Organellar Genomes of White Spruce (*Picea glauca*): Assembly and Annotation

Shaun D. Jackman[1], René L. Warren[1], Ewan A. Gibb[1], Benjamin P. Vandervalk[1], Hamid Mohamadi[1], Justin Chu[1], Anthony Raymond[1], Stephen Pleasance[1], Robin Coope[1], Mark R. Wildung[2], Carol E. Ritland[3], Jean Bousquet[4], Steven J. M. Jones[1,5,6], Joerg Bohlmann[3,7,8,*], and Inanç Birol[1,5,6,9,*]

[1]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada

[2]School of Molecular Biosciences, Washington State University

[3]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC, Canada

[4]Department of Forest and Environmental Genomics, Université Laval, Québec, QC, Canada

[5]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

[6]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada

[7]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

[8]Department of Botany, University of British Columbia, Vancouver, BC, Canada

[9]Department of Computer Science, University of British Columbia, Vancouver, BC, Canada

*Corresponding author: bohlmann@msl.ubc.ca; ibirol@bcgsc.ca.

Data deposition: The white spruce plastid and mitochondrial genomes and their annotations have been deposited at NCBI under the accessions KT634228 and LKAM00000000, respectively.

Accepted: November 28, 2015

Abstract

The genome sequences of the plastid and mitochondrion of white spruce (*Picea glauca*) were assembled from whole-genome shotgun sequencing data using ABySS. The sequencing data contained reads from both the nuclear and organellar genomes, and reads of the organellar genomes were abundant in the data as each cell harbors hundreds of mitochondria and plastids. Hence, assembly of the 123-kb plastid and 5.9-Mb mitochondrial genomes were accomplished by analyzing data sets primarily representing low coverage of the nuclear genome. The assembled organellar genomes were annotated for their coding genes, ribosomal RNA, and transfer RNA. Transcript abundances of the mitochondrial genes were quantified in three developmental tissues and five mature tissues using data from RNA-seq experiments. C-to-U RNA editing was observed in the majority of mitochondrial genes, and in four genes, editing events were noted to modify ACG codons to create cryptic AUG start codons. The informatics methodology presented in this study should prove useful to assemble organellar genomes of other plant species using whole-genome shotgun sequencing data.

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

Introduction

Most plant cells contain two types of organelles that comprise their own genomes, mitochondria, and plastids. In *Pinaceae*, mitochondrial genomes are inherited maternally, and plastid genomes are inherited paternally (Whittle and Johnston 2002).

Complete plastid genomes of the gymnosperms Norway spruce (*Picea abies*) (Nystedt et al. 2013), *Podocarpus lambertii* (Vieira Ldo, Faoro, Rogalski, et al. 2014), *Taxus chinensis* var. *mairei* (Zhang et al. 2014), and four *Juniperus* species (Guo et al. 2014) have recently been published in National Center for Biotechnology Information (NCBI) GenBank (Benson et al. 2014). These projects used a variety of strategies for isolating plastid DNA (cpDNA), using physical separation methods in the lab or computationally separating cpDNA sequences from nuclear sequences. They also used different approaches for sequencing and assembly (table 1).

The *P. abies* project used 454 GS FLX Titanium sequencing and Sanger sequencing of polymerase chain reaction (PCR)

Table 1

Methods of cpDNA Separation, Sequencing, and Assembly of Complete Plastid Genomes of Gymnosperms Published

Species	cpDNA Separation	Sequencing	Sequence Assembler Software Tool
Picea abies	BLAST in silico	454 GS FLX Titanium[a]	Newbler
Podocarpus lambertii	Saline Percoll gradient	Illumina MiSeq	Newbler
Juniperus bermudiana	Longer-range PCR	Illumina GAII[a]	Geneious
Other *Juniperus*	Unspecified	Illumina MiSeq	Velvet
Taxus chinensis	BLAT in silico	Illumina HiSeq 2000	SOAPdenovo

[a] Finished with PCR and Sanger sequencing.

amplicons for finishing, Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990) to isolate the cpDNA reads, and the software Newbler to assemble the reads. [AQ]The *Po. lambertii* project isolated the cpDNA with a saline Percoll gradient protocol (Vieira Ldo, Faoro, Fraga, et al. 2014), used Illumina MiSeq sequencing data, and assembled the reads using the Newbler software. The *Juniperus bermudiana* project used long-range PCR to amplify the plastid DNA, a combination of Illumina GAII and Sanger sequencing, and the software Geneious to assemble the reads using Camellia*japonica* as a reference genome. The other three *Juniperus* projects used Illumina MiSeq sequencing and the software Velvet (Zerbino and Birney 2008) to assemble the reads. The *T. chinensis* project used whole-genome Illumina HiSeq 2000 sequencing, BLAT (Kent 2002) to isolate the cpDNA reads, and SOAPdenovo (Luo et al. 2012) to assemble the isolated cpDNA reads. All of these projects used DOGMA (Wyman et al. 2004) to annotate their assemblies.

Only one complete mitochondrial genome of a gymnosperm has been published so far (*Cycas taitungensis* [Chaw et al. 2008]), whereas complete mitochondrial genome sequences of the angiosperms *Brassica maritima* (Grewe et al. 2014), *Brassica oleracea* (Grewe et al. 2014), *Capsicum annuum* (Jo et al. 2014), *Eruca sativa* (Wang et al. 2014), *Helianthus tuberosus* (Bock et al. 2014), *Raphanus sativus* (Jeong et al. 2014), *Rhazya stricta* (Park et al. 2014) and *Vaccinium macrocarpon* (Fajardo et al. 2014) have been deposited in NCBI GenBank. Six of these projects gave details of the sample preparation, sequencing, assembly, and annotation strategy. Three projects enriched organellar DNA using various laboratory methods (Kim et al. 2007;Keren et al. 2009;Chen et al. 2011), and the remainder used total genomic DNA. Three projects used Illumina HiSeq 2000 sequencing and Velvet for assembly, and three projects used Roche 454 GS-FLX sequencing and Newbler for assembly. Most projects used an aligner such as BLAST (Altschul et al. 1990) to isolate sequences with similarity to known mitochondrial sequence, either before or after assembly. Two projects used Mitofy (Alverson et al. 2010) to annotate the genome, and the remainder used a collection of tools such as BLAST, tRNAscan-SE (Lowe and Eddy 1997), and ORF Finder to annotate genes. Plant mitochondrial genomes can substantially vary in size, with some of the largest mitochondrial genomes reported for the basal angiosperm *Amborella trichopoda* (3.9 Mb) (Rice et al. 2013) and the two *Silene* species *S. noctiflora* and *S. conica* (6.7 and 11.3 Mb, respectively) (Sloan et al. 2012). The mitochondrial genome of the gymnosperm *Cycas* is relatively smaller with a length of 415 kb (Chaw et al. 2008).

The SMarTForests project (www.smartforests.ca) has recently published a set of stepwise improved assemblies of the 20-Gb white spruce (*Picea glauca*) genome (Birol et al. 2013; Warren, Keeling, et al. 2015; Warren, Yang, et al. 2015), a gymnosperm genome seven times the size of the human genome, sequenced using the Illumina HiSeq and MiSeq sequencing platforms. The whole-genome sequencing data contained reads originating from both the nuclear and organellar genomes. Although one copy of the diploid nuclear genome is found in each cell, hundreds of organelles are present, and thus hundreds of copies of the organellar genomes. This abundance results in an overrepresentation of the organellar genomes in whole-genome sequencing data.

Assembling low coverage white spruce whole-genome shotgun (WGS) sequencing data using the software ABySS (Simpson et al. 2009) yielded assemblies composed mainly of organellar sequences and nuclear repeat elements. The assembled sequences that originate from the organellar genomes were separated from those of nuclear origin by classifying the sequences using their length, depth of coverage and GC content. The plastid genome of white spruce was compared with that of Norway spruce (*P.abies*) (Nystedt et al. 2013), and the mitochondrial genome of white spruce was compared with that of prince sago palm (*C. taitungensis*) (Chaw et al. 2008). Notably, white spruce and Norway spruce belong to phylogenetically remote spruce lineages (Bouillé et al. 2011) and their split occurred at least 10 Ma (Bouillé and Bousquet 2005), resulting in a nuclear genome sequence divergence of approximately 3% (Warren, Keeling, et al. 2015).

Materials and Methods

DNA, RNA, and Software Materials

Genomic DNA was collected from the apical shoot tissues of a single interior white spruce tree, clone PG29 from the British

Columbia Ministry of Forests, Lands and Natural Resource Operations, and sequencing libraries constructed as described before (Birol et al. 2013). Because the original intention of this sequencing project was to assemble the nuclear genome, an organelle exclusion method was used to preferentially extract nuclear DNA. However, sequencing reads from both organellar genomes were present in sufficient depth to assemble their genomes.

RNA was extracted from eight samples, three developmental stages and five mature tissues: Megagametophyte, embryo, seedling, young bud, xylem, mature needle, flushing bud and bark, as described earlier (Warren, Keeling, et al. 2015). These samples were sequenced with the Illumina HiSeq 2000 (Warren, Keeling, et al. 2015). The RNA-seq data were used to quantify the transcript abundance of the annotated mitochondrial genes using the software Salmon (Patro et al. 2014).

The software used in this analysis and their versions are listed in supplementary table S1, Supplementary Material online. All software tools were installed using Homebrew (http://brew.sh, last accessed December 17, 2015).

Plastid Genome Assembly

A single lane of Illumina MiSeq paired-end sequencing (SRR525215) was used to assemble the plastid genome. Paired-end sequencing usually leaves a gap of unsequenced nucleotides in the middle of the DNA fragment. Because 300-bp paired-end reads were sequenced from a library of 500-bp DNA fragments, the reads are expected to overlap by 100 bp. These overlapping paired-end reads were merged using ABySS-mergepairs, a component of the software ABySS (Simpson et al. 2009). These merged reads were assembled using ABySS. Contigs that were putatively derived from the plastid genome were separated by length and depth of coverage using thresholds chosen by inspection of a scatter plot (see supplementary fig. S1, Supplementary Material online). These putative plastid contigs were assembled into scaffolds using ABySS-scaffold.

We ran the gap-filling application Sealer (Paulino et al. 2015) (options -v -j 12 -b 30G -B 300 -F 700 with -k from 18 to 108 with step size 6) on the ABySS assembly of the plastid genome, closing five of the remaining seven gaps, with a resulting assembly consisting of two large (~50 and ~70 kb) scaftigs. Given the small size of the plastid genome, we opted to manually finish the assembly using the software Consed 20.0 (Gordon and Green 2013). We loaded the resulting gap-filled assembly into Consed and imported Pacific Biosciences (PacBio) sequencing data (SRR2148116 and SRR2148117), 9,204 reads 500 bp and larger, into the assembly and aligned them to the plastid genome using cross_-match from within Consed. For each scaftig end, six PacBio reads were pulled out and assembled using the mini-assembly feature in Consed. Cross_match alignments of the resulting contigs to the plastid assembly were used to merge the two scaftigs and confirm that the complete circular genome sequence was obtained. In a subsequent step, 7,742 Illumina HiSeq reads were imported and aligned to the assembly using Consed. These reads were selected from the library of 133 million reads used to assemble the mitochondrial genome (see below) on the basis of alignment to our draft plastid genome using BWA 0.7.5a (Li 2013), focusing on regions that would benefit from read import by restricting our search to regions with ambiguity and regions covered by PacBio reads exclusively. The subset of Illumina reads was selected using samtools 0.1.18, mini-assembled with Phrap (Gordon and Green 2013) and the resulting contigs remerged to correct bases in gaps filled only by PacBio, namely one gap and sequence at edges confirming the circular topology. The starting base was chosen using the Norway spruce plastid genome sequence (NC_021456) (Nystedt et al. 2013). Our assembly was further polished using the Genome Analysis Toolkit (GATK) 2.8-1-g932cd3a FastaAlternateReferenceMaker (McKenna et al. 2010).

The assembled plastid genome was initially annotated using DOGMA (Wyman et al. 2004). Being an interactive web application, it is not convenient for automated annotation. We used the software MAKER (Campbell et al. 2014) to annotate the white spruce plastid using the Norway spruce plastid genome for both protein-coding and noncoding gene homology evidence. The parameters of MAKER are shown in supplementary table S2, Supplementary Material online. The inverted repeat was identified using MUMmer (Kurtz et al. 2004), shown in supplementary figure S3, Supplementary Material online.

The assembled plastid genome was aligned to the Norway spruce plastid using BWA-MEM (Li 2013). The two genomes were compared using QUAST (Gurevich et al. 2013) to confirm the presence and position of the annotated genes of the Norway spruce plastid in the white spruce plastid.

Mitochondrial Genome Assembly

Konnector (Vandervalk et al. 2015) was used to fill the gap between the paired-end reads of a single lane of Illumina HiSeq 2000 paired-end sequencing (SRR525196). These connected paired-end reads were assembled using ABySS. Putative mitochondrial sequences were separated from nuclear sequences by their length, depth of coverage and GC content using k-means clustering in R (see supplementary fig. S2, Supplementary Material online). The putative mitochondrial contigs were then assembled into scaffolds using ABySS-scaffold with a single lane of Illumina HiSeq sequencing of a mate-pair library.

The ABySS assembly of the white spruce mitochondrial genome resulted in 71 scaffolds. We ran the gap-filling application Sealer attempting to close the gaps between every combination of two scaffolds. This approach closed 10 gaps and yielded 61 scaffolds, which we used as input to the LINKS scaffolder 1.1 (Warren, Yang, et al. 2015) (options -k 15 -t 1 -l 3 -r 0.4, 19 iterations with -d from 500 to 6,000 with step size 250) in conjunction with long PacBio reads, further decreasing the number of scaffolds to 58. The Konnector pseudoreads were aligned to the 58 LINKS scaffolds with BWA 0.7.5a (bwa mem -a multimap), and we created links between two scaffolds when reads aligned within 1,000 bp of the edges of any two scaffolds. We modified LINKS to read the resulting SAM alignment file and link scaffolds satisfying this criterion (options LINKS-sam -e 0.9 -a 0.5), bringing the final number of scaffolds to 38. We confirmed the merges using mate-pair reads. The white spruce mate-pair libraries used for confirmation were presented earlier (Birol et al. 2013), and are available from DNAnexus (http://sra.dnanexus.com/studies/SRP014489 last accessed 17 Dec 2015). In brief, mate-pair reads from three fragment size libraries (5, 8, and 12 kb) were aligned to the 38-scaffold assembly with BWA-MEM 0.7.10-r789 and the resulting alignments parsed with a PERL script. A summary of this validation is presented in supplementary table S4, Supplementary Material online. Automated gap-closing was performed with Sealer 1.0 (options -j 12 -B 1000 -F 700 -P10 -k96 -k80) using Bloom filters built from the entire white spruce PG29 read data set (Warren, Keeling, et al. 2015) and closed 55 of the 182 total gaps (30.2%). We polished the gap-filled assembly using GATK, as described for the plastid genome.

The assembled scaffolds were aligned to the NCBI nucleotide (nt) database using BLAST to check for hits to published mitochondrial genomes, and to screen for contamination.

The mitochondrial genome was annotated using MAKER (parameters shown in supplementary table S3, Supplementary Material online) and Prokka (Seemann 2014), and the two sets of annotations were merged using BEDTools (Quinlan and Hall 2010) and GenomeTools (Gremme et al. 2013), selecting the MAKER annotation when the two tools had overlapping annotations. The proteins of all green plants (*Viridiplantae*) with complete mitochondrial genome sequences in NCBI GenBank (Benson et al. 2014), 142 species, were used for protein homology evidence, the most closely related of which is the prince sago palm (*C. taitungensis*; NC_010303) (Chaw et al. 2008), being the only gymnosperm with a complete mitochondrial genome. Transfer RNA (tRNA) were annotated using ARAGORN (Laslett and Canback 2004). Ribosomal RNA (rRNA) were annotated using RNAmmer (Lagesen et al. 2007). Prokka uses Prodigal (Hyatt et al. 2010) to annotate open-reading frames (ORFs). Repeats were identified using RepeatMasker and RepeatModeler (Smit et al. 1996).

The RNA-seq reads were aligned to the annotated mitochondrial genes using BWA-MEM and variants were called using samtools and bcftools requiring a minimum genotype quality of 50 to identify possible sites of C-to-U RNA editing. A gene with an abundance of at least ten transcripts per million as quantified by Salmon (Patro et al. 2014) was considered expressed.

Results

The White Spruce Plastid Genome

The assembly and annotation metrics for the white spruce plastid and mitochondrial genomes are summarized in table 2. The plastid genome was assembled into a single circular contig of 123,266 bp containing 114 identified genes: 74 protein-coding (mRNA) genes, 36 tRNA genes, and 4 rRNA genes (fig. 1).

The majority of the protein-coding genes and tRNAs was present in single copies, with the exception of the coding genes *psbI* and *ycf12*, which were found to be duplicated. Likewise, the tRNAs *trnH-GUG*, *trnI-CAU*, *trnS-GCU*, and *trnT-GGU* were found to have two copies each. All the rRNA genes were found to be single copy.

Most of the plastid protein-coding genes had no introns. However, like other conifer plastid genomes, we found that introns were prevalent in the white spruce plastid genome. We found the protein-coding genes *atpF*, *petB*, *petD*, *rpl2*, *rpl16*, *rpoC1*, and *rps12* each contained a single intron, whereas *ycf3* contained two introns. Like the protein-coding genes, many tRNA genes including *trnA-UGC*, *trnG-GCC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, and *trnV-UAC* were split by an intron. In total we observed 15 intron insertions, 11 of which had a group II intron signature as determined by the

Table 2
Sequencing, Assembly, and Annotation Metrics of the White Spruce Organellar Genomes

Metric	Plastid	Mitochondrion
Number of lanes	1 MiSeq lane	1 HiSeq lane
Number of read pairs	4.9 million	133 million
Read length	2 × 300 bp	2 × 150 bp
Number of merged reads	3.0 million	1.4 million
Median merged read length	492 bp	465 bp
Number of assembled reads	21,000	377,000
Proportion of organellar reads	1/140 or 0.7%	1/350 or 0.3%
Depth of coverage	80×	30×
Assembled genome size	123,266 bp	5.94 Mb
Number of contigs	1 contig	130 contigs
Contig N50	123 kb	102 kb
Number of scaffolds	1 scaffold	36 scaffolds
Scaffold N50	123 kb	369 kb
Largest scaffold	123 kb	1,222 kb
GC content	38.8%	44.7%
Number of genes without ORFs	114 (108)	143 (74)
Protein-coding genes (mRNA)	74 (72)	106 (51)
rRNA genes	4 (4)	8 (3)
tRNA genes	36 (32)	29 (20)
ORFs ≥ 300 bp	Not available	1,065
Coding genes containing introns	8	5
Introns in coding genes	9	7
tRNA genes containing introns	6	0

Note.—The number of distinct genes are shown in parentheses.

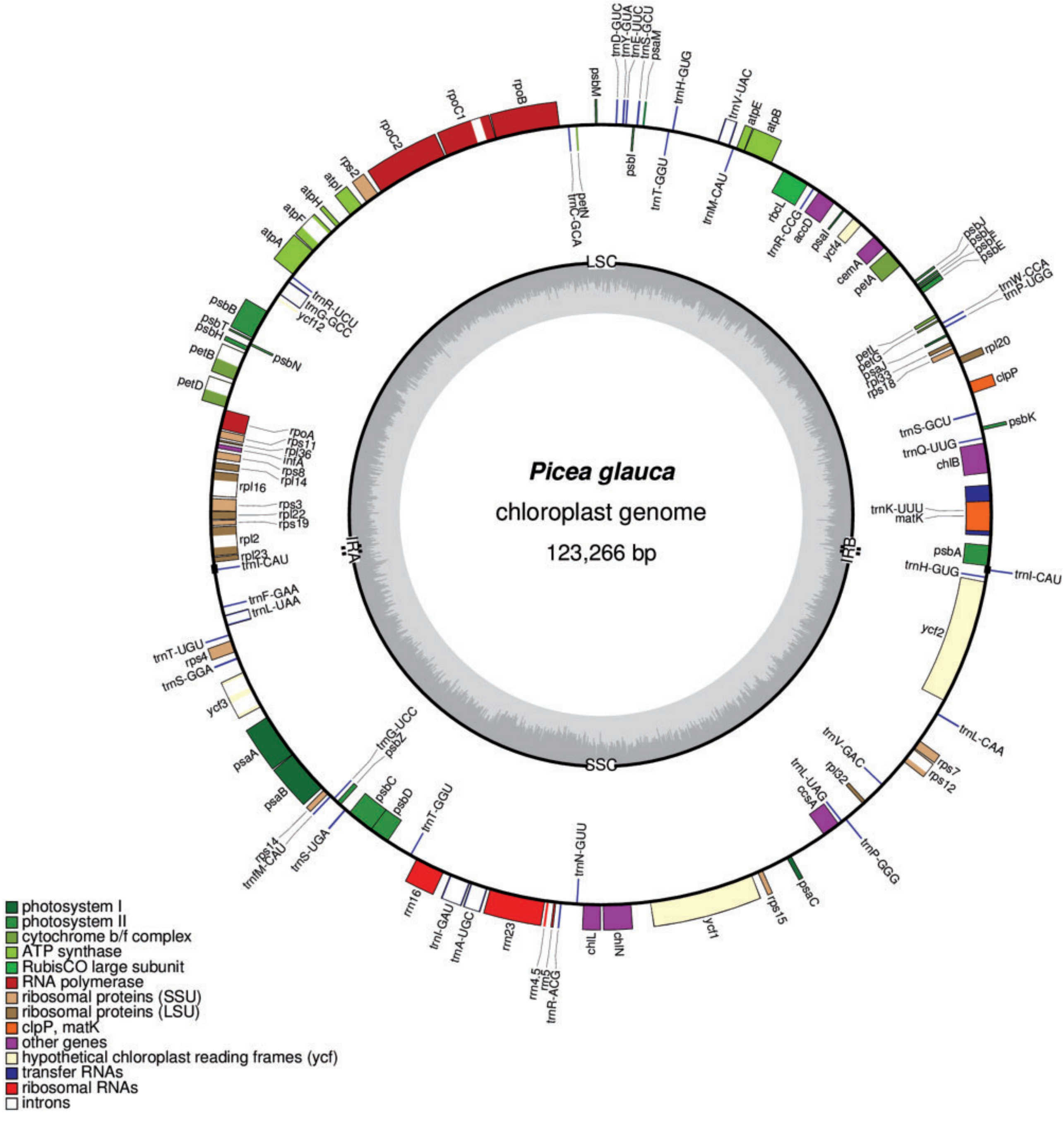

FIG. 1.—The complete plastid genome of white spruce. The PG29 white spruce chloroplast genome was annotated using MAKER and plotted using OrganellarGenomeDRAW (Lohse et al. 2007). The inner gray track depicts the G+C content of the genome.

online software RNAweasel (Lang et al. 2007). Group II introns are mobile, self-splicing ribozymes found inserted in bacterial and organellar genomes (Lambowitz and Zimmerly 2011).

Interestingly, some protein-coding genes were difficult to annotate using MAKER due to particularly small initial exons. The smallest observed exons were found for *petB*, *petD* and *rpl16* where the exons were 6, 8 and 9 bp, respectively. These genes, likely to belong to polycistronic transcripts (Barkan 1988), were annotated manually. Another gene we annotated manually was *rps12* (Hildebrand et al. 1988), as this gene is typically *trans*-spliced; where exons of two different primary transcripts are ligated together, making this difficult to annotate using MAKER. In the spruce plastid, *rps12* is composed of three exons and one *cis*-spliced intron.

The plastid genome was mostly free of repeat elements, with the exception of one class of inverted repeats (IR), present in two copies. Each copy of the IR was 445 bp in size, much smaller than most plants, typical of *Pinaceae* (Lin et al. 2010). Yet, atypically the two copies differed by a single base. In both cases, the IR contained a single gene, the tRNA *trnI-CAU*.

Having completely annotated the white spruce plastid genome, we sought to determine the similarity of this genome to other conifer plastids. Alignment of the white spruce plastid genome to that of the Norway spruce resulted in a sequence coverage of 99.7% and the sequence identity in aligned regions was 99.2%. Consistent with this similarity, we found all 114 genes of the Norway spruce plastid genome (Nystedt et al. 2013) were present in the white spruce plastid genome in perfect synteny and order. Altogether, these observations indicate that the congeneric spruce plastid genomes have not diverged significantly over evolutionary time. These data are in contrast to the level of nuclear genome conservation, which shows 3% sequence divergence between white spruce and Norway spruce (Warren, Keeling, et al. 2015). Higher sequence divergence is also frequently observed among the congeneric plastid genomes in the Angiosperms (Yang et al. 2013; Huang et al. 2014).

The White Spruce Mitochondrial Genome

The white spruce mitochondrial genome was assembled into 38 scaffolds (132 contigs) with a scaffold N50 of 369 kb (contig N50 of 102 kb). The largest scaffold was 1,222 kb (table 2). The scaffolds were aligned to the NCBI nucleotide (nt) database using BLAST. Of the 38 scaffolds, 26 scaffolds aligned to mitochondrial genomes, 3 small scaffolds (<10 kb) aligned to white spruce mRNA clones and BAC sequences, 7 small scaffolds (<10 kb) had no significant hits, and 2 small scaffolds (<5 kb) aligned to cloning vectors. These last two scaffolds were removed from the assembly.

The mitochondrial genome was rich with both annotated and putative protein-coding genes. A total of 106 protein-coding genes (51 distinct genes) were identified, which comprised 75 kb (1.3%) of the genome. In addition to 106 protein-coding genes, we found an additional 6,265 ORFs of least 90 bp (or 30 amino acids), which occupied 1.4 Mb (24%) of genome sequence, including 1,065 ORFs of at least 300 bp (100 amino acids), covering 413 kb (7%). We could not identify similar genes in the *Viridiplantae* mitochondrial genome (Benson et al. 2014) to annotate these ORFs. As such, these putative coding genes may represent novel proteins unique to the spruce or conifer mitochondrial genomes.

In addition to protein-coding genes, we identified 29 tRNAs and 8 rRNA genes. As with the spruce plastid, these ncRNAs showed variable copy number, with the majority being in a single copy. We found three copies of *trnD-GUC*, seven copies of *trnM-CAU*, and two copies of *trnY-GUA*. Five of the seven *trnM-CAU* genes share sequence similarity to the plastid translation initiator *trnfM-CAU*. The significance of the duplication of the remaining tRNA genes is unknown. Like the tRNAs, the rRNAs genes were variable with the *rrn5* present in four copies, *rrn18* in three copies, and *rrn26* in single copy. The relative order of the genes on the scaffolds and gene size is shown in figure 2. The size of each gene family is shown in figure 3. The precise position of each gene on its scaffold is shown in supplementary figure S4, Supplementary Material online.

As we observed for the plastid genome, introns were not particularly abundant, and were found to be inserted in only in a handful of genes. A total of seven intron insertions were found distributed among five protein-coding genes. These included *nad2*, *nad5*, and *nad7* which each contained one intron, and *nad4* and *rps3* which had two introns. All introns were determined to be group II introns using RNAweasel (Lang et al. 2007).

Unexpectedly, repeat elements comprised only 390 kb (6.6%) of the mitochondrial genome (fig. 4). The most commonly represented repeat elements were simple repeats and LTR Copia, ERV1, and Gypsy.

We compared the spruce mitochondrial genome with the closest sequenced spruce relative, the gymnosperm *C. taitungensis*. The *C. taitungensis* mitochondrial genome contains 39 protein-coding genes and 3 rRNA genes, all of which were also identified in white spruce. Of the 22 tRNA genes of *C. taitungensis*, we found 13 in white spruce, but the spruce mitochondrial genome also had an additional eight tRNA genes that were not observed in *C. taitungensis*.

Transfer of organellar DNA to the nucleus is common in plants (Kleine et al. 2009). To gather evidence for DNA transfer between the organellar and nuclear genome, we aligned the two organellar genomes to the WGS assembly of white spruce (Warren, Keeling, et al. 2015). As the WGS assembly may contain fragments of assembled organellar DNA, alignments with perfect identity were excluded. For aligned segments larger than 500 bp, we observed that 98% of the plastid genome and 54% of the mitochondrial genome were represented in the WGS assembly, with 7% and 4% mean DNA sequence divergence, respectively, suggesting that nearly all of the plastid genome and over half of the mitochondrial genome is represented in the white spruce nuclear genome. In comparison, 16% of the plastid genome of *Arabidopsis thaliana* is found in its nuclear genome, and 83% of the plastid genome of *Oryza sativa* is found in its nuclear genome (Shahmuradov et al. 2003). Although these results are intriguing, it is unclear whether the high sequence identity reflects true gene transfer events between the organellar and nuclear genomes of white spruce. Further investigation is warranted in future work.

The White Spruce Mitochondrial Transcriptome

Having assembled and annotated the mitochondrial genome of the white spruce, we next sought to identify the gene

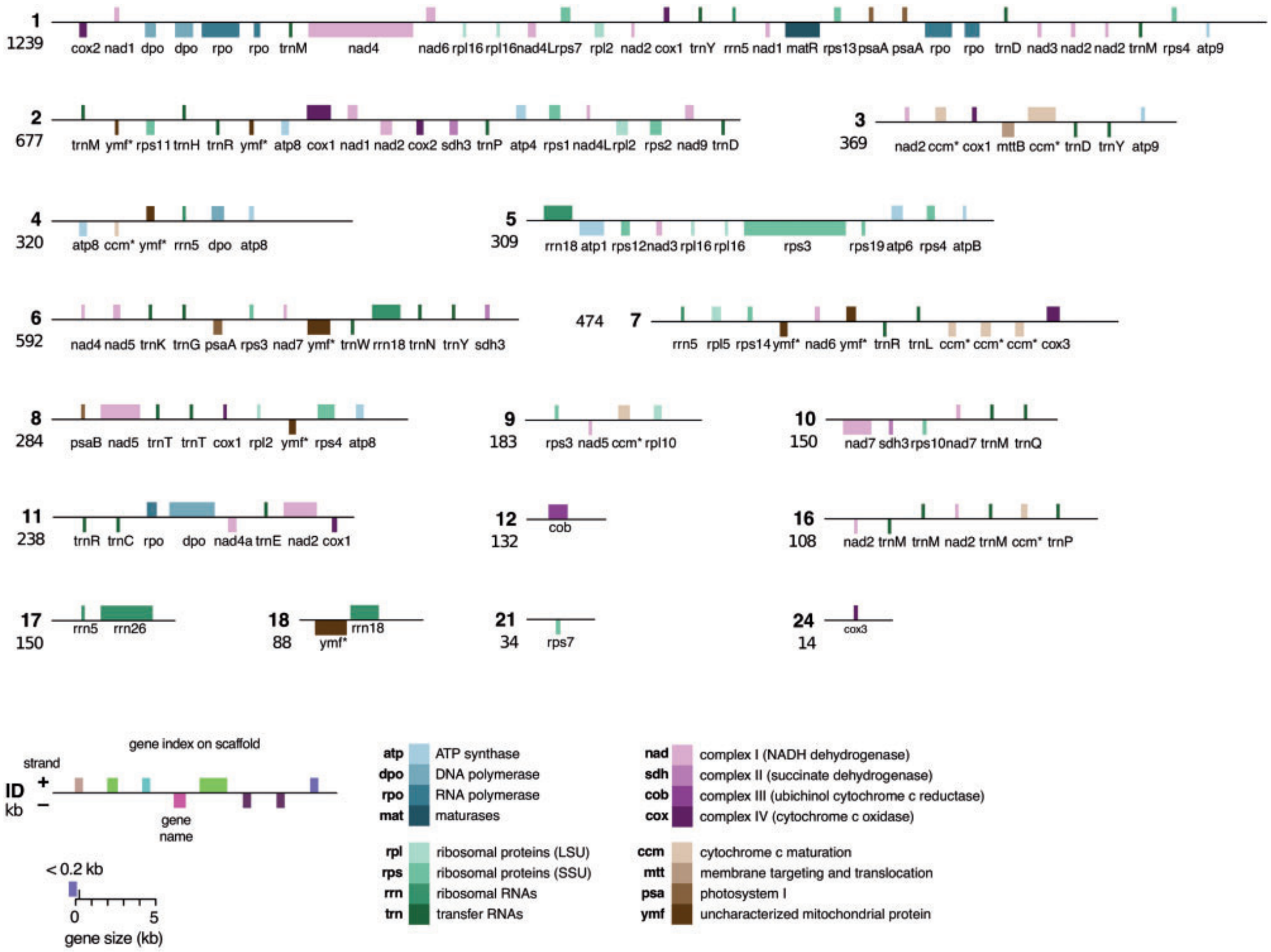
1
1239
cox2 nad1 dpo dpo rpo rpo trnM nad4 nad6 rpl16 rpl16 nad4L rps7 rpl2 nad2 cox1 trnY rrn5 nad1 matR rps13 psaA psaA rpo rpo trnD nad3 nad2 nad2 trnM rps4 atp9
2
677
trnM ymf* rps11 trnH trnR ymf* atp8 cox1 nad1 nad2 cox2 sdh3 trnP atp4 rps1 nad4L rpl2 rps2 nad9 trnD
3
369
nad2 ccm* cox1 mttB ccm* trnD trnY atp9
4
320
atp8 ccm* ymf* rrn5 dpo atp8
5
309
rrn18 atp1 rps12 nad3 rpl16 rpl16 rps3 rps19 atp6 rps4 atpB
6
592
nad4 nad5 trnK trnG psaA rps3 nad7 ymf* trnW rrn18 trnN trnY sdh3
474 7
rrn5 rpl5 rps14 ymf* nad6 ymf* trnR trnL ccm* ccm* ccm* cox3
8
284
psaB nad5 trnT trnT cox1 rpl2 ymf* rps4 atp8
9
183
rps3 nad5 ccm* rpl10
10
150
nad7 sdh3 rps10 nad7 trnM trnQ
11
238
trnR trnC rpo dpo nad4a trnE nad2 cox1
12
132
cob
16
108
nad2 trnM trnM nad2 trnM ccm* trnP
17
150
rrn5 rrn26
18
88
ymf* rrn18
21
34
rps7
24
14
cox3
gene index on scaffold
strand
ID +
kb −
gene
name
&lt; 0.2 kb
0 5
gene size (kb)
atp ATP synthase
dpo DNA polymerase
rpo RNA polymerase
mat maturases
rpl ribosomal proteins (LSU)
rps ribosomal proteins (SSU)
rrn ribosomal RNAs
trn transfer RNAs
nad complex I (NADH dehydrogenase)
sdh complex II (succinate dehydrogenase)
cob complex III (ubichinol cytochrome c reductase)
cox complex IV (cytochrome c oxidase)
ccm cytochrome c maturation
mtt membrane targeting and translocation
psa photosystem I
ymf uncharacterized mitochondrial protein

Fig. 2.—Relative order and size of genes on the scaffolds of the white spruce mitochondrial genome. Each box is proportional to the size of the gene including introns, except that genes smaller than 200 bp are shown as 200 bp. The space between genes is not to scale. An asterisk indicates that the gene name is truncated. Only scaffolds that harbor annotated genes are shown.

expression patterns of this organellar genome in the developing spruce and in the adult tissues.

First, we profiled the expression patterns of the 106 coding genes with known function across three developmental tissues and five mature tissues (fig. 5). Here, we found that at least 29 genes were expressed in at least one of the developing tissues, but not in a mature tissue, suggesting some mitochondrial transcripts may be expressed during spruce development. In contrast, we found 60 genes to be expressed in at least one of the mature tissues and a total of 17 genes were not expressed. The developing spruce megagametophyte and embryo were the most transcriptionally active tissues and clustered together using an unsupervised clustering algorithm.

Conversely, although 2,809/6,265 (45%) of ORFs at least 90 bp were expressed in at least one developing tissue, only 427/6,265 (7%) were expressed in at least one mature tissue (table 3 and fig. 6). Nearly half (3,029/6,265; 48%) did not have detectable expression in any tissue sampled.

Many conifer organellar transcripts are subject to C-to-U RNA editing (Chateigner-Boutin and Small 2010). To explore this possibility for the mitochondrial transcriptome of white spruce, we analyzed RNA-seq read alignments against the reconstructed mitochondrial genome using a custom pipeline (supplementary listing S1, Supplementary Material online). RNA editing events were predicted as nucleotide positions where the genomic sequence shows a C-residue, but the RNA-sequencing data suggest a T (supplementary table S5, Supplementary Material online). Correctly calling RNA editing events can be confounded by genomic single nucleotide variants and by false-positives originating from misaligned reads. However, we see an enrichment of C-to-T variants in 91% (1,601 of 1,751) positions (supplementary table S6, Supplementary Material online), suggesting a large fraction of these are true C-to-U RNA edits.

We find C-to-U RNA editing events occurred in 68 of the 106 coding genes (supplementary table S5, Supplementary Material online), with the most highly edited gene, *nad3*,

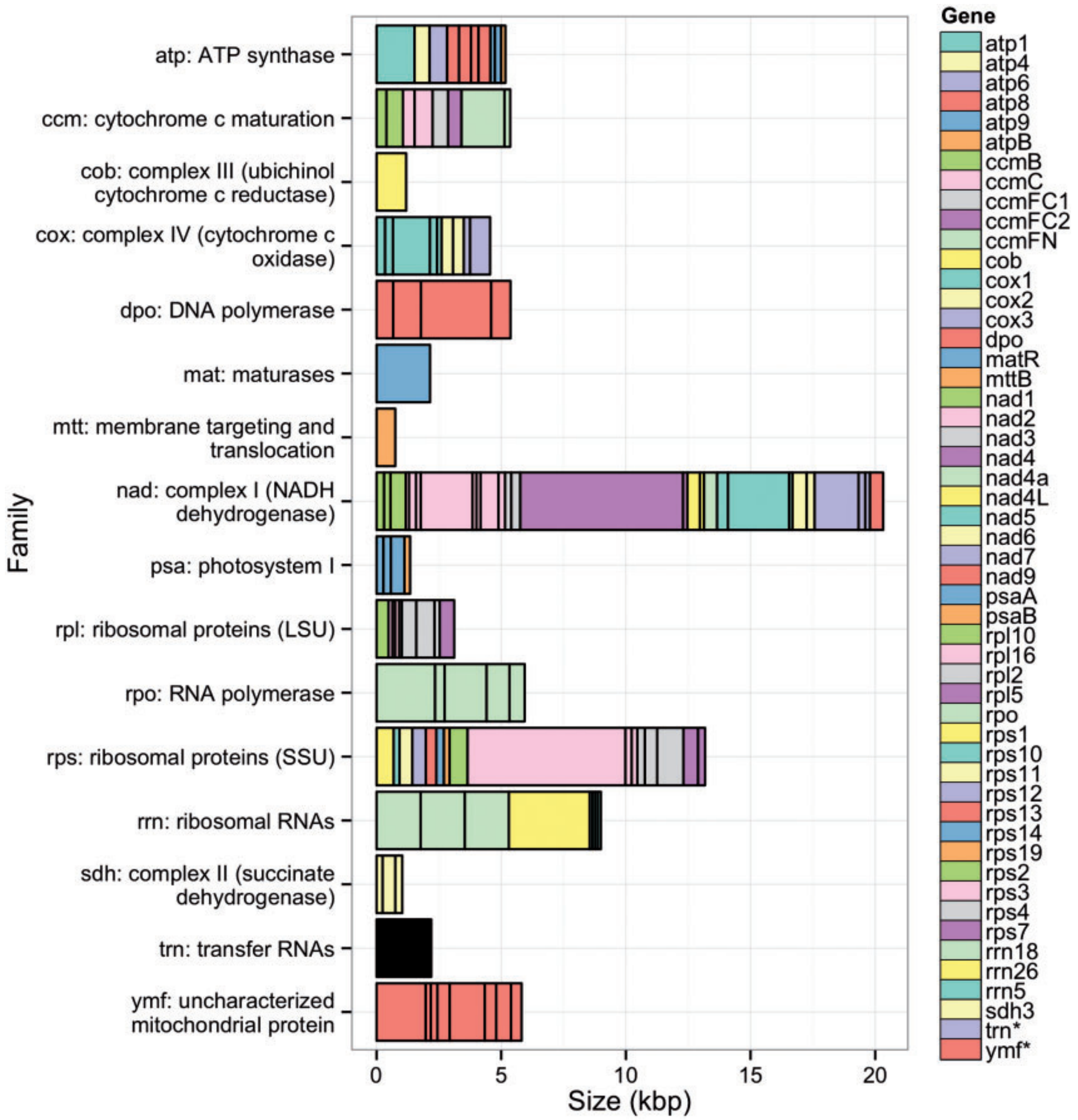

Fig. 3.—Gene content of the white spruce mitochondrial genome, grouped by gene family. Each box is proportional to the size of the gene including introns. The color of each gene is unique within its gene family.

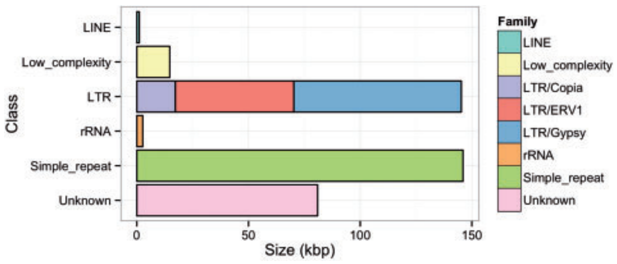

Fig. 4.—Repetitive sequence content of the white spruce mitochondrial genome, annotated using RepeatMasker and RepeatModeler.

edited at a rate of nine edits per 100 bp. Uniquely, C-to-U RNA editing can generate new start and stop codons, but it is unable to destroy existing start and stop codons. In organellar genomes, editing of the ACG (Thr) codon to AUG (Met), which creates a novel start codon, is frequently observed (Neckermann et al. 1994). In the white spruce mitochondrial genome, we observed four such editing events for the genes *mttB*, *nad1*, *rps3*, and *rps4*.

Discussion

In this work, we outline an informatics methodology to assemble organellar genomes of plant species using WGS

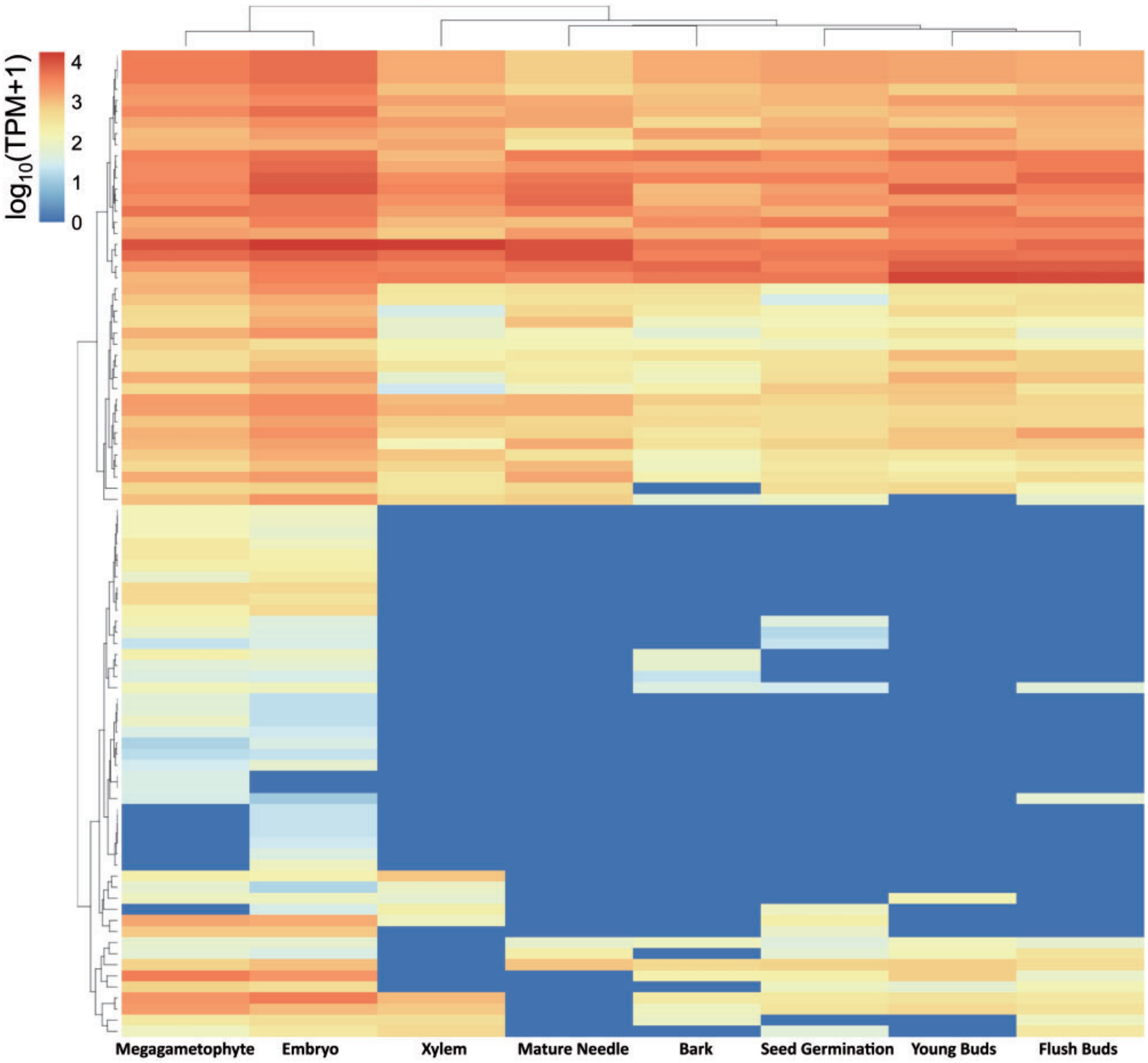

FIG. 5.—Heatmap of the transcript abundance of mitochondrial protein-coding genes of white spruce. Each column is a tissue sample. Each row is a gene. Each cell represents the transcript abundance of one gene in one sample. The color scale is $\log_{10}$(TPM+1), where TPM is transcripts per million as measured by Salmon (Patro et al. 2014).

Table 3
Number of Expressed Protein-Coding Genes and ORFs of the White Spruce Mitochondrial Transcriptome Tabulated by Developmental Stage

	Both	Mature Only	Developing Only	Neither	Sum
CDS	60	0	29	17	106
ORF	411	16	2,809	3,029	6,265
Sum	471	16	2,838	3,046	6,371

sequencing data. Usually plant genomics projects generate such data sets to reconstruct and study the nuclear genomes of their target species. Here we demonstrate that the same data sets can be mined for the organellar genomes, providing added value to those projects with no additional cost to experimental budgets.

One lane of MiSeq sequencing of whole-genome DNA (4.9 million read pairs) was sufficient to assemble the 123-kb plastid genome, and one lane of HiSeq sequencing of whole-genome DNA (133 million read pairs) was sufficient to assemble the 5.9-Mb mitochondrial genome of white spruce.

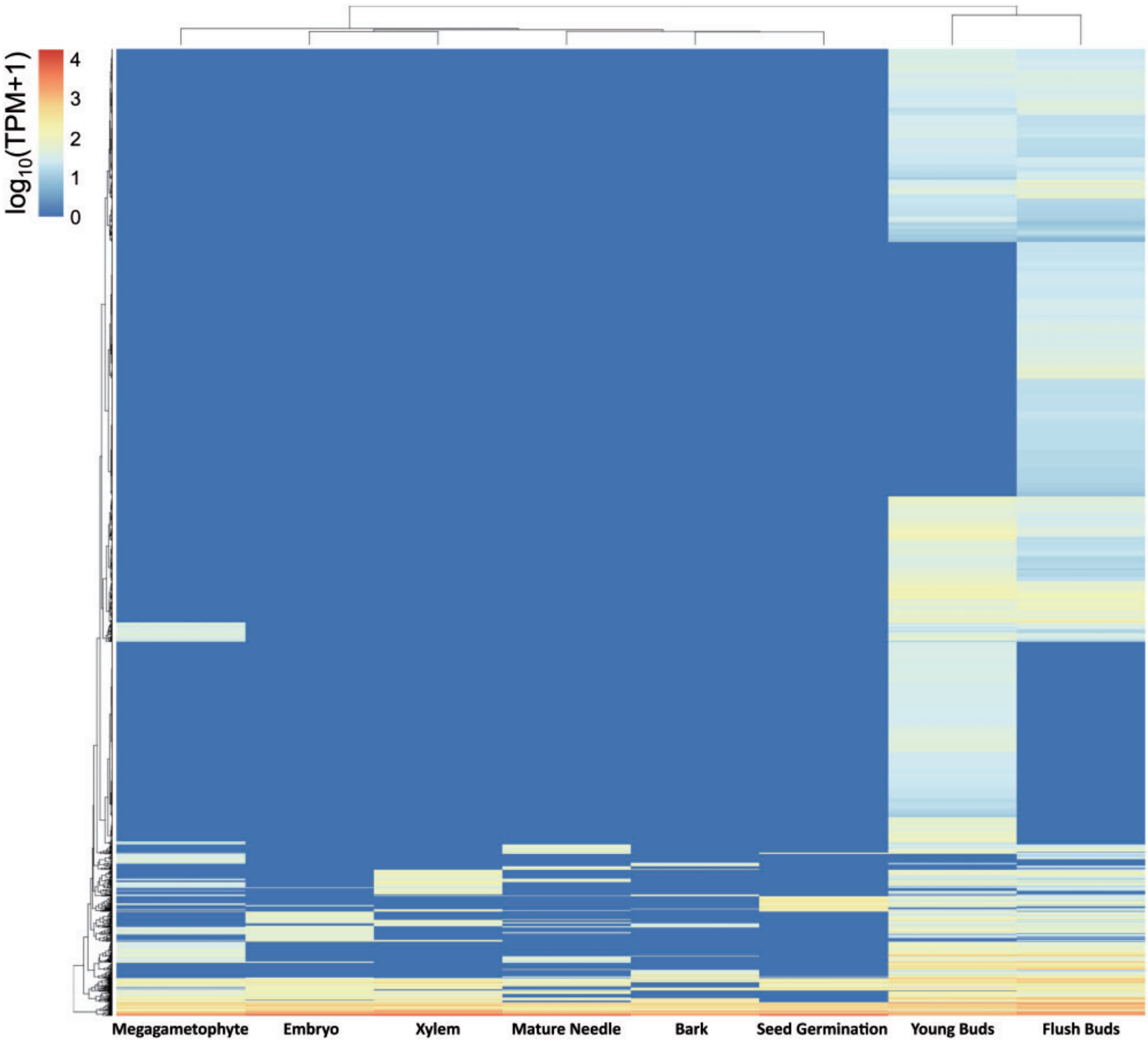

FIG. 6.—Heatmap of the transcript abundance of mitochondrial protein-coding genes of white spruce, including ORFs. Each column is a tissue sample. Each row is a gene. Each cell represents the transcript abundance of one gene in one sample. The color scale is $\log_{10}$(TPM+1), where TPM is transcripts per million as measured by Salmon (Patro et al. 2014).

Additional Illumina and PacBio sequencing was used to improve scaffold contiguity and to close scaffold gaps, after which the plastid genome was assembled in a single contig, and the largest mitochondrial scaffold was 1.2 Mb.

In previous studies, analysis of cpDNA was useful in reconstructing phylogenies of plants (Wu et al. 2007), in determining the origin of an expanding population (Aizawa et al. 2012), and in determining when distinct lineages of a species resulted from multiple colonization events (Jardon-Barbolla et al. 2011). The contrasting inheritance schemes of plastids and mitochondria can be useful in the characterization of species expanding their range. In the case of two previously allopatric species now found in sympatry and hybridizing, the mitochondrial DNA (mtDNA) would be contributed by the resident species, whereas introgression of the plastid genome into the expanding species would usually be limited, as pollen would be more readily dispersed than seeds (Du et al. 2011). Differential gene flow of cpDNA and mtDNA due to different modes of inheritance and dispersion would result in new assemblages of organellar genomes and an increase of genetic diversity after expansion from a refugium (Gerardi et al. 2010).

The white spruce plastid genome showed no structural rearrangements when compared with that of Norway spruce despite the divergence of these two species, estimated at more than 10 Ma (Bouillé and Bousquet 2005). All genes of the Norway spruce plastid were present and collinear in the white spruce plastid, and no new plastid genes were found in white spruce. The remarkable level of sequence conservation between these spruces for the plastid genome was also in contrast to the nuclear genome, suggesting strong evolutionary pressure to maintain a functional plastid genome. The plastids of the angiosperms demonstrate frequent rearrangements (Palmer and Herbon 1988; Knox 2014) or higher sequence divergence (Yang et al. 2013; Huang et al. 2014). Likewise, comparative genomics among the five extant gymnosperm group show that Pinaceae and non-Pinaceae conifers (cupressophytes) have lost a different copy of IR (Wu et al. 2011; Yi et al. 2013; Wu and Chaw 2014), and may explain the reduced diversity of cpDNA organizations in Pinaceae. It will be interesting to see whether high sequence conservation is observed for plastid genomes within the Pinaceae family, and particularly within the genus *Picea*.

All genes of the prince sago palm (*C. taitungensis*) mitochondrial genome were also present in white spruce, but mitochondrial ORFs were found that were unique to white spruce. The protein-coding gene content of the white spruce mitochondrial genome was quite sparse, with 106 protein-coding genes contained in 5.9 Mb, in comparison to the plastid genome with 74 protein-coding genes in 123 kb. The mitochondrial genome of white spruce appears quite large for a gymnosperm, compared with the 415-kb mitochondrial genome of *Cycas* (Chaw et al. 2008) or the estimate of 750 kb–1 Mb obtained from southern blot analysis for the conifer *Larix* (Kumar et al. 1995). Nearly 7% of the white spruce mitochondrial genome was composed of repeats, and roughly 1% was composed of coding genes. Thus, a significant portion of the unusually large white spruce mitochondrial genome remains to be characterized.

Sequencing and annotation of organellar genomes in spruce trees offers significant advancement in our understanding of conifer biology and evolution while providing a reference for further research. For instance, the remarkable level of structure and sequence conservation of the plastid genome between the distantly related white spruce and Norway spruce indicates strong selective pressures not only on genes but also on intergenic regions and overall genome structure, which should facilitate comparative evolutionary studies in the genus. Further investigations implicating several other Pinaceae genera appear necessary to assess the extent of this trend at a larger phylogenetic scale and how different it is from the trends seen in other plant groups. The present reference genomes should also be helpful in resequencing projects so to better identify islands of intraspecific sequence variation and how they vary among conifer taxa. Overall, these new reference genomes should help develop applications for the management and conservation of natural genetic diversity in this group of ecologically and economically important trees.

Supplementary Material

Supplementary listing S1, figures S1–S3, and tables S1–S6 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Genome Canada, Genome British Columbia, and Genome Quebec as part of the SMarTForests Project (www.smartforests.ca last accessed 17 Dec 2015). The authors thank Carson Holt for his help with the MAKER analysis and Martin Krzywinski for helping draft some of the figures.

Literature Cited

Aizawa M, Kim ZS, Yoshimaru H. 2012. Phylogeography of the Korean pine (*Pinus koraiensis*) in northeast Asia: inferences from organelle gene sequences. J Plant Res. 125:713–723.

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

Alverson AJ, et al. 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol. 27:1436–1448.

Barkan A. 1988. Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic mRNAs. Embo J. 7:2637–2644.

Benson DA, et al. 2014. GenBank. Nucleic Acids Res. 42:D32–D37.

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.

Bock DG, Kane NC, Ebert DP, Rieseberg LH. 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytol. 201:1021–1030.

Bouillé M, Bousquet J. 2005. Trans-species shared polymorphisms at orthologous nuclear gene loci among distant species in the conifer *Picea* (Pinaceae): implications for the long-term maintenance of genetic diversity in trees. Am J Bot. 92:63–73.

Bouillé M, Senneville S, Bousquet J. 2011. Discordant mtDNA and cpDNA phylogenies indicate geographic speciation and reticulation as driving factors for the diversification of the genus *Picea*. Tree Genet Genomes. 7:469–484.

Campbell MS, et al. 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164:513–524.

Chateigner-Boutin AL, Small I. 2010. Plant RNA editing. RNA Biol. 7:213–219.

Chaw SM, 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. 25:603–615.

Chen J, et al. 2011. Substoichiometrically different mitotypes coexist in mitochondrial genomes of *Brassica napus* L. PLoS One 6:e17662.

Du FK, et al. 2011. Direction and extent of organelle DNA introgression between two spruce species in the Qinghai-Tibetan Plateau. New Phytol. 192:1024–1033.

Fajardo D, et al. 2014. The American cranberry mitochondrial genome reveals the presence of selenocysteine (tRNA-Sec and SECIS) insertion machinery in land plants. Gene 536:336–343.

Gerardi S, Jaramillo-Correa JP, Beaulieu J, Bousquet J. 2010. From glacial refugia to modern populations: new assemblages of organelle genomes generated by differential cytoplasmic gene flow in transcontinental black spruce. Mol Ecol. 19:5265–5280.

Gordon D, Green P. 2013. Consed: a graphical editor for next-generation sequencing. Bioinformatics 29:2936–2937.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform. 10:645–656.

Grewe F, et al. 2014. Comparative analysis of 11 Brassicales mitochondrial genomes and the mitochondrial transcriptome of *Brassica oleracea*. Mitochondrion 19(Pt B):135–143.

Guo W, et al. 2014. Predominant and substoichiometric isomers of the plastid genome coexist within *Juniperus* plants and have shifted multiple times during cupressophyte evolution. Genome Biol Evol. 6:580–590.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075.

Hildebrand M, Hallick RB, Passavant CW, Bourque DP. 1988. Trans-splicing in chloroplasts: the rps 12 loci of *Nicotiana tabacum*. Proc Natl Acad Sci U S A. 85:372–376.

Huang DI, Hefer CA, Kolosova N, Douglas CJ, Cronk QC. 2014. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). New Phytol. 204:693–703.

Hyatt D, et al. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

Jardon-Barbolla L, Delgado-Valerio P, Geada-Lopez G, Vazquez-Lobo A, Pinero D. 2011. Phylogeography of *Pinus* subsection Australes in the Caribbean Basin. Ann Bot. 107:229–241.

Jeong YM, et al. 2014. The complete mitochondrial genome of cultivated radish WK10039 (*Raphanus sativus* L.). Mitochondrial DNA 12. 27(2):941–942.

Jo YD, Choi Y, Kim DH, Kim BD, Kang BC. 2014. Extensive structural variations between mitochondrial genomes of CMS and normal peppers (*Capsicum annuum* L.) revealed by complete nucleotide sequencing. BMC Genomics 15:561.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.

Keren I, et al. 2009. AtnMat2, a nuclear-encoded maturase required for splicing of group-II introns in Arabidopsis mitochondria. RNA 15:2299–2311.

Kim DH, Kang JG, Kim BD. 2007. Isolation and characterization of the cytoplasmic male sterility-associated orf456 gene of chili pepper (*Capsicum annuum* L.). Plant Mol Biol. 63:519–532.

Kleine T, Maier UG, Leister D. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol. 60:115–138.

Knox EB. 2014. The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. Proc Natl Acad Sci U S A. 111:11097–11102.

Kumar R, Lelu MA, Small I. 1995. Purification of mitochondria and mitochondrial nucleic acids from embryogenic suspension cultures of a gymnosperm, *Larix x leptoeuropaea*. Plant Cell Rep. 14:534–538.

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Lagesen K, et al. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35:3100–3108.

Lambowitz AM, Zimmerly S. 2011. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb Perspect Biol. 3:a003616.

Lang BF, Laforest MJ, Burger G. 2007. Mitochondrial introns: a critical view. Trends Genet. 23:119–125.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 32:11–16.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.

Lin CP, Huang JP, Wu CS, Hsu CY, Chaw SM. 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biol Evol. 2:504–517.

Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet. 52:267–274.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955–964.

Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Neckermann K, Zeltz P, Igloi GL, Kossel H, Maier RM. 1994. The role of RNA editing in conservation of start codons in chloroplast genomes. Gene 146:177–182.

Nystedt B, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584.

Palmer JD, Herbon LA. 1988. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol. 28:87–97.

Park S, et al. 2014. Complete sequences of organelle genomes from the medicinal plant *Rhazya stricta* (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. BMC Genomics 15:405.

Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 32:462–464.

Paulino D, et al. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16:230.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Rice DW, et al. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science 342:1468–1473.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

Shahmuradov IA, Akbarova YY, Solovyev VV, Aliyev JA. 2003. Abundance of plastid DNA insertions in nuclear genomes of rice and *Arabidopsis*. Plant Mol Biol. 52:923–934.

Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123.

Sloan DB, et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10:e1001241.

Smit AFA, Hubley R, Green P. 1996. RepeatMasker Open-3.0. Available from: http://www.repeatmasker.org.

Vandervalk BP, et al. 2015. Konnector v2.0: pseudo-long reads from paired-end sequencing data. BMC Med Genomics. 8(Suppl 3):S1.

Vieira Ldo N, Faoro H, Fraga HP, et al. 2014. An improved protocol for intact chloroplasts and cpDNA isolation in conifers. PLoS One 9:e84792.

Vieira Ldo N, Faoro H, Rogalski M, et al. 2014. The complete chloroplast genome sequence of *Podocarpus lambertii*: genome structure, evolutionary aspects, gene content and SSR detection. PLoS One 9:e90618.

Wang Y, et al. 2014. Complete mitochondrial genome of *Eruca sativa* Mill. (Garden rocket). PLoS One 9:e105748.

Warren RL, Keeling CI, et al. 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 83:189–212.

Warren RL, Yang C, et al. 2015. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. GigaScience 4:35.

Whittle CA, Johnston MO. 2002. Male-driven evolution of mitochondrial and chloroplastidial DNA sequences in plants. Mol Biol Evol. 19:938–949.

Wu CS, Chaw SM. 2014. Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. Plant Plant Biotechnol J. 12:344–353.

Wu CS, Lin CP, Hsu CY, Wang RJ, Chaw SM. 2011. Comparative chloroplast genomes of Pinaceae: insights into the mechanism of diversified genomic organizations. Genome Biol Evol. 3:309–319.

Wu CS, Wang YN, Liu SM, Chaw SM. 2007. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. Mol Biol Evol. 24:1366–1379.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.

Yang JB, Tang M, Li HT, Zhang ZR, Li DZ. 2013. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol Biol. 13:84.

Yi X, Gao L, Wang B, Su YJ, Wang T. 2013. The complete chloroplast genome sequence of *Cephalotaxus oliveri* (*Cephalotaxaceae*): evolutionary comparison of *Cephalotaxus* chloroplast DNAs and insights into the loss of inverted repeat copies in gymnosperms. Genome Biol Evol. 5:688–698.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.

Zhang Y, et al. 2014. The complete chloroplast genome sequence of *Taxus chinensis* var. *mairei* (*Taxaceae*): loss of an inverted repeat region and comparative analysis with related species. Gene 540:201–209.

Associate editor: Shu-Miaw Chaw