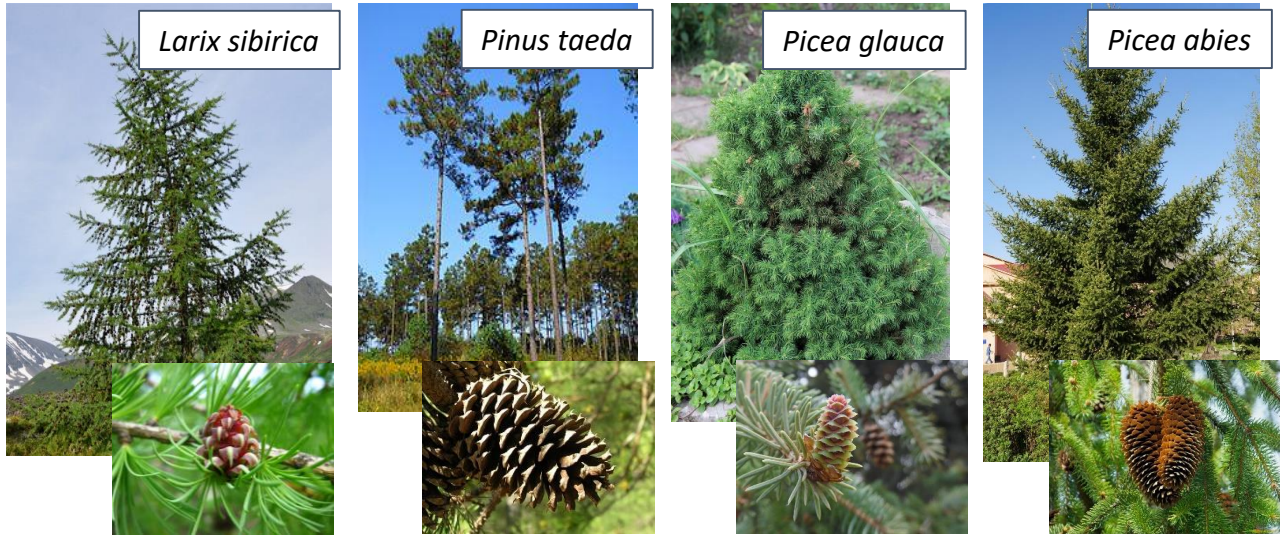
Genome-wide Prediction of Transcription Start Site in Four Conifer Species

E. Bondar, D. Kuzmin, K. Krutovsky, V. Sharov, T. Tatarinova

bondar.ev@ksc.krasn.ru

Introduction

The currently available annotations of several conifer mega-genomes although preliminary and limited, nevertheless provide opportunities for structural and functional analysis. Understanding of regulatory relationships between genome elements requires information on promoter sequences, which are usually located directly upstream or at the 5' end of the transcription start site (TSS). Here we present our attempts to improve the existing genome annotations by predicting TSSs and marking 5'-UTRs in the four relatively recently published species *Pinus taeda, Picea glauca, Picea abi*es, and *Larix sibirica*.



Methods

- All gene models retrieved from genomic annotations were aligned against the database of RNA-seq data, including ESTs and TSAs, of a corresponding species using hisat2.
- Prediction of putative TSSs was performed using TSSPlant program, in the promoter sequences of selected genes, which were defined as regions of -1000 and +250 bp around the start codon.
- The selection of the best prediction was based on the distribution of 5'-UTR lengths from the annotations of several model plants.
- Frequencies of CA and TATA motifs were calculated with a sliding window (width = 40 bp, increment step = 10 bp) using stringr package for R. CG-skew of a given sequence was defined as a proportion (C-G)/(C+G) and calculated with sliding window width of 50 bp and window increment step of 10 bp along the promoter sequence. GC3 was calculated using gene sequences with removed introns.

Laboratory of Forest Genomics, Siberian Federal University, Russia
Laboratory of Genomeic Research and Biotechnology, FRC KSC SB RAS, Russia

СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ | SIBERIAN FEDERAL UNIVERSITY

Federal Research Center
"Krasnoyarsk Science Center of the Siberian Branch of the Russian Academy of Sciences"

Genome-wide Prediction of Transcription Start Site in Four Conifer Species

E. Bondar, D. Kuzmin, K. Krutovsky, V. Sharov, T. Tatarinova

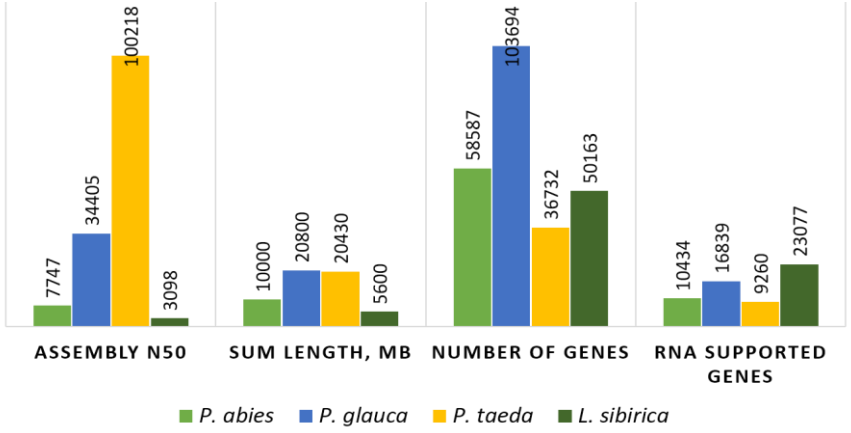
bondar.ev@ksc.krasn.ru

Results

We took currently available annotations for four conifer species and public RNA-seq and ESTs data, which allowed to retrieve 9260 evidence-supported gene models for *P. taeda*, 16853 for *P. glauca*, 7587 for *P. abies*, and 23077 for *L. sibirica*.

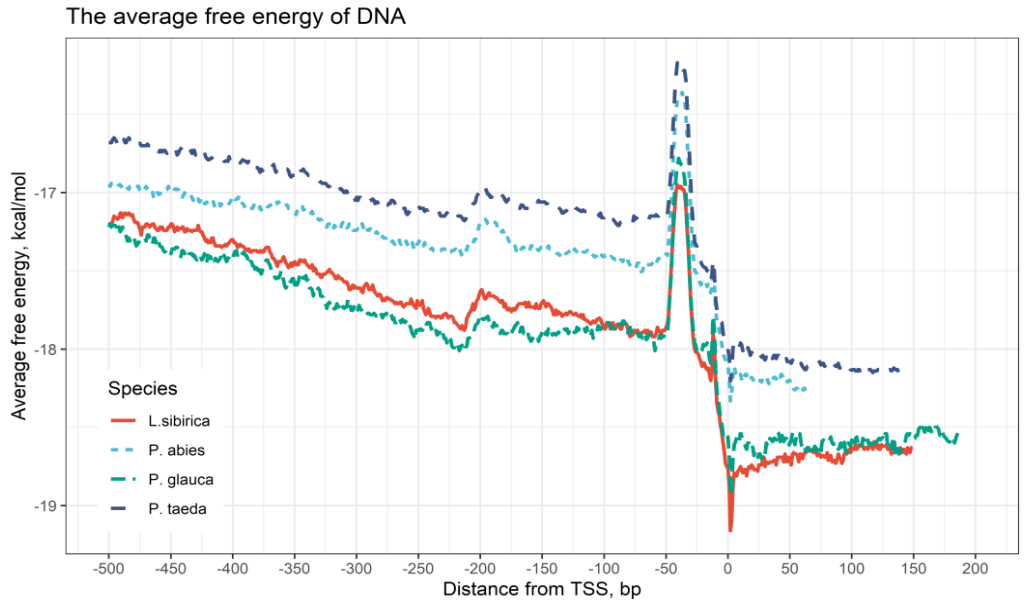
Summary of assemblies and annotations features

Assemly	Assembly N50	Sum length, Gb	Number of genes
P. abies v1.0	7,747	9.9	58,587
P. glauca v3	34,405	20.8	103,694
P. taeda v2.01	100,218	20.43	36,732
L. sibirica	3,098	5.6	50,163



Free energy of DNA

Change of standard free energy of a DNA duplex across genome sequence is considered to be a strong indicator of a promoter region and has been implemented successfully for promoter prediction. We used this as supporting evidence for promoters predicted by TSSPlant. The free energy profile shows a peak around -40 bp and a sharp decline around putative TSS.



Laboratory of Forest Genomics, Siberian Federal University, Russia
Laboratory of Genomeic Research and Biotechnology, FRC KSC SB RAS, Russia

Federal Research Center
"Krasnoyarsk Science Center of the Siberian Branch of the Russian Academy of Sciences"

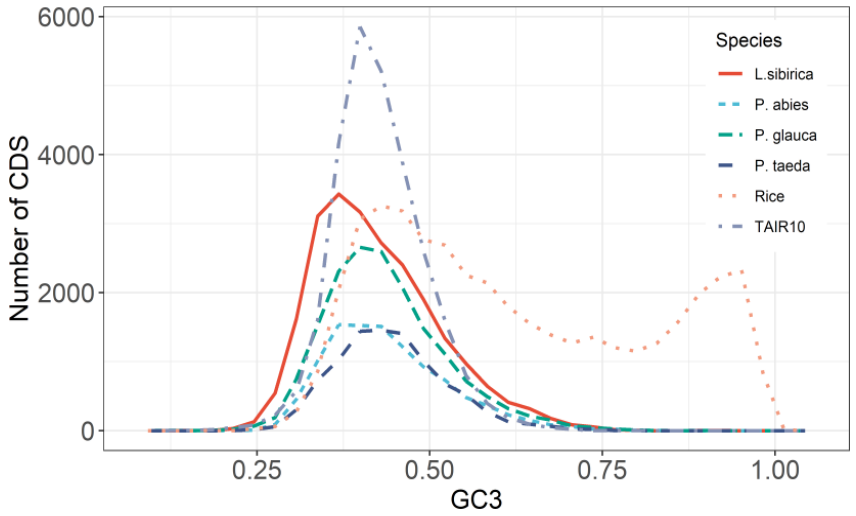
Genome-wide Prediction of Transcription Start Site in Four Conifer Species

E. Bondar, D. Kuzmin, K. Krutovsky, V. Sharov, T. Tatarinova

bondar.ev@ksc.krasn.ru

GC3 distribution

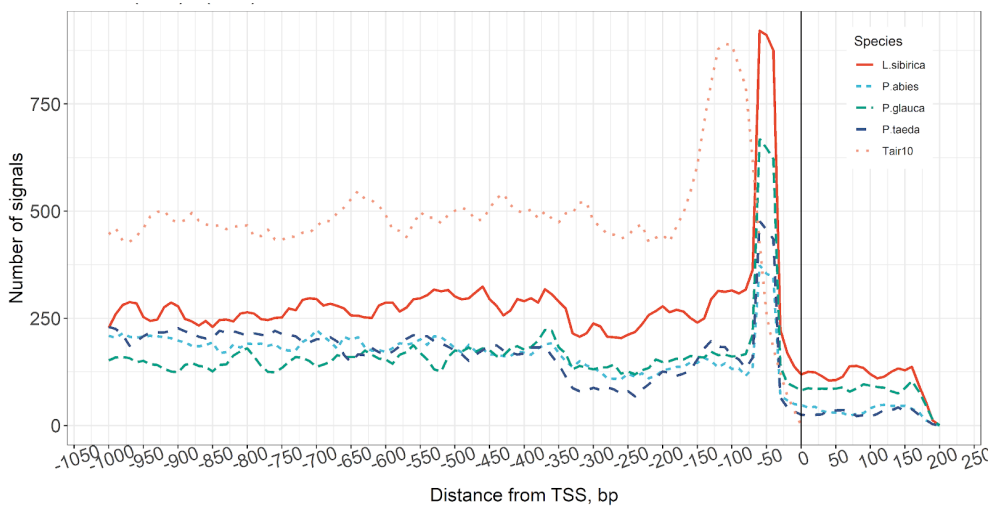
Due to the degeneracy of a genetic code, the third position in a codon is referred to as a wobble. It was observed that based on the frequency of G and C nucleotides at the third position (GC3) organisms can be classified into those having unimodal and bimodal GC3 distribution. We estimated GC3 over all CDSs retrieved from current annotations that had RNA-support. It appears that conifers, similar to dicot plants, possess unimodal GC3 distribution, with an average of 0.43 (sd= 0.087).



A – GC3 distribution across all coding sequences, B – GC3 gradient of coding sequences.

Position of TATA-box

Frequency of TATA(A/T)A(A/T) motif in the predicted TSS-centered promoter regions showed pronounced peak around -60 bp form TSS, which supports the predictions. Although the canonical location of TATA-box is considered to be about 40 bp upstream, it has been previously reported that in some plants, such as *Vinis vinifera*, TATA-box was observed within -70 bp relative to TSS.



Frequency of TATA(A/T)A(A/T) motif in the TSS-centered promoter region

Acknowledgments

We thank Dr. N. V. Oreshkova and Dr. S. I. Feranchuk for help with sequencing and draft annotation of the Larix sibirica genome, respectively.

Laboratory of Forest Genomics, Siberian Federal University, Russia
Laboratory of Genomeic Research and Biotechnology, FRC KSC SB RAS, Russia

Federal Research Center
"Krasnoyarsk Science Center of the Siberian Branch of the Russian Academy of Sciences"