Complete Chloroplast Genome Sequence of a Black Spruce (*Picea mariana*) from Eastern Canada

Theodora Lo,[a] Lauren Coombe,[a] Diana Lin,[a] René L. Warren,[a] Heather Kirk,[a] Pawan Pandoh,[a] Yongjun Zhao,[a] Richard A. Moore,[a] Andrew J. Mungall,[a] Carol Ritland,[b,e] Jean Bousquet,[c] Steven J. M. Jones,[a] Joerg Bohlmann,[b,e] Ashley Thomson,[d] Inanc Birol[a]

[a]Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
[b]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC, Canada
[c]Canada Research Chair in Forest Genomics, Université Laval, Quebec City, QC, Canada
[d]Faculty of Natural Resources Management, Lakehead University, Thunder Bay, ON, Canada
[e]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

ABSTRACT Here, we present the chloroplast genome sequence of black spruce (*Picea mariana*), a conifer widely distributed throughout North American boreal forests. This complete and annotated chloroplast sequence is 123,961 bp long and will contribute to future studies on the genetic basis of evolutionary change in spruce and adaptation in conifers.

Global climate change is predicted to impact the growth of *Picea mariana* (black spruce), a dominant species of significant ecological and economic importance in Canada's boreal forests (1, 2). Black spruce has demonstrated local adaptations to climate (3). Determining the genetic and molecular bases of these adaptations can provide valuable insights into mitigating climate change effects on Canada's forests (2, 3).

An unannotated black spruce chloroplast draft genome assembly with several gaps was submitted to GenBank (accession number LT727842.1) in 2018. Here, we present a complete and annotated black spruce chloroplast genome sequence from a different genotype.

A black spruce needle tissue sample (genotype 40-10-1) was collected in Thunder Bay, Ontario (50°57′39.96″N, 90°27′20.16″E; elevation 741 m). Following nucleus purification, genomic DNA was extracted by Bio S&T using a cetyltrimethylammonium bromide (CTAB)/chloroform method, yielding 60 μg of high-quality purified DNA (4, 5). A sequencing library was prepared using the Chromium linked-read platform from 10X Genomics (5) and sequenced with paired-end 150-base pair reads on an Illumina HiSeq X instrument at Canada's Michael Smith Genome Sciences Centre.

One lane of sequencing data, consisting of 428,820,113 read pairs, was used to assemble the chloroplast genome. After trimming adapters using Trimadap vr11 (6), subsets were sampled (n = 0.75, 1.5, 3, 6, 12, 25, 50, and 200 million read pairs) to reduce the noise from nuclear and mitochondrial DNA.

Each subsample was assembled with ABySS v2.1.0 (7) using various k-mer sizes (k = 64 to 104, step 8) and k-mer count thresholds (kc = 3 and 4). Chloroplast sequences in the assemblies were extracted from BWA-MEM v0.7.17 (8) alignments of scaffolds to the reference white spruce chloroplast genome (genotype WS77111; GenBank accession number MK174379) (9) and evaluated with QUAST v5.02 (10). The assembly with the highest NGA50 length of 42,639 bp (where NGA50 indicates the length of the shortest aligned scaffold, with all aligned scaffolds at least NG50 making up at least 50% of the target genome) and 0 misassemblies (25 million read pair subset; k = 104; kc = 4) was further scaffolded using ntJoin v1.0.1 (11), supplying the white spruce chloroplast genome as the reference and setting reference_weights="2".

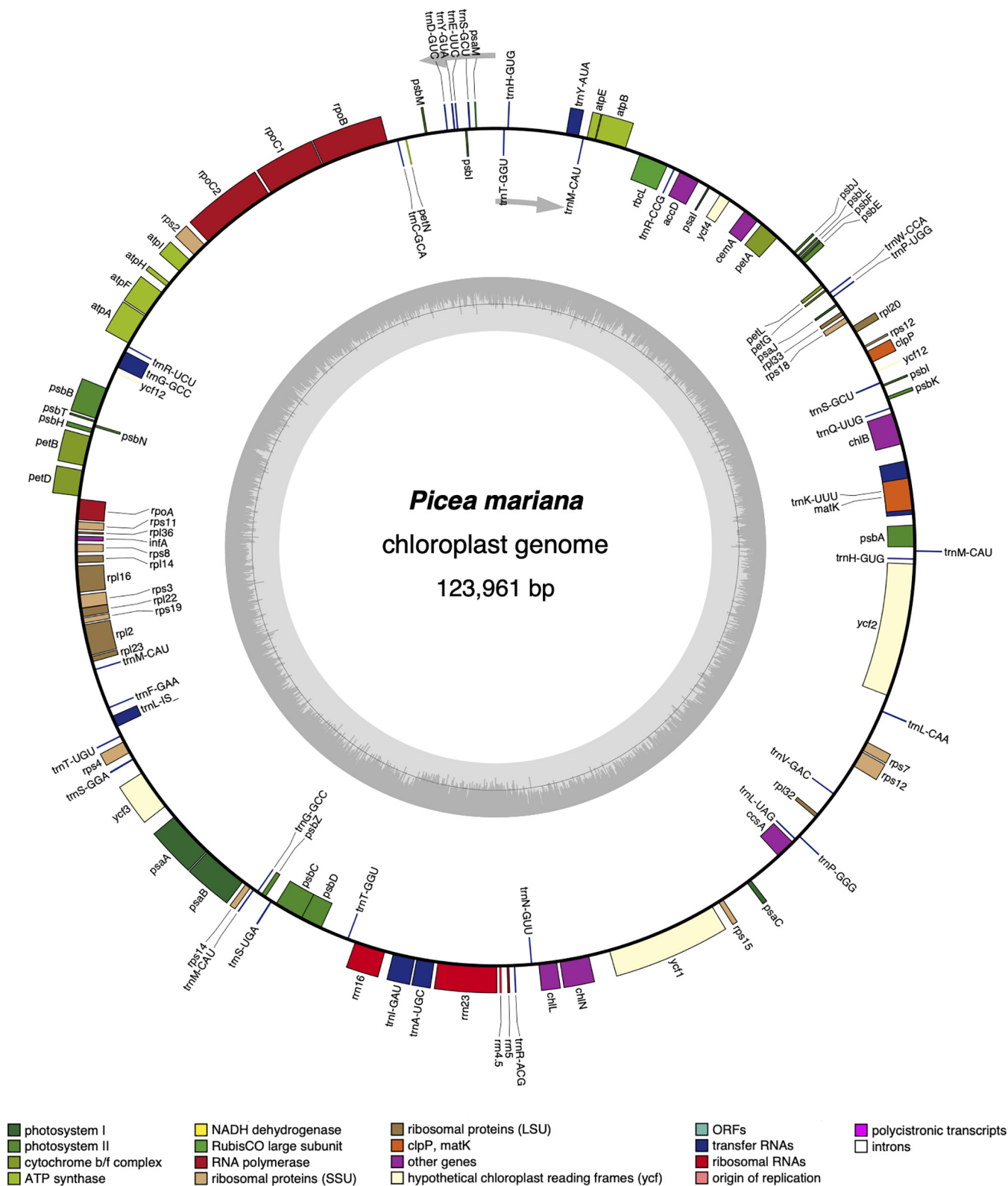**FIG 1** The complete chloroplast genome of *Picea mariana* genotype 40-10-1. The *Picea mariana* chloroplast genome was annotated using GeSeq and plotted using OGDRAW (16). The inner gray circle illustrates the GC content of the genome, and the outer circle shows the annotated genes as rectangular boxes with labels, colored by functional categories. The arrows indicate the direction of transcription for each DNA strand.

Remaining gaps in the resulting scaffold were filled using Sealer v2.2.3 (12) with multiple values of k (k = 70 to 120, step 10), and the assembly was polished using Pilon v1.23 (13) with --diploid --fix all options. Approximately 700 bp on the two ends of our assembly were successfully recovered by supplying the 3′ and 5′ ends of our draft to Sealer v2.2.3 (12) using the abovementioned parameters, yielding a complete chloroplast genome. Finally, BLAST v2.10.0 (14) was used to adjust the start position for consistency with previously published chloroplast genomes. Note that default parameters were used unless otherwise specified.

The complete *Picea mariana* chloroplast genome is 123,961 bp long with a GC content of 38.70%. Using GeSeq v1.79 (15), with several *Picea* sp. chloroplast genomes as references, we successfully annotated 114 genes, including 74 protein-coding, 36 tRNA-coding, and 4 rRNA-coding genes (Fig. 1). Due to a frameshift mutation, *psbZ* was annotated as a pseudogene. Also, the annotations of *petB*, *petD*, and *rpl16* were corrected manually.

Offering this chloroplast genome to the community will enrich public genomic repositories of spruce species, facilitate research on climate adaptation, and contribute to the development of forest management policies.

Data availability. The complete chloroplast genome sequence of *Picea mariana*, genotype 40-10-1, is available from GenBank under accession number MT261462, and the raw sequencing reads are available from the SRA under SRX7890468 and SRR11284755. The annotations used as references include those from *Picea abies* (NC_021456), *Picea asperata* (NC_032367), *Picea engelmannii* (NC_041067), *Picea glauca* genotype WS77111 (MK174379), *Picea morrisonicola* (NC_016069), *Picea sitchensis* (KU215903), *Picea chihuahuana* (NC_039584), *Picea crassifolia* (NC_032366), and *Picea jezoensis* (NC_029374).

REFERENCES

1. Girardin MP, Hogg EH, Bernier PY, Kurz WA, Guo XJ, Cyr G. 2016. Negative impacts of high temperatures on growth of black spruce forests intensify with the anticipated climate warming. Glob Chang Biol 22:627–643. https://doi.org/10.1111/gcb.13072.

2. Thomson AM, Parker WH, Riddell CL. 2009. Boreal forest provenance tests used to predict optimal growth and response to climate change: 2. black spruce. Can J For Res 39:143–153. https://doi.org/10.1139/X08-167.

3. Prunier J, Gérardi S, Laroche J, Beaulieu J, Bousquet J. 2012. Parallel and lineage-specific molecular adaptation to climate in boreal black spruce. Mol Ecol 21:4270–4286. https://doi.org/10.1111/j.1365-294X.2012.05691.x.

4. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MMS, Keeling CI, Brand D, Vandervalk BP, Kirk H, Pandoh P, Moore RA, Zhao Y, Mungall AJ, Jaquish B, Yanchuk A, Ritland C, Boyle B, Bousquet J, Ritland K, Mackay J, Bohlmann J, Jones SJ. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497. https://doi.org/10.1093/bioinformatics/btt178.

5. Taylor GA, Kirk H, Coombe L, Jackman SD, Chu J, Tse K, Cheng D, Chuah E, Pandoh P, Carlsen R, Zhao Y, Mungall AJ, Moore R, Birol I, Franke M, Marra MA, Dutton C, Jones SJ. 2018. The genome of the American brown bear or grizzly: Ursus arctos ssp. horribilis. Genes 9:598. https://doi.org/10.3390/genes9120598.

6. Li H. 2014. Trimadap: fast but inaccurate adapter trimmer for Illumina reads (r11). https://github.com/lh3/trimadap.

7. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27:768–777. https://doi.org/10.1101/gr.214346.116.

8. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

9. Lin D, Coombe L, Jackman SD, Gagalova KK, Warren RL, Hammond SA, Kirk H, Pandoh P, Zhao Y, Moore RA, Mungall AJ, Ritland C, Jaquish B, Isabel N, Bousquet J, Jones SJ, Bohlmann J, Birol I. 2019. Complete chloroplast genome sequence of a white spruce (*Picea glauca*, Genotype WS77111) from Eastern Canada. Microbiol Resour Announc 8:e00381-19. https://doi.org/10.1128/MRA.00381-19.

10. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34:i142–i150. https://doi.org/10.1093/bioinformatics/bty266.

11. Coombe L, Nikolić V, Chu J, Birol I, Warren RL. 2020. ntJoin: fast and lightweight assembly-guided scaffolding using minimizer graphs. Bioinformatics 36:3885–3887. https://doi.org/10.1093/bioinformatics/btaa253.

12. Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, Birol I. 2015. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics 16:230. https://doi.org/10.1186/s12859-015-0663-4.

13. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal.pone.0112963.

14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

15. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq—versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6–W11. https://doi.org/10.1093/nar/gkx391.

16. Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res 47:W59–W64. https://doi.org/10.1093/nar/gkz238.