Comparative study of pine reference genomes reveals transposable element interconnected gene networks

Angelika Voronova (angelika.voronova@silava.lv)
Latvian State Forest Research Institute https://orcid.org/0000-0002-4824-1706
Martha Rendón-Anaya
Swedish University of Agricultural Sciences https://orcid.org/0000-0002-8047-223X
Pär Ingvarsson
Swedish University of Agricultural Sciences https://orcid.org/0000-0001-9225-7521
Ruslan Kalendar
University of Helsinki https://orcid.org/0000-0003-3986-2460
Dainis Ruņģis
Latvian State Forest Research institute "Silava" https://orcid.org/0000-0001-5173-2912

Research

Keywords: pine reference genome, gene networks, gene regulation, node gene, transposable elements, retrotransposons, MITE

Posted Date: July 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-34803/v2

Version of Record: A version of this preprint was published on October 16th, 2020. See the published version at https://doi.org/10.3390/genes11101216.

Abstract

Background: Sequencing the giga-genomes of several pine species belonging to the ancient gymnosperm clade has enabled comparative genomic analyses of these widely distributed outcrossing tree species. Initial sequence studies have revealed the wide distribution and extraordinary diversity of transposable elements (TEs) that occupy the large intergenic spaces. Our previous investigations revealed significant variations of class I TEs within pine subpopulations, but inoculation with pathogen induced correlated expression of TE families in pine seedlings, suggesting TE co-localisation with stress-responsive genes. In this study, we analyzed the distribution of TEs in gene regions of the assembled genomes of Pinus taeda and Pinus lambertiana using high-performance computing resources.

Results: The quality of draft genomes and the genome annotation have significant consequences for the investigation of TEs and these aspects are discussed. Several TE families were identified in both species genomes frequently inserted into genes or their flanks. The non-autonomous MITE3321 family found in gene flanking regions and provide TATA boxes, several ARR1, DOF, WRKY and GT-binding sites, which are important signals in plant transcription activation and stress-response regulation. Distribution of MITE3321 across gene non-coding regions suggests action of selective pressure on this sequences and formation of gene sub-networks, depending on the location of MITE insertions. DNA TE DTX184 could potentially form microRNAs or provide its target site, thereby connect about 200 important stress-responsive genes in both pine species. Several retrotransposons propagated in gene regions were also identified, such as Copia-1813 containing important light-responsive regulative sequences, however full-length structures of low-copy-number TEs couldn't be verified in the current genome assemblies. Only rare gene homologs carried similar insertions, indicating that most transposition events occurred after separation of investigated pine species. Node genes that contain many types of interspersed repeats were identified and observed in multiple potential transposable element associated networks.

Conclusions: This study demonstrated the increased accumulation of TEs in the introns of stress-responsive genes of pines and the probability of rewiring them in responsive networks and sub-networks interconnected with node genes containing multiple TEs. Many such regulatory influences could lead to the adaptive environmental response clines that are characteristic of naturally spread pine populations.

Introduction

The functional role of transposable element (TE) insertions distributed throughout plant genomes is less studied and phenotypic changes are less obvious compared with protein-coding regions [1–3]. Transposition of TEs is linked to stress conditions and evolutionary change [3–12], but these sequences are usually controlled by the host organism [13]. Progress in transcriptome sequencing has revealed multiple types of non-coding RNAs that originate from TE sequences, nested elements, and their relicts remaining after purifying selection [3, 14–16]. Information from various plant species and genes where TE-derived insertions are linked to phenotype alterations is accumulating [17–24]. Reported influences of these insertions include: gene interruption by transposition; alteration of gene expression levels via

providing additional transcription initiation signals or downregulation by methylation; exon shuffling and alternate splicing; initiation of antisense transcription; non-coding RNA production or providing target sites; providing additional poly-A signals that change transcript stability and transport from nucleus; multiple insertions of similar TEs could establish dynamic gene networks [10, 25–30].

TEs are organized into classes (retrotransposons and DNA transposons), superfamilies (*Ty1-copia*, *Ty3-gypsy*, etc.), families, and subfamilies. Up to 20% nucleotide sequence variation is permitted between family members [31]. The most prevalent TEs in plants are long terminal repeat (LTR) retrotransposons (RLX), which are sequences that contain direct repeats that flank the internal sequence or body of the element. LTRs may contain transcription initiation (plant type II promoters) and termination signals (polyadenylation site), polypurine tracts, integrase-binding signals, tRNA primer binding sites (PBS), and cis-acting elements [32]. Several elements with extraordinarily short LTRs have been described for angiosperms [e.g. 85 bp, *FRetro129* [33]], while the LTRs of other RLXs can be over 5 kb in length [*Sukkula*, [34, 35], *Grande*, [36], and *Ogre*, [37]]. However, these are exceptions and LTRs are typically 0.1–2 kb in length [38, 39]. The internal region of autonomous RLXs contain gag- and polyprotein-coding domains that produce the proteins necessary for retrotransposition, namely protease, reverse transcriptase, and integrase. Non-autonomous elements contain empty, partial, or disrupted polyprotein sequences and use proteins produced by autonomous elements, sometimes integrating more effectively than their autonomous partner [35, 40–42]. Reverse transcription processes introduce mutations and can produce chimeric elements by template switching [43]. Non-homologous recombination often involves highly similar repeats and can result in deletions of portions of TEs or even neighboring genes [44]. Some genome regions contain multiple insertions of TEs into each other, forming nested repeat regions [45]. During transposition, RLXs proliferate via RNA transcript intermediates, but DNA transposons excise from their current location and migrate to a new genomic locus. Therefore, copy number proliferation is not as pronounced for DNA transposons. However, some nonautonomous DNA transposons have been found more frequently in plant gene regions, such as MITE elements [46–48]. Distributed TE families are broadly used as molecular markers in population genetic investigations and marker-assisted breeding [48–57].

Conifer genomes are greater than 15 Gbp in size and contain extended regions of non-coding DNA, much of which are derived from TEs. The genomes of pines represent one extreme in plants, with stable diploid genomes that are expanded by the proliferation of TEs, in contrast to frequent polyploidisation events in angiosperms [58–60]. TEs are rarely deleted via non-homologous recombination processes, therefore the ratio of solo LTR to full length elements is lower in conifers and most TEs are represented as full-length elements [60]. The genome sequences of several gymnosperm species have been published [60–63]. However, data quality, coverage, and gene models are continually improving [64]. The loblolly pine (*Pinus taeda*) genome contains a high proportion of TEs; there are approximately 1,500 families with the prevalence of LTR RLXs (42% of genome sequence [65, 66]. In conifer genomes, TE sequences are highly diverged, however, some conserved families are found in *Pinaceae* and even in more distant gymnosperms [67, 68]. Our previous studies on Scots pine (*Pinus sylvestris*) revealed a rapid expression of TE-containing sequences in response to stress conditions [69]. The composition of the studied RLXs

was found to be specific to pine lineages, while family proportions and copy numbers showed variation between and within pine species [68]. The more frequent distribution of a particular RLX family in the pine genome does not lead to its higher expression rate, but expression levels of different RLX were strongly correlated within individuals [70]. Therefore, we hypothesized that stress-responsive genes or their surrounding regions could be enriched with particular TE families, that are co-expressed with genes in unfavorable conditions. The aim of this study was to analyze genes containing TEs and the distribution of TEs in genomic sequences of genes and gene-flanking regions in the available pine reference genomes (*Pinus taeda* and *Pinus lambertiana*). We also explored the possibility of transferring this information to *Pinus sylvestris* genome studies. The obtained results were used to evaluate if the distribution of TEs in gene regions is random regarding different gene regions (e.g. flanks or introns), species or TE families; if genes containing similar TE families are involved in similar processes, and if the investigated TEs contain potential gene regulatory motifs.

Materials And Methods

Generation of datasets

Reference genomes for *P. taeda* v.1.01 and v.2.0 and *P. lambertiana* v.1.01 were downloaded from https://treegenesdb.org. The *PIER* v.2.0 (Pine Interspersed Element Resource) (Neale et al., 2014) database was used for TE identification. The *UPPMAX* (Uppsala Multidisciplinary Center for Advanced Computational Science) resource was initially used for analyses of genomic sequences. The RigaTechnical University (RTU) High Performance Computing (HPC) Centre was used for the further analysis in Latvia. The publicly available PIER 2.0 database contains conifer TEs and nested repeats that were recognized by automatic genome annotation from the *P. taeda* v.1.01 genome. A large portion of the database entries contain nested repeats with 2–4 pairs of direct LTRs rather than full-length TE sequences, which could negatively influence the analyses in this study. One TE insertion within a particular gene will have hits to hundreds of different database entries containing high sequence similarity to this TE; therefore, all full-length elements were clustered with *CD-Hit* v.4.6.4 (Fu et al., 2012) utilizing a sequence identity threshold of 0.8. This resulted in 15,622 unique TE representatives from the 19,700 entries originally found in the database. A further problem was the high sequence diversity of conifer TEs that contain many insertion-deletion polymorphisms. It is possible to apply strict similarity search criteria, but then two genes containing insertions of the same TE family will be not recognized as members of the same network. In this case, potentially important carriers of similar transcription signals (eg. LTR sequences) will be not recognized. Additionally, genome scaffolds are often truncated at repeat insertion loci, leaving numerous genes with only partial TE sequences unmasked. Therefore, all direct repeats were extracted from each defined mobile element in the *PIER* database, resulting in 24,591 repeats. The new data set of TE-derived repeats contained sequences from 79 bp to 13,295 bp in length, with an average length of 579 bp and a median length of 417 bp. The presence of nested repeats could lead to the extraction of direct repeats of the same TE family, or sequences that represent a body of nested TEs. This was verified by individual sequence inspection and all-to-all alignment. To reduce such

errors, extracted TE-derived repeats were *CD-Hit* clustered, resulting in 9,659 entries. Sequences of only 0.1–2 kb in length were used for further analyses, resulting in a final database of 9,107 entries (5.7% reduction).

Extraction of gene introns and flanking regions

Bedtools v.2.27.1 (Quinlan and Hall, 2010) were used to extract sequences of all genes containing introns from the reference genomes. Full-length and LTR TE similarity searches within the extracted gene regions for each reference genome were performed with local *BLASTn* v.2.2.26 (Altschul et al., 1990). *Samtools* (Li et al., 2009) were used for matching gene sequence extraction and *BEDOPS* v.2.4.35 (Neph et al., 2012) for extraction of gene-flanking regions, with 5-kb sequences from the 5’ and 3’ flanking regions of each gene extracted into separate databases. Flanking sequences were further divided into 1-kb fragments, resulting in 10 databases. If any of 1-kb region included a scaffold end, the shorter remaining sequence was included. Each flanking region had reference to its gene ID and position. Sequences were compared using *BLAST* with the following parameters: percent identity ≥80%, alignment length ≥100 bp, Query

Coverage Per HSP ≥90%. For analyses in *P. lambertiana*, the Query Coverage Per HSP parameter was lowered to ≥80% as the previous parameter set produced no hits. A *t*-test was calculated for the evaluation of TE enrichment significance to flanking regions in the vicinity of genes compared with other regions. The standard deviation of the distribution of differences between independent sample means was estimated for each TE-derived repeat (for the equal variance *t*-test). Enrichment significance of p=0.001 was considered if t was greater than 5.04 (df=8) and p=0.05 if t>2.31 (df=8).

To overcome inconsistencies found in genome files, *P. taeda* v.2.0 true genomic coordinates were evaluated for 15,534 transcripts annotated with any Gene Ontology term (from 36,730 entries annotated as genes). Correct genomic coordinates of transcripts were identified by running a computationally intense local NCBI short-blast algorithm with parameters of sequence identity >98% and 100% query coverage by HSP. The resulting genomic coordinates were further sorted and corrected as follows: scaffolds containing identical hits and gene coordinates were deleted, with genes larger than 20 kb manually verified and repetitions of partial exon structures or gene repeats on one scaffold deleted leaving only one conventional structure with all exons present. Further gene transcripts were screened against the *PIER* database and genes matching TEs with more than 50% query coverage and more than 80% sequence identity were filtered out. All extractions and analyses were repeated with the new gene set. After all steps, genes containing two or more types of repeats (filling less than 50% of template) were still present in the dataset and were later filtered from the results. Gene genomic sequences from the transcription start to termination site were extracted from the reference genomes into separate databases.

Analysis of LTR structure and TFBS

The database of LTR representatives was used for analysis of flanking and intron regions, which was generated by computational selection of any repeated region found in one RLX sequence defined by automated predictions. Then, 5-kb gene-flanking regions containing hypothetical distributed LTRs were extracted and full-length TE structures were identified if possible. Further, verified TE structures were repeatedly queried among all gene regions and the related information was updated. Hits with ≥1 kb query coverage and more than 80% nucleotide identity to full-length TE were considered as strong candidates. Consensus sequences were evaluated after multiple alignment of extracted TE sequences. Characteristic features of TEs were identified using NCBI Conserved Domain search, *REPFIND* repeat prediction tool (Betley et al., 2002), and *Repbase* (Bao et al., 2015). *Softberry* tools (ScanWM-PL, TSSPlant, NSITE-PL, POLYAH, http://www.softberry.com; (Solovyev et al., 2010)) and *PLACE* v.30.0 (Higo et al., 1999) were used for the identification of TFBS, plant promoters, and poly-A sites in LTRs that were found frequently near exons. *miRBase* (Kozomara et al., 2019) and *RNAfold* web server (Gruber et al., 2008) were used for microRNA prediction.

Gene networking and Gene Ontology analysis

The Gene Ontology (GO) classification file was obtained from the Gene Ontology Consortium (http://www.geneontology.org/) and *P. taeda* v.1.01 gene functional annotation was downloaded from *PLAZA* (Van Bel et al., 2012). Gene functional annotation for *P. taeda* v.2.0 was generated by considering gene transcript homology to *P. taeda* v.1.01, which enabled comparison between differing numberings of genome versions, with less than 50% of the genes categorized to any GO term for *P. taeda* v.2.0. For *P. lambertiana* v.1.01, a functional annotation file was downloaded from the *Treegenes* data repository (https://treegenesdb.org/). In total, 8,943 genes from 13,637 were assigned with any GO term. BINGO 3.0.3 (Maere et al., 2005) was used for gene networking analysis using the custom annotation available, with *Cytoscape* v.3.3.0 (Smoot et al., 2011) used for gene network visualization. Gene transcripts containing TE insertions of one family in their genomic sequences and with similar TE localization (gene introns or 0–1kb flanking regions) were extracted and annotated using a blastx search against NCBI reference protein database with *CLC Genomics Workbench* tools. Gene networks were formed from the GO annotations of genes that contain similar TE. Network edges represent connections of GO terms, but the node size depends on the gene count categorized to a particular term. A hypergeometric test with Bonferroni correction implemented in BINGO was used for the GO term overrepresentation test in one particular network compared to custom GO annotation of all pine genes. Comparative networks were made using *DyNet* (Goenawan et al., 2016). *GenAlEx* 6.05 software package (Peakall and Smouse, 2012) was used for TE insertions genotype analysis. Identified protein interactions were analyzed using *STRING v.11 (Szklarczyk et al., 2019). A graphic overview of analysis workflow is provided in Additional file 1.*

Results

Quality of assembled genomes and TE assay

Data analysis demonstrated that the quality of reference genomes and the repeat database used (full-length vs short repeats) play a major role when analyzing the presence of TE in gene regions. Analysis of the first version of the P. taeda genome with the full-length TE element database identified more genes containing TE sequences in introns (the top-ranked TE had unique matches to 200 genes) than in the P. taeda v.2.0 genome assembly (the top-ranked TE had matches to only 39 genes). A similar situation was observed when analyzing the 5-kb flanking regions using the full-length TE database. However, when the short repeat database was used for the analyses, the number of genes with repeats in the introns or flanking regions was similar in both P. taeda genome versions. The main reason for these differences may be because the repeat database was generated using the first genome version, but for P. taeda v.2.0, assembly longer reads were used and TEs were not predicted de novo. Therefore, full-length elements identified from v.1.0 did not align over their entire length with the v.2.0 genomic sequence but align over shorter regions. Additionally, observation of particular gene introns with similarity to interspersed repeats, but not to full-length elements, revealed that incomplete TE sequences in v.2.0 are frequently present with masked remaining sequences nearby. This suggests persistent problems with TE-containing reads assembly, even from longer reads, and indicates the need for validation of observed matches to confirm the presence of full-length or partial TE presence in gene introns or flanks. Due to the diversity and repetitive nature of conifer TEs, the differentiation of full-length elements from partial, chimeric, or nested copies is often not possible without resequencing. Therefore, utilizing shorter repeats as representatives was more suitable at this point for evaluation of prevalent TE in or near genes. The presence of any part of a TE could be indicative of the occurred insertion, and the partial or full-length structure could be confirmed further in additional experiments. For P. lambertiana, only one genome version is currently available and the most frequent TE matched 272 genes, which is comparable with the P. taeda v.1.01 genome. For the most frequent TE-derived repeat, approximately 500 matching genes were found in P. taeda v.1.01 and v2.0, while 532 genes were identified in the P. lambertiana genome. These results are in accordance with whole-genome TE content data, where the larger genome size of the pine subgenus Strobus (P. lambertiana) correlates with a higher gene and repetitive element content (Stevens et al., 2016).

Previously annotated pine RLXs were used to understand the common distribution and quality of genome assembly in gene regions. The ratio of LTR to internal sequences of RLXs should reach a value of 2 if LTR RLX are present as full-length elements (Additional file 2). However, incomplete sequences of the ancient and widely distributed IFG family were found in gene introns of P. lambertiana, with more internal sequences present than LTRs. In contrast, in the P. taeda v.2.0 genome, IFG solo LTRs prevailed. More rapid degradation of LTR sequences is common for RLXs in plant genomes; however, this process is relatively slow in conifers according to previous studies. Other common TEs, like Pinewoods and

Appalachian, were represented in both species' genomes as full-length sequences. The low-copy-number RLX Angelina was overrepresented by single LTRs in P. taeda introns; 90% similarity and only one complete internal sequence and 13 LTRs were found. At 80% sequence similarity, more solo LTRs were identified, which could be a result of fragmentation of RLX, structural differences, or degradation of some sequences in gene introns. The patterns of the most widely distributed TEs in the gene introns and flanking regions of both species were different (Additional file 3). These results indicate that the TE composition of the two pine species is variable and structural inconsistencies are present in the current assemblies, not only across the entire genome, but also in the functional regions of genes.

We aimed to identify TEs common to both genomes in an effort to transfer the obtained results to the non- model species P. sylvestris, which has limited genome sequence data available. More than 20 highly distributed TEs were initially identified. Genes containing identical TEs in the introns or in flanks were compared, but in most cases no homologous gene transcripts between P. taeda and P. lambertiana were found. Seven TEs distributed in the gene regions of both species were analyzed in more detail and only the structurally most well defined and informative are described separately in the next sections. TEs found distributed in gene regions of both species included DNA transposons DTX184 and MITE3321, Copia LTR RLX (Copia-930, Copia-1813, Copia-2602, Copia-25, Copia-17), and two RLX of the Gypsy superfamily (IFG7a, Appalachian). Evidence of the presence of other TEs in gene flanks and introns is also found from TE-derived repeat analysis, however, the fragmentation of the genome does not allow these data to be confirmed conclusively.

Gene Ontology analysis

GO analysis was performed for the extracted gene groups that had similar TEs insertions in gene introns or flanks. Network edges represent connections of GO terms, but the node size depends on the gene count categorized to a particular term, with approximately 50% of all pine genes not annotated to any GO category. Therefore, GO analysis was used only for the overall recognition of potentially important gene networks.

For flanking regions, GO analysis was done for genes with statistically higher frequency of TE-derived repeats (Additional file 4). Some members of the identified gene networks contained TE insertions in their vicinity with the highest t-significance of enrichment, but had few associated GO terms and were poorly annotated. TE-containing transcripts were annotated with GO terms such as DNA integration, RNA-dependent DNA biosynthesis, endopeptidase activity, or nucleic acid binding. Some gene groups contained only a few annotated genes, and their significance could not be established with certainty.

A higher diversity of TEs were identified within gene introns. Gene networks were built for genes having similar TE insertions in their introns without differentiation of exact location regarding intron number and position (Additional file 5). GO categories from biological processes and molecular function were determined and overall hits to unique genes and GO-annotated gene count in category were indicated. In the DyNet comparative trees, common GO categories are indicated in white (Additional file 5). Genes from the two species containing identical TEs were categorized into similar parental GO categories. However, no homologous genes were revealed in most of the analysed networks between pine species. This demonstrates that gene regions of pine species belonging to different subgenera contain different insertions of TEs. Revealed gene networks were often associated with defense and regulative responses, such as oxidation-reduction processes, transmembrane receptor biosynthesis, metal ion binding, hormone metabolic processes, and carbohydrate metabolic processes.

Topology of TEs in gene flanks

The P. taeda v.2.0 genome currently has the best scaffold quality regarding TE positioning. To analyze the distribution of TEs relative to distance from genes, the 5-kb flanking regions of P. taeda v.2.0 genome were divided into 1-kb regions (0–1; 1–2; 2–3; 3–4, and 4–5 kb from genes) and the position (5' or 3') relative to each gene was noted. For each repeat, t-statistics was calculated comparing average unique hit numbers to the 0–2 kb and 2–5 kb and to the 0–1 kb and 2–5 kb flanking regions; 135 TE-derived repeats present in gene-flanking regions were revealed. In total, each repeat had sequence similarity with 21 to 266 unique gene-flanking regions and occupancies in the 0-1 kb flanking regions of nine RLX families (p=0.001) and seven RLX families (p=0.05) were significantly increased. Significantly overrepresented repeat families were found in the vicinity of 27–48 genes (0–1 kb 5' or 3' region), with occupancies of two RLXs (p=0.001) and five RLXs (p=0.05) significantly increased in 0–2 kb flanking regions compared with the remaining regions (Additional file 4). False genes were identified and evaluated in subsequent detailed sequence analyses. These false gene transcripts consisted of several parts of different TEs that were not present in databases and therefore could not be filtered out in previous stages. Therefore, fewer protein-coding genes actually contained TEs in their vicinity than initially identified, and all gene groups were repeatedly screened against the NCBI database using a blastx search to confirm protein coding.

An overview of the analyses of 5' and 3' gene-flanking regions between the two pine species is presented in ***Table 1****. The ratio of hit number to the number of extracted flanking regions was similar for all regions in the P. taeda v.2.0 genome (0.1–0.11, except for the 0–1 kb region with ratio 0.16–0.18). After in-depth analysis of the P. taeda v.2.0 genome, global errors in gene coordinates were revealed such that many*

protein-coding gene transcript coordinates were mapped to TE genomic coordinates without any sequence similarity. This increased the number of LTRs found in the 0–1 kb flanks of false genes. After creating our own genome annotation, similar tendencies were revealed in the TE distribution in all studied genomes, namely an increase of hits to TE-derived repeats with increasing distance from genes.

*In the P. taeda v.1.01 and P. lambertiana v.1.01 genomes, all gene sets, especially high-quality gene sets, contained almost no TE-derived repeats in the 0–1 kb regions. Furthermore, the number of TE-derived repeats gradually increased with distance from genes, suggesting a slight elimination of TEs from gene regions or results of insertion preference. One of the aims was to identify and characterize common TEs associated with genes in pine species for the subsequent analysis of non-model species. Interspersed repeats that were identified as associated with genes in both the P. taeda and P. lambertiana genomes were manually reviewed to verify their full-length structure and proximity to genes. The P. taeda v.1.01 genome and P. lambertiana gene-flanking regions were highly enriched with only one repeat; this was later identified as the MITE3321 element (**Figure 1**). Additionally, there was evidence of the distribution of Copia RLX RYX6 LTRs family in a smaller P. taeda gene set. RLX short insertions were found in the 0–2 kb flanks of P. taeda. This element had 91% similarity to the PpRT6 partial RNaseH-like gene (EF102091) previously reported for Pinus pinaster (Miguel et al., 2008). The full-length element (3,367 bp) was present in only one gene flank while others contained single LTRs; most sequences between LTRs were masked.*

The identified potential LTR was rich in AT-containing motifs (AT-1, ARR1AT, CAATBOX1, MYB1AT, CIACADIANLELHC, MYCCONSENSUSAT, ROOTMOTIFTAPOX1, P1BS, "TATCCA" element, and others). The extracted 416-bp LTRs contained conventional 5'TG-CA-3'dinucleotides, TATA box (S000203), and a polypurine tract. The absence of full-length coverage of RYX6 due to masked regions prevented determination of a consensus sequence and verification of the results, indicating that sequence scaffolding problems persist in the case of longer repetitive elements. Insertions of less common TE-derived repeats in close vicinity from genes were also revealed.

The 5,767-bp RLX Copia-2602 with 160-bp LTRs was found in gene flanks of P. taeda 2.0; the hit count to the newly identified RLX body increased with distance from the gene (from 4–6 in 0–1kb to 28–36 in 4–5 kb). The relatively short LTR of this RLX contains three TATA box motifs on the negative strand and one on the positive strand and three CAAT boxes on the negative strand. Hits to the 3' flanking regions of the genes were found more frequently; insertion of this element could therefore promote antisense transcription. Only two P. lambertiana genes possessed a Copia-2602 insertion in 0–1 kb flanks, these were found in the 5' flank of the probable histone H2B.3 gene and in the 3' flank Piwi-like domain containing the argonaute family member. No genes containing LTR of Copia 2602 were found in the 5' 0–1 kb flanks of the P. taeda 2.0 genome, but two 5' flanks contained hits to the Copia-2602 body (serine/threonine protein kinase PEPKR2 and carnitine transporter 4). Hits of the LTRs to two 3' flanks were found in the vicinity of TMV resistance protein N-like and one unknown gene. In the 1–2 kb region, hits to the LTR were identified in the 5' flanks of flavonoid 3'5'- hydroxylase 2 gene and in the probable

disease resistance protein At4g33300 and in the 3' flanks of chitinase 2-like and cytochrome P450 CYP736A12 genes.

MITE3321 distribution in gene vicinity and introns

*The PtRXX_3321 TE-derived repeat was initially found as highly represented (found in 57 filtered genes) in P. taeda v.2.0 introns. Similarly, this repeat was the only highly represented sequence in the taeda v.1.0 gene flanks. In-depth analysis of the sequences obtained from the P. taeda v.2.0 genome revealed a MITE element that was 259 bp in length and contained 24-bp inverted repeats. The stem sequence was flanked with 40-bp direct repeats and was first identified in the probable pectinesterase/pectinesterase inhibitor gene first intron. A consensus sequence of the MITE3321 transposon from the P. lambertiana genome was built (74% similar to P.taeda MITE3321). Compared with P. taeda, the P. lambertiana MITE3321 contained a 10-bp insertion in the 3' stem sequence and was 265 bp in length with 28-bp TIRs with two mismatches. Newly revealed stem structures were used for an additional search within introns and gene flanks. The P. taeda MITE3321 element was found in 74 genes (sharing more than 80% sequence similarity), from which 58 unique genes contained MITE3321 insertions with 99–100% nucleotide identity (**Figure 2**). The P. lambertiana high-quality gene set contained 87 genes with MITE3321 within introns, with 53 genes containing the MITE insert with 99–100% query coverage; however, the diversity of these sequences was higher (most sequences share 94–96% nucleotide similarity with the consensus sequence). The MITE3321 element was also significantly propagated in the vicinity of P. taeda genes, with enrichment of 0–1 kb and 0–2 kb flanks being statistically significant for both species (p=0.001). In P. taeda, 191 MITE3321 insertions in the 0–1 kb gene flanks had on average 90% sequence identity. MITE3321 was inserted in 65 HQ genes in the 0–1 kb flanking regions in P. lambertiana. Some gene introns contained several MITE3321 insertions (**Table 2**). An unannotated P. taeda gene with a conserved phosphoglucosamine mutase family protein domain contained a maximum of seven MITE3321s at different locations, a subtilisin- like protease SBT5.3 coding gene carried four MITE3321s, a metal tolerance protein 11 coding gene contained three insertions, and eight other genes contained two insertions each. Some insertions were localized close to each other, as indicated by the genomic coordinates.*

Genes carrying MITE3321 in introns share GO categories such as glycotransferase activity; the P. lambertiana gene had 1.3-beta-D-glucan synthetase activity, but the P. taeda gene had xyloglycotransferase activity (Additional file 5e). Common GO categories included regulation of gene expression, response to stress, and transmembrane transport. However, gene nucleotide sequence comparisons revealed no homologous genes between the evaluated P. taeda and P. lambertiana transcripts containing similar MITE3321 insertions in genomic sequences. This, as well as the revealed species-specific structural differences, suggests that MITE3321 occurred in a common ancestor, but expansion of this element occurred after separation of species. A comparison of genes containing

*MITE3321 within introns with genes having identical MITE in their flanks did not reveal any common genes in either species, indicating that MITE3321 was inserted only into introns or in flanks of certain genes, but never in both sites of presumably transcriptionally active genes (**Figure 2**). Therefore, insertion of MITE3321 could not only be explained by random transposition into transcriptionally active chromatin. Better-annotated P. taeda genes containing MITE3321 in introns or in flanks were compared using GO terms (Additional file 6 b), revealing that genes containing MITE3321 in their flanks were involved in the regulation of developmental processes, such as regulation of cell division, pollination, negative regulation of macromolecule biosynthetic processes, nuclear division, and methylation. Genes containing MITE3321 in their introns were involved in innate immune response, positive regulation of defense responses (jasmonic acid-related responses), pigment metabolic processes, plastid organization, maturation of ribosomal proteins, potassium ion transmembrane transport, and proline biosynthetic processes. The frequency of GO terms involved in regulation was increased for the MITE3321 insertions into gene flanks, while genes containing MITE3321 in their introns were more often related to response (**Figure 2 A, B**).*

Both networks contained GO terms related to ion homeostasis, glucan metabolic processes, proteolysis, regulation of gene expression, post-embryonic development, and oxidation-reduction processes (Additional file 6b).

*MITE3321 elements from P. lambertiana and P. taeda contain one TATABOX on the positive strand, 7–10 ARR1-binding elements, 2–4 CAAT boxes, 4–5 DOF TFBS, 3 GT-1 binding sites; these are all important regulative motifs found in the promoter regions of plant genes. The P. lambertiana MITE contained a 10-bp insertion (AGAGAAATTA) that disrupted a site (TTTGACC) identical to several WRKY TFBS, but gained a site identical to co-dependent regulatory elements responsible for pollen- specific activation (**Figure 2C**). Differences in predicted TFBS presence in MITE3321 could explain the depletion of this TE in the P. lambertiana gene 0–1 kB flanks and enhanced distribution in gene introns.*

DNA transposon DTX184 forms a stress-responsive gene network

An 820-bp TE-derived repeat was initially found as moderately distributed within P. taeda v.1.0 gene introns and many genes contained extended GO annotations belonging to defense responses (Additional file 5 a-b). In the entire P. taeda v.2.0 genome, only 251 copies of DTX184 were found, indicating preferential distribution of this TE in genes. One of the identified genes was Nonexpresser of Pathogenesis-related proteins-1 (NPR1), which is involved in plant systemic acquired resistance and the salicylic acid-mediated signaling pathway. NPR1 contained several TE-derived repeats in the second intron and all repeats were tested with additional searches, but only DTX184 was distributed and found in an additional 200 gene introns from the non-filtered P. taeda v.2.0 genes and formed a stress-responsive

*gene network (**Figure 3**). Other identified genes included a histone- binding PHD1 finger protein ALFIN-like 4 coding gene, a COPII-coated ER to Golgi transport vehicle SNARE-like 13 gene, eukaryotic translation initiation complex 2B, ribosome biogenesis protein RPF1, and other important genes (Additional file 7 a-b).*

As the remaining intron was masked, only a partial sequence of the repeat was present and therefore a full-length structure insertion was not revealed for all genes. However, a longer repeated structure was isolated (1,978 bp) and a conserved domain search (Marchler-Bauer et al., 2017) revealed transposase-like protein (pfam05699), hAT family C-terminal dimerization region (pfam05699), and BED zinc finger domain (pfam02892), indicating that this was a DNA transposon. The shorter analyzed DTX184 repeat (820 bp) contained a TE dimerization region and the FindMiRNA tool (http://www.softberry.com/) predicted seven probable pre-microRNAs with free energy ranging from -53.75 to -45.82 kcal/mol. A search of MiRBase revealed homology to the mature microRNA sly-miR9472-3p from a drought tolerant tomato line (Candar-Cakir et al., 2016; Liu et al., 2017). The P. lambertiana genome contains more hits to the newly isolated DTX184 element; 143 genes contained hits longer than 1 kb with sequence identity >82%. DTX184 consensus sequences from both species were compared, but no species-specific structural polymorphisms were identified, suggesting ancient transposition events and probable distribution in genes of other pine species. Gene transcripts from P. taeda and P. lambertiana containing similar TE were compared and only one homologous gene was identified with more than 95% sequence identity and 100% query coverage. This gene was annotated as 26S proteasome non-ATPase regulatory subunit 4. DTX184 was not identified in the 0–1 kb gene vicinity of both species.

Distribution of the widespread IFG Gypsy RLX

The IFG RLX is a remarkable TE family as it is highly distributed in conifer genomes and is far more ancient than other RLXs but sequence homology is still maintained (Kossack and Kinlaw, 1999; Voronova et al., 2017). All matches in gene introns (LTR, internal sequence of IFG, or both) were considered in gene network analyses, resulting in 99 genes from the filtered P. taeda v.2.0 dataset and 317 genes from HQ genes in P. lambertiana (Additional file 7 m,o). The following three homologous protein kinase genes with IFG insertions were identified: plastidial pyruvate kinase coding gene, PTI1-like tyrosine protein kinase gene, and putative receptor-like protein kinase gene. The P. lambertiana tyrosine protein kinase had an 85,239-bp second intron with one single IFG insertion (internal part with 3' attached one matching LTR) and 7,987 bp in total from intron II was masked. In the P. taeda v.2.0 tyrosine protein kinase gene, IFG was inserted into the first intron, which is 106,013 bp long and only 1,639 bp was masked. Both protein kinase genes shared an identical exon-intron structure and 97% cDNA similarity. While IFG LTRs were 82% similar, the IFG body was interrupted by sequence masking. A receptor-like protein kinase gene contained exon duplication events in both species, with both introns containing one full-length IFG insertion on the minus strand. These insertions were 87% similar in the IFG body and 79–83% similar in their LTRs. The P.

taeda v 2.0 gene contained an additional IFG sequence with one attached LTR on the positive strand, two IFG on the positive strand, and four nested LTRs with surrounding masked regions. The lambertiana receptor-like protein kinase gene contained one full-length IFG and one single LTR on the positive strand.

IFG insertions were also found in gene flanks (Additional file 7 n,p), but RLX frequency tended to increase with distance from genes in both genomes. The IFG LTR insertion was in close vicinity to 14 filtered P. taeda genes (0–1 kb) in the 5' flanks. Genes encoding sugar transport protein 7 contained only the IFG body in the 5' 0–1 kb flank. In the 3'gene flanks, 12 genes contained IFG insertions, but only two annotated genes contained LTRs (abietadienol/abietadienal oxidase and protein DMR6-LIKE OXYGENASE 2-like). Similarly, in P. lambertiana, 18 genes contained IFG insertions in close vicinity to genes. LTRs were inserted in the 5' flanks of two annotated genes coding for DNA-directed RNA polymerase II second largest chain and serine/threonine protein kinase HT1 (not homologous to that previously found with an insertion in the intron). No genes were found that contained insertions of IFG in both the introns and flanking regions (5'and 3' 0–2 kb) in the P. lambertiana or P. taeda v2.0 genomes.

Copia-1813 RLX resides gene introns and flanks

A newly identified RLX was found frequently distributed within gene introns in the P. lambertiana and P. taeda genomes. The RLX isolated from a BAC clone was 6,884 bp long with two large 1,386 bp LTRs, starting with conventional TG and ending with CA sites. The RLX internal region was weakly similar at the protein level to Retrovirus-related Pol polyprotein from tobacco tnt1 RLX (39% identity, 59% positives). The PPT (AAGAGGGAG) site was identical for elements from P. taeda and P. lambertiana. Some inserts in other locations probably have shorter LTRs, but due to multiple masked regions, it was not possible to extract copies for the consensus structure. The P. lambertiana Photosystem II stability/assembly factor HCF136 coding gene contained a Copia-1813 RLX in the first intron in the positive orientation; the LTR was similar to a RLX from P. taeda with 34% query coverage and 82% sequence identity, and the RLX body was 89% similar with 62% coverage. However, some parts of this RLX were masked. The P. taeda LTR contained the following two AG-rich tracts: (AGNN)3(AG)3(NNAG)2 and (AGNN)2(AGN)4. The P. lambertiana LTR also contained polypurine-rich motifs 25 bp apart: AA(AGG)2A3(AGG)2GA3AGG and GAG(AGG)3AGA(AG)3. The P. taeda filtered gene set contained 80 strong hits to genes with Copia-1813 RLX in the introns (hits >1 kb to body and presence of at least one LTR). The P. lambertiana HQ gene set contained 688 genes with similar hit parameters, 243 genes were not annotated, and 67 were uninformative. However, a search in the NCBI blastx protein database enabled annotation of additional proteins (Additional file 7 c-f). For the network analyses, we considered only >4 kb hits to Copia-1813 RLX and strong hits to LTRs; 116 genes carrying such a combination of hits were found in P. lambertiana. There were 19 P. lambertiana HQ genes containing the Copia-1813 insertion in their flanks. Only the following two P. taeda genes contained Copia-1813 RLX in 0–1 kb flanks: cytochrome P450 78A7-like and secreted RxLR effector protein 161-like.

High TE diversity inside gene introns was observed in both pine species, and the presence of intron TE insertions could form gene networks with similar expression and response patterns. If each interspersed repeat introduced additional gene regulation signals in gene introns, then genes containing a single interspersed repeat might show specificity regarding their function. To test this assumption, genes that contained an insertion of the Copia-1813 RLX family (and lacks other TE insertions) were analyzed (Additional file 8). Five of 19 P. lambertiana proteins were found to have

homologs in the A. thaliana genome by STRING software and were connected to mitotic spindle assembly checkpoint (coexpressed, found interacting and mentioned together in other publications).

TE patterns embedded in gene introns

The evaluated Copia-1813 RLX network genes were further analyzed regarding all identified intronic TE insertions. The presence of unique short TE-derived repeats (95 loci) and unique >1-kb hits (22 loci) were used to form TE insertion patterns (suggested genotypes) present in introns. Intronic TE patterns common to P. taeda and P. lambertiana were identified using the GenAlEx

*6.05 software package Matching Multilocus Genotypes test and Principal Coordinates Analysis (**Figure 4**). Presence of Copia-1813 RLX and DTX184 TE was found within introns of seven P. lambertiana genes (Pattern type I, Additional file 7-c). Products of these genes were found in different cell compartments and are involved in protein folding in ER (oxidation), positive regulation of RNA export from the nucleus, protein heterodimerization, and SYM-1 stress responsive protein from yeast; the function of this protein is not yet described in plants. Two genes contained the Copia-2602 RLX in addition to Copia-1813 and DTX184. The P. lambertiana gene was involved in glycoprotein formation, but the P. taeda SFGH gene is involved in detoxification of formaldehyde. Several genes with the TE pattern type P (Copia-1813+Copia-2602+Copia-25) are involved in pH regulation in Golgi, tethering of vesicles to Golgi membranes, nuclear protein import, and intracellular protein transport. TE pattern type Y (Copia-1813+Copia-2602+Copia-25+IFG) was found in the following two genes: insulinase (involved in protein targeting to mitochondrion) and histone deacetylase 15 (tag for epigenetic repression). If short TE-derived repeats were considered (95 loci), a low number of matching genotypes were revealed. One specific intron pattern type C was revealed for three P. lambertiana genes: two were annotated as splicing factor 3A subunit 3 genes and one as a cleavage and polyadenylation specificity factor subunit 5- like coding gene. In addition to Copia-1813 RLX, these genes contained two additional TE-derived repeats and products of these genes are involved in pre-mRNA maturation and splicing according to the UniProt Knowledgebase. Genes are coexpressed (score 0.168) according to the STRING database. While these analyses provide some initial clues, it is not currently possible to identify probable connections and roles of intron TE pattern genotypes, as gene annotations and expression data for pine species are scarce.*

GC content was calculated for TEs and TE-derived repeats (Additional file 9). The TE-associated GC content for each gene involved in the Copia-1813 RLX network was evaluated (Additional file 7 c, e). However, GC content for full-length introns could be biased due to masked TE parts in the gene introns and flanks. Patterns of identified TEs were considered and GC content of each full- length TE was counted for each gene introns. The overall average GC content for introns considering 1-kb hits was 39% for P. taeda and 41% for P. lambertiana, respectively, which was even lower if only short hits to LTRs were considered (27% and 36% respectively). This could be explained by the lower GC content of LTRs of some RLXs distributed in genes; in addition, the database of conifer-interspersed repeats used for searches contained more sequences specific to the P. taeda genome. The mean GC content of the gene transcripts was 44% for P. lambertiana and for P. taeda, which was higher than any average estimate for introns and published estimates for whole BAC clones and whole genome sequences of gymnosperms (38%, (Gonzalez-Ibeas et al., 2016; Perera et al., 2018)).

Genes appearing in many TE-connected networks

Genes with broad GO annotations were frequently found across many predicted TE networks related to the TE distribution within introns of both species. Therefore, genes containing multiple TEs in their introns were isolated and analyzed. In the P. taeda v.2.0 genome, 75 genes were identified containing 8–65 unique TE-derived repeats in the introns, with the two-pore potassium channel coding gene containing the largest number of different TE-derived repeats (65). According to the PLAZA Gymnosperm database, P. taeda contains 14 members in two-pore potassium channel gene family, which contains 659 co-occurring terms, indicating the involvement of these gene products in many plant-cell processes (Huntley et al., 2009). Other repeat-reach genes identified in P. taeda v.2.0 were chloroplastic/amyloplastic 1,4-alpha-glucan-branching enzyme coding gene, GTP binding protein Der, S-formylglutathione hydrolase, cytochrome P450, B3 domain-containing transcription repressor VAL2, serine/threonine protein kinase GRIK1, ribosome production factor 1, jasmonic acid-amido synthetase JAR1, COP9 signalosome complex subunit, chaperone protein ClpB3, and homeobox-leucine zipper protein ATHB-15 (Additional file 10 a).

In the P. lambertiana HQ gene set, 59 genes containing 21–34 unique TE-derived repeats were analyzed. One gene that contained 34 TEs in introns was not annotated, but could be characterized as having with SpoT (COG0317i) and ubiquitin-like fold (cl28922) conserved domains. A search of the Uniprot database (The UniProt Consortium, 2019) with this gene identified an HD domain- containing protein with 63% identity and 75% positives e-value 0.0 that is involved in guanosine tetraphosphate metabolic processes (GO:0015969) with 107 co-occurring terms. A gene containing the second largest number of TEs in P. lambertiana was also an unknown protein containing the AMN1 domain (cl2816) and F-box domain

(pfam12937), annotated only with the parent term "protein binding". Other genes with a high amount of TEs in the P. lambertiana genome were DNA repair helicase XPD, mitochondrial substrate carrier family protein C, chloroplastic isoform of imidazole glycerol phosphate synthase hisHF, translation initiation factor elF-2B, syntaxin-81, ADB-ribosylation factor, chloroplastic stromal processing peptidase, and acyl-CoA dehydrogenase IBR3 (Additional file 10 b).

*Eight similar genes were identified in both P. taeda and P. lambertiana (**Table 3**), from which the following two gene homologs were found: plastidial pyruvate kinase 2 and phospholipid:diacylglycerol acyltransferase genes. Pyruvate kinase is involved in carbohydrate degradation and is associated with 26 GO terms. P. taeda and P. lambertiana introns contain seven common TE-derived repeats in pyruvate kinase homologous genes; O-acyltransferase activity (GO:0008374) contains 329 co-occurring terms, 12 TEs were similar in gene homologs. SrmB conserved domains were present in ATP-dependent RNA helicases, but these genes were not homologous. ATP-dependent helicase activity (GO:0003724) contained 929 co-occurring terms. A similar situation was found with proteins containing conserved domains of chromosome segregation protein SMC (cl37069) from nuclear pore complex protein NUP62, WD40 domain from protein WRAP73, RAE1 and actin-related proteins (found in a number of eukaryotic proteins that cover a wide variety of functions), alpha-tubulin suppressor (ATS1) domain-containing proteins (cl34932), and mitochondrial carrier protein domain (pfam14560, involved in localization, transmembrane transport, amide biosynthetic process, and translation).*

Discussion

Automated TE detection relies on several strategies, such as searches for sequences with homology to known elements, de novo evaluation of repeated elements, analysis of the presence of specific structural features, and combined techniques (Bergman and Quesneville, 2007). In the genome assembly process, repeated and highly similar sequences like TEs are the source of errors and gaps (Treangen and Salzberg, 2011; Tørresen et al., 2019). Properties of plant genomes, such as multiple gene families, pseudogenes, and chromosomal and plastid genome duplications, further complicate the process of genome assembly and annotation (Claros et al., 2012; Goodwin et al., 2016). Typically, gymnosperms are characterized by large genomes with proliferated TEs, high levels of heterozygosity, a constant chromosome number, and very rare polyploidy events (Morse et al., 2009; Pellicer et al., 2018). Several conifer genomes have been sequenced using short-read assembly methods (Nystedt et al., 2013; Wegrzyn et al., 2014; Guan et al., 2016; Stevens et al., 2016); the P. taeda v.2.0 genome currently has the highest quality conifer genome regarding scaffold length. Version 2.0 was improved by merging small contigs; while the first version of the P. taeda genome contained 16.5 million contigs, the second version contains only 2.9 million (Zimin et al., 2017). However, the average PacBio read length used to reconstruct genome v.2.0 was 9,665 bp with

12x coverage, which is shorter than many RLXs. The mean TE length in the PIER database was 6,273 bp (median 5,383 bp) but 3,046 TEs were longer than 10 Kbp. With the availability of better quality or longer-read genome assemblies, TE identification should be repeated de novo. Genome annotation of conifers is an ongoing process. Indeed, our study revealed errors in annotation files and datasets, where gene IDs of transcripts did not coincide with genomic sequences bearing identical gene IDs, and genes had noticeably varying intron lengths in different genome versions. Some additional errors involve automated annotation, nested repeat annotation, defined pools of identical reverse transcriptase domains as different genes, and intense masking of ambiguous regions, which are not discussed sufficiently in associated publications. Consequently, comparison of genomes assembled with different approaches and of varying quality should be performed with caution to avoid errors in interpreting the results. Additionally, widespread gene capture by TEs has often been described for large plant genomes (Takahashi et al., 1999; Zabala and Vodkin, 2005; Kalendar et al., 2008; Wei et al., 2009) and this phenomenon could introduce additional errors in short-read genome assemblies (Wei et al., 2009). Considering the high copy number and the structural and functional differences from protein-coding genes, it is advisable to annotate TE- associated sequences in separate data sets even though many TE families bear ORFs. Despite the fact that many conifer TEs only show weak similarity to annotated TE domains, complexity was effectively reduced in the P. lambertiana high-quality gene set (Crepeau et al., 2017). Evaluation of the only prevalent MITE element in the flanking regions of initial genome versions demonstrates the importance of read length in assembly and the association of TEs to particular genome locations.

Diverse conifer RLXs sharing high partial sequence similarities may not be classified in one family by full-length alignments used in bioinformatic studies, resulting in inflated family counts (Wegrzyn et al., 2014). Due to the high diversity of conifer TE sequences and patchy distribution within short read assemblies, we found that identification of short repeated regions could overcome these problems and more efficiently identify prevalent TE families. The use of clustered TE-derived repeats extracted from automatically predicted nested regions allowed not only for the identification of RLX, but also internal regions of other TEs. The use of TE-derived repeats allowed simple statistical tests of distribution relative to distance from genes. RLXs are the most widely distributed TE class in plants and gymnosperm genomes (Kumar and Bennetzen, 1999; Nystedt et al., 2013; Wegrzyn et al., 2014; Galindo-González et al., 2017). LTRs contain important regulatory signals that can influence gene expression even if the body of the element is deleted (Casacuberta and Grandbastien, 1993; Takeda et al., 1999; Butelli et al., 2012). The PIER v.2. database entries contain nested TEs from automated predictions, including repeated internal regions of different TEs. Chimeric TEs that could evolve from two RLXs by template switching (Sabot and Schulman, 2007) and display signals of independent transposition have been identified in other plant genomes, e.g. Veju (Sabot et al., 2005) and BARE-2 (Vicient et al., 2005). TE structures with several LTRs are found in plant genomes, for example, 13 Cassandra elements in pear contain 3 LTRs (Yin et al., 2013). Considering the large conifer genome sizes, such chimeric structures may also be found in conifer genomes. Therefore, each high-frequency repeat should be verified and the true full-length structure and

TSDs should be identified. Unfortunately, the pine genomes contain many low-copy-number TEs, and if all copies are masked, full-length structures or TSDs cannot be resolved. Additionally, similarity to known proteins and GO annotations are available for only approximately half of all pine genes, but species-specific genes could be more important in adaptation of species. Therefore, the data evaluated and the results obtained from this study could be expanded with improvement of genome quality, annotation, or other associated information.

In the current study, we identified several TE families that unite many pine stress-responsive genes and contain potentially important gene regulatory signals. The DTX184 DNA TE could provide microRNA target sites or produce microRNA, or both. However, further analysis is required to determine if these sequences represent microRNA precursors or, alternatively target sites (Ambros et al., 2003). This DNA TE insertion was found in the second intron of the P. taeda NPR1 gene, a key regulator of systemic acquired resistance (MAIER et al., 2011), the PSMD4 gene coding for the 26S proteasome (involved in the ATP-dependent degradation of ubiquitinated proteins), ubiquitin thioesterase OTU1 (plays an important regulatory role at the level of protein turnover by preventing degradation), S- formylglutathione hydrolase (detoxifies formaldehyde), a PHD finger protein ALFIN-LIKE 4 (a histone-binding component that recognizes H3K4me3), and other important genes. Based on the presence of short consensus sequences of the DTX184, approximately 200 genes could be involved in this network in the P. taeda genome. In P. lambertiana, the similarity of DTX184 with 143 high-quality genes with high coverage was found, many of which are important genes coding for transcription factors, chromatin modification enzymes, protein kinases, and receptors. The short MITE3321 family identified in proximal gene flanking regions and introns could provide TATA boxes and several ARR1, DOF, W-box, GT and MYB-binding sites, which are important signals in plant transcription activation and stress-response regulation (Yanagisawa, 2004; Eulgem and Somssich, 2007; Taniguchi et al., 2007). Ten different ARR1 binding sites (7 on positive and 3 on negative strand) were present in the consensus sequence of P.taeda MITE3321 element. The transcription factor-type response regulator ARR1 directs transcriptional activation of the ARR6 gene, which responds to cytokinins without de novo protein synthesis (Sakai et al., 2001). Cytokinins are an important class of phytohormones that are involved developmental processes and growth (Mok, 1994; Mok et al., 2000; Hwang et al., 2012), as well as in defense responses (Argueso et al., 2012). The (AG)4A motif is one of the most common TFBS for plant promoters (Liu et al., 2013), that regulate light-responsive phototransduction processes in plants (Parida et al., 2009). The Copia-1813 identified in this study contain longer AG-rich tracts, this RLX family is distributed in pine gene introns and several flanks and might form responsive gene network. Genome-wide approaches in mammalian genomes demonstrate that TEs contribute to rewiring and selection of gene networks (Feschotte et al., 2002; Sundaram et al., 2014). Approximately one-sixth of rice genes are associated with TEs (Krom et al., 2008). Comparative analysis of three inbred maize lines revealed that the expression of 33% of stress-responsive genes could be attributed to regulation by TEs (Makarevitch et al., 2015). Various maize TEs contain approximately 25% of all DNase I hypersensitivity sites within the genome, which are associated with open chromatin and cis- acting elements and are therefore essential transcriptional regulators (Zhao et

al., 2018a). In contrast to mammals, plant genes more commonly contain longer introns, which are expressed at higher levels (Ren et al., 2006; Colinas et al., 2008). However, in recent studies, expression in different tissues or expression breadth has also been considered and plant genes expressed across a wide range of tissues or conditions were found to have a higher intron density (Camiolo et al., 2009; Das and Bansal, 2019). In P. taeda and P. lambertiana, extensive transcriptome data is not currently available, therefore the role of TE insertions within introns in regulation of gene expression requires additional investigations. Using the Copia-1813 RLX gene network identified in this study, formation of specific patterns or intronic genotypes was tested. TE patterns could influence gene availability, responsiveness, stability, or higher order organization structures in the nucleus. Revealed genes with identical TE insertion patterns are involved in pre-mRNA maturation and splicing, while other genes with identical TE insertion genotypes are linked to protein metabolic processes and Golgi body homeostasis. Further investigation will enable more thorough analyses of relevance. The function of genes where multiple TEs were identified within introns (e.g. potassium channel coding genes and other receptors, protein kinases, cytochrome genes) suggests involvement in the maintenance of cell homeostasis under stress conditions. These genes were found to have many co-occurring GO terms, indicating that gene products are involved in many cellular processes and these genes may be expressed or retain stability in a broad range of conditions. We suggest that genes with many different types of TEs could act as node genes that are functional or stable across a range of conditions and could be important in early defense responses and rapid metabolome switching. This is also supported by the discovery of several homologous genes with large introns in both pine species. Cases of independent transpositions of different RLX families into homologous genes in varieties of differing origin resulting in similar phenotypes have been reported (Butelli et al., 2012). Additionally, the TE insertion patterns in investigated pine introns were found to have lower average GC content (39%) than nearby transcripts. The GC content of gene transcripts in the studied genes in P. taeda and P. lambertiana were comparable (44%) and higher than the reported genome average of 38% (Gonzalez-Ibeas et al., 2016; Perera et al., 2018). A lower GC content in plant gene introns has been reported for other plant species (Mizuno and Kanehisa, 1994; Singh et al., 2016), indicating that intron sequences may have a more relaxed DNA conformation and are more accessible to transcription and other regulatory factors (Schwartz et al., 2009; Gelfman and Ast, 2013).

It remains unclear if TE sequences have insertional preferences, but it has been reported that a relaxed chromatin configuration promotes TE insertions and leads to an increased mutation probability in tissue- and stage-specific genes (Muotri et al., 2005; Singer et al., 2010). In this aspect, TEs should be randomly distributed in different stress-responsive gene non-coding regions (flanks and introns). The non-autonomous MITE3321 element insertions were statistically significantly overrepresented in the proximity of pine genes (0–2 kb), a distance over which linkage equilibrium extends in P. taeda (Brown et al., 2004). MITEs are preferentially located within gene regions of many plants and could influence gene expression (Bureau and Wessler, 1994; Wessler et al., 1995; Casacuberta and Santiago, 2003; Liu et al., 2005; Lu et al., 2012). There is evidence that some MITE insertions located close to gene promoters could also

downregulate gene expression (To et al., 2015). MITE3321 was frequently inserted also in the introns of genes of the studied pine species. However, in this study, no genes with several MITE3321 insertions in its different non-coding regions (flanks or introns) were identified. This distribution suggests formation of differentially regulated gene sub-networks, depending on the location of MITE insertions. For example, it was determined that many genes containing MITE3321 in flanking regions were involved in the regulation of developmental processes and cell division, but genes having MITE3321 in their introns were associated with immune responses and cell-wall biosynthesis, among other activities. If MITE3321 cis-acting elements in the more accessible introns help activate a specific network of genes, then non-coding RNA products of splicing from stress- responsive genes could be involved in feedback loop mechanisms for blocking transcription of genes with proximal MITE3321. In this example, identical TEs could have a downregulatory effect on developmental gene transcription, switching to an energy-saving mode, while at the same time increasing transcription of defense genes. Similar strategies could be highly advantageous for the rapid activation of defense responses and switching of metabolic functions. Additionally, in the current study, MITE3321 insertions were found in both analyzed pine species, belonging to separate subgenera, suggesting similar distributions also in other pine species. Therefore, MITE3321 could be a useful molecular marker for genotyping of pine species, as shown for MITEs in other plant species (Casa et al., 2000; Singh et al., 2017; Stelmach et al., 2017).

TE distribution is linked with speciation (Jurka et al., 2011; Serrato-Capuchina and Matute, 2018). In pines, transposition activity accounts for the period following the divergence of the pine subgenera Strobus (P. lambertiana) and Pinus (P. taeda) and speciation (Kossack and Kinlaw, 1999; Friesen et al., 2001; Neale et al., 2014; Stevens et al., 2016; Voronova et al., 2017). Only a few homologous genes in both pine species were found with similar TE insertions, indicating that differences in TE family distribution between species are also found in more conservative gene regions. Insertion of DTX184 in homologous genes of both species was found in PSMD4, a 26S proteasome non-ATPase regulatory subunit gene (40% query coverage, 98% sequence identity of transcripts). Insertion of the ancient Gypsy RLX IFG (Kossack and Kinlaw, 1999) was found in three homologous P. taeda and P. lambertiana gene pairs, surprisingly, all encoding different protein kinases. However, some of the IFG insertions were located in different introns of kinase genes and therefore could also represent independent transpositions. Analyzed IFG sequences contained masked regions and therefore the age of the insertions was not evaluated. Notably, insertions of other TEs were also frequently found in different receptor-like protein kinases and tyrosine protein kinases in our study. Protein kinase genes belong to one of the most proliferated gene superfamilies in plants, whose members are linked with a range of key metabolic and various plant-specific adaptation processes (Stone and Walker, 1995; Lehti- Shiu et al., 2009; Lehti-Shiu and Shiu, 2012; White-Gloria et al., 2018).

In conclusion, the source of TE sequences expressed in response to stress conditions could be the transcription of introns of many stress-responsive genes, which could explain the observed highly correlated expression levels of RLX families within individuals (Voronova, 2019). Transfer of information about TE insertions in gene regions to non-model pine species is complicated, as common TE families were revealed, but they are generally located in non-homologous genes. This highlights the need for additional studies and sequencing of species of interest to investigate TE-associated polymorphisms, such as in P. sylvestris, which is an important species in northern Europe. In the two analyzed pine species, TE insertions were more often found in gene introns but less commonly in gene flanking regions, similar to other plant species (West et al., 2014; Le et al., 2015; Li et al., 2017a; Ma et al., 2019). Insertions of TEs in gene regions are associated with various epigenetic mechanisms, but it remains unknown how some actively transcribed plant genes are coping with large introns (To et al., 2015). Cultivation and plant breeding reduces genetic diversity and fitness to environmental stresses (Roger et al., 2012). A recent whole-genome comparison of wild and cultivated rice species revealed depletion of intronic TE insertions in cultivated species (Li et al., 2017b). Gymnosperms are outcrossing species that produce large quantities of pollen and seeds, generating a genetically diverse germplasm pool for subsequent natural selection of highly adaptable seedlings. Pine species are known for their strong adaptation to local growing conditions (Savolainen et al., 2007; Eriksson, 2008; Neale and Ingvarsson, 2008; Prunier et al., 2016). This study demonstrated the increased accumulation of TE sequences in stress-responsive gene introns and the probability of rewiring them in responsive networks interconnected with node genes containing multiple TEs. Preferential TE insertion in open chromatin has been reported for tissue-specific genes (Muotri et al., 2005), hence similar mechanisms followed by sustained natural selection could drive the accumulation of fitting TEs in gene non-coding sequences of plants. The inclusion or exclusion of genes from TE-mediated networks is an efficient way means of dynamic change in response to various environmental factors, including changing host-pathogen interactions and unresolved layered processes in plants such as communication and signaling organization. Therefore, many such regulatory influences could lead to the adaptive environmental response clines that are characteristic of naturally spread pine populations (Eckert et al., 2010; Cumbie et al., 2011; Chhatre et al., 2013; Lu et al., 2016).

Tables

Due to technical limitations, Tables 1-3 have been provided as a supplementary file.

Declarations

Ethics approval and consent to participate

Not applicable

Availability of data and materials

Please contact author for data requests.

Funding

This study was supported by the European Regional Development Fund Postdoctoral research aid 1.1.1.2/VIAA/1/16/094.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AV conceived and designed the analysis, collected data, performed the analyses, contributed to the interpretation of the results, and wrote the manuscript. DR helped supervise the project. PI and MAR contributed to mastering of data and analysis tools. RK verified the analytical methods and results.

All authors discussed the results, commented on the manuscript, and contributed to the final version for publication.

Acknowledgments

Not applicable

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., et al. (2003). A uniform system for microRNA annotation. RNA 9, 277–279. doi:10.1261/rna.2183803.

Argueso, C. T., Ferreira, F. J., Epple, P., To, J. P. C., Hutchison, C. E., Schaller, G. E., et al. (2012). Two-component elements mediate interactions between cytokinin and salicylic acid in plant immunity. PLoS Genet. 8. doi:10.1371/journal.pgen.1002448.

Baluska, F., Gagliano, M., and Witzany, G. (2018). Memory and Learning in Plants. doi:10.1007/978- 3-319-75596-0.

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11. doi:10.1186/s13100-015-0041-9.

Beguiristain, T., Grandbastien, M. A., Puigdomènech, P., and Casacuberta, J. M. (2001). Three Tnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. Plant Physiol. 127, 212–221. doi:10.1104/pp.127.1.212.

Bergman, C. M., and Quesneville, H. (2007). Discovering and detecting transposable elements in genome sequences. Brief. Bioinform. 8, 382–392. doi:10.1093/bib/bbm048.

Betley, J. N., Frith, M. C., Graber, J. H., Choo, S., and Deshler, J. O. (2002). A ubiquitous and conserved signal for RNA localization in chordates. Curr. Biol. 12, 1756–61. doi:10.1016/s0960- 9822(02)01220-4.

Brierley, C., and Flavell, A. J. (1990). The retrotransposon Copia controls the relative levels of its gene products post-transcriptionally be differential expression from its two major mRNAs. Nucleic Acids Res. 18, 2947–2951. doi:10.1093/nar/18.10.2947.

Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H., and Neale, D. B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. 101, 15255–15260. doi:10.1073/pnas.0404231101.

Bureau, T. E., and Wessler, S. R. (1994). Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell 6, 907–916. doi:10.1105/tpc.6.6.907.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., et al. (2012). Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. Plant Cell 24, 1242–1255. doi:10.1105/tpc.111.095232.

Camiolo, S., Rau, D., and Porceddu, A. (2009). Mutational biases and selective forces shaping the structure of Arabidopsis genes. PLoS One 4. doi:10.1371/journal.pone.0006356.

Candar-Cakir, B., Arican, E., and Zhang, B. (2016). Small RNA and degradome deep sequencing

reveals drought-and tissue-specific micrornas and their important roles in drought-sensitive and drought-tolerant tomato genotypes. Plant Biotechnol. J. 14, 1727–1746. doi:10.1111/pbi.12533.

Capy, P., Gasperi, G., Biémont, C., and Bazin, C. (2000). Stress and transposable elements: Co- evolution or useful parasites? Heredity (Edinb). 85, 101–106. doi:10.1046/j.1365- 2540.2000.00751.x.

Casa, A. M., Brouwer, C., Nagel, A., Wang, L., Zhang, Q., Kresovich, S., et al. (2000). The MITE family Heartbreaker (Hbr): Molecular markers in maize. Proc. Natl. Acad. Sci. 97, 10083–10089. doi:10.1073/pnas.97.18.10083.

Casacuberta, J. M., and Grandbastien, M. angèle (1993). Characterisation of LTR sequences involved in the protoplast specific expression of the tobacco Tnt1 retrotransposon. Nucleic Acids Res. 21, 2087–2093. doi:10.1093/nar/21.9.2087.

Casacuberta, J. M., and Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: Control of transposition and impact on the evolution of plant genes and genomes. Gene 311, 1–11. doi:10.1016/S0378-1119(03)00557-2.

Chhatre, V. E., Byram, T. D., Neale, D. B., Wegrzyn, J. L., and Krutovsky, K. V. (2013). Genetic structure and association mapping of adaptive and selective traits in the east Texas loblolly pine (Pinus taeda L.) breeding populations. Tree Genet. Genomes 9, 1161–1178. doi:10.1007/s11295- 013-0624-x.

Chu, C.-G., Tan, C. T., Yu, G.-T., Zhong, S., Xu, S. S., and Yan, L. (2011). A Novel Retrotransposon Inserted in the Dominant Vrn-B1 Allele Confers Spring Growth Habit in Tetraploid Wheat (Triticum turgidum L.). G3 (Bethesda). 1, 637–45. doi:10.1534/g3.111.001131.

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo,

59. (2012). Why assembling plant genome sequences is so challenging. Biology (Basel). 1, 439– 59. doi:10.3390/biology1020439.

Colinas, J., Schmidler, S. C., Bohrer, G., Iordanov, B., and Benfey, P. N. (2008). Intergenic and genic sequence lengths have opposite relationships with respect to gene expression. PLoS One 3. doi:10.1371/journal.pone.0003670.

Crepeau, M. W., Langley, C. H., and Stevens, K. A. (2017). From Pine Cones to Read Clouds: Rescaffolding the Megagenome of Sugar Pine (Pinus lambertiana). G3: Genes|Genomes|Genetics 7, 1563–1568. doi:10.1534/g3.117.040055.

Cumbie, W. P., Eckert, A., Wegrzyn, J., Whetten, R., Neale, D., and Goldfarb, B. (2011). Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of Pinus taeda L. Heredity (Edinb). 107, 105–114. doi:10.1038/hdy.2010.168.

Das, S., and Bansal, M. (2019). Variation of gene expression in plants is influenced by gene architecture and structural properties of promoters. PLoS One 14. doi:10.1371/journal.pone.0212678.

Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez,

1. C., et al. (2010). Patterns of Population Structure and Environmental Associations to Aridity

Across the Range of Loblolly Pine (Pinus taeda L., Pinaceae). Genetics 185, 969–982. doi:10.1534/genetics.110.115543.

Eriksson, G. (2008). Pinus sylvestris recent genetic research. Uppsala: Swedish University of Agricultural Science.

Eulgem, T., and Somssich, I. E. (2007). Networks of WRKY transcription factors in defense signaling.

Curr. Opin. Plant Biol. 10, 366–371. doi:10.1016/j.pbi.2007.04.020.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nat. Rev. Genet.

9, 397–405. doi:10.1038/nrg2337.

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: Where genetics meets genomics. Nat. Rev. Genet. 3, 329–341. doi:10.1038/nrg793.

Flavell, A. J., Knox, M. R., Pearce, S. R., and Ellis, T. H. N. (1998). Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J. 16, 643–650. doi:10.1046/j.1365-313X.1998.00334.x.

Fray, R. G., and Grierson, D. (1993). Identification and genetic analysis of normal and mutant phytoene synthase genes of tomato by sequencing, complementation and co-suppression. Plant Mol. Biol. 22, 589–602. doi:10.1007/BF00047400.

Friesen, N., Brandes, A., and Heslop-Harrison, J. S. (2001). Diversity, origin, and distribution of retrotransposons (gypsy and copia) in conifers. Mol. Biol. Evol. 18, 1176–1188. doi:10.1093/oxfordjournals.molbev.a003905.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next- generation sequencing data. Bioinformatics 28, 3150–3152. doi:10.1093/bioinformatics/bts565.

Galindo-González, L., Mhiri, C., Deyholos, M. K., and Grandbastien, M. A. (2017). LTR- retrotransposons in plants: Engines of evolution. Gene 626, 14–25. doi:10.1016/j.gene.2017.04.051.

Gao, D., Jimenez-Lopez, J. C., Iwata, A., Gill, N., and Jackson, S. A. (2012). Functional and Structural Divergence of an Unusual LTR Retrotransposon Family in Plants. PLoS One 7, 1–12. doi:10.1371/journal.pone.0048595.

Garcia-Martinez, J., and Martínez-Izquierdo, J. A. (2003). Study on the Evolution of the Grande Retrotransposon in the Zea Genus. Mol. Biol. Evol. 20, 831–841. doi:10.1093/molbev/msg095.

Gelfman, S., and Ast, G. (2013). When epigenetics meets alternative splicing: the roles of DNA methylation and GC architecture. Epigenomics 5, 351–353. doi:10.2217/epi.13.32.

Goenawan, I. H., Bryan, K., and Lynn, D. J. (2016). DyNet: visualization and analysis of dynamic molecular interaction networks. Bioinformatics 32, 2713–2715. doi:10.1093/bioinformatics/btw187.

Gonzalez-Ibeas, D., Martinez-Garcia, P. J., Famula, R. A., Delfino-Mix, A., Stevens, K. A., Loopstra,

1. *A., et al. (2016). Assessing the Gene Content of the Megagenome: Sugar Pine (Pinus*

lambertiana). G3: Genes|Genomes|Genetics 6, 3787–3802. doi:10.1534/g3.116.032805.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: Ten years of next- generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. doi:10.1038/nrg.2016.49.

Grandbastien, M. A. (2015). LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim. Biophys. Acta - Gene Regul. Mech. 1849, 403–416. doi:10.1016/j.bbagrm.2014.07.017.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008). The Vienna RNA websuite. Nucleic Acids Res. 36. doi:10.1093/nar/gkn188.

Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., et al. (2016). Draft genome of the living fossil Ginkgo biloba. Gigascience 5, 49. doi:10.1186/s13742-016-0154-1.

Hadjiargyrou, M., and Delihas, N. (2013). The intertwining of transposable elements and non-coding RNAs. Int. J. Mol. Sci. 14, 13307–28. doi:10.3390/ijms140713307.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 27, 297–300. doi:10.1093/nar/27.1.297.

Huntley, R. P., Binns, D., Dimmer, E., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO: A user tutorial for the web-based Gene Ontology browser. Database 2009, 1–19. doi:10.1093/database/bap010.

Hwang, I., Sheen, J., and Müller, B. (2012). Cytokinin Signaling Networks. Annu. Rev. Plant Biol. 63, 353–380. doi:10.1146/annurev-arplant-042811-105503.

Jurka, J., Bao, W., and Kojima, K. K. (2011). Families of transposable elements, population structure and the origin of species. Biol. Direct 6, 44. doi:10.1186/1745-6150-6-44.

Kalendar, R., Amenov, A., and Daniyarov, A. (2019). Use of retrotransposon-derived genetic markers to analyse genomic variability in plants. Funct. Plant Biol. 46, 15–29. doi:10.1071/FP18098.

Kalendar, R., Flavell, A. J., Ellis, T. H. N., Sjakste, T., Moisy, C., and Schulman, A. H. (2011). Analysis of plant diversity with retrotransposon-based molecular markers. Heredity (Edinb). 106, 520–530. doi:10.1038/hdy.2010.93.

Kalendar, R., and Schulman, A. H. (2006). IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nat. Protoc. 1, 2478–2484. doi:10.1038/nprot.2006.377.

Kalendar, R., Tanskanen, J., Chang, W., Antonius, K., Sela, H., Peleg, O., et al. (2008). Cassandra retrotransposons carry independently transcribed 5S RNA. Proc. Natl. Acad. Sci. U. S. A. 105, 5833–5838. doi:10.1073/pnas.0709698105.

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., and Schulman, A. H. (2000). Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. 97, 6603–6607. doi:10.1073/pnas.110587497.

Kalendar, R., Vicient, C. M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A. H. (2004). Large Retrotransposon Derivatives: Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes. Genetics 166, 1437–1450. doi:10.1534/genetics.166.3.1437.

Kashkush, K., Feldman, M., and Levy, A. A. (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat. Genet. 33, 102–106. doi:10.1038/ng1063.

Kazazian, H. H. (2004). Mobile Elements: Drivers of Genome Evolution. Science (80-.). 303, 1626– 1632. doi:10.1126/science.1089670.

Kentner, E. K., Arnold, M. L., and Wessler, S. R. (2003). Characterization of high-copy-number retrotransposons from the large genomes of the Louisiana iris species and their use as molecular markers. Genetics 164, 685–697.

Kobayashi, S., Goto-Yamamoto, N., and Hirochika, H. (2004). Retrotransposon-Induced Mutations in Grape Skin Color. Science (80-.). 304, 982. doi:10.1126/science.1095011.

Korf, I., Neale, D., Kovach, A., Wegrzyn, J., Parra, G., Holt, C., et al. (2010). The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11, 420. doi:10.1186/1471-2164-11-420.

Kossack, D. S., and Kinlaw, C. S. (1999). IFG, a gypsy-like retrotransposon in Pinus (Pinaceae), has an extensive history in pines. Plant Mol. Biol. 39, 417–426. doi:10.1023/A:1006115732620.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. 47, D155–D162. doi:10.1093/nar/gky1141.

Krom, N., Recla, J., and Ramakrishna, W. (2008). Analysis of genes associated with retrotransposons in the rice genome. Genetica 134, 297–310. doi:10.1007/s10709-007-9237-3.

Kumar, A., and Bennetzen, J. L. (1999). Plant Retrotransposons. Annu. Rev. Genet. 33, 479–532. doi:10.1146/annurev.genet.33.1.479.

Lai, Y., Cuzick, A., Lu, X. M., Wang, J., Katiyar, N., Tsuchiya, T., et al. (2019). The Arabidopsis RRM domain protein EDM 3 mediates race-specific disease resistance by controlling H3K9me2- dependent alternative polyadenylation of RPP 7 immune receptor transcripts. Plant J. 97, 646– 660. doi:10.1111/tpj.14148.

Le, T. N., Miyazaki, Y., Takuno, S., and Saze, H. (2015). Epigenetic regulation of intragenic transposable elements impacts gene transcription in Arabidopsis thaliana. Nucleic Acids Res. 43, 3911–3921.

doi:10.1093/nar/gkv258.

Lee, H., Zhang, Z., and Krause, H. M. (2019). Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? Trends Genet. 35, 892–902. doi:10.1016/j.tig.2019.09.006.

Lehti-Shiu, M. D., and Shiu, S.-H. (2012). Diversity, classification and function of the plant protein kinase superfamily. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367, 2619–39. doi:10.1098/rstb.2012.0003.

Lehti-Shiu, M. D., Zou, C., Hanada, K., and Shiu, S.-H. (2009). Evolutionary history and stress regulation of plant receptor-like kinase/pelle genes. Plant Physiol. 150, 12–26. doi:10.1104/pp.108.134353.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, S. F., Su, T., Cheng, G. Q., Wang, B. X., Li, X., Deng, C. L., et al. (2017a). Chromosome evolution in connection with repetitive sequences and epigenetics in plants. Genes (Basel). 8. doi:10.3390/genes8100290.

Li, X., Guo, K., Zhu, X., Chen, P., Li, Y., Xie, G., et al. (2017b). Domestication of rice has reduced the occurrence of transposable elements within gene coding regions. BMC Genomics 18, 55. doi:10.1186/s12864-016-3454-z.

Liu, B., Shan, X. H., Liu, Z. L., Dong, Z. Y., Wang, Y. M., Chen, Y., et al. (2005). Mobilization of the active MITE transposons mPing and Pong in rice by introgression from wild rice (Zizania latifolia Griseb.). Mol. Biol. Evol. 22, 976–990.

Liu, M., Yu, H., Zhao, G., Huang, Q., Lu, Y., and Ouyang, B. (2017). Profiling of drought-responsive microRNA and mRNA in tomato using high-throughput sequencing. BMC Genomics 18, 1–18.

doi:10.1186/s12864-017-3869-1.

Liu, Y., Yin, J., Xiao, M., Mason, A. S., Gao, C., Liu, H., et al. (2013). Characterization of Structure, Divergence and Regulation Patterns of Plant Promoters. J. Mol. Biol. Res. 3, 23–36. doi:10.5539/jmbr.v3n1p23.

Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W., and Kuang, H. (2012). Miniature Inverted-Repeat Transposable Elements (MITEs) Have Been Accumulated through Amplification Bursts and Play Important Roles in Gene Expression and Species Diversity in Oryza sativa. Mol. Biol. Evol. 29, 1005–1017. doi:10.1093/molbev/msr282.

Lu, M., Krutovsky, K. V., Nelson, C. D., Koralewski, T. E., Byram, T. D., and Loopstra, C. A. (2016). Exome genotyping, linkage disequilibrium and population structure in loblolly pine (Pinus taeda L.). BMC Genomics 17, 730. doi:10.1186/s12864-016-3081-8.

Ma, B., Xin, Y., Kuang, L., and He, N. (2019). Distribution and Characteristics of Transposable Elements in the Mulberry Genome. Plant Genome 12, 0. doi:10.3835/plantgenome2018.12.0094.

Macas, J., and Neumann, P. (2007). Ogre elements — A distinct group of plant Ty3/gypsy-like retrotransposons. Gene 390, 108–116. doi:10.1016/j.gene.2006.08.007.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21, 3448–3449. doi:10.1093/bioinformatics/bti551.

MAIER, F., Zwicker, S., Hückelhoven, A., Meissner, M., FUNK, J., Pfitzner, A. J. P. P., et al. (2011). Nonexpressor Of Pathogenesis-Related Proteins1 (NPR1) and some NPR1-related proteins are sensitive to salicylic acid. Mol. Plant Pathol. 12, 73–91. doi:10.1111/j.1364-3703.2010.00653.x.

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J., et al. (2015). Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. PLoS Genet. 11. doi:10.1371/journal.pgen.1004915.

McClintock, B. (1984). The significance of responses of the genome to challenge. Science (80-.). 226, 792–801. doi:10.1126/science.15739260.

Mita, P., and Boeke, J. D. (2016). How retrotransposons shape genome regulation. Curr. Opin. Genet.

Dev. 37, 90–100. doi:10.1016/j.gde.2016.01.001.

Mizuno, M., and Kanehisa, M. (1994). Distribution profiles of GC content around the translation initiation site in different species. FEBS Lett. 352, 7–10. doi:10.1016/0014-5793(94)00898-1.

Mok, M. C. (1994). "Cytokinin and Plant Development—An Overview.," in CytokininsChemistry, Activity, and Function, ed. M. C. Mok, David W. S.; Mok (Ann Arbor, Michigan: CRC Press), 155–166. Available at: https://books.google.lv/books?hl=lv&lr=&id=yov7iUL7OTAC&oi=fnd&pg=PA155&dq=Cytokinin+and+Plant+Development— An+Overview.+Mok+1994&ots=0dGZoSlvBj&sig=_qiXSJaSqmGC0e-KG-ECigKxfig&redir_esc=y#v=onepage&q=Cytokinin and Plant Development—An Overview. Mok 1994&f [Accessed June 10, 2020].

Mok, M. C., Martin, R. C., and Mok, D. W. S. (2000). Cytokinins: Biosynthesis, metabolism and perception. Vitr. Cell. Dev. Biol. - Plant 36, 102–107. doi:10.1007/s11627-000-0021-7.

Monden, Y., and Tahara, M. (2015). Plant Transposable Elements and Their Application to Genetic Analysis via High-throughput Sequencing Platform. Hortic. J. 84, 283–294. doi:10.2503/hortj.MI-IR02.

Morse, A. M., Peterson, D. G., Islam-Faridi, M. N., Smith, K. E., Magbanua, Z., Garcia, S. A., et al. (2009). Evolution of genome size and complexity in Pinus. PLoS One 4, 1–11. doi:10.1371/journal.pone.0004332.

Muotri, A. R., Chu, V. T., Marchetto, M. C. N., Deng, W., Moran, J. V., and Gage, F. H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435, 903–910. doi:10.1038/nature03663.

Neale, D. B., and Ingvarsson, P. K. (2008). Population, quantitative and comparative genomics of adaptation in forest trees. Curr. Opin. Plant Biol. 11, 149–155. doi:10.1016/j.pbi.2007.12.004.

Neale, D. B., Wegrzyn, J. L., Stevens, K. A., Zimin, A. V., Puiu, D., Crepeau, M. W., et al. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15. doi:10.1186/gb-2014-15-3-r59.

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919–1920. doi:10.1093/bioinformatics/bts277.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature 497, 579. Available at:

https://doi.org/10.1038/nature12211.

Parida, S. K., Dalal, V., Singh, A. K., Singh, N. K., and Mohapatra, T. (2009). Genic non-coding microsatellites in the rice genome: characterization, marker design and use in assessing genetic and evolutionary relationships among domesticated groups. BMC Genomics 10, 140. doi:10.1186/1471-2164-10-140.

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. Bioinformatics 28, 2537–9. doi:10.1093/bioinformatics/bts460.

Pellicer, J., Hidalgo, O., Dodsworth, S., and Leitch, I. J. (2018). Genome Size Diversity and Its Impact on the Evolution of Land Plants. Genes (Basel). 9. doi:10.3390/GENES9020088.

Perera, D., Magbanua, Z. V., Thummasuwan, S., Mukherjee, D., Arick, M., Chouvarine, P., et al. (2018). Exploring the loblolly pine (Pinus taeda L.) genome by BAC sequencing and Cot analysis. Gene 663, 165–177. doi:10.1016/J.GENE.2018.04.024.

Prunier, J., Verta, J.-P., and MacKay, J. J. (2016). Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. New Phytol. 209, 44–62. doi:10.1111/nph.13565.

Purugganan, M. D., and Wessler, S. R. (1995). Transposon signatures: species-specific molecular markers that utilize a class of multiple-copy nuclear DNA. Mol. Ecol. 4, 265–270. doi:10.1111/j.1365-294X.1995.tb00218.x.

Qin, S., Jin, P., Zhou, X., Chen, L., and Ma, F. (2015). The Role of Transposable Elements in the Origin and Evolution of MicroRNAs in Human. PLoS One 10, e0131365. doi:10.1371/journal.pone.0131365.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. doi:10.1093/bioinformatics/btq033.

Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu. Rev. Genet. 46, 21–42. doi:10.1146/annurev-genet-110711-155621.

Ren, X. Y., Vorst, O., Fiers, M. W. E. J., Stiekema, W. J., and Nap, J. P. (2006). In plants, highly expressed genes are the least compact. Trends Genet. 22, 528–532. doi:10.1016/j.tig.2006.08.008.

Rho, M., Choi, J.-H. H., Kim, S., Lynch, M., and Tang, H. (2007). De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 8, 1–16. doi:10.1186/1471-2164-8-90.

Roger, F., Godhe, A., and Gamfeldt, L. (2012). Genetic Diversity and Ecosystem Functioning in the Face of Multiple Stressors. PLoS One 7. doi:10.1371/journal.pone.0045007.

Sabot, F., and Schulman, A. H. (2007). Template switching can create complex LTR retrotransposon insertions in Triticeae genomes. BMC Genomics 8, 5–9. doi:10.1186/1471-2164-8-247.

Sabot, F., Sourdille, P., and Bernard, M. (2005). Advent of a new retrotransposon structure: The long form of the Veju elements. Genetica 125, 325–332. doi:10.1007/s10709-005-7926-3.

Sakai, H., Honma, T., Takashi, A., Sato, S., Kato, T., Tabata, S., et al. (2001). ARR1, a transcription factor for genes immediately responsive to cytokinins. Science (80-.). 294, 1519–1521. doi:10.1126/science.1065201.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., et al. (1996). Nested Retrotransposons in the Intergenic Regions of the Maize Genome. Science (80-.). 274, 765–768. doi:10.1126/science.274.5288.765.

Savolainen, O., Pyhäjärvi, T., and Knürr, T. (2007). Gene Flow and Local Adaptation in Trees. Annu.

Rev. Ecol. Evol. Syst. 38, 595–619. doi:10.1146/annurev.ecolsys.38.091206.095646.

Schulman, A. H. (2013). Retrotransposon replication in plants. Curr. Opin. Virol. 3, 604–614. doi:10.1016/j.coviro.2013.08.009.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure.

Nat. Struct. Mol. Biol. 16, 990–995. doi:10.1038/nsmb.1659.

Serrato-Capuchina, A., and Matute, D. R. (2018). The Role of Transposable Elements in Speciation.

Genes (Basel). 9. doi:10.3390/genes9050254.

Shang, Y., Yang, F., Schulman, A. H., Zhu, J., Jia, Y., Wang, J., et al. (2017). Gene Deletion in Barley Mediated by LTR-retrotransposon BARE. Sci. Rep. 7, 1–9. doi:10.1038/srep43766.

Shirasu, K., Schulman, A. H., Lahaye, T., and Schulze-Lefert, P. (2000). A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. 10, 908–915. doi:10.1101/gr.10.7.908.

Singer, T., McConnell, M. J., Marchetto, M. C. N., Coufal, N. G., and Gage, F. H. (2010). LINE-1 retrotransposons: Mediators of somatic variation in neuronal genomes? Trends Neurosci. 33, 345– 354. doi:10.1016/j.tins.2010.04.001.

Singh, R., Ming, R., and Yu, Q. (2016). Comparative Analysis of GC Content Variations in Plant Genomes. Trop. Plant Biol. 9, 136–149. doi:10.1007/s12042-016-9165-4.

Singh, S., Nandha, P. S., and Singh, J. (2017). Transposon-based genetic diversity assessment in wild and cultivated barley. Crop J. 5, 296–304. doi:10.1016/J.CJ.2017.01.003.

Slotkin, R. K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8, 272–285. doi:10.1038/nrg2072.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27, 431–432. doi:10.1093/bioinformatics/btq675.

Solovyev, V. V., Shahmuradov, I. A., and Salamov, A. A. (2010). "Identification of Promoter Regions and Regulatory Sites," in Methods in molecular biology (Clifton, N.J.), 57–83. doi:10.1007/978- 1-60761-854-

6_5.

Stelmach, K., Macko-Podgórni, A., Machaj, G., and Grzebelus, D. (2017). Miniature Inverted Repeat Transposable Element Insertions Provide a Source of Intron Length Polymorphism Markers in

the Carrot (Daucus carota L.). Front. Plant Sci. 8, 725. doi:10.3389/fpls.2017.00725.

Stevens, K. A., Wegrzyn, J. L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., et al. (2016). Sequence of the Sugar Pine Megagenome. Genetics 204, 1613–1626. doi:10.1534/genetics.116.193227.

Stone, J. M., and Walker, J. C. (1995). Plant protein kinase families and signal transduction. 108.

Available at: http://www.ncbi.nlm.nih.gov/pubmed/7610156 [Accessed December 10, 2019].

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nat. Genet. 43, 1160–1163. doi:10.1038/ng.942.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., et al. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 24, 1963– 1976. doi:10.1101/gr.168872.113.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613. doi:10.1093/nar/gky1131.

Takahashi, S., Inagaki, Y., Satoh, H., Hoshino, A., and Iida, S. (1999). Capture of a genomic HMG domain sequence by the En/Spm -related transposable element Tpn1 in the Japanese morning glory. Mol. Gen. Genet. MGG 261, 447–451. doi:10.1007/s004380050987.

Takeda, S., Sugimoto, K., Otsuki, H., and Hirochika, H. (1999). A 13-bp cis-regulatory element in the LTR promoter of the tobacco retrotransposon Tto1 is involved in responsiveness to tissue culture, wounding, methyl jasmonate and fungal elicitors. Plant J. 18, 383–393. doi:10.1046/j.1365- 313X.1999.00460.x.

Taniguchi, M., Sasaki, N., Tsuge, T., Aoyama, T., and Oka, A. (2007). ARR1 Directly Activates Cytokinin Response Genes that Encode Proteins with Diverse Regulatory Functions. Plant Cell Physiol. 48, 263–277. doi:10.1093/pcp/pcl063.

To, T. K., Saze, H., and Kakutani, T. (2015). DNA Methylation within Transcribed Regions. Plant Physiol. 168, 1219–1225. doi:10.1104/PP.15.00543.

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., et al. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. doi:10.1093/nar/gkz841.

Treangen, T. J., and Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36–46. doi:10.1038/nrg3117.

Tsuchiya, T., and Eulgem, T. (2013). An alternative polyadenylation mechanism coopted to the Arabidopsis RPP7 gene through intronic retrotransposon domestication. Proc. Natl. Acad. Sci. 110, E3535–E3543. doi:10.1073/pnas.1312545110.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., et al. (2012). Dissecting Plant Genomes with the PLAZA Comparative Genomics Platform. Plant Physiol. 158, 590–600. doi:10.1104/pp.111.189514.

Varagona, M. J., Purugganan, M., and Wessler, S. R. (1992). Alternative splicing induced by insertion of retrotransposons into the maize waxy gene. Plant Cell 4, 811–820. doi:10.1105/tpc.4.7.811.

Venkatesh, and Nandini, B. (2020). Miniature inverted-repeat transposable elements (MITEs), derived insertional polymorphism as a tool of marker systems for molecular plant breeding. Mol. Biol. Rep. 47. doi:10.1007/s11033-020-05365-y.

Vicient, C. M., Kalendar, R., and Schulman, A. H. (2005). Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. J. Mol. Evol. 61, 275–291. doi:10.1007/s00239- 004-0168-7.

Voronova, A. (2019). Retrotransposon expression in response to in vitro inoculation with two fungal pathogens of Scots pine (Pinus sylvestris L.). BMC Res. Notes 12, 243. doi:10.1186/s13104-019- 4275-3.

Voronova, A., Belevich, V., Jansons, A., and Rungis, D. (2014). Stress-induced transcriptional activation of retrotransposon-like sequences in the Scots pine (Pinus sylvestris L.) genome. Tree Genet. Genomes 10, 937–951. doi:10.1007/s11295-014-0733-1.

Voronova, A., Belevich, V., Korica, A., and Rungis, D. (2017). Retrotransposon distribution and copy number variation in gymnosperm genomes. Tree Genet. Genomes 13. doi:10.1007/s11295-017- 1165-5.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics.

Nat. Rev. Genet. 10, 57–63. doi:10.1038/nrg2484.

Waugh, R., McLean, K., Flavell, A. J., Pearce, S. R., Kumar, A., Thomas, B. B. T., et al. (1997). Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol. Gen. Genet. 253, 687–694. doi:10.1007/s004380050372.

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., et al. (2014). Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. Genetics 196, 891–909. doi:10.1534/genetics.113.159996.

Wei, F., Stein, J. C., Liang, C., Zhang, J., Fulton, R. S., Baucom, R. S., et al. (2009). Detailed analysis of a contiguous 22-Mb region of the maize genome. PLoS Genet. 5, e1000728. doi:10.1371/journal.pgen.1000728.

Wendel, J. F., Greilhuber, J., Doležel, J., and Leitch, I. J. (2012). Plant genome diversity volume 1: Plant genomes, their residents, and their evolutionary dynamics. Plant Genome Divers. Vol. 1 Plant Genomes, their Resid. their Evol. Dyn., 1–279. doi:10.1007/978-3-7091-1130-7.

Wessler, S. R. (1996). Plant retrotransposons: Turned on by stress. Curr. Biol. 6, 959–961. doi:10.1016/S0960-9822(02)00638-3.

Wessler, S. R., Bureau, T. E., and White, S. E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr. Opin. Genet. Dev. 5, 814–821.

West, P. T., Li, Q., Ji, L., Eichten, S. R., Song, J., Vaughn, M. W., et al. (2014). Genomic Distribution

of H3K9me2 and DNA Methylation in a Maize Genome. PLoS One 9, e105267. doi:10.1371/journal.pone.0105267.

White-Gloria, C., Johnson, J. J., Marritt, K., Kataya, A., Vahab, A., and Moorhead, G. B. (2018). Protein Kinases and Phosphatases of the Plastid and Their Potential Role in Starch Metabolism. Front. Plant Sci. 9, 1032. doi:10.3389/fpls.2018.01032.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982. doi:10.1039/b921331g.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., and van der Knaap, E. (2008). A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit. Science (80-.). 319,

1527–1530. doi:10.1126/science.1153040.

Yanagisawa, S. (2004). Dof Domain Proteins: Plant-Specific Transcription Factors Associated with Diverse Phenomena Unique to Plants. Plant Cell Physiol. 45, 386–391. doi:10.1093/pcp/pch055.

Yin, H., Liu, J., Xu, Y., Liu, X., Zhang, S., Ma, J., et al. (2013). TARE1, a Mutated Copia-Like LTR Retrotransposon Followed by Recent Massive Amplification in Tomato. PLoS One 8. doi:10.1371/journal.pone.0068587.

You, F. M., Cloutier, S., Shan, Y., and Ragupathy, R. (2015). LTR Annotator: Automated Identification and Annotation of LTR Retrotransposons in Plant Genomes. Int. J. Biosci. Biochem. Bioinforma. 5, 165–174. doi:10.17706/ijbbb.2015.5.3.165-174.

Zabala, G., and Vodkin, L. O. (2005). The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell 17, 2619–32. doi:10.1105/tpc.105.033506.

Zhao, H., Zhang, W., Chen, L., Wang, L., Marand, A. P., Wu, Y., et al. (2018a). Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. Plant Physiol. 176, pp.01467.2017. doi:10.1104/pp.17.01467.

Zhao, X., Li, J., Lian, B., Gu, H., Li, Y., and Qi, Y. (2018b). Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. Nat. Commun. 9, 5056. doi:10.1038/s41467-018-07500-7.

Zhou, H., Liu, Q., Li, J., Jiang, D., Zhou, L., Wu, P., et al. (2012). Photoperiod- and thermo-sensitive genic male sterility in rice are caused by a point mutation in a novel noncoding RNA that produces a small RNA. Cell Res. 22, 649–60. doi:10.1038/cr.2012.28.

Zimin, A., Stevens, K. A., Crepeau, M. W., Holtz-Morris, A., Koriabine, M., Marcais, G., et al. (2014). Sequencing and assembly of the 22-gb loblolly pine genome. Genetics 196, 875–890. doi:10.1534/genetics.113.159715.

Zimin, A. V., Stevens, K. A., Crepeau, M. W., Puiu, D., Wegrzyn, J. L., Yorke, J. A., et al. (2017). An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. Gigascience 6, 1–4. doi:10.1093/gigascience/giw016.

Figures

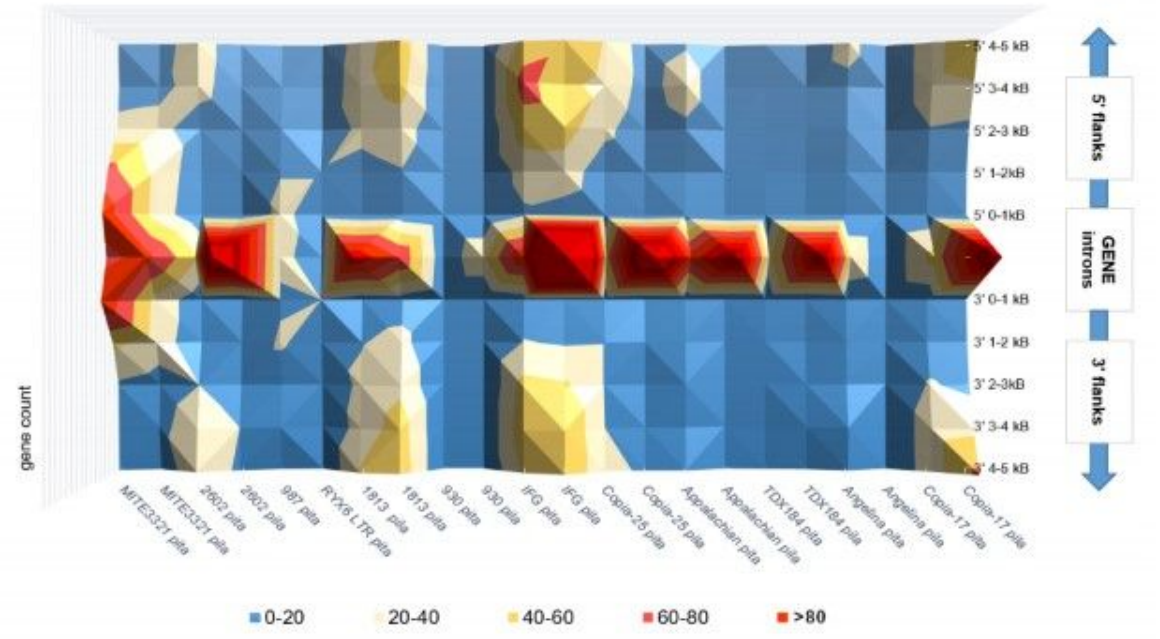

Figure 1

Comparison of TE distribution in gene non-coding regions. High-quality genes of P. lambertiana genome v.1.0 (pila) and filtered annotated gene set of P. taeda v.2.0 (pita).

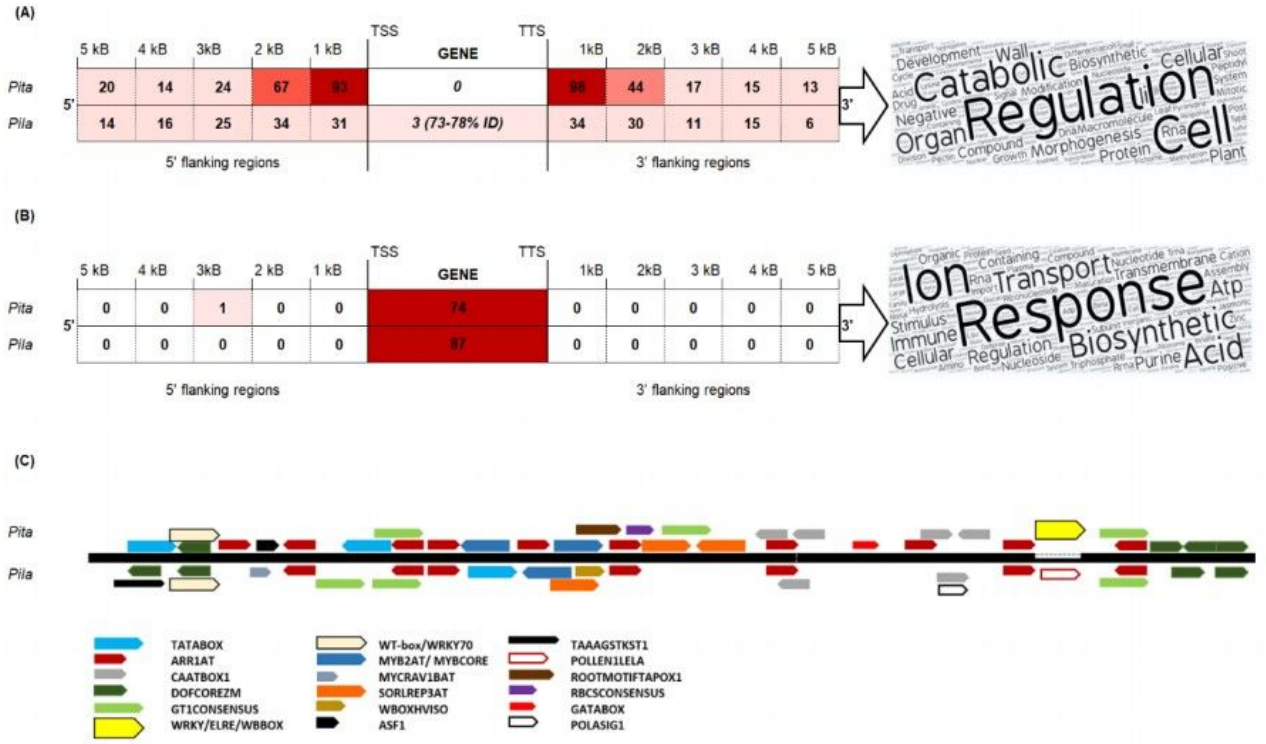

Figure 2

(A) Distribution of MITE3321 insertions across Pinus taeda (Pita) and Pinus lambertiana (Pila) gene flanking regions. (B) Distribution of MITE3321 insertions across Pinus taeda (Pita) and Pinus lambertiana (Pila) gene introns. World cloud generated from biological process GO terms of Pinus taeda genes involved in the networks using the online tool https://wordart.com/. (C) Alignment of Pinus taeda (Pita) and Pinus lambertiana (Pila) consensus sequences with predicted plant cis-acting regulatory elements.

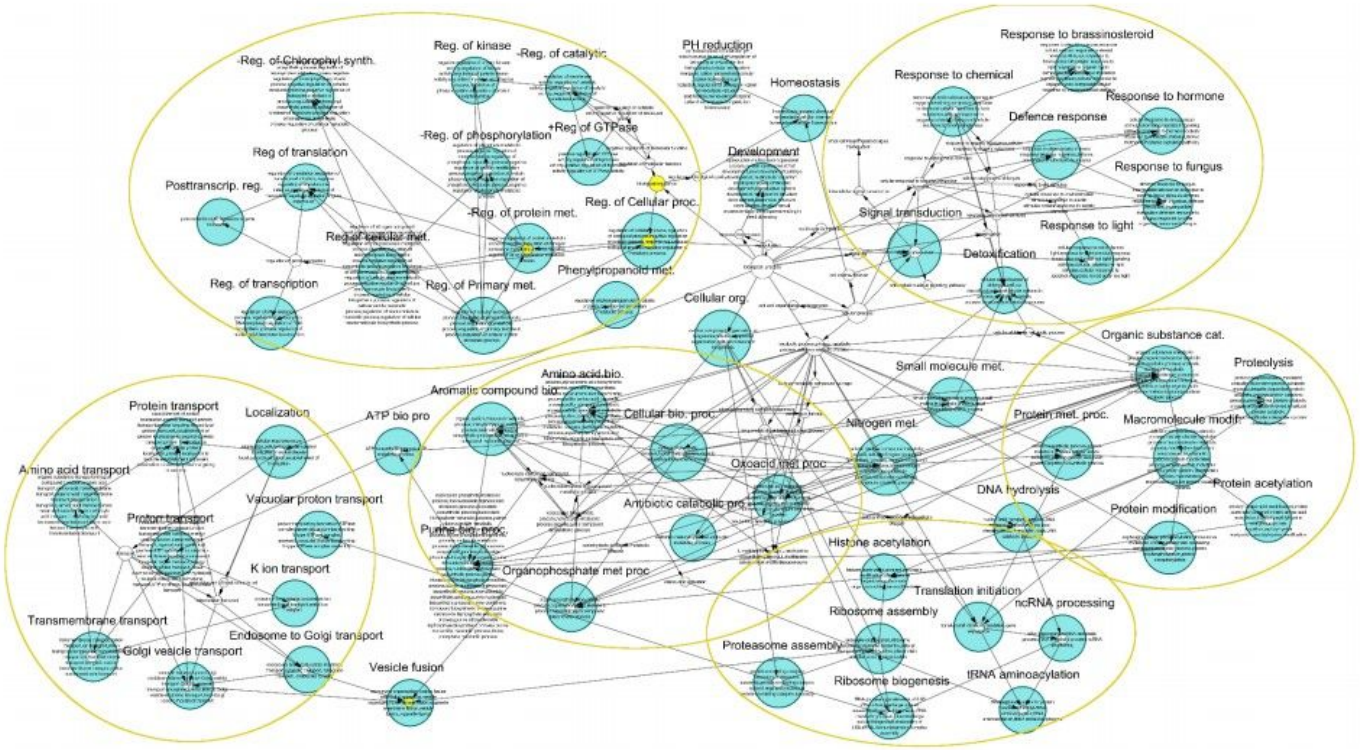

Figure 3

Gene network formed by DTX184 presence in the gene introns of P. taeda.

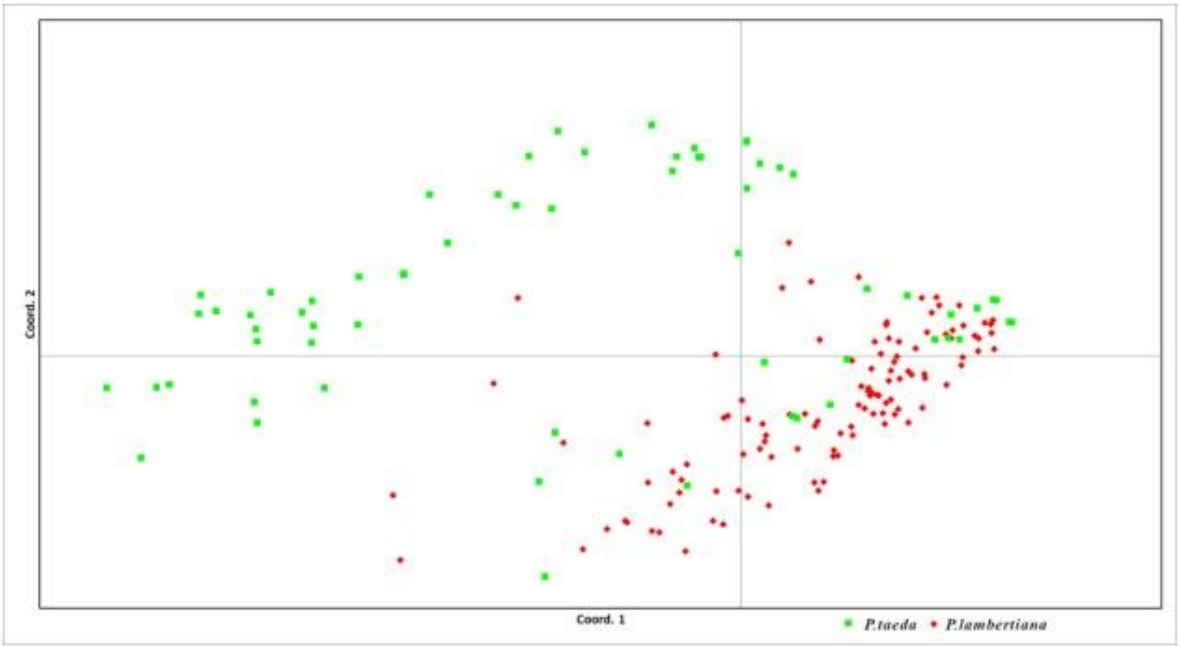

Figure 4

PCoA of RLX Copia-1813 gene network of all co-occurring TE patterns.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- *v2tables13.pdf*
- *v2additonalfiles1through10.pdf*