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

TBD

January 23, 2019

Genome Announcement

The *P. engelmannii* sample was collected from Kalamalka Forestry Centre in British Columbia, Canada (36°17' 60" N 105°24' 0" W; elevation of 2987 m). The tissue sample used originated from the needles of a 13-year-old Engelmann spruce tree grown in BC, where its seed originated from Don Fernando Mountain, New Mexico (50°14' 38.4" N 119°16' 40.8" W; elevation of 450 m). These samples (BioSample: SAMN10388286; BioProject: PRJNA504036) were then sequenced at the British Columbia Cancer's Genome Sciences Centre, using the protocol outlined below.

A modified version of TruSeq DNA PCR-Free kit (E6875-6877B-GSC, New England Biolabs) genome protocol was used to generate a 900bp gap Illumina library on a Microlab NIMBUS liquid handling robot (Hamilton). 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 concentrated with PCRClean DX magnetic beads (Aline Biosciences) and fractionated in 2 lanes of a 6% PAGE gel to recover fragments greater than 700bp for library preparation. The isolated DNA fragments were end-repaired and bead-purified with a 1.8:1 ratio of beads, then A-tailed and ligated with full length indexed TruSeq adapters, and bead-purified. For quality check, an aliquot of the constructed library DNA was PCR amplified with Illumina universal primers to estimate the library gap size using the Agilent 2100 Bioanalyzer HSDNA assay, while the library concentration was determined using KAPA qPCR Library Quantification kit (KK4824). The PCR-Free library was sequenced with paired-end 150 base reads on the Illumina HiSeqX platform using V4 chemistry according to manufacturer recommendations.

Similar to the assembly method of the *P. glauca* isolate WS77111 chloroplast (Genbank MK174379), the reads were subsampled in the following sizes of read pairs (in millions): 0.75, 1.5, 3, 6, 12, 25, 46. Each subset of the reads was then assembled using ABySS v2.1.0, where the size of the kmer was set to 128 (-k 128), and the minimum k-mer count threshold for Bloom filter assembly was set to 3 (-kc 3), using a 10G Bloom filter. Then, each assembly was filtered for the chloroplast genome by alignment to the reference genome, the *P. glauca* isolate PG29 complete chloroplast genome (Genbank KT634228), where only contigs greater or equal to 500bp that aligned were kept, using BWA v0.7.17, aligning intraspecies contigs to the reference (-x intractg). In the 3M subset assembly, there was only one resulting contig, where the chloroplast had already been assembled in one piece. Using QUAST v5.0.2 and the reference chloroplast genome,, it was assessed that this contig contained zero misassemblies, and zero gaps, with a length of 123,601bp, approximately the size of the other *Picea* chloroplast genomes.

To ensure that the Se404-851 chloroplast assembly had the same start and end as the PG29 chloroplast, the first 500bp and the last 500 bp of the reference chloroplast genome were aligned to the Se404-851 chloroplast assembly using BLAST v2.7.1, where the pairwise alignments showed that the two chloroplasts were on opposite strands. Additionally, it also revealed that the start of the PG29 chloroplast, aligned with position 21830. Consequently, the Se404-851cp chloroplast was modified the strand was split into two segment A, position 1 to 21,829 and segment B, position 21,830 to 123,601. Although the assembly did not contain any internal gaps, potentially there could be missing sequences at both ends of the assembly. To resolve this, a 'fake' gap of 50 N's were introduced between the two segments, where the resulting strand is sequentially composed of segment B, 50 N's, and segment A. Sealer (part of ABySS), with a sweep of k-values of 70 to 200 in intervals of 10, using a 5G bloom filter closed the 'fake' gap at k=70, removing overlapping sequences. Finally, Pilon v1.22 was run to polish the assembly, altering two nucleotides, and a final QUAST analysis was conducted.

The Se404-851 chloroplast genome is 123,601bp in length, with a GC content of 38.74%. It has a total of 114 genes: 74 protein-coding genes, 36 tRNA-coding genes, 4 rRNA-coding genes. This chloroplast genome was annotated using GeSeq, with all available *Picea* NCBI RefSeq chloroplast genomes as reference genomes: *Picea abies* (NCBI NC_021456), *Picea asperata* (NCBI NC_032367), *Picea glauca* (NCBI NC_028594), *Picea morrisonicola* (NCBI NC_016069), and *Picea sitchensis* (NCBI NC_011152). GeSeq annotated most genes without issue, except for four genes that required manual annotation: *rps12*, *petB*, *petD*, and *rpl16.* The gene *rps12* is a transpliced gene, whereas *petB*, *petD*, and *rpl16* have very short initial exons of 6, 7, and 8 bp respectively. In the case of the short exons the position that GeSeq chose to annotate as the start of the gene, was in actuality the start of the second, larger exon.