bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Chromosome-level genome assembly for the angiosperm *Silene conica*

Peter D. Fields[1,2,*], Melody M. Weber[1], Gus Waneka[1], Amanda K. Broz[1], Daniel B. Sloan[1,*]

[1]Department of Biology, Colorado State University, Fort Collins, Colorado, USA.

[2]Mammalian Genetics, The Jackson Laboratory, Bar Harbor, Maine, USA

*Authors for correspondence: peter.fields@jax.org and dan.sloan@colostate.edu

Genome Resource

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

The angiosperm genus *Silene* has been the subject of extensive study in the field of ecology and evolution, but the availability of high-quality reference genome sequences has been limited for this group. Here, we report a chromosome-level assembly for the genome of *Silene conica* based on PacBio HiFi, Hi-C and Bionano technologies. The assembly produced 10 scaffolds (one per chromosome) with a total length of 862 Mb and only ~1% gap content. These results confirm previous observations that *S. conica* and its relatives have a reduced base chromosome number relative to the genus's ancestral state of 12. *Silene conica* has an exceptionally large mitochondrial genome (>11 Mb), predominantly consisting of sequence of unknown origins. Analysis of shared sequence content suggests that it is unlikely that transfer of nuclear DNA is the primary driver of this mitochondrial genome expansion. More generally, this assembly should provide a valuable resource for future genomic studies in *Silene*, including comparative analyses with related species that recently evolved sex chromosomes.

Significance

Whole-genome sequences have been largely lacking for species in the genus *Silene* even though these flowering plants have been used for studying ecology, evolution, and genetics for over a century. Here, we address this gap by providing a high-quality nuclear genome assembly for *S. conica*, a species known to have greatly accelerated rates of sequence and structural divergence in its mitochondrial and plastid genomes. This resource will be valuable in understanding the coevolutionary interactions between nuclear and cytoplasmic genomes and in comparative analyses across this highly diverse genus.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Silene (Caryophyllaceae) is a diverse angiosperm genus that encompasses over 800 species and has been the subject of extensive study in ecology and evolutionary genetics (Bernasconi, et al. 2009; Jafari, et al. 2020). Species in this genus have served as models for investigating topics such as breeding system and sex chromosome evolution (Nicolas, et al. 2005; Kejnovsky and Vyskot 2010; Muyle, et al. 2012; Krasovec, et al. 2018; Martin, et al. 2019), host-pathogen interactions (Alexander and Antonovics 1988; Kephart, et al. 2006; Le Gac, et al. 2007), heavy metal tolerance (Antonovics, et al. 1971; Bringezu, et al. 1999; Papadopulos, et al. 2021), species invasion (Blair and Wolfe 2004; Keller and Taylor 2008; Keller, et al. 2014), and cytoplasmic/cytonuclear genetics (Taylor, et al. 2001; Touzet and Delph 2009; Sloan, Alverson, Chuckalovcak, et al. 2012; Fields, et al. 2014; Havird, et al. 2015). Even though many of these topics are directly rooted in areas of genome biology, the availability of high-quality genomic resources for *Silene* has remained limited. An extensive number of RNA-seq projects and transcriptome assemblies have been performed in *Silene* (e.g., Blavet, et al. 2011; Havird, et al. 2017; Jafari, et al. 2020; Warren, et al. 2022), but initial whole-genome sequencing efforts in *Silene* were limited to a mix of short-read and early-generation long-read technologies, resulting in highly fragmented and incomplete assemblies (Papadopulos, et al. 2015; Krasovec, et al. 2018; Williams, et al. 2021). Only recently was the first chromosome-level genome assembly produced for a member of this genus (Yue, et al. 2023).

To date, genome assemblies have been published for only two *Silene* species: *S. latifolia* (Papadopulos, et al. 2015; Krasovec, et al. 2018; Yue, et al. 2023) and *S. noctiflora* (Williams, et al. 2021). These two species are found near the high end of the known range of 0.7-3.3 Gb for genome sizes of diploid species in this genus (Pellicer and Leitch 2019). At the other end of the spectrum, *S. conica* has one of the smallest estimated genome sizes (0.9 Gb) in *Silene* (Williams, et al. 2021). *Silene conica* and its relatives within section *Conoimorpha* also have a reduced chromosome number (2n=20) compared to the ancestral state (2n=24) for *Silene* (Bari 1973). In contrast to its small nuclear genome size, *S. conica* has one of the largest mitochondrial genomes of any eukaryote at >11 Mb, containing over 99% non-coding content of largely unknown origin (Sloan, Alverson, Chuckalovcak, et al. 2012). The cytoplasmic genomes in *S. conica* also exhibit other distinctive features, including accelerated evolutionary rates, major structural changes, and extensive gene loss (Erixon and Oxelman 2008; Sloan, Alverson, Chuckalovcak, et al. 2012; Sloan, Alverson, Wu, et al. 2012). As such, this species has been a valuable model for studying how changes in cytoplasmic genomes can spur cytonuclear coevolution (Rockenbach, et al. 2016; Havird, et al. 2017; Abdel-Ghany, et al. 2022).

With the continuing improvement of DNA sequencing technologies, it is becoming increasingly possible to generate chromosome-level assemblies, even for complex eukaryotic

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

nuclear genomes like those of plants (Jiao and Schneeberger 2017; Belser, et al. 2018; Shirasawa, et al. 2021). In particular, the advent of Pacific Bioscience (PacBio) HiFi sequencing has resulted in a major step forward, providing single-molecule long reads (~15-25 kb) at high accuracy (>99%) (Wenger, et al. 2019). Here, we report a chromosome-level assembly of the *S. conica* genome generated from HiFi sequencing in combination with Bionano optical mapping (Lam, et al. 2012) and Hi-C scaffolding (Belton, et al. 2012).

Results & Discussion

Chromosome-Level Genome Assembly

PacBio HiFi sequencing of *S. conica* total-cellular DNA produced 2.23M circular consensus sequence (CCS) reads with an average length of 14.9 kb and a total yield 33.18 Gb. Our base assembly of these reads produced by *hifiasm* was a total of 938 Mb in length, with an N50 length of 14.95 Mb (n=20), a maximum contig length of 46.80 Mb, and a total of 1367 contigs. Following the application of purge_haplotigs, we saw improvement in assembly metrics, including an N50 length of 16.23 Mb (n=18), a total contig number of 217, and a total assembly length of 869 Mb, which is slightly shorter than the previous estimate of 930 Mb from flow cytometry (Williams, et al. 2021). This base assembly was then used as an input for Bionano Access.

A single round of hybrid scaffolding with a Bionano optical map resulted in a large improvement in overall genome contiguity. Specifically, while the overall genome did not change substantially in length (862 Mb), Bionano scaffolding produced 16 scaffolds with an N50 scaffold length of 74.70 Mb (n=5), a longest scaffold of 122.47 Mb, and 65 gaps totaling only 6.42 Mb. This small amount of gapped content (~1% of the assembly) compares favorably to the level of contiguity achieved for most chromosome-level assemblies of plant genomes (Shirasawa, et al. 2021). We then proceeded with analysis of Hi-C data to generate chromosomal scaffolds, although six of the scaffolds already appeared to represent near or whole chromosomes at this stage.

Visualization of the Hi-C based contact network (Figure S1) provided support for joining the above Bionano scaffolds into the expected number of 10 chromosomes (Bari 1973). However, this visualization also revealed that the Bionano scaffold containing Chr1 was misassembled and contained a large portion of Chr6 (Figure S1). The point at which these two regions were joined corresponded to a large gap in the Bionano scaffold that was associated with highly repetitive ribosomal DNA. After breaking this misassembly and joining the Bionano scaffolds based on the Hi-C contract map, our analysis resulted in 10 chromosome-level scaffolds with 71 gaps, a total length of 862 Mb, and 37.0% GC content (Figure S2).

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Genome Annotation

We used BUSCO (Manni, et al. 2021) to assess the biological completeness of our genome assembly. At present, there is no Caryophyllaceae-specific ancestral gene set, so we used the embryophyta_odb10 dataset for the BUSCO analysis, which produced a completeness score of 97.6. We detected 1586 of the 1614 BUSCO genes searched (1506 complete and single-copy, 70 complete but duplicated, and 10 fragmented), while 28 BUSCOs were missing. An ancestral gene set derived from species that better represent genomes of the Caryophyllaceae might result in a slightly higher overall BUSCO score.

In order to maximize our genome annotation completeness, we relied on a plurality of approaches. Specifically, we used a combination of MAKER2 (Holt and Yandell 2011), Funannotate (Palmer and Stajich 2020), and BRAKER (Hoff et al. 2019) individually, followed by the merging and collapsing of redundant annotations using AGAT (Dainat et al. 2023). For MAKER2, we utilized an iterative application of the pipeline to annotate the *S. conica* genome. Following the first round of MAKER2, which relied on protein and transcript hints alone, we identified a total of 51,311 putative gene models. Our second round of MAKER2, which included the application of AUGUSTUS and SNAP *ab initio* hints, as well as the gene models from MAKER2 round one, resulted in 57,309 putative gene models. Our third round of MAKER2, which incorporated GeneMark *ab initio* hints, as well as the gene models from MAKER2 round two, resulted in 56,305 putative gene models. Finally, the fourth round of MAKER2 included the application of AUGUSTUS and SNAP *ab initio* hints, which were trained off gene models resulting from MAKER2 round three, and also included gene models from MAKER2 round three. This final round resulted in 58,409 putative gene models, which were filtered based upon an annotation edit distance (AED) score threshold of ≤ 1, yielding a total of 47,262 putative gene models. Our Funannotate annotation resulted in a total of 55,798 putative gene models. Finally, BRAKER annotation resulted in a total of 56,992 putative gene models. Our merging and de-duplication of gene models using AGAT resulted in a total of 63,211 putative gene models. The resultant BUSCO score of our annotation is nearly as good as the overall genome with a completeness score of 93.2%, suggesting our automated annotation process was highly effective. Manual curation would likely improve the overall accuracy of individual gene models, and filter out spurious annotations that may have inflated the number of identified gene models.

Karyotyping has indicated that the *S. conica* chromosomes are metacentric or submetacentric (Bari 1973). Accordingly, the chromosomes show the typical pattern of higher gene densities and lower CpG methylation rates at the ends of chromosome arms relative to the middle of the chromosome (Figure 1). However, it should be noted that this expectation was used to orient scaffolds during the final Hi-C based joining for Chr3 and Chr8 (see Methods). Therefore, the gene density patterns for these two chromosomes do not provide any further independent evidence for a typical metacentric structure.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

An analysis of interspersed repeat content in the genome found that transposable elements constituted approximately three-quarters of all sequence content (Table S1). Indeed, long terminal repeat (LTR) retrotransposons by themselves were estimated to account for more than half of the genome.

DNA transfer between nuclear and cytoplasmic genomes

The insertion of mitochondrial and plastid DNA into the nuclear genome (known as *numts* and *nupts*, respectively) is a widespread phenomenon across eukaryotes (Hazkani-Covo, et al. 2010; Zhang, et al. 2020). In plants, the movement of DNA in the opposite direction – from the nucleus to cytoplasmic genomes – is also common for mitochondria (Goremykin, et al. 2012; Qiu, et al. 2014) but very rare for plastids (Smith 2014). To characterize the extent of intracellular DNA transfer in *S. conica*, we used BLAST searches to compare the mitochondrial and plastid genomes against our nuclear genome assembly. We found regions with significant similarity to the cytoplasmic genome were widely distributed across the nuclear genome (Figure 2). A total of 1938 kb of nuclear DNA sequence (0.22% of the nuclear genome) shared similarity with the mitochondrial genome, and 186 kb shared similarity with the plastid genome (0.02% of the entire nuclear genome). This amount of shared content is well within the range observed in other species. Angiosperm nuclear genomes have been estimated to share anywhere from 142 kb to 11,420 kb of sequence with the mitochondrial genome (covering between 0.03% and 2.08% of the assembled nuclear genomes) and 36 kb to 9830 kb with the plastid genome (0.01% to 1.49% of the assembled nuclear genomes) (Zhang, et al. 2020).

The shared sequences in *S. conica* represented a total of 2863 kb (25.3%) in the mitochondrial genome and 103 kb (70.3%) in the plastid genome. Note that the raw totals represented by these shared sequences differ between the nuclear and cytoplasmic genomes because of the differences in the extent to which the shared content is repeated within each of the genomes. There were two regions of the nuclear genome that were especially rich in DNA shared with the mitochondrial genome (Figure 2): the ~1405 kb region Chr3 from position 49.307 Mb to 50.712 Mb and the ~335 kb region on Chr10 from position 47.115 Mb to 47.450 Mb. Although synteny in these regions of shared sequence was highly fragmented relative to the mitochondrial genome, much of the sequence retained very high nucleotide identity (>99% in many cases) between the two genomes, implying that the transfer(s) responsible for the shared content occurred relatively recently.

Given the rarity with which foreign DNA is inserted into plastid genomes (Smith 2014), it is reasonable to interpret any regions of shared sequence between nuclear and plastid genomes as being ultimately plastid in origin (i.e. *nupts*). However, because transfer between mitochondrial and nuclear genomes is bidirectional in plants, it is more challenging to polarize the movement of DNA

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

shared between these two genomes. The fact that approximately one-quarter of the mitochondrial genome content in *S. conica* is shared with the nucleus could mean that nuclear DNA is a major contributor to mitochondrial genome expansion in this lineage (Sloan, Alverson, Chuckalovcak, et al. 2012), as suggested in other plant mitochondrial genomes (Goremykin, et al. 2012). However, it is possible or even likely that most of this shared content results from transfers in the opposite direction given the enormous quantity of *numts* often found in plant nuclear genomes (Zhang, et al. 2020; Fields, et al. 2022). In principle, the direction of transfer could be polarized by identifying homologous content in either the nuclear or mitochondrial genomes of other plant species. However, this is not currently feasible because only a tiny fraction of the *S. conica* mitochondrial genome has detectable homology with any known sequence (Sloan, Alverson, Chuckalovcak, et al. 2012). As such, deciphering the origins of these shared sequences will likely require additional genome sequences from close relatives within *Silene* section *Conoimorpha* or population genomic sampling within *S. conica*.

Materials and Methods

Plant material and growth conditions

Seeds from the *S. conica* ABR (Abruzzo, Italy) line were sown in ProMix BX soil mix. The seeds were collected from a sibling of the plant previously used for Iso-Seq analysis (Warren, et al. 2022), and they were the same batch used for ddPCR validation of genome copy number in Broz, et al. (2021). This line had gone through four generations of selfing in the lab and likely started with low levels of heterozygosity given our observation that *S. conica* readily self-fertilizes without any intervention. They were germinated in a growth room under 10-hr short-day lighting conditions and switched to 16-hr long-day conditions after 3 weeks of growth. Lighting was provided with Fluence LEDs at ~100 $\mu E\ m^{-1}\ s^{-1}$. Plants were initially each grown in a 2-inch pot and then transplanted to 4-inch pots after 4 months of growth. They were watered on an as-needed basis and treated with dilute Miracle-Gro after 5 weeks of growth. Spinosad and neem oil were applied to limit observed outbreaks of thrips and other potential plant pests. After 5 months of growth, 2.3 g of tissue from young rosette leaves was harvested from a single individual and flash frozen in liquid N_2 for subsequent DNA extraction and PacBio HiFi sequencing. The same plant was allowed to regrow leaf tissue for another 1.5 months, at which point 3.1 g of young rosette leaf tissue and developing shoots (the plant had begun to bolt) were harvested and flash frozen in liquid N_2 for subsequent Bionano optical mapping. Finally, after two additional weeks of regrowth, 0.6 g of young leaf tissue was harvested from the same plant and immediately used to perform Hi-C library preparation.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

PacBio HiFi sequencing

Flash-frozen leaf tissue (see above) was shipped on dry ice to the Arizona Genomics Institute at the University of Arizona. DNA was extracted with a modified CTAB protocol (Doyle and Doyle 1987). Extracted DNA was analyzed with pulsed-field gel electrophoresis to confirm that it was high molecular weight. A Covaris G-Tube was then used to shear 10 μg of DNA to a size range of ~10-30 kb followed by bead purification with PB Beads (PacBio). The HiFi sequencing library was constructed following manufacturer's protocols using SMRTbell Express Template Prep Kit 2.0. The final library was size selected with a range of 10-25 kb on a Blue Pippin (Sage Science) using S1 marker. The recovered final library was quantified with a Qubit HS dsDNA kit, and size was confirmed with an Agilent Femto Pulse system. Sequencing was performed on PacBio Sequel II, using standard manufacturer's protocols for the Sequel II Sequencing Kit 2.0. The library was sequenced on 2 SMRT Cells (8M) in CCS mode for 30 hours. Analysis was performed with SMRT Link v10.1 software, requiring a minimum of 3 passes for CCS generation.

Hi-C library preparation and sequencing

Fresh leaf tissue (see above) was used to generate a Hi-C library with the Proximo Hi-C Plant Kit from Phase Genomics (v4.0 protocol). Input material was ground in a mortar and pestle under liquid N_2. The library was amplified with 12 cycles of PCR, and a total of 234M read pairs (2×150 bp) were generated on an Illumina NovaSeq 6000 platform at the Genomics and Microarray Core at the University of Colorado Anschutz Medical Campus.

Bionano optical mapping

Flash-frozen leaf tissue (see above) was shipped on dry ice to the McDonnell Genome Institute at Washington University in St. Louis. DNA was isolated with the Bionano plant tissue hybrid protocol (liquid N_2 grinding and tissue ruptor), including density gradient purification of nuclei, which were then embedded in agarose plugs prior to DNA extraction. Labeling was performed with a Bionano DLS Kit followed by analysis on a Bionano Saphyr platform, generating an estimated genome coverage of 214×. Computational analysis was performed with Bionano Access software.

Hifiasm de novo assembly

We used the *hifiasm* v.0.15.2-r334 (Cheng, et al. 2021) assembler to generate contigs from PacBio HiFi sequencing data. Given that the focal genotype was relatively inbred, we included the '-l0' flag as part of the assembler configuration, thereby disabling automatic duplication purging. Additional purging of contigs that result from individual heterozygosity, so-called 'haplotigs', was done with the purge_haplotig v.1.1.1 package (Roach, et al. 2018). The assembly graph generated by *hifiasm* was converted to a set of contigs in multi-fasta format using AWK (Aho, et al. 1987) as described at

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

https://github.com/chhylp123/hifiasm. We assessed focal species containment by using BlobTools2 (Challis, et al. 2020) to detect assembly contamination by non-focal species. In order to quantify biological completeness of our contig set, we used the package BUSCO v.4.1.4 (Manni, et al. 2021) with the eudicotyledons_odb10 ancestral lineage dataset.

Scaffolding with Hi-C and Bionano

To scaffold contigs generated by *hifiasm*, we used a paired approach of Bionano optical map construction and Hi-C scaffolding. Research has suggested that higher accuracy scaffolding can be attained by first applying Bionano hybrid scaffolding (Bickhart, et al. 2017). A Bionano optical map as well as the construction of hybrid scaffolds was made using the Bionano Access software package. We used HiC-Pro v.3.1.0 (Servant, et al. 2015) and the HiTC v1.42.0 R package (Servant, et al. 2012) to visualize the Hi-C contact map for the resulting Bionano scaffolds. Inspection of this contact map identified Bionano scaffolds that could be joined into chromosome-level scaffolds (Figure S1), as well as one misassembly (see above). Although the Hi-C data provided a clear signal for connecting Bionano scaffolds, it did not provide compelling evidence for how those scaffolds should be oriented within chromosomes. Therefore, in this final assembly step, which connected 11 Bionano scaffolds into the remaining four chromosomes, we determined order and orientation based on the expectations that telomeric repeat sequences should be placed at chromosome ends and that annotated gene density should be higher at the chromosome ends than internal centromeric regions. The final Hi-C based scaffolding was performed with a custom Perl script. For Bionano scaffolding, gap sizes were estimated based on the optical map and include inferred locations of the interspersed 6-bp Bionano nicking sites. For the seven scaffold connections inferred from Hi-C data, we used an arbitrary gap size of 100 bp.

Genome annotation

For annotation, we relied on a combination of protein and transcript evidence constructed from PacBio Iso-Seq and bulk Illumina RNA-seq for our focal species as well as related species in the tribe *Sileneae*. Specifically, we used the Iso-Seq transcriptome data for *S. conica*, *S. latifolia*, *S. noctiflora*, *S. vulgaris*, and *Agrostemma githago* described in Williams, et al. (2021) and Warren, et al. (2023) and Illumina RNA-seq from Havird et al. (2017). Because we relied on multiple annotation approaches, the way these datasets were incorporated differed slightly. To generate protein evidence as input for the different annotation approaches, we used TransDecoder v.5.5.0 to identify the most likely protein-coding regions for individual transcripts in each of the Iso-Seq datasets. To reduce redundancy in our total protein hint dataset, we combined proteins for each individual species and ran CD-HIT v.4.8.1 (Li and Godzik 2006). The resulting protein fasta file was used as the protein hint dataset for all annotation approaches.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Our genome was first soft-masked using a *de novo* generated repeat library created with RepeatModeler2 (Flynn et al. 2020). Only instances of known transposable elements were masked in order to avoid the false masking of genic regions. Transposable element and interspersed repeat content was summarized for visualization with EDTA v2.0.1 (Ou, et al. 2019), using the --anno 1 and --sensitive 1 options, though these annotations were not used as part of the gene annotation process described below.

We provided Funannotate (Palmer and Stajich 2020) with both Illumina RNA-seq and PacBio Iso-Seq data as part of the *train* function, which utilizes a combination of Trinity (Grabherr et al. 2011) and PASA (Haas et al. 2003) to assemble high-quality transcripts. Next, we used the *predict* function, which utilizes the transcripts generated as part of the *train* function plus the protein evidence described above to parameterize the *ab initio* gene prediction software AUGUSTUS (Stanke et al. 2008), which, combined with the alignment of transcript evidence, is then used by Evidence Modeler (Haas et al. 2008) to generate high-quality, consensus gene models.

In order to identify gene models with the BRAKER (Hoff et al. 2019) pipeline we followed the tutorial described at https://github.com/Gaius-Augustus/BRAKER/blob/master/docs/long_reads/long_read_protocol.md. Specifically, after aligning our Illumina short-read data to our genome with STAR (Dobin et al. 2013), we used the resulting BAM file as an input to BRAKER1 which then uses a combination of GeneMark-ES/ET/EP (Bruna et al. 2020) and AUGUSTUS to generate gene models. Next, we used our protein evidence to generate a second set of gene models using BRAKER2. Finally, following the collapsing of redundant transcripts in our PacBio Iso-Seq data for *S. conica* using cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake), we used TSEBRA (Gabriel et al. 2021) to both compare the gene models generated with BRAKER1 and BRAKER2 to our Iso-Seq based gene models and also retain the best amongst the three sets of evidence.

Finally, we used the full set of transcripts for *S. conica* and the optimized AUGUSTUS models generated as part of the Funannotate pipeline and the protein evidence described before, as inputs for the MAKER2 pipeline (Holt and Yandell 2011). The full set of configuration files used for four separate iterations of MAKER2 are available at on Github (https://github.com/Sloan-Lab/Silene_conica_genome_project). We included a separate mapping iteration, two iterations with a combination of AUGUSTUS and SNAP (Johnson et al. 2008), and an iteration with GeneMark-ES/ET/EP (Bruna et al. 2020). The resultant gene models were filtered to retain those which had an AED score ≤ 1.

We used the software AGAT (Dainat et al. 2023), specifically the function *agat_sp_merge_annotations.pl*, in order to merge and de-duplicate annotations generated by the three separate approaches. We used the AGAT function *agat_sp_keep_longest_isoform.pl* to remove isoforms from our annotation.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Analysis of DNA transfer between nuclear and cytoplasmic genomes

To identify sequences transferred between the nuclear and cytoplasmic genomes, we searched published *S. conica* mitochondrial and plastid genome sequences (Sloan, Alverson, Chuckalovcak, et al. 2012; Sloan, Alverson, Wu, et al. 2012) against our nuclear genome assembly, using NCBI BLASTN v2.12.0+ with the *-task blastn* option. BLAST hit locations and the percentage of the nuclear and cytoplasmic genomes that were covered by hits (e-value threshold of 1e-6) were summarized with custom Perl scripts (https://github.com/Sloan-Lab/Silene_conica_genome_project). A sliding window analysis was also performed to summarize the percentage of shared sequence in 1-Mb windows (with a 250-kb step size) along the length of the nuclear chromosomes. Coverage values and individual BLAST hits were visualized with a custom R script (https://github.com/Sloan-Lab/Silene_conica_genome_project). Only hits with a minimum length of 300 bp were visualized with individual tick marks, but all hits meeting the e-value threshold (1e-6) were used for calculating and visualizing coverage percentages. These calculations were performed after removing hits from two very large regions of ribosomal DNA repeats (one on Chr6 and the other on Chr8) that result from ancient similarity between nuclear, mitochondrial, and plastid rRNA genes (rather than recent transfers between genomes).

Methylation analysis

PacBio HiFi sequencing data can also be used to detect some types of epigenetic modifications. We used the ccsmeth v.0.3.2 (Ni, et al. 2022) package to detect 5-methylcytosine base modifications in a CpG context (5mCpGs). Specifically, putative per-base modification information was first detected using the PacBio software CCS v.6.4.0 (flag '--hifi-kinetics'; https://github.com/PacificBiosciences/ccs) followed by alignment to the target genome using pbmm2 v1.7.0 (https://github.com/PacificBiosciences/pbmm2). ccsmeth then applies a deep-learning-based model (here, *model_ccsmeth_5mCpG_aggregate_attbigru_b11.v2.ckpt*) to infer methylation state across the target genome. Analyses with ccsmeth were done using a NVIDIA RTX 3090 graphical processing unit (GPU).

Data availability

The raw data (PacBio HiFi reads, Illumina-based Hi-C reads, and Bionano optical map data), genome assembly, and annotation can be accessed under the NCBI BioProject PRJNA904366 (assembly version 2: JAQQAY000000000.2). Annotation data can be found in the Zenodo repository at the following DOI: 10.5281/zenodo.8223290. All scripts for bioinformatic analyses are available at https://github.com/Sloan-Lab/Silene_conica_genome_project.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Acknowledgements

This project was supported bv the National Science Foundation (MCB-2048407) and the National Institutes of Health (R35 GM148134).

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S1. Interspersed repeat content in the *S. conica* genome

	No. of Elements	Total Length (Mb)	Genome %
LTR Retrotransposons	**543123**	**472.03**	**54.77**
Ty1/Copia	157792	185.57	21.53
Gypsy/DIRS1	167710	181.10	21.01
Other	217621	105.36	12.23
DNA Transposons	**584735**	**152.27**	**17.67**
Unclassified	**49601**	**8.28**	**0.96**
Total Interspersed Repeats	**1177459**	**632.58**	**73.40**

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

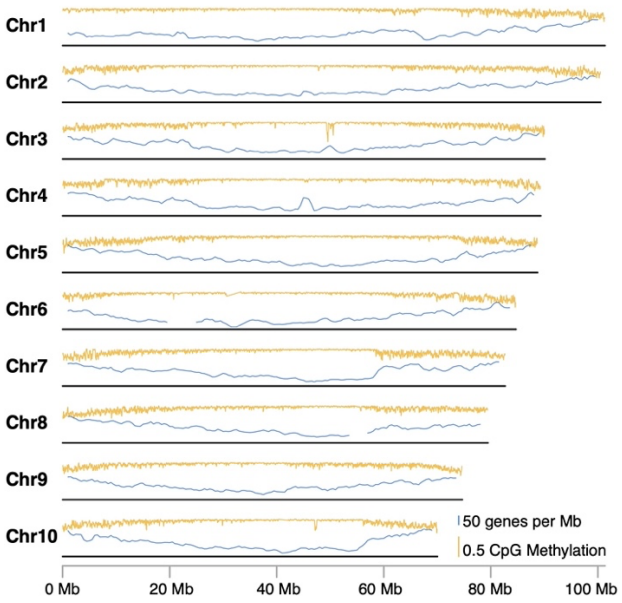


Figure 1. Summary of CpG methylation rates and gene density across the length of each chromosome in the *S. conica* genome. The orange trace represents a sliding window summary (100-kb window size and 25-kb step size) of the proportion of CpG sites inferred to be methylated from analysis of PacBio read data. The blue trace represents a sliding window summary (2-Mb window size and 500-kb step size) of the average number of annotated genes per Mb. The localized dips in estimated methylation rates on Chr3 and Chr10 correspond to the positions of large *numt* insertions (Figure 3). CpG methylation generally does not occur in mitochondrial DNA, so these dips are likely an artifact result from incorrectly mapping true mitochondrial reads to these *numt* regions (Fields, et al. 2022). Gene density estimates are not reported for the regions on Chr6 and Chr8 with tandemly repeated ribosomal DNA.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

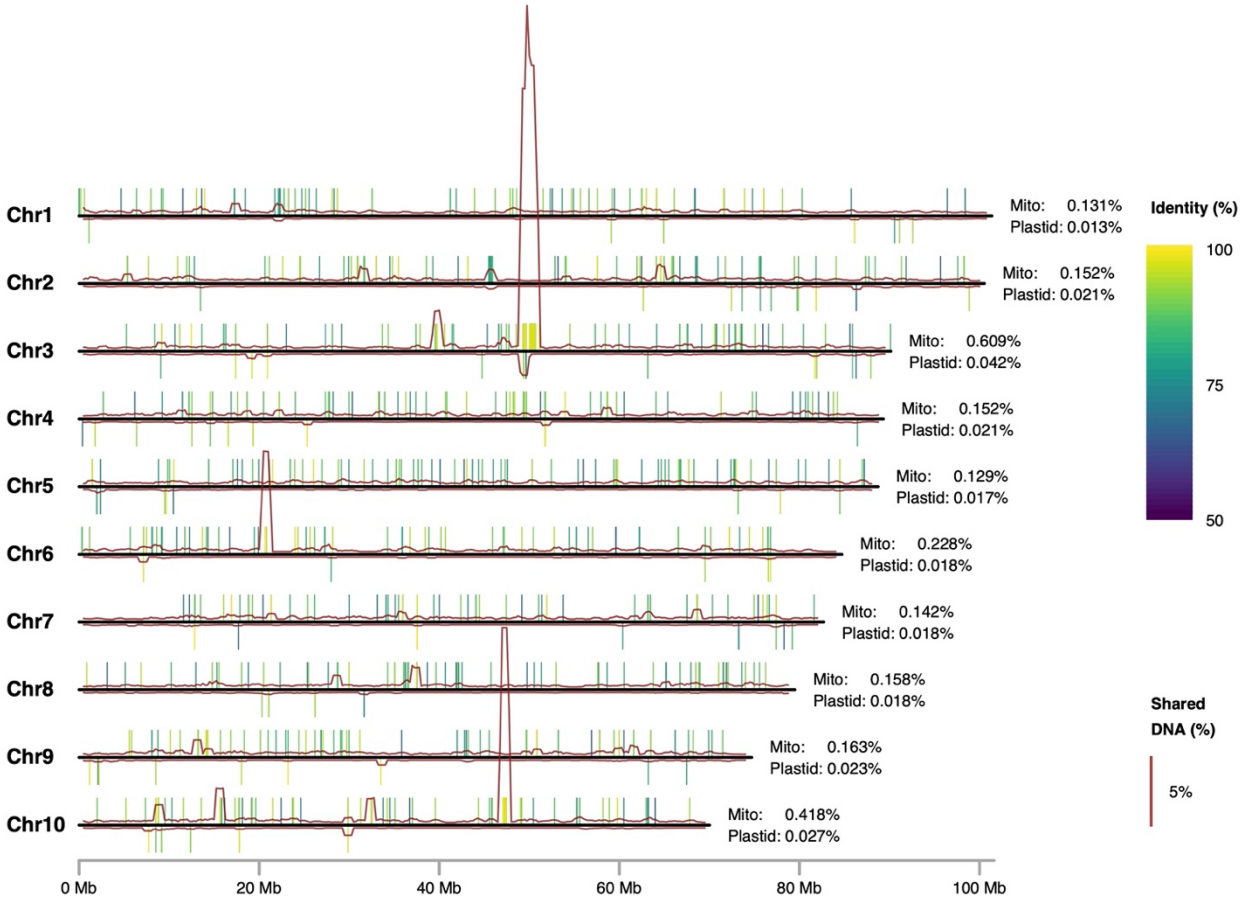


Figure 2. Summary of sequence content shared between the *S. conica* nuclear and cytoplasmic genomes. Tick marks above and below each nuclear chromosome indicate sequence content shared with the mitochondrial and plastid genomes, respectively, as identified by BLAST analysis (minimum hit length of 300 bp and e-value threshold of 1e-6). The color of each tick mark indicates the percent nucleotide identity of the BLAST hit. The red traces are from a sliding window analysis (1-Mb window size and 250-kb step size), indicating the percentage of sequence in the corresponding window that is shared with the mitochondrial genome (above the chromosome) or plastid genome (below the chromosome). Values on the right of each chromosome indicate the overall percentage of sequence shared with each cytoplasmic genome.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

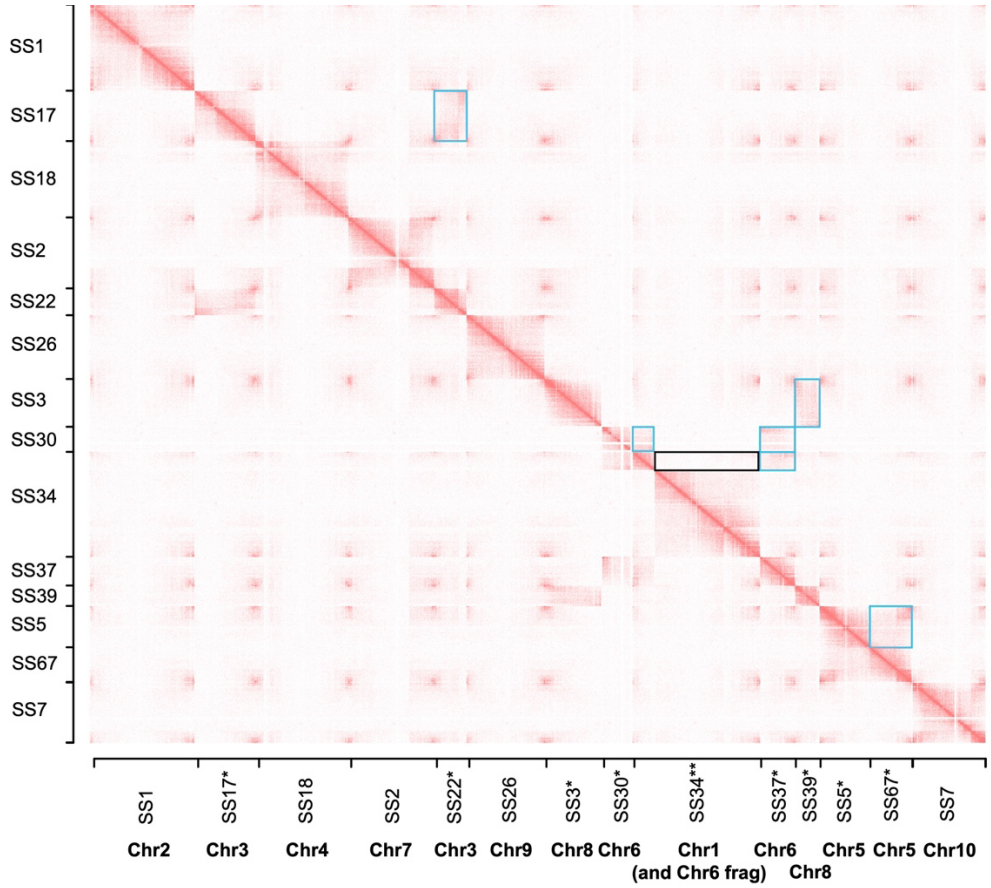
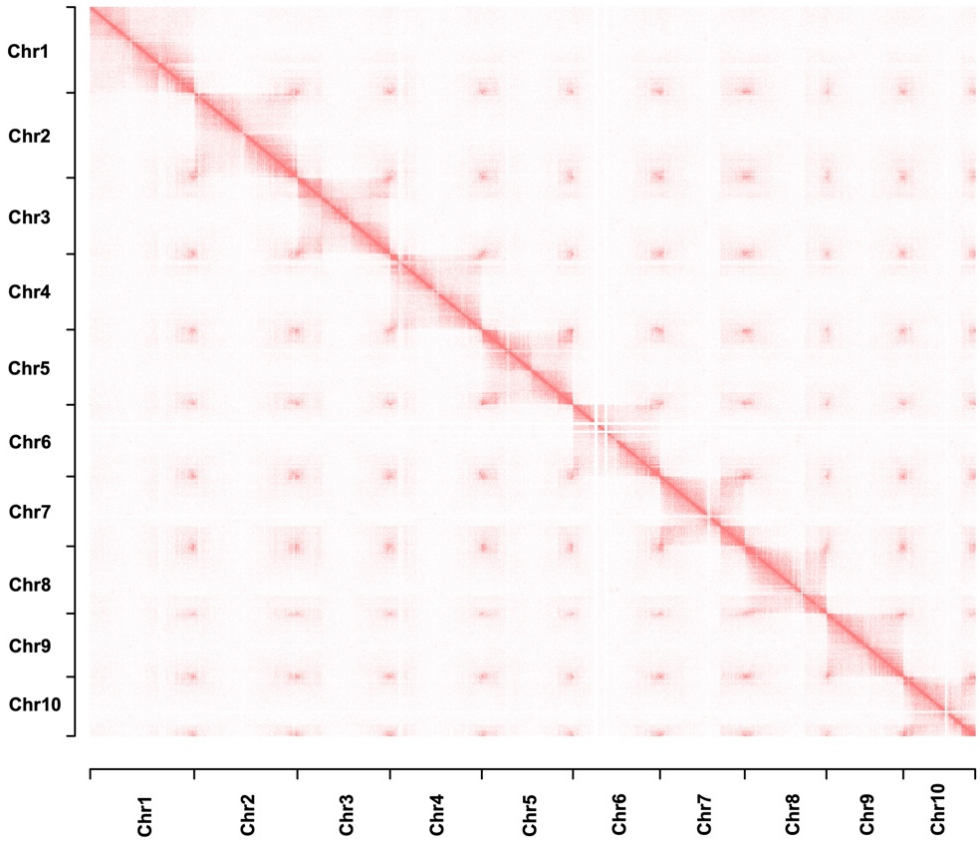


Figure S1. Heatmap generated with HiC-Pro and the HitC R package visualizing Hi-C interactions across the Bionano super-scaffolds (SS) from the *S. conica* genome. The chromosomes to which the scaffolds were eventually assigned are indicated on the x-axis. The outlined boxes above the primary diagonal highlight signal that led to joining scaffolds (teal boxes) or breaking apart a misassembly due to lack of Hi-C contacts within the scaffold (black box). The corresponding regions below the primary diagonal are left unhighlighted for visual comparison. Single asterisks (*) indicate Bionano scaffolds that were joined to form larger chromosome-level scaffolds. The double asterisk (**) indicates the misassembled Bionano SS34 scaffold that was subsequently separated into Chr1 and a portion of Chr6. Only 14 of the 16 Bionano scaffolds are visualized in this figure because the other two were too small and consisted of tandemly repeated ribosomal DNA regions. Those two scaffolds were joined with the other Chr6 scaffolds based on strength of Hi-C contact signal.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



395 **Figure S2.** Heatmap generated with HiC-Pro and the HitC R package visualizing Hi-C interactions
396 across the 10 chromosome-level scaffolds from the *S. conica* genome.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

REFERENCES

Abdel-Ghany SE, LaManna LM, Harroun HT, Maliga P, Sloan DB. 2022. Rapid sequence evolution is associated with genetic incompatibilities in the plastid Clp complex. Plant Molecular Biology 108:277-287.

Aho AV, Kernighan BW, Weinberger PJ. 1987. The AWK programming language: Addison-Wesley Longman Publishing Co.

Alexander HM, Antonovics J. 1988. Disease spread and population dynamics of anther-smut infection of Silene alba caused by the fungus Ustilago violacea. Journal of Ecology 76:91-104.

Antonovics J, Bradshaw AD, Turner R. 1971. Heavy metal tolerance in plants. In. Advances in Ecological Research: Elsevier. p. 1-85.

Bari EA. 1973. Cytological studies in the Genus Silene L. New Phytologist 72:833-838.

Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, Genete M, Berrabah W, Chèvre A-M, Delourme R. 2018. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nature Plants 4:879-887.

Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi–C: a comprehensive technique to capture the conformation of genomes. Methods 58:268-276.

Bernasconi G, Antonovics J, Biere A, Charlesworth D, Delph LF, Filatov D, Giraud T, Hood ME, Marais GAB, McCauley D. 2009. Silene as a model system in ecology and evolution. Heredity 103:5-14.

Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, Lee J, Lam ET, Liachko I, Sullivan ST. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nature genetics 49:643-650.

Blair AC, Wolfe LM. 2004. The evolution of an invasive plant: an experimental study with Silene latifolia. Ecology 85:3035-3042.

Blavet N, Charif D, Oger-Desfeux C, Marais G, Widmer A. 2011. Comparative high-throughput transcriptome sequencing and development of SiESTa, the Silene EST annotation database. BMC Genomics 12:376.

Bringezu K, Lichtenberger O, Leopold I, Neumann D. 1999. Heavy Metal Tolerance of Silene vulgaris. Journal of Plant Physiology 154:536-546.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

425 Broz AK, Waneka G, Wu Z, Sloan DB. 2021. Detecting de novo mitochondrial mutations in angiosperms
426 with highly divergent evolutionary rates. Genetics 218:iyab039.

427 Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit–interactive quality assessment
428 of genome assemblies. G3: Genes, Genomes, Genetics 10:1361-1374.

429 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using
430 phased assembly graphs with hifiasm. Nature Methods 18:170-175.

431 Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue.
432 Phytochemical bulletin 19:11-15.

433 Erixon P, Oxelman B. 2008. Whole-gene positive selection, elevated synonymous substitution rates,
434 duplication, and indel evolution of the chloroplast clpP1 gene. PloS one 3:e1386.

435 Fields PD, McCauley DE, McAssey EV, Taylor DR. 2014. Patterns of cyto-nuclear linkage disequilibrium
436 in Silene latifolia: genomic heterogeneity and temporal stability. Heredity 112:99-104.

437 Fields PD, Waneka G, Naish M, Schatz MC, Henderson IR, Sloan DB. 2022. Complete Sequence of a
438 641-kb Insertion of Mitochondrial DNA in the Arabidopsis thaliana Nuclear Genome. Genome Biology and
439 Evolution 14:evac059.

440 Goremykin VV, Lockhart PJ, Viola R, Velasco R. 2012. The mitochondrial genome of Malus domestica
441 and the import-driven hypothesis of mitochondrial genome expansion in seed plants. The Plant Journal
442 71:615-626.

443 Havird JC, Trapp P, Miller C, Bazos I, Sloan DB. 2017. Causes and consequences of rapidly evolving
444 mtDNA in a plant lineage. Genome Biology and Evolution 9:323-336.

445 Havird JC, Whitehill NS, Snow CD, Sloan DB. 2015. Conservative and compensatory evolution in
446 oxidative phosphorylation complexes of angiosperms with highly divergent rates of mitochondrial genome
447 evolution. Evolution 69:3069-3081.

448 Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in
449 sequenced nuclear genomes. PLoS Genetics 6:e1000834.

450 Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for
451 second-generation genome projects. BMC Bioinformatics 12:1-14.

452 Jafari F, Zarre S, Gholipour A, Eggens F, Rabeler RK, Oxelman B. 2020. A new taxonomic backbone for
453 the infrageneric classification of the species-rich genus Silene (Caryophyllaceae). Taxon 69:337-368.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

454 Jiao W-B, Schneeberger K. 2017. The impact of third generation genomic technologies on plant genome
455 assembly. Current opinion in plant biology 36:64-70.

456 Kejnovsky E, Vyskot B. 2010. Silene latifolia: the classical model to study heteromorphic sex
457 chromosomes. Cytogenetic and genome research 129:250-262.

458 Keller S, Fields P, Berardi A, Taylor D. 2014. Recent admixture generates heterozygosity–fitness
459 correlations during the range expansion of an invading species. Journal of Evolutionary Biology 27:616-
460 627.

461 Keller SR, Taylor DR. 2008. History, chance and adaptation during biological invasion: separating
462 stochastic phenotypic evolution from response to selection. Ecology Letters 11:852-866.

463 Kephart S, Reynolds RJ, Rutter MT, Fenster CB, Dudash MR. 2006. Pollination and seed predation by
464 moths on Silene and allied Caryophyllaceae: evaluating a model system to study the evolution of
465 mutualisms. New Phytologist 169:667-680.

466 Krasovec M, Chester M, Ridout K, Filatov DA. 2018. The mutation rate and the age of the sex
467 chromosomes in Silene latifolia. Current Biology 28:1832-1838.

468 Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M.
469 2012. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly.
470 Nature Biotechnology 30:771-776.

471 Le Gac M, Hood ME, Fournier E, Giraud T. 2007. Phylogenetic evidence of host-specific cryptic species
472 in the anther smut fungus. Evolution 61:15-26.

473 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or
474 nucleotide sequences. Bioinformatics 22:1658-1659.

475 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and
476 streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic,
477 prokaryotic, and viral genomes. Molecular Biology and Evolution 38:4647-4654.

478 Martin H, Carpentier F, Gallina S, Godé C, Schmitt E, Muyle A, Marais GA, Touzet P. 2019. Evolution of
479 young sex chromosomes in two dioecious sister plant species with distinct sex determination systems.
480 Genome Biology and Evolution 11:350-361.

481 Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais GA. 2012. Rapid de novo evolution of X
482 chromosome dosage compensation in Silene latifolia, a plant with young sex chromosomes. PLoS
483 biology 10:e1001308.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Ni P, Xu J, Zhong Z, Zhang J, Huang N, Nie F, Luo F, Wang J. 2022. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. bioRxiv.

Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Moneger F. 2005. A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. PLoS biology 3:47-56.

Ou S, Su W, Liao Y, Chougule K, Agda JR, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome biology 20:275.

Papadopulos AS, Chester M, Ridout K, Filatov DA. 2015. Rapid Y degeneration and dosage compensation in plant sex chromosomes. Proceedings of the National Academy of Sciences 112:13021-13026.

Papadopulos AS, Helmstetter AJ, Osborne OG, Comeault AA, Wood DP, Straw EA, Mason L, Fay MF, Parker J, Dunning LT. 2021. Rapid parallel adaptation to anthropogenic heavy metal pollution. Molecular Biology and Evolution 38:3724-3736.

Pellicer J, Leitch IJ. 2019. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytologist 226.

Qiu Y, Filipenko SJ, Darracq A, Adams KL. 2014. Expression of a transferred nuclear gene in a mitochondrial genome. Current Plant Biology 1:68-72.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19:460.

Rockenbach KD, Havird JC, Monroe JG, Triant DA, Taylor DR, Sloan DB. 2016. Positive selection in rapidly evolving plastid-nuclear enzyme complexes. Genetics 204:1507-1522.

Servant N, Lajoie BR, Nora EP, Giorgetti L, Chen C-J, Heard E, Dekker J, Barillot E. 2012. HiTC: exploration of high-throughput 'C'experiments. Bioinformatics 28:2843-2844.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome biology 16:259.

Shirasawa K, Harada D, Hirakawa H, Isobe S, Kole C. 2021. Chromosome-level de novo genome assemblies of over 100 plant species. Breeding science 71:117-124.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.05.556365; this version posted September 7, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, Taylor DR. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS biology 10:e1001241.

Sloan DB, Alverson AJ, Wu M, Palmer JD, Taylor DR. 2012. Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus Silene. Genome Biology and Evolution 4:294-306.

Smith DR. 2014. Mitochondrion-to-plastid DNA transfer: it happens. New Phytologist 202:736-738.

Taylor DR, Olson MS, McCauley DE. 2001. A quantitative genetic analysis of nuclear-cytoplasmic male sterility in structured populations of Silene vulgaris. Genetics 158:833-841.

Touzet P, Delph LF. 2009. The effect of breeding system on polymorphism in mitochondrial genes of Silene. Genetics 181:631-644.

Warren JM, Broz AK, Elowsky C, Martinez-Hottovy A, Christensen AC, Sloan DB. 2022. Rewiring of aminoacyl-tRNA synthetase localization and interactions in plants with extensive mitochondrial tRNA gene loss. bioRxiv 2022.01.27.478071v2.

Warren JM, Broz AK, Martinez-Hottovy A, Elowsky C, Christensen AC, Sloan DB. 2023. Rewiring of aminoacyl-tRNA synthetase localization and interactions in plants with extensive mitochondrial tRNA gene loss. Molecular Biology and Evolution 40:msad163.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 37:1155-1162.

Williams AM, Itgen MW, Lambert A, Broz AK, Mueller RL, Sloan DB. 2021. Long-read transcriptome and other genomic resources for the angiosperm Silene noctiflora. G3: Genes, Genomes, Genetics 11:jkab189.

Yue J, Krasovec M, Kazama Y, Zhang X, Xie W, Zhang S, Xu X, Kan B, Ming R, Filatov DA. 2023. The origin and evolution of sex chromosomes, revealed by sequencing of the Silene latifolia female genome. Current Biology 33:1-11.

Zhang G-J, Dong R, Lan L-N, Li S-F, Gao W-J, Niu H-X. 2020. Nuclear integrants of organellar DNA contribute to genome structure and evolution in plants. International Journal of Molecular Sciences 21:707.