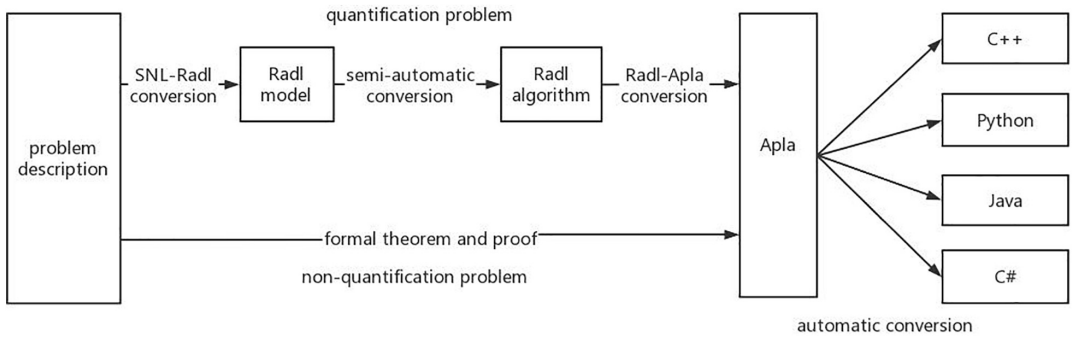


FIGURE 1 | The development algorithm flow of PAR method.

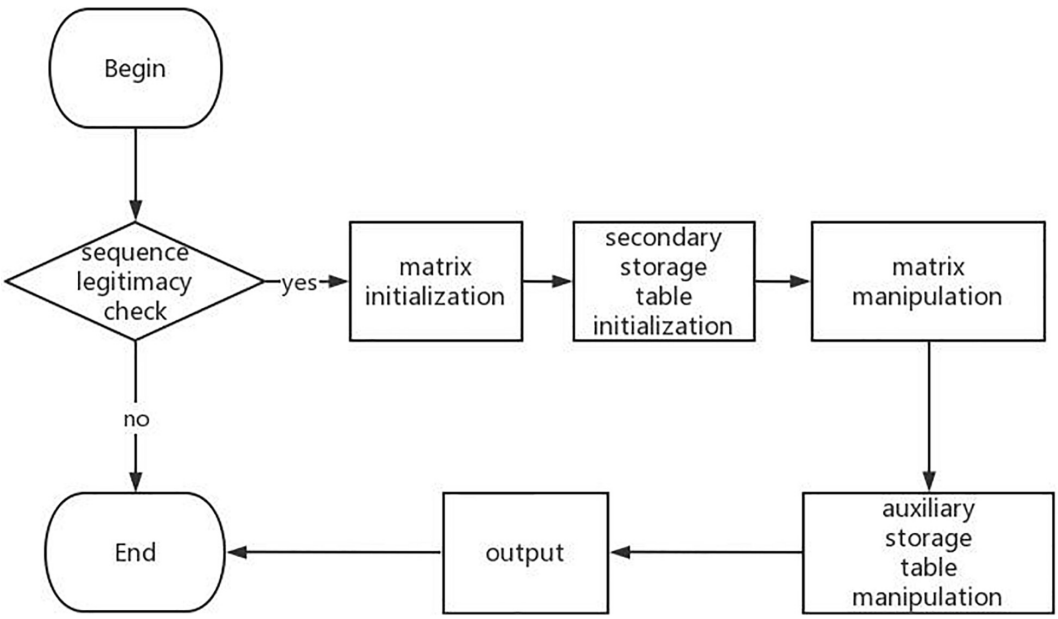


FIGURE 2 | Flowchart of RNA secondary structure prediction algorithm.

Domain Analysis and Abstraction of RNA Secondary Structure Prediction Algorithm

Here, we briefly analyze the core ideas of three typical dynamic programming algorithms.

(1) Nussinov algorithm

Given a sequence s, when the i-th base S_i is paired with the j-th base S_j in S, $\theta(i, j) = 1$, otherwise $\theta(i, j) = 0$. M(i, j) is used to represent the maximum matching base logarithm on the subsequence S_{ij}, and its value can be iterated by the following formula:

$$M(i, j) = \text{Max}\begin{cases} M(i+1, j) \\ M(i, j-1) \\ M(i+1, j-1) + \delta(i, j) \\ \text{Max}(M(i.k) + M(k+1, j)) \end{cases} \quad (1)$$

in which $i \leq k \leq j$, when $i = 1, 2, 3, \ldots, n$, $M(i, i) = 0$. When $i = 2, 3, 4, \ldots, n$, $M(i-1, i) = 0$.

The four terms in formula (1) correspond to the possible pairing between the i-th base and the j-th base in the sequence:

① S_i does not participate in base pairing, then the maximum number of base pairing in interval (i, j) is equal to the maximum number of base pairing in interval $(i+1, j)$.

② S_j does not participate in base pairing, then the maximum number of base pairing in interval (i, j) is equal to the maximum number of base pairing in interval $(i, j=1)$.

③ S_i is paired with S_j, and the maximum number of base pairs in interval (i, j) is equal to the maximum number of base pairs in interval $(i+1, j=1)$ plus 1.

④ S_i is paired with base S_k in interval (i, j), then the maximum number of base pairings in interval (i, j) is equal to the number of pairings in interval (i, k) plus the pairing number of interval (k + 1, j). Take k = i, k = i + 1, k = i + 2,..., j in turn, then the maximum number of base pairings in interval (i, j) is equal to the one with the largest number.

Each iteration takes the maximum of the above four cases, and the value of M(1, n) is the maximum number of base pairs. The secondary structure of sequence s can be obtained by backtracking from W(1, n).

(2) Zuker algorithm

Give a sequence s, fragment S_{ij} represents the subsequence from the i-th base to the j-th base in the s sequence, where $1 \leq i \leq j \leq n$. W(i, j) is the minimum free energy of all RNA secondary structures composed of subsequence S_{ij} (whether S_i and S_j are paired or not), V(i, j) is the minimum free energy of RNA secondary structure formed by pairing S_i and S_j. The calculation process of W(i, j) and V(i, j) is shown in formulas (2)–(6) as follows:

$$W(i, i) = 0 \tag{2}$$

$$W(i, j) = V(i, j) = \text{if } j - i < 4 \tag{3}$$

$$V(i, j) = \text{if i and j are not paired.} \tag{4}$$

$$V(i, j) = \begin{cases} E_1 = E_H(i, j) \\ E_2 = Es(i, j; i + 1, j - 1) + V(i + 1.j - 1) \\ E_3 = \min\{E_L(i, j; i', j') + V(i', j')\}, \\ \quad i < i' < j' < j, (i' - i) + (j - j') > 2 \\ E_4 = \min\{W_m(i + 1, k) + W_m(k + 1, j - 1)\}, \\ \quad i < k < j - 1 \end{cases} \tag{5}$$

$$W(i, j) = \min\{V(i, j)W(i + 1, j), W(i, j-1), \min\{W(i, k) + W(k + 1, j)\}, i < k < j-1\}, j-i \geq 4 \tag{6}$$

$E_H(i, j)$ in formula (5) represents the minimum free energy corresponding to the hairpin loop structure formed by pairing base S_i and base S_j, E_s represents the minimum free energy corresponding to the stem structure formed by the pairing of base S_i and base S_j, E_L represents the minimum free energy corresponding to the interior-loop or bulge-loop structure formed by the pairing of base S_i and base S_j, and E_4 represents the minimum free energy corresponding to the multibranched loop structure formed by the pairing of base S_i and base S_j.

By using formula (6) to iterate continuously, w(1, n) is the minimum free energy of sequence s. The secondary structure of sequence s can be obtained by backtracking from W(1, n).

(3) Helix-based algorithm

Given an RNA sequence s, all possible stem regions were calculated by using the INN-HB energy model. $E_{i,j}$ represents the minimum free energy of the subsequence S_{ij}, and its value can be obtained by using formula (7).

$$E_{i,j} = \begin{cases} E_{init} \\ E(H_{i,j,k} + E_{i+k,j-k}) \\ min[E_{i,k} + E_{k+1,j}] \end{cases} \tag{7}$$

Equation (7) corresponds to three situations: ① If $j - i < 8$, then $E_{i,j} = E_{init} = 0$; ② Otherwise, search the stem regions, if there is a stem region $H_{i,j,k}$ starting with the i-th base and ending with the j-th base, and $H_{i,j,k} + E_{i+k,j-k} < E_{i,j}$, then $E_{i,j} = E(H_{i,j,k} + E_{i+k,j-k})$. ③ For each k($i < k < j$), if $E_{i,k} + E_{k+1,j} < E_{i,j}$, then $E_{i,j} = E_{i,k} + E_{k+1,j}$. When $E_{1,n}$ is calculated, the minimum free energy of RNA is found, and the secondary structure with the minimum free energy can be found by backtracking.

By further analyzing a large number of RNA secondary structure prediction algorithms based on dynamic programming, we can know that the process of RNA DP-SSP can be summarized as shown in **Figure 2**.

Next, we analyze the DP-SSP domain with FODM method, and consider the Service, Function, and Behavior in the DP-SSP domain to build a feature model. The scope of the algorithm field is limited to an algorithm form with dynamic programming as the main strategy and RNA secondary structure prediction as the main prediction method in the field of RNA function analysis. RNA secondary structure prediction is the core service in this field. The user-defined RNA secondary structure prediction algorithm is realized by controlling the prediction mode, the execution priority, and the combination mode between algorithm features in the process of RNA secondary structure prediction.

(1) Sequence validity check (seq_check) is a preprocessing operation that must be performed on the input sequence before each algorithm runs, which is regarded as a common function.

(2) All algorithms in this field need to build matrices and auxiliary storage tables to store data and also need to operate matrices and auxiliary storage tables, so matrix manipulating (matrix_mani) and auxiliary storage table manipulating (auxiliary_storage_mani) are the required components. Further analysis shows that for auxiliary_storage_mani, the auxiliary storage mode (auxiliary_mode) is its behavior characteristic, and there are mainly the following ways: auxiliary matrix (matrix_op), auxiliary stem pool (stem pool_op), and auxiliary free energy parameter (free energy_op).

(3) In this field, the dynamic programming idea is used to predict the RNA secondary structure. Different RNA secondary structure prediction algorithms can be obtained by selecting different dynamic programming strategies. Therefore, dynamic programming pattern selection is regarded as a common component in this field. There are four behavior patterns: based on maximizes base pairs (Nussinov_op), based on minimizes free energy (Zuker_op), based on helix-based (helix_op), and based on partition function (partition function_op).

(4) Output function (result_op) as a common function in the field, it has two output modes: matching logarithmic output (pairing number_op) and matching interval output (pairing interval_op). Among them, the matching interval needs backtracking (backtrack) and remembering the source of elements (element source). Therefore, tracing back and remembering the source of elements are optional components.

The established feature model is shown in **Figure 3**.

A complete domain feature model also needs the interaction between features. In the feature model, the interaction between features is mainly reflected by the constraints and dependencies between features. Therefore, for the feature model established above, we design the feature interaction model in DP-SSP domain.

Through the establishment of DP-SSP feature model, it is analyzed that the algorithm mainly includes three change process features: matrix_mani, dp_mode, and output. In addition, the input of the algorithm in this field is gene sequence, so it is necessary to check the legality of the sequence information before the algorithm is executed. Therefore, the main components in this field are seq_check component, matrix_mani component, dp_mode component, and output component. Other features in the feature model are used as auxiliary components, and the interaction model of components is established according to the dependencies between components, as shown in **Figure 4**.

Among them, the nodes connected by the solid line represent the basic features that must be contained in the DP-SSP field. The direction indicated by the arrow indicates the execution priority of the four features from high to low. The arrow with dotted line represents the associated operations required in the algorithm assembly process. For example, when we choose dynamic programming mode or perform matrix operations, we need to use the information of auxiliary storage table operations. The dotted line indicates the interaction between the two features in the process of algorithm execution. For example, matrix operation features need to be used in result output or backtracking.

Here, we further analyze the feature model and algorithm component interaction model in the above DP-SSP field and package them into two ADT components and an RNA secondary structure prediction algorithm component. With the advantages of high abstraction, good support for ADT, easy formal derivation, and correctness verification of Apla program, the DP-SSP model is formally designed and implemented based on Apla code.

(1) Matrix-type component

define ADT matrix_mani (sometype elem);
type matrix_mani = private;
var:
matrix:list(array[0. . .n,elem])
aux:auxiliary_storage_mani
procedure apply_memory (m: matrix _ mani;length:integer);
procedure init_matrix(proc initial(); m:matrix_mani;matrix: list (array[0. . .n, elem]));
procedure setvalue (m:matrix_mani; matrix:list(array [0. . .n,elem]); i:integer; j:integer; aux:auxiliary_storage_mani);
function getvalue(m: matrix_mani; matri x:list(array [0. . .n,elem]);i:integer;j:integer):elem
function the_last_element():elem;
procedure traceback(m: matrix_mani; matrix:list(array [0. . .n,elem]); aux: auxiliary _ storage_ mani);
procedure output(m:matrix_mani; matrix: list(array [0. . .n,elem]);aux:auxiliary_storage_mani);
enddef;

The generic ADT named *matrix_mani* contains a type parameter *elem*, which can accept character type or other types of data. *Type matrix_mani = private* is the storage space description, which is used to indicate that the storage space used by the custom ADT is private. *Apply_memory (m:matrix_mani;length:integer)* is used to dynamically allocate memory space for *matrix_mami* according to the value of integer variable length; *init_matrix(proc initial(); m:matrix_mani; matrix: list (array[0. . .n,elem]))* has a generic process parameter. Different process parameters can be passed in to instantiate different matrices. The functions of *Procedure setvalue (m:matrix_mani;matrix:list (array[0. . .n,elem]); i:integer; j: integer; aux: auxiliary_ storage_mani)* and *getvalue (m:matrix_mani; matrix: list (array[0. . .n,elem]); i:integer; j:integer):elem* are to set element values and obtain element values, respectively; *the_last_element():elem* indicates the last element in *matrix_mani*, $i(0 \leq i \leq length)$,$j(0 \leq j \leq length)$ indicates the subscript of the corresponding element, and *length* means the length of RNA sequence. *Traceback (m:matrix_mani; matrix: list (array[0. . .n,elem]); aux:auxiliary_storage_mani)* means to backtrack the results. *Output (m:matrix_mani; matrix: list (array[0. . .n,elem]);aux:auxiliary_storage_mani)* is used to output the final result. By default, it outputs the interval of base pairing.

(2) Auxiliary storage table-type component

define ADT auxiliary_storage_mani(someproc inilization (sometype:elem);n:integer);
type auxiliary_storage_mani = private;
procedure set_value(a: auxiliary_ storage _mani;i:integer; j:integer);
function get_value(a: auxiliary_storage_ mani;i:integer;j: integer):elem;
procedure traceback(a: auxiliary_storage _mani);
enddef;

The ADT contains a process generic parameter *someproc initialization_ auxiliary()* and an integer parameter *n* so that generic programs can support instantiation of different dynamic programming patterns. T*ype auxiliary_storage_mani = private* is the storage space description, which is used to indicate that the storage space used by the custom ADT is private. The functions of *Procedure set_value (a: auxiliary_ storage_ mani; i:integer;j:integer)* and *Function get_value(a: auxiliary_storage_mani; i: integer; j:integer): elem* are to set element values and obtain element values, respectively. *Procedure*

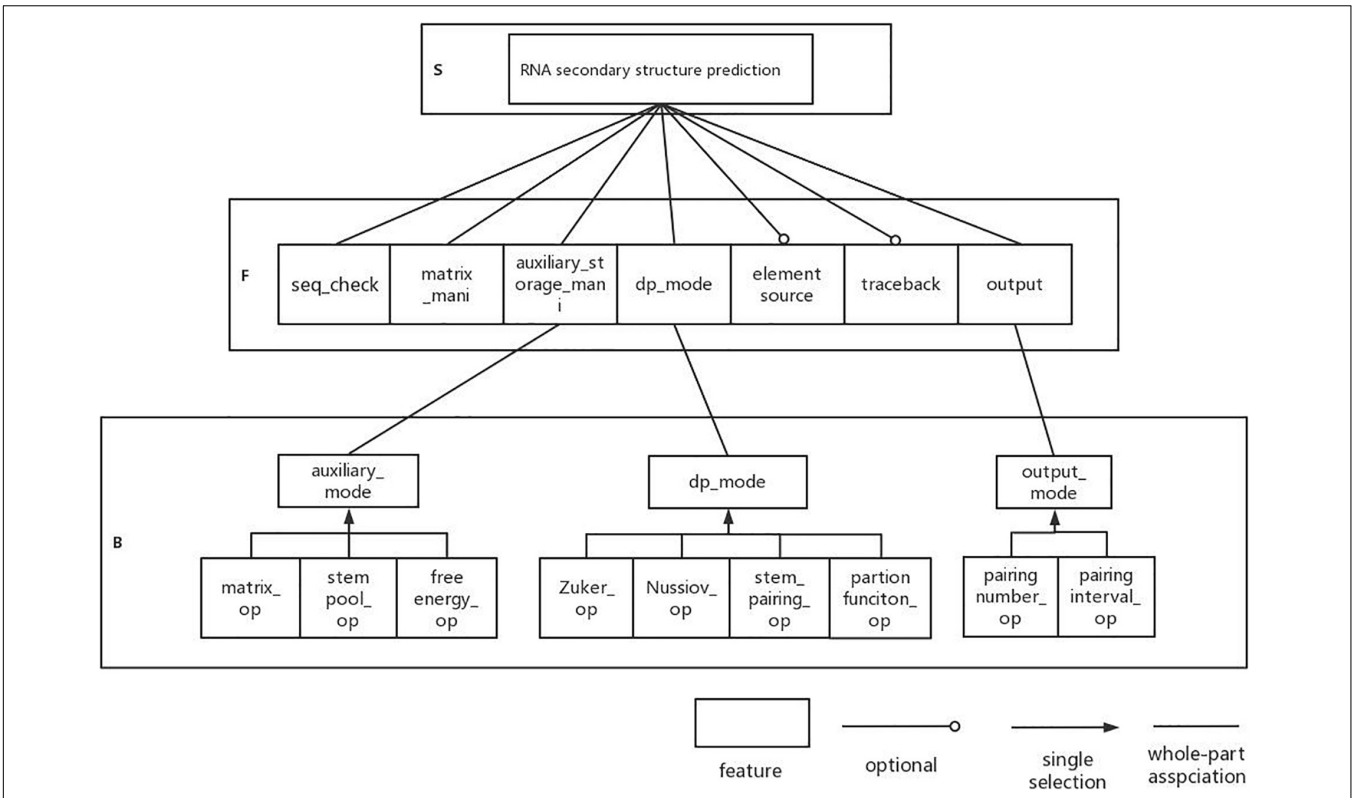


FIGURE 3 | DP-SSP feature model.

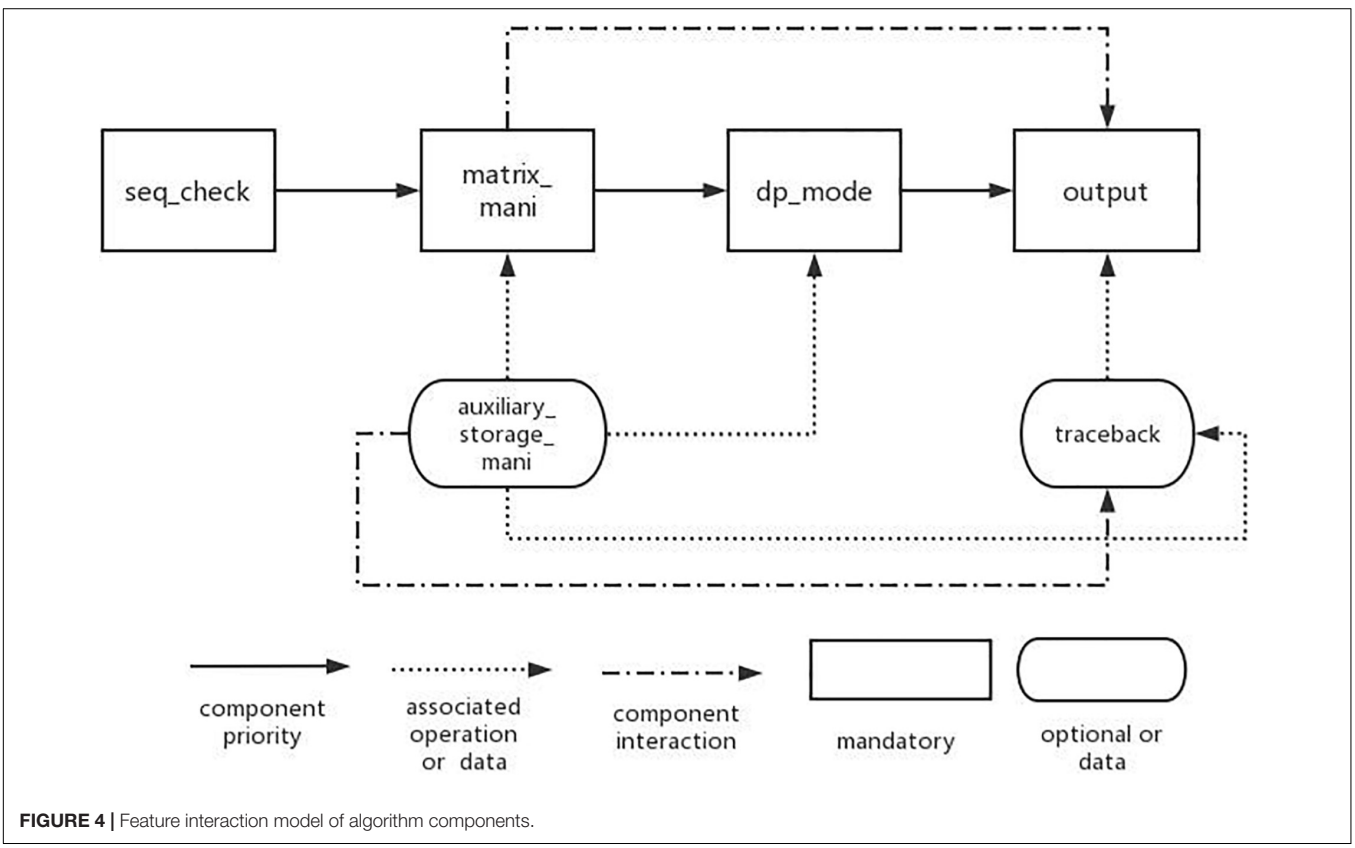


FIGURE 4 | Feature interaction model of algorithm components.

traceback(a:auxiliary_ storage_ mani) means backtracking *auxiliary_storage_mani*.

(3) Secondary structure prediction algorithm component

```
procedure RNA_prediction(m: matrix_mani; a:auxiliary_
storage_mani)
begin
m.apply_memory(m,length);
m.init_matrix();
  do
    i,j ≤ length
    →
    m.setvalue(m.matrix,i,j,a);
  od;
m.traceback(m,a);
M.output(m,matrix,a);
end;
```

The algorithm component contains two generic parameters *m* and *a*; corresponding to types *matrix_mani* and *auxiliary_storage_mani,* respectively, users can get different RNA secondary structure prediction algorithms by passing in different ADT parameters.

Development of Zuker Algorithm Based on ADT

```
program Zuker;
Procedure Zuker_auxiliary_initialization(char);. . . . . ①
var
i:integer;
length:integer;
symbol:char;
begin
open(D:\Zuker\sourcedata.txt)
foreach(i = 0;i< = length;i++)
. . . . . //Code segment, omitted.
end;
ADT Zuker_auxiliary:new auxiliary_ storage_mani(Zuker_
auxiliary_initialization,4);. . . . . ②
ADT Zuker_ matrix_mani:new matrix_ mami
(double);. . . . . ③
Zuker_RNA_prediction():Procedure RNA_prediction (Zuker_
matrix_mani; Zuker_auxiliary);. . . ④
Var:
m:Zuker_ matrix_mani
a:Zuker _ auxiliary
Begin // Main program code. . . . . ⑤
open(D:\Zuker\sourcedata.txt)
foreach(i = 0,j = 0;i <=length, j< = length;i++,j++)
. . . . . //Code segment, omitted.
Zuker_RNA_prediction(m,a);
end;
```

Code block ① indicates the dynamic planning mode of Zuker algorithm, ② indicates the instantiation of auxiliary storage table of Zuker algorithm, ③ indicates the main matrix of instantiating Zuker algorithm, ④ indicates the implementation of prediction code of Zuker algorithm, and ⑤ the following code blocks are the main programs. As the above Apla program cannot be run directly, we use Apla-C++ converter in PAR platform to convert Apla program into C++ program for experimental comparison.

RESULTS

Gutell laboratory provided a large number of real secondary structures of RNA, so we downloaded six real RNA sequences from http://www.rna.icmb.utexas.edu/ to run the assembly algorithm. **Figure 5** shows the prediction result of an RNA sequence named d.5.e.C.carpio. **Table 1** shows the comparative experiments of our assembled algorithm with partition function and Nussinov algorithm in this field.

At present, researchers often use sensitivity (X), specificity (Y), and Matthews correlation coefficient (MCC) to measure the prediction accuracy of the algorithm. Sensitivity refers to the percentage that the real base pairs in the secondary structure are correctly predicted. Specificity refers to the percentage of correct prediction among all predicted base pairs. It is difficult to take both into account in general prediction methods, so researchers often use MCC to compromise. The calculation formulas are as follows:

$$X = \frac{TP}{TP + FN} \tag{8}$$

$$Y = \frac{TP}{TP + FP} \tag{9}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(FN + FP)(TN + FN)}} \tag{10}$$

where TP represents the number of base pairs correctly predicted; FN indicates the logarithm of base pairs in all real structures that are not predicted; and FP represents the predicted logarithm of base pairs that do not exist in the real structure. The value range of MCC is −1 (TP = TN = 0, completely wrong) to 1(FP = FN = 0, absolutely right). Sometimes, for convenience, people often simplify formulas (10) and (11) to evaluate the prediction results.

$$MCC = \sqrt{XY} \tag{11}$$

In this experiment, formulas (8), (9), and (11) are used to evaluate the assembly algorithm.

According to the data in **Table 1**, when the sequence length is 120, 218, 380, 423, 543, and 670, respectively, the algorithm assembled in this article can obtain a better result. The X parameter, Y parameter, and MCC parameter are not inferior to the other two popular RNA secondary structure prediction algorithms. This shows that the algorithm generated by assembly has certain practicability. In addition, using the formal method PAR to develop the algorithm can also improve the development efficiency and reliability of the algorithm, which is convenient for researchers to maintain and optimize. Users only need to select different components for assembly according to the configuration knowledge to generate different specific algorithms. With the continuous

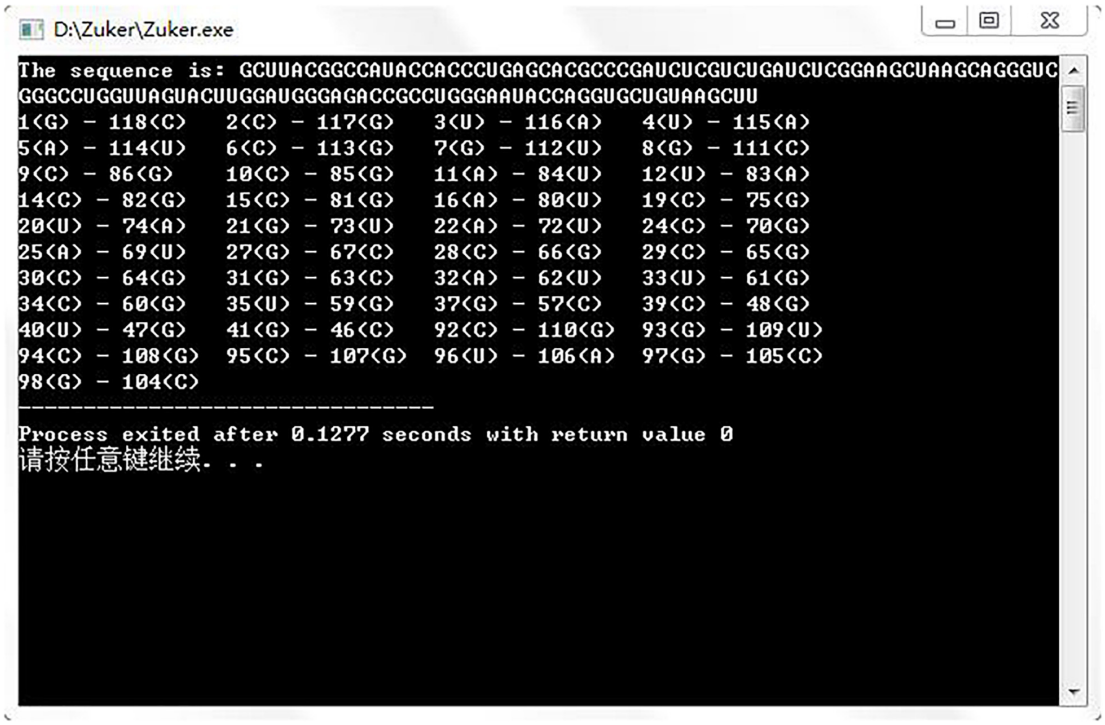


FIGURE 5 | Experimental results of assembly algorithm.

TABLE 1 | Experimental results of different input sequences.

RNA name	Sequence length	Partition function algorithm			Nussinov algorithm			Assembly algorithm		
		X	Y	MCC	X	Y	MCC	X	Y	MCC
d.5.e.C.carpio	120	0.66	0.68	0.67	0.64	0.59	0.61	0.61	0.63	0.62
a.l1.e.L.dispersa	218	0.49	0.44	0.46	0.46	0.43	0.44	0.62	0.63	0.62
a.l1.e.P.inouyei	380	0.53	0.59	0.56	0.36	0.29	0.32	0.68	0.65	0.66
a.l1.e.Staurastrum.sp	423	0.45	0.49	0.47	0.30	0.29	0.30	0.53	0.51	0.52
b.l1.e.H.rubra	543	0.42	0.29	0.35	0.19	0.17	0.17	0.51	0.46	0.48
a.16.m.L.tarentolae	670	0.16	0.21	0.18	0.15	0.18	0.16	0.23	0.24	0.23H

expansion of DP-SSP component library, we are expected to assemble a more efficient new RNA secondary structure prediction algorithm.

DISCUSSION

RNA secondary structure prediction is a hot research direction in bioinformatics, and its implementation algorithm has been widely studied. Because of the flexibility of its algorithm design strategy and the complexity of the problem, this kind of algorithm is full of diversity and complexity. In this article, the GP technology is used to deeply analyze the domain of RNA DP-SSP, find out the general features and variable features, design a highly abstract program component based on Apla language by using the formal method PAR, and use PAR platform to assemble and generate Zuker algorithm, thus improving the reliability and reusability of the algorithm component and reducing the development cost.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HS instructed the whole research work and revised the manuscript. XJ designed the experiments and developed the original manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (No. 62062039 and 61662035).

ACKNOWLEDGMENTS

We thank Xue Jinyun of Jiangxi Normal University for his PAR framework.

REFERENCES

Bowen, S. (2021). *Research on the Method of RNA Secondary Structure Prediction Based on Deep Learning*. Changchun: Jilin University.

Chastek, G., Donohoe, P., and Kang, K. C. (2001). *Produet line analysis:a Practical introduction.Technical Report,Pittsburgh*. Pittsburgh: Carnegie Mellon University. 31–42.

Czarnecki, K., Eisenecker, U. W., and Czarnecki, K. (2000). *Generative programming: methods, tools, and applications*. Boston: Addison Wesley.

Dowell, R. D., and Eddy, S. R. (2004). Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinform.* 5:14. doi: 10.1186/1471-2105-5-71

Eddy, R., and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Res.* 22, 2079–2088. doi: 10.1093/nar/22.11.2079

Fan, S. F., and Zhang, N. X. (2005). Survey of generative programming. *Comp. Sci.* 32, 12–16.

Hu, K. J., and Wei, C. J. (2008). Component-based domain engineering implementation. *Comp. Engineer. Sci.* 30, 92–94. doi: 10.1016/j.ijmedinf.2005.07.017

Huang, Y. Z., Li, D. Y., and Liu, W. (2014). Review of RNA secondary structure prediction algorithms. *Hunan Agricult. Sci.* 2014, 3–8.

Jiang, T., Xu, Y., and Zhang, M. Q. (2002). *Current Topics in Computational Molecular Biology*. Beijing: Tsinghua University Press.

Jinyun, X. (1993). Two new strategies for developing loop in variants and their applications. *J. Comp. Sci. Tech.* 8, 147–154. doi: 10.1007/BF02939477

Jinyun, X. (1998). Formal derivation of graph algorithmic programs using partition-and-recur. *J. Comp. Sci. Tech.* 13, 553–561. doi: 10.1007/bf02946498

Li, K. Q., Chen, Z. L., Mei, H., and Yang, F. Q. (1999). An outline of domain engineering. *Comp. Sci.* 5, 21–25.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29, 1105–1119. doi: 10.1002/bip.360290621

Neighbors, J. M. (1989). Draco: a method for engineering reusable software systems. *Softw. Reusabil.* 1, 295–319. doi: 10.1145/73103.73115

Nussinov, R., Pieczenik, G., and Griggs, J. R. (1978). Algorithms for looping matchings. *SIAM J. Appl. Math.* 35, 68–82. doi: 10.1137/0135006

Peter, C., and Rdf, B. (2000). *Computational molecular biology: an introduction*. New York, NY: John Wiley & Sons Inc, 202–213.

Shi, H. H., and Xue, J. Y. (2009). PAR-based formal development of algorithms. *Chin. J. Comp.* 32, 982–991. doi: 10.3724/sp.j.1016.2009.00982

Shuaimin, L. (2019). *Research on RNA Secondary Structure Prediction*. Wuhan: Huazhong University of Science and Technology.

Wang, C. J., and Xue, J. Y. (2009). *Formal Derivation of a Generic Algorithmic Program for Solving a Class of Extremum Problems*. Piscataway: IEEE, 100–105.

Wartrk, S. (1999). A Phased reuse adoption model. *J. Syst. Softw.* 46, 13–23. doi: 10.1016/s0164-1212(98)10085-7

Xia, P. M. (2008). *Research and Implementation of RNA Secondary Structure Prediction Algorithm*. Harbin: Harbin Institute of Technology.

Xue, J. (1997). A unified approach for developing efficient algorithmic programs. *J. Comp. Sci. Tech.* 12, 314–329. doi: 10.1007/bf02943151

Xue, J. (2015). *Genericity in PAR Platform// International Workshop on Structured Object - Oriented Formal Language and Method*. Cham: Springer, 3–14. doi: 10.1007/978-3-319-31220-0_1

Yang, H. (2013). *Research on Pseudoknot in RNA Secondary Structure Prediction.[master's thesis]*. Changchun: Jilin University.

Yu, D. (2009). *Analysis and Comparison on RNA Secondary Structure Prediction Algorithm.[master's thesis]*. Changchun: Jilin University.

Zhang, W. and Mei, H. (2003). A feature-oriented domain model and its modeling process. *J. Softw.* 14, 1345–1356.

Zhaokui, C., and Yuanchao, X. (2021). Research progresses of RNA structure probing technologies.chinese. *Bull. Life Sci.* 33, 281–291.

Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary infomation. *Nucleic Acids Res.* 9, 133–148. doi: 10.1093/nar/9.1.133

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

METHODS
published: 14 February 2022
doi: 10.3389/fpls.2022.845835

i6mA-Vote: Cross-Species Identification of DNA N6-Methyladenine Sites in Plant Genomes Based on Ensemble Learning With Voting

Zhixia Teng[1], Zhengnan Zhao[1], Yanjuan Li[2], Zhen Tian[3], Maozu Guo[4], Qianzi Lu[5] and Guohua Wang[1]**

[1] College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, [2] College of Electrical and Information Engineering, Quzhou University, Quzhou, China, [3] College of Information Engineering, Zhengzhou University, Zhengzhou, China, [4] College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [5] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Weihua Pan, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences (CAAS), China

Reviewed by:
Guohua Huang, Shaoyang University, China
Yungang Xu, Xi'an Jiaotong University, China
Hui Ding, University of Electronic Science and Technology of China, China

***Correspondence:**
Qianzi Lu
245360359@qq.com
Guohua Wang
ghwang@nefu.edu.cn

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *30 December 2021*
Accepted: *24 January 2022*
Published: *14 February 2022*

Citation:
Teng Z, Zhao Z, Li Y, Tian Z, Guo M, Lu Q and Wang G (2022) i6mA-Vote: Cross-Species Identification of DNA N6-Methyladenine Sites in Plant Genomes Based on Ensemble Learning With Voting. Front. Plant Sci. 13:845835. doi: 10.3389/fpls.2022.845835

DNA N6-Methyladenine (6mA) is a common epigenetic modification, which plays some significant roles in the growth and development of plants. It is crucial to identify 6mA sites for elucidating the functions of 6mA. In this article, a novel model named i6mA-vote is developed to predict 6mA sites of plants. Firstly, DNA sequences were coded into six feature vectors with diverse strategies based on density, physicochemical properties, and position of nucleotides, respectively. To find the best coding strategy, the feature vectors were compared on several machine learning classifiers. The results suggested that the position of nucleotides has a significant positive effect on 6mA sites identification. Thus, the dinucleotide one-hot strategy which can describe position characteristics of nucleotides well was employed to extract DNA features in our method. Secondly, DNA sequences of Rosaceae were divided into a training dataset and a test dataset randomly. Finally, i6mA-vote was constructed by combining five different base-classifiers under a majority voting strategy and trained on the Rosaceae training dataset. The i6mA-vote was evaluated on the task of predicting 6mA sites from the genome of the Rosaceae, Rice, and Arabidopsis separately. In Rosaceae, the performances of i6mA-vote were 0.955 on accuracy (ACC), 0.909 on Matthew correlation coefficients (MCC), 0.955 on sensitivity (SN), and 0.954 on specificity (SP). Those indicators, in the order of ACC, MCC, SN, SP, were 0.882, 0.774, 0.961, and 0.803 on Rice while they were 0.798, 0.617, 0.666, and 0.929 on Arabidopsis. According to the indicators, our method was effectiveness and better than other concerned methods. The results also illustrated that i6mA-vote does not only well in 6mA sites prediction of intraspecies but also interspecies plants. Moreover, it can be seen that the specificity is distinctly lower than the sensitivity in Rice while it is just the opposite in Arabidopsis. It may be resulted from sequence similarity among Rosaceae, Rice and Arabidopsis.

Keywords: N6-methyladenine, plant genomes, cross-species, feature encoding, ensemble learning

INTRODUCTION

DNA N6-methyladenine (6mA) is a methyl modification at the sixth position of the adenine ring, which was discovered by Vanyushin et al. (1968). 6mA is widely found in prokaryotes and eukaryotes (Fu et al., 2015; Greer et al., 2015; Zhang et al., 2015). It is reported that 6mA plays vital roles in DNA replication, repairing nucleotide dislocations, and preventing the invasion of foreign DNA (Wion and Casadesús, 2006). Although 6mA in animal genomes studies have been well studied, those of plants genomes have still known a little, which hampered to explore their functions. To better understand the molecular mechanism of 6mA in plants, it is the first step to determine the 6mA sites accurately.

To detect 6mA sites, several biochemical methods were developed, such as single-molecule real-time sequencing technology (SMRT-seq) (Davis et al., 2013) and restriction endonuclease-based 6mA sequencing (6MA-RE-seq) (Fu et al., 2015). In SMRT-seq, single-nucleotide molecules labeled by different fluorophores were paired with bases of a DNA sequence, and the fluorescence signals were recorded during the process of pairing. The fragment of DNA sequence may be methylated if it showed the continuous same signal during the process of pairing. 6mA-RE-seq explored restriction enzymes to fragment genomic DNA at "CATG" and "GATC" motifs that did not contain 6mA and then retained these motifs containing 6mA. In this way, after end-repair and other operations, the methylated motifs would be enriched in the internal positions of DNA fragments. However, these methods are hard to detect 6mA sites from high-throughput sequences because they are time-consuming and expensive.

Therefore, some machine learning models have been developed to identify 6mA sites in recent years because they are efficient and cheap. At first, iDNA6mA-PseKNC (Feng et al., 2019) was proposed to detect 6mA sites in the mouse genome. In this model, DNA sequences were represented by pseudo-k-tuple nucleotide composition incorporating the physicochemical properties of nucleotides, and then the sequences were classified by a support vector machine (SVM). Subsequently, i6mA-Pred (Chen et al., 2019) trained a novel SVM model to identify 6mA sites in the rice genome based on the chemical properties of nucleotide such as the loop structure, the hydrogen bond, and the amino groups, and the nucleotide frequency of DNA sequences. To avoid overfitting, i6mA-Pred used the maximum correlation maximum distance approach to select the most representative features. Afterward, iN6-methylate (Le, 2019), another novel SVM model, used FastText to generate feature vectors for DNA sequences based on the assumption that a DNA sequence is a sentence and a nucleotide is a word. Unlike previous models, MM-6mAPred (Pian et al., 2019) constructed Markov chains based on DNA sequences with 6mA sites (positive samples) and DNA sequences without 6mA sites (negative samples) in the training dataset. Based on the Markov chains, the positive and negative probabilities of a DNA sequence were calculated separately. It is considered that a sequence contained 6mA site if the ratio of positive probability against negative probability is greater than 1.

To improve the performance of above methods, ensemble learning has been increasingly applied to 6mA sites prediction. In the beginning, iDNA6mA-Rice (Lv et al., 2019), a rice 6mA site classification model based on random forest, encoded DNA sequences via three feature descriptors, namely the k-nucleotide frequency, the mono-nucleotide binary coding, and the natural vector containing the frequency, average position, and second-order central moment of mono-nucleotides. Soon afterward, on the basis of bagging with CART, i6mA-DNCP (Kong and Zhang, 2019) represented rice DNA sequences by two novel feature descriptors: dinucleotide frequency and dinucleotide physicochemical properties. In addition, i6mA-DNCP employed heuristic ideas to select the most representative features. Several months later, i6mA-Fuse (Hasan et al., 2020) was proposed to classify Rosaceae DNA sequences with random forest and linear regression. Subsequently, a random forest-based multi-species 6mA site prediction model 6mA-Finder (Xu et al., 2020) was developed, which contained three modules for mouse, rice, and a general species admixed by mouse and rice DNA sequences, respectively. i6mA-stack (Khanal et al., 2021) developed a two-level stacked ensemble classifier based on linear regression, random forest, support vector machine, and gaussian naive bayes to recognize Rosaceae 6mA sites.

With the development of deep learning, some neural network models were also developed for identifying 6mA sites. For example, iDNA6mA (Tahir et al., 2019) is composed of four layers: two convolution layers which extract features of DNA sequences, a dropout layer which is used to avoid overfitting, and a full-connection layer which performs classification tasks. Subsequently, SNNRice6mA (Yu and Dai, 2019) was improved iDNA6mA by adding a normalization layer and a pooling layer between the convolution layer and the dropout layer, which aimed to reduce redundant features of DNA sequences according to the correlation of the features. i6mA-DNC (Park et al., 2020) is similar with the above two models except it extracted features from nucleotide pairs of DNA sequences rather than from single nucleotides. It is worth noting that the three neural network models mentioned above were all developed for predicting 6mA sites in the rice genome.

Because the previously mentioned models are species-specific, Meta-i6mA (Hasan et al., 2021) was proposed for 6mA site prediction from multiple plants. Although Meta-i6mA has achieved encouraging results in intraspecies, it still has room for improvement in interspecific. To solve this problem, a novel classification model i6mA-vote was developed based on an ensemble learning strategy. In this model, DNA sequences were encoded by nucleotide position-based feature descriptors, and then these sequences were classified by an ensemble classifier integrating random forest, linear discriminant analysis, multi-layer perceptron, stochastic gradient descent, and extreme gradient boosting. The details of i6mA-vote will be introduced in the following sections.

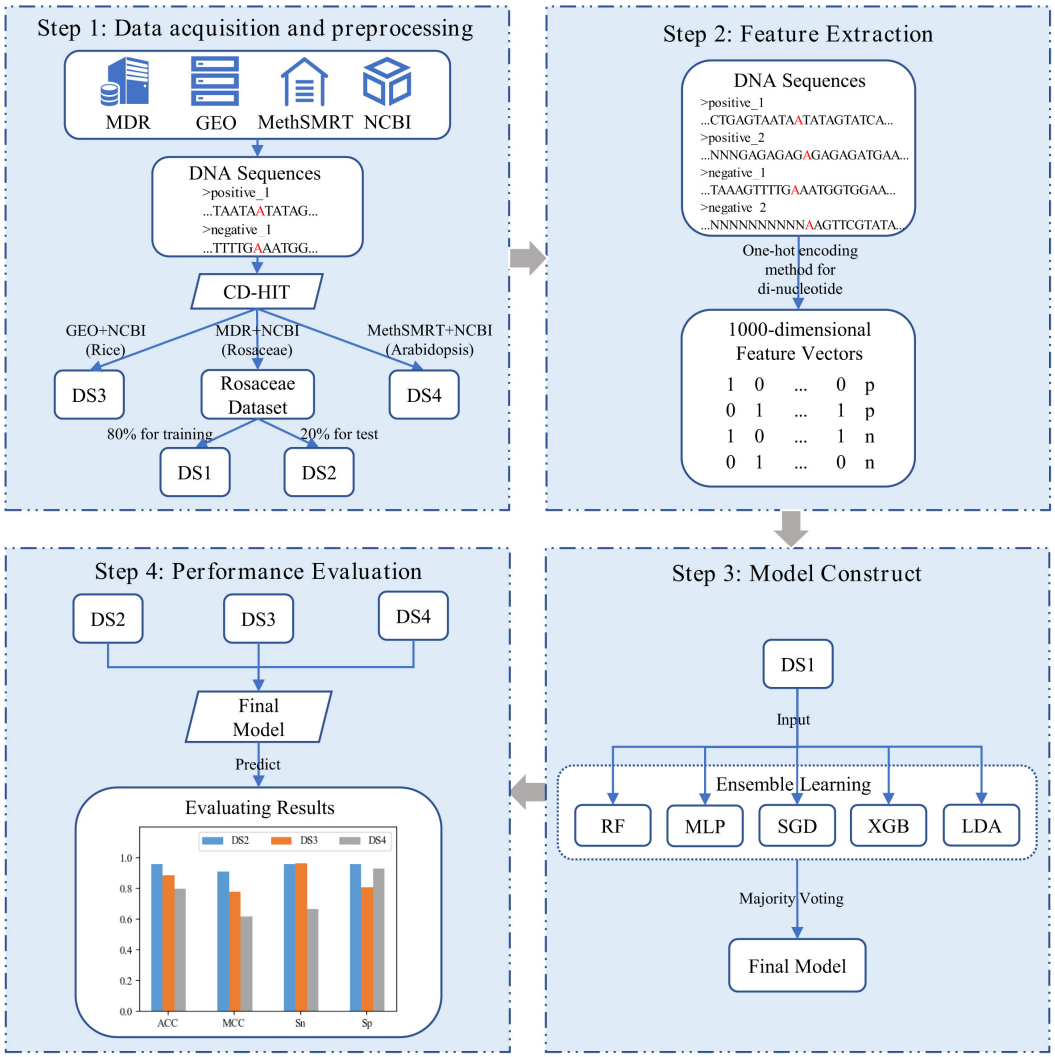


FIGURE 1 | Frame diagram of i6mA-vote. (DS1, DS2, DS3, and DS4 refer to Rosaceae training dataset, Rosaceae test dataset, Rice test dataset, and Arabidopsis test dataset; In the DNA sequences, the letter "A" marked in red refers to the possible 6mA site, and the letter "N" indicates the unidentified nucleotide; In the feature vectors, "p" and "n" are short for the positive sample and the negative sample).

MATERIALS AND METHODS

Framework of i6mA-Vote

In our study, as shown in **Figure 1**, i6mA-vote was constructed by four steps. Firstly, positive samples of Rosaceae, Rice, and Arabidopsis were derived from MDR (Liu et al., 2019), GEO (Edgar et al., 2002), and MethSMRT (Ye et al., 2017) databases, and negative samples of these plants were downloaded from NCBI. For each plant, the positive and negative samples were filtered by CD-HIT (Li and Godzik, 2006) to reduce high similar samples. Then all samples were divided into three datasets according to organisms for the subsequent experiments. The Rosaceae dataset was split into a training dataset and a test dataset, and datasets for the remaining two species were used as cross-species evaluation datasets. Secondly, to transform DNA sequences into feature vectors, one-hot encoding method was performed on dinucleotides (e.g., AA, AG, ...) of DNA sequences. Because the known nucleotides can be represented by four symbols (A, G, C, T) and other unknown nucleotides can be denoted by symbol N, in this way, there were twenty-five dinucleotide combinations. Thirdly, an ensemble learning model, named i6mA-vote, was built by integrating random forest (RF), multi-layer perceptron (MLP), stochastic gradient descent (SGD), linear discriminant analysis (LDA), extreme gradient boosting (XGB), based on majority voting strategy. Then all samples were represented by feature vectors and the ensemble learning model was trained on the samples. Finally, to evaluate the performance of the model, i6mA-vote was used to perform simulation a task on test datasets, and its superiority was demonstrated by accuracy, Matthew correlation coefficient, sensitivity, and specificity. In the following sections, the

TABLE 1 | Number of samples in each dataset.

Datasets	Number of positive samples	Number of negative samples	Total number
DS1	29237	29433	58670
DS2	7298	7300	14598
DS3	153635	153629	307264
DS4	31414	31843	63257

detail process of constructing the i6mA-vote model will be illustrated step by step.

Datasets

The quality of the dataset affects the performance of the classification model. In this study, four high-quality datasets that have been applied in the 6mA prediction domain were selected.

The Rosaceae dataset was collected, collated, and constructed by Hasan's team (Hasan et al., 2021). The part containing 6mA were derived from the MDR database (Liu et al., 2019). After removing similar sequences and excluding 90% sequence identity, 36,537 positive samples were obtained. The other part, including the same number of negative ones, was taken from NCBI, and it was generated by chromosomes with no 6mA detected. Finally, 80% of this dataset was randomly selected as the training dataset (DS1), and the remaining 20% was regarded as the test dataset (DS2).

The Rice dataset (DS3) was created by Lin's group (Lv et al., 2019). The positive portion and the negative one were obtained from the GEO database (Edgar et al., 2002) and NCBI. And they both included 154000 samples.

The Arabidopsis dataset (DS4) was also constructed by Hasan's team (Hasan et al., 2021). It extracted 31,873 6mA sites from the MethSMRT database (Ye et al., 2017) and replenished the same number of negative samples from NCBI using the same way as for the Rosaceae dataset.

Among them, DS1 was used for training the model, DS2, DS3, and DS4 were used to evaluate the generalization performance and cross-species prediction ability of the model.

All the above four datasets were downloaded from the online server of model Meta-i6mA (Hasan et al., 2021)[1]. In addition, these datasets were also processed as follows: (1) Sequences longer than 41bp were removed. (2) If a sequence was repeated multiple times, it would be deleted, leaving only one copy. (3) If a sequence was present in both positive and negative samples, it would be removed from both parts. Finally, the number of samples included in each dataset is shown in **Table 1**. Their sequences all consisted of 41 nucleotides with an "A" in the middle.

Feature Extraction

To convert DNA sequences into feature vectors, One-hot encoding method for dinucleotides was employed in our model. This strategy and other concerned strategies will be described in detail below.

[1]http://kurata14.bio.kyutech.ac.jp/Meta-i6mA/download_file/Meta-6mA-datasets.zip

Our Encoding Strategy

One-hot encoding method for dinucleotides (One-hot2) is based on the one-hot encoding method in natural language processing. The one-hot encoding method compiles a dictionary using the words in the sentences and then encodes each word into a 0-1 vector through this dictionary. The length of the vector is equal to that of the dictionary, and each bit in the vector corresponds to a word in the dictionary. When encoding a word, its corresponding bit is set to 1 in the vector, and the other bits are kept at 0. Similarly, One-hot2 treats DNA sequences as sentences and dinucleotides as words.

A DNA sequence is usually composed of four standard nucleotide symbols: A, C, G, and T. However, sometimes the DNA sequence also include non-standard nucleotide symbol N, which means that the nucleotide was not identified. Accordingly, a DNA sequence may consist of 5 symbols, and it contains 25 possible symbol combinations of dinucleotides like AA, AC, AN. In our method, the one-hot2 encoded each dinucleotide into a 25-dimensional 0-1 vector. The vector of each dinucleotide is shown in Formula (1).

$$\begin{cases} AA = (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \\ AC = (0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \\ AG = (0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) \\ \qquad\qquad\qquad\qquad \vdots \\ NT = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0) \\ NN = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1) \end{cases} \quad (1)$$

To show how One-hot2 encodes DNA sequences, an example is given below. DNA sequence $D = ACGTNA$ can be split into five dinucleotides (AC, CG, GT, TN, NA), and then they are replaced with their corresponding one-hot codes. In this way, a vector with the dimension of 125 is generated.

Because the length of the DNA sequences in our datasets are 41bp, the sequences can be spliced into 40 dinucleotides and thus the vectors of these dinucleotides were concatenated into a 1000-dimensional feature vector to describe their primary sequence.

There are three reasons why One-hot2 was chosen: (1) It can solve the problem that classifiers are not good at handling continuous data. In addition, it generates sparse vectors, allowing many machine learning problems to be linearly separated and models more efficient to be stored. (2) It considers the relationship between adjacent nucleotides as it is encoded in dinucleotide. (3) Some studies (Chen et al., 2019; Feng et al., 2019) found position-specific features can better represent sequences containing 6mA sites, and One-hot2 happens to be this kind of method.

The Concerned Encoding Strategies

Density-Based Approach

Accumulated Mono-Nucleotide Frequency (AMNF) represent the frequency of single nucleotides in the subsequence which ranges from the first nucleotide to the current nucleotide of the original sequence. Similarly, Accumulated Di-Nucleotide Frequency (ADNF) (Chen et al., 2017) denotes the nucleotide pairs which appears before current nucleotide. For example, DNA sequence

$D = ACGTNA$ can be encoded as (1, 0.5, 0.33, 0.25, 0.2, 0.33) and (1, 0.5, 0.33, 0.25, 0.2) by AMNF and ADNF, respectively.

Physicochemical-Properties-Based Approach

Dinucleotide Physical-Chemical Properties (DPCP) and Trinucleotide Physical-Chemical Properties (TPCP) (Manavalan et al., 2019; Wei et al., 2019) replace the DNA sequences with the vectors calculated by Equation (2) using the physicochemical-properties in **Supplementary Tables 1,2**. In **Supplementary Table 1**, the columns represent 15 physicochemical properties, and the rows represent 25 dinucleotides. In **Supplementary Table 2**, the columns represent 11 physicochemical properties, and the rows represent 125 trinucleotides.

$$xPCP_i = N_i \times xPC_{ij} \tag{2}$$

where $x = D$ refers to Dinucleotide and $x = T$ denotes Trinucleotide. When $x = D$, the values of i range from 1 to 25, the values of j range from 1 to 15, $DPCP_i$ is the DPCP value of the ith dinucleotides, N_i is the count of the ith dinucleotides in the DNA sequence, and DPC_{ij} is the jth properties of the ith dinucleotides; When $x = T$, the values of i range from 1 to 125, the values of j range from 1 to 11, $TPCP_i$ is the TPCP value of the ith trinucleotides, N_i is the count of the ith trinucleotides in the DNA sequence, and TPC_{ij} is the jth properties of the ith trinucleotides.

Position-Based Approach

One-hot encoding method for mononucleotide (One-hot1) is similar to One-hot2, except that its encoding unit is the mononucleotide. It converts a mononucleotide into a one-hot code with a length of five, corresponding to five mononucleotides (A, C, G, T, and N). For instance, the encoded vector of DNA sequence $D = ACGTNA$ is (1,0,0,0,0| 0,1,0,0,0| 0,0,1,0,0| 0,0,0,1,0| 0,0,0,0,1| 1,0,0,0,0).

Classifier

To train a classification model with stable and good performance, five machine learning algorithms was utilized to construct five base-classifiers. Subsequently, majority voting was adopted to integrate these five base-classifiers. Its detailed procedure is illustrated in the following steps.

(1) The processed training dataset was inputted into five machine learning algorithms, and five base-classifiers were generated. These five algorithms were random forest (RF), multi-layer perceptron (MLP), stochastic gradient descent (SGD), linear discriminant analysis (LDA), extreme gradient boosting (XGB). Among them, RF refers to one type of classifier that utilizes multiple decision trees to train and predict samples. MLP, as a simple neural network, contains three fully connected layers, the input layer, the hidden layer, and the output layer. SGD is a kind of support vector machine model. LDA is a classifier generated according to Bayes' rule. XGB is also based on trees, but unlike random forests, its trees are regressive, and it also optimizes the algorithm itself, the efficiency and robustness of the algorithm.

(2) The five base classifiers were combined into one ensemble classifier by majority voting. That is, when three or more base classifiers judge a sequence to be a positive (or negative) sample, then their combination also treats this sequence as a positive (or negative) sample.

It should be noted that the hyperparameters of the base-learners were optimized by grid search strategy. After manually specifying variation ranges of hyperparameters, this strategy adopted an exhaustive method-like approach to find the best-performing combination from these hyperparameters. In addition, all classifier algorithms in this paper were implemented by sklearn (Hinton, 1989; Belhumeur et al., 1997; Platt, 2000; Breiman, 2001; Bengio and Glorot, 2010; Pedregosa et al., 2011; Kingma and Ba, 2014; He et al., 2015; Chen and Guestrin, 2016).

Performance Evaluation

Our model was validated according to accuracy (ACC), Matthew correlation coefficient (MCC), Sensitivity (SN), Specificity (SP) which had been widely adopted in the field of bioinformatics (Huang and Gong, 2020; Liu et al., 2020; Smolarczyk et al., 2020; Wang H. et al., 2020; Wang J. et al., 2020; Shao and Liu, 2021; Zhang et al., 2021). These metrics can be calculated by equations (3) ∼ (6).

$$ACC = \frac{n_{TP} + n_{TN}}{n_{TP} + n_{FN} + n_{TN} + n_{FP}} \tag{3}$$

$$MCC = \frac{n_{TP} \times n_{TN} - n_{FN} \times n_{FP}}{\sqrt{(n_{TP} + n_{FP}) \times (n_{TP} + n_{FN}) \times (n_{TN} + n_{FP}) \times (n_{TN} + n_{FN})}} \tag{4}$$

$$SN = \frac{n_{TP}}{n_{TP} + n_{FN}} \tag{5}$$

$$SP = \frac{n_{TN}}{n_{TN} + n_{FP}} \tag{6}$$

Where TP and TN refer to correctly predicted 6mA and non-6mA; FP and FN denote incorrectly predicted non-6mA and 6mA; n_x means the number of x.

RESULTS AND DISCUSSION

DNA Sequence Logos

To find optimal features of samples, the DNA sequences of samples should be analyzed. Since these sequences were of equal length, they could be analyzed sequence logos (Schneider and Stephens, 1990). Two Sample Logo was employed (Vacic et al., 2006), which calculated the statistical difference between positive and negative samples at specific positions. The logo consists of three parts, the upper and lower parts represent the enriched and depleted nucleotides at specific positions, and the middle part denotes the consistent results of positive and negative samples. The x-axis indicates the position. The length of DNA sequences in our datasets is 41bp, so there are 41 scales on the x-axis. Additionally, as the middle nucleotide is consistent in both positive and negative samples, it is set to the 0th scale. The y-axis represents the amount of information at the position. The higher the symbol in a position, the more

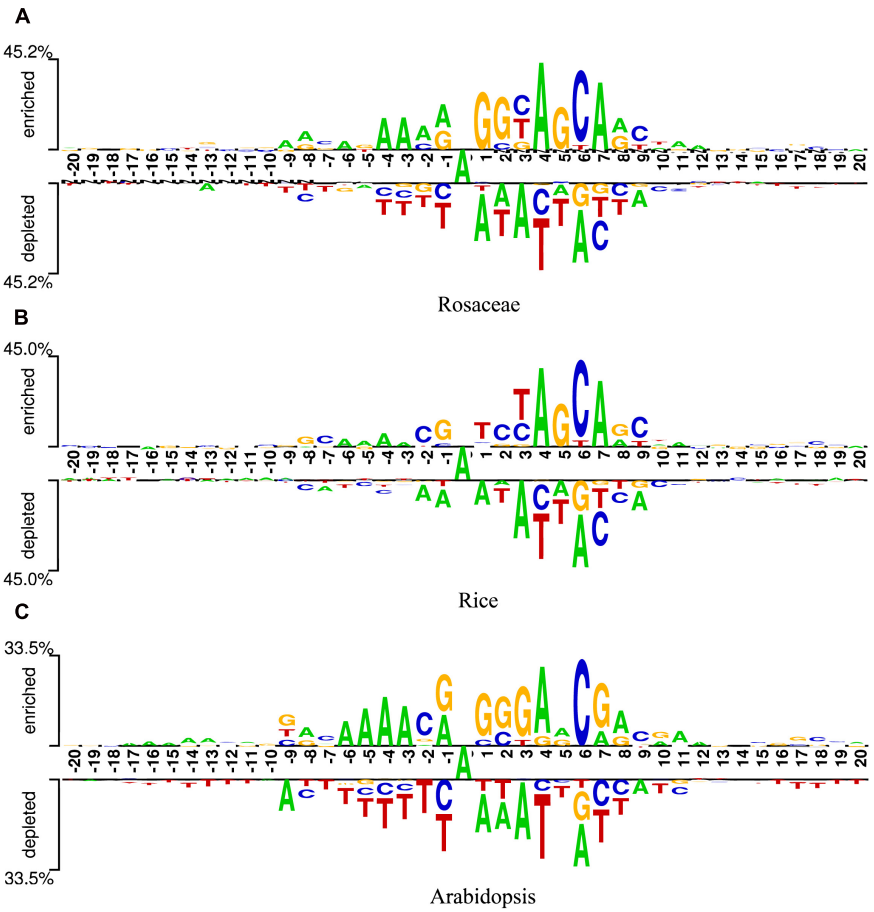


FIGURE 2 | Sequence logos of Rosaceae **(A)**, Rice **(B),** and Arabidopsis **(C)**.

information the position contains. In addition, the relative size of a base letter shows its relative frequency at one position. If a letter is larger than the other letters in the column, it has a high frequency in that position. At each position, the base letters are arranged in the order of dominance from top to bottom. Generally, the consensus motif can be found by reading the top of each position.

Figures 2A-C are the sequence logos established for Rosaceae, Rice, and Arabidopsis. From the three figures, it can be seen that the sequences have a length of 41bp with "A"s at the center. In addition, "A" enriched at positions −6, −4, −3, 4, 7, 8, 10, 11, 12, "C" enriched at positions −7, −2, 2, 6, 9, "G" enriched at positions −8, −1, 2, 3, 5, 8, and "T" enriched at positions 3. Since these sequences containing 6mA are enriched with some nucleotides at some positions, it is speculated that position-based approaches are more suitable for extracting information from the sequences in our datasets.

Performance Evaluation of Models

To verify the conjecture in the previous section, six methods were chosen to extract the datasets as features and then they were applied to five commonly used well-performing algorithms in sklearn. Since the conjecture is too intuitive and may lead to some significant features being overlooked, not only nucleotide position-based methods are compared, but also density-based and physicochemical property-based methods were also compared.

The experimental results of 5-fold cross-validation are displayed in **Table 2**. The columns indicate the feature extraction methods which have been introduced in the "Feature Extraction" section. The rows denote classifier algorithms and their evaluation metrics, and they have been briefly described in the "Classifiers" section and the "Performance Evaluation" section.

As can be seen in **Table 2**, whichever classifier algorithm is selected, the ACCs, SNs, SPs, and MCCs of AMNF, ADNF, DPCP and TPCP are all lower than 0.80, 0.79, 0.83, and 0.60, whereas them of One-hot1 and One-hot2 are all higher than 0.93, 0.93, 0.91, and 0.86. These illustrate that compared with density-based and physicochemical property-based approaches, position-based ways can better express the characteristics contained in DNA sequences in our datasets. XGB performed slightly better with one-hot1 than

TABLE 2 | Indicators of different features and classifier algorithms.

		AMNF	ADNF	DPCP	TPCP	One-hot1	One-hot2
Random forest	ACC	0.786	0.642	0.594	0.669	0.935	**0.938**
	SN	0.746	0.627	0.587	0.683	0.937	**0.939**
	SP	0.825	0.655	0.602	0.656	0.933	**0.937**
	MCC	0.573	0.283	0.189	0.339	0.870	**0.877**
Linear discriminant analysis	ACC	0.643	0.609	0.614	0.660	0.908	**0.931**
	SN	0.650	0.624	0.597	0.629	0.937	**0.945**
	SP	0.637	0.594	0.632	0.692	0.879	**0.917**
	MCC	0.287	0.219	0.228	0.321	0.818	**0.862**
Multi-layer perceptron	ACC	0.755	0.627	0.602	0.625	0.937	**0.939**
	SN	0.743	0.602	0.576	0.621	0.936	**0.942**
	SP	0.767	0.652	0.628	0.629	**0.939**	0.937
	MCC	0.510	0.255	0.204	0.250	0.875	**0.878**
Stochastic gradient descent	ACC	0.643	0.605	0.549	0.577	0.910	**0.931**
	SN	0.631	0.647	0.561	0.481	0.917	**0.936**
	SP	0.654	0.563	0.538	0.672	0.904	**0.926**
	MCC	0.287	0.212	0.099	0.157	0.821	**0.861**
Extreme gradient boosting	ACC	0.790	0.647	0.616	0.673	**0.944**	0.940
	SN	0.788	0.650	0.617	0.672	**0.948**	0.942
	SP	0.791	0.644	0.616	0.675	**0.939**	0.937
	MCC	0.579	0.294	0.233	0.346	**0.888**	0.880

Bold values indicate the best performance.

one-hot2. This may be because XGB may lose some valuable information when it was applied on high-dimensional one-hot2 features. Specifically, XGB divides the high-dimensional feature space into many small parts which may be treated as noise. In addition, if the feature descriptor is One-hot1 or One-hot2, all classifiers show good performance, which indicates that all these algorithms are appropriate for this classification task.

Moreover, to judge intuitively whether the above six feature extraction methods were good at distinguishing between positive and negative samples, the tSNE (van der Maaten and Hinton, 2008) technique in sklearn (Pedregosa et al., 2011) was used to project the sample points of these methods from the high-dimensional space to the two-dimensional space. If the positive and negative sample points can be well separated in the two-dimensional space, they are also separable in the high-dimensional space. The visualization plots of the projection are shown in **Figure 3**. It can be seen from **Figure 3** that the samples of the two labels are separated by certain dividing lines in **Figures 3E,F**, while in other subgraphs, the negative sample points are almost covered by the positive ones. These illustrate that One-hot1 and One-hot2 can better discriminate the sample points of the two labels in a high dimensional space than the other four methods.

Through these arguments, the nucleotide position-based methods are indeed more suitable for extracting features from DNA sequences in our datasets, and the assumptions that was made in the previous section are proved to be correct. Therefore, in the subsequent analysis, only One-hot1 and One-hot2 would be considered.

Comparison of Features

In the previous section, it has been learned that the position-based approaches express the information contained in our DNA sequences well. However, it is not sure which is the best among One-hot1, One-hot2, and their fusion. Therefore, in this subsection, they are compared. The comparison results are shown in **Figure 4**.

As can be seen from **Figure 4**, only when the classifier is XGB, the effect of the other two is slightly better than One-hot2; when the classifier is RF, LDA, MLP, or SGD, One-hot2 is significantly better than One-hot1 and slightly better than the fusion. The reason for this is that when encoding a dinucleotide, some information about the mononucleotide is involved. Therefore, in most cases, One-hot1 is not as good as One-hot2, and their fusion produces some redundant information. Consequently, One-hot2 is the best answer.

Efficiency of Ensemble Strategy

Using One-hot2 to extract features and take RF, LDA, MLP, SGD, and XGB as classifiers, five base models can be obtained. As shown in **Figure 5**, except for some differences between SN and SP of LDA and SGD, SN and SP for the other three classifiers do not differ much, as well as these base models are all with excellent performance, so they were tried to be combined with the majority voting strategy. The integrated results are also shown in **Figure 5**. It can be found that after voting, except for no enhancement in SP, all the other three metrics improved, which means that after this operation, the performance of the whole classification system has been risen to a higher level.

Comparison With Other Machine Learning Models

To evaluate the generalization capability and cross-species identification ability of our model, it was applied to three test datasets, DS2, DS3, and DS4. Moreover, the test results were compared with several other machine learning models to demonstrate the advantages of our model. **Table 3** shows the comparative results on Rosaceae, Rice, Arabidopsis. The columns indicate four evaluation indicators that have been introduced in the "Performance Evaluation" section. The rows represent the species and the models applied on these species. The models include Meta-i6mA (Hasan et al., 2021), i6mA-Fuse (Hasan et al., 2020), i6mA-stack (Khanal et al., 2021), i6mA-Pred (Chen et al., 2019), iDNA6mA-Rice (Lv et al., 2019), MM-6mAPred (Pian et al., 2019), and 6mA-Finder (Xu et al., 2020). Among them, i6mA-Fuse consists of two modules, which were trained by the datasets of Fragaria Vesca and Rosa Chinensis, respectively. To

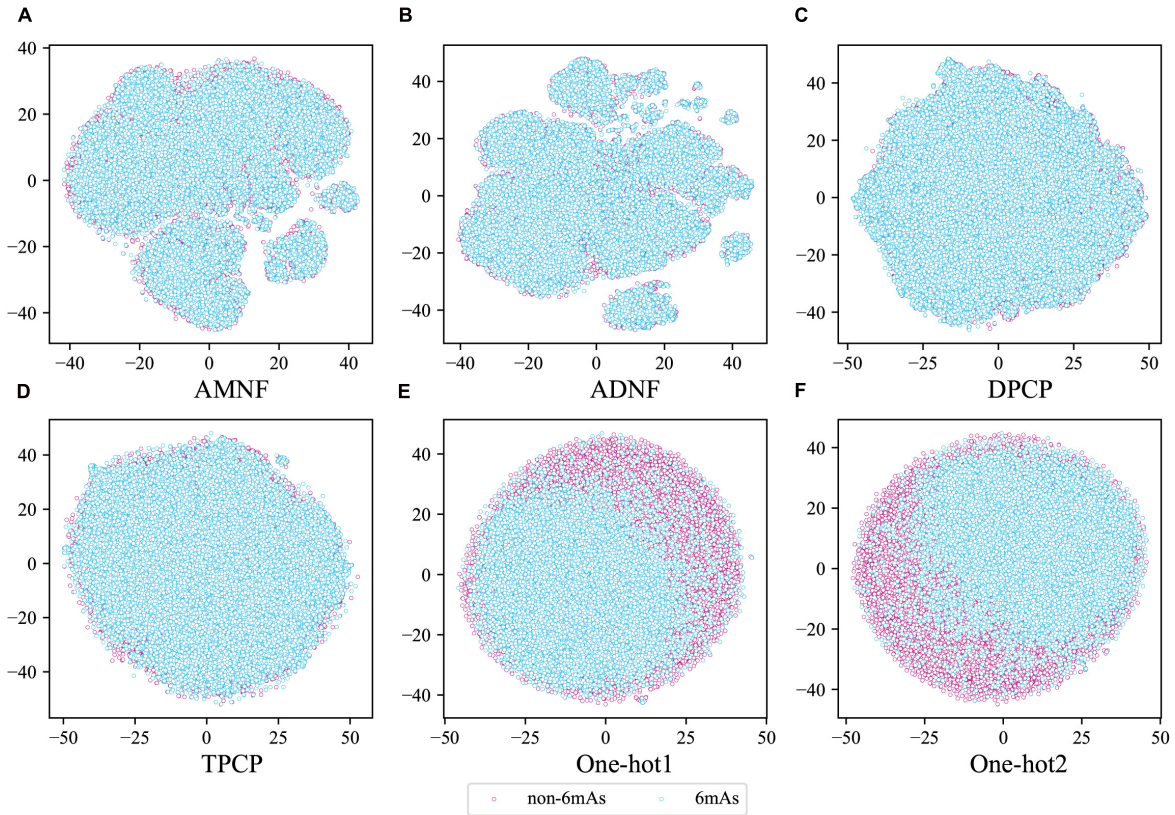


FIGURE 3 | The tSNE scatterplots of AMNF **(A)**, ADNF **(B)**, DPCP **(C)**, TPCP **(D)**, One-hot1 **(E)**, and One-hot2 **(F)**. (Blue and pink dots indicate DNA sequence samples with and without 6mA sites, respectively).

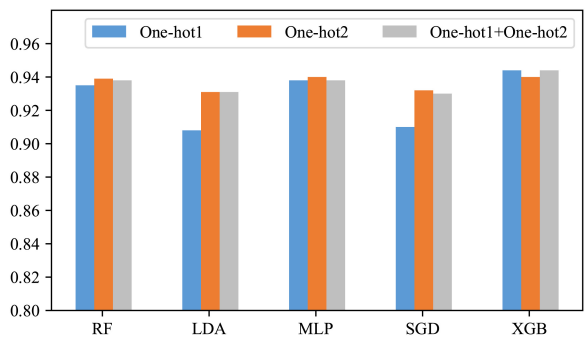


FIGURE 4 | Comparison before and after feature fusion.

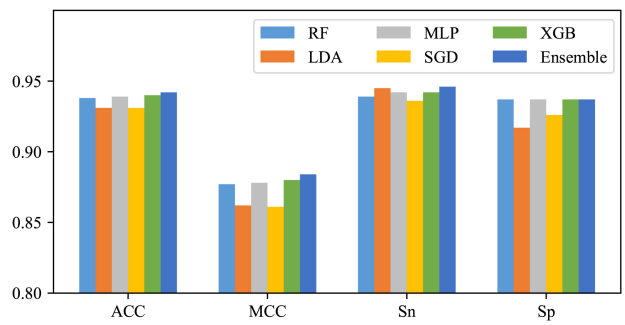


FIGURE 5 | Effects of the ensemble strategy.

better distinguish them, i6mA-Fuse_FV and i6mA-Fuse_RC are used instead. The same situation is true for i6mA-stack.

As can be seen from **Table 3**, when the species is Rosaceae, although our SN and SP values only rank second, our ACC and MCC values are the maximum, suggesting that our model has the best overall performance in Rosaceae. It can be concluded that our model can make cross-species predictions for Rice as all four metrics of our model rank at the top. And it can better find 6mA sites from unknown Rice sequences because our model has the highest SN value. Like Rosaceae, our model predicts 6mA sites well in Arabidopsis, and with the highest SP, our model can better screen out those sequences that do not contain 6mA sites. Considering the comparative results on the three species, our model has better generalization performance and cross-species prediction ability than other methods. This may be because only the best-performing feature descriptor was selected to represent the DNA sequences rather than the fusion of several well-performing features. Thereby, the risk of generating

TABLE 3 | Comparison with other machine learning models on Rosaceae, Rice, and Arabidopsis.

		ACC	MCC	SN	SP
Rosaceae	Meta-i6mA	0.953	0.905	0.954	0.951
	i6mA-Fuse_FV	0.943	0.887	0.924	**0.962**
	i6mA-Fuse_RC	0.893	0.786	0.890	0.895
	i6mA-stack_FV	0.928	0.856	0.928	0.927
	i6mA-stack_RC	0.899	0.798	0.920	0.877
	i6mA-Pred	0.840	0.684	0.897	0.782
	iDNA6mA-Rice	0.878	0.764	0.951	0.805
	MM-6mAPred	0.873	0.758	**0.961**	0.785
	6mA-Finder	0.846	0.701	0.928	0.764
	i6mA-vote	**0.955**	**0.909**	0.955	0.954
Rice	Meta-i6mA	0.880	0.768	0.957	0.802
	i6mA-Fuse_FV	**0.890**	**0.781**	0.921	**0.859**
	i6mA-Fuse_RC	0.775	0.571	0.907	0.644
	i6mA-stack_FV	0.876	0.756	0.938	0.815
	i6mA-stack_RC	0.813	0.640	0.915	0.712
	i6mA-Pred	0.791	0.592	0.878	0.705
	iDNA6mA-Rice	0.755	0.561	0.960	0.547
	MM-6mAPred	0.834	0.689	0.958	0.710
	6mA-Finder	0.809	0.636	0.928	0.690
	i6mA-vote	0.882	0.774	**0.961**	0.803
Arabidopsis	Meta-i6mA	0.787	0.600	0.636	0.936
	i6mA-Fuse_FV	0.749	0.542	0.545	**0.949**
	i6mA-Fuse_RC	0.757	0.534	0.615	0.897
	i6mA-stack_FV	0.770	0.570	0.604	0.933
	i6mA-stack_RC	0.751	0.514	0.634	0.865
	i6mA-Pred	0.730	0.462	0.679	0.780
	iDNA6mA-Rice	0.734	0.473	0.655	0.812
	MM-6mAPred	0.765	0.531	**0.784**	0.747
	6mA-Finder	0.724	0.448	0.741	0.706
	i6mA-vote	**0.798**	**0.617**	0.666	0.929

Bold values indicate the best performance.

irrelevant and redundant features is reduced so that our model has better predictive performance. Furthermore, for Rosaceae, SN is approximately equal to SP and greater than 0.9, indicating that our model has a good discrimination between 6mAs and non-6mAs in the same plant family. For Rice, the SN is greater than 0.9, while the SP is less than 0.9, which may be due to a strong similarity between Rice sequences and Rosaceae positive sequences, resulting in a high false-positive rate and a low true-negative rate when the model recognizes Rice. The situation for Arabidopsis is contrary to that for Rice. It may be because the similarity between Arabidopsis sequences and Rosaceae positive sequences is weak, leading to some 6mAs in Arabidopsis being identified as non-6mAs.

CONCLUSION

In this study, a plant cross-species 6mA site recognition model was constructed by ensemble learning. It has been applied on Rosaceae, Rice, and Arabidopsis and achieved good results. In the construction process, a hypothesis was put forward by analyzing the sequence logos of these three plants. The conjecture was that position-based approaches were more suitable for extracting information from the sequences in our datasets. Next, the hypothesis was verified by comparing different models and observing the tSNE visualization. Then, one-hot encoding for dinucleotide was chosen to represent the datasets by contrasting two nucleotide position-based feature extraction methods and their fusion. Finally, several well-performed models were integrated to form the final classifier by majority voting. To simulate a realistic prediction task, the model was trained on Rosaceae and tested on Rosaceae, Rice, and Arabidopsis. The experimental results showed that our model was adept at predicting the 6mA sites in homologous and heterologous species. In addition, it was also found that there might be a strong similarity between Rice sequences and Rosaceae positive sequences, and the similarity between Arabidopsis sequences and Rosaceae positive sequences is weak. The comparison with other models also showed the superiority of our model. In summary, i6mA-vote outperformed other concerned methods in predicting 6mA sites in the plant genomes. Meanwhile, our research also has the limitation that only three plants were considered. Therefore, future studies will focus on the 6mA site formation characteristics of more plants.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/zhaozhengnan/i6mA-vote/tree/master, github.

AUTHOR CONTRIBUTIONS

ZXT improved the model, designed experiments and drafted the manuscript. ZZ proposed the initial idea and implemented the experiments. YL prepared all datasets for experiments. ZT analyzed experimental results. MG revised the manuscript. QL designed experiments and revised the manuscript. GW conceived the whole research process and revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This manuscript was sponsored by National Natural Science Foundation of China (Grant Nos. 61901103, 61801432, and 61771165), Natural Science Foundation of Heilongjiang Province (Grant No. LH2019F002) and Postdoctoral Science Foundation of Heilongjiang Province of China (Grant No. LBH-Z19106).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.845835/full#supplementary-material

REFERENCES

Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 711–720. doi: 10.1109/34.598228

Bengio, Y., and Glorot, X. (2010). "Understanding the difficulty of training deep feed forward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, (Sardinia: Italy), 249–256.

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA: Association for Computing Machinery).

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 35, 2796–2800. doi: 10.1093/bioinformatics/btz015

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Davis, B. M., Chao, M. C., and Waldor, M. K. (2013). Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.* 16, 192–198. doi: 10.1016/j.mib.2013.01.011

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.-C. (2019). iDNA6mA-PseKNC: identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 111, 96–102. doi: 10.1016/j.ygeno.2018.01.005

Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., et al. (2015). N6-methyldeoxyadenosine marks active transcription start sites in *Chlamydomonas*. *Cell* 161, 879–892. doi: 10.1016/j.cell.2015.04.010

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., et al. (2015). DNA methylation on N6-adenine in *C. elegans*. *Cell* 161, 868–878. doi: 10.1016/j.cell.2015.04.005

Hasan, M. M., Basith, S., Khatun, M. S., Lee, G., Manavalan, B., and Kurata, H. (2021). Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief. Bioinform.* 22:bbaa202. doi: 10.1093/bib/bbaa202

Hasan, M. M., Manavalan, B., Shoombuatong, W., Khatun, M. S., and Kurata, H. (2020). i6mA-Fuse: improved and robust prediction of DNA 6 mA sites in the Rosaceae genome by fusing multiple feature representation. *Plant Mol. Biol.* 103, 225–234. doi: 10.1007/s11103-020-00988-y

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 1026–1034.

Hinton, G. E. (1989). Connectionist learning procedures. *Artif. Intell.* 40, 185–234. doi: 10.1016/0004-3702(89)90049-0

Huang, H., and Gong, X. (2020). A review of protein inter-residue distance prediction. *Curr. Bioinformatics* 15, 821–830. doi: 10.2174/1574893615999200425230056

Khanal, J., Lim, D. Y., Tayara, H., and Chong, K. T. (2021). i6mA-stack: a stacking ensemble-based computational prediction of DNA N6-methyladenine (6mA) sites in the Rosaceae genome. *Genomics* 113(Pt 2), 582–592. doi: 10.1016/j.ygeno.2020.09.054

Kingma, D., and Ba, J. (2014). "Adam: a method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, San Diego, United States.

Kong, L., and Zhang, L. (2019). i6mA-DNCP: computational identification of DNA N6-methyladenine sites in the rice genome using optimized dinucleotide-based features. *Genes* 10:828. doi: 10.3390/genes10100828

Le, N. Q. K. (2019). iN6-methylat (5-step): identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. *Mol. Genet. Genomics* 294, 1173–1182. doi: 10.1007/s00438-019-01570-y

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, Y., Ouyang, X.-H., Xiao, Z.-X., Zhang, L., and Cao, Y. (2020). A review on the methods of peptide-MHC binding prediction. *Curr. Bioinformatics* 15, 878–888. doi: 10.2174/1574893615999200429122801

Liu, Z.-Y., Xing, J.-F., Chen, W., Luan, M.-W., Xie, R., Huang, J., et al. (2019). MDR: an integrative DNA N6-methyladenine and N4-methylcytosine modification database for Rosaceae. *Hortic. Res.* 6:78. doi: 10.1038/s41438-019-0160-4

Lv, H., Dao, F.-Y., Guan, Z.-X., Zhang, D., Tan, J.-X., Zhang, Y., et al. (2019). iDNA6mA-Rice: a computational tool for detecting N6-methyladenine sites in rice. *Front. Genet.* 10:793. doi: 10.3389/fgene.2019.00793

Manavalan, B., Basith, S., Shin, T. H., Wei, L., and Lee, G. (2019). Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol. Ther. Nucleic Acids* 16, 733–744. doi: 10.1016/j.omtn.2019.04.019

Park, S., Wahab, A., Nazari, I., Ryu, J. H., and Chong, K. T. (2020). i6mA-DNC: prediction of DNA N6-methyladenosine sites in rice genome based on dinucleotide representation using deep learning. *Chemometr. Intell. Lab. Syst.* 204:104102. doi: 10.1016/j.chemolab.2020.104102

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pian, C., Zhang, G., Li, F., and Fan, X. (2019). MM-6mAPred: identifying DNA N6-methyladenine sites based on Markov model. *Bioinformatics* 36, 388–392. doi: 10.1093/bioinformatics/btz556

Platt, J. (2000). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, Vol. 10, eds A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (Cambridge, MA: MIT Press).

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100. doi: 10.1093/nar/18.20.6097

Shao, J., and Liu, B. (2021). ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm. *Brief. Bioinform.* 22:bbaa192. doi: 10.1093/bib/bbaa192

Smolarczyk, T., Roterman-Konieczna, I., and Stapor, K. (2020). Protein secondary structure prediction: a review of progress and directions. *Curr. Bioinformatics* 15, 90–107. doi: 10.2174/1574893614666191017104639

Tahir, M., Tayara, H., and Chong, K. T. (2019). iDNA6mA (5-step rule): identification of DNA N6-methyladenine sites in the rice genome by intelligent computational model via Chou's 5-step rule. *Chemometr. Intell. Lab. Syst.* 189, 96–101. doi: 10.1016/j.chemolab.2019.04.007

Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22, 1536–1537. doi: 10.1093/bioinformatics/btl151

van der Maaten, L. J. P., and Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Vanyushin, B. F., Belozersky, A. N., Kokurina, N. A., and Kadirova, D. X. (1968). 5-Methylcytosine and 6-methylaminopurine in bacterial DNA. *Nature* 218, 1066–1067. doi: 10.1038/2181066a0

Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt independence criterion. *Neurocomputing* 383, 257–269. doi: 10.1016/j.neucom.2019.11.103

Wang, J., Shi, Y., Wang, X., and Chang, H. (2020). A drug target interaction prediction based on LINE-RF learning. *Curr. Bioinformatics* 15, 750–757. doi: 10.2174/1574893615666191227092453

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Wion, D., and Casadesús, J. (2006). N6-methyl-adenine: an epigenetic signal for DNA–protein interactions. *Nat. Rev. Microbiol.* 4, 183–192. doi: 10.1038/nrmicro1350

Xu, H., Hu, R., Jia, P., and Zhao, Z. (2020). 6mA-Finder: a novel online tool for predicting DNA N6-methyladenine sites in genomes. *Bioinformatics* 36, 3257–3259. doi: 10.1093/bioinformatics/btaa113

Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., and Xie, Z. (2017). MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucleic Acids Res.* 45, D85–D89. doi: 10.1093/nar/gkw950

Yu, H., and Dai, Z. (2019). SNNRice6mA: a deep learning method for predicting DNA N6-methyladenine sites in rice genome. *Front. Genet.* 10:1071. doi: 10.3389/fgene.2019.01071

Zhang, D., Chen, H.-D., Zulfiqar, H., Yuan, S.-S., Huang, Q.-L., Zhang, Z.-Y., et al. (2021). iBLP: an XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Math. Methods Med.* 2021:6664362. doi: 10.1155/2021/6664362

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., et al. (2015). N6-methyladenine DNA modification in *Drosophila*. *Cell* 161, 893–906. doi: 10.1016/j.cell.2015.04.018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 14 February 2022
doi: 10.3389/fpls.2022.818206

A High-Quality, Chromosome-Level Genome Provides Insights Into Determinate Flowering Time and Color of Cotton Rose (*Hibiscus mutabilis*)

*Yuanzhao Yang[1], Xiaodan Liu[1], Xiaoqing Shi[1], Jiao Ma[1], Xinmei Zeng[1], Zhangshun Zhu[1], Fangwen Li[1], Mengyan Zhou[2], Xiaodan Guo[2] and Xiaoli Liu[1]**

[1] Chengdu Botanical Garden, Chengdu, China, [2] Novogene Bioinformatics Institute, Beijing, China

OPEN ACCESS

Edited by:
Surya Saha,
Boyce Thompson Institute,
United States

Reviewed by:
Kun_Li Xiang,
Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences (CAAS), China
Wenjuan Yu,
Humboldt University of Berlin, Germany
Naama Menda,
Boyce Thompson Institute,
United States

***Correspondence:**
Xiaoli Liu
673911390@qq.com

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: 19 November 2021
Accepted: 20 January 2022
Published: 14 February 2022

Citation:
Yang Y, Liu X, Shi X, Ma J, Zeng X, Zhu Z, Li F, Zhou M, Guo X and Liu X (2022) A High-Quality, Chromosome-Level Genome Provides Insights Into Determinate Flowering Time and Color of Cotton Rose (Hibiscus mutabilis). Front. Plant Sci. 13:818206. doi: 10.3389/fpls.2022.818206

Hibiscus mutabilis (cotton rose) is a deciduous shrub or small tree of the Malvaceae family. Here, we report a chromosome-scale assembly of the *H. mutabilis* genome based on a combination of single-molecule sequencing and Hi-C technology. We obtained an optimized assembly of 2.68 Gb with a scaffold N50 length of 54.7 Mb. An integrated strategy of homology-based, *de novo*, and transcriptome-based gene predictions identified 118,222 protein-coding genes. Repetitive DNA sequences made up 58.55% of the genome, and LTR retrotransposons were the most common repetitive sequence type, accounting for 53.15% of the genome. Through the use of Hi-C data, we constructed a chromosome-scale assembly in which Nanopore scaffolds were assembled into 46 pseudomolecule sequences. We identified important genes involved in anthocyanin biosynthesis and documented copy number variation in floral regulators. Phylogenetic analysis indicated that *H. mutabilis* was closely related to *H. syriacus*, from which it diverged approximately 15.3 million years ago. The availability of cotton rose genome data increases our understanding of the species' genetic evolution and will support further biological research and breeding in cotton rose, as well as other Malvaceae species.

Keywords: *Hibiscus mutabilis*, genome, Hi-C, phylogenetic affiliation, floral regulators

INTRODUCTION

Hibiscus mutabilis is one of the most popular tree species in the Malvaceae family, which includes species such as *Gossypium raimondii* and *Hibiscus syriacus* (Rose of Sharon). Some members of the Malvaceae have relatively high economic value. For example, cotton is the largest source of natural textile fibers in the world, and over 90% of its annual fiber production comes from allotetraploid cotton (*G. hirsutum* and *G. barbadense*) (Wang et al., 2019). Additionally, many Malvaceae species are used as ornamentals because of their flowers. *H. syriacus* is an important horticultural species whose attractive white, pink, red, lavender, or purple flowers are displayed over a long bloom period, although individual flowers last only a day in the landscape (Kim et al., 2016). This study of *H. mutabilis* ($2n = 92$) (Li et al., 2015) focuses primarily on its ornamental characters, including

its flower colors, long bloom time, and floral development and morphogenesis. In addition to its ornamental value, *H. mutabilis* is also an ingredient in local herbal remedies. It is thought to cool the blood, relieve toxins, reduce swelling, and alleviate pain, and it has long been used in the treatment of ulcers, swelling, herpes zoster, scalding, bruises, etc. (Liu et al., 2015). The complete genome sequence of an organism provides a large amount of information for subsequent biological studies (Yang et al., 2005). The *H. mutabilis* genome sequencing project is therefore extremely valuable for breeding, comparative genomics research, and other activities.

H. mutabilis has been cultivated for more than 2,000 years south of the Yangtse River; it is also the city flower of Chengdu and has great significance for the city. Commonly used as an ornamental species, its attractive purple, red, pink, or white flowers are displayed over a long bloom period (3–4 months or more), although its individual flowers last for less than 48 h (**Figure 1**). During flower development, the floral color of some varieties shows little change, but that of other varieties undergoes a marked change from white to pink within a single day. This interesting dynamic phenomenon can be seen in the cultivars 'Drunk girl' and 'Bairihuacai' and occurs during the process of individual flower development, unlike the color differences found in distinct cultivars of chrysanthemum (Ohmiya et al., 2006), *Narcissus pseudonarcissus* (Li et al., 2018), or *Brassica napus* (Zhang et al., 2015).

The development of genomic resources and molecular breeding technologies holds promise for targeted character improvement of *H. mutabilis* in the near future. Recently, a 1.75 Gb draft genome of *H. syriacus* was assembled, and a chromosome-scale genome of *H. cannabinus* was published in 2020 (Zhang et al., 2020). Many breeding systems and novel varieties have been produced using traditional methods to meet horticultural requirements, but a completed genomic sequence will accelerate the breeding of cotton rose.

The many flower colors of cotton rose give it high ornamental value and reflect the complexity of the underlying flavonoid metabolic pathway. One endpoint of flavonoid biosynthesis is the production of anthocyanins, pigments that produce the colors of many flowers, fruits, and other plant tissues (Koes et al., 1994). Chalcone synthase (CHS), chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), dihydroflavonol reductase (DFR), anthocyanidin synthase (ANS), and flavonoid 3-O-glucosyltransferase (UFGT) all function in the synthesis of anthocyanins and anthocyanidins, their aglycone counterparts. CHS represents the first committed step in the flavonoid pathway (Meer et al., 1993). The second step is performed by CHI, which acts on the yellow naringenin chalcone product of CHS, catalyzing its isomerization to the colorless flavanone naringenin (Moustafa and Wong, 1967). Dihydroflavonols are subsequently reduced to leucoanthocyanidins by DFR (Durbin et al., 2004). ANS catalyzes the formation of cyanidin from leucoanthocyanidin and is the penultimate step in the biosynthesis of the anthocyanin class of flavonoids (**Figure 2**). Despite recent progress in understanding *H. mutabilis* anthocyanidin biosynthesis, the lack of a genome sequence has hampered efforts to elucidate the molecular and genetic determinants of this trait, which underlies the dynamic phenomenon of flower color development. Genome and transcriptome sequences are needed in order to fully analyze the molecular mechanisms of anthocyanidin biosynthesis.

In the present study, we generated a reference genome for *H. mutabilis* using a combination of single-molecule sequencing and Hi-C technology. We identified functional genes involved in the biosynthesis of anthocyanins based on homology searches and functional annotations. We also investigated copy number variation in floral regulators among multiple species to gain insight into the evolution of flowering phenotypes in *H. mutabilis*. The genomic resources developed here will be useful for further experimentation, cultivation, and breeding of *H. mutabilis* and other Malvaceae species.

MATERIALS AND METHODS

Plant Materials and Whole-Genome Sequencing

The *H. mutabilis* material sequenced in this study was the stably heritable single-petal white color cultivars, which is cultivated in the nursery of the Chengdu Botanical Garden (CDBG), Sichuan, China. The breeding system of *H. mutabilis* belongs to allogamy. Seeds of 'single-petal white' were collected in the laboratory of the CDBG. Young leaves (∼3 cm width) were harvested to extract high-quality DNA for Illumina and Oxford Nanopore Technology (ONT) sequencing. For transcriptome sequencing, petals were manually collected from three color cultivars ('single-petal white,' 'single-petal pink,' and 'Purple silk') and at three stages of color development in 'Drunk girl' (white, blended white and pink, and fully pink). Flowers at the same stage from individual *H. mutabilis* plants were pooled and divided into three samples. These samples were immediately frozen in liquid nitrogen and then used for RNA sequencing.

High-quality *H. mutabilis* genomic DNA was extracted from young leaves with a DNA secure Plant Kit (TIANGEN, China) and used to construct long-read libraries for the ONT platform.[1] Libraries were prepared following the ONT'1D Genomic DNA by Ligation (Kit 9 chemistry)-PromethION' protocol and sequenced using the PromethION protocol. In addition, high-quality DNA was broken into random fragments, and an Illumina paired-end library was constructed with an insert size of 350 bp and sequenced using the Illumina HiSeq X Ten platform.

For Hi-C sequencing, leaves were fixed with 1% formaldehyde solution in MS buffer (10 mM potassium phosphate, pH 7.0; 50 mM NaCl; 0.1 M sucrose) at room temperature for 30 min in a vacuum. After fixation, the leaves were incubated at room temperature for 5 min under a vacuum in MC buffer with 0.15M glycine. Approximately 2 g of fixed tissue was homogenized with liquid nitrogen, resuspended in nuclei isolation buffer, and filtered with a 40-nm cell strainer. The procedures for enriching nuclei from flow through and subsequent denaturation were performed according to a 3C protocol. The chromatin extraction procedures were similar to those described previously. In brief,

[1] https://nanoporetech.com

FIGURE 1 | *Hibiscus mutabilis* floral morphology. **(A)** Dynamic change in the flower color of *H. mutabilis* f. *mutabilis* 'Drunk girl' from white to pink. **(B)** A variety of colors found in different color cultivars of *H. mutabilis*, and from left to right in turn are *H. mutabilis* 'Single-Petal Pink,' *H. mutabilis* 'Single-Petal White' and *H. mutabilis* 'Purple silk.'

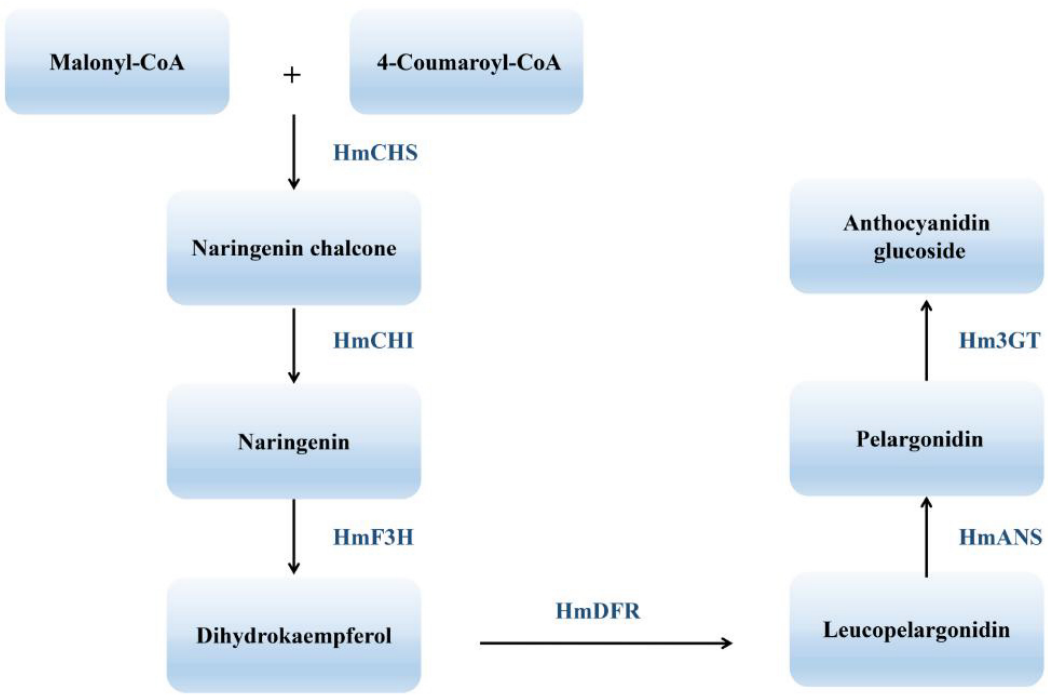


FIGURE 2 | An abbreviated diagram of the flavonoid pathway that produces anthocyanins. CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; DFR, dihydroflavonol 4-reductase; ANS, anthocyanidin synthase; 3GT, 3-O-glucosyl transferase.

chromatin was digested for 16 h with 400 U *Hin*dIII restriction enzyme (NEB) at 37°C. DNA ends were labeled with biotin and incubated at 37°C for 45 min, and the enzyme was inactivated with 20% SDS solution. DNA ligation was performed by the addition of T4 DNA ligase (NEB) and incubation at 16°C for 4–6 h. After ligation, proteinase K was added for reverse cross-linking during overnight incubation at 65°C. DNA fragments were purified and dissolved in 86 μL of water, and unligated ends were then removed. Purified DNA was fragmented to a size of 300–500 bp, and DNA ends were repaired. Finally, DNA fragments labeled with biotin were separated on Dynabeads M-280 Streptavidin (Life Technologies). Hi-C libraries were assessed for quality and sequenced on an Illumina HiSeq X Ten sequencer.

Genome Assembly and Chromatin Interaction Analysis Using Hi-C Technology

De novo assembly of all Nanopore long reads was performed using wtdbg2 v2.5 (Ruan and Li, 2020). Because Nanopore reads contain systematic errors in homopolymeric regions, we polished the consensus assembly three times using the Nanopore reads as input to Racon v1.3.1 (Vaser et al., 2017) and then three additional times using Illumina reads as input to Pilon v1.22 (Walker et al., 2014).

The Hi-C data were mapped to the original scaffold genome using BWA v0.7.7 (Li, 2009) and only reads with unique alignment positions were extracted to construct a chromosome-scale assembly using LACHESIS v201701 (Burton et al., 2013).

We used both CEGMA (Core Eukaryotic Gene Mapping Approach) (Parra et al., 2007; **Supplementary Table 4**) and BUSCO (Benchmarking Universal Single-Copy Orthologs) (Simão et al., 2015; **Supplementary Table 5**) to evaluate the completeness of the assembly.

Genome Annotation

TEs were identified in the genome assembly at both the DNA and protein levels. We used RepeatModeler, RepeatScout (Tarailo-Graovac and Chen, 2009), Piler (Edgar and Myers, 2005), and LTR_FINDER (Xu and Wang, 2007) to develop a *de novo* TE library. RepeatMasker (Tarailo-Graovac and Chen, 2009) was used for DNA-level identification with Repbase and the *de novo* TE library. Tandem repeats were identified using Tandem Repeats Finder (Benson, 1999). At the protein level, RepeatProteinMask (Tarailo-Graovac and Chen, 2009) was used to conduct WU-BLASTX searches against the TE protein database. Overlapping TEs that belonged to the same type of repeat were integrated together.

We used homology-based, *de novo*, and transcriptome-based approaches to predict protein-coding genes in the *H. mutabilis* genome. For homolog-based prediction, sequences of homologous proteins from six plants (*A. thaliana, C. capsularis, D. zibethinus, G. raimondii, H. umbratica, H. syriacus,* and *T. cacao*) were downloaded from Ensembl, NCBI, or JGI. Protein sequences were aligned to the genome using TBLASTN with an *E*-value cutoff of 1×10^{-5}. The blast hits were concatenated using solar (Yu et al., 2007). For each blast hit, GeneWise v2.4.1 (Birney and Clamp, 2004) was used to predict the exact gene structure in the corresponding genomic regions. The five *ab initio* gene prediction programs AUGUSTUS v3.0.2 (Stanke and Morgenstern, 2005), Genescan v1.0 (Aggarwal and Ramaswamy, 2002), GeneID (Parra et al., 2000), GlimmerHMM v3.0.2 (Majoros et al., 2004), and SNAP (Korf, 2004) were used for *de novo* protein prediction. To further optimize the genome annotation, RNA-seq data from floral, leaf, and stem tissues were aligned to the *H. mutabilis* genome using TopHat v2.0.13 (Trapnell et al., 2009) to identify exon regions and splice junctions. The alignment results were then used as input for Cufflinks v2.1.1 (Trapnell et al., 2010) in order to assemble transcripts into gene models. Trinity (Grabherr et al., 2011) was used with default parameters to assemble the RNA-seq data, and PASA (Haas et al., 2003) was used to improve the gene structures. A weighted and non-redundant gene set was generated by EVidenceModeler (EVM) (Haas et al., 2008), which merged all gene models predicted using the three approaches above. PASA adjusted the gene models generated by EVM based on information from the transcriptome assembly.

The functional annotation of protein-coding genes was evaluated by BLASTP (*E*-value $\leq 1 \times 10$-5) against two integrated protein sequence databases, Swiss-Prot (Bairoch and Apweiler, 2000) and the NCBI non-redundant (NR) database. Protein domains were annotated using InterProScan v4.8 to search InterPro v32.0 (Mulder and Apweiler, 2007), which includes the Pfam, PRINTS, PROSITE, ProDom, and SMART databases. Gene Ontology (GO) (Ashburner et al., 2000) terms for each gene were obtained from the corresponding InterPro descriptions. Putative pathway assignments for each gene were obtained by blasting against the KEGG (Kanehisa and Goto, 2006) database with an *E*-value cutoff of 1×10^{-5}.

tRNA genes were predicted by tRNAscan-SE (Lowe and Eddy, 1997), and miRNA and snRNA fragments were identified using Infernal (Nawrocki et al., 2009) with the Rfam (Griffiths-Jones et al., 2005) database. rRNA genes were identified using BLASTN (*E*-value $\leq 1 \times 10^{-10}$) against the plant rRNA database.

Genome Evolution Analysis

First, nucleotide and protein data from nine species (*A. trichopoda, A. thaliana, B. ceiba, C. capsularis, D. zibethinus, G. raimondii, H. syriacus, R. chinensis,* and *T. cacao*) were downloaded from Ensembl, NCBI, and JGI. The longest transcript was selected from the alternatively spliced transcripts of each gene, and genes with $\leq$ 50 amino acids were removed. Nucleotide and protein data from *H. mutabilis* and the other nine angiosperms were clustered into orthologous groups using BLASTP and OrthoMCL v2.0.9, and an MCL inflation of 1.5 was used as the cluster granularity setting (Li et al., 2003). A phylogenetic tree was constructed using shared single-copy orthologs. Protein sequences of the orthologs were aligned using MUSCLE (Edgar, 2004), and the protein alignments were transformed to CDS alignments. We then concatenated the CDS alignments into a "supermatrix" from which the phylogenetic tree was constructed using the maximum likelihood (ML) TREE algorithm in RAxML v8.1.13 (Stamatakis, 2006) with the best-scoring protein substitution model (GTRGAMMA)

and 1,000 bootstrap replicates. The MCMCtree program in the PAML package (Yang, 1997) was used to estimate divergence times among the ten species. Three fossil calibration points were used for restraining the age of the nodes: 23–48 Mya (Million years ago) for the MRCA of *T. cacao–G. raimondii,* 65–107 Mya for the MRCA of *G. raimondii–A. thaliana* (Wang et al., 2012), and 103–109 Mya for the MRCA of Malvales–Rosales (Wikström et al., 2001). CAFE was used to identify expansions and contractions within orthologous gene families by comparing cluster size differences between the ancestor and other species (De Bie et al., 2006). To estimate the synonymous substitutions per synonymous site (Ks), all paralogous gene pairs were analyzed with the ML method in PAML (Yang, 1997). MCscan (Tang et al., 2008) was used to analyze genome collinearity in *H. mutabilis*.

Identification of Nucleotide-Binding Site-Encoding Genes

To identify NBS-encoding genes, representative genes from each plant genome were screened using a raw Hidden Markov Model (HMM3.0) (Marchin et al., 2005) to search for the Pfam NBS family PF00931 domain with an *E*-value cut-off of 1.0. All putative NBS protein sequences were analyzed and manually curated based on a TBLASTN search against known R gene sequences in GenBank. To further identify TIR homologs and sequences that encoded CC and LRR motifs, candidate NBS-LRR protein sequences were characterized using SMART (Schultz et al., 1998), the Pfam database (Finn et al., 2013), and the COILS program (Lupas et al., 1991) with a threshold of 0.9 to specifically detect the CC domain.

Transcriptome Sequencing

For analysis of flowering gene(s), petals were manually collected from three color cultivars at the same time (2–3 p.m.) and at three stages of color development in 'Drunk girl' (white at the 9 a.m., blended white and pink at the 12 a.m., and fully pink at 6 p.m.) and these samples were immediately frozen in liquid nitrogen and then used for RNA sequencing, and total RNA was extracted using an RNAprep Pure Plant Kit (TIANGEN, China). The quality and quantity of the RNA samples were evaluated using a NanoPhotometer (Implen, CA, United States), a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, United States), and an Agilent 2,100 Bioanalyzer (Agilent Technologies, United States). All RNA samples with integrity values greater than 7.0 were used for cDNA library construction and sequencing. The cDNA libraries were prepared using the NEB Next Ultra RNA Library Prep Kit (E7350L, NEB, United States), and 150-bp paired-end sequencing was performed on the Illumina NovaSeq 6000 platform (Illumina, CA, United States).

RESULTS

Genome Sequencing and Assembly

We assembled the *H. mutabilis* genome using a combination of Illumina HiSeq X Ten and Oxford Nanopore PromethION sequencing. We generated 315.22 Gb (104-fold coverage) of raw 150-bp paired-end Illumina reads and 469.91 Gb (155-fold coverage) of raw Nanopore reads. The genome size was estimated to be 3032.98 Mb based on the 17-mer depth distribution (**Supplementary Table 1** and **Supplementary Figure 1**). Nanopore long reads were assembled into contigs and scaffolds using wtdbg2 v2.5,13 resulting in a final assembly of 2.68 Gb with 5,464 contigs and a contig N50 of 2.22 Mb (**Supplementary Table 2**). Its GC percentage was 35.36%, similar to that of the *H. syriacus* genome (34.04%). In total, 363.5 Gb of clean reads were obtained from Hi-C sequencing (over 121-fold coverage). We used these data to construct chromosome-scale scaffolds, resulting in a total of 5,598 contigs, with a scaffold N50 of 54.70 Mb and a total length of 2,676,237,573 bp (**Supplementary Table 10** and **Figure 3A**).

Next, The clustering of contig by hierarchical clustering of the Hi-C data was performed. Hi-C linkage was used as a criterion to measure the degree of tightness of the association between different contigs by standardizing the digestion sites of *Dpn*II on the genome sketch. The contigs were assembled into 46 pseudo-chromosomes using LACHESIS package tools. The Illumina paired-end reads were mapped to the assembled genome to assess assembly accuracy, resulting in a 98.81% mapping rate (**Supplementary Table 3** and **Figure 3B**). The genome assembly captured 96.77% of the core eukaryotic genes from CEGMA18 (**Supplementary Table 4**) and 92.6% of the Embryophyta OrthoDB gene set in BUSCO19 (**Supplementary Table 5**), indicating a high level of completeness.

Genome Annotation

We identified 1.56 Gb of non-redundant repetitive elements, representing approximately 55.85% of the *H. mutabilis* genome assembly. Because long terminal repeat retrotransposons (LTR-RTs) typically make a significant contribution to large genome size (Zhao and Ma, 2013), we estimated LTR-RT insertion time in *H. mutabilis*. We identified a round of LTR-RT burst approximately 2.5 million years ago (Mya), especially for the Ty3/Gypsy-del and Ty1/Copia-Retrofit families (**Supplementary Table 6**). The transposable elements (TEs) were primarily long terminal repeats (LTRs), which accounted for approximately 53% of the genome (**Supplementary Figure 2**).

We used *de novo* and homology-based gene prediction approaches and combined their results to annotate 118,222 protein-coding genes in the *H. mutabilis* genome. The average transcript length was 2,466.97 bp, with an average of 4.53 exons per gene and an average exon length of 218.86 bp. Compared with other model plants and Malvaceae species, the *H. mutabilis* genome contained a larger number of genes: *H. syriacus* (82,827 genes), *Arabidopsis thaliana* (26,869), *Theobroma cacao* (29,144), *G. raimondii* (35,526), *Corchorus capsularis* (29,356), *Herrania umbratica* (29,262), and *Durio zibethinus* (63,819) (**Supplementary Table 7**).

In addition to RNA-coding genes, we also identified 827 mature microRNAs (miRNAs), 3,604 transfer RNAs (tRNAs), 3,423 ribosomal RNAs (rRNAs), and 9,370 small nuclear RNAs (snRNAs) in the *H. mutabilis* genome (**Supplementary Table 11**).

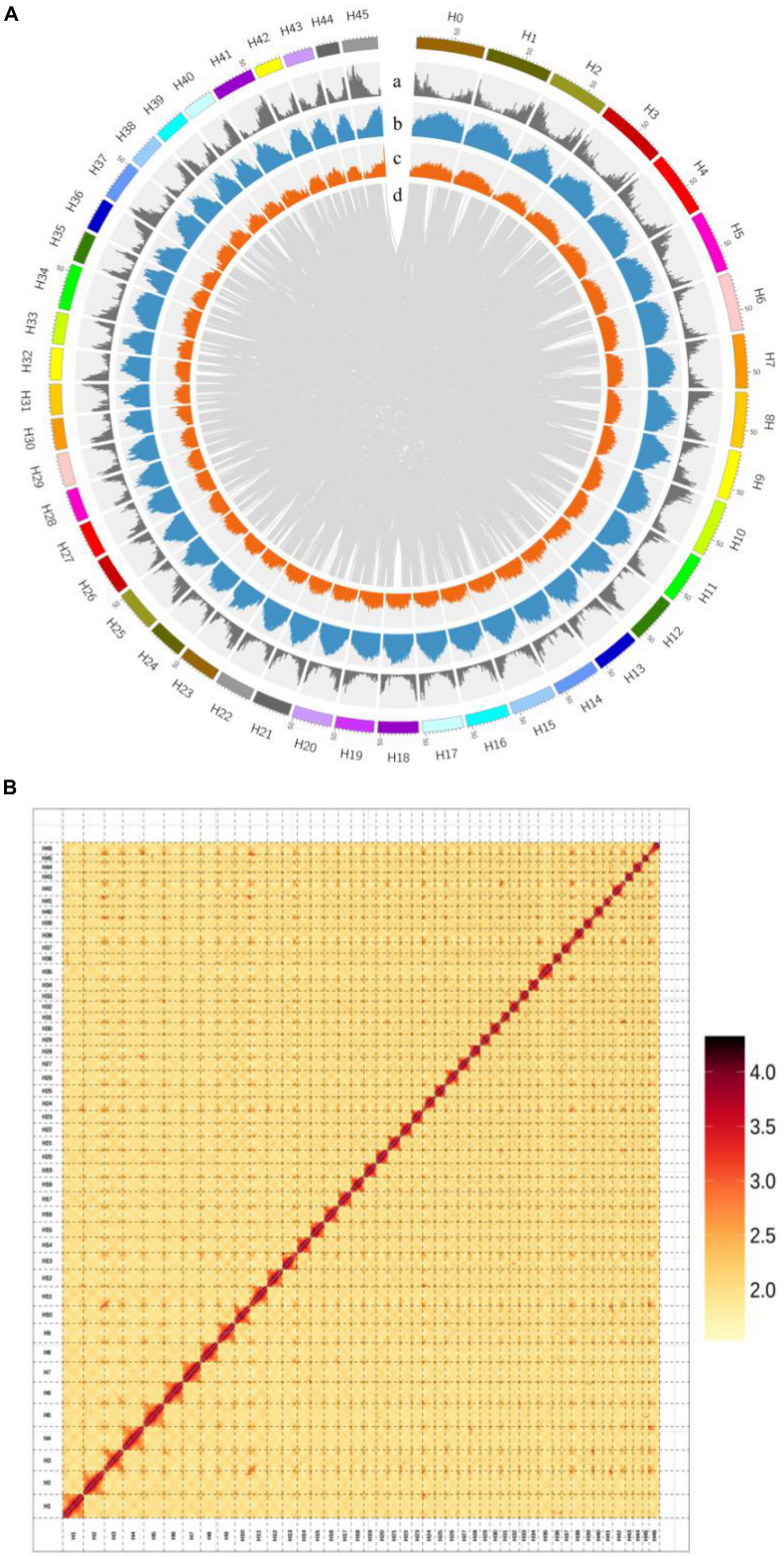


FIGURE 3 | (A) Chromosomal features of *H. mutabilis*. (a) Gene density; (b) Repeat density; (c) AT content; (d) Syntonic blocks. **(B)** Hi-C map of the *H. mutabilis* genome showing genome-wide all-by-all interactions. The map shows a high resolution of individual chromosomes that were scaffolded and assembled independently. Color intensity indicates the frequency of Hi-C interaction links from low (yellow) to high (red).

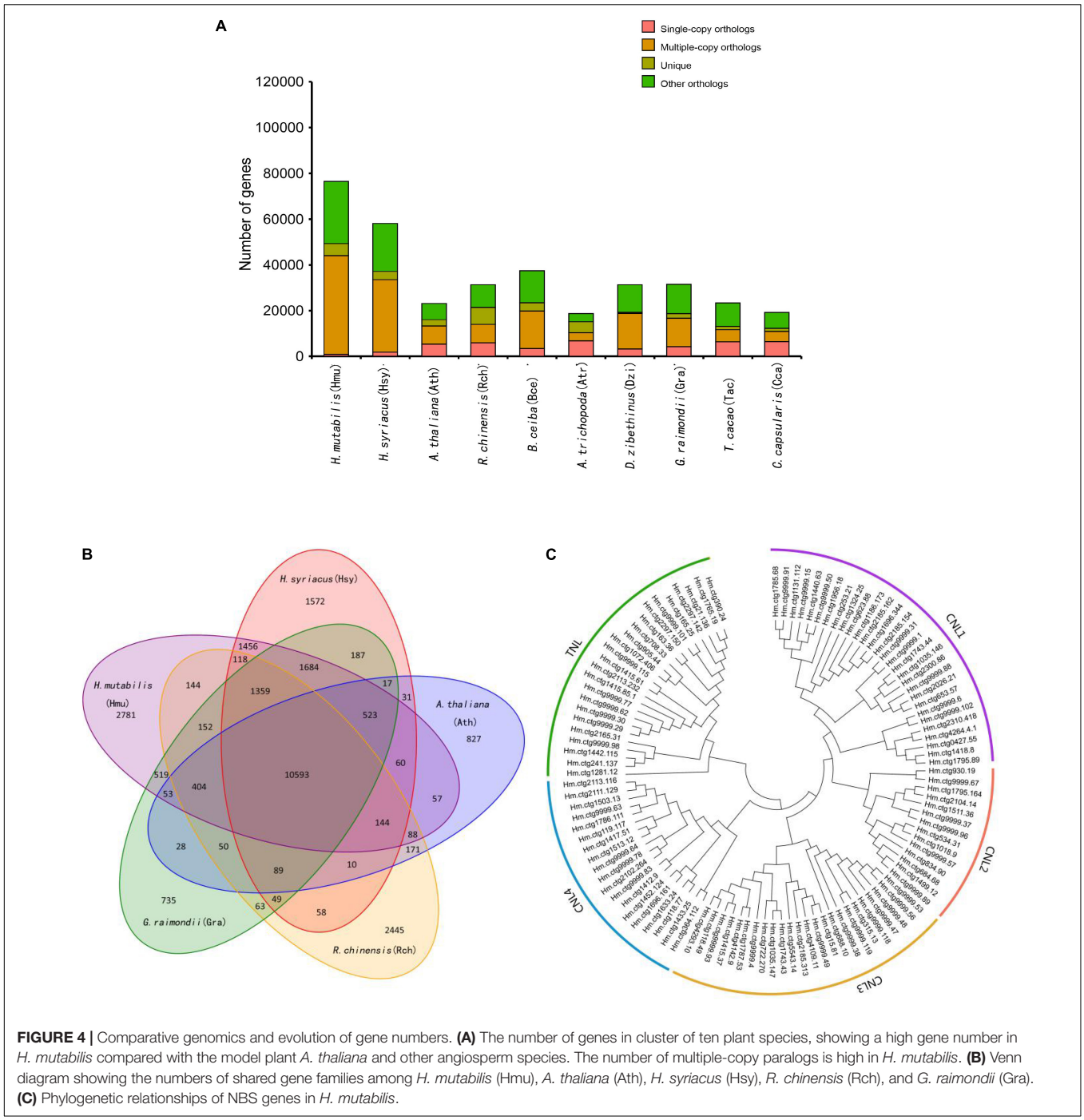


FIGURE 4 | Comparative genomics and evolution of gene numbers. **(A)** The number of genes in cluster of ten plant species, showing a high gene number in *H. mutabilis* compared with the model plant *A. thaliana* and other angiosperm species. The number of multiple-copy paralogs is high in *H. mutabilis*. **(B)** Venn diagram showing the numbers of shared gene families among *H. mutabilis* (Hmu), *A. thaliana* (Ath), *H. syriacus* (Hsy), *R. chinensis* (Rch), and *G. raimondii* (Gra). **(C)** Phylogenetic relationships of NBS genes in *H. mutabilis*.

The probable functions of the predicted genes were assessed by searching against public databases, including Swiss-Prot, NR, InterPro, and KEGG of 118,222 predicted genes in the *H. mutabilis* genome, 113,821 (96.3%) were assigned potential functions as a result of these database searches (**Supplementary Table 8**).

Genome Evolution

Although morphological investigations have placed *H. mutabilis* in the Malvaceae family, there is still no phylogenomic analysis of its evolutionary position within the family based on whole-genome data. Here, we compared the *H. mutabilis* genome with the genome sequences of nine other angiosperm plants (*H. syriacus*, *A. thaliana*, *Rosa chinensis*, *Bombax ceiba*, *Amborella trichopoda*, *G. raimondii*, *T. cacao*, and *C. capsularis*). Orthologous protein groups were identified within the genomes, yielding a total of 30,208 gene families and 198 single-copy orthologs across ten species. There were 2,781 gene families specific to *H. mutabilis*, and 10,593 gene families were shared among all species investigated (**Figure 4**). We detected 4,558

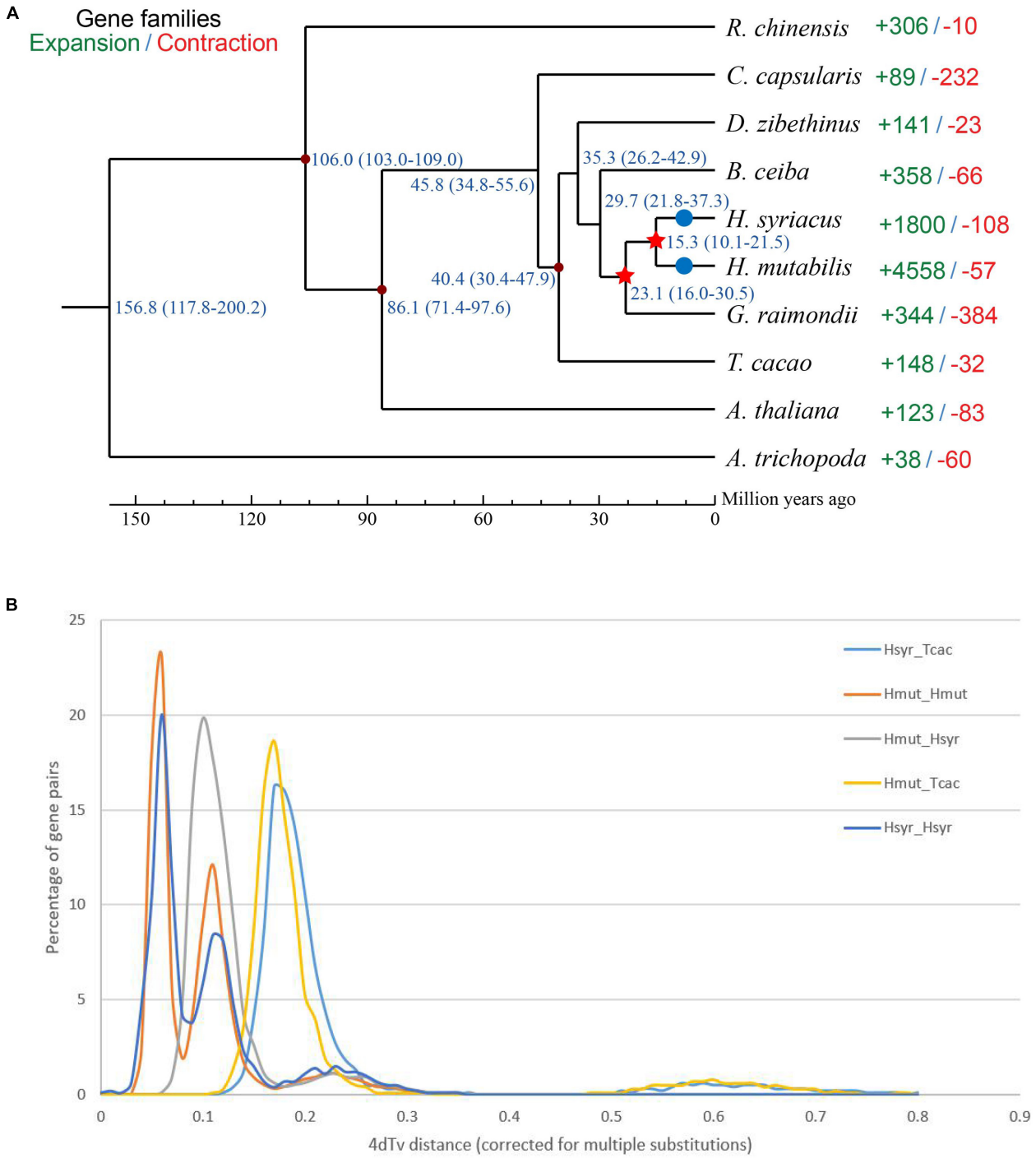


FIGURE 5 | (A) Phylogenetic tree showing the close relationship between *H. mutabilis* and *H. syriacus*. Numbers outside the parentheses represent the divergence times of species (Mya), and numbers inside the parentheses indicate the confidence intervals. The red points indicated calibration points, and the red stars indicate WGD events and blue circles indicate diploidization events. **(B)** Distribution of 4Dtv distances.

gene families expansion when *H. mutabilis* and *H. syriacus* have diverged (**Figure 5C**).

We constructed a phylogenetic tree based on single-copy genes using PAML and estimated the divergence times among the 10 species. The Malvaceae family appeared to have diverged from a Tiliaceae–Malvaceae most recent common ancestor (MRCA) approximately 45.8 (34.8–55.6) Mya, and the *Hibiscus-Gossypium* divergence was estimated at 23.1 (16.0–30.5) Mya.

TABLE 1 | Copy numbers of genes encoding flowering time regulators in five plant species.

Gene	*Arabidopsis* locus	Copy number				
		H. mutabilis	*H. syriacus*	*A. trichopoda*	*T. cacao*	*G. raimondii*
CO	AT5G15840	8	9	7	3	2
ELF4	AT2G40080	9	12	2	1	5
FCA	AT4G16280	4	0	1	2	1
FKE1	AT1G68050	0	3	1	1	2
FLK	AT3G04610	5	4	1	1	3
GI	AT1G22770	22	15	5	5	7
LFY	AT5G61850	4	4	1	1	1
LHY	AT1G01060	0	0	1	0	0
VIN3	AT5G57380	0	0	1	0	0
SOC1	AT2G45660	15	12	4	4	6
TFL	AT5G03840	24	13	6	5	7
SVP	AT1G24260	48	33	8	7	17
PHYA	AT1G09570	10	5	3	3	4
PHYB	AT2G18790	10	5	4	3	5
PHYC	AT5G35840	0	1	1	1	4
PHYE	AT4G18130	5	3	3	1	2

TABLE 2 | Numbers and classifications of genes encoding NBS-containing resistance proteins in five plant species.

Protein domain	Letter code	*H. mutabilis*	*H. syriacus*	*T. cacao*	*G. raimondii*	*A. thaliana*
CC-NBS-LRR	CNL	81	183	202	220	52
CC-NBS	CN	32	77	25	24	3
TIR-NBS-LRR	TNL	28	68	14	26	87
TIR-NBS	TN	10	9	3	1	17
NBS-LRR	NL	147	81	34	28	8
NBS	N	192	54	9	4	3
Total		490	472	287	303	170
% of total genes		0.41	0.53	0.97	0.81	0.63

H. mutabilis was most closely related to *H. syriacus*, with an estimated divergence time of approximately 15.3 (10.1–21.5) Mya (**Figure 5A**). In addition to the paleohexaploidization event shared by the eudicots, we observed three additional whole-genome duplication (WGD) events in *H. mutabilis*. *Hibiscus* and *Gossypium* shared a WGD event (13.3–20.0 Mya), *H. mutabilis* and *H. syriacus* shared a WGD event (10.76–21.51 Mya), and *H. mutabilis* and *H. syriacus* each underwent a WGD event (4.61–9.00 Mya) (**Figure 5B**).

Flowering Time and Disease Resistance Genes

Genetic and molecular mechanisms of floral development are highly conserved among different plant species (Schiessl et al., 2014) and include four major flowering pathways that have been well characterized in *A. thaliana*. *H. mutabilis* is similar to *H. syriacus*, a short-day flowering plant with a long bloom period that produces more than 30 blossoms per day. Like those of *H. syriacus*, the flowers of *H. mutabilis* open daily and last for less than 48 h. Because flowering time is frequently dependent on gene copy number (Grover et al., 2015), we investigated the copy numbers of genes involved in the four major flowering pathways in *A. thaliana*, *T. cacao*, *G. raimondii*, *A. trichopoda*, *H. syriacus*, and *H. mutabilis*. Copy numbers of most flowering-related genes were higher in *H. mutabilis* than in other plants, including *T. cacao*, *G. raimondii*, *A. trichopoda*, and *H. syriacus*. In particular, the copy number of the plant-specific nuclear protein GIGANTEA (GI) was three to four times greater in *H. mutabilis* than in *A. trichopoda*, *T. cacao*, or *G. raimondii* (**Table 1**).

Nucleotide-binding site (NBS) and carboxy-terminal LRR domains are found in the majority of R proteins (DeYoung and Innes, 2006; Takken et al., 2006). Based on resistance domain analyses in the *H. mutabilis* genome, a total of 490 NBS-containing resistance genes were identified and classified into six groups: CC-NBS-LRR, CC-NBS, TIR-NBS-LRR, TIR-NBS, NBS-LRR, and NBS. In total, their gene numbers were approximately three times greater in *H. mutabilis* than in *A. thaliana* (170). This trend was particularly striking for the NBS genes, whose numbers were much higher in *H. mutabilis* (192 genes) than in *H. syriacus* (54), *T. cacao* (9), *G. raimondii* (4), and *A. thaliana* (3). Although *H. mutabilis* had the highest number of NBS-containing resistance genes among the five angiosperms (**Table 2**), its number of NBS-containing genes as

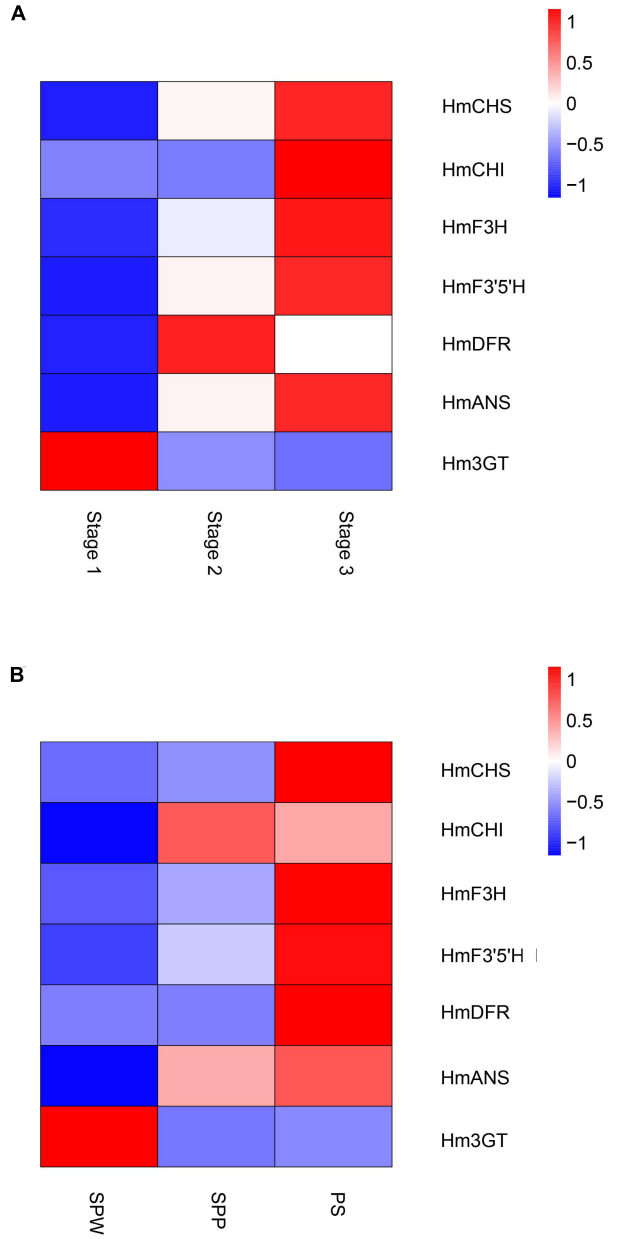


FIGURE 6 | Heatmaps show the expression of key genes in the anthocyanin biosynthesis pathway of *H. mutabilis*. **(A)** Gene expression changes in flowers of *H. mutabilis* 'Drunk girl' as their color transitions from white to fully pink. **(B)** Gene expression in flowers of three different color cultivars: 'single-petal white' (SPW), 'single-petal pink' (SPP), and 'Purple silk' (PS).

a percentage of total genes was the lowest. All six NBS-containing groups existed in each plant genome, but their distributions differed among species.

Transcriptome Sequencing Analysis

Global gene expression patterns were quantified in three stages of the floral color transition of *H. mutabilis* 'Drunk girl': white (Stage 1), blended white and pink (Stage 2), and fully pink (Stage 3). A total of 9,492 genes were up-regulated from Stage 1 to Stage 2, and more than 15,000 genes were up-regulated from Stage 1 to Stage 3. A total of 8,481 genes were down-regulated from Stage 1 to Stage 2, and 15,839 genes were down-regulated from Stage 1 to Stage 3. In particular, we analyzed expression changes in anthocyanin-related genes at the three flower stages and present the results in a heatmap (**Figure 6A**). A number of key anthocyanin biosynthesis-related genes, such as Hmchs, HmchI, and Hmans, increased in expression from Stage 1 to Stage 3, consistent with the pattern of floral color development.

Key anthocyanin biosynthesis-related genes also differed in expression among different color cultivars of *H. mutabilis*, including 'single-petal white,' 'single-petal pink,' and 'Purple silk.' The highest expression levels were generally found in 'Purple silk,' which is a deep purple color form (**Figure 6B**).

DISCUSSION

Completeness and continuity are important indicators of genome assembly quality. In this study, we took advantage of the longer read lengths offered by ONT sequencing that have proven advantageous in the assembly of other plant genomes such as *Solanum pennellii* (Schmidt et al., 2017) and *Chrysanthemum nankingense* (Song et al., 2018). Here, we report the first genome data for *H. mutabilis* and estimate its genome size to be 2.68 Gb, far larger than that of *H. syriacus*. Our *H. mutabilis* genome assembled using Nanopore reads had a contig N50 of 2.02 Mb. We then used Hi-C data to cluster the contigs into forty-six chromosomes with a final scaffold N50 of 54.70 Mb. The genome contained complete copies of 92.6% of the BUSCO orthologs examined. This genome sequence will contribute to our understanding of the biosynthesis of natural products such as anthocyanins, although additional research is needed to directly link specific genes to individual traits. Nonetheless, our high-quality, annotated genome sequence provides insights into determinate flowering time and flower color in *H. mutabilis*.

Compared with the *H. syriacus* genome, the *H. mutabilis* genome was larger and contained more protein-coding genes. *H. mutabilis* and *H. syriacus* share an MRCA approximately 15.3 (10.1–21.5) Mya, and investigation of WGD timing in the *H. mutabilis* genome showed that two WGDs occurred after *H. mutabilis*–*H. syriacus* divergence and *H. mutabilis* speciation. WGD events and tandem duplications are the most important determinants of genome size variation in angiosperms (Piegu et al., 2006; El Baidouri and Panaud, 2013). This recent WGD event not only caused genome expansion in *H. mutabilis*, but may also have contributed to the morphological and physiological diversity of *H. mutabilis*. We inferred that gene losses, which had different frequencies in *H. mutabilis* and *H. syriacus*, made the *H. syriacus* genome smaller than that of *H. mutabilis*. *H. mutabilis* has a long bloom period and high blossom turnover. The copy numbers of most flowering-related genes, such as GI, CONSTANS (CO), and SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1 (SOC1) were higher in *H. mutabilis* than in *T. cacao*, *A. thaliana*, and *H. syriacus*. These results show that

H. mutabilis preserved many copies of flowering-related genes during the transition from a polyploid to a diploid genome.

The dynamic color change from white to pink in *H. mutabilis* 'Drunk girl' flowers is reported to be caused by variations in anthocyanin contents (Liu et al., 2008). Flavonoids are the major molecules involved in plant pigmentation (Lai et al., 2014) and include anthocyanins, flavan-3-ols (catechins and proanthocyanidins), flavanonols, flavonols, flavones, and phenolic acid (Lou et al., 2014). To date, regulation of the flavonoid pathway has been shown to occur primarily at the transcriptional level (Mol et al., 1998). Different species have distinct regulated genes, and these appear to be among the most important candidate genes for flower color determination (Casimiro-Soriguer et al., 2016; Jiao et al., 2020). To investigate the expression of anthocyanin-related genes over the course of flower development and in different color forms, we combined the high-quality genome sequence generated here with RNA-seq data from *H. mutabilis* The expression levels of anthocyanin biosynthetic genes such as Hmchs, HmchI, and Hmans were correlated and increased as flowers transitioned from white to pink. The pink flower color in cotton rose is related to the synthesis of cyanidin-based pigments (Chen et al., 2014), and our results indicate that low CHS, CHI, and ANS expression may inhibit cyanidin production in white flowers. Thus, combined genomic and transcriptomic analysis of *H. mutabilis* flowers indicated that structural genes had important roles in anthocyanin biosynthesis during the transition from white to pink flower coloration. In maize, an MYB-related protein and a bHLH containing protein interact to activate genes in the anthocyanin biosynthetic pathway (Schwinn et al., 2006). However, the functions of transcription factors, including MYBs, bHLHs, and WD40s, are unknown in *H. mutabilis*. The investigation of these TFs in cotton rose will be a subject for further research.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, PRJNA717149.

AUTHOR CONTRIBUTIONS

YY and XLL: conceptualization. XG, MZ, and XDL: formal analysis. FL and ZZ: project administration. JM and XZ: resources. YY and XS: writing—original draft. XLL: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.818206/full#supplementary-material

REFERENCES

Aggarwal, G., and Ramaswamy, R. (2002). *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* 27, 7–14. doi: 10.1007/BF02703679

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. doi: 10.1093/nar/28.1.45

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Birney, E., and Clamp, M. (2004). GeneWise and GenomeWise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125. doi: 10.1038/nbt.2727

Casimiro-Soriguer, I., Narbona, E., Buide, M. L., Del Valle, J. C., and Whittall, J. B. (2016). Transcriptome and biochemical analysis of a flower color polymorphism in *Silene littorea* (Caryophyllaceae). *Front. Plant Sci.* 7:204. doi: 10.3389/fpls.2016.00204

Chen, Y., Mao, Y., Liu, H., Yu, F., Li, S., and Yin, T. (2014). Transcriptome analysis of differentially expressed genes relevant to variegation in peach flowers. *PLoS One* 9:e90842. doi: 10.1371/journal.pone.0090842

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

DeYoung, B. J., and Innes, R. W. (2006). Plant NBS-LRR proteins in pathogen sensing and host defense. *Nat. Immunol.* 7, 1243–1249. doi: 10.1038/ni1410

Durbin, M., Lundy, K., Morrell, P., Torres-Martinez, C., and Clegg, M. (2004). Genes that determine flower color: the role of regulatory changes in the evolution of phenotypic adaptations. *Mol. Phylogenet. Evol.* 29, 507–518. doi: 10.1016/S1055-7903(03)00196-9

Edgar, R., and Myers, E. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics* 21, 152–158. doi: 10.1093/bioinformatics/bti1003

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-Driven genome evolution. *Genome Biol. Evol.* 5, 954–965. doi: 10.1093/gbe/evt025

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2013). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33, 121–124. doi: 10.1093/nar/gki081

Grover, C. E., Gallagher, J. P., and Wendel, J. F. (2015). Candidate gene identification of flowering time genes in cotton. *Plant Genome* 8:eplantgenome2014.2012.0098. doi: 10.3835/plantgenome2014.12.0098

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Jiao, F., Zhao, L., Wu, X., Song, Z., and Li, Y. (2020). Metabolome and transcriptome analyses of the molecular mechanisms of flower color mutation in tobacco. *BMC Genom.* 21:611. doi: 10.1186/s12864-020-07028-5

Kanehisa, M., and Goto, S. (2006). KEGG: kyoto encyclopedia of genes and genomes. *Artif. Intell.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kim, Y. M., Kim, S., Koo, N., Shin, A. Y., Yeom, S. I., Seo, E., et al. (2016). Genome analysis of *Hibiscus syriacus* provides insights of polyploidization and indeterminate flowering in woody plants. *DNA Res.* 24, 71–80. doi: 10.1093/dnares/dsw049

Koes, R., Quattrocchio, F., and Mol, J. (1994). The flavonoid biosynthetic pathway in plants: Function and evolution. *Bioessays* 16, 123–132. doi: 10.1002/bies.950160209

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinf.* 5:59. doi: 10.1186/1471-2105-5-59

Lai, B., Li, X. J., Hu, B., Qin, Y. H., Huang, X. M., Wang, H. C., et al. (2014). LcMYB1 is a key determinant of differential anthocyanin accumulation among genotypes, tissues, developmental phases and ABA and light stimuli in *Litchi chinensis*. *PLoS One* 9:e86293. doi: 10.1371/journal.pone.0086293

Li, H. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, L., Stoeckert, C. J. Jr., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Li, X., Tang, D., Du, H., and Shi, Y. (2018). Transcriptome sequencing and biochemical analysis of perianths and coronas reveal flower color formation in *Narcissus pseudonarcissus*. *Int. J. Mol. Sci.* 19:4006. doi: 10.3390/ijms19124006

Li, Y. P., Zhang, X. L., Wu, W. T., Miao, S. X., and Chang, J. L. (2015). Chromosome and karyotype analysis of *Hibiscus mutabilis* f. *mutabilis*. *Front. Life Sci.* 8, 300–304. doi: 10.1080/21553769.2015.1041166

Liu, D., Mei, Q., Wan, X., Que, H., Li, L., and Wan, D. (2015). Determination of rutin and isoquercetin contents in *Hibiscus mutabilis* folium in different collection periods by HPLC. *J. Chromatogr. Sci.* 53, 1680–1684. doi: 10.1093/chromsci/bmv071

Liu, J. Q., Jin, H. Q., Yuan, H. Y., and Lu, X. P. (2008). Mechanism analysis of variety corolla from *Hibiscus mutabilis* L. *Northern Hortic.* 11, 113–116.

Lou, Q., Liu, Y., Qi, Y., Jiao, S., Tian, F., Jiang, L., et al. (2014). Transcriptome sequencing and metabolite analysis reveals the role of delphinidin metabolism in flower colour in grape hyacinth. *J. Exp. Bot.* 65, 3157–3164. doi: 10.1093/jxb/eru168

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. *Science* 252, 1162–1164. doi: 10.1126/science.252.5009.1162

Majoros, W., Pertea, M., and Salzberg, S. (2004). TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Marchin, M., Kelly, P. T., and Fang, J. (2005). Tracker: continuous HMMER and BLAST searching. *Bioinformatics* 21, 388–389. doi: 10.1093/bioinformatics/bti012

Meer, I. M., Stuitje, A., and Mol, J. N. M. (1993). "Regulation of general phenylpropanoid and flavonoid gene expression," in *Control of Plant Gene Expression*, ed. D. P. S. Verma (Boca Raton, FL: CRC Press), 125–155.

Mol, J., Grotewold, E., and Koes, R. (1998). How genes paint flowers and seeds. *Trends Plant Sci.* 3, 212–217. doi: 10.1016/S1360-1385(98)01242-4

Moustafa, E., and Wong, E. (1967). Purification and properties of chalconeflavone isomerase from soya bean seed. *Phytochemistry* 6, 625–632. doi: 10.1016/S0031-9422(00)86001-X

Mulder, N., and Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol. Biol.* 396, 59–70. doi: 10.1007/978-1-59745-515-2_5

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25, 1335–1337. doi: 10.1093/bioinformatics/btp157

Ohmiya, A., Kishimoto, S., Aida, R., Yoshioka, S., and Sumitomo, K. (2006). Carotenoid cleavage dioxygenase (CmCCD4a) contributes to white color formation in chrysanthemum petals. *Plant Physiol.* 142, 1193–1201. doi: 10.1104/pp.106.087130

Parra, G., Blanco, E., and Guigó, R. (2000). GeneId in *Drosophila*. *Genome Res.* 10, 511–515. doi: 10.1101/gr.10.4.511

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi: 10.1093/bioinformatics/btm071

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., et al. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* 16, 1262–1269. doi: 10.1101/gr.5290206

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* 17, 155–158. doi: 10.1038/s41592-019-0669-3

Schiessl, S., Samans, B., Hüttel, B., Reinhard, R., and Snowdon, R. J. (2014). Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species *Brassica napus*. *Front. Plant Sci.* 5:404. doi: 10.3389/fpls.2014.00404

Schmidt, M. H., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., et al. (2017). *De novo* assembly of a new *Solanum pennellii a*ccession using nanopore sequencing. *Plant Cell* 29, 2336–2348. doi: 10.1105/tpc.17.00521

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.* 95, 5857–5864. doi: 10.1073/pnas.95.11.5857

Schwinn, K., Venail, J., Shang, Y., Mackay, S., Alm, V., Butelli, E., et al. (2006). A small family of MYB-regulatory genes controls floral pigmentation intensity and patterning in the genus *Antirrhinum*. *Plant Cell* 18, 831–851. doi: 10.1105/tpc.105.039255

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Song, C., Liu, Y., Song, A., Dong, G., Zhao, H., Sun, W., et al. (2018). The *Chrysanthemum nankingense* genome provides insights into the evolution and diversification of chrysanthemum flowers and medicinal traits. *Mol. Plant* 11, 1482–1491. doi: 10.1016/j.molp.2018.10.003

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Takken, F. L., Albrecht, M., and Tameling, W. I. (2006). Resistance proteins: molecular switches of plant defence. *Curr. Opin. Plant Biol.* 9, 383–390. doi: 10.1016/j.pbi.2006.05.009

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 25, 4.10.1–4.10.14. doi: 10.1002/0471250953.bi0410s25

Trapnell, C., Pachter, L., and Salzberg, S. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi: 10.1101/gr.214270.116

Walker, B., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., et al. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* 44, 1098–1103. doi: 10.1038/ng.2371

Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum*

and *Gossypium barbadense*. *Nat. Genet.* 51, 224–229. doi: 10.1038/s41588-018-0282-x

Wikström, N., Savolainen, V., and Chase, M. W. (2001). Evolution of the angiosperms: calibrating the family tree. *Proc. Biol. Sci.* 268, 2211–2220. doi: 10.1098/rspb.2001.1782

Xu, Z., and Wang, H. (2007). LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, T. J., Kim, J. S., Lim, K. B., Kwon, S. J., Kim, J. A., Jin, M., et al. (2005). The Korea *Brassica* genome project: a glimpse of the *Brassica* genome based on comparative genome analysis with *Arabidopsis*. *Comp. Funct. Genom.* 6, 138–146. doi: 10.1002/cfg.465

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* 13, 555–556. doi: 10.1093/bioinformatics/13.5.555

Yu, X. J., Zheng, H. K., Wang, J., Wang, W., and Su, B. (2007). Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* 88, 745–751. doi: 10.1016/j.ygeno.2006.05.008

Zhang, B., Liu, C., Wang, Y., Yao, X., Wang, F., Wu, J., et al. (2015). Disruption of a *CAROTENOID CLEAVAGE DIOXYGENASE 4* gene converts flower colour from white to yellow in *Brassica* species. *New Phytol.* 206, 1513–1526. doi: 10.1111/nph.13335

Zhang, L., Xu, Y., Zhang, X., Ma, X., Zhang, L., Liao, Z., et al. (2020). The genome of kenaf (*Hibiscus cannabinus* L.) provides insights into bast fiber and leaf shape biogenesis. *Plant Biotechnol. J.* 18, 1796–1809. doi: 10.1111/pbi.13341

Zhao, M., and Ma, J. (2013). Co-evolution of plant LTR-retrotransposons and their host genomes. *Prot. Cell* 4, 493–501. doi: 10.1007/s13238-013-3037-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 March 2022
doi: 10.3389/fpls.2022.839044

CRIA: An Interactive Gene Selection Algorithm for Cancers Prediction Based on Copy Number Variations

*Qiang Wu and Dongxi Li**

College of Data Science, Taiyuan University of Technology, Taiyuan, China

Genomic copy number variations (CNVs) are among the most important structural variations of genes found to be related to the risk of individual cancer and therefore they can be utilized to provide a clue to the research on the formation and progression of cancer. In this paper, an improved computational gene selection algorithm called CRIA (correlation-redundancy and interaction analysis based on gene selection algorithm) is introduced to screen genes that are closely related to cancer from the whole genome based on the value of gene CNVs. The CRIA algorithm mainly consists of two parts. Firstly, the main effect feature is selected out from the original feature set that has the largest correlation with the class label. Secondly, after the analysis involving correlation, redundancy and interaction for each feature in the candidate feature set, we choose the feature that maximizes the value of the custom selection criterion and add it into the selected feature set and then remove it from the candidate feature set in each selection round. Based on the real datasets, CRIA selects the top 200 genes to predict the type of cancer. The experiments' results of our research show that, compared with the state-of-the-art related methods, the CRIA algorithm can extract the key features of CNVs and a better classification performance can be achieved based on them. In addition, the interpretable genes highly related to cancer can be known, which may provide new clues at the genetic level for the treatment of the cancer.

Keywords: gene selection, correlation-redundancy analysis, interaction analysis, copula entropy, copy number variations (CNVs), cancers prediction

OPEN ACCESS

Edited by:
Weihua Pan,
Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Lin Wang,
Tianjin University of Science and Technology, China
Jiazhou Chen,
South China University of Technology, China

***Correspondence:**
Dongxi Li
dxli0426@126.com

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *19 December 2021*
Accepted: *19 January 2022*
Published: *21 March 2022*

Citation:
Wu Q and Li D (2022) CRIA: An Interactive Gene Selection Algorithm for Cancers Prediction Based on Copy Number Variations. Front. Plant Sci. 13:839044. doi: 10.3389/fpls.2022.839044

INTRODUCTION

The occurrence of many diseases is associated with genome structural variations. Human genome variations include single nucleotide polymorphisms (SNPs), copy number variations (CNVs), etc. The copy number variations refer to the amplification, deletion, and more complex mutations in the genome of DNA fragments longer than 1 kb in length (Redon et al., 2006). SNPs account for 0.5% of the human genome, and nearly 12% of the human genome often undergoes copy number variations (Redon et al., 2006). Copy number variations have become an important genomic variation, and their role in the pathogenesis of complex human diseases is still being revealed.

The close relationship between CNVs and diseases has been widely recognized. Numerous studies have demonstrated that not a few human diseases involved copy number variations that could change the diploid status of particular locus of the genome (Zhang et al., 2016). The Flierl research team found that the higher vulnerability of Parkinson's disease and stress sensitivity of neuronal precursor cells carry an α-synuclein gene triplication (Flierl et al., 2014).

Grangeon et al. (2021) discovered that early-onset cerebral amyloid angiopathy and Alzheimer Disease (AD) were related to an amyloid precursor protein (App) gene triple amplification. Breunis et al. (2008) reported that the copy number variations of FCGR2C gene promoted idiopathic thrombocytopenic purpura. Zheng et al. (2017) found that the low copy number of FCGR3B was associated with lupus nephritis in a Chinese population. And Pandey et al. (2015) revealed that there was both direct and indirect evidence suggesting abnormalities of glycogen synthase kinase (GSK)-3β and β-catenin in the pathophysiology of bipolar illness and possibly schizophrenia (SZ). Moreover, several neuro-developmental relevant genes, such as A2BP1, IMMP2, and AUTS2, were reported with mutational CNVs (Elia et al., 2010). In 2006, a research team composed of researchers from the United Kingdom, Japan, the United States, Canada and other countries studied 270 individuals in 4 groups of the HapMap project, and constructed the first-generation copy number variations map of the human genome, and obtained 144 CNVs region (about 12% of the size of the human genome). Among them, 285 CNVs regions were related to the occurrence of known diseases (Redon et al., 2006). Compared with SNPs, CNVs regions contained more DNA sequences, disease sites and functional elements, which could provide more clues for disease research. The publication of this map has become an important tool for studying the complex structural variations of the human genome and human diseases.

Cancer is a kind of diseases which involves uncontrolled abnormal cell growth and can spread to other tissues (Du and Elemento, 2015). The formation and development of cancer are also associated with copy number variations (Frank et al., 2007). Van Bockstal et al. (2020) discovered that HER2 gene amplification had a relationship with a bad result in invasive breast cancer and the amplification of heterogeneous HER2 had been described in 5–41% of breast cancer. The experimental results of Buchynska et al. (2019) shown that assessment of copy number variations of HER-2/neu, c-MYC and CCNE1 genes revealed their amplification in the tumors of 18.8, 25.0 and 14.3% of endometrial cancer patients, respectively. Heo et al. (2020) pointed out that CNVs were related to the mechanism of lung cancer development through a comparative experiment. Moreover, Tian et al. (2020) found that CNVs of CYLD, USP9X and USP11 were significantly associated with the risk of colorectal cancer. A latest global cancer burden data released by the International Agency for Research on Cancer(IARC) of the WHO showed that the number of patients with new cancer and cancer deaths in China ranked first around the world with 4.57 million patients with new cancer and 3 million cancer deaths, accounting for 23.7 and 30%, respectively. It is of great significance to investigate cancer causes and its treatment. Because the gene expression patterns in cancer tumor have high specificity (Liang et al., 2020), studying the relationship between these genetic information and cancer can provide a new idea for investigating the causes of cancer and help in early cancer diagnosis.

However, few studies have utilized machine learning (ML) or deep learning (DL) methods to use copy number variations data for the prediction of various cancer types. Zhang et al. (2016) used the mRMR and IFS methods to select 19 features from the 24,174 gene features of the copy number variations data set, which contained a total of 3,480 samples of 6 cancer types. They applied the Dagging algorithm with ten-fold cross-validation to classify cancer. But the accuracy of final result only reached 75%. Liang et al. (2020) used CNA_origin for cancer classification on the same data set. CNA_origin was an intelligent combined deep learning network, which was composed of two parts—a stacked autoencoder and a one-dimensional convolutional neural network with multiscale convolutional kernels. CNA_origin eventually had an overall accuracy of 83.81% on ten-fold cross-validation. But it could not identify which gene features were more important and more closely associated with cancer classification.

Here, we present an improved novel computational algorithm named CRIA, which can successfully classify cancer based on the information of gene CNVs levels from the same dataset. CRIA can not only effectively perform dimensionality reduction operation on high-dimensional gene CNVs data, which can improve the efficiency of the experiment, but also selects specific gene features closely related to cancer, making it clear which genes are more important in cancer classification. And the final results had higher classification accuracy than the state-of-the-art methods.

The rest sections of this paper are structured as follows: Section Background describes the theoretical background and related work. Section The Proposed Method-CRIA introduces the collection of CNVs dataset, the implementation details and performance of the proposed algorithm. Section Results and Discussions demonstrates the experimental results on CNVs dataset and the performance comparison with the recent methods. In section Conclusions, we summarize the conclusions and point out our future work.

BACKGROUND

In section Information Theory, we introduce some basic information theory knowledge, which is the core of our proposed algorithm. Before proposing our algorithm, we summarize some related work on gene selection methods and point out their drawbacks in section Related Work.

Information Theory

As early as 1948, Shannon's information theory had been proposed (Shannon, 2001), providing an effective method for measuring random variables' information. The entropy can be understood as a measure of the uncertainty of a random variable (Cover and Thomas, 1991). The greater the entropy of a random variable, the greater its uncertainty. If $X = \{x_1, x_2, ..., x_l\}$ is a discrete random variable, its probability distribution is $p(x) = P(X = x), x \in X$. The entropy of X is defined as:

$$H(X) = -\sum_{i=1}^{l} p(x_i) \log p(x_i) \tag{1}$$

where $p(x_i)$ is the probability of x_i. Here the base of log is 2 and specified that $0 \log 0 = 0$.

If $Y = \{y_1, y_2, ..., y_m\}$ is a discrete random variable, $p(x_i, y_j)$ is the joint probability of X and Y. Then, their joint entropy is defined as:

$$H(X, Y) = -\sum_{i=1}^{l}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j) \quad (2)$$

If the random variable X is in a given situation, the uncertainty measure of the variable Y can be defined by conditional entropy as follows:

$$H(Y|X) = H(X, Y) - H(X) = -\sum_{i=1}^{l}\sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i) \quad (3)$$

where $p(y_j|x_i)$ is the conditional probability of Y under the condition of X.

Definition 1: Mutual information (MI) (Cover and Thomas, 1991) is a measure of useful information in information theory. It can be regarded as the amount of information shared by two random variables. MI can be defined as:

$$\begin{aligned} I(X;\, Y) &= \sum_{i=1}^{l}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \\ &= H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y) \end{aligned} \quad (4)$$

Definition 2: Conditional mutual information (CMI) (Cover and Thomas, 1991) can be defined as the amount of information that shared by variables X and Y, if a discrete random variable $Z = \{z_1, z_2, ..., z_n\}$ is known.

$$\begin{aligned} I(X;\, Y\,|Z) &= \sum_{i=1}^{l}\sum_{j=1}^{m}\sum_{k=1}^{n} p(z_k)p(x_i, y_j|z_k) \log \frac{p(x_i, y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)} \\ &= H(Y|Z) - H(Y\,|X, Z) \end{aligned} \quad (5)$$

Definition 3: Joint mutual information (JMI) (Cover and Thomas, 1991) measures the amount of information shared by a joint random variable $(X_1, X_2, \cdots X_q)$ and Y and it can be defined as:

$$\begin{aligned} &I(X_1, X_2, ..., X_q;\, Y) = \sum_{x_1 \in X_1}\sum_{x_2 \in X_2}\cdots\sum_{x_q \in X_q}\sum_{y \in Y} p(x_1, x_2, ..., x_q, y) \\ &\log \frac{p(x_1, x_2, ..., x_q, y)}{p(x_1, x_2, ..., x_q)p(y)} \\ &= H(X_1, X_2, ..., X_q) - H(X_1, X_2, ..., X_q|Y) \end{aligned} \quad (6)$$

Definition 4: Interaction gain (IG) had been introduced by Jakulin (2003), Jakulin and Bratko (2004) to measure the amount of information shared by three random variables at the same time. Mutual information can be regarded as a two-way interaction gain. IG is defined as follows:

$$IG(X;\, Y;\, Z) = I(X;\, Y;\, Z) = I(X, Y;\, Z) - I(X;\, Z) - I(Y;\, Z) \quad (7)$$

Related Work

The irrelevant features and redundant features existed in high-dimensional data will damage the performance of the learning algorithm and reduce the efficiency of the learning algorithm. Therefore, the dimensionality reduction of features is one of the most common methods of data preprocessing (Orsenigo and Vercellis, 2013) and its purpose is to reduce the training time of the algorithm and improve the accuracy of final results (Bennasar et al., 2015). In recent years, the research of gene selection methods based on mutual information has received wide attention from scholars. Best individual gene selection (BIF) (Chandrashekar and Sahin, 2014) is the simplest and fastest filtering gene selection algorithm, especially suitable for high-dimensional data.

Battiti utilized the mutual information (MI) between features and class labels $[I(f_i; c)]$ to measure the relevance and the mutual information between features $[I(f_i; f_s)]$ to measure the redundancy (Battiti, 1994). He proposed the Mutual Information Gene selection (MIFS) criterion and it is defined as:

$$J_{MIFS}(f_i) = I(f_i; c) - \beta \sum_{f_s \in \Omega_S} I(f_i; f_s),\ f_i \in F - \Omega_S \quad (8)$$

where F is the original feature set, Ω_S is the selected feature subset, $F - \Omega_S$ is the candidate feature subset and c is the class label. β is a configurable parameter to determine the trade–off between relevance and redundancy. However, β is set experimentally, which results in an unstable performance.

Peng et al. (2005) proposed the Minimum-Redundancy Maximum-Relevance (MRMR) criterion and its evaluation function is defined as:

$$J_{mRMR}(f_i) = I(f_i; c) - \frac{1}{|n_s|} \sum_{f_s \in \Omega_S} I(f_i; f_s),\ f_i \in F - \Omega_S \quad (9)$$

where $|n_s|$ is the number of selected features.

Similarly, other gene selection methods that consider relevance between features and the class label and redundancy between features are concluded, such as Normalized Mutual Information Gene selection (NMIFS) and Conditional Mutual Information (CMI), and they were proposed by Estévez et al. (2009) and Liang et al. (2019) respectively. Their evaluation function are defined as follows:

$$J_{NMIFS}(f_i) = I(f_i; c) - \frac{1}{|n_s|} \sum_{f_s \in \Omega_S} \frac{I(f_i; f_s)}{\min(H(f_i), H(f_s))}, f_i \in F - \Omega_S \quad (10)$$

$$J_{CMI}(f_i) = I(f_i; c) - \frac{H(f_i\,|c)}{H(f_i)} \sum_{f_s \in \Omega_S} \frac{I(f_s; c) I(f_i; f_s)}{H(f_s)H(c)},\ f_i \in F - \Omega_S \quad (11)$$

where $H(f_i)$ is the information entropy and $H(f_i\,|c)$ is the conditional entropy.

Many gene selection algorithms based on information theory tend to use mutual information as a measure of relevance, which will bring a disadvantage that mutual information tends

to select features with more discrete values (Foithong et al., 2012). Thus, the symmetrical uncertainty (Witten and Frank, 2002) (a normalized form of mutual information, *SU*) is adopted to solve this problem. The symmetrical uncertainty can be described as:

$$SU(f_i; c) = \frac{2I(f_i; c)}{H(f_i) + H(c)} \tag{12}$$

The *SU* can redress the bias of mutual information as much as possible and scale its values to [0,1] by penalizing inputs with large entropies. It will make the performance of gene selection better. Same as MI, for any two features f_{i1} and f_{i2}, if $SU(f_{i1}; c) > SU(f_{i2}; c)$, due to more information can be provided by the former, f_{i1} and c are more relevant. If $SU(f_{i1}; f_s) > SU(f_{i2}; f_s)$, owing to the information shared by f_{i1} and f_s being more and providing less information, f_{i1} and f_s have greater redundancy.

Additionally, these gene selection algorithms mentioned above fail to take the feature interaction into consideration. After relevance and redundancy analysis, one feature deemed useless may interact with other features to provide more useful information. Especially in complicated biology systems, molecules interacting with each other, they work together to express physiological and pathological changes. If we only consider relevance and redundancy but ignore the feature interaction in data analysis, we may miss some useful features and affect the analysis results (Chen et al., 2015).

Sun et al. (2013), Zeng et al. (2015), and Gu et al. (2020), respectively proposed a gene selection method using dynamic feature weights: Dynamic Weighting-based Gene selection algorithm (DWFS), Interaction Weight based Gene selection algorithm (IWFS) and Redundancy Analysis and Interaction Weight-based gene selection algorithm (RAIW). All of them employ the symmetric uncertainty to measure the relevance between features and the class label, and exploit the three-dimensional interaction information (mentioned at **Information Theory Definition 4**) to measure the interaction between two features and the class label. The evaluation functions are defined as follow:

$$J_{DWFS}(f_i) = SU(f_i; c) \times w_{DWFS}(f_i),\ f_i \in -\Omega_S \tag{13}$$

$$J_{IWFS}(f_i) = w_{IWFS}(f_i) \times [1 + SU(f_i; c)],\ f_i \in F - \Omega_S \tag{14}$$

$$J_{RAIW}(f_i) = SU(f_i; c) \times [1 - \alpha SU(f_i; f_s)] \times w_{RAIW}(f_i), f_i \epsilon F - \Omega_s \tag{15}$$

where $w(f_i)$ is the weight of each feature and its initial value is set to 1, α is a redundancy coefficient and the value is relevant to the number of dataset's features, f_s is one of features in the selected feature subset. In each round, the feature weight $w(f_i)$ is updated by their interaction weight factors.

$$\begin{aligned} w_{DWFS}(f_i) &= w_{DWFS}(f_i') \times [1 + CR(f_i, f_s)] = w_{DWFS}(f_i') \\ &\times [1 + 2\tfrac{I(f_i; c|f_s) - I(f_i; c)}{H(f_i) + H(c)}] \\ &= w_{DWFS}(f_i^{'}) \times [1 + 2\tfrac{I(f_i; f_s; c)}{H(f_i) + H(c)}] \end{aligned} \tag{16}$$

$$\begin{aligned} w_{IWFS}(f_i) &= w_{IWFS}(f_i^{'}) \times IW(f_i, f_s) \\ &= w_{IWFS}(f_i') \times [1 + \frac{I(f_i; f_s; c)}{H(f_i) + H(f_s)}] \end{aligned} \tag{17}$$

$$\begin{aligned} w_{RAIW}(f_i) &= w_{RAIW}(f_i^{'}) \times [1 + If(f_i, f_s, c)] \\ &= w_{RAIW}(f_i') \times [1 + \frac{2I(f_i; f_s; c)}{H(f_i) + H(f_s) + H(c)}] \end{aligned} \tag{18}$$

where $w(f_i')$ denotes the feature weight of the previous round, $I(f_i; c\,|f_s)$ is the conditional mutual information of f_i and c when f_s is given. $I(f_i; f_s; c)$ is three-dimensional interaction information. However, we can find that although DWFS and IWFS take into account relevance and interaction, they ignore the redundancy between features. Correlation, redundancy and interaction are all taken into account by RAIW, but there is a no reasonable value for α in a specific dataset.

Furthermore, some other gene selection methods about three-way mutual information are listed and their evaluation function are defined as follows, such as Composition of Feature Relevance (CFR) (Gao et al., 2018a), Joint Mutual Information Maximization (JMIM) (Bennasar et al., 2015), Dynamic Change of Selected Feature with the class (DCSF) (Gao et al., 2018b) and Max-Relevance and Max-Independence (MRI) (Wang et al., 2017).

$$J_{CFR}(f_i) = \sum_{f_s \in \Omega_S} I(f_i; c\,|f_s) + \sum_{f_s \in \Omega_S} I(f_i; f_s; c),\ f_i \in F - \Omega_S \tag{19}$$

$$J_{JMIM}(f_i) = \max[\min_{f_s \in \Omega_S}(I(f_i, f_s; c))],\quad f_i \in F - \Omega_S \tag{20}$$

$$J_{DCSF}(f_i) = \sum_{f_s \in \Omega_S} I(f_i; c\,|f_s) + \sum_{f_s \in \Omega_S} I(f_s; c\,|f_i) - \sum_{f_s \in \Omega_S} I(f_i; f_s),\quad f_i \in F - \Omega_S \tag{21}$$

$$J_{MRI}(f_i) = I(f_i; c) + \sum_{f_s \in \Omega_S} I(f_i; c\,|f_s) + \sum_{f_s \in \Omega_S} I(f_s; c\,|f_i),\quad f_i \in F - \Omega_S \tag{22}$$

where $I(f_i, f_s; c)$ is the joint mutual information of f_i, f_s and c. $I(f_s; c\,|f_i)$ is the conditional mutual information of f_s and c when f_i is given. However, these algorithms only take into account three-way mutual information among features and the class label, and none of them considers relevance, redundancy and three-dimensional mutual information between features at the same time, which will affect the performance of these algorithms.

THE PROPOSED METHOD-CRIA

In section CNVs Dataset, we firstly introduce the collection of datasets and the process of data processing specifically. Subsequently, we redress other methods' shortcomings and propose an improved gene selection algorithm called CRIA

TABLE 1 | The number of samples for each cancer type in this dataset.

Class label	Histology	Samples	Percentage
1	UCEC (Uterine corpus endometrial carcinoma)	443	12.73%
2	KIRC (Kidney renal clear cell carcinoma)	490	14.08%
3	OV (Ovarian serous cystadenocarcinoma)	562	16.15%
4	GBM (Glioblastoma multiforme)	563	16.18%
5	COAD/READ (Colon adenocarcinoma/Rect-um adenocarcinoma)	575	16.52%
6	BRCA (Breast invasive carcinoma)	847	24.34%
Total		3,480	100%

in section The Proposed Algorithm and give it a specific implementation in section Algorithm Implementation. Finally, in section Verify the Performance of CRIA, we verify the performance of CRIA by comparing the experimental results of CRIA and other 8 algorithms on 5 datasets.

CNVs Dataset

The datasets of copy number variations in different cancer types used in this paper comes from the cBioPortal for Cancer Genomics (http://cbio.mskcc.org/cancergenomics/pancan_tcga/, Release 2/4/2013) (Cerami et al., 2012; Ciriello et al., 2013; Gao et al., 2013). The copy number values in the dataset are generated by Affymetrix SNP 6.0 arrays for the set of samples in the cancer genome atlas (TCGA) study (Liang et al., 2020). The preprocessing analysis of the dataset is performed with GISTIC (Beroukhim et al., 2007). There are 11 cancer types in the cBioPortal database with the largest sample number was 847 and the smallest sample was 135. In order to avoid affecting the experimental results due to the large difference in the number of samples of cancer types, we only select six cancer types with more than 400 samples as our experimental data. The details of six cancer types are listed in **Table 1**, and totally there are 3480 samples in our experimental dataset.

In this dataset, each sample consists of labels for 24174 genetic cytobands. The CNV spectrum is divided into five regions/labels by setting four thresholds in cancer algorithm (Mermel et al., 2011). Then, the CNV values are discretized into 5 different values—"-2," "-1," "0," "1," "2," where "-2" denotes the deletion of both copies (possibly homozygous deletion), "−1" means the deletion of one copy (possibly heterozygous deletion), "0" corresponds to exactly two copies, i.e., no gain/loss (diploid), "1" denotes a low-level copy number gain and "2" means a high-level copy number amplification (Ciriello et al., 2013).

The CNVs values are preprocessed to the range of [−1,1] with Equation (23).

$$val' = \frac{val}{|val|_{\max}} \tag{23}$$

where val is the value of gene copy number variations of each sample, $|val|_{\max}$ is the maximum absolute value of gene CNVs among samples and val' is the recalculated value.

The Proposed Algorithm

In section Related Work, we analyze the 11 gene selection methods and point out their shortcomings. In view of the defects of these algorithms, we propose an improved gene selection algorithm to redress their shortcomings: Correlation-Redundancy and Interaction Analysis based gene selection algorithm (CRIA). This method uses the symmetric uncertainty (SU) to measure the correlation between features and the class label and the redundancy among features. In addition, copula entropy is introduced to measure the feature interaction information. Different from the three-way interaction of DWFS, IWFS and RAIW, the proposed algorithm considers the interaction between the candidate feature and the entire set of selected features, instead of being limited to the three-dimensional interaction.

As we know, Shannon's definition of mutual information aims at a pair of random variables, and it measures the correlation between two random variables. Therefore, naturally, many researchers have tried to study how to extend the definition of mutual information from two variables to multivariate situations. In 2011, Ma and Sun published a paper (Ma and Sun, 2011), which contributed to the entropy of information theory. They defined a new concept of entropy in that paper, called Copula Entropy. Copula Entropy is defined on a set of random variables and conformed to symmetry. Therefore, it is a multivariate extension of mutual information, which can be utilized to measure the full-order, non-linear correlation among random variables. They proved the equivalence between copula entropy and the concept of mutual information, which was, mutual information was equal to negative copula entropy (Ma and Sun, 2011).

The copula entropy of $\vec{x} = (x_1, x_2, ..., x_N) \in R^N$ is defined as:

$$H_c(\vec{x}) = -\int c(\vec{u}) \log c(\vec{u}) d\vec{u} \tag{24}$$

where $\vec{x}$ are random variables with marginal functions $\vec{u} = [F_1, F_2, ..., F_N]$ and copula density $c(\vec{u}) = \frac{d^N C(\vec{u})}{du_1 du_2 ... du_N}$.

Thus, we can use interaction factor IF_{CRIA}, which is defined in Equation (25) to measure the interaction between the candidate feature and the selected feature subset. The meaning of IF_{CRIA} is that, after adding a random candidate feature f_i into the selected feature subset Ω_S, the amount of interaction information increased relative to the original selected feature subset. So, the bigger value of IF_{CRIA}, the bigger value of interaction between f_i and Ω_S. In each round of calculation, we are supposed to choose the variable that maximizes the IF_{CRIA} value.

$$IF_{CRIA} = \frac{H_c(\Omega_S, f_i, c)}{H_c(\Omega_S, c)} \tag{25}$$

Where c is the target class label.

Integrating the correlation between the features and the class label and the redundancy between features that we improved, and the interaction factor we proposed, we can define the evaluation

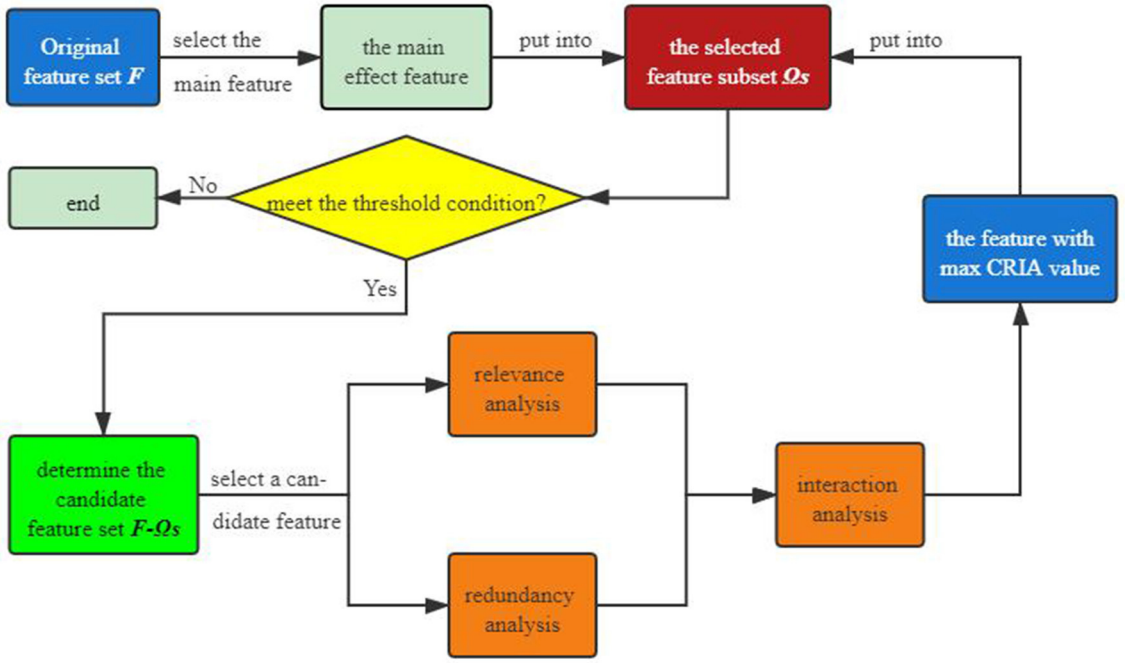


FIGURE 1 | General flow chart of the proposed algorithm.

criterion of a candidate feature as follows:

$$J_{CRIA}(f_i) = \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \frac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times IF_{CRIA}\} = \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \frac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times \frac{H_c(\Omega_S,f_i,c)}{H_c(\Omega_S,c)}\} \quad (26)$$

For the equation (26), we can see that the proposed algorithm can take into account the relevance between the candidate feature and the class label, redundancy and multi-dimensional interaction among the candidate feature and the selected features at the same time. The formula $SU(f_i,c)$ can denote the relevance and $\frac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)$ calculates the redundancy. Also, the formula $\frac{H_c(\Omega_S,f_i,c)}{H_c(\Omega_S,c)}$ denotes the interaction among the features.

According to the definition of copula entropy and Equation (24), there is a theorem.

Theorem 1: The mutual information of random variables is equivalent to their negative copula entropy (Ma and Sun, 2011):

$$I(\vec{x}) = -H_c(\vec{x}) \quad (27)$$

According to Theorem 1, the value of copula entropy can be calculated by the MI of multivariates. The definition of mutual information extended from two variables to multivariate is described as follows:

$$I(X_m,c) = \int\int p(X_m,c)\log\frac{p(X_m,c)}{p(X_m)p(c)}dX_m dc = \int\int p(X_{m-1},x_m,c)\log\frac{p(X_{m-1},x_m,c)}{p(X_{m-1},x_m)p(c)}dX_{m-1}dx_m dc = \int\ldots\int p(x_1,\ldots,,x_m,c)\log\frac{p(x_1,\ldots,x_m,c)}{p(x_1,\ldots,x_m)p(c)}dx_1,\ldots,dx_m dc \quad (28)$$

TABLE 2 | Datasets for comparison between CRIA algorithm and other algorithms.

Datasets type	No.	Datasets	Samples	Features	Classes	Types
Biological data	1	leukemia	72	7,070	2	Discrete
	2	Carcinoma	174	9,182	11	Continuous
	3	colon	62	2,000	2	Discrete
	4	TOX_171	171	5,748	4	Continuous
Digit recognition	5	Gisette	7,000	5,000	2	Continuous

where $X_m = \{x_1, x_2, ..., x_{m-1}, x_m\} = \{X_{m-1}, x_m\}$.

According to the Equation (28), we have:

$$H(X_{m-1},x_m) = H(X_m) = \sum_{i=1}^{m} H(x_i) - I(X_m)$$
$$H(X_{m-1},x_m,c) = H(X_m,c) = H(c) + \sum_{i=1}^{m} H(x_i) - I(X_m,c) \quad (29)$$

Therefore,

$$I(X_m,c) = H(x_1) + \ldots + H(x_m) + H(c) - H(x_1,\ldots,x_m,c)$$
$$I(X_m,\hat{X}_s,c) = H(x_1) + \ldots + H(x_m) + H(\hat{X}_s) + H(c) - H(x_1,\ldots,x_m,\hat{X}_s,c) \quad (30)$$

According to Equation (26), (27) and (30), we have:

$$J_{CRIA}(f_i) = \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \frac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times IF_{CRIA}\} = \max_{f_i \in F-\Omega_S} \{[SU(f_i,c) - \frac{1}{n_s}\sum_{f_s \in \Omega_S} SU(f_i,f_s)] \times \frac{I(\Omega_S,f_i,c)}{I(\Omega_S,c)}\} \quad (31)$$

Let $\Omega_S = \{f_1, f_2, ..., f_m\}$, Since,

$$\frac{I(\Omega_S, f_i, c)}{I(\Omega_S, c)} = \frac{\sum_{k=1}^{m} H(f_k) + H(f_i) + H(c) - H(\Omega_S, f_i, c)}{\sum_{k=1}^{m} H(f_k) + H(c) - H(\Omega_S, c)} \quad (32)$$

Therefore,

$$J_{CRIA}(f_i) = \max_{f_i \in F - \Omega_S} \{[SU(f_i, c) - \frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)] \times \frac{\sum_{k=1}^{m} H(f_k) + H(f_i) + H(c) - H(\Omega_S, f_i, c)}{\sum_{k=1}^{m} H(f_k) + H(c) - H(\Omega_S, c)}\} \quad (33)$$

The general flow chart of the proposed algorithm is presented in **Figure 1** we can see that an original feature set F is first given, from which we select the main effect feature that maximizes the value of (12). Then the main feature is put into the selected feature subset Ω_S. For each feature in the candidate feature set, after conducting correlation and redundancy analysis, we are next supposed to use (25) to perform interaction analysis on it. Choose the feature that maximizes the value of (33), which then is put into the selected feature set. If the number of the selected features meets the threshold condition, the above steps will be executed again, otherwise the program ends directly.

Algorithm Implementation

We propose a gene selection method based on correlation-redundancy and interaction analysis. The pseudo code of CRIA algorithm is described as follows.

Here, for the CNVs dataset, we set the value of the threshold M to be 200 to reduce the calculation time and avoid curse of dimensionality. In addition, we need to control the number of selected features to be same as the method proposed by Zhang et al. (2016).

The CRIA algorithm consists of two stages:

Stage 1 (lines 1–7): In this part, the selected feature subset Ω_S and the original feature set F are first initialized. For each feature in the original feature set f_i, the symmetrical uncertainty $SU(f_i; c)$ between f_i and class label c is calculated. The feature whose value of symmetrical uncertainty with class label is the maximum is selected out and added into the selected features subset Ω_S, which we name "the main effect feature."

Stage 2 (lines 8-18): The second stage mainly calculates the correlation measure $SU(f_i; c)$ and the redundancy measure $\frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)$. Then the interaction value IF_{CRIA} between Ω_S, f_i and c is updated. $J_{CRIA}(f_i)$ is calculated and the feature with the maximum value is added into the selected feature subset Ω_S. This procedure terminates until the number of selected features is no less than predefined threshold M.

According to **Algorithm 1**, when the size of the feature subset reaches the set threshold M, the procedure will be terminated. The value of the threshold setting should be determined by different datasets. A small M can reduce the amount of calculation but may also lose many effective features that are

TABLE 3 | Comparision (mean ± std.dev.) of performance between CRIA and other 8 algorithms with J48 classifier.

Datasets	CRIA (proposed)	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCFS
leukemia	**95.00 ± 1.17 (1)**	93.08 ± 0.67 (4)	93.09 ± 0.64 (3)	92.51 ± 1.14 (7)	92.05 ± 0.73 (8)	93.36 ± 0.59 (2)	92.70 ± 0.98 (6)	93.07 ± 1.10 (5)	91.50 ± 1.44 (9)
Carcinoma	**75.17 ± 1.67 (1)**	71.73 ± 1.80 (3)	71.78 ± 2.26 (2)	69.79 ± 1.58 (4)	64.47 ± 1.82 (9)	64,84 ± 1.68 (5)	68.65 ± 2.20 (6)	68.01 ± 1.97 (7)	65.53 ± 1.89 (8)
colon	**79.68 ± 2.54 (1)**	76.55 ± 3.86 (5)	77.14 ± 3.31 (4)	74.32 ± 2.48 (8)	76.32 ± 2.07 (6)	77.28 ± 3.77 (3)	77.30 ± 4.73 (2)	74.71 ± 3.96 (7)	73.31 ± 3.77 (9)
TOX_171	62.01 ± 1.61 (4)	62.21 ± 2.04 (2)	56.61 ± 2.21 (8)	60.94 ± 3.08 (5)	62.09 ± 2.14 (3)	57.49 ± 3.01 (9)	59.78 ± 2.02 (7)	60.18 ± 1.91 (6)	**62.26 ± 2.71 (1)**
gisette	**93.71 ± 0.17 (1)**	92.40 ± 0.08 (6)	92.02 ± 0.08 (8)	92.66 ± 0.10 (5)	92.05 ± 0.12 (7)	91.18 ± 0.12 (9)	92.84 ± 0.08 (3)	92.82 ± 0.07 (4)	93.35 ± 0.08 (2)
Avg.acc	81.11	79.19	78.53	78.04	77.40	77.63	78.25	77.76	77.19
Avg.rank	1.60	4.00	5.00	5.80	6.60	5.60	4.80	5.80	5.80
Improved rate	–	2.42%	3.29%	3.93%	4.79%	4.48%	3.65%	4.31%	5.08%

The meaning of the bold values represent the best performance achieved on a certain dataset for the nine methods.

TABLE 4 | Comparision (mean ± std.dev.) of performance between CRIA and other 8 algorithms with IB1 classifier.

Datasets	CRIA (proposed)	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCFS
leukemia	**99.44 ± 0.97 (1)**	97.03 ± 0.87 (2)	96.16 ± 0.43 (4)	95.56 ± 0.90 (7)	88.75 ± 1.88 (9)	96.61 ± 0.78 (3)	96.13 ± 0.90 (5)	95.73 ± 0.95 (6)	94.22 ± 0.80 (8)
Carcinoma	**86.84 ± 0.50 (1)**	82.45 ± 1.33 (2)	81.39 ± 1.02 (4.5)	82.32 ± 1.07 (3)	76.88 ± 1.97 (9)	81.10 ± 1.06 (7)	81.39 ± 1.16 (4.5)	81.35 ± 1.08 (6)	80.96 ± 1.18 (8)
colon	**86.77 ± 1.27 (1)**	78.60 ± 2.10 (2)	78.22 ± 1.54 (3)	75.94 ± 2.32 (5)	70.87 ± 1.97 (9)	76.69 ± 2.35 (4)	71.77 ± 3.41 (8)	72.24 ± 2.17 (7)	74.87 ± 1.80 (6)
TOX_171	84.56 ± 0.52 (3)	85.13 ± 1.26 (2)	78.14 ± 1.36 (8)	**85.19 ± 1.30 (1)**	82.59 ± 1.78 (5)	76.68 ± 1.68 (9)	81.69 ± 1.64 (7)	82.05 ± 1.28 (6)	84.04 ± 1.48 (4)
gisette	**93.75 ± 0.14 (1)**	91.88 ± 0.08 (6)	91.26 ± 0.09 (7)	92.26 ± 0.07 (5)	91.05 ± 0.15 (8)	90.20 ± 0.10 (9)	92.70 ± 0.06 (3)	92.58 ± 0.05 (4)	93.13 ± 0.10 (2)
Avg.acc	90.27	87.02	85.03	86.25	82.03	84.26	84.74	84.79	85.44
Avg.rank	1.40	2.80	5.30	4.20	8.00	6.40	5.50	5.80	5.60
Improved rate	–	3.73%	6.16%	4.66%	10.05%	7.13%	6.53%	6.46%	5.65%

The meaning of the bold values represent the best performance achieved on a certain dataset for the nine methods.

TABLE 5 | Comparision (mean ± std.dev.) of performance between CRIA and other 8 algorithms with Naïve Bayes classifier.

Datasets	CRIA (proposed)	RAIW	mRMR	DWFS	IWFS	JMIM	MRI	CFR	DCFS
leukemia	**99.58 ± 0.67 (1)**	97.44 ± 0.66 (2)	96.27 ± 0.30 (7)	97.03 ± 0.70 (4)	95.48 ± 1.96 (9)	96.18 ± 0.39 (8)	97.15 ± 0.80 (3)	96.79 ± 0.71 (5)	96.70 ± 0.58 (6)
Carcinoma	81.61 ± 1.11 (2)	**82.02 ± 0.83 (1)**	80.23 ± 1.33 (7)	81.58 ± 0.95 (3)	76.38 ± 1.85 (9)	80.19 ± 1.28 (8)	80.41 ± 0.86 (5)	80.34 ± 0.87 (6)	80.46 ± 1.09 (4)
colon	**88.71 ± 0.00 (1)**	82.97 ± 1.34 (2)	82.70 ± 1.20 (3)	80.72 ± 1.47 (6)	74.39 ± 4.37 (9)	81.66 ± 1.45 (5)	79.84 ± 2.40 (7)	78.95 ± 1.72 (8)	82.32 ± 2.08 (4)
TOX_171	69.53 ± 0.70 (3)	**70.74 ± 1.07 (1)**	63.73 ± 1.61 (8)	68.68 ± 1.26 (4)	65.64 ± 1.63 (7)	60.41 ± 2.35 (9)	66.96 ± 1.56 (6)	67.04 ± 1.74 (5)	70.28 ± 1.60 (2)
gisette	**93.16 ± 0.05 (1)**	90.46 ± 0.13 (2)	88.26 ± 0.02 (4)	87.69 ± 0.08 (5)	86.23 ± 0.24 (8)	86.01 ± 0.05 (9)	87.60 ± 0.04 (6)	87.48 ± 0.03 (7)	89.46 ± 0.05 (3)
Avg.acc	86.52	84.73	82.24	83.14	79.62	80.89	82.39	82.12	83.84
Avg.rank	1.60	1.80	5.80	4.40	8.40	7.80	5.40	6.20	3.80
Improved rate	–	2.11%	5.20%	4.07%	8.67%	6.96%	5.01%	5.36%	3.20%

The meaning of the bold values represent the best performance achieved on a certain dataset for the nine methods.

Algorithm 1 | CRIA: correlation-redundancy and interaction analysis based gene selection algorithm.

Input N: the number of original features, M: the number of features to be selected, n_s: the number of selected features.
Output: the selected feature subset ($\Omega_S \subseteq \mathbb{F}$).
1 First initializes $\Omega_S = \varnothing, \mathbb{F} = \{f_1, f_2, ..., f_N\}$;
2 for each $f_i \in \mathbb{F}$ do:
3 calculate $SU(f_i, c)$;
4 end for
5 select the feature $f_{i\max} \in \mathbb{F}$ with the largest value of $SU(f_i, c)$;
6 $\Omega_S = \Omega_S \cup \{f_{i\max}\}$;
7 $F = F - \{f_{i\max}\}$;
8 while $n_s \leq M$ do:
9 for $f_i \in \mathbb{F}$ do:
10 calculate $SU(f_i, c) - \frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)$;
11 calculate $IF_{CRIA} = \frac{H_c(\Omega_S, f_i, c)}{H_c(\Omega_S, c)}$;
12 calculate $J_{CRIA}(f_i) = [SU(f_i, c) - \frac{1}{n_s} \sum_{f_s \in \Omega_S} SU(f_i, f_s)] \times IF_{CRIA}$ and append it into a list;
13 end for
14 select the feature $f_{k\max} \in \mathbb{F}$ with the largest value of $J_{CRIA}(f_i)$ from list;
15 $\Omega_S = \Omega_S \cup \{f_{k\max}\}$;
16 $F = F - \{f_{k\max}\}$;
17 end while
18 output Ω_S.

useful; a large M will increase the amount of calculation but may improve the accuracy of final result (Foithong et al., 2012). Actually, when the threshold exceeds a certain value, the accuracy of the final result will not only not increase, but may decrease, and it will bring computational complexity. The selected features are ranked according to the value of the evaluation function $J_{CRIA}(f_i)$ from largest to smallest.

Verify the Performance of CRIA

Eight gene selection algorithms—JMIM (Bennasar et al., 2015), mRMR (Peng et al., 2005), DWFS (Sun et al., 2013), IWFS (Zeng et al., 2015), RAIW (Gu et al., 2020), CFR (Gao et al., 2018a), DCSF (Gao et al., 2018b), and MRI (Wang et al., 2017) are used to compare with CRIA to examine the performance of our proposed method.

The datasets used in validation experiment come from Arizona State University (ASU) datasets (Li et al., 2017), which include four biological data and one other type of data (digit recognition). They are all high-dimensional data. The smallest feature number is 2000 and the largest feature number is 9182 among them. The specific details of these datasets are shown in **Table 2**. We only use minimum description length method (Fayyad and Irani, 1993) for gene selection and utilize it to convert these numerical features.

The number of features N used in the experiment is reduced to 50 and three classifiers—IB1, J48 and Naïve Bayes are exploited. The parameters of the classifiers are set to the default parameters of Waikato Environment for Knowledge Analysis (WEKA) (Hall et al., 2009). We use 10 times of ten-fold cross-validation to avoid the influence of randomness on experimental results. Then mean value and Standard Deviation (STD) are taken as the comparison indices of performance of each algorithm and STD is defined as follows:

$$STD = \sqrt{\frac{1}{n_{run}} \sum_{i=1}^{N} (ACC_i - u)} \tag{34}$$

where n_{run} is the number of times of our experiments, here we set $n_{run} = 10$, ACC is the classification accuracy, u represents the average value of ACC, and N denotes the number of samples. The bigger ACC, the better performance, and the smaller STD, the higher stability.

The comparison results between the proposed algorithm and other gene selection algorithms are shown in **Tables 3–5**. As shown in **Table 3**, for the five data sets in the experiment, we can see that the classification results of CRIA in four data sets are better than other eight algorithms, which ranking first, except TOX_171, ranking fourth. Compared with other algorithms, the average accuracy of CRIA is increased by 2.42–5.08%. In **Table 4**, CRIA also outperforms the other 8 algorithms on four data sets except TOX_171, on which the experimental results of CRIA ranking third. The biggest improved rate of the proposed algorithm is 10.05% and the smallest one is 3.73%. From **Table 5**, we can find that the results of CRIA on the three data sets are superior to other algorithms, ranking first. However, on the datasets of Carcinoma and TOX_171, compared with the maximum values, the experimental accuracies of CRIA are slightly decreased by 0.50 and 1.71%, ranking second and third respectively. From the perspective of average accuracy, CRIA's result is better than other algorithms, and it is improved by 2.11–8.67%.

RESULTS AND DISCUSSIONS

Evaluation Metrics of Experimental Results

Four evaluation metrics—precision, recall, accuracy and F1-score are utilized to evaluate the performance of the corresponding method and values of these criteria are defined as equation (35).

$$\begin{aligned} precision &= \frac{T_P}{T_P + F_P} \\ recall &= \frac{T_P}{T_P + F_N} \\ accuracy &= \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \\ F1 - score &= \frac{2 \times precision \times recall}{precision + recall} \end{aligned} \tag{35}$$

TABLE 6 | The top 15 feature genes chosen by CRIA defined as equation (26).

Ranked order	Official name	Official full gene name	Category	CRIA value
1	RPS15	ribosomal protein S15	Protein Coding	0.168
2	TBC1D5	TBC1 Domain Family Member 5	Protein Coding	0.089
3	CUL2	Cullin 2	Protein Coding	0.093
4	SMPD3	Sphingomyelin Phosphodiesterase 3	Protein Coding	0.089
5	CTAGE10P	CTAGE Family Member 10, Pseudogene	Pseudogene	0.071
6	C1orf98	Chromosome 1 Open Reading Frame 98	Protein Coding	0.043
7	ZNF281	Zinc Finger Protein 281	Protein Coding	0.061
8	CDKN2A	Cyclin Dependent Kinase Inhibitor 2A	Protein Coding	0.161
9	EGFR	Epidermal Growth Factor Receptor	Protein Coding	0.121
10	TMEM98	Transmembrane Protein 98	Protein Coding	0.103
11	CTBP2	C-Terminal Binding Protein 2	Protein Coding	0.083
12	SEMA6A	Semaphorin 6A	Protein Coding	0.081
13	MIR1208	MicroRNA 1208	RNA Gene	0.077
14	RBFOX1	RNA Binding Fox-1 Homolog 1	Protein Coding	0.069
15	CDC25A	Cell Division Cycle 25A	Protein Coding	0.066

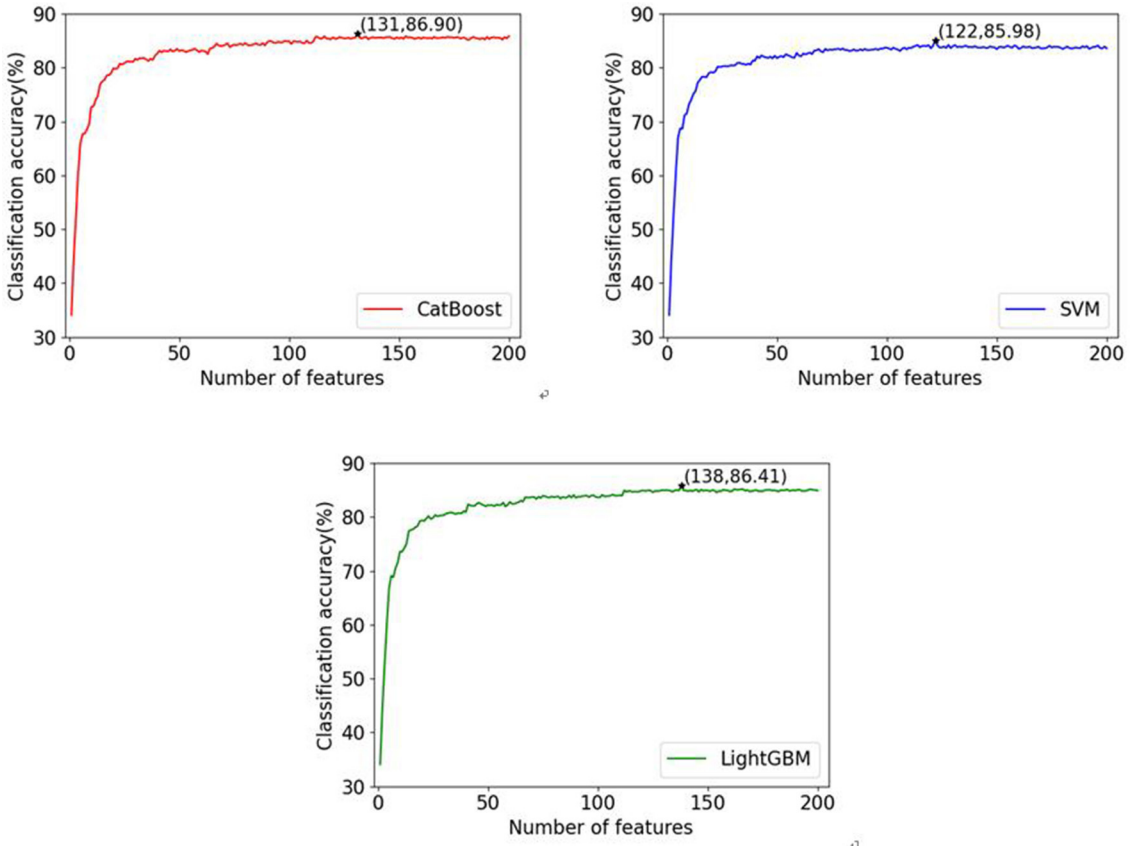


FIGURE 2 | Classification accuracies of three classifiers (CatBoost, LightGBM and SVM) with different numbers of features during the IFS procedure. The top 200 feature genes are selected by CRIA method.

where T_P, T_N, F_P, and F_N denotes the numbers of true positives, true negatives, false positives, and false negatives respectively.

The CRIA and IFS Results

As mentioned in section Evaluation Metrics of Experimental Results, each sample is represented by 24,174 features, each of

TABLE 7 | Average performance of precision, recall and F1-score on 10 test datasets with three classifiers *via* ten-fold cross-validation (%).

Metrics		UCEC	KIRC	OV	GBM	COAD/READ	BRCA
Precision	CRIA_CatBoost	74.31	93.74	84.59	94.63	89.67	84.48
	CRIA_SVM	70.47	90.33	85.40	95.64	88.54	84.76
	CRIA_LightGBM	71.46	93.37	82.84	95.72	90.23	84.71
Recall	CRIA_CatBoost	73.14	91.63	87.90	90.76	86.09	88.67
	CRIA_SVM	73.81	89.59	88.43	89.70	83.30	87.96
	CRIA_LightGBM	72.91	92.04	89.32	91.30	83.48	87.01
F1-score	CRIA_CatBoost	73.72	92.67	86.21	92.65	87.84	86.52
	CRIA_SVM	72.13	89.96	86.89	92.57	85.84	86.33
	CRIA_LightGBM	72.18	92.70	85.96	93.46	86.72	85.84

which indicates the expression level of genes. The 24174 feature genes are sorted by CRIA value in descending order. However, we only select the top 200 features in this work for the consideration of computational time and curse of dimensionality. The top 15 key feature genes chosen by CRIA defined by equation (26) are listed in **Table 6**.

We use the Incremental Gene selection (IFS) (Yang et al., 2019) to determine the optimal feature set. The first 200 features are added one by one to a feature subset in order. Each time a feature is added, a classifier is trained and examined. So, 200 classifiers are constructed. We use the criteria of accuracy to evaluate the performance of all the 200 classifiers and then we choose the classifier with the highest accuracy as the final one. The corresponding feature subset that the final classifier used is deemed to be the optimal feature set.

In this paper, three commonly used classifiers are adopted to verify the generalization performance of the proposed gene selection method on different classifiers. ten-fold cross-validation is used to evaluate our algorithm with the selected features. The complete data set is randomly split into 10 parts of approximately equal size. The three classifiers are trained 10 times; nine of the 10 subsets are used as the training datasets, and the remaining one is the test dataset. The average values of accuracy for each classifier are calculated and the IFS results are shown in **Figure 2**. Here, we name our methods as CRIA_CatBoost, CRIA_SVM and CRIA_LightGBM. From **Figure 2**, it can be seen that the highest accuracy of 86.90% for CRIA_CatBoost method followed by 86.41% for CRIA_LightGBM and 85.98% for CRIA_SVM method, with only using the CNVs of 131 genes, 138 genes and 122 genes respectively.

The Proposed Algorithm Performance

For the different classifiers used in this work, after determining the optimal numbers of features according to the CRIA and IFS results, the classification performance can be further analyzed. The average values of three metrics-precision, recall and F1-score defined in Equation (35) on 10 test datasets are listed in **Table 7**.

Performance Comparison With Other Methods

After selecting important features, we use three common classifiers—CatBoost, SVM and LightGBM to predict cancer samples. The performance of our methods are compared with other two classification methods published before whose experimental dataset is the same as ours. Liang et al. (2020) used a method called CNA_origin which was composed of

TABLE 8 | Performance comparison of the proposed algorithm predictions with those of other methods (%).

Cancer	Predictor	Precision	Recall	F1-score
UCEC	CRIA_CatBoost	**74.31**	**73.14**	**73.72**
	CRIA_SVM	70.47	**73.81**	**72.13**
	CRIA_LightGBM	71.46	**72.91**	**72.18**
	CNA_origin	67.92	72.00	69.90
	mRMR_Dagging	74.19	46.93	57.50
KIRC	CRIA_CatBoost	**93.74**	91.63	**92.67**
	CRIA_SVM	**90.33**	89.59	89.96
	CRIA_LightGBM	**93.37**	92.04	**92.70**
	CNA_origin	88.89	**96.00**	92.31
	mRMR_Dagging	80.85	92.68	86.36
OV	CRIA_CatBoost	84.59	**87.90**	86.21
	CRIA_SVM	85.40	**88.43**	86.89
	CRIA_LightGBM	82.84	**89.32**	85.96
	CNA_origin	**89.80**	86.72	**88.00**
	mRMR_Dagging	84.61	75.86	80.00
GBM	CRIA_CatBoost	**94.63**	**90.76**	**92.65**
	CRIA_SVM	**95.64**	**89.70**	**92.57**
	CRIA_LightGBM	**95.72**	**91.30**	**93.46**
	CNA_origin	93.10	84.38	88.52
	mRMR_Dagging	88.70	85.93	87.30
COADREAD	CRIA_CatBoost	**89.67**	**86.09**	**87.84**
	CRIA_SVM	**88.54**	**83.30**	**85.84**
	CRIA_LightGBM	**90.23**	**83.48**	**86.72**
	CNA_origin	81.58	73.81	77.50
	mRMR_Dagging	60.00	73.46	66.05
BRCA	CRIA_CatBoost	84.48	88.67	86.52
	CRIA_SVM	84.76	87.96	86.33
	CRIA_LightGBM	84.71	87.01	85.84
	CNA_origin	**87.50**	**92.31**	**89.84**
	mRMR_Dagging	79.16	87.35	83.06

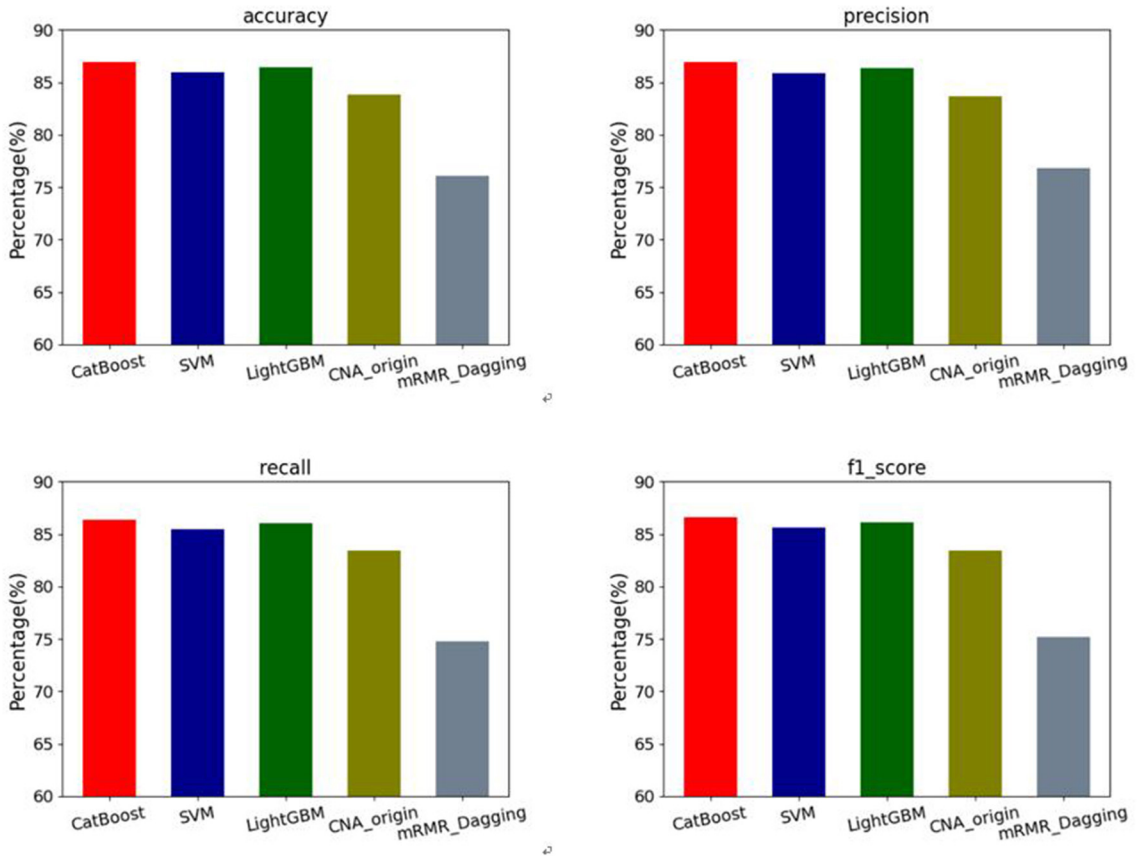


FIGURE 3 | Performance comparison of 4 evaluation metrics: accuracy, precision, recall and F1-score among our methods and the other two algorithms (CNA_origin and mRMR_Dagging).

a stacked autoencoder and an one-dimensional convolutional neural network. The 24,174 gene features were extracted to 100 genes by the autoencoder, and then these 100 gene features were put into the 1D CNN for classification (Liang et al., 2020). A computationally method for cancer types classification proposed by Zhang et al. (2016) was named as mRMR_Dagging here because there was no specific method name given by authors. It first used mRMR and IFS to select 19 of the 24,174 genes as classification features, and then used the Dagging algorithm to give the final results.

In **Table 8**, it can be seen that if the results of our methods are superior to CNA_origin and mRMR_Dagging, they are marked in bold. Similarly, if the largest of CNA_origin and mRMR_Dagging results is better than our method, it is also marked in bold. **Table 8** demonstrated that the performance of our methods is superior to CNA_origin and mRMR_Dagging for UCEC, KIRC, GBM, and COADREAD. For UCEC, the recall and F1-score of our methods (CRIA_Cat- Boost, CRIA_SVM and CRIA_LightGBM) are all superior to CNA_origin and mRMR_Dagging. The best precision of our methods is 0.12 percentage points higher than mRMR_Dagging. SVM and LightGBM are slightly worse than mRMR_Dagging with reductions of 5.28 and 3.82% in precision respectively. For KIRC, the precision and F1-score are all superior to CNA_origin and mRMR_Dagging except the F1-score of SVM, which performs slightly worse than the CNA_origin with reductions of 2.61%. Compared with the best, CNA_origin, the recall of our methods are decreased by 4.77% for CatBoost, 7.15% for SVM and 4.30% for LightGBM. For OV, compared with CNA_origin, the recall of our methods is at least increased by 1.36%. The precision and F1-score are slightly worse than CNA_origin, with reductions at most of 8.40, and 2.37%, respectively. For GBM and COADREAD, our methods are better than CNA_origin and mRMR_Dagging on all evaluation indicators. Compared with the best of the other two algorithms, the worst precision of our methods is increased by 1.64 and 8.53%, respectively, the worst recall is increased by 4.39 and 12.86%, respectively, and the worst F1-score is increased by 4.58 and 10.76%, respectively. For BRCA, the worst among our methods performs slightly worse than the best CNA_origin algorithm, with reductions of 3.57% in precision, 6.09% in recall and 4.66% in F1-score respectively.

In addition, the macro-average results of four evaluation metrics: accuracy, precision, recall and F1-score are used to assess our methods and the other two algorithms on the datasets of six types of cancers. The results can be seen in **Figure 3**. For accuracy, our methods have mean values of

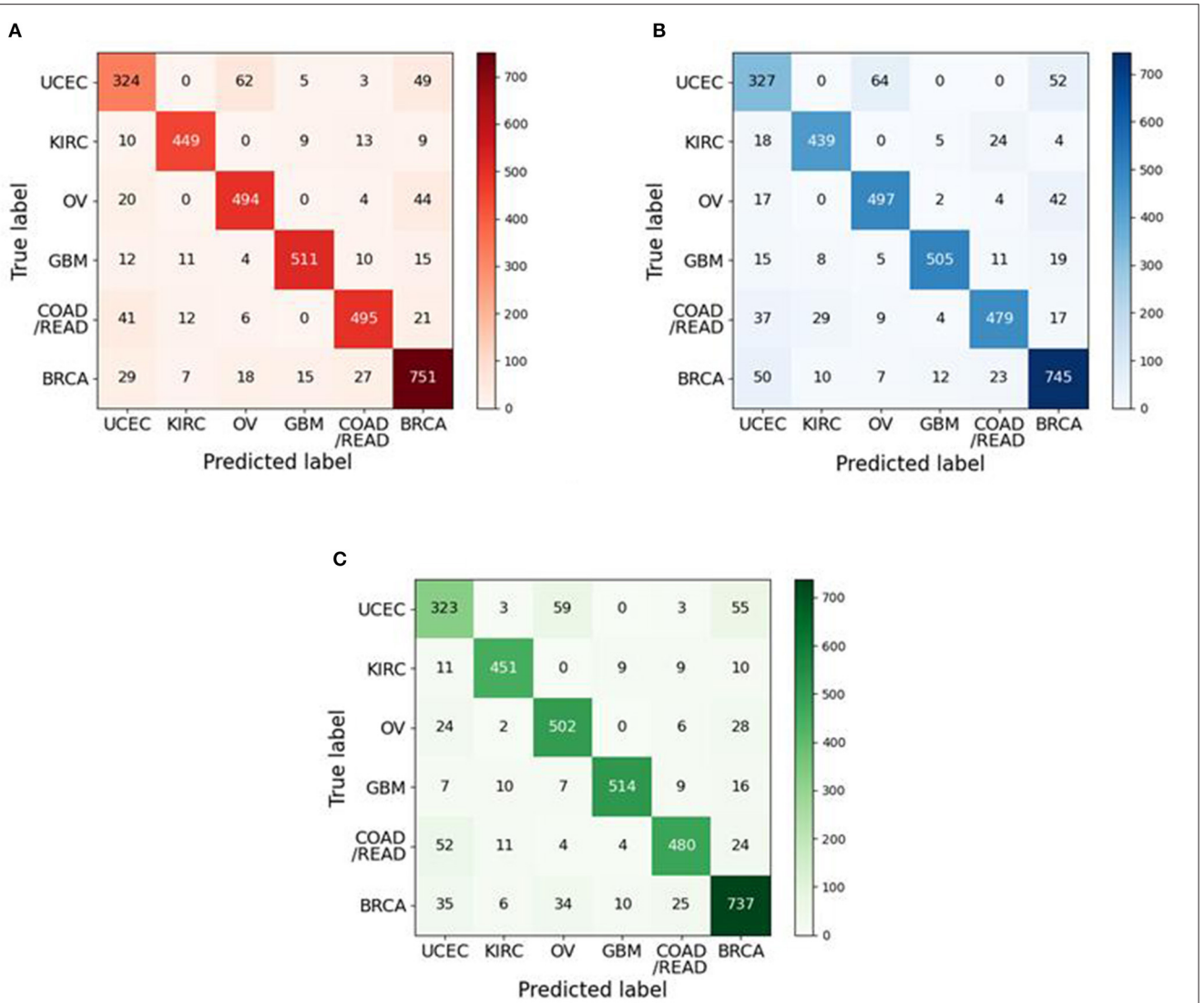


FIGURE 4 | Confusion matrices on the test groups: **(A)** CRIA_CatBoost; **(B)** CRIA_SVM; **(C)** CRIA_LightGBM.

86.90% for CatBoost, 86.41% for LightGBM and 85.98% for SVM respectively, which are increased by 3.69, 3.10, and 2.59% compared with CNA_origin. For precision, the average values of our methods are 86.61, 86.39, and 85.86%, which are increased by 3.49, 3.23, and 2.59%, respectively compared with the best among CNA_origin and mRMR_Dagging. For recall, our methods' mean values are 86.37, 86.01, and 85.47%, which are 2.92, 2.56 and 2.02 percentage points higher than CNA_origin, respectively. For F1-score, compared with our methods, whose average values are 86.60, 86.14, and 85.62%, CNA_origin is decreased by 3.71, 3.19, and 2.60%, respectively.

Further Discussion

In order to study the relationship between the classes, we also summarize the confusion matrices in **Figure 4** for class predictions using our methods. From **Figure 4**, we can find that there existed a high error rate when predicting the samples of UCEC. Regardless of whether it is CRIA_CatBoost, CRIA_SVM or CRIA_LightGBM, more than 10% of the UCEC samples are incorrectly predicted as OV and BRCA. In **Figure 4A**, 14.00% of UCEC samples are predicted as OV, while 11.06% of UCECsamples are predicted to be BRCA. In **Figure 4B**, 14.45 and 11.74% of UCEC samples are predicted as OV and BRCA respectively. In **Figure 4C**, 13.32 and 12.42% of UCEC samples are predicted as OV and BRCA respectively. The reasons may be that UCEC, OV and BRCA are hormone-dependent tumors and they relate closely in tumorigenesis. The 16 and 27 risk regions were identified by an independent genome-wide association study (GWAS) on endometrial cancer and ovarian cancer, respectively (Glubb et al., 2020). Studies have shown that mutations in breast cancer susceptibility genes (BRCA1, BRCA2) have a relationship in hereditary ovarian cancer. Mutations

at either end of the BRCA1 gene increase a person's risk of breast cancer, and its probability is higher than ovarian cancer. However, mutations in the middle of the BRCA1 gene put a person at a higher risk of ovarian cancer than breast cancer (Shi et al., 2017). In addition, there is also a study indicated that UCEC, OV and BRCA all have a relationship with the changes in estrogen and estrogen receptors (Rodriguez et al., 2019).

CONCLUSIONS

In this paper, we introduce a gene selection algorithm—CRIA. We firstly apply this algorithm to 5 datasets and verify the effective performance of CRIA through comparison with other eight gene selection algorithms. The proposed algorithm can select features which are closely related to the class label. Then, we use this algorithm to select 200 genes that have a close relationship with cancer types from 24,174 genes features based on the value of copy number variations in the samples, and then combine three common classifiers—CatBoost, SVM and LightGBM to predict the type of cancer. Our experimental results show that our methods have higher accuracies than the state-of-the-art methods for solving this problem. Our research has a certain degree of interpretability for cancer-related researches at the genetic level. As we all know, cancer is closely related to gene structural variations and the appearance of cancer is often accompanied by abnormalities in the deoxyribonucleic acid (DNA) sequence. Because CNVs is one of the most crucial structural variations of genes, studying the relationship between cancers and CNVs is of great significance. Many studies have tried to utilize the genetic information of cancers to predict cancer type, which can provide significant guidance for patient care and cancer therapy in promptly.

The future direction of this work can continue to develop from two aspects. First of all, because we only use the datasets of six cancer types and the total number of samples is only 3,480 in this paper, by collecting data sets of other cancer types and optimizing the proposed algorithm, we can continue to conduct further research in the field of cancer classification based on copy number variations. Moreover, integrating non-CNVs features for the samples can be taken into consideration. In addition to using CNVs for cancer prediction, we can also apply other genetic information for cancer prediction, or combine several biomarkers to reduce the error rate of classification as much as possible.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

QW conducted the experiments and wrote the manuscript. DL conceived and provided the main direction of the manuscript and guided the writing and modification of this manuscript. Both authors read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 11571009) and Applied Basic Research Programs of Shanxi Province (Grant No. 201901D111086).

REFERENCES

Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* 5, 537–550. doi: 10.1109/72.298224

Bennasar, M., Hicks, Y., and Setchi, R. (2015). Gene selection using Joint Mutual Information Maximisation. *Expert Syst. Appl.* 42, 8520–8532. doi: 10.1016/j.eswa.2015.07.007

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., et al. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and app- lication to glioma. *Proc. Natl. Acad. Sci.* 104, 20007–20012. doi: 10.1073/pnas.0710052104

Breunis, W. B., van Mirre, E., Bruin, M., Geissler, J., de Boer, M., Peters, M., et al. (2008). Copy number variation of the activating FCGR2C gene predisposes to idiopathic thrombocytopenic purpura. *Blood* 111, 1029–1038. doi: 10.1182/blood-2007-03-079913

Buchynska, L. G., Brieieva, O. V., and Iurchenko, N. P. (2019). Assessment of HER-2/neu, α-MYC and CCN- E1 gene copy number variations and protein expression in endometrial carcinomas. *Exp. Oncol.* 41. doi: 10.32471/exp-oncology.2312-8852.vol-41-no-2.12973

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer geno- mics data: figure 1. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Chandrashekar, G., and Sahin, F. (2014). A survey on gene selection methods. *Comput. Electr. Eng.* 40, 16–28. doi: 10.1016/j.compeleceng.2013.11.024

Chen, Z., Wu, C., Zhang, Y., Huang, Z., Ran, B., Zhong, M., et al. (2015). Gene selection with redundancy-complementariness dispersion. *Knowl. Based Syst.* 89, 203–217. doi: 10.1016/j.knosys.2015.07.004

Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133. doi: 10.1038/ng.2762

Cover, T. M., and Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: John Wiley and Sons.

Du, W., and Elemento, O. (2015). Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. *Oncogene* 34, 3215–3225. doi: 10.1038/onc.2014.291

Elia, J., Gai, X., Xie, H. M., Perin, J. C., Geiger, E., Glessner, J. T., et al. (2010). Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmen- tal genes. *Mol. Psychiatry* 15, 637–646. doi: 10.1038/mp.2009.57

Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. A. (2009). Normalized Mutual Information Gene selection. *IEEE Trans. Neural Netw.*20, 189–201. doi: 10.1109/TNN.2008.2005601

Fayyad, U. M., and Irani, K. B. (1993). "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," in *Pro-ceedings of International Joint Conference on Artificial Intel- ligence, pp* 1022–1027

Flierl, A., Oliveira Luís, M. A., Falomir-Lockhart Lisandro, J., Mak Sally, K., Hesley, J., Soldner, F., et al. (2014). Higher vulnerability and stress sensitivity of neuronal precursor cells carrying an alpha-synuclein gene triplication. *PLoS ONE* 9, e112413. doi: 10.1371/journal.pone.0112413

Foithong, S., Pinngern, O., and Attachoo, B. (2012). Feature subset selection wrapper based on mutual information and rough sets. *Expert Syst. Appl.* 39, 574–584. doi: 10.1016/j.eswa.2011.07.048

Frank, B., Bermejo, J. L., Hemminki, K., Sutter, C., Wappenschmidt, B., Meindl, A., et al. (2007). Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. *Carcinogenesis* 28, 1442–1445. doi: 10.1093/carcin/bgm033

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the Cbioportal. *Sci. Signaling* 6, pl1–pl1. doi: 10.1126/scisignal.2004088

Gao, W., Hu, L., and Zhang, P. (2018b). Class-specific mutual information variation for gene selection. *Pattern Recogn.* 79, 328–339. doi: 10.1016/j.patcog.2018.02.020

Gao, W., Hu, L., Zhang, P., and He, J. (2018a). Gene selection considering the composition of feature relevancy. *Pattern Recogn. Lett.* 112, 70–74. doi: 10.1016/j.patrec.2018.06.005

Glubb, D. M., Thompson, D. J., Aben, K. K., Alsulimani, A., Amant, F., Annibali, D., et al. (2020). Cross-cancer genome-wide association study of endometrial cancer and epithelial ovarian cancer identifies genetic risk regions associated with risk of both cancers. *Cancer Epidemiol. Biomarkers Prev.* 30, 217–28. doi: 10.1158/1055-9965.EPI-20-0739

Grangeon, L., Cassinari, K., Rousseau, S., Croisile, B., Formaglio, M., Moreaud, O. et al. (2021). Early-onset cerebral amyloid angiopathy and alzheimer disease related to an app locus triplication. *Neurol. Genet.* 7, e609–e609. doi: 10.1212/NXG.0000000000000609

Gu, X., Guo, J., Li, C., and Xiao, L. (2020). A gene selection algorithm based on redundancy analysis and interaction weight. *Appl. Intell.* 51, 2672–2686. doi: 10.1007/s10489-020-01936-5

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor Newsl.* 11, 10–18. doi: 10.1145/1656274.1656278

Heo, Y., Heo, J., Han, S., Kim, W. J., Cheong, H. S., and Hong, Y. (2020). Difference of copy number variation in blood of patients with lung cancer. *Int. J. Biol. Markers* 36, 3–9. doi: 10.1177/1724600820980739

Jakulin, A. (2003). *Attribute Interactions in Machine Learning (Master thesis). Computer and Information Science, University of Ljubljana.*

Jakulin, A., and Bratko, I. (2004). "Testing the significance of attribute interactions," in *Proceedings of the Twenty-first international conference on Machine learning - ICML'04.* Banff, AL: ACM Press. pp. 409–416.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et al. (2017). Gene selection: a data perspective. *ACM Comput. Surv.* 50, 1–45. doi: 10.1145/3136625

Liang, J., Hou, L., Luan, Z., and Huang, W. (2019). Gene selection with conditional mutual information considering feature interaction. *Symmetry* 11, 858. doi: 10.3390/sym11070858

Liang, Y., Wang, H., Yang, J., Li, X., Dai, C., Shao, P., et al. (2020). A deep learning framework to predict tumor tissue-of-origin based on copy number alteration. *Front. Bioeng. Biotech.* 8, 701. doi: 10.3389/fbioe.2020.00701

Ma, J., and Sun, Z. (2011). Mutual information is copula entropy. *Tsinghua Sci. Technol.* 16, 51–54. doi: 10.1016/S1007-0214(11)70008-6

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number al teration in human cancers. *Genome Biol.* 12, 4. doi: 10.1186/gb-2011-12-4-r41

Orsenigo, C., and Vercellis, C. (2013). A comparative study of non-linear manifold learning methods for cancer microarray data classification. *Expert Syst. Appl.* 40, 2189–2197. doi: 10.1016/j.eswa.2012.10.044

Pandey, G. N., Rizavi, H. S., Tripathi, M., and Ren, X. (2015). Region-specific dysregulation of glycogen synthase kinase-3β and β-catenin in the postmortem brains of subjects with bipolar disorder and schizophrenia. *Bipolar Disord.* 17, 160–171. doi: 10.1111/bdi.12228

Peng, H., Long, F., and Ding, C. (2005). Gene selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE T. Pattern Anal.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454. doi: 10.1038/nature05329

Rodriguez, A. C., Blanchard, Z., Maurer, K. A., and Gertz, J. (2019). Estrogen signaling in endometrial cancer: a key oncogenic pathway with several open questions. *HORM. CANC.* 10, 51–63. doi: 10.1007/s12672-019-0358-9

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.* 5, 3–55. doi: 10.1145/584091.584093

Shi, T., Wang, P., Xie, C., Yin, S., Shi, D., Wei, C., et al. (2017). *BRCA1* and *BRCA2* mutations in ovarian cancer patients from China: ethnic-related mutations in *BRCA1* associated with an increased risk of ovarian cancer: *BRCA1/2* mutation in Chinese ovarian cancer. *Int. J. Cancer* 140, 2051–2059. doi: 10.1002/ijc.30633

Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., and Wang, K. (2013). Gene selection using dynamic weights for classification. *Knowl. Based Syst.* 37, 541–549. doi: 10.1016/j.knosys.2012.10.001

Tian, T., Bi, H., Liu, Y., Li, G., Zhang, Y., Cao, L., et al. (2020). Copy number variation of ubiquitin- specific proteases genes in blood leukocytes and colorectal cancer. *Cancer Biol. Ther.* 21, 637–646. doi: 10.1080/15384047.2020.1750860

Van Bockstal, M. R., Agahozo, M. C., van Marion, R., Atmodimedjo, P. N., Sleddens, H. F. B. M., Dinjens, W. N. M., et al. (2020). Somatic mutations and copy number variations in breast cancers with heterogeneous *HER2* amplification. *Mol. Oncol.* 14, 671–685. doi: 10.1002/1878-0261.12650

Wang, J., Wei, J. M., Yang, Z., and Wang, S. Q. (2017). Gene selection by Maximizing Independent Classification Information. *IEEE Trans. Knowl. Data Eng.* 29, 828–841. doi: 10.1109/TKDE.2017.2650906

Witten, I. H., and Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *SIGMOD Rec.* 31, 76–77. doi: 10.1145/507338.507355

Yang, Y., Song, S., Chen, D., and Zhang, X. (2019). Discernible neighborhood counting based incremental gene selection for heterogeneous data. *Int. J. Mach. Learn. Cybern.* 11, 1115–1127. doi: 10.1007/s13042-019-00997-4

Zeng, Z., Zhang, H., Zhang, R., and Yin, C. (2015). A novel gene selection method considering feature interaction. *Pattern Recogn.* 48, 2656–2666. doi: 10.1016/j.patcog.2015.02.025

Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochim. Biophys. Acta, Gen. Subj.* 1860, 2750–2755. doi: 10.1016/j.bbagen.2016.06.003

Zheng, Z., Yu, R., Gao, C., Jian, X., Quan, S., Xing, G., et al. (2017). Low copy number of FCGR3B is associated with lupus nephritis in a Chinese population. *Exp. Ther. Med.* 14, 4497–4502. doi: 10.3892/etm.2017.5069

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 April 2022
doi: 10.3389/fpls.2022.841855

The Use and Limitations of Exome Capture to Detect Novel Variation in the Hexaploid Wheat Genome

Amanda J. Burridge[1], Mark O. Winfield[1], Paul A. Wilkinson[2], Alexandra M. Przewieslik-Allen[1], Keith J. Edwards[1] and Gary L. A. Barker[1]*

[1] School of Life Sciences, University of Bristol, Bristol, United Kingdom, [2] Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, United Kingdom

The bread wheat (*Triticum aestivum*) pangenome is a patchwork of variable regions, including translocations and introgressions from progenitors and wild relatives. Although a large number of these have been documented, it is likely that many more remain unknown. To map these variable regions and make them more traceable in breeding programs, wheat accessions need to be genotyped or sequenced. The wheat genome is large and complex and consequently, sequencing efforts are often targeted through exome capture. In this study, we employed exome capture prior to sequencing 12 wheat varieties; 10 elite *T. aestivum* cultivars and two *T. aestivum* landrace accessions. Sequence coverage across chromosomes was greater toward distal regions of chromosome arms and lower in centromeric regions, reflecting the capture probe distribution which itself is determined by the known telomere to centromere gene gradient. Superimposed on this general pattern, numerous drops in sequence coverage were observed. Several of these corresponded with reported introgressions. Other drops in coverage could not be readily explained and may point to introgressions that have not, to date, been documented.

Keywords: wheat, *Triticum aestivum*, introgression, exome capture, exome capture sequencing, sequence variation

OPEN ACCESS

Edited by:
Surya Saha, Boyce Thompson Institute (BTI), United States

Reviewed by:
Weilong Kong, Wuhan University, China
Zhenyang Liao, Agricultural Genomics Institute at Shenzhen (CAAS), China
Dong Xu, Laboratory of Genome Analysis, Agricultural Genomics Institute at Shenzhen (CAAS), China

****Correspondence:***
Amanda J. Burridge
amanda.burridge@bristol.ac.uk

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *22 December 2021*
Accepted: *28 February 2022*
Published: *12 April 2022*

Citation:
Burridge AJ, Winfield MO, Wilkinson PA, Przewieslik-Allen AM, Edwards KJ and Barker GLA (2022) The Use and Limitations of Exome Capture to Detect Novel Variation in the Hexaploid Wheat Genome. Front. Plant Sci. 13:841855. doi: 10.3389/fpls.2022.841855

INTRODUCTION

The bread wheat (*Triticum aestivum*) pangenome is a patchwork containing translocations and introgressions from wheat's wild relatives (Przewieslik-Allen et al., 2021) as well as numerous deletions. Some of these features may be present in only a handful of accessions coming from a limited geographic area whilst others may be prevalent and present in varying combinations across many accessions. Some variable regions may have occurred naturally by mutation or as a consequence of promiscuous pollination events between wheat and one of its primary relatives (He et al., 2019). Others are the result of breeding efforts (Schneider et al., 2008) using traditional methods to introduce segments from progenitors and close relatives or, more recently, using more advanced methods to perform wide crosses (Cseh et al., 2019; Devi et al., 2019; King et al., 2019; Xu et al., 2020). Regardless of their origin, the number of these variable regions that have been documented is probably not a genuine reflection of their true number; breeding companies may not have reported, and indeed may not know, all the introgressed regions in their elite lines, and chance events in landrace accessions are unlikely to have been documented at all. It would seem

highly likely, therefore, that there are numerous unknown, introgressions present in modern wheat accessions (Przewieslik-Allen et al., 2021).

With this in mind, and with modern techniques allowing for wide crossing with increasing success, increasingly diverse wheat accessions are becoming available for pre-breeding (Hao et al., 2020). To be of use to research and breeding programs, such material needs to be tracked using either targeted molecular markers (Singh et al., 2018; Rasheed and Xia, 2019) or sequencing. The former has most frequently been used because it offers low cost and high throughput (Zhang J. et al., 2017; Zhang W. et al., 2017; Przewieslik-Allen et al., 2019). However, marker probes will only hybridize to, and so provide a signal for, the sequences for which they were designed. Thus, wheat genotyping markers intended for introgression detection need to be designed using sequences from a combination of wheat and the progenitors and relatives thought to have been the source of those introgressions (Wang et al., 2014; Zhang J. et al., 2017; Przewieslik-Allen et al., 2019). Where the source of introgressed material is unknown, and so not included in probe design, genotyping is unlikely to track such regions.

Sequencing, having no requirement for prior knowledge of the target, does not suffer from such a problem. However, the size and complexity of the wheat genome create problems in this regard. *T. aestivum* has a large (∼17 Gb) polyploid and highly repetitive genome of which the exome constitutes less than 5% (International Wheat Genome Sequencing Consortium [IWGSC], 2014). To sidestep these issues, targeted sequencing approaches, such as exome capture, are used (Kaur and Gaikwad, 2017). In wheat, several exome capture systems that incorporate capture probe sets derived from both hexaploid wheat and its relatives have been proposed (Winfield et al., 2012; Gardiner et al., 2019; He et al., 2019). The capture probes themselves can tolerate some degree of mismatch thus allowing the capture of sequences outside the immediate confines of the species from which they are derived. The Roche SeqCap EZ system can tolerate up to 10% (Roche pers com) and the Arbor Biosciences myBaits system can tolerate up to 20% divergence from the target sequence (Arbor Biosciences, 2021). This property is highly beneficial where the exact source of the material is unknown and has been exploited to capture sequences from diverse origins in the wild relative species of cotton (Salmon et al., 2012), cows (Cosart et al., 2011), and humans (Jin et al., 2012) as well as in wheat (Saintenac et al., 2011; Henry et al., 2014; He et al., 2019).

We recently described the variable sequence coverage of the wheat variety 'Player' when exome capture data were aligned to the 'Chinese Spring' reference sequence (Przewieslik-Allen et al., 2021). Distinct drops in sequence coverage were evident in chromosomes 2A and 2B which correlated with introgressions from *Aegilops ventricosa* and *Triticum timopheevii*, respectively. As the use of exome capture prior to sequencing followed by alignment to a standard reference is common practice, the potential for this to be disrupted by introgressions is a concern, especially as many interesting, rare, and novel alleles may be located in regions derived from wild relatives. This was investigated using 10 elite *T. aestivum* cultivars and 2 *T. aestivum* landrace accessions used in breeding.

RESULTS

Sequence Coverage

Using gene and promoter sequence capture (Gardiner et al., 2019), 12 *T. aestivum* accessions (10 elite varieties and 2 landrace accessions) were sequenced and total coverage compared. Total reads were between 48,255,718 and 145,897,760 per accession; after quality trimming and alignment to the IWGSC RefSeq v1.0 'Chinese Spring' reference (International Wheat Genome Sequencing Consortium [IWGSC], 2018), there were between 20,973,857 and 63,662,179 uniquely mapped, paired reads per accession (**Table 1**).

Sequence coverage across chromosomes displayed a characteristic pattern; that is, there was a greater depth of coverage toward the ends of chromosome arms and lower coverage across centromeres (**Figure 1A**). However, this overall pattern was, in places, interrupted by regions of pronounced reduction in sequence coverage. These regions were not seen on all chromosomes or simultaneously in all accessions (**Supplementary File 1**). The most pronounced of these reductions in coverage was observed in 'Bacanora', 'Bobwhite', and 'KWS Kielder' and extended across the whole of the short arm of chromosome 1B (c. 240 Mb; **Figure 1B**) in line with the well documented and prevalent 1RS/1BL *Secale cereale* translocation (Rabinovich, 1998). There was no reduction in capture probe density across 1BS (**Figure 1C**) and the nine accessions without the 1RS translocation do not show a reduction in read coverage across this chromosome arm (**Supplementary File 1**: 1B).

Other large drops in coverage were seen on 2BL, 2DL, and 5BL (**Figure 2**) which extended over approximately 85, 45, and 40 Mb, respectively. Additional, smaller drops in coverage were also observed in telomeric regions, such as 2AS (**Figure 3**), 7DL, and an additional region in 2DL (**Supplementary File 1**).

Cluster Analysis of Accessions

To determine whether there was any relationship between the lines that shared read coverage profiles, cluster analysis was performed with Axiom 35K Wheat Breeders' Array genotyping data (Allen et al., 2016). Analysis was performed on markers specific to the chromosomes 2B (2,083 markers), 2D (2,237 markers), and 5B (1,749 markers). Accessions showed a pattern of clustering that corresponded with the drops in coverage (**Figure 2B** and **Supplementary File 2**). For chromosome 5B, for example, the 12 accessions separated into two main clusters; the accessions thought to contain the deletion fell into one cluster while those with even sequence coverage fell into the other. The separation into two clusters was driven by the markers spanning the drop. Across the interval corresponding to the decline in read coverage on chromosome 5B (position 499,569,304–534,345,241), there were 141 single nucleotide polymorphism (SNP) markers; for these markers, the mean percentage similarity between the genotype calls for 'Chinese Spring' and those of the eight accessions displaying the drop in coverage was only 13.3%. This compares to a mean similarity of 59.1% for the SNP calls across the rest of the chromosome (**Figure 2C**).

TABLE 1 | Read statistics before and after trimming with alignment statistics for total mapped and uniquely mapped reads.

Variety	Total reads	Trimmed reads	Total mapped paired reads	Uniquely mapped paired reads
Apogee	66,890,848	64,719,732 (96.8%)	33,445,424	28,796,984 (86.1%)
Bacanora	80,558,016	77,878,454 (96.7%)	38,939,227	33,411,894 (85.8%)
Bobwhite	57,245,246	55,371,124 (96.7%)	27,685,562	24,139,596 (87.2%)
Boregar	48,255,718	46,603,262 (96.6%)	24,127,859	20,973,857 (86.9%)
Cadenza	72,469,144	70,237,432 (96.9%)	36,234,572	31,033,245 (85.6%)
KWS Kielder	65,891,042	63,791,268 (96.8%)	32,945,521	28,236,689 (85.7%)
Maris Huntsman	52,673,928	50,665,506 (96.2%)	25,332,753	21,992,928 (86.8%)
Pavon 76	145,897,760	141,207,878 (96.8%)	72,948,880	63,662,179 (87.3%)
Renan	56,921,314	54,670,680 (96.0%)	28,460,657	24,854,489 (87.3%)
Riband	68,965,154	66,392,980 (96.3%)	33,196,490	28,910,396 (87.0%)
Watkins 141	51,300,352	49,480,812 (96.5%)	24,740,406	21,238,652 (85.8%)
Watkins 777	99,274,852	95,527,090 (96.2%)	49,637,426	43,055,487 (86.7%)

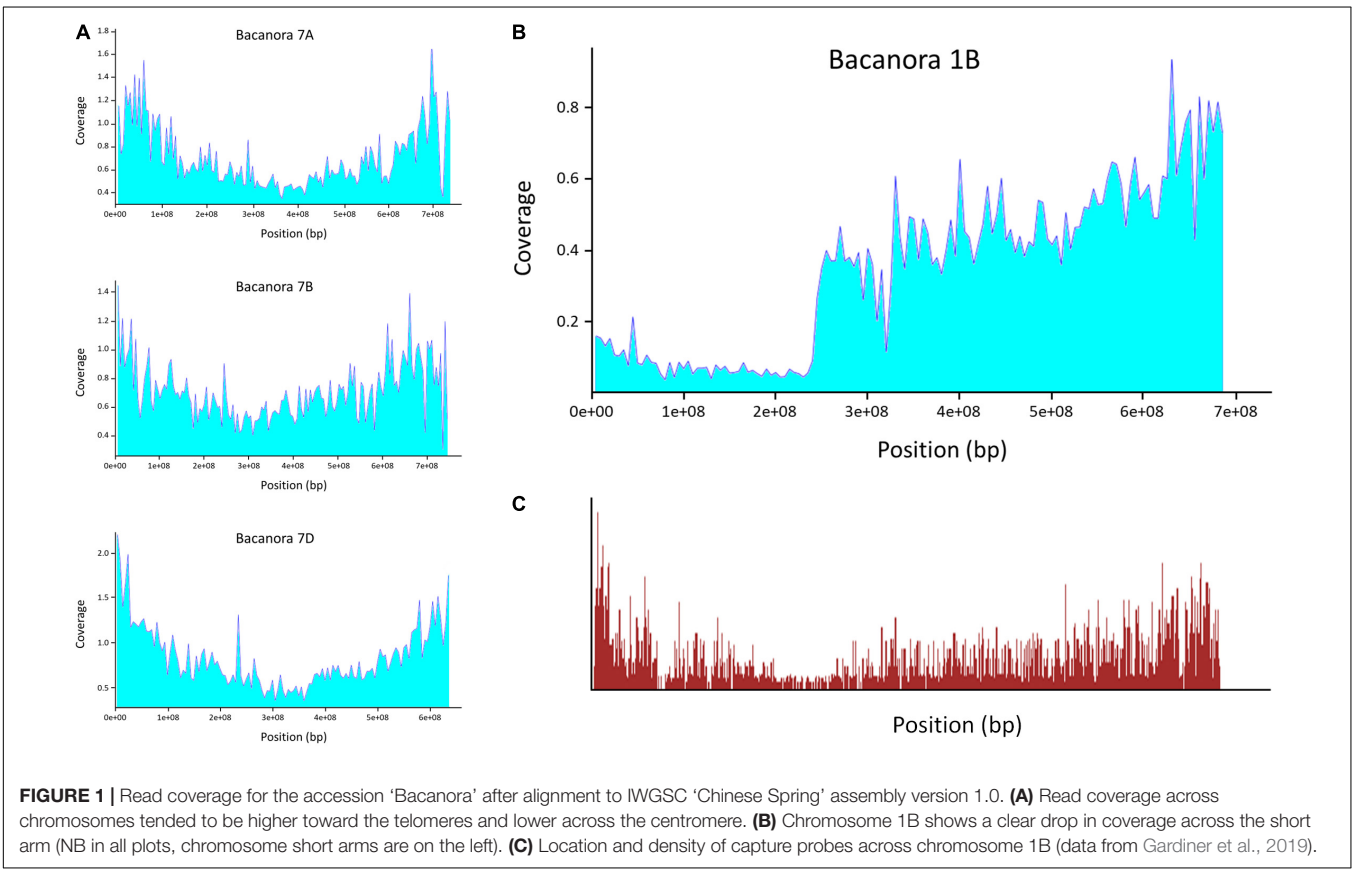


FIGURE 1 | Read coverage for the accession 'Bacanora' after alignment to IWGSC 'Chinese Spring' assembly version 1.0. **(A)** Read coverage across chromosomes tended to be higher toward the telomeres and lower across the centromere. **(B)** Chromosome 1B shows a clear drop in coverage across the short arm (NB in all plots, chromosome short arms are on the left). **(C)** Location and density of capture probes across chromosome 1B (data from Gardiner et al., 2019).

Bibliographic Search for Introgressions

A number of wheat introgressions reported in the literature were assembled (**Table 2**) to determine whether there was any relationship between them and the patterns of reduced sequence coverage observed in this study. The large drop in coverage on 1BS, for example, is present in those varieties (Bacanora, Bobwhite, and KWS Kielder) known to possess a whole arm translocation from *S. cereale*; we have previously reported this ourselves based on genotyping results using the Axiom High-Density Array (Winfield et al., 2015). Other chromosomal regions with reduced read coverage were also related to regions of known introgressions. However, not all the reports of introgressions that we found in the literature had a corresponding drop in sequence coverage, and in some cases, there was a drop in sequence coverage for which no source was found. Notable deletions, such as that on 1DL of 'Cadenza', highlight the similarity between deletions and introgressions in sequence coverage.

Efficacy of Sequence Capture

The accessions containing the 1RS.1BL translocation ('Bacanora', 'Bobwhite', and 'KWS Kielder') displayed a clear drop in read coverage across the short arm of 1B; we hypothesized that this was

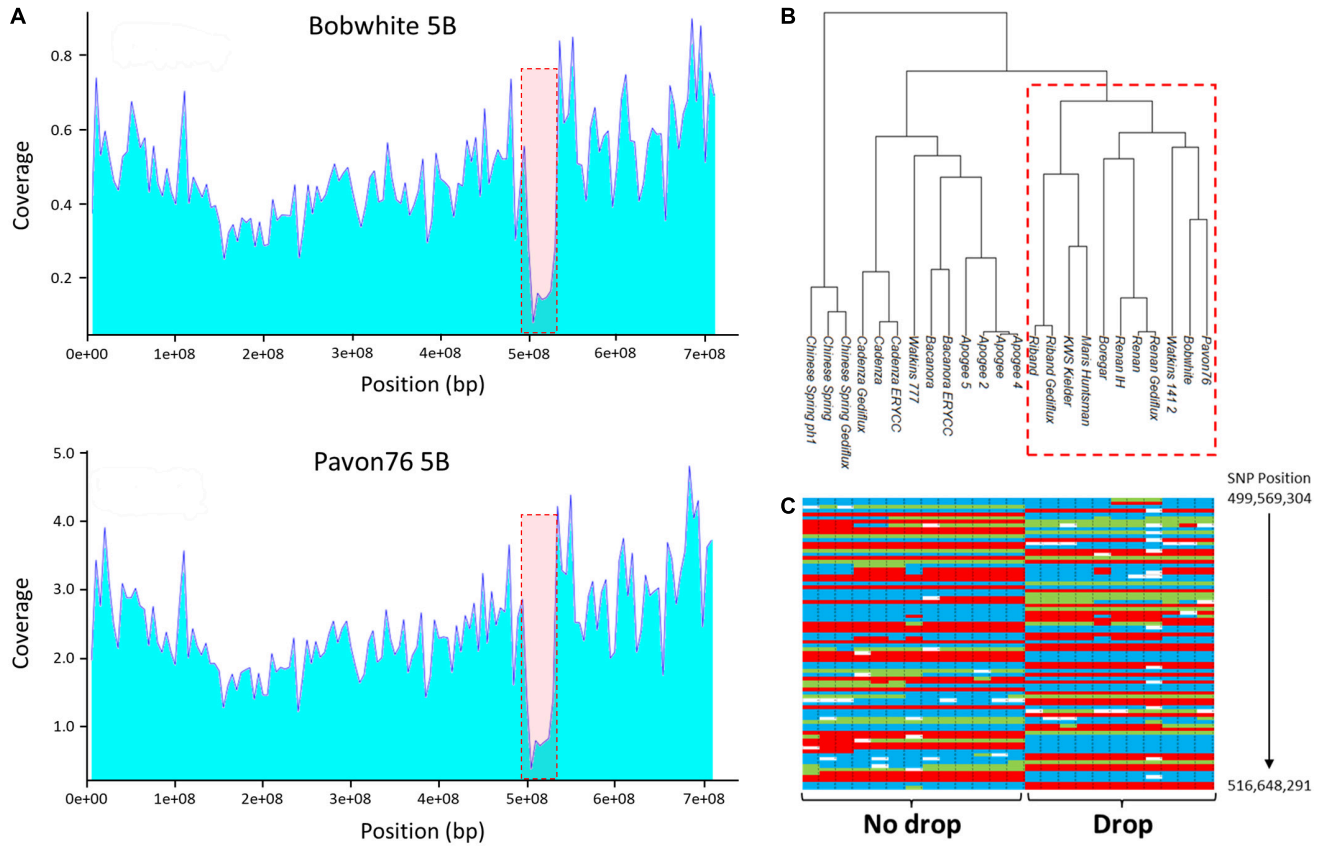


FIGURE 2 | (A) Average depth of coverage for chromosome 5B in the accessions 'Bobwhite' and 'Pavon 76'; both show a drop in read coverage on the long arm at approximately position 490,000,000–540,000,000. **(B)** Dendrogram based on the 1,749 Axiom markers mapped to chromosome 5B; the 8 varieties ('Bobwhite', 'Boregar', 'KWS Kielder', 'Maris Huntsman', 'Pavon 76', 'Renan', 'Riband', and 'Watkins 141') with the drop in read coverage cluster. **(C)** A sample of the SNP calls across the interval 499,569, 304–534,345,241 highlighting the difference between the two groups (blue and red are the alternative homozygote calls; green indicates heterozygote calls).

due to capture efficacy in the different backgrounds. The potential efficacy of probes to capture sequences from either 'Chinese Spring' or *S. cereale* was assessed by BLASTing their sequences to their respective assemblies. Capture probe sequences for chromosome 1BS (26,985 sequences) were BLASTed against the 1B pseudomolecule of 'Chinese Spring' and 1R of *S. cereale*. This resulted in 29,652 hits to 'Chinese Spring' 1BS and 12,120 hits to *S. cereale* 1RS. To both assemblies, some probes had multiple hits. The number of probe sequences that had a hit was 26,222 and 8,419, respectively. Those with a single hit were 23,969 and 5,822, respectively (**Figure 4A**), and the percentage similarity between probe sequences and their target was 99.8 and 95.6%, respectively (**Figure 4B**). That is, a greater number of probes matched the 'Chinese Spring' sequence and with greater percentage similarity.

In contrast, the known *Ae. tauschii* introgression into 5DS of the variety 'Maris Huntsman' (Wang et al., 2005) was not evidenced by a drop in read coverage. The probe sequences for chromosome 5DS (20,253 sequences) were BLASTed against the assemblies of both 'Chinese Spring' and *Ae. tauschii* 5DS resulted in 24,300 hits to the former and 24,173 hits to the latter. The number of probe sequences that had a hit was 20,082 and 19,872, respectively. Those with a single hit were 17,550 and 17,358, respectively (**Figure 4A**). The percentage similarity between probe sequences and their target was 99.1 and 98.9%, respectively (**Figure 4B**). Thus, it would appear, the sequences of wheat and *Ae. tauschii* are sufficiently similar over this region that capture probes are equally efficient at capturing sequences from them. To confirm this hypothesis, the sequences surrounding *Pm2*, were compared. Based on the alignment, the 'Chinese Spring' and *Ae. tauschii* reference assemblies were highly similar across the 2 Mb of sequence centered on the *Pm2* gene (99.1% similarity); in each, there were 21 annotated genes and synteny appears to be maintained apart from the presence of an inverted repeat of TraesCS5D02G044500 (position 43,382,967–43,386,355) to the upstream position 42,989015–42,992,511 – TraesCS5D02G043600 (**Supplementary Table 1**). The sequences from 'Maris Huntsman' also aligned well to both assemblies. However, within the coding sequence of the *Pm2* gene itself, two indels, one particularly relevant, supported the hypothesis that 'Maris Huntsman' is more similar to *Ae. tauschii* than to 'Chinese Spring'. That is, relative to 'Chinese Spring', both *Ae. tauschii* and

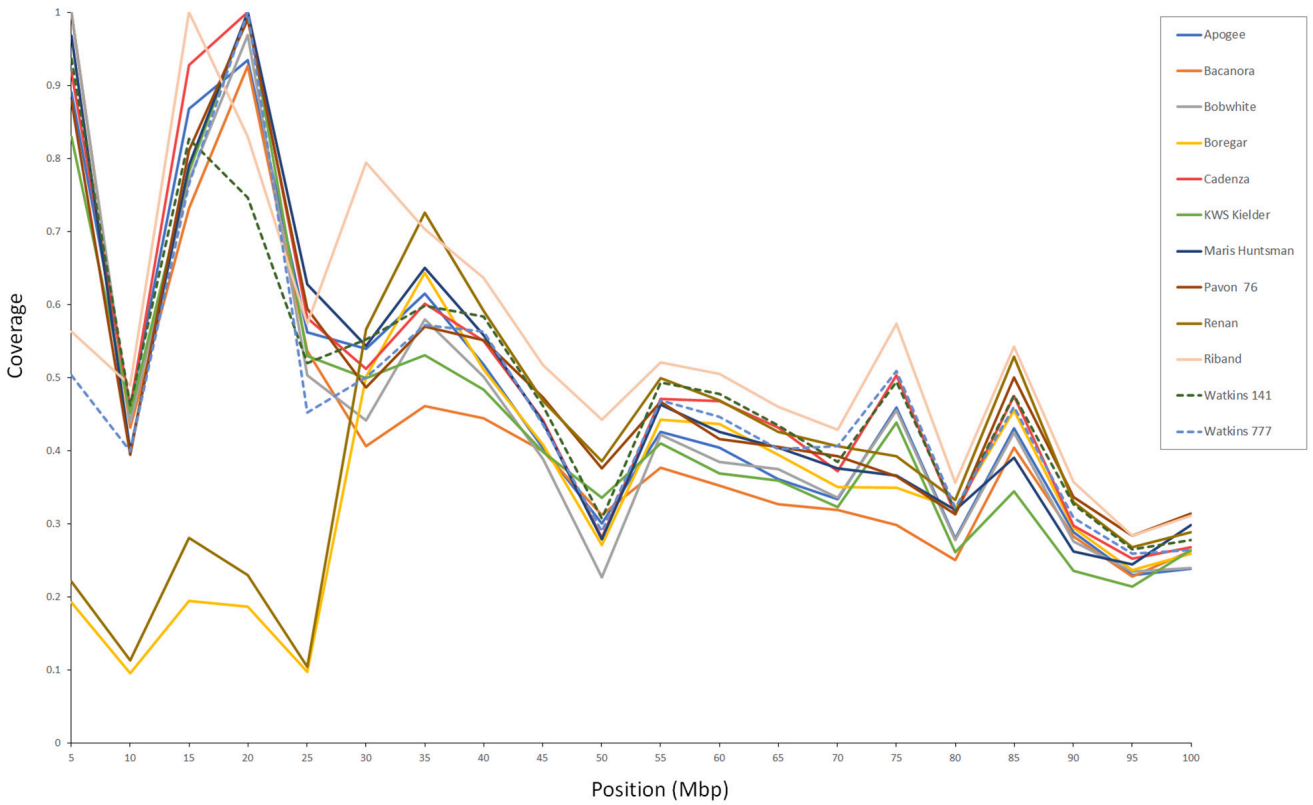


FIGURE 3 | Sequence coverage across the first 100 Mb of chromosome 2AS. The two accessions, 'Boregar' and 'Renan', show reduced coverage across the first 25–30 Mb which corresponds with the size of the known introgression from *Ae. ventricosa* (Robert et al., 1999).

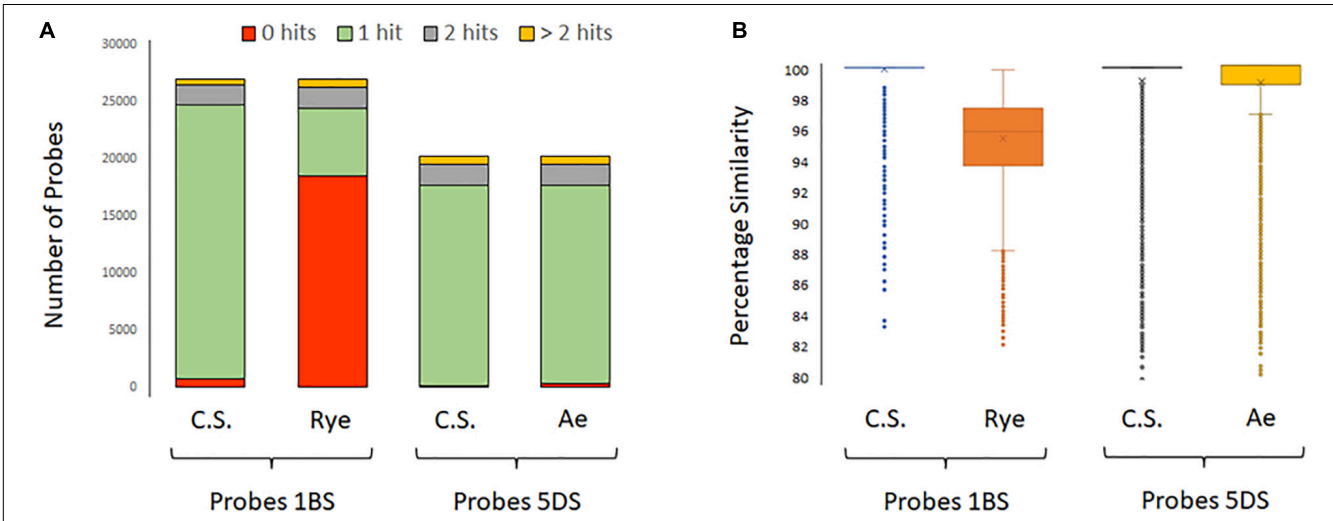


FIGURE 4 | (A) Bar graphs showing the number of capture probes that had BLAST hits to 'Chinese Spring' chromosome 1BS (IWGSC v1), *S. cereale* chromosome 1RS (JADQCU000000000 v1 of the cultivar Weining), 'Chinese Spring' chromosome 5DS (IWGSC v1), and *Ae. tauschii* chromosome 5DS (PRJNA341983 assembly of *Ae. tauschii* subsp. *strangulata*). The number of probe sequences for chromosomes 1BS and 5DS was 26,985 and 20,253, respectively. The number of probes that produced a hit was 26,222 to 'Chinese Spring' 1BS, 8,419 to *S. cereale* 1RS, 20,082 to 'Chinese Spring' 5DS, and 19,872 to *Ae. tauschii* 5DS. There were more hits than probe sequences as some probes had multiple hits. **(B)** Box and whisker plots showing the percentage similarity between the probe sequences and their respective targets.

TABLE 2 | Introgressions and deletions reported in the literature for the accessions in this study.

Cultivar	Gene	Chromosome	Source	References
Bacanora		1BS	*S. cereale*	Driever et al., 2014
	Ppd1	2DS		Driever et al., 2014
Bobwhite		1BS	*S. cereale*	Warburton et al., 2002
	Rht8a	2DS		Worland et al., 1998
Boregar	*Pch1*	7DL	*Ae. ventricosa*	Burt and Nicholson, 2011
Cadenza	*Eps*	1DL	Deletion	Zikhali et al., 2016
	Yr7	2BL	*T. durum*	Marchal et al., 2018
	Yr5	2BL	*T. spelta*	Marchal et al., 2018
	Sbm1	5DL		Kanyuka et al., 2004
	Yr6	7BS	*T. aestivum*	Ma et al., 2015
KWS Kielder		1BS	*S. cereale*	Przewieslik-Allen et al., 2021
Maris Huntsman	*Yr3a*	1B	*T. aestivum*	Bai et al., 2014
	Pm6	2BL	*T. timopheevii*	Wang et al., 2005
	Yr13	2BS	*T. aestivum*	Bai et al., 2014
	Lr13	2BS	*T. aestivum*	McIntosh et al., 1995
	Yr34	5AL	*T. monococcum*	Chen et al., 2021
	Yr4a	6B	*T. aestivum*	Bai et al., 2014
	Yr2	7BL	*T. aestivum*	Bai et al., 2014
	Pm2	5DS	*Ae. tauschii*	Wang et al., 2005
Pavon 76	*Lr10*	1A	*T. aestivum*	Singh and Rajaram, 1991
	Yr29	1BL	*T. aestivum*	Cobo et al., 2019
	Lr46	1BL	*T. aestivum*	Singh and Rajaram, 1991
	Yr29	1BL	*T. aestivum*	William et al., 2003
	Yr7	2BL	*T. durum*	Durbin et al., 1989
	Lr13	2BS	*T. aestivum*	Singh and Rajaram, 1991
	Yr30	3BS	*T. aestivum*	Boyd, 2005
	Sr2	3BS	*T. dicoccum*	Mago et al., 2014
	Lr1	5D	*T. aestivum*	Singh and Rajaram, 1991
	Yr6	7BS	*T. aestivum*	Wellings, 1986
Renan	*Pm4b*	2AL	*T. turgidum*	Chantret et al., 1999
	Yr17	2AS	*Ae. ventricosa*	Robert et al., 1999
	Ppd-B1b	2B		Kiseleva et al., 2007
	Pch1	7DL	*Ae. ventricosa*	Burt and Nicholson, 2011
Riband	*Lr17b*	2AS	*T. aestivum*	Pathan and Park, 2006
	Stb15	6AS		Arraiano et al., 2007
	Pm4b	2AL	*T. turgidum*	United Kingdom Cereal Pathogen Virulence Survey [UKCVS], 2004
	Pm6	2BL	*T. timopheevii*	United Kingdom Cereal Pathogen Virulence Survey [UKCVS], 2004
	Pm2	5DS	*Ae. tauschii*	United Kingdom Cereal Pathogen Virulence Survey [UKCVS], 2004

'Maris Huntsman' carry a 3 bp insertion at position 43,405,954 and a 7 bp insertion at position 43,407,045 (**Figure 5**[1]).

Efficacy of Alignment to the Reference Assembly

To further investigate the role of sequence alignment in the regions of reduced sequence coverage, a BLAST search was performed using the mapped and unmapped reads from 'Bacanora' against a database containing both *T. aestivum* and *S. cereale* sequences. Of the 1,959 unmapped reads, 709 (36.2%) hit sequences in the BLAST database: 654 (33.4%) to the *S. cereale* 1R sequence and 55 (2.8%) to the *T. aestivum* 1B sequence. Conversely, for the 1,421 reads that had successfully mapped to the *T. aestivum* 'Chinese Spring' reference sequence, there were only 167 (11.8%) hits to the *S. cereale* 1R sequence while 1,242 (87.4%) hits to the wheat 1B reference sequence.

For unknown introgressions, it is not possible to compare the unmapped reads to the source sequence. To better understand from where these reads came, an assembly of unmapped reads for all 12 accessions was created and then compared with a database of *Poaceae*/*S. cereale* protein sequences (**Figure 6**). The unmapped sequences were predominantly (62.1%) found in the progenitor accessions *Triticum turgidum* (AABB genome), *Ae. tauschii* (DD), and *Triticum urartu* (AA). There were also additional hits

[1]https://plants.ensembl.org/Triticum_aestivum/Location/Compara_Alignments?align=9814--Aegilops_tauschii--5D:46999777-47002088;db=core;g=TraesCS5D02G044600;r=5D:43405783-43407148;t=TraesCS5D02G044600.1

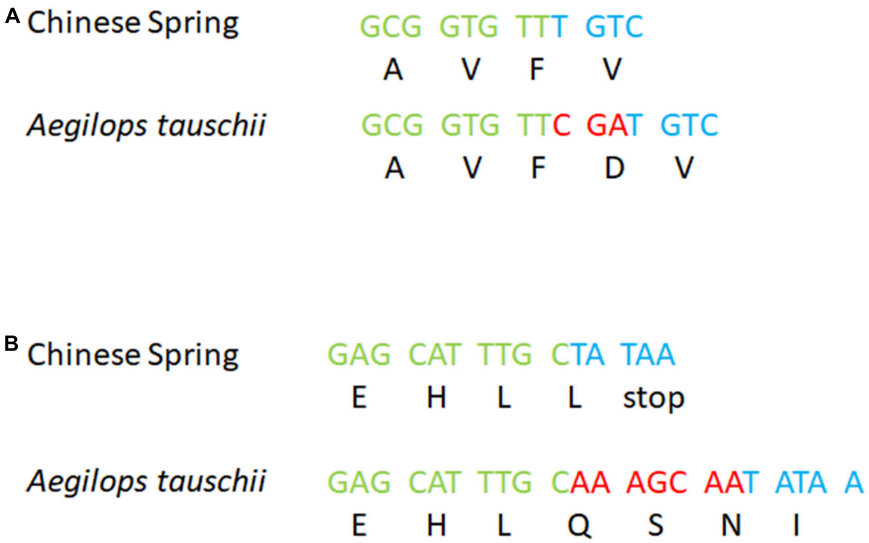


FIGURE 5 | Details of the *Pm2* gene in 'Chinese Spring' and *Ae. tauschii*: **(A)** a 3 bp insertion and **(B)** a 7 bp insertion. Respectively, green and blue bases are 'Chinese Spring' reference sequences before and after the indel. Red bases are the insertion (found in both *Ae. tauschii* and 'Maris Huntsman').

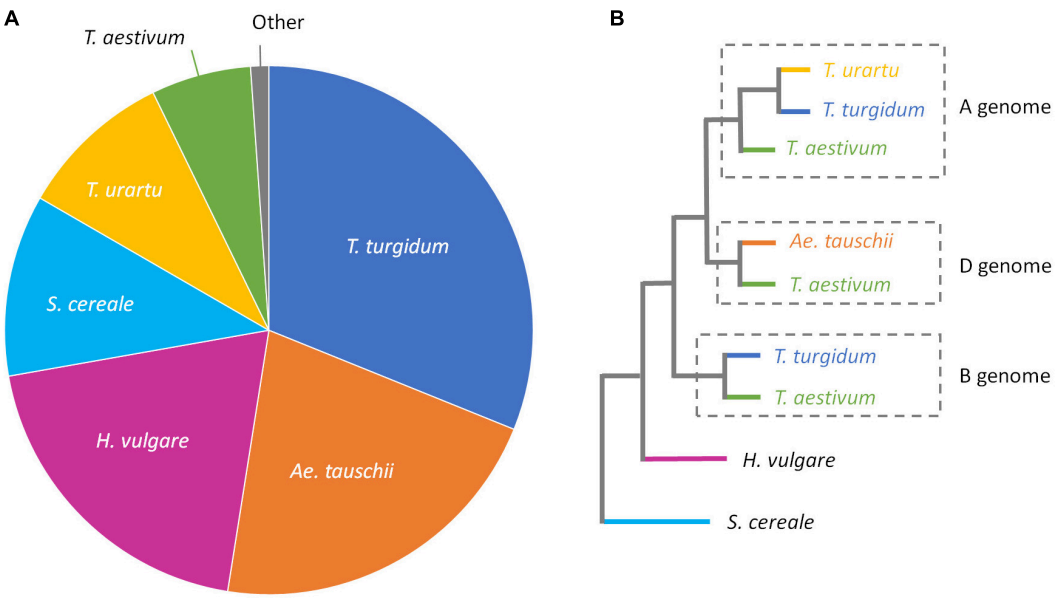


FIGURE 6 | (A) Pie chart showing the best BLAST hits against a combined *Poaceae/S. cereale* database for captured reads that didn't map to the IWGSC 'Chinese Spring' assembly v1. **(B)** Phylogenetic tree (redrawn from Zhou et al., 2017), showing the relationship of the species used in our *Poaceae/S. cereale* database.

to the more distant relatives *Hordeum vulgare* (HH) and *S. cereale* (RR).

DISCUSSION

Exome Capture

The 'Gene Capture v1' and 'Promoter Capture v1' probes are based on sequences not only from *T. aestivum* but also *Ae. tauschii* and *T. turgidum* and, thus, should capture sequence from bread wheat and its progenitors (Gardiner et al., 2019). In this study, the exome capture protocol proved effective at capturing a representative genome sample from each of the 12 accessions examined with sequence coverage in distal regions of chromosomes being greater than that across centromeres (**Figure 1A**); this pattern reflects capture probe distribution which itself is determined by the known telomere to centromere gene gradient (Pingault et al., 2015; Gardiner et al., 2019). Probes appear to have successfully captured wheat sequence and that of introgressions from progenitors as demonstrated by the

capture of sequence from the 5DS, *Ae. tauschii* introgression in 'Maris Huntsman' (**Supplementary Table 1**). A review of the literature reporting primary genepool introgression into bread wheat, further indicated that probes were effectively capturing sequence from these introgressions and, thus, resulting in even sequence coverage across such introgressions and the host sequences flanking them. For example, an introgression from *T. turgidum* subsp. *carthlicum* has been reported on 2AL of 'Renan' and 'Riband' (Chantret et al., 1999; United Kingdom Cereal Pathogen Virulence Survey [UKCVS], 2004); we saw no decrease in sequence coverage for either accession indicating successful capture and alignment. Importantly, this was not just the case for the primary relatives (*T. turgidum* and *Ae. tauschii*) that had been included in the design of the capture probes. An introgression from the primary relative, *Triticum monococcum*, has been reported to be present in 5AL of 'Maris Huntsman' (Chen et al., 2021; **Supplementary File 1**), a *Triticum spelta* introgression has been reported in 2BL of 'Cadenza' (Marchal et al., 2018; **Supplementary File 1**) and introgression from *Triticum dicoccum* has been reported in 3BS of 'Pavon 76' (Mago et al., 2014; **Supplementary File 1**) and none had a corresponding decrease in coverage suggesting adequate capture of these sequences.

The region associated with the *Pm2* gene in 'Maris Huntsman' was used as a case study to confirm that sequence diversity present in regions with successful capture and alignment were, indeed, from a wild relative source. Alignment of the captured sequences from 'Maris Huntsman' to both the 'Chinese Spring' *T. aestivum* reference (IWGSC v1.0) and *Ae. tauschii* (*Ae. tauschii* v4.0 GCF_002575655.1) assemblies showed them to be highly similar. However, two small insertions, with respect to 'Chinese Spring', in 'Maris Huntsman' and *Ae. tauschii* give support to the hypothesis that 'Maris Huntsman' harbors an *Ae. tauschii* introgression (**Figure 5**). The successful capture of this region is hardly surprising considering that *Ae. tauschii* sequence was used to guide capture probe design (Gardiner et al., 2019) and given the high degree of similarity between the two species, *T. aestivum* and *Ae. tauschii*, *across* the *Pm2* region. Indeed, capture probes designed exclusively from bread wheat sequence may well have proved equally efficacious at capturing sequence from this introgressed region.

The design of the probes, then, has allowed the capture of sequences beyond those belonging exclusively to *T. aestivum*. However, one must expect that beyond a certain level of sequence diversity, a reflection of the evolutionary distance of donors of introgressed segments, probes will no longer capture sequence. Such wide introgressions will not be captured, and coverage of the target will drop. This is a serious limitation if novel regions from more distant relatives are the aim of the capture sequencing and other sequencing methods will need to be employed.

Alignment to the Reference Assembly

In addition to successful capture and sequencing, one must be able to realign the sequence to the reference (in this case 'Chinese Spring' IWGSC v1) for it to be identified as present. There is the potential for the mapping parameters to under-utilize the available sequence as the stringency of the parameters used to align the captured sequences to the 'Chinese Spring' reference genome result in some successfully captured sequences being unable to align. Not all variation present in sequencing data is a true reflection of the sequence present and as the alignment stringency is relaxed, sequencing errors may enter the data. To preserve the high-quality sequences, it seems inevitable that diverse sequences will be lost by data processing.

Some mapping protocols, such as the mapping of non-unique hits, can allow for homoeologous sequences to mask gaps in coverage due to deletions or introgressions. In addition, as the mapping of zero in read coverage is not a standard protocol, the gaps seen as a result of diverse sequences are not made apparent (**Supplementary Figure 1**) and the inability to align diverse sequences to the reference is not reported.

Efficacy of Alignment to the Reference Assembly

For all 12 accessions, the captured sequences that could not be mapped to the reference were BLASTed against a *Poaceae*/*S. cereale* protein database (**Figure 6**). Of the sequences that had a hit to the protein database, 62.1% had a match to a sequence derived from a progenitor species (**Figure 6**). This indicates that some sequences were captured and sequenced but had no corresponding sequence in the 'Chinese Spring' reference. Given an alternative reference, some of these sequences may have aligned. The failure of almost 40% of the captured sequences that did not map to the reference probably reflects the limitations of the created *Poaceae*/*S. cereale* protein database since we recognize that there is limited sequence data available for many wheat relatives; the major crop species *T. aestivum*, *T. turgidum*, and *H. vulgare* are well represented in nucleotide databases, but this is not the case for wild relatives. Indeed, we chose to compare our un-mapped sequences to a protein database, rather than a nucleotide database, to maximize the amount of sequence data available. The *Poaceae*/*S. cereale* protein database contained 472,031 sequences. Through this approach, we were able to identify sequences potentially originating from secondary and tertiary genepool species. However, some sequences remained completely unidentified emphasizing that, probably, some diversity is regularly omitted from standard sequencing and alignment. As such, exome capture followed by alignment to a hexaploid reference is not a reliable tool for the identification of introgressions within hexaploid wheat. Where exome capture has been performed and an introgression is suspected, identification is limited by the current availability of wheat relative sequences.

Diverse sequences, such as the *Ae. tauschii* introgression, described in 'Maris Huntsman' were successfully captured, sequenced, and aligned in part due to the presence of *Ae. tauschii* sequences in the capture probe set and in part due to the similarity of the progenitor sequence to the D genome of the reference assembly. For the more distant wild relatives, both capture and alignment were less successful. The reduction in mapped sequences was most pronounced in the accessions containing the 1RS.1BL translocation (**Figure 1**). This is a known introgression that is from a tertiary source. When the 1BS capture

probe sequences (26,985) were BLASTed against 1RS of the rye genome assembly (JADQCU000000000 v1), 31% had a hit (**Figure 4A**), suggesting that some capture would occur, but the percentage similarity between probe sequence and its target was lower in rye than in wheat, suggesting that it might not map back to the reference. This *in silico* assessment was reflected in the captured but un-mapped sequences. By performing a BLAST search against a *T. aestivum* and *S. cereale* database, a number of the unmapped reads in the 1RS containing accession 'Bacanora' were found to have matches to the *S. cereale* sequences (33.4%), considerably higher than the *S. cereale* sequences found within the mapped reads of the same accession (11.8%). This suggests that some of the unmapped reads were from regions of 1RS.1BL that were successfully captured but could not be successfully mapped back to the reference. As *S. cereale* sequences are poorly represented in the BLAST database (there were 25,214 out of 472,031 in total), the full extent of *S. cereale* sequences captured is not known and the ratio present may be higher. While it seems that this tertiary relative introgression was not captured to the same extent as a primary genepool relative, it is important to note that some sequences were captured despite the dissimilarity between *S. cereale* and the *T. aestivum* target but the presence of the sequenced further limited by alignment to the reference.

Each of the 12 accessions used in this study, showed reduced read coverage across some regions of at least one of its chromosomes. Most of these drops in coverage were common to several of the accessions studied and, in many cases, they co-located with documented introgressions or with regions where genotyping data had highlighted extensive variability. The accessions 'Renan' and 'Boregar' had reduced coverage at the end of the short arm of chromosome 2A corresponding to the known introgression from *Ae. ventricosa* associated with rust resistance (*Lr37*, Hanzalová et al., 2007; *Yr17* Dedryver et al., 2009). The size of this introgression has been reported to be c. 33 Mb (Gao et al., 2021) which corresponds with the size of the decline in coverage observed in this study. The eyespot resistance gene, *Pch1* located on the distal end of 7DL, also introduced from *Ae. ventricosa* (Leonard et al., 2007) corresponded to the terminal drop in coverage seen in 'Boregar' and 'Renan', both reported containing the *Pch1* gene (Burt and Nicholson, 2011). The powdery mildew resistance gene *Pm6* from *T. timopheevii* on 2BL was reported in both 'Riband' and 'Maris Huntsman' (United Kingdom Cereal Pathogen Virulence Survey [UKCVS], 1996; Wang et al., 2005) and reveals itself as a distinct decrease in coverage in both accessions. Interestingly, this dip is also found in 'Boregar' which hasn't been reported to carry the 2BL introgression but, on the basis of evidence here, probably does. The presence of unreported introgressions is thought to be quite common. For example, several accessions ('Bacanora', 'Boregar', 'Cadenza', 'KWS Kielder', 'Maris Huntsman', 'Renan', and 'Riband') shared a large region (c. 45 Mb) with reduced read coverage, which we assume might indicate an introgression, but for which we could find no documentary evidence. This region spans over 640 genes with a range of functions, such as ion channel regulation, phosphorylation, and electron transfer (**Supplementary File 4**).

Here we demonstrate that there is a relationship between drops in sequence coverage and sequence similarity of the introgression sequence to the region it replaced. That is, introgressions from primary relatives, such as *Ae. tauschii* or *T. dicoccum* (**Table 2**), are unlikely to fail capture and thus be sequenced and aligned. On the other hand, introgressions from secondary and tertiary genepool species, such as *S. cereale*, *Ae. ventricosa*, and *T. timopheevii*, are likely to avoid capture (**Figure 4**) and, if captured, fail to align to the reference (**Figure 6**); such failures are characterized by reduced sequence coverage across the introgressions. The degree of sequence similarity between a wheat relative sequence and the *T. aestivum* equivalent reflects the evolutionary distance. The observations of this study agree with the study in which human exome capture probes were used to capture exome sequences in non-human primates; "*specificity of the capture decreased as evolutionary divergence from humans increase*" (Jin et al., 2012). Exome capture probes designed for *T. aestivum* efficiently captured genic sequences from the D genome progenitor species, *Ae. tauschii*, but performed much less well against *S. cereale*, an evolutionary more distant species belonging to the tertiary genome.

Modern elite wheat varieties carry numerous introgressions which provide genes of important agronomic traits (**Table 2**), but exome capture may limit the ability to sequence these novel and interesting regions. Introgressions from the primary genepool were successfully captured. Those from more distantly related species, members of the secondary and tertiary genepool, however, were poorly represented in the mapped sequences data (**Table 2**). While there was evidence that some sequences from secondary and tertiary genepool relatives were present amongst the captured sequences (**Figure 6**) their number was small and did not map to the reference. Localized reduction in sequence coverage was observed in all 12 accessions studied, including the landrace accessions. Many of these regions of low coverage were collocated with documented introgressions or deletions, while others remain unknown. The method of sequencing used here has essentially limited the diversity of sequence that could be reported. The careful design of capture probes is critically important as lack of capture probe diversity will lead to failure to capture sequence introgressed from distantly related species. The reference genome used will also strongly bias the sequences that can be aligned and so reported as present.

EXPERIMENTAL PROCEDURES

Sample Preparation and Sequencing

Genomic DNA from 12 wheat accessions (14 days after germination) was extracted, RNase treated, and purified as described in Burridge et al. (2017).

Individual aliquots in a total volume of 55 μl were sheared to an average of 300 bp using an E220 Focused-ultrasonicator (Covaris, Woburn, MA, United States). SeqCap EZ HyperCap Workflow User's Guide (Version 2.0) was used with the following modifications. The starting material was increased to 2 μg

DNA. The A-tailing reaction was changed to 20°C for 30 min, followed by 65°C for 30 min. Size selection of the pre-capture libraries was replaced with a 0.9 bead: sample ratio. The precapture amplification was changed to nine cycles followed by immediate clean-up. COT human DNA was replaced with 1 μl of Developer Reagent Plant Capture Enhancer (NimbleGen) per 100 ng of DNA.

Exome capture was performed using 'Gene Capture v1, 4000026820' and 'Promoter Capture v1, 4000030160' wheat capture probes (Gardiner et al., 2019). Gene and Promoter capture probes were not lyophilized but capture reactions performed separately and products combined after post-capture amplification. For the capture wash, the first Wash Buffer I and both Stringent Wash Buffer steps used buffer preheated to 57°C. Fragment size distribution throughout was determined by TapeStation (Agilent) analysis.

Capture probe enriched sequencing libraries were sequenced at the Bristol Genomics Facility using NextSeq 500 and NextSeq500 2 × 150 bp High−Output v2 kit (Illumina). A final library concentration of 0.8 pM was used with a 5% PhiX control library. The full library preparation and capture method are described in detail in **Supplementary File 3**. All reads are available from the NCBI sequencing read archive using project ID: PRJNA789931.

Data Analysis

Fastq files for each wheat variety were subjected to quality control using FastQC1 (Babraham Bioinformatics, 2020) and were pre-processed using Fastp (Chen et al., 2018) to trim adaptor sequence and for quality filtering. Paired-end reads were aligned to the 'Chinese Spring' reference sequence (IWGSV v1.0) using Burrow-Wheeler Aligner (BWA) (Li and Durbin, 2009) (version 0.7.7-r441), and uniquely mapped reads were identified using sambamba (Tarasov et al., 2015) (version v0.4.4).

Coverage for each chromosome was calculated using samtools (Li et al., 2009) (version 0.1.19-44428cd) using the depth option. Custom perl scripts (available on request) were used to calculate the average depth of coverage for 5 million base pair bins across each chromosome and exome coverage graphs were generated using R (version 3.2.5) (R Core Team, 2013).

Capture probe coverage diagrams were generated with the R package chromPlot using unique location hits and including 0 reads (Verdugo and Oróstica, 2016).

All unmapped reads for the 'Bacanora' were extracted from the bam file using samtools (Version: 1.10-24-g383a31b), along with all reads that mapped to the chromosome 1B IWGSC v1.0 reference from physical mapping positions 1–230,000,000 bp (spanning the putative 1B/1RS introgression. These unmapped and mapped reads were then separately queried against a local BLAST database that contained the wheat 1B sequence and the *S. cereale* 1R sequence, using default BLASTN parameters. The top BLAST hit was then parsed from the BLAST output files using custom perl scripts.

Several gnome assemblies were required for this study: IWGSC v1 Chinese Spring assembly; Rye assembly of the Chinese rye cultivar Weining (Li et al., 2021); *Ae. tauschii* subsp. *strangulata* (Luo et al., 2017).

Exome Capture Probes to 1BS and 5DS

The browser extensible data (BED) file containing the genomic coordinates of the gene capture probes, Wheat_gene_capture_probes.bed, from Gardiner et al. (2019) was downloaded from the Grassroots Data Repository.[2] From this file, the coordinates for the TGAC v1 probes to chromosomes 1BS and 5DS were extracted. Using the python package pysam, the sequences for these probes were extracted from the TGAC version 1 genome assembly of 'Chinese Spring' (Triticum_aestivum.TGACv1.30.dna.genome.fa). The gene capture probe sequences for chromosome 1BS were BLASTed against the chromosome 1BS sequence from the IWGSC v1 assembly and to the chromosome 1RS sequence of the genome assembly of the cultivar 'Weining' rye (JADQCU000000000 v1), an elite Chinese *S. cereale* variety (Li et al., 2021). Likewise, the capture probe sequences for chromosome 5DS were BLASTed against the chromosome 5DS sequence from the IWGSC v1 assembly and to the chromosome 5DS from *Ae. tauschii* subsp. *strangulata* assembly, Aet v4.0 (GCA_002575655.1).

Gene Sequences Surrounding *Pm2* Gene

The putative 5D introgression in 'Maris Huntsman' containing the powdery mildew resistance gene *Pm2*, was used as the point of reference. The *Pm2* gene (TraesCS5D02G044600.1) sequence downloaded from EnsemblPlants is 1,266 bp long and produces a protein of 421 aa. To obtain the *Ae. tauschii* homolog, the 'Chinese Spring' *Pm2* sequence was BLASTed against the NCBI Triticeae database; the top hit, with 99.3% identity (1,255/1,264), was the *Ae. tauschii* subsp. *strangulata* sequence on 5D (sequence id MW538911.1). The full length of this sequence was 4,421 bp.

To compare sequence similarity of 'Chinese Spring' and *Ae. tauschii* coding sequences around the *Pm2* gene, we identified, using the gff3 file for IWGSC v1 (Ensembl Plants genome browser), all the annotated genes within 1 Mb up- and downstream; in 'Chinese Spring', 21 genes were present within this interval (**Supplementary Table 1**). The sequences of these 21 genes were BLASTed against the NCBI Triticeae database to obtain their homologs in *Ae. tauschii*. These were then BLASTED against the *Ae. tauschii* v 4.0 (GCF_002575655.1) assembly to find their positions.

Using BWA, (Li and Durbin, 2009) we aligned the 'Maris Huntsman' captured sequences against both 'Chinese Spring' (IWGSC v1.0) and *Ae. tauschii* (Aet V4.0) assemblies. Both assemblies and the 'Maris Huntsman' BAM files were indexed using Samtools. The gff3 file of the 'Chinese Spring' assembly was also downloaded. An equivalent gff3 file for *Ae. tauschii* was created based on the positions obtained by BLAST and the regions viewed in IGV (Robinson et al., 2011).

The 'Maris Huntsman' captured sequences were aligned, using BWA, to both CS and Aet the sequences around the *Pm2* gene (TraesCS5D02G044600 in 'Chinese Spring' and AET5Gv20114600 in *Ae. tauschii*) in the accession 'Maris Huntsman' to that of the 'Chinese Spring' assembly (IWGSC v1.0) and *Ae. tauschii* v4.0. For both assemblies, using pysam,

[2]https://opendata.earlham.ac.uk/wheat/under_license/toronto/Gardiner_2018-07-04_Wheat-gene-promoter-capture/

pulled out the sequence for the *Pm2* gene (c. 4,420 bp) plus 1 Mb both up and downstream from it. In both assemblies, this region contains 21 genes.

We were interested to see whether the exome captured sequences from 'Maris Huntsman' 5DS had greater similarity to the gene sequences of 'Chinese Spring' or those of *Ae. tauschii*. Because we believed that the putative introgression contained the powdery mildew resistance gene *Pm2*, we used this gene as our point of reference. We began by pulling down the *Pm2* gene (TraesCS5D02G044600.1) sequence from EnsemblPlants; this is 1,266 bp long and produces a protein of 421 aa. To obtain the homolog from *Ae. tauschii*, we BLASTed the 'Chinese Spring' *Pm2* sequence (TraesCS5D02G044600) against the NCBI Triticeae database; the top hit was the homologous gene on 5D of *Ae. tauschii* subsp. *strangulata* (AET5Gv20114600). This sequence was then BLASTed against the NCBI Triticeae database. With 99.3% identity (1,255/1,264), it hit the *Ae. tauschii Pm2* sequence (MW538911.1), which has a full-length functional gene of 4,421 bp.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA789931. The SNP markers and genotypes are available from the CerealsDB website (Wilkinson et al., 2016): https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/array_info.php.

AUTHOR CONTRIBUTIONS

AB prepared the samples and took the lead in writing the manuscript. MW carried out the computational analyses and contributed to the manuscript. PW and GB carried out the computational analyses. KE and GB secured the required funding and supervised the project. All authors helped to interpret the data and contributed to the final manuscript.

FUNDING

This study was funded by the Biotechnology and Biological Sciences Research Council (Grant BBS/E/C/000I0250) as part of the Designing Future Wheat (DFW) program.

ACKNOWLEDGMENTS

Sequencing was performed by the Bristol Genomics Facility. We would like to thank Karim Gharbi, Leah Catchpole, and Thomas Brabbs at the Earlham Institute for assistance and advice regarding the optimization of the exome capture protocol and Jane Coghill and Christy Waterfall at the Bristol Genomics Facility for assistance with library preparation optimization.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.841855/full#supplementary-material

Supplementary Figure 1 | Sequence coverage diagrams for chromosome 2B of the accession 'Riband' using different alignment parameters. **(A)** Average depth of coverage across 5 Mb bins using any good hit to the 2B reference. **(B)** Average depth of coverage across 5 Mb bins using only sequences that give a unique hit to the 2B reference and allowing the display of zero reads.

REFERENCES

Allen, A. M., Winfield, M. O., Burridge, A. J., Downie, R. C., Benbow, H. R., and Barker, G. L. A. (2016). Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid wheat (Triticum aestivum). *Plant Biotechnol. J.* 15, 390–401. doi: 10.1111/pbi.12635

Arbor Biosciences (2021). *myBaits®* - *Hybridization Capture for Targeted NGS* – Manual v.5.01 - Long Insert Protocol. Ann Arbor, MI: Arbor Biosciences.

Arraiano, L. S., Chartrain, L., Bossolini, E., Slatter, H. N., Keller, B., and Brown, J. K. M. (2007). A gene in European wheat cultivars for resistance to an African isolate of Mycosphaerella graminicola. *Plant Pathol.* 56, 73–78.

Babraham Bioinformatics (2020). *FastQC* Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed June 9, 2020).

Bai, B., Du, J. Y., Lu, Q. L., He, C. Y., Zhang, L. J., Zhou, G., et al. (2014). Effective Resistance to Wheat Stripe Rust in a Region with High Disease Pressure. *Plant Dis.* 98, 891–897. doi: 10.1094/PDIS-09-13-0909-RE

Boyd, L. (2005). Can Robigus defeat an old enemy? – Yellow rust of wheat. *J. Agricult. Sci.* 143, 233–243. doi: 10.1017/S0021859605005095

Burridge, A. J., Winfield, M. O., Allen, A. M., Wilkinson, P. A., Barker, G. L., and Coghill, J. (2017). "High-Density SNP Genotyping Array for Hexaploid Wheat and Its Relatives," in *Wheat Biotechnology: Methods and Protocols*, eds P. Bhalla and M. Singh (Totowa, NJ: Humana Press), 293–306. doi: 10.1007/978-1-4939-7337-8_19

Burt, C., and Nicholson, P. (2011). Exploiting co-linearity among grass species to map the Aegilops ventricosa-derived Pch1 eyespot resistance in wheat and establish its relationship to Pch2. *Theoret. Appl. Genet.* 123, 1387–1400. doi: 10.1007/s00122-011-1674-9

Chantret, N., Pavoine, M. T., and Doussinault, G. (1999). The race specific resistance gene to powdery mildew, MIRE, has a residual effect on adult plant resistance of winter wheat line RE714. *Phytopathology* 89, 533–539. doi: 10.1094/PHYTO.1999.89.7.533

Chen, S., Hegarty, J., Shen, T., Hua, L., Li, H., Luo, J., et al. (2021). Stripe rust resistance gene Yr34 (synonym Yr48) is located within a distal translocation of Triticum monococcum chromosome 5AmL into common wheat. *Theoret. Appl. Genet.* 134, 2197–2211. doi: 10.1007/s00122-021-03816-z

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Cobo, N., Wanjugi, H., Lagudah, E., and Dubcovsky, J. (2019). A High-Resolution Map of Wheat QYr.ucw-1BL, an Adult Plant Stripe Rust Resistance Locus in the Same Chromosomal Region as Yr29. *Plant Genome* 12:180055. doi: 10.3835/plantgenome2018.08.0055

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S. B., Shendure, J., and Luikart, G. (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* 12:347. doi: 10.1186/1471-2164-12-347

Cseh, A., Yang, C., Hubbart-Edwards, S., Scholefield, D., Ashling, S. S., and Burridge, A. J. (2019). Development and validation of an exome-based SNP

marker set for identification of the St Jr and Jvs genomes of Thinopyrym intermedium in a wheat background. *Theoret. Appl. Genet.* 132, 1555–1570. doi: 10.1007/s00122-019-03300-9

Dedryver, F., Paillard, S., Mallard, S., Robert, O., Trottet, M., Nègre, S., et al. (2009). Characterization of genetic components involved in durable resistance to stripe rust in the bread wheat 'Renan'. *Phytopathology* 99, 968–973. doi: 10.1094/PHYTO-99-8-0968

Devi, U., Grewal, S., Yang, Cy, Hubbart-Edwards, S., Scholefield, D., and Ashling, S. S. (2019). Development and characterisation of interspecific hybrid lines with genome-wide introgressions from Triticum timopheevii in a hexaploid wheat background. *BMC Plant Biol.* 19:183. doi: 10.1186/s12870-019-1785-z

Driever, S. M., Lawson, T., Andralojc, P. J., Raines, C. A., and Parry, M. A. J. (2014). Natural variation in photosynthetic capacity, growth, and yield in 64 field-grown wheat genotypes. *J. Exp. Bot.* 65, 4959–4973. doi: 10.1093/jxb/eru253

Durbin, H. J., Johnson, R., and Stubbs, R. W. (1989). Postulated genes to stripe rust in selected CIMMYT and related wheats. *Plant Dis.* 73, 472–475. doi: 10.1094/pd-73-0472

Gao, L., Koo, D. H., Juliana, P., Rife, T., Singh, D., Lemes, et al. (2021). The Aegilops ventricosa 2NvS segment in bread wheat: cytology, genomics and breeding. *Theoret. Appl. Genet.* 134, 529–542. doi: 10.1007/s00122-020-03712-y

Gardiner, L.-J., Brabbs, T., Akhunov, A., Jordan, K., Budak, H., Richmond, T., et al. (2019). Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. *GigaScience* 8:4. doi: 10.1093/gigascience/giz018

Hanzalová, A., Dumalasová, V., Sumíkova, T., and Bartoš, P. (2007). Rust resistance of the French wheat cultivar Renan. *Czech J. Genet. Plant Breed.* 43, 53–60. doi: 10.17221/1912-CJGPB

Hao, M., Zhang, L., Ning, S., Huang, L., Yuan, Z., Wu, B., et al. (2020). The Resurgence of Introgression Breeding, as Exemplified in Wheat Improvement. *Front. Plant Sci.* 11:252. doi: 10.3389/fpls.2020.00252

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., et al. (2019). Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* 51, 896–904. doi: 10.1038/s41588-019-0382-2

Henry, I. M., Nagalakshmi, U., Lieberman, M. C., Ngo, K. J., Krasileva, K. V., and Vasquez-Gross, H. (2014). Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing. *Plant Cell* 26, 1382–1397. doi: 10.1105/tpc.113.121590

International Wheat Genome Sequencing Consortium [IWGSC] (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 18:345. doi: 10.1126/science.1251788

International Wheat Genome Sequencing Consortium [IWGSC] (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 17:361. doi: 101126/science.aar7191

Jin, X., He, M., Ferguson, B., Meng, Y., Ouyang, L., Ren, J., et al. (2012). An effort to use human-based exome capture methods to analyze chimpanzee and macaque exomes. *PLoS One* 7:e40637. doi: 10.1371/journal.pone.0040637

Kanyuka, K., Lovell, D. J., Mitrofanova, O. P., Hammond-Kosack, K., and Adams, M. J. (2004). A controlled environment test for resistance to Soil-borne cereal mosaic virus (SBCMV) and its use to determine the mode of inheritance of resistance in wheat cv. Cadenza and for screening Triticum monococcum botypes for sources of SBCMV resistance. *Plant Pathol.* 53, 154–160. doi: 10.1111/j.0032-0862.2004.01000.x

Kaur, P., and Gaikwad, K. (2017). From genomes to GENE-omes: exome sequencing concept and applications in crop improvement. *Front. Plant Sci.* 8:2164. doi: 10.3389/fpls.2017.02164

King, J., Newell, C., Grewal, S., Hubbart-Edwards, S., Yang, C.-Y., Scholefield, D., et al. (2019). Development of Stable Homozygous Wheat/Amblyopyrum muticum (Aegilops mutica) Introgression Lines and Their Cytogenetic and Molecular Characterization. *Front. Plant Sci.* 10:34. doi: 10.3389/fpls.2019.00034

Kiseleva, A. A., Potokina, E. K., and Salina, E. A. (2007). Features of Ppd-B1 expression regulation and their impact on the flowering time of wheat near-isogenic lines. *BMC Plant Biol.* 17:172. doi: 10.1186/s12870-017-1126-z

Leonard, J. M., Watson, C. J. W., Carter, A. H., Hansen, J. L., Zemetra, R. S., and Santra, D. K. (2007). Identification of a candidate gene for the wheat endopeptidase Ep-D1 locus and two other STS markers linked to the eyespot resistance gene Pch1. *Theoret. Appl. Genet.* 116, 261–270. doi: 10.1007/s00122-007-0664-4

Li, G., Wang, L., Yang, J., He, H., Jin, H., Li, X., et al. (2021). A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. *Nat. Genet.* 53, 574–584. doi: 10.1038/s41588-021-00808-z

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Luo, M. C., Gu, Y., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., et al. (2017). Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature* 551, 498–502. doi: 10.1038/nature24486

Ma, J., Wingen, L. U., Orford, S., Fenwick, P., Wang, J. K., and Griffiths, S. (2015). Using the UK reference population Avalon x Cadenza as a platform to compare breeding strategies in elite Western European bread wheat. *Mol. Breed.* 35:70. doi: 10.1007/s11032-015-0268-7

Mago, R., Tabe, L., Vautrin, S., Šimková, H., Kubaláková, M., Upadhyaya, N., et al. (2014). Major haplotype divergence including multiple germin-like protein genes at the wheat Sr2 adult plant stem rust resistance locus. *BMC Plant Biol.* 14:379. doi: 10.1186/s12870-014-0379-z

Marchal, C., Zhang, J., Zhang, P., Fenwick, P., Steuernagel, B., and Adamski, N. M. (2018). BED-domain-containing immune receptors confer diverse resistance spectra to yellow rust. *Nat. Plants* 4, 662–668. doi: 10.1038/s41477-018-0236-4

McIntosh, R. A., Wellings, C. R., and Park, R. F. (1995). *Wheat Rusts: An Atlas of Resistance Genes*. Dordrecht: Kluwer Academic Publishers.

Pathan, A. K., and Park, R. F. (2006). Evaluation of seedling and adult plant resistance to leaf rust in European wheat cultivars. *Euphytica* 149, 327–342. doi: 10.1007/s10681-005-9081-4

Pingault, L., Choulet, F., Alberti, A., Glover, N., Wincker, P., Feuillet, C., et al. (2015). Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biol.* 16:29. doi: 10.1186/s13059-015-0601-9

Przewieslik-Allen, A. M., Burridge, A. J., Wilkinson, P. A., Winfield, M. O., and Shaw, D. S. (2019). Developing a High-Throughput SNP-Based Marker System to Facilitate the Introgression of Traits from Aegilops Species into Bread Wheat (Triticum aestivum). *Front. Plant Sci.* 9:1993. doi: 10.3389/fpls.2018.01993

Przewieslik-Allen, A. M., Wilkinson, P. A., Burridge, A., Winfield, M., Dai, X., and Beaumont, M. (2021). The role of gene flow and chromosomal instability in shaping the bread wheat genome. *Nat. Plants* 7, 172–183. doi: 10.1038/s41477-020-00845-2

R Core Team (2013). *R: A language and environment for statistical computing R Foundation for Statistical Computing Vienna Austria*. Austria: R Core Team.

Rabinovich, S. V. (1998). Importance of wheat-rye translocations for breeding modern cultivars of Triticum aestivum L. *Euphytica* 100, 323–340.

Rasheed, A., and Xia, X. (2019). From markers to genome-based breeding in wheat. *Theoret. Appl. Genet.* 132, 767–784. doi: 10.1007/s00122-019-03286-4

Robert, O., Abelard, C., and Dedryver, F. (1999). Identification of molecular markers for the detection of the yellow rust resistance gene Yr17 in wheat. *Mol. Breed.* 5, 167–175. doi: 10.1023/A:1009672021411

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., and Getz, G. (2011). Intergrative Genomics Viewer. *Nat. Biotechnol.* 29, 24–26.

Saintenac, C., Jiang, D., and Akhunov, E. D. (2011). Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* 12:R88. doi: 10.1186/gb-2011-12-9-r88

Salmon, A., Udall, J. A., Jeddeloh, J. A., and Wendel, J. (2012). Targeted capture of homoeologous coding and noncoding sequence in polyploid cotton. *G3* 2, 921–930. doi: 10.1534/g3.112.003392

Schneider, A., Molnár, I., and Molnár-Láng, M. (2008). Utilisation of Aegilops (goatgrass) species to widen the genetic diversity of cultivated wheat. *Euphytica* 163, 1–19. doi: 10.1007/s10681-007-9624-y

Singh, R. P., and Rajaram, S. (1991). Resistance to Puccinia recondita f.sp. tritici in 50 Mexican bread wheat cultivars. *Crop Sci.* 31, 1472–1479. doi: 10.2135/cropsci1991.0011183x003100060016x

Singh, S., Vikram, P., Sehgal, D., Burgueño, J., Sharma, A., Singh, S. K., et al. (2018). Harnessing genetic potential of wheat germplasm banks through

impact-oriented-prebreeding for future food and nutritional security. *Scient. Rep.* 21:12527. doi: 10.1038/s41598-018-30667-4

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31, 2032–2034. doi: 10.1093/bioinformatics/btv098

United Kingdom Cereal Pathogen Virulence Survey [UKCVS] (1996). *Annual Report. Accessed from AHDB archives: ahdb.org.uk/ukcpvs*. Telford, UK: UKCVS.

United Kingdom Cereal Pathogen Virulence Survey [UKCVS] (2004). *Annual Report. Accessed from AHDB archives*. Telford, UK: UKCVS.

Verdugo, R. A., and Oróstica, K. Y. (2016). chromPlot: Global visualization tool of genomic data R package version 1160. *Bioinformatics* 32, 2366–2368. doi: 10.1093/bioinformatics/btw137

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genome diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Wang, Z. L., Li, L. H., He, Z. H., Duan, X. Y., Zhou, Y. L., Chen, X. M., et al. (2005). Seedling and adult plant resistance to powdery mildew in Chinese bread wheat cultivars and lines. *Plant Dis.* 89, 457–463. doi: 10.1094/PD-89-0457

Warburton, M., Skovmand, B., and Mujeeb-Kazi, A. (2002). The molecular genetic characterization of the 'Bobwhite' bread wheat family using AFLPs and the effect of the T1BL.1RS translocation. *Theoret. Appl. Genet.* 104, 868–873. doi: 10.1007/s00122-001-0816-x

Wellings, C. R. (1986). *Host: Pathogen Studies of Wheat Stripe Rust in Australia*. Ph.D thesis, Australia: University of Sydney.

Wilkinson, P. A., Winfield, M. O., Barker, G. L. A., Tyrell, S., Bian, X., Allen, A. M., et al. (2016). CerealsDB 3.0: expansion of resources and data integration. *BMC Bioinform.* 17:256. doi: 10.1186/s12859-0161139-x

William, M., Singh, R. P., Huerta-Espino, J., Islas, S. O., and Hoisington, D. (2003). Molecular marker mapping of leaf rust resistance gene Lr46 and its association with stripe rust resistance gene Yr29 in wheat. *Phytopathology* 93, 153–159. doi: 10.1094/PHYTO.2003.93.2.153

Winfield, M. O., Allen, A. M., Burridge, A. J., Barker, G. L. A., Benbow, H. R., and Wilkinson, P. A. (2015). High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* 14, 1195–1206. doi: 10.1111/pbi.12485

Winfield, M. O., Wilkinson, P. A., Allen, A. M., Barker, G. L. A., Coghill, J. A., and Burridge, A. (2012). Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol. J.* 10, 733–742. doi: 10.1111/j.1467-7652.2012.00713.x

Worland, A. J., Korzun, V., Roder, M. S., Ganal, M. W., and Law, C. N. (1998). Genetic analysis of the dwarfing gene Rht8 in wheat. Part II. The distribution and adaptive significance of allelic variants at the Rht8 locus of wheat as revealed by microsatellite screening. *Theoret. Appl. Genet.* 96, 1110–1120. doi: 10.1007/s001220050846

Xu, J., Wang, L., Deal, K. R., Zhu, T., Ramasamy, R. K., Luo, Mc, et al. (2020). Genome-wide introgression from a bread wheat×Lophopyrum elongatum amphiploid into wheat. *Theoret. Appl. Genet.* 133, 1227–1241. doi: 10.1007/s00122-020-03544-w

Zhang, J., Liu, W., Lu, Y., Liu, Q., Yang, X., Li, X., et al. (2017). A resource of large-scale molecular markers for monitoring Agropyron cristatum chromatin introgression in wheat background based on transcriptome sequences. *Scient. Rep.* 7:11942. doi: 10.1038/s41598-017-12219-4

Zhang, W., Cao, Y., Zhang, M., Zhu, X., Ren, S., Long, Y., et al. (2017). Meiotic Homoeologous Recombination-Based Alien Gene Introgression in the Genomics Era of Wheat. *Crop Sci.* 57, 1189–1198. doi: 10.2135/cropsci2016.09.0819

Zhou, S., Yan, B., Li, F., Zhang, J., Zhang, J., Ma, H., et al. (2017). RNA-Seq Analysis Provides the First Insights into the Phylogenetic Relationship and Interspecific Variation between Agropyron cristatum and Wheat. *Front. Plant Sci.* 8:1644. doi: 10.3389/fpls.2017.01644

Zikhali, M., Wingen, L. U., and Griffiths, S. (2016). Delimitation of the Earliness per se D1 (Eps-D1) flowering gene to a subtelomeric chromosomal deletion in bread wheat (Triticum aestivum). *J. Exp. Bot.* 67, 1287–1299. doi: 10.1093/jxb/erv458

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fpls.2022.890980

A High-Quality Haplotype-Resolved Genome of Common Bermudagrass (*Cynodon dactylon* L.) Provides Insights Into Polyploid Genome Stability and Prostrate Growth

*Bing Zhang[1,2], Si Chen[2], Jianxiu Liu[3], Yong-Bin Yan[1], Jingbo Chen[3], Dandan Li[3] and Jin-Yuan Liu[1]**

[1]School of Life Sciences, Tsinghua University, Beijing, China, [2]College of Animal Science and Technology, Yangzhou University, Yangzhou, China, [3]Institute of Botany, Jiangsu Province and Chinese Academy of Sciences, Nanjing, China

OPEN ACCESS

Edited by:
Surya Saha, Boyce Thompson Institute (BTI), United States

Reviewed by:
Liangsheng Zhang, Zhejiang University, China
Kai Wang, Nantong University, China
Kevin Andrew Bird, Michigan State University, United States

***Correspondence:**
Jin-Yuan Liu
liujy@mail.tsinghua.edu.cn

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *07 March 2022*
Accepted: *04 April 2022*
Published: *25 April 2022*

Citation:
Zhang B, Chen S, Liu J, Yan Y-B, Chen J, Li D and Liu J-Y (2022) A High-Quality Haplotype-Resolved Genome of Common Bermudagrass (Cynodon dactylon L.) Provides Insights Into Polyploid Genome Stability and Prostrate Growth. Front. Plant Sci. 13:890980. doi: 10.3389/fpls.2022.890980

Common bermudagrass (*Cynodon dactylon* L.) is an important perennial warm-season turfgrass species with great economic value. However, the reference genome is still deficient in *C. dactylon*, which severely impedes basic studies and breeding studies. In this study, a high-quality haplotype-resolved genome of *C. dactylon* cultivar Yangjiang was successfully assembled using a combination of multiple sequencing strategies. The assembled genome is approximately 1.01 Gb in size and is comprised of 36 pseudo chromosomes belonging to four haplotypes. In total, 76,879 protein-coding genes and 529,092 repeat sequences were annotated in the assembled genome. Evolution analysis indicated that *C. dactylon* underwent two rounds of whole-genome duplication events, whereas syntenic and transcriptome analysis revealed that global subgenome dominance was absent among the four haplotypes. Genome-wide gene family analyses further indicated that homologous recombination-regulating genes and tiller-angle-regulating genes all showed an adaptive evolution in *C. dactylon*, providing insights into genome-scale regulation of polyploid genome stability and prostrate growth. These results not only facilitate a better understanding of the complex genome composition and unique plant architectural characteristics of common bermudagrass, but also offer a valuable resource for comparative genome analyses of turfgrasses and other plant species.

Keywords: ***Cynodon dactylon*, common bermudagrass, genome, haplotype, tiller angle**

INTRODUCTION

Common bermudagrass (*Cynodon dactylon* L., 2n = 4x = 36) is an important warm-season turfgrass species and is widely used to produce beautiful and uniform turf for public parks, home lawns, golf courses, and sport fields in warm regions around the world (Yang et al., 2018; Zhang et al., 2018a). In some regions, *C. dactylon* is also used as forage, medicinal, and energy plants (Hill et al., 2001; Nagori and Solanki, 2011; Xu et al., 2011). Since its origination

from Africa or Indo-Malaysian, *C. dactylon* was spread to tropical and subtropical areas worldwidely (Kneebone, 1966; Harlan and de Wet, 1969). As a cross-pollinating plant, wild germplasms of *C. dactylon* collected at different regions usually show enormous genetical and morphological variations (Wu et al., 2004, 2007; Farsani et al., 2012; Tan et al., 2014; Zheng et al., 2017). Karyotype and molecular marker analyses not only revealed that polyploidy and aneuploidy events exist in *C. dactylon* but also pointed out the genome of *C. dactylon* is highly heterozygous (Wu et al., 2006; Chaves et al., 2019; Grossman et al., 2021). These characteristics make *C. dactylon* an interesting plant material to investigate genome stability, variability, and evolution (Khanal et al., 2017).

Unlike domesticated cereal grasses including rice, wheat, maize, and sorghum, *C. dactylon* has typical plant architectural characteristics of wild grasses that its stems are differentiated into shoots, stolons, and rhizomes (Dong and de Kroon, 1994; Zhang et al., 2019; Ma et al., 2021). Shoots grow erectly and are widely seen in other plants, whereas stolons and rhizomes are two types of prostrate stems that grow aboveground and underground, respectively (Guo et al., 2021). Through regeneration of new seedlings at stolon nodes, *C. dactylon* plants are asexually reproduced in a colonial growth mode (Zhang and Liu, 2018). The high efficiency to build turf using commercial *C. dactylon* cultivars is mainly derived from this virtue. During cold days in winter, the aboveground parts of common bermudagrass plants withered and died, whereas the underground rhizomes remain alive and new plants will regenerate from rhizome nodes at warm days next year (Satorre et al., 1996). By repeating the cycle of growth at aboveground and dormancy at underground, *C. dactylon* maintains a perennial life style, which also contribute to its usage as an eminent turfgrass. Development of asexual reproductive and perennial versions of important grain crops is an attractive measure to sustainably meet the increasing global food demand (Glover et al., 2010; Ozias-Akins and Conner, 2020). Elucidating the mechanism how *C. dactylon* possesses its unique plant architectural characteristics could simultaneously provide new insights into turf breeding and crop improvement.

In this study, we reported a haplotype-resolved assembly of the highly heterozygous *C. dactylon* genome through the combined application of Pacific Biosciences (PacBio) single-molecule sequencing, Illumina paired-end sequencing, Bionano optical mapping, and chromosome conformation capture (Hi-C) technologies. With the assembled genome dataset and annotation information, we further analyzed the subgenomic composition and adaptive evolution of *C. dactylon*. Results of this study not only expand our understanding of genome structure and plant architectural regulation in *C. dactylon*, but also provide a valuable resource for genetic studies and breeding of turfgrasses.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

Cynodon dactylon cultivar Yangjiang was used for genome sequencing and assembly in this study. The bermudagrass turf were grown in turfgrass plots of Yangzhou University (longitude and latitude: 32°35'N, 119°40′E; average annual temperatures: 22.4°C; average annual precipitation: 1, 106 mm; annual average sunshine hours: 1, 960 h; soil type: 80% river sand; and 20% peat soil) under routine management conditions (irrigation: keep the soil moist as required; fertilization: four times/year; and mowing: one times/month) for 3 years. Healthy leaves were randomly collected from the turf plots. Half of the leaf samples were frozen and used for *de novo* sequencing, whereas another half of fresh leaf samples were used for Bionano and Hi-C sequencing. *Oryza sativa* subspecies *indica* cultivar 93–11 was grown in growth chamber at 24°C under 16 h/8 h light/dark conditions.

Flow Cytometry Estimation of Genome Size

The genome size of *C. dactylon* cultivar Yangjiang was estimated using flow cytometry as previously described (Zhang et al., 2020). Specifically, *O. sativa* cv. 93–11 with a genome size of 430 Mb was used as an internal standard. Young leaves of *C. dactylon* and *O. sativa* were homogenized on ice in Galbraith's buffer (45 mM $MgCl_2$, 30 mM sodium citrate, 20 mM MOPS, and 0.1% (v/v) Triton X-100, pH 7.0) with 50 μg mL^{-1} propidium iodide. After filtration with 40 μm nylon cell strainer (BD Biosciences, Franklin Lakes, United States), samples were analyzed on a FACSCanto™ II flow cytometer (BD Biosciences). The flow cytometry data were analyzed using BD Spectrum Viewer.

Illumina Sequencing and K-mer Analysis

Genomic DNA was isolated from the frozen leave samples using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). The DNA quality and concentration were tested by 1% agarose gel electrophoresis and Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, United States). Two paired-end libraries with short insert size of 270 bp and 500 bp were constructed using the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, Ipswich, United States) and sequenced on the Illumina HiSeq X Ten platform (Illumina, San Diego, United States). The raw Illumina sequencing reads were processed with SOAPnuke v2.1.6[1] to remove adapters and low-quality reads (Chen et al., 2018). The obtained 161.1 Gb high-quality sequencing reads were used to generate a k-mer depth distribution curve adopting the Jellyfish v2.3.0.[2] The obtained peak k-mer number ($k = 27$) and corresponding peak depth were calculated by GenomeScope v2.0[3] to estimate the genome size and heterozygosity (Marçais and Kingsford, 2011).

PacBio Sequencing and Preliminary Genome Assembly

High-molecular weight (HMW) DNA fragments were separated from the extracted genomic DNA samples using BluePippin

[1]https://github.com/BGI-flexlab/SOAPnuke
[2]https://github.com/gmarcais/Jellyfish
[3]http://qb.cshl.edu/genomescope

Size Selection System (Sage Science, Beverly, United States) through pulse-field gel electrophoresis and eight 20-kb sequencing libraries were constructed using SMRTbell Template Prep Kit (Pacific Biosciences, Menlo Park, United States) following the manufacturer's instructions. The libraries (16 SMRT cells) were sequenced on the PacBio RSII platform (Pacific Biosciences). Contig sequences were assembled from the 151.99 Gb PacBio sequencing reads using Hifiasm v0.12[4] and polished by Racon v1.4.3[5] (Cheng et al., 2021). The Illumina sequencing reads were aligned to the assembled contigs using Bwa-mem v2.2.1[6] and the draft assembly was corrected by the aligned short sequences using Pilon v1.24[7] (Walker et al., 2014).

Bionano Optical Genome Mapping

HMW DNA was extracted from the agarose-embedded cell nuclei fractions, which were isolated from fresh leaf samples, using the Bionano Prep™ Plant DNA Isolation Kit (Bionano Genomics, San Diego, United States) following the manufacturer's instructions. The DNA was digested by the single-stranded nicking endonuclease Nt.BspQI, fluorescently labeled, loaded into a Saphyr Chip®, and imaged on a Saphyr Optical Genome Mapping Instrument (Bionano Genomics). The 395.4 Gb image data were filtered using a molecule length cutoff of 100 kb and a label number cutoff of 6, and assembled to 954 genome maps. To assist genome assembly, contigs obtained from the above-mentioned PacBio sequencing were transformed into *in silico Bsp*QI-digested reference genome maps and compared with the optical genome maps. The aligned and merged genome maps were further transformed into scaffold sequences using the Bionano Solve™ v3.6.1.[8]

Hi-C Sequencing and Pseudochromosome Construction

Fresh leaf samples were fixed in 1% formaldehyde to maintain the 3-D structure of genome. Genomic DNA was extracted and digested with restriction endonuclease MboI. The digested DNA fragments were biotin-labeled at the ends and ligated to each other randomly. The ligated DNA was sheared into 300–600 bp fragments, blunt-end repaired, and purified using streptavidin pull-down. The purified DNA was also sequenced on the Illumina HiSeq X Ten platform, which yielded 231.38 Gb of data with 771 million paired-end reads. The paired-end reads were mapped to the assembled scaffold sequences using Juicer v1.6[9] to discriminate valid and invalid interaction pairs (Durand et al., 2016). The obtained 185 million valid interaction pairs (55.5 Gb data) were further used to adjust the relative locations of the scaffolds and cluster the scaffolds into pseudochromosomes using 3D-DNA[10] (Kronenberg et al., 2021).

[4]https://github.com/chhylp123/hifiasm
[5]https://github.com/isovic/racon
[6]https://github.com/bwa-mem2/bwa-mem2
[7]https://github.com/broadinstitute/pilon
[8]https://bionanogenomics.com/support/software-downloads
[9]https://github.com/aidenlab/juicer
[10]https://github.com/aidenlab/3d-dna

Annotation and Analysis of Repetitive Sequences

Repetitive sequences were annotated by combining the homology alignment and *de novo* prediction approaches (Zhang et al., 2021). For the homology alignment approach, the assembled genome sequence was blast searched against the RepBase repeat sequence collection[11] using RepeatMasker v4.0.9[12] (Tempel, 2012). For the *de novo* prediction approach, five softwares, including RepeatModeler,[13] PILER,[14] RepeatScout,[15] LTR_Finder,[16] and Tandem Repeats Finder,[17] were used to find the possible repeat sequences (Price et al., 2005; Flynn et al., 2020). The identified repetitive sequences were manually checked and classified according to the nomenclature system of transposons. The insertion time of different families of long-terminal repeat retrotransposons (LTR-RTs) were calculated using the formula $T = k/2r$, where k is the divergence distance between the 5′ LTR and 3′ LTR of intact LTR-RTs and r is the base substitution rate (1.38×10^{-8} substitutions/site/year for grasses; Ma et al., 2004). The LTR Assembly Index (LAI) scores of assembled pseudo chromosomes and whole genome were calculated using LTR_retriever v2.9.0[18] with default parameters (Ou et al., 2018). Putative centromeric repeat arrays were specifically identified using Tandem Repeats Finder with searching parameters "1 1 2 80 5 200 2000 -d –h" as previously described (VanBuren et al., 2020). The identified centromeric repeat array sequences were used to construct a maximum likelihood phylogenetic tree using MEGA v10.0.5 with a bootstrap of 1,000.

Prediction and Annotation of Protein-Coding Genes

Protein-coding genes were identified by combining the homology alignment prediction, *ab initio* prediction, and transcriptome-assisted prediction approaches (Zhang et al., 2021). For the homology alignment approach, protein sequences of *Arabidopsis thaliana* and five grass species, including *O. sativa*, *Brachypodium distachyon*, *Zea mays*, *Sorghum bicolor*, and *Oropetium thomaeum*, were downloaded from the Phytozome database[19] and blast searched against the assembled genome sequence to identify the homologous proteins, which were then aligned to the genome by GeneWise[20] to annotate gene structures (Birney et al., 2004). *Ab initio* gene prediction was conducted using five softwares, including Augustus v3.4.0,[21] geneid v1.4.4,[22]

[11]https://www.girinst.org/server/RepBase
[12]http://www.repeatmasker.org
[13]http://www.repeatmasker.org/RepeatModeler
[14]http://www.drive5.com/piler
[15]http://bix.ucsd.edu/repeatscout
[16]http://tlife.fudan.edu.cn/tlife/ltr_finder
[17]https://tandem.bu.edu/trf/trf.html
[18]https://github.com/oushujun/LTR_retriever
[19]https://phytozome-next.jgi.doe.gov
[20]https://www.ebi.ac.uk/Tools/psa/genewise
[21]https://github.com/Gaius-Augustus/Augustus
[22]https://genome.crg.es/software/geneid/

FgeneSH,[23] GlimmerHMM v3.0.4,[24] and Genscan[25] with default parameters (Yao et al., 2005; Nachtweide and Stanke, 2019). For transcriptome-assisted prediction, the PacBio single-molecule transcriptome sequencing data of mixed organ samples (Zhang et al., 2018a) were aligned to the assembled genome using GMAP[26] and the gene structures were modeled using PASA,[27] whereas Illumina transcriptome sequencing data of six different organs (Chen et al., 2021) were aligned to the genome using TopHat v2.1.1[28] and the gene structures were modeled using Cufflinks v2.2.1[29] (Wu and Watanabe, 2005; Ghosh and Chan, 2016). A non-redundant reference gene set was generated by merging the predicted genes using EVidenceModeler v1.1.1[30] (Haas et al., 2008). Functional annotations of the reference gene set were obtained through orthology assignment of the eggNOG v5.0 database[31] using eggNOG-mapper v2[32] (Cantalapiedra et al., 2021). Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation of the reference gene set were obtained through BLAST searching against the GO database[33] and the KEGG pathway database,[34] respectively. KEGG enrichment analysis were performed using KEGG-Orthology Based Annotation System (KOBAS; Xie et al., 2011).[35] Transcription factors (TFs) were annotated using iTAK[36] incorporated with PlantTFDB database[37] (Zheng et al., 2016).

Prediction of Non-coding RNA Genes

rRNA and tRNA genes were predicted using the programs Barrnap[38] and tRNAscan-SE-2.0,[39] respectively (Chan et al., 2021). miRNA, snoRNA, and snRNA genes were all identified by searching against the Rfam database *via* Infernal v1.1.4[40] with default parameters (Nawrocki and Eddy, 2013).

BUSCO Assessment

The completeness and accuracy of the assembled genome and predicted reference gene set were both assessed using the embryophyta_odb10 core gene collect (1,375 genes) of the Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.2.2 database[41] (Simão et al., 2015). The number of single-copy and duplicated genes with complete coverage, genes with fragment coverage, and missing genes were all counted.

[23]http://www.softberry.com
[24]http://ccb.jhu.edu/software/glimmerhmm
[25]http://argonaute.mit.edu/GENSCAN.html
[26]http://research-pub.gene.com/gmap
[27]https://anaconda.org/bioconda/pasa
[28]http://ccb.jhu.edu/software/tophat
[29]http://cole-trapnell-lab.github.io/cufflinks
[30]http://evidencemodeler.github.io
[31]http://eggnog5.embl.de
[32]http://eggnog-mapper.embl.de
[33]http://www.geneontology.org
[34]http://www.genome.jp/kegg
[35]http://bioinfo.org/kobas
[36]http://bioinfo.bti.cornell.edu/tool/itak
[37]http://planttfdb.gao-lab.org
[38]https://github.com/tseemann/barrnap
[39]http://trna.ucsc.edu/tRNAscan-SE
[40]http://eddylab.org/infernal
[41]https://busco.ezlab.org

Gene Family Identification, Phylogenetic Analysis, and Divergence Time Estimation

The protein sequences of *A. thaliana*, *O. sativa*, *B. distachyon*, *Z. mays*, *S. bicolor*, *O. thomaeum*, *Panicum hallii*, *Setaria viridis*, *Hordeum vulgare*, and *Tritcum urartu* were downloaded from the Phytozome database. Orthologous gene families were clustered using OrthoFinder v2.5.4[42] through all-against-all blast alignment of these protein sequences and predicted protein sequences of *C. dactylon* (Emms and Kelly, 2015). The identified 112 single-copy orthologous gene families were aligned using MUSCLE[43] and the alignments of each gene family were concatenated to a super-alignment matrix. A phylogenetic tree was then constructed using OrthoFinder with *A. thaliana* as the outgroup. PAML v4.9[44] was used to estimate the divergence time of *C. dactylon* using recorded divergence times of other 10 species in the TimeTree database[45] as calibrations (Yang, 2007).

Synteny and WGD Analysis

Homologous pairs of *C. dactylon* proteins were identified using the all-to-all search in BLASTP v2.12.0[46] with an E-value cutoff of 10^{-5}. Syntenic blocks with at least 50 collinear gene pairs were then identified using MCScanX[47] with default parameters (Wang et al., 2012). The same method was used to identify the collinear blocks between *C. dactylon* and *O. thomaeum*/*B. distachyon*. Synonymous substitutions per site values (Ks) of syntenic gene pairs were calculated using PAML v4.9 and the distribution of Ks values was plotted to infer the time for speciation or whole-genome duplication (WGD) events using the formula $T = \mathrm{Ks}/2\lambda$, where Ks is peak Ks value and λ is the average substitution rate (6.5×10^{-9} substitutions/site/year for grasses; Gaut et al., 1996).

Transcriptome-Based Gene Expression Analyses

The Illumina transcriptome sequencing data of six different organs of bermudagrass cultivar Yangjiang were aligned to the assembled genome using HISAT v2.1.1[48] with default parameters (Kim et al., 2015). The numbers of mapped reads for each genes were converted to RPKM (reads per kilobase of transcript per million mapped fragments) values. The $\log_2$ transformed RPKM values were applied to perform Hierarchical clustering using Pearson's correlation distance (Chen et al., 2021). The significantly expressed genes were defined as RPKM value > 1, the organ-enhanced genes were defined as RPKM value is 5-fold above the average RPKM values of other organs, whereas organ-enriched genes were defined as RPKM value is 5-fold above the RPKM values of any other organs (Uhlén et al., 2016; Nautiyal et al., 2020).

[42]https://github.com/davidemms/OrthoFinder
[43]https://www.ebi.ac.uk/Tools/msa/muscle
[44]http://abacus.gene.ucl.ac.uk/software/
[45]http://www.timetree.org
[46]https://ftp.ncbi.nlm.nih.gov/blast/executables/
[47]https://github.com/wyp1125/MCScanX
[48]http://daehwankimlab.github.io/hisat2

TABLE 1 | Statistics of *Cynodon dactylon* genome assembly.

	Illumina + PacBio	Illumina + PacBio + BioNano	Illumina + PacBio + BioNano + Hi-C
Assembly size (Mb)	1294.65	1258.08	1005.67
Scaffold number	3,703	241	36
N50 Scaffold length (Mb)	2.65	9.38	28.85
Longest scaffold (Mb)	13.42	34.64	52.77
Mean scaffold length (Mb)	0.35	5.22	27.94
Complete BUSCOs	97.80%	97.67%	96.20%

Analyses of Homologous Recombination-Regulating Genes and Tiller-Angle-Regulating Genes

To obtain ZMM (acronym for Zip1-4, Msh4-5, and Mer3) protein-coding genes in *C. dactylon* and other 10 plant species, ZMM genes from *A. thaliana* were used as baits to search against the assembled genome of *C. dactylon* and other plant species recorded in Phytozome database or Ensembl Plants database[49] using BLASTP v2.12.0 with an E-value cutoff of 10^{-5}. The gene copy numbers and chromosome locations of different genes were manually summarized based on their identities. For *PROG1*, *LA1*, and *TAC1* genes, PROG1, LA1, and TAC1 proteins from *O. sativa* were used as baits to search against the assembled genome of *C. dactylon* and other seven species of *Oryza* genus recorded in Ensembl Plants database as described above. The amino acid sequences of proteins encoded by each gene families were searched against the Pfam database[50] for domain comparisons (Mistry et al., 2021).

RESULTS

Assembly of the *Cynodon dactylon* Genome

The *C. dactylon* cultivar Yangjiang was used for genome sequencing. As a national authorized *C. dactylon* cultivar., cultivar Yangjiang is a typical turf-type common bermudagrass and is widely used for turf planting in China (Zhang et al., 2018a; **Supplementary Figure S1**). Based on the K-mer genome survey result, the estimated genome size of *C. dactylon* cultivar Yangjiang is approximately 984 Mb, which is in line with the flow cytometry genome size estimation result of 1.02 Gb (**Supplementary Figure S2**). K-mer analysis also revealed that the genome of *C. dactylon* cultivar Yangjiang has a high heterozygosity (1.92%) with a repeat frequency of 56.91%.

To overcome the impact of high heterozygosity on the genome assembly, we adopted an integrated assembly strategy incorporating PacBio sequencing, Illumina sequencing, and Bionano and Hi-C techniques with the haplotype-resolving Hifiasm algorithm (Cheng et al., 2021; **Supplementary Figure S3**). Firstly, 151.99 Gb PacBio long reads (about 150× coverage of the genome) were *de novo* assembled into contigs, which were polished by 161.1 Gb Illumina paired-end reads (about 160× coverage of the genome; **Supplementary Tables S1 and S2**). Totally, 3,703 contigs with a N50 contig length of 2.65 Mb and a sum contig length of 1.295 Gb were obtained (**Table 1**). Secondly, 395.4 Gb Bionano optical maps (about 390× coverage of the genome) were used to integrate the contigs into scaffolds (**Supplementary Figure S4**; **Supplementary Table S3**). This procedure generated 241 scaffolds with a N50 scaffold length of 9.38 Mb and a sum scaffold length of 1.26 Gb (**Table 1**). Lastly, 231.3 Gb Hi-C data (24% useful information, about 55× coverage of the genome) were used to further cluster the scaffolds into pseudo chromosomes (**Supplementary Figure S5**; **Supplementary Table S4**). The finally obtained genome assembly (1.01 Gb) contained 36 chromosome-level superscaffolds, among which the longest and the shortest are 52.77 Mb and 14.32 Mb, respectively (**Figure 1**; **Table 1**). The assembly size was consistent with the estimated genome size. Furthermore, BUSCO analysis against the 1,375 *Embryophyta* gene sets indicated that 96.2% complete genes were successfully identified in the genome assembly, among which 88.1% were duplicated genes (**Supplementary Table S5**). These results collectively suggested that the assembled *C. dactylon* genome is high quality and complete.

Annotation of the *Cynodon dactylon* Genome

A total of 76,879 protein-coding genes with an average gene length of 3,535 bp and an average transcript number per gene of 1.9 were successfully predicted from the assembled genome (**Table 2**). The predicted gene model was also evaluated by BUSCO analysis. The result indicated that 1,324 (96.3%) complete core *Embryophyta* genes were identified and the majority (1,272, 96.07%) was duplicated genes (**Supplementary Table S5**). Among the predicted 146,743 transcripts, 87.89% (128, 966) were annotated by various functional database (**Supplementary Figure S6**; **Supplementary Table S6**). Functional classification further indicated that signal transduction mechanism, post-translation modification/protein turnover/chaperones, and transcription are the top three categories containing the largest number of transcripts (**Supplementary Figure S7**). Specifically, 4,888 transcription factors (TFs) belonging to 65 classes were successfully identified. Compared with other grass species, gene numbers of HSF, WRKY, NF-X1, NF-YA, NF-YC, CPP, GARP-G2-like, and DDT TF families were greatly increased in common bermudagrass (**Supplementary Table S7**). In addition, 6,265 non-protein-coding genes were also identified, including 1349 rRNAs, 2047 tRNAs, 1025 miRNAs, 1441 snoRNAs, and 403 snRNAs (**Table 2**).

[49]http://plants.ensembl.org
[50]https://pfam.xfam.org

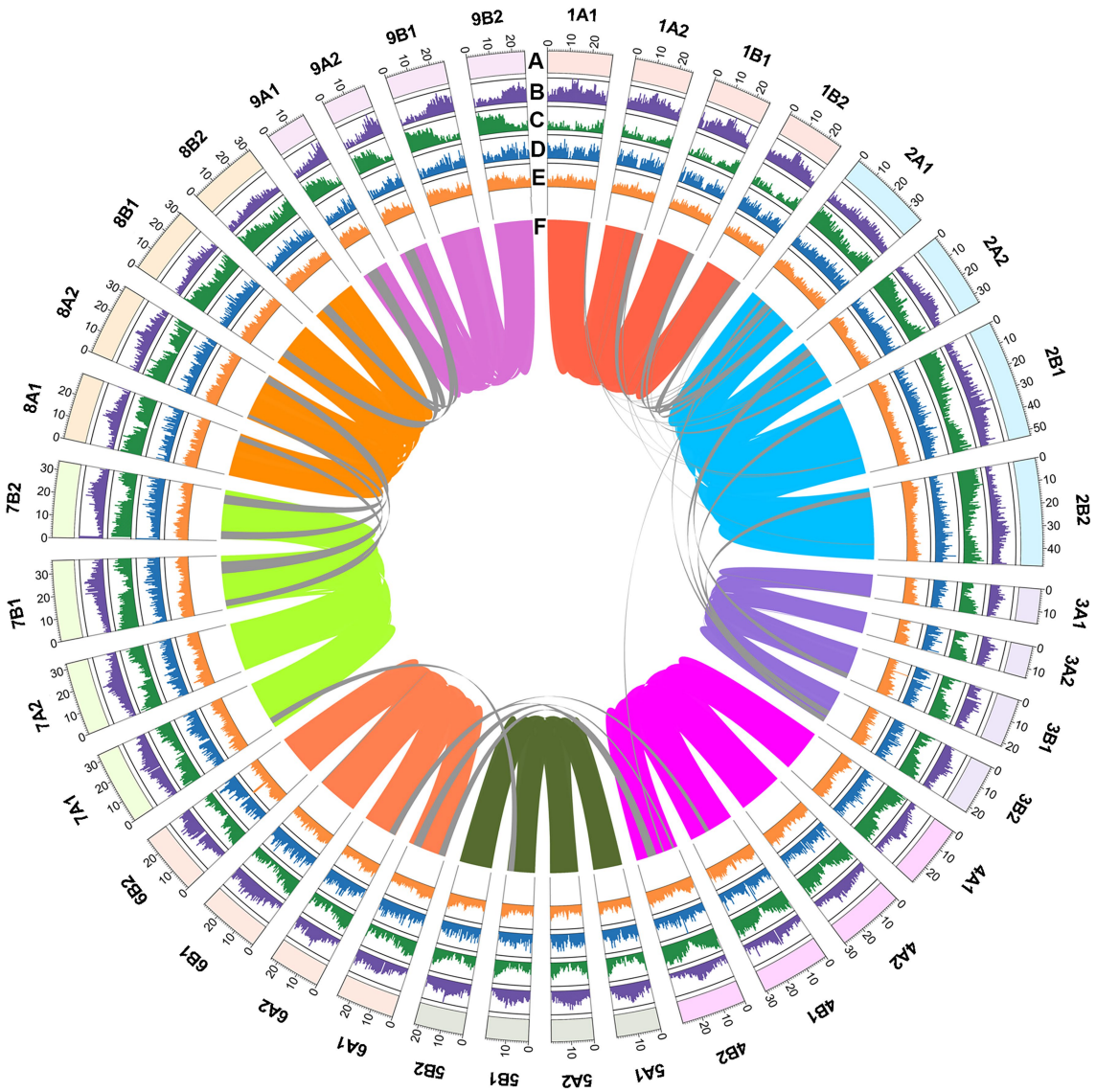


FIGURE 1 | Genome features of *C. dactylon* cultivar Yangjiang. **(A)** Circular representation of the 36 pseudo chromosomes with scale mark labeling each 10 Mb. The density of **(B)** long-terminal repeat retrotransposons (LTR-RT), **(C)** protein-coding genes, **(D)** tandem repeat sequences, and **(E)** GC contents were calculated using 500 kb non-overlap window. **(F)** Inter-chromosomal synteny was illustrated with color lines.

Orthologous clustering of protein-coding genes of *C. dactylon* with other ten plant species totally identified 32,695 orthologous gene families, including 7,792 commonly shared gene families and 3,173 bermudagrass-specific gene families consisting of 9,152 genes (**Supplementary Tables S8 and S9**). KOBAS enrichment analysis indicated that these bermudagrass-specific genes were enriched in glutathione metabolism, zeatin biosynthesis, ubiquitin mediated proteolysis, and other eight pathways (q value <0.05; **Supplementary Table S10**). In agreement with the BUSCO analysis result, orthologous gene clustering further revealed that as many as 91.2% (70117) of *C. dactylon* genes are members of 17,632 multiple-copy gene families, which is much higher than that of other ten species (**Figure 2A**; **Supplementary Table S8**). A phylogenetic tree was constructed based on the 112 shared single-copy orthologous genes (**Figure 2A**). The result indicated that *O. thomaeum* was closest to *C. dactylon* and the estimated divergence time of the two species was between 17.85 to 29.19 (midvalue of 23.52) million years ago (MYA). In line with phylogenic relationships, *C. dactylon* shared more orthologous gene families with members of the PACMAD (acronym for Panicoideae, Aristidoideae, Chloridoideae, Micrairoideae, Arundinoideae, and Danthonioideae) clade of grasses, including *O. thomaeum*, *S. bicolor*, and *S. viridis*, compared with *O. sativa* belonging to the BEP (acronym for Bambusoideae, Ehrhartoideae, and Pooideae) clade of grasses (**Supplementary Figure S8**).

A total of 381.3 Mb of repetitive sequences were also annotated in the assembled *C. dactylon* genome (**Table 2** and

TABLE 2 | Statistics of *C. dactylon* genome annotation.

Number of non-protein-coding genes	6,264
Number of rRNA genes	1,349
Number of tRNA genes	2047
Number of miRNA genes	1,025
Number of snoRNA genes	1,340
Number of snRNA genes	503
Number of protein-coding genes	76,879
Mean gene length (bp)	3534.75
Percentage in genome (%)	27.02
Mean transcript number per gene	1.91
Total transcript number	146,743
Mean transcript length (bp)	1680.90
Mean 5′UTR length (bp)	145.36
Mean 3′UTR length (bp)	217.54
Mean coding sequence length (bp)	1392.22
Mean exon number per gene (bp)	7.29
Mean exon length (bp)	230.63
Mean intron number per gene (bp)	6.29
Mean intron length (bp)	409.82
Number of repetitive sequences	529,092
Mean repetitive sequence length (bp)	720.67
Percentage in genome (%)	37.91

Supplementary Table S11). The most abundant repetitive sequences are retrotransposons (70.95% of repetitive sequences and 26.9% of genome assembly), among which LTR-RTs and non-LTR-RT represent 84.73 and 15.27%, respectively. DNA transposons make up 11.29% of the repetitive sequences (4.28% of genome assembly), whereas tandem repeats and unclassified repeat sequences account for 2.79 and 3.75% of the assembled genome, respectively. Interestingly, the total repetitive sequence content and retrotransposon content in *C. dactylon* (37.91 and 26.9%, respectively) were similar to those of closely related species, including *O. thomaeum* (45 and 26%, respectively), *P. hallii* (36 and 23%, respectively), and *S. viridis* (46 and 29%, respectively), but much lower than those of distantly related species, including *Z. mays* (82 and 76%, respectively), *T. urartu* (81 and 72%, respectively), and *H. vulgare* (80 and 75%, respectively; **Figure 2B**). It is also noteworthy that genes are unevenly distributed in different chromosomes (39.87 to 104.80 Mb^{-1} in density), whereas similar distributions of repetitive sequences were found on all chromosomes of *C. dactylon* (482.23 to 559.93 Mb^{-1} in density; **Supplementary Table S12**). The annotated LTR-RTs were further used to calculate the LAI of the assembled genome. The total LAI score of *C. dactylon* genome is 13.63, implying that the current assembly of *C. dactylon* genome reached the reference genome level (**Supplementary Table S13**; Ou et al., 2018).

Subgenome Composition of *Cynodon dactylon*

Intra-genomic syntenic analysis totally detected 845 syntenic blocks containing 84,649 pairs of homoeologous genes in the *C. dactylon* genome, whereas 643 syntenic blocks containing 52,590 pairs of homoeologous genes were found between *C. dactylon* and *O. thomaeum* through inter-genomics syntenic analysis (**Figure 3A** and **Supplementary Figure S9**). Interestingly, the syntenic depth ratios of *C. dactylon* versus *O. thomaeum* and *C. dactylon* itself were 4:1 and 4:4, respectively, implying that *C. dactylon* genome is composed of four haplotypes containing the same number of chromosomes (**Supplementary Figure S9**). To distinguish homoeologous chromosomes from the four haplotypes of *C. dactylon*, putative centromeric array tandem repeat sequences were identified from the 36 chromosomes and were used to construct a maximum likelihood phylogenetic tree as previously described (**Supplementary Table S14**; VanBuren et al., 2020). The result indicated that the 36 centromeric array sequences showed distinguishing polymorphisms and could be clustered in four clades (**Supplementary Figure S10**). Based on this classification result and chromosome length variance, the four haplotypes, which were named as A1, A2, B1 and B2, respectively, were successfully resolved in the *C. dactylon* genome (**Supplementary Table S15**). In addition, syntenic analysis also revealed that chromosome 2, 3, and 10 of *O. thomaeum* are split and merged into chromosome 2 and 7 in four haplotypes of *C. dactylon*, whereas other chromosomes all have one-to-one correspondence (**Figure 3A** and **Supplementary Figure S9**).

Calculation of Ks of homologous gene pairs in the inter-genomic and intra-genomic synteny blocks not only confirmed the phylogenic analysis result that *C. dactylon* and *O. thomaeum* diverged at approximately 21.54 MYA (Ks = 0.28), but also indicated that two rounds of WGD events occurred in the evolutionary history of *C. dactylon* (**Supplementary Figure S11**). Specifically, the first WGD event occurred at approximately 5.38 MYA (Ks = 0.07), whereas the second WGD event occurred lately at about 0.77 MYA (Ks = 0.01; **Supplementary Figure S11**). Interestingly, the two WGD time points equivalent exactly to the divergence time of haplotypes A1/A2 with haplotypes B1/B2 and haplotype A1 with haplotype A2 (the same as haplotype B1 with haplotype B2), respectively (**Figure 3B**).

In combination with the orthologous gene clustering result, syntenic analysis totally identified 20,849 alleles in *C. dactylon* (**Supplementary Table S16**). Among these alleles, 11,614 have four allelic copies in all haplotypes, 2,711 have three allelic copies in three of four haplotypes, and 6,524 have two allelic copies in two of four haplotypes (**Figure 3C** and **Supplementary Figure S12**; **Supplementary Table S17**). Meanwhile, 3559, 4954, 1716, and 3530 orphan genes that exist as single-copy genes were also identified from haplotype A1, A2, B1, and B2, respectively (**Figure 3C** and **Supplementary Figure S12**; **Supplementary Table S17**). KOBAS enrichment analyses indicated that alleles were enriched in valine/leucine/isoleucine degradation, proteasome, brassinosteroid biosynthesis, and other eight pathways, whereas orphan genes were enriched in plant-pathogen interaction, base excision repair, DNA replication, and other nine pathways (q value <0.05; **Supplementary Table S18**). The expression abundance of alleles and orphan genes were further analyzed using the organ-specific transcriptome sequencing data of *C. dactylon* cultivar Yangjiang (Chen et al., 2021; **Supplementary Table S19**). The result indicated that similar portions of alleles and orphan genes in the four haplotypes were significantly expressed in six organs of bermudagrass; however, gene numbers of alleles and orphan genes enhance- or enrich-expressed in different organs, especially

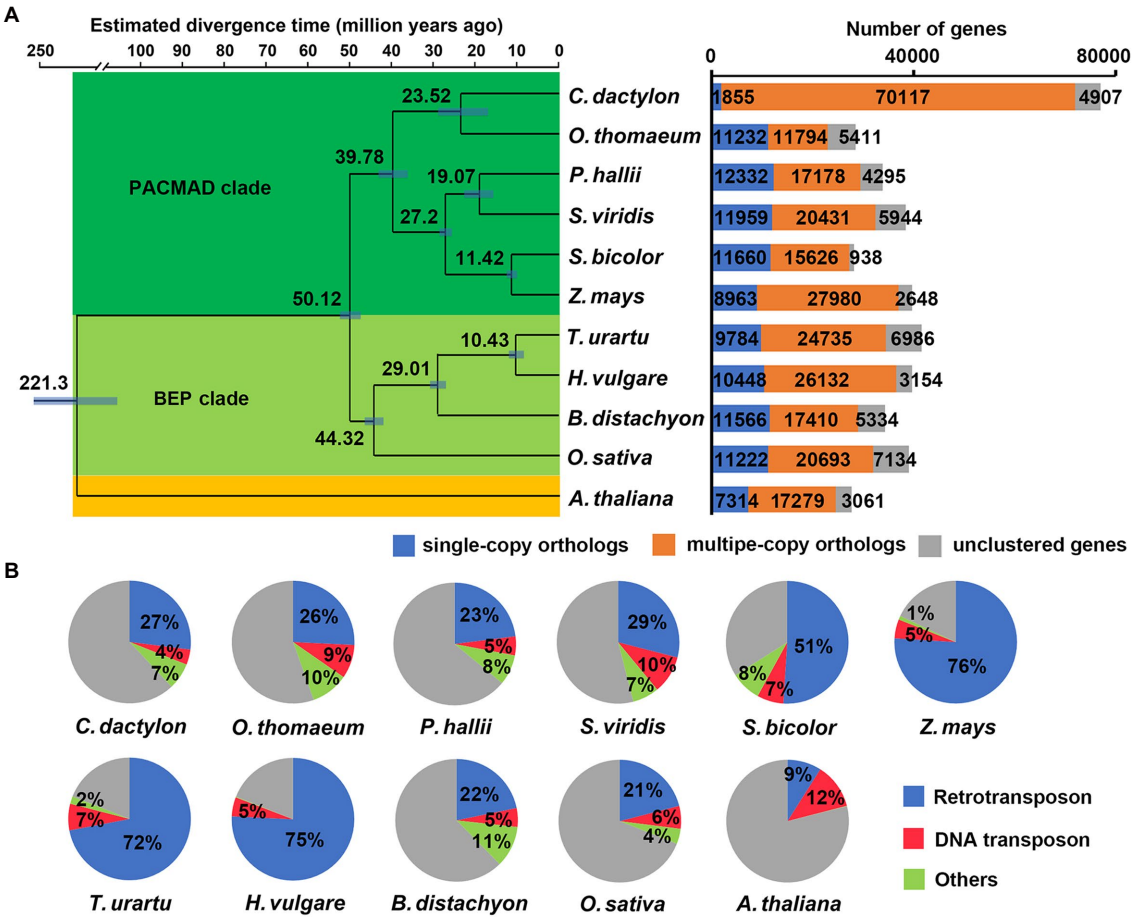


FIGURE 2 | Comparative genomic analysis among *C. dactylon* and other plant species. **(A)** Phylogenetic relationship, divergence time, and gene family clustering of *C. dactylon* and other ten plant species. Left panel, Maximum parsimony (MP) species tree was constructed using protein sequences of 112 shared single-copy orthologous genes. The numbers in the brackets indicate the estimated divergence time of each node, and the blue bars show the 95% confidence interval of divergence time. All the nodes are 100% bootstrap support. Right panel, Orthologous gene families of *C. dactylon* and other ten plant species. **(B)** Comparison of repetitive sequences in *C. dactylon* and other ten plant species.

the three types of stems, varied greatly in the four haplotypes (**Supplementary Figure S13**; **Supplementary Tables S20 and S21**). Accordingly, the 11,614 four-copy alleles of the four haplotypes showed similar total expression abundance in the six organs (**Figure 3D**).

The distribution of repeat sequences in *C. dactylon* was also analyzed at the haplotype level. Among the four haplotypes, haplotype A2 and B1 has the minimum and maximum number of RTs, respectively (**Supplementary Figure S14**). By contrast, maximum number of four types of DNA transposons, including Tcl/mariner, EnSpm/CACTA, hAT, and muDR, was observed in haplotype B2, while haplotype A1 has the fewest muDR- and Helitron-type of DNA transposons (**Supplementary Figure S14**). Notably, total sequence length of Ty3-Gypsy LTR-RTs in haplotype B1 was 2.2 Mb larger than that of haplotype B2, which contributed approximately 40% of size variance between the two haplotypes, whereas another type of LTR-RTs, Ty1-Copia, showed similar sequence length in the two haplotypes (**Supplementary Figure S14**; **Supplementary Tables S22 and S23**). Moreover, 5,066 intact LTR-RTs were further used to estimate the insertion time of different families of LTR-RTs in *C. dactylon* genome (**Supplementary Table S24**). The results indicated that four families of LTR-RTs, including Athila, SIRE, TAR, and Tork, inserted into the four haplotypes of *C. dactylon* genome at different time, whereas other nine families showed similar insertion time range in the four haplotypes (**Figure 3E**). Interestingly, among the 244 active LTR-RTs with an insertion time of zero, 153 were located in three chromosomes of haplotype B1, 47 were located in two chromosomes of haplotype A1, whereas only 29 and 16 were located in single chromosome of haplotype A2 and B1, respectively (**Supplementary Table S24**).

Adaptive Evolution of *Cynodon dactylon*

As a polyploid plant species with four sets of chromosomes, *C. dactylon* might develop a mechanism to control proper pairing and segregation of chromosomes during meiosis thus

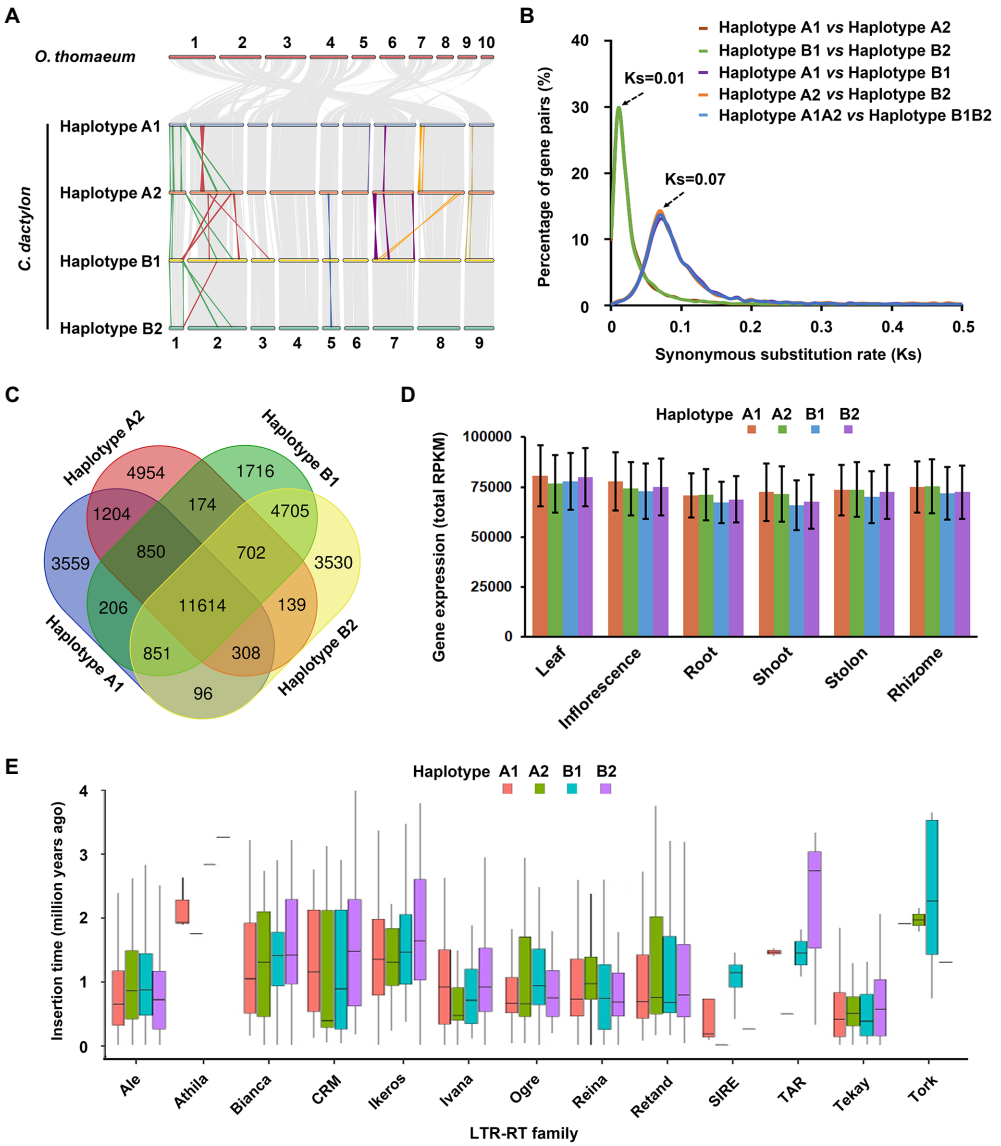


FIGURE 3 | Subgenomic organization and variation of *C. dactylon*. **(A)** Schematic representation of syntenic genes among *O. thomaeum* and four haplotypes of *C. dactylon*. Gray lines depict homologous genome blocks. Color lines indicate inversion and translocation on the homologous chromosomes. **(B)** Distribution of synonymous nucleotide substitution levels (Ks) of syntenic gene pairs between different haploptypes of *C. dactylon*. **(C)** Venn diagram of alleles and orphan genes in the four haploptypes of *C. dactylon*. **(D)** Total gene expression level of the 11,614 four-copy alleles based on their relative expression level in six organs of *C. dactylon*. Error bars represent SE of the three sequencing replicates. **(E)** Box plots showing the insertion dynamics of 13 LTR-RT families in four haploptypes of *C. dactylon*. Box boundaries indicate the 25th and 75th percentiles of the insertion time and whiskers extend to 1.5 times the interquartile range.

maintain its genome stability (Svačina et al., 2020). ZMM proteins, which stabilize the D-loop crossover intermediate of synapsis, are important homologous recombination regulators in all eukaryotes (Pyatnitskaya et al., 2019; **Figure 4A**). Previous studies have illustrated that gene copy number reduction of a ZMM protein, MSH4, could prevent meiotic crossovers between non-homologous chromosomes and stabilize the genome in allotetraploid *Brassica napus* (Gonzalo et al., 2019). Similar gene copy number reduction of MSH4 was also observed in other two polyploidy plants, allotetraploid *Gossypium hirsutum* and hexaploid *Triticum aestivum* (**Figure 4B**; **Supplementary Table S25**). However, *MSH4* and other four *ZMM* genes, including *ZYP1*, *MER3*, *SHOC1*, and *MSH5*, all existed as four-copy alleles in *C. dactylon* genome. By contrast, two *ZMM* genes, *PTD* and *HEL10*, existed as two-copy alleles, and another *ZMM* gene, *ZIP4*, existed as single-copy orphan gene in haplotype B1 of *C. dactylon* genome (**Figure 4B**; **Supplementary Tables S25 and S26**). Syntenic analysis further indicated that different sizes of chromosomal fragments containing *ZIP4* were lost in other three haplotypes (**Figure 4C**). These results collectively implied that *C. dactylon* also evolved a ZMM-dependent regulatory

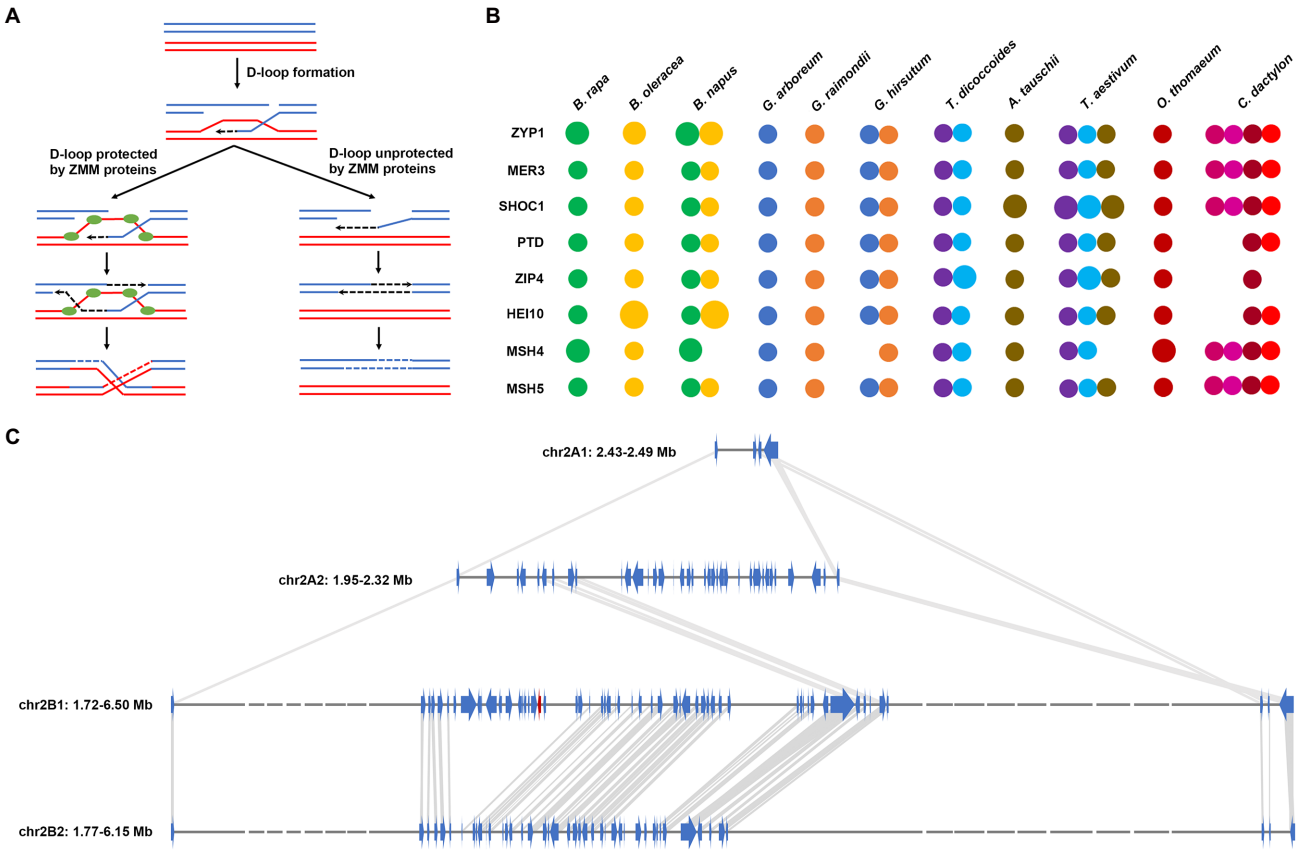


FIGURE 4 | Evolution of homologous recombination-regulating genes in *C. dactylon*. **(A)** Diagram depicting the essential roles of ZMM proteins in regulating homologous recombination. **(B)** Comparison of the gene copy number of eight ZMM proteins in *C. dactylon*, *O. thomaeum*, and other nine species with different ploidy levels. **(C)** Enlarged chromosomal gene location map showing the loss of *ZIP4* gene and other contiguous genes in haplotype A1, A2, and B2 of *C. dactylon*. The *ZIP4* gene in haplotype B1 was shown in red color.

mechanism to maintain its genome stability as other polyploidy plants did; however, the key regulator might be ZIP4 rather than MSH4.

As a widely used turfgrass species with two types of specialized stems, stolon and rhizome, *C. dactylon* exhibits a prostrate plant architecture owing to increased tiller angles of the two specialized stems (Dong and de Kroon, 1994). Previous studies have successfully identified several tiller-angle-regulating genes, including *PROG1*, *TAC1*, and *LA1*, in rice and other plants (Li et al., 2007; Yu et al., 2007; Jin et al., 2008; Tan et al., 2008; **Figure 5A**). Eight *PROG1*-like genes, four *TAC1*-like genes, and two *LA1*-like genes were also identified in *C. dactylon* (**Figure 5B**). Similar to semi-prostrate and prostrate growing *Oryza* genus plants, the family/genome gene number ratio of two prostrate growth-promoting genes, *PROG1*-like and *TAC1*-like, were higher than that of erect-growth-promoting *LA1*-like gene in *C. dactylon* (**Figure 5B**). Syntenic and phylogenic analysis revealed that six of the eight *PROG1*-like genes existed as three-copy alleles and the remaining two genes existed as two-copy alleles, the four *TAC1*-like genes existed as four-copy alleles, whereas two *LA1*-like genes existed as two-copy alleles (**Figure 5C** and **Supplementary Figures S15, S16**). Pfam domain analysis further indicated that all eight PROG1-like proteins have the conserved C_2H_2-type zinc finger domain identified in the functional OsPROG1 and OgPROG7 proteins, however, both two LA1-like proteins of *C. dactylon* lack the functional C-terminal conserved region V (**Figure 5D** and **Supplementary Figure S15**). Moreover, the LA1-like protein encoded by the allele of haplotype A2 further lack the functional N-terminal conserved region I and two other conserved regions II and III (Yoshihara and Spalding, 2020; **Figure 5D**). In combination with the observation that large chromosome fragments containing the *LA1-like* gene locus were lost in other two haplotypes (**Supplementary Figure S16**), sequence variation of *LA1-like* genes in the two residual alleles suggested that LA1 protein activity was inhibited in *C. dactylon*. In addition, both two *LA1*-like genes were weakly expressed in stolon and rhizome, whereas three of four *TAC1*-like genes were preferentially expressed in the two specialized stems (**Supplementary Figure S16**; **Supplementary Table S27**). These results collectively implied that different tiller-angle-regulating genes were synergistically evolved to promote a prostrate plant architecture in *C. dactylon*.

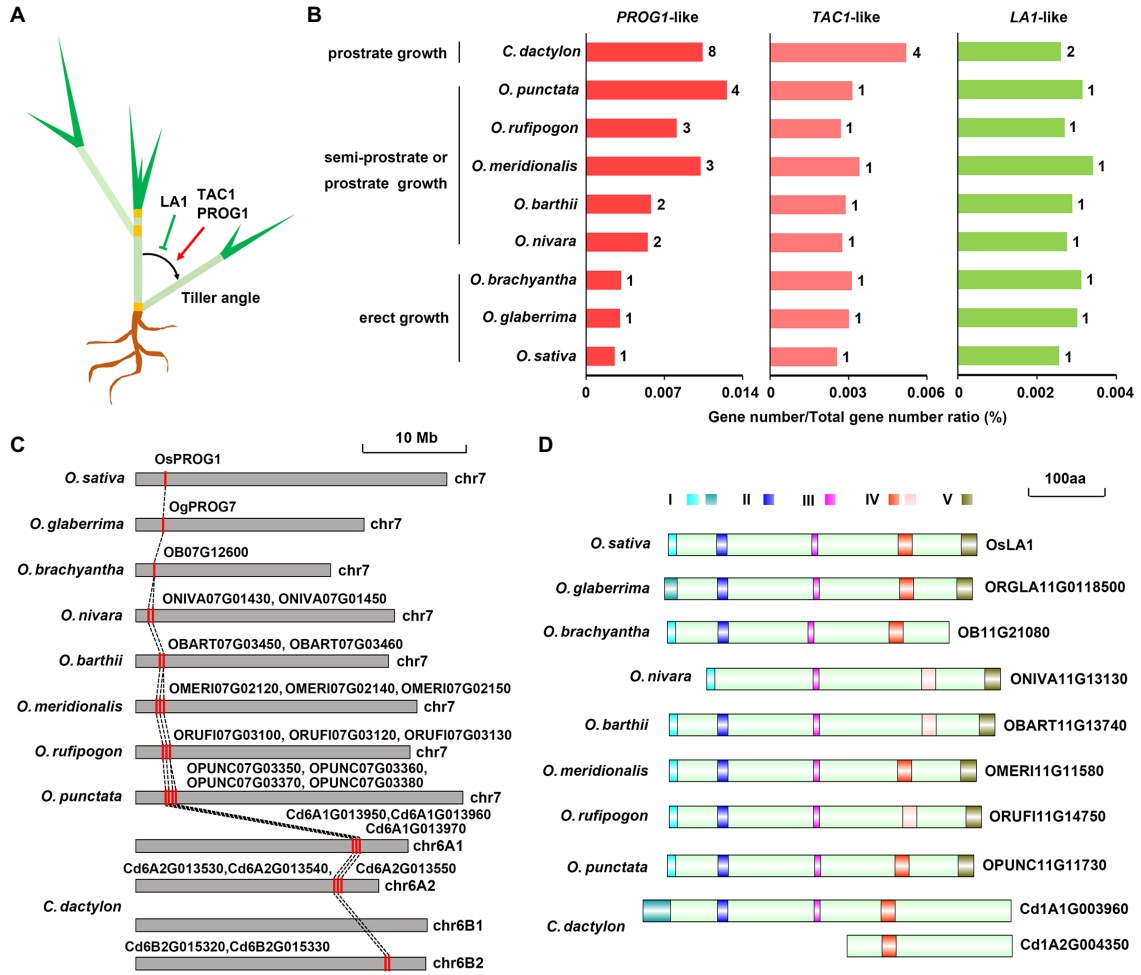


FIGURE 5 | Evolution of tiller angle-regulating genes in *C. dactylon*. **(A)** Diagram depicting the positive/negative regulatory roles of *PROG1*, *LAZY1*, and *TAC1* genes in tiller angle control of grasses. **(B)** Comparison of the gene number of *PROG1*-, *LAZY1*-, and *TAC1*-like genes in *C. dactylon* and eight species of *Oryza* genus with different growth habits. **(C)** Syntenic relationship of *PROG1*-like genes in *C. dactylon* and eight species of *Oryza* genus with different growth habits. **(D)** Diagram showing the deficiency of key functional motifs in two LAZY1-like proteins of *C. dactylon* compared with those of eight species of *Oryza* genus.

DISCUSSION

Turfgrasses are important groups of grass species serving essential functions, including soil stabilization, water conservation, filtration of air, and water borne pollutants, in urban and suburban landscapes (Huang, 2021). In the past several years, genome sequences of many turfgrass species, including zoysiagrasses (*Zoysia japonica* and *Zoysia matrella*), perennial ryegrass (*Lolium perenne*), centipedegrass (*Eremochloa ophiuroides*), and African bermudagrass (*C. transvaalensis*), were successfully sequenced and assembled using different techniques (Tanaka et al., 2016; Cui et al., 2021; Frei et al., 2021; Wang et al., 2021). In this study, we reported a high-quality haplotype-resolved genome of another important turfgrass species, common bermudagrass (*C. dactylon*), consisting of 36 pseudo chromosomes with a contig N50 of 2.65 Mb and a LAI score of 13.63 (**Figure 1**; **Table 1** and **Supplementary Table S13**). The assembled nome of *C. dactylon* not only offers a solid foundation to study the molecular basis of valuable agronomic traits as well as molecular breeding of this important turfgrass species, but also provides an essential resource for comparative genomic analysis among different turfgrasses and other grasses.

The most prominent characteristics of *C. dactylon* genome are the presence of four haplotypes, named as A1, A2, B1, and B2, respectively. As an allotetraploid plants with high heterozygosity (1.92%), *C. dactylon* has four sets of chromosomes with significant differences that could be discriminated as different haplotypes using the newest haplotype-resolving Hifiasm algorithm; thus, an A1A2B1B2 genome assembly with 36 chromosomes, rather than an AB genome assembly with 18 chromosomes, was finally obtained (Kyriakidou et al., 2018). Similar result was also observed in the haplotype-phased genome assembly of tetraploid blueberry (2n = 4x = 48), which also reported a four haplotype-resolved genome containing 48 pseudo chromosomes (Colle et al., 2019). Notably, all the four haplotypes of *C. dactylon* have nine chromosomes; however, the total

chromosome size showed variance among different haplotypes (**Figure 3A**; **Supplementary Table S15**). Specifically, haplotype A1 and A2 have a similar size (236.96 Mb and 231.36 Mb), whereas haplotype B1 and B2 have another similar size (271.19 Mb and 266.17 Mb; **Supplementary Table S15**). Accordingly, genes and repeat sequences in haplotype A1 and A2 are fewer than those in haplotype B1 and B2 (**Supplementary Table S15**). Interestingly, the size of four haplotypes are similar to the genome size of *O. thomaeum* (10 chromosomes, 243 Mb) and the monoploid genome size of *Eragrostis tef* (10 chromosomes, 288 Mb), two grass species belonging to the Chloridoideae subfamily of PACMAD clade of grasses as *C. dactylon* does, but much smaller than the genome size of African bermudagrass *C. transvaalensis* (nine chromosomes, 444 Mb), which is classified along with *C. dactylon* in the same *Cynodon* genus (VanBuren et al., 2015, 2020; Cui et al., 2021). Similar chromosome size variation was also observed between *Morus notabilis* and *M. alba*, both of which belongs to the same *Morus* genus (Xuan et al., 2022). These findings collectively suggested that genome size variation among different plant species might not be simply correlated with their phylogenic relationships.

Whole-genome duplication is an extreme mechanism of gene duplication that leads to a sudden increase in both genome size and the entire gene set thus plays important roles in plant genome evolution (Clark and Donoghue, 2018). Ks analysis revealed that two rounds of WGD events occurred in *C. dactylon*, which is in correspondence to the divergence time of haplotypes A1/A2 with haplotypes B1/B2 at 5.38 MYA and haplotype A1 with haplotype A2 at 0.77 MYA (the same as haplotype B1 with haplotype B2), respectively (**Figure 3B** and **Supplementary Figure S11**). These results collectively implied a complex evolutionary history of *C. dactylon*. At approximately 5.38 MYA, the ancestor of haplotype A1 and A2, named as A, might hybridized with B, the ancestor of haplotype B1 and B2, to form an AB hybrid species. At about 0.77 MYA, either an autopolyploidization event occurred in the AB hybrid species that doubled the genome to AABB or a secondary hybridization event occurred between two AB hybrid species to form an ABAB hybrid species through allopolyploidization, both of which could finally evolved into the present A1A2B1B2 genome of *C. dactylon*. The latter allopolyplodization mechanism seems more possible because the ratio of coupling to repulsion linkage phase of nondistorted mapped loci was approximately 1: 1 in an SSR-maker based linkage mapping of the first-generation selfed population of *C. dactylon* (Guo et al., 2017). Similar two rounds of WGD events were also observed in the formation of the polyploidy genome of *Miscanthus floridulus* and *Saccharum spontaneum*, suggesting a conserved evolution mechanism might exist in different genus of polyploid grasses (Zhang et al., 2018b, 2021).

A dominant subgenome often emerges immediately following the WGD event in the genome of allopolyploids (Liang and Schnable, 2018). However, some recent allopolyploids, including the above-mentioned *M. floridulus* and *S. spontaneum*, display indistinguishable or slight subgenome dominance (Zhang et al., 2018b, 2021). Orthologous gene clustering analysis indicated that four haplotypes of *C. dactylon* shared similar number of gene families with *O. thomaeum* (**Supplementary Figure S8**). Syntenic analysis further revealed that the four haplotypes have 12,197 (68.50% of 17,805), 12,039 (68.40% of 17,600), 14,406 (69.20% of 20,818), and 14,347 (69.46% of 20,656) syntenic orthologs to *O. thomaeum*, respectively (**Supplementary Figure S9**). These results suggested that four subgenomes of *C. dactylon* did not experience biased gene loss during evolution. Moreover, although a few genes from different haplotypes showed biased expression in different organs, overall gene expression levels showed high similarity among the four haplotypes (**Figure 3D** and **Supplementary Figure S13**). In addition, similar distribution and insertion time of LTRs were also observed in the four haplotypes (**Figure 3E** and **Supplementary Figure S14**). Taken together, these analyses collectively implied subgenome dominance is also unobvious in *C. dactylon*.

Polyploidy brings many advantages to polyploid plants. Heterosis could foster a greater biomass and accelerated development, whereas gene redundancy could mask deleterious mutations and diversify the functions of extra gene copies (Comai, 2005). As a worldwidely distributed grass species inhabiting diverse and harsh environments, allotetraploid *C. dactylon* undoubtedly benefits from these advantages. However, long-term survival of polyploid plants also require a mechanism to withstand the extensive genomic instability that accompanies with the presence of multiple pairing chromosomes in meiosis (Mason and Mason and Wendel, 2020). As a clonal plant with stolons and rhizomes, *C. dactylon* reproduces asexually through regenerating new plants from axillary buds of stolon and rhizome node (Dong and de Kroon, 1994), thus bypasses meiosis and recombination in gamete generation process. On the other hand, a ZMM-dependent regulatory mechanism to maintain genome stability during meiosis was also identified in *C. dactylon* (**Figure 4**). Owing to these belt and braces strategies, four unbiased haplotypes of subgenome are stably maintained in *C. dactylon* genome.

Tiller angle (branch angle in eudicot plants) is an important plant architectural trait affecting the density of growing plants (Wang et al., 2022). Cereal grasses often have compact and erect plant architecture characteristics with small tiller angles, which is essential for high yields. Specifically, successful domestication of cultivated rice from wild rice ancestors depended on the transition from prostrate growth to erect growth, in which process the tiller angle was greatly reduced (Li et al., 2007; Yu et al., 2007; Jin et al., 2008; Tan et al., 2008). However, for turfgrasses including *C. dactylon*, prostrate growth mode with large tiller angle is more preferable because it could accelerate turf formation, increase soil coverage, and diminish mowing frequency (Wang et al., 2021). Blast searches indicated that key tiller-angle-regulating genes reported in rice and other plants, including *PROG1*, *LA1*, and *TAC1*, were highly conserved in *C. dactylon* (**Figure 5B**). Similar to prostrate growing wild rice species, clustering of *PROG1*-like C_2H_2 transcription factor genes in adjacent positions of chromosomes were observed in *C. dactylon* (Wu et al., 2018; Huang et al., 2020; **Figure 5C** and **Supplementary Figure S15**). By contrast, *LA1*-like genes that promote erect growth not only experienced gene copy lost due to large chromosomal fragment deletions but also

mutated to form truncated proteins (**Figure 5D** and **Supplementary Figure S16**). These results strongly suggested that similar selection pressure might also exist in *C. dactylon* to form the prostrate plant architecture characteristics as the domestication of rice from wild rice; however, the selection target might be *LA1* rather than *PROG1*.

CONCLUSION

The genome of a widely used warm-season turfgrass species, *C. dactylon*, was sequenced and annotated in this study. The assembled genome contains 36 pseudo chromosomes, includes 37.91% genome size of repeat sequences, and encodes 76,879 protein-coding genes. The polyploid *C. dactylon* genome is consists of four haplotypes derived from two rounds of WGD events. Although a few haplotype-specific genes and transposons were identified, no global subgenome dominance was detected among the four haplotypes. A ZMM-dependent regulatory mechanism to maintain the genome stability was successfully identified. Furthermore, synergistic evolution of tiller-angle-regulating genes was also observed. In summary, the extensive datasets and analyses presented in this study not only offer an essential resource for basic studies and breeding researches of turfgrasses, but also provide new insights into regulation mechanisms underlying polyploid genome stability and prostrate growth.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: National Center for Biotechnology Information (NCBI) BioProject database under accession numbers PRJNA430136, PRJNA685207, and PRJNA805105.

AUTHOR CONTRIBUTIONS

BZ and J-YL planned and managed the project and wrote the manuscript. J-YL provided the research fund. BZ, SC, JC, and DL conducted the research and analyzed the data. JL provided the plant material and helped to write the manuscript. Y-BY helped to analyze the data and write the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by the Science and Technology Development Foundation of Tsinghua University.

ACKNOWLEDGMENTS

The authors want to appreciate Professor Hong-Wei Wang at Tsinghua University for his invaluable suggestions and Qiang Gao and Zhaoyang Wang at BGI-Shenzhen for their helps in accomplishment of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.890980/full#supplementary-material

REFERENCES

Birney, E., Clamp, M., and Durbin, R. (2004). GENEWISE and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293

Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 49, 9077–9096. doi: 10.1093/nar/gkab688

Chaves, A. L. A., Chiavegatto, R. B., Benites, F. R. G., and Techio, V. H. (2019). Comparative karyotype analysis among cytotypes of *Cynodon dactylon* (L.) Pers. (Poaceae). *Mol. Biol. Rep.* 46, 4873–4881. doi: 10.1007/s11033-019-04935-z

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., et al. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* 7, 1–6. doi: 10.1093/gigascience/gix120

Chen, S., Xu, X., Ma, Z., Liu, J., and Zhang, B. (2021). Organ-specific transcriptome analysis identifies candidate genes involved in the stem specialization of bermudagrass (*Cynodon dactylon* L.). *Front. Genet.* 12:678673. doi: 10.3389/fgene.2021.678673

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi: 10.1038/s41592-020-01056-5

Clark, J. W., and Donoghue, P. C. J. (2018). Whole-genome duplication and plant macroevolution. *Trends Plant Sci.* 23, 933–945. doi: 10.1016/j.tplants.2018.07.006

Colle, M., Leisner, C. P., Wai, C. M., Ou, S., Bird, K. A., Wang, J., et al. (2019). Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *Gigascience* 8:giz012. doi: 10.1093/gigascience/giz012

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Cui, F., Taier, G., Li, M., Dai, X., Hang, N., Zhang, X., et al. (2021). The genome of the warm-season turfgrass African bermudagrass (*Cynodon transvaalensis*). *Horticult. Res.* 8:93. doi: 10.1038/s41438-021-00519-w

Dong, M., and de Kroon, H. (1994). Plasticity in morphology and biomass allocation in *Cynodon dactylon*, a grass species forming stolons and rhizomes. *Oikos* 70, 99–106. doi: 10.2307/3545704

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Farsani, T. M., Etemadi, N., Sayed-Tabatabaei, B. E., and Talebi, M. (2012). Assessment of genetic diversity of bermudagrass (*Cynodon dactylon*) using ISSR markers. *Int. J. Mol. Sci.* 13, 383–392. doi: 10.3390/ijms13010383

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable

element families. *Proc. Natl. Acad. Sci. U. S. A.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Frei, D., Veekman, E., Grogg, D., Stoffel-Studer, I., Morishima, A., Shimizu-Inatsugi, R., et al. (2021). Ultralong Oxford Nanopore reads enable the development of a reference-grade perennial ryegrass genome assembly. *Genome Biol. Evol.* 13, evab159. doi: 10.1093/gbe/evab159

Gaut, B. S., Morton, B. R., McCaig, B. C., and Clegg, M. T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci. U. S. A.* 93, 10274–10279. doi: 10.1073/pnas.93.19.10274

Ghosh, S., and Chan, C. K. (2016). Analysis of RNA-seq data using TopHat and Cufflinks. *Methods Mol. Biol.* 1374, 339–361. doi: 10.1007/978-1-4939-3167-5_18

Glover, J. D., Reganold, J. P., Bell, L. W., Borevitz, J., Brummer, E. C., Buckler, E. S., et al. (2010). Increased food and ecosystem security via perennial grains. *Science* 328, 1638–1639. doi: 10.1126/science.1188761

Gonzalo, A., Lucas, M. O., Charpentier, C., Sandmann, G., Lloyd, A., and Jenczewski, E. (2019). Reducing MSH4 copy number prevents meiotic crossovers between non-homologous chromosomes in *Brassica napus*. *Nat. Commun.* 10:2354. doi: 10.1038/s41467-019-10010-9

Grossman, A. Y., Andrade, M. H. M. L., Chaves, A. L. A., Mendes Ferreira, M. T., Techio, V. H., Lopez, Y., et al. (2021). Ploidy level and genetic parameters for phenotypic traits in bermudagrass (*Cynodon* spp.) germplasm. *Agronomy* 11:912. doi: 10.3390/agronomy11050912

Guo, L., Plunkert, M., Luo, X., and Liu, Z. C. (2021). Developmental regulation of stolon and rhizome. *Curr. Opin. Plant Biol.* 59:101970. doi: 10.1016/j.pbi.2020.10.003

Guo, Y., Wu, Y., Anderson, J. A., Moss, J. Q., Zhu, L., and Fu, J. (2017). SSR marker development, linkage mapping, and QTL analysis for establishment rate in common bermudagrass. *Plant Genome* 10:1. doi: 10.3835/plantgenome2016.07.0074

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Harlan, J. R., and de Wet, J. M. J. (1969). Sources of variation in *Cynodon dactylon* (L). *Pers. Crop Sci* 9, 774–778. doi: 10.2135/cropsci1969.0011183X000900060031x

Hill, G. M., Gates, R. N., and West, J. W. (2001). Advances in bermudagrass research involving new cultivars for beef and dairy production. *J. Anim. Sci.* 79, E48–E58. doi: 10.2527/jas2001.79E-SupplE48x

Huang, B. R. (2021). Grass research for a productive, healthy and sustainable society. *Grass Res.* 1, 1–2. doi: 10.48130/GR-2021-0001

Huang, L., Liu, H., Wu, J., Zhao, R., Li, Y., Melaku, G., et al. (2020). Evolution of plant architecture in *Oryza* driven by the *PROG1* locus. *Front. Plant Sci.* 11:876. doi: 10.3389/fpls.2020.00876

Jin, J., Huang, W., Gao, J. P., Yang, J., Shi, M., Zhu, M. Z., et al. (2008). Genetic control of rice plant architecture under domestication. *Nat. Genet.* 40, 1365–1369. doi: 10.1038/ng.247

Khanal, S., Kim, C., Auckland, S. A., Rainville, L. K., Adhikari, J., Schwartz, B. M., et al. (2017). SSR-enriched genetic linkage maps of bermudagrass (*Cynodon dactylon* × *transvaalensis*), and their comparison with allied plant genomes. *Theor. Appl. Genet.* 130, 819–839. doi: 10.1007/s00122-017-2854-z

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kneebone, W. R. (1966). Bermuda grass-worldly, wily, wonderful weed. *Econ. Bot.* 20, 94–97. doi: 10.1007/BF02861931

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., et al. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat. Commun.* 12:1935. doi: 10.1038/s41467-020-20536-y

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2018). Current strategies of polyploid plant genome sequence assembly. *Front. Plant Sci.* 9:1660. doi: 10.3389/fpls.2018.01660

Li, P., Wang, Y., Qian, Q., Fu, Z., Wang, M., Zeng, D., et al. (2007). LAZY1 controls rice shoot gravitropism through regulating polar auxin transport. *Cell Res.* 17, 402–410. doi: 10.1038/cr.2007.38

Liang, Z., and Schnable, J. C. (2018). Functional divergence between subgenomes and gene pairs after whole genome duplications. *Mol. Plant* 11, 388–397. doi: 10.1016/j.molp.2017.12.010

Ma, Z., Chen, S., Wang, Z., Liu, J., and Zhang, B. (2021). Proteome analysis of bermudagrass stolons and rhizomes provides new insights into the adaptation of plant stems to aboveground and underground growth. *J. Proteome* 241:104245. doi: 10.1016/j.jprot.2021.104245

Ma, J., Devos, K. M., and Bennetzen, J. L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* 14, 860–869. doi: 10.1101/gr.1466204

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Mason, A. S., and Wendel, J. F. (2020). Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. *Front. Genet.* 11:1014. doi: 10.3389/fgene.2020.01014

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Nachtweide, S., and Stanke, M. (2019). Multi-genome annotation with AUGUSTUS. *Methods Mol. Biol.* 1962, 139–160. doi: 10.1007/978-1-4939-9173-0_8

Nagori, B. P., and Solanki, R. (2011). *Cynodon dactylon* (L.) Pers.: a valuable medicinal plant. *Res. J. Med. Plant* 5, 508–514. doi: 10.3923/rjmp.2011.508.514

Nautiyal, A. K., Gani, U., Sharma, P., Kundan, M., Fayaz, M., Lattoo, S. K., et al. (2020). Comprehensive transcriptome analysis provides insights into metabolic and gene regulatory networks in trichomes of *Nicotiana tabacum*. *Plant Mol. Biol.* 102, 625–644. doi: 10.1007/s11103-020-00968-2

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* 46:e126. doi: 10.1093/nar/gky730

Ozias-Akins, P., and Conner, J. A. (2020). Clonal reproduction through seeds in sight for crops. *Trends Genet.* 36, 215–226. doi: 10.1016/j.tig.2019.12.006

Price, A. L., Jones, N. C., and Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics* 21, i351–i358. doi: 10.1093/bioinformatics/bti1018

Pyatnitskaya, A., Borde, V., and De Muyt, A. (2019). Crossing and zipping: molecular duties of the ZMM proteins in meiosis. *Chromosoma* 128, 181–198. doi: 10.1007/s00412-019-00714-8

Satorre, E. H., Rizzo, F. A., and Arias, S. P. (1996). The effect of temperature on sprouting and early establishment of *Cynodon dactylon*. *Weed Res.* 36, 431–440. doi: 10.1111/j.1365-3180.1996.tb01672.x

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Svačina, R., Sourdille, P., Kopecký, D., and Bartoš, J. (2020). Chromosome pairing in polyploid grasses. *Front. Plant Sci.* 11:1056. doi: 10.3389/fpls.2020.01056

Tan, L., Li, X., Liu, F., Sun, X., Li, C., Zhu, Z., et al. (2008). Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* 40, 1360–1364. doi: 10.1038/ng.197

Tan, C., Wu, Y., Taliaferro, C. M., Bell, G. E., Martin, D. L., Smith, M. W., et al. (2014). Selfing and outcrossing fertility in common bermudagrass under open-pollinating conditions examined by SSR markers. *Crop Sci.* 54, 1832–1837. doi: 10.2135/cropsci2013.12.0816

Tanaka, H., Hirakawa, H., Kosugi, S., Nakayama, S., Ono, A., Watanabe, A., et al. (2016). Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Res.* 23, 171–180. doi: 10.1093/dnares/dsw006

Tempel, S. (2012). Using and understanding RepeatMasker. *Methods Mol. Biol.* 859, 29–51. doi: 10.1007/978-1-61779-603-6_2

Uhlén, M., Hallström, B. M., Lindskog, C., Mardinoglu, A., Pontén, F., and Nielsen, J. (2016). Transcriptomics resources of human tissues and organs. *Mol. Syst. Biol.* 12:862. doi: 10.15252/msb.20155865

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* 527, 508–511. doi: 10.1038/nature15714

VanBuren, R., Ching, M. W., Wang, X., Pardo, J., Yocca, A. E., Wang, H., et al. (2020). Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nat. Commun.* 11:884. doi: 10.1038/s41467-020-14724-z

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963

Wang, W., Gao, H., Liang, Y., Li, J., and Wang, Y. (2022). Molecular basis underlying rice tiller angle: current progress and future perspectives. *Mol. Plant* 15, 125–137. doi: 10.1016/j.molp.2021.12.002

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wang, J., Zi, H., Wang, R., Liu, J., Wang, H., Chen, R., et al. (2021). A high-quality chromosome-scale assembly of the centipedegrass [*Eremochloa ophiuroides* (Munro) Hack.] genome provides insights into chromosomal structural evolution and prostrate growth habit. *Horticult. Res.* 8:201. doi: 10.1038/s41438-021-00636-6

Wu, Y. Q., Taliaferro, C. M., Bai, G. H., and Anderson, M. P. (2004). AFLP analysis of *Cynodon dactylon* (L.) Pers. var. *dactylon* genetic variation. *Genome* 47, 689–696. doi: 10.1139/g04-032

Wu, Y. Q., Taliaferro, C. M., Bai, G. H., Martin, D. L., Anderson, J. A., Anderson, M. P., et al. (2006). Genetic analyses of Chinese *Cynodon* accessions by flow cytometry and AFLP markers. *Crop Sci.* 46, 917–926. doi: 10.2135/cropsci2005.08.0256

Wu, Y. Q., Taliaferro, C. M., Martin, D. L., Anderson, J. A., and Anderson, M. P. (2007). Genetic variability and relationships for adaptive, morphological, and biomass traits in Chinese bermudagrass accessions. *Crop Sci.* 47, 1985–1994. doi: 10.2135/cropsci2007.01.0047

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. doi: 10.1093/bioinformatics/bti310

Wu, Y., Zhao, S., Li, X., Zhang, B., Jiang, L., Tang, Y., et al. (2018). Deletions linked to *PROG1* gene participate in plant architecture domestication in Asian and African rice. *Nat. Commun.* 9:4157. doi: 10.1038/s41467-018-06509-2

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322. doi: 10.1093/nar/gkr483

Xu, J., Wang, Z., and Cheng, J. J. (2011). Bermuda grass as feedstock for biofuel production: a review. *Bioresour. Technol.* 102, 7613–7620. doi: 10.1016/j.biortech.2011.05.070

Xuan, Y., Ma, B., Li, D., Tian, Y., Zeng, Q., and He, N. (2022). Chromosome restructuring and number change during the evolution of *Morus notabilis* and *Morus alba*. *Horticult. Res* 9:uhab030. doi: 10.1093/hr/uhab030

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, L., Wu, Y. Q., Moss, J. Q., Zhong, S., and Yang, B. (2018). Molecular identification and characterization of seeded turf bermudagrass cultivars using simple sequence repeat markers. *Agron. J.* 110, 2142–2150. doi: 10.2134/agronj2018.01.0068

Yao, H., Guo, L., Fu, Y., Borsuk, L. A., Wen, T. J., Skibbe, D. S., et al. (2005). Evaluation of five *ab initio* gene prediction programs for the discovery of maize genes. *Plant Mol. Biol.* 57, 445–460. doi: 10.1007/s11103-005-0271-1

Yoshihara, T., and Spalding, E. P. (2020). Switching the direction of stem gravitropism by altering two amino acids in AtLAZY1. *Plant Physiol.* 182, 1039–1051. doi: 10.1104/pp.19.01144

Yu, B., Lin, Z., Li, H., Li, X., Li, J., Wang, Y., et al. (2007). TAC1, a major quantitative trait locus controlling tiller angle in rice. *Plant J.* 52, 891–898. doi: 10.1111/j.1365-313X.2007.03284.x

Zhang, B., Fan, J., and Liu, J. (2019). Comparative proteomic analysis provides new insights into the specialization of shoots and stolons in bermudagrass (*Cynodon dactylon* L.). *BMC Genomics* 20:708. doi: 10.1186/s12864-019-6077-3

Zhang, G., Ge, C., Xu, P., Wang, S., Cheng, S., Han, Y., et al. (2021). The reference genome of *Miscanthus floridulus* illuminates the evolution of Saccharinae. *Nat. Plants* 7, 608–618. doi: 10.1038/s41477-021-00908-y

Zhang, B., and Liu, J. (2018). Molecular cloning and sequence variance analysis of the *TEOSINTE BRANCHED1* (*TB1*) gene in bermudagrass [*Cynodon dactylon* (L.) Pers]. *J. Plant Physiol.* 229, 142–150. doi: 10.1016/j.jplph.2018.07.008

Zhang, B., Liu, J., Wang, X., and Wei, Z. (2018a). Full-length RNA sequencing reveals unique transcriptome composition in bermudagrass. *Plant Physiol. Bioch.* 132, 95–103. doi: 10.1016/j.plaphy.2018.08.039

Zhang, W., Liu, J., Zhang, Y., Qiu, J., Li, Y., Zheng, B., et al. (2020). A high-quality genome sequence of alkaligrass provides insights into halophyte stress tolerance. *Sci. China Life Sci.* 63, 1269–1282. doi: 10.1007/s11427-020-1662-x

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018b). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* 50, 1565–1573. doi: 10.1038/s41588-018-0237-2

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zheng, Y., Xu, S., Liu, J., Zhao, Y., and Liu, J. (2017). Genetic diversity and population structure of Chinese natural bermudagrass [*Cynodon dactylon* (L.) Pers.] germplasm based on SRAP markers. *PLoS One* 12:e0177508. doi: 10.1371/journal.pone.0177508

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/fpls.2022.859386

Identification and Analysis of bZIP Family Genes in *Sedum plumbizincicola* and Their Potential Roles in Response to Cadmium Stress

Zhuchou Lu[1,2†], Wenmin Qiu[1†], Kangming Jin[1], Miao Yu[1], Xiaojiao Han[1], Xiaoyang He[3], Longhua Wu[4], Chao Wu[5] and Renyin Zhuo[1*]*

[1] State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Tree Breeding of Zhejiang Province, Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou, China, [2] Faculty of Forestry, Nanjing Forestry University, Nanjing, China, [3] Agricultural Technology Extension Centre of Dongtai, Yancheng, China, [4] Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, China, [5] Institute of Horticulture, Zhejiang Academy of Agricultural Science, Hangzhou, China

OPEN ACCESS

Edited by:
Weihua Pan,
Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Youxiong Que,
Fujian Agriculture and Forestry University, China
Qibin Ma,
South China Agricultural University, China
Bobin Liu,
Yancheng Teachers University, China

***Correspondence:**
Chao Wu
semporna@126.com
Renyin Zhuo
zhuory@gmail.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: 21 January 2022
Accepted: 29 March 2022
Published: 27 April 2022

Citation:
Lu Z, Qiu W, Jin K, Yu M, Han X, He X, Wu L, Wu C and Zhuo R (2022) Identification and Analysis of bZIP Family Genes in Sedum plumbizincicola and Their Potential Roles in Response to Cadmium Stress. Front. Plant Sci. 13:859386. doi: 10.3389/fpls.2022.859386

Sedum plumbizincicola (Crassulaceae), a cadmium (Cd)/zinc (Zn)/lead (Pb) hyperaccumulator native to Southeast China, is potentially useful for the phytoremediation of heavy metal-contaminated soil. Basic leucine zipper (bZIP) transcription factors play vital roles in plant growth, development, and abiotic stress responses. However, there has been minimal research on the effects of Cd stress on the bZIP gene family in *S. plumbizincicola*. In this study, 92 *SpbZIP* genes were identified in the *S. plumbizincicola* genome and then classified into 12 subgroups according to their similarity to bZIP genes in *Arabidopsis*. Gene structure and conserved motif analyses showed that *SpbZIP* genes within the same subgroup shared similar intron–exon structures and motif compositions. In total, eight pairs of segmentally duplicated *SpbZIP* genes were identified, but there were no tandemly duplicated *SpbZIP* genes. Additionally, the duplicated *SpbZIP* genes were mainly under purifying selection pressure. Hormone-responsive, abiotic and biotic stress-responsive, and plant development-related *cis*-acting elements were detected in the *SpbZIP* promoter sequences. Expression profiles derived from RNA-seq and quantitative real-time PCR analyses indicated that the expression levels of most *SpbZIP* genes were upregulated under Cd stress conditions. Furthermore, a gene co-expression network analysis revealed that most edge genes regulated by hub genes were related to metal transport, responses to stimuli, and transcriptional regulation. Because its expression was significantly upregulated by Cd stress, the hub gene *SpbZIP60* was selected for a functional characterization to elucidate its role in the root response to Cd stress. In a transient gene expression analysis involving *Nicotiana benthamiana* leaves, *SpbZIP60* was localized in the nucleus. The overexpression of *SpbZIP60* enhanced

the Cd tolerance of transgenic *Arabidopsis* plants by inhibiting ROS accumulation, protecting the photosynthetic apparatus, and decreasing the Cd content. These findings may provide insights into the potential roles of the bZIP family genes during the *S. plumbizincicola* response to Cd stress.

Keywords: bZIP gene family, *Sedum plumbizincicola*, Cd stress, expression profiles, *SpbZIP60*

INTRODUCTION

Heavy metal pollution has become a global environmental problem (Ali et al., 2013). Cadmium (Cd) is a major heavy metal pollutant that is released into the environment because of human industrial and agricultural production activities (Tchounwou et al., 2012). Cd contamination leads to decreased soil quality and suppressed crop production. Heavy metal stress results in changes to various physiological and metabolic processes. For example, the expression of many genes is induced in plants under heavy metal stress conditions, and the upregulated expression of stress-responsive genes, which is usually mediated by transcription factors, may increase plant survival rates (Yao et al., 2018; Zhang et al., 2020; Xu et al., 2021b).

The basic leucine zipper (bZIP) family is one of the largest and most diverse transcription factor families in eukaryotes (Pérez-Rodríguez et al., 2010). These transcription factors contain a highly conserved bZIP domain with two different functional regions, one of which is a sequence-specific DNA-binding alkaline region (N-x7-R/K-x9), whereas the other is a leucine zipper consisting of several heptapeptide repeats comprising Leu or other large hydrophobic amino acids (e.g., Ile, Val, Phe, or Met) that influence dimerization specificity (Jakoby et al., 2002; Nijhawan et al., 2008). The bZIP gene family was first identified and classified in *Arabidopsis* at the genome-wide level (Jakoby et al., 2002). The current study is related to earlier research, during which 78 *AtbZIP* genes were identified and divided into 13 subclasses (A–M) (Dröge-Laser et al., 2018). Additionally, analyses of the bZIP gene family in diverse species resulted in the identification of 64 genes in cucumber (Corrêa et al., 2008), 85 genes in rice (Nijhawan et al., 2008), 86 genes in poplar (Zhao et al., 2021), 96 genes in buckwheat (Liu et al., 2019b), 112 genes in apple (Zhao et al., 2016), 125 genes in maize (Wei et al., 2012), and 160 genes in soybean (Zhang et al., 2018).

There is considerable evidence that bZIP transcription factors in plants play crucial roles in various biological processes, including seed maturation (Izawa et al., 1994), organ differentiation (Pautler et al., 2015), photomorphogenesis (Huang et al., 2012), and floral development (Abe et al., 2005; Muszynski et al., 2006). They also contribute to responses to various abiotic stresses, including salinity (Bi et al., 2021), drought (Wang et al., 2017; Tu et al., 2020), heat (Deng et al., 2011; Liu et al., 2012), osmotic stress (Xu et al., 2013), and oxidative stress (Choi et al., 2021). Most of these responses are abscisic acid (ABA) signal transduction-dependent processes (Banerjee and Roychoudhury, 2017). As a key member of the ABA signal transduction pathway, bZIP proteins are activated by kinases, such as SnRK2, and then bind to an ABA-responsive element (ABRE) to regulate the expression of downstream genes. In rice, *OsbZIP46* positively regulates ABA signal transduction and drought stress tolerance (Tang et al., 2012). The stress-induced expression of the activated form of *AtbZIP17* protects *Arabidopsis* from salt stress (Liu et al., 2008). In poplar, a loss-of-function mutation to *PtabZIP1* enhances lateral root formation under osmotic stress conditions (Dash et al., 2017). As the most dangerous pollutant, heavy metals have been regarded as new stress factors.

Similar to other abiotic stress responses, there has been increasing interest in the relationship between bZIP transcription factors and heavy metal stress responses. The *BjCdR15/TGA3* transcription factor gene encodes an important regulator of Cd uptake by roots and the subsequent long-distance root-to-shoot transport. The overexpression of this gene in *Arabidopsis* and tobacco enhances Cd tolerance and accumulation. This is related to the regulation of the synthesis of phytochelatin synthase and the expression of several metal transporter genes (Farinati et al., 2010). In *Arabidopsis*, ABI5 (ABA-Insensitive 5), which is a central ABA signaling molecule, represses Cd accumulation in plants by physically interacting with MYB49 and preventing it from binding to the downstream genes *bHLH38*, *bHLH101*, *HIPP22*, and *HIPP44*, resulting in the inactivation of *IRT1* and decreased Cd uptake (Zhang et al., 2019). The subgroup F bZIP transcription factors *AtbZIP19* and *AtbZIP23* are Zn sensors that regulate *Arabidopsis* responses to Zn deficiency *via* the binding between Zn^{2+} ions and their Zn sensor motif (Assunção et al., 2010; Lilay et al., 2019, 2021). Thus, the bZIP transcription factors appear to participate in plant responses to heavy metal stress.

Current research on the heavy metal homeostasis in plants primarily focuses on model plants or crop plants. Hyperaccumulator plants are valuable research materials because of their potential utility for remediating heavy metal-contaminated soil. Moreover, they are useful for investigating plant adaptation and evolution in extreme environments. The Cd, Pb, and Zn hyperaccumulator *Sedum plumbizincicola* (Wu et al., 2013), which is also known as the hyperaccumulating ecotype of *S. alfredii* (Yang et al., 2002), can tolerate, transport, and accumulate large amounts of Cd (Li et al., 2018), with a shoot Cd concentration as high as 9,000 mg/kg (Yang et al., 2004). Its efficient Cd absorption, transport, and detoxification systems are necessary for its growth in highly contaminated soils. Some genes related to Cd absorption, resistance, and hyperaccumulation, such as *SpHMA3* (Liu et al., 2017), *SpMTL* (Peng et al., 2017), *SaNramp6* (Chen et al., 2017; Lu et al., 2020), *SaCAX2* (Zhang et al., 2016), *SaHsfA4c* (Chen et al., 2020), *SaCAD* (Qiu et al., 2018), *SaREF* (Liu et al., 2016), and *SaPCR2* (Lin et al., 2020), have been characterized. However, there has yet to be a systematic analysis of the transcription factor families (e.g., bZIP) in *S. plumbizincicola* to clarify their roles in response to heavy metal stress.

In this study, we identified 92 bZIP genes in the *S. plumbizincicola* genome and then analyzed their structures, motifs, *cis*-acting elements, and phylogenetic relationships. On the basis of RNA sequencing (RNA-seq) and quantitative real-time PCR (qRT-PCR) methods, we explored their expression profiles in response to Cd stress. Furthermore, the bZIP60 function related to plant responses to Cd stress was investigated. The results of this study will be useful for the future functional characterization of the *SpbZIP* genes in terms of their roles during plant responses to Cd stress.

MATERIALS AND METHODS

Identification of the Basic Leucine Zipper Family Genes in *Sedum plumbizincicola*

To identify all members of the bZIP gene family in *S. plumbizincicola*, HMMER3.0 was used to screen for candidate proteins in the *S. plumbizincicola* genome database (unpublished work) on the basis of the Hidden Markov Model profile of the bZIP domain (PF00170).[1] A BLASTP search was performed using 78 *Arabidopsis* protein sequences that were annotated according to previously published methods from TAIR.[2] Subsequently, Pfam, SMART,[3] and CDD[4] were used to confirm the presence of the bZIP domain in candidate proteins. All putative bZIP genes were named according to their homologs in *Arabidopsis*. The encoded protein sequences were analyzed using the online tool ProtParam[5] to predict the amino acid composition, molecular weight, and isoelectric point (Gasteiger et al., 2005). Additionally, PSORT prediction[6] was used to predict the subcellular localization of the proteins.

Multiple Sequence Alignment and Phylogenetic Analysis

ClustalX2 was used to align the full-length SpbZIP and AtbZIP amino acid sequences. Phylogenetic trees were constructed using the maximum-likelihood criteria in MEGA 5.0, with 1,000 bootstrap replicates. The identified *SpbZIP* genes were divided into different groups according to the *AtbZIP* classification scheme. The phylogenetic tree was visualized using iTOL.[7]

Analysis of *cis*-Acting Elements in *SpbZIP* Promoters

The *cis*-acting elements in the promoter region 2 kb upstream of the *SpbZIP* genes were identified and then submitted to the PlantCARE database[8] (Lescot et al., 2002). The position of the identified elements was graphically displayed using the TBtools software.[9]

Analysis of *SpbZIP* Gene Structures and Encoded Motifs

The exon/intron structure of *SpbZIP* genes was analyzed and displayed using the GSDS platform.[10] The conserved motifs in the SpbZIP proteins were identified using the MEME program (version 5.0.5),[11] with the following parameters: optimum motif width range of 6–50 amino acid residues and a maximum of 22 motifs (Bailey and Elkan, 1994).

Synteny Analysis and Chromosomal Distribution of *SpbZIP* Genes

The default parameters of the Multiple Collinearity Scan (MCScanX) toolkit were used to analyze gene duplication events (Wang et al., 2012). Diagrams were generated using the Circos program (version 0.69)[12] (Krzywinski et al., 2009). Non-synonymous (ka) and synonymous (ks) substitutions in each duplicated *SpbZIP* gene were calculated using KaKs_Calculator 2.0 (Wang et al., 2010).

Plant Materials and Cd Stress Treatments

Sedum plumbizincicola plants were collected from an old Pb/Zn mine in Huiping town, Quzhou city, Zhejiang province, China. The shoots from a single genotype were asexually propagated and cultivated in water in an artificial climate chamber at 25°C with a 16-h light/8-h dark cycle. The plants were grown in a half-strength Hoagland solution for about 4 weeks. Similarly growing plants were then treated with 400 μM $CdCl_2$. The roots, stems, and leaves were sampled at 0, 0.5, 2, 4, 8, and 12 h after the Cd stress treatment. Three biological replicates were collected for all samples.

SpbZIP Expression Profiles in Response to Cd Stress

The Total RNA Purification kit (NORGEN, Thorold, ON, Canada) was used to extract total RNA from the roots, stems, and leaves. First-strand cDNA was generated using PrimeScript™ RT Master Mix (TaKaRa, Dalian, China). The qRT-PCR analysis was performed in triplicate using the 7,300 Real-Time PCR System (Applied Biosystems, CA, United States) and the SYBR® Premix Ex Taq™ reagent (TaKaRa, Dalian, China). Gene-specific primers were designed using the "Genes" module of the SPDE software (Xu et al., 2021a). The UBC gene was selected as the internal reference (Sang et al., 2013). Primers used are listed in **Supplementary Table 5**. Relative expression levels were calculated according to the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001). The FPKM values for the *SpbZIP* genes were derived from the RNA-seq data (Han et al., 2016). Expression

[1] http://pfam.xfam.org/
[2] http://www.arabidopsis.org/
[3] http://smart.embl-heidelberg.de/
[4] https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
[5] https://web.expasy.org/protparam/
[6] http://psort1.hgc.jp/form.html
[7] https://itol.embl.de/
[8] http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
[9] https://github.com/CJ-Chen/TBtools
[10] http://gsds.cbi.pku.edu.cn/
[11] http://meme-suite.org/tools/meme
[12] http://circos.ca/

values were normalized *via Z*-score normalization. An expression profile heatmap was generated using the pheatmap package in R (4.0.2).

SpbZIP Co-expression Regulatory Network

The weighted gene co-expression network analysis (WGCNA) R package was used to construct a co-expression regulatory network on the basis of the expression profiles of differentially expressed genes under Cd stress conditions (Han et al., 2016). The *SpbZIP* genes among the co-expressed genes with strong interconnections were designated as hub genes. The Pearson's correlation coefficient threshold was set as 0.40 according to the FPKM values for each gene pair using the R (version 4.0.2) program (Han et al., 2016). We screened for co-expression edge genes associated with the *SpbZIP* hub genes and performed Gene Ontology (GO) analyses using the Gene Annotation Software for Plants (GFAP) (Xu et al., 2022). Subsequently, we classified the related genes according to their functions and visualized the relationships between nodes and edges using Cytoscape (version 3.6.1).

Subcellular Localization of *SpbZIP60*

The *SpbZIP60* coding sequence without the stop codon was fused to the mGFP-encoding sequence in the pCAMBIA1302 expression vector using the ClonExpress II One Step Cloning Kit (Vazyme, Nanjing, China). *Agrobacterium tumefaciens* GV3101 cells were transformed with the recombinant plasmid, which was then transferred into healthy *Nicotiana benthamiana* leaves for a transient gene expression analysis; the empty vector was used as a control. After co-culturing for 3 days, the leaves were soaked in a 4,6-diamidino-2-phenylindole (DAPI) staining solution to visualize nuclear DNA. The LSM 710 confocal laser-scanning microscope (Zeiss, Germany) was used to detect the fluorescence of the fusion protein.

Ectopic Expression of *SpbZIP60* in *Arabidopsis* and Cd Treatment

The *SpbZIP60* coding sequence was amplified by PCR and inserted into the pCAMBIA1300 vector. The recombinant plasmid was inserted into *Arabidopsis* (Col-0) plants *via A. tumefaciens* (EHA105)-mediated transformation (Zhang et al., 2006). The T_3 homozygous transgenic lines and wild-type (WT) plants were grown in a half-strength Hoagland solution. The seedlings were transferred to a solution containing 30 μM $CdCl_2$ after 4 weeks and grown for 7 days. The roots of the treated seedlings were immersed in a 10-mM EDTA solution for 0.5 h to remove Cd from the surface. The samples were dried and then digested with a solution comprising HNO_3 and perchloric acid (9:1 v/v) at 120–200°C in a microwave-accelerated reaction system (CEM, Matthews, NC, United States). The Cd content was determined using the 7500a inductively coupled plasma mass spectrometry system (Agilent, Santa Clara, CA, United States). Previously described 3,3'-diaminobenzidine (DAB) and nitroblue tetrazolium (NBT) staining methods were used to reveal the presence of H_2O_2 and O_2^- *in situ* (Chen et al., 2020). The chlorophyll content was measured according to an acetone ethanol extraction method (Li et al., 2000). Chlorophyll fluorescence was analyzed using the Dual-PAM-100 system (Walz, Effeltrich, Germany); the parameters were set, and the data were analyzed as previously described (Su et al., 2020).

RESULTS

Identification and Characterization of Putative Basic Leucine Zipper Transcription Factors

Following a search of the *S. plumbizincicola* genome database using HMMER3.0, the identified candidate sequences were examined using CDD, Pfam, and SMART to confirm the presence of the bZIP domain (*E*-value $< 1e^{-5}$). A total of 92 non-redundant genes were identified as bZIP genes in the *S. plumbizincicola* genome. They were named according to the corresponding *Arabidopsis* homologs. The subsequent analysis indicated that the SpbZIP proteins comprise 117–707 amino acids (average of 303 amino acids), with a molecular weight of 13.7–77.3 kDa (average of 33.7 kDa) and a predicted isoelectric point of 5.05–10.26 (average of 7.02). Most of the identified SpbZIP proteins were predicted to localize in the nucleus, which is a characteristic of transcription factors (**Supplementary Table 1**).

Phylogenetic Analysis of *SpbZIP* Genes

To classify the *SpbZIP* genes into subgroups and elucidate the evolutionary relationships between *S. plumbizincicola* and *Arabidopsis* genes, we constructed an unrooted phylogenetic tree using the maximum-likelihood method and the protein sequences encoded by 78 *AtbZIP* genes and the 92 identified *SpbZIP* genes (**Figure 1**). On the basis of the phylogenetic tree, the *SpbZIP* genes were divided into 12 of 13 subgroups; the exception was subgroup M. There were no individual clades among the *SpbZIP* genes, suggesting that they were relatively conserved. Similar to the *Arabidopsis* homologs, most of the *SpbZIP* genes were classified into subgroups S and A. Subgroups J and K had the fewest genes, each with only two *SpbZIP* genes.

SpbZIP Gene Structure and Protein Motif Composition

To gain insights into the structures of *SpbZIP* genes, their introns and exons were analyzed. Of the *SpbZIP* genes in subgroup S, 20 (21.7%) lacked introns. In contrast, three (3.3%) and seven (7.6%) genes contained one and two introns, respectively. Three or more introns were detected in 62 genes (68.5%) (**Figure 2C**). An examination using the MEME online program detected 22 conserved motifs in the SpbZIP proteins. The conserved motifs comprised 20–50 amino acids. Details regarding the 22 putative motifs are provided in **Supplementary Table 2**. Motif 1 (leucine zipper region of bZIP) was identified as the core conserved domain. A few subgroup-specific motifs were identified, including motifs 10 and 15 (subgroup A) and motifs 11, 12, and 16 (subgroup G). Most of the SpbZIP proteins in the same subgroup in the phylogenetic tree had common motifs,

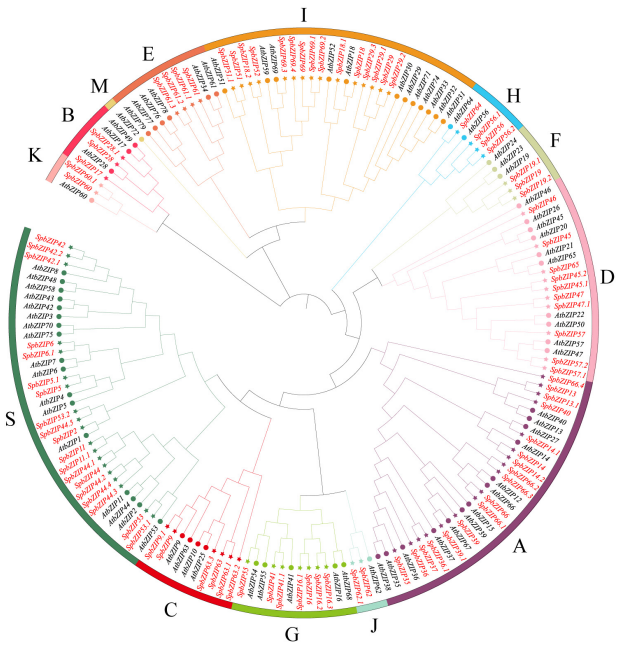


FIGURE 1 | Phylogenetic relationships among the bZIP genes from *Sedum plumbizincicola* and *Arabidopsis*. The phylogenetic tree was constructed on the basis of the alignment of *S. plumbizincicola* and *Arabidopsis* bZIP proteins according to the maximum-likelihood method, with 1,000 bootstrap replicates.

indicating a close evolutionary relationship and a high degree of conservation.

Chromosomal Locations and Collinearity Analysis of *SpbZIP* Genes

The 92 *SpbZIP* genes were distributed unequally among 30 *S. plumbizincicola* chromosomes (**Figure 3**). Segmental duplications of multiple genes are caused by chromosomal rearrangements (Yu et al., 2005), whereas tandem duplications, which mainly occur in the recombination region of chromosomes, usually result in the formation of a cluster of genes with similar sequences and functions (Ramamoorthy et al., 2008). During evolution, segmental and tandem duplications are the two main drivers of the expansion of plant gene families. In the *S. plumbizincicola* genome, eight segmental duplication events involving 16 *SpbZIP* genes (i.e., 17.4% of the *SpbZIP* genes) were detected. Among the segmentally duplicated gene pairs, *SpbZIP42.1*/*SpbZIP42* and *SpbZIP45.2*/*SpbZIP45.1* were distributed on chromosomes 4 and 14, respectively, whereas *SpbZIP60*/*SpbZIP60.1* and *SpbZIP61.3*/*SpbZIP61.2* were distributed on chromosomes 5 and 6, respectively. Additionally, *SpbZIP36.1*/*SpbZIP36*, *SpbZIP53*/*SpbZIP53.1*, *SpbZIP44*/*SpbZIP44.1*, and *SpbZIP52*/*SpbZIP18.2* resulted from gene duplication events. Of these gene pairs, six were assigned to subgroup S. Furthermore, none of the genes were the result of tandem duplications. Thus, we speculated that segmental duplications were important for the expansion of the *SpbZIP* family in *S. plumbizincicola*. Moreover, the Ka/Ks ratios for all eight duplicated *SpbZIP* gene pairs were less than 0.5 (**Supplementary Table 3**), indicating that the *SpbZIP* family paralogs were primarily under purifying selection.

Next, we created two comparative syntenic maps of the association between *S. plumbizincicola* and *Arabidopsis* or *Kalanchoe fedtschenkoi*, which is a representative Crassulaceae plant species, to further clarify the origin and evolution of the *S. plumbizincicola* bZIP family (**Figure 4**). A total of 15 *SpbZIP* genes had a syntenic relationship with 17 and 48 genes in *Arabidopsis* and *K. fedtschenkoi*, respectively (**Supplementary Table 4**). Additionally, 20 orthologous gene pairs were detected between *S. plumbizincicola* and *Arabidopsis*, which was fewer than the 54 orthologous gene pairs between *S. plumbizincicola* and *K. fedtschenkoi*. There were more collinear gene pairs between *S. plumbizincicola* and *K. fedtschenkoi* than between *S. plumbizincicola* and *Arabidopsis*, which is in accordance with the fact *S. plumbizincicola* is phylogenetically closer to *K. fedtschenkoi* than to *Arabidopsis*. Some collinear gene pairs (involving 11 *SpbZIP* genes) among all three species were identified, implying that the orthologous gene pairs may have existed before ancestral divergence. These orthologous genes were also under intense purifying selection.

Analysis of *cis*-Acting Elements in *SpbZIP* Promoters

To clarify the regulatory mechanisms underlying *SpbZIP* expression, the *cis*-acting elements in the promoter sequences were analyzed using PlantCARE. The identified *cis*-acting elements (**Figure 5**) were divided into three categories (stress-responsive, plant development-related, and phytohormone responsive). The following seven abiotic stress-responsive elements were detected: ARE (important for anaerobic induction), MBS (MYB-binding site associated with drought-inducible expression), TC-rich repeat (stress-responsive element), WUN-motif (wound-responsive element), LTR (low temperature-responsive element), G-box, and W-box. At least one stress-responsive *cis*-acting element was detected in the promoter of all *SpbZIP* genes, with the exception of *SpbZIP66*, reflecting the importance of *SpbZIP* expression for plant responses to various abiotic stresses. Among the phytohormone-responsive *cis*-acting elements, ABRE was the most common, with 251 ABREs detected in 72 *SpbZIP* promoters (enrichment level of 3.49), followed by MeJA-responsive *cis*-acting elements (TGACG-motif and CGTCA-motif) (enrichment level of 2.49).

SpbZIP Expression Profiles Under Cd Stress Conditions

We used our previously published RNA-seq data to determine *SpbZIP* expression patterns (Han et al., 2016), which were revealed in terms of FPKM values, in the roots, stems, and leaves. The *SpbZIP* expression trends in the roots during the Cd treatment period were divided into four categories (**Figure 6**). The expression levels of 32 *SpbZIP* genes gradually decreased or increased over the entire treatment period. In contrast, the expression levels of 18 genes peaked at 1 day after initiating the Cd treatment. However, the genes whose expression in the roots

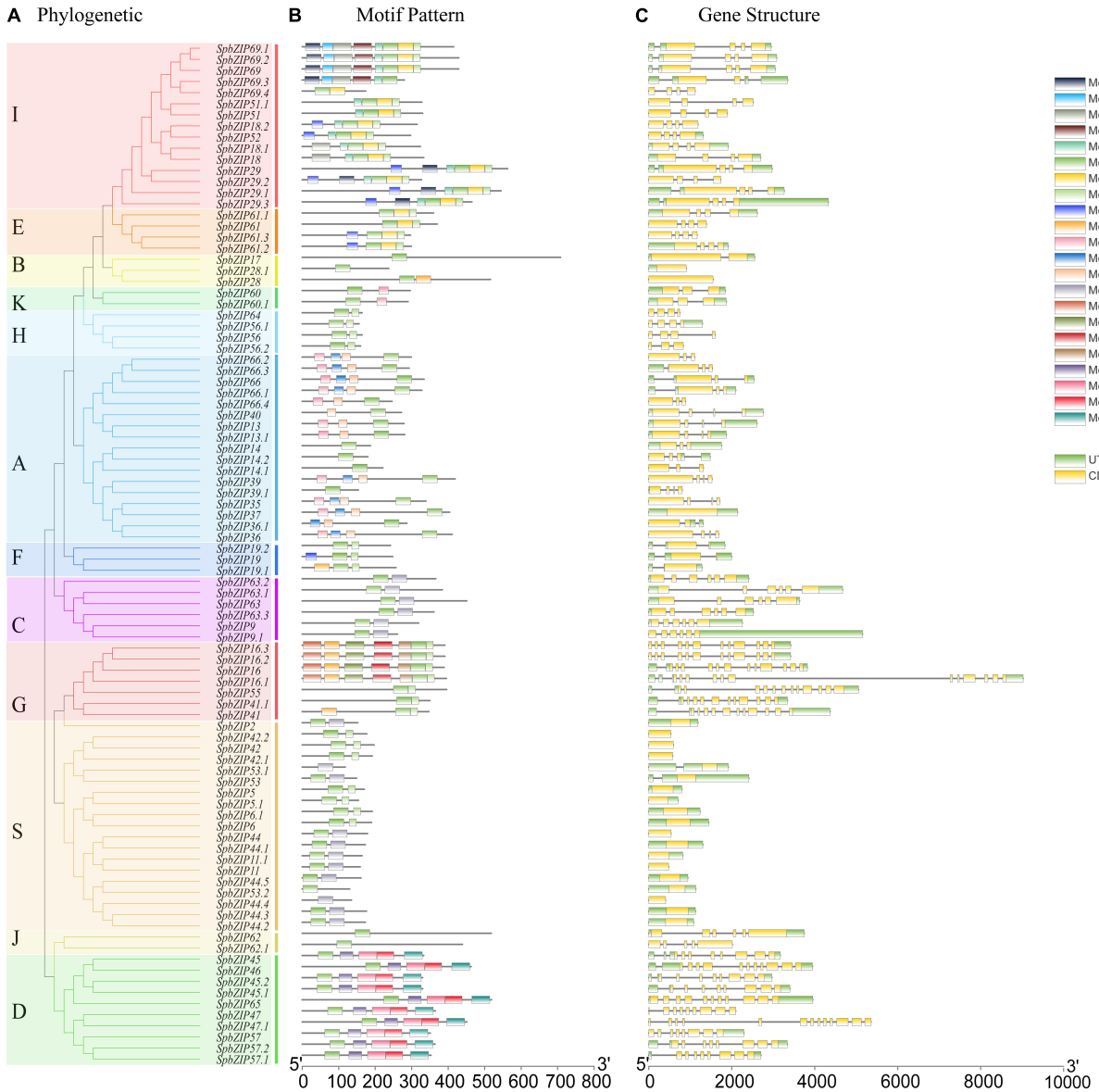


FIGURE 2 | Phylogenetic relationships, motif compositions, and gene structures of *SpbZIP* genes in *S. plumbizincicola*. **(A)** Phylogenetic analysis of *S. plumbizincicola* bZIP family members. **(B)** All conserved motifs in the SpbZIP proteins were identified using the MEME program. Different motifs are highlighted with different colored boxes (numbered 1–22). **(C)** Gene structures. Exons and 5'/3' untranslated regions are indicated by green and yellow bars, respectively, whereas gray lines represent introns.

was not induced by Cd stress had upregulated or downregulated expression levels in the stems (27) or leaves (8) in response to the Cd treatment. These results suggested that SpbZIP transcription factors may play a major role in the roots as part of the initial response to Cd stress. Transcription factors often rapidly respond to environmental cues. We further shortened and refined the treatment time and then performed qRT-PCR analysis to investigate the expression levels of 25 hub genes selected from the co-expression network. As expected, for most of the *SpbZIP* genes, the expression levels peaked earlier in the roots (4 h) than in the stems (8 h) and leaves (12 h) (**Figure 7**).

SpbZIP Co-expression Network

To further clarify the regulatory effects of bZIP family members on the expression of Cd-responsive genes, a co-expression regulatory network was constructed on the basis of the expression profiles of differentially expressed genes under Cd stress conditions determined in an earlier transcriptome analysis, in which 11 *SpbZIP* genes were annotated as hub genes. The nodes associated with hub genes were clustered according to functional categories, which reflected their association with metabolic processes, cellular activities, membranes, cells, binding, and catalytic activities (**Supplementary Table 6**).

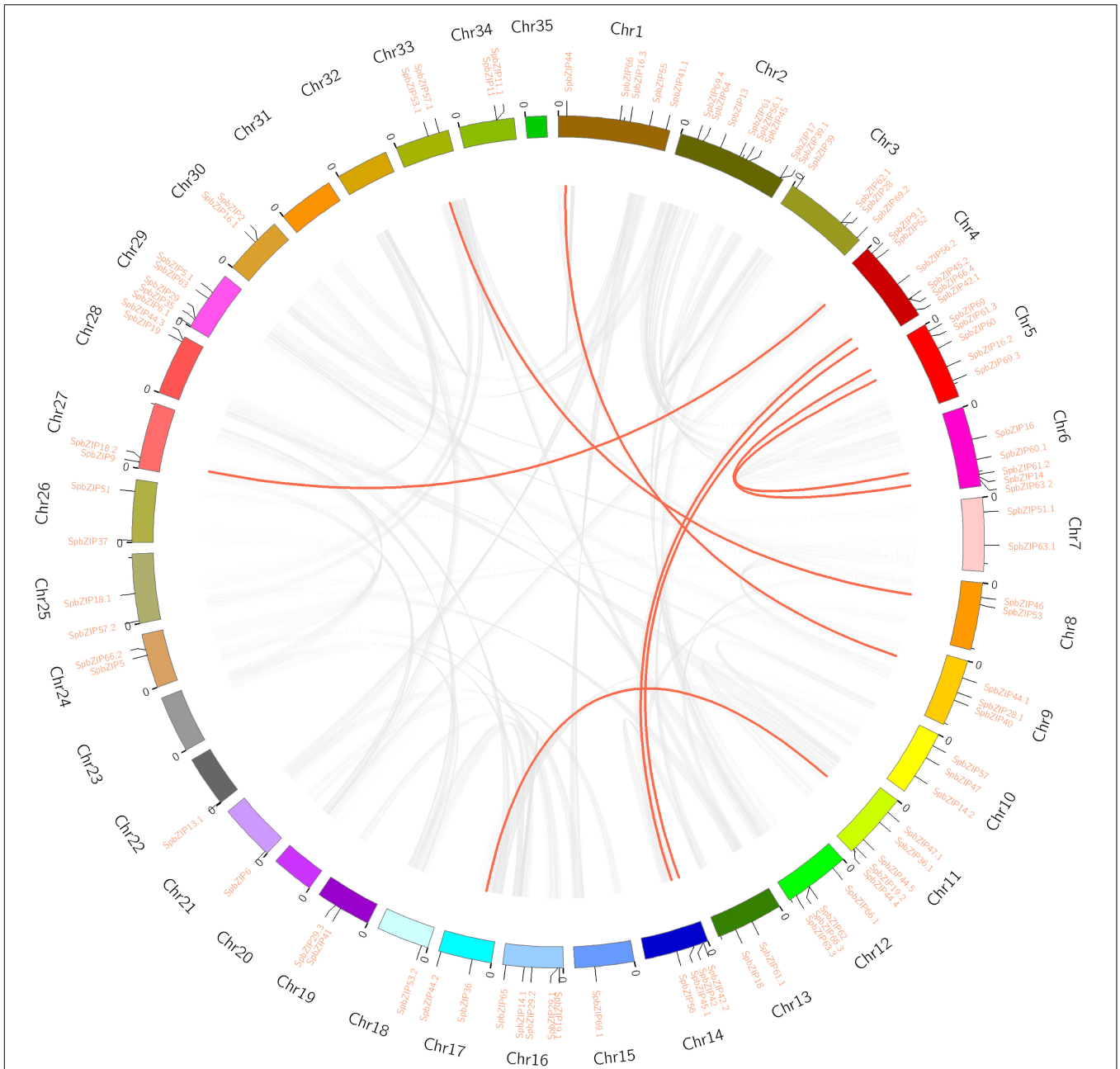


FIGURE 3 | Genome location and synteny of bZIP genes in *S. plumbizincicola*. Gray lines indicate syntenic blocks in the *S. plumbizincicola* genome, whereas the red lines between chromosomes indicate segmentally duplicated gene pairs.

The Cd-responsive gene co-expression network had 189 nodes (**Figure 8**). The major categories included transcription factor (59 edges), transporter activity (52 edges), stimulus-response (43 edges), signaling (19 edges), and antioxidant activity (8 edges). The hub gene *SpbZIP60.1* was associated with the most nodes (59), including 19 transcription factor nodes, 12 transporter activity nodes, 4 stimulus-response nodes, and 4 signaling nodes, followed by *SpbZIP69.2* (34 nodes) and *SpbZIP63.3* (21 nodes). Accordingly, in response to Cd stress, SpbZIP transcription factors appear to regulate the expression of downstream genes associated with diverse functions.

SpbZIP60 Was Localized in the Nucleus

In this study, *SpbZIP60* was one of the hub genes in the co-expression regulatory network, and its expression level was significantly upregulated in the roots during the Cd stress treatment. Hence, the subcellular localization of SpbZIP60 was analyzed to elucidate the potential functions of bZIP transcription factors in *S. plumbizincicola*. The control

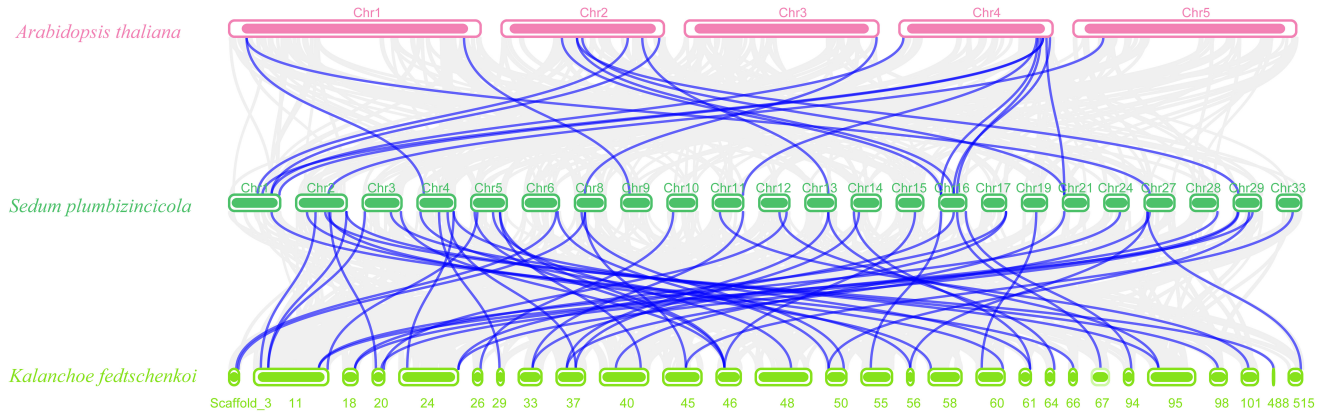


FIGURE 4 | Synteny between *SpbZIP* genes and genes in other species (*Arabidopsis* and *K. fedtschenkoi*). Gray lines in the background represent collinear blocks in *S. plumbizincicola* and the other species, whereas blue lines indicate syntenic bZIP gene pairs.

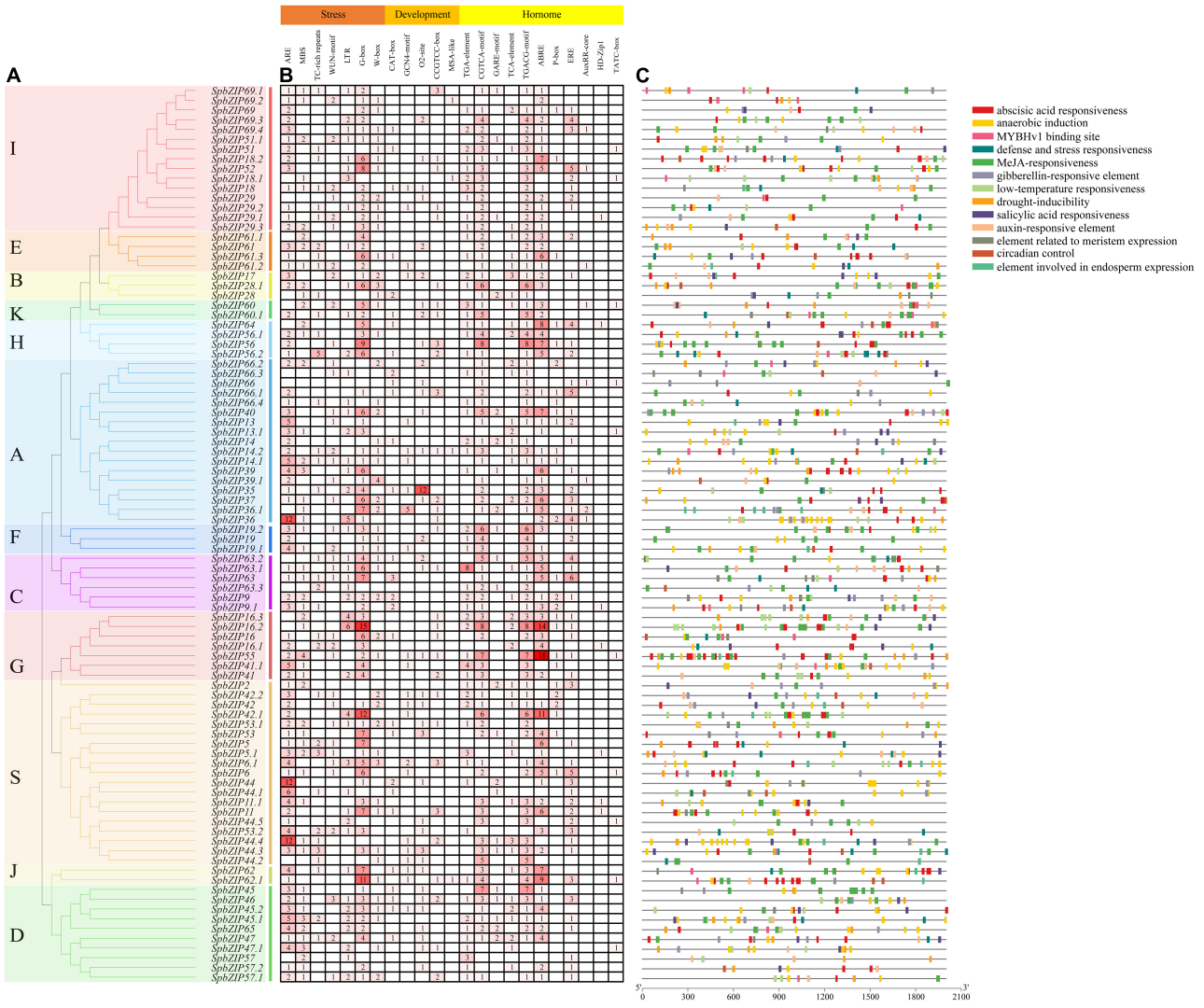


FIGURE 5 | Analysis of *cis*-acting elements in the *SpbZIP* promoter region. **(A)** Phylogenetic analysis of *SpbZIP* genes. **(B)** The number of each *cis*-acting element in the promoter region (2 kb upstream of the translation start site) of *SpbZIP* genes. **(C)** Distribution of related *cis*-acting elements in *SpbZIP* promoters.

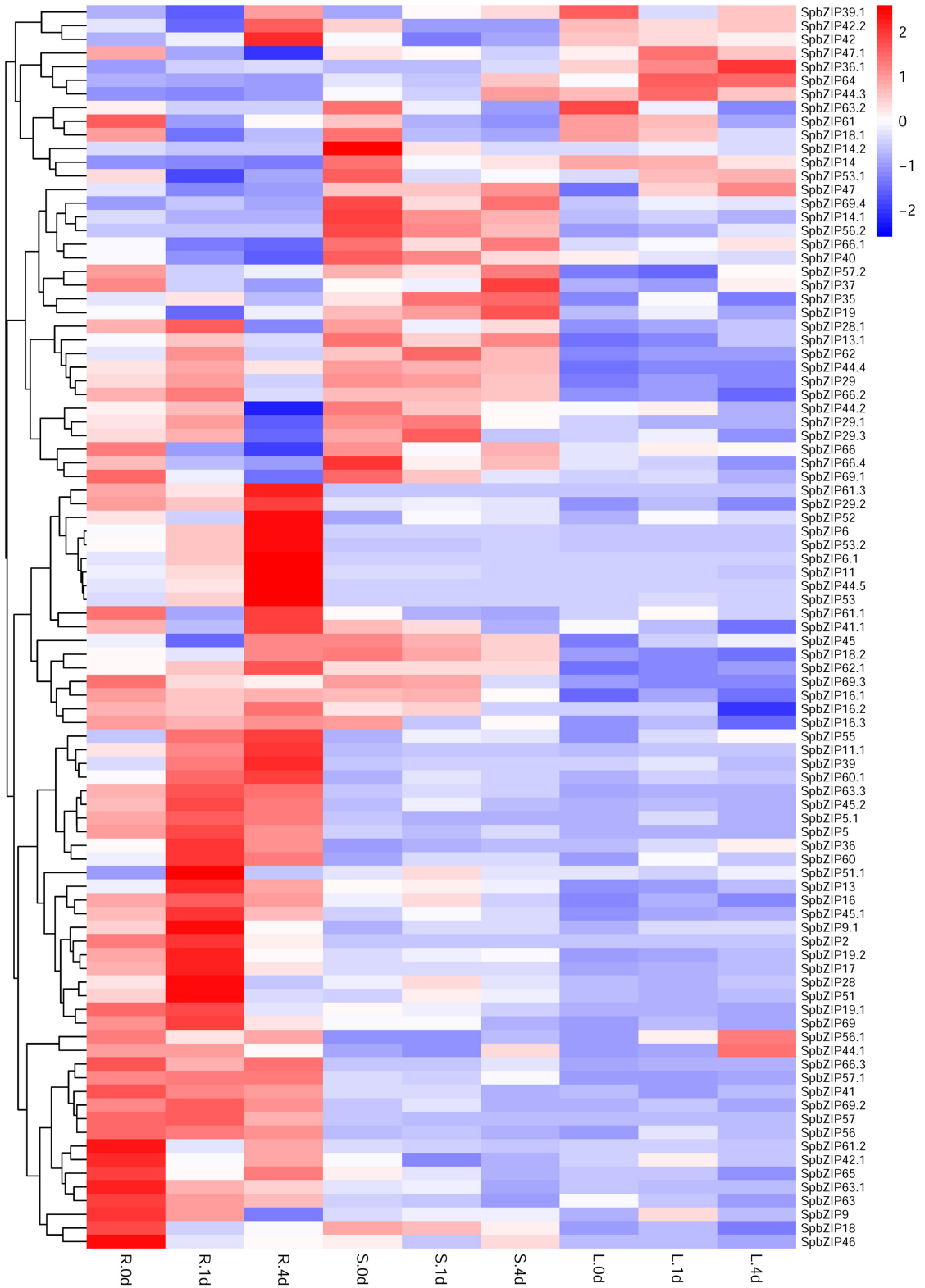


FIGURE 6 | Expression profiles of *SpbZIP* genes in plant tissues under Cd stress conditions. Gene expression data at 0, 1, and 4 days after the 400 μM $CdCl_2$ treatment were retrieved from an RNA-seq database and visualized using R (version 4.0.2). Expression levels are indicated by a gradient from low (blue) to high (red). L, S, and R represent leaves, stems, and roots, respectively.

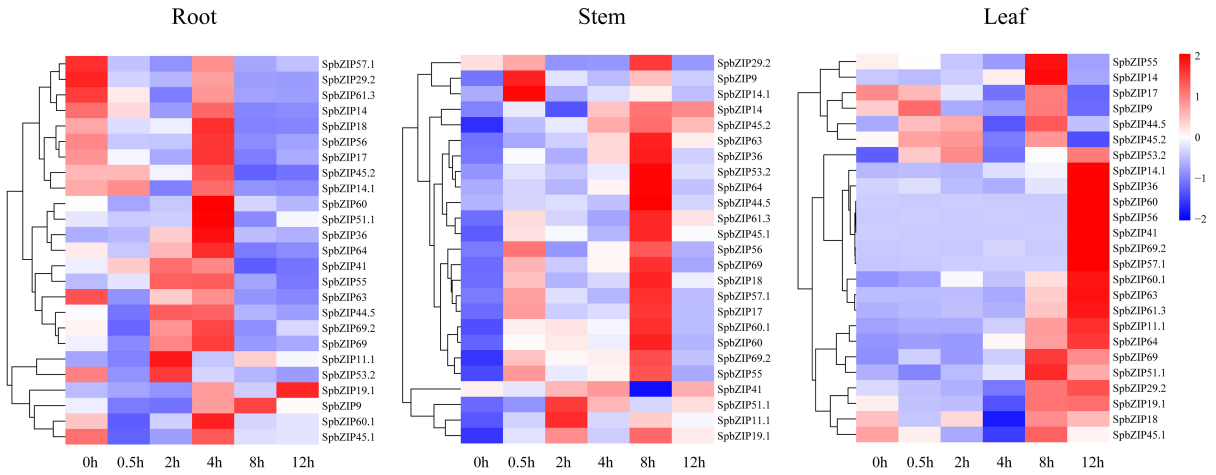


FIGURE 7 | Expression profiles of hub *SpbZIP* genes in different tissues of *S. plumbizincicola* soon after the exposure to Cd stress.

GFP signal was distributed throughout the cell, whereas the fluorescence of the SpbZIP60-mGFP fusion protein was detected only in the nucleus (**Figure 9**). Thus, SpbZIP60 likely functions as a nuclear protein that regulates transcription.

Overexpression of *SpbZIP60* Enhanced the Cd Tolerance of *Arabidopsis*

To further explore the function of SpbZIP60 under Cd stress conditions, transgenic *Arabidopsis* plants overexpressing *SpbZIP60* were generated. The T_0 transgenic lines were verified by PCR using genomic DNA as the template. After analyzing the *SpbZIP60* expression levels by semi-RT-PCR, the transgenic lines were cultivated to produce the homozygous T_3 lines used for the subsequent analyses (**Supplementary Figure 1**).

Leaf chlorosis and damages to the photosynthetic apparatus are observable symptoms of Cd toxicity. The degree of chlorosis in leaves at 7 days after initiating the Cd stress treatment was higher in the WT plants than in the *SpbZIP60*-overexpressing plants (**Figure 10A**). Histochemical staining revealed that less H_2O_2 and O_2^- accumulated in the transgenic *Arabidopsis* lines (OE#5 and OE#8) than in the WT control following the Cd treatment (**Figures 10B,C**). Meanwhile, the total chlorophyll content of the *SpbZIP60*-overexpressing plants where significantly higher than those of WT (**Figure 10D**). Chlorophyll fluorescence properties, which reflect the photochemical processes of PSII, are a useful indicator of the effects of heavy metal stress, especially Cd stress, on the photosynthetic apparatus. In the WT *Arabidopsis* plants, the Fv/Fm decreased, which was indicative of photoinhibition. Moreover, the inactivation or destruction of PSII resulted in an increase in the initial fluorescence (F_0). Additionally, the relative PSII electron transport rate was higher in the *SpbZIP*-overexpressing plants than in the WT plants (**Supplementary Table 7**). These results suggested that in response to Cd stress, the photosynthetic apparatus was damaged less in the *SpbZIP60*-overexpressing plants than in the WT plants. Next, we analyzed the Cd concentrations in hydroponically grown *SpbZIP60*-overexpressing lines. The Cd concentrations in the leaves and roots decreased substantially in the transgenic lines (**Figure 10E**). Therefore, SpbZIP60 significantly decreased the Cd concentration in the roots of the transgenic *Arabidopsis* plants, likely by inhibiting Cd uptake.

DISCUSSION

Sedum plumbizincicola has undergone long-term evolution and natural selection in heavy metal-contaminated soil (Wu et al., 2013; Yang et al., 2017). The *S. plumbizincicola* proteins involved in the absorption, transport, sequestration, and detoxification of heavy metals have been thoroughly studied, especially the heavy metal transporters (Liu et al., 2016, 2017, 2019a; Peng et al., 2017; Chen et al., 2020; Zhu et al., 2022). However, systematic analyses of the transcriptional regulation of the genes encoding these proteins have not been conducted. Transcription factors in the bZIP family modulate various physiological processes and abiotic stress responses (Corrêa et al., 2008). Thus, characterizing the *S. plumbizincicola* bZIP family is critical for clarifying the mechanism underlying the responses of *S. plumbizincicola* plants to environmental factors, especially heavy metal stress.

In this study, we conducted a genome-wide analysis of the *S. plumbizincicola* bZIP transcription factor family and explored the potential functions in response to Cd stress. On the basis of the encoded motifs, 92 *SpbZIP* genes were identified in the *S. plumbizincicola* genome. The number of bZIP genes in *S. plumbizincicola* is higher than that in some plant species but lower than that in other plant species (Corrêa et al., 2008; Nijhawan et al., 2008; Wei et al., 2012; Zhao et al., 2016, 2021; Dröge-Laser et al., 2018; Zhang et al., 2018; Liu et al., 2019b). We then divided the 92 *SpbZIP* genes into 12 subgroups after comparing the encoded protein sequences with the corresponding sequences in *Arabidopsis*. The classification of the bZIP genes was relatively consistent between *S. plumbizincicola* and *Arabidopsis*. However, *AtbZIP72*

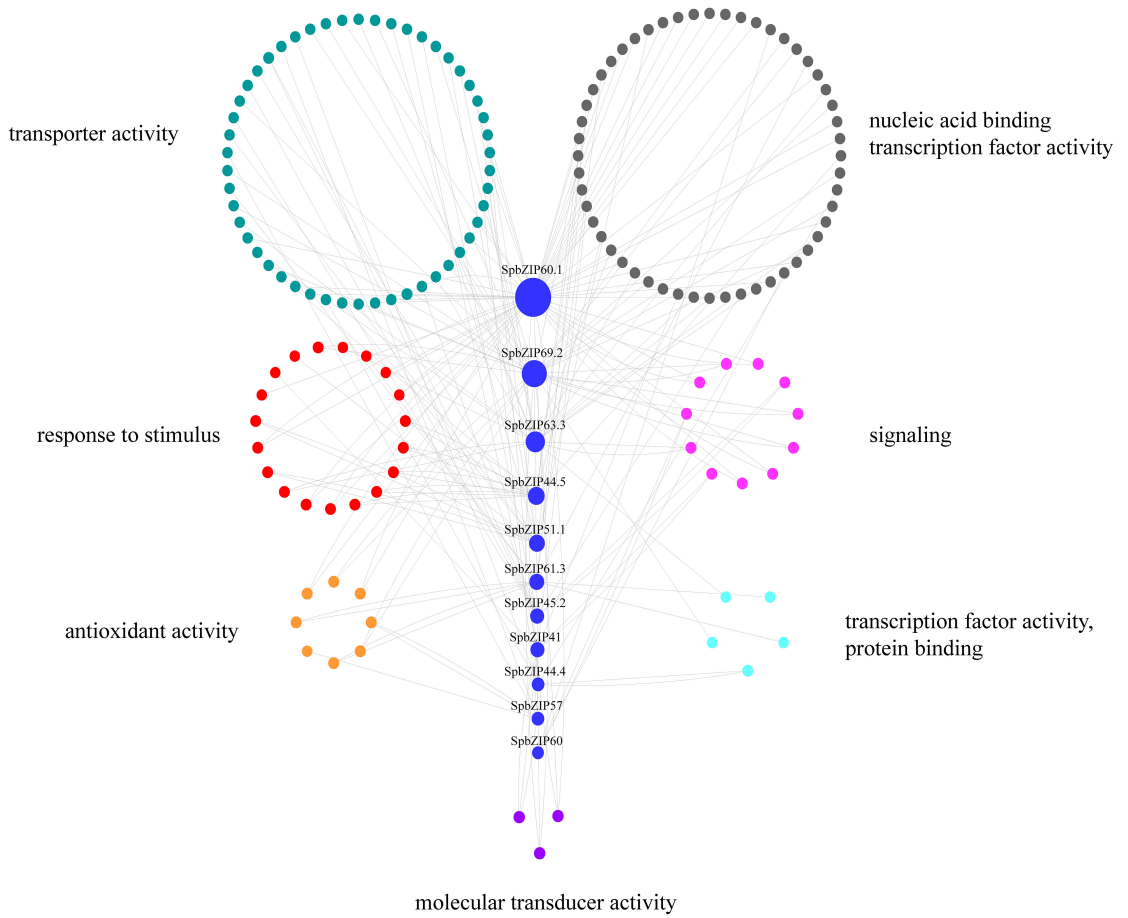


FIGURE 8 | *SpbZIP* gene co-expression network. The genes are divided on the basis of the following seven GO terms, which are represented by different colors: transporter activity, nucleic acid binding transcription factor activity, response to stimulus, signaling, antioxidant activity, protein binding transcription factor activity, and molecular transducer activity.

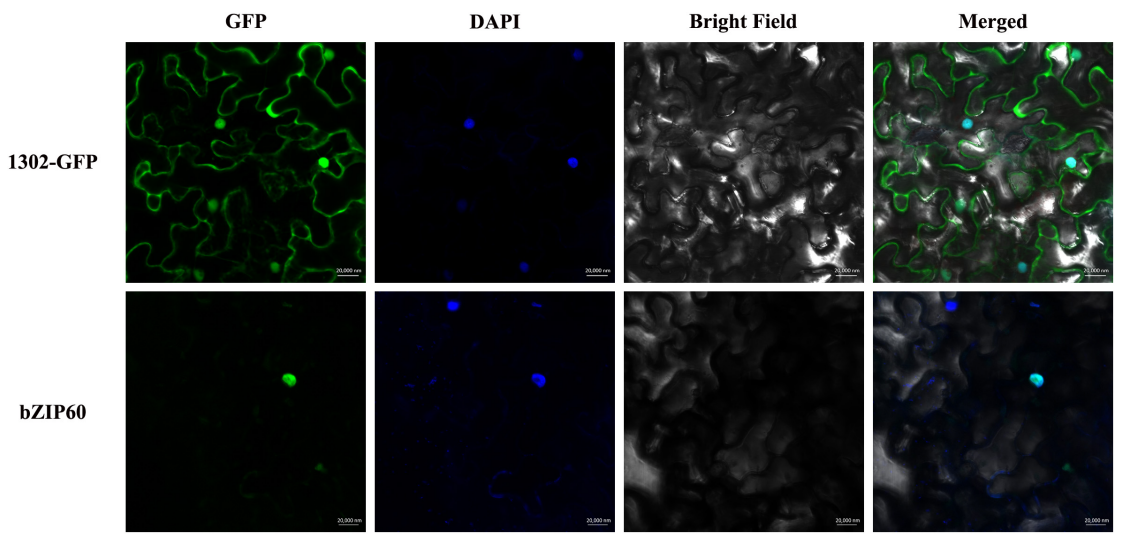


FIGURE 9 | Subcellular localization of SpbZIP60. The SpbZIP60-GFP fusion construct and the GFP gene driven by the CaMV 35S promoter were transiently expressed in tobacco. The nucleus was visualized using the DAPI staining solution.

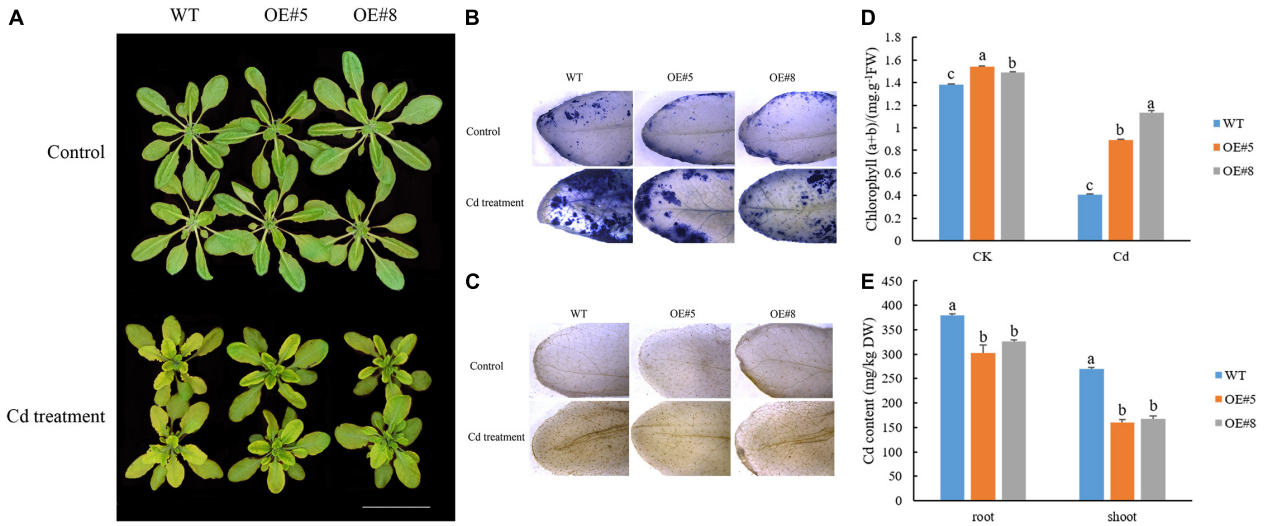


FIGURE 10 | Effects of Cd stress treatments on the growth of *SpbZIP60*-overexpressing *Arabidopsis* plants. **(A)** Phenotypes of *SpbZIP60*-overexpressing transgenic lines and wild-type (WT) plants under normal conditions or in response to the Cd treatment. Bar = 5 cm. **(B)** NBT staining results. **(C)** DAB staining results. **(D)** Chlorophyll contents of the WT and transgenic lines before and after the Cd treatment. **(E)** Cd contents of the WT and transgenic lines. Bars represent the mean ± standard deviation (SD) of at least three independent biological replicates. Significant differences according to a one-way analysis of variance are denoted as follows: $p < 0.05$ (Duncan's test).

was included in a separate clade (subgroup M), which lacked *SpbZIP* genes, suggesting that this clade is specific to *Arabidopsis*. In the phylogenetic tree constructed in this study, there were no branches that were exclusive to *S. plumbizincicola*, suggesting the *SpbZIP* genes are evolutionarily conserved (**Figure 1**). Moreover, genes belonging to the same subgroup were revealed to share similar gene structures and encode common motifs (**Figure 2**). For example, subgroup S consisted of small proteins encoded by genes lacking introns, which is in accordance with the results of earlier studies (Dröge-Laser et al., 2018; Wang et al., 2021).

Tandem and segmental duplication events are crucial for the expansion of gene families and the diversification of gene functions, which have enabled plants to adapt to environmental conditions (Cannon et al., 2004). We detected eight pairs of segmentally duplicated genes on 11 chromosomes, but no tandemly duplicated genes. Therefore, the expansion of the bZIP gene family in *S. plumbizincicola* was mainly the result of segmental duplications. The calculated Ka/Ks ratios for all gene pairs were less than 0.5, implying these genes might have experienced strong purifying selection pressure during evolution. Furthermore, we analyzed the collinearity between the *SpbZIP* genes and genes in *Arabidopsis* and *K. fedtschenkoi*. There were more collinear gene pairs between *S. plumbizincicola* and *K. fedtschenkoi*, which has a relatively close evolutionary relationship with *S. plumbizincicola*, than between *S. plumbizincicola* and *Arabidopsis*. A comparison between *S. plumbizincicola* and *Arabidopsis* detected 20 orthologous pairs of bZIP genes. As putative orthologs of *SpbZIP19.1*, both AT4G35040.1 (*AtbZIP19*) and AT2G16770.1 (*AtbZIP23*), which belong to subgroup F, encode Zn sensors that contain a motif that binds Zn^{2+} ions, enabling them to regulate plant responses to zinc deficiency (Lilay et al., 2021). Additionally, the following four G-box-binding factors (GBFs) were identified: GBF1 (SpbZIP41.1/AT4G36730.1), GBF2 (SpbZIP55/AT4G01120.1), GBF3 (SpbZIP55/AT2G46270.1), and GBF6 (SpbZIP16.3 and SpbZIP44.4/AT4G34590.1). Previous research indicated that GBFs participate in abiotic stress responses (Sun et al., 2015). For example, the expression of *AtGBF3* induces drought and pathogen stress tolerance by activating ABA-mediated signaling (Ramegowda et al., 2017; Dixit et al., 2019). Interestingly, the promoter of *SpbZIP55*, which is orthologous to *AtGBF3*, was revealed to contain the most ABREs among the examined *SpbZIP* genes, suggesting that *SpbZIP55* may also be related to ABA signaling and stress responses.

We further explored the *SpbZIP* expression patterns in response to Cd stress. Most of the *SpbZIP* genes were responsive to Cd stress, especially in the roots. This finding may be related to the fact that plants first perceive Cd stress in the roots, which take up Cd from the soil. The Cd is then transported to the stems and leaves. Therefore, the response to Cd stress will likely be greater in the roots than in the other plant tissues (Pan et al., 2019). Transcription factors may regulate metal ion transport in the stem. For example, in *Brassica juncea*, BjCdR15/TGA3 is a transcription factor that is crucial for the regulation of Cd uptake by the roots and the root-to-shoot transport of Cd (Farinati et al., 2010). Moreover, bZIP genes encode transcription factors that respond rapidly to stimuli. A co-expression regulatory network analysis is useful for identifying closely co-regulated and functionally related genes or genes affecting the same signaling pathway or physiological process. To identify the core *SpbZIP* genes responsive to Cd stress, we constructed a co-expression network and identified 11 hub *SpbZIP* genes that are co-expressed, with strong interconnections to edges (Han et al., 2016). These genes may encode proteins that

sense specific signals, respond to stimuli, regulate the expression of other transcription factor genes, and ultimately affect metal transport or oxidative elimination.

The hub gene *SpbZIP60* was selected for functional analysis because its expression was observed to be upregulated by Cd stress. The overexpression of *SpbZIP60* in transgenic *Arabidopsis* resulted in increased Cd tolerance. More specifically, the photosynthetic apparatus was damaged more in the WT plants than in the transgenic plants following the Cd treatment. Furthermore, Cd accumulated less in the transgenic plants than in the WT controls. These results indicate that SpbZIP60 may affect the uptake or transport of Cd. However, it is unclear whether the increased Cd resistance is also the result of enhancements to other detoxification-related processes. The *Chlamydomonas* bZIP transcription factor BLZ8 confers oxidative stress tolerance by inducing a carbon-concentrating mechanism (Choi et al., 2021). In *Arabidopsis*, AtbZIP60 responds to endoplasmic reticulum stress through the IRE1-bZIP60 mRNA splicing pathway (Deng et al., 2011). Briefly, AtIRE1 selectively recognizes and cleaves the unspliced *bZIP60* mRNA that normally exists in the ER membrane, and the resulting spliced *bZIP60* mRNA can be translated into an active bZIP transcription factor (Howell, 2013). The subcellular localization experiment conducted in the current study demonstrated that SpbZIP60 is a nuclear protein, but whether this means SpbZIP60 contributes to the ER stress response remains to be determined. At present, there are relatively few studies on Cd-mediated ER stress in plants.

CONCLUSION

In this study, we identified 92 bZIP genes in *S. plumbizincicola* and analyzed their evolutionary relationships. These genes were divided into 12 subgroups, and the members of each subgroup had common gene structures and motif compositions. An analysis of the *S. plumbizincicola* bZIP genes revealed eight segmental duplication events, but no tandem duplication events, suggesting that segmental duplication events were the main force driving the evolution of the bZIP gene family in *S. plumbizincicola.* A collinearity analysis involving *S. plumbizincicola* and other species and a comparison between the *S. plumbizincicola* genes and the genes encoding bZIP transcription factors with known functions in model plants will provide new clues regarding SpbZIP functions. We also characterized the *SpbZIP* expression profiles under Cd stress conditions and constructed a co-expression network comprising 11 *SpbZIP* hub genes. The results of this study reflect the importance of SpbZIP transcription factors for regulating plant responses to Cd stress. The expression of the hub gene *SpbZIP60* was induced by Cd stress and enhanced the Cd tolerance of transgenic *Arabidopsis*. Overall, these findings may provide new insights into the stress response-related functions of SpbZIP transcription factors in *S. plumbizincicola*.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ZL and RZ designed the experiments. ZL performed the experiments, analyzed the data, and wrote the manuscript. KJ and MY analyzed the data and prepared the display items. ZL, WQ, XJH, and RZ helped revise the manuscript. LW and CW designed the work and provided materials. XYH provided the culture room. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (31872168) and the National Non-profit Institute Research Grant of CAF (RISFZ-2021-01 and CAFYBB2020SY016).

ACKNOWLEDGMENTS

We thank Liwen Bianji (Edanz) (www.liwenbianji.cn) for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.859386/full#supplementary-material

REFERENCES

Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., et al. (2005). *FD*, a bZIP protein mediating signals from the floral pathway integrator *FT* at the shoot apex. *Science* 309, 1052–1056. doi: 10.1126/science.1115983

Ali, H., Khan, E., and Sajad, M. A. (2013). Phytoremediation of heavy metals—concepts and applications. *Chemosphere* 91, 869–881. doi: 10.1016/j.chemosphere.2013.01.075

Assunção, A. G., Herrero, E., Lin, Y.-F., Huettel, B., Talukdar, S., Smaczniak, C., et al. (2010). *Arabidopsis thaliana* transcription factors *bZIP19* and *bZIP23* regulate the adaptation to zinc deficiency. *Proc. Natl. Acad. Sci.* 107, 10296–10301. doi: 10.1073/pnas.1004788107

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Banerjee, A., and Roychoudhury, A. (2017). Abscisic-acid-dependent basic leucine zipper (bZIP) transcription factors in plant abiotic stress. *Protoplasma* 254, 3–16. doi: 10.1007/s00709-015-0920-4

Bi, C., Yu, Y., Dong, C., Yang, Y., Zhai, Y., Du, F., et al. (2021). The bZIP transcription factor *TabZIP15* improves salt stress tolerance in wheat. *Plant Biotechnol. J.* 19, 209–211. doi: 10.1111/pbi.13453

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4:10. doi: 10.1186/1471-2229-4-10

Chen, S., Han, X., Fang, J., Lu, Z., Qiu, W., Liu, M., et al. (2017). Sedum alfredii SaNramp6 metal transporter contributes to cadmium accumulation in transgenic *Arabidopsis thaliana*. *Sci. Rep.* 7, 1–13. doi: 10.1038/s41598-017-13463-4

Chen, S., Yu, M., Li, H., Wang, Y., Lu, Z., Zhang, Y., et al. (2020). *SaHsfA4c* From *Sedum alfredii* hance enhances cadmium tolerance by regulating ROS-scavenger activities and heat shock proteins expression. *Front. Plant Sci.* 11:142. doi: 10.3389/fpls.2020.00142

Choi, B. Y., Kim, H., Shim, D., Jang, S., Yamaoka, Y., Shin, S., et al. (2021). The *Chlamydomonas* bZIP transcription factor BLZ8 confers oxidative stress tolerance by inducing the carbon-concentrating mechanism. *Plant Cell* 34, 910–926. doi: 10.1093/plcell/koab293

Corrêa, L. G. G., Riaño-Pachón, D. M., Schrago, C. G., Vicentini dos Santos, R., Mueller-Roeber, B., and Vincentz, M. (2008). The role of bZIP transcription factors in green plant evolution: adaptive features emerging from four founder genes. *PLoS One* 3:e2944. doi: 10.1371/journal.pone.0002944

Dash, M., Yordanov, Y. S., Georgieva, T., Tschaplinski, T. J., Yordanova, E., and Busov, V. (2017). Poplar PtabZIP1-like enhances lateral root formation and biomass growth under drought stress. *Plant J.* 89, 692–705. doi: 10.1111/tpj.13413

Deng, Y., Humbert, S., Liu, J.-X., Srivastava, R., Rothstein, S. J., and Howell, S. H. (2011). Heat induces the splicing by IRE1 of a mRNA encoding a transcription factor involved in the unfolded protein response in *Arabidopsis*. *Proc. Natl. Acad. Sci.* 108, 7247–7252. doi: 10.1073/pnas.1102117108

Dixit, S. K., Gupta, A., Fatima, U., and Senthil-Kumar, M. (2019). *AtGBF3* confers tolerance to *Arabidopsis thaliana* against combined drought and *Pseudomonas syringae* stress. *Environ. Exp. Bot.* 168:103881. doi: 10.1016/j.envexpbot.2019.103881

Dröge-Laser, W., Snoek, B. L., Snel, B., and Weiste, C. (2018). The *Arabidopsis* bZIP transcription factor family—an update. *Curr. Opin. Plant Biol.* 45, 36–49. doi: 10.1016/j.pbi.2018.05.001

Farinati, S., DalCorso, G., Varotto, S., and Furini, A. (2010). The *Brassica juncea BjCdR15*, an ortholog of *Arabidopsis TGA3*, is a regulator of cadmium uptake, transport and accumulation in shoots and confers cadmium tolerance in transgenic plants. *New Phytol.* 185, 964–978. doi: 10.1111/j.1469-8137.2009.03132.x

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005). "Protein identification and analysis tools on the ExPASy server," in *The proteomics protocols handbook*, ed. J. M. Walker (Totowa, NJ: Humana Press), 571–607. doi: 10.1385/1-59259-890-0:571

Han, X., Yin, H., Song, X., Zhang, Y., Liu, M., Sang, J., et al. (2016). Integration of small RNAs, degradome and transcriptome sequencing in hyperaccumulator *Sedum alfredii* uncovers a complex regulatory network and provides insights into cadmium phytoremediation. *Plant Biotechnol. J.* 14, 1470–1483. doi: 10.1111/pbi.12512

Howell, S. H. (2013). Endoplasmic reticulum stress responses in plants. *Annu. Rev. Plant Biol.* 64, 477–499. doi: 10.1146/annurev-arplant-050312-120053

Huang, X., Ouyang, X., Yang, P., Lau, O. S., Li, G., Li, J., et al. (2012). *Arabidopsis FHY3* and *HY5* positively mediate induction of *COP1* transcription in response to photomorphogenic UV-B light. *Plant Cell* 24, 4590–4606. doi: 10.1105/tpc.112.103994

Izawa, T., Foster, R., Nakajima, M., Shimamoto, K., and Chua, N.-H. (1994). The rice bZIP transcriptional activator *RITA-1* is highly expressed during seed development. *Plant Cell* 6, 1277–1287. doi: 10.1105/tpc.6.9.1277

Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., et al. (2002). bZIP transcription factors in *Arabidopsis*. *Trends Plant Sci.* 7, 106–111. doi: 10.1016/S1360-1385(01)02223-3

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Li, H., Sun, Q., and Zhao, S. (2000). *Principle and Technology of Plant Physiological and Biochemical Experiments*. Beijing: Higher Education Press.

Li, J.-T., Gurajala, H. K., Wu, L.-H., van der Ent, A., Qiu, R. l, Baker, A. J., et al. (2018). Hyperaccumulator plants from China: a synthesis of the current state of knowledge. *Environ. Sci. Technol.* 52, 11980–11994. doi: 10.1021/acs.est.8b01060

Lilay, G. H., Castro, P. H., Campilho, A., and Assunção, A. G. (2019). The *Arabidopsis bZIP19* and *bZIP23* activity requires zinc deficiency–insight on regulation from complementation lines. *Front. Plant Sci.* 9:1955. doi: 10.3389/fpls.2018.01955

Lilay, G. H., Persson, D. P., Castro, P. H., Liao, F., Alexander, R. D., Aarts, M. G., et al. (2021). *Arabidopsis bZIP19* and *bZIP23* act as zinc sensors to control plant zinc status. *Nat. Plants* 7, 137–143. doi: 10.1038/s41477-021-00856-7

Lin, J., Gao, X., Zhao, J., Zhang, J., Chen, S., and Lu, L. (2020). Plant cadmium resistance 2 (*SaPCR2*) facilitates cadmium efflux in the roots of hyperaccumulator *Sedum alfredii* Hance. *Front. Plant Sci.* 11:568887. doi: 10.3389/fpls.2020.568887

Liu, G.-T., Wang, J.-F., Cramer, G., Dai, Z.-W., Duan, W., Xu, H.-G., et al. (2012). Transcriptomic analysis of grape (*Vitis vinifera* L.) leaves during and after recovery from heat stress. *BMC Plant Biol.* 12:174. doi: 10.1186/1471-2229-12-174

Liu, H., Zhao, H., Wu, L., Liu, A., Zhao, F.-J., and Xu, W. (2017). Heavy metal ATPase 3 (HMA3) confers cadmium hypertolerance on the cadmium/zinc hyperaccumulator *Sedum plumbizincicola*. *New Phytol.* 215, 687–698. doi: 10.1111/nph.14622

Liu, J.-X., Srivastava, R., and Howell, S. H. (2008). Stress-induced expression of an activated form of AtbZIP17 provides protection from salt stress in *Arabidopsis*. *Plant Cell Environ.* 31, 1735–1743. doi: 10.1111/j.1365-3040.2008.01873.x

Liu, M., He, X., Feng, T., Zhuo, R., Qiu, W., Han, X., et al. (2019a). cDNA library for mining functional genes in *Sedum alfredii* hance related to cadmium tolerance and characterization of the roles of a novel *SaCTP2* gene in enhancing cadmium hyperaccumulation. *Enviro. Sci. Technol.* 53, 10926–10940. doi: 10.1021/acs.est.9b03237

Liu, M., Wen, Y., Sun, W., Ma, Z., Huang, L., Wu, Q., et al. (2019b). Genome-wide identification, phylogeny, evolutionary expansion and expression analyses of bZIP transcription factor family in tartary buckwheat. *BMC Genomics* 20:483. doi: 10.1186/s12864-019-5882-z

Liu, M., Qiu, W., He, X., Zheng, L., Song, X., Han, X., et al. (2016). Functional characterization of a Gene in *Sedum alfredii* hance resembling rubber elongation factor endowed with functions associated with cadmium tolerance. *Front. Plant Sci.* 7:965. doi: 10.3389/fpls.2016.00965

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2– ΔΔCT method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lu, Z., Chen, S., Han, X., Zhang, J., Qiao, G., Jiang, Y., et al. (2020). A single amino acid change in *Nramp6* from *Sedum Alfredii* hance affects cadmium accumulation. *Int. J. Mol. Sci.* 21:3169. doi: 10.3390/ijms21093169

Muszynski, M. G., Dam, T., Li, B., Shirbroun, D. M., Hou, Z., Bruggemann, E., et al. (2006). Delayed flowering1 encodes a basic leucine zipper protein that mediates floral inductive signals at the shoot apex in maize. *Plant Physiol.* 142, 1523–1536. doi: 10.1104/pp.106.088815

Nijhawan, A., Jain, M., Tyagi, A. K., and Khurana, J. P. (2008). Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. *Plant Physiol.* 146, 333–350. doi: 10.1104/pp.107.112821

Pan, C., Lu, H., Yu, J., Liu, J., Liu, Y., and Yan, C. (2019). Identification of Cadmium-responsive Kandelia obovata SOD family genes and response to Cd toxicity. *Environ. Exp. Bot.* 162, 230–238. doi: 10.1016/j.envexpbot.2019.02.018

Pautler, M., Eveland, A. L., LaRue, T., Yang, F., Weeks, R., Lunde, C., et al. (2015). *FASCIATED EAR4* encodes a bZIP transcription factor that regulates shoot meristem size in maize. *Plant Cell* 27, 104–120. doi: 10.1105/tpc.114.132506

Peng, J.-S., Ding, G., Meng, S., Yi, H.-Y., and Gong, J.-M. (2017). Enhanced metal tolerance correlates with heterotypic variation in SpMTL, a metallothionein-like protein from the hyperaccumulator *Sedum plumbizincicola*. *Plant Cell Environ.* 40, 1368–1378. doi: 10.1111/pce.12929

Pérez-Rodríguez, P., Riano-Pachon, D. M., Corrêa, L. G. G., Rensing, S. A., Kersten, B., and Mueller-Roeber, B. (2010). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.* 38(Suppl._1), D822–D827. doi: 10.1093/nar/gkp805

Qiu, W., Song, X., Han, X., Liu, M., Qiao, G., and Zhuo, R. (2018). Overexpression of *Sedum alfredii* cinnamyl alcohol dehydrogenase increases the tolerance and accumulation of cadmium in *Arabidopsis*. *Environ. Exp. Bot.* 155, 566–577. doi: 10.1016/j.envexpbot.2018.08.003

Ramamoorthy, R., Jiang, S.-Y., Kumar, N., Venkatesh, P. N., and Ramachandran, S. (2008). A comprehensive transcriptional profiling of the WRKY gene family in rice under various abiotic and phytohormone treatments. *Plant Cell Physiol.* 49, 865–879. doi: 10.1093/pcp/pcn061

Ramegowda, V., Gill, U. S., Sivalingam, P. N., Gupta, A., Gupta, C., Govind, G., et al. (2017). GBF3 transcription factor imparts drought tolerance in *Arabidopsis thaliana*. *Sci. Rep.* 7, 1–13. doi: 10.1038/s41598-017-09542-1

Sang, J., Han, X., Liu, M., Qiao, G., Jiang, J., and Zhuo, R. (2013). Selection and validation of reference genes for real-time quantitative PCR in hyperaccumulating ecotype of *Sedum alfredii* under different heavy metals stresses. *PLoS One* 8:e82927. doi: 10.1371/journal.pone.0082927

Su, L., Lv, A., Wen, W., Zhou, P., and An, Y. (2020). Auxin is involved in magnesium-mediated photoprotection in photosystems of alfalfa seedlings under aluminum stress. *Front. Plant Sci.* 11:746. doi: 10.3389/fpls.2020.00746

Sun, Y., Xu, W., Jia, Y., Wang, M., and Xia, G. (2015). The wheat *TaGBF1* gene is involved in the blue-light response and salt tolerance. *Plant J.* 84, 1219–1230. doi: 10.1111/tpj.13082

Tang, N., Zhang, H., Li, X., Xiao, J., and Xiong, L. (2012). Constitutive activation of transcription factor OsbZIP46 improves drought tolerance in rice. *Plant Physiol.* 158, 1755–1768. doi: 10.1104/pp.111.190389

Tchounwou, P. B., Yedjou, C. G., Patlolla, A. K., and Sutton, D. J. (2012). Heavy metal toxicity and the environment. *Mol. Clin. Environ. Toxicol.* 101, 133–164. doi: 10.1007/978-3-7643-8340-4_6

Tu, M., Wang, X., Yin, W., Wang, Y., Li, Y., Zhang, G., et al. (2020). Grapevine *VlbZIP30* improves drought resistance by directly activating *VvNAC17* and promoting lignin biosynthesis through the regulation of three peroxidase genes. *Hortic. Res.* 7, 1–15. doi: 10.1038/s41438-020-00372-3

Wang, B., Du, H., Zhang, Z., Xu, W., and Deng, X. (2017). *BhbZIP60* from resurrection plant *Boea hygrometrica* is an mRNA splicing-activated endoplasmic reticulum stress regulator involved in drought tolerance. *Front. Plant Sci.* 8:245. doi: 10.3389/fpls.2017.00245

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteomics Bioinformat.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wang, Q., Guo, C., Li, Z., Sun, J., Wang, D., Xu, L., et al. (2021). Identification and analysis of bZIP family genes in potato and their potential roles in stress responses. *Front. Plant Sci.* 12:637343. doi: 10.3389/fpls.2021.637343

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wei, K., Chen, J., Wang, Y., Chen, Y., Chen, S., Lin, Y., et al. (2012). Genome-wide analysis of bZIP-encoding genes in maize. *DNA Res.* 19, 463–476. doi: 10.1093/dnares/dss026

Wu, L. H., Liu, Y. J., Zhou, S. B., Guo, F. G., Bi, D., Guo, X. H., et al. (2013). *Sedum plumbizincicola* X.H. Guo et S.B. Zhou ex L.H. Wu (Crassulaceae): a new species from Zhejiang Province, China. *Plant Syst. Evol.* 299, 487–498. doi: 10.1007/s00606-012-0738-x

Xu, D., Jin, K., Jiang, H., Gong, D., Yang, J., Yu, W., et al. (2022). GFAP: ultra-fast and accurate gene functional annotation software for plants. *bioRxiv* Available online at: https://www.biorxiv.org/content/10.1101/2022.01.05.475154v1. doi: 10.1101/2022.01.05.475154[accessed on January 6, 2022]

Xu, D., Lu, Z., Jin, K., Qiu, W., Qiao, G., Han, X., et al. (2021a). SPDE: a multi-functional software for sequence processing and data extraction. *Bioinformatics* 37, 3686–3687. doi: 10.1093/bioinformatics/btab235

Xu, D., Lu, Z., Qiao, G., Qiu, W., Wu, L., Han, X., et al. (2021b). Auxin-Induced *SaARF4* down regulates *SaACO4* to inhibit lateral root formation in *Sedum alfredii* Hance. *Int. J. Mol. Sci.* 22:1297. doi: 10.3390/ijms22031297

Xu, Z.-Y., Kim, S. Y., Hyeon, D. Y., Kim, D. H., Dong, T., Park, Y., et al. (2013). The Arabidopsis NAC transcription factor ANAC096 cooperates with bZIP-type transcription factors in dehydration and osmotic stress responses. *Plant Cell* 25, 4708–4724. doi: 10.1105/tpc.113.119099

Yang, Q., Shohag, M. J. I., Feng, Y., He, Z., and Yang, X. (2017). Transcriptome comparison reveals the adaptive evolution of two contrasting ecotypes of Zn/Cd hyperaccumulator *Sedum alfredii* Hance. *Front. Plant Sci.* 8:425. doi: 10.3389/fpls.2017.00425

Yang, X., Long, X. X., Ni, W. Z., and Fu, C. X. (2002). *Sedum alfredii* H: a new Zn hyperaccumulating plant first found in China. *Chin. Sci. Bull.* 47, 1634–1637. doi: 10.1360/02tb9359

Yang, X. E., Long, X. X., Ye, H. B., He, Z. L., Calvert, D. V., and Stoffella, P. J. (2004). Cadmium tolerance and hyperaccumulation in a new Zn-hyperaccumulating plant species (*Sedum alfredii* Hance). *Plant Soil* 259, 181–189. doi: 10.1023/B:PLSO.0000020956.24027.f2

Yao, X., Cai, Y., Yu, D., and Liang, G. (2018). *bHLH104* confers tolerance to cadmium stress in *Arabidopsis thaliana*. *J. Integr. Plant Biol.* 60, 691–702. doi: 10.1111/jipb.12658

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. doi: 10.1371/journal.pbio.0030038

Zhang, M., Liu, Y., Shi, H., Guo, M., Chai, M., He, Q., et al. (2018). Evolutionary and expression analyses of soybean basic Leucine zipper transcription factor family. *BMC Genomics* 19:159. doi: 10.1186/s12864-018-4511-6

Zhang, M., Zhang, J., Lu, L. L., Zhu, Z. Q., and Yang, X. E. (2016). Functional analysis of CAX2-like transporters isolated from two ecotypes of *Sedum alfredii*. *Biol. Plant.* 60, 37–47. doi: 10.1007/s10535-015-0557-3

Zhang, P., Wang, R., Ju, Q., Li, W., Tran, L.-S. P., and Xu, J. (2019). The R2R3-MYB transcription factor *MYB49* regulates cadmium accumulation. *Plant Physiol.* 180, 529–542. doi: 10.1104/pp.18.01380

Zhang, Q., Cai, W., Ji, T.-T., Ye, L., Lu, Y.-T., and Yuan, T.-T. (2020). *WRKY13* enhances cadmium tolerance by promoting D-cysteine desulfhydrase and hydrogen sulfide production. *Plant Physiol.* 183, 345–357. doi: 10.1104/pp.19.01504

Zhang, X., Henriques, R., Lin, S.-S., Niu, Q.-W., and Chua, N.-H. (2006). Agrobacterium-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nat. Protocols* 1, 641–646. doi: 10.1038/nprot.2006.97

Zhao, J., Guo, R., Guo, C., Hou, H., Wang, X., and Gao, H. (2016). Evolutionary and expression analyses of the apple basic leucine zipper transcription factor family. *Front. Plant Sci.* 7:376. doi: 10.3389/fpls.2016.00376

Zhao, K., Chen, S., Yao, W., Cheng, Z., Zhou, B., and Jiang, T. (2021). Genome-wide analysis and expression profile of the bZIP gene family in poplar. *BMC Plant Biol.* 21:122. doi: 10.1186/s12870-021-02879-w

Zhu, Y., Qiu, W., Li, Y., Tan, J., Han, X., Wu, L., et al. (2022). Quantitative proteome analysis reveals changes of membrane transport proteins in *Sedum plumbizincicola* under cadmium stress. *Chemosphere* 287:132302. doi: 10.1016/j.chemosphere.2021.132302

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fpls.2022.884443

TCP Transcription Factors Involved in Shoot Development of Ma Bamboo (*Dendrocalamus latiflorus* Munro)

Kangming Jin[1,2†], Yujun Wang[1†], Renying Zhuo[1], Jing Xu[1], Zhuchou Lu[1], Huijin Fan[1], Biyun Huang[1] and Guirong Qiao[1]*

[1] State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Tree Breeding of Zhejiang Province, Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou, China, [2] Forestry Faculty, Nanjing Forestry University, Nanjing, China

OPEN ACCESS

Edited by:
Weihua Pan,
Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Alex Goldshmidt,
Agricultural Research Organization, Volcani Center, Israel
Qiang Zhu,
Fujian Agriculture and Forestry University, China

***Correspondence:**
Guirong Qiao
gr_q1982@163.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: 26 February 2022
Accepted: 08 April 2022
Published: 10 May 2022

Citation:
Jin K, Wang Y, Zhuo R, Xu J, Lu Z, Fan H, Huang B and Qiao G (2022) TCP Transcription Factors Involved in Shoot Development of Ma Bamboo (Dendrocalamus latiflorus Munro). Front. Plant Sci. 13:884443. doi: 10.3389/fpls.2022.884443

Ma bamboo (*Dendrocalamus latiflorus* Munro) is the most widely cultivated clumping bamboo in Southern China and is valuable for both consumption and wood production. The development of bamboo shoots involving the occurrence of lateral buds is unique, and it affects both shoot yield and the resulting timber. Plant-specific TCP transcription factors are involved in plant growth and development, particularly in lateral bud outgrowth and morphogenesis. However, the comprehensive information of the *TCP* genes in Ma bamboo remains poorly understood. In this study, 66 TCP transcription factors were identified in Ma bamboo at the genome-wide level. Members of the same subfamily had conservative gene structures and conserved motifs. The collinear analysis demonstrated that segmental duplication occurred widely in the TCP transcription factors of Ma bamboo, which mainly led to the expansion of a gene family. Cis-acting elements related to growth and development and stress response were found in the promoter regions of *DlTCPs*. Expression patterns revealed that *DlTCPs* have tissue expression specificity, which is usually highly expressed in shoots and leaves. Subcellular localization and transcriptional self-activation experiments demonstrated that the five candidate TCP proteins were typical self-activating nuclear-localized transcription factors. Additionally, the transcriptome analysis of the bamboo shoot buds at different developmental stages helped to clarify the underlying functions of the TCP members during the growth of bamboo shoots. *DlTCP12-C*, significantly downregulated as the bamboo shoots developed, was selected to further verify its molecular function in *Arabidopsis*. The *DlTCP12-C* overexpressing lines exhibited a marked reduction in the number of rosettes and branches compared with the wild type in *Arabidopsis*, suggesting that *DlTCP12-C* conservatively inhibits lateral bud outgrowth and branching in plants. This study provides useful insights into the evolutionary patterns and molecular functions of the TCP transcription factors in Ma bamboo and provides a valuable reference for further research on the regulatory mechanism of bamboo shoot development and lateral bud growth.

Keywords: TCP transcription factors, Ma bamboo, expression profiles, bamboo shoot development, transcriptome analysis

INTRODUCTION

Transcription factors play a critical role in plant growth and development and can regulate development by transmitting external environmental factors, mediating hormone signal pathways, and responding to gene regulatory networks in plants (Chen et al., 2021). TCP genes, a kind of plant-specific transcription factors, are typically involved in the plant developmental process, such as seed germination (Zhang et al., 2019a), floral organ development (Wang et al., 2015), leaf morphogenesis (Sarvepalli and Nath, 2011), axillary meristem development (Aguilar-Martínez et al., 2007; Nicolas et al., 2015; Min et al., 2021), and hormone signal transduction (Kosugi and Ohashi, 2002; González-Grandío et al., 2017). The abbreviation "TCP" comes from its earliest discovered members: *Teosinte Branched1* (*TB1*) from maize (*Zea mays*), *CYCLOIDEA* (*CYC*) from snapdragon (*Antirrhinum majus*), and the *PROLIFERATING CELL FACTORS* (*PCF*) from rice (*Oryza sativa*) (Luo et al., 1996; Doebley et al., 1997; Kosugi and Ohashi, 1997). They all have a conserved TCP domain, a basic helix–loop–helix (bHLH) structure, which is primarily related to DNA binding, protein interaction, and protein nuclear localization (Cubas et al., 1999). It can also be divided into Class I (also known as PCF) and Class II according to the characteristics of the conserved domain of TCP proteins. Additionally, Class II can be divided into CIN subfamily and CYC/TB1 subfamily (Martín-Trillo and Cubas, 2010). The most obvious difference between Class I and Class II is the absence of four conservative amino acids in the Class I basic TCP domain. The R domain is rich in polar amino acids such as lysine, glutamic acid, and arginine and is only found in Class II members, which is used to mediate protein interaction (Cubas et al., 1999).

So far, the majority of TCP family members have been determined to play a role in plant growth and development. Previous studies have shown that the *TCP* members of Class I primarily promotes leaf cell proliferation, thereby regulating plant growth and development, and plays an important role in the response to stress (Aguilar Martinez and Sinha, 2013). In *Arabidopsis thaliana*, *TCP14* and *TCP15* are involved in seed germination by acting downstream of the gibberellin and vernalization pathways (Resentini et al., 2015; Xu et al., 2020). Through binding to the homologous GCCCR elements, AtTCP20 regulates the expression of cyclin and ribosomal protein genes. Furthermore, it acts as a flexible regulator to coordinate growth and division pathways in post-embryonic plant development (Li et al., 2005). Additionally, the downregulation of the expression of *OsPCF6* and *OsTCP21* enhances the tolerance of rice under cold stress by changing the scavenging of reactive oxygen species (Wang et al., 2014). *TCP21* is particularly important during plant growth. Its low expression makes rice more susceptible to rice ragged stunt virus (RRSV) (Zhang et al., 2016; Wang et al., 2021). Overexpression lines of *TCP21* exhibited increased tiller bud length, biomass, and tiller number in rice (Wang et al., 2021). *PeTCP10* is induced by drought and ABA treatment and plays a vital role in plant growth and development and response to environmental stress, which can be seen due to its effects in the evident effects on drought tolerance and the lateral root growth of transgenic lines (Liu et al., 2020). However, the *TCP* members of Class II primarily plays an important role in morphological construction and organ development (Manassero et al., 2013; Sarvepalli and Nath, 2018). Overexpression of *AtTCP1*, the homolog of *CYC*, directly stimulates the expression of the target gene *DWF4* to actively regulate brassinosteroid (BR) biosynthesis, which affects leaf development and leaf shape regulation (Guo et al., 2010; An et al., 2011). It was found that *OsTCP17* (*REP1*), a *CYC* homologous gene in rice, regulates the attributes and development of palea and controls the flower symmetry of the inner and outer lemma axis (Yuan et al., 2009). In wild rice, the *OsTCP15* (*TIG1*) gene is specifically highly expressed on the distal side of the tiller base, which promotes cell elongation by activating the expression of downstream genes such as *EXPA3*, *EXPB5,* and *SAUR39*, which helps to maintain a large tillering angle (Zhang et al., 2019b).

Branching affects plant morphogenesis and growth to a great extent and determines plant architecture, yield, and ecological sustainability (Richards, 2000; Wang and Li, 2008; Wang et al., 2019). *TB1*, an inhibitory regulator of the development of axillary buds, realizes the transformation from lateral branch growth to apical dominance, transforming teosinte with more tiller numbers into commonly cultivated maize (Doebley et al., 1997; Wang et al., 2019). It has been found in many species due to its conservative function of negatively regulating axillary bud growth (Takeda et al., 2003; Kebrom et al., 2006; Dixon et al., 2018; Shen et al., 2019). *TCP18* (*BRANCHED1, BRC1*) and *TCP12* (*BRANCHEND2, BRC2*) in *Arabidopsis* are the two orthologous homologs of *ZmTB1* in maize, which are highly expressed in axillary buds and negatively regulate the growth of axillary buds, whereas SML 6, 7, and 8 promote branching through transcriptional inhibition of *BRC1* and the non-transcriptional regulation of auxin (Aguilar-Martínez et al., 2007; Wang et al., 2020). Strigolactones (SLs) are a new plant hormone that inhibits the germination of lateral branches and suppresses the growth of tiller buds in rice by regulating the transcriptional level of *OsTB1* (Takeda et al., 2003; Nicolas and Cubas, 2016; Wang et al., 2018). Meanwhile, *OsMADS57* and *OsTB1* jointly regulate the transcription of its target genes *OsWRKY94* and *D14*, realizing the transition between organogenesis and cold adaptation defense in rice at different temperatures (Chen et al., 2018). As wild cucumbers were domesticated, it was found that two light response elements inserted in the promoter region promoted the expression of *CsBRC1*, which directly restrained the auxin outflow from the lateral buds mediated by *PIN3* (encoding auxin transporter) and reduced the production of branches due to the accumulation of auxin in axillary buds (Shen et al., 2019). To date, homologous genes of *TB1* were subsequently identified in numerous gramineous plants such as rice (Takeda et al., 2003), wheat (Dixon et al., 2018), and sorghum (Kebrom et al., 2006). They are very conservative in function, which can cause bud dormancy, and effectively regulate the growth of axillary buds in response to hormones and the external environment which controls the tiller number (Wang et al., 2019).

Bamboo is one of the fast-growing non-timber forest resources and has significant economic, cultural, and ecological

value (Zhao et al., 2017). Ma bamboo (*Dendrocalamus latiflorus* Munro) is the most widely cultivated clumping bamboo in Southern China. Its bamboo shoots taste good and are high in nutritional value. Mature bamboo can be used as building materials, decorations, and ornamental planting. Therefore, it is a fast-growing and environmentally friendly clump bamboo species that can be cultivated for its shoots and timber. Bamboo shoots have a longer period of emergence because they are often exposed to the soil surface, which is easily frozen in winter. When cultivating bamboo forests, seasonal shooting and the low germination rate of shoot buds directly affect the yield of bamboo shoots and timber. The formation, growth, and development of bamboo shoots involved in the morphogenesis of lateral branches are unique and affect the yield of bamboo shoots and timber. Therefore, it is important to explore the role of genes related to lateral branch formation in the outgrowth of bamboo shoots, which will help to clarify the molecular mechanism of bamboo shoot development.

Whereas, the TCP family has been characterized in many species, such as *Arabidopsis* (Yao et al., 2007), rice (Yao et al., 2007), sorghum (Francis et al., 2016), and Moso bamboo (Liu et al., 2018), less is known about the TCP transcription factors in Ma bamboo. Recently released genomic data about Ma bamboo, including a representative of hexaploid clumping bamboo (AABBCC, 2n = 6x = 72), allowed us to perform a genome-wide analysis of the TCP transcription factors in Ma bamboo (Zheng et al., 2022). In this study, 66 TCP transcription factor members of Ma bamboo were identified, and the phylogenetic relationship, gene structure and motif information, collinearity, tissue differential expression analysis, subcellular localization analysis, and transcriptional self-activation analysis were analyzed. Additionally, we performed transcriptome analysis of bamboo shoot buds at different developmental stages to clarify the function of the TCP family during bamboo shoot outgrowth. Through the phenotypic observation of transgenic *Arabidopsis*, we concluded that *DlTCP12-C* plays a significant role in controlling the number of branches. Our study reveals the basic information and evolutionary relationship of plant-specific TCP transcription factors in Ma bamboo, clarifies the role of candidate genes in bamboo shoot growth and development by transcriptome analysis, and preliminarily outlines on the function of candidate TCPs, all of which provides valuable insights into future investigations.

MATERIALS AND METHODS

Genome-Wide Identification of Putative *DlTCPs*

To identify TCP transcription factors, the detailed information of *D. latiflorus* genome was obtained through the website (http://forestry.fafu.edu.cn/pub/Dla/). On the Pfam website (http://pfam.xfam.org/), the Hidden Markov Model (HMM) of the conserved TCP domain (PF03634) was downloaded. With a threshold: *e*-values $< 10^{-5}$, all putative TCP members with conserved TCP domain were obtained in our protein dataset through the HMMsearch module in SPDE software (Xu et al., 2021). Subsequently, the putative TCP genes (TCPs) were further checked the integrity of its domain through the NCBI (https://www.ncbi.nlm.nih.gov/), InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/), and SMART (http://smart.embl-heidelberg.de/) databases. The genes without complete TCP domain will be manually eliminated. The ExPaSy (https://web.expasy.org/compute_pi/) and Plant-mPLoc (http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/#) were used to predict their molecular weight (MW) and isoelectric point (pI) and subcellular localization, respectively.

Phylogenetic Tree Construction and Sequence Alignment

All protein sequences of *Arabidopsis*, rice, and Moso bamboo were downloaded from TAIR (https://www.arabidopsis.org/), China Rice Data Center (https://www.ricedata.cn/gene/), and Bamboo databases (http://forestry.fafu.edu.cn/db/PhePacBio/phe/Jbnest.php), respectively. Then, we obtained the sequences of TCP proteins in *Arabidopsis*, rice, and Moso bamboo from previous studies (**Supplementary Table S1**). To explore the evolutionary relationships, ClustalW was used to perform the multiple sequence alignments between 24 *Arabidopsis*, 22 rice, 16 Moso bamboo, and 66 Ma bamboo TCP proteins with default parameters (Thompson et al., 2003), and MEGA 7.0 was subsequently used to construct a neighbor-joining (NJ) phylogenetic tree with the following parameters: NJ tree method, complete deletion, and 1,000 bootstrap replicates (Kumar et al., 2016). Additionally, the multiple sequences' alignment of DlTCP proteins was performed using DNAMAN software (version 9.0), and the conserved TCP domain regions with 55–60 amino acids were intercepted to further investigate the conservation and diversity. The online tool RNA22 v2 microRNA target detection (https://cm.jefferson.edu/rna22/Interactive/) was used to predict miRNA target sites (Miranda et al., 2006).

Gene Structure, Conserved Motifs, Chromosome Distribution, cis-Regulatory Element Analysis, Synteny, and Gene Duplication Analysis

The exon–intron structures of *DlTCPs* were mapped using the TBtools software (Chen et al., 2020) based on the obtained coding sequence (CDS) and genomic sequences of Ma bamboo. The online MEME program version 5.4.1 (http://meme-suite.org/tools/meme) was used to identify and analyze the conserved motifs of TCP proteins in Ma bamboo. The 2,000-bp upstream promoter sequences of the *DlTCPs* were submitted to the online program Plant CARE (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) to search the predicted cis-regulatory elements, and these results were visualized using the online tool Gene Structures Display Server (http://gsds.gao-lab.org/).

MCScanX was used to analyze the duplication events of the *DlTCPs* with the default parameters. The diagram of chromosomal location and synteny relationships was generated by the program Circos (Krzywinski et al., 2009) version 0.69 (http://circos.ca/) based on the information about collinear pairs and genetic location. Meanwhile, collinearity analysis between

Ma bamboo and the three other species of *TCPs* was performed using the dual synteny plot module in TBtools. The non-synonymous (Ka) replacement rate and synonymous (Ks) rate were calculated by Ka/Ks calculator to analyze gene duplication events (Wang et al., 2010).

Plant Materials, Growth Conditions, and qRT-PCR

The seedlings of Ma bamboo were cultured in a culture room at 25°C (16-h light, 8-h dark) with stable humidity. About 6-week-old seedlings in similar growth status were selected to collect the samples of young roots, stems, leaves, and bamboo shoots emerging from the bottom. A total of three biological repetitive samples were collected from each tissue to reduce the experimental error.

Tiangen RNAprep plant kit (Tiangen) were used to isolate total RNA from above-mentioned plant samples. Before dissolving RNA, RNase-free DNaseI (Tiangen, Beijing, China) was used to eliminate any contaminating genomic DNA. Then, 1 μg RNA was reverse-transcribed into first-strand cDNA using Takara PrimeScript First-Strand cDNA Synthesis kit (Takara, Dalian, China). All first-Strand cDNA samples were diluted 5 times and stored at −20°C for real-time quantitative PCR (qRT-PCR) experiments. Gene-specific primers were designed using Primer 5 software and shown in **Supplementary Table S2**. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as used internal reference (Liu et al., 2014). qRT-PCR was performed using TB Green™ Premix Ex Taq™ (Tli RNaseH Plus) kit (Takara) with the QuantStudio™ 7 Flex Real-Time PCR instrument (Applied Biosystems). A total of three biological replicates were carried out to eliminate errors. The relative expression level was estimated based on the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001).

Transcriptome Sequencing and Expression Analysis of *DlTCPs*

The samples used for transcriptome sequencing were collected from Hua'an County, Zhangzhou City, Fujian Province (117°24′-117°35′ *E*, 24°65′-25°02′ *N*) in July 2021. A total of four representative developmental stages of bamboo shoot buds were selected for sampling. In the Stage 1, the largest shoot buds located at the lowest layer of mature bamboo shoots were mainly collected, which were still in the dormant stage. The bamboo shoots were expanding, and the top of the shoots begins to twist and grow upward, which are the main characteristics of bamboo shoots in the Stage 2 (the height is about 8 cm). In the Stage 3, the bamboo shoots grew completely upright, accompanied by the further expansion of the bamboo bodies. The bamboo shoots in the Stage 4 (about 45 cm) have entered the high growth stage, and the height of bamboo shoots is significantly higher than that in the Stage 3 (about 20 cm). After removing the bamboo shoot sheath, the bamboo shoot bud samples around the top 0.4 cm were immediately collected and frozen in liquid nitrogen to prevent RNA degradation. A number of four biological repetitive samples were taken in each stage, and each biological repeat consists of about seven apical buds. The separation of total RNA, quality evaluation, and construction of sequencing library were performed as described in the previous studies (Zou et al., 2021).

Illumina Hiseq 2500 platform (Novogene, Beijing, China) was used for sequencing, and HISAT 2.0.5 software was used to map all clean reads to the reference genome of *D. latiflorus*. The differentially expressed genes (DEGs) were screened by DESeq2 package according to the threshold of fold change ≥1.5 and the adjusted p-value < 0.05. Compared with the two groups, the differentially expressed genes (DEGs) were screened by pairwise comparison. The above RNA-seq and bioinformatic analysis were carried out by BioMarker Technologies Illumina, Inc. (Shanghai, China). To further screen the differential genes in the transcriptome, all the differential genes were annotated using Gene Annotation Software for Plants (GFAP) (Xu et al., 2022), which quickly obtained the detailed information of candidate differential expressed genes in the transcriptome.

Subcellular Localization and Transactivation Activity

To analyze the subcellular localization of candidate genes (*DlTCP5-C*, *DlTCP7-B*, *DlTCP9-A*, *DlTCP12-C,* and *DlTCP23-C*), we designed specific primers to amplify the full-length coding sequence of candidate TCP genes and then inserted them into mGFP fusion expression vector pMDC43 (**Supplementary Table S2**). Then, the location signal was analyzed in the leaf tissue of *Nicotiana benthamiana* after transforming *Agrobacterium tumefaciens* GV3101 as described in the previous research (Sparkes et al., 2006). The empty vector was used as the control. About 72 h later, the transient expression of GFP fusion protein was observed by LSM900 confocal microscope imaging system (Zeiss, Germany). The nucleus was visualized with mCherry-labeled nuclear markers. Subsequently, the candidate TCP genes were inserted into the pGBKT7 vector to study the transcriptional activity of DlTCP proteins in yeast (**Supplementary Table S2**). Then, the above recombinant vector, positive control pGBKT7-p53 + pGADT7-T, and negative control pGBKT7 empty plasmid were transformed with lithium acetate method into yeast strain AH109. The transformed strains were cultured on SD medium lacking Trp (SD-Trp) and further selected on defective medium SD-Trp-His-Ade supplemented with X-α-gal for 3–5 days. The transcriptional activation activity was evaluated according to the growth status.

Cloning of the *DlTCP12-C* Gene and Phenotypic Analysis of Transgenic *Arabidopsis*

The full-length CDS of *DlTCP12-C* was amplified from *D. latiflorus* cDNA and constructed into the binary vector pCAMBIA1300 driven by the CaMV 35S promoter using ClonExpress II One Step Cloning Kit (Vazyme, Nanjing, China) (**Supplementary Table S2**). Then, *Agrobacterium tumefaciens* (EHA105)-mediated transformation of *Arabidopsis* was used. The transgenic *Arabidopsis* seeds were collected from T1 plants alone, and positive lines were screened on MS medium with 50 mg L^{-1} kanamycin until homozygous transgenic *Arabidopsis* lines

of the T3 generation were obtained. The methods of phenotypic observation refer to the previous studies (Li et al., 2021).

RESULTS

Identification of *TCP* Genes in Ma Bamboo

A total of 66 TCP members were identified in the *D. latiflorus* genome, among which the number of subgenomes A, B, and C is 22, 20, and 24, respectively. We renamed and classified the DlTCPs according to their subgenomic attribution and chromosomal position. Consistent with *Arabidopsis thaliana*, rice, and Moso bamboo, most DlTCPs belong to the PCF subfamily (35 members), whereas the CIN and CYC/TB1 subfamilies have 21 and 10 members, respectively. The molecular weights (MWs) varied from 13.51 (DlTCP13-C) to 81.80 (DlTCP10-B) kDa, and the lengths varied from 128 (DlTCP13-C) to 768 (DlTCP10-B) amino acid (aa) of DlTCP proteins. The isoelectric point (pI) of the DlTCP proteins varied from 5.16 (DlTCP22-C) to 10.60 (DlTCP9-C). As a typical family of transcription factors, predicted subcellular localization results indicated that the majority of DlTCP proteins were located on the nucleus using the online software Plant-mPLoc. The details of the putative DlTCPs can be found in **Supplementary Table S3**.

Phylogenetic Analysis of TCPs in the Four Different Plant Species

To further explore the phylogenetic relationship and evolutionary process of the TCP gene family, MEGA7.0 was used to construct phylogenetic trees of the TCP members of Ma bamboo, Moso bamboo, rice, and *Arabidopsis thaliana*. The tree constructed by the neighbor-joining (NJ) method divided the TCP family into two clades with 1,000 bootstraps replication: Class I and Class II. A total of 35 DlTCPs were identified in the Class I (PCF) subfamily, and Class II was further categorized into CIN and CYC/TB1 subfamilies (**Figure 1**). In addition, the phylogenetic tree showed that the DlTCPs have a close evolutionary relationship between Ma bamboo and Moso bamboo, both of which belong to *Gramineae*. For example, DlTCP11-C and PeTCP14, and DlTCP8-C and PeTCP13 were clustered in the same small subfamily. We also counted the number of TCP family members in different species (**Table 1**). The results demonstrate that the number of TCP family members in hexaploid Ma bamboo is approximately three times that of diploid plants, which demonstrates the important role of the TCP family during the growth and development of Ma bamboo.

Gene Structure and Conserved Protein Motifs

Gene structure analysis highlighted the differences in conserved domains and motif compositions among subfamilies. However, members of the same subfamily often possess similar structures, for example, motifs 1 and 3 exist in all Class II members (**Figure 2**). The consistency and difference in this structure can be further confirmed in the results of the multiple sequence alignment of TCP proteins in Ma bamboo. Detailed information of conserved motifs is displayed in **Supplementary Table S4**. The deletion of four conservative amino acids in Class I members is the most obvious difference compared with Class II (**Supplementary Figure S1**). Additionally, most TCP members, especially the CYC/TB1 subfamily, showed consistent conservatism with only one exon. In the CIN subfamily, five *DlTCPs* (*DlTCP2-B*, *DlTCP3-C*, *DlTCP12-B*, *DlTCP14-B,* and *DlTCP15-A*) were found to have putative binding sites for *miR319* of Ma bamboo (**Supplementary Table S5**).

Chromosome Distribution and Synteny Analysis

As expected, synteny analysis demonstrated that the members of the TCP gene family had a very complex colinear relationship in Ma bamboo, suggesting that polyploidization was the main source of the expansion of the TCP gene family (**Figure 3**). As shown in **Figure 3**, TCP members have only not been identified on chromosomes 29.1, 30.1, and 31.1, which belongs to the three subgenomes A, B, and C, respectively. The chromosome distribution of TCP members in Ma bamboo was uneven. Chr18.1 contained the largest number of TCP members (6), including 3 PCF members, 1 CYC/TB1 member, and 2 CIN members. Tandem duplication events can drive the renewal of the biological functions of genes. Only one tandem duplication gene pair, *DlTCP5-B* and *DlTCP6-B*, was found in this study.

Collinearity Analysis

The collinearity of TCPs was analyzed in Ma bamboo, *Arabidopsis*, rice, and Moso bamboo by MCscanX. The results demonstrate that Ma bamboo and Moso bamboo, which are both Bambusoideae, had a more conservative evolutionary relationship compared with *Arabidopsis* (**Figure 4**). However, due to polyploidization, several genes are differentiated during the evolutionary process, producing new gene functions to adapt to the environment. The number of genes with a collinear relationship in Ma bamboo is ∼3 times that in rice, indicating that the hexaploid Ma bamboo has a very conservative duplication process during evolution. Homologous gene analysis demonstrated that duplication events occur during genomic evolution. Using phylogenetic tree analysis, 3 putative orthologous gene pairs (*Dl-Os*), 3 putative orthologous gene pairs (*Dl-Pe*), and 14 putative paralogous gene pairs (*Dl-Dl*) were obtained (**Supplementary Table S6**). The Ks value and Ka/Ks ratio of all putative orthologous and paralogous pairs were calculated to analyze the evolutionary selection and divergence pattern of *TCPs* (**Supplementary Table S6**). Generally, a Ka/Ks ratio >1, $=1$, and <1 indicates positive selection, which will be conducive to genetic variation of natural adaptation, neutral selection, and purifying selection to reduce amino acid mutation, respectively. The Ka/Ks ratios of orthologous gene pairs and paralogous gene pairs of TCP members were all <1, indicating that these genes have undergone a strong purifying selection during evolution (**Supplementary Figure S2**).

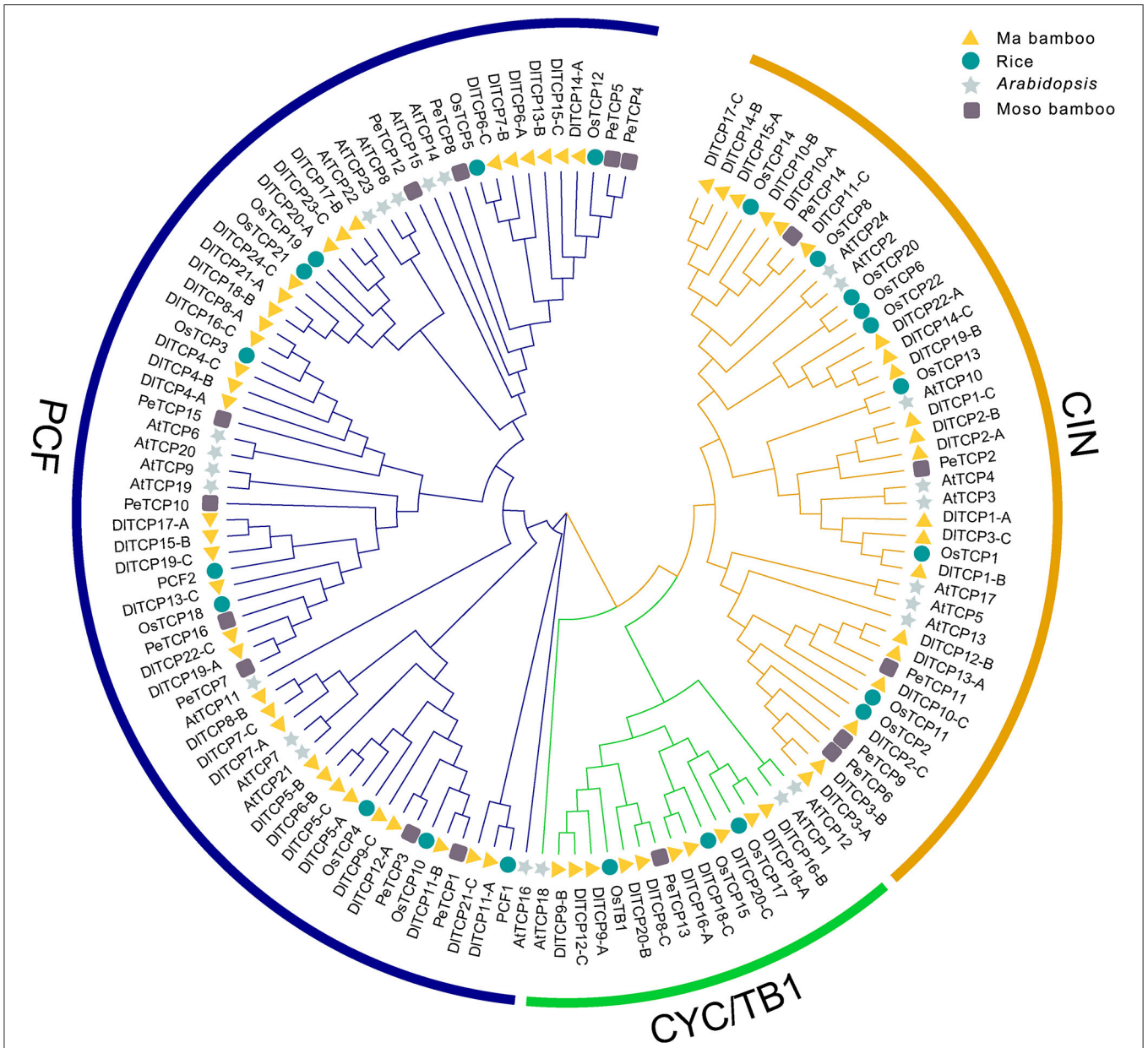


FIGURE 1 | Phylogenetic analysis of TCP members in Ma bamboo, Moso bamboo, rice, and Arabidopsis. MEGA7.0 was used to construct phylogenetic tree by neighbor-joining (NJ) method divided the TCP family into two clades with 1,000 bootstraps replication. Triangle, round, star, and square represent the TCP members of Ma bamboo, rice, *Arabidopsis,* and Moso bamboo, respectively. Different colored geometric shapes were used to mark TCP members of different species.

Detection of cis-Regulatory Elements in the Promoter Regions of *TCPs* in Ma Bamboo

To analyze the cis-acting elements of the promoter regions, sequences of the 2,000-bp upstream of 66 TCP members were extracted from the Ma bamboo genome and predicted on the Plant CARE website (**Figure 5**). A total of 14 cis-acting elements related to hormone response, plant growth and development, and stress have been discovered, including low-temperature-responsive elements, drought-inducible elements, seed-specific regulatory elements, ABA-responsive elements, auxin-responsive elements, and so on. The different cis-elements on promoters may lead to functional differentiation between family members. The number of seed-specific regulatory elements on the promoters of CYC/TB1 subfamily members exceeds that of other members, which suggests that these members have specific functions during seed germination. The results of the phylogenetic tree showed that TCP members that are closely related to each other tend to cluster in a small branch whose promoter region has similar cis-regulatory elements, such as *DlTCP21-A*, *DlTCP18-B,* and *DlTCP24-C*. This suggests that they have a conserved molecular function. Additionally, the promoter

TABLE 1 | The number of TCP family members in five plant species.

Species	PCF	CIN	CYC/TB1	Total	References
Arabidopsis	13	8	3	24	Yao et al., 2007
Rice	10	9	3	22	Yao et al., 2007
Sorghum	9	8	3	20	Francis et al., 2016
Moso bamboo	10	5	1	16	Liu et al., 2018
Ma bamboo	35	21	10	66	

regions of most TCP members have an MYB transcription factor-binding site and WRKY-binding site, especially *DlTCP12-C*, which contains 5 MYB-binding sites and 1WRKY-binding sites.

Tissue-Specific Expression Patterns

The tissue expression pattern can deeply reflect the primary role of TCPs, highlighting the need to further study its specific function. According to the above analysis, 16 putative genes were evenly selected from each subfamily of TCP members for tissue expression analysis. The expression levels of putative TCP members were analyzed in bamboo shoots, roots, stems, and leaves of the 6-week-old seedlings (**Figure 6**). The results demonstrated that TCP family members had obvious tissue expression specificity and were particularly highly expressed in bamboo shoots and leaves. Numerous *TCPs* were highly expressed in shoots, such as *DlTCP5-C*, *DlTCP15-B,* and *DlTCP9-A*. *DlTCP10-B* was specifically expressed in leaves, whereas *DlTCP-4A* was a transcription factor specifically expressed in the roots. We then attempted to clarify the potential functions of TCP members in shoots by other means.

Transcriptome Analysis of Shoot Buds at Different Developmental Stages in Ma Bamboo

Bamboo shoot buds with a top of ∼0.4 cm at four developmental stages were selected for transcriptome sequencing (**Figure 7A**). We used transcriptome analysis to demonstrate that many growth- and development-related transcription factors were differentially expressed genes, including MYB, NAC, WRKY, and TCP. RNA-seq results indicated that the number of differentially expressed genes (DEGs) in the S1–S2, S1–S3, and S1–S4 groups were higher than in other groups, whereas the number of DEGs significantly decreased between S2, S3, and S4 (**Figure 7B**). There were only a few differential genes between S3 and S4, which indicates that the morphological structure of the top has been basically developed in later stages of bamboo shoot bud development, making the gene expression very similar. Therefore, we focused on the DEGs in the S1–S2, S1–S3, and S1–S4 comparative groups and selected the common differential genes as the candidate genes. As shown in **Figure 7C**, 3,444 and 3,206 DEGs were consistently upregulated and downregulated in the S1–S2, S1–S3, and S1–S4 groups, respectively. The GO enrichment revealed that the DEGs were involved in biological processes, cellular components, and molecular functions (**Supplementary Figure S4**). Genes related to hormone signaling pathways, transcription factors (MYB, NAC, and WRKY), and developmental processes (expansin, growth-regulating factor, and Dof zinc finger proteins) showed significant differential expression during four developmental stages in Ma bamboo shoots (**Figure 7D**; **Supplementary Table S7**).

Expression Analysis of TCPs in Shoot Buds at Different Developmental Stages

The expression levels of TCP family members in different developmental stages of bamboo shoot buds were obtained to further explore their roles in shoot bud growth and development using transcriptome data. A total of 29 TCP family members were found as DEGs, and 9 and 20 differentially upregulated and downregulated genes, respectively. The expression levels of *DlTCP9-A* and *DlTCP12-C* belonging to the CYC/TB1 subfamily were significantly decreased during the shoot germination, indicating that they play an important role in the growth and development of bamboo shoots. This is consistent with the molecular function of the *TB1* gene in inhibiting branch growth (**Figure 8A**). We further counted the differences in DEGs in various stages, and the results indicated that the number of downregulated genes gradually decreased whereas the upregulated genes gradually increased from S1 to S4 in bamboo shoots. This indicates that some members of the TCP family primarily function in the early stage of shoot bud germination and development, while others are highly expressed in the later stages of development (**Figure 8B**). This emphasized that TCP family members are thoroughly involved in the growth and development of bamboo shoots. Meanwhile, we selected 8 TCP members to verify the reliability of the transcriptome results by qRT-PCR (**Figure 8C**). GO annotation indicated that TCP transcription factors were involved in growth and development, hormone signal transduction, and stimulus response (**Supplementary Figure S3**).

Subcellular Localization and Transactivation Activity

The protein structure of transcription factors is typically composed of four functional domains: DNA-binding domain, transcriptional regulatory domain (including activation domain or inhibitory domain), oligomerization site, and nuclear localization signal. Transcription factors normally function in the nucleus to regulate the expression of their target genes by binding to corresponding binding elements on their promoters. A total of four putative *DlTCPs* highly expressed in bamboo shoots were further selected for subcellular localization and transcriptional self-activation experiment from the 16 genes used for tissue expression analysis. The results showed that *DlTCP5-C*, *DlTCP7-B*, *DlTCP9-A,* and *DlTCP23-C* are all nuclear-localized TCP transcription factors (**Figure 9**). In addition, the yeast transforms with the four candidate genes can grow on the defective medium of SD-Trp-His-Ade supplemented with X-α-gal and turned blue,

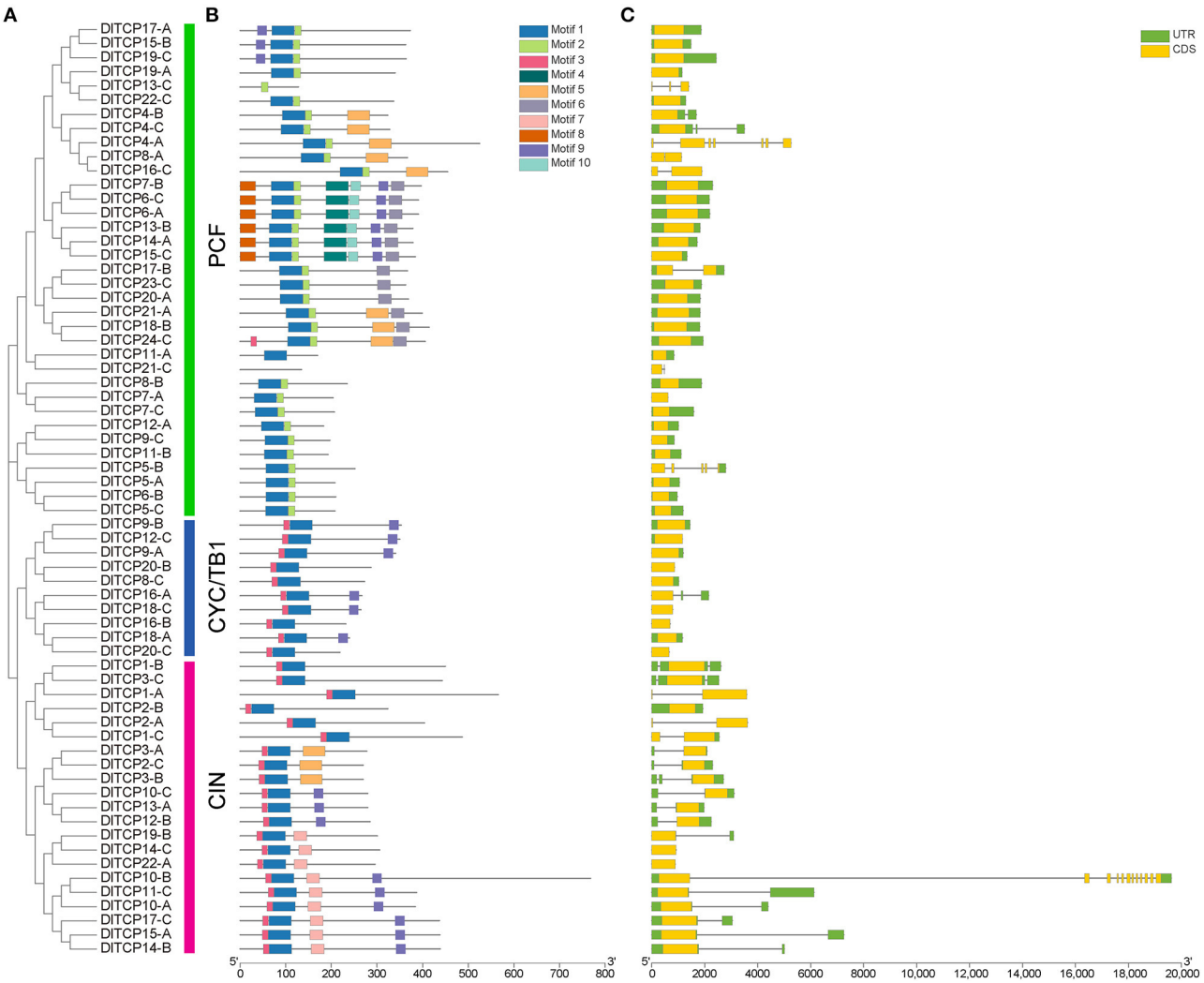


FIGURE 2 | Phylogenetic relationships, motif compositions, and gene structure of TCP genes in Ma bamboo. **(A)** Phylogenetic relationships of 66 TCP members in Ma bamboo. They were further classified into three subfamilies: PCF, CYC/TB1, and CIN. **(B)** Different motif compositions of TCP members in Ma bamboo were detected using MEME. The conserved motifs were represented by boxes with different colors. **(C)** Gene structures of TCPs in Ma bamboo. Green indicates 5′UTR and 3′UTR, yellow indicates exons, and black lines indicate introns.

indicating that *DlTCP5-C*, *DlTCP7-B*, *DlTCP9-A*, and *DlTCP23-C* all have transcriptional self-activation activity (**Figure 10**).

Cloning and Relevant Analysis of *DlTCP12-C*

DlTCP12-C, the homologous gene of *OsTB1*, is the key node gene for lateral bud outgrowth, which was significantly differentially expressed according to the transcriptome analysis. This suggests that *DlTCP12-C* could play an important role in the growth of bamboo shoots. Then, the full-length CDS of the *DlTCP12-C* was cloned. The *DlTCP12-C* gene has no intron and contains SP, TCP, and R conserved domains (**Figure 11**). The original transcriptional self-activating activity was lost after the deletion of 285 bp at the 5' end. The results of the expression pattern and subcellular localization showed that it was a nuclear-localized TCP transcription factor highly expressed in bamboo shoots.

Phenotypic Assay of the *DlTCP12-C* Overexpression Transgenic *Arabidopsis*

Due to the long period of genetic transformation of Ma bamboo, preliminary functional verification was performed in *Arabidopsis*. Transgenic *Arabidopsis* plants overexpressing *DlTCP12-C* were obtained, and the phenotype of the T3 generation *Arabidopsis* lines was observed after 35 days of culture (**Figure 12**). Compared with the wild type, the overexpressed lines had significantly fewer rosette leaves and branch numbers. Obviously, no new lateral branches were found in *DlTCP12-C* overexpression transgenic *Arabidopsis* after 35 days of cultivation, whereas there were typically 3–4 branches in WT lines, indicating that *DlTCP12-C*

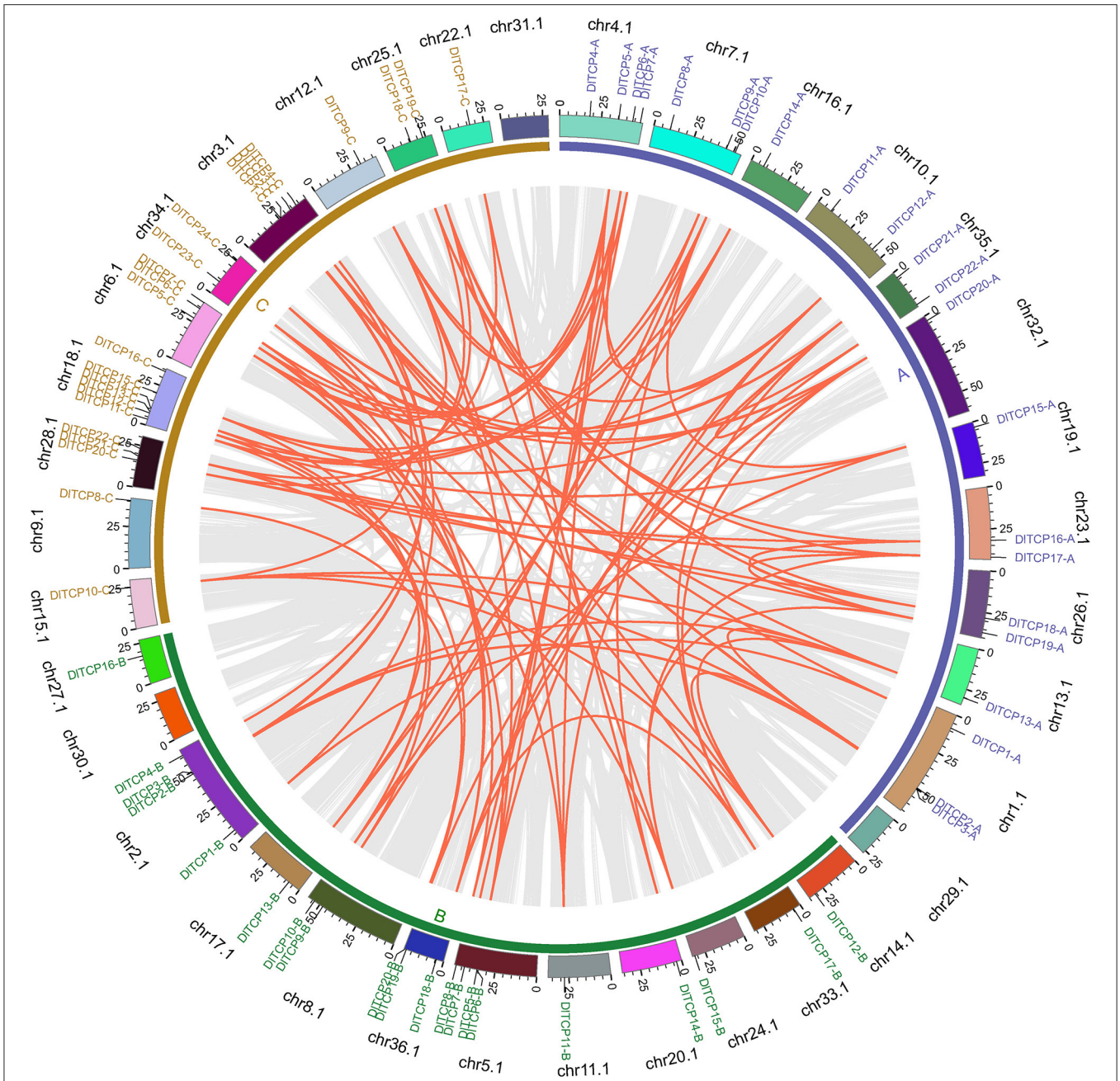


FIGURE 3 | Chromosome distribution and synteny analysis of TCP genes in Ma bamboo. Different chromosomes were showed in different colors. The inner gray lines were used to mark all the collinearity relationships in Ma bamboo and the inner red lines were represented the collinearity relationships of TCP members in Ma bamboo. Gene names of different colors were used to mark the TCP members from different subgenomes.

plays an important role in the development of branch outgrowth in *Arabidopsis*. The results of transgene identification and expression detection are shown in **Supplementary Figure S4**.

DISCUSSION

As we all know, members of gene families tend to expand through several evolutionary mechanisms, including tandem duplication, large-scale chromosome segmental duplication, and translocation, This leads to the evolution of complex phenotypes (Cannon et al., 2004; McCarthy et al., 2015). In Ma bamboo, numerous genes have undergone triple genome duplication due to whole-genome duplication events (Zheng et al., 2022). Approximately 280 gene families have undergone expansion during their evolution from paleotropical woody bamboos to allohexaploid Ma bamboo, which could be the reason for some of

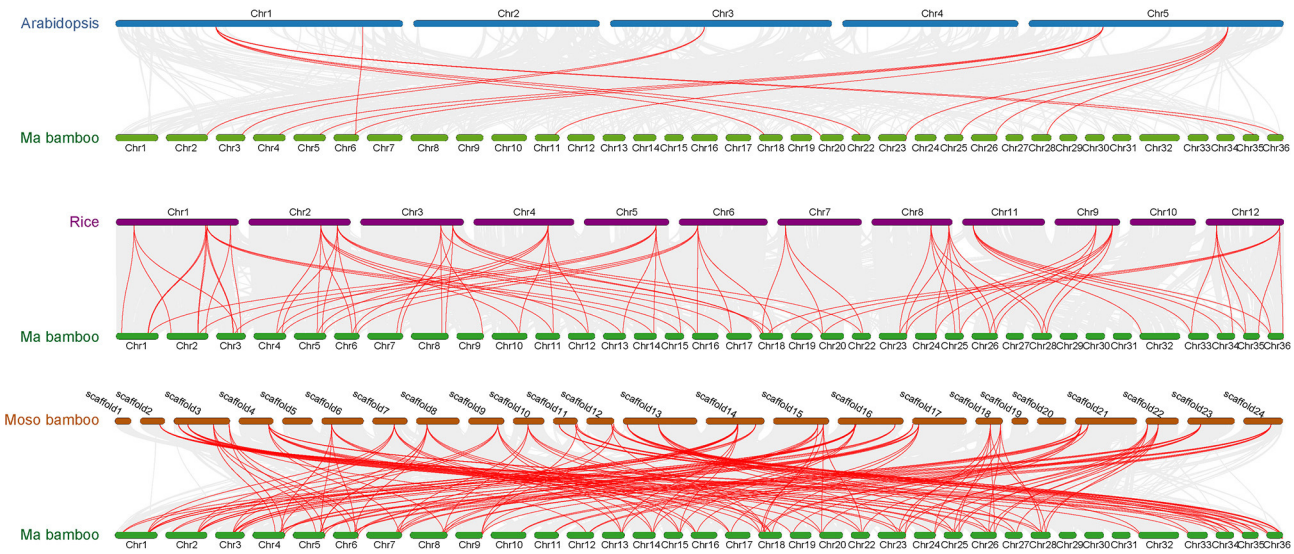


FIGURE 4 | Collinear relationships of genes pairs from Ma bamboo, Moso bamboo, rice, and *Arabidopsis*. Gray lines indicate that all the collinear gene pairs between Ma bamboo and other species, whereas red lines indicate the collinear TCP gene pairs.

bamboos' unique traits, including their rapid vegetative growth and high biomass (Zheng et al., 2022). In this study, a total of 66 plant-specific TCP transcription factors were identified in Ma bamboo, and the number of genes was about three times that of rice, Arabidopsis, and Moso bamboo (**Table 1**). In other words, all *DlTCPs* possessed 3–5 homologous genes, indicating that they are preferentially retained during polyploidization. Similarly, there were three closely related homologous genes in the TCP family of hexaploid wheat. However, their cis-acting elements were quite different, which could lead to the sub-functionalization of wheat homologous genes (Zhao et al., 2018). Collinear analysis demonstrated that the main reason for the expansion of the TCP gene family is chromosome polyploidization (whole-genome duplication) and large-scale chromosome segment duplication, which primarily occurs between three subgenomes A, B, and C of *D. latiflorus*, and only a few chromosomes do not possess TCP transcription factors (**Figure 3**). The TCP family has expanded due to large-scale segment duplication events that also occurred in upland cotton, whereas 74 *GhTCP* genes were identified in the allotetraploid plant upland cotton genome (AADD) (Li et al., 2017). Tandem duplication events, which are important events driving the occurrence of new biological functions, have only been found one time between *DlTCP5-B* and *DlTCP6-B*. This suggests that TCP transcription factors have conservative and irreplaceable functions in Ma bamboo (Shang et al., 2012). The TCP family could have a continually increasing role in the development of the typical hexaploid bamboo species *D. latiflorus*. While they may have functional redundancy, the sub-functionalization of these homologous genes and gene dosage could make Ma bamboo more adaptable during growth and development (He et al., 2022).

Plants have several complex regulatory mechanisms and signal networks, which can quickly perceive the external environment and regulate gene expression to adapt to unpredictable environmental changes and resist several biotic and abiotic stresses in the long-term evolutionary process. Transcription factors can regulate the occurrence of biological processes such as plant morphology, developmental patterns, and stress responses to varying degrees. As the plant-specific transcription factors, the TCP family plays a vital role in plant growth and development. A group of functionally redundant phylogenetic-related class I *TCP* genes (*AtTCP7*, *AtTCP8*, *AtTCP22,* and *AtTCP23*) had similar expression patterns in young leaves, regulating leaf development by controlling cell proliferation (Aguilar Martinez and Sinha, 2013). In *Arabidopsis*, *AtTCP14* directly activates the growth potential of the embryo during seed germination, whose expression level is highest before seed germination (Tatematsu et al., 2008). In addition, the *AtTCP14* mutant was highly sensitive to abscisic acid and gibberellin synthase inhibitors, indicating that *AtTCP14* regulates seed germination by regulating hormone response (Tatematsu et al., 2008; Manassero et al., 2013). The senescence phenotype of *TCP19* and *TCP20* double mutants was significantly enhanced in *Arabidopsis*, and classic genetic and molecular methods have been used to demonstrate that *TCP19* and *TCP20* are involved in controlling leaf senescence in *Arabidopsis*, despite functional redundancy (Danisman et al., 2013). *OsPCF7* is primarily expressed at the tillering stage and plays an important role in the tillering and heading process of rice seedlings. It significantly affects the panicle numbers, the number of filled grains per plant, and the grain yield per plant (Li et al., 2020). *CsTCP3* is induced by gibberellin, photoperiod, and temperature and directly participates in the development of axillary buds by controlling the content of auxin in axillary buds, which affects the number of lateral branches in cucumber (Wen et al., 2020). It can integrate upstream environmental factors and hormone

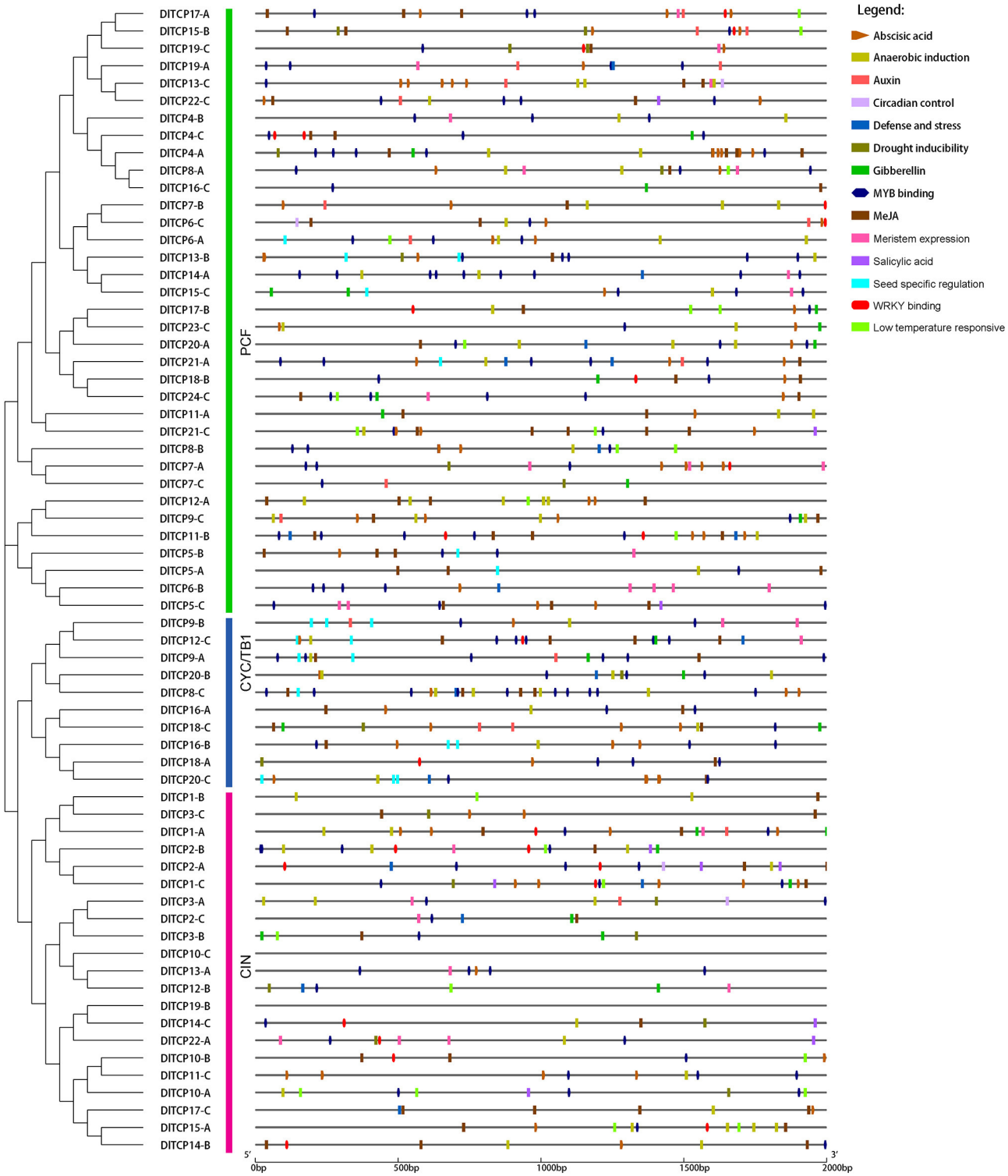


FIGURE 5 | Predicted cis-elements of TCP gene promoters in Ma bamboo. Plant CARE was used to predict and analyze the promoter region of the 2,000-bp upstream of 66 TCP members. Different colored rectangles represented different cis-elements and especial cis-elements were highlighted in different shapes.

signals and further affect the development of axillary meristem, to adapt the plant architecture to environmental conditions. Biological processes from shoot bud germination to growth is crucial for the growth and development of bamboo (Shou et al., 2020), all of which directly affect the yield of bamboo shoots and timber in Ma bamboo. Therefore, we collected

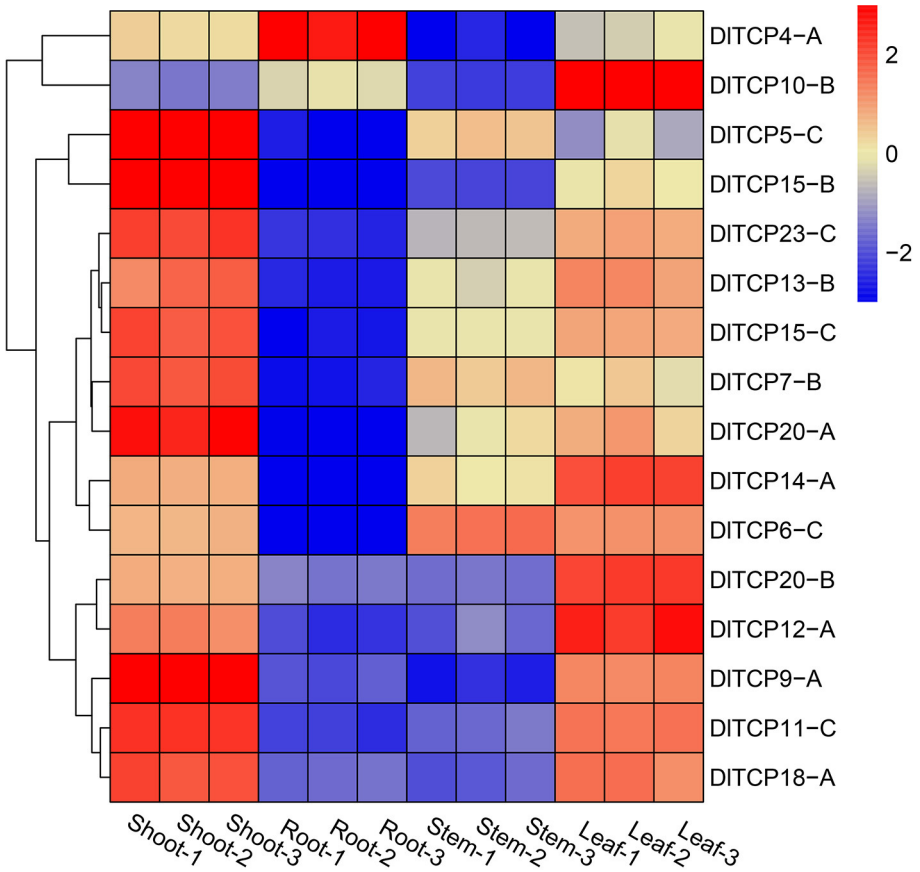


FIGURE 6 | Tissue expression pattern of TCP genes in Ma bamboo. The expression levels of putative TCP members in bamboo shoots, roots, stems, and leaves were normalized and visualized by R (4.0.2). Red and blue represent high and low expression levels, respectively. 1, 2, and 3 represent three biological replicates, respectively.

the apical buds of bamboo shoots from four representative developmental stages for transcriptome analysis to find out the relevant gene sets involved in this process (**Figure 7A**). A total of 29 TCP transcription factors were subsequently found as differentially expressed genes from RNA-seq data. Interestingly, TCP transcription factors exhibited spatiotemporal expression specificity during the development of bamboo shoots (**Figure 8A**). From the dormancy stage (S1) to the rapid high growth stage (S4), the number of downregulated genes gradually decreased; in contrast, the number of upregulated genes gradually increased (**Figure 8B**). These results suggested that there is an alternation and switching mechanism in the function of TCP transcription factors to better adapt to the growth and development of bamboo shoots in Ma bamboo. *TB1* is the key node gene for lateral bud outgrowth, which plays a conservative role in many species (Takeda et al., 2003; Dixon et al., 2018; Li et al., 2021). Whereafter, the preliminary functional verification results confirmed the critical role of *DlTCP12-C* in inhibiting axillary bud growth and lateral branch growth in overexpressed transgenic *Arabidopsis* (**Figure 12**). Future research will verify the biological function and regulation pathway of *DlTCP12-C* in shoot buds development using genetic transformation in Ma bamboo and assess whether there is functional redundancy among other members of the CYC/TB1 subfamily.

Transcription factors have the binding activity of specific DNA sequences or the characteristics of known DNA-binding domains, so they bind to cis-acting elements on the target site to ensure that the target gene is expressed at a specific intensity, in a specific time and place. In our study, a large number of cis-acting regulatory elements related to plant hormone signals, organ development, stress response, *MYB* transcription factors, and *WRKY* transcription factor-binding sites accumulated in the promoter of *DlTCPs*, indicating that TCP transcription factors could act as a central regulatory integrin regulated by environmental factors, hormone signals, and upstream transcription factors to affect plant growth and development. *GhTCP14a/22* is involved in controlling cotton fiber growth through the gibberellin, brassinosteroids, and auxin signal transduction pathways, which play a remarkable role in the development of cotton fiber and are primarily expressed during fiber initiation and elongation (Li et al., 2017). *DWARF27* (*D27*) is a key gene involved in strigolactone synthesis, which can sense strigolactone signaling and activate downstream TB1-like TCP transcription factors by recruiting

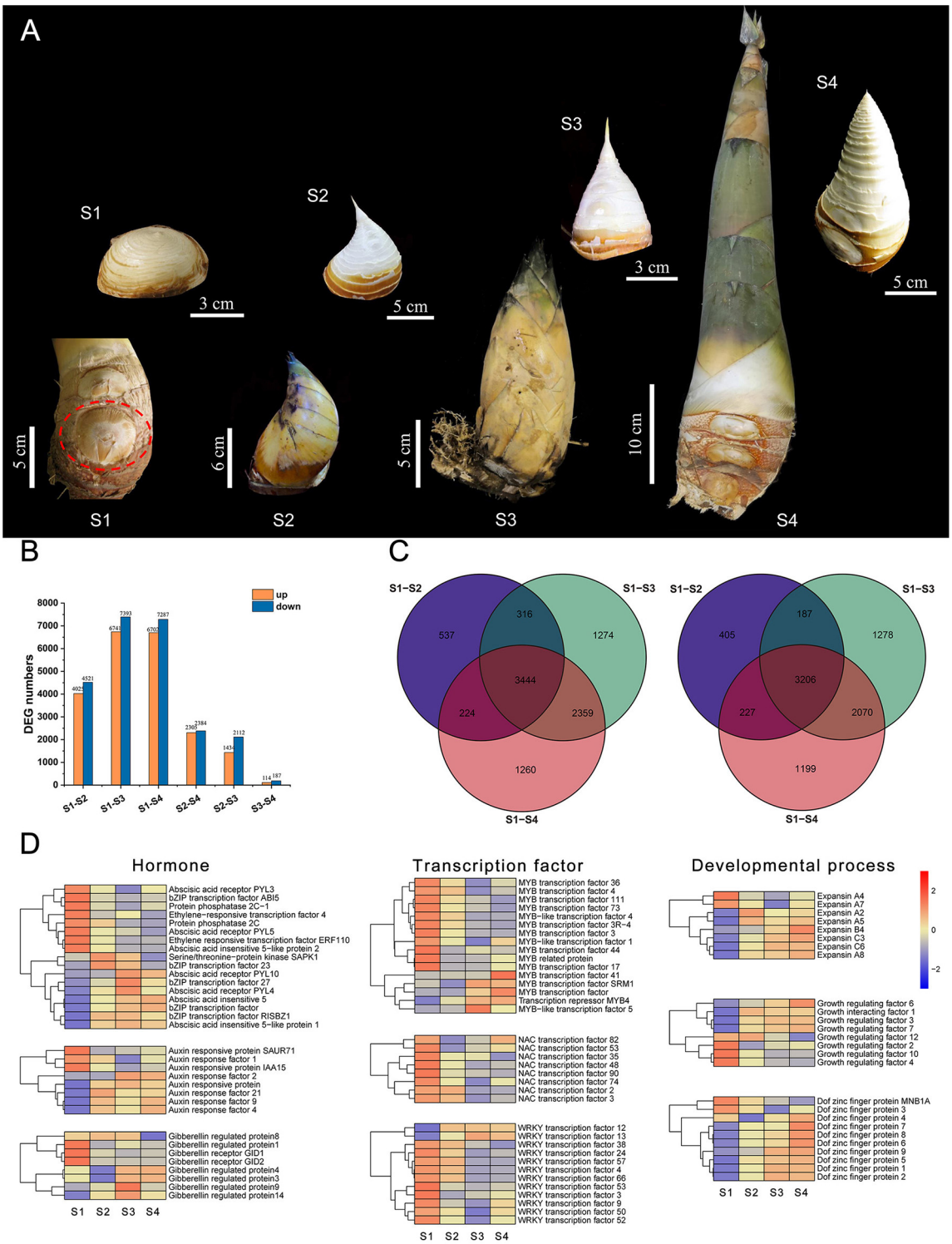


FIGURE 7 | Transcriptome analysis of shoot buds at different developmental stages in Ma bamboo. **(A)** A total of four representative developmental stages of Ma bamboo apical bud were characterized by the length of bamboo shoot. The red dotted line represents the position of shoot buds in S1. A separate scale of each image is shown separately. **(B)** The number of differentially expressed genes between different stages was counted by transcriptome analysis. Yellow and blue represent upregulated and downregulated differentially expressed genes, respectively. **(C)** Venn diagrams of differentially expressed genes in three group. Left and right represent upregulated and downregulated, respectively. **(D)** The expression levels of three kinds of differentially expressed genes in transcriptome data were related to hormones, transcription factors, and developmental process, respectively. Yellow and blue represent upregulated and downregulated differential expression, respectively.

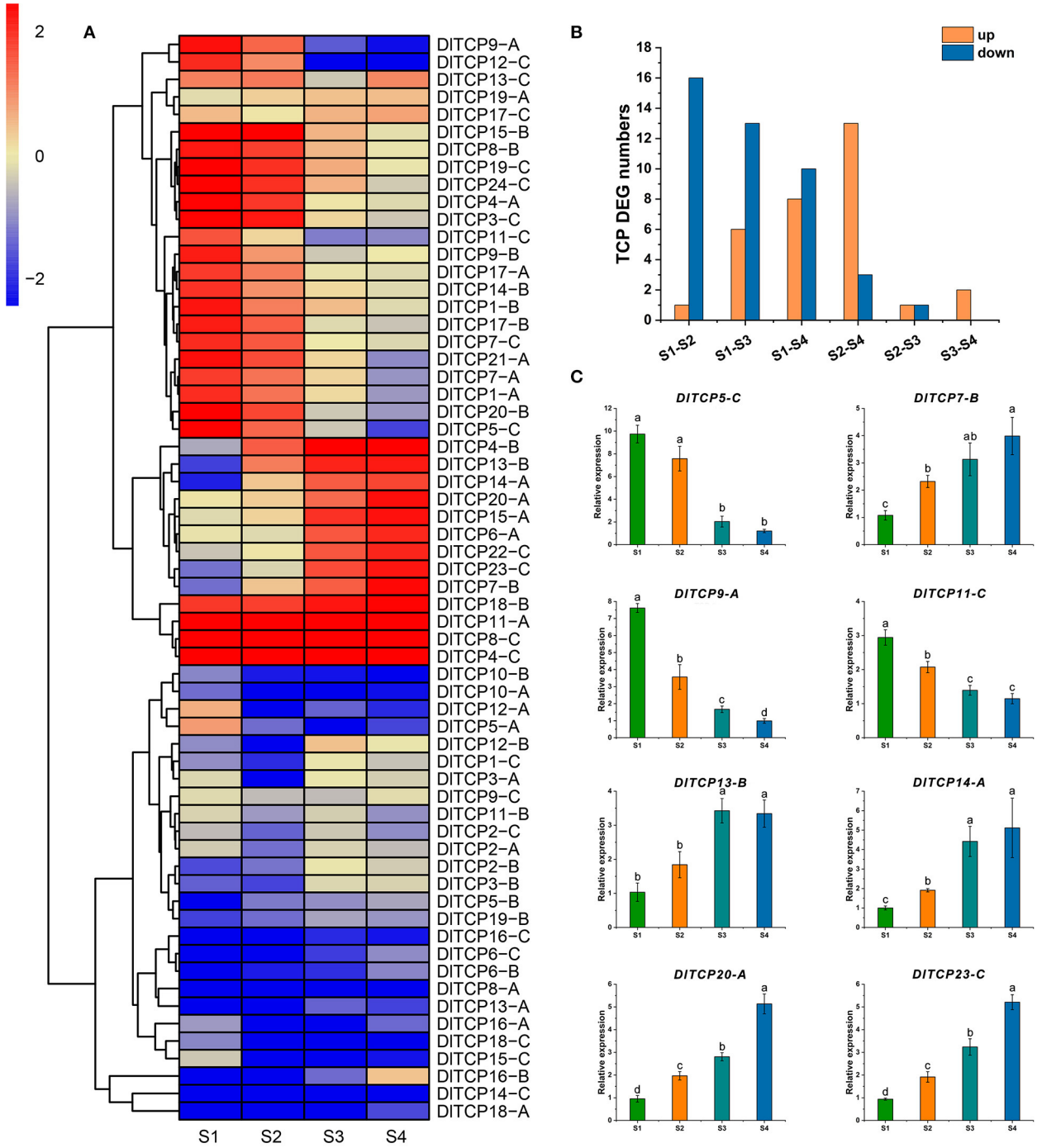


FIGURE 8 | Expression analysis of TCPs in shoot buds at different developmental stages. **(A)** The expression data of TCP family members in different developmental stages of bamboo shoot buds were retrieved from transcriptome analysis. The results were normalized and visualized by R (4.0.2). **(B)** The number of TCP differentially expressed genes between different stages. Yellow and blue represent upregulated and downregulated differential expression, respectively. **(C)** qRT-PCR was used to verify the results of the transcriptome analysis. Error bars were obtained from three replicates. Statistically significant differences between the expression level of different stage were analyzed by Student's *t*-test. Those with different marked letters were represented significantly different: $p < 0.05$.

SCF complexes to stimulate the ubiquitination and degradation of DWARF53 (D53) repressor proteins (Kerr and Beveridge, 2017). Cytokinins and sugars also inhibit the expression of *TB1* (Mason et al., 2014; Patil et al., 2021). Auxin upregulates the expression of MAX3 and MAX4 through the AXR1-AFB-mediated signaling pathway, but downregulates the members of the IPT family, promoting strigolactone biosynthesis, and the inhibition of cytokinin biosynthesis, which further

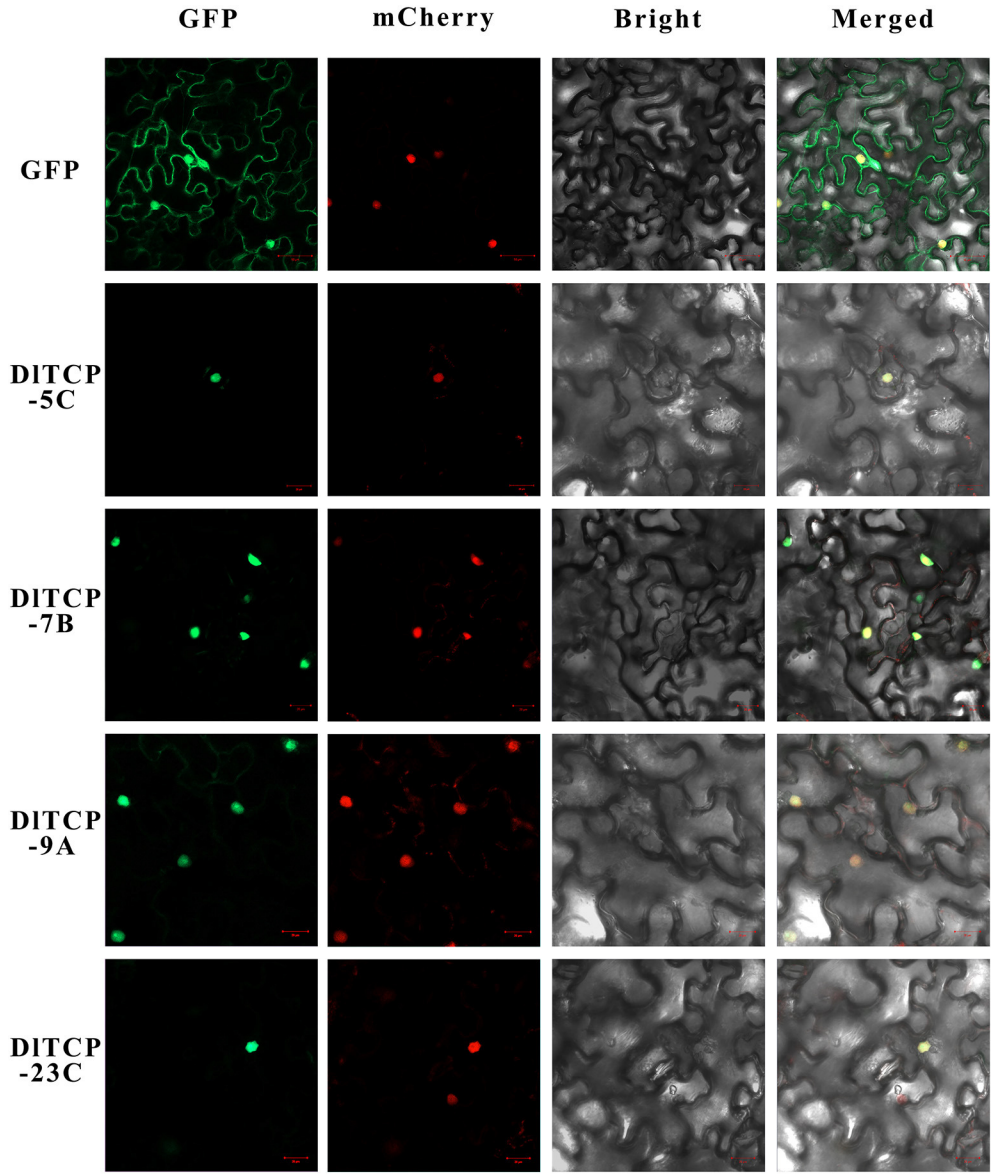


FIGURE 9 | Subcellular localization of four mGFP-fused TCP proteins in Ma bamboo. The four candidate TCP proteins (DlTCP5-C, DlTCP7-B, DlTCP9-A, and DlTCP23-C) and GFP as a control were transiently expressed in *Nicotiana benthamiana* leaves and observed under a fluorescence microscope. The nucleus was visualized with mCherry-labeled nuclear markers.

promotes the expression of *TB1* (Nordström et al., 2004; Tanaka et al., 2006). A number of two closely related TCP transcription factors *TCP14* and *TCP15* affect the development of foliage and trichomes, participate in cytokinin-regulated signal pathways, and stimulate the expression of cytokinin-regulated gene RESPONSE REGULATOR 5 through interaction with SPINDLY (SPY) (Steiner et al., 2012). Jasmonic acid is a kind of plant hormone of lipids (oxylipins), which is involved in plant development, abiotic stress response, and the interaction between plants and microorganisms. *AtTCP4* reportedly directly targets *LIPOXYGENASE2* (*LOX2*), encoding a chloroplast enzyme gene involved in α-linolenic acid biosynthesis and jasmonic acid synthesis, and is involved in the regulation of jasmonic acid biosynthesis and leaf development (Vick and Zimmerman, 1983; Danisman et al., 2012). The LsAP2 transcription factor further regulates leaf morphology in lettuce by inhibiting the activity of the CIN-like TCP transcription factor (Luo et al., 2021). The TCP-mediated complicated hormone signal regulatory network further emphasizes the important role of TCP in affecting plant growth patterns and biological processes, and more in-depth research is needed to clarify the pathway of TCP transcription factors. Our transcriptome and promoter element analyses indicate that TCP transcription factors regulate the expression of target genes through transcriptional regulation and hormone signals, which affect related biological processes in plants. Recent studies

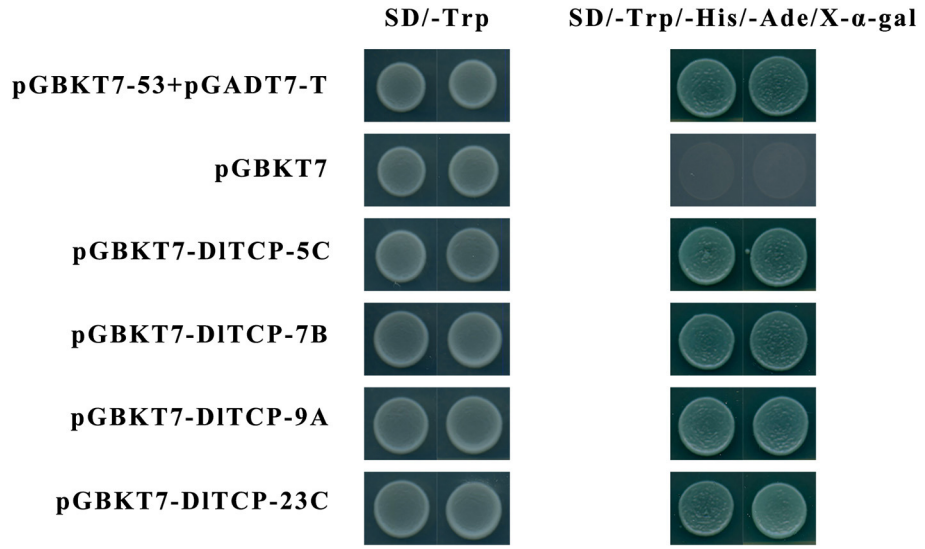


FIGURE 10 | Transactivational analyses of DlTCP proteins in yeast. The positive control pGBKT7-p53 + pGADT7-T, negative control pGBKT7 empty plasmid, and four candidate pGBKT7-DlTCPs plasmids were transformed into yeast AH109, and the strains were further cultured on the yeast medium of SD-Trp and SD-Trp-His-Ade supplemented with X-α-gal to analyze their transactivation activity.

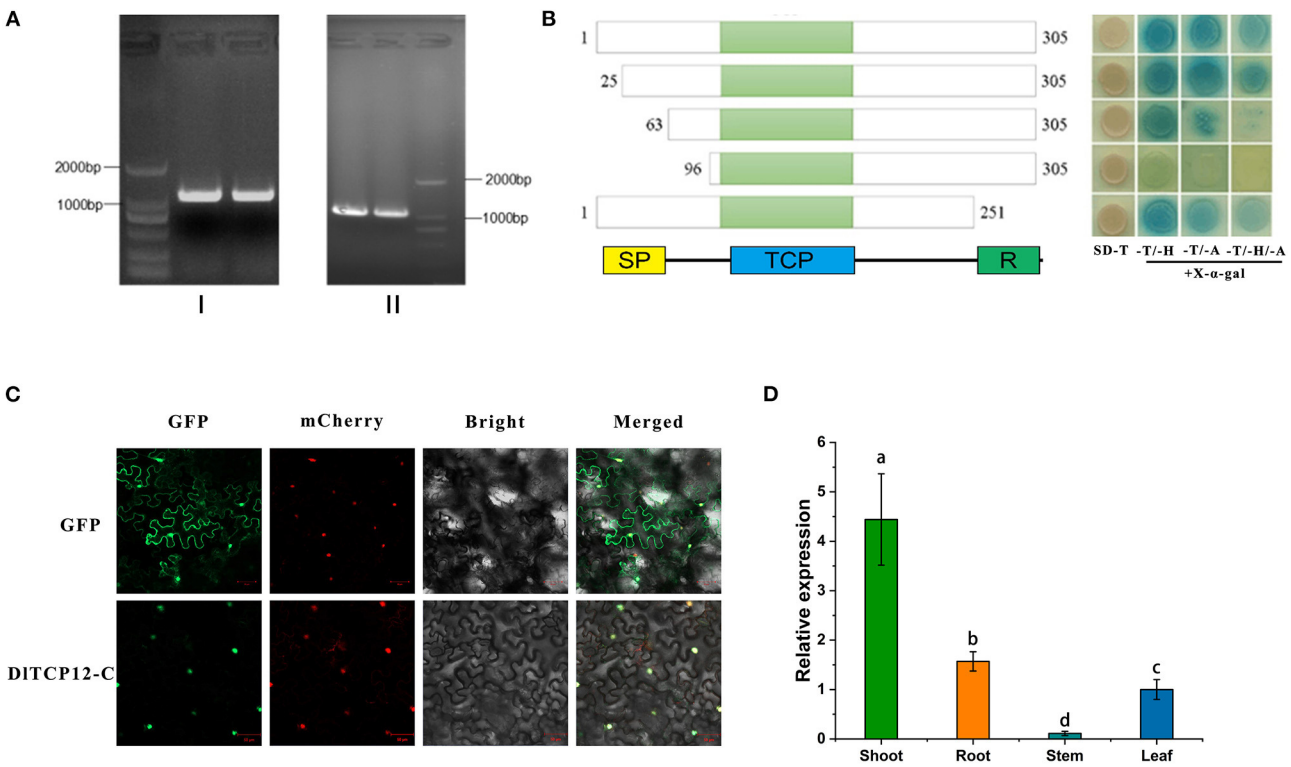


FIGURE 11 | Gene structure, transactivational analyses, subcellular localization, and expression patterns of the *DlTCP12-C*. **(A)** Using genomic DNA (I) and cDNA (II) as template, gel electrophoresis amplification results of *DlTCP12-C*. **(B)** Transcriptional self-activation experiments of five truncated forms of DlTCP12-C, SD-T, -T/-H, -T/-A, and -T/-H/-A represents SD medium lacking Trp, Trp and His, Trp and Ade, Trp, and His and Ade, respectively. **(C)** Subcellular localization of DlTCP12-C protein. **(D)** The expression levels of *DlTCP12-C* in bamboo shoot, root, stem, and leaf; those with different marked letters were represented significantly different: $p < 0.05$.

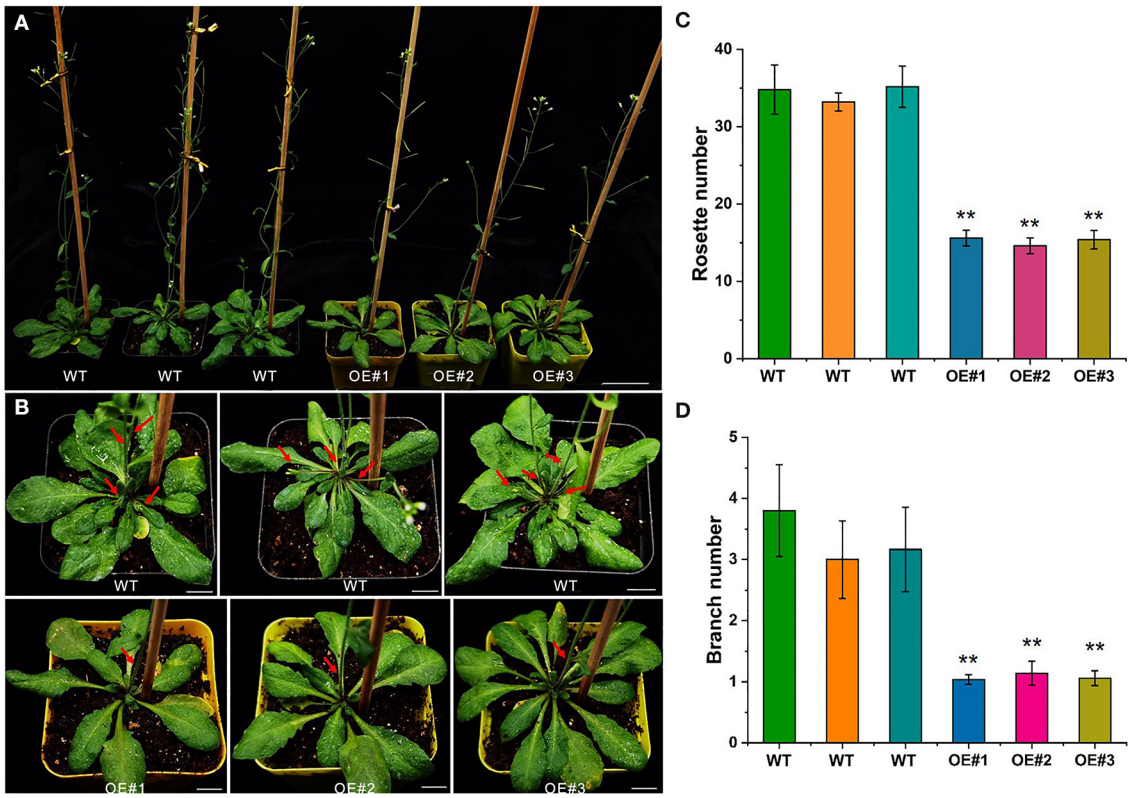


FIGURE 12 | Phenotypic assay of the *DlTCP12-C* overexpression transgenic *Arabidopsis*. **(A)** Branching phenotypes of 35-day-old wild-type and *DlTCP12-C* overexpressing transgenic *Arabidopsis* lines. Scale bar = 4 cm. **(B)** Close-up views of the rosettes of the plants in **(A)**. Red arrow indicates branch. Scale bar = 1 cm. **(C)** Number of rosettes. Error bars were obtained from five replicates. **(D)** Number of branches. Error bars were obtained from five replicates. Significant differences compared with the WT were analyzed by Student's *t*-test: $^{**}p < 0.01$.

have made clarified the mechanism of the direct targeting regulation of TCP transcription factors by miRNA. Some TCP members directly targeted by *miR319* are widely involved in plant hormone signal transduction, leaf development, vascular formation, and response to abiotic stress (Fang et al., 2021). Comprehensive analysis of catechin metabolism profiles and *TCP* gene expression profiles in different plant tissues at different developmental stages indicated that the *CsmiR319b*/*CsTCP3-4* module was not only related to shoot tip development, but also played a potential role in catechin biosynthesis in tea plants (Yu et al., 2021). Genetic and molecular analyses indicated that *PtoTCP20*, the direct target gene of *miR319a*, regulated the proliferation of vascular cambium along with *PtoWOX4a* and promoted the differentiation of secondary xylem by activating the transcription of *PtoWND6*, thereby regulating the secondary growth of *Populus tomentosa* stem (Hou et al., 2020). A total of five *TCP* genes were found to contain miR319 directly targeted binding sites in 3′ UTR (**Supplementary Table S6**). These miR319-targeted *DlTCPs* were the members of the CIN subfamily. The regulation mechanism of DlTCPs related to the bud growth of bamboo shoots requires further study.

In short, this study identified 66 plant-specific TCP transcription factors in the *D. latiflorus* genome using bioinformatics and analyzed their evolutionary relationship, duplication events, promoter cis-elements, tissue expression patterns, subcellular localization, and self-activating transcriptional activity. Transcriptome analysis of different developmental stages of bamboo shoot buds was used to preliminarily study the function of TCP transcription factors in Ma bamboo, providing a series of differentially expressed genes that could be involved in the growth and development of bamboo shoots. Subsequently, the conservative function of *DlTCP12-C*, which negatively regulates axillary bud development and lateral branch growth, was confirmed in overexpressed transgenic *Arabidopsis*. This comprehensive study of TCP transcription factors in Ma bamboo provides several candidate genes worthy of further analysis, including the regulatory mechanism of bamboo shoot bud growth and development.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: National Center for Biotechnology Information (NCBI) BioProject database under Accession Number PRJNA792061.

AUTHOR CONTRIBUTIONS

GQ and RZ conceived this project. KJ and YW designed experiments, interpreted the results, and wrote the manuscript. JX, ZL, HF, and BH performed the experiments and analyzed the data. YW provided technical guidance for the experiment. All authors read and approved the submission of this manuscript.

FUNDING

This research was funded by the Zhejiang Science and Technology Major Program on Agricultural New Variety Breeding, Grant Number 2021C02070-4.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.884443/full#supplementary-material

Supplementary Figure S1 | Multiple sequence alignment of TCP proteins in Ma bamboo.

Supplementary Figure S2 | Distribution of Ka and Ks from paralogous (*Dl-Dl*) and orthologous (*Dl-Os*, *Dl-Pe*) gene pairs.

Supplementary Figure S3 | GO annotation results of all differential genes and TCP members in the transcriptome.

Supplementary Figure S4 | The results of transgene identification and expression detection.

Supplementary Table S1 | The detailed information of TCP members in rice, *Arabidopsis thaliana*, and Moso bamboo.

Supplementary Table S2 | List of primer sequences used in this study.

Supplementary Table S3 | The detailed information of putative TCP members in Ma bamboo.

Supplementary Table S4 | The detailed information of conserved motifs in DlTCPs.

Supplementary Table S5 | Putative microRNA319-targeted binding sites of DlTCPs.

Supplementary Table S6 | The detailed information and Ka/Ks ratio of putative orthologous and paralogous pairs.

Supplementary Table S7 | The detailed information of three kinds of differentially expressed genes.

REFERENCES

Aguilar Martinez, J. A., and Sinha, N. R. (2013). Analysis of the role of Arabidopsis class I TCP genes AtTCP7, AtTCP8, AtTCP22, and AtTCP23 in leaf development. *Front. Plant Sci.* 4, 406. doi: 10.3389/fpls.2013.00406

Aguilar-Martínez, J. A., Poza-Carrión, C., and Cubas, P. (2007). Arabidopsis BRANCHED1 acts as an integrator of branching signals within axillary buds. *Plant Cell* 19, 458–472. doi: 10.1105/tpc.106.048934

An, J., Guo, Z., Gou, X., and Li, J. (2011). TCP1 positively regulates the expression of DWF4 in *Arabidopsis thaliana*. *Plant Signal. Behav.* 6, 1117–1118. doi: 10.4161/psb.6.8.15889

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4, 1–21. doi: 10.1186/1471-2229-4-10

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, H.-C., Chien, T.-C., Chen, T.-Y., Chiang, M.-H., Lai, M.-H., and Chang, M.-C. (2021). Overexpression of a novel ERF-X-type transcription factor, OsERF106MZ, reduces shoot growth and tolerance to salinity stress in rice. *Rice* 14, 1–18. doi: 10.1186/s12284-021-00525-5

Chen, L., Zhao, Y., Xu, S., Zhang, Z., Xu, Y., Zhang, J., et al. (2018). Os MADS 57 together with Os TB 1 coordinates transcription of its target Os WRKY 94 and D14 to switch its organogenesis to defense for cold adaptation in rice. *New Phytol.* 218, 219–231. doi: 10.1111/nph.14977

Cubas, P., Lauter, N., Doebley, J., and Coen, E. (1999). The TCP domain: a motif found in proteins regulating plant growth and development. *Plant J.* 18, 215–222. doi: 10.1046/j.1365-313x.1999.00444.x

Danisman, S., Van der Wal, F., Dhondt, S., Waites, R., de Folter, S., Bimbo, A., et al. (2012). Arabidopsis class I and class II TCP transcription factors regulate jasmonic acid metabolism and leaf development antagonistically. *Plant Physiol.* 159, 1511–1523. doi: 10.1104/pp.112.200303

Danisman, S., van Dijk, A. D., Bimbo, A., van der Wal, F., Hennig, L., de Folter, S., et al. (2013). Analysis of functional redundancies within the Arabidopsis TCP transcription factor family. *J. Exp. Botany* 64, 5673–5685. doi: 10.1093/jxb/ert337

Dixon, L. E., Greenwood, J. R., Bencivenga, S., Zhang, P., Cockram, J., Mellers, G., et al. (2018). TEOSINTE BRANCHED1 regulates inflorescence architecture and development in bread wheat (*Triticum aestivum*). *Plant Cell* 30, 563–581. doi: 10.1105/tpc.17.00961

Doebley, J., Stec, A., and Hubbard, L. (1997). The evolution of apical dominance in maize. *Nature* 386, 485–488. doi: 10.1038/386485a0

Fang, Y., Zheng, Y., Lu, W., Li, J., Duan, Y., Zhang, S., et al. (2021). Roles of miR319-regulated TCPs in plant development and response to abiotic stress. *Crop J.* 9, 17–28. doi: 10.1016/J.CJ.2020.07.007

Francis, A., Dhaka, N., Bakshi, M., Jung, K.-H., Sharma, M. K., and Sharma, R. (2016). Comparative phylogenomic analysis provides insights into TCP gene functions in Sorghum. *Sci. Rep.* 6, 1–13. doi: 10.1038/srep38488

González-Grandío, E., Pajoro, A., Franco-Zorrilla, J. M., Tarancón, C., Immink, R. G., and Cubas, P. (2017). Abscisic acid signaling is controlled by a BRANCHED1/HD-ZIP I cascade in Arabidopsis axillary buds. *Proc. Natl. Acad. Sci.* 114, E245–E254. doi: 10.1073/pnas.1613199114

Guo, Z., Fujioka, S., Blancaflor, E. B., Miao, S., Gou, X., and Li, J. (2010). TCP1 modulates brassinosteroid biosynthesis by regulating the expression of the key biosynthetic gene DWARF4 in *Arabidopsis thaliana*. *Plant Cell* 22, 1161–1173. doi: 10.1105/tpc.109.069203

He, F., Wang, W., Rutter, W. B., Jordan, K. W., Ren, J., Taagen, E., et al. (2022). Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat. *Nat. Commun.* 13, 1–15. doi: 10.1038/s41467-022-28453-y

Hou, J., Xu, H., Fan, D., Ran, L., Li, J., Wu, S., et al. (2020). MiR319a-targeted PtoTCP20 regulates secondary growth *via* interactions with PtoWOX4 and PtoWND6 in *Populus tomentosa*. *New Phytol.* 228, 1354–1368. doi: 10.1111/nph.16782

Kebrom, T. H., Burson, B. L., and Finlayson, S. A. (2006). Phytochrome B represses Teosinte Branched1 expression and induces sorghum axillary bud outgrowth in response to light signals. *Plant Physiol.* 140, 1109–1117. doi: 10.1104/pp.105.074856

Kerr, S. C., and Beveridge, C. A. (2017). IPA1: a direct target of SL signaling. *Cell Res.* 27, 1191–1192. doi: 10.1038/cr.2017.114

Kosugi, S., and Ohashi, Y. (1997). PCF1 and PCF2 specifically bind to cis elements in the rice proliferating cell nuclear antigen gene. *Plant Cell* 9, 1607–1619. doi: 10.1105/tpc.9.9.1607

Kosugi, S., and Ohashi, Y. (2002). DNA binding and dimerization specificity and potential targets for the TCP protein family. *Plant J.* 30, 337–348. doi: 10.1046/j.1365-313x.2002.01294.x

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Li, C., Potuschak, T., Colón-Carmona, A., Gutiérrez, R. A., and Doerner, P. (2005). Arabidopsis TCP20 links regulation of growth and cell division control pathways. *Proc. Natl. Acad. Sci.* 102, 12978–12983. doi: 10.1073/pnas.0504039102

Li, G., Tan, M., Ma, J., Cheng, F., Li, K., Liu, X., et al. (2021). Molecular mechanism of MdWUS2–MdTCP12 interaction in mediating cytokinin signaling to control axillary bud outgrowth. *J. Exp. Botany* 72, 4822–4838. doi: 10.1093/jxb/erab163

Li, W., Chen, G., Xiao, G., Zhu, S., Zhou, N., Zhu, P., et al. (2020). Overexpression of TCP transcription factor OsPCF7 improves agronomic trait in rice. *Mol. Breed.* 40, 1–13. doi: 10.1007/s11032-020-01129-5

Li, W., Li, D.-D., Han, L.-H., Tao, M., Hu, Q.-Q., Wu, W.-Y., et al. (2017). Genome-wide identification and characterization of TCP transcription factor genes in upland cotton (*Gossypium hirsutum*). *Sci. Rep.* 7, 1–14. doi: 10.1038/s41598-017-10609-2

Liu, H., Gao, Y., Wu, M., Shi, Y., Wang, H., Wu, L., et al. (2020). TCP10, a TCP transcription factor in moso bamboo (*Phyllostachys edulis*), confers drought tolerance to transgenic plants. *Environ. Exp. Botany* 172, 104002. doi: 10.1016/j.envexpbot.2020.104002

Liu, H.-L., Wu, M., Li, F., Gao, Y.-M., Chen, F., and Xiang, Y. (2018). TCP transcription factors in moso bamboo (*Phyllostachys edulis*): genome-wide identification and expression analysis. *Front. Plant Sci.* 9, 1263. doi: 10.3389/fpls.2018.01263

Liu, M., Jiang, J., Han, X., Qiao, G., and Zhuo, R. (2014). Validation of reference genes aiming accurate normalization of qRT-PCR data in *Dendrocalamus latiflorus* Munro. *PLoS ONE* 9, e87417. doi: 10.1371/journal.pone.0087417

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2– ΔΔCT method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Luo, C., Wang, S., Ning, K., Chen, Z., Wang, Y., Yang, J., et al. (2021). LsAP2 regulates leaf morphology by inhibiting CIN-like TCP transcription factors and repressing LsKAN2 in lettuce. *Horticult. Res.* 8:15. doi: 10.1038/s41438-021-00622-y

Luo, D., Carpenter, R., Vincent, C., Copsey, L., and Coen, E. (1996). Origin of floral asymmetry in Antirrhinum. *Nature* 383, 794–799. doi: 10.1038/383794a0

Manassero, N. G. U., Viola, I. L., Welchen, E., and Gonzalez, D. H. (2013). TCP transcription factors: architectures of plant form. *Biomol. Concepts* 4, 111–127. doi: 10.1515/bmc-2012-0051

Martín-Trillo, M., and Cubas, P. (2010). TCP genes: a family snapshot ten years later. *Trends Plant Sci.* 15, 31–39. doi: 10.1016/j.tplants.2009.11.003

Mason, M. G., Ross, J. J., Babst, B. A., Wienclaw, B. N., and Beveridge, C. A. (2014). Sugar demand, not auxin, is the initial regulator of apical dominance. *Proc. Natl. Acad. Sci.* 111, 6092–6097. doi: 10.1073/pnas.1322045111

McCarthy, E. W., Mohamed, A., and Litt, A. (2015). Functional divergence of APETALA1 and FRUITFULL is due to changes in both regulation and coding sequence. *Front. Plant Sci.* 6, 1076. doi: 10.3389/fpls.2015.01076

Min, Z., Chen, L., Zhang, Y., Li, Z., Liu, M., Li, W. P., et al. (2021). VvBRC inhibits shoot branching in grapevine. *Sci. Horticult.* 289, 110370. doi: 10.1016/j.scienta.2021.110370

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., et al. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126, 1203–1217. doi: 10.1016/j.cell.2006.07.031

Nicolas, M., and Cubas, P. (2016). TCP factors: new kids on the signaling block. *Curr. Opin. Plant Biol.* 33, 33–41. doi: 10.1016/j.pbi.2016.05.006

Nicolas, M., Rodríguez-Buey, M. L., Franco-Zorrilla, J. M., and Cubas, P. (2015). A recently evolved alternative splice site in the BRANCHED1a gene controls potato plant architecture. *Curr. Biol.* 25, 1799–1809. doi: 10.1016/j.cub.2015.05.053

Nordström, A., Tarkowski, P., Tarkowska, D., Norbaek, R., and Åstot, C., Dolezal, K., et al. (2004). Auxin regulation of cytokinin biosynthesis in *Arabidopsis thaliana*: a factor of potential importance for auxin–cytokinin-regulated development. *Proc. Natl. Acad. Sci.* 101, 8039–8044. doi: 10.1073/pnas.0402504101

Patil, S. B., Barbier, F. F., Zhao, J., Zafar, S. A., Uzair, M., Sun, Y., et al. (2021). Sucrose promotes D53 accumulation and tillering in rice. *New Phytol.* 234, 122–136. doi: 10.1111/nph.17834

Resentini, F., Felipo-Benavent, A., Colombo, L., Blazquez Rodriguez, M. A., Alabadí Diego, D., and Masiero, S. (2015). TCP14 and TCP15 mediate the promotion of seed germination by gibberellins in *Arabidopsis thaliana*. *Mol. Plant* 8, 482–485. doi: 10.1016/j.molp.2021.09.012

Richards, R. (2000). Selectable traits to increase crop photosynthesis and yield of grain crops. *J. Exp. Botany* 51(Suppl. 1), 447–458. doi: 10.1093/jexbot/51.suppl_1.447

Sarvepalli, K., and Nath, U. (2011). Interaction of TCP4-mediated growth module with phytohormones. *Plant Signal. Behav.* 6, 1440–1443. doi: 10.4161/psb.6.10.17097

Sarvepalli, K., and Nath, U. (2018). CIN-TCP transcription factors: transiting cell proliferation in plants. *Iubmb Life* 70, 718–731. doi: 10.1002/iub.1874

Shang, Q.-M., Li, L., and Dong, C.-J. (2012). Multiple tandem duplication of the phenylalanine ammonia-lyase genes in *Cucumis sativus* L. *Planta* 236, 1093–1105. doi: 10.1007/s00425-012-1659-1

Shen, J., Zhang, Y., Ge, D., Wang, Z., Song, W., Gu, R., et al. (2019). CsBRC1 inhibits axillary bud outgrowth by directly repressing the auxin efflux carrier CsPIN3 in cucumber. *Proc. Natl. Acad. Sci.* 116, 17105–17114. doi: 10.1073/pnas.1907968116

Shou, Y., Zhu, Y., and Ding, Y. (2020). Transcriptome analysis of lateral buds from *Phyllostachys edulis* rhizome during germination and early shoot stages. *BMC Plant Biol.* 20, 1–18. doi: 10.1186/s12870-020-02439-8

Sparkes, I. A., Runions, J., Kearns, A., and Hawes, C. (2006). Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nat. Protocols* 1, 2019–2025. doi: 10.1038/nprot.2006.286

Steiner, E., Efroni, I., Gopalraj, M., Saathoff, K., Tseng, T.-S., Kieffer, M., et al. (2012). The Arabidopsis O-linked N-acetylglucosamine transferase SPINDLY interacts with class I TCPs to facilitate cytokinin responses in leaves and flowers. *Plant Cell* 24, 96–108. doi: 10.1105/tpc.111.093518

Takeda, T., Suwa, Y., Suzuki, M., Kitano, H., Ueguchi-Tanaka, M., Ashikari, M., et al. (2003). The OsTB1 gene negatively regulates lateral branching in rice. *Plant J.* 33, 513–520. doi: 10.1046/j.1365-313x.2003.01648.x

Tanaka, M., Takei, K., Kojima, M., Sakakibara, H., and Mori, H. (2006). Auxin controls local cytokinin biosynthesis in the nodal stem in apical dominance. *Plant J.* 45, 1028–1036. doi: 10.1111/j.1365-313X.2006.02656.x

Tatematsu, K., Nakabayashi, K., Kamiya, Y., and Nambara, E. (2008). Transcription factor AtTCP14 regulates embryonic growth potential during seed germination in *Arabidopsis thaliana*. *Plant J.* 53, 42–52. doi: 10.1111/j.1365-313X.2007.03308.x

Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protocols Bioinform.* 2, 2–3. doi: 10.1002/0471250953.bi0203s00

Vick, B. A., and Zimmerman, D. C. (1983). The biosynthesis of jasmonic acid: a physiological role for plant lipoxygenase. *Biochem. Biophys. Res. Commun.* 111, 470–477. doi: 10.1016/0006-291x(83)90330-3

Wang, B., Smith, S. M., and Li, J. (2018). Genetic regulation of shoot architecture. *Ann. Rev. Plant Biol.* 69, 437–468. doi: 10.1146/annurev-arplant-042817-040422

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteom. Bioinform.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wang, H., Mao, Y., Yang, J., and He, Y. (2015). TCP24 modulates secondary cell wall thickening and anther endothecium development. *Front. Plant Sci.* 6, 436. doi: 10.3389/fpls.2015.00436

Wang, L., Wang, B., Yu, H., Guo, H., Lin, T., Kou, L., et al. (2020). Transcriptional regulation of strigolactone signalling in Arabidopsis. *Nature* 583, 277–281. doi: 10.1038/s41586-020-2382-x

Wang, M., Le Moigne, M.-A., Bertheloot, J., Crespel, L., Perez-Garcia, M.-D., Og,é, L., et al. (2019). BRANCHED1: a key hub of shoot branching. *Front. Plant Sci.* 10, 76. doi: 10.3389/fpls.2019.00076

Wang, R., Yang, X., Guo, S., Wang, Z., Zhang, Z., and Fang, Z. (2021). MiR319-targeted OsTCP21 and OsGAmyb regulate tillering and grain yield in rice. *J. Integr. Plant Biol.* 63, 1260–1272. doi: 10.1111/jipb.13097

Wang, S.-t., Sun, X.-l., Hoshino, Y., Yu, Y., Jia, B., and Sun, Z.-w., et al. (2014). MicroRNA319 positively regulates cold tolerance by targeting OsPCF6 and OsTCP21 in rice (*Oryza sativa* L.). *PLoS ONE* 9, e91357. doi: 10.1371/journal.pone.0091357

Wang, Y., and Li, J. (2008). Molecular basis of plant architecture. *Annu. Rev. Plant Biol.* 59, 253–279. doi: 10.1146/annurev.arplant.59.032607.092902

Wen, H., Chen, Y., Du, H., Zhang, L., Zhang, K., He, H., et al. (2020). Genome-wide identification and characterization of the TCP gene family in cucumber (*Cucumis sativus* L.) and their transcriptional responses to different treatments. *Genes* 11, 1379. doi: 10.3390/genes11111379

Xu, D., Jin, K., Jiang, H., Gong, D., Yang, J., Yu, W., et al. (2022). GFAP: ultra-fast and accurate gene functional annotation software for plants. *bioRxiv*. [preprint]. doi: 10.1101/2022.01.05.475154

Xu, D., Lu, Z., Jin, K., Qiu, W., Qiao, G., Han, X., et al. (2021). SPDE: a multi-functional software for sequence processing and data extraction. *Bioinformatics* 37, 3686–3687. doi: 10.1093/bioinformatics/btab235

Xu, H., Lantzouni, O., Bruggink, T., Benjamins, R., Lanfermeijer, F., Denby, K., et al. (2020). A molecular signal integration network underpinning Arabidopsis seed germination. *Curr. Biol.* 30, 3703–3712.e3704. doi: 10.1016/j.cub.2020.07.012

Yao, X., Ma, H., Wang, J., and Zhang, D. (2007). Genome-wide comparative analysis and expression pattern of TCP gene families in *Arabidopsis thaliana* and *Oryza sativa*. *J. Integr. Plant Biol.* 49, 885–897. doi: 10.1111/j.1672-9072.2007.00509.x

Yu, S., Li, P., Zhao, X., Tan, M., Ahmad, M. Z., Xu, Y., et al. (2021). CsTCPs regulate shoot tip development and catechin biosynthesis in tea plant (*Camellia sinensis*). *Horticul. Res.* 8, 104. doi: 10.1038/s41438-021-00538-7

Yuan, Z., Gao, S., Xue, D.-W., Luo, D., Li, L.-T., Ding, S.-Y., et al. (2009). RETARDED PALEA1 controls palea development and floral zygomorphy in rice. *Plant Physiol.* 149, 235–244. doi: 10.1104/pp.108.128231

Zhang, C., Ding, Z., Wu, K., Yang, L., Li, Y., Yang, Z., et al. (2016). Suppression of jasmonic acid-mediated defense by viral-inducible microRNA319 facilitates virus infection in rice. *Mol. Plant* 9, 1302–1314. doi: 10.1016/j.molp.2016.06.014

Zhang, W., Cochet, F., Ponnaiah, M., Lebreton, S., Mathéron, L., Pionneau, C., et al. (2019a). The MPK 8-TCP 14 pathway promotes seed germination in Arabidopsis. *Plant J.* 100, 677–692. doi: 10.1111/tpj.14461

Zhang, W., Tan, L., Sun, H., Zhao, X., Liu, F., Cai, H., et al. (2019b). Natural variations at TIG1 encoding a TCP transcription factor contribute to plant architecture domestication in rice. *Mol. Plant* 12, 1075–1089. doi: 10.1016/j.molp.2019.04.005

Zhao, H., Zhao, S., Bamboo, International Network for Bamboo and Rattan, Fei, B., Liu, H., et al. (2017). Announcing the Genome Atlas of Bamboo and Rattan (GABR) project: promoting research in evolution and in economically and ecologically beneficial plants. *GigaScience* 6, gix046. doi: 10.1093/gigascience/gix046

Zhao, J., Zhai, Z., Li, Y., Geng, S., Song, G., Guan, J., et al. (2018). Genome-wide identification and expression profiling of the TCP family genes in spike and grain development of wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 9, 1282. doi: 10.3389/fpls.2018.01282

Zheng, Y., Yang, D., Rong, J., Chen, L., Zhu, Q., He, T., et al. (2022). Allele-aware chromosome-scale assembly of the allopolyploid genome of hexaploid Ma Bamboo (*Dendrocalamus latiflorus* Munro). *J. Integr. Plant Biol.* 64, 649–670. doi: 10.1111/jipb.13217

Zou, X., Du, M., Liu, Y., Wu, L., Xu, L., Long, Q., et al. (2021). CsLOB1 regulates susceptibility to citrus canker through promoting cell proliferation in citrus. *Plant J.* 106, 1039–1057. doi: 10.1111/tpj.15217

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 20 May 2022
doi: 10.3389/fpls.2022.897843

Organization, Phylogenetic Marker Exploitation, and Gene Evolution in the Plastome of *Thalictrum* (Ranunculaceae)

Kun-Li Xiang[1,2], Wei Mao[3], Huan-Wen Peng[2,4], Andrey S. Erst[5,6], Ying-Xue Yang[1], Wen-Chuang He[1]* and Zhi-Qiang Wu[1,7]**

[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China, [2]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China, [3]College of Ecology and Environment, Hainan University, Haikou, China, [4]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China, [5]Central Siberian Botanical Garden, Russian Academy of Sciences, Novosibirsk, Russia, [6]Laboratory Herbarium (TK), Tomsk State University, Tomsk, Russia, [7]Kunpeng Institute of Modern Agriculture at Foshan, Chinese Academy of Agricultural Sciences, Foshan, China

OPEN ACCESS

Edited by:
Jianyu Zhou,
Nankai University, China

Reviewed by:
Xiaoguo Xiang,
Nanchang University, China
Weishu Fan,
Kunming Institute of Botany (CAS), China

***Correspondence:**
Zhi-Qiang Wu
wuzhiqiang@caas.cn
Wen-Chuang He
hewenchuang@caas.cn
Ying-Xue Yang
yyxue32@163.com

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *16 March 2022*
Accepted: *11 April 2022*
Published: *20 May 2022*

Citation:
Xiang K-L, Mao W, Peng H-W, Erst AS, Yang Y-X, He W-C and Wu Z-Q (2022) Organization, Phylogenetic Marker Exploitation, and Gene Evolution in the Plastome of Thalictrum (Ranunculaceae). Front. Plant Sci. 13:897843. doi: 10.3389/fpls.2022.897843

Thalictrum is a phylogenetically and economically important genus in the family Ranunculaceae, but is also regarded as one of the most challengingly difficult in plants for resolving the taxonomical and phylogenetical relationships of constituent taxa within this genus. Here, we sequenced the complete plastid genomes of two *Thalictrum* species using Illumina sequencing technology *via de novo* assembly. The two *Thalictrum* plastomes exhibited circular and typical quadripartite structure that was rather conserved in overall structure and the synteny of gene order. By updating the previously reported plastome annotation of other nine *Thalictrum* species, we found that the expansion or contraction of the inverted repeat region affect the boundary of the single-copy regions in *Thalictrum* plastome. We identified eight highly variable noncoding regions—*infA-rps8*, *ccsA-ndhD*, *trnS*^{UGA}*-psbZ*, *trnH*^{GUG}*-psbA*, *rpl16-rps3*, *ndhG-ndhI*, *ndhD-psaC*, and *ndhJ-ndhK*—that can be further used for molecular identification, phylogenetic, and phylogeographic in different species. Selective pressure and codon usage bias of all the plastid coding genes were also analyzed for the 11 species. Phylogenetic relationships showed *Thalictrum* is monophyly and divided into two major clades based on 11 *Thalictrum* plastomes. The availability of these plastomes offers valuable genetic information for accurate identification of species and taxonomy, phylogenetic resolution, and evolutionary studies of *Thalictrum*, and should assist with exploration and utilization of *Thalictrum* plants.

Keywords: *Thalictrum*, plastid genome, genome structure, molecular markers, phylogeny

INTRODUCTION

Thalictrum L., comprising *ca.* 200 species, is a phylogenetically and economically important genus in the family Ranunculaceae (Tamura, 1995) and is worldwide with main distribution in northern temperate regions. *Thalictrum* plants are rich in benzylisoquinoline-derived alkaloids; at least 250 such compounds have been isolated from 60 species, and most of them show strong biological activities (Zhu and Xiao, 1991). *Thalictrum* plants are used in folk medicine for the treatment of many kinds of diseases by various ethnic groups of China, which has a long history (Wang and Xiao, 1979; Zhu and Xiao, 1989; Wu et al., 1998; Wang et al., 2001). In some place, roots of *Thalictrum* were used as substitutes for *Rhizoma coptidis* to treat enteritis and dysentery (Wu et al., 1998). Furthermore, bearing luxuriant foliage, extended branches, and attractive flowers, *Thalictrum* species have previously been mainly applied as perennial garden plants. At present, the horticultural values of *Thalictrum* plants, such as *Thalictrum delavayi*, *Thalictrum reniforme*, and *Thalictrum grandiflorum* have been widely paid attention with great commercial prospects (Wang and Xiao, 1979).

Thalictrum is taxonomically and phylogenetically regarded as one of the most challengingly difficult taxa in plants. Traditionally, *Thalictrum* was classified into 14 sections based on morphological traits such as leaf, flower, and fruit characteristics (Tamura, 1995). Molecular phylogenetic analyses have consistently suggested only that *Thalictrum* is a monophyletic group containing two major clades, based on the nuclear ribosomal internal transcribed spacer (ITS) region (ITS1, ITS2, and 5.8S) and the chloroplast DNA (cpDNA) *rpl16* intron (Soza et al., 2012). Then, a revised phylogeny yielded better resolution based on nuclear ribosomal ITS region, external transcribed spacer (ETS) region, and the cpDNA 3'*trnV-ndhC* (*trnV-ndhC*) intergenic region (Soza et al., 2013). Nonetheless, none of the sections traditionally assigned to the genus (Tamura, 1995) are monophyletic (Soza et al., 2012, 2013). Moreover, numerous species and varieties in *Thalictrum* are poorly defined owing to insufficient field studies and lack of consistent characteristics for diagnostic methods in the literature (Wang et al., 2001). Therefore, further exploiting more stable genetic variations and effective molecular markers in *Thalictrum* species is greatly important for conservation and utilization of the plants from this genus.

The popularity of the ITS region for infrageneric studies within angiosperms is well-known (Baldwin et al., 1995; Hughes et al., 2006; Mort et al., 2007). Levels of ITS sequence divergence within *Thalictrum* are relatively high (Soza et al., 2012, 2013). However, *Thalictrum* exhibits an enormous range of ploidy, from $2n = 2x = 14$ to $2n = 24x = 168$ (Löve, 1982; Tamura, 1995), with very small chromosomes known as the T-type in Ranunculaceae (Langlet, 1927). In *Thalictrum*, the ITS region is often presented as more than one copy (Soza et al., 2012, 2013). Owing to their haploidy, maternal inheritance, and high conservation in gene content and genome structure, the plastomes have been popular in researches on evolutionary relationships at almost any taxonomic level in plants. Although sequence divergence among the interspecific cpDNAs is generally less than ITS (Hughes et al., 2006; Mort et al., 2007), it is necessary to utilize cpDNA regions that exhibit relatively high rates of substitution in *Thalictrum*. With the advent of high-throughput sequencing technologies, it is now more practical and inexpensive to obtain plastome sequences and to upgrade cp-based phylogenetics to phylogenomics.

In the present study, we sequenced the complete plastid genomes of two *Thalictrum* species by using the next-generation sequencing platform and performed the first comprehensive analysis of *Thalictrum* plastomes by combining these data with previously reported plastomes of other nine species (Park et al., 2015; He et al., 2019, 2021b; Morales-Briones et al., 2019). Our study aims were as follows: (1) to investigate global structural patterns of the 11 *Thalictrum* plastomes; (2) to identify the most variable regions of these plastomes as prospective DNA barcodes for future species identification; (3) to choose more effective molecular markers *via* reconstruction of phylogenetic relationships among the 11 *Thalictrum* species using various makers; and (4) to test for the presence of adaptive evolution in all genes located in the two single-copy regions, and one of the two inverted-repeat (IR) regions by analyses of selective pressure and codon usage bias. The results will provide abundant information for further species identification, phylogenetic, and phylogeographic studies on *Thalictrum*, and will assist in exploration and utilization of *Thalictrum* plants.

MATERIALS AND METHODS

Sample Preparation, Sequencing, Assembly, and Annotation

The sequenced two *Thalictrum* species (*Thalictrum minus* var. *hypoleucum* and *Thalictrum simplex*) are growing in the Beijing Botanical Garden, Beijing, China. Genomic DNA was extracted from fresh leaves and purified using the Tiangen Isolation/Extraction/Purification Kit [Tiangen Biotech (Beijing) Co., Ltd.]. Short insert of 300–500 bp libraries were prepared for sequencing on the Illumina HiSeq X-Ten platform.

Before assembly of the short reads, plastome original reads were extracted by mapping all short reads to the nine plastomes as reference with BWA (Li and Durbin, 2009) and SAMtools (Danecek et al., 2021). Then the two plastomes were *de novo* assembled with SPAdes v3.15.2 (Bankevich et al., 2012) as described in He et al. (2021a). Highly accurate annotation of organelle genomes was performed by using the Organellar Genome GeSeq tool (Tillich et al., 2017) with subsequent manual correction. Three chloroplast genomes from *Thalictrum coreanum* (GenBank accession No. NC_026103), *Thalictrum minus* (NC_041544), and *Thalictrum thalictroides* (NC_039433) were used as reference sequences. The circular plastomes were visualized by using OGDRAW v1.3.1 (Greiner et al., 2019), with subsequent manual editing. We also updated the annotation of plastomes for the other 11 species in this study.

Detection and Annotation for Plastid Genomic Variations

Multiple sequence alignments of whole plastome sequences from the 11 *Thalictrum* species that have the representatives

of the two major clades of this genus in previous studies (Soza et al., 2012, 2013), as well as *Paraquilegia anemonoides* and *Leptopyrum fumarioides* in Thalictreae as outgroups were implemented using MAFFT v7 (Katoh and Toh, 2010) with standard parameters, and further adjusted manually in Geneious v8.0.4 (Kearse et al., 2012). For comparison, the gene order and structure of the 13 plastomes were compared by using IRscope.[1]

To identifying hypervariable regions, the sequence alignment of *Thalictrum* plastomes without outgroups was subjected to a sliding window analysis in DNAsp v6.12.03 (Rozas et al., 2017) to evaluate nucleotide diversity (π) of all genes, genes without introns, and intergenic spacer (IGS) regions. Functional annotations for the nucleotide variations were conducted by using snpEff v5.1 (Cingolani, 2012).

Phylogenetic Analysis

Phylogenetic analyses of *Thalictrum* were performed with maximum likelihood (ML) method in RAxML v8.2.11 (Stamatakis, 2014) with 1,000 replicates under GTRGAMMA model. The analyses were carried out based on the following nine data sets, including the complete plastid DNA sequences, concatenation of 115 IGS regions, concatenation of 114 gene sequences, and six genes and/or their introns and spacers (*rpl16* intron, *ndhC-trnV*^{UAC}, *ndhA* intron, *trnL*^{UAA}-*trnF*^{GAA}, *rpl32-trnL*^{UAG}, and *rbcL*) that have been employed in previous studies on *Thalictrum* (Soza et al., 2012, 2013; Wang et al., 2019).

Selective Pressure Analysis

Selective pressures were detected throughout the phylogenetic tree of *Thalictrum* for each plastid gene. Nonsynonymous (d_N) and synonymous (d_S) substitution rates of each plastid gene were assessed by using the CODEML program in PAML v4.9 (Yang, 2007). We tested different hypotheses *via* branch models, H0: the one-ratio model (m0), assumes the same d_N/d_S ratio (ω ratio) for all branches in the phylogeny, HA: the free-ratio model (m1) that assumes an independent ω ratio for each branch. Likelihood ratio tests were used to test each model's fit. The double log-likelihood difference between the two models (2ΔL) was compared to a chi-square distribution with N−1 degrees of freedom, where N is the number of branches in the phylogeny (Whelan and Goldman, 1999).

Codon Usage Analysis

The program DNAsp v6.12.03 (Rozas et al., 2017) was used to examine the synonymous codon usage of 79 protein-coding genes in the plastome of *Thalictrum* and to calculate several related parameters such as the effective number codons (ENC), codon bias index (CBI), and relative synonymous codon usage (RSCU). The ENC and CBI are often used to evaluate codon bias at the level of an individual gene (Frank, 1990). RSCU is the observed codon frequency divided by the expected frequency. An RSCU value close to 1.0 indicates that the deviation is not significant (Sharp et al., 1986). Amino acid (AA) frequency was calculated as the percentage of codons encoding the same amino acid divided by the total codons.

[1]https://irscope.shinyapps.io/irapp/

RESULTS

Genome Features

The 11 plastomes of the *Thalictrum* species ranged in size from 154,924 bp (*T. thalictroides*) to 156,258 bp (*T. minus* var. *hypoleucum*). All these plastomes displayed the typical quadripartite structure of nearly all land plants, consisting of a pair of inverted repeats (IRs, 26,273–26,521 bp) separated by a single-copy (LSC) region (84,733–85,700 bp) and a small single-copy (SSC) region (17,479–17,655 bp; **Table 1**). The average GC content was ~38.39%, which is almost identical with each other among the 11 complete *Thalictrum* plastomes. In the IR region, the GC content (43.22%) was found to be much higher than that in the LSC (36.62%) and SSC regions (32.45%). Although overall genomic structure including gene number and gene order were well-conserved (**Figure 1**), the 11 *Thalictrum* plastomes exhibited obvious differences in the IR-SC boundary regions (**Figure 2**). The gene *ycf1* spanned the SSC-IR$_B$ region while a pseudogene fragment ψ*ycf1* was located at the IR$_A$ region with a length range of 1,144–1,152 bp. The gene *rps19* spanned the LSC-IR$_A$ region and a pseudogene fragment ψ*rps19* (100–122 bp) was located in the IR$_B$ region of all *Thalictrum* species except *T. thalictroides*. At the junction of IR$_A$ and SSC regions in most species, the distance between ψ*ycf1* and *ndhF* ranged from 0 to 752 bp, except for that of *Thalictrum foeniculaceum* with an overlap region of 39 bp between ψ*ycf1* and *ndhF*. At the junction of IR$_B$ and LSC regions, the distances between ψ*rps19* and *trnH* ranged from 42 to 81 bp.

All the 11 plastomes each identically encoded 131 predicted functional genes and three pseudo genes, of which seven protein-coding genes, seven tRNA genes, four rRNA genes, and two pseudo genes were duplicated in the IR regions (**Figure 1**). Two introns were detected in each of two protein-coding genes (*clpP* and *ycf3*) while a single intron was detected in each of 11 protein-coding genes (*atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl2*, *rpl16*, *rpoC1*, *rps12*, *rps16*, and *ycf15*) and six tRNA genes (*trnA*^{UGC}, *trnG*^{UCC}, *trnI*^{GAU}, *trnK*^{UUU}, *trnL*^{UAA}, and *trnV*^{UAC}; **Supplementary Table S1**). Among 79 protein-coding genes, 75 contained standard AUG as the initiation codon, while three genes (*ndhD*, *rps19*, and *ycf15*) contained GUG instead, and the *rpl2* started with ACG.

Polymorphic Variation and Hypervariable Regions

Nucleotide variations among the complete plastid genomes of the 11 *Thalictrum* species were identified to elucidate the level of sequence divergence (**Figure 3**). The aligned matrix of the 11 *Thalictrum* plastomes (159,334 bp) contained 2,957 single-nucleotide polymorphisms (SNPs) and 1,016 insertion-deletions (indels). The vast majority of SNPs from coding genes were functionally silent (synonymous), while 594 SNPs (43.8%) and six SNPs (0.4%), from altogether 79 coding genes, were missense

TABLE 1 | Summary of characteristics of plastome sequnences used in the study.

Species	GenBank numbers	Total genome size (GC content)	LSC size (GC content)	IR size (GC content)	SSC size (GC content)	No. total gene (unique gene)	No. protein-coding gene (unique gene)	No. tRNA gene (unique gene)	No. rRNA gene (unique gene)	No. pseudo gene
Thalictrum aquilegiifolium L.	MZ442608	156,253 (38.35%)	85,695 (36.55%)	26,480 (43.23%)	17,598 (32.41%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum baicalense Turcz. ex Ledeb.	MW133265	155,859 (38.39%)	85,258 (36.63%)	26,482 (43.22%)	17,637 (32.41%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum coreanum H. Lév.	NC_026103	155,088 (38.44%)	84,733 (36.68%)	26,403 (43.25%)	17,549 (32.49%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum foeniculaceum Bunge	NC_053570	155,923 (38.34%)	85,323 (36.57%)	26,486 (43.21%)	17,628 (32.30%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum foliolosum DC.	MZ196217	155,764 (38.46%)	85,086 (36.71%)	26,521 (43.22%)	17,636 (32.58%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum minus var. *hypoleucum* (Siebold & Zucc.) Miq.	OM501079	156,258 (38.35%)	85,700 (36.55%)	26,480 (43.23%)	17,598 (32.41%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum petaloideum L.	MK253449	155,876 (38.42%)	85,326 (36.64%)	26,480 (43.23%)	17,590 (32.55%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum simplex L.	OM501080	156,211 (38.36%)	85,662 (36.56%)	26,481 (43.22%)	17,587 (32.46%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum tenue Franch.	MK253448	156,103 (38.37%)	85,507 (36.59%)	26,504 (43.2%)	17,588 (32.43%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Thalictrum thalictroides (L.) A. J. Eames & B. Boivin	NC_039433	154,924 (38.43%)	84,899 (36.66%)	26,273 (43.26%)	17,479 (32.5%)	133 (114)	86 (79)	37 (30)	8 (4)	2
Thalictrum viscosum W. T. Wang & S. H. Wang	MZ442609	155,984 (38.38%)	85,339 (36.63%)	26,495 (43.2%)	17,655 (32.36%)	134 (114)	86 (79)	37 (30)	8 (4)	3
Leptopyrum fumarioides (L.) Rchb.	NC_041542	157,448 (38.41%)	84,907 (36.41%)	27,821 (43.34%)	16,899 (32.19%)	133 (113)	86 (79)	37 (30)	8 (4)	2
Paraquilegia microphylla (Royle) J. R. Drumm. & Hutch.	NC_041479	164,383 (38.87%)	84,925 (36.62%)	30,979 (43.77%)	17,500 (32.42%)	134 (114)	86 (79)	37 (30)	8 (4)	3

and nonsense variations (**Supplementary Table S2**). A total of 549 simple sequence repeats (SSRs) were identified in the 11 *Thalictrum* plastomes with a range of 39 (*Thalictrum petaloideum*) to 60 (*Thalictrum baicalense*) SSRs were detected in each species (**Supplementary Table S3**), indicating rich polymorphism of the SSRs among plastomes of different species. The SSC regions showed the highest nucleotide diversity ($\pi = 0.01381$), followed by the LSC ($\pi = 0.00803$) and IR ($\pi = 0.00154$) regions. In the 114 unique genes, the nucleotide diversity for each locus ranged from 0 (e.g., *rps7*, *rrn16*, and *trnC*^{GCA}) to 0.02608 (*infA*) with an average of 0.00438, whereby 10 regions (i.e., *infA*, *rpl32*, *ycf1*, *rpl20*, *ccsA*, *rpl22*, *rpl16*, *rps15*, *rps16*, and *accD*) had remarkably high values ($\pi > 0.0096$; **Supplementary Table S1**; **Figure 3A**). For exons in genes, the nucleotide diversity ranged from 0 (e.g., *rps7*, *rrn16*, and *trnA-UGC*) to 0.02608 (*infA*) with an average of 0.00373, while for the 115 IGS regions it ranged from 0 (e.g., *atpE-atpB*, *rpl23-trnI*^{CAU}, *rrn16-trnI*^{GAU}, and *trnI*^{GAU}-*trnA*^{UGC}) to 0.03486 (*rpoC1-rpoB*) with an average of 0.01025, except for the *rpoC1-rpoB* ($\pi > 0.07171$) with a targetable sequence of only 5 bp. Additionally, 10 of those regions showed considerably high values ($\pi > 0.0217$; i.e., *ndhF-rpl32*, *infA-rps8*, *ccsA-ndhD*, *rpl32-trnL*^{UAG}, *trnS*^{UGA}-*psbZ*, *trnH*^{GUG}-*psbA*, *rpl16-rps3*, *ndhG-ndhI*, *ndhD-psaC*, and *ndhJ-ndhK*; see **Supplementary Table S4; Figure 3B**).

Phylogenetic Analysis

Three datasets, the whole complete plastid genome sequences, IGS regions, and gene sequences were constructed to investigate the phylogenetic relationships among the 11 *Thalictrum* species, with *P. anemonoides* and *Leptopyrum fumarioides* as two outgroups. By using ML method, three phylogenetic trees were built based on the three respective datasets, whose topologies were found to be highly concordant between one another (**Figures 4A–C**). The *Thalictrum* was strongly supported as a monophyletic group [bootstrap support (bs) = 100%], and contained two major clades that are strongly supported as sister groups: clades I (bs = 100%) and II (bs = 100%;

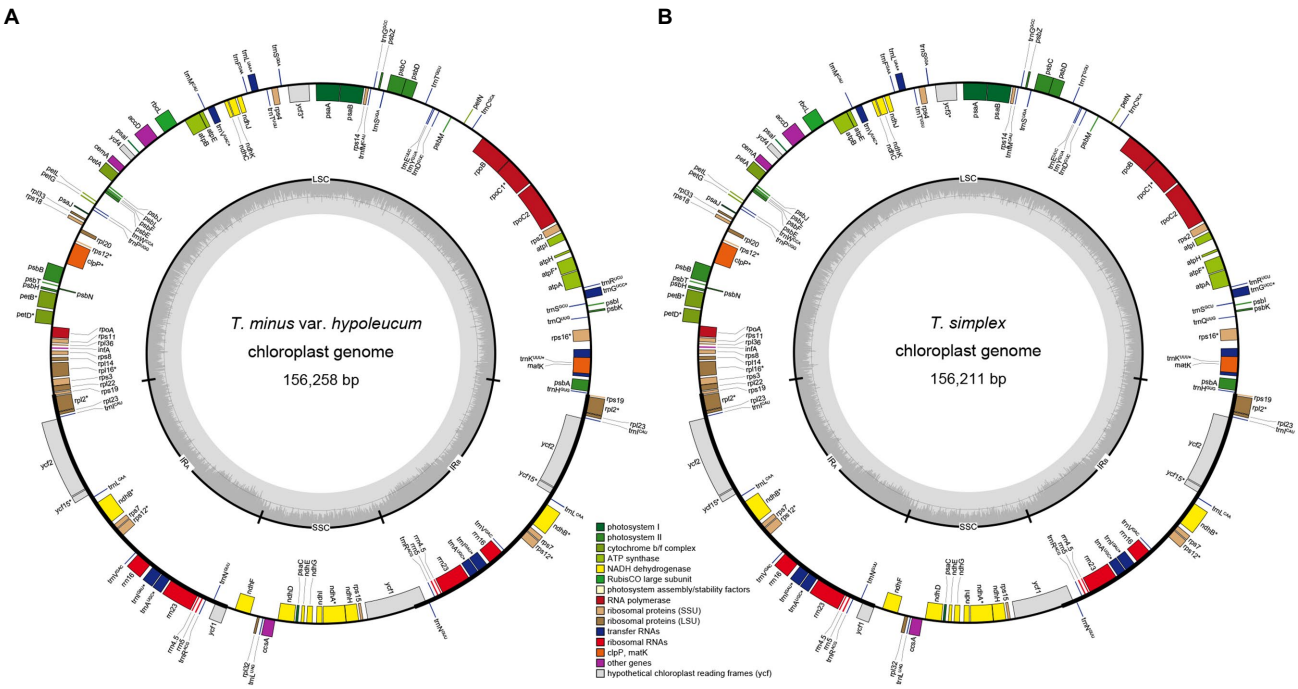


FIGURE 1 | Plastome of *Thalictrum minus* var. *hypoleucum* **(A)** and *Thalictrum simplex* **(B)**. The genes inside and outside of the circle are transcribed in clockwise and counterclockwise directions, respectively. Genes belonging to different functional groups are shown in different colors. The thick lines indicate the extent of the inverted repeats (IR_A and IR_B) that separate the genomes into small single-copy (SSC) and large single-copy (LSC) regions.

Figures 4A–C). The resolution of previously used six molecular fragments was also evaluated for *Thalictrum* species. Five genes and/or their introns and spacers yielded similar results except for the *rpl32-trnL*^{UAG} (**Figures 4D–I**). However, different supporting values were observed from the nodes based on different sequence dataset. For example, two nodes in clades II derived from the dataset of gene sequences both showed weaker supports (bs = 54% and bs = 68%; **Figure 4C**) than those derived from complete plastid genome sequences (bs = 100% and bs = 95%; **Figure 4A**) and IGS regions (bs = 100% and bs = 95%; **Figure 4B**). Additionally, the *rpl16* intron had the strongest support within clades II (**Figure 4D**), while *rbcL* had the weakest support in them (**Figure 4I**). These results indicated a much stronger resolving power of complete plastid genome sequences as well as IGS and intron regions as compared to the exon regions, which may serve as a reliable source of phylogenetic information in *Thalictrum*.

Selective Pressure and Codon Usage Analysis

Selective pressure analysis was conducted for CDS of all the 79 plastid protein-coding genes. A total of 66 genes are fit of m1 model in which *atpF* showed the highest ω ratio (1.13) except for *rpl23* ($\omega = 999$), while other 13 genes (*psbL*, *psaC*, *rps12*, *rps19*, *petB*, *psbN*, *psbF*, *psaJ*, *psbE*, *rpl36*, *psbZ*, *petN*, and *rps7* are fit of m0 model; **Table 2**). Among the 66 genes, most (50/66) were located in LSC region following by IR (7/66) and SSC (9/66) regions. The values of ω are significantly different ($p < 0.05$) between *Thalictrum* species for *ndhG* (SSC), *petA* (LSC), and *rpl22* (LSC) gene based on likelihood ratio tests, within some species have positive selection (e.g., *ndhG* in *T. coreanum*, *T. foeniculaceum*, *Thalictrum foliolosum*, and *T. thalictroides*; *petA* in *T. minus* var. *hypoleucum*). No genes in IR regions were detected significantly different between different species. However, 12 genes (LSC: *atpF*, *rpl33*, *rpl20*, *rps16*, *rps18*, *petG*, *rpl2*, *petL*, *psbJ*, *psbM*; IR: *rpl23*; SSC: *rps15*) were subject to positive selection in most species (median of $\omega > 1$; see **Table 2**; **Supplementary Tables S6, S7**), although their values of ω are not significantly different between different species.

We further analyzed the codon usage bias of the 79 protein coding genes in the plastomes of the 11 *Thalictrum* species. Most codons (55/64) were found to be used without bias or with only a slight bias ($0.5 \leq RSCU \leq 1.5$) in the protein-coding genes (**Supplementary Table S8**). The effective number of codons (ENC) and codon bias index (CBI) of all the 79 genes varied within a wide range, e.g., from 25.02 to 61.00 and from 0.28 to 0.85, respectively, with a median value of 49.0 and 0.50, respectively (**Figure 5**; **Supplementary Figure S1**; **Supplementary Table S8**). The data indicated that these genes were probably expressed in different levels due to their different usage frequencies of the rare and optimal codons, although they are all highly conserved in the plastomes. Most genes in SSC region (80.0%) showed relatively strong bias in the codon usage

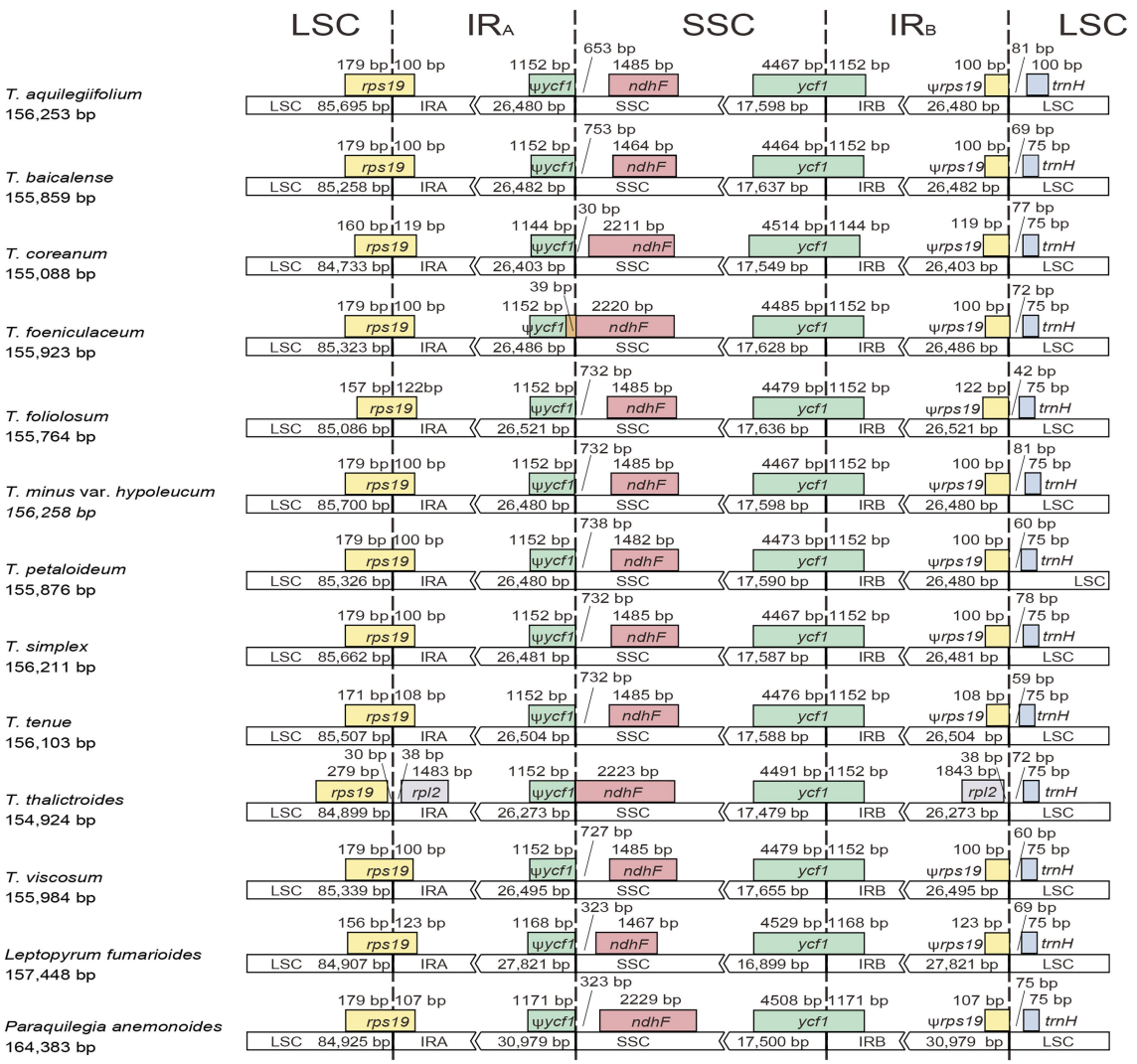


FIGURE 2 | Comparison of LSC, inverted-repeats (IRs), and SSC junction positions among *Thalictrum* plastomes.

($ENC \leq ENC_{median} = 49.0$ or $CBI \geq CBI_{median} = 0.5$), while 67.2% of genes in LSC region and 50.0% of genes in IR region performed relatively strong codon usage bias. Notably, almost all genes under positively selective pressures in more than half species performed relatively strong bias in the codon usage ($ENC \leq ENC_{median} = 49.0$ or $CBI \geq CBI_{median} = 0.5$), e.g., *atpF* featured a relatively strong codon usage bias with a low ENC of 41.61. This finding suggested that those important genes with higher expression levels may played important roles in the evolution and divergence of *Thalictrum* plastomes.

DISCUSSION

Plastome Characteristics of *Thalictrum*

In the present study, complete plastome sequences were firstly assembled for *T. minus* var. *hypoleucum* and *T. simplex* in the *Thalictrum* genus, with a total length of 156,211 and 156,258 bp, respectively (**Table 1**). The two plastomes are also highly similar in overall structure and gene order when compared to the majority of previously published plastomes of other nine species in *Thalictrum* (Park et al., 2015; He et al., 2019, 2021b; Morales-Briones et al., 2019). However, there was obvious variation in the IR-SC boundary regions among the 11 *Thalictrum* plastomes (**Figure 2**). The variations in IR-SC boundary regions in the 11 *Thalictrum* plastomes led to their length variation of the four regions and whole genome sequences. The expansion and contraction of the IR-SC boundary regions was considered as a primarily mechanism causing the length variation of angiosperm plastomes (Kim and Lee, 2004). In general, such expansions or contractions of the IRs into or out of adjacent single-copy regions are frequently observed in angiosperm plastomes (e.g., Yang et al., 2016; Zhang et al., 2016; Ye et al., 2018).

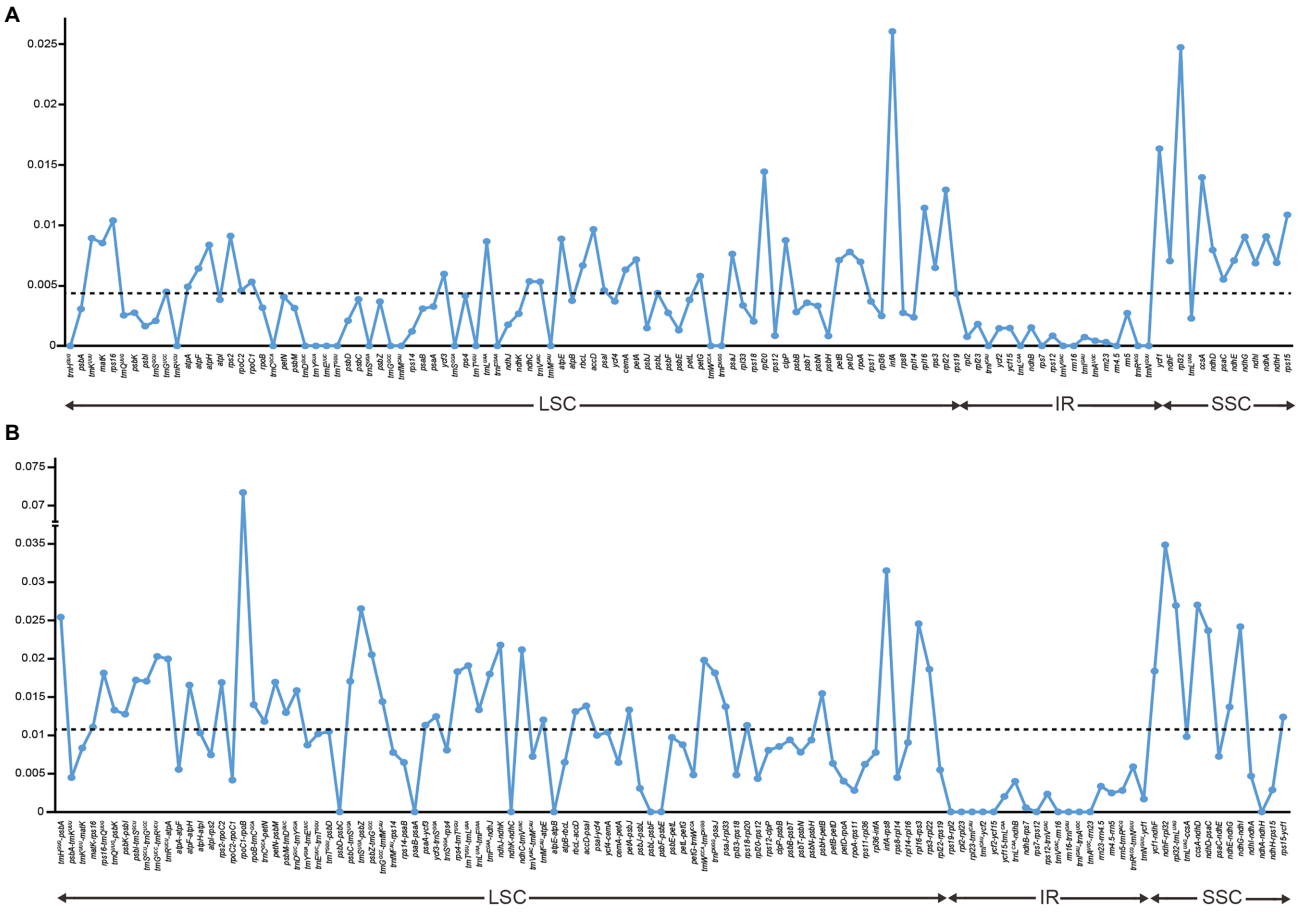


FIGURE 3 | Comparison of nucleotide variability (π) values in *Thalictrum* plastomes. **(A)** Pi values among genes, **(B)** Pi values among intergenic spacer (IGS) regions. The break in the middle of the bars indicated that other regions and genes are omitted here. The dot line denoted the average value.

Nonetheless, there are particular genes, especially *ycf1*, *rps19*, *ndhF*, *ycf15*, and ψ*rpl32*, which deserve closer scrutiny. For instance, in various members of *Thalictrum*, *ycf1* is duplicated, with a shorter copy (ψ*ycf1*, 1,144–1,152 bp) and a larger copy (*ycf1*, 5,616–5,658 bp) located at the SSC-IR$_A$ and SSC-IR$_B$ boundaries, respectively (**Figure 2**). Similarly, the *rps19* is present as two copies including ψ*rps19* (100–122 bp) and *rps19* (279 bp) at the SSC-IR$_B$ and SSC-IR$_A$ boundaries respectively except in *T. thalictroides* (**Figure 2**). Both shorter copies apparently resulted from incomplete duplication. Similar pseudogenizations of *ycf1* and locations of ψ*ycf1* copies are known from other plants (Yang et al., 2013, Szczecińska and Sawicki, 2015; Ye et al., 2018), and two copies of *rps19* have been found in Podophylloideae (Berberidaceae; Ye et al., 2018). As for the *ndhF*, the coding sequence was unexpectedly terminated by a stop-codon-gained event caused by nucleotide variation of a poly-A region in eight *Thalictrum* species except for *T. coreanum*, *T. foeniculaceum*, and *T. thalictroides*. For the *ycf15*, an intact copy and an interrupted gene have been found in other plants, with lengths of c. 150–300 bp (Raubeson et al., 2007, Shi et al., 2013). By contrast, an interrupted *ycf15* gene has been annotated in the sequenced chloroplast genomes in *Thalictrum* species. Additionally, ψ*rpl32* is incomplete because the *rpl32* gene was found to be transferred to the nucleus in the ancestor of the subfamily Thalictroideae (Park et al., 2015).

Regarding the initiation codon, *ndhD*, *rps19*, and *ycf15* used GUG, while *rpl2* used ACG in *Thalictrum*. The ACG codon may be restored to the canonical start codon (AUG) by RNA editing (Hoch et al., 1991; Takenaka et al., 2013), whereas GUG has been detected in in other plastomes (Kuroda et al., 2007; Gao et al., 2009; Zhang et al., 2016).

Noncoding Regions as a Source of Phylogenetic Information in *Thalictrum*

Given that the nuclear-genome coded ITS region is often presented as more than one copy in *Thalictrum*, sequences of cpDNA intergenic spacers have been employed to uncover intraspecific variability in *Thalictrum* (Soza et al., 2012, 2013). The IRs usually showed lower sequence divergence than the SC regions in most of higher plants and possibly due to copy correction between IR sequences by gene

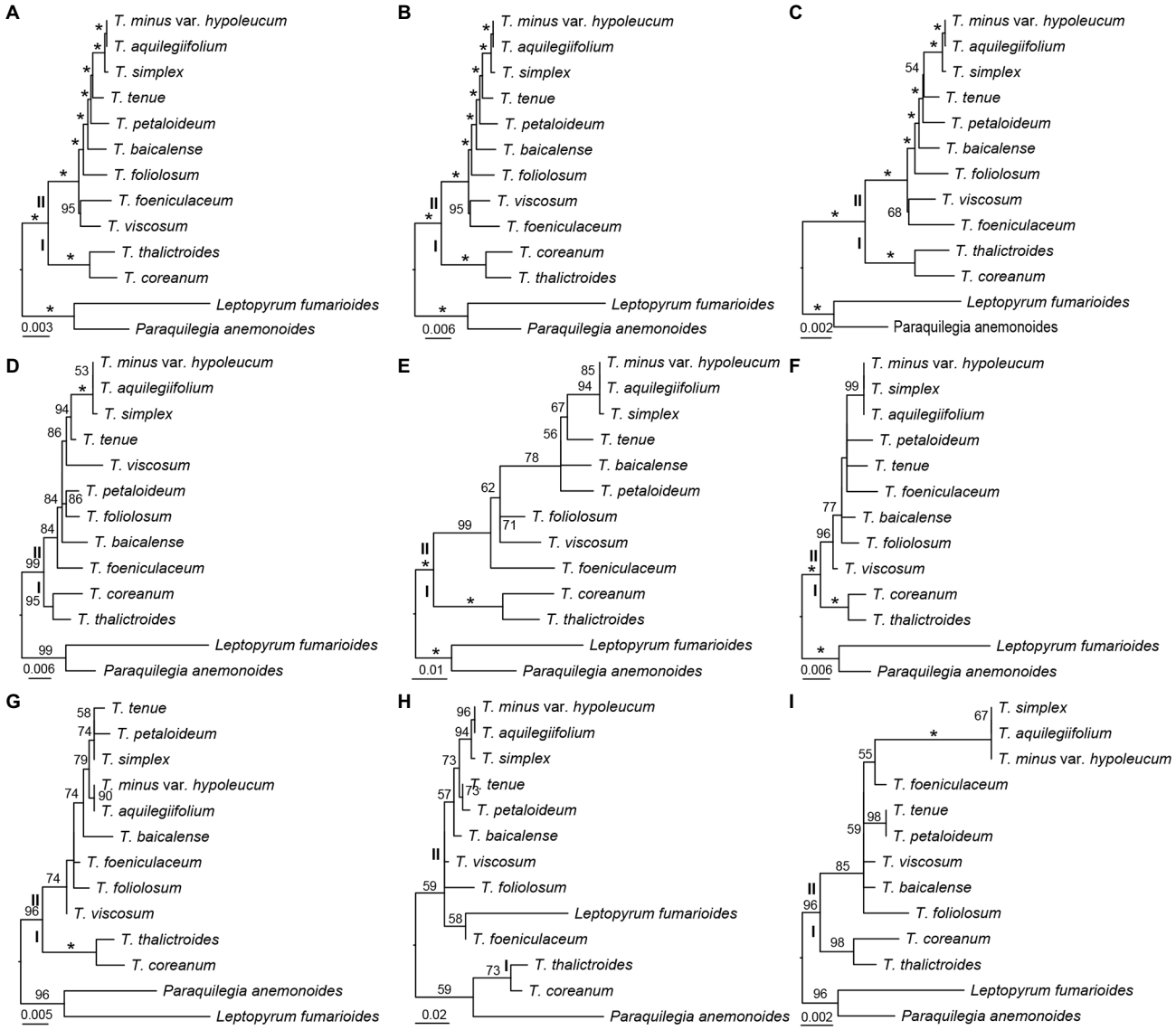


FIGURE 4 | Phylogenetic relationships of *Thalictrum* inferred from maximum likelihood (ML) analysis. **(A)** All sequence, **(B)** concatenation of 115 IGS regions, **(C)** concatenation of 114 gene sequences, **(D)** *rpl16* (with intron, Soza et al., 2012), **(E)** *ndhC-trnV*^{UAC} (Soza et al., 2013), **(F)** *ndhA* intron (Wang et al., 2019), **(G)** *trnL*^{UAA}-*trnF*^{GAA} (Wang et al., 2019), **(H)** *rpl32-trnL*^{UAG} (Wang et al., 2019), and **(I)** *rbcL* (Wang et al., 2019). The numbers above the branches indicate bootstrap support (%), and the asterisk indicates 100% bootstrap support in ML tree.

conversion (Khakhlova and Bock, 2006; Zhang et al., 2016). In the present study, the whole genome and IGS regions manifested higher sequence divergence than genes did, and genes with introns showed higher sequence divergence than genes without introns in *Thalictrum* species (**Figure 3**). In general, the non-coding regions (introns and spacers) had higher variability proportions than coding regions, which was also true for most higher plants (Shaw et al., 2014; Zhang et al., 2016).

In some studies, eight noncoding regions (*ndhF-rpl32*, *rpl32-trnL*^{UAG}, *ndhC-trnV-UAC*, *rps16-trnQ*^{UUG}, *psbE-petL*, *trnT*^{GGU}-*psbD*, *petA-psbJ*, and *rpl16* intron) have been identified as the best possible choices for low-level phylogenetic studies on angiosperms (Shaw et al., 2014). Among these regions, *ndhF-rpl32*, *rpl32-trnL*^{UAG}, and the *rpl16* intron were also identified as highly divergent loci among *Thalictrum* species in the present study. Nonetheless, two IGS regions related to *rpl32* are not suitable as molecular markers in *Thalictrum* because the *rpl32* gene is often transferred to the nucleus (Park et al., 2015). Aside from these loci, we also observed high nucleotide diversity in *infA-rps8*, *ccsA-ndhD*, *trnS*^{UGA}-*psbZ*, *trnH*^{GUG}-*psbA*, *rpl16-rps3*, *ndhG-ndhI*, and *ndhD-psaC* regions. Additionally, an intron of *rps16* also showed highly variable here, similarly to

TABLE 2 | Summary of models H0 and HA analyzed in the study. d_N, d_S, and ω are presented as medians.

Gene	H0: m0				HA: m1				2*(HA-H0)	*p*-value
	d_N	d_S	$\omega 0$	lnL	d_N	d_S	ω	lnL		
accD	0.0005	0.0019	0.2489	−2437.30	0.000832	0.00428	0.193579	−2433.08	8.44	0.75
atpA	0.0001	0.0021	0.0705	−2272.75	0	0.002509	0.0001	−2266.30	12.90	0.38
atpB	0.0002	0.0022	0.0699	−2221.92	0	0.002701	0.0001	−2213.35	17.14	0.14
atpE	0	0	0.1804	−642.17	0.000001	0.000005	0.0372454	−635.33	13.69	0.32
atpF	0	0	1.1265	−866.78	0.000002	0.000002	124.242	−862.26	9.02	0.70
atpH	0	0	0.0001	−366.55	0.000002	0.000005	0.0001	−366.55	0.000258	1.00
atpI	0.0004	0.0043	0.0857	−1137.90	0.000001	0.004595	0.0001	−1130.12	15.55	0.21
ccsA	0.0022	0.0071	0.3153	−1836.83	0.002604	0.010913	0.169018	−1831.85	9.97	0.62
cemA	0.001	0.0034	0.2884	−1096.08	0.000001	0.000006	0.0001	−1086.23	19.71	0.07
clpP	0	0	0.0694	−935.05	0.000001	0.000004	0.0001	−929.55	11.01	0.53
infA	0.0096	0.0123	0.7783	−269.79	0.013353	0.000013	74.8743	−264.81	9.97	0.62
matK	0.0027	0.0058	0.4631	−2572.82	0.002523	0.006385	0.911524	−2562.74	20.15	0.06
ndhA	0.0008	0.0053	0.1467	−1710.61	0.000002	0.003808	0.329949	−1704.63	11.97	0.45
ndhB	0	0	0.1178	−2133.67	0	0.000006	0.0001	−2129.17	8.99	0.70
ndhC	0	0	0.0916	−566.08	0	0.000005	0.0001	−562.34	7.48	0.82
ndhD	0.0011	0.0108	0.0998	−2601.20	0.00089	0.011321	0.0627638	−2593.10	16.20	0.18
ndhE	0	0	0.0235	−456.81	0	0.000006	0.0001	−454.86	3.92	0.98
ndhF	0.0009	0.0071	0.1292	−2463.23	0.001765	0.007192	0.147131	−2455.95	14.55	0.27
ndhG	0.001	0.005	0.1954	−909.26	0.000003	0.000074	0.0001	−896.00	26.52	0.01
ndhH	0.0005	0.006	0.0866	−1923.38	0.001108	0.003891	0.0698539	−1919.64	7.48	0.82
ndhI	0.0007	0.0061	0.113	−908.57	0.000001	0.008613	0.0001	−902.89	11.35	0.50
ndhJ	0	0	0.3505	−688.53	0	0.000005	0.0001	−686.35	4.38	0.98
ndhK	0	0	0.3155	−1048.40	0	0.000006	0.0001	−1041.06	14.69	0.26
petA	0.0004	0.003	0.1455	−1568.16	0.000002	0.004392	0.0001	−1555.79	24.74	0.02
petD	0	0	0.0307	−741.85	0.000001	0.000005	0.0102435	−739.71	4.29	0.98
petG	0	0	0.8509	−156.93	0.000002	0	999	−155.55	2.77	1.00
petL	0	0	0.1008	−136.46	0.000002	0	313.224	−135.59	1.74	1.00
psaA	0	0.0018	0.0084	−3309.32	0	0.001795	0.0001	−3306.01	6.62	0.88
psaB	0.0002	0.0031	0.0672	−3293.87	0	0.003743	0.0001	−3286.29	15.16	0.23
psaI	0	0	0.8811	−177.10	0	0.000005	0.106004	−174.86	4.48	0.97
psbA	0.0001	0.0037	0.0157	−1554.95	0	0.003857	0.0001	−1552.80	4.31	0.98
psbB	0.0002	0.0021	0.0878	−2235.59	0	0.002695	0.0001	−2229.89	11.39	0.50
psbC	0.0001	0.0024	0.0259	−2138.08	0	0.002509	0.0001	−2131.91	12.33	0.42
psbD	0	0	0.075	−1533.73	0	0.000005	0.0001	−1529.39	8.68	0.73
psbH	0	0	0.12	−321.61	0.000002	0.000006	0.0978274	−319.64	3.93	0.98
psbI	0	0	0.0001	−152.19	0	0.000003	0.0001	−152.19	0.005686	1
psbJ	0	0	0.1454	−163.03	0.000002	0	113.399	−162.51	1.05	1.00
psbK	0	0	0.3253	−268.48	0	0.000005	0.0001	−265.68	5.59	0.94
psbM	0	0	0.0001	−126.66	0.000002	0	212.117	−126.66	0.000884	1.00
psbT	0	0	0.245	−142.20	0	0.000004	0.0001	−141.67	1.05	1.00
rbcL	0.0004	0.0018	0.1906	−2281.62	0	0.002963	0.0001	−2274.83	13.59	0.33
rpl14	0	0	0.037	−529.92	0.000001	0.000006	0.0001	−526.77	6.30	0.90
rpl16	0	0	0.1546	−631.29	0.000001	0.000005	0.0001	−624.98	12.62	0.40
rpl2	0	0	0.183	−1139.62	0.000002	0	31.3706	−1138.57	2.10	1.00
rpl20	0	0	0.4022	−628.11	0.000002	0.000004	9.55435	−623.99	8.24	0.77
rpl22	0	0	0.4957	−931.12	0	0.000007	0.0001	−919.97	22.31	0.03
rpl23	0	0	999	−396.07	0.000002	0.000001	55.6975	−396.07	0.00	1.00
rpl33	0	0	0.1364	−307.30	0.000002	0	491.843	−303.04	8.51	0.74
rpoA	0.001	0.006	0.1692	−1751.50	0.000001	0.010121	0.0001	−1742.58	17.85	0.12
rpoB	0.0003	0.0029	0.1019	−4816.32	0.000418	0.001825	0.0461232	−4807.91	16.82	0.16
rpoC1	0.0007	0.004	0.1796	−3224.37	0.000638	0.002151	0.0743791	−3216.52	15.71	0.20
rpoC2	0.0012	0.0042	0.3139	−6585.10	0.001256	0.004513	0.215304	−6576.06	18.07	0.11
rps11	0	0	0.0797	−634.65	0	0.000005	0.0001	−631.31	6.68	0.88
rps14	0	0	0.1102	−420.73	0	0.000005	0.0001	−418.47	4.51	0.97
rps15	0	0	0.6425	−434.79	0.000002	0	76.3582	−430.76	8.08	0.78
rps16	0	0	0.4244	−380.25	0.000002	0	15.7794	−377.71	5.07	0.96
rps18	0	0	0.0918	−411.02	0.000002	0	192.024	−408.76	4.52	0.97
rps2	0	0	0.4331	−1203.16	0.000002	0.000006	0.134231	−1198.26	9.80	0.63
rps3	0.0005	0.005	0.1081	−1071.79	0.000001	0.005949	0.0001	−1064.71	14.16	0.29
rps4	0.0007	0.0051	0.1334	−944.70	0.000001	0.000006	0.0001	−937.89	13.61	0.33

(Continued)

TABLE 2 | Continued

Gene	H0: m0				HA: m1				2*(HA-H0)	*p*-value
	d_N	d_S	$\omega 0$	lnL	d_N	d_S	ω	lnL		
rps8	0	0	0.2502	−588.72	0	0.000006	0.0001	−586.69	4.08	0.98
ycf1	0.0035	0.0068	0.5214	−11057.96	0.003664	0.004828	0.495128	−11050.66	14.59	0.26
ycf2	0	0	0.3282	−246.36	0	0.000005	0.0001	−244.96	2.79	1.00
ycf3	0	0	0.1983	−696.68	0.000001	0.000004	0.20605	−693.90	5.54	0.94
ycf4	0	0	0.036	−802.98	0	0.000005	0.0001	−799.72	6.51	0.89
ycf15	0.0001	0.0002	0.552	−9655.78	0.000185	0.000006	0.402698	−9648.33	14.90306	0.246781
petB	0	0.0058	0.0001	−991.11	0.000001	0.005832	0.0001	−991.11	0.000242	1
petN	0	0	0.2706	−116.86	0.000002	0	198.276	−116.86	0.001416	1
psaC	0	0	0.0001	−375.50	0	0.000005	0.0001	−375.50	0.00008	1
psaJ	0	0	0.0001	−208.38	0	0.000004	0.0001	−208.38	0.00036	1
psbE	0	0	0.0001	−350.96	0	0.000005	0.0001	−350.96	0.000432	1
psbF	0	0	0.0001	−159.00	0	0.000005	0.0001	−159.00	0.000294	1
psbL	0	0	0.0001	−162.62	0	0.000007	0.0001	−162.62	0	1
psbN	0	0	0.0001	−180.08	0	0.000004	0.0001	−180.08	0.000258	1
psbZ	0	0	0.0001	−247.62	0	0.000005	0.0001	−247.62	0.000616	1
rpl36	0	0	0.0001	−152.86	0.000001	0.000002	0.520692	−152.86	0.00048	1
rps12	0	0	0.0001	−496.22	0.000002	0	60.2915	−496.22	0.000092	1
rps19	0	0	0.0001	−404.64	0.000001	0.000005	0.0001	−404.64	0.000202	1
rps7	0	0	0.0001	−612.36	0.000002	0	78.7221	−612.36	0.002164	1

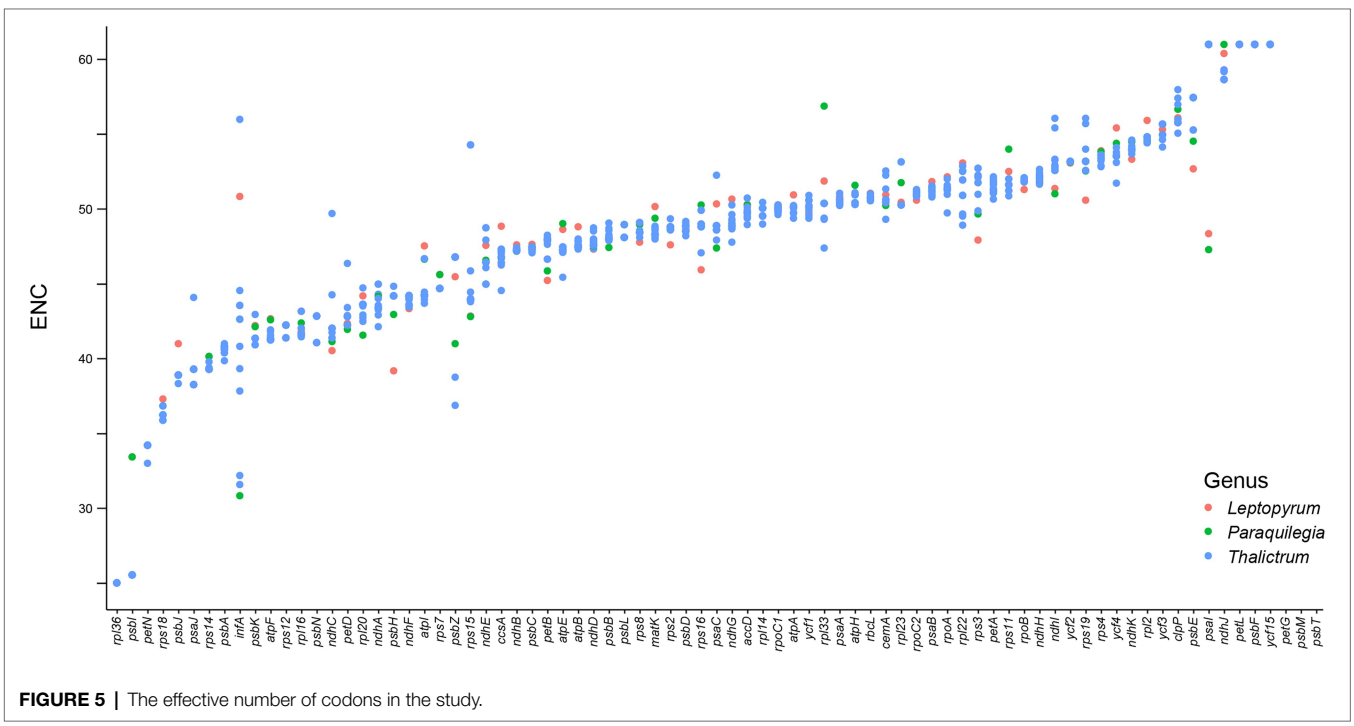


FIGURE 5 | The effective number of codons in the study.

Podophylloideae (Berberidaceae; Ye et al., 2018). These divergence hotspot regions of the 11 *Thalictrum* plastid genome sequences provided abundant information for developing effective molecular markers to the phylogenetic analyses and plant identification of *Thalictrum* species. Besides, the resolution and efficiency of chloroplast markers can be strongly affected by the length of target fragment. The *rpoC1-rpoB* region has a relatively high nucleotide diversity among different plastomes, but cannot be a good molecular marker as its target length is only 5 bp.

Phylogenetic Relationships

The plastid genome sequences have been utilized successfully for the phylogenetic studies on angiosperms (Jansen et al., 2007; Huang et al., 2014; Kim et al., 2015; Li et al., 2019).

Our phylogenetic trees based on whole complete plastid genome sequences, 116 IGS regions, and 114 gene sequences revealed that *Thalictrum* contains two major clades that is consistent with previous studies (**Figures 4A–C**; Soza et al., 2012, 2013; Morales-Briones et al., 2019; Wang et al., 2019). However, the relationships along the backbone of the clades are not well-supported in their studies. None of the sections traditionally circumscribed for this genus (Tamura, 1995) is monophyletic. It is necessary to apply more samplings and find more efficient molecular markers for *Thalictrum*.

Our phylogenetic trees indicated that 116 IGS regions had stronger support than 114 gene sequences (**Figures 4B,C**). Additionally, the *rpl16* intron—that was used by Soza et al. (2012) with high sequence divergence in the studies—showed also strong support in clades II here (**Figure 4D**). While the coding regions of *rbcL* employed by Wang et al. (2019) showed lower supports within clades II in our analysis (**Figure 4I**). The non-coding regions (introns and spacers) are more variable molecular markers. For the ML tree of *rpl32-trnL*^{UAG} used by Wang et al. (2019), the outgroups are embedded in *Thalictrum* probably because the matrix of *rpl32-trnL*^{UAG} contained lots of indels (**Figure 4H**). The *rpl32* gene is often transfers to the nucleus (Park et al., 2015) that make the *ndhF-rpl32*, *rpl32-trnL*^{UAG}, and *rpl32* regions not reliable to be markers for phylogeny in *Thalictrum*.

Positive Selection in Different Genes

It is believed that selection is the most probable components of the evolutionary forces acting on most highly expressed genes, although all genes are basically subjected to a certain degree of natural selection (Gouy and Gautier, 1982; Sueoka, 1999; Sharp et al., 2010). And the degeneracy of genetic code leads to the expression of variation contained in a gene through its manifestation in protein, which varied among different species (Edelman and Gally, 2001; Wan et al., 2004; Chakraborty et al., 2020). In the present study, we observed different codon usage frequency on different genes under positive pressure. For example, 12 plastid genes (*atpF*, *rpl33*, *rps15*, *rpl20*, *rps16*, *rps18*, *petG*, *rpl2*, *petL*, *psbJ*, *psbM*, and *rpl23*) were observed under positive selective pressure in most of the 11 *Thalictrum* species among which 11 showed relatively higher CBI values (>0.5) suggesting high expression level *in vivo*; while three plastid genes that are relative with NADH oxidoreductase (*ndhG*), cytochrome b6/f complex (*petA*), and ribosomal proteins (*rpl22*) were observed under significantly strong positive selective pressure ($p < 0.05$ based on likelihood ratio tests) in only 1–4 *Thalictrum* species, showing relatively lower CBI (<0.5). The former and latter genes performed different codon usage bias suggesting different expression levels due to different usage frequency of the rare and optimal codons, which could further affected the functional patterns of those genes during their evolution process. Additionally, it also indicated potential functional divergence among plastid genomes of different *Thalictrum* species, according to abundant differences observed between selective pressures and usage codon frequencies for different plastid genes in these species.

CONCLUSION

This is the first report to describe a comprehensive landscape of plastomic variations among *Thalictrum* species on the basis of 11 complete plastomes. Comparison between these plastomes uncovered not only high similarities in overall structure, gene order, and content but also some structural variations caused by the expansion or contraction of the IR regions into or out of adjacent single-copy regions. DNA sequence divergence across 11 *Thalictrum* plastomes revealed that *infA-rps8*, *ccsA-ndhD*, *trnS*^{UGA}*-psbZ*, *trnH*^{GUG}*-psbA*, *rpl16-rps3*, *ndhG-ndhI*, *ndhD-psaC*, and *ndhJ-ndhK* are among the fastest-evolving loci and are promising molecular markers. Therefore, these highly variable loci should be valuable for future phylogenetic and phylogeographic studies on *Thalictrum*. Our phylogenomic analyses based on whole complete plastid genome sequences, 116 IGS regions and 114 gene sequences were all supported the monophyly of *Thalictrum* and two major clades within this genus. Furthermore, among 79 plastome-derived protein-coding genes (CDSs), 15 genes were identified as fast evolving genes, which were all proved to be under positive selection but showed different bias in their codon usage frequencies. Overall, our results demonstrate the ability of plastid phylogenomics to improve phylogenetic resolution, and will expand the understanding of plastid gene evolution in *Thalictrum*.

DATA AVAILABILITY STATEMENT

All raw sequencing reads generated in the study have been deposited in NCBI under the BioProject accession PRJNA817687. The complete sequences and annotations of plastomes have also been deposited at GenBank under the accessions OM501079 and OM501080. The updated annotations of plastomes for the other 11 species in this study have been deposited to the Figshare online database (https://doi.org/10.6084/m9.figshare.19108097.v1).

AUTHOR CONTRIBUTIONS

W-CH and Z-QW conceived the research. H-WP carried out taxon sampling and generated all the data. K-LX and W-CH performed the data analyses. K-LX, W-CH, and WM wrote the manuscript with help from Z-QW. Y-XY revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the China Postdoctoral Science Foundation (grant number 2021M703540), the Training of

Excellent Science and Technology Innovation talents in Shenzhen-Basic Research on Outstanding Youth (grant number RCYX20200714114538196), the National Natural Science Foundation of China (grant number 32011530072), the Initial fund of Shenzhen Agricultural Genome Research Institute, Chinese Academy of Agricultural Sciences (grant number SJXW19073), and the Russian Science Foundation [grant number 19-74-10082 (preparation of material)], within state assignments for CSBG SB RAS [grant number AAAA-A21-121011290024-5 (study of herbarium collections)].

ACKNOWLEDGMENTS

The authors are grateful for Jian-Fei Ye for kind advices on identification of the species used in the present study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.897843/full#supplementary-material

REFERENCES

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., and Donoghue, M. J. (1995). The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. *Ann. Missouri Bot. Gard.* 82, 247–277. doi: 10.2307/2399880

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Chakraborty, S., Yengkhom, S., and Uddin, A. (2020). Analysis of codon usage bias of chloroplast genes in *Oryza* species. *Planta* 252:67. doi: 10.1007/s00425-020-03470-7

Cingolani, P. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Flying* 6, 80–92. doi: 10.4161/fly.19695

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, 1–4. doi: 10.1093/gigascience/giab008

Edelman, G. M., and Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13763–13768. doi: 10.1073/pnas.231499798

Frank, W. (1990). The 'effective number of codons' used in a gene. *Gene* 87, 23–29. doi: 10.1016/0378-1119(90)90491-9

Gao, L., Yi, X., Yang, Y. X., Su, Y. J., and Wang, T. (2009). Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: insights into evolutionary changes in fern chloroplast genomes. *BMC Evol. Biol.* 9:130. doi: 10.1186/1471-2148-9-130

Gouy, M., and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10, 7055–7074. doi: 10.1093/nar/10.22.7055

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

He, W., Chen, C., Xiang, K., Wang, J., Zheng, P., Tembrock, L. R., et al. (2021a). The history and diversity of rice domestication as resolved from 1464 complete plastid genomes. *Front. Plant Sci.* 12:781793. doi: 10.3389/fpls.2021.781793

He, Y., Wang, R., Gai, X., Lin, P., and Wang, J. (2021b). The complete chloroplast genome of *Thalictrum baicalense* turcz. ex ledeb. *Mitochondrial DNA B Resour.* 6, 437–438. doi: 10.1080/23802359.2020.1870896

He, J., Yao, M., Lyu, R. D., Lin, L. L., Liu, H. J., Pei, L. Y., et al. (2019). Structural variation of the complete chloroplast genome and plastid phylogenomics of the genus *Asteropyrum* (Ranunculaceae). *Sci. Rep.* 9:15285. doi: 10.1038/s41598-019-51601-2

Hoch, B., Maier, R. M., Appel, K., Igloi, G. L., and Kössel, H. (1991). Editing of a chloroplast mRNA by creation of an initiation codon. *Nature* 353, 178–180. doi: 10.1038/353178a0

Huang, H, Shi, C., Liu, Y., Mao, S. Y., and Gao, L. Z. (2014). Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol. Biol.* 14:151. doi: 10.1186/1471-2148-14-151

Hughes, C. E., Eastwood, R. J., and Bailey, C. D. (2006). From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 361, 211–225. doi: 10.1098/rstb.2005.1735

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Leebens-Mack, J., Müller, K. F., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19369–19374. doi: 10.1073/pnas.0709121104

Katoh, K., and Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 26, 1899–1900. doi: 10.1093/bioinformatics/btq224

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Khakhlova, O., and Bock, R. (2006). Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant J.* 46, 85–94. doi: 10.1111/j.1365-313X.2006.02673.x

Kim, K. J., and Lee, H. L. (2004). Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res.* 11, 247–261. doi: 10.1093/dnares/11.4.247

Kim, K., Lee, S.-C., Lee, J., Yu, Y., Yang, K., Choi, B.-S., et al. (2015). Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Sci. Rep.* 5:15655. doi: 10.1038/srep15655

Kuroda, H., Suzuki, H., Kusumegi, T., Hirose, T., Yukawa, Y., and Sugiura, M. (2007). Translation of *psbC* mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended Shine–Dalgarno sequence in tobacco chloroplasts. *Plant Cell Physiol.* 48, 1374–1378. doi: 10.1093/pcp/pcm097

Langlet, O. F. I. (1927). Beitrage zur zytologie der Ranunculazeen. *Sven. Bot. Tidskr.* 21, 1–17.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H. T., Yi, T. S., Gao, M. L., Ma, P. F., Zhang, T., Yang, J. B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants.* 5, 461–470. doi: 10.1038/s41477-019-0421-0

Löve, Á. (1982). IOPB chromosome number reports LXXIV. *Taxon* 31, 119–128. doi: 10.1002/j.1996-8175.1982.tb02346.x

Morales-Briones, D. F., Arias, T., Di Stilio, V. S., and Tank, D. C. (2019). Chloroplast primers for clade-wide phylogenetic studies of *Thalictrum*. *Appl. Plant Sci.* 7:e11294. doi: 10.1002/aps3.11294

Mort, M. E., Archibald, J. K., Randle, C. P., Levsen, N. D., O'Leary, T. R., Topalov, K., et al. (2007). Inferring phylogeny at low taxonomic levels: utility of rapidly evolving cpDNA and nuclear ITS loci. *Am. J. Bot.* 94, 173–183. doi: 10.3732/ajb.94.2.173

Park, S., Jansen, R. K., and Park, S. (2015). Complete plastome sequence of *Thalictrum coreanum* (Ranunculaceae) and transfer of the *rpl32* gene to the nucleus in the ancestor of the subfamily Thalictroideae. *BMC Plant Biol.* 15:40. doi: 10.1186/s12870-015-0432-6

Raubeson, L. A., Peery, R., Chumley, T., Dziubek, C., Fourcade, H. M., Boore, J. L., et al. (2007). Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* 8, 174–200. doi: 10.1186/1471-2164-8-174

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Librado, P. G., Ramos-Onsins, S. E., and Sánchez-Gracia, A. (2017). DnaSP v6: DNA

sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Sharp, P. M., Emery, L. R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 365, 1203–1212. doi: 10.1098/rstb.2009.0305

Sharp, P. M., Tuohy, T. M., and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14, 5125–5143. doi: 10.1093/nar/14.13.5125

Shaw, J., Shafer, H. L., Leonard, O. R., Kovach, M. J., Schorr, M., and Morris, A. B. (2014). Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. *Am. J. Bot.* 101, 1987–2004. doi: 10.3732/ajb.1400398

Shi, C., Liu, Y., Huang, H., Xia, E. H., Zhang, H. B., and Gao, L. Z. (2013). Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: an exemplary study of *ycf15* function and evolution in angiosperms. *PLoS One* 8:e59620. doi: 10.1371/journal.pone.0059620

Soza, V. L., Brunet, J., Liston, A., Smith, P. S., and Di Stilio, V. S. (2012). Phylogenetic insights into the correlates of dioecy in meadow-rues (*Thalictrum*, Ranunculaceae). *Mol. Phylogenet. Evol.* 63, 180–192. doi: 10.1016/j.ympev.2012.01.009

Soza, V. L., Haworth, K. L., and Di Stilio, V. S. (2013). Timing and consequences of recurrent polyploidy in meadow-rues (*Thalictrum*, Ranunculaceae). *Mol. Biol. Evol.* 30, 1940–1954. doi: 10.1093/molbev/mst101

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Sueoka, N. (1999). Two aspects of DNA base composition: G + C content and translation-coupled deviation from intra-strand rule of A = T and G = C. *J. Mol. Evol.* 49, 49–62. doi: 10.1007/PL00006534

Szczecińska, M., and Sawicki, J. (2015). Genomic resources of three *Pulsatilla* species reveal evolutionary hotspots, species-specific sites and variable plastid structure in the family Ranunculaceae. *Int. J. Mol. Sci.* 16, 22258–22279. doi: 10.3390/ijms160922258

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B., and Brennicke, A. (2013). RNA editing in plants and its evolution. *Annu. Rev. Genet.* 47, 335–352. doi: 10.1146/annurev-genet-111212-133519

Tamura, M. (1995). "Ranunculaceae," in *Die Naturlichen Pflanzenfamilien. Vol. 17a*. ed. P. Hiepko (Germany: Duncker & Humblot, Berlin), 223–497.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq—versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Wan, X. F., Xu, D., Kleinhofs, A., and Zhou, J. (2004). Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.* 4:19. doi: 10.1186/1471-2148-4-19

Wang, T. N., Clifford, M. R., Martínez-Gómez, J., Johnson, J. C., Riffell, J. A., and Di Stilio, V. S. (2019). Scent matters: differential contribution of scent to insect response in flowers with insect vs. wind pollination traits. *Ann. Bot.* 123, 289–301. doi: 10.1093/aob/mcy131

Wang, W. C., Fu, D. Z., Li, L. Q., Bartholomew, B., Brach, A. R., Dutton, B. E., et al. (2001). "Ranunculaceae," in *Flora of China. Vol. 6*. eds. Z. Y. Wu, P. H. Raven and D. Y. Hong (Beijing: Science Press), 133–438.

Wang, W. C., and Xiao, P. G. (1979). "Ranunculaceae," in *Flora Reipublicae Popularis Sinicae. Vol. 27*. (Beijing: Science Press), 502–592.

Whelan, S., and Goldman, N. (1999). Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 16, 1292–1299. doi: 10.1093/oxfordjournals.molbev.a026219

Wu, Z. Y., Zhou, T. Y., and Xiao, P. G. (1998). *Xin Hua Compendium of Materia Medica*. Shanghai: Science and Technology Press. *1*, 133–139.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, J. B., Yang, S. X., Li, H. T., Yang, J., and Li, D. Z. (2013). Comparative chloroplast genomes of *Camellia* species. *PLoS One* 8:e73053. doi: 10.1371/journal.pone.0073053

Yang, Y., Zhou, T., Duan, D., Yang, J., Feng, L., and Zhao, G. (2016). Comparative analysis of the complete chloroplast genomes of five *Quercus* species. *Front. Plant Sci.* 7:959. doi: 10.3389/fpls.2016.00959

Ye, W., Yap, Z., Li, P., Comes, H. P., and Qiu, Y. X. (2018). Plastome organization, genome-based phylogeny and evolution of plastid genes in Podophylloideae (Berberidaceae). *Mol. Phylogenet. Evol.* 127, 978–987. doi: 10.1016/j.ympev.2018.07.001

Zhang, Y., Du, L., Liu, A., Chen, J., Wu, L., Hu, W., et al. (2016). The complete chloroplast genome sequences of five *Epimedium* species: lights into phylogenetic and taxonomic analyses. *Front. Plant Sci.* 7:306. doi: 10.3389/fpls.2016.00306

Zhu, M., and Xiao, P. G. (1989). Study on resource utilization of germander (*Thalictrum*). *Chin. Trad. Herb Drugs* 20, 29–31.

Zhu, M., and Xiao, P. G. (1991). Chemosystematic studies on *Thalictrum* L. in China. *Acta Phytotaxon. Sin.* 29, 358–369.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 12 July 2022
doi: 10.3389/fpls.2022.931721

Genome-Wide Identification, Characterization, and Expression Profile Analysis of *CONSTANS-like* Genes in Woodland Strawberry (*Fragaria vesca*)

Xinyong Zhao[1,2†], Fuhai Yu[1,3†], Qing Guo[1,2], Yu Wang[1,2], Zhihong Zhang[1,2] and Yuexue Liu[1,2]*

[1] College of Horticulture, Shenyang Agricultural University, Shenyang, China, [2] Liaoning Key Laboratory of Strawberry Breeding and Cultivation, Shenyang Agricultural University, Shenyang, China, [3] TieLing Academy of Agricultural Science, Tieling, China

OPEN ACCESS

Edited by:
Weihua Pan, Agricultural Genomics Institute at Shenzhen (CAAS), China

Reviewed by:
Jingbo Chen, Jiangsu Province and Chinese Academy of Sciences, China
Jianchao Ma, Henan University, China

***Correspondence:**
Yuexue Liu
yuexueliu@aliyun.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *29 April 2022*
Accepted: *15 June 2022*
Published: *12 July 2022*

Citation:
Zhao X, Yu F, Guo Q, Wang Y, Zhang Z and Liu Y (2022) Genome-Wide Identification, Characterization, and Expression Profile Analysis of CONSTANS-like Genes in Woodland Strawberry (Fragaria vesca). Front. Plant Sci. 13:931721. doi: 10.3389/fpls.2022.931721

CONSTANS-like (*CO-like*) gene is one of the most important regulators in the flowering process of the plant, playing a core role in the photoperiodic flowering induction pathway. In this study, we identified 10 distinct *CO-like* genes (*FveCOs*) in woodland strawberry (*Fragaria vesca*). They were classified into three groups with specific gene structure characteristics or protein domains in each group. The effect of selection pressure on the *FveCOs* in the woodland strawberry was tested by Ka/Ks, and it was shown that the evolution rate of *FveCOs* was controlled by purification selection factors. Intraspecific synteny analysis of woodland strawberry *FveCOs* showed that at least one duplication event existed in the gene family members. Collinearity analysis of woodland strawberry genome with genomes of *Arabidopsis*, rice (*Oryza sativa*), and apple (*Malus* × *domestica*) showed that *CO-like* genes of *F. vesca* and *Malus* × *domestica* owned higher similarity for their similar genomes compared with those of other two species. The *FveCOs* showed different tissue-specific expression patterns. Moreover, real-time quantitative PCR results revealed that the expressions of the most *FveCOs* followed a 24-h rhythm oscillation under both long-day (LD) and short-day (SD) conditions. Further expression analysis showed that the individual expression changing profile of *FveCO3* and *FveCO5* was opposite to each other under both LD and SD conditions. Moreover, the expression of *FveCO3* and *FveCO5* was both negatively correlated with the flowering time variation of the woodland strawberry grown under LD and SD conditions, indicating their potential vital roles in the photoperiodic flowering regulation. Further protein interaction network analysis also showed that most of the candidate interaction proteins of FveCO3 and FveCO5 were predicted to be the flowering regulators. Finally, LUC assay indicated that both FveCO3 and FveCO5 could bind to the promoter of *FveFT1*, the key regulator of flowering regulation in the woodland strawberry, and thus activate its expression. Taken together, this study laid a foundation for understanding the exact roles of FveCOs in the reproductive development regulation of the woodland strawberry, especially in the photoperiodic flowering process.

Keywords: ***CONSTANS-like*, woodland strawberry, flowering, expression profiles, photoperiod**

INTRODUCTION

Flowering is a critical growth transition period during growth and development of the plant. To ensure the continuation of species, plants regulate their flowering process accompanying with the external environmental factors such as the temperature and day length. Moreover, plants also adjust their endogenous hormones in response to the external environmental factors to ensure their vegetative and reproductive growth (Song et al., 2013). Flowering regulation is a sophisticated biological process involving various signaling pathways, which have been extensively revealed in the past decades (Perrella et al., 2020). Traditional signaling pathways of flowering regulation are normally known as the photoperiod, vernalization, autonomy, and gibberellin pathway (Mouradov et al., 2002; Hayama and Coupland, 2003; Michaels et al., 2005; Wenkel et al., 2006; Domagalska et al., 2010). The emerging signaling pathways include age pathway, thermosensory pathway, sugar pathway, stress pathway, and hormonal signals to control floral transition (Izawa, 2021). Photoperiodic flowering dominates among these pathways, especially in *Arabidopsis* and rice (Putterill et al., 1995; Yano et al., 2000). According to the day-length requirements in flowering, photoperiodic condition can be described as long-day (LD) condition, short-day (SD) condition, and neutral condition.

CONSTANS (CO) protein is a core transcriptional regulator in the photoperiodic pathway, which is firstly reported in *Arabidopsis* mutant studies. AtCO in *Arabidopsis* accelerated flowering only under long-day (LD) condition via activating the transcriptions of *AtFT* and *AtSOC1* (Valverde, 2011). However, the *CO* homologs can also promote flowering under SD condition in other species. For example, in rice, the *CO* homolog *HEADING DATE 1* (*Hd1*) promoted flowering under SD condition, but suppressed flowering under LD condition, which is completed by regulating the expression of the rice *FT* ortholog *HEADING DATE 3a* (*Hd3a*) (Yano et al., 2000).

The CO homologs were characterized as the important zinc-finger transcription factors (TFs) belonging to a subset of BBX protein family with the specific B-box and CCT domains. The B-box domain located in the N-terminal was found to participate in the protein–protein interaction, while the CCT domain located in the C-terminal owned the nuclear localization function (Putterill et al., 1995; Yano et al., 2000; Robson et al., 2001; Griffiths et al., 2003; Valverde, 2011). The CONSTANS/CONSTANS-like proteins (CO/COLs) can be divided into three major groups according to the divergence of conserved domains (Crocco and Botto, 2013). The CCT domain in the C-terminal can be found in all group members. According to the basis of the consistency of amino acid sequences and the specificity of zinc-binding amino acid residues, the B-box domain can be divided into two types, named as B-box1 (B1) and B-box2 (B2) individually. Group I members owned both B-box1 domain and B-box2 domain, while group II members possessed only B-box1 domain. Group III members owned B-box1 domain and one diverged zinc-finger structure.

Diverse numbers of *CO-like* gene were detected in different plant species, such as 17 members of *CO* family initially identified in *Arabidopsis* (Robson et al., 2001; Crocco and Botto, 2013; Romero-Campero et al., 2013; Wang et al., 2013) and 16 members identified in rice (Griffiths et al., 2003). For horticulture plants, *CO-like* genes have also been widely detected, such as 11 members in *Medicago truncatula* (Wong et al., 2014), 16 members in potato (*Solanum tuberosum*) (Talar et al., 2017), 20 members in grape (*Vitis vinifera*) (Wang et al., 2019), 23 members in tomato (*Solanum lycopersicum*) (Yang et al., 2020), 25 members in banana (*Musa acuminata*) (Chaurasia et al., 2016), and 25 members in Chinese cabbage (*Brassica campestris*) (Song et al., 2015).

The *CO* gene was initially identified as the transcriptional activator of *FLOWERING LOCUS T* (*FT*) using its B-box domain to form a multimeric binding to the *FT* promoter (Samach et al., 2000; Tiwari et al., 2010). FT protein plays the "florigen" role and can directly target the second exon of *LFY*, the master regulator of flower fate, to enhance its expression (Zhu et al., 2020). In *Arabidopsis*, AtCO is considered to be inactive under SD conditions. Flowering time of the *co* mutant is the same under either LD or SD condition. However, cutting down the expression of *AtCO* leads to late flowering under LD condition, while overexpression promotes the flowering process under both LD and SD conditions (Putterill et al., 1995; Robson et al., 2001; Kotake et al., 2003).

In *Arabidopsis*, the expression of *AtFT* was increased by AtCO and AtCOL5 (Hassidim et al., 2009), which were positive regulators of *FT*. However, not all the CO members function as the flowering activator. *AtCOL3* and *AtCOL4* acted as flowering repressors under both LD and SD conditions, while *AtCOL8* and *AtCOL9* delayed flowering only under LD condition (Cheng and Wang, 2005). *Atcol3* mutant could flower earlier under both LD and SD conditions. Further researches demonstrate that AtCOL3 may regulate the expression of *AtFT* by interacting with AtBBX32 to control *Arabidopsis* flowering (Yang et al., 2019).

Functions of CO/COL members were not only restricted in the photoperiodic flowering regulation; remarkable different roles of *CO*/*COL* genes had been reported in various species. For example, in rice, OsBBX5 plays a role in downstream of phytochrome-B receptor, while OsK accelerates the leaf senescence. The ectopic expression of *AtCO* gene in potato can inhibit the potato tuber expansion, while silencing of potato *StCO* gene can promote the potato tuber expansion. In *Glycine max*, *GmCO9* affects root development and is closely related to seed maturation (Huang et al., 2011). *CO* family also mediates various aspects functions of the plant given as follows: *AtCOL3* regulates root growth (Datta et al., 2006), *VviCOL1* plays a major role in bud dormancy (Almada et al., 2009), *CrCO* regulates star synthesis and cell division (Deng et al., 2015), *GmCO9* is closely related to seed competition (Liu et al., 2011), *AtCOL7* regulates branching (Wang et al., 2013), *Ghd2* confers drought sensitivity (Liu et al., 2016), *StCO1* is involved in tuberization (González-Schain et al., 2012), and *MaCOL1* regulates fruit ripening (Chen et al., 2012).

Strawberry (*Fragaria* × *ananassa*), cultivated in different arable regions all over the world, is one of the most important

berries characterized by its unique flavor and nutritional value. Fruit quantity and quality directly determine the economic value of strawberry. Early flowering is one of the great advantages of strawberry cultivation with the reduced production time. Compared with other plants, interesting flowering habit is reported in strawberry. For example, while most cultivars of cultivated strawberry are June-bearing SD plants, there are also strawberry cultivars that flower perpetually with no requirement for SD or low-temperature condition, known as the everbearing types (EB). A similar flowering habit also exists in the wild diploid strawberry, whose genome is much simpler than that of the cultivated octoploid strawberry.

At present, studies on strawberry flowering mainly focus on the effects of ambient temperature, photoperiod, or hormone on flowering regulation (Koskela et al., 2012; Sønsteby et al., 2017). Molecular mechanisms research about flowering regulation in strawberry is mainly limited in the function illustration of its FT homolog and TFL1 homolog, which were found function as the florigen and antiflorigen individually (Zhu et al., 2020). Few reports mentioned the possible roles of *CO/COL* genes in woodland strawberry flowering. *FvCO*, a homologous gene of *AtCO* in *F. vesca,* is found to be indispensable for the generation of the bimodal rhythm expression profile of *FvFT1* and thus plays its role in the photoperiodic development of strawberry (Kurokura et al., 2017).

The roles of strawberry *CO/COL* genes in the flowering process have not yet been well elucidated. Hence, in this study, we performed the genome-wide identification of the *CO-like* gene family members in the woodland strawberry based on the high-quality *Fragaria vesca* v4.0.a1 genome database. We provide the detailed molecular information about the *Fragaria vesca CO-like* gene family, including the chromosomal location, sequence homology, introns distribution, motif composition, and evolutionary relationships. The expression of *FveCOs* in different tissues and organs was checked. Meanwhile, their diurnal expression changes in leaf treated with different photoperiodic conditions (LD or SD) were also analyzed. Our results would be valuable for understanding the roles of strawberry *CO-like* genes, especially in the photoperiodic flowering of strawberry.

MATERIALS AND METHODS

Identification of *CO/COL* Genes in Woodland Strawberry

A BLAST search (E-value < $1E^{-5}$) was performed against woodland strawberry *(F. vesca)* genome data v4.0.a1 in the Genome Database for Rosaceae (GDR[1]) using the full-length amino acid sequences of COs and COLs of *Arabidopsis*. The amino acid sequences of 17 CO/COL proteins in *Arabidopsis* were obtained from NCBI[2]. Then, the sequences of the retrieved woodland strawberry *(F. vesca*) CO/COL candidates (FveCOs) were submitted to the PFAM database[3] to annotate the unique and conserved protein domains. The amino acid sequences of the B-box and CCT domains in *Arabidopsis* CO/COL proteins peculiar to the members of this family were then used as the query sequences for further confirmation.

The detailed information of identified woodland strawberry *CO-like* genes, including chromosomal location, cDNA length, ORF, and the amino acid (AA), was then downloaded from the GDR. Physicochemical properties of the identified FveCOs proteins, including molecular weight (MV) and isoelectric point (pI), were counted in the ProtParam database[4]. The prediction of subcellular localization was implemented through the PSORT[5].

Chromosomal Mapping, Gene Structure, and Multiple Sequence Alignment

Chromosomal positions of *FveCOs* were plotted with MapInspect software[6].

Gene structure was drawn with Gene Structure Display Server[7] following the DNA and CDS information of *FveCOs*.

Multiple sequence alignment of FveCOs and other homologs was performed by ClustalW in the MEGA7 software package (Saitou and Nei, 1987; Tamura et al., 2011). Then, the alignment result was illustrated with JalView[8].

Phylogenetic Analyzes and Motif Analyzes

Phylogenetic analysis was performed using the CO/COL homologous sequences of several plant species together with the FveCOs. The CO/COL protein sequences used were obtained from the GDR database[9], Ensembl Plants[10], NCBI[11], and Phytozomev10[12]. MEGA7 software was used to construct the phylogenetic tree using the neighbor-joining (NJ) method and Jones–Taylor–Thornton (JTT) model by partial deletion with 2000 bootstrap replications.

The main motifs of FveCOs were characterized by the MEME program[13]. Then, the schematic diagrams of protein domain structure and conserved motif were illustrated with the TBtools software (Chen et al., 2020)[14].

Computation of Ka/Ks Values

ParaAT2.0 program[15] was used to perform the nucleotide sequence alignment of *FveCOs*. Non-synonymous and synonymous substitution rates (denoted as Ka and Ks,

[1]https://www.rosaceae.org/

[2]https://www.ncbi.nlm.nih.gov/

[3]http://pfam.xfam.org/

[4]https://www.expasy.org/

[5]https://www.genscript.com/psort.html

[6]http://mapinspect.software.informer.com/

[7]http://gsds.cbi.pku.edu.cn/

[8]https://www.jalview.org/

[9]https://www.rosaceae.org/

[10]https://plants.ensembl.org/index.html

[11]http://www.ncbi.nlm.nih.gov

[12]https://phytozome-next.jgi.doe.gov/

[13]http://meme-suite.org/tools/meme, v5.3.3

[14]https://github.com/twdb/tbtools

[15]https://ngdc.cncb.ac.cn/tools/paraat

respectively) were implemented by Ka/Ks_Calculator program[16]. The ratio of Ka/Ks was used to detect natural selection pressure.

Collinearity Analysis

Genome data of *F. vesca*, *A. thaliana*, *O. sativa,* and *Malus* × *domestic*a were used to analyze their collinearity and synteny relationships. The genome sequences and genome annotation files were downloaded from Phytozome v10[17]. MCScanX software was used to perform the whole gene collinearity analysis of the three species, and CO-like gene collinearity of the species was also stood out. The chart was manufactured with Circos 0.69, drawing software developed by Perl[18].

Analysis of the *cis*-Acting Elements

The upstream sequences (2000 bp) of *FveCOs* were collected for the analysis of cis-acting elements distributed in their promoter regions. The corresponding analysis was performed by online tools PlantCARE (Lescot et al., 2002)[19], and then, the results were exported with TBtools software[20].

Prediction of *FveCOs* Interaction Proteins

The interaction networks of FveCO proteins were predicted and constructed by the STRING v11.0[21]. The active interaction sources include text mining, experiments, databases, co-expression, neighborhood, gene fusion, and co-occurrence. The minimum required interaction score was set as 0.400.

Expression Detection of *FveCOs*

Tissue-specific expression analyzes and diurnal expression analyzes were carried out in the woodland strawberry LD-flowering accession "Ruegen." For other analyzes, the plants were field-grown in a greenhouse under natural LD conditions during the spring in ShenYang (Liaoning, China; 41°N, 123°E).

Tissue-specific expression detection was carried out with the samples of root (R), petiole (P), leaf (L), flower (F), shoot apex (SA), green fruit (GF), white fruit (WF), turning red fruit (TF), and red fully fruit (RF). All the samples were frozen in liquid nitrogen and laid at –80°C before total RNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) method as described in Koskela et al. (2012). The full-length cDNAs were synthesized using the PrimeScript RT reagent Kit (TaKaRa).

For diurnal expression analysis, the woodland strawberry "Ruegen" plants with three true leaves were moved to the artificial illumination incubators under 12-h light and 12-h dark conditions for 10 days. Then, the plants were moved into two artificial illumination incubators with different photoperiodic treatments, 25°C/18°C in day/night under LD (16-h light) and SD (8-h light) conditions, respectively. Leaves were then collected to detect the diurnal expression profiles of *FveCOs* at the beginning of the light phase (zeitgeber time 0, ZT0) under different photoperiodic conditions. The leaves were collected as materials every 4 h over 24 h, and the last time point is ZT24.

Pearson's correlation analyzes were performed with SPSS Statistics 22.0 (IBM Corporation, Armonk, USA) to explore the relationships between the flowering time and relative expression of *FveCO3* and *FveCO5*. Leaves of three plants under each mentioned photoperiod were sampled for expression detection of *FveCO3* and *FveCO5* when the first inflorescence appeared.

qRT-PCR was performed on the CFX96 Real-Time PCR System (Applied Biosystems, Foster City, CA, United States) using the SYBR Premix Ex Taq Kit (TaKaRa) according to the manufacturer's protocol. The *FveActin* served as an internal control. The relative expression of genes was presented by the $2^{-\Delta\Delta Ct}$ method. All of the above samples were executed independently in triplicate. Primers used in this study are listed in **Supplementary Table 1**.

Dual-Luciferase Assays

The dual-luciferase reporter assay was carried out. The 35S:FveCO3 and 35S:FveCO5 vectors were constructed (**Supplementary Figure 1**) and used as effectors, and the 2-Kb *FveFT1* promoter was inserted into the pGreen II 0800 vector and used as the reporter. The constructed vectors were transformed into *Agrobacterium* strain GV1301. *Agrobacterium* strains were introduced into tobacco (*Nicotiana tabacum*) leaves. The luciferase fluorescence and luciferase signal intensity were imaged and measured after three days using a living fluorescence imager (Lb985, Berthold, Germany).

RESULTS

Identification of *CO-like* Genes in Woodland Strawberry

To survey the *CO*-like members in the woodland strawberry, a genome-wide search against the GDR database was performed via selecting the typical B-box and CCT domains. A total of 10 distinct genes were identified as putative members of the woodland strawberry *CO-like* members. The gene names were entitled according to the order in which they were found. The detailed information of these genes is shown in **Table 1**.

The cDNA length of the *FveCOs* ranged from 1020 bp to 2376 bp, following the polypeptide sequences varying from 312 aa to 478 aa and the molecular weights of 51.77 kDa to 34.5 kDa. The prediction of FveCOs proteins showed that they were all located to the cell nucleus (**Table 1**).

Chromosomal distribution detection of *FveCOs* determined by the chromosome mapping indicated that these genes were unevenly distributed on five chromosomes (**Figure 1**). In detail, *FveCO4* and *FveCO6* were both located on chromosome 2, *FveCO1* and *FveCO7* were independently located on

[16]https://ngdc.cncb.ac.cn/tools/kaks

[17]https://phytozome-next.jgi.doe.gov/

[18]http://www.circos.ca/software/download/circos/

[19]http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

[20]https://github.com/twdb/tbtools

[21]https://string-db.org/

TABLE 1 | Sequence analysis of *FveCOs*.

Gene names	Gene ID	Length of cDNA (bp)	Length of ORF (bp)	AA	Chromosome	Position	MW (kDa)	pI	Prediction of protein location
FveCO1	gene00355	2376	1437	478	LG7	90927..92118	51.77	5.66	Nuclear
FveCO2	gene03742	1473	1353	450	LG4	962863..964335	50.33	5.56	Nuclear
FveCO3	gene04172	1863	1158	385	LG6	31553425..31557231	42.35	5.35	Nuclear
FveCO4	gene14981	1020	939	312	LG2	35460923..35461942	34.5	6.33	Nuclear
FveCO5	gene15552	2368	1254	417	LG6	23430145..23432512	45.25	5.45	Nuclear
FveCO6	gene24941	1664	1182	393	LG2	562983..564196	43.58	5.74	Nuclear
FveCO7	gene25171	1726	1374	456	LG5	26696242..26697967	51.23	5.56	Nuclear
FveCO8	gene27383	1191	942	356	LG4	10584825..10586015	38.99	5.63	Nuclear
FveCO9	gene03650	1269	912	303	LG4	24743480..24758335	33.28	6.25	Nuclear
FveCO10	gene03651	2554	1416	471	LG4	24759237..24764662	52.96	5.62	Nuclear

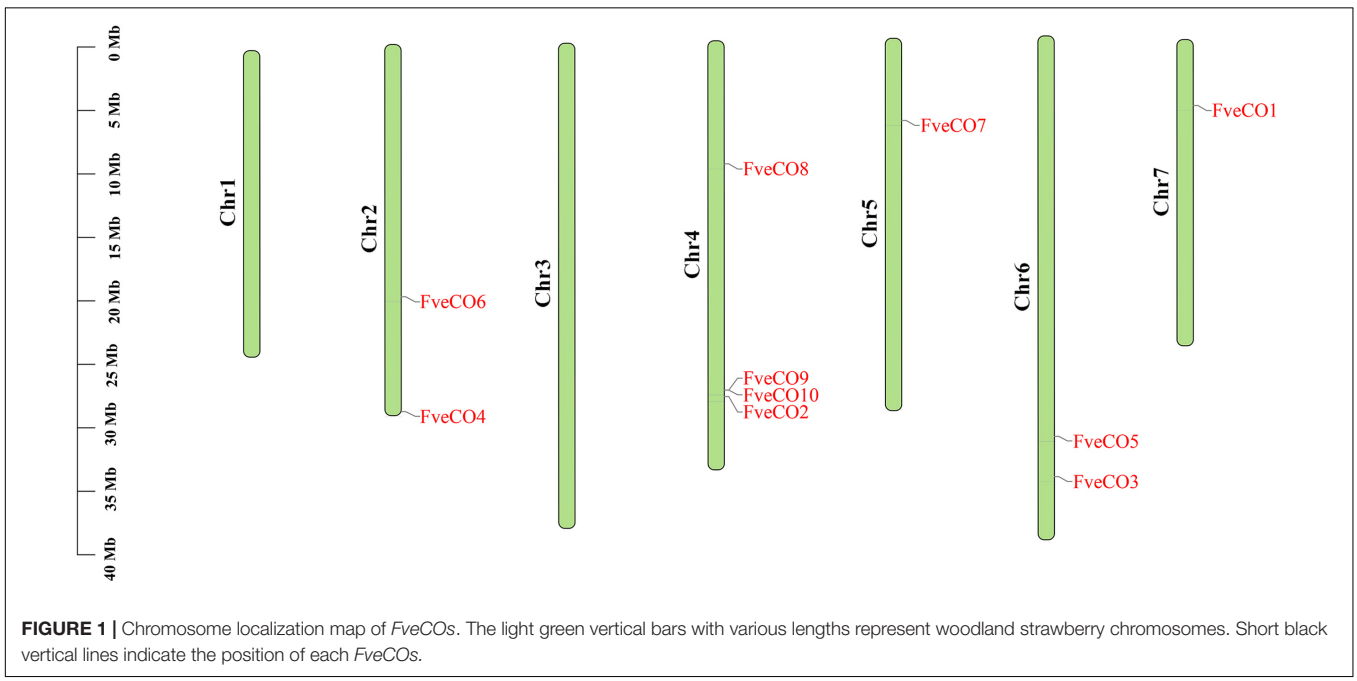


FIGURE 1 | Chromosome localization map of *FveCOs*. The light green vertical bars with various lengths represent woodland strawberry chromosomes. Short black vertical lines indicate the position of each *FveCOs*.

chromosomes 5 and 7, *FveCO2*, *FveCO8*, *FveCO9,* and *FveCO10* were located on chromosome 4, and *FveCO3* and *FveCO5* were located on chromosome 6.

Phylogeny and Multiple Sequence Alignment Analyzes of *FveCOs*

To fully identify the evolutionary relationship of CO homologs belonging to the woodland strawberry and other plant species, a phylogenetic tree was constructed with 121 CO-like amino acid sequences of 20 plant species, including 17 of *Arabidopsis*, 13 of tomato, 10 of tobacco, 16 of rice, 2 of apple, and so on. The results (**Figure 2**) suggested that these CO-like proteins could be subdivided into three groups, which are consistent with the previous findings. As shown in the tree, three *Fragaria vesca* CO homologs, namely, FveCO3, FveCO4, and FveCO8, belong to group I, two homologs, namely, FveCO2 and FveCO7, belong to group II, while the other five homologs, namely, FveCO1, FveCO5, FveCO6, FveCO9, and FveCO10, belong to group III.

Protein structures and evolutionary relationships could be elucidated by multiple sequence alignment. The alignment results of the amino acid sequences of FveCOs showed that all 10 FveCOs contained at least one B-box domain and one CCT domain. The identity of 10 FveCOs amino acid sequences ranged from 15.4% to 49.9% (**Supplementary Figure 2**).

According to the consistency difference of the amino acid sequence of B-box domain and the specificity of zinc ion-binding amino acid residues, the B-box domain can be further divided into two types: B-box1 (B1) and B-box2 (B2). Besides the B1 and B2 domains, an additional diverged zinc-finger structure (DZF) was also found in some CO-like proteins of *Fragaria vesca*.

B-box1 domain is the most conserved region in all the FveCOs proteins and is composed of approximately 40 residues, which owned the consensus C-X2-C-X7-9-C-X2-D-X4-C-X2-C-X3-4-H-X4-8-H, where X can be any amino acid. CCT domains are also highly conserved in FveCOs containing 43–45 amino acid residues.

Thus, according to the domain combination, FveCOs can be grouped into three types, similar to the three clades shown in

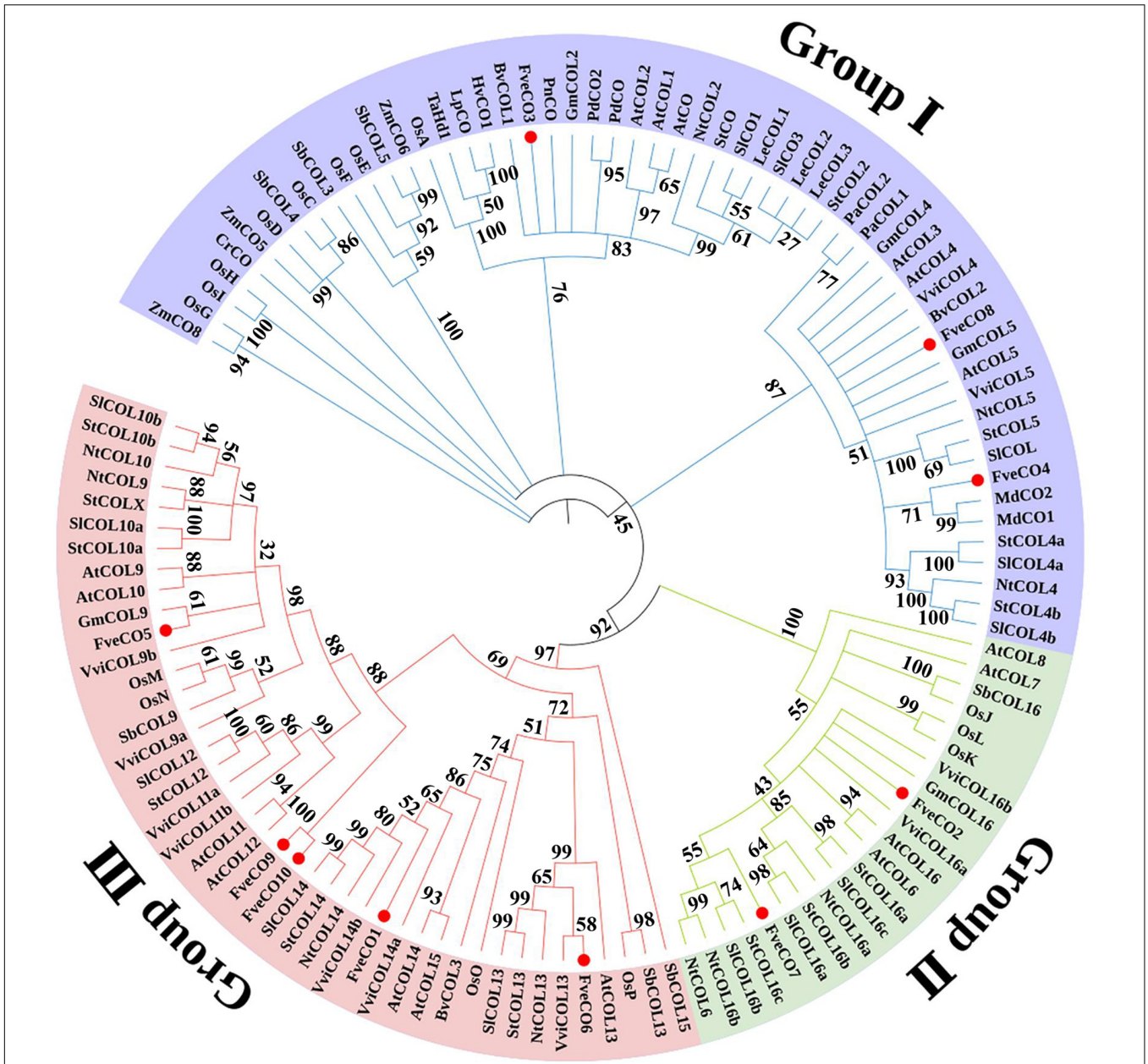


FIGURE 2 | Phylogenetic analysis of CONSTANS/CONSTANS-like (CO/COL) homologs in different species. The clades of groups I, II, and III are marked in green, blue, and pink, respectively. *Fragaria vesca constans-like genes* (FveCOs) are indicated by red points. At, *Arabidopsis thaliana*; Ca, *Capsicum annuum*; Gm, *Glycine max*; Hv, *Hordeum vulgare*; Lt, *Lolium temulentum*; Md, *Malus domestica*; Nt, *Nicotiana tabacum*; Os, *Oryza sativa*; Ph, *Petunia hybrida*; Sb, *Sorghum bicolor*; Sl, *Solanum lycopersicum*; St, *Solanum tuberosum*; Ta, *Triticum aestivum*; Vv, *Vitis vinifera*; Zm, *Zea mays*.

the phylogenetic tree (**Figure 3D**). Members in type I (FveCO3, FveCO4, and FveCO8) own two B-box domains and one CCT domain (B1 + B2 + CCT), members in type II (FveCO2 and FveCO7) own only one B-box domain (B1) and one CCT domain (B1 + CCT), while for members in type III (FveCO1, FveCO5, FveCO6, FveCO9, and FveCO10), besides a typical B1-type B-box domain and one CCT domain, the diverged zinc-finger structure domain was also detected (B1 + DZF + CCT).

Meme online software was used to further analyze the amino acid sequence similarity of those domains. The B-box domain is rich in amino acids such as A, C, D, and L, while the CCT domain is rich in arginine (R) and lysine (K) (**Supplementary Figure 3**). Although the B1, B2, and the DZF domains are all belonging to the zinc-finger domain, their amino acid sequences are variant, even in B1 and B2 (**Supplementary Figure 3**).

Gene Structure Analysis of *FveCOs*

Intron distribution is important for the gene selective splicing, allowing a gene to produce different proteins. The distribution of the *FveCOs* introns was identified by the comparative

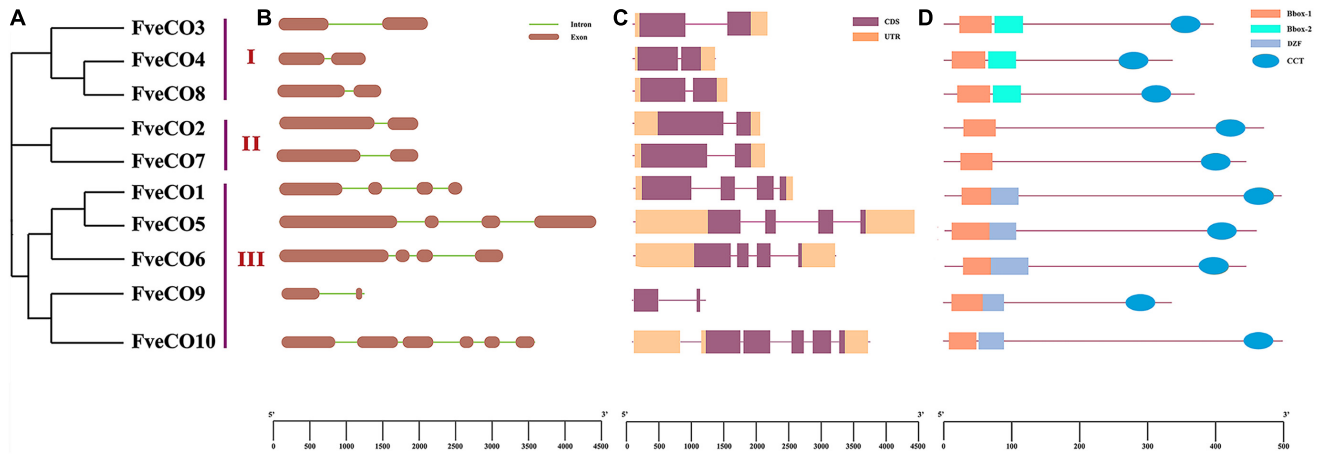


FIGURE 3 | Genomic and protein structures of *FveCOs*. **(A)** Phylogenetic analysis of 10 *FveCO* genes which are categorized them into three groups. **(B)** The intron distribution. The exons and introns are represented by dark ochre boxes and light green lines, respectively. **(C)** The CDS and UTR are represented by purple boxes and orange boxes, respectively. **(D)** The domain distribution of FveCOs. The CCT, B-box1, B-box2, and DZF are represented by light blue ovals, orange boxes, light green boxes, and blue–gray boxes, respectively.

analyzes of *FveCOs* DNA sequences with their coding sequences. Two exons and one intron were detected in all the members belonging to group I (*FveCO3, FveCO4,* and *FveCO8*) and group II (*FveCO2* and *FveCO7*). For members in group III, most of them own four exons and three introns, except *FveCO9* and *FveCO10*. *FveCO9* has only one intron just like members in groups I and II, while five introns were detected in *FveCO10* (**Figure 3B**).

The introns of *FveCO*s are 81 bp–1114 bp in length separately, leading to a large variation in their genomic length. The CDS and UTR information of *FveCOs* is listed in **Figure 3C**.

Synteny Analysis of *FveCOs*

To estimate the evolutionary character of woodland strawberry *FveCO* genes, the replication events about this gene family in the intraspecific and interspecific genomes were also analyzed. The results implied that only one pair of duplicate genes (*FveCO2*/*FveCO7*) was found in the genome of the woodland strawberry, which may be the result of tandem replication or whole-genome replication (WGD) (**Figure 4**).

To further explore the selection pressure between *FveCO* duplicate genes, we calculated the Ka, Ks, and Ka/Ks values of paralogous genes (**Table 2**). The divergence time of the woodland strawberry was estimated as 9.4 Mya (million years ago). Moreover, with the Ks value, we also calculate the substitutions rate of per site per year as $5.8 \times e^{-8}$.

Moreover, to explore the evolution mechanism and biochemical features of the *FveCOs*, a collinearity comparison of *Fragaria vesca* genome with genomes of *Arabidopsis*, rice, and apple, belonging to dicotyledon plant, monocotyledon plant, and Rosaceae plant, respectively, was also performed. The results showed that there are 12 collinear gene pairs between *F. vesca* and *A. thaliana*, 17 pairs between *M* × *domestica* and *F. vesca*, and 5 pairs between *O. sativa* and *F. vesca* (**Figure 5**). Orthologous gene numbers identified between *F. vesca* and *M* × *domestica*, which are all belonging to Rosaceae, were much larger than those identified between woodland strawberry and *Arabidopsis* and rice.

Both *FveCO4* and *FveCO6* were located on the second chromosome, and their orthologous gene pairs were detected in the rice genome. Orthologous gene pairs of *FveCO1*, *FveCO2*, *FveCO3*, *FveCO4*, *FveCO5*, *FveCO6*, *FveCO7,* and *FveCO8* were identified in both apple and *Arabidopsis*, and the remaining *FveCOs* were not present in any of the duplicated blocks. These results suggested that those genes are highly conservative and the collinear gene pairs may come from the common ancestor before evolution.

Cis-Element Analysis of Woodland Strawberry *FveCOs*

Promoters with 2 kb in length of each *FveCO* members were used to identify the putative cis-acting regulatory elements (CREs) with the PlantCARE database. Totally, 43 kinds of CREs were identified (**Supplementary Table 2**). Among them, there are 283 core promoter elements (TATA box) and 109 common *cis*-acting elements (CAAT box).

The other *cis*-acting elements identified might be divided into three main kinds, namely hormone response element, light response element, and stress response element.

Fragaria vesca constans-like genes may also play crucial roles in the regulation of photoperiod flowering just like their reported homologs; thus, we further analyzed the light response elements distributed in their promoters. As shown in **Figure 6**, a large number of light-responsive elements were detected in all the promoter regions of 10 *Fve*CO genes, while the kinds and amounts are various. Promoter of *FveCO3* owns 21 light-responsive *cis*-elements with the maximum quantity, while promoter of *FveCO6* only has three light-responsive *cis*-element with the minimum number.

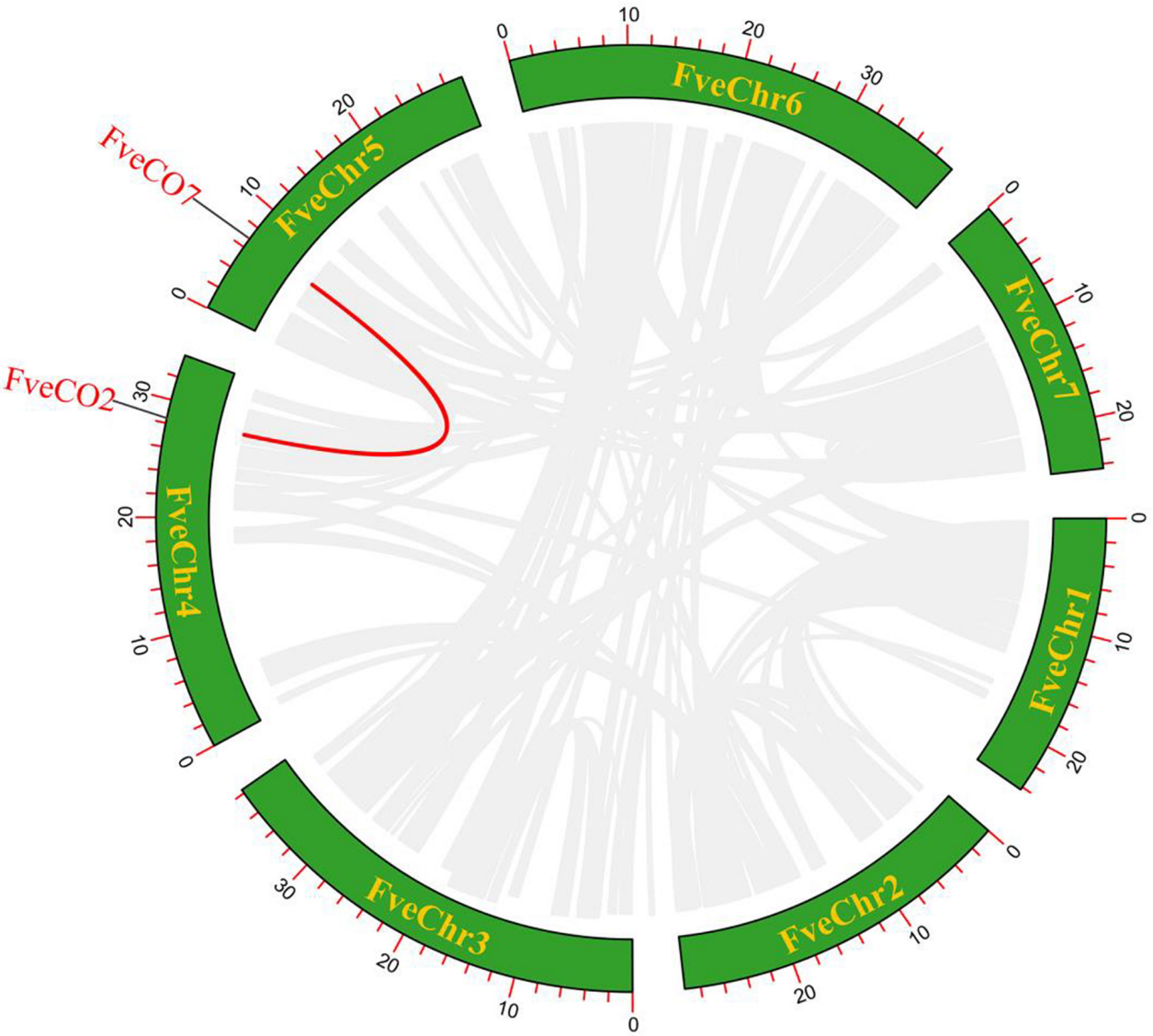


FIGURE 4 | Schematic representation of the interchromosomal relationships between *FveCOs* in the *Fragaria vesca* genome. Red line indicates the colinear gene pair, and gray lines indicate the syntenic blocks.

Expression of *FveCOs* in Different Tissues

The qRT-PCR results showed that *FveCO*s were variously expressed in different tissues (**Figure 7**). All *FveCO* genes exhibited higher expression levels in leaves and petioles than those in other tissues. For fruits in different development stages, *FveCOs* are mainly expressed in green fruit, especially for the expression of *FveCO2*, *FveCO4*, *FveCO7*, *FveCO8,* and *FveCO10*. It was also found that the lowest *FveCO* expression level was detected in the fully red fruit.

Among all of these genes, *FveCO3* and *FveCO4* showed similar high expression levels in different tissues, while the expression of *FveCO9* was border on the minimum in each tested tissue. The specific and varied expression profiles of *FveCO*s in different tissues suggest that they may play diverse roles.

Expression Profile of *FveCOs* in Photoperiodic Flowering

According to the previous studies, the expression of the *CO-like* genes exhibits a circadian rhythm profile in most plant species (Suárez-López et al., 2001; Wang et al., 2013; Fu et al., 2015; Chaurasia et al., 2016). Therefore, to evaluate the potential functions of *FveCOs* in photoperiodic flowering, we detected their diurnal expression profiles over a 24-h period at 4-h intervals under LD and SD conditions, separately.

As shown in **Figure 8**, three genes, including *FveCO1*, *FveCO2*, and *FveCO5*, owned similar expression patterns under

TABLE 2 | Ka/Ks analysis for the duplicated *FveCO* paralogs.

	Ka	Ks	Ka/Ks	Purifying selection	Duplicate type
FveCO2/FveCO7	0.4225	2.9676	0.1423	Yes	Segmental

both photoperiodic conditions, with the highest expression level that appeared at ZT16 h under LD condition and then gradually reduced at night. However, the expression levels of these three genes slightly increased during the day of SD and peaked at midnight. In addition, the peak expressions of these genes were higher in plants grown under LD condition than under SD condition.

FveCO4, together with *FveCO6, FveCO7,* and *FveCO10,* showed the consistent expression pattern under SD and LD conditions, which were slightly similar to that of *FveCO1*, *FveCO2*, and *FveCO5*. The expression of *FveCO4*, *FveCO6*, *FveCO7,* and *FveCO10* under SD condition was detected higher separately than that under LD condition. The expression peaks appeared at ZT20 h and ZT16 h for SD and LD conditions, respectively. Under SD condition, their expression decreased rapidly to the lowest level.

Correlation of Woodland Strawberry Flowering Time and the *FveCOs* Expression

To further explore the potential function of *FveCOs* on strawberry flowering time, we also investigated the correlation between the flowering time and the expression levels of *FveCOs*. In all the *FveCOs*, only the expression of *FveCO3* and *FveCO5* showed the correlation with the flowering time. The results showed that the expression levels of *FveCO3* and *FveCO5* were all negatively correlated with the flowering time under both LD and SD conditions (r = −0.949, −0.964, −0.936, and −0.891, respectively) (**Figure 9**). Plants under LD condition showed earlier flowering time with higher expression levels of both *FveCO3* and *FveCO5,* while plants grown under SD condition owned the lower expression levels of both *FveCO3* and *FveCO5*.

FveCO3 and *FveCO5* Activate the Expression of *FveFT1*

To explore how *FveCO3* and *FveCO5* regulate photoperiodic flowering in the woodland strawberry, luciferase reporter assay was carried out. The promoter sequence of 2000-bp upstream of the ATG codon of *FveFT1* was cloned and inserted into the upstream of LUC reporter gene. The effector plasmid containing 35S-FveCO3 and 35S-FveCO5 construct was co-transfected into tobacco leaves, respectively. The luciferase assays indicated that co-expression of 35S-FveCO3 and *proFveFT1-LUC* or 35S-FveCO5 and *proFveFT1-LUC* resulted in much stronger luminescence signals than any other points (**Figure 10**). These results showed that FveCO3 and FveCO5 could bind to the *FveFT1* promoter individually and thus activate the transcription of the corresponding gene.

Moreover, the interaction networks of *FveCO3* and *FveCO5* with other flowering-related proteins were predicted and constructed by the STRING v11.0 (see the footnote 21). As shown in **Figure 11**, multiple genes were screened as the candidate interactors of FveCO3 or FveCO5, including FveSOC1, FveHOS1, FveAGLs, FveGI, FveFBH, and FveAP1.

DISCUSSION

Molecular Characteristics of *FveCOs* in Woodland Strawberry

Normally, plants can be classified as either SD or LD type according to their flowering response to the photoperiod. In strawberry, two types of flowering habit exist in different varieties, known as the June-bearing SD type and perpetually flowering everbearing type. For the conserved function of CO-FT module in plant photoperiodic flowering regulation, the characterization of the roles of *CO-like* genes in strawberry flowering habit variation is important for its breeding.

Actually, *CO-like* genes of strawberry had been reported in some previous studies. Kurokura reported one woodland strawberry *CO* homolog (gene 04172) and another nine putative CO-like proteins in *F. vesca* genome (V1.1) using the B-box domain as the search query (Kurokura et al., 2017). Among those putative CO-like proteins, six proteins are included in the FveCOs we reported here, while the CCT domain cannot be detected in the remaining four proteins they reported, so those four proteins should not be classified as CO-like proteins.

Using the updated *Fragaria vesca* genome database v4.0.a1, here we reported the 10 non-redundant FveCO candidates, including the new four ones which had not been reported. All the FveCO candidates contain both B-box domain and CCT domain (**Figure 3**). The 10 identified FveCOs can be divided into three groups depending on the phylogenetic analysis (**Figures 2**, **3**), which was consistent with the reports in other plants (Putterill et al., 1995; Cheng and Wang, 2005; Hassidim et al., 2009; Song et al., 2015).

Although the *Arabidopsis* genome size (130 Mb) is much smaller than that of the woodland strawberry (230 Mb), the number of *Arabidopsis COL* member is much larger than that in the woodland strawberry (17:10). The gene structure and conserved domain of FveCOs are similar to the homologs of other species. Conserved domain-based CO proteins can also be divided into three subfamilies in the woodland strawberry. However, the gene structure classification of these FveCOs proteins is quite complex. In group III, the gene structures of *FveCO9* and *FveCO10* were different, but their amino acid sequences encoded were similar to each other. Such variation may come from the duplication, variation and recombination of exons, or the insertion and loss of introns in the process of genome evolution.

The existing of duplicated genes implied that independent duplication events existed besides the whole-genome duplication event. Only one pair of *FveCOs* (*FveCO2/7*) is identified to be involved in fragment replication. It should be the reason that the woodland strawberry genome experienced few replication

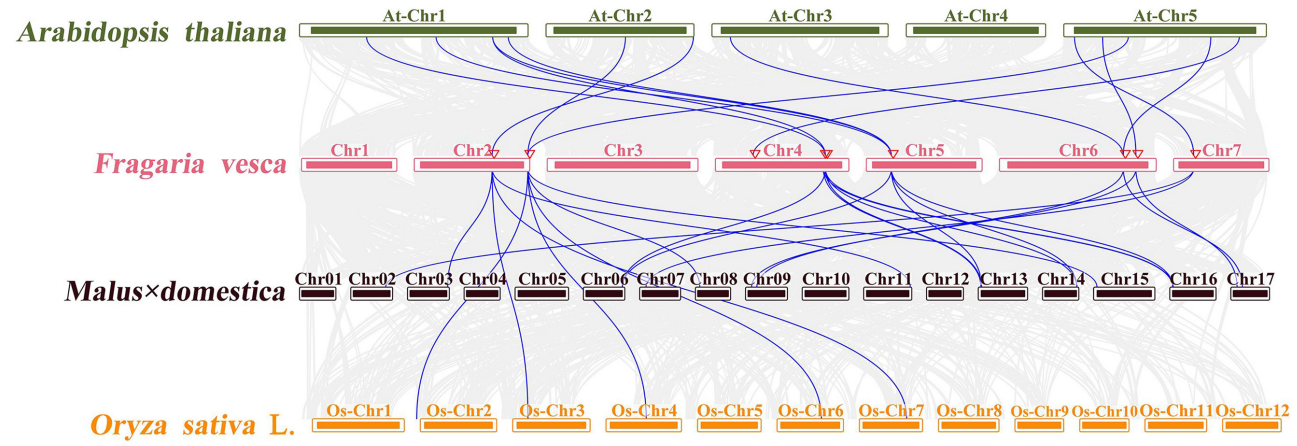


FIGURE 5 | Microsynteny of CONSTANS-like *(COL)* regions across *Fragaria vesca*, *M* × *domestica*, *O. sativa,* and *A. thaliana*. The *F. vesca*, *M* × *domestica*, *O. sativa,* and *A. thaliana* chromosomes are shown in different colors. The number on each chromosome box indicates the number of chromosomes. Blue lines represent the syntenic relationships between *AtCOLs*, *MdCOs*, *OsCOLs,* and *FveCOs* regions. Light gray lines represent the syntenic relationships between genome wide of *Arabidopsis*, woodland strawberry, apple, and rice.

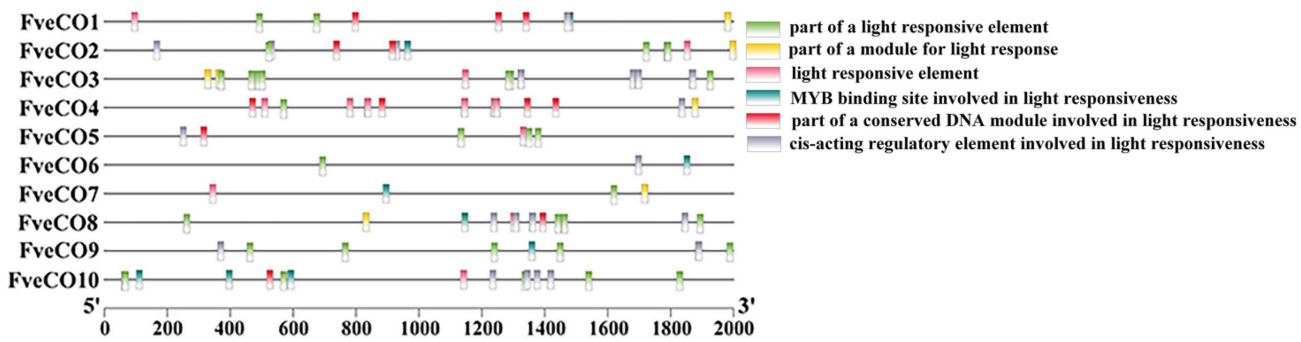


FIGURE 6 | Distribution of the major light-related *cis*-elements in the promoter sequences of 10 *FveCO* genes. Different *cis*-elements are represented by different symbols as indicated.

events as a kind of diploid variety. Segmental duplication was a kind of genome replication. Tandem duplications were characterized as multiple members of one family occurring within the same intergenic region or in neighboring intergenic regions. The most representative tandem replication genes are adjacent homologous genes on a single chromosome (Moore and Purugganan, 2003; Yu et al., 2005). In this study, *FveCO9* and *FveCO10* are adjacent genes on chromosome 4. Therefore, FveCO9 and FveCO10 are probably the results of tandem replication.

Collinearity analysis between the genomes of the woodland strawberry and other three plants suggested that the *FveCOs* own the highest relationship with their homologs in apple, followed by *Arabidopsis thaliana*, and the lowest relationship with rice orthologous gene. Such results are consistent with their taxonomical relationship: Strawberry and apple are both Rosaceae plants.

The expansion of all *CO* genes occurs with the divergence of the plants of the same family. The subclasses of *CO* genes extend from the common ancestral genes in a species-specific manner, which existed before the diversification of the same family lineage.

Expression Profile of *FveCOs* Indicates Their Potential Functions in the Photoperiodic Flowering Regulation

Most of the *FveCO* genes are preferentially expressed in leaf and petiole (**Figure 7**), which are consistent with the findings in other species (Almada et al., 2009; Tan et al., 2016; Wu et al., 2017). As leaf is the organ in which the plant percepts the light, such result strongly demonstrated that *FveCOs* might also involve in the photoperiodic sensitive response. Besides the leaf and petiole, the expression of *FveCOs* could also be detected in many tissues except the root and red fully fruit, which implies that *FveCO* members might also function in other multiple developmental aspects of the woodland strawberry. On the contrary, the expression of the different *FveCO* members could be detected in the same tissue, which might result in their functional redundancy.

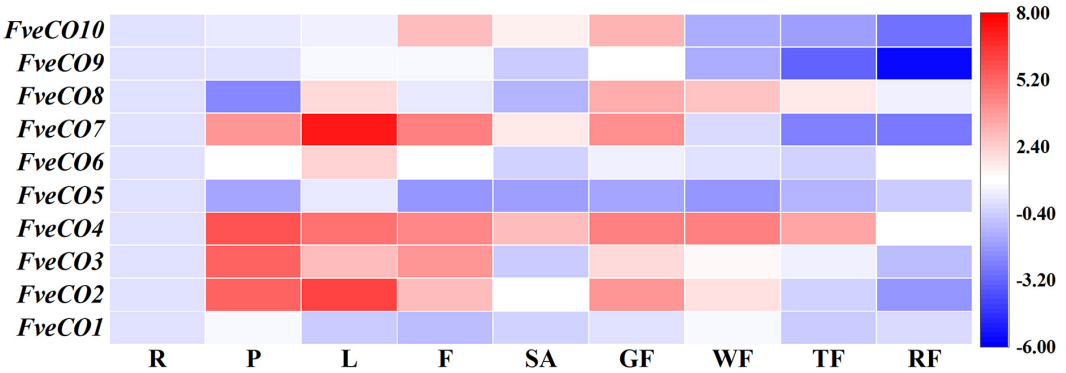


FIGURE 7 | Expression of *FveCOs* in different woodland strawberry tissues. Different colors of heat map represent the expression levels. The tissues of the samples are as follows: R, root; P, petiole; L, leaf; F, flower; SA, shoot apex; GF, green fruit; WF, white fruit; TF, turning red fruit; RF, red fully fruit. For each gene, the expression level was set to 1 in the root, and the corresponding fold changes were calculated in other tissues. The gene expression heatmap was generated on the log base 2 average expression fold values.

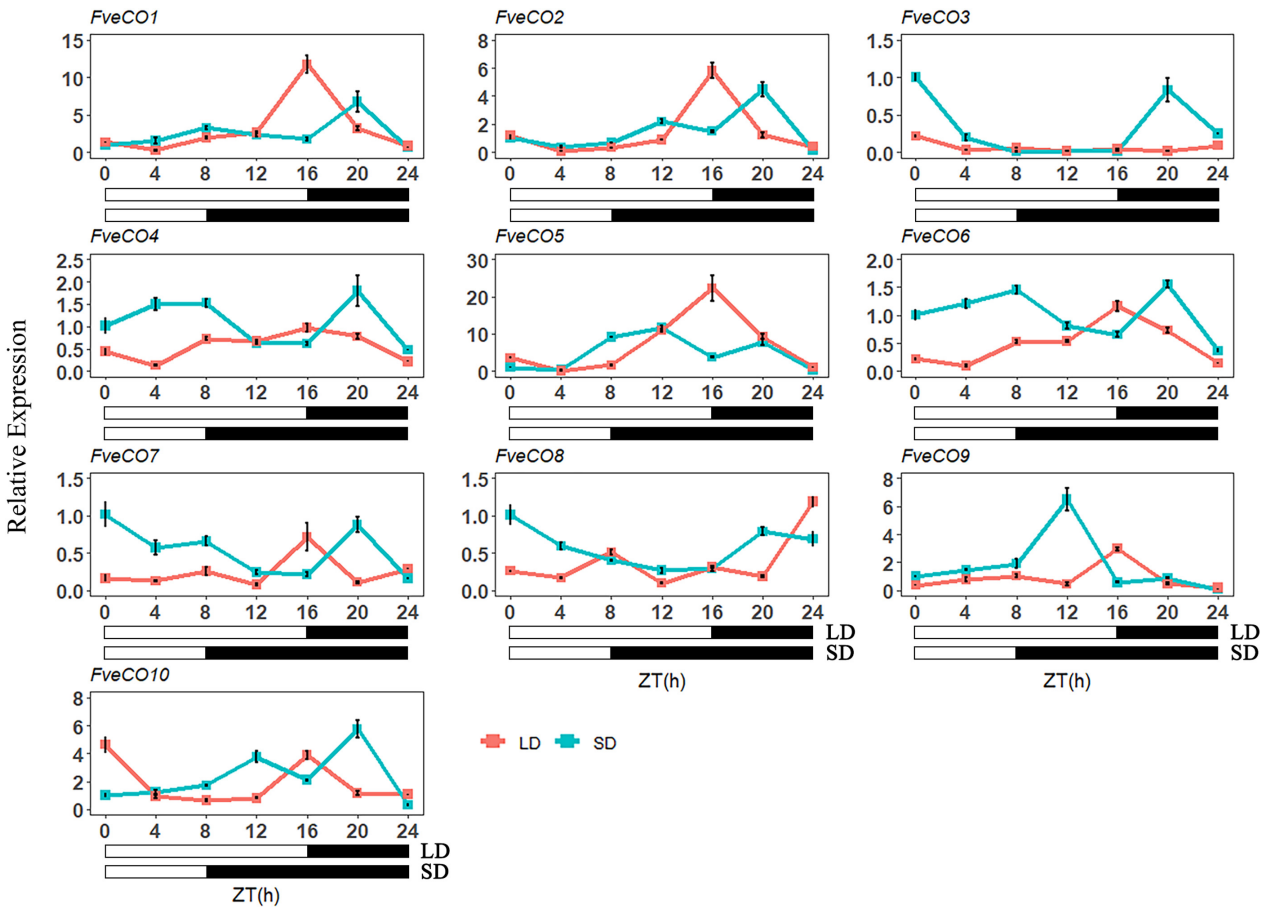


FIGURE 8 | Diurnal expression pattern of *FveCOs* under long-day (LD) and short-day (SD) conditions. The collection of samples was started at the beginning of the light phase (zeitgeber time 0, ZT0) and continued every 4 h over 24 h under LD (16-h light/8-h dark) and SD (8-h light/16-h dark) conditions. Each value is the mean ± SD of three biological replicates.

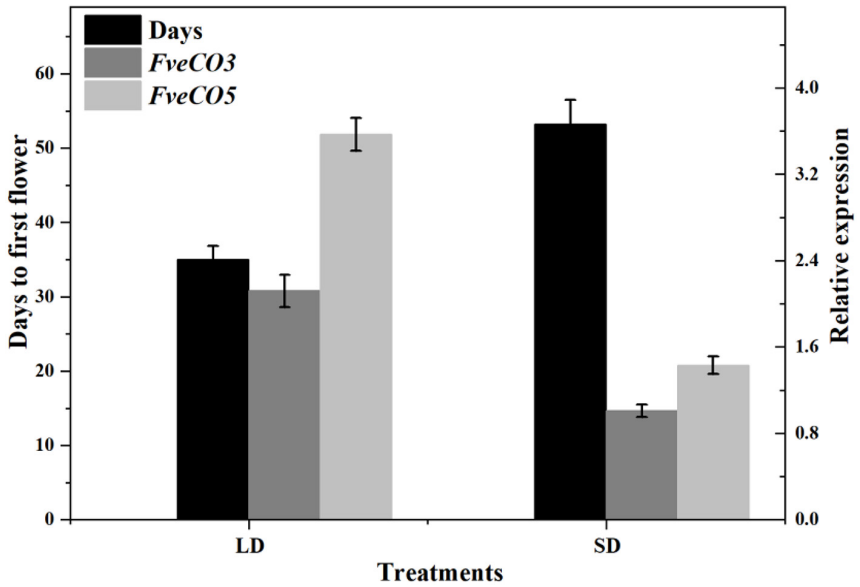


FIGURE 9 | Negative correlation of *FveCO3* and *FveCO5* expression with flowering time in woodland strawberry "Ruegen." The black bars indicate the days needed to flower (flowering time); the gray bar and light gray bar indicate the expression of *FveCO3* and *FveCO5*, respectively. Ten plantlets were used for each. Three independent samples were used in the expression analysis. Values are the mean ± SD.

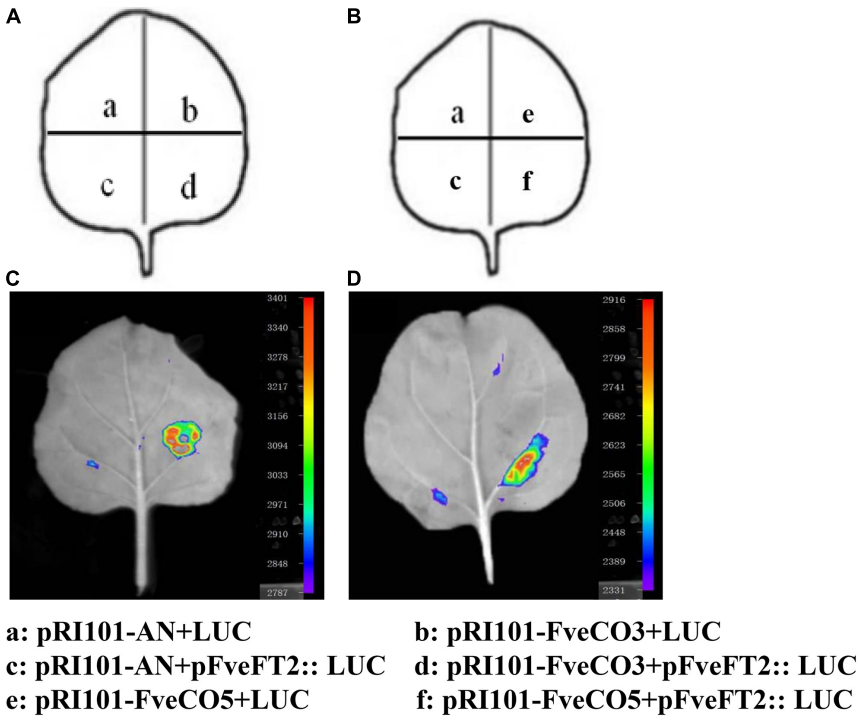


FIGURE 10 | *FveCO3* and *FveCO5* regulate *FveFT1* expression by binding to its promoter. **(A,B)** Schematic diagram of the reporter and effectors used in luciferase reporter assay. Luciferase reporter assay showing that FveCO3 and FveCO5 regulate *FveFT1* expression **(C,D)**, respectively. The FveCO3 and FveCO5 effector and the *proFveFT1* reporter were coinfiltrated into tobacco leaves, and the luciferase signal was measured.

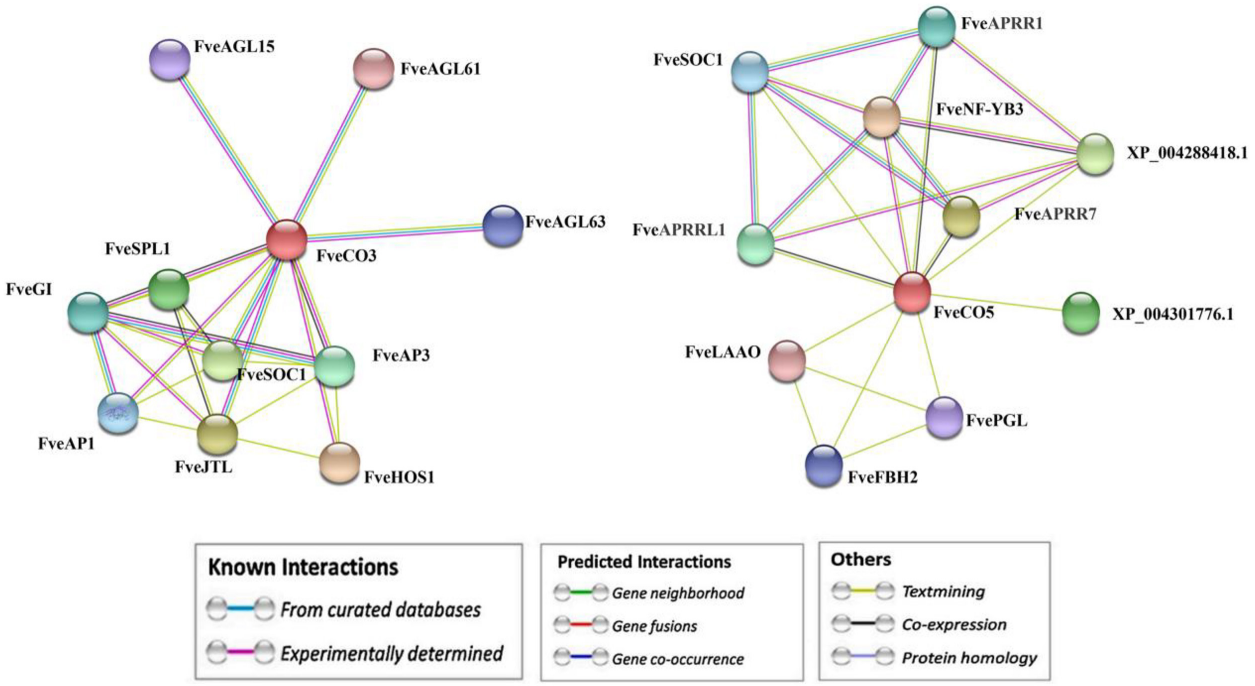


FIGURE 11 | Interaction network prediction of *FveCO3* and *FveCO5* with other proteins. The *Fragaria vesca constans-like genes* (*FveCOs*) are marked in red.

At the same time, our result also showed that the expression of most *FveCO* genes can be regulated by circadian clock (**Figure 8**), which resulted in their diurnal expression changes. The one-day periodic changing expression profile is the prominent character of almost all the CO-like genes (Martin et al., 2004; Pan et al., 2021). When treated with different photoperiods, the *FveCOs* showed different diurnal expression patterns under SD or LD condition. *FveCO* genes showed different circadian rhythms, including at least four peaks at different ZT time points. Among them, the expression of *FveCO1*, *FveCO2*, *FveCO5,* and *FveCO9* showed notable expressional amplitudes during the circadian rhythms at the time points of measurement in 24 h under both LD and SD conditions.

Interestingly, it was found that the individual diurnal expression profile of *FveCO3* and *FveCO5* was opposite to each other under both LD and SD conditions. Further work should be performed to clarify the mechanism about such different expression profiles, such as the transcriptional or posttranscriptional regulation research of the upstream regulators on *FveCO3* and *FveCO5*.

FveCO3 and *FveCO5* Function as Potential Flowering Promoters in Woodland Strawberry

Our results showed that, under either LD or SD condition, the expression levels of *FveCO3* and *FveCO5* were both negatively correlated with flowering time (**Figure 9**). *FveCO3* is the woodland strawberry homolog of *AtCO* belonging to group I type, while *FveCO5* is the homolog of *AtCOL9* which negatively regulates the expression of *AtCO* and *AtCOL9* to inhibit the flowering in LD. Previous studies have shown that the expression of *AtCO* is induced by SD rather than by LD (Putterill et al., 1995; Robson et al., 2001; Kotake et al., 2003; Jang et al., 2008), while the expression of *FveCO3* in SD or LD was similar to each other, and the same expression profile was also detected in *FveCO5*. Further work about the protein interaction networks prediction of FveCO3 and FveCO5 showed that they could interact with many other reported flowering regulators, such as FveSOC1, FveHOS1, FveAGLs, FveGI, FveFBH, and FveAP1(**Figure 11**). In *Arabidopsis*, SOC1 is one of the direct targets of CO and directly regulates the flowering process (Samach et al., 2000). GI and FBH can both bind to the upstream CO negative regulatory transcription factor CDFs and thus to degrade the CDFs and positively regulates the stability of CO (Toledo-Ortiz et al., 2003; Fornara et al., 2009). GI also promotes the FT expression to regulate flowering (Imaizumi et al., 2005; Sawa and Kay, 2011; Ito et al., 2012; Fornara et al., 2015). Such finding implied that *FveCO3* and *FveCO5* should both function in the woodland strawberry flowering and the mechanism should be more complex than that of *Arabidopsis*. The function of CO/COL proteins mainly depends on their regulation of the expression of the *FT*-like genes, the master regulators in plant flowering process. *RcCO* is the homolog of *FveCO3* in rose, and it regulates the photoperiodic flowering under long-day condition (Lu et al., 2020). RcCO could bind to the CORE motif of the *RcFT* promoter so as to enhance the *RcFT* expression. In the woodland strawberry, *FveFT1* has been identified as the key gene to determine the flowering time

(Koskela et al., 2012). In this study, luciferase reporter assay suggested that both FveCO3 and FveCO5 could directly bind to the *FveFT1* promoter individually and thus may promote the flowering process by transcriptional regulation. Similar upregulation of *FveFT1* was also reported in the strawberry *FveCO* overexpression lines (Kurokura et al., 2017). Those findings suggested that the CO-FT module also functions in the photoperiodic flowering of strawberry.

CONCLUSION

Totally, 10 distinct *CO*-like genes (*FveCOs*) in the woodland strawberry (*F. vesca*) were identified. The expression analysis indicated that multiple *FveCO* genes were highly responsive to the photoperiodic induction. Both *FveCO3* and *FveCO5* are potential positive regulators for photoperiodic flowering, which is different from their individual homologs in *Arabidopsis*. FveCO3 and FveCO5 can bind to the promoter of *FveFT1*, the key reported flowering regulator. The mechanism of *FveCO3* and *FveCO5* in strawberry flowering regulation should be deeply clarified with further work.

BRIEF SUMMARY

Ten distinct *CO-like* genes (*FveCOs*) were identified in the woodland strawberry (*F. vesca*). The expression of *FveCO3* and *FveCO5* was both negatively correlated with the flowering time variation of the woodland strawberry grown under both long-day and short-day conditions. FveCO3 and FveCO5 may function in flowering induction *via* the photoperiodic regulation in the woodland strawberry.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

YL designed the research. XZ, FY, QG, and YW performed the experiments. XZ and YL analyzed the data. XZ wrote the manuscript. YL and ZZ assisted with interpretation of results, manuscript writing, and revision. All authors read and approved the manuscript.

FUNDING

This work was financially supported by the National Natural Science Foundation of China (No. 31872073) and the Natural Science Foundation of Liaoning Province (Nos. 2014027015 and 20180550431).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.931721/full#supplementary-material

REFERENCES

Almada, R., Cabrera, N., Casaretto, J. A., Ruiz-Lara, S., and Villanueva, E. G. (2009). VvCO and VvCOL1, two CONSTANS homologous genes, are regulated during flower induction and dormancy in grapevine buds. *Plant Cell Rep.* 28, 1193–1203. doi: 10.1007/s00299-009-0720-4

Chaurasia, A. K., Patil, H. B., Azeez, A., Subramaniam, V. R., Krishna, B., Sane, A. P., et al. (2016). Molecular characterization of CONSTANS-Like (COL) genes in banana (*Musa acuminata* L. AAA Group, cv. Grand Nain). *Physiol. Mol. Biol. Plants.* 22, 1–15. doi: 10.1007/s12298-016-0345-3

Chen, C. J., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant.* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, J., Chen, J. Y., Wang, J. N., Kuang, J. F., Shan, W., and Lu, W. J. (2012). Molecular characterization and expression profiles of MaCOL1, a CONSTANS-like gene in banana fruit. *Gene* 496, 110–117. doi: 10.1016/j.gene.2012.01.008

Cheng, X. F., and Wang, Z. Y. (2005). Overexpression of COL9, a CONSTANS-LIKE gene, delays flowering by reducing expression of CO and FT in *Arabidopsis thaliana*. *Plant J.* 43, 758–768.

Crocco, C. D., and Botto, J. F. (2013). BBX proteins in green plants: insights into their evolution, structure, feature and functional diversification. *Gene* 531, 44–52. doi: 10.1016/j.gene.2013.08.037

Datta, S., Hettiarachchi, G. H., Deng, X. W., and Holm, M. (2006). *Arabidopsis* CONSTANS-LIKE3 is a positive regulator of red light signaling and root growth. *Plant Cell* 18, 70–84. doi: 10.1105/tpc.105.038182

Deng, X. D., Fan, X. Z., Li, P., and Fei, X. W. (2015). A photoperiod-regulating gene CONSTANS is correlated to lipid biosynthesis in *Chlamydomonas reinhardtii*. *Biomed Res. Int.* 2015:715020. doi: 10.1155/2015/715020

Domagalska, M. A., Elzbieta, S., Ferenc, N., and Davis, S. J. (2010). Genetic analyses of interactions among gibberellin, abscisic acid, and brassinosteroids in the control of flowering time in *Arabidopsis thaliana*. *PLoS One* 5:e14012. doi: 10.1371/journal.pone.0014012

Fornara, F., de, Montaigu, A., Sánchez-Villarreal, A., Takahashi, Y., Ver, Loren. Van. Themaat, E., Huettel, B., et al. (2015). The GI-CDF module of *Arabidopsis* affects freezing tolerance and growth as well as flowering. *Plant J.* 81, 695–706. doi: 10.1111/tpj.12759

Fornara, F., Panigrahi, K. C., Gissot, L., Sauerbrunn, N., Rühl, M., and Jarillo, J. A. (2009). *Arabidopsis* DOF transcription factors act redundantly to reduce CONSTANS expression and are essential for a photoperiodic flowering response. *Dev. Cell* 17, 75–86. doi: 10.1016/j.devcel.2009.06.015

Fu, J. X., Yang, L. W., and Dai, S. L. (2015). Identification and characterization of the CONSTANS-like gene family in the short-day plant *Chrysanthemum lavandulifolium*. *Mol. Genet. Genomics* 290, 1039–1054. doi: 10.1007/s00438-014-0977-3

González-Schain, N. D., Díaz-Mendoza, M., Zurczak, M., and Suárez-López, P. (2012). Potato CONSTANS is involved in photoperiodic tuberization in a graft-transmissible manner. *Plant J.* 70, 678–690. doi: 10.1111/j.1365-313X.2012.04909.x

Griffiths, S., Dunford, R. P., Coupland, G., and Laurie, D. A. (2003). The evolution of CONSTANS-like gene families in barley, rice, and *Arabidopsis*. *Plant Physiol.* 131, 1855–1867. doi: 10.1104/pp.102.016188

Hassidim, M., Harir, Y., Yakir, E., Kron, I., and Green, R. M. (2009). Over-expression of CONSTANS-LIKE 5 can induce flowering in short-day grown *Arabidopsis*. *Planta* 230, 481–491. doi: 10.1007/s00425-009-0958-7

Hayama, R., and Coupland, G. (2003). Shedding light on the circadian clock and the photoperiodic control of flowering. *Curr. Opin. Plant Biol.* 6, 13–19. doi: 10.1016/s1369-5266(02)00011-0

Huang, G. W., Ma, J. H., Han, Y. Z., Chen, X. J., and Fu, Y. F. (2011). Cloning and expression analysis of the soybean CO-Like gene GmCOL9. *Plant Mol. Biol. Rep.* 29, 352–359. doi: 10.1007/s11105-010-0240-y

Imaizumi, T., Schultz, T. F., Harmon, F. G., Ho, L. A., and Kay, S. A. (2005). FKF1 F-box protein mediates cyclic degradation of a repressor of *CONSTANS* in *Arabidopsis*. *Science* 309, 293–297. doi: 10.1126/science.1110586

Ito, S., Song, Y. H., Josephson-Day, A. R., Miller, R. J., Breton, G., Olmstead, R. G., et al. (2012). FLOWERING BHLH transcriptional activators control expression of the photoperiodic flowering regulator CONSTANS in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3582–3587. doi: 10.1073/pnas.1118876109

Izawa, T. (2021). What is going on with the hormonal control of flowering in plants? *Plant J.* 105, 431–445. doi: 10.1111/tpj.15036

Jang, S., Marchal, V., Panigrahi, K. C., Wenkel, S., Soppe, W., Deng, X. W., et al. (2008). *Arabidopsis* COP1 shapes the temporal pattern of CO accumulation conferring a photoperiodic flowering response. *EMBO J.* 27, 1277–1288. doi: 10.1038/emboj.2008.68

Koskela, E. A., Mouhu, K., Albani, M. C., Kurokura, T., Rantanen, M., Sargent, D. J., et al. (2012). Mutation in TERMINAL FLOWER1 reverses the photoperiodic requirement for flowering in the wild strawberry *Fragaria vesca*. *Plant Physiol.* 159, 1043–1054. doi: 10.1104/pp.112.196659

Kotake, T., Takada, S., Nakahigashi, K., Ohto, M., and Goto, K. (2003). *Arabidopsis* TERMINAL FLOWER 2 gene encodes a heterochromatin protein 1 homolog and represses both FLOWERING LOCUS T to regulate flowering time and several floral homeotic genes. *Plant Cell Physiol.* 44, 555–564. doi: 10.1093/pcp/pcg091

Kurokura, T., Samad, S., Koskela, E., Mouhu, K., and Hytönen, T. (2017). *Fragaria vesca* CONSTANS controls photoperiodic flowering and vegetative development. *J. Exp. Bot.* 68, 4839–4850. doi: 10.1093/jxb/erx301

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Liu, J. H., Shen, J. Q., Xu, Y., Li, X. H., Xiao, J. H., and Xiong, L. Z. (2016). Ghd2, a CONSTANS-like gene, confers drought sensitivity through regulation of senescence in rice. *J. Exp. Bot.* 67, 5785–5798. doi: 10.1093/jxb/erw344

Liu, L., Ma, J., Han, Y., Chen, X., and Fu, Y. F. (2011). The isolation and analysis of a soybean CO homologue GmCOL10. *Russ. J. Plant Physiol.* 58, 330–336. doi: 10.1134/S1021443711020105

Lu, J., Sun, J., Jiang, A., Bai, M., Fan, C., Liu, J., et al. (2020). Alternate expression of CONSTANS-LIKE 4 in short days and CONSTANS in long days facilitates day-neutral response in *Rosa chinensis*. *J. Exp. Bot.* 71, 4057–4068.

Martin, J., Storgaard, M., Andersen, C. H., and Nielsen, K. K. (2004). Photoperiodic regulation of flowering in perennial ryegrass involving a CONSTANS-like homolog. *Plant Mol. Biol.* 56, 159–169. doi: 10.1007/s11103-004-2647-z

Michaels, S. D., Himelblau, E., Sang, Y. K., and Amasino, S. (2005). Integration of flowering signals in winter-annual *Arabidopsis*. *Plant Physiol.* 137, 149–156. doi: 10.1104/pp.104.052811

Moore, R. C., and Purugganan, M. D. (2003). The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15682–15687. doi: 10.1073/pnas.2535513100

Mouradov, A., Crèmer, F., and Coupland, G. (2002). Control of flowering time: interacting pathways as a basis for diversity. *Plant Cell* 14, 111–130. doi: 10.1105/tpc.001362

Pan, G., Li, Z., Yin, M., Huang, S. Q., Tao, J., Chen, A. G., et al. (2021). Genome-wide identification, expression, and sequence analysis of CONSTANS-like gene family in cannabis reveals a potential role in plant flowering time regulation. *BMC Plant Biol.* 21:142. doi: 10.1186/s12870-021-02913-x

Perrella, G., Vellutini, E., Zioutopoulou, A., Patitaki, E., Headland, L. R., and Kaiserli, E. (2020). Let it bloom: cross-talk between light and flowering signaling in *Arabidopsis*. *Physiol. Plant.* 169, 301–311. doi: 10.1111/ppl.13073

Putterill, J., Robson, F., Lee, K., Simon, R., and Coupland, G. (1995). The CONSTANS gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc-finger transcription factors. *Cell* 80, 847–857.

Robson, F., Costa, M. M., Hepworth, S. R., Vizir, I., Pineiro, M., Reeves, P. H., et al. (2001). Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants. *Plant J.* 28, 619–631. doi: 10.1046/j.1365-313x.2001.01163.x

Romero-Campero, F. J., Lucas-Reina, E., Said, F. E., Romero, J. M., and Valverde, F. (2013). A contribution to the study of plant development evolution based on gene co-expression networks. *Front. Plant Sci.* 4:291.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425. doi: 10.1093/oxfordjournals.molbev.a040454

Samach, A., Onouchi, H., Gold, S. E., Ditta, G. S., Schwarz-Sommer, Z., Yanofsky, M. F., et al. (2000). Distinct roles of CONSTANS target genes in reproductive development of *Arabidopsis*. *Science* 288, 1613–1616.

Sawa, M., and Kay, S. A. (2011). GIGANTEA directly activates Flowering Locus T in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11698–11703. doi: 10.1073/pnas.1106771108

Song, J., Irwin, J., and Dean, C. (2013). Remembering the prolonged cold of winter. *Curr. Biol.* 23, 807–811. doi: 10.1016/j.cub.2013.07.027

Song, X. M., Duan, W. K., Huang, Z. N., Liu, G. F., Wu, P., Liu, T. K., et al. (2015). Comprehensive analysis of the flowering genes in Chinese cabbage and examination of evolutionary pattern of CO-like genes in plant kingdom. *Sci. Rep.* 5:14631. doi: 10.1038/srep14631

Sønsteby, A., Roos, U. M., and Heide, O. M. (2017). Phenology, flowering and yield performance of 13 diverse strawberry cultivars grown under Nordic field conditions. *Acta. Agric. Scand Sect. B Soil Plant Sci.* 67, 278–283.

Suárez-López, P., Wheatley, K., Robson, F., Onouchi, H., Valverde, F., and Coupland, G. (2001). CONSTANS mediates between the circadian clock and the control of flowering in *Arabidopsis*. *Nature* 410, 1116–1120. doi: 10.1038/35074138

Talar, U., Kiełbowicz-Matuk, A., Czarnecka, J., and Rorat, T. (2017). Genome-wide survey of B-box proteins in potato (*Solanum tuberosum*)-identification, characterization and expression patterns during diurnal cycle, etiolation and de-etiolation. *PLoS One* 12:e0177471. doi: 10.1371/journal.pone.0177471

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tan, J. J., Jin, M. N., Wang, J. C., Wu, F. Q., Sheng, P. K., Cheng, Z. J., et al. (2016). OsCOL10, a CONSTANS-Like gene, functions as a flowering time repressor downstream of Ghd7 in rice. *Plant Cell Physiol.* 57, 798–812. doi: 10.1093/pcp/pcw025

Tiwari, S. B., Shen, Y., Chang, H. C., Hou, Y. L., Harris, A., Ma, S. F., et al. (2010). The flowering time regulator CONSTANS is recruited to the FLOWERING LOCUS T promoter via a unique cis-element. *New Phytol.* 187, 57–66. doi: 10.1111/j.1469-8137.2010.03251.x

Toledo-Ortiz, G., Huq, E., and Quail, P. H. (2003). The *Arabidopsis* basic/helix-loop-helix transcription factor family. *Plant Cell* 15, 1749–1770. doi: 10.1105/tpc.013839

Valverde, F. (2011). CONSTANS and the evolutionary origin of photoperiodic timing of flowering. *J. Exp. Bot.* 62, 2453–2463. doi: 10.1093/jxb/erq449

Wang, H. G., Zhang, Z. L., Li, H. Y., Zhao, X. Y., Liu, X. M., Ortiz, M., et al. (2013). CONSTANS-LIKE 7 regulates branching and shade avoidance response in *Arabidopsis*. *J. Exp. Bot.* 64, 1017–1024. doi: 10.1093/jxb/ers376

Wang, L., Xue, J., Dai, W., Tang, Y., Gong, P., Wang, Y., et al. (2019). Genome-wide identification, phylogenetic analysis, and expression profiling of CONSTANS-like (COL) genes in *Vitis vinifera*. *J. Plant Growth Regul.* 38, 631–643. doi: 10.1007/s00344-018-9878-8

Wenkel, S., Turck, F., Singer, K., Gissot, L., Le Gourrierec, J., Samach, A., et al. (2006). Constans and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of *Arabidopsis*. *Plant Cell* 18, 2971–2984. doi: 10.1105/tpc.106.043299

Wong, A. C., Hecht, V. F., Picard, K., Diwadkar, P., Laurie, R. E., Wen, J., et al. (2014). Isolation and functional analysis of CONSTANS-LIKE genes suggests that a central role for CONSTANS in flowering time control is not

evolutionarily conserved in *Medicago truncatula*. *Front. Plant Sci.* 5:486. doi: 10.3389/fpls.2014.00486

Wu, W. X., Zheng, X. M., Chen, D. B., Zhang, Y. X., Ma, W. W., Zhang, H., et al. (2017). OsCOL16, encoding a CONSTANS-like protein, represses flowering by up-regulating Ghd7 expression in rice. *Plant Sci.* 260, 60–69. doi: 10.1016/j.plantsci.2017.04.004

Yang, N., Cong, Q., and Cheng, L. J. (2019). BBX transcriptional factors family in plants a review. *Chin. J. Biotech.* 36, 666–677. doi: 10.13345/j.cjb.190302

Yang, T., He, Y., Niu, S., Yan, S., and Zhang, Y. (2020). Identification and characterization of the CONSTANS (CO)/CONSTANS-like (COL) genes related to photoperiodic signaling and flowering in tomato. *Plant Sci.* 301:110653. doi: 10.1016/j.plantsci.2020.110653

Yano, M., Katayose, Y., Ashikari, M., Yamanouchi, U., Monna, L., Fuse, T., et al. (2000). Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering time gene CONSTANS. *Plant Cell* 12, 2473–2484. doi: 10.1105/tpc.12.12.2473

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.

Zhu, Y., Klasfeld, S., Jeong, C. W., Jin, R., Goto, K., Yamaguchi, N., et al. (2020). TERMINAL FLOWER 1-FD complex target genes and competition with FLOWERING LOCUS T. *Nat. Commun.* 11:5118. doi: 10.1038/s41467-020-18782-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 22 August 2022
DOI 10.3389/fpls.2022.995634

OPEN ACCESS

EDITED BY
Weihua Pan,
Agricultural Genomics Institute at Shenzhen (CAAS), China

REVIEWED BY
Jingping Fang,
Fujian Normal University, China
Luomiao Yang,
Northeast Agricultural University, China
Showkat Ganie,
Royal Holloway, University of London, United Kingdom

*CORRESPONDENCE
Zhaohai Wang
xzhaohai_wang@163.com
Yangsheng Li
lysh2001@whu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 16 July 2022
ACCEPTED 01 August 2022
PUBLISHED 22 August 2022

CITATION
Kong W, Deng X, Liao Z, Wang Y, Zhou M, Wang Z and Li Y (2022) *De novo* assembly of two chromosome-level rice genomes and bin-based QTL mapping reveal genetic diversity of grain weight trait in rice.
Front. Plant Sci. 13:995634.
doi: 10.3389/fpls.2022.995634

De novo assembly of two chromosome-level rice genomes and bin-based QTL mapping reveal genetic diversity of grain weight trait in rice

Weilong Kong[1†], Xiaoxiao Deng[1†], Zhenyang Liao[2], Yibin Wang[2], Mingao Zhou[1], Zhaohai Wang[3*] and Yangsheng Li[1*]

[1]State Key Laboratory of Hybrid Rice, College of Life Sciences, Wuhan University, Wuhan, China, [2]Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, China, [3]Key Laboratory of Crop Physiology, Ecology and Genetic Breeding (Jiangxi Agricultural University), Ministry of Education of the People's Republic of China, Nanchang, China

Following the "green revolution," *indica* and *japonica* hybrid breeding has been recognized as a new breakthrough in further improving rice yields. However, heterosis-related grain weight QTLs and the basis of yield advantage among subspecies has not been well elucidated. We herein *de novo* assembled the chromosome level genomes of an *indica/xian* rice (Luohui 9) and a *japonica/geng* rice (RPY geng) and found that gene number differences and structural variations between these two genomes contribute to the differences in agronomic traits and also provide two different favorable allele pools to produce better derived recombinant inbred lines (RILs). In addition, we generated a high-generation (> F_{15}) population of 272 RILs from the cross between Luohui 9 and RPY geng and two testcross hybrid populations derived from the crosses of RILs and two cytoplasmic male sterile lines (YTA, *indica* and Z7A, *japonica*). Based on three derived populations, we totally identified eight 1,000-grain weight (KGW) QTLs and eight KGW heterosis loci. Of QTLs, *qKGW-6.1* and *qKGW-8.1* were accepted as novel KGW QTLs that have not been reported previously. Interestingly, allele genotyping results revealed that heading date related gene (*Ghd8*) in *qKGW-8.1* and *qLH-KGW-8.1*, can affect grain weight in RILs and rice core accessions and may also play an important role in grain weight heterosis. Our results provided two high-quality genomes and novel gene editing targets for grain weight for future rice yield improvement project.

KEYWORDS

rice, genome sequencing, yield improvement, 1,000-grain weight, heterosis loci

Introduction

Rice is one of the most important food crops in the world, providing food for more than half of the world's population (Qin et al., 2021). Over the past two decades, multiple high-quality genomes of *indica* and *japonica* subspecies have been assembled, such as Nipponbare (Goff et al., 2002), 9311 (Yu et al., 2002), ZS97 (Zhang et al., 2016a,b), MH63 (Zhang et al., 2016a,b), R498 (Du et al., 2017), IR64 (Tanaka et al., 2020), TN1 (Panibe et al., 2021), Huazhan (Zhang H. et al., 2022), Tianfeng (Zhang H. et al., 2022), etc. Recently, several gap-free reference genomes were completed, namely, ZS97, MH63, PR106, LIMA, LARHAMUGAD, KETANNANGKA, NATELBORO, XL628S, LK638S, J4155S, and HZ (Song et al., 2021; Zhang F. et al., 2022; Zhang Y. et al., 2022). Advances in third-generation sequencing and assembly algorithms have continuously updated the accuracy of the rice pan-genome, revealing some important functionally related structural variations (SVs) and gene copy number variations (gCNVs) (Zhao et al., 2018; Qin et al., 2021; Zhang F. et al., 2022). However, there are dramatically different genetic backgrounds among thousands of rice cultivars, especially between subspecies, including cultivar-specific genes, different alleles of one gene, or gene family expansions (Li et al., 2021). Differences in rice agronomic traits are closely related to these genome variations (Stein et al., 2018; Zhao et al., 2018; Qin et al., 2021). Therefore, the discovery of new genes/alleles related to agronomic traits is inseparable from the comparative analysis of the genomes of elite varieties and the fine mapping in derived populations. For instance, we previously found that the hybrid progeny of the Luohui 9 (*xian*/*indica*) and RPY geng (*geng*/*japonica*) cross had significant heterosis in yield and resistance traits, and multiple recombinant inbred lines (RILs) derived from Luohui 9 X RPY geng aggregated the advantages of both parents (Kong et al., 2022a). Based on the high-density genetic map, we also obtained some QTLs related to plant height, salt stress tolerance, submerged germination, and grain shape (Kong et al., 2021b, 2022a,b; Deng et al., 2022). But the gene number differences and large structural variations between Luohui 9 and RPY geng, and the effects of these variations/differences on traits and heterosis, remain unclear.

Owing to the impacts of human population growth and limited arable land, breeders and scientists faced the challenge of breeding higher yield potential crops. Rice yield is a complex agronomic trait composed of four main factors including effective panicle number, grain number per panicle, seed setting rate and 1,000-grain weight (KGW) (Zuo and Li, 2014). In addition, heterosis refers to the phenomenon that the phenotype of the hybrid progeny surpasses their parents in biomass, yield, growth vigor, resistance, etc., (Birchler et al., 2010). Yield heterosis between *indica* and *japonica* subspecies has been widely used to improve rice yield, causing a worldwide yield revolution (Li et al., 2018). As statistics, hybrid rice shows a 20–30% increase in yield than inbred rice and has effectively solved world food crisis (Xu et al., 2016). Therefore, analyzing the mechanism of rice grain weight (GW) and mining GW-related QTLs and GW-related heterosis loci are important foundations for improving rice yield. Based on different populations and QTL mapping methods, more than 600 QTLs related to grain weight and grain shape have been identified on all 12 chromosomes in rice to date[1] (Chan et al., 2021), and more than 20 QTLs have been cloned, including *GW2* (Yan S. et al., 2011), *GS3* (Liu et al., 2018), *TGW6* (Ishimaru et al., 2013), *GW6a* (Song et al., 2015), *WTG1* (Huang et al., 2017), *GL7* (Wang et al., 2015), and *gw5* (Wan et al., 2005). In fact, multiple reported QTLs may belong to one QTL, but there are differences in the size of the interval. Meta-analysis of QTLs was used to merge multiple QTLs from different rice genetic populations and to identify consensus and stable QTLs (Arcade et al., 2004; Kong et al., 2020), which narrowed down the confidence intervals of QTLs (Martinez et al., 2016; Zhang et al., 2017). Recently, 339 published GW QTLs merged into 34 Meta-QTLs (MQTLs) in rice (Khahani et al., 2020). The new GW QTLs/genes must be urgently explored in *indica* X *japonica* derived populations to further improve rice yields.

In this study, we performed the chromosome-level *de novo* assembly of the Luohui 9 and RPY geng genomes and characterized their genomic differences in genome-wide scale. Additionally, the KGW traits of derived RIL populations from Luohui 9 X RPY geng in four environments were used for QTL mapping. Two testcross populations derived from the crosses of RILs and Z7A (*japonica*) or YTA (*indica*) were used to explore the heterosis loci of KGW. These results provided a new insight into the diversity mechanism of grain weight in rice.

Materials and methods

Materials and sequencing

The highly homozygous *O. sativa* ssp. *indica*/*xian* (Luohui 9, 2n = 2 × = 24) and *O. sativa* ssp. *japonica*/*geng* (RPY geng, 2n = 2 × = 24) were planted in the field in Wuhan, China, in 2016, These two subspecies have many significant differences in important agronomic traits, including plant height, number of tillers, and heading date, for example, Luohui 9 has excellent agronomic traits, and RPY geng has an ideal plant architecture (**Figure 1A**).

The genomic DNA of these two subspecies was extracted from young leaves using a modified CTAB method and tested using Qubit Quantitation Starter Kit (Invitrogen, United States) and a 1% agarose gel electrophoresis, respectively. Libraries for Illumina short-read and single-molecule real-time (SMRT) sequencing (Pacific Biosciences, United States) were prepared

1 http://www.gramene.org/

10.3389/fpls.2022.995634

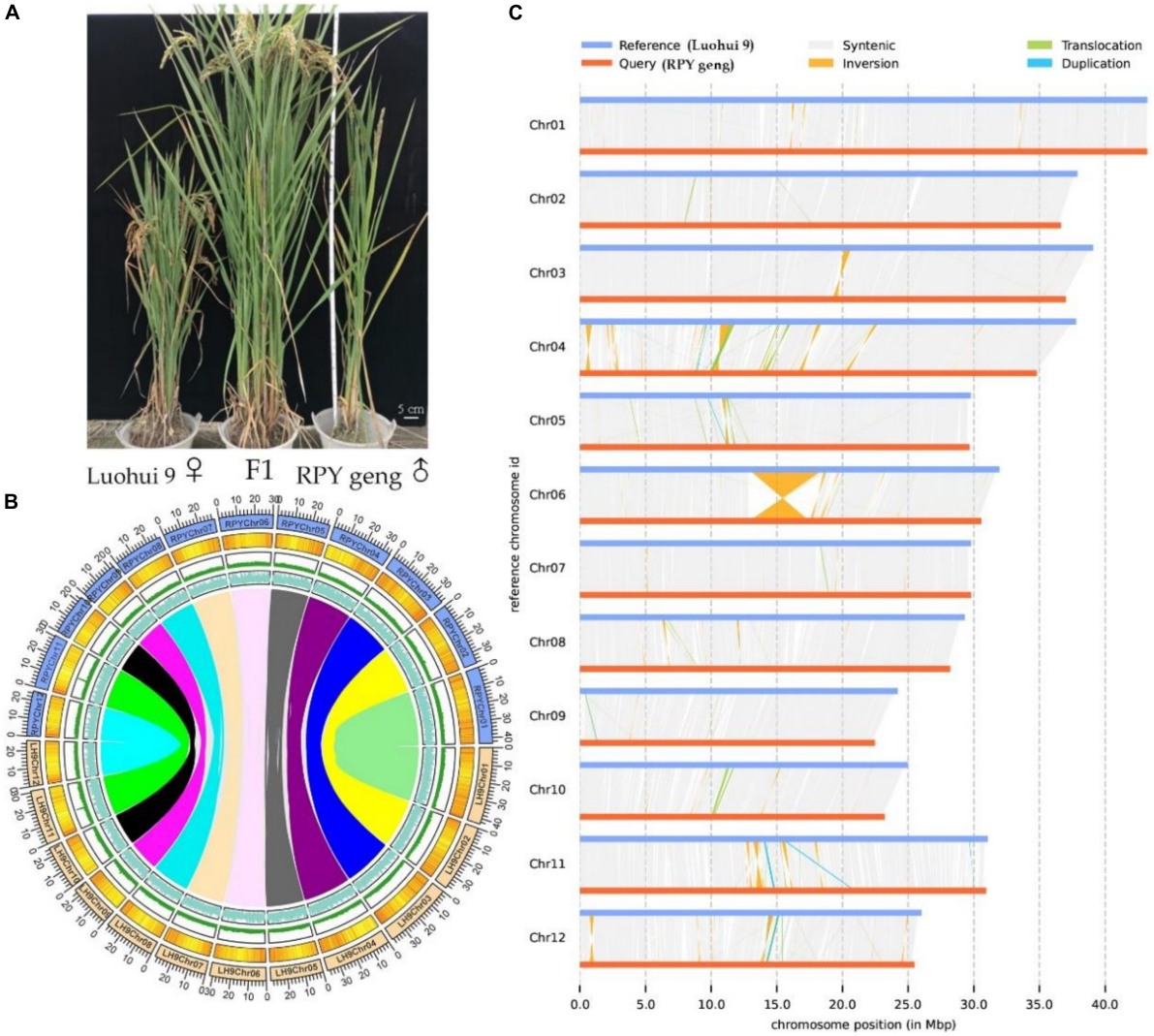


FIGURE 1
Whole plant phenotype **(A)**, genomic characteristics **(B)**, and large structural variations **(C)** between RPY geng and Luohui 9. The tracks from outside to inside are the chromosome, gene density, repeat sequence density, and GC content and the different color links represent orthologous gene pairs among chromosomes in **(B)**.

according to the respective manufacturer's instructions. The short-read DNA libraries were sequenced by paired-end (2 × 150 bp) method on Illumina HiSeqTM 2,500 (Illumina, United States) and the SMRT sequencing were sequenced on the PacBio RS II platform. In addition, Hi-C reads of DNA of young leaves from F1 progeny of Luohui 9 and RPY geng were sequenced on Illumina HiSeqTM 2,500 paired-end (2 × 150 bp) sequencing according to standard protocol. The total RNA of mixed tissues (root, stem, leaf, and young panicle) was extracted for RNA-seq libraries following the manufacturer's standard protocol. Then, RNA-seq libraries were sequenced on an Illumina HiSeqTM 2,500 paired-end (2 × 150 bp) and raw reads were filtered using Trimmomatic software as described previously (Kong et al., 2020).

Genome assembly and annotation

PacBio RSII sub-reads were filtered by the PacBio SMRT-Analysis package including removing the adapters, low quality, and short length reads (parameters: readScore, 0.75; minSubReadLength, 500) and sub-reads after filtering were corrected by Illumina reads using an error correction module embedded in Canu v1.5 (Koren et al., 2017). The high-quality PacBio sub-reads were used for genome contigs assembly by using Canu v1.5 with default parameters. The contig-level genome was polished by Plion with these parameters: –mindepth 10; –changes; –threads 4; –fix bases (Walker et al., 2014). Hi-C data were used to assist in constructing chromosome-level genome assemblies. The Hi-C data were

mapped to the contig-level genomes using BWA aligner software (Li and Durbin, 2009). A total of 46.73 Gb clean Hi-C data was mapped to the genome, with a coverage of 116.83 X. These uniquely mapped Hi-C reads were retained for chromosome-level genome assemblies using LACHESIS software with these parameters: CLUSTER MIN RE SITES = 22; CLUSTER MAX LINK DENSITY = 2; CLUSTER NON-INFORMATIVE RATIO = 2; ORDER MIN N RES IN TRUNK = 10; ORDER MIN N RES IN SHREDS = 10 (Burton et al., 2013). Finally, the chromosome-level genome was further improved and corrected by the high-density binmaps genetic map constructed in this study using all maps software according to the method described (Tang H. B. et al., 2015).

Repeat sequence annotation was performed by EDTA with default parameters (Ou et al., 2019). Coding genes were predicted by *de novo*, homolog-based, and transcriptome-based strategies. Augustus v2.4 (Stanke and Waack, 2003), Genscan (Burge and Karlin, 1997), GeneID v1.4 (Alioto et al., 2018), GlimmerHMM v3.0.4 (Majoros et al., 2004), and SNAP (version 2006-07-28) (Korf, 2004) were used for *de novo* prediction. GeMoMa v1.3.1 (Keilwagen et al., 2016) was used for homolog-based prediction. In the transcriptome-based prediction, we used Hisat v2.0.4 and Stringtie v1.2.3 for sequence assembly based on a reference genome (Pertea et al., 2016), and applied TransDecoder v2.0[2] and GeneMarkS-T v5.1 (Tang S. Y. Y. et al., 2015) for gene prediction; On the other hand, PASA v2.0.2 software (Campbell et al., 2006) was used to perform unigene sequence prediction without reference assembly based on transcriptome sequencing data. Finally, we used the EVM v1.1.1 software (Haas et al., 2008) to integrate all gene prediction results from these three analysis methods. The predicted coding genes were annotated according to alignments against (*E* value 1e–5) databases including GO, KEGG, KOG, TrEMBL, and Nr databases using BLAST v2.2.31 (Altschul, 2012).

Orthologous clusters analysis and structural variant identification

We extracted the longest-protein sequences from Luohui 9 and RPY geng genomes for orthologous clusters identification using OrthoVenn2, E-value was set 1e–10, and other parameters with default (Xu et al., 2019). Genomic structural variants between Luohui 9 and RPY geng genomes were identified by SyRI (Goel et al., 2019). Luohui 9 genome was used as the reference genome, "nucmer – mum" for sequence alignment, with parameter, -c 100 -l 50 -g 1,000. Then, "delta-filter" was used to filter the comparison results, with parameter, -1 -q -r -i 89 -l 50. Finally, "show-coords," "syri -c," and "syri plotsr" steps were done with default parameters.

2 http://transdecoder.github.io

TABLE 1 Summary of genome assembly and annotation of RPY geng and Luohui 9.

	RPY geng	Luohui 9
Genome assembly		
Assembly size (Mb)	383.45	394.43
Number of contigs	684	569
N50 size of contigs (Mb)	2.81	2.84
Anchored contigs (Mb)	369.49	382.54
Anchored contigs (%)	96.35	96.99
Complete assessment of 456 core genes in CEGMA v2.5 (%)	99.56	100
Complete assessment of 238 core genes in CEGMA v2.5 (%)	95.97	97.18
Complete assessment of 1,614 core genes in BUSCO v3.0.2 (%)	98.9	99.1
The long terminal repeat (LTR)-assembly index (LAI)	19.44	19.37
Genome annotation		
Percentage of repeat sequences (%)	44.29	46.96
Number of predicted genes	39,255	39,440

QTL mapping of KGW

A 272 RILs and their parents were planted in the Ezhou (30°N, 114°E) Experimental Base of Wuhan University, Wuhan City, Hubei Province in April 2017 (2017EZ) and in April 2018 (2018EZ), the Hybrid Rice Experimental Base of Wuhan University in Lingshui City (18°N, 110°E), Hainan Province in November 2019 (2019LS), and Breeding Experimental Base of Wuhan University Tianyuan Co., Ltd in Hannan District (30°N, 114°E), Wuhan City, Hubei Province in May 2019 (2019HN). Two testcross populations (Z7A-TCF$_1$ and YTA-TCF$_1$) were developed by crossing RILs (F$_{14}$) with Z7A (*japonica*) and YTA (*indica*) in 2019LS. KGWs of Z7A-TCF$_1$, YTA-TCF$_1$, and RILs were investigated in 2019HN.

All plants were planted under standard agricultural management practice (Kong et al., 2022a). KGW was surveyed in the above four environments. Each inbred line counted five individual plants, the average KGW value of five individual plants was considered as the KGW value of each inbred line.

The genetic linkage map of 272 RILs including 4,578 bin blocks with the total bin-map distance 2,356.41 cM was previously constructed in our lab (Kong et al., 2022a). The QTL mapping of KGW was analyzed by R/qtl (Arends et al., 2010), the CIM interval mapping method was adopted and the LOD threshold was set by 3.0. The confidence interval was calculated with the function "lodint" (Dupuis and Siegmund, 1999) and the drop value was set to 1.5.

QTL mapping of KGW heterosis loci

Heterosis related indexes for the KGW trait were calculated by the formulas:

$$\mathrm{MPH} = [F1 - (P1 + P2)/2\,]/[\,(P1 + P2)/2]\;X\;100\%$$

$$\mathrm{BPH} = (F1 - P1)/P1\;X\;100\%$$

$$\mathrm{LPH} = (F1 - P2)/P2\;\mathrm{X}\;100\%$$

where MPH is middle-parent heterosis, BPH is better-parent heterosis, LPH is lower-parent heterosis, P1 is the high parent, and P2 is the low parent.

The QTL mapping of KGW heterosis related indexes was analyzed by R/qtl (Arends et al., 2010), the CIM interval mapping method was adopted and the LOD threshold was set by 2.5. The confidence interval was calculated with the function "lodint" (Dupuis and Siegmund, 1999) and the drop value was set to 1.5.

Results and discussion

De novo assembly and annotation of RPY geng and Luohui 9 genomes

A total of 25.1 / 36.5 Gb Illumina short reads with 62 / 91 X coverage of the genome and 15.6 / 19.5 Gb PacBio RSII long reads with 39 / 48 X coverage of the genome of RPY geng / Luohui 9 was obtained (**Supplementary Table 1**). The long reads were polished by the Illumina paired read and the polished long reads were assembled into contigs by Canu V1.5. After three rounds of contig polish by Pilon v1.22 and Hi-C data correction, we obtained a 383.45 Mb RPY geng genome and a 394.43 Luohui 9 genome, with the contig N50 of 2.81 and 2.84 Mb, respectively (**Table 1**). Luohui 9 was ∼10.98 Mb larger than the genome of RPY geng. Finally, 96.35 and 96.99 % of contigs were anchored onto 12 pseudo-chromosomes of RPY geng and Luohui 9 based on Hi-C interactions and linkage map from RPY geng x Luohui 9 derived population (Kong et al., 2022a), respectively (**Table 1** and **Supplementary Figure 1**).

The percentages of repeat sequences in the genomes of RPY geng and Luohui 9 were 44.29 and 46.96% based on EDTA with default parameters (**Supplementary Table 2**). A combination of prediction strategies (*de novo*, homologous based and RNA-seq based) totally identified 39,255 and 39,440 gene models among RPY geng and Luohui 9 genomes, of which 96.81 and 94.75% had at least one annotation result in GO, KEGG, KOG, TrEMBL, or Nr database (**Supplementary Tables 3**, **4**). The results of CEGMA and BUSCO showed that the assembly of RPY geng and Luohui 9 was complete, with more than 95.0% of the core genes. The long terminal repeat (LTR)-assembly index (LAI) of RPY geng and Luohui 9 was 19.44 and 19.37, which is close to the gold genome level (LAI $\geq$ 20) (**Table 1**). All the above-mentioned genome indices indicated that the newly assembled genomes of RPY geng and Luohui 9 was of high quality.

Global genome differences between RPY geng and Luohui 9

RPY geng and Luohui 9 showed obvious differences in yield, grain shape (Deng et al., 2022), plant height (Kong et al., 2022a), and abiotic stress resistances (Kong et al., 2020, 2021a,b, 2022b), and the hybrid progeny of RPY geng X Luohui 9 had the excellent heterosis (**Figure 1A**). These essential agronomic differences are inseparable from the number and structural variation of genes between the two subspecies genomes (Zhao et al., 2018; Qin et al., 2021). Benefiting from the completion of the genomes of RPY geng and Luohui 9, we compared their gene numbers and large structural variations at the genome-wide level and highlighted some genes that have potential impact on agronomic traits. A total of 32,720 orthologous clusters including 32,509 orthologous gene pairs were identified (**Supplementary Figure 2** and **Figure 1B**). Luohui 9 unique orthologous clusters were enriched with multiple essential life GO terms, while RPY geng unique orthologous clusters were enriched with multiple stress-related GO terms (**Supplementary Table 5**), namely, defense response, cellular response to amino acid stimulus, positive regulation of hydrogen peroxide, as well as response to osmotic stress, suggesting that RPY geng has more tolerance-related genes to abiotic stress than Luohui 9. These results are consistent with our previous findings that RPY geng has stronger resistance to salt stress and cold stress and carries important stress tolerance genes (Kong et al., 2020, 2021a,b, 2022b).

We further found 190 inversions, 6,852 translocations, 1,279 (Luohui 9)/1,212 (RPY geng) duplications between RPY geng and Luohui 9 involving 2,234 SV-related genes in Luohui 9 and 1,544 SV-related genes in RPY geng (**Supplementary Table 6**). Notably, at the position of 12.8–18.6 Mb on Luohui 9 chromosome 6 showed a sequence inversion with a length of about 5.7 Mb compared with the RPY geng genome (**Figure 1C**) and this inversion has also been reported in previous comparative genomic studies between subspecies (Du et al., 2017; Li et al., 2021; Xie et al., 2021), suggesting that this may be an important structural difference between subspecies. To study the potential roles of these SV-related genes in important agronomic traits, we collected 283 important known genes with functional function verifications (**Supplementary Table 7**) as query sequences to find homologous genes against SV-related genes by BlastP (value E-10). Totally, 337 SV-related genes were identified as homologous genes of 138 known functional genes

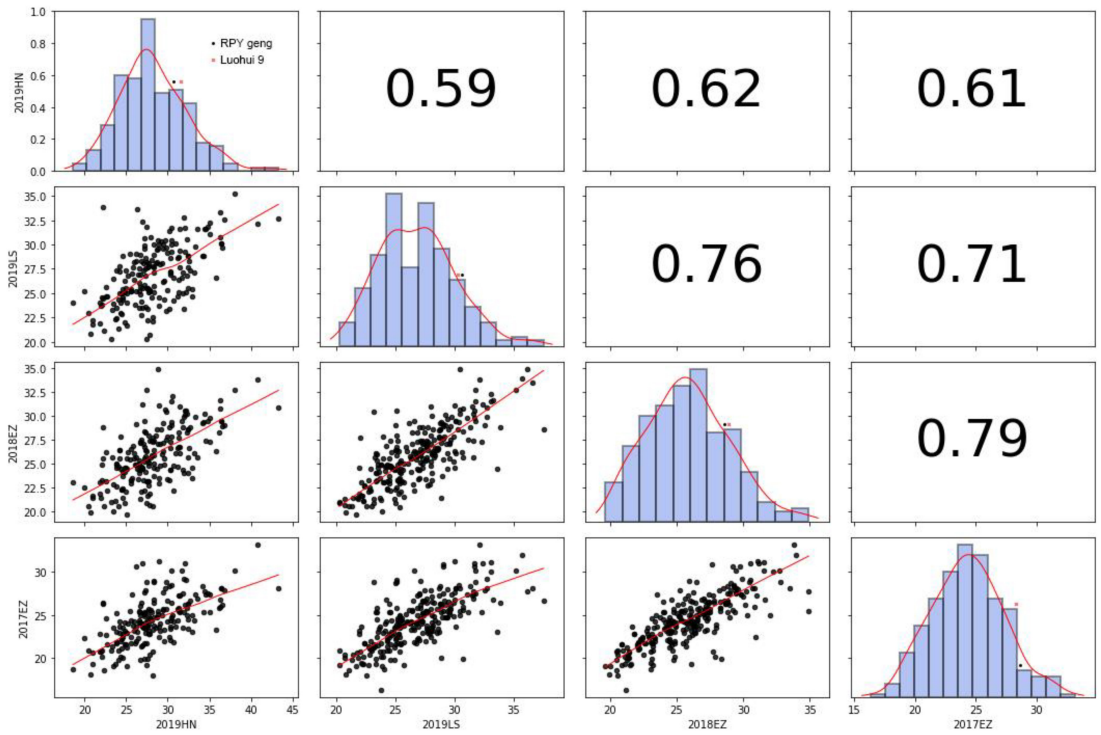


FIGURE 2
Thousand-grain weight in 2019HN, 2019LS, 2018EZ and 2017EZ.

TABLE 2 Details of thousand-grain weight QTLs.

QTL	Environment	Chr	QTL position	LOD	Size of QTL (Mb)	Phenotypic variation (%)
qKGW-2.1	2019LS	Chr 2	22532020–29979073	3.77	7.45	6.18
qKGW-3.1	2019HN	Chr 3	8166100–10764126	7.74	2.60	12.28
qKGW-3.2	2018EZ; 2019LS	Chr 3	8166100–11054115	6.49; 8.45	2.89	10.41; 13.33
qKGW-5.1	2019LS	Chr 5	2493875–5085048	3.1	2.59	5.11
qKGW-5.2	2018EZ	Chr 5	3542240–4415859	4.33	0.87	7.07
qKGW-6.1	2019_HN	Chr 6	2215678–2577924	4.83	0.36	7.85
qKGW-8.1	2017EZ; 2018EZ; 2019LS	Chr 8	4158052–4833970	3.42; 6.2; 4.41	0.68	5.63; 9.96; 6.20
qKGW-10.1	2017EZ	Chr 10	13311698–19472799	3.78	6.16	6.2

belonging to cold tolerance, heat tolerance, salt tolerance, insect resistance, disease resistance, drought tolerance, fertility, grain quality, grain shape, heading date, panicle architecture, nutrient utilization, and panicle architecture (**Supplementary Table 8**). The above results suggested that the genome-wide number and structural differences play essential roles in the trait differences of RPY geng and Luohui 9, which were consistent with their differential trait characteristics.

QTLs of KGW

RPY geng and Luohui 9 belonged to *japonica*/*geng* and *indica*/*xian* subspecies, respectively, yield traits of their F1 and many RILs showed obvious over-parent dominance. To resolve yield-related genes, we here conducted trait surveys and linkage analysis of KGW based on the previously constructed high-density genetic map (Kong et al., 2022a).

KGW of RILs were investigated in Lingshui, Hannan, or Ezhou among 2017 – 2019. There was extensive variance of KGW in RIL population while there was a minor difference of KGW between the parents (**Figure 2** and **Supplementary Table 9**). KGW of the RIL population showed a normal distribution with high Pearson coefficients in four different environments and transgressive segregations were observed in the RIL population (**Figure 2**), which indicated that KGW were controlled by multiple genes and

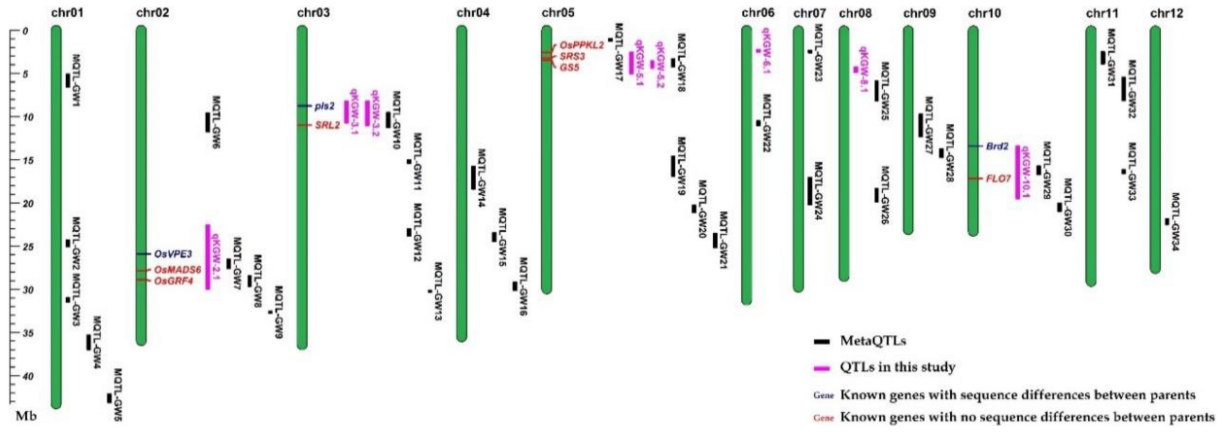


FIGURE 3
The positions of QTLs, MetaQTLs, and known KGW-related genes.

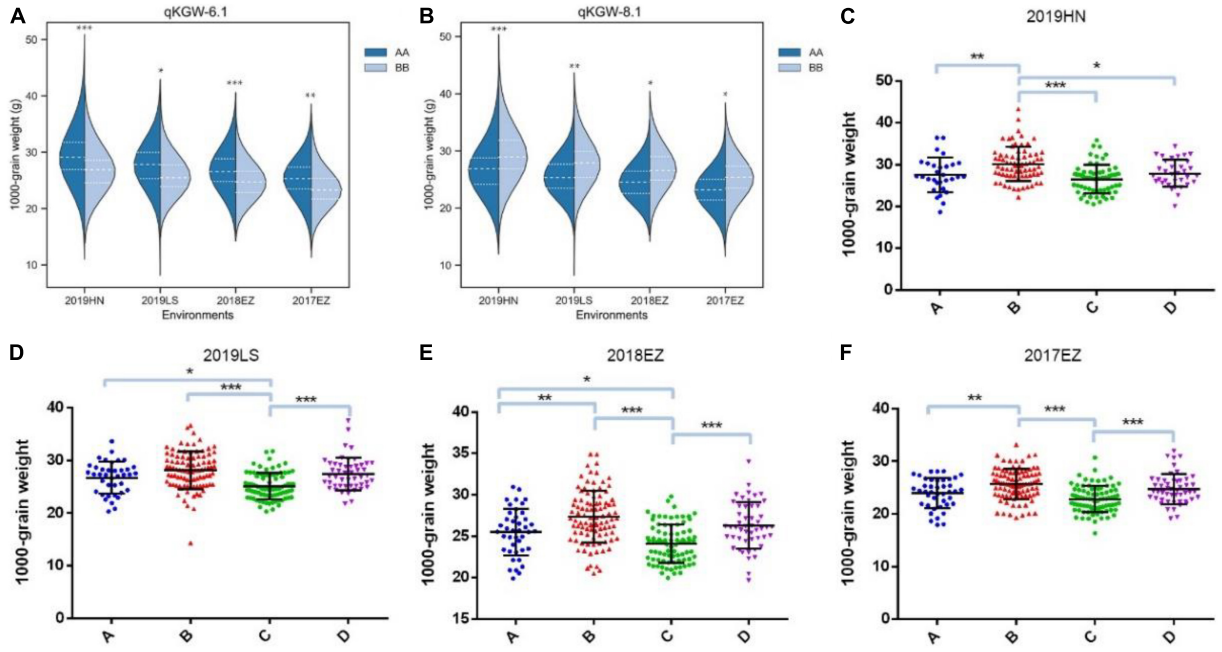


FIGURE 4
1000-grain weight of different allele combinations of *qKGW-6.1* and *qKGW-8.1*. In **(A)**, **(C–F)**: *qKGW-6.1* AA + *qKGW-8.1* AA, **(B)**: *qKGW-6.1* AA + *qKGW-8.1* BB; **(C)**: *qKGW-6.1* BB + *qKGW-8.1* AA; **(D)**: *qKGW-6.1* BB + *qKGW-8.1* AA.

indica x *japonica* hybrid breeding strategy can breed high-yielding rice materials.

We totally identified eight KGW QTLs on Chr 2, Chr 3, Chr5, Chr6, Chr8, and Chr10 (Table 2). Of QTLs, *qKGW-8.1* was repeatedly detected in 2017EZ, 2018EZ, and 2019LS. *qKGW-3.1* and *qKGW-3.2* had almost the same interval and *qKGW-5.2* was fully contained by *qKGW-5.1*. These results suggested that these three QTLs had relatively stable effects on KGW in multiple different environments. The remaining QTLs (*qKGW-2.1*, *qKGW-6.1*, and *qKGW-10.1*) were only detected in one specific ecological environment, and were possibly environment-specific KGW QTLs.

Identification and function confirmation of two novel KGW QTLs

To distinguish the new TWG QTLs first discovered in this study, 34 Meta-QTLs from 339 original GW QTLs (Supplementary Table 10) and 126 known GW genes

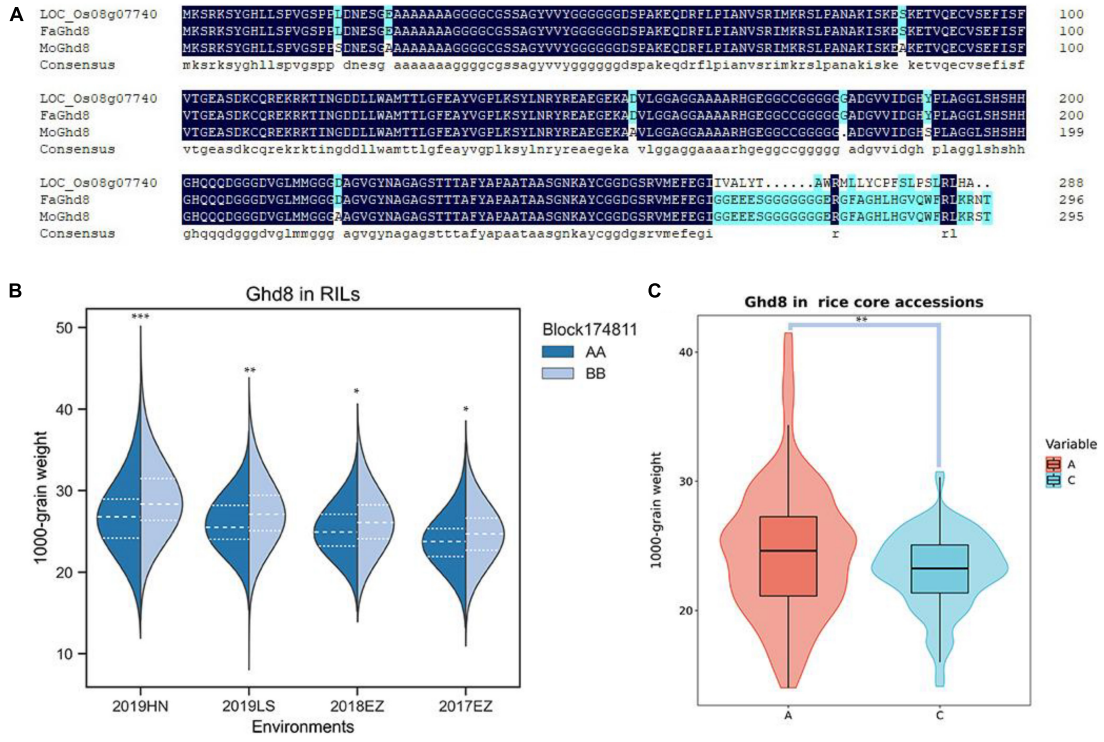


FIGURE 5
The candidate gene prediction of *qKGW-8.1*. **(A)**. Protein sequence alignment results of Nipponbare, RPYgeng (*FaGhd8*), Luohui 9 (*MoGhd8*). **(B)** The 1,000-grain weight of RPY geng (AA) and Luohui 9 (BB) allele recombinant inbred lines (RILs). **(C)** The 1,000-grain weight of different allele rice core accessions.

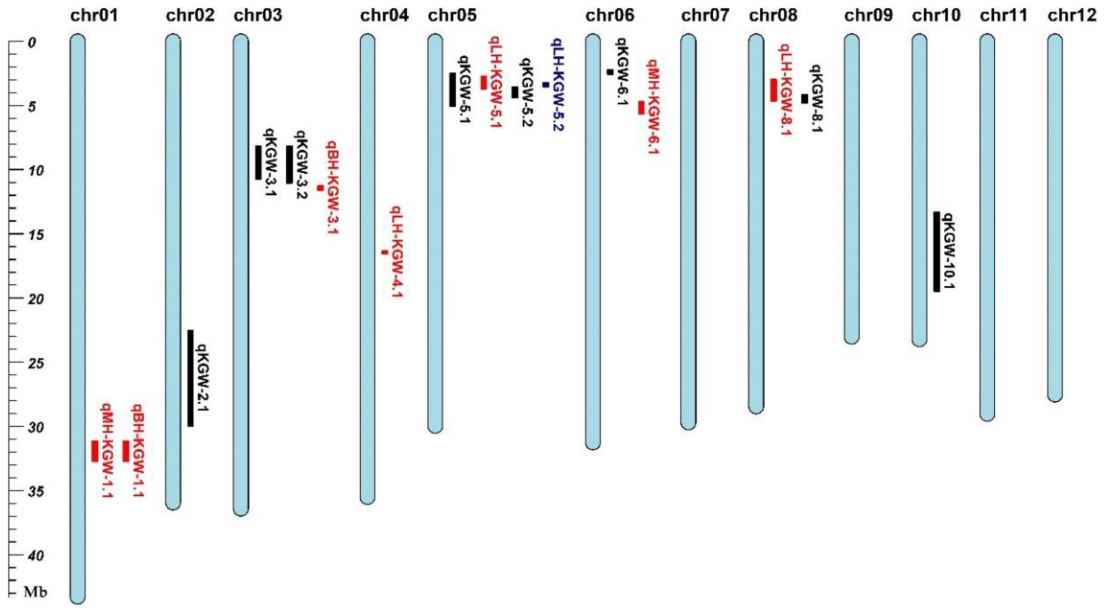


FIGURE 6
Grain weight heterosis loci in two testcross populations. Black represents QTL loci in recombinant inbred lines (RILs). Red and blue loci represent heterosis-related QTLs of YTA and Z7A testcross population, respectively.

(**Supplementary Table 11**) were collected (Khahani et al., 2020). Except for *qKGW-6.1* and *qKGW-8.1*, the remaining QTLs all had complete or partial overlap with Meta-QTLs, or contained the known KGW genes in these intervals (**Figure 3**). For example, *qKGW-3.1* and *qKGW-3.2* fully covered *MQTL-GW18* and included *pls2* and *SRL2*. Sequence alignment result showed that *SRL2* had sequence differences in the coding sequence regions (**Supplementary Table 12**). Similarly, *qKGW-2.1* containing *OsVPE3*, *OsMADS6*, and *OsGRF4*, overlapped *MQTL-GW7* and *MQTL-GW8*; *qKGW-5.1* and *qKGW-5.2* carrying *OsPPKL2*, *SRS3*, and *GS5* overlapped *MQTL-GW18*; *qKGW-10.1* including *FLO7* overlapped *MQTL-GW29*.

Therefore, *qKGW-6.1* and *qKGW-8.1* were accepted as novel KGW QTLs. To further confirm the KGW regulation function of *qKGW-6.1* and *qKGW-8.1*, all RILs were divided into different allelic combination based on peak maker genotyping results in genetic map. RPY geng allele (AA) RILs of *qKGW-6.1* showed greater KGW than Luohui 9 (BB) RILs (**Figure 4A**). Interestingly, *qKGW-8.1* showed the opposite result relative to *qKGW-6.1* (**Figure 4B**). This suggested that the favorable alleles of these two QTLs are derived from RPY geng and Luohui 9, respectively, and the favorable allele aggregation may enhance KGW of some RILs. As expected, RILs that aggregated favorable alleles (*qKGW-6.1* AA + *qKGW-8.1* BB) had the largest KGW in all tested environments (**Figures 4C–F**). These results demonstrated that *qKGW-6.1* and *qKGW-8.1* are two new KGW loci that can be used to improve rice yield.

Ghd8 has a potential function in regulating rice grain weight

qKGW-8.1 could be detected repeatedly in three environments with phenotypic interpretation rates of 5.11–9.96. However, no known genes directly related to grain weight were found in *qKGW-8.1*. We therefore traversed the 92 gene annotation results in *qKGW-8.1* and tried to correspond them to the phenotypic differences between the parents. We found that *Ghd8* is located within the *qKGW-8.1* interval, a gene reported to be closely associated with heading date and yield (Yan W. H. et al., 2011; Dai et al., 2012), which is consistent with parental heading date differences. Whether in Hainan or Hubei, both parents maintained the heading date difference of more than 10 days. We extracted the Ghd8 protein sequences from our newly assembled genome and sequence alignment revealed multiple sequence variations, including seven amino acid substitutions, one amino acid deletion, and a complex C-terminal amino acid variation (**Figure 5A**).

To test whether *Ghd8* has a potential effect on grain weight as a pleiotropic gene, we observed grain weight at different alleles in our RILs and in 532 rice core accessions from RiceVarMap v2.0[3] (Zhao et al., 2021). In RILs, the allele types of *Ghd8* was determined based on a bin maker (Block174811) because *Ghd8* is the only gene within block174811. In 532 rice core accessions, a functional snp (vg0804334484) was found in *Ghd8* gene, containing A, C, and N alleles, and the N allele was eliminated in further phenotypic comparisons due to uncertainty about its base type. We found that different allele RILs or core accessions displayed significantly different KGW, suggesting that *Ghd8* may play a role in KGW regulation (**Figures 5B,C**).

KGW heterosis loci

We totally identified two QTLs for KGW BPH, two QTLs for KGW MPH, and four QTLs for KGW LPH (**Figure 6** and **Supplementary Table 13**). Three of the eight heterosis-related QTLs overlapped KGW QTLs: *qLH-KGW-5.1* and *qLH-KGW-5.2* were covered by *qKGW-5.1* and *qLH-KGW-8.1* overlapped with *qKGW-8.1*. Interestingly, *qLH-KGW-8.1* coincided with a reported yield heterosis locus, *RH8* (*rice heterosis 8*) (Li et al., 2016). *Ghd8*, as a major gene in *RH8*, was also located in the *qLH-KGW-8.1* interval. This suggested that *Ghd8* plays an important role in rice yield heterosis. In addition, two GW-related genes were in heterosis-related QTLs, namely, *SRS3* and *GS5* in *qLH-KGW-5.1*, *GS5* in *qLH-KGW-5.2*. Whether these GW-related genes play a role in heterosis remains to be further explored.

Conclusion

In the present study, we *de novo* assembled genomes of an *indica* rice (Luohui 9) and a *japonica* rice (RPY geng) at the chromosome level and analyzed the KGW trait of their derived RIL populations. We concluded that the substantial genetic diversity of KGW in RILs were closely related to genome variations and allele aggregation difference of KGW QTLs. Importantly, we identified two novel KGW-related QTLs (*qKGW-6.1* and *qKGW-8.1*) and several KGW heterosis loci in three derived population. Based on the genotyping results in RILs and 532 rice core accessions, *Ghd8* in *qKGW-8.1* was presumed to play an important role in GW regulation.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: ngdc, PRJCA010706.

3 http://ricevarmap.ncpgr.cn/

Author contributions

YL and WK conceived and designed the experiments. WK performed genome assembly, analyzed the data, prepared the figures and tables, and wrote the manuscript. XD conducted a field survey of agronomic traits, QTL mapping, and allele genotyping in RILs and rice core accessions. ZW participated in the construction of the recombinant inbred lines and revision of the manuscript. YW provided help in genome annotation. ZL performed repetitive sequence annotation of RPY geng and Luohui 9 genomes. MZ collected Meta-QTLs and known genes performed parts of figures and tables. All authors read and approved the final version of the manuscript.

Funding

This work was supported by the National Key Research and Development Program of China (2016YFD0100400), the National Special Key Project for Transgenic Breeding (Grant No. 2016ZX08001001), and the National Natural Science Foundation of China (Grant No. 31760380).

Acknowledgments

We thank Professor Qian Qian (State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou, Zhejiang, China) for giving us RPY geng materials.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.995634/full#supplementary-material

SUPPLEMENTARY TABLE 1
The statistic of sequencing data.

SUPPLEMENTARY TABLE 2
Repeat sequence statistics of Luohui 9 and RPY geng.

SUPPLEMENTARY TABLE 3
Function annotations of RPY geng genes.

SUPPLEMENTARY TABLE 4
Function annotations of Luohui 9 genes.

SUPPLEMENTARY TABLE 5
GO annotation of RPY geng / Luohui 9 unique orthologous clusters.

SUPPLEMENTARY TABLE 6
SV-related genes in Luohui 9 and RPY geng.

SUPPLEMENTARY TABLE 7
A 283 important known genes for important agronomic traits.

SUPPLEMENTARY TABLE 8
Homologous SV-related genes of 138 known functional genes.

SUPPLEMENTARY TABLE 9
The KGW of the RIL population, RPY geng, and Luohui 9.

SUPPLEMENTARY TABLE 10
The Meta-QTLs of KGW.

SUPPLEMENTARY TABLE 11
The known genes of KGW.

SUPPLEMENTARY TABLE 12
The known genes in KGW QTLs.

SUPPLEMENTARY TABLE 13
The heterosis-related QTLs of KGW.

References

Alioto, T., Blanco, E., Parra, G., and Guigó, R. (2018). Using geneid to identify genes. *Curr. Protoc. Bioinform.* 64:e56.

Altschul, S. F. (2012). Basic local alignment search tool (BLAST). *J. Mol. Biol.* 215, 403–410.

Arcade, A., Labourdette, A., Falque, M., Mangin, B., Chardon, F., Charcosset, A., et al. (2004). BioMercator: Integrating genetic maps and QTL towards discovery of candidate genes. *Bioinformatics* 20, 2324–2326. doi: 10.1093/bioinformatics/bth230

Arends, D., Prins, P., Jansen, R. C., and Broman, K. W. (2010). R/qtl: High-throughput multiple QTL mapping. *Bioinformatics* 26, 2990–2992.

Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D., and Veitia, R. A. (2010). Heterosis. *Plant Cell* 22, 2105–2112.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R. L., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31:1119.

Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M., and Buell, C. R. (2006). Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics* 7:327. doi: 10.1186/1471-2164-7-327

Chan, A. N., Wang, L. L., Zhu, Y. J., Fan, Y. Y., Zhuang, J. Y., and Zhang, Z. H. (2021). Identification through fine mapping and verification using CRISPR/Cas9-targeted mutagenesis for a minor QTL controlling grain weight in rice. *Theor. Appl. Genet.* 134, 327–337. doi: 10.1007/s00122-020-03699-6

Dai, X., Ding, Y., Tan, L., Fu, Y., Liu, F., Zhu, Z., et al. (2012). LHD1, an allele of DTH8/Ghd8, controls late heading date in common wild rice (*Oryza rufipogon*). *J. Integr. Plant Biol.* 54, 790–799. doi: 10.1111/j.1744-7909.2012.01166.x

Deng, X., Kong, W., Sun, T., Zhang, C., Zhong, H., Zhao, G., et al. (2022). Bin mapping-based QTL analyses using three genetic populations derived from *indica-japonica* crosses uncover multiple grain shape heterosis-related loci in rice. *Plant Genome* 15:e20171. doi: 10.1002/tpg2.20171

Du, H. L., Yu, Y., Ma, Y. F., Gao, Q., Cao, Y. H., Chen, Z., et al. (2017). Sequencing and de novo assembly of a near complete *Indica* rice genome. *Nat. Commun.* 8:12. doi: 10.1038/ncomms15324

Dupuis, J., and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* 151, 373–386.

Goel, M., Sun, H. Q., Jiao, W. B., and Schneeberger, K. (2019). SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 20:277. doi: 10.1186/s13059-019-1911-0

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. L., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). *Science* 296, 92–100.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:22. doi: 10.1186/gb-2008-9-1-r7

Huang, K., Wang, D., Duan, P., Zhang, B., Xu, R., Li, N., et al. (2017). WIDE AND THICK GRAIN 1, which encodes an otubain-like protease with deubiquitination activity, influences grain size and shape in rice. *Plant J.* 5, 849–860. doi: 10.1111/tpj.13613

Ishimaru, K., Hirotsu, N., Madoka, Y., Murakami, N., Hara, N., Onodera, H., et al. (2013). Loss of function of the IAA-glucose hydrolase gene *TGW6* enhances rice grain weight and increases yield. *Nat. Genet.* 45, 707–711. doi: 10.1038/ng.2612

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* 44:11.

Khahani, B., Tavakol, E., Shariati, V., and Fornara, F. (2020). Genome wide screening and comparative genome analysis for Meta-QTLs, ortho-MQTLs and candidate genes controlling yield and yield-related traits in rice. *BMC Genomics* 21, 294–318. doi: 10.1186/s12864-020-6702-1

Kong, W. L., Zhang, C. H., Qiang, Y. L., Zhong, H., Zhao, G. Q., and Li, Y. S. (2020). Integrated RNA-seq analysis and Meta-QTLs mapping provide insights into cold stress response in rice seedling roots. *Int. J. Mol. Sci.* 21:14. doi: 10.3390/ijms21134615

Kong, W., Deng, X., Yang, J., Zhang, C., Sun, T., Ji, W., et al. (2022a). High-resolution bin-based linkage mapping uncovers the genetic architecture and heterosis-related loci of plant height in *Indica-japonica* derived populations. *Plant J.* 110, 814–827. doi: 10.1111/tpj.15705

Kong, W., Li, S., Zhang, C., Qiang, Y., and Li, Y. (2022b). Combination of quantitative trait locus (QTL) mapping and transcriptome analysis reveals submerged germination QTLs and candidate genes controlling coleoptile length in rice. *Food Energy Security* 11:e354.

Kong, W., Sun, T., Zhang, C., Deng, X., and Li, Y. (2021a). Comparative Transcriptome analysis reveals the mechanisms underlying differences in salt tolerance between *Indica* and japonica rice at seedling stage. *Front. Plant Sci.* 12:725436. doi: 10.3389/fpls.2021.725436

Kong, W., Zhang, C., Zhang, S., Qiang, Y., Zhang, Y., Zhong, H., et al. (2021b). Uncovering the novel qtls and candidate genes of salt tolerance in rice with linkage mapping, RTM-Gwas, and RNA-seq. *Rice* 14:93. doi: 10.1186/s12284-021-00535-3

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5:9. doi: 10.1186/1471-2105-5-59

Li, D., Huang, Z., Song, S., Xin, Y., Mao, D., Lv, Q., et al. (2016). Integrated analysis of phenome, genome, and transcriptome of hybrid rice uncovered multiple heterosis-related loci for yield increase. *Proc. Natl. Acad. Sci. U.S.A.* 113, E6026–E6035. doi: 10.1073/pnas.1610115113

Li, F., Gao, Y., Wu, B., Cai, Q., and Wang, S. (2021). High-quality de novo genome assembly of Huajingxian 74, a receptor parent of single segment substitution lines. *Rice Sci.* 28, 109–113.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Li, X. K., Wu, L., Wang, J. H., Sun, J., Xia, X. H., Geng, X., et al. (2018). Genome sequencing of rice subspecies and genetic analysis of recombinant lines reveals regional yield- and quality-associated loci. *BMC Biol.* 16:12. doi: 10.1186/s12915-018-0572-x

Liu, Q., Han, R., Wu, K., Zhang, J., Ye, Y., Wang, S., et al. (2018). G-protein βγ subunits determine grain size through interaction with MADS-domain transcription factors in rice. *Nat. Commun.* 9:852. doi: 10.1038/s41467-018-03047-9

Majoros, W., Pertea, M., and Salzberg, S. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Martinez, A. K., Soriano, J. M., Tuberosa, R., Koumproglou, R., Jahrmann, T., and Salvi, S. (2016). Yield QTLome distribution correlates with gene density in maize. *Plant Sci.* 242, 300–309. doi: 10.1016/j.plantsci.2015.09.022

Ou, S. J., Su, W. J., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20:275.

Panibe, J. P., Wang, L., Li, J., Li, M. Y., Lee, Y. C., Wang, C. S., et al. (2021). Chromosomal-level genome assembly of the semi-dwarf rice Taichung Native 1, an initiator of Green Revolution. *Genomics* 113, 2656–2674. doi: 10.1016/j.ygeno.2021.06.006

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* 11, 1650–1667. doi: 10.1038/nprot.2016.095

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184, 3542–3558.e16. doi: 10.1016/j.cell.2021.04.046

Song, J. M., Xie, W. Z., Wang, S., Guo, Y. X., Koo, D. H., Kudrna, D., et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. *Mol. Plant* 14, 1757–1767. doi: 10.1016/j.molp.2021.06.018

Song, X. J., Kuroha, T., Ayano, M., Furuta, T., Nagai, K., Komeda, N., et al. (2015). Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. *Proc. Natl. Acad. Sci. U.S.A.* 112, 76–81. doi: 10.1073/pnas.1421127112

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19, II215–II225. doi: 10.1093/bioinformatics/btg1080

Stein, J. C., Yu, Y., Copetti, D., Zwickl, D. J., Zhang, L., Zhang, C., et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* 50, 285–296. doi: 10.1038/s41588-018-0040-0

Tanaka, T., Nishijima, R., Teramoto, S., Kitomi, Y., Hayashi, T., Uga, Y., et al. (2020). De novo genome assembly of the *Indica* rice variety IR64 using linked-read sequencing and nanopore sequencing. *G3 (Bethesda)* 10, 1495–1501. doi: 10.1534/g3.119.400871

Tang, H. B., Zhang, X. T., Miao, C. Y., Zhang, J. S., Ming, R., Schnable, J. C., et al. (2015). ALLMAPS: Robust scaffold ordering based on multiple maps. *Genome Biol.* 16:15. doi: 10.1186/s13059-014-0573-1

Tang, S. Y. Y., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* 43:10.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:14. doi: 10.1371/journal.pone.0112963

Wan, X. Y., Wan, J. M., Weng, J. F., Jiang, L., Bi, J. C., Wang, C. M., et al. (2005). Stability of QTLs for rice grain dimension and endosperm chalkiness characteristics across eight environments. *Theor. Appl. Genet.* 110, 1334–1346. doi: 10.1007/s00122-005-1976-x

Wang, Y., Xiong, G., Hu, J., Jiang, L., Yu, H., Xu, J., et al. (2015). Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.* 47, 944–948. doi: 10.1038/ng.3346

Xie, X., Du, H., Tang, H., Tang, J., Tan, X., Liu, W., et al. (2021). A chromosome-level genome assembly of the wild rice *Oryza rufipogon* facilitates tracing the origins of Asian cultivated rice. *Sci. China Life Sci.* 64, 282–293. doi: 10.1007/s11427-020-1738-x

Xu, L., Dong, Z. B., Fang, L., Luo, Y. J., Wei, Z. Y., Guo, H. L., et al. (2019). OrthoVenn2: A web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 47, W52–W58. doi: 10.1093/nar/gkz333

Xu, S. Z., Xu, Y., Gong, L., and Zhang, Q. F. (2016). Metabolomic prediction of yield in hybrid rice. *Plant J.* 88, 219–227.

Yan, S., Zou, G., Li, S., Wang, H., Liu, H., Zhai, G., et al. (2011). Seed size is determined by the combinations of the genes controlling different seed characteristics in rice. *Theor. Appl. Genet.* 123, 1173–1181.

Yan, W. H., Wang, P., Chen, H. X., Zhou, H. J., Li, Q. P., Wang, C. R., et al. (2011). A major QTL, *Ghd8*, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. *Mol. Plant* 4, 319–330. doi: 10.1093/mp/ssq070

Yu, J., Hu, S. N., Wang, J., Wong, G. K. S., Li, S. G., Liu, B., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp *Indica*). *Science* 296, 79–92.

Zhang, F., Xue, H., Dong, X., Li, M., Zheng, X., Li, Z., et al. (2022). Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res* 32, 853–863. doi: 10.1101/gr.276015.121

Zhang, H., Wang, Y., Deng, C., Zhao, S., Zhang, P., Feng, J., et al. (2022). High-quality genome assembly of Huazhan and Tianfeng, the parents of an elite rice hybrid Tian-you-hua-zhan. *Sci. China Life Sci.* 65, 398–411. doi: 10.1007/s11427-020-1940-9

Zhang, J. W., Chen, L. L., Sun, S., Kudrna, D., Copetti, D., Li, W. M., et al. (2016a). Building two *Indica* rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. *Sci. Data* 3:160076. doi: 10.1038/sdata.2016.76

Zhang, J. W., Chen, L. L., Xing, F., Kudrna, D. A., Yao, W., Copetti, D., et al. (2016b). Extensive sequence divergence between the reference genomes of two elite *Indica* rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. U.S.A.* 113, E5163–E5171. doi: 10.1073/pnas.1611012113

Zhang, X. C., Shabala, S., Koutoulis, A., Shabala, L., and Zhou, M. X. (2017). Meta-analysis of major QTL for abiotic stress tolerance in barley and implications for barley breeding. *Planta* 245, 283–295. doi: 10.1007/s00425-016-2605-4

Zhang, Y., Fu, J., Wang, K., Han, X., Yan, T., Su, Y., et al. (2022). The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. *Plant Biotechnol. J.* doi: 10.1111/pbi.13880

Zhao, H., Li, J. C., Yang, L., Qin, G., Xia, C. J., Xu, X. B., et al. (2021). An inferred functional impact map of genetic variants in rice. *Mol. Plant* 14, 1584–1599.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50, 278–284.

Zuo, J., and Li, J. (2014). Molecular genetic dissection of quantitative trait loci regulating rice grain size. *Annu. Rev. Genet.* 48, 99–118.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fpls.2022.1028885

OPEN ACCESS

EDITED BY
Jianyu Zhou,
Nankai University, China

REVIEWED BY
Xiaojiao Han,
Chinese Academy of Forestry, China
Bing Zhang,
Yangzhou University, China

*CORRESPONDENCE
Chen Bai
nmgnkybc@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 26 August 2022
ACCEPTED 20 September 2022
PUBLISHED 13 October 2022

CITATION
Li X, He W, Fang J, Liang Y, Zhang H, Chen D, Wu X, Zhang Z, Wang L, Han P, Zhang B, Xue T, Zheng W, He J and Bai C (2022) Genomic and transcriptomic-based analysis of agronomic traits in sugar beet (*Beta vulgaris* L.) pure line IMA1. *Front. Plant Sci.* 13:1028885. doi: 10.3389/fpls.2022.1028885

Genomic and transcriptomic-based analysis of agronomic traits in sugar beet (*Beta vulgaris* L.) pure line IMA1

Xiaodong Li[1†], Wenjin He[2†], Jingping Fang[2], Yahui Liang[1,3], Huizhong Zhang[1,3], Duo Chen[2], Xingrong Wu[1,3], Ziqiang Zhang[1,3], Liang Wang[1,3], Pingan Han[1,3], Bizhou Zhang[1,3], Ting Xue[2], Wenzhe Zheng[1,3], Jiangfeng He[1,3] and Chen Bai[1,3]*

[1]Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences, Hohhot, China, [2]Life Science College of Fujian Normal University, Fuzhou, China, [3]Inner Mongolia Key Laboratory of Sugarbeet Genetics & Germplasm Enhancement, Inner Mongolia Academy of Agricultural and Animal Husbandry Sciences, Hohhot, China

Sugar beet (*Beta vulgaris* L.) is an important sugar-producing and energy crop worldwide. The sugar beet pure line IMA1 independently bred by Chinese scientists is a standard diploid parent material that is widely used in hybrid-breeding programs. In this study, a high-quality, chromosome-level genome assembly for IMA1was conducted, and 99.1% of genome sequences were assigned to nine chromosomes. A total of 35,003 protein-coding genes were annotated, with 91.56% functionally annotated by public databases. Compared with previously released sugar beet assemblies, the new genome was larger with at least 1.6 times larger N50 size, thereby substantially improving the completeness and continuity of the sugar beet genome. A Genome-Wide Association Studies analysis identified 10 disease-resistance genes associated with three important beet diseases and five genes associated with sugar yield per hectare, which could be key targets to improve sugar productivity. Nine highly expressed genes associated with pollen fertility of sugar beet were also identified. The results of this study provide valuable information to identify and dissect functional genes affecting sugar beet agronomic traits, which can increase sugar beet production and help screen for excellent sugar beet breeding materials. In addition, information is provided that can precisely incorporate biotechnology tools into breeding efforts.

KEYWORDS

***Beta vulgaris*, whole-gene sequencing, whole genome duplications (WGD), gene family, genome wide association study (GWAS), male sterility**

Introduction

Sugar beet (*Beta vulgaris* L.) is in the Caryophyllales in the family Chenopodiaceae. The chromosome number of cultivated sugar beet is2n =2x= 18, with a predicted genome size of714 to 758 Mb (Arumuganathan and Earle, 1991). Sugar beet is an important biennial root crop cultivated in temperate climate regions with outstanding sugar-producing capability. Sugar beet was originated by selecting lines with high sugar content in the storage root from hybridizations between typical fodder beet and chardin the late eighteenth century (Eberhard, 1989) and thus is one of the most recently domesticated crops.

Sugar beet productivity is threatened by various pathogens, including bacteria, fungi, viruses, and nematodes (Larson et al., 2006; Saleh et al., 2011; Strausbaugh and Eujayl, 2018). Molecular breeding approaches have been used to create resistant or high taproot-yield sugar beet germplasms to increase production while greatly decreasing time, effort, and costs (Boyd et al., 2013). Many genes associated with important agronomic traits have been identified in sugar beet, including those responsible for nematode resistance (Cai et al., 1997), life cycle adaptation (Pin et al., 2012), cytoplasmic male sterility (Matsuhira et al., 2012), bolting tolerance (Hébrard et al., 2016), and salt tolerance (Sahashi et al., 2019). A wide range of sequence-based genetic and genomic resources are emerging for sugar beet. Single Nucleotide Polymorphism based genetic and physical maps have been constructed (Dohm et al., 2012; Holtgräwe et al., 2014), and transcriptome profiles have been analyzed to reveal important metabolic pathways and stress-responsive genes (Mutasa-Göttgens et al., 2012; Lv et al., 2018; Geng et al., 2019; Zou et al., 2020). Several sugar beet genomes have been assembled, including chromosome-level assemblies of double-haploid line RefBeet (Dohm et al., 2014) and the five-generation inbred line EL10 (Funk et al., 2018). Genome-wide identification and characterization of various important functional genes have also been reported (Stracke et al., 2014; Funk et al., 2018; Wang et al., 2019; Wu et al., 2019a; Wu et al., 2019b).

However, insufficient publicly available genetic resources and innovative germplasms are two major factors that limit the development of superior sugar beet cultivars. In this study, the chromosome-level genome assembly of the first Chinese native sugar beet line IMA1 was built by combining IlluminaHiseq, PacBio SEQUEL, and Hi-C sequencing platforms. Compared with previously released sugar beet assemblies, the new genome was 220 Mblarger with N50 size that was at least 1.6 times larger, thereby greatly improving the completeness and continuity of the sugar beet genome. Seven important beet agronomic traits and disease-resistance characteristics were also assessed by resequencing 114 accessions. In addition, a group of candidate genes associated with male sterility in sugar beet were selected based on q-PCR and transcriptome sequencing.

In conclusion, sequencing, assembly, and annotation of the sugar beet IMA1line provide the foundation for future comparative genomics efforts and phylogenetic reconstructions in the Caryophyllales and eudicots. Furthermore, valuable information is provided to identify and dissect functional genes affecting agronomic traits, which can be used to create breeding materials and to precisely incorporate biotechnology tools into breeding efforts.

Materials and methods

Sample collection and processing

Beta vulgaris IMA1, an inbreeding line with low level of heterozygosity, was selected for sequencing. Scientists from the Inner Mongolia Academy of Agricultural and Animal Science (IMAAAHS, Hohhot, Inner Mongolia, China) independently developed line IMA1. The line is standard diploid parent material with good combining ability that is widely used in creating sugar beet parent materials and hybrid breeding.

Seeds of IMA1 were planted in one gallon flowerpots filled with organic loamon August 16, 2018, and placed in a greenhouse at IMAAAHS. Greenhouse temperatures were 26°C(day) and 21°C (night). Two months after planting, tender, young, healthy leaf samples were collected and immediately flash-frozen in liquid nitrogen for one hour and then stored at −80°C until DNA and RNA extraction. Voucher specimens of IMA1 were deposited at IMAAAHS with collection number 14.S4006C.

Sampling germplasms of 114 sugar beet accessions

Test materials were 114 accessions randomly selected from the sugar beet gene bank stored at the Special Crop Research Institute of IMAAAHS. All test materials were planted in the experimental field of IMAAAHS (longitude 40°46′19.43″N, latitude 111°39′44.96″E)in Hohhot, Inner Mongolia, China. The complete data set contained three years (2017 to 2019) of agronomic traits collected in the field. Sugar beets were planted at the beginning of May and harvested at the beginning of October. Each plot was 6 m in length and 55-cm in width. The 114 sugar beet accessions were randomly sampled during the lush growth period. Newly emerged leaves were removed, put into zip lock bags, quickly frozen in a sample box with liquid nitrogen, and placed in a freezer at −80°C.

Selection of beet accessions for transcriptome analysis

Two pairs of beet lines with differences in male fertility were selected for transcriptome analysis: two male-sterile beet lines MS137 and MS301 and two beet maintainer lines OT152 and OT302. Beet roots that had undergone vernalization were planted in a test field arranged for beet breeding and isolation. On June 20, during the sugar beet budding stage, beet inflorescences with unopened, mature flower buds were selected and snap-frozen in liquid nitrogen.

DNA sequencing

To extract DNA and total RNA from young and healthy sugar beet leaf tissues, a DNeasy Plant Mini Kit (Qiagen, Germany) and an RNAprep pure Plant Kit (Tiangen, Beijing, China) were used, respectively. The DNA-seq was used to assist genome assembly, and the RNA-seq was used for gene model prediction. Low-quality reads and adaptor sequences were filtered out with the HTQC utility (Yang et al., 2013).

To obtain long reads for genome assembly, long read libraries were constructed using the extracted high-quality DNA in PacBio sequencing. Five SMRT (Single-Molecule Real Time Sequencing) cells were sequenced, and roughly 65.67 Gb of data were generated on a PacBio SEQUEL platform (Menlo Park, CA, USA) (Supplementary Table 1). With a genome size of 700 Mb assumed for sugar beet, the sequencing result theoretically represented 94-fold coverage. The average subread length was 10,727 bp, and the N50 length was 17,047 bp. The PacBio sequencing was combined with Illumina sequencing to generate longer scaffold genome assemblies.

Scaffold-level genome assembly of *Beta vulgaris* IMA1

High-quality Illumina sequences with a *K*-mer size of 17 were counted using the JELLYFISH program (Marçais and Kingsford, 2011). The PacBio sequencing subreads were assembled using Canu v1.7 (Koren et al., 2017). There were two steps of genome assembly polishing to correct random sequencing errors. Aquiver algorithm (Chin et al., 2013) was used to polish the Canu assembly using 50× long PacBio subreads. Next Generation Sequencing (NGS) short reads deliver a read accuracy of over 99% (Dohm et al., 2008). By contrast, with PacBio long reads, the error rate isas high as 15% to 20% (Ono et al., 2013; Ross et al., 2013). Therefore, two rounds of polishing were conducted with 67.22 Gb of Illumina short reads recruited with Pilonv 1.21 (Altschul et al., 1990; Walker et al., 2014). Organellar contigs were also removed by BLAST searches against organellar genomes of sugar beet (chloroplast genome: accession number KR230391.1; mitochondria genome: accession number BA000024.1).

High-throughput chromatin conformation capture library construction and chromosome assembly

In the current study, the Hi-C approach was used for chromosome-level assembly of sugar beet (Zhang et al., 2018; Chen et al., 2019; Zhang et al., 2020). To construct a Hi-C library, young leaves were cross-linked with formaldehyde and digested with *Dpn*II restriction enzyme overnight. Chimeric junctions were formed followed by biotinylating and proximity ligating sticky ends and then sheared and enriched for fragment sizes from 500 to 700 bp. Chimeric fragments were subjected to PE sequencing on an Illumina HiSeq X ten system (San Diego, CA, United States) with the PE 150 nt mode.

After mapping the clean sequencing reads against the polished sugar beet genome with Bowtie2 software (Langmead and Salzberg, 2012), over 369.4 million PE reads matched unique genomic locations, which were assessed and filtered by the hiclib Python library (Imakaev et al., 2012) and HiC-Pro program (Servant et al., 2015). Mis-joined contigs were corrected with the 3D-DNA pipeline (Dudchenko et al., 2017), and Hi-C-corrected contigs were grouped into pseudo-chromosomes by the ALLHIC pipeline (Zhang et al., 2019) on the basis of relations among valid reads.

Genome annotation

With gene model parameter strained from *Arabidopsis thaliana*, ab initio predictions were conducted using AUGUSTUS (Stanke and Morgenstern, 2005). Previously published sugar beet genome RefBeet-1.2.2 of sugar beet line RefBeet (Dohm et al., 2014) with accession number GCA_000511025.2 was selected as the reference genome to perform homology annotation. The protein sequences of the RefBeet genome were aligned with those of the new genome by TBLASTN software (Winsor et al., 2016). Gene structures were further predicted by GeneWise (Birney and Durbin, 2000) on the basis of TBLASTN results. The RNA-seq data sampled from leaf tissues were used for Trinity (Haas et al., 2013) *de novo* assembly. Transcript abundance was calculated with RNA-Seq by Expectation-Maximization (RSEM) (Li and Dewey, 2011), and transcripts with Fragments Per Kilobase Million (FPKM) <1 and iso-percentage <3% were filtered out. The PASA program (Haas et al., 2003) was used to construct comprehensive

transcripts using the filtered transcripts. Sugar beet transcripts were compared with the UniProt to identify candidates covering ≥95% of any target protein. Homology-based annotation, ab initio, and transcriptome-based gene prediction were combined to generate a protein-coding gene set by using the Evidence Modeler pipeline (Haas et al., 2008).Tandem Repeats Finder (Benson, 1999) and LTR_FINDER (Xu and Wang, 2007) were used to predict repeat elements. Subsequently, assembled genome sequences were aligned to the Repbase TE database (Bao et al., 2015) using Repeat Masker (Tarailo-Graovac and Chen, 2004) to search for sequences of repeat elements. The tRNAscan-SE (Schattner et al., 2005) and rRNAmmer (Lagesen et al., 2007) were used to detect reliable transfer RNA(tRNA) and ribosomal RNA(rRNA) positions, respectively. The small RNAs (sRNAs), microRNAs (miRNAs), and small nuclear RNAs(snRNAs)were predicted by searching the RFAM databases (Gardner et al., 2009) using INFERNAL software (Nawrocki et al., 2009) with the default parameters. For functional annotations, sequence-similarity searches were performed using Blast with *E*-value of 10^{-5} in available protein databases [(Non-Redundant Proteins (NR), Swiss-Prot, Clusters of Orthologous Groups (COGs), Kyoto Encyclopedia of Genes and Genomes(KEGG), and Gene Ontology(GO)].

Phylogenetic analysis and divergence time estimation

Phylogenetic analysis was conducted using the protein-coding genes of IMA1 and 25 other species. Protein sequence alignments and phylogenetic tree construction were conducted using OrthoFinder software (Emms and Kelly, 2019). Reconstruction of phylogenetic trees was inferred by maximum likelihood (ML), and the estimated divergence time of plant species based on the TimeTree database (Puttick, 2019) (http://www.time.org/) was used to recalibrate the divergence time for the 26 plant species. To identify the expansion and contraction of gene families, CAFE was used (Lu et al., 2017).

Synteny analysis and whole-genome duplication

The paralogous genes of IMA1 were identified in a BLASTP search (*E*-value cutoff of 1E−5). Synteny and collinearity blocks of those genes were analyzed using MCScanX (Wang et al., 2012). Gene synteny, gene density, and GC content on individual pseudo-chromosomes were mapped by using Circos software (http://www.circos.ca). The synonymous substitution rate (Ks) was calculated using KaKs_Calculatorand the Nei–Gojobori method (Wei and Zhang, 2014).

Single nucleotide polymorphisms and insertion and deletion calling

Trimmed reads were mapped to the new genome using BWA-MEM (Li, 2013). Average mapping rates were 99.33%, and average genome coverage was 7.72-fold of the reference genome. Mapping results were sorted and duplicate reads marked based on Sambamba (Tarasov et al., 2015). SNPs and InDels of the 114 accessions were called by GATK HaplotypeCaller (Hasanl et al., 2015). The results were calculated using the following parameters: QD < 2.0; MQ < 40.0; FS > 60.0; QUAL <30.0; MQrankSum <−12.5; Read PosRankSum <−8.0 -clusterSize 2 –cluster Window Size 5. The identified SNPs were filtered. High-quality SNPs were defined as only those with a minor allele frequency >0.05 and missing data rate<0.8. SNPs were annotated based on the genome with snpEff (Cingolani et al., 2012). Furthermore, SNPs were classified as coding synonymous SNPs and non-synonymous SNPs, and InDels in exons were grouped based on whether they led to a frameshift.

Genome wide association study analysis

Genome wide association study was performed by using FaST-LMM (v2.07.20140723) or EMMAX (Kang et al., 2010). A total of3,738,500 SNPs with a minor allele frequency of 0.05 or greater and a missing data rate of 80% or less in the entire population were used for GWAS. A Bonferroni correction was used to determine the genome-wide significance thresholds of the GWAS, based on a nominal level of –log10(P) values of 5.

Results

Sequencing and assembly of IMA1 genome

The Illumina resequencing reads combined length was 67.22 Gb, which was 96× the estimated genome size. The RNA-seq generated a clean dataset of 15.93 Gb consisting of over 98.9 million Paired-end reads. Quality of Illumina resequencing reads was high (92.06% with Phred quality score >30). In total, 448 million high-quality, 150-bp clean paired-end reads were retained for use in the following analysis (Supplementary Table 2). The 17-mer analysis-based genome size of sugar beet was estimated at 720.5 Mb. A single main peak indicated the nature of the isolated genomic material, with heterozygosity of only 0.6% (Supplementary Figure 1). For accurate homozygous assembly, Illumina, Pacbio, and Hi-C sequences were combined to perform the sequencing. Approximately 120.75 Gb of clean data consisting of 805 million PE reads were produced from the

Hi-C library sequencing (Supplementary Table 2). An initial 786-Mb genome sequence was obtained consisting of 4,824 contigs, with contig N50 of 367.5 kb. The longest contig was 5.91 Mb (Table 1). Additionally, 4,576 contigs from the Canu assembly were successfully clustered, ordered, and oriented to nine pseudo-chromosomes. In the IMA1 genome, 171 syntenic blocks were detected, which involved 3,508 genes (Figures 1, 2). The results indicated the quality of the genome assembly for IMA1was high. The interaction signals were enriched in chromosomes, and the intensity of interaction along the diagonal was relatively smooth, showing well-organized contig orderings. The anchor rate was 99.1%, and only 248 contigs (7.1 Mb) were not anchored. The scaffold N50 was 93.06 Mb, and the longest chromosome values reached 112.63 Mb (Supplementary Table 3).

TABLE 1 Assembly statistics of *B. vulgaris* IMA1nuclear genome.

	Canu	HiC
Assembly genome size (Mb)	786.13	786.59
Genomic G+C content	35.85%	35.85%
Number of assembled scaffolds	4,824	257
Number of scaffolds (> 2 kb)	4,824	257
Max Length (Mb)	5.91	112.64
Scaffolds N50 (kb)	367.5	93.06

Evaluation of the genome assembly

Assembled genomes were further validated by mapping NGS short reads, which indicated that 446.7 million (99.23%) Illumina reads were reliably aligned, which covered 96.84% of the assembly (Supplementary Table 4). Additionally, 96.8% to 98.08% of RNA-seq clean reads were reliably aligned to the assembled genome. Genome completeness was assessed based on the viridiplantae_odb9 database in the BUSCO program (Jia et al., 1997). A total of 1,326 (96.4%) complete single-copy orthologs among 1,375 conserved plant genes were recalled in the assembly (Supplementary Table 5). We assessed the coherence of the IMA1 genome assembly with LAI (Long terminal repeat assembly index). LAI score was assessed by

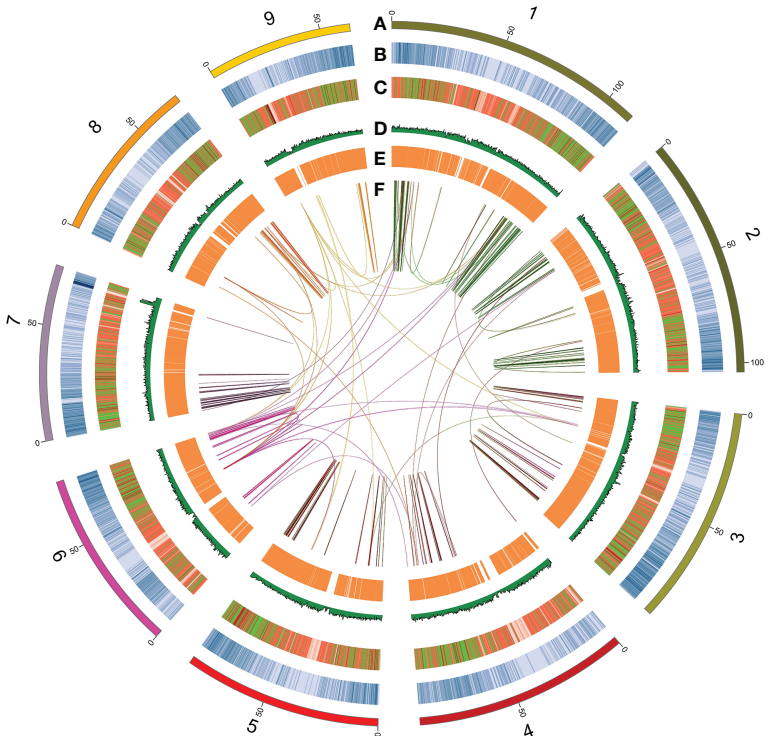


FIGURE 1
Circos plot showing the distribution of genomic features along the IMA1 genome. The rings from outermost to innermost indicate **(A)** nine pseudo-chromosomes of Beta vulgaris IMA1genome; **(B)** gene density distributed inside 200-kb sliding windows; **(C)** transposable element abundance; **(D)** distribution of GC content; **(E)** expression values of leaf-expressed genes; and **(F)** schematic presentation of major inter-chromosomal relations in the *B vulgaris* IMA1 genome. Each line represents a syntenic block; block size = 3 kb. Chromosomes in the outer ring are ordered by chromosomes length as follow:1, chr5; 2, chr4; 3, chr3; 4, chr7; 5, chr6; 6,chr9; 7, chr1; 8, chr8; 9, chr2.

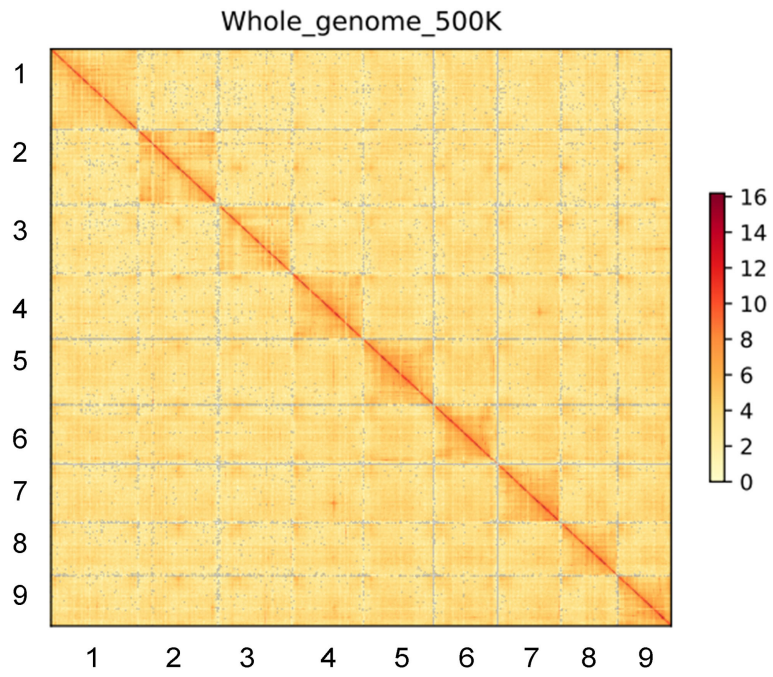


FIGURE 2
Integrated Hi-C interaction heatmap of *B. Vulgaris* IMA1 genome. The heatmap displays high-resolution single pseudo-chromosomes, which were scaffolded and composed independently. Lines are ordered by chromosomes length as follow:1, chr5; 2, chr4; 3, chr3; 4, chr7; 5, chr6; 6, chr9; 7, chr1; 8, chr8; 9, chr2.

LTR_RETRIEVER (v2.9.0) (Ou and Jiang, 2017). The LAI value of the IMA1 genome was 13.4, which was at the Reference level.

Gene prediction and functional annotation

In the IMA1 genome, 35,003 genes encoding proteins were annotated. Average gene length was 1,121 bp. Total combined length of all genes was 39.23 Mb, which accounted for 4.99% of the assembled genome. According to the BUSCO assessment, 86.2% of core eukaryotic genes were complete in the assembly. Totals of 32,043; 27,574; 20,155; 10,157; and 21,351 genes were annotated in Nr, GO, COG, KEGG, and Swiss-Prot databases, respectively, and 32,047 (91.56%) genes had at least one hit to the databases (Supplementary Table 6). There were 8,725 genes annotated in all five databases, representing 24.93% of all protein-coding genes. Based on KEGG annotation (Supplementary Figure 2), 10,157 genes were involved in 33 pathways. There were 1,442 tRNAs, 945 5S rRNAs, 138 18S rRNAs, 139 28S rRNAs, 410 snRNAs, and 56 miRNAs in the IMA1 genome (Supplementary Table 7). The IMA1 genome contained a total of 512.72 Mb of repetitive sequences, with more than 284,501 tandem repeats identified (Supplementary Table 8).

Comparisons of the IMA1 genome assembly with previously reported sugar beet genome assembly

The new sugar beet IMA1 assembled genome (~786 Mb) was compared with the two previously released chromosome-level assemblies of *B. vulgaris*: line RefBeet (~566 Mb, accession numbers: GCA_000511025.2) (Dohm et al., 2014) and EL10 (~540 Mb, accession numbers: GCA_002917755.1) (Funk et al., 2018). The new genome was much larger than those previously reported. In addition, the IMA1 genome had the longest chromosome length of 112.63Mb and the largest number of genes identified, with 35,003 genes. The two previous genome assemblies of *B. vulgaris* had scaffold N50 of 57.94 Mb and 2.01 Mb, respectively, which were much shorter than the 93.06 Mb in the current assembly (Supplementary Table 3). There were 257 scaffolds in the new genome assembly, and longer scaffold N50s were obtained than those in the EL10 and RefBeet genome, which was the best assembled genome to date. The completeness and continuity of the new assembly might be attributed to the high-sequencing depth of PacBio and Hi-C reads and the extremely low heterozygosity of the sugar beet line.

The IMA1 genome contained a total of 512.72 Mb of repetitive sequences, which were 65.18% of the IMA1 genome.

It was higher than the previously released genome of sugar beet line EL10 and RefBeet (62.91% and 51.75%, respectively) (Dohm et al.,2014; Funk et al., 2018). The most abundant repetitive sequences in the IMA1 genome are Class I retroelement (66.65% of total TEs and 43.44% of genome). The Long terminal repeat retrotransposons (LTR-RTs) of IMA1 accounted for 31.24% of the assembly, while those of EL10 and RefBeet accounted for 28.07% and 21.82%, respectively. Over 284,501 tandem repeats were identified, representing 10.34% of the genome. (Supplementary Table 8). Compared with RefBeet and EL10, IMA1 annotated the highest proportion and number of repetitive sequences with significant improvements in the continuity and integrity of repeat regions.

The synteny analysis showed that the *B. vulgaris* IMA1 assembly shared 17,462 and 14,551 common gene pairs with EL10 and RefBeet, respectively, indicating a high ratio of the syntenicregion. Most sequences in RefBeet and EL10 genomes aligned with corresponding counterparts in the IMA1 assembled genome; whereas the IMA1 assembly had extended sequences, especially in Chr1, Chr3, Chr4, and Chr7. Some genomic arrangements were also observed in the IMA1genome compared with RefBeet and EL10 (Figure 3).

Evolution and gene family analysis of the *Beta vulgaris* IMA1 genome

To analyze genome evolution and divergence time of IMA, some genome sequences of plant species were selected. Gene family expansions were greater than contractions in *Nymphaea colorata*, *Brassica napus*, *Chenopodium quinoa*, *B. vulgaris* IMA1, *Malusbaccata*, *Rosa chinensis*, *Cannabis sativa*, *Juglansregia*, *Quercussuber*, *Duriozibethinus*, and *Camellia sinensis*, compared with the other species. In the phylogenetic tree, published *B. vulgaris* and IMA1 phylogenetically diverged into the Betoideae branch approximately 11 million years ago (Mya). Results also showed that published *B. vulgaris* and IMA1 were sisters in coccolithophores, which is consistent with the findings of phylogenetic analysis (Figure 4A).

Age distribution of duplicated genes was determined, followed by using a mixture model implemented in the mixtools R package (Benaglia et al, 2009) to identify significant gene duplication peaks consistent with whole genome duplications (WGDs). The median replication peak for IMA1 was around 0.55, which was younger than the ortholog divergence of IMA1 and *A. thaliana* (Ks, ~2.56) (Figure 4B). The distribution of ks values indicated that only one recent WGD event occurred in the IMA1 genome, whereas an ancient WGD event occurred 29 Mya ago.

From the 26 species, orthologous protein groups were delineated, and 35,818 orthologous groups were obtained (Figure 4A). In the IMA1 genome, 2,907 gene families expanded and 2,781 contracted. The 2,907 expanded gene families were annotated in KEGG and GO databases. In the GO analysis, the expanded orthologous groups were associated with biological regulation, growth, reproductive process, and signaling. In the KEGG analysis, most of the expanded genes were enriched to the categories of cell growth and death, plant hormone signal transduction, and environmental adaptation. The 2,781 contracted gene families were associated with signal transduction and steroid biosynthesis, as well as metabolism of pyruvate, terpenoids, polyketides, or lipids. In KEGG and GO analysis, contracted genes were also involved in developmental process and regulation of biological process.

In the comparison of IMA1, RefBeet, *A. thaliana*, *B. napus*, *C. quinoa*, *C. sativa*, *O. sativa*, and *S. oleracea*, 9,128 gene families were shared among these species (Figure 4C). According to the GO analysis, functions of those genes were primarily in growth, reproductive process, stimulus response, developmental process, and immune system. According to the KEGG analysis, enriched pathways for the genes included phenylpropanoid biosynthesis, purine metabolism, pyrimidine metabolism, and arginine biosynthesis.

Phylogenetic analysis of *SWEET, SUT, SPS* and *SUS* gene families

To analyze evolutionary relations, a phylogenetic tree was constructed with *SWEET* (sugars will eventually be exported transporters) gene family members from *A. thaliana* (17), *B. vulgaris* IMA1 (9), RefBeet (16), and EL10 (10) (Figure 5A). Nine *SWEET* genes were found in the IMA1 genome, and they were grouped into four clusters: 1, 2, 3, and 4. In cluster 1,there were more subfamily genes of the *SWEET* family in IMA1 than in *B. vulgaris* RefBeet and EL10. Therefore, cluster 1 members from the *SWEET* family in IMA1 might have a more important role in sugar export transportation. In the *SUT* (sucrose transporters) gene family, a transmembrane transporter was involved in the absorption and transport of sucrose.

Evolutionary relations among *SUT* gene proteins from *B. vulgaris* IMA1(11), RefBeet(12), and EL10 (8) and *C. quinoa* (9), *S.oleracea* (12), and *A. thaliana* (9) were also determined *via* phylogenetic tree analysis (Figure 5B). The *SUT* gene proteins were classified into three groups, including clusters 1, 2, and 3 of subfamily genes. In cluster 3, *B. vulgaris* IMA1 had eight *SUT* genes, which was higher than that of RefBeet (6) and EL10 (6). It was hypothesized that the cluster 3 gene proteins have a key role in sucrose accumulation in IMA1.

Evolutionary relations among *SPS* (sucrose phosphate synthase) gene proteins from *B. vulgaris* IMA1(3), RefBeet (1), and EL10 (2) and *C. quinoa* (4),*Cucumissativus* (3), *B. napus* (5),

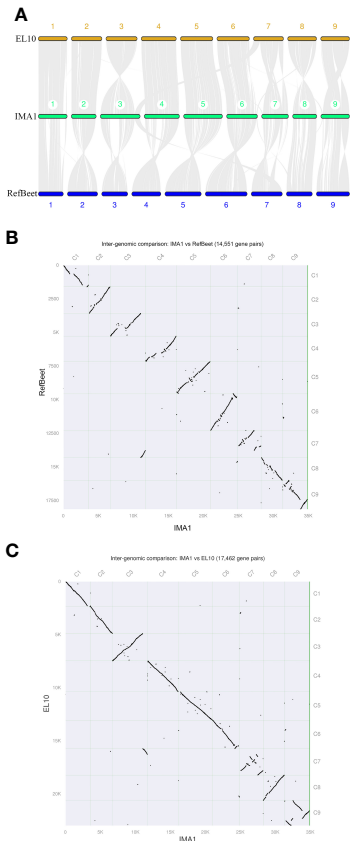


FIGURE 3
Genomic alignment among three genome assemblies of sugar beet lines IMA1, EL10, and RefBeet. **(A)** Schematic representation of synteny among IMA1, EL10, and RefBeet genomes. Gray lines connect matching gene pairs. **(B)** Scatter plot of syntenic blocks of conserved genes between *Beta vulgaris* IMA1 and RefBeet genomes. **(C)** Scatter plot of syntenic blocks of conserved genes between *B. vulgaris* IMA1 and EL10 genomes. Chromosome order in the new assembly was determined by length (from largest to smallest). Rightward and downward are 5′ to 3′ on assembly plus strands.

O. sativa (4), *S.oleracea* (2), and *A. thaliana* (4) were determined *via* phylogenetic tree analysis (Supplementary Figure 3). The *SPS* gene proteins were divided into clusters 1, 2, and 3. In IMA1, clusters 3 and 2 had two and one *SPS* genes, respectively. In addition to *SWEET*, *SUT*, and *SPS* gene families, evolutionary relations of the *SUS* (sucrose synthase) gene family were also analyzed (Supplementary Figure 4). The *SUS* gene proteins were analyzed in *B. vulgaris* IMA1 (4), RefBeet (6), and EL10 (4) and *A. thaliana* (6), *C. quinoa* (7), and *S. oleracea* (4). Numbers in *SPS* and *SUS* gene families in IMA1 were fewer than those in other species. However, because IMA1 accumulated higher sugar content than that in other species, it was hypothesized that *SPS* and *SUS* gene family members in IMA1 had higher sugar accumulation efficiency than that in the other species.

Genome wide association study of seven agronomic traits in *Beta vulgaris* IMA1

Phenotyping data of seven major agronomic traits of 114 *B. vulgaris* samples were used to perform GWAS (Supplementary Table 9). Sucrose content is an important economic trait for superior individuals of *B. vulgaris*. Nine strong GWAS signals were detected, including *BvNR* (IMABv01g023663), *BvGN4* (IMABv01g023668), *BvMYST1* (IMABv01g023671), *BvPGD* (IMABv01g018581), *BvSNAT* (IMABv01g018582), *BvCDK12_13* (IMABv01g018584), *BvGBF* (IMABv01g018599), *BvPOD*(IMABv01g018569), and *BvTOGT1*(IMABv01g018570) genes (Figure 6A; Supplementary Figure 5A; Supplementary Table 10).

Some strong GWAS signals on Chr6 and Chr7 were significantly associated with sugar yield per hectare, which is an important target in sugar beet breeding. For example, three genes were located in the strong association peaks, including *BvPGD* (IMABv01g018581), *BvE3.2.1.6* (IMABv01g018526), and *BvSLC35F1-2* (IMABv01g015264), which participate in the carbohydrate metabolism. The gene*BvYGK1* (IMABv01g018527) is associated with purine metabolism, and there was a strong GWAS signal on Chr7 for *BvACP7* (IMABv01g015268), which is associated with purple-acid phosphates (Figure 6B; Supplementary Figure 5B; Supplementary Table11). In addition, genes were also identified that were associated with root yield per hectare, including *BvPGD* (IMABv05g004241), *BvSTP* (IMABv04g007036), *BvHPGT* (IMABv05g004244), and glucose-6-phosphate 1-epimerase (IMABv05g004245), which were associated with the pentose phosphate pathway and galactose and monosaccharide transport. Serine/threonine protein kinase (STPK), a type of eukaryotic cell-like protein kinase, is involved in the transport of glucose and glutamine (Jia et al., 1997). Genes *BvULK4* (*IMABv03g010205*), *BvTMK1* (IMABv09g022186), and *BvPTO* (IMABv03g013119) code serine/threonine kinases and were also associated with root yield per hectare (Figure 6C; Supplementary Figure 5C; Supplementary Table 12).

Root rot, damping off, and rhizomaniaare emerging serious threats to sugar beet production. In the GWAS, several genes associated with disease defense were identified, including *BvTSSK6* (IMABv02g031103), *BvCLCN7* (IMABv09g022351), *BvPRPS* (IMABv01g024513), *BvEXO1* (IMABv08g027895), *BvFAR1* (IMABv04g006805), *BvSERK1* (IMABv09g023194), *BvLRR* (IMABv03g010906), *BvPTI1* (IMABv03g010905), *WRKY1* (IMABv09g020695), and *BvDELLA* (IMABv09g020694) (Figures 6D–F; Supplementary Figures 5D–F; Supplementary Tables 13–15).

In the GWAS analysis on pollen scale types of different beet varieties, there were some strong signals on Chr2, Chr3, and Chr4. Genes were identified that were related to pollen number, including *BvHSFF* (IMABv03g011676), *BvQUA3* (IMABv03g011680), *BvARR-B* (IMABv03g011683), *BvBRI1* (IMABv02g031269),

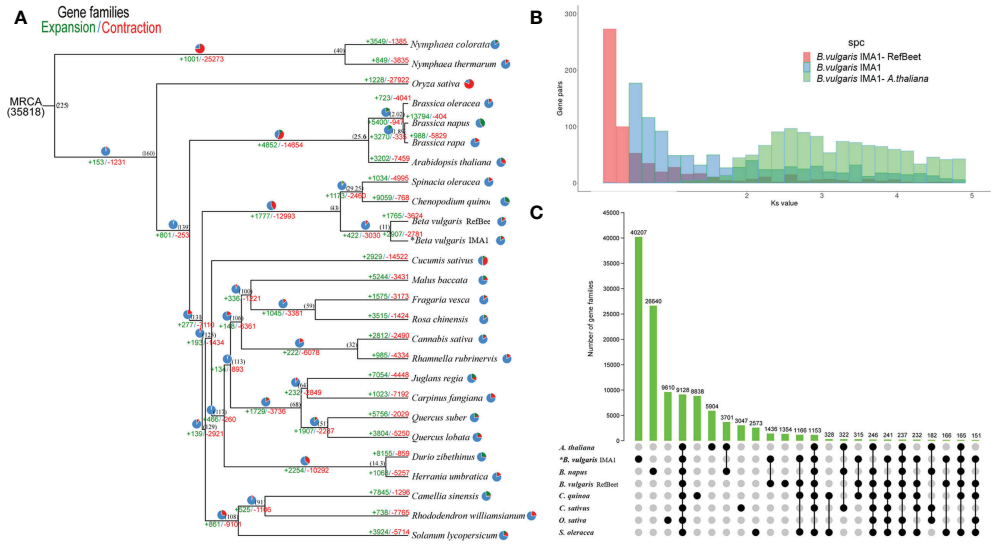


FIGURE 4
(A) Phylogenetic tree of gene families number unveiling expansion (green) and contraction (red) among 26 species. Pie diagrams represent the ratio of expanded (green), contracted (red), and conserved (blue) genes among whole gene families. The estimated divergence time (million years ago) is shown in black next to the phylogenetic tree. MRCA: most recent common ancestor. **(B)** Ks distributions for duplicated gene pairs in Beta vulgaris IMA1, RefBeet, and Arabidopsis thaliana. **(C)** UpSet plot of gene families intersection in *B. vulgaris* IMA1, RefBeet, *A. thaliana*, Brassica napus, Chenopodiumquinoa, Cannabis sativa, Oryza sativa, and Spinacia oleracea. Gene family numbers (clusters) are marked for each species and species intersection.

BvNXN (IMABv02g031270), *BvERAL1*(IMABv02g031271), *BvFRK1*(IMABv04g005357), *BvDELLA*(IMABv04g005358), and *BvANKRD44* (IMABv04g005362). The gene *BvQUA3*, a putative homo-galacturonan methyl-transferase, is involved in regulating cell wall biosynthesis in *Arabidopsis* suspension-cultured cells (Miao et al., 2011). The gene *BvARR-B* is a member of the two-component response regulator ARR-B family, which is a partially redundant negative regulator of cytokinin signaling (To et al., 2004; Mason et al., 2005). The gene *BvBRI1*, protein brassinosteroid insensitive 1, is another gene associated with plant hormone signal transduction, which can transfer phosphorus-containing groups (Zipfel, 2008). The gene *BvFRK1*, a target of AtWRKY6 regulation during plant senescence, is a senescence-induced receptor-like serine/threonine-protein kinase (Robatzek and Somssich, 2002) (Figure 6G; Supplementary Figure 5G; Supplementary Table 16). The results provide valuable information on the characteristic genes associated with *B. vulgaris* pollen fertility, which can be used in molecular breeding.

Gene ontology and KyotoEncyclopedia of genes and genomes pathway analysis of differential expressed genes

Transcriptomes of two pairs of sugarbeet cytoplasmic male sterility (CMS) lines were compared (MS137 vs. OT152 and MS301 vs. OT302). MS137 and MS301 are sugar beet sterile lines, and OT152 and OT302 are sugar beet maintainer lines. There were 2,032 and 2,090 significant DEGs identified in MS137 vs. OT152 and MS301 vs. OT302 comparisons, respectively (Supplementary Figure 6; Supplementary Tables 17, 18). Six hundred and twenty-one DEGs were identified in both MS137 vs. OT152 and MS301 vs. OT302 comparisons (Supplementary Tables 19, 20). In the KEGG analysis, the 621 shared genes were enriched in plant–pathogen interaction [two up-regulated genes, including *FRK1* (IMABv09g022061) and *RPS2* (IMABv03 g008950)], glycolysis/gluconeogenesis [two down-regulated genes, including *pdhC* (newGene_5384) and *gapN* (IMABv04g008381)], photosynthesis [one down-regulated gene, *petF* (IMABv02g032492)), and MAPK signaling pathway (one up-regulated gene, *FRK1* (IMABv09g022061)]. "Binding" (GO:0005488, four up- and one down-regulated genes) and "catalytic activity" (GO:0003824, five up- and two down-regulated genes) were the two most enriched GO terms in the molecular function ontology. "Cell" (GO:0005623, two up- and one down-regulated genes) and "membrane part" (GO:0044425, three up- and two down-regulated genes) were the two most enriched GO termsin the cellular component. In addition, there were 997 up-regulated and 1,035 down-regulated DEGs in MS137 vs. OT152 (Supplementary Table 17) compared with 997 up-regulated and 1,093 DEGs in MS301 vs. OT302 (Supplementary Table 18). Among those DEGs, 334 were up-regulated and 285

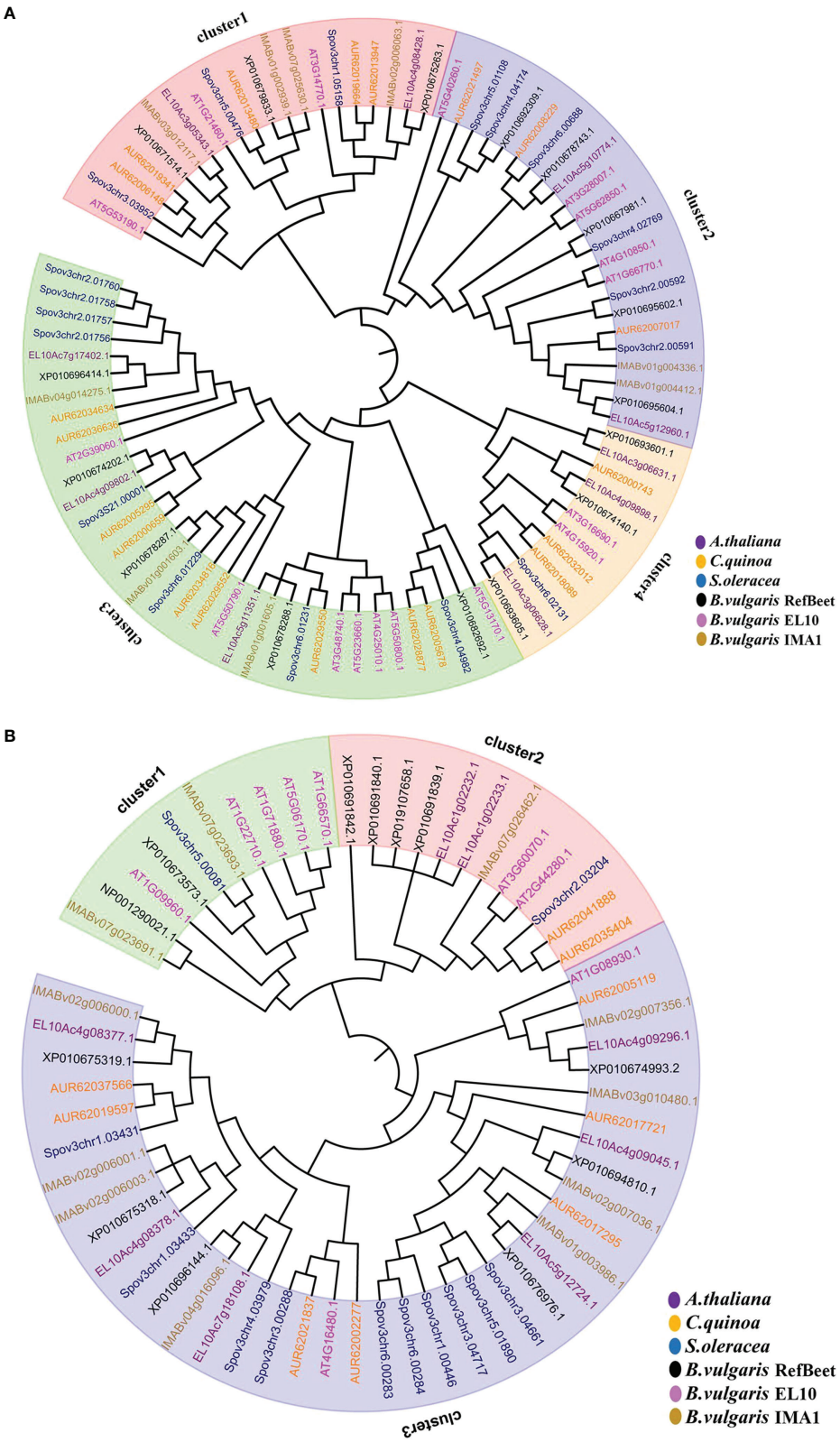


FIGURE 5
Genes in the *SWEET* and *SUT* families were clustered by neighbor-joining method. **(A)** Evolutionary tree of *SWEET* genes in *Arabidopsis thaliana*, *Chenopodium quinoa*, *Spinacia oleracea*, and *Beta vulgaris* IMA1, EL10, and RefBeet. **(B)** Evolutionary tree of *SUT* genes in *A. thaliana*, *C. quinoa*, *S.oleracea*, and *B. vulgaris* IMA1, EL10, and RefBeet.

were down-regulated between MS137 vs. OT152 and MS301 vs. OT302 (Supplementary Figure 7).

Based on GO and KEGG analyses of differential expression, six genes with significant differential expression were selected for a real-time fluorescence quantitative PCR test for verification (Supplementary Figure 8; Supplementary Table 21). The SGNH hydrolase gene (IMABv04g006046), GDSL esterase gene (IMABv05g001851), galacturonase gene (IMABv04g007649), and pectinlyase gene (IMABv07g016298) were up-regulated in maintainers. Genes for UDP-glucosyltransferase (IMABv06g016651) and cytochrome P450 (IMABv09g022885) were up-regulated in sterile lines. The results showed that expression profiles of the genes were consistent with transcriptome results.

Discussion

The newly assembled genome was compared with the two previously released chromosome-level assemblies of *B. vulgaris*: line RefBeet (Dohm et al., 2014) and EL10 (Funk et al., 2018) (accession numbers: GCA_000511025.2 and GCA_002917755.1, respectively). The covered genome size of ~786 Mb was very close to the estimated sugar beet genome of 714 to 758 Mb (Arumuganathan and Earle, 1991) and was much larger than that of previous reports (RefBeet about ~540 Mb and EL10 about ~566 Mb). Compared with EL10, the best previously assembled genome, the new genome contained fewer scaffolds (257) and had a longer scaffold N50 (93.06Mb), indicating a significant improvement in sequence continuity.

When sugar beet IMA1 assembly and RefBeet genome were compared, the synteny analysis revealed that part of segments in Chr6 of IMA1 had inverted compared with the counterpart in Chr9 of RefBeet. Inherited variation between the two sugar beet cultivars and the much more accurate and complete assembly of IMA1 genome might be major reasons for differences. Overall, the quality of the new genome assembly of *B. vulgaris* IMA1was higher than that of the RefBeet genome, and therefore, it will be valuable in genetic analyses of sugar beet and related species.

The *SUS* and *SPS* gene families are well documented in plants, and gene family members vary from species to species (Castleden et al., 2004). In the metabolism of uridine diphosphate glucose, it is catalyzed and hydrolyzed to sucrose, and SPS is the key rate-limiting enzyme in the process (Lunn and Macrae, 2003). Changes in sugar content are closely related to expression levels of *SUS* and *SPS* genes (Lv et al., 2018). For example, increases inactivities of SUS and SPS enzymes are correlated with increases in sucrose content in the high sucrose-accumulating Japanese pear 'Chojuro'. By contrast, activity of the enzymes does not increase in the low sucrose-accumulating pear cultivar 'Yali' during fruit ripening (Moriguchi et al., 1992). In addition, in the early stages of fruit development, Asian pear cultivars 'Niitaka' and 'Whangkeumbae' have relatively low sucrose content with relatively low activities of SUS and SPS enzymes, but when sucrose content reaches the peak value, SUS and SPS enzymes have the highest activities (Choi et al., 2009).

Cluster analysis of the *SUS* gene family in six dicotyledons was performed, and 31 genes were categorized into three different clusters. Six and four *SUS* genes were identified in sugar beet RefBeet and sugar beet IMA1, respectively. Similarly, *SPS* gene families were compared in nine dicotyledons, and 28 genes were categorized into three different clusters. Three and two *SPS* genes were identified in IMA1 and EL10, respectively, which were categorized to clusters 2 and 3, respectively. In addition, the number of *SPS* genes was species-related, and sugar metabolism regulation was related to the activity of SPS enzymes but was not affected by the quantity of genes.

In the distribution and transport of sucrose from source to sink in plants, sucrose transporters (*SUTs*) are important genes (Chao et al., 2020). However, the molecular mechanisms of SUT function in the sugar metabolism pathway are not fully understood. Three different SUT clusters have been identified in the analysis of *SUT* gene family clusters in eight dicot species (Chen et al., 2010).

The *SUS, SPS*, and *SUT* gene families are involved in sucrose synthesis, transport, and accumulation. Although there are fewer *SUS*, *SPS*, and *SUT* genes, sugar beet accumulates much more sugar in storage tissues than that of other dicots. Therefore, it was hypothesized that compared with other species, members of those gene families in sugar beet have more important roles in sugar catalysis and sugar transport efficiency or some strong transcription regulatory factors regulate those functional genes. As a result, sugarbeet has strong capability to synthesize, transport, and accumulate sugar.

The *SWEET* gene family in plants is categorized into four different clusters. SWEETs in cluster1 are mainly responsible for glucose transport. For example, AtSWEET1 of *Arabidopsis* can mediate the absorption and transport of glucose (Chong et al., 2014; Tao et al., 2015). SWEETs in cluster 2 are mainly responsible for monosaccharide transport (Chong et al., 2014). Most of the SWEETs in cluster3 are associated with sucrose transportation (Kryvoruchko et al., 2016). In *Arabidopsis*, AtSWEET11 and AtSWEET12 are responsible for transporting intracellular sucrose to the apoplast and then moving it into the phloem for long-distance transport (Chen et al., 2012). In cluster 4, AtSWEET16 is associated with transport of glucose, fructose, and sucrose (Klemens et al., 2013). In IMA1, 11 *SWEET* family genes were identified, including four in cluster1, one in cluster2, four in cluster3, and two in cluster4. It was hypothesized that the *SWEET* family genes in IMA1 are involved in transporting sucrose, fructose, and glucose, as well as long-distance transport from mesophyll cells into the phloem.

In summary, the data collected from gene sequencing of IMA1 were used to identify the members of *SUS*, *SPS*, *SUT*, and *SWEET*

gene families that are generally considered to be crucial genes involved in plant sugar metabolism. Genes related to disease resistance were also identified. Candidate genes were nominated with the potential to regulate sugar metabolism and improve sugar productivity. Genes were also nominated that were related to disease-resistance, which could be targets for genetic improvement.

In this study, GWAS was performed for a set of sugar beet agronomic traits. Ten disease-resistance genes significantly associated with root rot, damping off, and rhizomania were identified. Five genes were identified that had significant relations with sugar yield per hectare of sugar beet. Among those genes, *BvSLC35F1-2* is involved in carbohydrate metabolism, whereas gene *BvACP7* codes a purple-acid phosphatase in a family of binuclear metallohydrolases identified in plants, animals, and fungi (Flanagan et al., 2006). In addition, nine highly expressed genes associated with sugar beet pollen fertility were identified. Those genes were involved in regulating cell wall biosynthesis, plant hormone signal transduction, and plant senescence. Among six significant DEGs, SGNH hydrolase, GDSL esterase, and pectinlyase were associated with another development, pollen wall development, and pollen tube growth (Guan et al., 2008; Wang et al., 2018; An et al., 2019). Those genes were down-regulated in sugar beet sterile lines, which might be related to sugar beet pollen abortion and male sterility. Plant auxin metabolism involves cytochrome P450 (Feldmann, 2001), and excessive auxin content can lead to stunting and sterility of plants. Cytochrome P450 was significantly up-regulated in sugar beet sterile lines, which might be related to sugar beet fertility. Yuan long Wu (Wu et al., 2022) recently identified a galacturan 1, 4-alpha-galacturonidase [EC:3.2.1.67] gene that controls the formation of cotton pollen outer cell wall. They revealed the important role of galacturan 1, 4-alpha-galacturonidaseis to de-esterify homogalacturonan in the formation of the outer wall of cotton

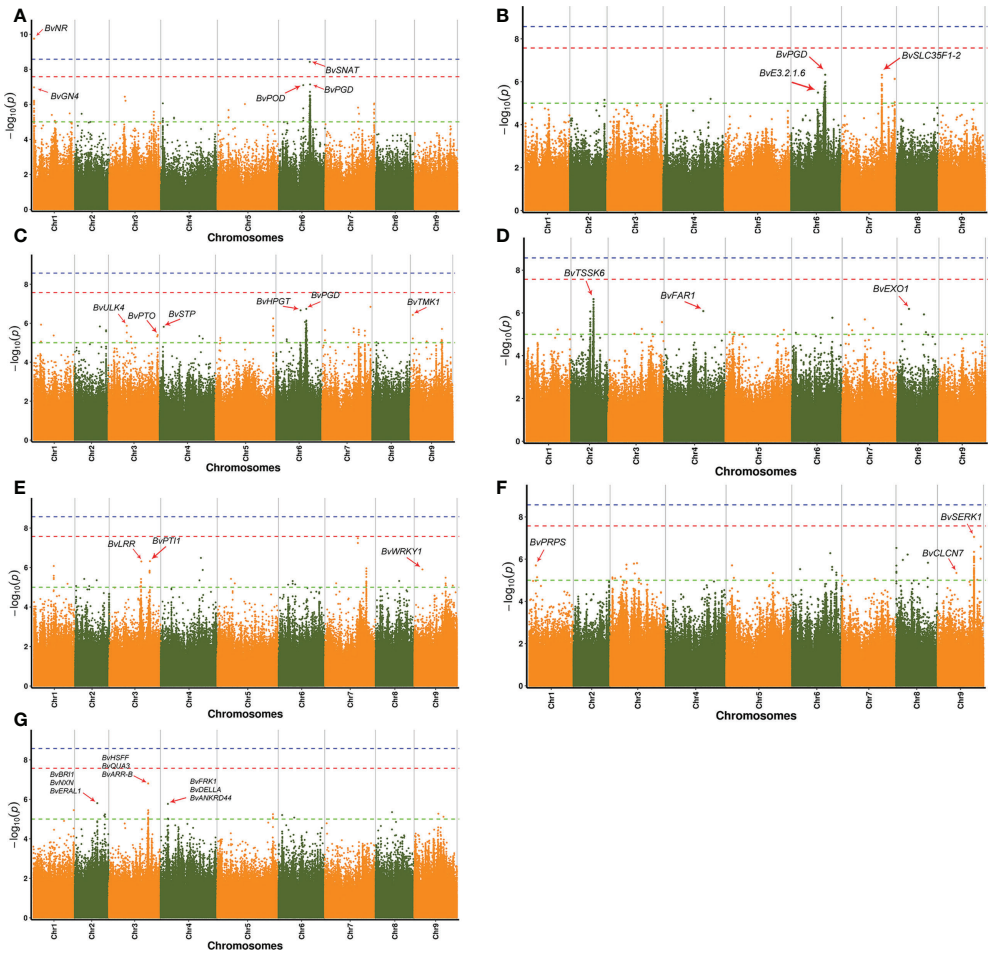


FIGURE 6
Manhattan plots for seven agronomic traits of 114 sugar beet lines. **(A)** Sugar content, **(B)** sugar yield per hectare, **(C)** root yield per hectare, **(D)** root rot of sugar beet, **(E)** damping off of sugar beet, **(F)** rhizomania of sugar beet, and **(G)** pollen fertility of sugar beet.

pollen. In this study, galacturan1, 4-alpha-galacturonidase gene expression was up-regulated in MS301 in the MS301 vs. OT302 DEG analysis (Supplementary Table 22). There were multiple copies of the gene, and expression of all copies was up-regulated (IMABv01g025187, IMABv01g025166, IMABv01g025186, IMABv01g025168). However, in the analysis of MS137 *vs.* OT152 DEGs, there was no difference in expression of a galacturan1, 4-alpha-galacturonidase gene, suggesting that the mechanism of male sterility might be diverse. The results suggested that secondary metabolism regulates the expression of male sterility genes. The results also provide a valuable resource to study male sterility related pathways in sugar beet.

Data availability statement

The data presented in the study are deposited in the Genome Sequence Archive (GSA) in National Genomics Data Center, Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, accession number CRA002683. The final chromosome-level genome sequence data reported in this paper have been deposited in the Genome Warehouse (GWH) in National Genomics Data Center under accession number GWHAMMD00000000 that is publicly accessible at https://bigd.big.ac.cn/gwh.

Author contributions

XL: investigation, resources, funding acquisition, supervision, and writing - review and editing. WH: methodology, supervision, funding acquisition, supervision, and writing - review and editing. JF: software, visualization, data curation, and writing – original draft. YL: resources. HZ: investigation, resources, and verification. DC: software and formal analysis. XW: investigation. ZZ: investigation and verification. LW: resources and data curation. PH: resources. BZ: investigation. TX: software and visualization. WZ: investigation. JH: formal analysis. CB: methodology, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Inner Mongolia Autonomous Region "the open competition mechanism to select the best candidates" project entitled "Creation of Elite Beet Germplasm and Breeding of Varieties Suitable for Mechanized Operation" (2022JBGS0029). This work was also funded by China Agriculture Research System of MOF and MARA, (CARA-170104 & CARA-170501).

Acknowledgments

We would like to thank the editor and reviewers for their helpful comments on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1028885/full#supplementary-material

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

An, X., Dong, Z., Tian, Y., Xie, K., Wu, S., Zhu, T., et al. (2019). ZmMs30 encoding a novel GDSL lipase is essential for Male fertility and valuable for hybrid breeding in maize. *Mol. Plant* 12, 343–359. doi: 10.1016/j.molp.2019.01.011

Arumuganathan, K., and Earle, E. D. (1991). Nuclear DNA content of some importantplant species. *Plant Mol. Biol. Rep.* 9, 208–218. doi: 10.1007/BF02672069

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile. DNA.* 6, 11. doi: 10.1186/s13100-015-0041-9

Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2009). Mixtools: An r package for analyzing mixture models. *J. Stat. Software.* 32 (6), 1–29. doi: 10.18637/jss.v032.i06

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Birney, E., and Durbin, R. (2000). Using GeneWise in the drosophila annotation experiment. *Genome Res.* 10, 547–548. doi: 10.1101/gr.10.4.547

Boyd, L. A., Ridout, C., O'Sullivan, D. M., Leach, J. E., and Leung, H. (2013). Plant-pathogen interactions: disease resistance in modern agriculture. *Trends Genet.* 29, 233–240. doi: 10.1016/j.tig.2012.10.011

Cai, D., Kleine, M., Kifle, S., Harloff, H. J., Sandal, N. N., Marcker, K. A., et al. (1997). Positional cloning of a gene for nematode resistance in sugar beet. *Science.* 275, 832–834. doi: 10.1126/science.275.5301.832

Castleden, C. K., Aoki, N., Gillespie, V. J., MacRae, E. A., Quick, W. P., Buchner, P., et al. (2004). Evolution and function of the sucrose-phosphate synthase gene families in wheat and other grasses. *Plant Physiol.* 135, 1753–1764. doi: 10.1104/pp.104.042457

Chao, M. N., Wang, B., Chen, Y., Zhang, J. B., Sun, X. K., and Wang, Q. L. (2020). Identification and expression analysis of sucrose transporter gene family in upland cotton (*Gossy pium hirsutum l.*). *Acta Botanica Boreali-Occidentalia. Sin.* 40, 1303–1312. doi: 10.7606/j.issn.1000-4025

Chen, L. Q., Hou, B. H., Lalonde, S., Takanaga, H., Hartung, M. L., Qu, X. Q., et al. (2010). Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature4.* 68, 527–532. doi: 10.1038/nature09606

Chen, L. Q., Qu, X. Q., Hou, B. H., Sosso, D., Osorio, S., Fernie, A. R., et al. (2012). Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. *Science* 335, 207–211. doi: 10.1126/science.1213351

Chen, L. Y., VanBuren, R., Paris, M., Zhou, H., Zhang, X., Wai, C. M., et al. (2019). The bracteatus pineapple genome and domestication of clonally propagated crops. *Nat. Genet.* 51, 1549–1558. doi: 10.1038/s41588-019-0506-8

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Choi, J. H., Choi, J. J., Bang, C. S., Lee, J. S., Choi, D. W., Cho, H. S., et al. (2009). Changes of sugar composition and related enzyme activities during fruit development of Asian pear cultivars 'Niitaka' and 'Whangkeumbae'. *Hortic. Environ. Biotechnol.* 50, 582–587.

Chong, J., Piron, M. C., Meyer, S., Merdinoglu, D., Bertsch, C., and Mestre, P. (2014). The SWEET family of sugar transporters in grapevine: VvSWEET4 is involved in the interaction with botrytis cinerea. *J. Exp. Bot.* 65, 6589–6601. doi: 10.1093/jxb/eru375

Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., and Wang, L. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Dohm, J. C., Lange, C., Holtgräwe, D., Sörensen, T. R., Borchardt, D., Schulz, B., et al. (2012). Palaeohexaploid ancestry for caryophyllales inferred from extensive gene-based physical and genetic mapping of the sugar beet genome (*Beta vulgaris*). *Plant J.* 70, 528–540. doi: 10.1111/j.1365-313X.2011.04898.x

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, e105. doi: 10.1093/nar/gkn425

Dohm, J. C., Minoche, A. E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., et al. (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 505, 546–549. doi: 10.1038/nature12817

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the aedes aegypti genome using Hi-c yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Eberhard, F. H. (1989). Origin of the 'Weisse schlesische rübe' (white silesian beet) and resynthesis of sugar beet. *Euphytica* 41, 75–80. doi: 10.1007/bf00022414

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Feldmann, K. A. (2001). Cytochrome P450s as genes for crop improvement. *Curr. Opin.Plant. Biol.* 4, 162–167. doi: 10.1016/s1369-5266(00)00154-0

Flanagan, J. U., Cassady, A. I., Schenk, G., Guddat, L. W., and Hume, D. A. (2006). Identification and molecular modeling of a novel, plant-like, human purple acid phosphatase. *Gene* 377, 12–20. doi: 10.1016/j.gene.2006.02.031

Funk, A., Galewski, P., and McGrath, J. M. (2018). Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. *Plant J.* 95, 659–671. doi: 10.1111/tpj.13977

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., et al. (2009). Rfam: updates to the RNA families database. *Nucleic Acids Res.* 37, D136–D140. doi: 10.1093/nar/gkn766

Geng, G., Lv, C., Stevanato, P., Li, R., Liu, H., Yu, L., et al. (2019). Transcriptome analysis of salt-sensitive and tolerant genotypes reveals salt-tolerance metabolic pathways in sugar beet. *Int. J. Mol. Sci.* 20, 5910. doi: 10.3390/ijms20235910

Guan, Y. F., Huang, X. Y., Zhu, J., Gao, J. F., Zhang, H. X., and Yang, Z. N. (2008). RUPTURED POLLEN GRAIN1, a member of the MtN3/saliva gene family, is crucial for exine pattern formation and cell integrity of microspores in arabidopsis. *Plant Physiol.* 147, 852–863. doi: 10.1104/pp.108.118026

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K.Jr., Hannick, L. I., et al. (2003). Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat.Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7

Hasanl, M. S., Wu, X., and Zhang, L. (2015). Performance evaluation of indel calling tools using real short-read data. *Hum. Genomics* 9, 2–14. doi: 10.1186/s40246-015-0042-2

Hébrard, C., Peterson, D. G., Willems, G., Delaunay, A., Jesson, B., Lefèbvre, M., et al. (2016). Epigenomics and bolting tolerance in sugar beet genotypes. *J. Exp. Bot.* 67, 207–225. doi: 10.1093/jxb/erv449

Holtgräwe, D., Sörensen, T. R., Viehöver, P., Schneider, J., Schulz, B., Borchardt, D., et al. (2014). Reliable in silico identification of sequence polymorphisms and their application for extending the genetic map of sugar beet (*Beta vulgaris*). *PLoS One* 9, e110113. doi: 10.1371/journal.pone.0110113

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., et al. (2012). Iterative correction of Hi-c data reveals hallmarks of chromosome organization. *Nat.Methods* 9, 999–1003. doi: 10.1038/nmeth.2148

Jia, Y., Loh, Y. T., Zhou, J., and Martin, G. B. (1997). Alleles of pto and fen occur in bacterial speck-susceptible and fenthion-insensitive tomato cultivars and encode active protein kinases. *Plant Cell* 9, 61–73. doi: 10.1105/tpc.9.1.61

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat.Genet.* 42, 348–354. doi: 10.1038/ng.548

Klemens, P. A., Patzke, K., Deitmer, J., Spinner, L., Le Hir, R., Bellini, C., et al. (2013). Overexpression of the vacuolar sugar carrier AtSWEET16 modifies germination, growth, and stress tolerance in arabidopsis. *Plant Physiol.* 163, 1338–1352. doi: 10.1104/pp.113.224972

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Kryvoruchko, I. S., Sinharoy, S., Torres-Jerez, I., Sosso, D., Pislariu, C. I., Guan, D., et al. (2016). MtSWEET11, a nodule-specific sucrose transporter of medicago truncatula. *Plant Physiol.* 171, 554–565. doi: 10.1104/pp.15.01910

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie2. *Nat.Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Larson, R. L., Hill, A. L., Fenwick, A., Kniss, A. R., Hanson, L. E., and Miller, S. D. (2006). Influence of glyphosate on rhizoctonia and fusarium root rot in sugar beet. *Pest Manag.Sci.* 62, 1182–1192. doi: 10.1002/ps.1297

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs withBWA-MEM. *Quantitative Biol.* 5, 26. doi: doi.org/10.48550/arXiv.1303.3997

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinform.* 12, 323. doi: 10.1186/1471-2105-12-323

Lunn, J. E., and MacRae, E. (2003). New complexities in the synthesis of sucrose. *Curr. Opin. Plant Biol.* 6, 208–214. doi: 10.1016/s1369-5266(03)00033-5

Lu, Y. Y., Tang, K., Ren, J., Fuhrman, J. A., Waterman, M. S., and Sun, F. (2017). CAFE: aCcelerated alignment-FrEe sequence analysis. *Nucleic Acids Res.* 45, W554–W559. doi: 10.1093/nar/gkx351

Lv, X., Jin, Y., and Wang, Y. (2018). *De novo* transcriptome assembly and identification of salt-responsive genes in sugar beet m14. *Comput. Biol. Chem.* 75, 1–10. doi: 10.1016/j.compbiolchem

Lv, J. H., Wang, Y. Z., Cheng, R., Wang, G. M., Zhang, S. L., Wu, J., et al. (2018). Genome-wide identification and expression analysis of sucrose synthase (*SUS*) and sucrose phosphate synthase (*SPS*) gene families in pear. *Acta Hortic. Sin.* 45, 421–435. doi: 10.16420/j.issn.0513-353x.2017-0474

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinform.* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Mason, M. G., Mathews, D. E., Argyros, D. A., Maxwell, B. B., Kieber, J. J., Alonso, J. M., et al. (2005). Multiple type-b response regulators mediate cytokinin signal transduction in arabidopsis. *Plant Cell* 17, 3007–3018. doi: 10.1105/tpc.105.035451

Matsuhira, H., Kagami, H., Kurata, M., Kitazaki, K., Matsunaga, M., and Hamaguchi, Y. (2012). Unusual and typical features of a novel restorer-of-fertility gene of sugar beet (*Beta vulgaris l.*). *Genetics* 192, 1347–1358. doi: 10.1534/genetics.112.145409

Miao, Y., Li, H. Y., Shen, J., Wang, J., and Jiang, L. (2011). QUASIMODO 3 (QUA3) is a putative homogalacturonan methyltransferase regulating cell wall biosynthesis in arabidopsis suspension-cultured cells. *J. Exp. Bot.* 62, 5063–5078. doi: 10.1093/jxb/err211

Moriguchi, T., Abe, K., Sanada, T., and Yamaki, S. (1992). Levels and role of sucrose synthase, sucrose-phosphate synthase, and acid invertase in sucrose accumulation in fruit of asian pear. *J.Amer.Soc.Hort* 117, 274–278. doi: 10.21273/JASHS.117.2.274

Mutasa-Göttgens, E. S., Joshi, A., Holmes, H. F., Hedden, P., and Göttgens, B. (2012). A new RNASeq-based reference transcriptome for sugar beet and its application in transcriptome-scale analysis of vernalization and gibberellin responses. *BMC Genom.* 13, 99. doi: 10.1186/1471-2164-13-99

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25, 1335–1337. doi: 10.1093/bioinformatics/btp157

Ono, Y., Asai, K., and Hamada, M. (2013). PBSIM: PacBio reads simulator–toward accurate genome assembly. *Bioinformatics* 29, 119–121. doi: 10.1093/bioinformatics/bts649

Ou, S., and Jiang, N. (2017). LTR_retriever: A highly accurate and sensitive program for identification of long terminal-repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Pin, P. A., Zhang, W., Vogt, S. H., Dally, N., Büttner, B., Schulze-Buxloh, G., et al. (2012). The role of a pseudo-response regulator gene in life cycle adaptation and domestication of beet. *Curr. Biol.* 22, 1095–1101. doi: 10.1016/j.cub.2012.04.007

Puttick, M. N. (2019). MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics* 35, 5321–5322. doi: 10.1093/bioinformatics/btz554

Robatzek, S., and Somssich, I. E. (2002). Targets of AtWRKY6 regulation during plant senescence and pathogen defense. *Genes Dev.* 16, 1139–1149. doi: 10.1101/gad.222702

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., et al. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.* 14, R51. doi: 10.1186/gb-2013-14-5-r51

Sahashi, K., Yamada-Kato, N., Maeda, T., Kito, K., Cha-Um, S., Rai, V., et al. (2019). Expression and functional characterization of sugar beet phosphoethanolamine /phosphocholine phosphatase under salt stress. *Plant Physiol. Biochem.* 142, 211–216. doi: 10.1016/j.plaphy.2019.07.011

Saleh, M. M., Draz, K. A., Mansour, M. A., Hussein, M. A., and Zawrah, M. F. (2011). Controlling the sugar beet fly pegomyia mixta vill. with entomopathogenic nematodes. *Commun. Agric. Appl. Biol. Sci.* 76, 297–305. doi: 10.1007/s10340-009-0253-1

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, W686–W689. doi: 10.1093/nar/gki366

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., et al. (2015). HiC-pro: an optimized and flexible pipeline for Hi-c data processing. *Genome Biol.* 16, 259. doi: 10.1186/s13059-015-0831-x

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Stracke, R., Holtgräwe, D., Schneider, J., Pucker, B., Sörensen, T. R., and Weisshaar, B. (2014). Genome-wide identification and characterisation of R2R3-MYB genes in sugar beet (*Beta vulgaris*). *BMC Plant Biol.* 14, 249. doi: 10.1094/PDIS-10-17-1575-RE

Strausbaugh, C. A., and Eujayl, I. A. (2018). Influence of beet necrotic yellow vein virus and freezing temperatures on sugar beet roots in storage. *Plant Dis.* 102, 932–937. doi: 10.1094/PDIS-10-17-1575-RE

Tao, Y., Cheung, L. S., Li, S., Eom, J. S., Chen, L. Q., Xu, Y., et al. (2015). Structure of a eukaryotic SWEET transporter in a homotrimeric complex. *Nature* 527, 259–263. doi: 10.1038/nature15391

Tarailo-Graovac, M., and Chen, N. (2004). Using repeat masker to identify repetitive elements in genomic sequences.Curr.Protoc. *Bioinformatics* 5, 4. doi: 10.1002/0471250953.bi0410s25

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics.* 12, 2032–2034. doi: 10.1093/bioinformatics/btv098

To, J. P. C., Haberer, G., Ferreira, F. J., Deruère, J., Mason, M. G., Schaller, G. E., et al. (2004). Type-a arabidopsis response regulators are partially redundant negative regulators of cytokinin signaling. *Plant Cell* 16, 658–671. doi: 10.1105/tpc.018978

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963. doi: 10.1371/journal.pone.0112963

Wang, W., Sun, Y. Q., Li, G. L., and Zhang, S. Y. (2019). Genome-wide identification, characterization, and expression patterns of the BZR transcription factor family in sugar beet (*Beta vulgaris l.*). *BMC Plant Biol.* 19, 191. doi: 10.1186/s12870-019-1783-1

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wang, D., Yeats, T. H., Uluisik, S., Rose, J., and Seymour, G. B. (2018). Fruit softening: Revisiting the role of pectin. *Trends Plant Sci.* 23, 302–310. doi: 10.1016/j.tplants.2018.01.006

Wei, X., and Zhang, J. (2014). A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biol. Evol.* 7, 381–390. doi: 10.1093/gbe/evu294

Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., and Brinkman, F. S. (2016). Enhanced annotations and features for comparing thousands of pseudomonas genomes in the pseudomonas genome database. *Nucleic Acids Res.* 44, D646–D653. doi: 10.1093/nar/gkv1227

Wu, G. Q., Li, Z. Q., Cao, H., and Wang, J. L. (2019a). Genome-wide identification and expression analysis of the WRKY genes in sugar beet (*Beta vulgaris l.*) under alkaline stress. *PeerJ.* 7, e7817. doi: 10.7717/peerj.7817

Wu, Y., Li, X., Li, Y., Ma, H., Chi, H., Ma, Y., et al. (2022). Degradation of de-esterified pctin/homogalacturonan by the polygalacturonase GhNSP is necessary for pollen exine formation and male fertility in cotton. *Plant Biotechnol. J.* 20, 1054–1068. doi: 10.1111/pbi.13785

Wu, G. Q., Wang, J. L., and Li, S. J. (2019b). Genome-wide identification of Na^+/H^+ antiporter (NHX) genes in sugar beet (*Beta vulgaris l.*) and their regulated expression under salt stress. *Genes* 10, 401. doi: 10.3390/genes10050401

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., et al. (2013). HTQC: a fast quality control toolkit for illumina sequencing data. *BMC Bioinform.* 14, 33. doi: 10.1186/1471-2105-14-33

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., et al. (2020). The water lily genome and the early evolution of flowering plants. *Nature* 577, 79–84. doi: 10.1038/s41586-019-1852-5

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane saccharum spontaneum l. *Nat. Genet.* 50, 1565–1573. doi: 10.1038/s41588-018-0237-2

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-c data. *Nat. Plants* 5, 833–845. doi: 10.1038/s41477-019-0487-8

Zipfel, C. (2008). Pattern-recognition receptors in plant innate immunity. *Curr. Opin. Immunol.* 20, 10–16. doi: 10.1016/j.coi.2007.11.003

Zou, C., Liu, D., Wu, P., Wang, Y., Gai, Z., Liu, L., et al. (2020). Transcriptome analysis of sugar beet (*Beta vulgaris l.*) in response to alkaline stress. *Plant Mol. Biol.* 102, 645–657. doi: 10.1007/s11103-020-00971-7

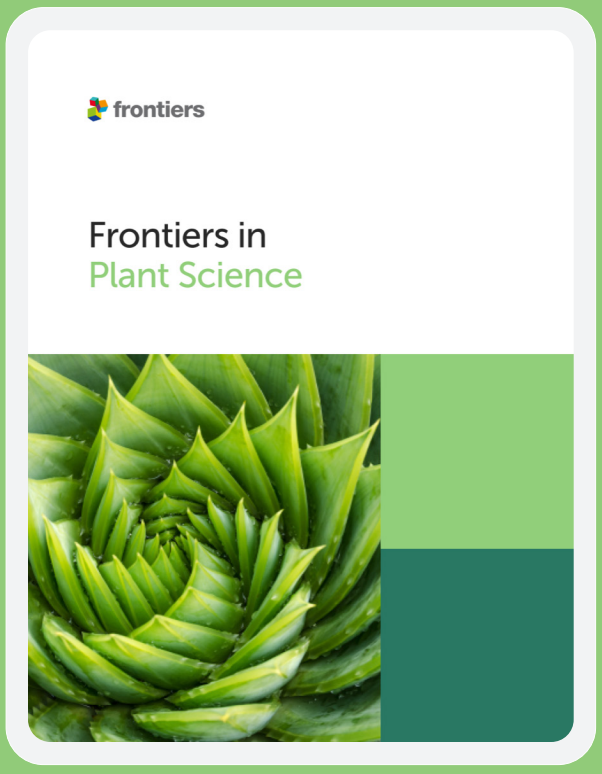
Frontiers in
Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact