The complete chloroplast genome of *Colobanthus apetalus* (Labill.) Druce: genome organization and comparison with related species

Piotr Androsiuk *Corresp.*, Jan Paweł Jastrzębski ¹, Łukasz Paukszto ¹, Adam Okorski ², Agnieszka Pszczółkowska ², Katarzyna Joanna Chwedorzewska ³, Justyna Koc ¹, Ryszard Górecki ¹, Irena Gielwanowska ¹

¹ Department of Plant Physiology, Genetics and Biotechnology, University of Warmia and Mazury in Olsztyn, Olsztyn, Poland
² Department of Entomology, Phytopathology and Molecular Diagnostics, University of Warmia and Mazury in Olsztyn, Olsztyn, Poland
³ Department of Antarctic Biology, Institute of Biochemistry and Biophysics Polish Academy of Sciences, Warszawa, Poland

*Corresponding Author: Piotr Androsiuk*
Email address: piotr.androsiuk@uwm.edu.pl

*Colobanthus apetalus* is a member of the genus *Colobanthus*, one of the 86 genera of the large family Caryophyllaceae which groups annual and perennial herbs (rarely shrubs) that are widely distributed around the globe, mainly in the Holarctic. The genus *Colobanthus* consists of 25 species, including *Colobanthus quitensis*, an extremophile plant native to the maritime Antarctic. Complete chloroplast (cp) genomes are useful for phylogenetic studies and species identification. In this study, next-generation sequencing (NGS) was used to identify the cp genome of *C. apetalus*. The complete cp genome of *C. apetalus* has the length of 151228 bp, 36.65% GC content, and a quadripartite structure with a large single copy (LSC) of 83380 bp and a small single copy (SSC) of 17206 bp separated by inverted repeats (IRs) of 25321 bp. The cp genome contains 131 genes, including 112 unique genes and 19 genes which are duplicated in the IRs. The group of 112 unique genes features 73 protein coding genes, 30 tRNA genes, 4 rRNA genes and 5 conserved chloroplast open reading frames (ORFs). A total of 12 forward repeats, 10 palindromic repeats, 5 reverse repeats and 3 complementary repeats were detected. In addition, a simple sequence repeat (SSR) analysis revealed 41 (mono-, di-, tri-, tetra-, penta- and hexanucleotide) SSRs, most of which were AT-rich. A detailed comparison of *C. apetalus* and *C. quitensis* cp genomes revealed identical gene content and order. A phylogenetic tree was built based on the sequences of 76 protein coding genes that are shared by the eleven sequenced representatives of Caryophyllaceae and *C. apetalus*, and it revealed that *C. apetalus* and *C. quitensis* form a clade that is closely related to *Silene* species and *Agrostemma githago*. Moreover, the genus *Silene* appeared as a polymorphic taxon. The results of this study expand our knowledge about the evolution and molecular biology of Caryophyllaceae.
The complete chloroplast genome of *Colobanthus apetalus* (Labill.) Druce: genome organization and comparison with closely related species

Piotr Androsiuk¹*, Jan Paweł Jastrzębski¹, Łukasz Paukszto¹, Adam Okorski², Agnieszka Pszczólkowska², Katarzyna J. Chwedorzewska³, Justyna Koc¹, Ryszard Górecki¹ and Irena Giełwanowska¹

¹Department of Plant Physiology, Genetics and Biotechnology, Faculty of Biology and Biotechnology, University of Warmia and Mazury in Olsztyn, Poland.

²Department of Entomology, Phytopathology and Molecular Diagnostics, Faculty of Environmental Management and Agriculture, University of Warmia and Mazury in Olsztyn, Poland.

³Department of Antarctic Biology, Institute of Biochemistry and Biophysics Polish Academy of Sciences Warsaw, Poland

* corresponding author
e-mail: piotr.androsiuk@uwm.edu.pl
Abstract

Colobanthus apetalus is a member of the genus Colobanthus, one of the 86 genera of the large family Caryophyllaceae which groups annual and perennial herbs (rarely shrubs) that are widely distributed around the globe, mainly in the Holarctic. The genus Colobanthus consists of 25 species, including Colobanthus quitensis, an extremophile plant native to the maritime Antarctic. Complete chloroplast (cp) genomes are useful for phylogenetic studies and species identification. In this study, next-generation sequencing (NGS) was used to identify the cp genome of C. apetalus. The complete cp genome of C. apetalus has the length of 151,228 bp, 36.65% GC content, and a quadripartite structure with a large single copy (LSC) of 83,380 bp and a small single copy (SSC) of 17,206 bp separated by inverted repeats (IRs) of 25,321 bp. The cp genome contains 131 genes, including 112 unique genes and 19 genes which are duplicated in the IRs. The group of 112 unique genes features 73 protein coding genes, 30 tRNA genes, 4 rRNA genes and 5 conserved chloroplast open reading frames (ORFs). A total of 12 forward repeats, 10 palindromic repeats, 5 reverse repeats and 3 complementary repeats were detected. In addition, a simple sequence repeat (SSR) analysis revealed 41 (mono-, di-, tri-, tetra-, penta- and hexanucleotide) SSRs, most of which were AT-rich. A detailed comparison of C. apetalus and C. quitensis cp genomes revealed identical gene content and order. A phylogenetic tree was built based on the sequences of 76 protein coding genes that are shared by the eleven sequenced representatives of Caryophyllaceae and C. apetalus, and it revealed that C. apetalus and C. quitensis form a clade that is closely related to Silene species and Agrostemma githago. Moreover, the genus Silene appeared as a polymorphic taxon. The results of this study expand our knowledge about the evolution and molecular biology of Caryophyllaceae.

1. Introduction

Chloroplasts are organelles whose main function is the photosynthetic fixation of carbon. They contain the complete enzymatic system for energy production in many metabolic pathways, including the biosynthesis of glucose, amino acids and fatty acids (Neuhaus and Emes 2000). Chloroplasts are uniparentally inherited: they are inherited maternally in most angiosperms and gymnosperms (Palmer et al. 1988), but are transmitted in the male line in some gymnosperms (Sears 1980). Their ultrastructure and characteristic genome features indicate that chloroplasts have evolved from free-living cyanobacteria through endosymbiosis (Gray 1989). In a typical...
terrestrial plant, the chloroplast (cp) genome is a circular DNA molecule with a conserved
quadripartite structure composed of a small single copy (SSC), a large single copy (LSC) and
two copies of inverted repeat (IR) regions. The cp genome is the smallest of the plant genomes,
and it ranges from 120 kb to 165 kb in most species (Ma et al. 2017). The variations in size can
be attributed mostly to the expansion, contraction or even loss of IRs as well as changes in the
length of intergenic spacers (Palmer et al. 1988). Due to their compact size, highly conserved
status, uniparental inheritance and haploid nature, cp genomes can be effectively used in studies
of plant taxonomy and evolution and in species identification (Kress et al. 2005; Chase et al.
2007; Parks et al. 2009; Yang et al. 2013).

The genus Colobanthus, a member of the family Caryophyllaceae, consists of 25 species
(Plant List v1.0) of tufted, mainly cushion-forming perennials from the Pacific region,
Australasia to southern South America, sub-Antarctic islands, maritime Antarctic and Hawaiian
mountains. These species are related to Spergula and have very small, narrow and dense leaves
and solitary, petalless, greenish flowers with four to six, but usually five prominent sepals
(Giełwanowska et al. 2011; AGS online Plant Encyclopedia). The available literature is nearly
exclusively devoted to only one species of the genus Colobanthus, namely Colobanthus quitensis
(Kunth) Bartl. Colobanthus quitensis is a species of special concern as the only representative of
Dicotyledoneae in the maritime Antarctic (Skottsberg 1954). Colobanthus quitensis has been
extensively studied to explore the morphological, physiological and biochemical features that
constitute the basis of adaptation to extreme Antarctic conditions (Bravo et al. 2007;
Giełwanowska et al. 2011, 2014; Bascunan-Godoy et al. 2012; Navarrete-Gallegos et al. 2012;
Pastorczyk et al. 2014; Cuba-Diaz et al. 2017). In contrast, very little is known about the genetic
diversity of this species and the genus Colobanthus (Androsiuk et al. 2015; Koc et al. in press).
The complete sequence of C. quitensis cp genome has been recently published (Kang et al.
2016), and it paved the way for more sophisticated genomic studies. However, the significance
of those studies will be limited without information about the genome composition of other
members of the genus Colobanthus. The complete chloroplast genome of eleven
Caryophyllaceae species has been sequenced to date (NCBI - The National Center for
Biotechnology Information), including eight species of the genus Silene (Silene capitata, S.
chalcedonica, S. conica, S. conoidea, S. latifolia, S. noctiflora, S. paradoxa, S. vulgaris), one
species of the genus *Lychnis* (*Lychnis wilfordii*), one species of the genus *Agrostemma* (*Agrostemma githago*) and one species of the genus *Colobanthus* (*Colobanthus quitensis*).

*Colobanthus apetalus* (Labill.) forms tufts with soft and grassy leaves (1.5-3 cm in length), stems of up to 3 cm, and small greenish flowers (5 mm in diameter) that are more obvious than in other *Colobanthus* species. The sepals often have purple borders, and seeds have low rounded papillae. The species has been recorded in New Zealand, south-eastern Australia, including Tasmania (Allan 1961), and in southern regions of South America (Skottsberg 1915).

In this study, the complete cp genome of *Colobanthus apetalus* was sequenced with the use of the Illumina MiSeq sequencing technology and compared with other Caryophyllaceae species, in particular *C. quitensis*.

2. Materials and Methods

2.1. DNA extraction and chloroplast genome sequencing

Fresh leaves of *C. apetalus* were harvested from greenhouse-grown plants (Department of Plant Physiology, Genetics and Biotechnology, University of Warmia and Mazury in Olsztyn). The seeds of *C. apetalus* were collected on the south-eastern shore of Lago Roca, near Lapataia Bay, in the Tierra del Fuego National Park in Argentina. Total genomic DNA was extracted from the fresh tissue of a single plant using the Syngen Plant DNA Mini Kit. The quality of DNA was verified on 1 % (w/v) agarose gel and visualized by staining with 0.5 µg/ml ethidium bromide. The amount and purity of DNA samples were assessed spectrophotometrically.

Genome libraries were prepared from high-quality genomic DNA using the Nextera XT kit (Illumina Inc., San Diego, CA). The prepared libraries were sequenced on the Illumina MiSeq platform (Illumina Inc., San Diego, CA) with a 150 bp paired-end read, version 3.

2.2. Annotation and genome analysis

The quality of the obtained row reads was checked with the FastQC tool. Row reads were trimmed (5 bp of each read end, regions with more than 5% probability of error per base) and mapped to the reference cp genome of *C. quitensis* (NC_028080) using Geneious v.R7 software (Kearse et al. 2012) with default medium-low sensitivity settings. The mapped reads were retrieved from the mapping file and used for *de novo* Velvet preassembly (K-mer – 23-41, low
coverage cut-off – 5, minimum contig length – 300). Preassembled contigs were extended by
mapping row reads using customized settings (minimum sequence overlap of 60 bp and 99%
overlap identity) with 30 iterations steps. Elongated contigs were de novo assembled after each
iteration step to reduce the number and increase the length of sequences and, finally, to create a
circular chloroplast genome. The cp genome was annotated using PlusMapper (Dong et al. 2004)
with manual adjustment. The annotated cp genome was used to draw gene maps with the
OrganellarGenome DRAW tool (Lohse et al. 2007).

2.3. Repeat and SSR analysis

The REPuter program (Kurtz et al. 2001) was used to detect and assess genomic repeats,
including forward, reverse, palindromic and complementary sequences with a minimal length of
30 bp, Hamming distance of 3, and 90% sequence identity. Chloroplast simple sequence repeats
(SSR) or microsatellites were identified in Phobos v.3.3.12 (Mayer 2006-2010) with default
settings for perfect SSRs with motif size of 1 to 6 nucleotide units. Standard thresholds for the
identification of chloroplast SSRs were applied (Sablok et al. 2015), i.e. minimum 12 repeat
units for mononucleotide SSRs, 6 repeat units for dinucleotide SSRs, 4 repeat units for
trinucleotide SSRs, and 3 repeat units for tetra-, penta- and hexanucleotide SSRs. A single IR
region was used to eliminate the influence of IR regions, and redundant results in REPuter were
deleted manually. The cp genomes of C. apetalus and C. quitensis were analyzed simultaneously
to compare their genomic repeats and identify their SSRs. The NC_028080 sequence
downloaded from NCBI represented C. quitensis chloroplast genomic data.

2.4. Intra-individual single nucleotide polymorphism

The “Find Variations/SNPs (Single Nucleotide Polymorphism)” feature in Geneious
software was used to reveal cp genome loci with more than one nucleotide type as candidates for
intraspecific SNPs and to assign major and minor alleles. The candidate SNPs were identified on
the following conditions: (1) the number of aligned reads for each selected locus must be above
30; (2) the percentage of the minor genotype must be above 10%; (3) maximum variant p-value:
10^{-9}; (4) minimum strand-bias p-value: 10^{-9}.

2.5. Synonymous (Ks) and non-synonymous (Ka) substitution rate analysis
The complete sequence of the *C. apetalus* chloroplast genome was compared with the chloroplast genome sequences of all members of the Caryophyllaceae family currently available in GenBank. The coding sequences of the same protein coding genes were extracted and aligned separately using MAFFT v7.310 (Katoh and Standley 2013) to estimate synonymous (Ks) and non-synonymous (Ka) substitution rates. The Ks and Ka in the shared genes were estimated in DnaSP (Rozas et al. 2017). The analysis was performed in two variants: (1) *C. apetalus* was compared with all representatives of the Caryophyllaceae family (Table 1), and (2) the differences in the genes shared by *C. apetalus* and *C. quitensis* were precisely described.

### 2.6. Phylogenetic analysis

Phylogenetic analyses were performed on 76 sequences of protein coding genes shared by *C. apetalus* and eleven species belonging to four genera of the family Caryophyllaceae and *Arabidopsis thaliana* (Sato et al. 1999) as an outgroup. The appropriate sequences were downloaded from the NCBI database (Table 1). The chosen sequences were aligned in MAFFT v7.310. Bayesian Inference (BI) and Maximum-Likelihood (ML) methods were used for genome-wide phylogenetic analyses in MrBayes v.3.2.6 (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) and PhyML 3.0 (Guindon et al., 2010). Before BI and ML analysis, the best fitting substitution model was searched in Mega 7 (Kumar et al. 2016), and the model GTR + G + I was selected. A BI partitioning analysis was carried out to develop a majority rule consensus tree with $1 \times 10^7$ generations using the Markov Chain Monte Carlo (MCMC) method. Tree sampling frequency was 1000 generations. The first 2500 trees were discarded as burn-in, with a random starting tree. The ML analysis was performed in PhyML 3.0 with 1000 bootstrap replicates.

### 3. Results

#### 3.1. Genome organization and content

The Illumina MiSeq platform generated around 4,004,432 high-quality reads (~150 bp in length each, 144.2 nt on average, SD = 19.1; PHRED > 30 for 96% of reads) for the *C. apetalus* cp genome, which were then assembled into a complete sequence based on the reference of *C.*
quitensis cp genome. The full length of the C. apetalus cp genome sequence is 151,228 bp (GenBank accession number MF687919). The genome’s circular, quadripartite structure is composed of SSCs with the length of 17,206 bp and LSC with the length of 83,380 bp, separated by a pair of IR elements (IRa and IRb) with the length of 25,321 bp (Fig. 1). The overall GC content of the C. apetalus cp genome was 36.65%. The entire C. apetalus cp genome contained 131 genes, including 112 unique genes and 19 genes which were duplicated in the IR regions.

The group of 112 unique genes consisted of 73 protein coding genes, 30 transfer RNAs, 4 ribosomal RNAs and 5 conserved chloroplast ORFs of various structure and function (ycf1, ycf2, ycf3, ycf4, ycf68) (Table 2). Most of the 131 genes in the C. apetalus cp genome did not contain introns, 17 contained one intron (atpF, ndhA, ndhB, petB, petD, rpl16, rpoC1, rps16, trnI-GAU, trnA-UGC, trnA-UGC, trnI-GAU, trnK-UUU, trnG-UCC, trnL-UAA, trnV-UAC, ycf3), and only 2 contained 2 introns (rps12 and clpP1). The smallest intron was found in the clpP1 gene (2nd intron, 569 bp), whereas the biggest intron (2 538 bp) – in trnK-UUU. The latter was long enough to contain the coding sequence for another gene, matK. The rps12 gene appeared to be trans-spliced because the exon from the 5’ end was located in LSC, whereas the remaining two exons from the 3’ end of the gene were located in the IR region. Out of the 19 duplicated genes in the IR regions, 7 were tRNA, 4 were rRNA and 8 were protein coding genes, including 3 conserved chloroplast ORFs, ycf2, ycf68 and ycf1, on the border between Ira/IRb and SSC. One ycf1 from the IRb/SSC border was a functional pseudogene. The SSC region contained 11 protein coding genes and one tRNA gene, whereas LSC contained 58 protein coding genes, 22 tRNA genes and 2 conserved chloroplast ORFs (ycf3 and ycf4). The ycf68 gene was classified as a protein coding gene in IR; however, according to some authors (Raubeson et al. 2007), it probably does not encode a protein.

3.2. Repeat sequences and SSRs

The analysis of repeat sequences from the cp genomes of C. apetalus and C. quitensis included forward, reverse, palindromic and complementary repeats. A total of 30 repeated sequences with lengths ranging from 30 to 169 bp and sequence identity greater than 90% (Supplemental Table S1) were identified in C. apetalus in the REPuter application. They included 12 forward repeats, 10 palindromic repeats, 5 reverse repeats and only 3 complementary repeats. Most of the repeated sequencers (23) were dispersed in the intergenic
regions (IGS), and only 7 were localized within genes. The LSC region was most abundant in repeated sequences (24), whereas three such elements were found in SSR and IR each. An analysis of the *C. quitensis* cp genome revealed 15 genomic repeats of similar size, from 30 to 168 bp (Supplemental Table S2). They included 7 forward repeats, 6 palindromic repeats and 2 reverse repeats, distributed almost equally in the intergenic regions (8) and genes (7), mostly in LSC (10) and, to a lesser extent, in IR (4) and SSC (1).

The distribution and type of SSRs and microsatellites was also studied in *C. apetalus* and *C. quitensis* cp genomes. Out of the 41 SSRs identified in *C. apetalus*, 33 (80.5%) were located in the LSC region, 7 (17.1%) in the SSC region, and 1 (2.4%) in the IR regions (Fig 2). The chloroplast SSRs of *C. quitensis* were also identified mainly in the LSC region (79.2%, i.e. 38 SSRs), whereas 8 (16.7 %) and 2 (4.2 %) were located within SSC and IR, respectively. The SSRs can be distributed across three different regions: exons, introns and intergenic spacers. In the analyzed SSRs from the *C. apetalus* cp genome, 28 (68.3%) repeats were located in the intergenic spacer regions, 7 (17.1%) in exons, and 6 (14.6%) in introns. At this point of the analysis, both species shared a similar pattern of variation where 36 (75.0%) of *C. quitensis* chloroplast SSRs were found in the intergenic spacer regions, 5 (10.4%) in exons and 7 (14.6%) in introns. A more detailed analysis of the SSRs in exons revealed that they were located within the coding sequences of 5 genes, *rpoC2*, *ndhF*, *ycf1*, *atpA* and *rrn23*, in both *C. apetalus* and *C. quitensis*. The majority of the microsatellites detected in *C. apetalus* (48.8%) and *C. quitensis* (54.2%) were mononucleotide repeats with one mononucleotide motif (A/T). In di- and trinucleotide SSRs, one microsatellite repeat motif was observed for both species (AT/TA and AAT/TTA, respectively). AAAT/TTTA (25.0%) and AATT/TTAA (25.0%) were the most common tetranucleotide SSRs in both species. In pentanucleotide SSRs, three tandem repeat motifs were observed: AAATT/TTTAA, AAATC/TTTAG and AATCT/TTAGA (Supplemental Table S3 and S4).

### 3.3. Intra-individual single nucleotide polymorphism.

A total of four potential interspecies SNPs were detected in intergenic or intronic regions (Supplemental Table S5). Three were transversions and one was a transition, with minor allele frequency below 12.1%. The detected SNPs were not randomly distributed across the entire cp genome, and they were clustered in LSC region.
3.4. Synonymous and non-synonymous substitution rate analysis

3.4.1. *C. apetalus* vs members of the Caryophyllaceae family

A total of 76 protein coding genes from the chloroplast genome of *C. apetalus* were used to analyze synonymous and non-synonymous substitution rates against 11 members of Caryophyllaceae family. Genes with non-applicable (NA) Ka/Ks ratios were changed to 0. The Ka/Ks ratio for most genes was less than 1, with certain exceptions (Supplemental Table S6a). The Ka/Ks ratio was highest in the *rps7* gene (2.587 for *S. paradoxa* and 2.372 for *S. conica*), whereas it was determined at 1.167 for *S. conoidea* and *S. latifolia*. The Ka/Ks ratio was also high in *rps11* (2.209 for *S. conica*), *rps18* (2.211 for *S. vulgaris*, 1.237 for *S. capitata*), *rps12* (1.416 for *S. paradoxa*, 1.068 for *S. conica*, *S. conoidea* and *S. latifolia*), *rps16* (1.161 for *S. vulgaris*), *clpP1* (1.314 for *S. chalcedonica*) and *ycf2* (1.245 for *S. conica*, 1.208 for *S. paradoxa*, 1.179 for *S. latifolia*, 1.163 *S. conoidea*, and 1.114 for *S. capitata*). A comparative analysis of the genes in each functional group revealed that the substitution rate varied widely across genes, with Ka and Ks values ranging from 0 to 0.757 and from 0 to 0.675, respectively (Supplemental Table S6a). The highest synonymous substitution rate (average Ks = 0.180) was observed in genes with various functions gathered in the ‘other genes’ group, whereas the lowest average Ks was noted in genes related to cytochrome b/f complex, photosystem II and the large subunit of RubisCO (0.106, 0.112 and 0.115, respectively). The highest average non-synonymous (Ka) substitution rate was also observed in the ‘other genes’ group (average Ka = 0.101), whereas the lowest average Ka was noted in photosystem I and the large subunit of RubisCO (0.006 and 0.007, respectively). Based on Ka/Ks values, 69 genes indicative of purifying selection (Ka/Ks < 1) were identified in the analyzed cp genomes. The Ka/Ks ratio was higher than 1.0 in 7 genes in at least one analyzed species, which was indicative of positive selection.

3.4.2 *C. apetalus* vs. *C. quitensis*

More detailed analyses of synonymous and non-synonymous substitution rate were performed for the chloroplast genomes of *C. apetalus* and *C. quitensis* based on 78 protein coding genes shared by the two species. The analysis revealed additional genes (*accD* and *ycf68*)
which were excluded from the previous study because they were not shared by all cp genomes of
the family Caryophyllaceae.

The Ka/Ks ratio for all analyzed genes was less than 1 in the range of 0 to 0.388, with the
maximum value in \textit{atpH}. However, this value resulted from only one synonymous and one non-
synonymous nucleotide substitution in this relatively short (246 nt) sequence. A comparative
study of the genes in each functional group revealed minor variations in the substitution rate
across genes, with Ka and Ks values in the range of 0 to 0.0084 and 0 to 0.0502, respectively
(Supplemental Table S6b). The highest synonymous substitution rate (average Ks = 0.0162) was
observed for the genes related to the large ribosome subunit, whereas a complete absence of
synonymous substitutions was noted in the genes related to the cytochrome b/f complex and the
large subunit of RubisCO (average Ks = 0). The highest non-synonymous (Ka) substitution rate
was observed for the gene of the large subunit of RubisCO (Ka = 0.0028). The average Ka =
0.0021 was noted for the genes with various functions in the ‘other genes’ group, and the average
Ka = 0.0019 – for the genes related to the large ribosome subunit. The lowest average non-
synonymous (Ka) substitution rate was observed for the genes related to photosystem I and
photosystem II (0.0001 for both groups). The noted Ka/Ks values were indicative of purifying
selection in all studied genes.

In general, 36 genes had identical sequences (Ka = 0, Ks = 0), whereas the remaining 42
genes showed 99% identity. In 18 analyzed genes, the differences at the nucleotide level were
not transmitted to amino acid sequences (Ka = 0). The highest number of nucleotide substitutions
was found in \textit{ycfl} and \textit{rpoC2} genes (28 and 15, respectively) which were also characterized by
the highest number of non-synonymous amino acid substitutions (15 and 7, respectively). A
pairwise alignment of \textit{rps16} in \textit{C. apetalus} and \textit{C. quitensis} demonstrated that in the \textit{C. apetalus}
sequence, the substitution of one nucleotide (T→C) in position 201 bp changed the STOP codon
(TGA) to CGA encoding arginine (R). The above gave rise to a new “long” allele for \textit{rps16} with
8 additional amino acids (RFKQIKFN). A comparison of cp genomes in \textit{C. apetalus} and \textit{C.
quitenis} revealed that the nucleotide substitution in \textit{rpoB} and \textit{rpl16} genes was accompanied by
indel polymorphism. In addition to the substitution of three nucleotides, which was responsible
for the shift of two amino acids, the \textit{rpoB} gene encoding the \(\beta\) subunit of RNA polymerase
appeared to be shorter in \textit{C. apetalus} due to the absence of one amino acid (glutamine deletion in
position 636 of the amino acid sequence). In the \textit{rpl16} gene in \textit{C. apetalus}, the insertion of
thymine in position 396 bp shifted the reading frame and, as a result, the protein sequence was
elongated by three additional amino acids (isoleucine, glycine and threonine) at the 3’-end. A
synonymous nucleotide substitution was observed simultaneously within that sequence.

Generally the comparison of cp genomes in *C. apetalus* and *C. quitensis* revealed little or
no variation in the genes associated with photosynthesis, whereas information protein genes,
including RNA polymerase subunits and ribosomal proteins, underwent greater changes (Fig. 3).
Other plastid genes characterized by rapid structural evolution were also identified: *ycf1*, which
is essential for cell survival, but whose function has not yet been fully elucidated (Drescher et al.
2000), and *accD* which is involved in fatty acid biosynthesis.

### 3.5. A comparative analysis of the organization of *C. apetalus* and *C. quitensis* chloroplast
\textbf{genomes}

A comparison of *C. apetalus* and *C. quitensis* cp genomes revealed considerable
similarities in genome composition and size between the species. Both species have the same
gene content and order, but *C. apetalus* has a slightly smaller genome (Table 3). A detailed
analysis of protein-coding sequences revealed low levels of differentiation and demonstrated that
the nucleotide substitution was almost exclusively responsible for the variation, with scarce indel
representation. Therefore, the differences in the length and organization of intergenic spacers
were responsible for the observed variations in the size of the cp genome.

The LSC/IRb/SSC/IRa boundary regions of the *C. apetalus* cp genome were also
compared to the corresponding regions of *C. quitensis*, and they were found in the same
positions within the same genomic elements (Fig. 4). The border between IRa and SSC was
located within the coding region of *ycf1*, placing 1819 bp from the 5’ -end within the IRa, which
resulted in the \(\Psi_{ycf1}\) pseudogene in the IRb region. The IRb/SSC border was localized within
*ndhF*, and a 45 bp section from its 3’-end overlapped \(\Psi_{ycf1}\) within IRb. The IRb/LSC border
was located within *rps19*, leaving 160 bp from the 5’-end within the IRb region, which resulted
in the \(\Psi_{rps19}\) pseudogene in IRa. The *trnH* gene was located in the LSC region next to the
IRa/LSC border.

### 3.6. Phylogenetic analysis
Both BI and ML methods generated phylogenetic trees with uniform topology. The BI tree revealed that only two nodes had posterior probability values below 1 (0.9978 and 0.9737, respectively). The reconstructed phylogeny revealed that *C. apetalus* and *C. quitensis* formed a monophyletic group that was closely related to a group of *Silene* species and *Agrostemma githago* which formed a solitary branch (Fig. 5). The genus *Silene* appeared as a polymorphic taxon with the two main clades: clade I containing *S. conica*, *S. conoidea*, *S. noctiflora*, *S. latifolia*, *S. vulgaris* and *S. capitate*, and clade II where the only member of the genus *Lychnis* (*L. wilfordii*) was merged with *S. chalcedonica* and *S. paradoxa*.

4. Discussion

The *Colobanthus apetalus* chloroplast genome is the second reference quality cp genome to have been sequenced for the genus *Colobanthus* and the 12th reference genome for the family Caryophyllaceae. The *Colobanthus apetalus* cp genome (151 228 bp) is typical in size relative to other angiosperms. It belongs to the group of medium-sized cp genomes in the family Caryophyllaceae: it is only 48 bp smaller than the cp genome of its close relative *Colobanthus quitensis* (151 276 bp), and more than 1000 bp smaller than the biggest cp genome of *Lychnis wilfordii* (152 320 bp). At the same time, the analyzed genome is more than 4000 bp bigger than the smallest known cp genome in Caryophyllaceae, which belongs to *Silene conica* (147 208 bp).

The observed differences in the size of cp genomes result from rearrangements in the genome structure of its non-coding regions, rather than from changes in the number of genes because *S. conica* has 111 genes (protein genes + tRNA genes + rRNA genes) and *L. wilfordii* has 110 genes due to the pseudogenization of the *accD* gene (Sloan et al. 2012; Kang et al. 2017). The variations in the size of cp genomes can also be explained by changes in IR structure, such as contractions and expansions (Raubeson et al. 2007; Ravi et al. 2008; Wang et al. 2008).

Although the detailed location of IR boundaries changes frequently in angiosperms (Goulding et al. 1996), they are generally found within *ycf1* and *rps19* genes. The location of the IR boundary followed the above rule in both analyzed species of the genus *Colobanthus*. Moreover, the IR boundaries appeared to be identical in both species. The IR were very similar in size in *C. apetalus* and *C. quitensis* (25 321 bp and 25 303 bp, respectively), and were located approximately in the middle of the IR size range for the family Caryophyllaceae, with the longest IR in *S. noctiflora* (29 891 bp), and the shortest IR in *S. chalcedonica* (23 540 bp) (Kang et al...
A detailed comparison of cp genomes in Caryophyllaceae (Kang et al. 2017) revealed the coexistence of three types of cp genomes: I) the ‘common’ type found in most Caryophyllaceae (Agrostemma githago, Silene capitata, S. latifolia, S. vulgaris and Colobanthus quitensis); II) cp genomes with an inversion of the ycf3-psal region (S. paradoxa, S. conoidea, S. conica); III) cp genomes with the most complex structure identified to date among Caryophyllaceae, with several transpositions and/or inversions (S. noctiflora). The C. apetalus cp genome was highly similar to the C. quitensis cp genome, and it was classified as belonging to the first type.

The repeat regions of genomes play an important role in recombination and rearrangement (Cavalier-Smith 2002). In angiosperms, repeat regions are generally found in non-coding regions, and their variations result mainly from illegitimate recombination and slipped-strand mispairing (Timme et al. 2007). Similar observations were made in C. apetalus where repeats of ≥ 30bp were generally found (76.7%) in intergenic regions and introns. The above values are relatively high in comparison with C. quitensis (53.3%) and other Caryophyllaceae, such as Silene capitata (56.0%) and Lychnis wilfordii (69.2%) (Kang et al. 2017).

Microsatellites (SSRs) are particularly important repetitive elements of the genome. Due to their high reproducibility, ease of scoring and fast throughput, they are the markers of choice for many population and evolutionary analyses (Rajwant et al. 2011). Microsatellite sequences were abundant in the cp genomes of C. apetalus and C. quitensis, and mononucleotide SSRs were most frequent (48.8% and 54.2%, respectively) with a predominance of the A/T motif. A predominance of A/T in mononucleotide SSRs was previously reported in Caryophyllaceae (Kang et al. 2017), Magnoliaceae (Kuang et al. 2011), Poaceae (Sonah et al. 2011) and Rhamnaceae (Ma et al. 2017). Microsatellites have a unique structure and play an important role in genomic rearrangement and sequence variation, including chloroplast genomes (Yang et al. 2013). In our study, most tandem repeats in C. apetalus and C. quitensis (82.9% and 89.6%, respectively) were found in intergenic spacers and introns which are often divergent hotspots (Huang et al. 2014). This observation suggests that these regions can be potentially used to develop new DNA markers for species identification and phylogenetic studies of Colobanthus and, possibly, other taxa that are closely related to the family Caryophyllaceae. Other Caryophyllaceae species were characterized by congruent distribution of SSRs which were also found mainly in non-coding regions (62.7% in Lychnis wilfordii and 73.3% in Silene capitata) (Kang et al. 2017). Moreover, large numbers of SSRs in C. apetalus and C. quitensis were found
within the coding sequences of only five genes, including \textit{ycf1} which is one of the most rapidly evolving sequences in many groups of plants, including Caryophyllaceae (Erixon et al. 2008; Sloan et al. 2014), Campanulaceae, Geraniaceae and Poaceae (Jansen et al. 2007).

Second generation high-throughput sequencing technologies generate massive amounts of data with dozens of possible applications, such as the identification of interspecific polymorphism within cp genomes. However, the detection of a polymorphic site within a genome and its separation into major and minor genotype is only the first step in interspecies SNP identification, and further confirmation is required based on extensive sampling. Interspecific polymorphisms point to the heterogeneous nature of the chloroplast population in a given species and could be indicative of heteroplasmy. Heteroplasmy in plastids has been detected in many flowering plants, and it is more common than previously thought (Tilney-Bassett et al. 1981; Moon et al. 1987; Lee et al. 1988; Frey et al. 1999; Sabir et al. 2014). Two mechanisms could be responsible for the development of heteroplasmy in plastids. The more common mechanism which is found in around 20% of angiosperms (Corriveau and Coleman 1988; Zhang et al. 2003) is biparental inheritance, where each parent transmits organelles to the zygote. The second mechanism is found in plants with uniparental plastid inheritance, where plastid sorting in parents is incomplete resulting in heteroplasmic gametes. Incomplete sorting could be the mechanism underlying heteroplasmy in the genus \textit{Colobanthus}. A microscopic analysis of reproductive biology in \textit{C. quitensis} revealed that the male germ unit is differentiated into smaller cell containing mainly mitochondria, and bigger one with plastids.

In the process of fertilization in \textit{C. quitensis} only one nucleus of the sperm cell, without cytoplasm fragments of pollen tube, entered the egg cell, and the proembryo developed according to the Caryophyllad type (Giełwanowska et al. 2011). A similar mechanism is likely in \textit{C. apetalus}. \textit{C. quitensis} develops two types of bisexual flowers: opening chasmogamous flowers and closed cleistogamous flowers, where the latter type is favored by low temperatures, high air humidity and strong winds (Giełwanowska et al. 2011). The above leads to high selfing rates and the loss of genetic variation in individuals. Considerable similarities in the reproductive biology of \textit{C. quitensis} and \textit{C. apetalus} offer the best explanation for the very small number of SNPs in the \textit{C. apetalus} cp genome, which should be regarded as a minor symptom of heteroplasmy.
Synonymous and non-synonymous nucleotide substitution patterns are a very important element in gene evolution studies (Kimura 1983). Non-synonymous nucleotide substitution is less frequently observed than synonymous substitution (Makalowski and Boguski 1998). In this study, most differences in the chloroplast genes of *C. apetalus* and *C. quitensis* were found in the second and third position of the codon rather than in the first position. However, *ycf1* and *rpoC2* appeared to have the most variable nucleotide sequence and the highest number of non-synonymous substitutions. Similar observations have been made in members of genus *Silene* (Caryophyllaceae), where *ycf1* was one of the most varied coding regions (Erixon et al. 2008; Sloan et al. 2012, 2014; Kang et al. 2017). In *rpoB* and *rpl16*, insertion/deletion polymorphisms were also responsible for sequence variation. In addition to three nucleotide substitutions, a comparison of *rpoB* sequences in *C. apetalus* and *C. quitensis* revealed the deletion of one amino acid (glutamine; Q) (KKGQQLLA in *C. quitensis* → KKGQLLA in *C. apetalus*). Glutamine was also deleted from the above amino acid sequence in other Caryophyllaceae. However, an additional conservative substitution of one leucine (L) with isoleucine (I) was noted in the KKGQLLA sequence characteristic for *C. apetalus*, and the new sequence (KKGQILA) was observed in the cp genomes of all Caryophyllaceae that have been sequenced to date. Moreover, the same pattern can be found in more distant relatives of the order Caryophyllales, such as *Spinacia oleracea*, *Dianthus longicalyx* or *Mesembryanthemum crystallinum*. The insertion of an additional nucleotide in the *rpl16* ribosomal protein gene shifted the reading frame and, consequently, added three amino acids (isoleucine, glycine and threonine) to the gene sequence at the carboxyl terminus of its protein. This type of rearrangement can influence protein structure; however, further research is needed to accurately predict its consequences for protein structure and function (Berezovsky et al. 1999). An analysis of *rpl16* in the cp genomes of the sequenced Caryophyllaceae did not produce any evidence to suggest that the gene’s length was expanded based on the mechanism observed in *C. apetalus*. The only exception was *Silene paradoxa*, where the C-terminus of *rpl16* was expanded by three, albeit different, amino acids (leucine, glycine and methionine). Comparative analyses of cp genomes in *Amelopsis*, *Vitis* and *Liquidambar* also demonstrated high variation within ribosomal protein (*rpl22* and *rps19*) sequences due to a high non-synonymous substitution rate (Raman and Park 2016).

An interesting difference in the sequence of *rpl22* and *rps16* genes was observed in the cp genomes of *C. apetalus* and *C. quitensis*. A comparison of the *rpl22* gene revealed a substitution
of four nucleotides in *C. apetalus*, which led to changes in three amino acids, the appearance of a premature STOP codon and gene contraction by eleven amino acids. A similar ‘short’ allele of *rpl22* (length of 453 bp = 151 aa) was found in other members of the family Caryophyllaceae (*Agrostemma githago, Lychnis wilfordii, Silene capitata, S. conica, S. conoidea, S. latifolia, S. noctiflora S. paradoxa, S. vulgaris*), and other alleles of the *rpl22* gene were found only in *S. paradoxa* (153aa) and *S. chalcedonica* (161aa). In contrast, the substitution of only one nucleotide in *rps16* in *C. apetalus* changed the STOP codon to arginine, which resulted in protein elongation by eight additional amino acids. The ‘long’ allele of *rps16* is also found in other Caryophyllaceae (*Silene vulgaris, S. noctiflora, S. latifolia, S. chalcedonica, Agrostemma githago*) and Caryophyllales (*Dianthus longicalyx, Spinacia oleracea Mesembryanthemum crystallinum*), but with possible amino acid substitutions. A ‘short’ allele similar to that found in *C. quitensis* was characteristic also for *Silene conica*. The above examples of unique molecular evolution patterns in the genus *Colobanthus* provide valuable inputs for further studies into the evolution and phylogeography of this plant group and the order Caryophyllales. In conclusion, both *C. apetalus* and *C. quitensis* revealed a high degree of sequence conservation in genes that are directly involved in photosynthesis and considerable variations in other genes, including ribosomal proteins and rapidly evolving genes such as *ycf1*. The same conclusions can be drawn from the analysis of synonymous and non-synonymous substitution rates based on the sequences of 76 protein genes shared by all of the studied Caryophyllaceae species. The average Ks values between *C. apetalus* and selected Caryophyllaceae species were estimated at 0.1383, 0.0922 and 0.1789 for the LSC, IR and SSC regions, respectively, with an average Ks of 0.1399 across all regions. The lowest Ks values were found mostly in the IR region, and in two genes (*rpl19* and *ycf1*), the average Ks exceeded the estimate for this group of genes. The LSC region harbored only 12 genes where the average Ks was below 0.0922, whereas no such genes were identified in the SSC region. The distribution of Ks values indicates that the IR region is generally more conserved than LSC and SSC where higher evolution rates are observed. Similar observations have been made by other authors (Cho et al. 2015; Fu et al. 2016).

The phylogenetic tree presented in this paper was similar to the trees that have been previously developed for Caryophyllaceae based on complete cp genomes (Kang et al. 2016) and protein-coding genes (Kang et al. 2017). In all cases, the separate character of the genera *Colobanthus* (so far, represented only by the cp genomes of *C. apetalus* and *C. quitensis*) and
Agrostemma was established, despite similarities with the heterogeneous genus Silene and the genus Lychnis nested within it. Silene and Lychnis were regarded as sister genera within the tribe Sileneae; however, the taxonomic identities and limitations between these two genera remain unclear (Lidén et al. 2000). Although some data is available for the genera Silene and Lychnis (Greenberg and Donoghue 2011; Kang et al. 2017), further research is needed to resolve the relationship between these genera and within the genus Silene.

5. Conclusions

The development of a reference cp genome for C. apetalus will be valuable for comparative studies of the family Caryophyllaceae and/or the order Caryophyllales which contain halophytic, drought-tolerant and cold-resistant species, including C. quitensis. The availability of cp genome sequences for such an interesting group of plants will contribute to the development of new applications in biotechnology, such as chloroplast gene transformation. The reference chloroplast genome is also highly useful for accurate assembly and annotation of cp genomes in other plants within the studied group, identification and analysis of interspecies hybridization events, monitoring their present spread and reconstructing their historical dispersal. The reference cp genome will be particularly useful for monitoring the spread of the genus Colobanthus throughout the Southern Hemisphere. The history of the genus Colobanthus, its historical dispersal routes and the location of glacial refugia for particular species are fascinating areas of research, but progress in this area is hampered by the lack of sufficient molecular data.

6. References

Allan, H.H. (1961). Flora of New Zealand. Vol. 1. Indigenous Tracheophyta. Wellington: Government Printer.

Androsiuk, P., Chwedorzewska, K.J., Szandar, K., Giełwanowska, I. (2015). Genetic variability of Colobanthus quitensis from King George Island (Antarctica). Pol. Polar Res. 36, 281-295.

Bascunan-Godoy, L., Sanhueza, C., Cuba, M., Zuniga, G.E., Corcuera, L.J., Bravo, L.A. (2012). Cold-acclimation limits low temperature induced photoinhibition by promoting a higher photochemical quantum yield and a more effective PSII restoration in darkness in the Antarctic rather than the Andean ecotype of Colobanthus quitensis Kunt Bartl (Caryophyllaceae). BMC Plant Biol. 12, 114.
Berezovsky, I.N., Kilosanidze, G.T., Tumanyan, V.G., Kisselev, L.L. (1999). Amino acid composition of protein termini are biased in different manners. Protein Eng. Des. Sel. 12, 23–30.

Bravo, L.A., Saavedra-Mella, F.A., Vera, F., Guerra, A., Cavieres, L.A., Ivanov, A.G., Huner N.P., Corcuera L.J. (2007). Effect of cold acclimation on the photosynthetic performance of two ecotypes of *Colobanthus quitensis* (Kunth) Bartl. J. Exp. Bot. 58, 3581-3590.

Cavalier-Smith, T. (2002). Chloroplast evolution: secondary symbiogenesis and multiple losses.Curr. Biol. 12, R62-R64.

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., van den Berg, C., Madriñán, S., Petersen, G., Seberg, O., Jørgensen, T., Cameron, K.M., Carine, M., Pedersen, N., Hedderson, T.A.J., Conrad, F., Salazar, G.A., Richardson, J.E., Hollingsworth, M.L., Barraclough, T.G., Kelly, L., Wilkinson, M. (2007). A proposal for standardized protocol to barcode all land plants. Taxon. 56, 295-299.

Cho, K.S., Yun, B.K., Yoon, Y.H., Hong S.Y., Mekapogu M., Kim K.H., Yang T.J. (2015) Complete chloroplast genome sequence of tartary buckwheat (*Fagopyrum tataricum*) and comparative analysis with common buckwheat (*F. esculentum*). PloS One, 10, e0125332.

Corriveau, J.L., Coleman, A.W. (1988) Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperms. Amer J Bot 75, 1443–1458

Cuba-Díaz, M., Castel, K., Acuña, D., Machuca, Á., Cid, I. (2017). Sodium chloride effect on *Colobanthus quitensis* seedling survival and in vitro propagation. Antarct. Sci. 29, 45-46.

Dong, X., Stothard, P., Forsythe, I.J., Wishart, D.S. (2004). PlasMapper: a web server for drawing and auto-annotating plasmid maps. Nucleic Acids Res. 32, W660-664.

Drescher, A., Ruf, S., Calsa, T.J., Carrer, H., Bock, R. (2000). The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 22, 97-104.

Erixon, P., Oxelman, B. (2008). Reticulate or tree-like chloroplast DNA evolution in *Sileneae* (Caryophyllaceae)? Mol. Phylogenet. Evol. 48, 313-325.

Frey, J.E., Müller-Schärer, H., Frey, B., Frei, D. (1999) Complex relation between triazine-susceptible phenotype and genotype in the weed *Senecio vulgaris* may be caused by chloroplast DNA polymorphism. Theor Appl Genet 99, 578–586.
Fu, P.C., Zhang, Y.Z., Geng, H.M., Chen S.L. (2016) The complete chloroplast genome sequence of *Gentiana lawrencei* var. farreri (Gentianaceae) and comparative analysis with its congeneric species. PeerJ 4, e2540.

Giełwanowska, I., Bochenek, A., Gojlo, E., Gorecki, R., Kellmann, W., Pastorczyk, et al. (2011). Biology of generative reproduction of *Colobanthus quitensis* (Kunth) Bartl. from King George Island, South Shetland Islands. Pol. Polar Res. 32, 139-155.

Giełwanowska, I., Pastorczyk, M., Lisowska, M., Wegrzyn, M., Gorecki, R.J. (2014). Cold stress effects on organelle ultrastructure in polar Caryophyllaceae species. Pol. Polar Res. 35, 627-646.

Goulding, S.E., Olmstead, R.G., Morden, C.W., Wolfe, K.H. (1996). Ebb and flow of the chloroplast inverted repeat. Mol. Gen. Genet. 252, 195-206.

Gray, M.W. (1989). The evolutionary origins of organelles. Trends Genet. 5, 294-299.

Greenberg, A.K., Donoghue, M.J. (2011). Molecular systematics and character evolution in Caryophyllaceae. Taxon. 60, 1637-1652.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59, 307-321 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.

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

Huelsenbeck, J.P., Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee S.B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L. (2007). Analysis of 81 genes from 64
plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. USA. 104, 19369–19374.

Kang, J.S., Lee, B.Y., Kwak, M. (2017). The complete chloroplast genome sequences of *Lychnis wilfordii* and *Silene capitata* and comparative analyses with other Caryophyllaceae genomes. PLoS One. 12(2), e0172924.

Kang, Y., Lee, H., Kim, M.K., Shin, S.C., Park, H., Lee, J. (2016). The complete chloroplast genome of Antarctic pearlwort *Colobanthus quitensis* (Kunth) Bartl. (Caryophyllaceae). Mitochondr. DNA Part A. 27, 4677-4678.

Katoh, K., Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772-780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjes, P., Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647-1649.

Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., Janzen, D.H. (2005). Use of DNA barcodes to identify flowering plants. Proc. Natl. Acad. Sci. USA. 102, 8369-74.

Koc J., Androsiuk P., Chwedorzewska K.J., Cuba-Díaz M, Górecki R., Giełwanowska I. 2018. Range-wide pattern of genetic variation in *Colobanthus quitensis*. Polar Biology (in press)

Kuang, D.Y., Wu, H., Wang, Y.L., Gao, L.M., Zhang, S.Z., Lu, L. (2011). Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implications for DNA barcoding and population genetics. Genome. 54(8), 663-673.

Kumar, S., Stecher, G., Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.. Mol Biol Evol. 33, 1870-4.

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

Lee, D.J., Blake, T.K., Smith, S.E. (1988) Biparental inheritance of chloroplast DNA and the existence of heteroplasmic cells in alfalfa. Theor Appl Genet 76, 545–549.
Lidén, M., Popp, M., Oxelman, B.A. (2000). A revised generic classification of the tribe Sileneae (Caryophyllaceae). Nord. J. Bot. 20, 513-518.

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. Curr. Genet. 52, 267-274.

Ma, Q., Li, S., Bi, C., Hao, Z., Sun, C., Ye, N. (2017). Complete chloroplast genome sequence of major economic species, Ziziphus jujuba (Rhamnaceae). Curr. Genet. 63, 117-129.

Makalowski, W., Boguski, M.S. (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA. 95, 9407-9412.

Mayer, C. (2006–2010). Phobos 3.3.11. Available from: http://www.rub.de/spezzoo/cm/cm_phobos.htm.

Moon, E., Kao, T.H., Wu, R. (1987) Rice chloroplast DNA molecules are heterogeneous as revealed by DNA sequences of a cluster of genes. Nucl Acids Res 15, 611–630

Moore, M.J., Bell, C.D., Soltis, P.S., Soltis, D.E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. USA. 104, 19363–19368.

Navarrete-Gallegos, A.A., Bravo, L.A., Molina-Montenegro, M.A., Corcuera, L.J. (2012). Antioxidant responses in two Colobanthus quitensis (Caryophyllaceae) ecotypes exposed to high UV-B radiation and low temperature. Rev. Chil. Hist. Nat. 85, 419-433.

Neuhaus, H.E., Emes, M.J. (2000). Nonphotosynthetic metabolism in plastids. Annu. Rev. Plant Biol. 51, 111-140.

Palmer, J.D., Jansen, R.K., Michaels, H.J., Chase, M.W., Manhart, J.R. (1988). Chloroplast DNA variation and plant phylogeny. Ann. Missouri. Bot. Gard. 75, 1180-1206.

Parks, M., Cronn, R., Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 7, 84. doi:10.1186/1741-7007-7-84.

Pastorczyk, M., Gielwanowska, I., Lahuta, L.B. (2014). Changes in soluble carbohydrates in polar Caryophyllaceae and Poaceae plants in response to chilling. Acta Physiol. Plant. 36, 1771-1780.
Rajwant, K.K., Manoj, K.R., Sanja, K., Rohtas, S., Dhawan, A.K. (2011). Microsatellite markers: an overview of the recent progress in plants. Euphytica. 177, 309-334.

Raman, G., Park, S. (2016). The complete chloroplast genome sequence of *Amelopsis*: gene organization, comparative analysis and phylogenetic relationships to other angiosperms. Front. Plant Sci. 7, 341.

Raubeson, L.A., Peery, R., Chumley, T.W., Dziubek, C., Fourcade, H.M., Boore, J.L., Jansen, R.K. (2007). Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genom. 8, 174.

Ravi, V., Khurana, J.P., Tyagi, A.K., Khurana, P. (2008). An update on chloroplast genomes. Plant Syst. Evol. 271, 101-122.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Rozas, J., Ferrer-Mata, A., Sánchez-Del Barrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sánchez-Gracia, A. (2017) DnaSP 6: DNA sequence polymorphism analysis of large datasets. Mol. Biol. Evol. 34, 3299–3302.

Sabir, J.S.M., Arasappan, D., Bahieldin, A., Abo-Aba, S., Bafeel, S., Zari, T.A., Edris, S., Shokry, A.M., Gadalla, N.O., Ramadan, A.M., Atef, A., Al-Kordy, M.A., El-Domyati, F.M., Jansen, R.K. (2014) Whole mitochondrial and plastid genome SNP analysis of nine date palm cultivars reveals plastid heteroplasmy and close phylogenetic relationships among cultivars. PLoS ONE, 9, e94158.

Sablok, G., Raju, G.V.P., Mudunuri, S.B., Prabha, R., Singh, D.P., Baev, V., Yahubyan, G., Ralph, P.J., La Porta, N. (2015). ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and on-the-fly repeat detection. Database. bav084.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Tabata, S. (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res. 6, 283-90.

Sears, B.B. (1980). Elimination of plastids during spermatogenesis and fertilization in the plant kingdom. Plasmid. 4, 233-255.

Skottsberg, C. (1915). Notes on the relations between the floras of sub-Antarctic America and New Zealand. The Plant World. 18, 129-142.

Skottsberg, C. (1954). Antarctic flowering plants. Svensk Botanisk Tidskrift. 51, 330-338.
Sloan, D.B., Alverson, A.J., Wu, M., Palmer, J.D., Taylor, D.R. (2012). Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus *Silene*. Genome Biol. Evol. 4, 294-306.

Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M., Taylor, D.R. (2014). A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe *Sileneae* (Caryophyllaceae). Mol. Phylogenet. Evol. 72, 82-89.

Sonah, H., Deshmukh, R.K., Sharma, A., Singh, V.P., Gupta, D.K. (2011). Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. PLoS One. 6(6), e21298.

The AGS online Plant Encyclopaedia. Available from: http://encyclopaedia.alpinegardensociety.net/plants/Colobanthus Cited 9th June 2017.

The Plant List Version 1.0. 2010; Available from: http://www.theplantlist.org/browse/A/Caryophyllaceae/Colobanthus Cited 9th June 2017.

Tilney-Bassett, R.A.E., Birky, C.W. Jr. (1981) The mechanism of the mixed inheritance of chloroplast genes in *Pelargonium*. Evidence from gene frequency distributions among the progeny of crosses. Theor Appl Genet 60, 43–53.

Timme, R.E., Kuehl, J.V., Boore, J.L., Jansen, R.K. (2007). A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: identification of diverged regions and categorization of shared repeats. Am. J. Bot. 94, 302-312.

Wang, R.J., Cheng, C.L., Chang, C.C., Wu, C.L., Su, T.M., Chaw, S.M. (2008). Dynamics and evolution of the inverted repeat-large scale copy junction in the chloroplast genome of monocots. BMC Evol. Biol. 8, 36.

Yang, J.B., Yang, S.X., Li, H.T., Jang, J., Li, DZ. (2013). Comparative chloroplast genomes of *Camellia* species. PLoS One. 8(8), e73053.

Zhang, Q., Liu, Y., Sodmergen (2003) Examination of the cytoplasmic DNA in male reproductive cells to determine the potential for cytoplasmic inheritance in 295 angiosperm species. Plt Cell Physiol 44, 941–951.
**Table legend:**

Table 1. GenBank accession numbers and references for cp genomes used in this study. Species list arranged alphabetically.

Table 2. List of genes present in chloroplast genome of *Colobanthus apetalus*, genes list arranged alphabetically.

Table 3. Summary of chloroplast genome characteristic of *C. apetalus* and *C. quitensis*.

**Figure legend**

Figure 1 Gene map of the *Colobanthus apetalus* chloroplast genome. Genes drawn inside the circle are transcribed clockwise, and those outside are transcribed counterclockwise (indicated by arrows). Differential functional gene groups are color-coded. GC content variations is shown in the middle circle.

Figure 2 The distribution, type and presence of SSRs in cp genome of *Colobanthus apetalus* and *Colobanthus quitensis*. (A) Presence of SSRs in the LSC, SSC and IR regions in *C. apetalus* cp genome; (B) Presence of SSRs in the LSC, SSC and IR regions in *C. quitensis* cp genome; (C) Presence of SSRs in the protein-coding regions, intergenic spacers and introns in cp genome of *C. apetalus* (C. a) and *C. quitensis* (C. q)

Figure 3 Nucleotide and amino acid substitutions in cp genome of *C. apetalus* when compared with *C. quitensis* plastid genome. \( N_{aa} \) - number of changed amino acids, \( N_n \) - number of changed nucleotides in particular group of genes: (A) genes for photosynthesis; (B) self-replication genes; (C) other genes

Figure 4 Border position of LSC, SSC and IR regions for *C. apetalus* and *C. quitensis*. Genes are indicated by boxes and the gaps between the genes and the boundaries are indicated by number of bases unless the gene coincides with the boundary. Extension of the genes are also indicated above the boxes.

Figure 5 Phylogeny of *Colobanthus apetalus* and other 11 representatives of Caryophyllaceae based on complete chloroplast genomes using maximum likelihood method.
Supplemental files:

Table S1. List of repeated sequences in the chloroplast genome of *Colobanthus apetalus*.

Table S2. List of repeated sequences in the chloroplast genome of *Colobanthus quitensis*.

Table S3. Distribution of SSR in the *Colobanthus apetalus* cp genome.

Table S4. Distribution of SSR in the *Colobanthus quitensis* cp genome.

Table S5. SNPs within cp genome of *C. apetalus*.

Table S6a. Ks and Ka values of *C. apetalus* chloroplast genome vs. selected representatives of Caryophyllaceae.

Table S6b. Ks and Ka values of *C. apetalus* chloroplast genome vs. *C. quitensis* cp genome.
Table 1 (on next page)

GenBank accession numbers and references for cp genomes used in this study.

Species list arranged alphabetically.
| Species                  | Accession number | Length     | Reference         |
|-------------------------|------------------|------------|-------------------|
| *Agrostemma githago*    | NC_023357        | 151 733 bp| Sloan et al. 2014 |
| *Arabidopsis thaliana*  | NC_000932        | 154 478 bp| Sato et al. 1999  |
| *Colobanthus apetalus*  | MF687919         | 151 228 bp| in this study     |
| *Colobanthus quitensis* | NC_028080        | 151 276 bp| Kang et al. 2016  |
| *Lychnis wilfordii*     | NC_035225        | 152 320 bp| Kang et al 2017   |
| *Silene capitata*       | NC_035226        | 150 224 bp| Kang et al 2017   |
| *Silene chalcedonica*   | NC_023359        | 148 081 bp| Sloan et al. 2014 |
| *Silene conica*         | NC_016729        | 147 208 bp| Sloan et al. 2012 |
| *Silene conoidea*       | NC_023358        | 147 896 bp| Sloan et al. 2014 |
| *Silene latifolia*      | NC_016730        | 151 736 bp| Sloan et al. 2012 |
| *Silene noctiflora*     | NC_016728        | 151 639 bp| Sloan et al. 2012 |
| *Silene paradoxa*       | NC_023360        | 151 632 bp| Sloan et al. 2014 |
| *Silene vulgaris*       | NC_016727        | 151 583 bp| Sloan et al. 2012 |
List of genes present in chloroplast genome of *Colobanthus apetalus*.

Genes list arranged alphabetically.
| Category                                      | Group of gene          | Name of genes                                                                 |
|----------------------------------------------|------------------------|------------------------------------------------------------------------------|
| Photosynthesis                               | Photosystem I          | *psaA, psaB, psaC, psaI, psaJ*                                               |
|                                              | Photosystem II         | *psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ* |
|                                              | Cytochrome complex     | *petA, petB, petD, petG, petL, petN*                                         |
|                                              | ATP synthase           | *atpA, atpB, atpE, atpF, atpH, atpI*                                        |
|                                              | NADH dehydrogenase     | *ndhA, ndhB (x2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*    |
|                                              | Large subunit of RUBISCO| *rbcL*                                                                       |
| DNA replication and protein synthesis        | Ribosomal RNA          | *rrn4.5 (x2), rrn5 (x2), rrn16 (x2), rrn23 (x2)*                            |
|                                              | Small subunit ribosomal proteins | *rps2, rps3, rps4, rps7 (x2), rps8, rps11, rps12 (x2), rps14, rps15, rps16, rps18, rps19 (x2)* |
|                                              | Large subunit ribosomal proteins | *rpl2 (x2), rpl14, rpl16, rpl20, rpl22, rpl32 (x2), rpl33, rpl36*     |
|                                              | RNA polymerase subunits | *rpoA, rpoB, rpoC1, rpoC2*                                                 |
|                                              | Transfer RNA           | *trnA-UGC (x2), trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnM-CAU, trnG-GCC, trnG-UCG, trnH-GUG, trnI-CAU (x2), trnI-GAU (x2), trnK-UUU, trnL-CAA (x2), trnL-UAA, trnL-UAG, trnM-CAU, trnN-GUU (x2), trnP-UGG, trnQ-UGG, trnR-ACG (x2), trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC (x2), trnV-UAC, trnW-CCA, trnY-GUA* |
|                                              | Conserved chloroplast ORF | *ycf1 (x2), ycf2, ycf3*, ycf4*, ycf68 (x2)*                                |
|                                              | Other proteins         | *accD, ccsA, cemA, clpP, matK*                                              |

1 * genes associated with Photosystem I
2
Table 3 (on next page)

Summary of chloroplast genome characteristic of *C. apetalus* and *C. quitensis*. 
| Genome features                  | C. apetalus   | C. quitensis  |
|---------------------------------|---------------|--------------|
| Size (bp)                       | 151 228       | 151 276      |
| LSC length (bp)                 | 83 380        | 83 462       |
| SSC length (bp)                 | 17 206        | 17 208       |
| IR length (bp)                  | 25 321        | 25 303       |
| Number of genes                 | 112           | 112          |
| Protein-coding genes            | 78            | 78           |
| tRNA genes                      | 30            | 30           |
| rRNA genes                      | 4             | 4            |
| Number of genes duplicated in IR| 19            | 19           |
| Overall GC content (%)          | 36.65         | 36.7         |
Figure 1

Gene map of the *Colobanthus apetalus* chloroplast genome.

Genes drawn inside the circle are transcribed clockwise, and those outside are transcribed counterclockwise (indicated by arrows). Differential functional gene groups are color-coded. GC content variations is shown in the middle circle.
The distribution, type and presence of SSRs in cp genome of *Colobanthus apetalus* and *Colobanthus quitensis*.

(A) Presence of SSRs in the LSC, SSC and IR regions in *C. apetalus* cp genome; (B) Presence of SSRs in the LSC, SSC and IR regions in *C. quitensis* cp genome; (C) Presence of SSRs in the protein-coding regions, intergenic spacers and introns in cp genome of *C. apetalus* (C. a) and *C. quitensis* (C. q).
Figure 3

Nucleotide and amino acid substitutions in cp genome of *C. apetalus* when compared with *C. quitensis* plastid genome.

$N_{aa}$ - number of changed amino acids, $N_n$ - number of changed nucleotides in particular group of genes: (A) genes for photosynthesis; (B) self-replication genes; (C) other genes.
Figure 4

Border position of LSC, SSC and IR regions for *C. apetalus* and *C. quitensis*.

Genes are indicated by boxes and the gaps between the genes and the boundaries are indicated by number of bases unless the gene coincides with the boundary. Extension of the genes are also indicated above the boxes.
Figure 5

Phylogeny of *Colobanthus apetalus* and other 11 representatives of Caryophyllaceae based on complete chloroplast genomes using maximum likelihood method.