Complete chloroplast genome sequence of *Picea engelmannii*, isolate Se404-851 from western Canada

Diana Lin[a], Lauren Coombe[a], Shaun D. Jackman[a], Kristina K. Gagalova[a], René L. Warren[a], S. Austin Hammond[a*], Helen McDonald[a], Heather Kirk[a], Pawan Pandoh[a], Yongjun Zhao[a], Richard A. Moore[a], Andrew J. Mungall[a], Carol Ritland[b,e], Trevor Doerksen[c], Barry Jaquish[c], Jean Bousquet[d], Steven J.M. Jones[a], Joerg Bohlmann[b,e], and 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]British Columbia Ministry of Forests, Lands, and Natural Resource Operations, Victoria, BC, Canada
[d]Canada Research Chair in Forest Genomics, Université Laval, QC, Canada
[e]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

Running Head: Complete chloroplast genome of a *Picea engelmannii*

#Address correspondence to Inanc Birol, ibirol@bcgsc.ca

*Current address: S. Austin Hammond, Next-Generation Sequencing Facility, University of Saskatchewan, Saskatoon, SK, Canada.

1

Abstract

Engelmann spruce (*Picea engelmannii*) is a conifer found primarily on the west coast of North America. Here, we present the complete chloroplast genome sequence of *Picea engelmannii*, isolate Se404-851. This chloroplast sequence will benefit future conifer genomic research and contribute resources to further species conservation efforts.

Genome Announcement

We sequenced, assembled, and annotated the complete chloroplast genome of Engelmann spruce (*Picea engelmannii*, isolate Se404-851). Engelmann spruce dominates much of the large spruce forests of interior British Columbia, where it has been reported to hybridize with *Picea glauca* and *Picea sitchensis* (1), and its range extends southward to New Mexico. The tree has three different genomes, a nuclear genome, a mitochondrial genome, and a plastid (i.e. chloroplast) genome. In general, chloroplast genomes are derived from the ancestral genomes of the microbial endosymbiont from which these organelles originated (2).

A needle tissue sample was collected from a 13-year-old Engelmann spruce grown at the Kalamalka Forestry Centre in British Columbia (50°14'38.4"N 119°16'40.8"W; elevation of 450m), planted from a seed from Don Fernando Mountain, New Mexico (36°17'60"N 105°24'0"W; elevation: 2987m). The sample was sequenced at Canada's Michael Smith Genome Sciences Centre.

To sequence the sample, a 900 bp whole genome library was constructed following the previously described protocol (3-4) with minor modifications. Briefly, 5 μg of genomic DNA was subjected to shearing by sonication (Covaris LE220) using a Duty Factor of 5 and Peak Incident Power of 450 for 70 seconds. The sonicated DNA products were fractionated in a 6% PAGE gel to recover fragments greater than 700 bp for library preparation. These PCR-Free libraries were sequenced with paired-end 150 base reads on the Illumina HiSeqX platform using V4 chemistry according to manufacturer recommendations.

To assemble the chloroplast genome, we subsampled the whole genome shotgun sequencing reads to subsets of 0.75, 1.5, 3, 6, 12, 25, 46 million read pairs, then assembling each subset with ABySS v2.1.1 (5) ($k = 128$, $kc = 3$). The ABySS assembly of the 3M read pair subset resulted in a single 123,601 contig that aligned to the reference chloroplast sequence (*Picea glauca* admix genotype PG29, NCBI accession NC_028594; (6)), with zero misassemblies and internal gaps, based on QUAST v5.0.0 (7) analysis.

Using BLAST v2.7.1 (8), we aligned our assembly to the reference sequence (PG29), previously, modifying start and stop positions for consistency with previously published conifer chloroplast genomes. To ensure there were no missing sequences at the ends of our assembly we introduced a gap at the end, circularized the sequence, and ran Sealer (9), closing the 'end' gap, and removing overlapping sequences, as previously done (10). Finally, the resulting assembly was polished using Pilon v1.22 (11).

The complete *P. engelmannii* Se404-851 chloroplast genome is 123,542 bp long, with 38.74% GC content. Using GeSeq (12) and other *Picea* chloroplast genomes as references

(6,10), we annotated 114 genes: 74 protein-coding, 36 tRNA-coding and 4 rRNA-coding. We note that four genes (*rps12*, *petB*, *petD*, and *rpl16*) in this list were manually annotated. We used OGDRAW v1.2 (13) to generate the map in Figure 1.

The introduction of this new chloroplast genome will benefit conifer genomic research and inform future evolutionary studies.

Accession number(s). The complete chloroplast genome sequence of *Picea engelmannii*, isolate Se404-851 is available under Genbank acession MK241981, and the raw reads in the SRA under SRX5070635. The annotation used as references were from *Picea abies* (NC_021456), *Picea asperata* (NC_032367), *Picea glauca* isolate PG29 (NC_028594), *Picea morrisonicola* (NC_016069), and *Picea sitchensis* (NC_011152, KU215903).

Acknowledgements

References

1. Sutton BCS, Flanagan DJ, Gawley JR, Newton CH, Lester DT, El-Kassaby YA. 1991. Inheritance of chloroplast and mitochondrial DNA in *Picea* and composition of hybrids from introgression zones. Theor Appl Genet 82: 242-248.

2. Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-Covo E, McInerney JO, Landan G, Martin WF. 2015. Endosymbioitic origin and differential loss of eukaryotic genes. Nature 524:427-432.

3. Jones MR, Schrader KA, Shen Y, Pleasance E, Chng C, Dar N, Yip S, Renouf DJ, Schein JE, Mungall AJ, Zhao Y, Moore R, Ma Y, Sheffield BS, Ng T, Jones SJM, Marra MA, Laskin J, Lim HJ. 2016. Response to angiotensin blockade with irbesartan in a patient with metastatic colorectal cancer. Ann Oncol 27:801–806.

4. Tsang ES, Shen Y, Chooback N, Ho C, Jones M, Renouf DJ, Lim HJ, Sun S, Yip S, Pleasance E, Ma Y, Zhao Y, Mungall AJ, Moore R, Jones S, Marra M, Laskin JJ. 2017. Clinical outcomes after whole genome sequencing in patients with metastatic non-small cell lung cancer. J Clin Oncol 35.

5. 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.

6. Jackman SD, Warren RL, Gibb EA, Vandervalk BP, Mohamadi H, Chu J, Raymond A, Pleasance S, Coope R, Wildung MR, Ritland CE, Bousquet J, Jones SJM, Bohlmann J, Birol I. 2015. Organellar Genomes of White Spruce (*Picea glauca*): Assembly and Annotation. Genome Biol Evol 8:29–41.

7. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34:i142–i150.

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

9. 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.

10. Coombe L, Warren RL, Jackman SD, Yang C, Vandervalk BP, Moore RA, Pleasance S, Coope RJ, Bohlmann J, Holt RA, Jones SJM, Birol I. 2016. Assembly of the Complete Sitka Spruce Chloroplast Genome Using 10X Genomics' GemCode Sequencing Data. PLoS One 11.

11. 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-e112963.

12. 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.

13. Lohse M, Drechsel O, Kahlau S, Bock R. 2013. OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41:W575-W581.

Figure 1: The complete chloroplast genome of *Picea engelmannii*, isolate Se404-851. The *Picea engelmannii* chloroplast genome was annotated using GeSeq (12) and plotted using OGDRAW (13). The inner grey circle illustrates the GC content of the genome.