CONIFER GENOMES AND IMPLICATIONS FOR PINE GENETICS AND IMPROVEMENT

Jill Wegrzyn[1], Sumaira Zaman[1,] Alyssa Ferreira[1], Madison Caballero[1], Ross Whetten[2]

[1]Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT; [2]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC

The emergence of improved sequencing technologies, coupled with decreasing costs, inspired innovative assembly methods for large and complex genomes, such as the conifer megagenomes, that range from 16 to 40 Gbp in size. Although these megagenomes are increasing in contiguity, accurate genome annotations remain challenging. Questions surrounding genome evolution are answered by interrogating the genome and its associated annotation. The accuracy of these products impacts estimates of genome duplication, gene family expansion/contraction, and functional assessments. Applications related to genomic selection, classification of hybrids, and pangenome approaches also require robust annotations. Among conifer genome assemblies, the gene space annotations are complicated by the presence of repetitive elements, large gene families, numerous pseudogenes, and long introns. Existing annotation packages are challenged to differentiate among these features and provide high quality results. Recent efforts have focused on improving strategies for the annotation of several gymnosperms, including five conifer species. We examine the impact of using assembled transcriptomic evidence (full length transcript and protein sequences) versus RNA read alignments to train *ab initio* gene predictors to annotate these genomes. These approaches are evaluated with assays for accessible chromatin, such as ATAC-Seq, which can improve the detection of true gene models, and distinguish prevalent pseudogenes. The final loblolly pine genome annotation improves on both the estimated completeness and structural metrics of the proposed gene models. A total of 51,200 genes were annotated with a novel pipeline integrating RNA-Seq and protein alignments with Braker2 and two in-house developed pieces of software, EnTAP and gFACs. The ATAC-Seq data assisted in filtering of the mono-exonic genes which are frequently inflated in conifer genomes. This approach was benchmarked against previous annotations as well as those resulting from standard standalone pipelines (MAKER and Braker2). The most recent release of the loblolly pine genome annotation can be retrieved from the TreeGenes database.