Transposable elements interconnect genes into networks via non-coding RNAs and other regulatory factors in pine

Angelika Voronova[1]*, Martha Rendón-Anaya[2], Pär Ingvarsson[2], Ruslan Kalendar[3], Dainis Ruņģis[1]

[1]Forest Molecular Biology and Genetics group, Latvian State Forest research institute "Silava", Salaspils, Latvia; [2] Linnean Centre for Plant Biology, Department of Plant Biology, Swedish University of Agricultural Sciences, Uppsala, Sweden; [3] Department of Agricultural Sciences, Viikki Plant Science Centre and Helsinki Sustainability Centre, University of Helsinki, Helsinki, Finland

Introduction

Conifer genomes are large (*P.sylvestris* (2C) = 46,96 pg or 44 949 Mbp, Fuchs *et al*. 2008), are characterised by multiple gene families and pseudogenes, contain large inter-gene regions and a high proportion of interspersed repeats. Up to 62% of the sequenced loblolly pine genome (*Pinus taeda*) consists of retrotransposon (RE) sequences and 70% of these are Long Terminal Repeat (LTR) REs (Neale *et al*. 2014). Transcription and transposition of REs is associated with stress conditions and/or meristematic tissues in various plant species. However, expression of the RE does not directly imply further transposition. In conifer genomes, it is possible to detect RE transcripts level increase in response to fungi pathogens and other stressors, where RE are probably co-expressed with stress associated genes (Voronova 2019; Voronova *et al*. 2013). It has been reported that transposable element (TE) composition varies considerably between individuals and can influence gene function by disruption of gene functional sequences, influencing of transcription, large insertions in introns could affect gene splicing, impact heterochromatin formation in the gene region, and play a part in functional non-coding RNA formation (Rebollo *et al.* 2012; Lisch 2013). TEs contribute to regulation of gene networks by embedding transcription factor binding sites (Feschotte 2008; Sundaram *et al.* 2014; Zhao *et al.* 2018). LTRs could contain transcription initiation and termination signals, cis-acting elements, polypurine tract (PPT), integrase binding signals, tRNA primer binding sites (Kumar, Bennetzen 1999). The aim of this study was the analysis of genes containing LTRs in flanking regions and gene introns in the *Pinus taeda v.2.0. and Pinus lambertiana v.1.01.* genomes. We also explored the possibility of transferring this information to *Pinus sylvestris* genome studies.

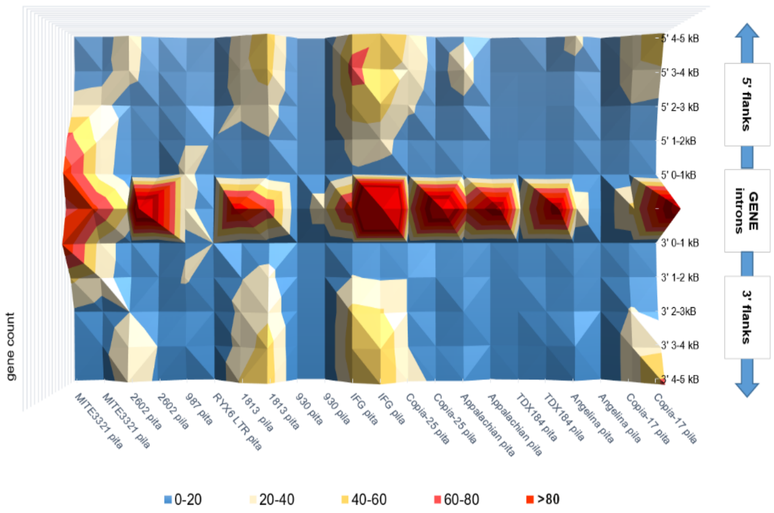

Figure 1. Comparison of TE distribution in gene non-coding regions. Explored data sets from high-quality genes of *P. lambertiana* genome v.1.0 and filtered annotated gene set of *P. taeda* v.2.0.

Table 3. Node genes containing several TE insertions and found to be homologous or containing identical domains containing genes between *P. taeda* and *P. lambertiana*.

LTR Nb.pita	LTR Nb.pila	Best hit/ BLASTX 2.9.0+ /refseq_protein	Accession, [b]. homologous genes	Common conserved domain name	Domain Accession	GO-terms
24	19	plastidial pyruvate kinase 2	XP_006843356.1[b]	PLN02623	PLN02623	reproduction; ATP generation from ADP; seed maturation;
23	26	DEAD-box ATP-dependent RNA helicase 20 isoform X2/helicase 58, chloroplastic isoform X3	XP_025888827.1/ XP_021667141.1	SrmB	COG0513	RNA secondary structure unwinding
21	21	phospholipid:diacylglycerol acyltransferase 1	XP_006849611.1[b]	PLN02517	PLN02517	acylglycerol biosynthetic process
18	20	nuclear pore complex protein NUP62-like/GPCR-type G protein 1 isoform X2	XP_024396806.1/ XP_007029700.2	SMC_prok_B super family	cl37069	RNA export from nucleus; protein import/export into/from nucleus; nucleocytoplasmic transport,
13	23	WD repeat-containing protein WRAP73	XP_008798782.1		COG2319	
	19	protein RAE1	XP_028076289.1	WD40 super family	cl29593	-
	24	actin-related protein 2/3 complex subunit 1A	XP_011627051.1		cl29593	
12	19	uncharacterized protein LOC109715170/probable E3 ubiquitin-protein ligase HERC4 isoform X1	XP_020095639.1	ATS1 super family	cl34932	-
11	31	peroxisomal adenine nucleotide carrier 1/mitochondrial substrate carrier family protein C-like	XP_006841423.1	Mito_carr	pfam00153	establishment of localization; transmembrane transport; amide biosynthetic process; translation; nitrogen compound metabolic process.

Main findings: node genes and TE patterns

- TE patterns embedded in gene introns could influence gene availability, responsiveness, stability, or higher order structure in the nucleus. Two evaluated genes with identical TE insertion patterns are involved in pre-mRNA maturation and splicing, with other genes with identical TE insertion genotypes linked to protein metabolic processes and Golgi body homeostasis. Further investigation will enable more thorough analyses of these processes.

- 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, so and these genes may be expressed in a broad range of conditions. If gene networks are formed via TE insertions, then genes with many different types of TEs could act as node genes that are functional or stable across a range of conditions and they could be important in early defense responses and metabolome switching. Several homologous genes with large introns containing similar protein domains were found in both pine species (**Table 3**)

Main findings: TE distribution in pine genes

- The quality of reference genomes and the repeat database used play a major role when analyzing the presence of TE in gene regions. The absence of full-length coverage of some retrotransposons, masked regions and ambiguous nesting structures, prevented determination of a consensus sequence and verification of some results, indicating that sequence scaffolding problems persist in the case of longer repetitive elements.

- Utilizing short repeats (LTRs) as TE representatives was considered more suitable at this point for evaluation of prevalent TE inside or near genes.

- No homologous genes were revealed in the most studied networks between pine species, indicating that most transposition events occurred after separation of the species.

- 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 process etc.

- The number of TE-derived repeats gradually increase with distance from genes, suggesting a slight elimination of TEs from gene regions (**Table 1**). Most TE diversity was observed in gene introns (Figure 1).

- Insertions of the DNA transposon *DTX184* carrying microRNA was found in the introns of important stress-responsive genes. One of the identified genes was NPR1 (Nonexpresser of Pathogenesis-related proteins-1), which is involved in plant systemic acquired resistance, and the salicylic acid-mediated signaling pathway. 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 (**Figure 3**).

- *IFG* retrotransposon is highly distributed in conifer genomes and it is far more ancient, but sequence homology is still maintained (Kossack and Kinlaw, 1999; Voronova et al., 2017). 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.

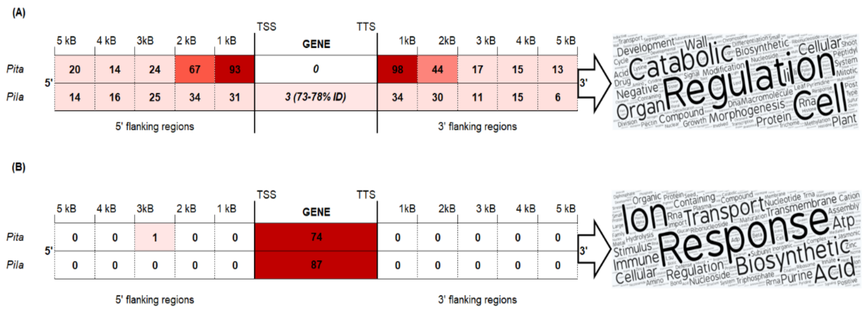

Figure 2. Distribution of *MITE3321* element insertions across *Pinus taeda* (Pita) and *Pinus lambertiana* (Pila) gene flanking regions and introns. World cloud generated from biological process GO terms of *Pinus taeda* genes involved in the networks using online tool https://wordart.com/.https://wordart.com/. (A) Gene count with *MITE3321* insertions in their flanks; (B) Gene count with *MITE3321* insertions in their introns.

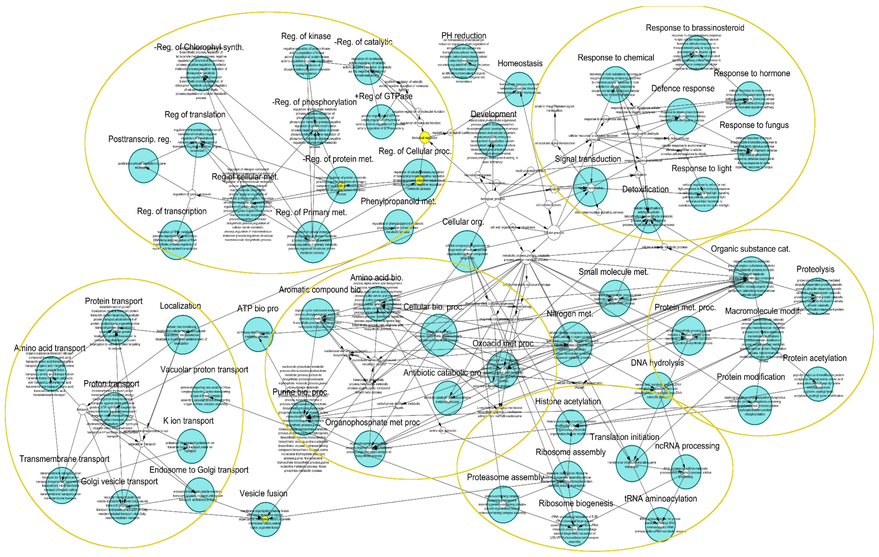

Figure 3. GO-based network constructed from 34 genes containing *DTX184* in *P.taeda* gene introns. Homolog of the *NPR1* gene was not present in current network, but an important regulator of plant systemic acquired resistance, contains the same TE insertion in the second intron according to *P.taeda* v.1.0.

Table 1 Total number of extracted gene flanking regions and total number of hits to predicted LTRs.

Genome & gene set		Flanking region from the gene start/ end coordinates									
		5' 0-1Kb	3' 0-1kB	5' 1-2 kB	3' 1-2 kB	5' 2-3kB	3' 2-3kB	5' 3-4kB	3' 3-4kB	5' 4-5kB	3' 4-5kB
P.taeda v.2.0. all genes	Nb of extr.seq.	36726	36728	34711	34063	33184	32310	31767	30838	30349	29479
	Nb of hqh to LTRs	5851	6450	4362	3901	3750	3628	3310	3069	3202	2924
	ratio	0,16	0,18	0,13	0,11	0,11	0,11	0,1	0,1	0,11	0,1
	>50*	17	22	10	10	4	2	1	0	0	0
	>100*	8	9	1	0	0	0	0	0	0	0
P.taeda v.2.0. annotated genes	Nb of extr.seq.	15084	15057	14114	13793	13371	12912	12713	12192	11985	11569
	Nb of hqh to LTRs	816	773	800	732	875	968	1161	991	901	1000
	ratio	0,05	0,05	0,06	0,05	0,07	0,07	0,09	0,08	0,08	0,09
	>50	0	0	0	0	0	0	0	0	0	0
	>100	0	0	0	0	0	0	0	0	0	0
P.taeda v.1.0. HQ genes	Nb of extr.seq.	4298	4239	4177	4128	4130	4091	4081	4028	4023	3967
	Nb of hqh to LTRs	784	779	2258	1890	3151	2693	3593	3222	3816	3539
	ratio	0,18	0,18	0,54	0,46	0,76	0,66	0,88	0,8	0,95	0,89
	>50	1	1	1	0	0	0	0	0	0	0
	>100	0	0	0	0	0	0	0	0	0	0
P.taeda v.1.0. LQ genes	Nb of extr.seq.	75425	75459	72840	72797	71554	71470	70002	69836	68237	68017
	Nb of hqh to LTRs	2317	2540	4188	4243	4979	5070	5256	5387	5645	5382
	ratio	0,03	0,03	0,06	0,06	0,07	0,07	0,08	0,08	0,08	0,08
	>50	2	2	5	5	6	5	4	7	7	6
	>100	1	1	3	4	1	1	0	1	0	0
P.lambertiana v.1.0 HQ genes	Nb of extr.seq.	8779	8778	8746	8742	8719	8708	8692	8673	8660	8640
	Nb of hqh to LTRs	71	55	163	187	278	277	315	296	355	357
	ratio	0,01	0,01	0,02	0,02	0,03	0,03	0,04	0,03	0,04	0,04
	>50	0	0	0	0	0	0	0	0	0	0
	>100	0	0	0	0	0	0	0	0	0	0
P.lambertiana v.1.0 LQ genes	Nb of extr.seq.	71162	71157	70386	70475	69773	69909	69217	69344	68660	68836
	Nb of hqh to LTRs	470	466	1063	1011	1556	1508	1789	1368	2038	1999
	ratio	0,01	0,01	0,02	0,01	0,02	0,15	0,03	0,02	0,03	0,03
	>50	0	0	1	0	4	3	6	1	7	7
	>100	0	0	0	0	0	0	0	0	0	0

Main findings: MITE element

- The *P. taeda* v.1.01 genome and *P. lambertiana* gene-flanking regions were highly enriched with only one repeat, that; this was later identified as the *MITE3321* element (**Figure 1, 2**).

- The short *MITE3321* family identified in proximal gene flanking regions and introns could provide TATA boxes, and several DOF, ARR1, W-box, and GT-binding sites, which are important signals in plant transcription activation and stress-response regulation.

- *MITE3321* was inserted only into introns or in flanking gene regions, but never in both sites of any transcriptionally active gene within one genome. Therefore, insertion of *MITE3321* could not be only be explained by random transposition into transcriptionally active chromatin.

- Differences in predicted TFBS presence in *MITE3321* (10 bp insertion that disrupt W-box) could explain depletion of this TE in *P. lambertiana* gene 0–1 kB flanks and enhanced distribution in gene introns.

Table 2. *P.taeda* v.2.0 and *P.lambertiana* v.1.01 genes containing several *MITE3321* insertions.

Species	Genes ID with multiple *3321MITEs*	Insertion count	Description
P.taeda v.2.0.	PITA_12742	7	uncharacterized protein with domain of phosphoglucosamine mutase family protein
	PITA_21987	4	subtilisin-like protease SBT5.3
	PITA_00114	3	metal tolerance protein 11
	PITA_24114	2	probable xyloglucan endotransglucosylase/hydrolase protein B
	PITA_21327	2	60S ribosomal protein L8-1-like
	PITA_17959	2	TMV resistance protein N-like
	PITA_34859	2	3-oxoacyl-[acyl-carrier-protein] synthase I, chloroplastic-like isoform X1
	PITA_28894	2	L-gulonolactone oxidase 2 isoform X2
	PITA_00539	2	probable potassium transporter 11
	PITA_33316	2	plasma membrane intrinsic protein 2;8
	PITA_09881	2	cytokinin hydroxylase
P.lambertiana v.1.01. HQ genes	S/hiseq/c38458_g1_i1|m.23006	2	bifunctional phosphatase IMPL2, chloroplastic
	PILAhq_048992	2	putative clathrin assembly protein At4g40080
	PILAhm_002002	2	histone deacetylase 15 isoform X3

This research was supported by The State Education Development Agency 1.1.1.2. "Post-doctoral Research Aid". Nb. *1.1.1.2/VIAA/1//16/094.*

PostDoc Latvia
NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020
EIROPAS SAVIENĪBA
Eiropas Reģionālās attīsitības fonds
IEGULDĪJUMS TAVĀ NĀKOTNĒ