TÜBİTAK

**Turkish Journal of Botany**

**http://journals.tubitak.gov.tr/botany/**

**Research Article**

Turk J Bot
(2019) 43: 444-457
© TÜBİTAK
doi:10.3906/bot-1811-13

# Comparative plastomes analysis reveals the first infrageneric evolutionary hotspots of *Orthotrichum* s.l. (Orthotrichaceae, Bryophyta)

**Patryk MIZIA[1,*], Kamil MYSZCZYŃSKI[1], Monika ŚLIPIKO[1], Katarzyna KRAWCZYK[1], Vítězslav PLÁŠEK[2], Monika SZCZECIŃSKA[1], Jakub SAWICKI[1]**

[1]Department of Botany and Nature Protection, University of Warmia and Mazury in Olsztyn, Olsztyn, Poland
[2]Department of Biology and Ecology, University of Ostrava, Ostrava, Czech Republic

**Received:** 09.11.2018 • **Accepted/Published Online:** 24.02.2019 • **Final Version:** 08.07.2019

**Abstract:** Plastomes serve as the main resources in the study of plant phylogeny, phylogeography, and barcode development. To date, only 5 complete plastomes of acrocarpous mosses are known. This study analyzes the complete plastome sequence of 5 mosses from Orthotrichaceae: *Nyholmiella obtusifolia* (Brid.) Holmen & E. Warncke; *Orthotrichum rogeri* Brid.; *Stoneobryum bunyaense* D.H. Norris & H. Rob.; *Lewinskya incana* F. Lara, Garilleti & Goffinet; and *Ulota bruchii* Hornsch. The plastomes of all studied species are similar in length (ca. 123 kbp) and have a quadripartite structure with 118 unique genes. A total of 278 simple sequence repeats were found, and the mononucleotide repeats were the most abundant. Dinucleotide, trinucleotide, and tetranucleotide repeats were also observed. Tandem repeats were observed in all studied genomes in similar amounts. A total of 6331 single nucleotide polymorphisms and 779 indels were found in the analyzed genomes. The most variable spacer was located between *rpl14* and *rps8* ($\pi = 0.366$), and *matK* ($\pi = 0.046$) was the most variable of the genes. Phylogenetic reconstruction based on protein coding sequences confirmed the recent division of the tribe.

**Key words:** Bryophyta, acrocarpous mosses, Orthotricheae, chloroplast genome, evolutionary hotspots, repetitive sequences

## 1. Introduction

The tribe Orthotricheae is a widespread, cosmopolitan moss group within the family Orthotrichaceae. Species grow in several biomes, only avoiding wet tropical forests and deserts. It is represented on all continents in a wide range of habitats, with a special prominence in epiphytic communities (Lewinsky, 1993). However, several species are characterized as epilithic plants and grow on rocks and stones.

The plants have orthotropic, mostly branched stems, and they typically exhibit xerophytic morphological features. The characteristic shape of capsules and the peristome structure is coupled with large, campanulate-mitrate calyptrae, and this unique combination of morphological traits makes the tribe Orthotricheae one of the most distinctive and readily recognizable of all moss groups.

Intra- and intergeneric evolutionary relationships were analyzed based on these morphological characters (Vitt, 1971; Lewinsky-Haapasaari and Hedenäs, 1998); however, the molecular data (Goffinet et al., 1998, 2004; Sawicki et al., 2009a, 2010) revealed paraphyly in both of the genera *Orthotrichum* s.l. and *Ulota* s.l., enabling division of the former into genera *Lewinskya*, *Nyholmiella*, *Pulvigera*, and *Orthotrichum* s.s. and the latter into *Plenogemma* and *Ulota* s.s.

The latest taxonomical arrangement based on the combining of morphological and molecular data divides the tribe into 8 genera. In addition to the above-mentioned genera, these include *Stoneobryum* and *Sehnemobryum* (Goffinet et al., 2004; Plášek et al., 2015; Sawicki et al., 2017); however, only 5 of these are not monotypic. The smallest, nonmonotypic genera contain 2 species: *Nyholmiella* (*N. obtusifolia* (Brid.) Holmen & E. Warncke and *N. gymnostoma* (Bruch ex Brid.) Holmen & E. Warncke), and *Stoneobryum* (*S. bunyaense* D.H. Norris & H. Rob. and *S. mirum* D.H. Norris & H. Rob.). The largest genera of the tribe, *Lewinskya*, *Orthotrichum*, and *Ulota*, comprise 70 taxa, 103 taxa, and 69 taxa, respectively (Caparrós et al., 2016; Lara et al., 2016).

Since the 1990s, plastid genomes have served as a valuable source of data for phylogenetics, population genetics, and phylogeography studies (Wu et al., 2018). Comparative plastid genomics serves as a useful resource

* Correspondence: patryk.mizia@uwm.edu.pl

for developing highly variable phylogenetics markers, molecular barcodes (Shaw et al., 2016), and simple sequence repeats (SSRs) used in population and phylogeography studies (Szczecińska and Sawicki, 2015). Complete plastome sequences have also been used as super-barcodes to delimit closely related species where standard single or multilocus barcodes have failed (Krawczyk et al., 2018).

The chloroplast genetic resources for bryophytes are limited compared to vascular plants (Park et al., 2018). To date, complete plastome sequences are known in only 2 hornworts, 11 liverworts, and 46 mosses, mostly belonging to Sphagnales (38 species) (Shaw et al., 2016; Park et al., 2018). Since most studies are focused on a descriptive analysis of obtained genomes, variation at the lower taxonomic level is poorly explored. In this study, we provide complete plastome sequences of the 5 species belonging to the tribe Orthotricheae of the family Orthotrichaceae: *N. obtusifolia*; *Orthotrichum rogeri* Brid.; *S. bunyaense*; *Lewinskya incana* F. Lara, Garilleti & Goffinet; and *Ulota bruchii* Hornsch. The aims of the present study were to establish and characterize the organization of the plastome genome sequences of members from 5 genera of the tribe Orthotricheae, to determine the evolutionary hotspots of the Orthotricheae plastomes for further phylogenetics studies, and to conduct a comparative analysis of previously known moss plastomes.

## 2. Materials and methods

### 2.1. DNA isolation and assembly

#### 2.1.1. DNA isolation

For genomic library construction, we used extracted DNA from previous studies stored at –20 °C (Sawicki et al., 2009a, 2009b, 2010, 2017). The specimens and sequencing result details are given in Table S1.

#### 2.1.2. Library preparation

A genomic library for the MiSeq sequencer was developed using a Nextera XT kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol. The details for library preparation, validation, quantification, and sequencing were given in previous studies of the genus *Orthotrichum* (Sawicki et al., 2017).

#### 2.1.3. Genome assembly

The obtained raw reads were cleaned by removing low-quality and short (<50 bp) reads. The remaining reads were assembled de novo using a Velvet assembler, as implemented in Geneious R8 software (Biomatters, Auckland, New Zealand). A previously designed bioinformatic pipeline (Szczecińska et al., 2014) was implemented for the assembly of Orthotricheae plastomes.

The junctions between single-copy and inverted repeat regions were confirmed by long-read Nanopore sequencing using SpotON Flow Cell Mk1 and the MinION Mk1B device (Oxford Nanopore Technologies, UK). The genomic library was prepared using the Rapid Sequencing Kit R9, Version SQK RAD001, and sequenced using the SQK-MAP005 Nanopore Sequencing Kit (Oxford Nanopore Technologies) according to the manufacturer's protocol. The reads were assembled using Canu software (Koren et al., 2017).

The assembled chloroplast genomes were annotated using DOGMA (Wyman et al., 2004) and Geneious R 7.1 (Kearse et al., 2012). Annotated chloroplast genome sequences were used to draw gene maps using the OGDRAW tool (Lohse et al., 2007).

### 2.2. Simple sequence repeats and tandem repeats detection

Monomeric-to-tetrameric microsatellite repeats were identified with MSATCOMMANDER 1.0.8 software (Faircloth, 2008). The minimum lengths of mono-, di-, tri-, tetra-, penta-, and larger nucleotide repeats were 10, 5, 4, 3, and 2, respectively.

Tandem repeat sequences (forward, reverse, palindromic, and complementary repeats) in the plastomes of *N. obtusifolia*, *O. rogeri*, *S. bunyaense*, *L. incana*, and *U. burchii* were found using REPuter (Kurtz et al., 2001). Tandem repeats were identified using the following conditions: sequence identity ≥90% (Hamming distance: 3) and repeat size exceeding 30 bp.

### 2.3. Plastome variation analyses

The 5 complete plastid genomes were aligned using the MAFFT genome aligner (Katoh and Standley, 2013). The alignment regions containing SSRs were manually corrected. Polymorphism analyses that identify single nucleotide polymorphism (SNP) and insertion or deletion (indel) were performed separately for each CDS, intron, and intergenic spacer using a custom Python script. Additionally, each SNP within a protein coding sequence was defined as a synonymous or nonsynonymous SNP. The obtained data were visualized using Circos software (Krzywinski et al., 2009).

The $\pi$ values of both genetic and intergenetic spacers were calculated using the TASSEL 5.0 program (Bradbury et al., 2007).

### 2.4. Phylogenetics analysis

The sequence identity among all 5 analyzed plastomes was plotted using the mVISTA program (Frazer et al., 2004),with the annotation of *N. obtusifolia* as a reference. Bayesian inference phylogenetic analyses for 11 plastid genomes of bryophytes were estimated for 1,500,000 generations, with a burn-in of 200,000 generations and sampling 1 out of every 100 generations of random trees using MrBayes v. 3.2.6 (Ronquist et al., 2012). Phylogenetic reconstruction was based on a partitioned dataset, including protein-coding genes. The *Takakia lepidozioides*

S. Hatt. & Inoue plastome was used as the root of the generated phylogenetic tree.

## 3. Results and discussion

### 3.1. Structure of the chloroplast genome

The circular plastomes of the 5 studied species (*Nyholmiella obtusifolia*, *Orthotrichum rogeri*, *Stoneobryum bunyaense*, *Lewinskya incana*, and *Ulota bruchii*) have a typical quadripartite structure with 1 small single-copy (SSC), 1 large single-copy (LSC), and 2 inverted repeats (IR) (Table 1).

The total length of the plastomes ranged from 122,895 bp (*N. obtusifolia* NC_026979) to 123,536 bp (*U. bruchii* MK521876).

The plastome genome lengths of all studied species are very similar to most of the plastomes of acrocarpous mosses deposited in the GenBank database: 122,630 bp in *Syntrichia ruralis* (Hedw.) F. Weber & D. Mohr to 127,489 bp in *Tetraphis pellucida* Hedw., although they are smaller than Sphagnales (136,219 bp in *Eosphagnum* sp. (H.A. Crum) A.J. Shaw to 140,117 bp in *Sphagnum portoricense* Hampe) and Takakiales (149,016 bp in *Takakia lepidozioides*).

The difference in length is mostly due to the presence of additional genes in Sphagnales (*rpoA*, *tufA*, and *ccsA*) (Shaw et al., 2016) and Takakiales (*cysA*, *petN*, *rpoA*, and *tufA*).

The plastome (Figure 1) contains 82 protein-coding genes (including the hypothetical chloroplast reading frames: *ycf1*, *2*, *3*, *4*, *12*, and *66*), 32 tRNA, and 4 rRNA (Table 2). The *rps12* gene is divided into 2 independent transcription units (*5'-rps12* and *3'-rps12*) whose transcripts are trans-spliced. The gene structure and order are almost identical to those of other acrocarpous mosses (Oliver et al., 2010; Bell et al., 2014; Lewis et al., 2016; Park et al., 2018), with the exception of *Physcomitrella patens* (Hedw.) Bruch & Schimp., which contains a large inversion in the LSC region and the functional *petN* gene (Sugiura et al., 2003).

The overall G/C content of the analyzed cpDNA is 28.4%, with the exception of *S. bunyaense* (28.1%). The G/C content is similar to other bryophyte plastomes (28.4% to 29.3%) (Table 1), although it is smaller than that of Sphagnales (36.9%) and Takakiales (37.1%) with a G/C content similar to that of seed plants (~34%–40%) (Yang et al., 2014).

The plastome *ndh* genes encode components of the thylakoid *ndh* complex. This complex acts as an electron feeding valve to adjust the redox level of the cyclic photosynthetic electron transporters.

In most plants, the *ndhC-ndhk* genes overlap by 10 bp at the 3'-end of *ndhC* and the 5'-end of *ndhK* (Martín and Sabater, 2010). In the case of Orthotricheae, 2 insertions, 10 bp and 4 bp long, were observed. Due to those insertions, a premature stop codon appeared in the *ndhK* gene. We assumed that the start codon is shifted, and the *ndhC* and *ndhK* genes do not overlap. This situation is exclusive to Orthotrichaceae when compared to other known bryophyte plastomes (Figure 2).

Introns were observed in 16 unique genes. The largest intron is located in *trnK-UUU*, and it also encodes maturase *matK*. The *matK* gene is required for the group II intron splicing (Hausner et al., 2006). The length of the *matK* gene in *S. bunyaense* (1529 bp) is shorter than in other Orthotrichales species (1557 bp to 1600 bp) (Figure 3A). The difference is caused by a 30-bp deletion involving 10 amino acids (Figure 3B). The deletion is located near the 5'-end of the sequence, thus omitting the reverse transcriptase (RT) and X domain located near the 3'-end of the *matK* gene (Mohr et al., 1993; Duffy et al., 2008). The X domain shows high similarity to the conserved functional domain found in mitochondrial group intron

**Table 1.** Size and GC content of the plastome genomes and their regions.

| Type | *Nyholmiella obtusifolia* | *Orthotrichum rogeri* | *Stoneobryum bunyaense* | *Lewinskya incana* | *Ulota bruchii* |
|---|---|---|---|---|---|
| Plastome [bp] | 122,895 | 123,363 | 123,041 | 123,464 | 123,536 |
| LSC [bp] | 85,353 | 85,143 | 84,832 | 85,222 | 85,338 |
| SSC [bp] | 18,414 | 18,472 | 18,439 | 18,432 | 18,446 |
| IR [bp] | 9564 | 9874 | 9885 | 9905 | 9876 |
| %GC | 28.4 | 28.4 | 25.2 | 28.4 | 28.4 |
| %GC in LSC | 25.6 | 25.6 | 25.2 | 25.6 | 25.6 |
| %GC in SSC | 25.0 | 25.0 | 24.7 | 25.0 | 25.0 |
| %GC in IR | 44.2 | 43.5 | 43.5 | 43.6 | 43.6 |

LSC: Large single copy, SSC: small single copy, and IR: inverted region.

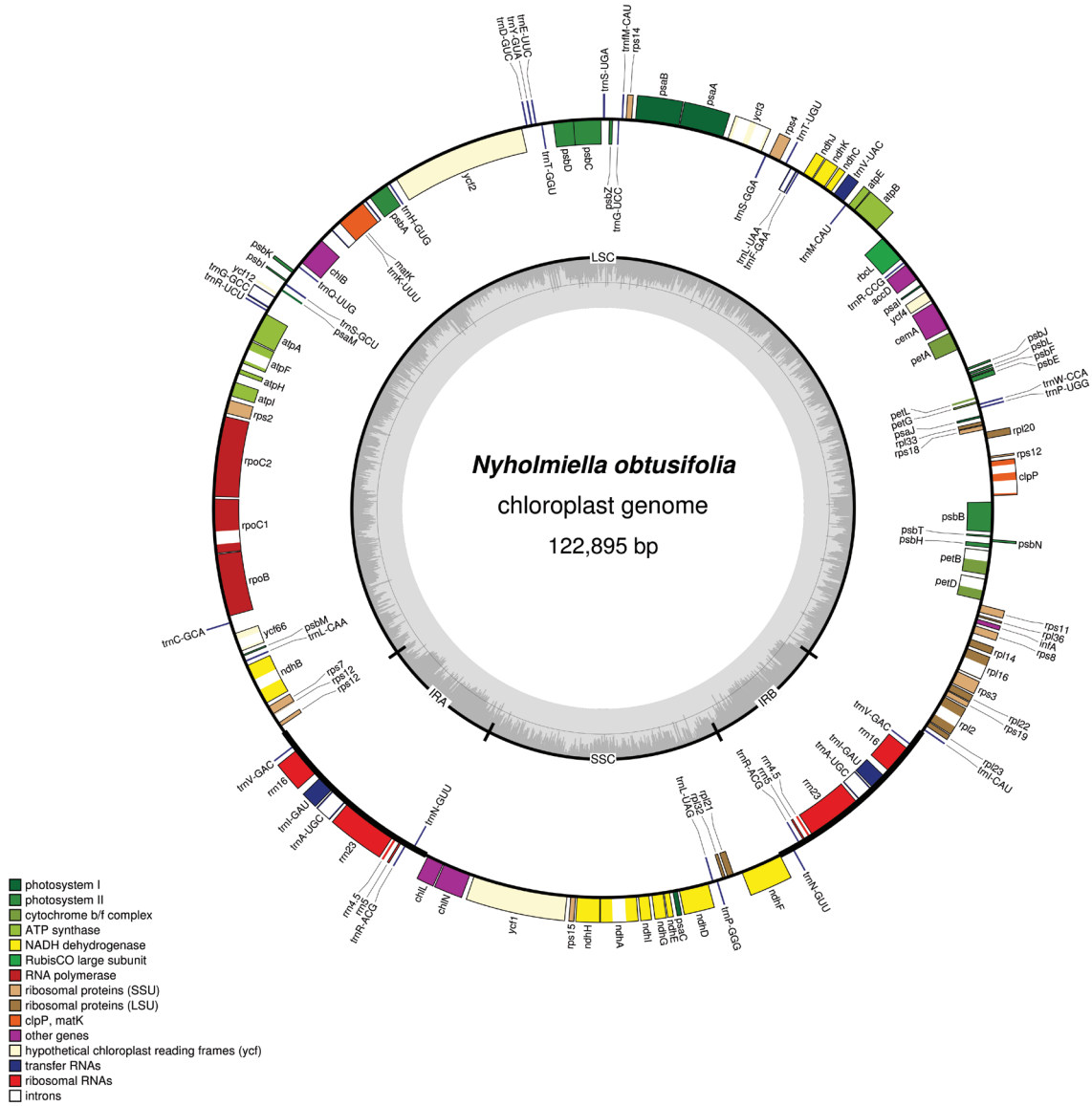


**Figure 1.** Chloroplast genome of *Nyholmiella obtusifolia*. Genes inside the circle are transcribed clockwise, and genes outside are transcribed counterclockwise. The dark gray inner circle corresponds to the GC content and the light gray circle corresponds to the AT content.

II maturases (Sugita et al., 1985). The deletion observed in *S. bunyaense* does not affect the reading frame of the gene or the 2 domains crucial for its role in splicing. It seems plausible that the protein coded by the *matK* gene is still functional, and that it is under the influence of positive selection. Due to its high polymorphism, *matK* is one of the 4 most frequently used plant barcodes (*rbcL*, *matK*, *trnH-psbA*, and *ITS2*) (Kress, 2017).

### 3.2. Repetitive sequences

#### 3.2.1. Microsatellite analysis

A total of 278 microsatellites were identified in the 5 analyzed species: 77 in genes, 48 in introns, and 153 in noncoding regions. Fifty-six microsatellite repeats are shared among all analyzed species. Thirty-three of the shared microsatellites are located in the gene coding region. The longest motif, (T)^22, is observed in noncoding DNA of *N. obtusifolia*, although the length of most motifs varied between 10 and 15 bp. The number of microsatellites per plastome varied from 158 (*S. bunyaense*) to 129 (*L. incana*).

Microsatellites were classified based on the length of the repeat motif into mononucleotide, dinucleotide, trinucleotide, and tetranucleotide repeats. Two hundred sixty-eight mononucleotide repeats (A/T), 7 dinucleotide repeats (AT/TA), 1 trinucleotide repeat (ATT), and 2

**Table 2.** Genes contained within chloroplast genome of *Nyholmiella obtusifolia*, *Orthotrichum rogeri*, *Stoneobryum bunyaense*, *Lewinskya incana*, and *Ulota bruchii.*

| Category | Gene group | Gene name |
|---|---|---|
| Self-replication | Ribosomal RNAs | *rrn4.5*[d], *5*[d], *16*[d], *23*[d] |
| | Transfer RNA genes | *trnA(UGC)*[ad], *C(GCA)*, *D(GUC)*, *E(UUC)*, *F(GAA)*, *fM(CAU)*, *G(GCC)*, *G(UCC)*[a], *H(GUG)*, *I(CAU)*, *I(GAU)*[ad], *K(UUU)*[a], *L(CAA)*, *L(UAA)*[a], *L(UAG)*, *M(CAU)*, *N(GUU)*[d], *P(GGG)*, *P(UGG)*, *Q(UUG)*, *R(ACG)*, *R(CCG)*, *R(UCU)*, *S(GCU)*, *S(GGA)*, *S(UGA)*, *T(GGU)*, *T(UGU)*, *V(GAC)*[d], *V(UAC)*[a], *W(CCA)*, *Y(GUA)* |
| | Small ribosomal protein units | *rps2*, *3*, *4*, *7*, *8*, *11*, *12*[ab], *14*, *15*, *18*, *19* |
| | Large ribosomal protein units | *rpl2*[a], *14*, *16*[a], *20*, *21*, *22*, *23*, *32*, *33*, *36* |
| | RNA polymerase | *rpoB*, *C1*[a], *C2* |
| | Translation initiation factor | *infA* |
| Genes for photosynthesis | Photosystem I | *psaA*, *B*, *C*, *I*, *J*, *M* |
| | Photosystem II | *psbA*, *B*, *C*, *D*, *E*, *F*, *H*, *I*, *J*, *K*, *L*, *M*, *N*, *T*, *Z* |
| | NADH dehydrogenase | *ndhA*[a], *B*[a], *C*, *D*, *E*, *F*, *G*, *H*, *I*, *J*, *K* |
| | Cytochrome b/f complex | *petA*, *B*[a], *D*[a], *G*, *L* |
| | ATP synthase | *atpA*, *B*, *E*, *F*[a], *H*, *I* |
| | Rubisco large subunit | *rbcL* |
| Other genes | Maturase | *matK* |
| | Envelope membrane protein | *cemA* |
| | Acetyl-Co A carboxylase | *accD* |
| | Light-independent protochlorophyllide oxidoreductase subunit | *chlB*, *L*, *N* |
| | Clp protease | *clpP*[c] |
| Conserved open reading frames | Hypothetical genes | *ycf1*, *2*, *3*[c], *4*, *12*, *66*[a] |

[a] Gene containing a single intron, [b] gene divided into 2 independent transcription units, [c] gene containing 2 introns, [d] 2 gene copies due to the IR.


**Figure 2.** Comparison of the *ndhC-ndhK* genes overlapping region in different acrocarpous mosses. First line indicates nucleotides, second translation, and yellow the CDS region. The translation in overlapping region is set for the *ndhK* CDS.

tetranucleotide repeats (ATTT/AAAT) were found in the plastomes of the studied species. This type of motif distribution is common in the plastomes of most plants (Wu et al., 2018), although longer decanucleotide motifs can also be found (Schlautman et al., 2017). The majority of microsatellites were found in the LSC region of the plastome and almost none in the IR region. Microsatellites were present in both coding and noncoding regions of the plastome as well as in the introns (Figure 4). Neither cytosine nor guanine was present in microsatellite repeats. The high A and T ratios observed in plastid genomes in the microsatellites is common among plants (Wu et al., 2018).

The bias among the high A/T ratios observed in microsatellites might contribute to high overall A-T richness in plant plastomes (Yang et al., 2014). The bias might be correlated with easier A-T than G-C changes in the genome (Li et al., 2013).

Microsatellites occur in all plant genomes and provide useful markers for studies of genetic diversity and structure. Nuclear microsatellites are equally dispersed in the genome. In contrast, plastomes have a higher concentration of microsatellites in the noncoding regions (Fajardo et al., 2013). The relatively high concentration of repeats might be responsible for maintaining the stability of plastid genomes (Wu et al., 2018), although plastomes with significantly fewer microsatellites were also observed (Szczecińska and Sawicki, 2015). Chloroplast SSRs have frequently been used for species identification, mainly due to their high reproductivity, high degree of length polymorphism, high polymorphism, abundance, and codominant inheritance, and they are more easily isolated than nuclear microsatellites (Yang et al., 2014).

### 3.2.2. Tandem repeats

A total of 224 tandem repeats (TRs) were found in all studied species: 136 palindromic repeats, 45 forward

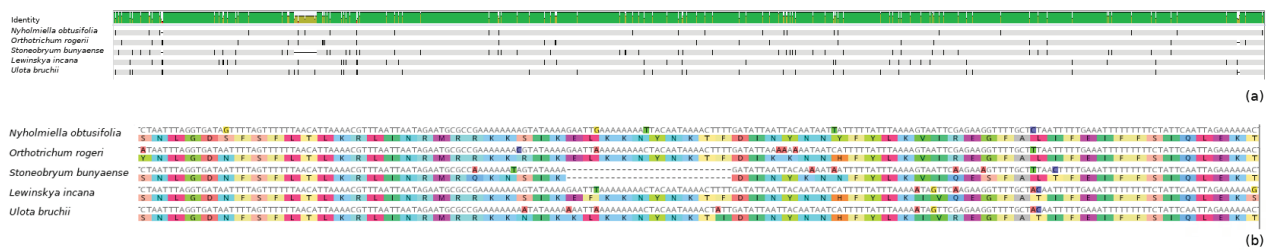


**Figure 3.** The structure and alignment of the matK gene in tribe Orthotricheae: (a) *matK* gene. Green color indicates 100% identity, yellow 50%, and red <50%; gray areas indicate lack of differences, horizontal lines indicates different nucleotides, and vertical lines indicate indels. (b) Indel in the *S. bunyaense matK* gene with translation. N – *N. obtusifolia*, O – *O. rogeri*, S – *S. bunyaense*, L – *L. incana*, and U – *U. bruchii*.

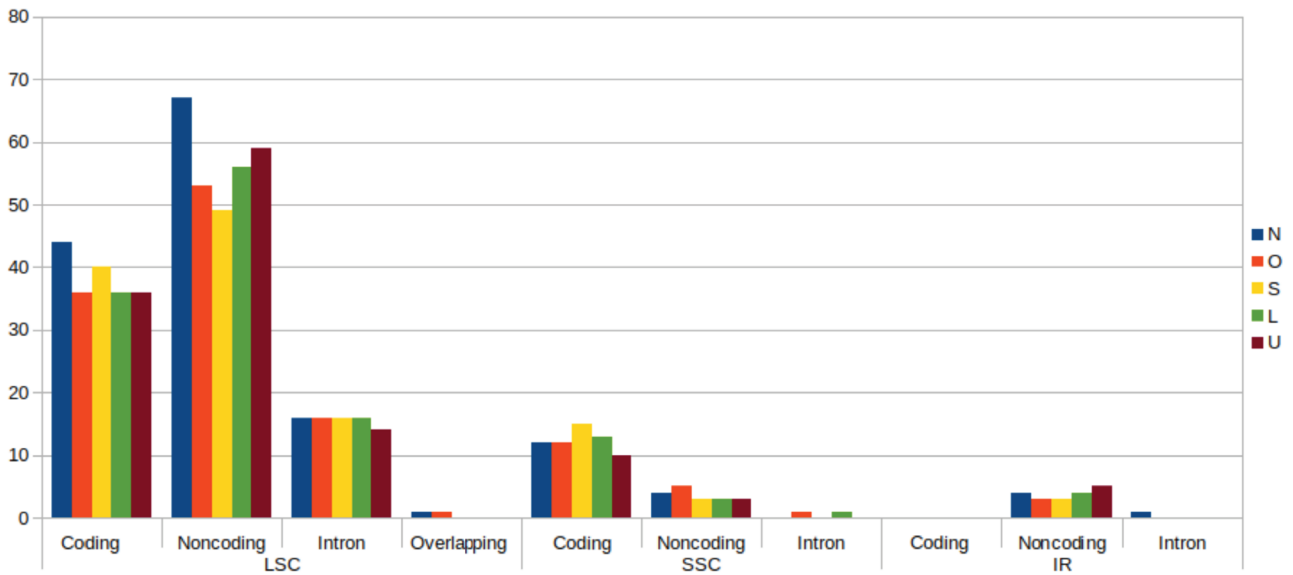


**Figure 4.** Simple sequence repeats (SSRs) in chloroplast genomes of the tribe Orthotricheae. Dispersion of the repeats in the coding/noncoding, intron, and overlapping parts of the plastome LSC, SSC, and IR regions. N – *N. obtusifolia*, O – *O. rogeri*, S – S. *bunyaense*, L – *L. incana*, and U – *U. bruchii*.

repeats, 25 reverse repeats, and 18 complementary repeats.

The number of TRs is very similar among analyzed plastomes and ranges from 43 in *U. bruchii* to 46 in *L. incana* and *S. bunyaense*. Most of the repeats were found in the LSC region. Only 1 (F) repeat was found in the IR region in the *U. bruchii* plastome (Figure 5).

The majority of TRs were 30–40 bp long (25 in *S. bunyaense* and *L. incana* to 27 in *O. rogeri*) (Figure 6).

The longest repeat (138 bp) is a palindrome found in the *S. bunyaense* noncoding region of the LSC. This large repeat is localized almost entirely in the *atpF–atpH* spacer (286 bp) and overlaps 3 nucleotides of the *atpH* gene. TRs were found in coding DNA, noncoding DNA, and introns. Repeats that overlap different groups or have repeats in different groups were marked as mixed. TRs in noncoding regions were the most abundant: 33 in *L. incana*, 32 in *U. bruchii*, and 27 in *N. obtusifolia*, *O. rogeri*, and *S. bunyaense*.

The total length of repeats in comparison to the whole plastome varied from 1.326% (*N. obtusifolia*) to 1.422% in *S. bunyaense*. The plastomes of *S. bunyaense* and *L. incana* (1.414%) are more abundant in repeats than those of *N. obtusifolia*, *O. rogeri*, and *U. bruchii* (1.326%, 1.360%, and 1.361%, respectively).

The analysis of tandem-repeats-to-length ratio in LSC, SSC, and IR showed that the LSC region is the most TR-rich. Only *O. rogeri* showed a different pattern of distribution, with SSC (1.580%) more abundant in TR than the LSC (1.803%). *Ulota bruchii* is the only case in which repeats were found in the IR region (Figure 7).

Palindrome repeats were the most frequent types of repeats in the plastomes of all studied species (Figure 8). They were located in the noncoding regions, mixed

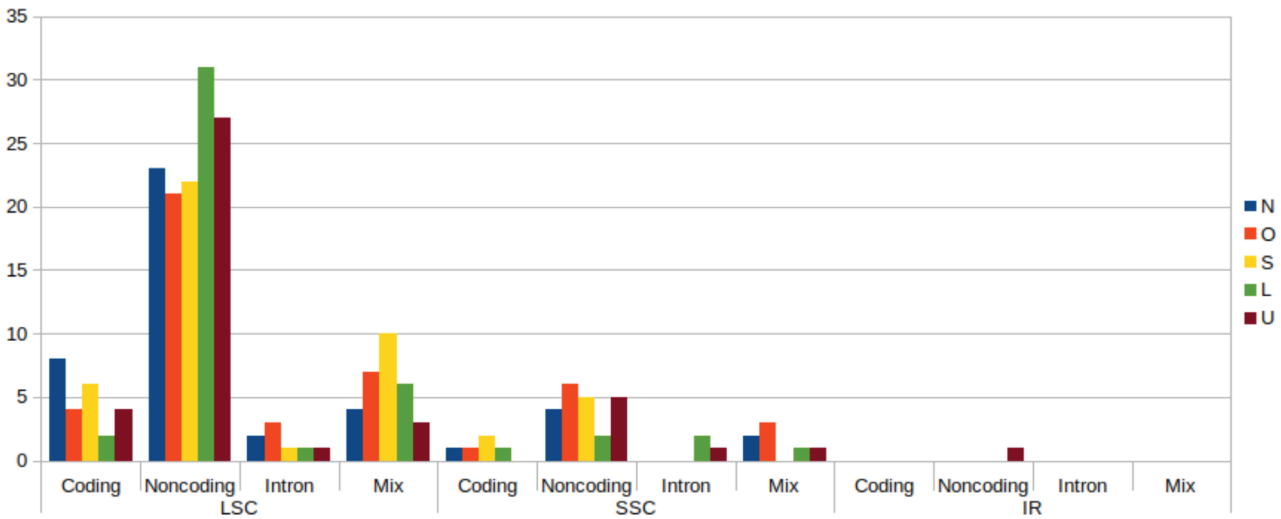


**Figure 5.** Tandem repeats (TRs) in chloroplast genomes of the tribe Orthotricheae. Dispersion of the repeats in the coding/noncoding, intron, and overlapping parts of the plastome LSC, SSC, and IR regions. *N – N. obtusifolia*, *O – O. rogeri*, *S – S. bunyaense*, *L – L. incana*, and *U – U. bruchii*.

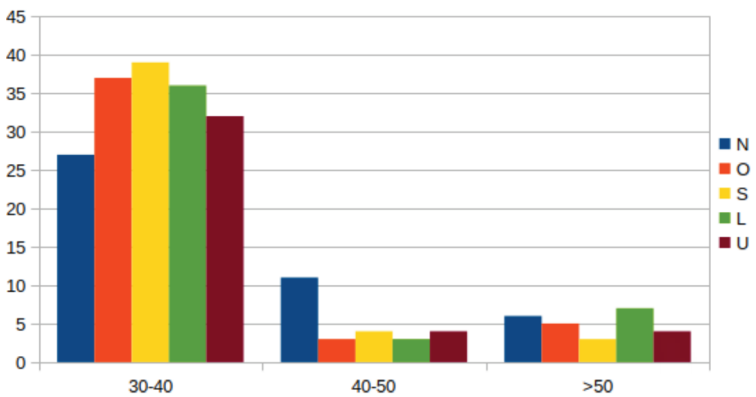


**Figure 6.** Number and length of TRs. *N – N. obtusifolia*, *O – O. rogeri*, *S – S. bunyaense*, *L – L. incana*, and *U – U. bruchii*.

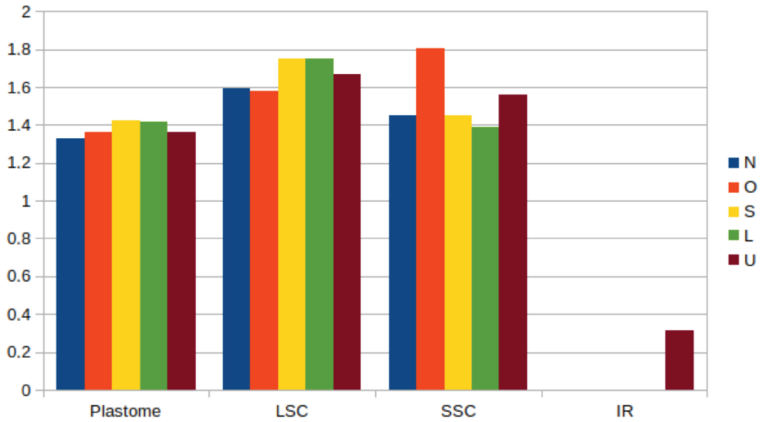


**Figure 7.** The percentage of TRs in different regions of the plastome. *N* – *N. obtusifolia*, *O* – *O. rogeri*, *S* – *S. bunyaense*, *L* – *L. incana*, and *U* – *U. bruchii*.

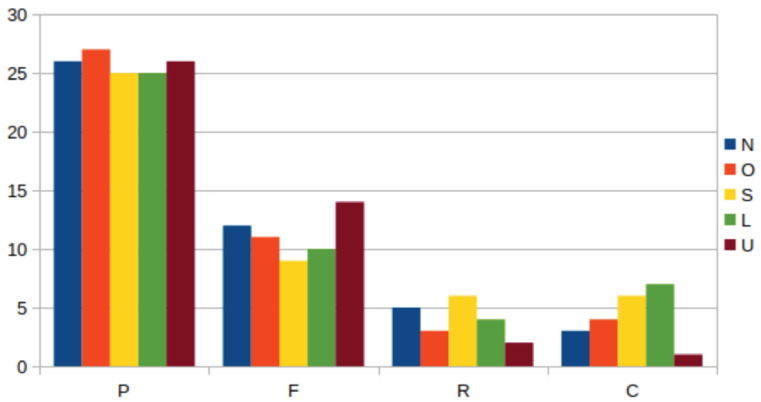


**Figure 8.** Number of different types of TRs. P: Palindrome, F: forward, R: reverse, and C: complementary. *N* – *N. obtusifolia*, *O* – *O. rogeri*, *S* – *S. bunyaense*, *L* – *L. incana*, and *U* – *U. bruchii*.

regions, (except in *N. obtusifolia*), and coding regions (except in *L. incana*), as well as in the introns of *O. rogeri* and *U. bruchii*. Forward repeats were found in coding, noncoding, mixed regions, and introns (except introns in *O. rogeri*). Reverse regions showed similar distribution except in the coding regions of *O. rogeri* and *U. bruchii* and noncoding regions of *S. bunyaense*, *L. incana*, and *U. bruchii*. *Nyholmiella obtusifolia* had no reverse repeats in mixed regions. Complementary regions were the least frequent. They were observed in noncoding regions of all studied species, 4 mixed regions (except in *U. bruchii*), and in the coding region in *O. rogeri*.

The nonrandom distribution of repeats is associated with their function in the plastome (Zhao et al., 2016). Most of the repeats may fold into secondary structures (Jiang et al., 2018). This phenomenon can be observed in the noncoding regions, where they regulate expression of the neighboring genes, increase the plastome plasticity, and take part in intragenic recombination (Zuker, 2003). For repeats in coding sequences and introns, both DNA and mRNA structures can be modified, thus affecting stability (Zuker, 2003; Delihas, 2011).

Although tandem repeats play a crucial role in plastome evolution, it is believed that the accumulation of too many might negatively affect the structure and function of the genome.

### 3.3. Single nucleotide polymorphism (SNPs)

A total of 177 regions were identified in the plastome genome of the 5 analyzed species: 82 protein coding sequences (CDS), 13 introns, and 82 spacers. A total of 6331 SNPs and 779 indels were found in plastomes of the analyzed Orthotricheae. Forty-five indels and 3085 SNPs were identified in coding sequences, and among these, 1291 are nonsynonymous and 1794 are synonymous (Figure 9). The mean $\pi$ value of the plastome was 0.024.

Among noncoding regions longer than 100 bp (for bias elimination), the highest $\pi$ value (0.366) is observed in the *rpl14/rps8* region. This 309-bp-long fragment has 41 SNPs

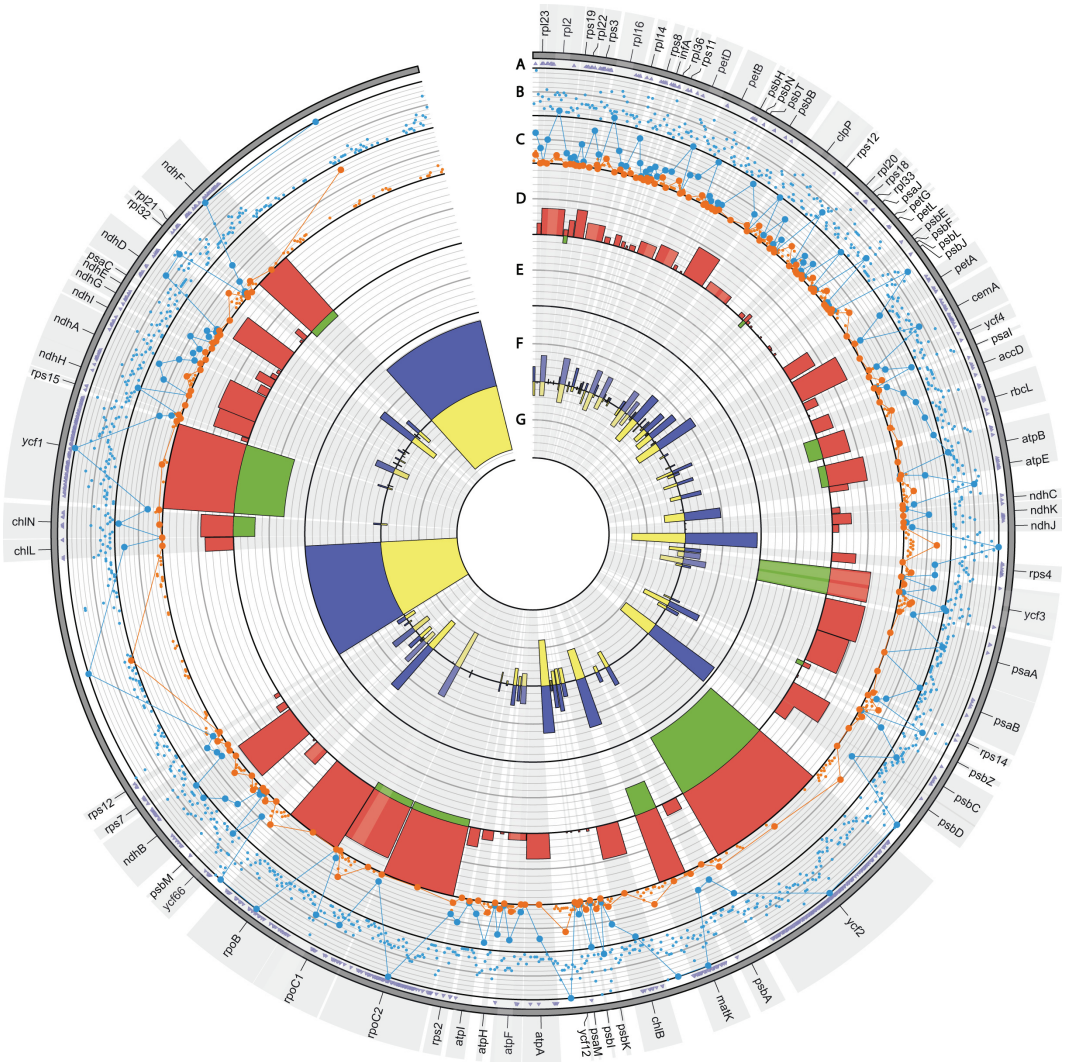


**Figure 9.** Circos graph representing SNP and indel variation among plastomes of 5 species of Orthotrichaceae. Track A shows nonsynonymous SNP occurrence within genes. Tracks B and C represent identified SNPs (light blue spots) and indels (purple spots) per 100 bp, respectively. Track D represents percent of SNPs per CDS length, while track E represents percent of indels per CDS length. Black plot represents π value (maximum value = 0.06) for each CDS. Track F represents percentage of SNPs per noncoding region length, while track G represents percentage of indels per noncoding region length. Black plot represents π value (maximum value = 0.2) for each noncoding region.

and 3 indels. The second most polymorphic fragment (π = 0.301) is located between *rps11* and *petD*. This 209-bp spacer contains 17 SNPs and 9 indels.

The most polymorphic intron is located in the *rpoC1* gene (π = 0.080). This 772-bp-long fragment has 71 SNPs and 25 indels. The second most polymorphic intron (π = 0.078) is the second intron in the *ycf3* gene. The third most polymorphic intron (π = 0.077) is located in the *ycf66* gene. Those 685- and 564-bp-long fragments have 49 and 56 SNPs and 15 and 8 indels, respectively.

Only 2 genes (*psbL* and *psbN*) showed no variation. Among other coding regions, the diversity varied from 0.004 to 0.046. Ten genes (*petG*, *psbZ*, *atpH*, *psaI*, *ycf12*, *psaC*, *psbK*, *psbF*, *rpl33*, and *psbI*) did not have any nonsynonymous substitutions.

*matK* (π = 0.046), *ycf2* (π = 0.041), and *ycf1* (π = 0.036) were the most polymorphic genes in the studied plastomes. The SNP numbers in those genes were 115, 507, and 326, respectively, with 73, 351, and 206 nonsynonymous substitutions.

Both *matK* and *ycf1* are used in phylogenetic studies and barcoding (Dong et al., 2015). Due to high variation, the *ycf2* gene is also considered a valuable candidate for plant barcoding (Celiński et al., 2017). Surprisingly,

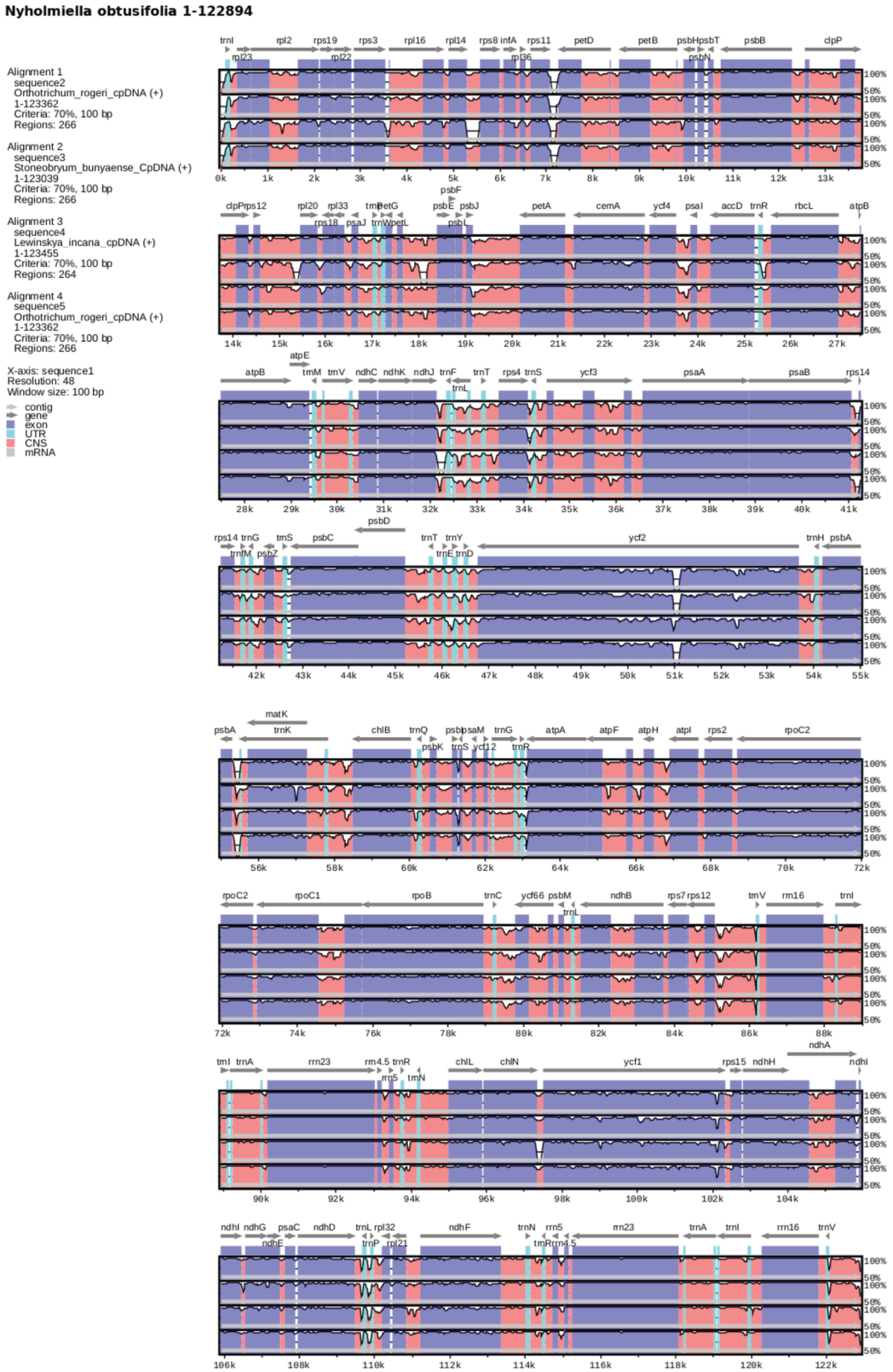


**Figure 10.** Sequence identity plots based on 5 Orthotrichaceae chloroplast genomes, with *Nyholmiella obtusifolia* as a reference. Sequences of chloroplast genomes were aligned and compared using the mVISTA program. Annotated genes are displayed along the top. The vertical scale indicates the percentage identity, ranging from 50% to 100%.

the number of nonsynonymous mutations in the *ycf2* gene of other acrocarpous mosses, such as *S. ruralis* and *Tetraplodon fuegianus* Besch., was smaller (3 and 1, respectively), suggesting that the product of the *ycf2* gene might play a crucial role in their plastids.

Our results confirmed that most mutations accumulated in both spacers and introns. Most of the introns and spacers have a neutral effect on the genome; thus, they lack the evolutionary pressure that purges mutations.

### 3.4. Phylogenetic reconstruction

The sequence identity between all 5 analyzed plastomes was plotted using the mVISTA program, with the annotation of *N. obtusifolia* as a reference (Figure 10). The alignment showed high similarity in both coding and noncoding regions, although variation is higher in the noncoding regions. This result is in accordance with the stability of plastome structure among the studied species.

The alignments of 72 shared CDS regions of 11 Bryophyta species were used to construct the phylogenetic trees using the Bayesian inference (BI) method (Figure 11). The *T. lepidozioides* CDS from the plastome was used as the root of the tree. All studied Orthotricheae species formed a monophyletic clade with significant bootstrap support. The presence of 2 internal clades (A and B) was also observed (100% BI). Clade A consists of 3 species: dioecious *N. obtusifolia*, monoecious *O. rogeri* (98% BI), and dioecious *S. bunyaense*, as a sister lineage to the *Nyholmiella-Orthotrichum* lineage (100% BI). *Orthotrichum rogeri*, as well as *S. bunyaense*, had cryptoporous stomata. *S. bunyaense* remains the only known member of dioecious *Orthotrichum* with cryptoporous stomata (Sawicki et al., 2017). The B clade consists of 2 monoecious species with cryptoporous stomata, *L. incana* and *U. bruchii* (100%), and *S. unicinata* is the closest relative to the studied species (100% BI).

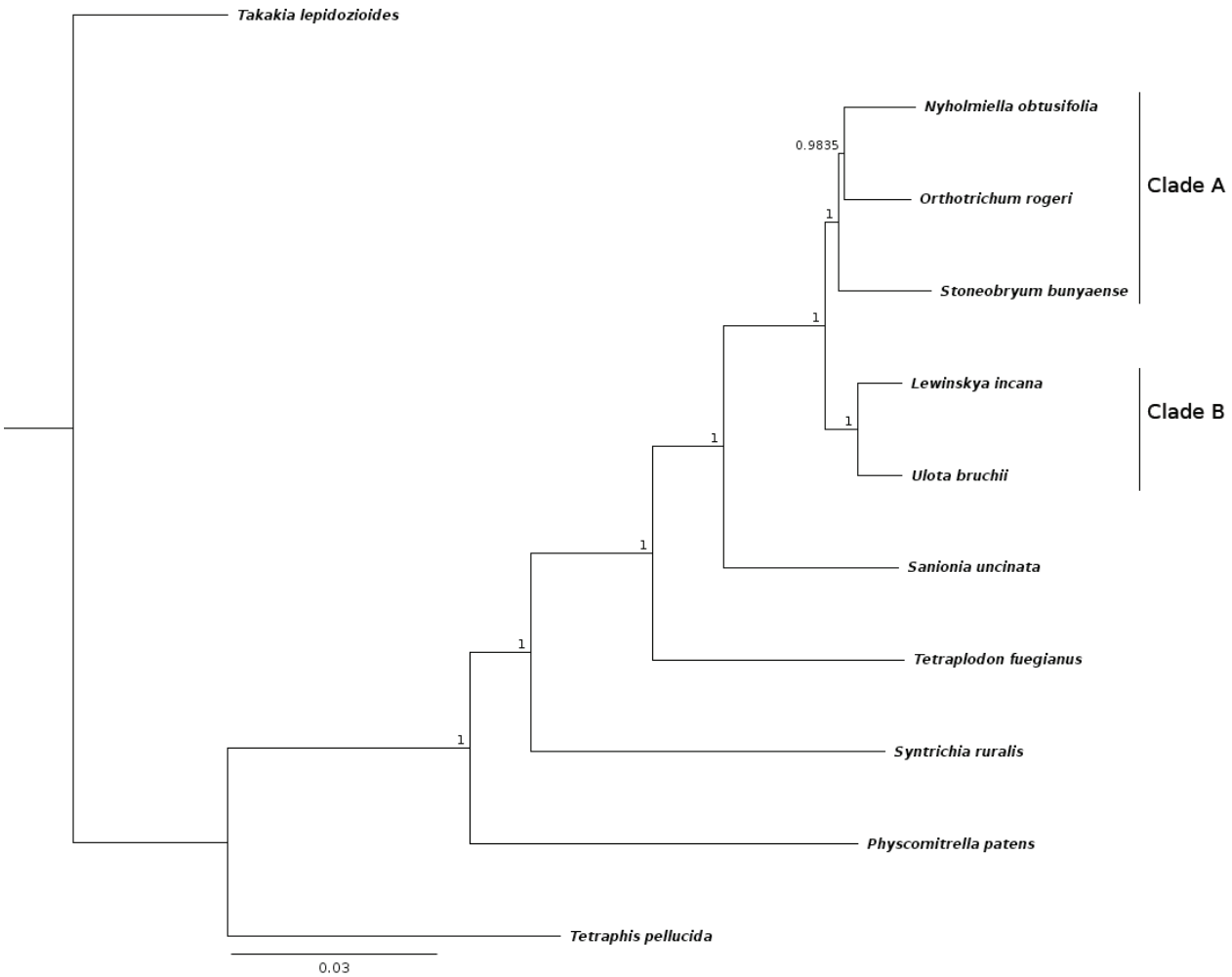


**Figure 11.** Phylogenetic relations among Orthotrichaceae species and other acrocarpous mosses based on the protein-coding regions of shared plastid genes, with *Takakia lepidozioides* as the outgroup.

Our results are in agreement with previous studies indicating a lack of monophyly among monoecious species in clade A, which suggests that it appeared independently.

The main division into 2 clades corresponds to the ML and BI data obtained during the comparison of mitogenomes of the Orthotrichaceae species (Sawicki et al., 2017). The relations in the *Lewinskya-Ulota* clade are also supported. However, relations among *Orthotrichum* s.s., *Nyholmiella*, and *Stoneobryum* were only partially resolved. The ML analysis resolved *Stoneobryum* as the sister clade to *Orthotrichum* s.s., with *Nyholmiella* as the basal genus for the clade. The mitochondrion analysis was in agreement with previous studies indicating that *Nyholmiella* was the sister clade to the *Orthotrichum* species with cryptoporous stomata (Lewinsky, 1993). Surprisingly, our results suggest that the taxa with immersed stomata (*Orthotrichum* and *Stoneobryum*) do not form a single clade; rather, they are sister clades with *Stoneobryum* as the basal genus for the clade, as proposed by Goffinet et al. (2004).

### 3.5. Conclusions

Chloroplast genomes sequenced for this study almost doubled the number of known plastomes of mosses. This study confirms the stability in the gene order and structure of mosses plastomes, although in Orthotricheae, the *ndhC* and *ndhK* genes do not overlap. Comparative analysis enabled identification of evolutionary hotspots, which can be used in further studies on the family Orthotrichaceae. Our study is the first attempt to resolve phylogenetic relations within the tribe Orthotricheae based on the plastome; however, obtained preliminary data are partly incongruent with the mitophylogenomics study (Sawicki et al., 2017). Still, the phylogenomic position of the genus *Stoneobryum* stands in agreement with the observations made by Goffinet et al. (2004). Further analysis, including increased sampling, is required. The microsatellites described in this study are the first example of cpSSR in Bryopsida and the second example in Bryophyta (Shaw et al., 2016). The SSRs and variable genetic and intergenetic sequences and TRs may serve as a source for phylogenetic studies and the development of barcodes.

**Acknowledgments**

This study was made possible through financial support from the Polish Ministry of Science and Higher Education (Iuventus Plus Grant, IP2010-037070) for the seventh author. Bioinformatic analysis was made possible through financial support from the Polish Ministry of Science and Higher Education for the first author (Grant No. 2016/21/B/NZ8/03325) and the second author (Grant No. 2015/19/B/NZ8/03970).

### References

Bell NE, Boore JL, Mishler BD, Hyvönen J (2014). Organellar genomes of the four-toothed moss, *Tetraphis pellucida*. BMC Genomics 15: 383. doi: 10.1186/1471-2164-15-383

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y et al. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635. doi: 10.1093/bioinformatics/btm308

Caparrós R, Lara F, Draper I, Mazimpaka V, Garilleti R (2016). Integrative taxonomy sheds light on an old problem: the *Ulota crispa* complex (Orthotrichaceae, Musci). Botanical Journal of the Linnean Society 180: 427-451. doi: 10.1111/boj.12397

Celiński K, Kijak H, Wojnicka-Półtorak A, Buczkowska-Chmielewska K, Sokołowska J et al. (2017). Effectiveness of the DNA barcoding approach for closely related conifers discrimination: a case study of the *Pinus mugo* complex. Comptes Rendus Biologies 340: 339-348. doi: 10.1016/j.crvi.2017.06.002

Delihas N (2011). Impact of small repeat sequences on bacterial genome Evolution. Genome Biology and Evolution 3: 959-973. doi: 10.1093/gbe/evr077

Dong W, Xu C, Li C, Sun J, Zuo Y et al. (2015). Ycf1, the most promising plastid DNA barcode of land plants. Scientific Reports 5: 8348. doi: 10.1038/srep08348

Duffy AM, Kelchner SA, Wolf PG (2009). Conservation of selection on matK following an ancient loss of its flanking intron. Gene 438: 17-25. doi: 10.1016/j.gene.2009.02.006

Faircloth BC (2008). Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. Molecular Ecology Resources 8: 92-94. doi: 0.1111/j.1471-8286.2007.01884.x

Fajardo D, Senalik D, Ames M, Zhu H, Steffan S et al. (2013). Complete plastid genome sequence of *Vaccinium macrocarpon*: structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genetics & Genomes 9: 489-498. doi: 10.1007/s11295-012-0573-9

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004). VISTA: Computational tools for comparative genomics. Nucleic Acids Research 32: w273-w279. doi: 10.1093/nar/gkh458

Goffinet B, Bayer RJ, Vitt DH (1998). Circumscription and phylogeny of the Orthotrichales (Bryopsida) inferred from rbcL sequence analyses. American Journal of Botany 85: 1324-1337.

Goffinet B, Shaw AJ, Cox CJ, Wickett NJ, Boles S (2004). Phylogenetic inferences in the Orthotrichoideae (Orthotrichaceae: Bryophyta) based on variation in four loci from all genomes. Monographs in Systematic Botany from the Missouri Botanical Garden 98: 270-289.

Hausner G, Olson R, Simon D, Johnson I, Sanders ER et al. (2006). Origin and evolution of the chloroplast trnK (matK) intron: a model for evolution of group II intron RNA structures. Molecular Biology and Evolution 3: 380-391. doi: 10.1093/molbev/msj047

Jiang M, Chen H, He S, Wang L, Chen AJ et al. (2018). Sequencing, characterization, and comparative analyses of the plastome of *Caragana rosea* var. *rosea*. International Journal of Molecular Sciences 9: 1419. doi: 10.3390/ijms19051419

Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30: 772-780. doi: 10.1093/molbev/mst010

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung 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

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH et al. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research 27: 722-736. doi: 10.1101/gr.215087.116

Krawczyk K, Nobis M, Myszczyński K, Klichowska E, Sawicki J ( 2018). Plastid super-barcodes as a tool for species discrimination in feather grasses (Poaceae: *Stipa*). Scientific Reports 8: 1924. doi: 10.1038/s41598-018-20399-w

Kress JW (2017). Plant DNA barcodes: applications today and in the future. Journal of Systematics and Evolution 55: 291-307. doi: 10.1111/jse.12254

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R et al. (2009). Circos: an information aesthetic for comparative genomics. Genome Research 19: 1639-1645. doi: 10.1101/gr.092759.109

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J et al. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research 29: 4633-4642.

Lara F, Garilleti R, Goffinet B, Draper I, Medina R et al. (2016). *Lewinskya*, a new genus to accommodate the phaneroporous and monoicous taxa of *Orthotrichum* (Bryophyta, Orthotrichaceae). Cryptogamie Bryologie 37: 361-382. doi: 10.7872/cryb/v37.iss4.2016.361

Lewinsky J (1993). A synopsis of the genus *Orthotrichum* Hedw. (Musci, Orthotrichaceae). Bryobrothera 2: 1-59.

Lewinsky-Haapasaari J, Hedenäs L (1998). A cladistic analysis of the moss genus *Orthotrichum*. Bryologist 101: 519-555.

Lewis LR, Liu Y, Rozzi R, Goffinet B (2016). Infraspecific variation within and across complete organellar genomes and nuclear ribosomal repeats in a moss. Molecular Phylogenetics and Evolution 96: 195-199. doi: 10.1016/j.ympev.2015.12.005

Li XW, Gao HH, Wang YT, Song JY, Henry R et al. (2013). Complete chloroplast genome sequence of Magnolia grandiflora and comparative analysis with related species. Science China-Life Sciences 56: 189-198. doi: 10.1007/s11427-012-4430-8

Lohse M, Drechsel O, Bock R (2007). OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Current Genetics 52: 267-274. doi: 10.1007/s00294-007-0161-y

Martín M, Sabater B (2010). Plastid ndh genes in plant evolution. Plant Physiology and Biochemistry 48: 636-645. doi: 10.1016/j.plaphy.2010.04.009

Mohr G, Perlman PS, Lambowitz AM (1993). Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Research 21: 4991-4997.

Oliver MJ, Murdock AG, Mishler BD, Kuehl JV, Boore JL et al. (2010). Chloroplast genome sequence of the moss *Tortula ruralis*: gene content, polymorphism and structural arrangement relative to other green plant chloroplast genomes. BMC Genomics 11: 143. doi: 10.1186/1471-2164-11-143

Park M, Park H, Lee H, Lee BH, Lee J (2018). The complete plastome sequence of an Antarctic bryophyte *Sanionia uncinata* (Hedw.) Loeske. International Journal of Molecular Sciences 19: 709. doi: 10.3390/ijms19030709

Plášek V, Sawicki J, Ochyra R, Szczecińska M, Kulik T (2015). New taxonomical arrangement of the traditionally conceived genera *Orthotrichum* and *Ulota* (Orthotrichaceae, Bryophyta). Acta Musei Silesiae, Scientiae Naturales 6: 169-174. doi: 10.1515/cszma-2015-0024

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61: 539-542. doi: 10.1093/sysbio/sys029

Sawicki J, Plášek V, Ochyra R, Szczecińska M, Ślipiko M et al. (2017). Mitogenomic analyses support the recent division of the genus *Orthotrichum* (Orthotrichaceae, Bryophyta). Scientific Reports 7: 4408. doi: 10.1038/s41598-017-04833-z

Sawicki J, Plášek V, Szczecińska M (2009a). Preliminary studies on the phylogeny of the genus Orthotrichum inferred from nuclear ITS sequences. Annales Botanici Fennici 46: 507-515. doi: 10.5735/085.046.0603

Sawicki J, Plášek V, Szczecińska M (2009b). Molecular evidence do not support the current division of *Orthotrichum* subgenus *Gymnoporus*. Plant Systematics and Evolution 279: 125-137. doi: 10.1007/s00606-009-0153-0

Sawicki J, Plášek V, Szczecińska M (2010). Molecular studies resolved *Nyholmiella* (Orthotrichaceae) as separated genus. Journal of Systematics and Evolution 48: 183-194. doi: 10.1111/j.1759-6831.2010.00076.x

Schlautman B, Covarrubias-Pazaran G, Fajardo D, Steffan S, Zalapa Z (2017). Discriminating power of microsatellites in cranberry organelles for taxonomic studies in *Vaccinium* and Ericaceae. Genetic Resources and Crop Evolution 64: 451-466. doi: 10.1007/s10722-016-0371-6

Shaw JA, Devos N, Liu Y, Cox CJ, Goffinet B et al. (2016). Organellar phylogenomics of an emerging model system: *Sphagnum* (peatmoss). Annals of Botany 118: 185-196. doi: 10.1093/aob/mcw086

Sugita M, Shinozaki K, Sugiura M (1985). Tobacco chloroplast tRNA Lys (UUU) gene contains a 2.5-kilobasepair intron: an open reading frame and a conserved boundary sequence in the intron. Proceedings of the National Academy of Sciences of the United States of America 82: 3557-3561.

Sugiura C, Kobayashi Y, Aoki S, Sugita C, Sugita M (2003). Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of rpoA from the chloroplast to the nucleus. Nucleic Acids Research 31: 5324-5331. doi: 10.1093/nar/gkg726

Szczecińska M, Gomolińska A, Szkudlarz P, Sawicki J (2014). Plastid and nuclear genomic resources of a relict and endangered plant species: *Chamaedaphne calyculata* (L.) Moench (Ericaceae). Turkish Journal of Botany 38: 1229-1238. doi:10.3906/bot-1405-80

Szczecińska M, Sawicki J (2015). Genomic resources of three *Pulsatilla* species reveal evolutionary hotspots, species-specific sites and variable plastid structure in the family Ranunculaceae. International Journal of Molecular Sciences 16: 22258-22279. doi: 10.3390/ijms160922258

Vitt DH (1971). The infrageneric evolution, phylogeny, and taxonomy of the genus *Orthotrichum* (Musci) in North America. Nova Hedwigia 21: 683-711.

Wu Y, Liu F, Yang DG, Li W, Zhou XJ et al. (2018). Comparative chloroplast genomics of Gossypium species: insights into repeat sequence variations and phylogeny. Frontiers in Plant Science 9: 376. doi: 10.3389/fpls.2018.00376

Wyman SK, Jansen RK, Boore JL (2004). Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252. doi: 10.1093/bioinformatics/bth352

Yang Y, Yuanye D, Qing L, Jinjian L, Xiwen L et al. (2014). Complete chloroplast genome sequence of poisonous and medicinal plant *Datura stramonium*: organizations and implications for genetic engineering. PLoS ONE 9: e110656. doi: 10.1371/journal.pone.0110656

Zhao CX, Zhu RL, Liu Y (2016). Simple sequence repeats in bryophyte mitochondrial genomes. Mitochondrial DNA Part A 27: 191-197. doi: 10.3109/19401736.2014.880889

Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research 31: 3406-3415.

**Supplementary Table S1.** Sample species, voucher specimens, and sequencing results of species used in this study.

| Species | Voucher | Sequencing results | GenBank ID |
|---|---|---|---|
| *Nyholmiella obtusifolia* | Olsztyn, Poland, OL-M 10109 | 6,234,126 2 × 250 bp pair-end reads | NC_026979 |
| *Orthotrichum rogeri* | Pyrenees Mts. | 3,557,043 2 × 300 bp pair-end reads | KP119739 |
| *Stoneobryum bunyaense* | Kiangaro Mt., Australia, US 00070599 | 3,668,008 2 × 300 bp pair-end reads | MK521875 |
| *Lewinskya incana (Orthotrichum incanum)* | El Rosario, Chile, NYBG 1010874 | 2,643,409 2 × 300 bp pair-end reads | MK521877 |
| *Ulota bruchii* | Czech Republic Herb. OP | 2,841,009 2 × 300 bp pair-end reads | MK521876 |