Complete chloroplast genome sequence of a white spruce (*Picea glauca*) genotype from eastern Canada

TBD

January 23, 2019

Genome Announcement

The *P. glauca* isolate WS77111 sample was collected from Otonabee-South Monaghan (44°19' 48" N 78°9' 0" W; elevation of 250 m) in southern Ontario. The tissue sample collected was from the needles of the white spruce tree, supplied by John MacKay of Universite Laval (BioSample: SAMN02736786; BioProject: PRJNA242552). These tissue samples were then sequenced at the British Columbia Cancer's Genome Sciences Centre, using the protocol delineated below.

Genomic DNA libraries were constructed according to British Columbia Cancer's Genome Sciences Centre plate-based and paired-end library protocols on a Microlab NIMBUS liquid handling robot (Hamilton, USA). Briefly, 1 μg of high molecular weight genomic DNA was sonicated (Covaris LE220) in 62.5 μL volume to 400 bp. Sonicated DNA was purified with PCRClean DX magnetic beads (Aline Biosciences). The DNA fragments were end-repaired, phosphorylated and bead purified in preparation for A-tailing using a custom NEB Paired-End Sample Prep Premix Kit (New England Biolabs). Illumina sequencing adapters were ligated overnight at 16oC and adapter ligated products bead purified and enriched with 6 cycles of PCR using primers containing a hexamer index that enables library pooling. Pooled libraries were sequenced with paired-end 250 bp reads on an Illumina HiSeq2500 instrument in rapid mode.

To assemble the complete chloroplast genome, these reads were subsampled in the following sizes of read pairs (in millions): 0.75, 1.5, 3, 6, 12, 25, 50, 200. Each subset of read pairs was then assembled using ABySS v2.1.0, where the size of the k-mer 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 20G Bloom filter. Then, each assembly was filtered for the chloroplast genome by alignment to the reference chloroplast genome, the Picea glauca isolate PG29 complete chloroplast genome (Genbank KT634228), where only contigs greater or equal to 500 bp that aligned were kept, using BWA v0.7.17, aligning intraspecies contigs to the reference (-x intractg). The resulting contigs were pieces of WS77111 chloroplast, assessed using QUAST v5.0.2 and the reference chloroplast genome. Of all the subsets, the 1.5M subset had the least number of misassemblies, with its chloroplast genome in 14 pieces, while the 6M subset had the least number of pieces at 6, but contained 1 misassembly. For this reason, the 1.5M subset, 6M subset and the 3M subset (subset in between) were chosen to advance for a parameter sweep. For each chosen subset, ABySS was rerun using k values of 96, 112, 128, 144, and 160, as well as kc values of 3 and 4. Of these consequent assemblies, two were chosen to advance to the next step: 1.5M and 6M read pairs subset. LINKS v1.8.5 with a k-mer value of 26 (-k 26), sweep of distance between k-mer pairs (-d) from 500 to 2000 at intervals of 250 and from 2000 to 8500 at intervals of 500 revealed that the 1.5M read pairs subset with k-mer size of 96 (-k 96) and minimum k-mer count threshold for Bloom filter assembly of 3 (-kc 3) contained the fewest pieces of that subset of the parameter sweep, with no misassemblies. Similarly, the 6M read pairs subset with k-mer size of 144 (-k 144) and minimum k-mer count threshold for Bloom filter assembly of 3 (-kc 3) contained the fewest pieces of the other 2 subsets, with only 1 misassembly.

After LINKS, both assemblies had combined the chloroplast genome into one single piece (the largest resulting contig), and therefore re-evaluated using QUAST. This analysis showed that the 1.5M read pairs subset assembly contained 12 gaps, and the 6M one contained 5 gaps. Both assemblies were then run through Sealer (part of ABySS) to seal the gaps. After running Sealer, the assemblies were re-assessed using QUAST, revealing that Sealer closed 10 out of 12 gaps for the 1.5M read pairs subset assembly, but only sealed 1 of 5 gaps for the 6M read pairs subset assembly. Thus, the 1.5 M read pairs subset assembly was selected to advance to the next step, since it contained fewer gaps after this Sealer run than the 6M read pairs subset assembly. This Sealer run, the initial Sealer run, closed real gaps in the chloroplast assembly, but would be unable to detect missing end 'gaps'. In order to seal these gaps as well, a fake gap of 10 N's was introduced between the last 200 bp and the first 200 bp of the chloroplast assembly. This sequence was run through Sealer, and was successfully sealed, where 9 bases were removed from the start of the sequence, and 246 bases were added to the end. The chloroplast assembly was then modified to reflect these changes.

To ensure that the WS77111 chloroplast assembly had the same start and end as the PG29 chloroplast, the first 300bp and last 300bp of the assembly were aligned to the PG29 chloroplast using BLAST v2.7.1, where the pairwise alignments showed that the two chloroplasts were the same strand and already aligned with respect to start and end. Finally, Pilon v1.22 was run to polish the chloroplast genome and final QUAST analysis revealed a 150bp insertion in WS77111 chloroplast when compared to PG29 chloroplast.

The WS77111 chloroplast genome is 123,421bp 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 five genes that required manual annotation: rps12, petB, petD, rpl16, and psbZ. The gene rps12 is a transpliced gene, whereas petB, petD, and rpl16 have very short initial exons of 6, 7, and 8 bp respectively. The gene psbZ contained the very large insertion not found in PG29. 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.