International Journal of
Molecular Sciences

MDPI

Article

Genome-Wide Prediction of Transcription Start Sites in Conifers

Eugeniya I. Bondar [1,2], **Maxim E. Troukhan** [3], **Konstantin V. Krutovsky** [1,4,5,6,7,8,*] and **Tatiana V. Tatarinova** [8,9,10,11,*]

[1] Laboratory of Forest Genomics, Institute of Fundamental Biology and Biotechnology, Siberian Federal University, 660036 Krasnoyarsk, Russia; ebondar@sfu-kras.ru
[2] Laboratory of Genomic Research and Biotechnology, Federal Research Center "Krasnoyarsk Science Center" Siberian Branch, Russian Academy of Sciences, 660036 Krasnoyarsk, Russia
[3] Persephone Software LLC, Agoura Hills, CA 91301, USA; mtroukhan@persephonesoft.com
[4] Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, 37077 Göttingen, Germany
[5] Center for Integrated Breeding Research, Georg-August University of Göttingen, 37075 Göttingen, Germany
[6] Laboratory of Population Genetics, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, 119333 Moscow, Russia
[7] Scientific and Methodological Center, G. F. Morozov Voronezh State University of Forestry and Technologies, 394087 Voronezh, Russia
[8] Department of Genomics and Bioinformatics, Institute of Fundamental Biology and Biotechnology, Siberian Federal University, 660074 Krasnoyarsk, Russia
[9] Department of Biology, University of La Verne, La Verne, CA 91750, USA
[10] Functional Genomics Group, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, 119333 Moscow, Russia
[11] A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, 127051 Moscow, Russia
* Correspondence: konstantin.krutovsky@forst.uni-goettingen.de (K.V.K.); ttatarinova@laverne.edu (T.V.T.)

Abstract: The identification of promoters is an essential step in the genome annotation process, providing a framework for gene regulatory networks and their role in transcription regulation. Despite considerable advances in the high-throughput determination of transcription start sites (TSSs) and transcription factor binding sites (TFBSs), experimental methods are still time-consuming and expensive. Instead, several computational approaches have been developed to provide fast and reliable means for predicting the location of TSSs and regulatory motifs on a genome-wide scale. Numerous studies have been carried out on the regulatory elements of mammalian genomes, but plant promoters, especially in gymnosperms, have been left out of the limelight and, therefore, have been poorly investigated. The aim of this study was to enhance and expand the existing genome annotations using computational approaches for genome-wide prediction of TSSs in the four conifer species: loblolly pine, white spruce, Norway spruce, and Siberian larch. Our pipeline will be useful for TSS predictions in other genomes, especially for draft assemblies, where reliable TSS predictions are not usually available. We also explored some of the features of the nucleotide composition of the predicted promoters and compared the GC properties of conifer genes with model monocot and dicot plants. Here, we demonstrate that even incomplete genome assemblies and partial annotations can be a reliable starting point for TSS annotation. The results of the TSS prediction in four conifer species have been deposited in the Persephone genome browser, which allows smooth visualization and is optimized for large data sets. This work provides the initial basis for future experimental validation and the study of the regulatory regions to understand gene regulation in gymnosperms.

Keywords: transcription start site; transcription factor binding site; TATA-box; conifer; gymnosperms; promoter prediction

check for updates

Citation: Bondar, E.I.; Troukhan, M.E.; Krutovsky, K.V.; Tatarinova, T.V. Genome-Wide Prediction of Transcription Start Sites in Conifers. *Int. J. Mol. Sci.* **2022**, *23*, 1735. https://doi.org/10.3390/ijms23031735

Academic Editor: Pablo Minguez

Received: 30 December 2021
Accepted: 1 February 2022
Published: 3 February 2022

1. Introduction

Transcription is a mechanism of information transmission encoded in protein-coding genes conducted by RNA Polymerase II, resulting in the production of messenger RNAs (mRNAs). This process is subject to complex regulation via binding of transcription factors (TFs) to appropriate genomic sites consisting of regulatory nucleotide motifs typically located within the 1000 bp region upstream of the transcription start sites (TSSs). This region is called the promoter. The TSS position corresponds to the first nucleotide transcribed by RNA Pol II. Eukaryotic genes can have multiple alternative TSSs [1,2].

The core promoter is a stretch of DNA up to 250 bp long located immediately upstream of the TSS and required for transcription initiation. There are two types of transcription initiation: narrow, generally associated with the regulation of tissue-specific and stress-response genes, and broad, typically occurring in housekeeping genes under a constitutive expression pattern [3]. The corresponding area is called a transcription start region (TSR) when the transcription initiation region is broad. The best-known regulatory motif in core promoter regions is the TATA-box, a recognition site for the TATA-binding protein (TBP). This motif has a highly conserved consensus sequence TATA(A/T)A(A/T) found in 5–60% of all RNA Pol II promoters [1,4–8]. Another common motif is Inr (initiator) with the consensus sequence YYA+1NT/AYY, which occurs at the start of transcription. Inr is more widespread than any other sequence motif [3] and is commonly found in housekeeping genes, whose transcription is initiated not with a single start but with positional clusters of TSSs referred to as TSRs [9]. In contrast, TATA-containing promoters are narrower and associated with tissue- or context-specific gene expression [10]. Other common core promoter elements are the TFIIB recognition element (BREu, consensus G/CG/CG/ACGCC, and BREd, consensus G/ATT/AT/GT/GT/GT/G [11,12]), the downstream promoter element (DPE, consensus RGWYV [13,14]), and the downstream core element (DCE, consensus CTTC, CTGT, AGC [15]).

The identification of promoters is a crucial step in genome annotation, providing a framework for understanding gene regulatory networks and their role in transcription regulation [8]. In recent years, high-throughput methods for identifying TSS and TFBS have advanced considerably. Such techniques as chromatin immunoprecipitation combined with microarray or sequencing analysis (ChIP-chip and ChIP-seq, respectively), identification of DNase I-hypersensitive sites (DHS), cap analysis of gene expression (CAGE), and paired-end analysis of TSS (PEAT) have allowed the accumulation of a substantial amount of data on plant regulatory regions [16,17]. TF studies in agriculturally important species, such as the *Prunus* genus, have utilized numerous resources and techniques, including gene expression analysis of different agronomic traits, quantitative RT-PCR, cDNA-AFLP, LC-ESI-MS, RNA, and DNA blotting, to build a database of genus-specific TFs and to provide a comprehensive source for further functional studies and breeding programs [18]. However, those experimental methods of identifying the TSS and promoter regions are time-consuming, labor-intensive, and expensive. Several computational approaches have been developed to provide fast and accurate ways to predict the location of TSSs and regulatory motifs on the whole-genome scale. These include Bayesian classification based on positional densities of oligonucleotides for detecting TSS in human genomic sequences [19], neural networks for predicting TSS in plant promoters [20,21], and conditional random fields for identifying TSS in eukaryotic promoters [22]. Strategies for genome-wide discovery of novel cis-regulatory motifs using position weight matrices (PWMs) and expression data were successfully implemented for rice and *Arabidopsis* [6,23], hop [24], and grapevine [25]. Genome-wide analysis of core promoter elements using PWMs and orthologous-based prediction were performed for several monocot and dicot species by Kumari and Ware [26].

It has been shown that promoters differ from the rest of the genome in several measurable properties: low DNA stability, high bendability, curvature, etc. [7,27–29]. DNA stability is associated with the melting of the double-stranded molecule before transcription initiation and is commonly calculated as the standard free energy of a DNA duplex. It has been implemented successfully for promoter identification in various eukaryote species [30]. Numerous studies have reported a curved DNA region upstream of TSSs and higher bendability of the area that interacts with DNA binding proteins [28,31–33]. Other features of promoter regions include CpG islands, GC-skew, and decreased genetic variability [7,29,34]. The excess of Cs over Gs (GC skew) in the sense strand around TSSs was reported for several plant species [35,36] and metazoans [37]. The peak of the GC skew around TSSs can be explained by cytosine deamination during the transcription due to RNA polymerase preferential protection of nucleotides on the non-transcribed strand [35].

It may seem surprising, but the properties of coding and promoter regions are correlated. The frequency of guanine and cytosine nucleotides at the third position GC_3 is one of the critical properties of coding regions. Nucleotides at the third position are less subjected to selection than at the first two due to the degeneracy of a genetic code. It was observed that based on GC_3, the genomes could be classified into those having unimodal and bimodal GC_3 distributions. For instance, all currently sequenced grass genomes have a bimodal distribution of GC_3, while the CDS of dicot plants shows a unimodal distribution of GC_3 [38,39]. It was previously thought that bimodal GC_3 is a specific feature of grass genomes. Later, it was demonstrated that GC content in other monocot species, such as *Curcuma longa, Zingiber officinale, Elaeis guineensis*, and *Zantedeschia aethiopica*, also exhibits a bimodal GC_3 distribution [40–42]. It has been shown that genes with a higher GC_3 content also have a higher frequency of TATA-boxes and are more likely to be stress-related [39].

Conifers are an ancient group of dicot plants represented by more than 600 species that play a significant role in boreal forest ecosystems. Due to their enormous size and highly repetitive nature, deciphering conifer genomes takes more time and effort than many other plant species. Several mega-genomes of conifer species have been sequenced and assembled to the draft state recently, resulting in multiple contigs and gaps in chromosome coverage. Although such annotations are preliminary, they provide an opportunity for structural and functional analysis. Even an incomplete genome contains keys to understanding regulatory relationships between genome elements, and this analysis requires knowledge of the precise locations of promoter sequences.

The aim of this study was to enhance and expand the existing genome annotations by predicting TSSs for the four recently published conifer species: loblolly pine (*Pinus taeda* L.), white spruce (*Picea glauca* (Moench) Voss), Norway spruce (*Picea abies* (L.) H. Karst.), and Siberian larch (*Larix sibirica* Ledeb.). Siberian larch is a cold-resistant fast-growing tree known for its rot-resistant timber, making it especially valuable in construction. Norway spruce had the first genome sequenced among gymnosperms. It is widely cultivated as an ornamental tree worldwide and is a source of timber for paper and construction lumber production. White spruce is another cold-resistant tree native to northern parts of North America with tremendous economic value in Canada. Loblolly pine has one of the largest genome sizes (22 Gbp) among sequenced plant species and is considered one of the most significant sources of timber in the U.S.

2. Results

2.1. Prediction of TSS in Four Conifer Species

Alignment of RNA-seq, ESTs, and RefSeq protein to the four genomes data allowed the identification of 9260 evidence-supported gene models for *Pinus taeda*, 16,853 for *Picea glauca*, 7587 for *Picea abies*, and 23,077 for *Larix sibirica* (Table 1, Supplementary Figure S1). For promoter prediction, we used TSSPlant [20], which utilizes neural networks to estimate up to 17 features, such as the presence of classic motifs, nucleotide composition variation, and others (more details in Methods or [20]). The use of TSSPlant generated predictions of 22,633 TSS positions in *P. taeda*, 25,889 in *P. abies*, 44,651 in *P. glauca*, and 62,420 in *L. sibirica*. From 13.3% to 14.3% of identified TSS positions occurred within the coding parts of their respective gene models (Figure S2) and were excluded.

Table 1. Summary of genome assemblies and annotations for four conifer species.

Assembly and Annotations Parameters	***Picea abies* [43]**	***Picea glauca* [44]**	***Pinus taeda* [45]**	***Larix sibirica* [46]**
Estimated genome size, Gbp	19.57	15.79	20.15	12.03
Assembly length, Gbp	12.30	25.47	22.10	12.34
Scaffold N50, Kbp	4.869	34.405	107.821	6.443
GC content, %	38.81	37.08	38.06	35.41
Repeat content, %	70.0		81.8	65–80 *
Total predicted gene models	58,587	103,694	36,732	50,163 *
Average CDS length, bp	287.21	283.56	419.81	291.01 *
Average intron length, bp	997.94	642.73	1146.12	351.13 *
Maximum intron length, bp	68,268	44,113	568,968	10,152 *
RNA/RefSeq supported genes *	10,434	16,839	9260	23,077
Predicted TSS positions *	25,889	44,651	22,633	62,420
Unique TSSs filtered by 5′-UTR distribution *	10,367	16,629	9149	23,016

* authors' data.

To select the most likely TSS among the multiple predictions for a given gene, we compared the length of each 5′ UTR to the distribution of 5′ UTR lengths in four plant species, two dicots, *A. thaliana* and *P. trichocarpa*, and two monocots, *O. sativa* and *S. bicolor* (Figure S3A,B). Two parameters, *k* and *theta*, which determine the shape and scale of the gamma distribution, were computed as follows: *theta* = *v*/*m*, *k* = *m*/*theta*. Using *k* = 0.62 and *theta* = 238.99, we selected predictions that better fit the theoretical 5′ UTR length distribution (Figure S3C). After filtering out predictions within the respective coding regions and selecting the highest-scoring positions, 10,367 *P. abies*, 16,629 for *P. glauca*, 9149 for *P. taeda*, and 23,016 for *L. sibirica* were identified as putative TSSs (Table 1). All gene models with corresponding predicted TSSs were deposited in the Persephone genome browser and are available at https://web.persephonesoft.com (accessed on 31 January 2022).

Genome annotations of the Siberian larch and white spruce were performed with the MAKER [47] pipeline using transcriptome data and available ESTs and RNA-seq data for these and other closely related species. Therefore, it became possible to conduct an automated prediction of 13,228 UTRs for *L. sibirica* and 14,056 UTRs for *P. glauca* based on the available ESTs within the annotation pipeline. We compared the TSSs predicted by the maker pipeline and TSSs predicted by the TSSPlant algorithm with filtering based on the 5′ UTR length distribution of model plant species. We showed that the positional distribution of TSSs predicted by the de novo method was similar to that predicted using RNA support (Figure S4).

In the predicted promoters, the occurrence of the TATA(A/T)A(A/T) motif shows a pronounced peak approximately 20 bp upstream of the predicted TSS position for all four species (Figure 1), which corresponds well to the canonical location of the TATA-box, since 30 to 50% of eukaryotic promoters contain a TATA-box 40 to 15 bp upstream of the TSS.

When comparing the number of promoters containing TATA-box or CA initiator motif, approximately half of the analyzed sequences (46–53%) had CA motif within the area [−2; +2] around the TSS, while TATA-box was found in 5–8% of promoters in the [–40; –20] region relative to the TSS. Among TATA-containing promoters, approximately half of them (1.6–2 ratio of TATA-containing to TATA-and-CA-containing) contained both TATA and CA motifs (see more in Table S3 and Figure S5).

Change in the standard free energy of a DNA duplex across the genome sequence is 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 (Figure 2).

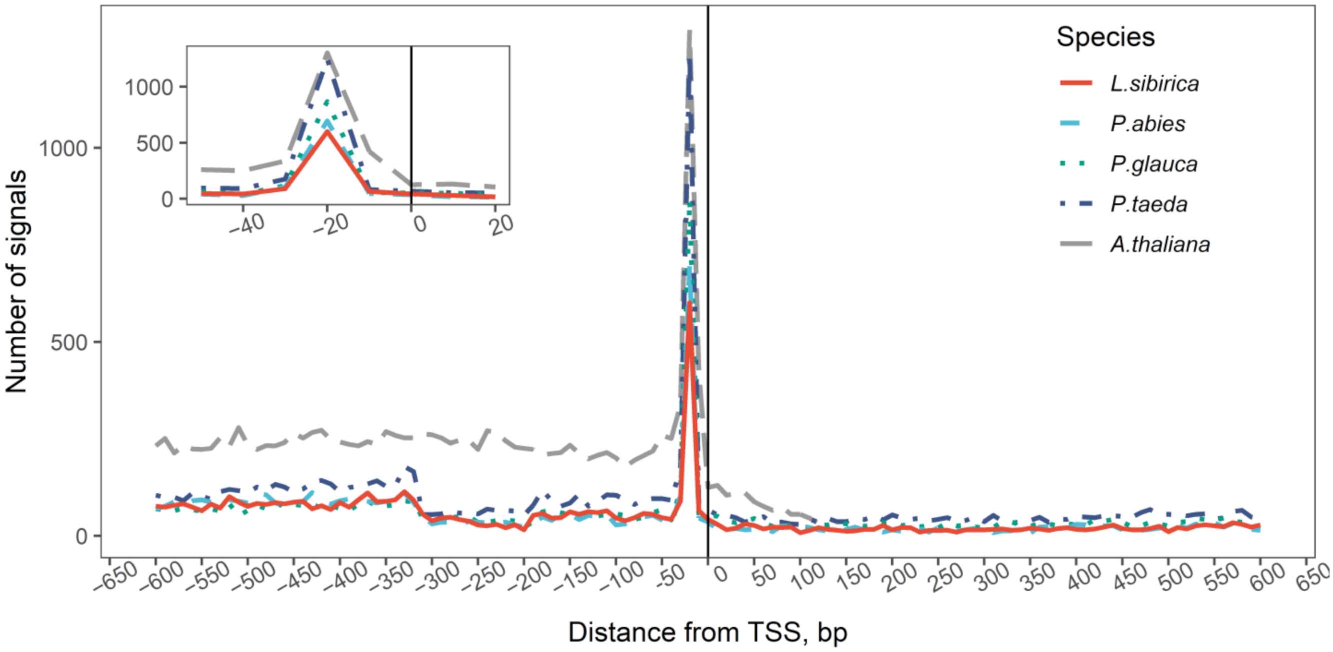

Figure 1. Frequency of the TATA(A/T)A(A/T) motif in the TSS-centered promoter region.

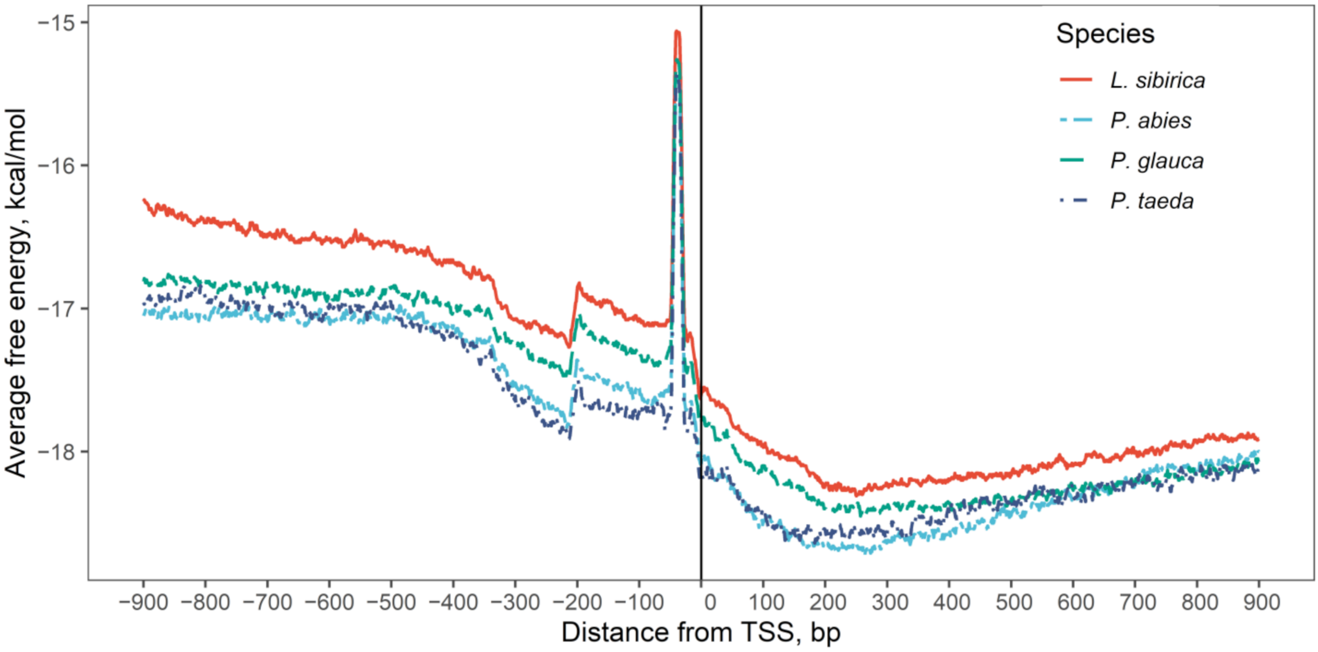

Figure 2. Distribution of DNA free energy around TSS position predicted by TSSPlant.

To determine the positional distribution of TFBS, we scanned the identified promoter regions for the presence of several developmental and stress-related TFBS using TRANSF AC and MATCH. PWMs that belong to the Homeodomain, Heat shock, and Myb TFs show two peaks in their positional distribution (Figure 3b–d), while AP2/EREBP TFs have an apparent decrease near the TATA-box region (Figure 3a).

Additionally, the promoters of two developmental genes, FLORICAULA/LEAFY and WLIM2a, were scanned for the presence of conserved sequence motifs. For orthologs of LEAFY, a helix-turn-helix transcription factor regulating inflorescence development in many flowering plant species [48], there are two predicted TSS positions (Figure 4) in the upstream regions of *P. abies* and *P. glauca*, and one prediction for *P. taeda* and *L. sibirica*.

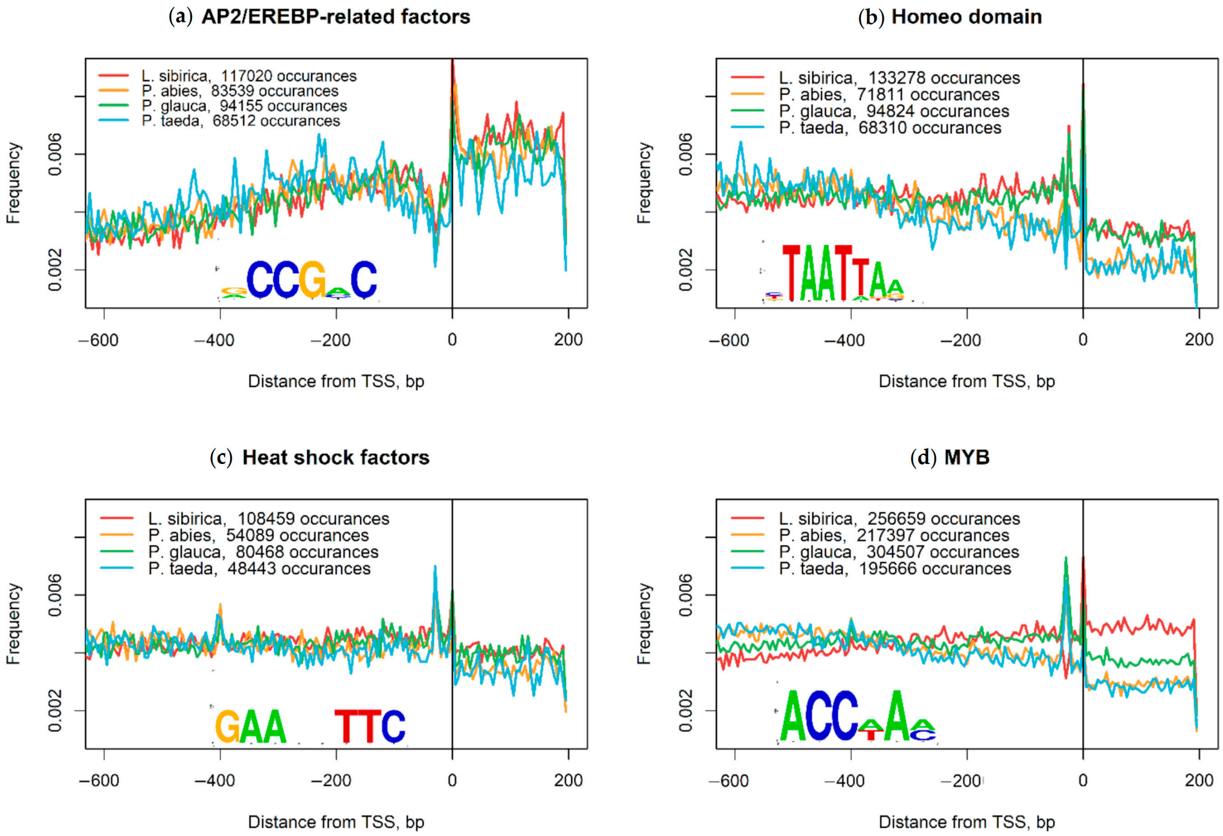

Figure 3. Positional distribution of transcription factor binding sites (TFBS) in *Larix sibirica*, *Picea abies*, *Picea glauca*, and *Pinus taeda* based on PWM scanning using TRANSFAC. (**a**) AP2/EREBP-related factors; (**b**) Homeodomain; (**c**) Heat shock transcription factors; (**d**) Myb transcription factors.

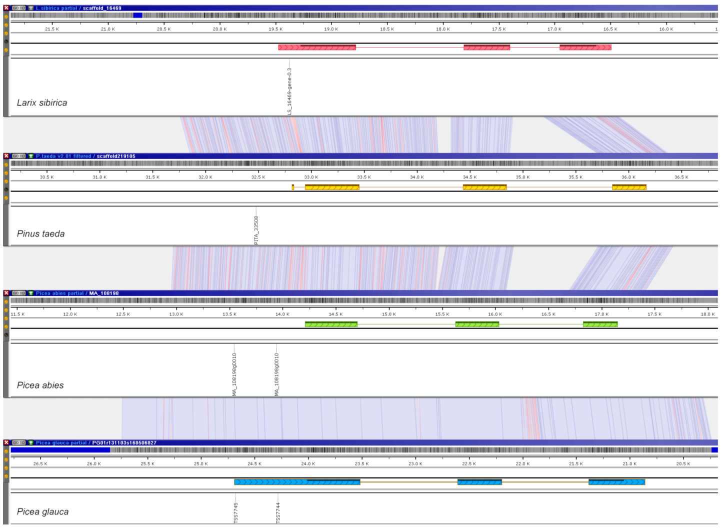

Figure 4. Orthologous genes of FLORICAULA/LEAFY-like proteins in *L. sibirica*, *P. taeda*, *P. abies*, and *P. glauca* with corresponding predicted TSS positions (depicted by the vertically-oriented labels) in their upstream regions are aligned using the genome browser Persephone. Red, yellow, green, and blue boxes represent exons. Light blue ribbon-like connectors indicate identical areas, blue lines mark nucleotide substitutions, and red lines indicate indels. The visualization is available at https://web.persephonesoft.com/?bookmark=43C6DEFD15C23F5F40A8AFF25F844042 (accessed on 31 January 2022).

For LIM domain-containing WLIM2a, a regulatory protein that triggers the formation of actin bundles playing an essential role in actin cytoskeleton organization [49], three orthologs were found in *L. sibirica*, *P. abies*, and *P. glauca*, each of which had two predicted TSS sites (Figure 5).

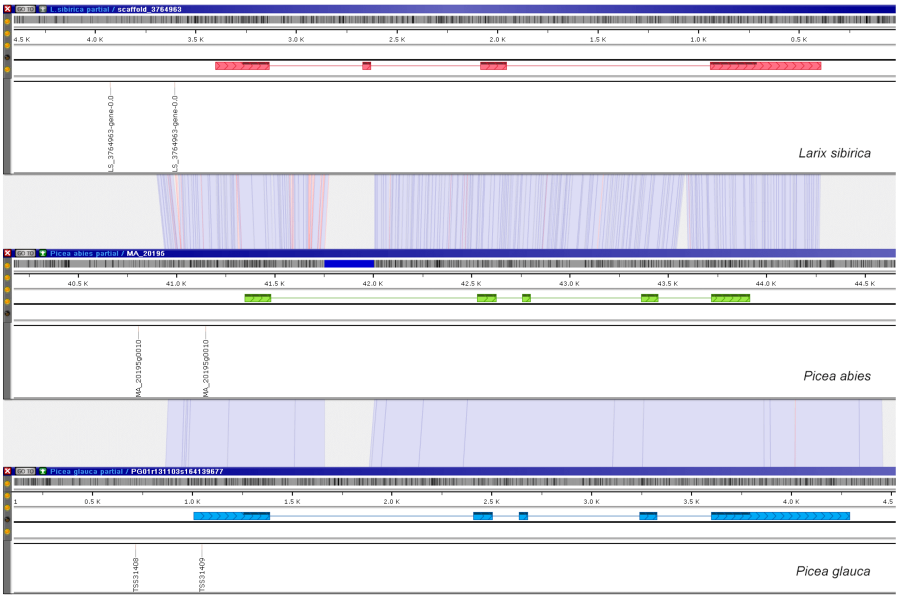

Figure 5. Orthologous genes of WLIM2a in *L. sibirica*, *P. abies*, and *P. glauca* with corresponding predicted TSS positions (depicted by the vertically-oriented labels) in their upstream regions. Red, green, and blue boxes represent exons. Light blue ribbon-like connectors indicate identical areas, blue lines mark nucleotide substitutions, and red lines indicate indels. The visualization is available at https://web.persephonesoft.com/?bookmark=4239E3155493E8E21C61A9932BD502EE (accessed on 31 January 2022).

2.2. Nucleotide Composition Analysis of Promoter and Coding Sequences

We computed GC3 for all coding regions retrieved from current annotations and had RNA-SEQ support. Similar to other dicot plants, conifers possess a unimodal GC_3 distribution, with a mean of 0.43 (sd = 0.087, Figure 6c). Analyzing coding sequences in several plant species has indicated a GC_3 gradient from the 5′ to 3′ end of a gene [39,50]. All four analyzed species had a similar GC_3 gradient that gradually decreased 250 bp after the start of the coding sequence (Figure 6a). We divided the genes into GC_3-poor and GC_3-rich categories using 10% and 90% quantiles of GC_3 to define the GC_3-rich and GC_3-poor gene sets. We determined the relationship between the position of the codon in the coding sequence and the GC_3 content for both GC_3 categories, applying linear regression to the first 1000 nucleotides of the coding sequence (Figure 6b). In Siberian larch and rice, the slope of the regression line is more prominent than in loblolly pine and thale cress. These results agree with a previous report on GC distribution patterns in gymnosperms [41].

Similar to *A. thaliana* and *O. sativa* [35,36], the CG-skew in the four examined conifer species exhibited a distinct peak around the TSS (Figure 6d). The height of the peak in the four conifer species is lower than that in the thale cress. It can be due to biological differences or the lower quality of genome assembly.

To test whether the difference in gene length between GC_3-poor and GC_3-rich genes can be observed in gymnosperm genomes, we compared these two classes of genes (Figure 7 and Figure S6). A Mann–Whitney U test indicated that the CDS length in GC_3-poor genes was significantly longer than that in GC_3-rich genes ($2.20 \times 10^{-16} < p < 6.09 \times 10^{-12}$, see Table S2).

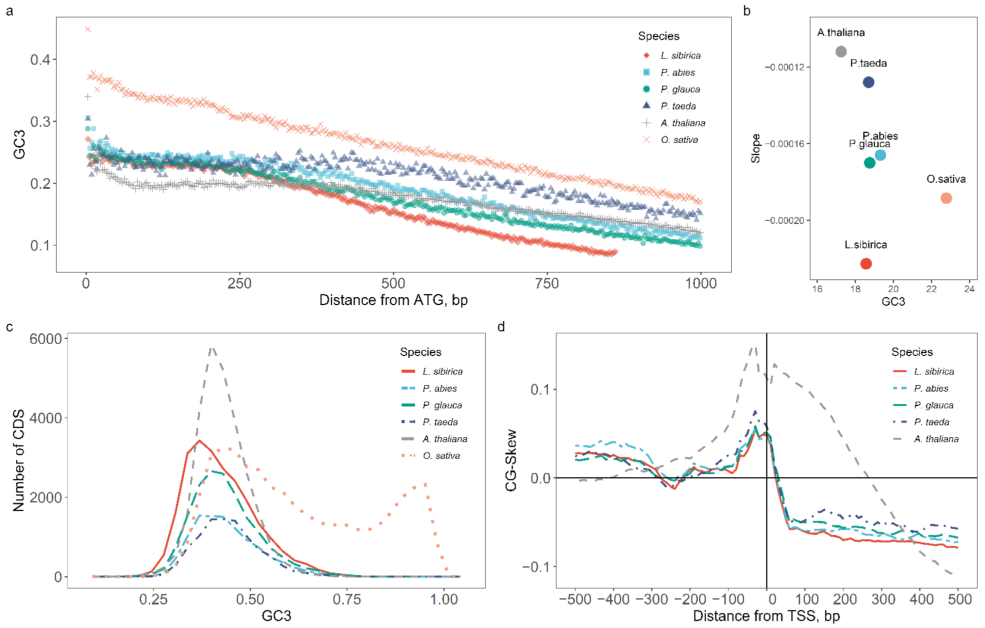

Figure 6. Some GC statistics for four conifer species, *Larix sibirica*, *Picea abies*, *Picea glauca*, *Pinus taeda*, and two model plant species, *Arabidopsis thaliana* and *Oryza sativa*: (**a**) GC_3 gradient of coding sequences, (**b**) GC_3 gradient slope, (**c**) GC_3 distribution across all CDSs, (**d**) CG-skew around TSSs.

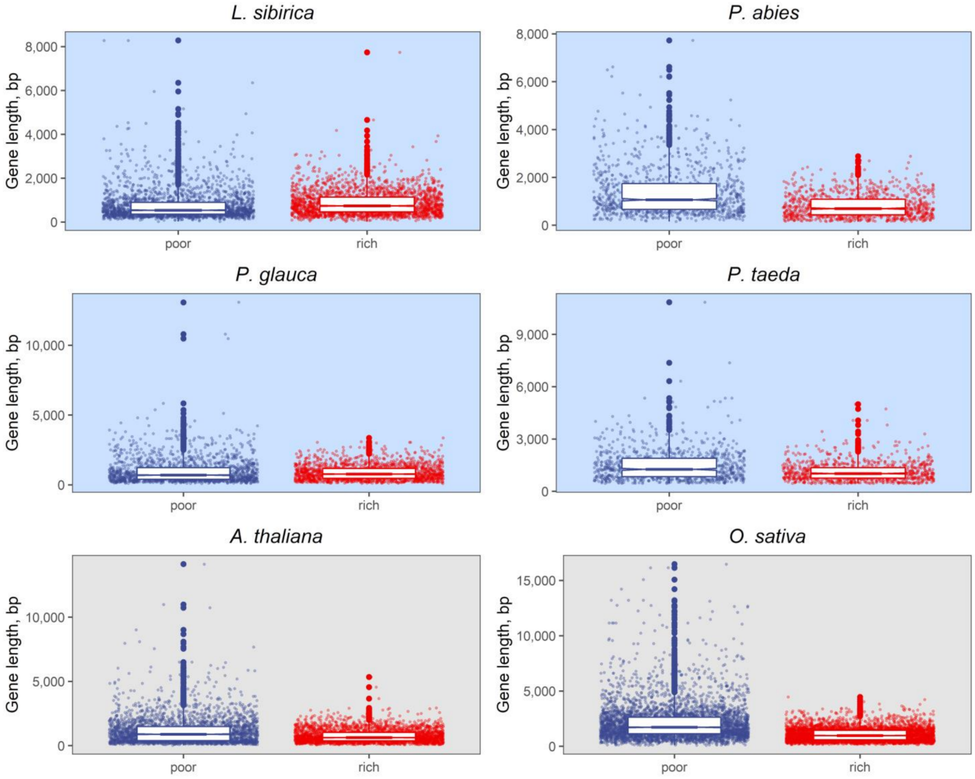

Figure 7. The difference in coding sequence length between GC_3-poor and GC_3-rich genes; 10% and 90% quantiles were used to divide genes into GC_3-poor and GC_3-rich classes (blue and red, respectively).

There was a significant difference in the number of exons between the two classes of genes; this trend holds for all studied gymnosperms and angiosperms. According to the current genome annotations, GC-rich genes tend to have between two and four exons in Siberian larch and loblolly pine and two or fewer exons in Norway and white spruces (Figure 8). Genes with more than five exons are generally GC-poor in all species.

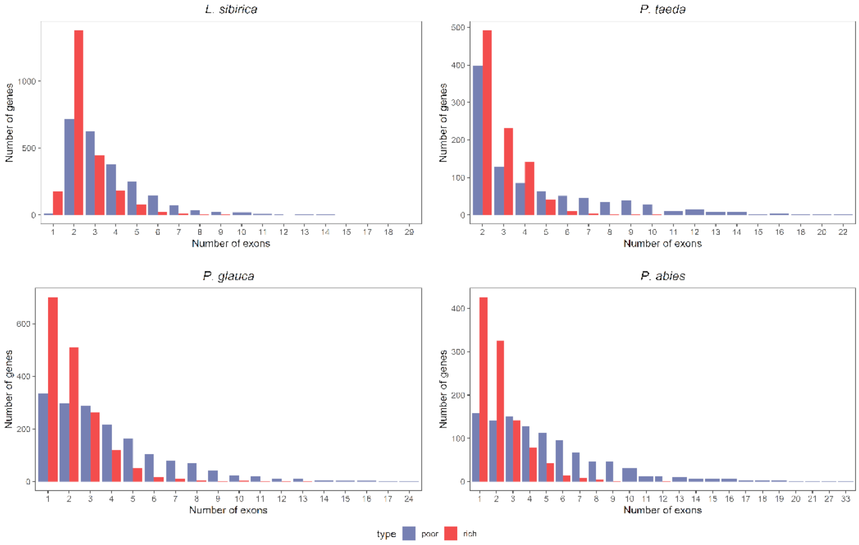

Figure 8. Distribution of the exon number per gene in GC_3-poor and GC_3-rich genes in *L. sibirica*, *P. abies*, *P. glauca*, and *P. taeda*. The number of genes in the GC3-poor and GC3-rich categories was the same within each organism.

3. Discussion

While extensive studies on mammalian regulatory regions have been conducted, plant promoters were left out of the limelight and have remained poorly investigated. Promoter prediction programs have been mainly developed for and trained on a limited range of model organisms such as humans, *Drosophila*, thale cress, or rice. Nevertheless, the existing annotations for sequenced conifer genomes allow for computational prediction of biologically relevant elements, such as TSS and TFBS, and meaningful comparative analysis.

3.1. Prediction of TSS

Plants rely on TATA-box to initiate the transcription of most genes. The TATA-box is located approximately 40 bp upstream of the TSSs in the conifers. Although the most common location for the TATA-box is from 20 to 40 bp upstream of the TSSs, it has been previously reported that in some plants, such as *Vinis vinifera*, the TATA-box was observed within −70 bp relative to the TSSs [26]. The height of the peak of TATA frequency (Figure 1) directly measures the accuracy of the TSS prediction.

In addition to the TATA-box, a standard plant core promoter model [51] includes the initiator motif at the TSSs, AGGA-box (YA2-5KNGA2-4YY, ~80 bp upstream of the TSSs, [51,52]), and the downstream promoter element, DPE (RGWYVT, ~28–32 bp downstream from the TSSs, [51]). There are crucial differences in the promoter organization of plants and animals. Mammalian promoters commonly have a CAAT-box [53] instead of the plant AGGA. The TATA-box appears in less than 10% of mammalian promoters. The DPE motif typically occurs in TATA-less promoters; however, TATA and DPE may co-occur [51]. Animal promoters may also have BRE and motif ten (MTE) elements not found in plant promoters [53]. It has been reported that genes may have bidirectional promoters [54,55]; this feature has been thoroughly studied in mammals and less extensively in plants. Morton et al. [17] analyzed *A. thaliana* root samples to find precise TSS locations. They found that most promoters were TATA-less but contained a specific combination of transcription factor binding sites regulating gene expression. Yamamoto et al. [56] studied core promoters in *A. thaliana*, such as TATA and GA. They observed that the promoter architecture was related to the gene structure. The length of the 5′ UTR (a distance from the TSS to the start of translation) is also negatively correlated with the expression level of respective genes [39].

One of the standard computational strategies in predicting transcriptional targets in promoter regions is to search for a consensus model of a TF binding site (TFBS). Such consensus sequences are derived from experimentally identified binding sites and stored as a position weight matrix (PWM), which can be used to scan a sequence of interest. Several curated databases provide collections of PWMs, such as JASPAR [57], PlantRegMap [58], or TRANSFAC [59]. Despite the continuous improvement in PWM matching methods, they yield many false-positive predictions. Short length and high variability of the actual TF binding motifs cause the majority of matrix matches to miss functional binding sites (a so-called "futility theorem" [60]). However, PWM scanning can serve as a useful preliminary step in generating a list of candidate TFBSs that can be further filtered using other methods. Motif conservation across homologs, identification of overrepresented motifs in co-expressed genes, or clustering of multiple closely positioned motifs (cis-regulatory modules, CRMs) can be among these filters. The consistency in the location of specific motifs can also indicate an accurately predicted promoter.

The AP2/EREBP superfamily is one of the largest plant TFs, and it can be classified into three families: the AP2, RAV, and ERF (ethylene response factor) families. The most abundant ERF family is further divided into ERF and DREB subfamilies. ERF TFs can bind to a GCC-box element (AGCCGCC) and are involved in hormone signaling pathways and the regulation of pathogenesis-related genes. DREB TFs bind to the dehydration-responsive element (DRE) with the A/GCCGAC motif and regulate the expression of stress-responsive genes [61,62]. Homeodomain TFs are widely conserved proteins accounting for approximately 15–30% of all TFs in eukaryotes that drive the transcription of genes responsible for cell differentiation, morphogenesis, and stem cell pluripotency maintenance. They possess a DNA-binding domain containing a helix-turn-helix (HTH) structure that recognizes a short 5′-TAAT-3′ motif with very moderate specificity [63,64]. Cold, salinity, drought, and other protein-damaging stress factors induce activation and trimerization of HSF, allowing binding of each HSF monomer to a heat shock element (HSE). HSE is located at the TSS of HSP genes and includes at least two inverted repeats with a 5′-nGAAn-3′ (5′-nGAAnnTTCn-3′) consensus motif upstream of the TATA-box [65,66]. MYB-like proteins control plant metabolism, development, cell fate, and stress response. TFs containing the R2R3-type MYB domain, typical for plants, usually bind to an AC-enriched DNA motif (AC-elements), such as 5′-ACC(A/T)A(A/C)-3′ [67,68]. PWMs belonging to Homeodomain, Heat shock, and Myb TFs have two peaks (Figure 2b–d) that match AT-rich positions at 0 and –40 bp corresponding to TSS and TATA-box, respectively. Similarly, AP2/EREBP TFs with their GC-rich binding motifs had a pronounced drop near the TATA-box region (Figure 2a).

TSSPlant [20] combines the EM algorithm and neural networks to estimate a comprehensive set of features. The prediction accuracy is similar to another classic method that calculates the free energy of the free-energy change of the DNA duplex implemented in PromPredict [69]. To the best of our knowledge, there are practically no experimental studies verifying 5′-UTR on a genome-wide level or for a large subset of genes in conifers. However, experiments with individual genes have been carried out in *Picea glauca* [70–72]. A trans-activation assay with *Agrobacterium* transient transformation method was used to evaluate promoter–TF interactions in 12 genes involved in lignin biosynthesis enzymes, cell-wall synthesis and remodeling, and transcriptional regulation [70]. The promoter of cellulose synthase gene PgCesA3, containing MYB cis-regulatory elements [71], and cinnamyl alcohol dehydrogenase (CAD), containing cis-elements matching MYB, WRKY, and bHLH [72], have been confirmed to induce expression of GUS reporter gene in transgenic spruce in differentiating xylem and foliar guard cells, and in lignifying tissues, respectively. We suggest that on the datasets coming from non-model plant organisms, these two algorithms can be used together for higher accuracy of the TSS prediction. The results obtained in this study can be further validated in the future using chromatin immunoprecipitation analysis (ChIP-chip and ChIP-seq), identification of DNase I-hypersensitive sites, CAGE sequencing, or PEAT. In addition, the in silico predictions may be used as complementary support in cases of ambiguous signals that can be captured by ChIP-seq or CAGE-seq analysis.

3.2. Nucleotide Composition of Promoters and Coding Regions

Most angiosperm genomes have a distinct 5′ to 3′ decreasing gradient of the GC content of coding regions. This effect is manifested the most at the third codon position. There is a possible connection between recombination and the 5′–3′ GC gradient, as the recombination rate is higher around the TSSs, which creates a 5′–3′ recombination gradient [73,74]. It was proposed that a 5′–3′ GC gradient can indicate recombination initiation at TSSs [40].

Enrichment of DNA in guanine and cytosine nucleotides is associated with higher gene compactness and density and higher recombination rates than less GC-enriched regions [40]. It has been observed in multiple species that genes can be grouped into two classes based on the GC content in the third nucleotide position of the coding sequences [75–77]. As was reported by Serres-Giardi et al. [41] and Tatarinova et al. [39], in some plants, GC_3-poor and GC_3-rich genes differ significantly in length, with longer coding sequences tending to have a lower frequency of G+C nucleotides in the third position. It is also believed that the prevalence of GC nucleotides in shorter genes is the result of their length, as the GC content of a gene is an average of the existing GC gradient. The shorter GC-rich genes tend to be either mono-exonic or have fewer exons and introns in general, directly affecting the average GC content, as introns have a lower GC content than exons. According to Glémin et al. [40], the unimodal distribution of GC_3 content indicates a smaller GC gradient within the genes and a lower recombination rate. It was also previously noticed that GC_3-rich genes could show more variable expression, more frequently have TATA-dependent promoters, and are commonly involved in stress response pathways. An observed peak in the CG skew around the predicted TSSs (Figure 6d) had been previously linked to the transcriptional efficiency and methylation status of the GC_3-rich genes [42].

The CG-skews in the four conifer genomes were lower than that reported for thale cress (*A. thaliana*), but to conclude whether it was due to the quality of conifer genome assembly or due to differences between gymnosperms and angiosperms (or monocots and dicots for that matter) many more genomes should be analyzed, which is beyond the scope of this study. However, it can be done using the tools presented here and will be an exciting avenue for prospective studies.

4. Materials and Methods

4.1. Genome Assemblies and Annotations

Pinus taeda genome assembly and annotation Pita v2_01 [45,78,79] were taken from https://treegenesdb.org/FTP/Genomes/Pita/v2.01 (accessed on 21 January 2020). Genome assembly PG29 v3.0 and the corresponding annotation for *Picea glauca* [44,80] were taken from ftp://plantgenie.org/Data/ConGenIE/Picea_glauca/PG29/v3.0/ (accessed on 16 January 2020); to our knowledge, assembly v4.0 has not yet been fully annotated. However, we also considered the manual annotation for assembly v4.0 added to annotation v3.0 *Picea abies* [43]; genome data Pabies_v1.0 were retrieved from ftp://plantgenie.org/Data/ConGenIE/Picea_abies/v1.0 (accessed on 16 January 2020). Corresponding annotations for the High Confidence set (predicted gene models with more than 70% homology with reference proteins) and the Medium Confidence set (predicted gene models with between 30% and 70% homology with reference proteins) as provided by the authors were grouped. For *Larix sibirica* [46], genome assembly under NCBI GenBank accession NWUY0000000000 (BioProject PRJNA393226) and unpublished draft annotation data were used. The detailed parameters of the genome assemblies used in the study are presented in Table 1.

4.2. Gene Filtering

To filter out possible pseudogenes and putative predicted coding sequences that do not represent functional genes, all gene models retrieved from genomic annotations were aligned against the database of RNA-seq data, including ESTs and TSAs (Table S1) of a related species using HISAT2 [81] (Table S4). To verify the selected gene models, we aligned the corresponding protein products to RefSeq plant protein sequences using BLASTp (Table S4).

4.3. Prediction of TSS

Prediction of putative TSSs was performed in the upstream sequences of selected genes, which were defined as regions of –1000 and +250 bp around the start codon, using the TSSPlant program [20]. TSSPlant utilizes the expectation-maximization (EM) algorithm and neural networks to estimate 17 and 15 features for predicting TATA and TATA-less promoters, respectively. The complex set of estimated parameters includes such features as the presence of classic plant promoter motifs (TATA, INR, DPE, YP), variation in nucleotide composition (CG- and AT-skews), oligomer scoring, and others. As the algorithm determines several possible start sites with the best scores, selecting the best prediction was utilized to leave one TSS per gene. Assuming that a gamma distribution can describe the length of the 5′-UTR, we compiled the pool of 5′-UTR lengths from the annotations of several model plants (*Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor*, and *Populus trichocarpa*). We computed k and theta parameters that determine the shape and scale of the distribution of the 5′-UTR lengths. The probability density function was used to select the most likely TSS positions using these parameters (Table S4).

4.4. Nucleotide Composition Analysis

Nucleotide frequency analysis of promoters was performed in TSS-centered sequences (–1000, +200 around TSS). CA and TATA motif frequencies were calculated with a sliding window (width = 20, increment step = 10) using the *stringr* R package. The CG-skew of a given sequence was defined as a proportion $(C-G)/(C+G)$ and calculated with a sliding window width of 50 bp and a window increment step of 10 bp along the promoter sequence (Table S4). GC_3 was calculated using CDSs and the R package *seqinr*. The slope of the GC_3 gradient was estimated using linear regression between GC_3 and the position relative to the first coding nucleotide (ATG) based on the first 1000 bp of a spliced transcript sequence. Genes were divided into GC_3-poor and GC_3-rich sets using 10% and 90% quantiles of GC_3. DNA duplex stability was estimated using PromPredict [69] in a 15 bp sliding window. All manipulations were performed using bedtools and custom R and C scripts. A TFBS search was performed using the TRANSFAC database and MATCH [82].

4.5. Genome Visualization

The genomic sequences, tracks with gene models, predicted TSS, and RNA-seq coverage data were deposited at the Persephone genome viewer at https://web.persephonesoft.com (accessed on 31 January 2022). The choice of using this visualization solution was based on its ability to align and analyze genomic sequences in real time. Persephone is a state-of-the-art genome browser specifically designed to show and compare multiple sequences and genetic maps on one screen (see also help files at https://help.persephonesoft.com, accessed on 31 January 2022). The aligned maps can be linked using common markers, orthologous gene pairs, or regions of sequence similarity. The sequence maps can be aligned at a specific zoom level by running a real-time BLASTn search that visualizes structural variations by displaying identical sequence regions with highlighted nucleotide substitutions and indels.

5. Conclusions

This work is the first genome-wide prediction of TSS in genomes larger than 10 Gbp. Ancient origin, massive genome size, not associated with recent polyploidization or duplication, and extensive gene families (with higher copy number than in most angiosperms) distinguish conifers from other plants. The predicted TSSs and their putative promoter regions provide the basis for future experimental verification and present a valuable resource for better understanding gene regulation and investigating the evolutionary relationships between gymnosperm and angiosperm clades. Identification of TSS can find its implementation in genetic-assisted breeding and genome editing, providing opportunities for more precise mapping and the targeting of SNPs in functional genomic regions and quantitative trait loci associated with adaptive traits, such as growth rate, cold- and drought-resistance, and resistance to pathogen invasion.

We predicted promoter regions for several conifer species using computational strategies based on the expectation-maximization and neural network classification method utilized by the TSSPlant algorithm. The predictions were ranked using the probabilistic distributions of UTR lengths in model plants. The predicted TSSs were assessed using the profiles of standard free energy of a DNA duplex and the distribution of a CG-skew, both of which showed peaks around putative TSS positions and near the TBP binding site. The positional distributions of TFBS for several abundant transcription factor families also support the predicted promoters.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23031735/s1.

Author Contributions: Conceptualization, T.V.T.; methodology, T.V.T. and M.E.T.; software, M.E.T.; validation, E.I.B., M.E.T., K.V.K. and T.V.T.; formal analysis, E.I.B.; investigation, E.I.B.; resources, M.E.T. and K.V.K.; data curation, E.I.B. and M.E.T.; writing—original draft preparation, E.I.B.; writing—review and editing, E.I.B., M.E.T., K.V.K. and T.V.T.; visualization, E.I.B., M.E.T. and T.V.T.; supervision, K.V.K. and T.V.T.; project administration, K.V.K. and T.V.T.; funding acquisition, K.V.K. and T.V.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by research grant No. 14.Y26.31.0004 from the Government of the Russian Federation. The funding agency played no role in the design of the study and collection material, analysis, and interpretation of data, and in writing the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during the current study are available from the corresponding author on request.

Acknowledgments: We thank Dmitry Kuzmin and Vadim Sharov from the Department of High-Performance Computing at the Siberian Federal University (Krasnoyarsk, Russia) for their help with computing using their high-performance computer cluster. We thank three anonymous reviewers for thoroughly and promptly reviewing our manuscript and their valuable recommendations and comments that helped us greatly improve the manuscript. We acknowledge support from the Open Access Publication Funds of the University of Göttingen.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Tatarinova, T.; Kryshchenko, A.; Triska, M.; Hassan, M.; Murphy, D.; Neely, M.; Schumitzky, A. NPEST: A nonparametric method and a database for transcription start site prediction. *Quant. Biol.* **2013**, *1*, 261–271. [CrossRef] [PubMed]
2. Reyes, A.; Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **2018**, *46*, 582–592. [CrossRef] [PubMed]
3. Juven-Gershon, T.; Kadonaga, J.T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* **2010**, *339*, 225–229. [CrossRef]
4. Alexandrov, N.N.; Troukhan, M.E.; Brover, V.V.; Tatarinova, T.; Flavell, R.B.; Feldmann, K.A. Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. *Plant Mol. Biol.* **2006**, *60*, 69–85. [CrossRef]

5. Alexandrov, N.N.; Brover, V.V.; Freidin, S.; Troukhan, M.E.; Tatarinova, T.V.; Zhang, H.; Swaller, T.J.; Lu, Y.-P.; Bouck, J.; Flavell, R.B.; et al. Insights into corn genes derived from large-scale cDNA sequencing. *Plant Mol. Biol.* **2009**, *69*, 179–194. [CrossRef]
6. Troukhan, M.; Tatarinova, T.; Bouck, J.; Flavell, R.B.; Alexandrov, N.N. Genome-wide discovery of cis-elements in promoter sequences using gene expression. *OMICS* **2009**, *13*, 139–151. [CrossRef] [PubMed]
7. Tatarinova, T.V.; Chekalin, E.; Nikolsky, Y.; Bruskin, S.; Chebotarov, D.; McNally, K.L.; Alexandrov, N. Nucleotide diversity analysis highlights functionally important genomic regions. *Sci. Rep.* **2016**, *6*, 35730. [CrossRef]
8. Triska, M.; Ivliev, A.; Nikolsky, Y.; Tatarinova, T.V. Analysis of cis-regulatory elements in gene co-expression networks in cancer. *Methods Mol. Biol.* **2017**, *1613*, 291–310. [CrossRef]
9. Roy, A.L.; Singer, D.S. Core promoters in transcription: Old problem, new insights. *Trends Biochem. Sci.* **2015**, *40*, 165–171. [CrossRef]
10. Sandelin, A.; Carninci, P.; Lenhard, B.; Ponjavic, J.; Hayashizaki, Y.; Hume, D.A. Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. *Nat. Rev. Genet.* **2007**, *8*, 424–436. [CrossRef]
11. Lagrange, T.; Kapanidis, A.N.; Tang, H.; Reinberg, D.; Ebright, R.H. New core promoter element in RNA polymerase II-dependent transcription: Sequence-specific DNA binding by transcription factor IIB. *Genes Dev.* **1998**, *12*, 34–44. [CrossRef] [PubMed]
12. Deng, W.; Roberts, S.G.E. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.* **2005**, *19*, 2418–2423. [CrossRef] [PubMed]
13. Kadonaga, J.T. The DPE, a core promoter element for transcription by RNA polymerase II. *Exp. Mol. Med.* **2002**, *34*, 259–264. [CrossRef]
14. Burke, T.W.; Kadonaga, J.T. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* **1996**, *10*, 711–724. [CrossRef] [PubMed]
15. Lee, D.-H.; Gershenzon, N.; Gupta, M.; Ioshikhes, I.P.; Reinberg, D.; Lewis, B.A. Functional characterization of core promoter elements: The downstream core element is recognized by TAF1. *Mol. Cell Biol.* **2005**, *25*, 9674–9686. [CrossRef]
16. Franco-Zorrilla, J.M.; Solano, R. Identification of plant transcription factor target sequences. *Biochim. Biophys. Acta Gene Regul. Mech.* **2017**, *1860*, 21–30. [CrossRef]
17. Morton, T.; Petricka, J.; Corcoran, D.L.; Li, S.; Winter, C.M.; Carda, A.; Benfey, P.N.; Ohler, U.; Megraw, M. Paired-end analysis of transcription start sites in *Arabidopsis* reveals plant-specific promoter signatures. *Plant Cell* **2014**, *26*, 2746–2760. [CrossRef]
18. Bianchi, V.J.; Rubio, M.; Trainotti, L.; Verde, I.; Bonghi, C.; Martínez-Gómez, P. *Prunus* transcription factors: Breeding perspectives. *Front. Plant Sci.* **2015**, *6*, 443. [CrossRef]
19. Narang, V.; Sung, W.-K.; Mittal, A. Computational modeling of oligonucleotide positional densities for human promoter prediction. *Artif. Intell. Med.* **2005**, *35*, 107–119. [CrossRef]
20. Shahmuradov, I.A.; Umarov, R.K.; Solovyev, V.V. TSSPlant: A new tool for prediction of plant Pol II promoters. *Nucleic Acids Res.* **2017**, *45*, e65. [CrossRef]
21. Pachganov, S.; Murtazalieva, K.; Zarubin, A.; Sokolov, D.; Chartier, D.R.; Tatarinova, T.V. TransPrise: A novel machine learning approach for eukaryotic promoter prediction. *PeerJ* **2019**, *7*, e7990. [CrossRef] [PubMed]
22. De Medeiros Oliveira, M.; Bonadio, I.; Lie de Melo, A.; Mendes Souza, G.; Durham, A.M. TSSFinder—Fast and accurate ab initio prediction of the core promoter in eukaryotic genomes. *Brief. Bioinform.* **2021**, *22*, bbab198. [CrossRef]
23. Ibraheem, O.; Botha, C.E.J.; Bradley, G. In silico analysis of cis-acting regulatory elements in 5′ regulatory regions of sucrose transporter gene families in rice (*Oryza sativa* Japonica) and *Arabidopsis thaliana*. *Comput. Biol. Chem.* **2010**, *34*, 268–283. [CrossRef] [PubMed]
24. Duraisamy, G.S.; Mishra, A.K.; Kocabek, T.; Matoušek, J. Identification and characterization of promoters and cis-regulatory elements of genes involved in secondary metabolites production in hop (*Humulus lupulus* L.). *Comput. Biol. Chem.* **2016**, *64*, 346–352. [CrossRef] [PubMed]
25. Wong, D.C.J.; Lopez Gutierrez, R.; Gambetta, G.A.; Castellarin, S.D. Genome-wide analysis of cis-regulatory element structure and discovery of motif-driven gene co-expression networks in grapevine. *DNA Res.* **2017**, *24*, 311–326. [CrossRef] [PubMed]
26. Kumari, S.; Ware, D. Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots. *PLoS ONE* **2013**, *8*, e79011. [CrossRef] [PubMed]
27. Gan, Y.; Guan, J.; Zhou, S. A Comparison study on feature selection of DNA structural properties for promoter prediction. *BMC Bioinform.* **2012**, *13*, 4. [CrossRef] [PubMed]
28. Kanhere, A.; Bansal, M. Structural properties of promoters: Similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res.* **2005**, *33*, 3165–3175. [CrossRef] [PubMed]
29. Triska, M.; Solovyev, V.; Baranova, A.; Kel, A.; Tatarinova, T.V. Nucleotide patterns aiding in prediction of eukaryotic promoters. *PLoS ONE* **2017**, *12*, e0187243. [CrossRef] [PubMed]
30. Yella, V.R.; Kumar, A.; Bansal, M. Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. *Sci. Rep.* **2018**, *8*, 4520. [CrossRef]
31. Kozobay-Avraham, L.; Hosid, S.; Bolshoy, A. Curvature distribution in prokaryotic genomes. *Silico Biol.* **2004**, *4*, 361–375.
32. Kumar, A.; Bansal, M. Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression. *DNA Res.* **2017**, *24*, 25–35. [CrossRef] [PubMed]
33. Pandey, S.P.; Krishnamachari, A. Computational analysis of plant RNA Pol-II Promoters. *Biosystems* **2006**, *83*, 38–50. [CrossRef] [PubMed]

34. Zuo, Y.-C.; Li, Q.-Z. Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-skew and DNA geometric flexibility. *Genomics* **2011**, *97*, 112–120. [CrossRef] [PubMed]
35. Tatarinova, T.; Brover, V.; Troukhan, M.; Alexandrov, N. Skew in CG content near the transcription start site in *Arabidopsis thaliana*. *Bioinformatics* **2003**, *19*, i313–i314. [CrossRef]
36. Fujimori, S.; Washio, T.; Tomita, M. GC-compositional strand bias around transcription start sites in plants and fungi. *BMC Genom.* **2005**, *6*, 26. [CrossRef] [PubMed]
37. Aerts, S.; Thijs, G.; Dabrowski, M.; Moreau, Y.; De Moor, B. Comprehensive analysis of the base composition around the transcription start site in metazoa. *BMC Genom.* **2004**, *5*, 34. [CrossRef]
38. Carels, N.; Bernardi, G. Two classes of genes in plants. *Genetics* **2000**, *154*, 1819–1825. [CrossRef] [PubMed]
39. Tatarinova, T.V.; Alexandrov, N.N.; Bouck, J.B.; Feldmann, K.A. GC3 biology in corn, rice, sorghum and other grasses. *BMC Genom.* **2010**, *11*, 308. [CrossRef]
40. Clément, Y.; Fustier, M.-A.; Nabholz, B.; Glémin, S. The bimodal distribution of genic GC content is ancestral to monocot species. *Genome Biol. Evol.* **2014**, *7*, 336–348. [CrossRef] [PubMed]
41. Serres-Giardi, L.; Belkhir, K.; David, J.; Glémin, S. Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell* **2012**, *24*, 1379–1397. [CrossRef] [PubMed]
42. Chan, K.-L.; Tatarinova, T.V.; Rosli, R.; Amiruddin, N.; Azizi, N.; Halim, M.A.A.; Sanusi, N.S.N.M.; Jayanthi, N.; Ponomarenko, P.; Triska, M.; et al. Evidence-based gene models for structural and functional annotations of the oil palm genome. *Biol. Direct* **2017**, *12*, 21. [CrossRef]
43. Nystedt, B.; Street, N.R.; Wetterbom, A.; Zuccolo, A.; Lin, Y.-C.; Scofield, D.G.; Vezzi, F.; Delhomme, N.; Giacomello, S.; Alexeyenko, A.; et al. The Norway spruce genome sequence and conifer genome evolution. *Nature* **2013**, *497*, 579–584. [CrossRef] [PubMed]
44. Warren, R.L.; Keeling, C.I.; Yuen, M.M.S.; Raymond, A.; Taylor, G.A.; Vandervalk, B.P.; Mohamadi, H.; Paulino, D.; Chiu, R.; Jackman, S.D.; et al. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J.* **2015**, *83*, 189–212. [CrossRef] [PubMed]
45. Zimin, A.V.; Stevens, K.A.; Crepeau, M.W.; Puiu, D.; Wegrzyn, J.L.; Yorke, J.A.; Langley, C.H.; Neale, D.B.; Salzberg, S.L. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience* **2017**, *6*, giw016. [CrossRef]
46. Kuzmin, D.A.; Feranchuk, S.I.; Sharov, V.V.; Cybin, A.N.; Makolov, S.V.; Putintseva, Y.A.; Oreshkova, N.V.; Krutovsky, K.V. Stepwise large genome assembly approach: A case of Siberian larch (*Larix sibirica* Ledeb). *BMC Bioinform.* **2019**, *20*, 37. [CrossRef] [PubMed]
47. Holt, C.; Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform.* **2011**, *12*, 491. [CrossRef]
48. Yamaguchi, N. LEAFY, a pioneer transcription factor in plants: A mini-review. *Front. Plant Sci.* **2021**, *12*, 701406. [CrossRef]
49. Papuga, J.; Hoffmann, C.; Dieterle, M.; Moes, D.; Moreau, F.; Tholl, S.; Steinmetz, A.; Thomas, C. Arabidopsis LIM proteins: A family of actin bundlers with distinct expression patterns and modes of regulation. *Plant Cell* **2010**, *22*, 3034–3052. [CrossRef]
50. Wong, G.K.-S.; Wang, J.; Tao, L.; Tan, J.; Zhang, J.; Passey, D.A.; Yu, J. Compositional gradients in Gramineae genes. *Genome Res.* **2002**, *12*, 851–856. [CrossRef]
51. Biłas, R.; Szafran, K.; Hnatuszko-Konka, K.; Kononowicz, A.K. Cis-regulatory elements used to control gene expression in plants. *Plant Cell Tiss. Organ. Cult.* **2016**, *127*, 269–287. [CrossRef]
52. Messing, J.; Geraghty, D.; Heidecker, G.; Hu, N.-T.; Kridl, J.; Rubenstein, I. Plant gene structure. In *Genetic Engineering of Plants: An Agricultural Perspective*; Kosuge, T., Meredith, C.P., Hollaender, A., Wilson, C.M., Eds.; Springer: Boston, MA, USA, 1983; pp. 211–227, ISBN 978-1-4684-4544-2.
53. Porto, M.S.; Pinheiro, M.P.N.; Batista, V.G.L.; dos Santos, R.C.; de Albuquerque Melo Filho, P.; de Lima, L.M. Plant promoters: An approach of structure and function. *Mol. Biotechnol.* **2014**, *56*, 38–49. [CrossRef] [PubMed]
54. Dhadi, S.R.; Krom, N.; Ramakrishna, W. Genome-wide comparative analysis of putative bidirectional promoters from rice, *Arabidopsis* and *Populus*. *Gene* **2009**, *429*, 65–73. [CrossRef]
55. Krom, N.; Ramakrishna, W. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, *Arabidopsis*, and *Populus*. *Plant Physiol.* **2008**, *147*, 1763–1773. [CrossRef] [PubMed]
56. Yamamoto, Y.Y.; Yoshioka, Y.; Hyakumachi, M.; Obokata, J. Characteristics of core promoter types with respect to gene structure and expression in *Arabidopsis thaliana*. *DNA Res.* **2011**, *18*, 333–342. [CrossRef]
57. Portales-Casamar, E.; Thongjuea, S.; Kwon, A.T.; Arenillas, D.; Zhao, X.; Valen, E.; Yusuf, D.; Lenhard, B.; Wasserman, W.W.; Sandelin, A. JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **2010**, *38*, D105–D110. [CrossRef] [PubMed]
58. Tian, F.; Yang, D.-C.; Meng, Y.-Q.; Jin, J.; Gao, G. PlantRegMap: Charting functional regulatory maps in plants. *Nucleic Acids Res.* **2020**, *48*, D1104–D1113. [CrossRef]
59. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* **2008**, *9*, 326–332. [CrossRef]
60. Wasserman, W.W.; Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **2004**, *5*, 276–287. [CrossRef]

61. Liu, C.; Zhang, T. Expansion and stress responses of the AP2/EREBP superfamily in cotton. *BMC Genom.* **2017**, *18*, 118. [CrossRef]
62. Dietz, K.-J.; Vogel, M.O.; Viehhauser, A. AP2/EREBP transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling. *Protoplasma* **2010**, *245*, 3–14. [CrossRef] [PubMed]
63. Svingen, T.; Tonissen, K.F. Hox transcription factors and their elusive mammalian gene targets. *Heredity* **2006**, *97*, 88–96. [CrossRef] [PubMed]
64. Noyes, M.B.; Christensen, R.G.; Wakabayashi, A.; Stormo, G.D.; Brodsky, M.H.; Wolfe, S.A. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **2008**, *133*, 1277–1289. [CrossRef]
65. Guo, M.; Liu, J.-H.; Ma, X.; Luo, D.-X.; Gong, Z.-H.; Lu, M.-H. The plant heat stress transcription factors (HSFs): Structure, regulation, and function in response to abiotic stresses. *Front. Plant Sci.* **2016**, *7*, 114. [CrossRef] [PubMed]
66. Miller, G.; Mittler, R. Could heat shock transcription factors function as hydrogen peroxide sensors in plants? *Ann. Bot.* **2006**, *98*, 279–288. [CrossRef] [PubMed]
67. Prouse, M.B.; Campbell, M.M. Interactions between the R2R3-MYB transcription factor, AtMYB61, and target DNA binding sites. *PLoS ONE* **2013**, *8*, e65132. [CrossRef]
68. Wang, B.; Luo, Q.; Li, Y.; Yin, L.; Zhou, N.; Li, X.; Gan, J.; Dong, A. Structural insights into target DNA recognition by R2R3-MYB transcription factors. *Nucleic Acids Res.* **2020**, *48*, 460–471. [CrossRef]
69. Rangannan, V.; Bansal, M. High-quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics* **2010**, *26*, 3043–3050. [CrossRef]
70. Duval, I.; Lachance, D.; Giguère, I.; Bomal, C.; Morency, M.-J.; Pelletier, G.; Boyle, B.; MacKay, J.J.; Séguin, A. Large-scale screening of transcription factor–promoter interactions in spruce reveals a transcriptional network involved in vascular development. *J. Exp. Bot.* **2014**, *65*, 2319–2333. [CrossRef]
71. Sena, J.S.; Lachance, D.; Duval, I.; Nguyen, T.T.A.; Stewart, D.; Mackay, J.; Séguin, A. Functional analysis of the PgCesA3 white spruce cellulose synthase gene promoter in secondary xylem. *Front. Plant Sci.* **2019**, *10*, 626. [CrossRef] [PubMed]
72. Bedon, F.; Levasseur, C.; Grima-Pettenati, J.; Séguin, A.; MacKay, J. Sequence analysis and functional characterization of the promoter of the *Picea glauca* cinnamyl alcohol dehydrogenase gene in transgenic white spruce plants. *Plant Cell Rep.* **2009**, *28*, 787–800. [CrossRef] [PubMed]
73. Choi, K.; Zhao, X.; Kelly, K.A.; Venn, O.; Higgins, J.D.; Yelina, N.E.; Hardcastle, T.J.; Ziolkowski, P.A.; Copenhaver, G.P.; Franklin, F.C.H.; et al. *Arabidopsis* meiotic crossover hotspots overlap with H2A.Z nucleosomes at gene promoters. *Nat. Genet.* **2013**, *45*, 1327–1336. [CrossRef]
74. Hellsten, U.; Wright, K.M.; Jenkins, J.; Shu, S.; Yuan, Y.; Wessler, S.R.; Schmutz, J.; Willis, J.H.; Rokhsar, D.S. Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 19478–19482. [CrossRef]
75. Fortes, G.G.; Bouza, C.; Martínez, P.; Sánchez, L. Diversity in isochore structure among cold-blooded vertebrates based on GC content of coding and non-coding sequences. *Genetica* **2007**, *129*, 281–289. [CrossRef] [PubMed]
76. Jørgensen, F.G.; Schierup, M.H.; Clark, A.G. Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of *Apis mellifera*. *Mol. Biol. Evol.* **2007**, *24*, 611–619. [CrossRef] [PubMed]
77. Lynch, D.B.; Logue, M.E.; Butler, G.; Wolfe, K.H. Chromosomal G + C content evolution in yeasts: Systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biol. Evol.* **2010**, *2*, 572–583. [CrossRef] [PubMed]
78. Neale, D.B.; Wegrzyn, J.L.; Stevens, K.A.; Zimin, A.V.; Puiu, D.; Crepeau, M.W.; Cardeno, C.; Koriabine, M.; Holtz-Morris, A.E.; Liechty, J.D.; et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* **2014**, *15*, R59. [CrossRef] [PubMed]
79. Wegrzyn, J.L.; Liechty, J.D.; Stevens, K.A.; Wu, L.-S.; Loopstra, C.A.; Vasquez-Gross, H.A.; Dougherty, W.M.; Lin, B.Y.; Zieve, J.J.; Martínez-García, P.J.; et al. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* **2014**, *196*, 891–909. [CrossRef]
80. Birol, I.; Raymond, A.; Jackman, S.D.; Pleasance, S.; Coope, R.; Taylor, G.A.; Yuen, M.M.S.; Keeling, C.I.; Brand, D.; Vandervalk, B.P.; et al. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* **2013**, *29*, 1492–1497. [CrossRef] [PubMed]
81. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915. [CrossRef] [PubMed]
82. Kel, A.E.; Gössling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O.V.; Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**, *31*, 3576–3579. [CrossRef] [PubMed]