Complete Mitochondrial Genome of a Gymnosperm, Sitka Spruce (*Picea sitchensis*), Indicates a Complex Physical Structure

Shaun D. Jackman1,†, Lauren Coombe1,†, René L. Warren1, Heather Kirk1, Eva Trinh1, Tina MacLeod1, Stephen Pleasance1, Pawan Pandoh1, Yongjun Zhao1, Robin J. Coope1, Jean Bousquet2, Joerg Bohlmann3, Steven J.M. Jones1, and Inanc Birol1,*

1Genome Sciences Centre, BC Cancer, Vancouver, British Columbia, Canada
2Forest Genomics, Institute for Systems and Integrative Biology, Université Laval, Quebec, Quebec, Canada
3Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada

*Corresponding author: E-mail: ibirol@bcgsc.ca.

Accepted: 20 May 2020

†These authors contributed equally to this work.

Data deposition: The Oxford Nanopore MinION sequencing data are available at the SRA under accessions SRX5081713 and SRR8264752–SRR8264769. The Illumina HiSeq sequencing data have SRA accession SRR5028199. The Sitka spruce (*Picea sitchensis*) genotype Q903 complete mitochondrial genome is available at NCBI GenBank under accessions MK697696–MK697708.

Abstract

Plant mitochondrial genomes vary widely in size. Although many plant mitochondrial genomes have been sequenced and assembled, the vast majority are of angiosperms, and few are of gymnosperms. Most plant mitochondrial genomes are smaller than a megabase, with a few notable exceptions. We have sequenced and assembled the complete 5.5-Mb mitochondrial genome of Sitka spruce (*Picea sitchensis*), to date, one of the largest mitochondrial genomes of a gymnosperm. We sequenced the whole genome using Oxford Nanopore MiniION, and then identified contigs of mitochondrial origin assembled from these long reads based on sequence homology to the white spruce mitochondrial genome. The assembly graph shows a multipartite genome structure, composed of one smaller 168-kb circular segment of DNA, and a larger 5.4-Mb single component with a branching structure. The assembly graph gives insight into a putative complex physical genome structure, and its branching points may represent active sites of recombination.

Key words: gymnosperms, Sitka spruce, organelle, genome assembly, sequencing, ABySS.

Introduction

Plant mitochondrial genomes are amazingly diverse and complex (Mower et al. 2012). Land plant mitochondrial genomes range in size from 66 kilobases (kb) for the parasitic angiosperm *Viscum scuruloides* (Skippington et al. 2015) to >11 megabases (Mb) in the case of the flowering plant *Silene conica* (Sloan et al. 2012). Although their genome structure is often portrayed as a circle, the true physical structure of their genome appears to be a variety of circles, linear molecules, and complex branching structures (Backert et al. 1997; Backert and Börner 2000). Although many species have a single master circle representation of their mitochondrial genome, others are composed of more than a hundred circular chromosomes (Sloan et al. 2012). The precise mechanism of how plant mitochondria replicate and maintain their DNA is not yet fully understood (Cupp and Nielsen 2014). It is hypothesized that recombination-dependent replication plays a role, giving a functional role to the repeat sequences often observed in mitochondria (Gualberto et al. 2014). This model does not fully explain how genomic copy number is regulated and maintained (Oldenburg and Bendich 2015), particularly in multipartite genomes (Vlcek et al. 2011). Although angiosperm mitochondrial genomes are well studied with numerous complete genomes available, gymnosperm mitochondrial
Genomes were scarcer until recently: One from each of the cycads (Chaw et al. 2008), ginkgos, gnetales (Guo et al. 2016), and conifers (Jackman et al. 2016). This year saw a number of mitochondrial genomes published in short succession (Guo et al. 2020; Kan et al. 2020; Sullivan et al. 2020). Although other gymnosperm mitochondrial genomes are smaller than a megabase, conifer mitochondrial genomes can exceed five megabases (Jackman et al. 2016), larger than many bacteria. Compared with other plants, conifers also tend to have very large nuclear genomes (De La Torre et al. 2014), particularly the spruce species (Birol et al. 2013; Nystedt et al. 2013; Warren et al. 2015). As plant mitochondrial genomes typically have <100 genes, what role this expanse of DNA serves, if any, remains mysterious.

Assembling plant mitochondrial genomes is difficult due to the presence of large (up to 30 kb) perfect repeats, which may be involved in active recombination, and hypothesized recombination-dependent replication (Gualberto et al. 2014). A hybrid assembly of both long reads, which are able to span most repeats, and accurate short sequencing reads, which correct indel errors, is well suited to tackle these challenging genome features. Hybrid assembly of long and short reads has been applied to assemble the plastid genome of Eucalyptus pauciflora (Wang et al. 2018), as well as plant mitochondrial genomes (Kovar et al. 2018; Kozik et al. 2019).

Annotation plant mitochondrial genomes is also challenging, due to numerous features of plant mitochondria that are not typical of most organisms. For one, RNA editing of C-to-U is pervasive, and this process creates AUG start codons by editing ACG to AUG (Hiesel et al. 1989) or by editing GCG to GUG, an alternative start codon used by some plant mitochondrial genes (Sakamoto et al. 1997). RNA editing can also create stop codons in a similar fashion. Further complicating annotation using available bioinformatics pipelines, the typical GU-AG splice site expected by most splice-acceptor alignments tools is instead GNGCG-AY (Y denotes C or T) for group II introns (Lambowitz and Zimmerly 2011). Also, trans-spliced genes are common in mitochondrial genomes (Kamikawa et al. 2016; Guo et al. 2020), and no purpose-built software tool exists for identifying and annotating trans-spliced genes. To add further difficulty, trans-spliced exons may be as small as 22 bp, as is nad5 exon 3 of gymnosperms (Guo et al. 2016) and other vascular plant mitochondria (Knoop et al. 1991). For these reasons, annotating a plant mitochondrial genome remains a laborious and manual task.

In this study, we report on the sequencing, assembly and annotation of the mitochondrial genome of Sitka spruce (Picea sitchensis [Bong.] Carrère, genotype Q903) needles, as described in Coombe et al. (2016). We constructed 18 Oxford Nanopore 1D sequencing libraries, 16 by ligation of 1–7 µg of lightly needle-sheared genomic DNA and 2 by rapid transposition of 0.6 µg of unsheared genomic DNA, and sequenced these on 18 MinION R9.4 flow cells. This whole genome sequencing produced 98 Gb in 9.6 million reads (SRA accession SRX5081713), yielding 5-fold depth of coverage of the roughly 20 Gb nuclear genome, and 26-fold depth of coverage of the mitochondrial genome. We first obtained a rough but computationally efficient assembly using Miniasm (Li 2016), after trimming adapter sequences with Porechop (Wick et al. 2017a) and polishing the resulting assembly with Racon (Vaser et al. 2017). We selected contigs homology to the white spruce (Picea glauca, interior white spruce genotype PG29) mitochondrial genome (Jackman et al. 2016) using Bandage (Wick et al. 2015), retaining contigs with at least one ≥5-kb alignment to the white spruce mitochon- drion by BlastN (Altschul et al. 1990). We then proceeded to align back the nanopore reads to our draft Sitka spruce assembly with minimap2 (Li 2018), segre- gate aligned reads and assemble them de novo first with Unicycler (Wick et al. 2017b), and then with Flye (Kolmogorov et al. 2019), as described (supplementary methods, Supplementary Material online). The final genome sequence was obtained by polishing with Pilon (Walker et al. 2014) using Illumina Hiseq sequencing reads of the same DNA extraction (supplementary fig. S1, Supplementary Material online).

**Materials and Methods**

**Genome Sequencing and Assembly**

Genomic DNA was extracted from young Sitka spruce (P. sitchensis [Bong.] Carrère, genotype Q903) needles, as described in Coombe et al. (2016). We constructed 18 Oxford Nanopore 1D sequencing libraries, 16 by ligation of 1–7 µg of lightly needle-sheared genomic DNA and 2 by rapid transposition of 0.6 µg of unsheared genomic DNA, and sequenced these on 18 MinION R9.4 flow cells. This whole genome sequencing produced 98 Gb in 9.6 million reads (SRA accession SRX5081713), yielding 5-fold depth of coverage of the roughly 20 Gb nuclear genome, and 26-fold depth of coverage of the mitochondrial genome. We first obtained a rough but computationally efficient assembly using Miniasm (Li 2016), after trimming adapter sequences with Porechop (Wick et al. 2017a) and polishing the resulting assembly with Racon (Vaser et al. 2017). We selected contigs homology to the white spruce (Picea glauca, interior white spruce genotype PG29) mitochondrial genome (Jackman et al. 2016) using Bandage (Wick et al. 2015), retaining contigs with at least one ≥5-kb alignment to the white spruce mitochon- drion by BlastN (Altschul et al. 1990). We then proceeded to align back the nanopore reads to our draft Sitka spruce assembly with minimap2 (Li 2018), segre- gate aligned reads and assemble them de novo first with Unicycler (Wick et al. 2017b), and then with Flye (Kolmogorov et al. 2019), as described (supplementary methods, Supplementary Material online). The final genome sequence was obtained by polishing with Pilon (Walker et al. 2014) using Illumina Hiseq sequencing reads of the same DNA extraction (supplementary fig. S1, Supplementary Material online).

**Annotation**

We annotated coding genes and non-coding rRNA and tRNA genes using automated methods where possible, and performed manual inspection to refine these automated annota- tions. We used Prokka (Seemann 2014), which uses Prodigal (Hyatt et al. 2010) to identify single-exon coding genes and open reading frames (ORFs). We used MAKER (Holt and Yandell 2011), which uses BLASTP and Exonerate (Slater and Birney 2005) to identify cis-spliced coding genes. Prokka used ab initio predictions from Prodigal. MAKER used evidence alignments only. Genes in GenBank from Viridiplantae mitochondria were used as evidence. The annota- tions of Prokka and Maker were combined, along with the manually annotated cis-spliced and trans-spliced genes. No species-specific TE libraries were used.

We used tRNAscan-SE (Lowe and Eddy 1997) and Aragorn (Laslett 2004) to identify tRNA. We used RNAmmer (Lagesen et al. 2007) and Barrnap (Seemann 2014) to identify rRNA. We used RNAweasel (Lang et al. 2007) and Infernal (Nawrocki
et al. 2009) to identify group II introns. RFAM motif Domain-V (RM00007) represents domain V of group II introns, and RFAM family Intron_gpII (RF00029) represents both domains V and VI (Kalvari et al. 2018).

Results and Discussion

Complete Genome Assembly

The complete mitochondrial genome of Sitka spruce is 5.52 Mb assembled in 13 segments, and has a GC content of 44.7%. The assembly statistics are summarized in table 1.

The genome assembly is composed of two components: A 168-kb circular segment (labeled 10), and a larger 5.36-Mb component composed of 12 segments, as visualized in figure 1 using Bandage (Wick et al. 2015). The eleven largest segments, ranging in size from 84 kb to 1.65 Mb, have similar depth of coverage, assumed to represent single-copy genomic segments. The two smallest segments (27 and 24 kb labeled 12 and 13 in figure 1, respectively, representing <1% of the mitochondrial genome) exhibit an estimated copy number of two based on their depth of sequencing coverage. No sequence variation is evident in these two repeats. An absence of variation in the repeat implies that they may be involved in active recombination (Maréchal and Brisson 2010). Though 10% of reads are >24 kb, no reads fully span these repeats.

The complete mitochondrial genome assembly of Sitka spruce assembled from Oxford Nanopore sequencing is composed of 13 contigs >20 kb with an N50 length of 547 kb. The use of long reads was critical in achieving this contiguity and completeness. In contrast, the draft mitochondrial genome assembly of white spruce was assembled from paired-end and mate-pair Illumina sequencing and is composed of 117 contigs >2 kb arranged in 36 scaffolds with a contig and scaffold N50 length of 102 and 369 kb, respectively (Jackman et al. 2016). The fragmented state of the white spruce mitochondrial assembly provided little information as to the structure of the mitochondrial genome, whereas the Sitka spruce assembly graph (fig. 1) suggests a multipartite genome structure.

The complete genome is composed of 1.7% (93 kb) of genes with known function, 28.0% (1,545 kb) of 6,806 ORFs (each of at least 90 bp), 3.7% (205 kb) of repeats, and 66.6% unclassified sequences. Of the ORFs, 1,039 are at least 300 bp (100 amino acids) in size and compose 7.2% (400 kb) of the genome. Aligning the ORFs with BLASTP, 63 ORFs (17 ORFs of at least 300 bp) have a significant

| Table 1 |
| --- |
| **Summary of Assembly Statistics** |
| **Assembly Statistics** |
| Number of contigs | 13 |
| Assembly N50 | 547 kb |
| Undetermined bases | None |
| Largest contig | 1.65 Mb |
| Total reconstruction | 5.53 Mb |
| GC content | 44.7% |

![Fig. 1](https://academic.oup.com/gbe/article/12/7/1174/5843792) The assembly graph of the mitochondrial genome of Sitka spruce. Each colored segment is labeled with its size and named 01–13 by rank of size. Only segments 12 and 13 representations are inferred as repeats. All segment adjacencies are supported by the long reads, indicating a complex branching genomic structure.
Plastid-derived sequences compose 0.25% (14 kb) of the genome spread across 24 segments. The nuclear repeats LTR/Gypsy compose 51% of the repeat sequence, and LTR/Copia compose 7%. The genome also has non-transposable element repeats; simple repeat sequences compose 34%, low complexity sequences compose 3%, and 5% are other repeat sequences. The 36-bp Bpu repeat sequence is present in roughly 500 copies in *Cycas taitungensis* and roughly 100 copies in *Ginkgo biloba* (Guo et al. 2016). We find only a single full-length copy with four mismatches in Sitka spruce, similar to *Welwitschia mirabilis*.

**Genes**

The mitochondrial genome of Sitka spruce has 41 distinct protein coding genes with known function, 3 distinct rRNA genes (supplementary table S1, Supplementary Material).
online), and 27 distinct tRNA genes representing 18 distinct anticodons (supplementary table S2, Supplementary Material online). The relative order and orientation of these genes are shown in figure 2. The 41 known protein coding genes found in the gymnosperm mitochondria C. taitungensis (Chaw et al. 2008) and G. biloba (Guo et al. 2016) are also found in Sitka spruce. The 29 introns, 16 cis-spliced, and 13 trans-spliced, are found in 10 protein coding genes, two pseudogenes, and one plastid-derived tRNA (supplementary table S3, Supplementary Material online).

Four ORFs (937 [Q903MT_gene558], 507 [Q903MT_gene3172], 371 [Q903MT_gene4186], and 221 [Q903MT_gene4187] amino acids) contain an organellar DNA polymerase type B (DNA_pol_B_2) Pfam family (El-Gebali et al. 2019). The largest one also contains a segment homologous to a putative cell-wall binding lipoprotein (YkyA) Pfam domain. These ORFs have homology to DNA polymerase genes found in the mitochondria of P. glauca and the angiosperms Cocos nucifera, Daucus carota, Helianthus annuus, and Silene vulgaris. Two ORFs (781 [Q903MT_gene4185] and 560 [Q903MT_gene6929] amino acids) contain an RNA polymerase (RNA_pol) domain. These ORFs have homology to DNA-dependent RNA polymerase genes found in the mitochondria of P. glauca and the angiosperms Beta vulgaris, C. nucifera, D. carota, and Phoenix dactylifera. The two largest genes of the Sitka spruce mitochondrial genome are these putative DNA and RNA polymerase genes.

The matR gene [Q903MT_gene5018] and three additional ORFs (476 [Q903MT_gene679], 197 [Q903MT_gene3613], and 163 [Q903MT_gene5889] amino acids) contain a reverse transcriptase, or RNA-dependent DNA polymerase, (RT_like) NCBI conserved protein domain with similarity to the group II intron reverse transcriptase/maturase (group II RT_mat) NCBI conserved protein domain (Marchler-Bauer et al. 2017). We hypothesize that these ORFs may be additional maturases involved in splicing (Matsuura 2001). These ORFs have homology to mitochondrial genes of P. glauca and the angiosperm Utricularia reniformis.

**Conclusion**

The 5.5-Mb mitochondrial genome of Sitka spruce is among the largest ones in plants, and is the largest complete mitochondrial genome reported to date for a gymnosperm. It follows the trend seen for spruce and conifer nuclear genomes, which are also among the largest in plants (De La Torre et al. 2014). The physical structure of the Sitka spruce mitochondrial genome is not the typical circularly mapping single chromosome, but multipartite. The larger component of the assembly graph exhibits a rosette-like structure, mirroring the rosette-like structures observed in electron micrographs of mitochondrial DNA (Backert and Börner 2000). The intricate structure and the discovery of large repeat elements suggest the presence of active sites for hypothesized recombination-dependent replication of the mitochondrial genome (Gualberto et al. 2014; Sullivan et al. 2020). Considering heteroplasmy resulting from naturally hybridizing species of spruce and paternal leakage of mitochondria, intermolecular recombination would result in interspecific hybridization of the mitochondrial genome, which has been reported to occur in natural spruce populations (Jaramillo-Correa and Bousquet 2005). The complete mitochondrial genome of Sitka spruce, a gymnosperm, should prove invaluable to future investigations into the genome structure and mechanism of replication of conifer mitochondrial genomes.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**Acknowledgments**

This study was supported by Genome Canada, Genome Quebec, Genome British Columbia, and Genome Alberta (Grant No. 243FOR) for the Spruce-Up project (http://www.spruce-up.ca, last accessed June 1, 2020).

**Literature Cited**

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Backert S, Börner T. 2000. Phage T4-like intermediates of DNA replication and recombination in the mitochondria of the higher plant Chenopodium album (L.). Curr Genet. 37(5):304–314.

Backert S, Lynn Nielsen B, Börner T. 1997. The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. Trends Plant Sci. 2(12):477–483.

Birol I, et al. 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29(12):1492–1497.

Chaw S-M, et al. 2008. The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol. 25(3):603–615.

Coombe L, et al. 2016. Assembly of the complete Sitka spruce chloroplast genome using 10X genomics’ GemCode sequencing data. PLoS One 11(9):e0163059.

Cupp JD, Nielsen BL. 2014. Minireview: DNA replication in plant mitochondria. Mitochondrion 19:231–237.

De La Torre AR, et al. 2014. Insights into conifer giant-genomes. PLANT Physiol. 166(4):1724–1732.

El-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432.

Gualberto JM, et al. 2014. The plant mitochondrial genome: dynamics and maintenance. Biochimie 100:107–120.

Guo W, et al. 2020. Ginkgo and Welwitschia mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. Mol Biol Evol. 33(6):1448–1460.

Guo W, et al. 2020. Extensive shifts from cis to trans splicing of gymnosperm mitochondrial introns. Mol Biol Evol. 37(6):1615–1620.

Hiesel R, Wissinger B, Schuster W, Brennicke A. 1989. RNA editing in plant mitochondria. Science 246(4937):1632–1634.

**References**

Aldrich J, et al. 2008. Insights into conifer gigagenomes. PLANT Physiol. 166(4):1724–1732.

Backert S, Lynn Nielsen B, Börner T. 1997. The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. Trends Plant Sci. 2(12):477–483.

Birol I, et al. 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29(12):1492–1497.

Chaw S-M, et al. 2008. The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol. 25(3):603–615.

Coombe L, et al. 2016. Assembly of the complete Sitka spruce chloroplast genome using 10X genomics’ GemCode sequencing data. PLoS One 11(9):e0163059.

Cupp JD, Nielsen BL. 2014. Minireview: DNA replication in plant mitochondria. Mitochondrion 19:231–237.

De La Torre AR, et al. 2014. Insights into conifer giant-genomes. PLANT Physiol. 166(4):1724–1732.

El-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432.

Gualberto JM, et al. 2014. The plant mitochondrial genome: dynamics and maintenance. Biochimie 100:107–120.

Guo W, et al. 2020. Ginkgo and Welwitschia mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. Mol Biol Evol. 33(6):1448–1460.

Guo W, et al. 2020. Extensive shifts from cis to trans splicing of gymnosperm mitochondrial introns. Mol Biol Evol. 37(6):1615–1620.

Hiesel R, Wissinger B, Schuster W, Brennicke A. 1989. RNA editing in plant mitochondria. Science 246(4937):1632–1634.
