The Complete Chloroplast Genome of *Trichopus zeylanicus*, And Phylogenetic Analysis with Dioscoreales

Vadakkemukadiyil Chellappan Biju,* Shidhi P.R., Sheethal Vijayan, Veena S. Rajan, Anu Sasi, Akhil Janardhanan, and Achuthsankar S. Nair

Dep. of Computational Biology and Bioinformatics, Univ. of Kerala, Thiruvananthapuram, Kerala, India.

**ABSTRACT** In this study, we determined the complete sequence of the chloroplast genome of an important, rare, and endangered medicinal plant, *Trichopus zeylanicus*. The analysis of the genome showed that the complete chloroplast genome of *Trichopus zeylanicus* is 153,497 bp in size, and has a quadripartite structure with a large single copy of 81,091 bp and a small single copy of 17,512 bp separated by inverted repeats of 27,447 bp. Sequence analysis revealed that the chloroplast genome encodes 112 unique genes, including 78 protein-coding genes, 30 rRNA genes, and four tRNA genes. We also identified 95 simple sequence repeats and 54 long repeats including 34 forward repeats, seven inverted repeats, nine palindromes, three reverse repeats, and one complementary repeat within the chloroplast genome of *Trichopus zeylanicus*. Whole chloroplast genome comparison with those of other Dioscoreales indicated that the inverted regions are more conserved than large single copy and small single copy regions. In the phylogenetic trees based on complete chloroplast genome and 78 shared chloroplast protein-coding genes in 15 monocot species, including 14 Dioscoreales, *Trichopus zeylanicus* formed a distinct clade. In summary, the first chloroplast genome from the genus *Trichopus* reported in this study gave a better insight into the phylogenetic relationships of different genera within the order Dioscoreales. Moreover, the present data will be a valuable chloroplast genomic resource for population genetics.

**CORE IDEAS**

- We presents the first chloroplast genome from the genus *Trichopus*.
- Comparative analysis revealed that the IR regions are more conserved than the SC regions.
- Highly divergent sequence hot spots were identified, which could be used as molecular markers.
- Phylogenetic analysis gave insight into the evolutionary history of *Trichopus zeylanicus*.

Chloroplasts are important organelles that are responsible for photosynthesis and play a vital role in plant physiology and development in green plants and algae (Raven and Allen, 2003). Chloroplasts have their own DNA, and for most of the land plants, the chloroplast genome exists in circular form with a quadripartite structure comprising of a large single copy (LSC) and a small single copy (SSC) region separated by two inverted repeats (IR; Sugiura, 2003; Wicke et al., 2011). The size of the chloroplast genome varies from 120 to 170 Kb and possessed 60 to 130 genes which are primarily involved in photosynthesis and other metabolic process (Sugiura, 2003; Wicke et al., 2011). The haploid nature, low level of recombination, and maternal inheritance of the chloroplast genome, as well as its low substitution rate compared with the nuclear genome, present chloroplast genomes as valuable sources of genetic information.

Abbreviations: AA, amino acid; APG, Angiosperm Phylogeny Group; AT, adenine-thymine; BI, Bayesian inference; GC, guanine-cytosine; IR, inverted repeat; JBL, junction between IRb and LSC; LSC, large single copy; MCMC, Markov Chain Monte Carlo; ML, maximum likelihood; MP, maximum parsimony; Nc, number of codons; Pi, nucleotide diversity; PCG, protein coding gene; RSCU, relative synonymous codon usage; SMRT, Single Molecule Real Time; SSC, small single copy; ssp., sub species; SSR, simple sequence repeat.


Received 26 Apr. 2019. Accepted 25 Sept. 2019.
*Corresponding author (bijuvcd@gmail.com).

© 2019 The Author(s). This is an open access article distributed under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
resource of genetic markers for resolving the evolutionary history of many closely related plant taxa (Greiner et al., 2015; Smith, 2015; Martin et al., 2005; Freitag et al., 2018).

The genus *Trichopus* contains two species, *Trichopus sempervirens* and *Trichopus zeylanicus* (Sivarajan et al., 1990; Sasikala and RamaSubbu, 2019). Both species have limited geographical distribution. *Trichopus zeylanicus* is distributed in southern part of India, Sri Lanka, the Malay Peninsula, Singapore, and Thailand, whereas *Trichopus sempervirens* is endemic to Madagascar. *Trichopus zeylanicus* is divided into three sub species (ssp.), namely: *Trichopus zeylanicus* ssp. angustifolius, *Trichopus zeylanicus* ssp. *travancoricus*, and *Trichopus ssp. Zeylanicus* (Sivarajan et al., 1990; Sasikala and RamaSubbu, 2019). The sub species present in India is *T. zeylanicus* ssp. *travancoricus* (here onward it is referred to as *T. zeylanicus*) which is endemic to southern part of Western Ghats. It is an important medicinal plant with proven therapeutic properties (Push-pangadan, 1988). It is famous for its traditional use as an instant energy stimulant by local tribal peoples settled in Agastya Hills, the extreme tip of Western Ghats (Push-pangadan, 1988). Experiments using different extracts of *T. zeylanicus* showed that this plant possessed diverse therapeutic properties such as anti-fatigue, anti-inflammatory, anti-oxidant, anti-ulcer, anti-cancer, anti-diabetic, etc. (Biju et al., 2019). A recent survey revealed that the *T. zeylanicus* is facing a high risk of extinction in Agastya Hills and the surrounding area (Sasikala and RamaSubbu, 2019). Therefore, proper conservation methods and propagation techniques have to be developed for the sustainable utilization of this valuable plant.

The genus *Trichopus* is currently assigned to the taxonomical order *Dioscoreales* and family *Dioscoreaceae* according to the Angiosperm Phylogeny Group II system of plant classification (Angiosperm Phylogeny Group, 2003). However, its familial position was a matter of debate because it was moved many times from one family to another. The genus was initially assigned under *Aristolochiaceae* by Lindley (1832) and Thwaites (1864). Bentham and Hooker (1883) and Knuth (1924) treated it under *Dioscoreaceae*, but later Hutchinson (1934) moved it from *Dioscoreaceae* and placed into a monogeneric family *Trichopodaceae* (Brenan, 2007). The separation of *Trichopus from Dioscoreaceae* and its inclusion in its own family *Trichopodaceae* was later supported by many scientists based on cytological and anatomical evidence (Ayensu, 1966; Ramachandran, 1968; Kale and Pai, 1979). Recently, Caddick et al. (2002a, b) placed it in *Dioscoreaceae* again based on a phylogenetic analysis using a combined dataset containing gene sequences of *rbcL*, *atpB*, and nuclear 18S rDNA genes. However, the recircumscription by Caddick et al. (2002b) is unclear because, in general, the partial or short sequences of DNA markers might not contain sufficient information to provide the high resolution necessary to distinguish closely related taxa. The advent of high-throughput sequencing technologies, which is a promising tool to decipher the phylogenetic relationship among closely related species, has now accelerated the sequencing of entire chloroplast genomes of many important plant species (Henry et al., 2014; Freitag et al., 2018).

Single Molecule Real Time technology (SMRT) is a third generation sequencing technology in PacBio systems recently applied in many chloroplast genome sequencing projects (Li et al., 2019; Kang et al., 2018; Lin et al., 2018). An advantage of this technology in chloroplast genome sequencing is the generation of long reads (average length over 10 Kb), which facilitate de novo assembly, especially in the four chloroplast junctions between the IR and single-copy regions (Ferrarini et al., 2013). However, PacBio single pass reads contain high random errors which can be corrected by combining with short Illumina reads (Mahmoud et al., 2019).

In this study, we sequenced the complete chloroplast genome from *T. zeylanicus* using Illumina and PacBio sequencing technologies. The entire chloroplast genome was de novo assembled into a single contig using error-corrected long PacBio reads. We describe the structure, gene content, and organization of the chloroplast genome, and compared with other chloroplast genomes from other *Dioscoreales* to understand its evolutionary history. This study reports the first chloroplast genome sequence from the genus *Trichopus*, and will be a valuable reference genome for future comparative and genetic breeding studies.

**MATERIALS AND METHODS**

**Plant Collection and Genomic DNA Isolation**

The plant *T. zeylanicus* ssp. *travancoricus* was collected from Agastya Hills, Trivandrum, Kerala (India). The voucher specimens were deposited in the Herbarium of Department of Botany, University of Kerala, India. Genomic DNA was isolated from tender leaf tissues using CTAB method (Healey et al., 2014).

**Library Preparation, Sequencing, and Assembly**

1 µg of high quality genomic DNA was fragmented into 300 bp using S220 Focused-ultrasonicator system (Covaris, Woburn, MA). Library preparation was conducted using NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA). Samples were sequenced using an Illumina HiSeq 2500 platform (2 times 100 bp; Illumina, San Diego, CA). For PacBio sequencing, 30 µg of genomic DNA was mechanically sheared using Covaris g-TUBE (Covaris, Woburn, MA). Single molecule real time bell templates with size ranging from 15 to 50 kb were prepared according to the protocols from the manufactures (Pacific Biosciences, Menlo Park, CA). The samples were sequenced on PacBio Sequel instrument with 5 SMRT cells using the P6 polymerase-C4 chemistry combination. All raw reads were processed with AdapterRemoval v2 software to remove adapter sequences short reads (length <50 bp), and to trim low quality bases (Q-value <20; Schubert et al., 2016). The PacBio raw reads (polymerase reads) were filtered by discarding low-quality polymerase reads (Q-value <0.80).
short reads (length <100 bp), short sub reads (length <500 bp), and adapters. Since we performed whole genome sequencing on total DNA isolated from leaf tissues, both Illumina reads and PacBio reads were mapped to the chloroplast reference genomes of 13 species belonging to the taxonomical order Dioscoreales (Supplemental Table S1) to extract the chloroplast reads. Short-reads were aligned to the reference set using Bowtie2 v2.2.6, and long-reads were aligned to the reference set using Blsr v5.1 (Langmead et al., 2009; Chaisson and Tesler, 2012). Prior to the assembly, all filtered PacBio reads were error corrected with Illumina data using Proovread v2.12 (Hackl et al., 2014). All error-corrected PacBio reads were assembled into a single contig using Canu v1.8 (Koren et al., 2017). To estimate the quality and coverage of the assembled genome, we mapped all Illumina and PacBio reads to the assembly using Bowtie and Blsr, respectively.

Genome Annotation
Annotation was performed using GeSeq with chloroplast genome of 13 Dioscoreales (Supplemental Table S1) and the genome map was drawn with OGDRAW (Tillich et al., 2017; Lohse et al., 2013). The repetitive structures, repeat sizes, and locations of forward match, reverse match, palindromic match, and complementary match nucleotide repeat sequences were identified by REPuter v2.74 (Kurtz, 2001). Simple sequence repeats (SSRs) were detected using the MISA-Web with the default parameters (Beier et al., 2017).

Codon Usage
Synonymous codon usage analysis was performed using the program CodonW (version 1.4.4; http://codonw.sourceforge.net, accessed 25 Jan. 2019). In total, 52 protein coding genes (PCGs) of length >300 bp were selected for the analysis. Amino acid (AA) frequency was also calculated and expressed by the percentage of the codons encoding the same amino acid divided by the total codons.

Genome Comparison
For the comparative analysis, chloroplast genomes of 13 Dioscoreales were obtained from Genbank (Supplemental Table S1). The mVISTA program (http://genome.lbl.gov/vista/mvista/submit.shtml, accessed 28 Jan. 2019) with Shuffle-LAGAN mode was used to compare the T. zeylanicus chloroplast genome with chloroplast genomes of other selected species (Frazer et al., 2004). The junction sites of these chloroplast genomes were analyzed and visualized using IRscope online program (Amiryousefi et al., 2018). The nucleotide diversity (Π) and sequence polymorphism of Dioscoreales were analyzed using DnaSP (version 5.10.1; Librado and Rozas, 2009).

Phylogenetic Analysis
For phylogenetic analysis, the chloroplast genomes of 15 species were selected (Supplemental Table S1). The ingroup contains the genomes of 14 Dioscoreales including nine Dioscoreaceae (six Dioscorea species, two Taccaceae species, and one T. zeylanicus), two Burmanniaceae, and three Nartheciaceae. Ananascomosus belonging to the family Bromeliaceae was selected as the out-group. The following datasets were used for the phylogenetic analysis: (i) The complete chloroplast genome and (ii) The dataset comprised sequences of 78 shared PCGs (Supplemental Table S2). The sequences were aligned using the program MAFFT (version 5) and adjusted manually where necessary. For the second data set, the genes were aligned separately and all the alignments were concatenated to form a “supergene” alignment. Three methods were employed to construct phylogenetic trees: Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian Inference (BI). The MP analyses was conducted using PAUP*(version 4.0a165, https://paup.phylsolutions.com, accessed 21 Feb. 2019). Heuristic search was performed with 1000 replicates of random addition sequence, tree bisection-reconnection (TBR) branch swapping, collapse of zero-length branches, and multiple tree option in effect. The ML analyses were performed by MEGA (version 7.0.26) with (GTR+G+I) model selected by jModelTest (version 2.1.1) and tested with 1000 bootstrap replicates (Kumar et al., 2015). The BI was executed with MrBayes program (version 3.2.6; Huelsenbeck and Ronquist, 2001). The number of substitution types was fixed to six. The four by four model was used for substitution, while rates variation across sites was fixed to “invgamma.” Four Markov Chain Monte Carlo (MCMC) chains were run for 10,000 generations, sampling every 10 generations, with the first 100 sampled trees discarded as “burn-in.” Finally, a 50% majority-rule consensus tree was constructed.

Divergence Time Estimation
To estimate the species divergence time, we used the Bayesian method implemented in BEAST (version 1.10.1) with GTR+GAMMA substitution model, with a strict molecular clock and Yule tree prior (Drummond et al., 2006). The analysis was performed on the datasets used in the phylogenetic analysis with previously published calibration times (divergence between D. elephantipes and A. comosus was 120 million years ago; Mennes et al., 2013). The BEAST MCMC simulations were run for 10,000 generations (Whidden and Matsen, 2015). TreeAnnotator (version v1.6.1) software was used to annotate the phylogenetic results generated by BEAST and the FigTree (version v1.3.1) was used to visualize the BEAST maximum clade credibility (MCC) tree.

RESULTS AND DISCUSSION

Trichopus zeylanicus Chloroplast Genome Sequencing, Assembly, and Features
The de novo assembly using error-corrected PacBio reads resulted in a single contig of size 153,497 bp spanning the entire chloroplast genome of T. zeylanicus. The size of the genome is within the previously reported angiosperm chloroplast genomes (Freitag et al., 2018). A circular representation of the chloroplast genome is shown in Fig. 1. The complete chloroplast genome sequence of
The plant genome

The chloroplast genome of *T. zeylanicus* is available at NCBI (Genbank accession number: NC_044084.1). In total, 7,879,378 Illumina reads (2.5 Gb) and 17,639 PacBio reads (63.8 Mb) were mapped to the entire genome representing a coverage of 17,589 times and 415.6 times, respectively. Similar to other angiosperm chloroplast genomes, *T. zeylanicus* chloroplast genome also possessed a quadripartite structure which comprised of a LSC (81,091 bp) and an SSC (17,512 bp) separated by two copies of an inverted repeat (IRA and IRb, 27,447 bp; Fig. 1). The genome contains 112 unique genes including 78 PCGs, 30 tRNAs, and four rRNAs (Table 1). The LSC region possessed 60 PCGs and 21 tRNA genes. The size of the *T. zeylanicus* chloroplast genome and its gene content is similar to other angiosperm chloroplast genomes (Sugiura, 2003). The SSC region contained 11 protein coding genes and one tRNA gene (trnL-UAG). Four rRNAs, eight tRNAs, and seven PCGs were duplicated in the IR region, yielding a total of 131 genes including one pseudo gene (*ycf1*). Among the PCGs, six each contains a single intron (*atpF*, *rpoC1*, *rpl2*, *ndhA*, *ndhB* and *rpsl2*) and two genes contain two introns (*ycf3* and *clpP*; Table 1). The *matK* gene was located within the intron of *trnK-UUU*. The major portion of the *T. zeylanicus* chloroplast genome (63.8%) encompassed gene coding regions including protein coding (52.2%), tRNA (5.3%), and rRNA (6.2%) regions. The intron and intergenic spacer region covered 36.1% of the genome (intron 7%, and intergenic spacer 29%). Overall guanine–cytosine (GC) content of *T. zeylanicus* chloroplast genome was 37.24% which is similar to closely related species (Zhao et al., 2018; Ma et al., 2018). The GC contents of the LSC and SSC regions are 31.5% and 31.4%, respectively, whereas that of the IR regions is 42.2%. Similar to other closely related species, the high GC content in IR region is due to the high GC content in four rRNA genes in this region (54.6%; Zhao et al., 2018).

**Codon Usage Analysis**

To examine codon usage bias, the effective number of codons (Nc) of 52 PCGs was calculated. Number of codons is often used to evaluate the codon bias at the individual gene level, in a range from 20 (extremely biased) to 61 (totally unbiased; Wright, 1990; Sun et al., 2013). The Nc values for each PCG in *T. zeylanicus* are shown in Supplemental Table S3. Our results indicated that the Nc values ranged from 33.85 (*rps18*) to 56.25 (*ycf3*) in all the selected PCGs. Most Nc values were greater than 44 which suggested a weak gene codon bias in the *T. zeylanicus* chloroplast genome. A similar range of Nc value was observed in many plants, for instance in *Forsythia suspense* and *Zizania latifolia* (Wang et al., 2017; Zhang et al., 2016). The *rps18* gene was detected to exist in the most biased codon usage with the lowest mean Nc value of 33.85 (Supplemental Table S3). The 52 unique PCGs comprised 59,785 bp that encoded 19,855 codons. Of these codons, leucine (2064 codons, 10.40% in total) and cysteine (279 codons, 1.40% in total) were found to be the most and least abundant amino acids (Supplemental Table S3), respectively. In many...
Angiosperm chloroplast genomes, leucine and cysteines were reported as the most and least abundant amino acids (Kaila et al., 2016; Chen et al., 2015). The codon usage bias was also measured by calculating the Relative Synonymous Codon Usage (RSCU; Sharp and Cowe, 1991). Relative Synonymous Codon Usage is the observed frequency of a codon divided by the expected frequency. The values close to 1.0 indicate a lack of bias. There were 31 codons with RSCU >1, among these 30 codons were adenine- or uracil-ending codons (Supplemental Table S4). In contrast, the guanine- or cytosine-ending codons mostly exhibited the RSCU values <1, indicating that they are less common in *T. zeylanicus* chloroplast genes (Supplemental Table S4). The most biased stop codon was TAA (Supplemental Table S4). Similar A- or T-ending codon usage bias was also found in other species (Wang et al., 2017; Zhou et al., 2008).

**Simple Sequence Repeats and Long Repeats in *T. zeylanicus* Chloroplast Genome**

Simple sequence repeats are small stretches of DNA sequences in which the same short nucleotide is repeated many times. Because of high levels of polymorphism even among the closely related species, SSRs are widely used as DNA markers in many plant genetic breeding programs (Al-Faifi et al., 2016). In the present study, we have identified 95 SSRs, including 91 mononucleotides, two dinucleotides, four trinucleotides, and six tetra nucleotides in *T. zeylanicus* chloroplast genome (Supplemental Table S5). Among the 91 mononucleotides, 95.6% were consisted of repeats of either A or T. All dinucleotide SSRs were comprised of the combination of A and T. Adenine–thymine (AT) content of other SSRs (tri- and tetra-nucleotides) were also high, consistent with the observation that SSRs in chloroplast genomes are generally composed of short polyadenine (polyA) or polythymine (polyT; dos Reis et al., 2016). Seventy-two SSRs, including 66 mononucleotide and two di-, tri-, and tetra-nucleotide SSRs were located in LSC, and 13 mononucleotide and two tetra-nucleotide SSRs were located in SSC. Six mononucleotides, one tri- and tetra-nucleotide SSRs were located in IR, and thus they were duplicated. By considering the duplicated SSRs in the IR region, total SSRs in *T. zeylanicus* plastid genome were counted as 103. Among all SSRs, 20 SSRs were located in PCGs and one is in tRNA (trnG-UCC; Supplemental Table S5). The *ndhD* and *cemA* genes possessed each tetranucleotide SSRs of length 12 bp. In addition, repeat analysis also identified 54 repeats including 34 forward repeats, seven inverted repeats, nine palindromes, three reverse repeats, and one complementary repeat. The length and distance between these repeats are varied from 16 bp to 50 bp and 1 bp to 47,587 bp, respectively. Among these repeats, 22 forward repeats, seven reverse repeats, four inverted repeats, 15 palindromes, and one complementary repeat were identified in intergenic spacers, five forward repeats, two reverse repeats, and one inverted repeat were in PCGs, and five forward repeats and two inverted repeats were in tRNA genes (Supplemental Table S6). Among the forward repeats, six were in the IR region, and thus were duplicated and counted 65 total repeats in *T. zeylanicus* plastid genome.

**Phylogenetic Analysis, Divergence Time, and Taxonomical Position of *T. zeylanicus***

At present, only three families are included within the order Dioscoreales: Dioscoreaceae, Nartheciaceae, and

---

Table 1. List of genes found in the *T. zeylanicus* chloroplast genome.

<table>
<thead>
<tr>
<th>Category for genes</th>
<th>Group of gene</th>
<th>Name of gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photosynthesis related genes</td>
<td>Rubisco</td>
<td>rbcL</td>
</tr>
<tr>
<td>Photosystem I</td>
<td>psaA, psaB, psaC, psaD, psaI</td>
<td></td>
</tr>
<tr>
<td>Photosystem II</td>
<td>rpoA, rpoB, rpoC1, rpoC2</td>
<td></td>
</tr>
<tr>
<td>ATP synthase</td>
<td>petA, petB, petG, petH, petI, petN</td>
<td></td>
</tr>
<tr>
<td>Cytochrome b and f-complex</td>
<td>ccsA</td>
<td></td>
</tr>
<tr>
<td>NADPH dehydrogenase</td>
<td>ndhA1, ndhB1, ndhC, ndhD, ndhF, ndhG, ndhH, ndhJ, ndhK</td>
<td></td>
</tr>
<tr>
<td>Transcription and translation related genes</td>
<td>transcription</td>
<td>rpoA, rpoB, rpoC1, rpoC2</td>
</tr>
<tr>
<td>ribosomal proteins</td>
<td>rps2, rps3, rps4, rps7, rps8, rps11, rps121, rps14, rps15, rps18, rps19, rpl21, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36</td>
<td></td>
</tr>
<tr>
<td>translation initiation factor</td>
<td>infA</td>
<td></td>
</tr>
<tr>
<td>RNA genes</td>
<td>ribosomal RNA</td>
<td>trnAUGC, trnCGCA, trnDGUC, trnEUUC, trnFGAA, trnGGCC, trnGUCC, trnH6UG, trnI4AU, trnI4GU, trnK1UU, trnL1CA, trnL1UA, trnLUAG, trnMCAU, trnM1GU, trnN1GU, trnP1UG, trnP1UG, trnR1CU, trnS1GU, trnS1GG, trnS2UG, trnT1GU, trnT1UG, trnV1UC, trnWCCA, trnY1GU</td>
</tr>
<tr>
<td>transfer RNA</td>
<td>trn5, trn5.5, trn6, trn23</td>
<td></td>
</tr>
<tr>
<td>Other genes</td>
<td>RNA processing</td>
<td>trnAUGC, trnCGCA, trnDGUC, trnEUUC, trnFGAA, trnGGCC, trnGUCC, trnH6UG, trnI4AU, trnI4GU, trnK1UU, trnL1CA, trnL1UA, trnLUAG, trnMCAU, trnM1GU, trnN1GU, trnP1UG, trnP1UG, trnR1CU, trnS1GU, trnS1GG, trnS2UG, trnT1GU, trnT1UG, trnV1UC, trnWCCA, trnY1GU</td>
</tr>
<tr>
<td>carbon metabolism</td>
<td>trnAUGC, trnCGCA, trnDGUC, trnEUUC, trnFGAA, trnGGCC, trnGUCC, trnH6UG, trnI4AU, trnI4GU, trnK1UU, trnL1CA, trnL1UA, trnLUAG, trnMCAU, trnM1GU, trnN1GU, trnP1UG, trnP1UG, trnR1CU, trnS1GU, trnS1GG, trnS2UG, trnT1GU, trnT1UG, trnV1UC, trnWCCA, trnY1GU</td>
<td></td>
</tr>
<tr>
<td>fatty acid synthesis</td>
<td>accD</td>
<td></td>
</tr>
<tr>
<td>proteolysis</td>
<td>dhfP</td>
<td></td>
</tr>
<tr>
<td>Genes of unknown function</td>
<td>conserved reading frames</td>
<td>ycf1, ycf2</td>
</tr>
</tbody>
</table>

† Genes with one intron.
‡ Genes with two introns.
The Bayesian relaxed molecular clock approach was used to estimate the divergence time between the different lineages within the order Dioscoreales (dos Reis et al., 2016). The analysis was performed on datasets used in the phylogenetic analysis, with previously calibrated divergence time between Poales (A. comosus) and Dioscoreales which suggests that both these lineages had diverged by 120 million years ago (Mennes et al., 2013). Our results show that the first speciation event within Dioscoreales dates back to 98.8 million years ago, where the Clade I (Nartheciaceae) had diverged from the common ancestor of the rest of Dioscoreales. The second speciation event occurred around 91 million years ago, where the common ancestor of members in Clade II (Burmanniaceae) had evolved (Fig. 2). T. zeylanicus (Clade III) was found to be the ancestral species of both Tacca species (Clade IV) and Dioscoreaceae (Clade IV), and its divergence happened around 78.5 million years ago (Fig. 2). Later, Tacca species diverged from Dioscoreaceae approximately 75 million years ago (Fig. 2). Overall, our phylogenetic analysis suggests that T. zeylanicus and Tacca species should be separated from Dioscoreaceae, and should be considered as separate families (Trichopodaceae and Taccaceae, respectively) since they formed distinct clades compared to other Dioscoreaceae species (Fig. 2). Moreover, the estimated divergence time between species in Dioscoreaceae and T. zeylanicus was 81.13 million years ago which is longer than that of the divergence time (77.5 million years ago) between species in Taccaceae and Dioscoreaceae family (Fig. 2), indicating that T. zeylanicus is an ancestor species connecting both Taccaceae and Dioscoreaceae.

Fig. 2. Phylogenetic tree of 14 Dioscoreales chloroplast genomes inferred from Bayesian Inference (BI), Maximum Parsimony (MP), and Maximum Likelihood (ML) based on 78 shared chloroplast genes. Numbers above nodes are support values with Bayesian posterior probabilities values on the left, MP bootstrap values in the middle, and ML bootstrap values on the right. Numbers below the nodes are divergence times of 14 Dioscoreales which was estimated using BEAST software. Values at nodes indicate divergence dates in millions of years.
Comparative Analysis

The comparison of chloroplast genome of three families of Dioscoreales (Dioscoreaceae, Burmanniaceae, and Nartheciaceae) and their evolutionary relationship was previously reported (Zhao et al., 2018; Ma et al., 2018). In this study, we extended the comparative analysis by including 11 Dioscoreales representing five families (Dioscoreaceae, Nartheciaceae, Taccaceae, Burmanniaceae, and Trichopodaceae; Table 2). To estimate the sequence divergence, a multiple sequence alignment between these species was performed using the program mVISTA with an annotated T. zeylanicus chloroplast genome as a reference (Fig. 3). As shown in the previous studies, our comparison also confirmed that the IR region was more conserved than the LSC and SSC regions (Fig. 3; Zhao et al., 2018; Ma et al., 2018).

Furthermore, the DNA polymorphism analysis detected high variable sites in the chloroplast genome among 14 Dioscoreales species (Fig. 4). The average Pi was 0.064. The IR regions showed lower variability than the LSC and SSC regions. The regions that showed highest Pi (>0.15) were ycf1-ndhF-rpl32-trnL-UAG (Pi = 0.18), followed by rpoB-trnC-GCA-petN (Pi = 0.16), and MatK-trnK-UUU-trnQ-UUG-psbK-psbl-trns-GCU-trnG-UCC (Pi = 0.156; Fig. 4). Due to the high variability, these regions could be used to develop molecular markers for plant identification and phylogenetic analysis.

The chloroplast genome size was also slightly variable among these species which was mostly attributed to expansion and contraction of IR regions (Table 2). T. chantrieri had the largest plastid genome (163,007 bp) and D. alata had the smallest (153,161 bp). The length of LSC region was almost conserved in all species, which varied from 81,091 bp (T. zeylanicus) to 85,600 bp (D. rotundata; Table 2). A significant difference in the size of SSC region was noticed which vary from 10,092 bp (T. chantrieri) to 19,038 (D. zingiberensis; Table 2). The smallest sizes of SSC region in T. chantrieri and T. leontopetaloides (both belong to the family Taccaceae) was because of their expanded IRs (33,837 bp; Table 2). The IR was found to be smallest in D. alata (25,464 bp). Overall GC content among these species was 37% except in B. disticha and B. coelestis where it reduced to 35% (Table 2). The gene number was conserved in all species except rps16 gene which was found only in species belonging to the Nartheciaceae family (A. fauriei and A. spicata; Table 2). The junction between IR and LSC and SSC is also varied among these species (Fig. 5). As previously shown, in all selected Dioscoreaceae species, except in D. zingiberensis, the junction between IRb and LSC (JLB) located within rps19 (IRb expanded to 2 to 279 bp of the 5′-end of this gene; Ma et al., 2018). Difference in JLB boundary was also observed within Nartheciaceae species. In A. fauriei, the JLB is located 55 bp upstream of trnH-GUG, whereas in A. spicata and M. luteoviride it is located within rps19 (IR expanded to 103-311 bp).
bp of the 5’-end of this gene). In Burmanniaceae species, IR harbors \textit{rps19}, and the JLB is located within \textit{rpl22} (IR expanded to 39–42 bp of the 5’-end of this gene). In \textit{Tacca} species, JLB is located 145 to 169 bp upstream of \textit{trnH-GUG} and 4 to 8 bp upstream of \textit{rps19}. In \textit{T. zeylanicus} chloroplast genome, similar to Burmanniaceae species, IR harbors \textit{rps19} and the JLB is located 35 bp downstream of \textit{rpl22}. In all selected Dioscoreaceae, Nartheciaceae species, one Burmanniaceae species (\textit{B. coelestis}), and \textit{T. zeylanicus}, the SSC–IR junction was located in 286 to 1675 bp 5’ end of \textit{ycf1}, resulting in the duplication in this region in IR regions. As shown previously, in \textit{B. disticha}, SSC–IR junction is located 351 bp downstream of \textit{ycf1}, resulting in the complete duplication of this gene in IR regions. In \textit{Taccaceae} species, IR expanded further and the SSC–IR junction is located 1242 bp 5’ end of \textit{ndhA}. In these species, the \textit{ycf1}, \textit{rps15}, and \textit{ndhH} are located within IR, and thus they are duplicated. In all species except \textit{T. chantrieri} and \textit{T. leontopetaloides} (\textit{Taccaceae}), 12 genes (excluding partial duplication) were presented in the SSC region (\textit{ndhF, rpl32, trnL-UAG, ccsA, ndhD, psaC, ndhE, ndhG, ndhI, ndhA, ndhH, and rps15}) whereas in \textit{Taccaceae} species, due to IR expansion, only nine genes were located (\textit{ndhF, rpl32, trnL-UAG, ccsA, ndhD, psaC, ndhE, ndhG, and ndhH}). Moreover, in \textit{T. zeylanicus}, \textit{B. disticha}, \textit{B. coelestis}, \textit{A. fauriei}, \textit{Aletris spicata}, and \textit{T. leontopetaloides} chloroplast genomes, the order of genes in SSC region is completely reversed compared with those in \textit{D. rotundata}, \textit{D. alata}, \textit{D. zingiberensis}, and \textit{T. chantrieri}, suggesting a complete inversion of the SSC region in the latter species (Fig. 5). The difference in the orientation of SSC among these species might not be a lineage-specific event, because in many previous studies it has been reported that chloroplast exists in two forms, called inversion isomers, differing only in the relative orientation of their SSC regions (Martin et al., 2013; Palmer, 1983). The coexistence of two orientation-forms of chloroplast genome in \textit{T. zeylanicus} and other \textit{Dioscoreales} needs to be resolved.

**CONCLUSIONS**

In this study we reported the chloroplast genome of \textit{T. zeylanicus}, the first chloroplast genome from the genus \textit{Trichopus}. The size, genome structure, gene number, and gene order of \textit{T. zeylanicus} chloroplast genome was similar to other angiosperm chloroplast genomes. We identified

---

**Table 2. Characteristics of the chloroplast genomes of 11 Dioscoreales.**

<table>
<thead>
<tr>
<th>Genome features</th>
<th>\textit{T. zeylanicus}</th>
<th>\textit{D. lata}</th>
<th>\textit{D. rotundata}</th>
<th>\textit{D. villosa}</th>
<th>\textit{D. zingiberensis}</th>
<th>\textit{B. disticha}</th>
<th>\textit{B. coelestis}</th>
<th>\textit{A. fauriei}</th>
<th>\textit{A. spicata}</th>
<th>\textit{T. chantrieri}</th>
<th>\textit{T. leontopetaloides}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size, bp</td>
<td>153,547</td>
<td>153,161</td>
<td>155,418</td>
<td>153,919.0</td>
<td>153,970</td>
<td>157,480</td>
<td>156,223</td>
<td>154,440</td>
<td>163,007</td>
<td>162,477</td>
<td></td>
</tr>
<tr>
<td>LSC length, bp</td>
<td>81,091</td>
<td>83,414</td>
<td>85,600</td>
<td>83,865</td>
<td>83,950</td>
<td>81,231</td>
<td>81,123</td>
<td>83,510</td>
<td>83,511</td>
<td>84,094</td>
<td></td>
</tr>
<tr>
<td>IR length, bp</td>
<td>27,447</td>
<td>25,466</td>
<td>25,484</td>
<td>25,576</td>
<td>25,491</td>
<td>31,616</td>
<td>27,328</td>
<td>26,374</td>
<td>26,662</td>
<td>33,837</td>
<td>34,140</td>
</tr>
<tr>
<td>SSC length, bp</td>
<td>17,512</td>
<td>18,819</td>
<td>18,850</td>
<td>18,902</td>
<td>19,038</td>
<td>13,017</td>
<td>18,444</td>
<td>18,191</td>
<td>18,164</td>
<td>10,092</td>
<td>10,103</td>
</tr>
<tr>
<td>Total genes</td>
<td>112</td>
<td>112</td>
<td>112</td>
<td>112</td>
<td>112</td>
<td>111</td>
<td>112</td>
<td>112</td>
<td>112</td>
<td>112</td>
<td>112</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>79</td>
<td>79</td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>tRNA genes</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>29</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>rRNA genes</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Overall GC content, %</td>
<td>37.2</td>
<td>37.0</td>
<td>37.2</td>
<td>37.2</td>
<td>37.2</td>
<td>34.9</td>
<td>34.6</td>
<td>37.5</td>
<td>37.5</td>
<td>36.7</td>
<td>36.9</td>
</tr>
<tr>
<td>GC content in LSC, %</td>
<td>31.5</td>
<td>34.8</td>
<td>35.2</td>
<td>35.0</td>
<td>35.1</td>
<td>32.3</td>
<td>31.6</td>
<td>35.4</td>
<td>35.4</td>
<td>34.6</td>
<td>34.9</td>
</tr>
<tr>
<td>GC content in SSC, %</td>
<td>31.4</td>
<td>31.0</td>
<td>30.9</td>
<td>31.2</td>
<td>31.2</td>
<td>28.8</td>
<td>27.6</td>
<td>31.2</td>
<td>31.2</td>
<td>30.5</td>
<td>30.4</td>
</tr>
<tr>
<td>GC content in IR, %</td>
<td>42.2</td>
<td>43.0</td>
<td>42.9</td>
<td>43.0</td>
<td>43.0</td>
<td>39.5</td>
<td>41.6</td>
<td>42.8</td>
<td>42.8</td>
<td>40.2</td>
<td>40.3</td>
</tr>
</tbody>
</table>

1. LSC, large single copy; IR, inverted repeat; SSC, small single copy; GC, guanine–cytosine.
95 SSRs that could be used for population genetics and genetic breeding studies within *Trichopus*. The chloroplast genome comparison to other *Dioscoreales* revealed high conservation within IR regions compared with SSC and LSC regions, indicating that more species DNA barcodes can be developed from these regions for the authentication of *Dioscoreales*. Phylogenetic studies based on 78 shared genes and complete chloroplast genomes of 14 *Dioscoreales* revealed that *T. zeylanicus* formed a separate clade within *Dioscoreales*, supporting the previous opinion of separating *Trichopus* from the family *Dioscoreaceae* and its inclusion within its own family *Trichopodaceae*. Moreover, divergence time analysis indicated that *Trichopus* is an ancestral genus connecting *Taccaceae* and *Dioscoreaceae*.

**Conflict of Interest Disclosure**
The authors declare that there is no conflict of interest.

**Author Contributions**
VCB and ASN designed the study. VCB, SPR, SV, VR, AS, and AJ conducted the experiments and prepared the materials. VCB coordinate the project and wrote the manuscript.

**Funding**
This work is supported by State Inter University Centre of Excellence in Bio Informatics (SIUCEB), Department of Computational Biology and Bioinformatics, University of Kerala.

**ACKNOWLEDGMENTS**
We thank Mr. Anoop PK, Advisory Committee Member; and Kerala Kani, Community Welfare Trust, Kottoor, Trivandrum, Kerala, India, for providing the samples of *T. zeylanicus*. We thank AgriGenome Labs, Cochin, Kerala, India for performing both illumina and PacBio sequencing.

**REFERENCES**


**Fig. 5.** Comparison of the borders of the large single copy (LSC), small single copy (SSC), and inverted repeat (IR) regions among nine *Dioscoreales* chloroplast genomes. JLA, junction between LSC and inverted repeat (IRa); JLB, junction between LSC and IRb; JSA, junction between SSC and IRa; JSB, junction between SSC and IRb.
Thwaites, G.H.K. 1864. Enumeratio Plantarum Zeylaniae: and Enumeration of Ceylon Plants, with descriptions of the new and little-known genera and species, observations on their habitats, uses, native names, etc. Dulau & Co., London.