Long-read transcriptome and other genomic resources for the angiosperm *Silene noctiflora*

Alissa M. Williams,[*,1] Michael W. Itgen,[*] Amanda K. Broz,[*] Olivia G. Carter,[*] Daniel B. Sloan[*]

[*]Department of Biology, Colorado State University, Fort Collins, Colorado 80523

[1]Corresponding author: Alissa.Williams@colostate.edu

Abstract

The angiosperm genus *Silene* is a model system for several traits of ecological and evolutionary significance in plants, including breeding system and sex chromosome evolution, host-pathogen interactions, invasive species biology, heavy metal tolerance, and cytonuclear interactions. Despite its importance, genomic resources for this large genus of approximately 850 species are scarce, with only one published whole-genome sequence (from the dioecious species *S. latifolia*). Here, we provide genomic and transcriptomic resources for a hermaphroditic representative of this genus (*S. noctiflora*), including a PacBio Iso-Seq transcriptome, which uses long-read, single-molecule sequencing technology to analyze full-length mRNA transcripts and identify paralogous genes and alternatively spliced genes. Using these data, we have assembled and annotated high-quality full-length cDNA sequences for approximately 17,000 *S. noctiflora* genes and 27,000 isoforms. We demonstrated the utility of these data to distinguish between recent and highly similar gene duplicates by identifying novel paralogous genes in an essential protease complex. Further, we provide a draft assembly for the approximately 2.7-Gb genome of this species, which is near the upper range of genome-size values reported for diploids in this genus and three-fold larger than the 0.9-Gb genome of *S. conica*, another species in the same subgenus. Karyotyping confirmed that *S. noctiflora* is a diploid, indicating that its large genome size is not due to polyploidization. These resources should facilitate further study and development of this genus as a model in plant ecology and evolution.

Introduction

Silene is the largest genus in the angiosperm family Caryophyllaceae and serves as a model system in many fields of ecology and evolutionary biology (Bernasconi *et al.* 2009; Jafari *et al.* 2020). For instance, *Silene* is used to study breeding system evolution, as the genus includes hermaphroditic, gynodioecious, gynomonoecious, monoecious, and dioecious species (Desfeux *et al.* 1996; Charlesworth 2006). Gynodioecy (the coexistence of both hermaphroditic and male-sterile individuals) is thought to be the ancestral state of the genus (Desfeux *et al.* 1996) and is found in many extant *Silene* species as a result of cytoplasmic male sterility (CMS) factors (Taylor *et al.* 2001; Garraud *et al.* 2011). Dioecy, however, has evolved at least two times independently within *Silene*, including both ZW and XY sex determination systems (Mrackova *et al.* 2008; Slancarova *et al.* 2013; Balounova *et al.* 2019). Despite the diversity of *Silene* sexual systems, there is only one available whole genome sequence for the entire genus—from the dioecious species *S. latifolia*, which has heteromorphic XY sex chromosomes (Papadopulos *et al.* 2015; Krasovec *et al.* 2018). Whole genome resources are not available for any of the hermaphroditic species, which has limited comparative genomic studies into the evolution of dioecy within this genus. *Silene noctiflora* (**Figure 1**) is largely hermaphroditic but can produce a mixture of hermaphroditic and male-sterile flowers on the same plant (gynomonoecy) (Davis and Delph 2005). Also known as the night-flowering catchfly, this annual species is native to Eurasia and introduced throughout much of the world (McNeill 1980; Davis and Delph 2005).

Silene is also used as a model system for investigating the coevolution between nuclear and cytoplasmic genomes (i.e., cytonuclear interactions), including in CMS systems (Olson and Mccauley 2002; Städler and Delph 2002; Klaas and Olson 2006; Garraud *et al.* 2011). In addition, there is considerable variation across the genus in organelle genome evolution. *Silene conica* and *S. noctiflora* have two of the largest known plant mitochondrial genomes at 11 Mb and 7 Mb, respectively (Sloan *et al.* 2012a). In contrast, the mitochondrial genome of *S. latifolia* is only 0.25 Mb, about 45 times smaller than that of *S. conica* (Sloan *et al.* 2012a). Interestingly, the *Silene* species with expanded mitogenomes also display unusually high evolutionary rates and stark structural changes in the mitochondrial genome (Mower *et al.* 2007; Sloan *et al.* 2012a). *Silene noctiflora*, for example, has a mitochondrial genome made up of around 60

circular-mapping chromosomes, and these chromosomes are rapidly gained and lost in different lineages (Wu and Sloan 2019). The plastid genomes in *S. conica* and *S. noctiflora* exhibit a correlated pattern of increased evolutionary rate—however, this pattern is found only in a subset of genes, and changes in plastid genome size and structure are more limited (Sloan *et al.* 2014). The natural variation in organelle genome evolution found in this genus has been used to study how these differences affect cytonuclear interactions (Havird *et al.* 2015; Williams *et al.* 2019).

The ability to use *Silene* as a model for cytonuclear evolution is still limited by the lack of extensive nuclear genome resources. Previous work has characterized *Silene* nuclear genome size and chromosome number. Nuclear genome sizes in the genus vary considerably, although not as starkly as mitochondrial genome sizes, ranging roughly 4.5-fold among diploids (haploid sizes of 0.71 to 3.23 Gb) and 8-fold when the tetraploid *S. stellata* (5.77 Gb) is included (Kruckeberg 1960; Siroký *et al.* 2001; Bai *et al.* 2012; Dagher-Kharrat *et al.* 2013; Pellicer and Leitch 2020). Most diploids in the genus, including *S. noctiflora,* have a chromosome number of 2n=24, which is likely the ancestral number (Bari 1973; McNeill 1980; Yildiz *et al.* 2008; Kemal *et al.* 2009; Gholipour and Sheidai 2010; Ghasemi *et al.* 2015; Mirzadeh Vaghefi and Jalili 2019). There are also numerous polyploid *Silene* species, including tetraploid, hexaploid, and octaploid forms (Kruckeberg 1960; Popp and Oxelman 2001, 2007; Popp *et al.* 2005; Bai *et al.* 2012). Most of the available nuclear sequence data comes from short-read RNA sequencing, which has been conducted on multiple *Silene* species (Blavet *et al.* 2011; Sloan *et al.* 2012b; Muyle *et al.* 2012; Casimiro-Soriguer *et al.* 2016; Havird *et al.* 2017; Bertrand *et al.* 2018; Balounova *et al.* 2019). These datasets have provided an important resource for molecular studies of *Silene*, but are limited because of the challenges associated with assembling short-read sequences, especially in distinguishing similar sequences arising from gene duplication, heterozygosity, and/or alternative splicing (Alkan *et al.* 2011; Schatz *et al.* 2012; Hahn *et al.* 2014; Lan *et al.* 2017).

Pacific Bioscience (PacBio) offers a long-read technology that involves sequencing single molecules, often leading to high error rates (Au *et al.* 2012; Rhoads and Au 2015; Hestand *et al.* 2016). However, these high error rates can be drastically reduced using circular consensus sequencing (CCS). CCS reads are generated by using hairpin adapters on each end of a double-stranded molecule, creating a circular, single-stranded topology (Wenger *et al.* 2019). This

topology allows the polymerase to read the same full-length molecule multiple times over, generating an accurate consensus sequence (Ono *et al.* 2013; Wang *et al.* 2019). The application of PacBio CCS technology to reverse transcribed RNA (i.e., cDNA) samples is known as Iso-Seq and has been used to study the transcriptomes of many organisms, often in the context of identifying splice variants (Xu *et al.* 2015; Gordon *et al.* 2015; Rhoads and Au 2015; Guo *et al.* 2016; Abdel-Ghany *et al.* 2016; Wang *et al.* 2016; Weirather *et al.* 2017). Splice variants can be identified using CCS because this technology obtains consensus sequences for full-length single transcripts (Zhao *et al.* 2019). In the same way, CCS can also be used to distinguish paralogs or gene duplicates.

We have generated genomic resources critical for investigations into *S. noctiflora*, a species of interest due to its extremely unusual organelle evolution and resultant use as a model for cytonuclear interactions, as well as its status as a hermaphrodite in a genus representing many types of breeding system. We include a high-quality transcriptome using long-read PacBio Iso-Seq technology, genome size estimates, and a draft nuclear genome assembly. These resources will expand opportunities for molecular and ecological studies within the genus.

Materials and Methods

Plant growth conditions, tissue sampling, and nucleic acid extractions

Plants used for genome sequencing, Iso-Seq, and flow cytometry estimates of genome size were grown under standard greenhouse conditions with 16-hr light/8-hr dark at Colorado State University (**Table 1**). DNA for short-insert paired-end Illumina libraries was extracted from leaf tissue from a 7-week old *S. noctiflora* OPL individual using a Qiagen Plant DNeasy kit. Additional DNA was extracted from the same individual 6 weeks later using a modified CTAB protocol (Doyle and Doyle 1987) for construction of Illumina mate-pair libraries. For Iso-Seq library construction, RNA was extracted from a single 12-week old *S. noctiflora* OPL individual (grown from seed of the plant used for DNA extraction), using a Qiagen Plant RNeasy kit. RNA extractions were performed for four different tissue samples: 1) a large flower bud with calyx removed, 2) an entire smaller flower bud including calyx, 3) the most recent (top-most) pair of cauline leaves, and 4) one leaf from the second most recent pair of cauline leaves. The four RNA

116 extractions were quantified with Qubit RNA BR kit (Thermo Fisher Scientific). Purity and
117 integrity were assessed with a NanoDrop 2000 (Thermo Fisher Scientific) and TapeStation 2200
118 (Agilent Technologies).
119
120 **PacBio Iso-Seq transcriptome sequencing and analysis**
121 The four *S. noctiflora* RNA extractions (1.5 µg each) were pooled into a single sample and sent
122 to the Arizona Genomics Institute for PacBio Iso-Seq library construction and sequencing.
123 Library construction followed the standard PacBio Iso-Seq protocol (dated September 2018), and
124 the library was sequenced with a PacBio Sequel (first generation) platform on two SMRT Cells.
125
126 Raw movie files of long-read, single-molecule sequences (one per SMRT cell) were processed
127 using the PacBio Iso-Seq v3.1 pipeline (Anvar *et al.* 2018; Pacific Biosciences 2020). Circular
128 consensus sequence calling was performed on each movie file separately using the command *ccs*
129 with the recommended parameters *--noPolish* and *--minPasses 1*. Next, primer removal and
130 demultiplexing was performed on each dataset by running the command *lima* with parameters --
131 *isoseq* and *--no-pbi*. Poly(A) tails were trimmed and concatemers were removed using the *refine*
132 command with the parameter *--require-polya*. Data from the two movies were merged at this
133 point using the commands *dataset create --type TranscriptSet* and *dataset create --type*
134 *SubreadSet*. Finally, the merged data were run through the *cluster* and *polish* commands.
135
136 Trinotate v3.2.0 (Bryant *et al.* 2017) was used to annotate the final polished sequences produced
137 by the Iso-Seq pipeline. To complete this process, we used Transdecoder v5.5.0
138 (https://github.com/TransDecoder/TransDecoder/wiki), SQLite v3 (Kreibich 2010), NCBI
139 BLAST + v2.2.29 (Camacho *et al.* 2009), HMMER v3.2.1 (including RNAMMER) (Lagesen *et*
140 *al.* 2007; Potter *et al.* 2018), signalP v4 (Petersen *et al.* 2011), and tmhmm v2 (Krogh *et al.*
141 2001). The Pfam (Bateman *et al.* 2004) and UniProt ("UniProt" 2015) databases were included
142 in the Trinotate installation. The transcripts and Transdecoder-predicted peptides were searched
143 against the respective databases, following the standard Trinotate pipeline. All of these results
144 were loaded into a Trinotate SQLite database.
145

Cogent v4.0.0 (https://github.com/Magdoll/Cogent/wiki), minimap2 v2.17 (Li 2018), and cDNA_Cupcake Py2 v8.7 (https://github.com/Magdoll/cDNA_Cupcake/wiki) were used to conduct family finding on the final sequences outputted by the Iso-Seq pipeline by partitioning sequences into gene families based on similarity. Next, coding genome reconstruction was performed on each gene family from the above step. Finally, a transcript-based genome was used to collapse redundant isoforms.

The Cogent family finding output was used with cDNA_Cupcake scripts (https://github.com/Magdoll/cDNA_Cupcake/wiki) to perform a rarefaction analysis (i.e. "collector's curve"). First, a modified form of the script *make_file_for_sampling_from_collapsed.py* was run with the parameter *--include_single_exons* in order to include all transcripts in the analysis. Using the resultant file, *subsample.py* was run twice: once at the gene level (using the *pbgene* column) and once at the transcript level (using the *pbid* column). In both cases, parameters were kept as the default. Results of the rarefaction analysis were plotted in R using a modified version of the relevant cDNA_Cupcake script.

We used genes from the plastid caseinolytic protease (Clp) as a case study to assess the ability of Iso-Seq dataset to detect gene duplication events of various ages. To identify nuclear-encoded plastid Clp core genes in our dataset, we used blastn in conjunction with the Cogent family finding output. There are eight nuclear-encoded plastid Clp core genes in *Arabidopsis thaliana*: *CLPP3-6* and *CLPR1-4* (Nishimura and van Wijk 2015). Additionally, the genus *Silene* shares a duplication of *CLPP5*, denoted *CLPP5A* and *CLPP5B* (Rockenbach *et al.* 2016). We obtained the sequences of all nine of these genes from a previous study (Rockenbach *et al.* 2016) and used them as queries in blastn searches against the *S. noctiflora* Iso-Seq transcriptome. We then identified the BLAST hits in the Cogent output and based on those groups, we determined that eight of the nine nuclear-encoded Clp core subunits in *Silene* (including *CLPP5A* and *CLPP5B*) are single copy. However, in the case of *CLPR2*, two different Cogent families contained relevant transcripts, indicating a possible case of gene duplication. Sequence alignment of the transcripts within each Cogent family revealed that one family contained two unique sequences. These data, along with sequencing results from a separate cloning project, suggested that there are actually three distinct *CLPR2* sequences in *S. noctiflora*. We examined the other eight

nuclear-encoded Clp gene Cogent families and found no evidence of additional duplications. In the subsequent phylogenetic analysis of *CLPR2*, we used the longest sequences from each of the three identified groups.

A phylogenetic tree was constructed using sequences from the three different *S. noctiflora CLPR2* genes. In addition to the three *S. noctiflora* sequences, we also included *Agrostemma githago*, *S. conica, S. latifolia, S. paradoxa,* and *S. vulgaris CLPR2* sequences from a previous study (Rockenbach *et al.* 2016), as well as three *S. undulata CLPR2* sequences identified using blastn against the *S. undulata* TSA database (accession GEYX00000000). All 11 sequences were aligned using the *einsi* option in MAFFT v7.222 (Katoh and Standley 2013), and trimmed at the 5′ end based on the trimming conducted in Rockenbach *et al*. (2016). The resultant sequence file was run through jModelTest v2.1.10 (Darriba *et al.* 2012) to choose a model of sequence evolution. We chose the top model based on the Bayesian Information Criterion (K80+I) and ran PhyML v3.3 (Guindon *et al.* 2010) with 1000 bootstrap replicates and 100 random starts.

Genome size estimates by flow cytometry

Leaf or seedling samples were collected from multiple individuals of varying age (between 2 and 14 weeks) for each of our target *Silene* species and shipped fresh to Plant Cytometry Services (Schijndel, Netherlands). Genome sizes were determined using the CyStain PI Absolute P reagent kit (05-5502). Samples were chopped with a razor blade in 500 µl of ice-cold Extraction Buffer in a plastic petri dish, along with *Pachysandra terminalis* tissue as an internal standard (3.5 pg/2C). After 30-60 sec of incubation, 2 ml of Staining Buffer was added. Each sample was then passed through a nylon filter of 50 µm mesh size, and then incubated for 30+ min at room temperature. The filtered solution was then sent through a CyFlow ML flow cytometer (Partec GmbH). The fluorescence of the stained nuclei, which passed through the focus of a light beam with a 50 mW, 532 nm green laser, was measured by a photomultiplier and converted into voltage pulses. The voltage pulses were processed using Flomax version 2.4d (Partec) to yield integral and peak signals. Genome sizes were reported in units of pg/2C. The conversion used to report each size (x) in units of Gb was (x/2)*0.978 (Gregory *et al.* 2007).

Karyotyping

Silene noctiflora OPL seeds were germinated on wet filter paper and grown for 5 days. Radicles were trimmed off and transferred to ice water for 24 hrs. The radicles were then fixed in a 3:1 solution of absolute ethanol and glacial acetic acid and stored at -20°C. Chromosomes were visualized using a squash preparation with Feulgen staining. Fixed radicles were rinsed in distilled water for 5 min at 20°C. Radicles were then hydrolyzed in 5M HCl at 20°C for 60 min followed by three rinses in distilled water. The hydrolyzed radicles were transferred to Schiff's reagent to stain the DNA for 120 min at 20°C and were then destained by rinsing in SO_2 water at 20°C three times for 2 min, two times for 10 min, once for 20 min, and then transferred to distilled water. Squashes were prepared by placing a piece of tissue in 45% acetic acid for 10 minutes and then minced on glass. A coverslip was placed over the minced tissue and pressed with enough pressure to produce a monolayer of nuclei. Slides were placed on dry ice for 1 min, and the coverslip was removed. The slides were transferred to 96% ethanol for 2 min, air dried, and mounted with mounting medium. Chromosomes were observed using a compound light microscope at 100× magnification.

Genome sequencing and assembly

Extracted *S. noctiflora* OPL DNA samples were used for Illumina library construction and sequencing. A paired-end library with a target insert size of 275-bp was constructed at the Yale Center for Genome Analysis and sequenced on a 2×150-bp HiSeq 2500 run (three lanes). Two mate-pair libraries (with target insert sizes of 3-5 kb and 8-11 kb) were generated at GeneWiz and sequenced on a 2×150-bp HiSeq 2500 run (one lane each). Approximately 480M, 250M, and 230M read pairs were generated for the 275-bp, 3-5 kb, and 8-11 kb libraries, respectively. These reads are available via the NCBI SRA (accessions SRR9591157-SRR9591159). Reads were trimmed for quality and to remove 3′ adapters, using cutadapt v1.3 (Martin 2011) under the following paramters*: -n 3 -O 6 -q 20 -m 30 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC --paired-output*. The trimmed reads were assembled with ALLPATHS-LG release 44837 (Gnerre *et al.* 2011). Estimates of mean insert size and standard deviation for each library were provided as input for the assembly by first mapping a sample of reads to the published *S. noctiflora* plastid genome (GenBank accession JF715056.1). These estimates were as follows: 274 bp (± 22 bp), 3752 bp (± 419 bp), and 9873 bp (± 1283 bp).

Data availability

The original subread bam files and final transcript sequences longer than 199 bp from the PacBio Iso-Seq transcriptome are available at NCBI Sequence Read Archive (SRA accession SRR11784995) and NCBI Transcriptome Shotgun Assembly Sequence Database (TSA accession GIOF01000000), respectively. The genome assembly has been deposited in GenBank (accession VHZZ00000000.1). Additional data have been provided at GitHub (https://github.com/alissawilliams/Silene_noctiflora_IsoSeq): 1) the full transcriptome as outputted by the PacBio Iso-Seq pipeline, 2) the annotation report for the transcriptome, 3) a custom script used to create a gene_trans_map file for our data in order to use Trinotate on non-Trinity-derived data, 4) the Cogent family finding output, and 5) the set of trimmed, aligned sequences used in the *CLPR2* phylogenetic analysis.

Results and Discussion

Silene noctiflora Iso-Seq transcriptome: Gene content and duplication

Sequencing of the Iso-Seq library on two Sequel SMRT Cells produced 711,625 and 686,576 reads for the first and second cells, respectively, where each read was derived from a single molecule. The two SMRT Cells differed substantially in data yield, with totals of 12,765,109 and 21,844,543 subreads, corresponding to subread counts of 17.9 and 31.8 per read, respectively. These reads were merged into 65,642 distinct high-quality transcripts according to the thresholds of the Iso-Seq 3.1 *merge* and *polish* commands. Of these transcripts, only 14 were found to be non-plant sequences, all of which were derived from *Frankliniella occidentalis* (the western flower thrip), a common greenhouse pest that likely contaminated our tissue samples.

We used the Cogent (https://github.com/Magdoll/Cogent/wiki) family finding algorithm to further collapse the 65,642 transcripts into 11,677 “gene families,” and then used the Cogent data along with Cupcake (https://github.com/Magdoll/cDNA_Cupcake/wiki) to conduct a rarefaction analysis. The rarefaction analysis, or “collector’s curve”, uses random sampling of reads to determine whether the extent of sequencing was sufficient to detect most of the genes and

isoforms in our RNA sample. Based on this analysis, the Iso-Seq transcriptome contains 16,230 *S. noctiflora* genes and 27,860 isoforms (**Figure 2**). In both cases, the rarefaction analysis converged on a single estimate at 560,000 reads out of 594,988, indicating that we sequenced enough reads to essentially saturate our detection ability (which is also evident in the fact that the curves plateaued).

We wanted to test the ability of Iso-Seq to detect and distinguish paralogs of varying levels of divergence using the Cogent family finding output. To this end, we used a sample gene family—the core subunit genes of the plastid Clp complex, as they have a rich history of paralogy. In *E. coli* and most other bacteria, the core of the Clp complex, which is responsible for proteolysis, contains 14 identical subunits (Yu and Houry 2007). In cyanobacteria, gene duplication has led to four different core subunit-encoding genes (Stanne *et al.* 2007). Continued gene duplication in the land plant lineage has further reshaped this complex in plastids; the 14 core subunits are encoded by nine different genes in *A. thaliana*, eight of which are nuclear encoded (*CLPP3-6, CLPR1-4*), and one of which is plastid encoded (*clpP1*) (Nishimura and van Wijk 2015). Further, we had previously identified a more recent duplication of *CLPP5* in *Silene*, as well as duplications of the plastid-encoded *clpP1* in a small number of angiosperm species (Erixon and Oxelman 2008; Rockenbach *et al.* 2016; Williams *et al.* 2019).

We used the Cogent family finding output to examine the nine nuclear-encoded Clp core genes in *S. noctiflora*. The core genes *CLPP3, CLPP4, CLPP5A, CLPP5B, CLPP6, CLPR1, CLPR3,* and *CLPR4* were each represented by a single gene family in the Cogent output, whereas *clpR2* was represented by two gene families. Upon further examination, one of these families actually represented two different genes, yielding a total of three *CLPR2* genes in *S. noctiflora*. Thus, *CLPR2* was duplicated in this lineage, and then one paralog underwent a second gene duplication. Based on a phylogenetic analysis (**Figure 3**), these two duplications are shared with *S. undulata* but none of the other sampled *Silene* species. Thus, these duplications likely occurred after the *Silene* section Elisanthe (including *S. noctiflora*, *S. undulata*, and *S. turkestanica*) diverged from the other members of the genus (Jafari *et al.* 2020).

The Iso-Seq data allowed us to identify transcripts from every known nuclear-encoded Clp core gene in *S. noctiflora*, including the closely related *CLPP5A* and *CLPP5B* subunits, as well as an additional, previously unreported triplication of *clpR2*. This result demonstrates that the Iso-Seq transcriptome provides highly accurate sequences, even for closely related paralogs that can be used in further study.

Silene genome size estimates and chromosome number

Genome sizes of *S. noctiflora, S. conica, S. vulgaris,* and *S. latifolia* were determined using flow cytometry. Our estimates for *S. vulgaris* and *S. latifolia* (1.07 and 2.67 Gb, respectively; **Table 1**) were concordant with previously published estimates for these two species of 1.11 and 2.64 Gb (Costich *et al.* 1991; Siroký *et al.* 2001). Interestingly, despite their similar and extreme patterns of organelle evolution (Sloan *et al.* 2012a, 2014), including large mitochondrial genomes, *S. noctiflora* and *S. conica* have very different nuclear genome sizes. We found their respective genome sizes to be approximately 2.74 and 0.93 Gb, respectively (**Table 1**), which are on opposite ends of the spectrum for *Silene* diploids (Pellicer and Leitch 2020). The *S. noctiflora* nuclear genome is almost three-fold larger than that of *S. conica* suggesting that mitochondrial genome size is not necessarily correlated with nuclear genome size.

S. noctiflora has been previously reported as a diploid (2n=24) (McNeill 1980; Yildiz *et al.* 2008; Ghasemi *et al.* 2015). Given its relatively large genome size, we sought to confirm this result in our sampled population with a karyotype analysis (**Figure 4**), which indeed supported the conclusion that that *S. noctiflora* OPL is diploid.

The *Silene noctiflora* nuclear genome

Illumina sequencing produced ~50× coverage of the *S. noctiflora* genome for a 275-bp paired-end library and ~15-20× for each of two mate-pair libraries. By performing a *de novo* assembly of these reads, we obtained a total assembly length (including estimated scaffold gaps) of 2.58 Gb, which is generally consistent with our estimate based on flow cytometry for *S. noctiflora* OPL (2.71 Gb). Given that we relied entirely on short-read sequencing technology, it was not surprising that the resulting assembly of this large genome was highly fragmented (79,768

scaffolds with a scaffold N50 of 59 kb). Moreover, assembly gaps made up 73% of the total scaffold length, presumably representing the highly repetitive content that is typical of plant nuclear genomes. As such, the assembled gap-free sequences amount to only about a quarter of the genome (702 Mb). This assembly should provide a useful resource to query for sequences of interest, especially in genic regions, and to compare against *S. latifolia* and other members of this genus. However, a more complete assembly that includes repetitive regions of the genome will require additional data from long-read technologies such as PacBio or nanopore sequencing.

Acknowledgements

We thank Jocelyn Cuthbert and Zhiqiang Wu for assistance with plant growth and DNA extraction, Suzanne Royer for preliminary investigations into *Silene* karyotyping, and Joel Sharbrough for assistance with PacBio data analysis. This work was supported by a National Science Foundation (NSF) grant (MCB-1733227), start-up funds from Colorado State University, and graduate fellowships from NSF (DGE-1321845) and the National Institutes of Health (T32-GM132057).

Literature Cited

Abdel-Ghany, S. E., M. Hamilton, J. L. Jacobi, P. Ngam, N. Devitt *et al.*, 2016 A survey of the sorghum transcriptome using single-molecule long reads. Nature Communications 7: 1–11.

Alkan, C., S. Sajjadian, and E. E. Eichler, 2011 Limitations of next-generation genome sequence assembly. Nat Methods 8: 61–65.

Anvar, S. Y., G. Allard, E. Tseng, G. M. Sheynkman, E. de Klerk *et al.*, 2018 Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. Genome Biology 19: 46.

Au, K. F., J. G. Underwood, L. Lee, and W. H. Wong, 2012 Improving PacBio Long Read Accuracy by Short Read Alignment. PLoS One 7:.

359 Bai, C., W. S. Alverson, A. Follansbee, and D. M. Waller, 2012 New reports of nuclear DNA
360 content for 407 vascular plant taxa from the United States. Ann. Bot. 110: 1623–1629.
361 Balounova, V., R. Gogela, R. Cegan, P. Cangren, J. Zluvova *et al.*, 2019 Evolution of sex
362 determination and heterogamety changes in section Otites of the genus Silene. Scientific
363 Reports 9: 1–13.
364 Bari, E. A., 1973 Cytological Studies in the Genus Silene L. New Phytologist 72: 833–838.
365 Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich *et al.*, 2004 The Pfam protein families
366 database. Nucleic Acids Res 32: D138–D141.
367 Bernasconi, G., J. Antonovics, A. Biere, D. Charlesworth, L. F. Delph *et al.*, 2009 Silene as a
368 model system in ecology and evolution. Heredity 103: 5–14.
369 Bertrand, Y. J. K., A. Petri, A.-C. Scheen, M. Töpel, and B. Oxelman, 2018 De novo
370 transcriptome assembly, annotation, and identification of low-copy number genes in the
371 flowering plant genus Silene (Caryophyllaceae). bioRxiv 290510.
372 Blavet, N., D. Charif, C. Oger-Desfeux, G. A. Marais, and A. Widmer, 2011 Comparative high-
373 throughput transcriptome sequencing and development of SiESTa, the Silene EST
374 annotation database. BMC Genomics 12: 376.
375 Bryant, D. M., K. Johnson, T. DiTommaso, T. Tickle, M. B. Couger *et al.*, 2017 A Tissue-
376 Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration
377 Factors. Cell Reports 18: 762–776.
378 Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+:
379 architecture and applications. BMC Bioinformatics 10: 421.

380 Casimiro-Soriguer, I., E. Narbona, M. L. Buide, J. C. del Valle, and J. B. Whittall, 2016
381 Transcriptome and Biochemical Analysis of a Flower Color Polymorphism in Silene
382 littorea (Caryophyllaceae). Front. Plant Sci. 7:.
383 Charlesworth, D., 2006 Evolution of Plant Breeding Systems. Current Biology 16: R726–R735.
384 Costich, D. E., T. R. Meagher, and E. J. Yurkow, 1991 A rapid means of sex identification
385 inSilene latifolia by use of flow cytometry. Plant Mol Biol Rep 9: 359–370.
386 Dagher-Kharrat, M. B., N. Abdel-Samad, B. Douaihy, M. Bourge, A. Fridlender *et al.*, 2013
387 Nuclear DNA C-values for biodiversity screening: Case of the Lebanese flora. Plant
388 Biosystems - An International Journal Dealing with all Aspects of Plant Biology 147:
389 1228–1237.
390 Darriba, D., G. L. Taboada, R. Doallo, and D. Posada, 2012 jModelTest 2: more models, new
391 heuristics and parallel computing. Nature Methods 9: 772–772.
392 Davis, S. L., and L. F. Delph, 2005 Prior Selfing and Gynomonoecy in Silene noctiflora L.
393 (Caryophyllaceae): Opportunities for Enhanced Outcrossing and Reproductive
394 Assurance. International Journal of Plant Sciences 166: 475–480.
395 Desfeux, C., S. Maurice, J. P. Henry, B. Lejeune, and P. H. Gouyon, 1996 Evolution of
396 reproductive systems in the genus Silene. Proc. Biol. Sci. 263: 409–414.
397 Doyle, J. J., and J. L. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh
398 leaf tissue. PHYTOCHEMICAL BULLETIN.
399 Erixon, P., and B. Oxelman, 2008 Whole-Gene Positive Selection, Elevated Synonymous
400 Substitution Rates, Duplication, and Indel Evolution of the Chloroplast clpP1 Gene.
401 PLOS ONE 3: e1386.

Garraud, C., B. Brachi, M. Dufay, P. Touzet, and J. A. Shykoff, 2011 Genetic determination of male sterility in gynodioecious Silene nutans. Heredity 106: 757–764.

Ghasemi, F. S., A. Jalili, and S. S. Mirzadeh Vaghefi, 2015 CHROMOSOME REPORT OF THREE SPECIES OF FLORA OF IRAN. 21: 165–168.

Gholipour, A., and M. Sheidai, 2010 Karyotype analysis and new chromosome number reports in Silene species (sect. Auriculatae, Caryophyllaceae). Biologia 65: 23–27.

Gnerre, S., I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton *et al.*, 2011 High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U.S.A. 108: 1513–1518.

Gordon, S. P., E. Tseng, A. Salamov, J. Zhang, X. Meng *et al.*, 2015 Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One 10:.

Gregory, T. R., J. A. Nicol, H. Tamm, B. Kullman, K. Kullman *et al.*, 2007 Eukaryotic genome size databases. Nucleic Acids Res 35: D332–D338.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk *et al.*, 2010 New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 59: 307–321.

Guo, W., F. Grewe, W. Fan, G. J. Young, V. Knoop *et al.*, 2016 Ginkgo and Welwitschia Mitogenomes Reveal Extreme Contrasts in Gymnosperm Mitochondrial Evolution. Mol Biol Evol 33: 1448–1460.

Hahn, M. W., S. V. Zhang, and L. C. Moyle, 2014 Sequencing, Assembling, and Correcting Draft Genomes Using Recombinant Populations. G3 (Bethesda) 4: 669–679.

Havird, J. C., P. Trapp, C. M. Miller, I. Bazos, and D. B. Sloan, 2017 Causes and Consequences of Rapidly Evolving mtDNA in a Plant Lineage. Genome Biol Evol 9: 323–336.

425 Havird, J. C., Whitehill Nicholas S., Snow Christopher D., and Sloan Daniel B., 2015
426 Conservative and compensatory evolution in oxidative phosphorylation complexes of
427 angiosperms with highly divergent rates of mitochondrial genome evolution. Evolution
428 69: 3069–3081.
429 Hestand, M. S., J. V. Houdt, F. Cristofoli, and J. R. Vermeesch, 2016 Polymerase specific error
430 rates and profiles identified by single molecule sequencing. Mutation
431 Research/Fundamental and Molecular Mechanisms of Mutagenesis 784–785: 39–45.
432 Jafari, F., S. Zarre, A. Gholipour, F. Eggens, R. K. Rabeler *et al.*, 2020 A new taxonomic
433 backbone for the infrageneric classification of the species-rich genus Silene
434 (Caryophyllaceae). TAXON 69: 337–368.
435 Katoh, K., and D. M. Standley, 2013 MAFFT Multiple Sequence Alignment Software Version 7:
436 Improvements in Performance and Usability. Mol Biol Evol 30: 772–780.
437 Kemal, Y., E. Minareci, and A. Çirpici, 2009 Karyotypic study on Silene, section Lasiostemones
438 species from Turkey. Caryologia 62: 134–141.
439 Klaas, A. L., and M. S. Olson, 2006 Spatial Distributions of Cytoplasmic Types and Sex
440 Expression in Alaskan Populations of Silene acaulis. International Journal of Plant
441 Sciences 167: 179–189.
442 Krasovec, M., M. Chester, K. Ridout, and D. A. Filatov, 2018 The Mutation Rate and the Age of
443 the Sex Chromosomes in Silene latifolia. Curr. Biol. 28: 1832-1838.e4.
444 Kreibich, J. A., 2010 *Using SQLite*. O'Reilly Media, Inc.
445 Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer, 2001 Predicting transmembrane
446 protein topology with a hidden Markov model: application to complete genomes. J. Mol.
447 Biol. 305: 567–580.

448 Kruckeberg, A. R., 1960 CHROMOSOME NUMBERS IN SILENE (CARYOPHYLLACEAE).
449 II. Madroño 15: 205–215.
450 Lagesen, K., P. Hallin, E. A. Rødland, H.-H. Stærfeldt, T. Rognes *et al.*, 2007 RNAmmer:
451 consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100–
452 3108.
453 Lan, T., T. Renner, E. Ibarra-Laclette, K. M. Farr, T.-H. Chang *et al.*, 2017 Long-read
454 sequencing uncovers the adaptive topography of a carnivorous plant genome. Proc Natl
455 Acad Sci U S A 114: E4435–E4441.
456 Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–
457 3100.
458 Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads.
459 EMBnet.journal.
460 McNeill, J., 1980 THE BIOLOGY OF CANADIAN WEEDS.: 46. Silene noctiflora L. Can. J.
461 Plant Sci. 60: 1243–1253.
462 Mirzadeh Vaghefi, S. S., and A. Jalili, 2019 CHROMOSOME NUMBERS OF SOME
463 VASCULAR PLANT SPECIES FROM IRAN. The Iranian Journal of Botany 25: 140–
464 144.
465 Mower, J. P., P. Touzet, J. S. Gummow, L. F. Delph, and J. D. Palmer, 2007 Extensive variation
466 in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol. Biol.
467 7: 135.
468 Mrackova, M., M. Nicolas, R. Hobza, I. Negrutiu, F. Monéger *et al.*, 2008 Independent origin of
469 sex chromosomes in two species of the genus Silene. Genetics 179: 1129–1133.

Muyle, A., N. Zemp, C. Deschamps, S. Mousset, A. Widmer *et al.*, 2012 Rapid De Novo Evolution of X Chromosome Dosage Compensation in Silene latifolia, a Plant with Young Sex Chromosomes. PLOS Biology 10: e1001308.

Nishimura, K., and K. J. van Wijk, 2015 Organization, function and substrates of the essential Clp protease system in plastids. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1847: 915–930.

Olson, M. S., and D. E. Mccauley, 2002 Mitochondrial Dna Diversity, Population Structure, and Gender Association in the Gynodioecious Plant Silene Vulgaris. Evolution 56: 253–262.

Ono, Y., K. Asai, and M. Hamada, 2013 PBSIM: PacBio reads simulator—toward accurate genome assembly. Bioinformatics 29: 119–121.

PacificBiosciences, 2020 *IsoSeq*. Pacific Biosciences.

Papadopulos, A. S. T., M. Chester, K. Ridout, and D. A. Filatov, 2015 Rapid Y degeneration and dosage compensation in plant sex chromosomes. PNAS 112: 13021–13026.

Pellicer, J., and I. J. Leitch, 2020 The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytologist 226: 301–305.

Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen, 2011 SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8: 785–786.

Popp, M., P. Erixon, F. Eggens, and B. Oxelman, 2005 Origin and Evolution of a Circumpolar Polyploid Species Complex in Silene (Caryophyllaceae) Inferred from Low Copy Nuclear RNA Polymerase Introns, rDNA, and Chloroplast DNA. Systematic Botany 30: 302–313.

492 Popp, M., and B. Oxelman, 2001 Inferring the History of the Polyploid Silene aegaea
493 (Caryophyllaceae) Using Plastid and Homoeologous Nuclear DNA Sequences. Molecular
494 Phylogenetics and Evolution 20: 474–481.
495 Popp, M., and B. Oxelman, 2007 Origin and evolution of North American polyploid Silene
496 (Caryophyllaceae). American Journal of Botany 94: 330–349.
497 Potter, S. C., A. Luciani, S. R. Eddy, Y. Park, R. Lopez *et al.*, 2018 HMMER web server: 2018
498 update. Nucleic Acids Res 46: W200–W204.
499 Rhoads, A., and K. F. Au, 2015 PacBio Sequencing and Its Applications. Genomics, Proteomics
500 & Bioinformatics 13: 278–289.
501 Rockenbach, K., J. C. Havird, J. G. Monroe, D. A. Triant, D. R. Taylor *et al.*, 2016 Positive
502 Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. Genetics 204: 1507–
503 1522.
504 Schatz, M. C., J. Witkowski, and W. R. McCombie, 2012 Current challenges in de novo plant
505 genome sequencing and assembly. Genome Biol 13: 243.
506 Siroký, J., M. A. Lysák, J. Dolezel, E. Kejnovský, and B. Vyskot, 2001 Heterogeneity of rDNA
507 distribution and genome size in Silene spp. Chromosome Res. 9: 387–393.
508 Slancarova, V., J. Zdanska, B. Janousek, M. Talianova, C. Zschach *et al.*, 2013 Evolution of Sex
509 Determination Systems with Heterogametic Males and Females in Silene. Evolution 67:
510 3669–3677.
511 Sloan, D. B., A. J. Alverson, J. P. Chuckalovcak, M. Wu, D. E. McCauley *et al.*, 2012a Rapid
512 Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria
513 with Exceptionally High Mutation Rates. PLOS Biology 10: e1001241.

514 Sloan, D. B., S. R. Keller, A. E. Berardi, B. J. Sanderson, J. F. Karpovich *et al.*, 2012b De novo
515 transcriptome assembly and polymorphism detection in the flowering plant Silene
516 vulgaris (Caryophyllaceae). Molecular Ecology Resources 12: 333–343.
517 Sloan, D. B., D. A. Triant, N. J. Forrester, L. M. Bergner, M. Wu *et al.*, 2014 A recurring
518 syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae
519 (Caryophyllaceae). Molecular Phylogenetics and Evolution 72: 82–89.
520 Städler, T., and L. F. Delph, 2002 Ancient mitochondrial haplotypes and evidence for intragenic
521 recombination in a gynodioecious plant. PNAS 99: 11730–11735.
522 Stanne, T. M., E. Pojidaeva, F. I. Andersson, and A. K. Clarke, 2007 Distinctive Types of ATP-
523 dependent Clp Proteases in Cyanobacteria. J. Biol. Chem. 282: 14394–14402.
524 Taylor, D. R., M. S. Olson, and D. E. McCauley, 2001 A Quantitative Genetic Analysis of
525 Nuclear-Cytoplasmic Male Sterility in Structured Populations of Silene vulgaris.
526 Genetics 158: 833–841.
527 UniProt: a hub for protein information, 2015 Nucleic Acids Res 43: D204–D212.
528 Wang, B., V. Kumar, A. Olson, and D. Ware, 2019 Reviving the Transcriptome Studies: An
529 Insight Into the Emergence of Single-Molecule Transcriptome Sequencing. Front. Genet.
530 10:.
531 Wang, B., E. Tseng, M. Regulski, T. A. Clark, T. Hon *et al.*, 2016 Unveiling the complexity of
532 the maize transcriptome by single-molecule long-read sequencing. Nature
533 Communications 7: 1–13.
534 Weirather, J. L., M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano *et al.*, 2017 Comprehensive
535 comparison of Pacific Biosciences and Oxford Nanopore Technologies and their
536 applications to transcriptome analysis. F1000Res 6:.

537 Wenger, A. M., P. Peluso, W. J. Rowell, P.-C. Chang, R. J. Hall *et al.*, 2019 Accurate circular
538 consensus long-read sequencing improves variant detection and assembly of a human
539 genome. Nature Biotechnology 37: 1155–1162.
540 Williams, A. M., G. Friso, K. J. van Wijk, and D. B. Sloan, 2019 Extreme variation in rates of
541 evolution in the plastid Clp protease complex. The Plant Journal 98: 243–259.
542 Wu, Z., J. M. Cuthbert, D. R. Taylor, and D. B. Sloan, 2015 The massive mitochondrial genome
543 of the angiosperm Silene noctiflora is evolving by gain or loss of entire chromosomes.
544 PNAS 112: 10185–10191.
545 Wu, Z., and D. B. Sloan, 2019 Recombination and intraspecific polymorphism for the presence
546 and absence of entire chromosomes in mitochondrial genomes. Heredity 122: 647–659.
547 Xu, Z., R. J. Peters, J. Weirather, H. Luo, B. Liao *et al.*, 2015 Full-length transcriptome
548 sequences and splice variants obtained by a combination of sequencing platforms applied
549 to different root tissues of Salvia miltiorrhiza and tanshinone biosynthesis. The Plant
550 Journal 82: 951–961.
551 Yildiz, K., E. Minareci, A. Çirpici, and M. Y. Dadandı, 2008 A karyotypic study on Silene,
552 section Siphonomorpha species of Turkey. Nordic Journal of Botany 26: 368–374.
553 Yu, A. Y. H., and W. A. Houry, 2007 ClpP: A distinctive family of cylindrical energy-dependent
554 serine proteases. FEBS Letters 581: 3749–3757.
555 Zhao, L., H. Zhang, M. V. Kohnen, K. V. S. K. Prasad, L. Gu *et al.*, 2019 Analysis of
556 Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-
557 Based Direct RNA Sequencing. Front Genet 10: 253.

Table 1: Genome sizes determined by flow cytometry

				Mean Genome Size	
Species	**Population**	**Location**	**Samples (pg/2C)**	**pg/2C**	**Gb/1C**
Silene noctiflora	OPL*	Opole, Poland	5.65, 5.61, 5.46, 5.44	5.54	2.71
	OSR	Giles County, VA	5.75, 5.61	5.68	2.78
	BRP	Nelson County, VA	5.63, 5.57	5.60	2.74
Silene conica	ABR	Abruzzo, Italy	1.92, 1.92, 1.88	1.91	0.93
Silene vulgaris	S9L	Giles County, VA	2.19, 2.16	2.18	1.07
Silene latifolia	UK2600	Bedford County, VA	5.46, 5.45	5.46	2.67

*The *S. noctiflora* OPL population was used for Iso-Seq, genome assembly, and karyotyping

Figure 1: *Silene noctiflora*, also known as the night-flowering catchfly.

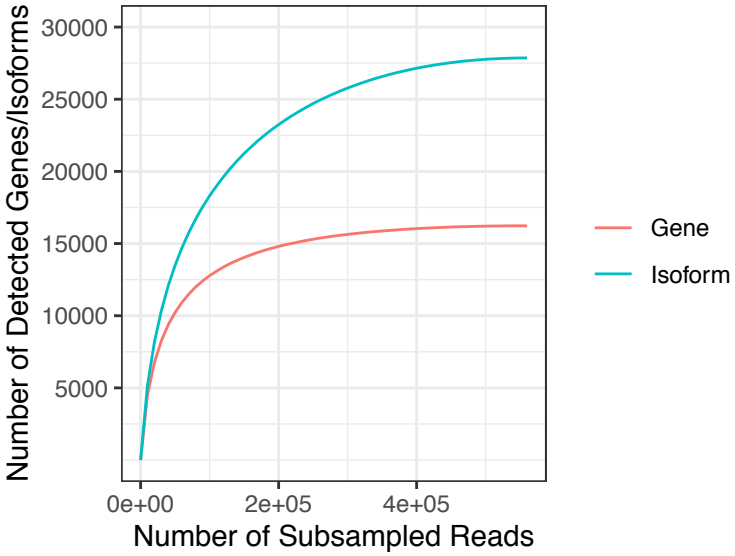

Figure 2: Rarefaction analysis of the *S. noctiflora* Iso-Seq transcriptome. Curves for both genes (red) and isoforms (blue) are depicted.

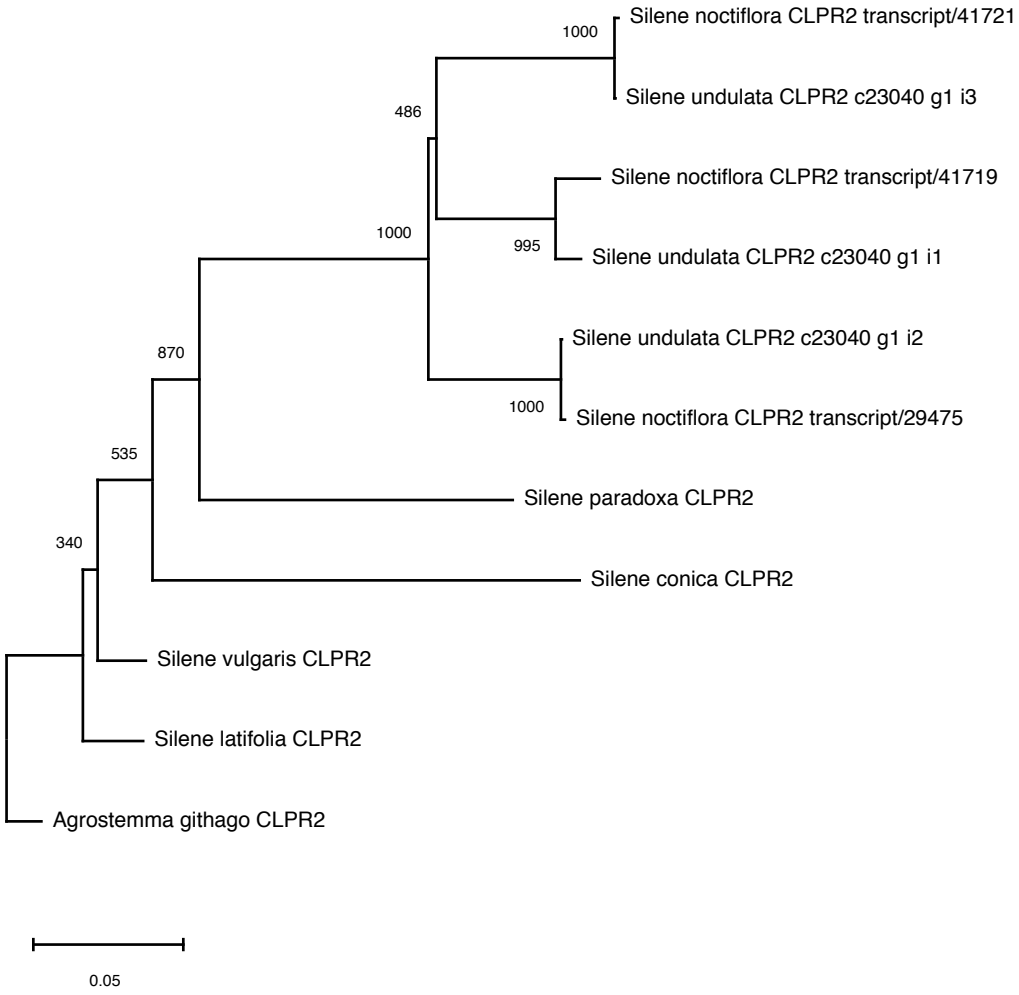

Figure 3: Phylogenetic analysis of *CLPR2* genes in *S. noctiflora* and related species. Branch lengths represent nucleotide sequence divergence. This tree was rooted on the *Agrostemma githago* sequence. The placement of *S. paradoxa* is in conflict with the species tree (Jafari *et al.* 2020), likely due to long branch attraction and the multiple independent evolutionary rate accelerations in this protein across *Silene* (Rockenbach *et al.* 2016).

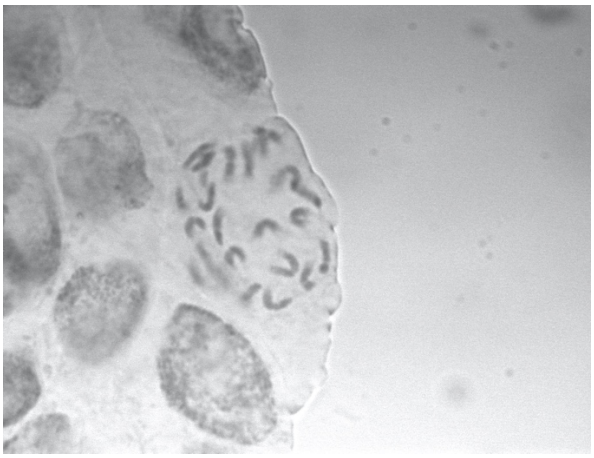

Figure 4: Micrograph verifying the diploidy of *Silene noctiflora* at 100× magnification.