Latvian State Forest Research Institute «Silava»
Genetic Resource Centre

Variation of mobile genetic elements in pine genes and flanking regions

***Angelika Voronova*, Martha Rendon, Pär Ingvarsson, Dainis Ruņģis**

PostDoc Latvia

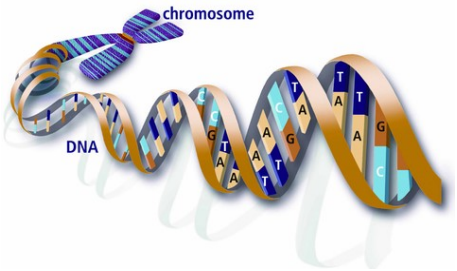
NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020

EIROPAS SAVIENĪBA
Eiropas Reģionālās attīsītības fonds

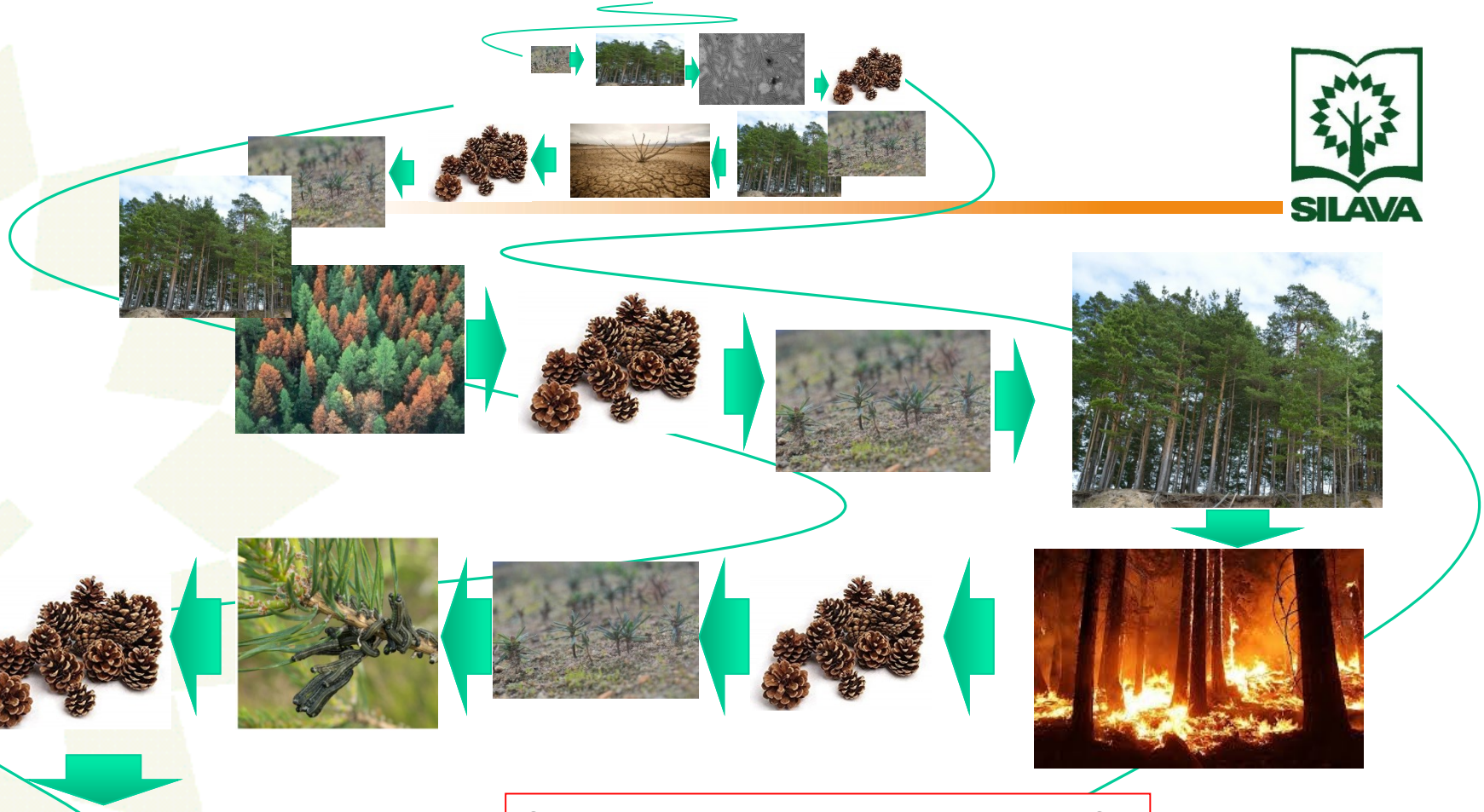
IEGULDĪJUMS TAVĀ NĀKOTNĒ

The 62nd International Scientific Conference of Daugavpils University, 29.05.2020

Hidden inheritance: defence against stress



STRESS

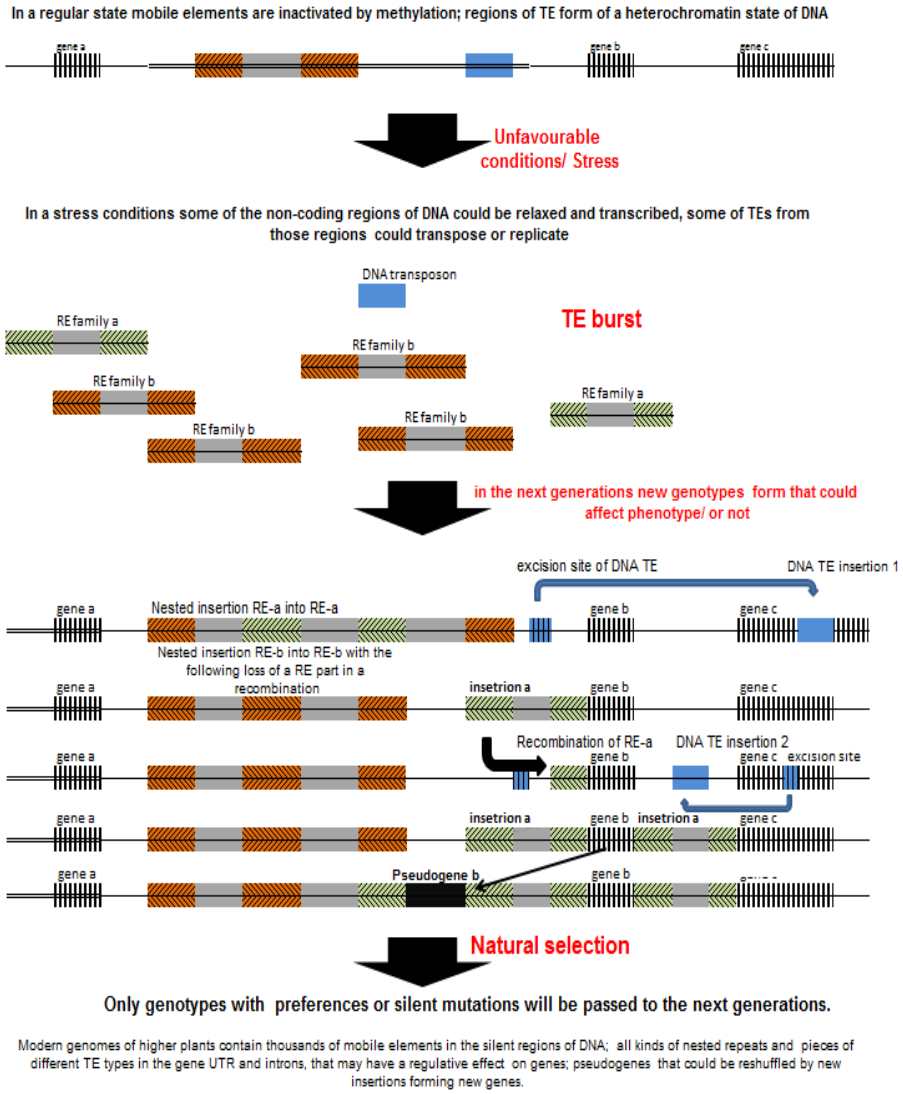


? What will be the new stressor?

? How to select the best genotype without reducing genetic diversity?

? Is it possible to find molecular markers for selection of plants with increased adaptability?

Transposable elements are drivers of genome evolution & dynamic change

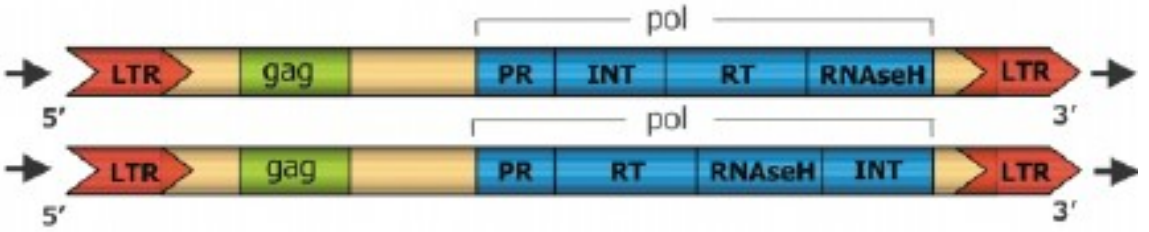


Class I transposable elements or Retrotransposons

LTR Retrotransposons

Ty1-*copia* group

Ty3-*gypsy* group



Non-LTR Retrotransposons

LINE

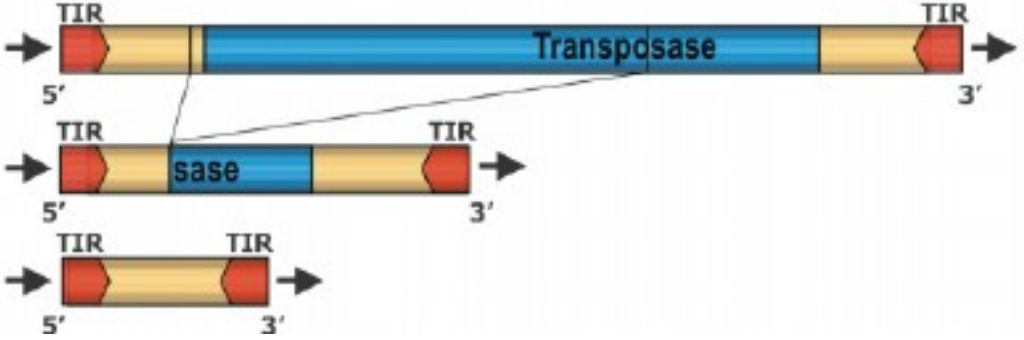
SINE

Class II transposable elements

Autonomous element

Non-autonomous element

MITE



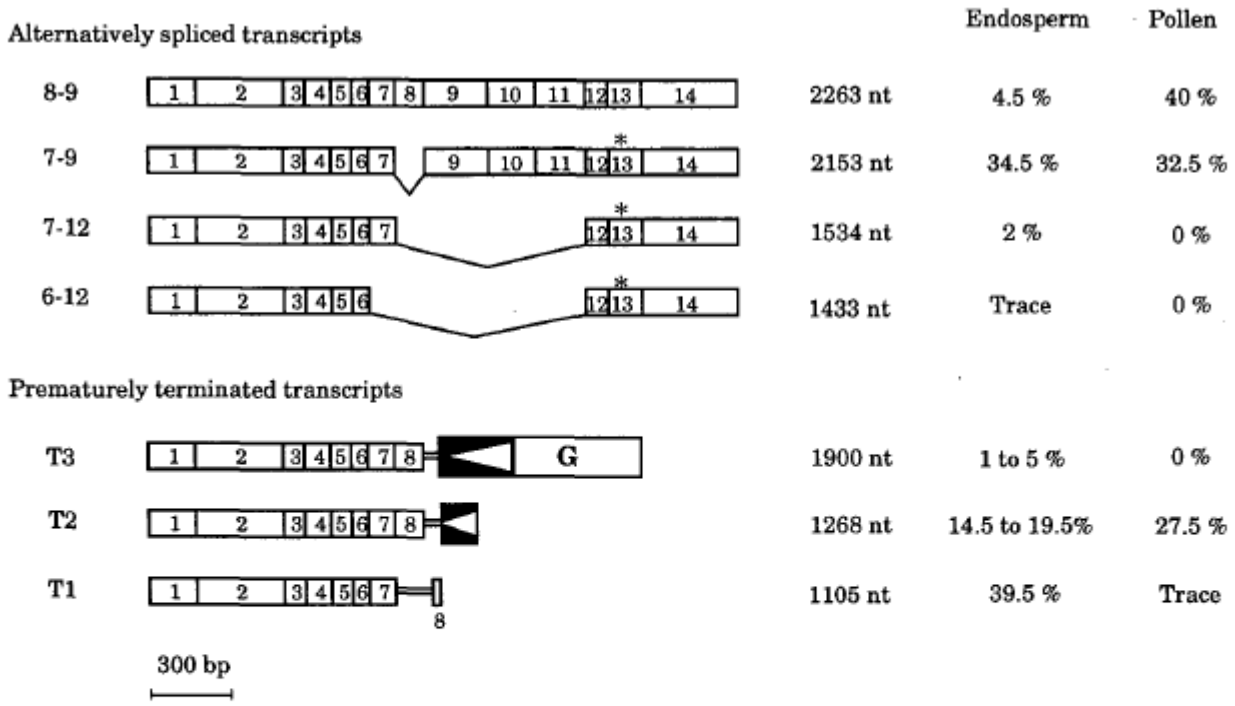
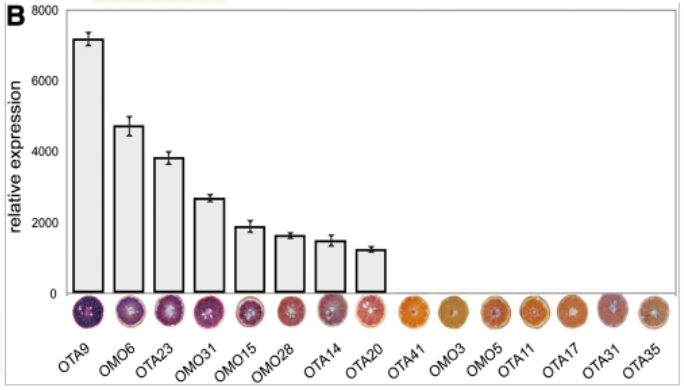


Figure 5. Summary of *wx*G-Encoded Transcripts in Endosperm and Pollen.

wx exon sequences are represented by open boxes with numbers. Filled boxes with open triangles represent LTR sequences from the G element. The open box labeled G represents internal element sequences. Asterisks denote the position of premature stop codons. The relative amount of the transcripts in each tissue is shown at right. These values were obtained from RNase protection analysis (transcripts 8 to 9 and T1 in endosperm; transcripts 8 to 9, 7 to 9, and T2 in pollen), from RNA gel blot analysis, or 3′ RACE products (transcripts 7 to 12 and 6 to 12 in endosperm; transcript T1 in pollen), by estimating the relative intensity of the 3′ RACE products from *wx*G tissues after agarose gel electrophoresis (transcript T3 in endosperm; transcripts 7 to 12, 6 to 12, and T3 in pollen), by subtracting 2% (the amount for alternatively spliced transcripts 7 to 12) from 36.5% (the sum of transcripts 7 to 9 and 7 to 12, obtained by RNase protection analysis; transcript 7 to 9 in endosperm), and by subtracting 1 to 5% (amount of transcript T3 in endosperm) from 19.5% (the sum of transcripts T2 and T3 as estimated from RNase protection analysis; transcript T2 in endosperm).

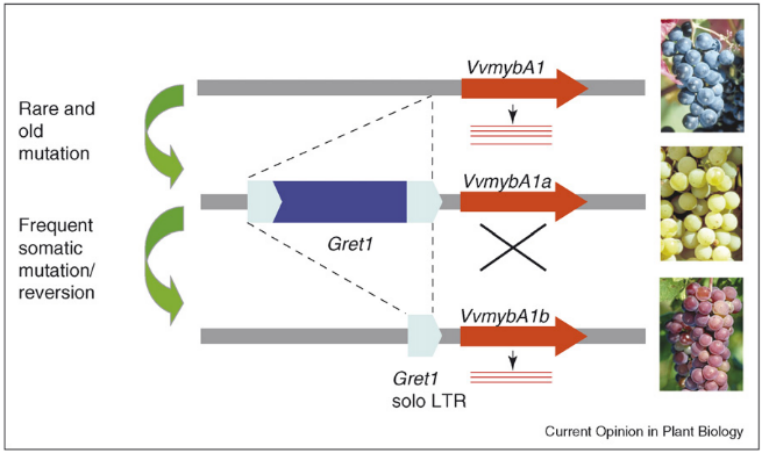
Sylvestre Marillonnet and Susan R. Wessler Retrotransposon Insertion into the Maize waxy Gene Results in Tissue-Specific RNA Processing. The Plant Cell, Vol. 9, 967-978, June 1997

Retrotransposons influence expression of important genes

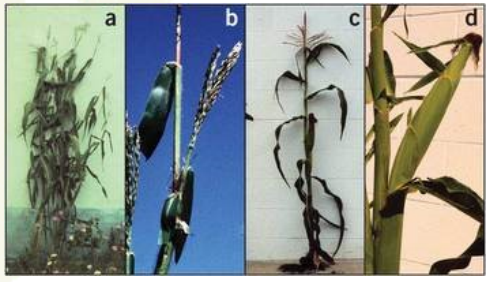


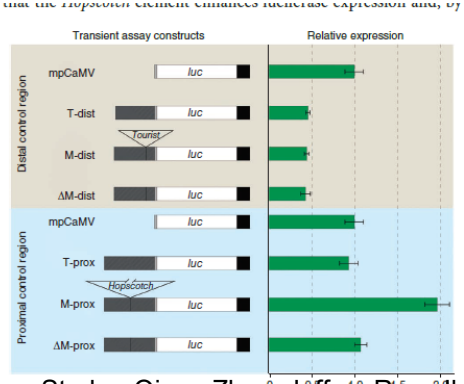
Expression Analysis of *Ruby*.

Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges (Butelli et.al. 2012), The Plant Cell, 2012 July 2012, 24 (7)



Kobayashi S, Goto-Yamamoto N, Hirochika H: Retrotransposon- induced mutations in grape skin color.Science2004,304:982.





Nature Genetics 43, 1160–1163 (2011) Anthony Studer, Qiong Zhao, Jeffrey Ross-Ibarra & John Doebley

MITEs short non-autonomous transposable elements that are often found populated in genes

Table 2 MITEs derived trait variations in plants

Crop	Trait	References
Pea	Seed shape	[32]
Maize	Flowering time	[33, 34]
Sorghum	Aluminium tolerance	[35]
Groundnut	Oleic acid	[10]
Potato	Tuber skin colour	[36]
Gentian	Petals colour	[37]
Rice	Leaf angle and seed size	[45]
Rice	Disease resistance	[44]
Rice	Glume shape	[17, 38]
Wheat	Heat tolerance	[39]
Maize	Seedling drought tolereance	[40]
Rice	Agronomic traits	[41]

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.

TE could form stress-responsive gene networks

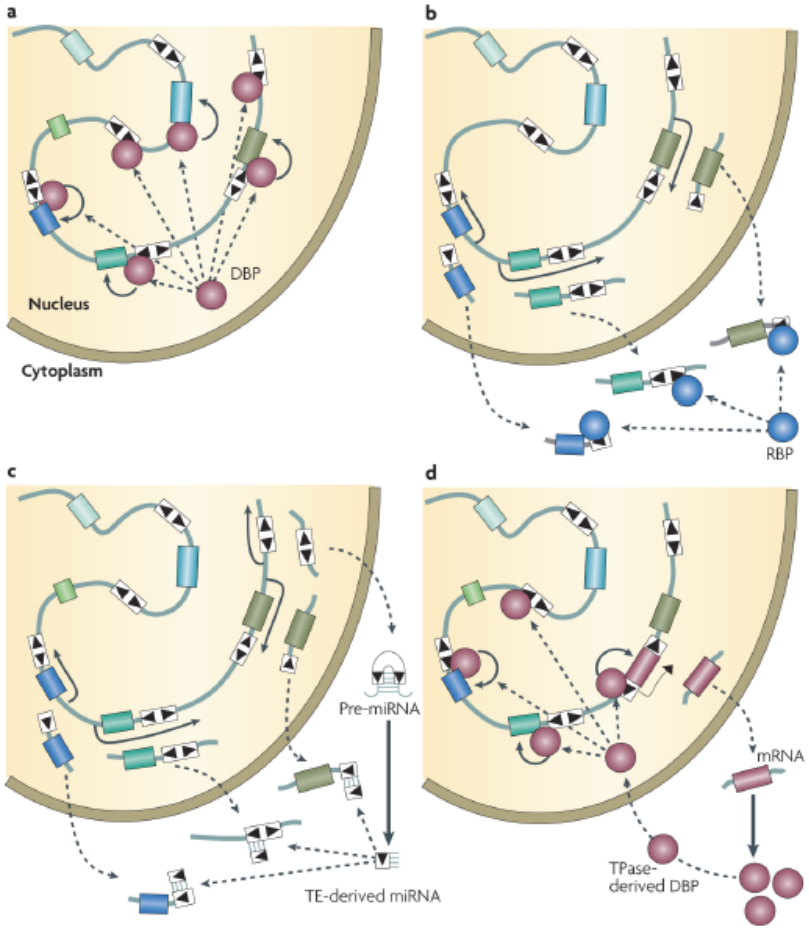
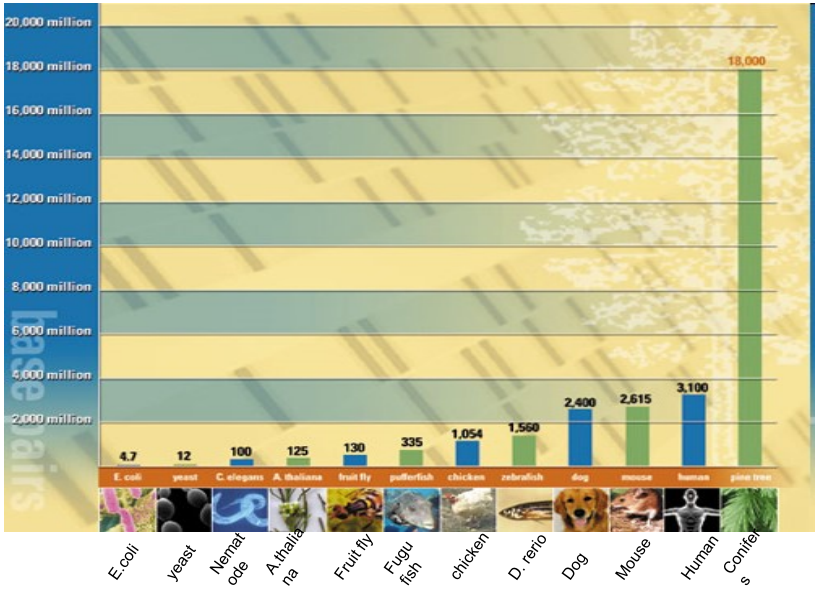


Figure 2. Building regulatory systems with transposable elements
A family of DNA transposons is shown, with its multiple copies (white boxes) delimited by terminal inverted repeats (black triangles) and interspersed with genes (color boxes) in the genome. For panels A and B, the TE family could be also a retrotransposon family. *A: Wiring of a transcriptional regulatory network by TE-derived cis-elements.* A binding site for a DNA binding protein (DBP) has been dispersed throughout the genome as part of the TE. If the DBP

Feschotte *et al.* 2008

Genome expanded by the prolifiration of TEs



http://www.genome.duke.edu/research/highlights/environmental/forestry-genomics.php

*Under stringent conditions (99% identical), only 24% of the *P.taeda* genome was estimated to be repetitive, while under more permissive conditions (75% identical), 80% of the genome was estimated to be repetitive (Kovach *et al.* 2010).

Mobile genetic elements in the genomes of gymnosperms

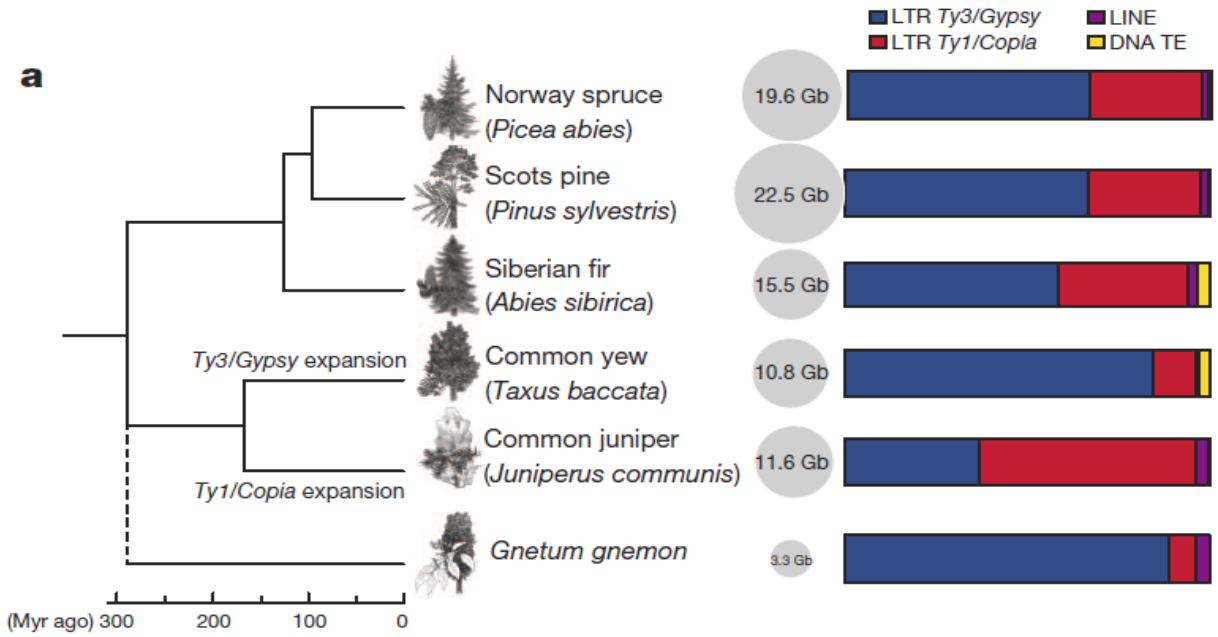
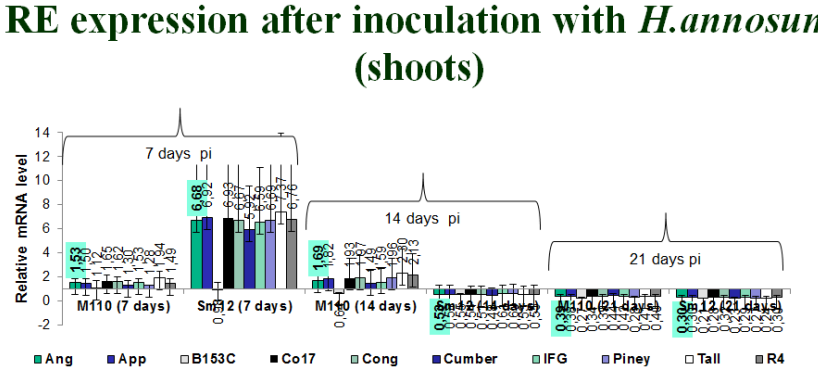
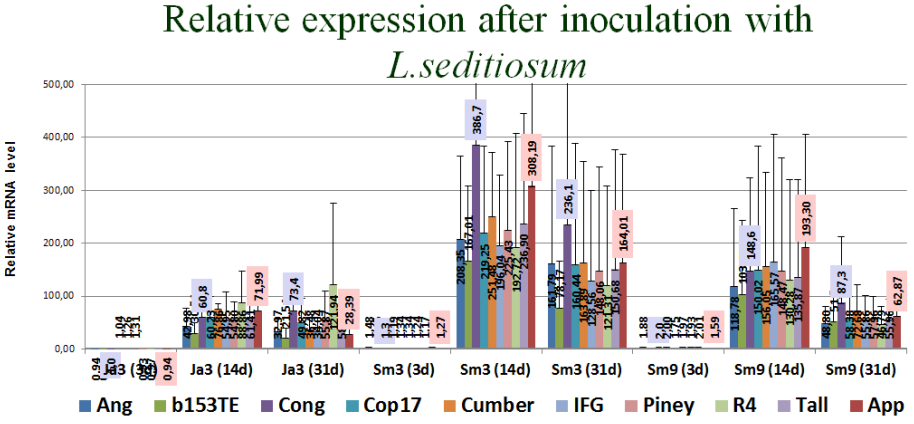
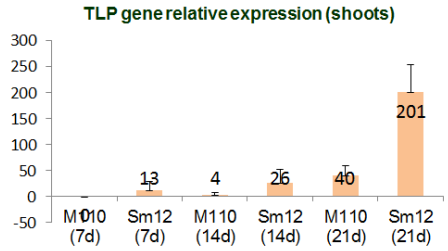


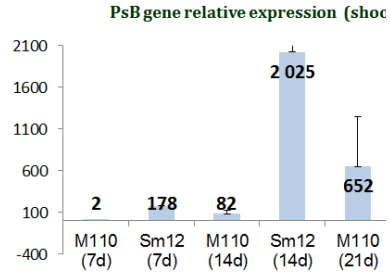
Figure 2. Conifer genomes contain expansions of a diverse set of LTR-RTs.
Distribution of different classes of transposable elements from six gymnosperm species. The figure is based on the total fraction of transposable elements (TE) identified and grouped into different classes from the different species. Genome sizes of the six species are given in circles and their phylogenetic relationship is shown, with tentative dating of divergence times (x-axis) based on 64 chloroplast genes over 39 species and five fossil calibration points. (Nystedt *et al.* 2013).

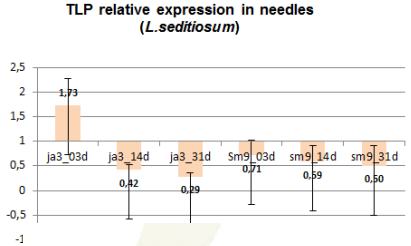
Retrotransposons expression











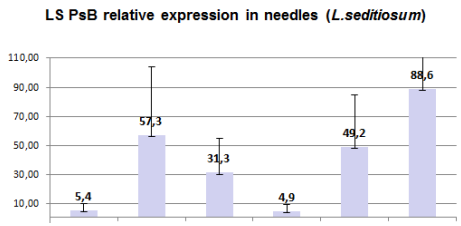


Figure 1. a) example of *in vitro* inoculation with *H.annosum* culture suspension. b) Scots pine seedlings (Sm12) 21 days after inoculation with *H.annosum* (the first 4 tubes represent uninnoculated controls)

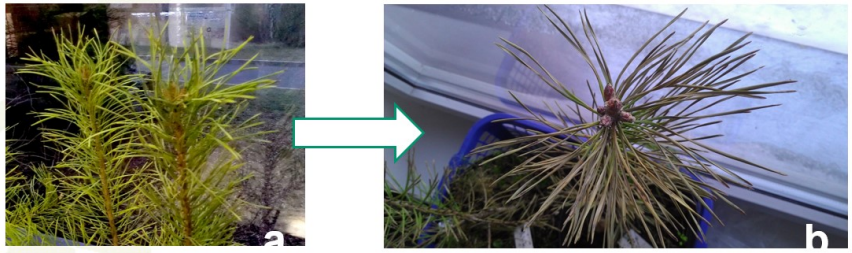


Figure 2. a) two year old grafted *P.sylvestris* clones were used in the *L.seditiosum* inoculation experiment. b) Scots pine ramets one month after inoculation with *L.seditiosum.*

CNVs among eight Scots pine tree genomes

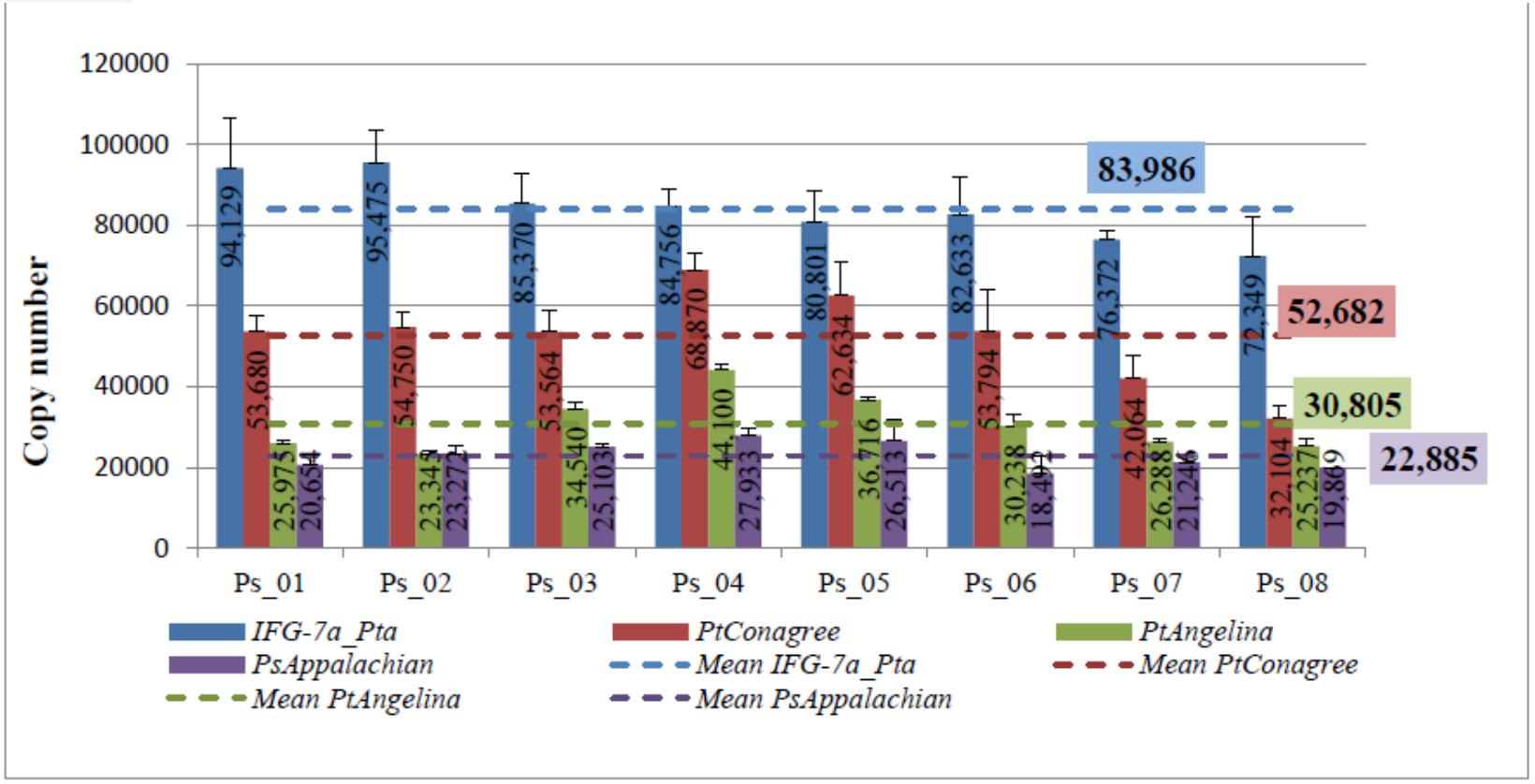


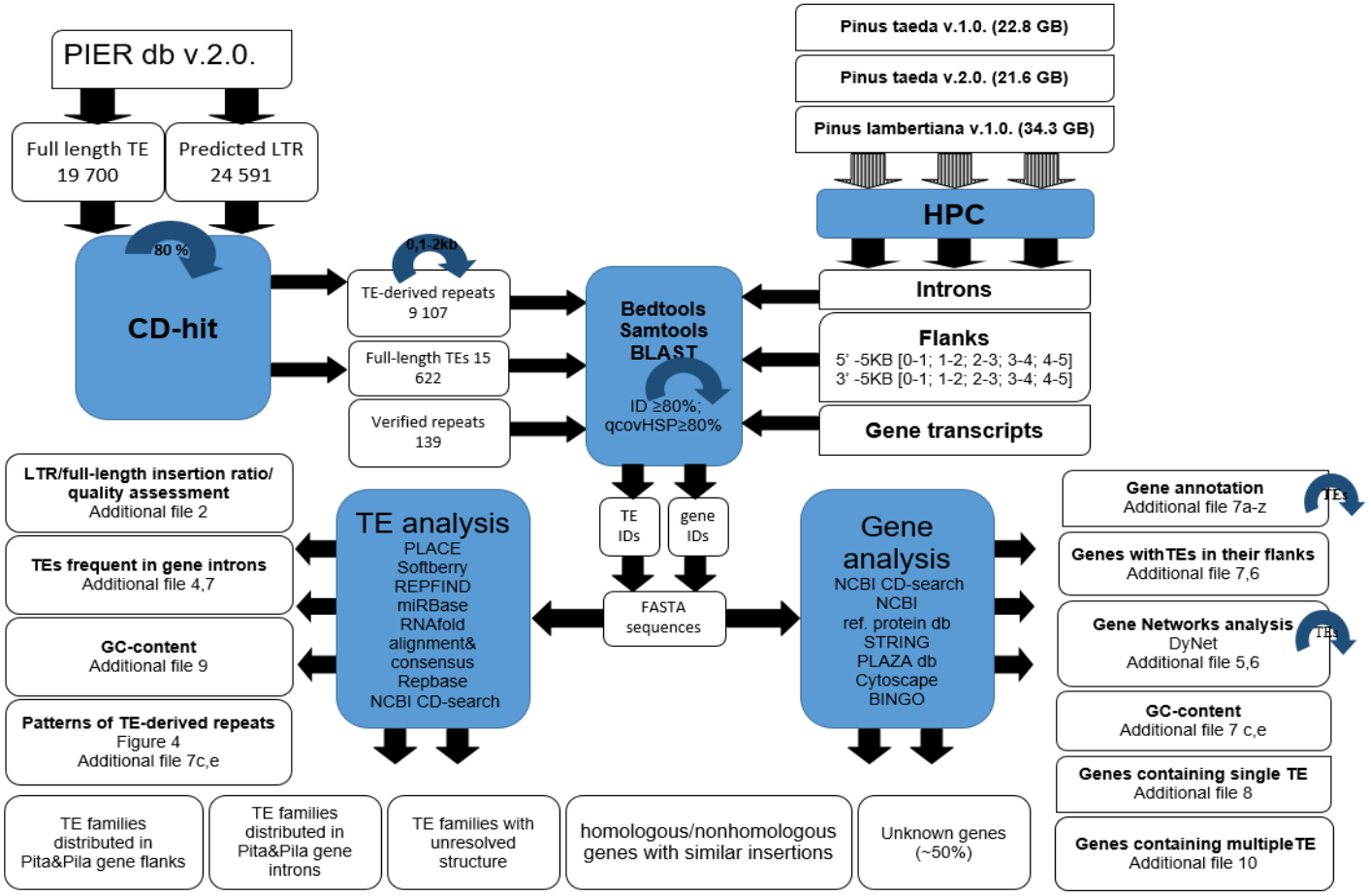
Figure 6. Copy number variation of high copy RE families among eight Scots pine tree genomes. Mean copy numbers among individuals indicated on the right.

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.

Aim of the study

- analyze the distribution of TEs in genes and gene-flanking regions in the available pine reference genomes (*Pinus taeda* and *Pinus lambertiana*).
- explore the possibility of transferring this information to non-model pine species (*P. sylvestris)* genome studies.
- evaluate if the distribution of TEs in gene regions is random regarding different gene regions (e.g. flanks or introns).
- whether genes containing similar TE families are involved in similar processes.
- whether found TEs contain potential gene regulatory motifs.

Overview of analysis workflow



* blue arrows indicate data filtering steps with parameters noted

PIER- Pine Interspersed Element Resource; HPC- High Capacity Computer Resources; LTR- Long Terminal Repeat; TE-Transposable elements

Pinus taeda/ P. lambertiana/ P.sylvestris

Pinus taeda v.2.01: 6,58 GB; 36 730 genes; 2.9 million contings

Pinus taeda v.1.0: 16.5 million contings

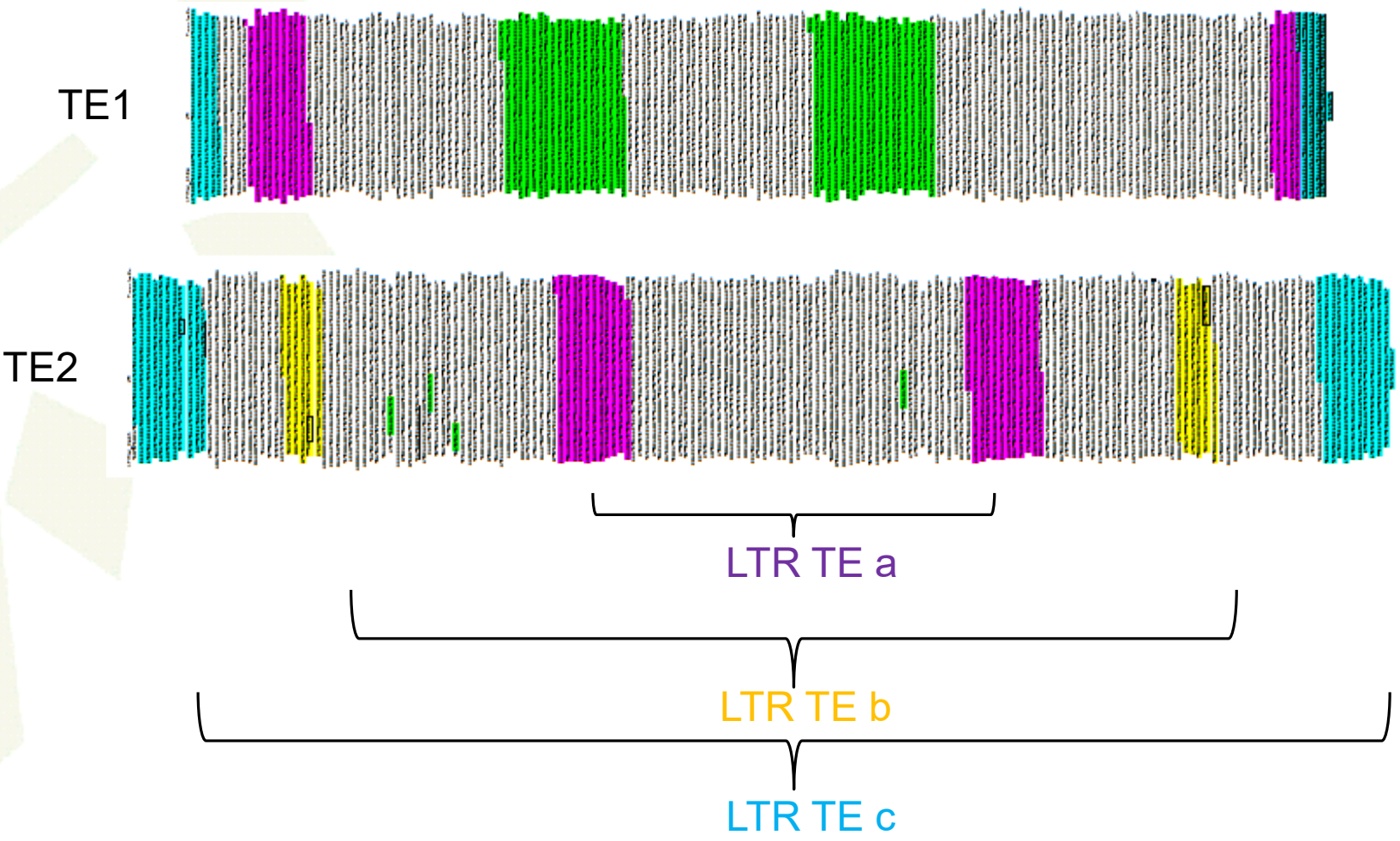
1 scaffold=1gene +“non-coding” sequences

Pinus lambertiana v.1.0.: HQ genes-8 779; LQ genes- 71 167

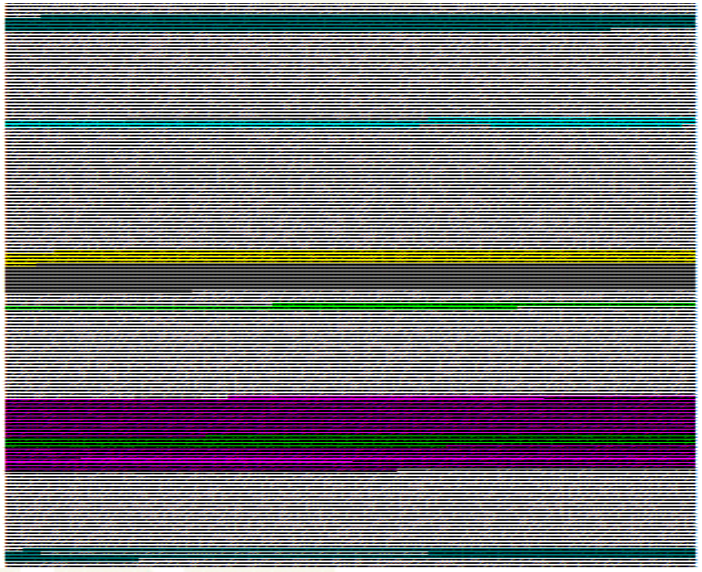
Pinus sylvestris unannotated scaff (no repeats, 12 737 exons, from them 2021 -whole contig)

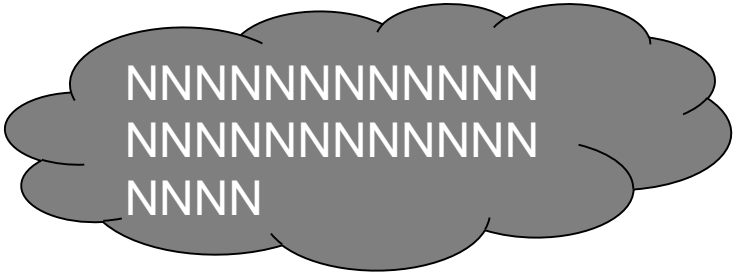
Nested repeats

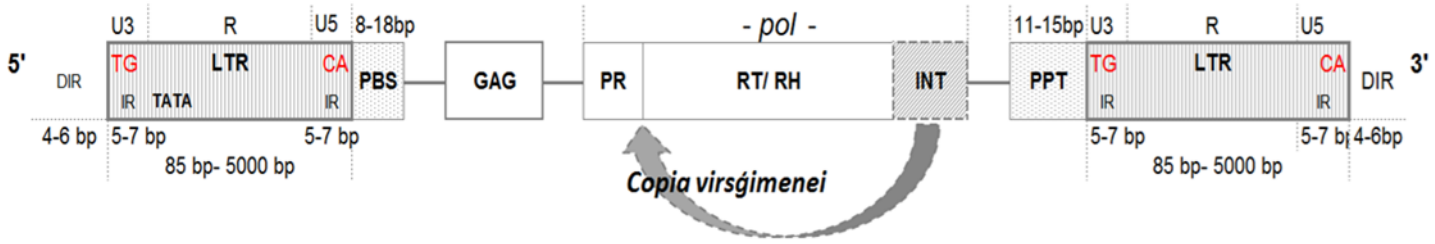
TE db (19 700→15 622, length from 257 to 35 042 bp)
Total CPU time for clustering 126419.09
Predicted LTR db (24 591-- 9 659) Total CPU time 1515.90



Full-length or only a part-?





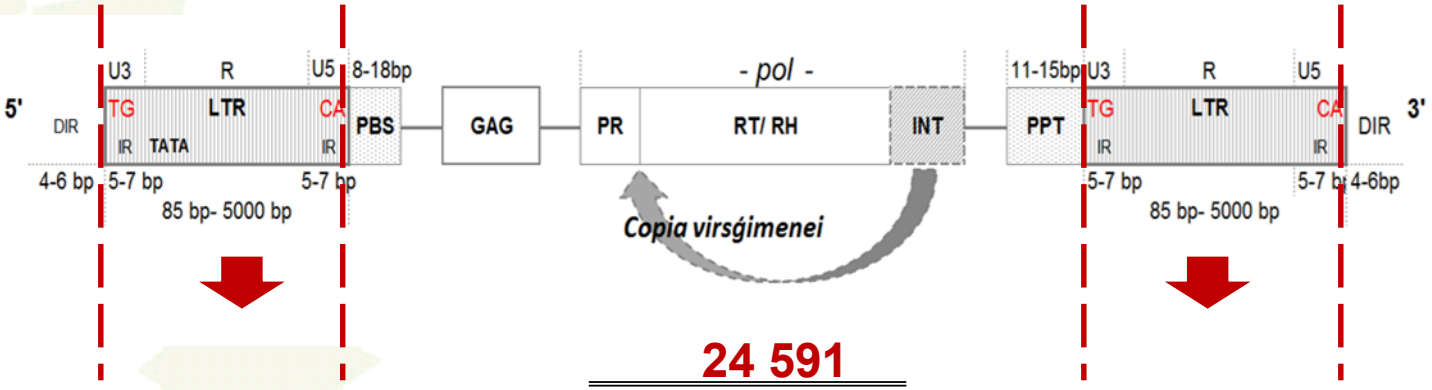


LTR/internal sequence ratio	RE name	>70% qq	>80%qq	>90%qq	Common distribution within introns
P.lambertiana v.1.01. HQ genes	IFG7_I	0,45	0,14	0,13	internal
	IFG-7a_PTa-I	0,36	0,09	0,02	internal
	PtAppalachian_I	2,47	2,31	2,07	full length
	PtPineywoods_I	3,33	2,80	0,48	single LTR/full length
P.taeda v.2. genes	IFG7_I	2,49	3,08	3,81	single LTR/full length
	IFG-7a_PTa-I	2,29	2,66	2,35	full length
	PtAngelina_I	3,60	4,33	13,00	single LTR
	PtAppalachian_I	2,47	2,19	1,88	full length
	PtBastrop_I	2,60	2,60	3,33	single LTR/full length
	PtCumberland_I	0,80	0,77	0,86	internal
	PtOuachita_I	1,25	-	-	internal
	PtPineywoods_I	2,82	1,91	0,20	full length

Giga-genome & repetitive transposable element analysis

- Nature of conifer genomes-large, full of divergent repetitive sequences, pseudogenes and gene families;
- Different research groups apply different workflow&quality indicators for the assembly & annotations;
- Two versions of *P.taeda* genome contain TEs with differing structure due to the technical (conting length) differencies.
- Automated annotation results in overestimated TE families nb., nested repeats
- Short-read sequencing &assembly did not allowed for correct TE assembly, contings are ending in the repeats
- Gene annotation files (genomic coordinates of a gene) could contain errors;
- Based on sequence simmilarity with known plant genes only 50% of genes could be annotated.

New strategy:
searching for the short TE derived repeats



24 591

CD-HIT

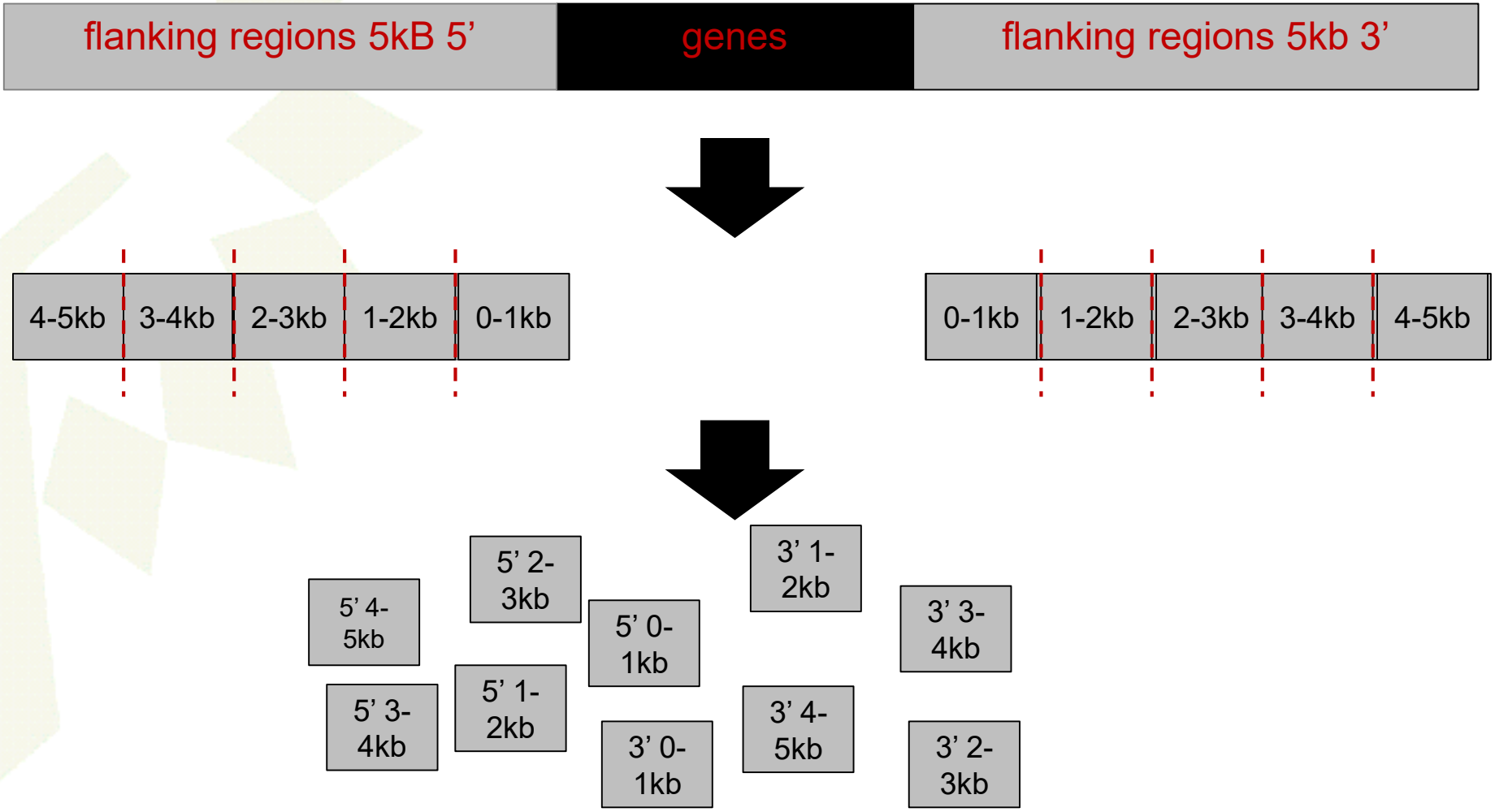
9 659

0.1–2 kb

9 107

Not only LTRs!!!

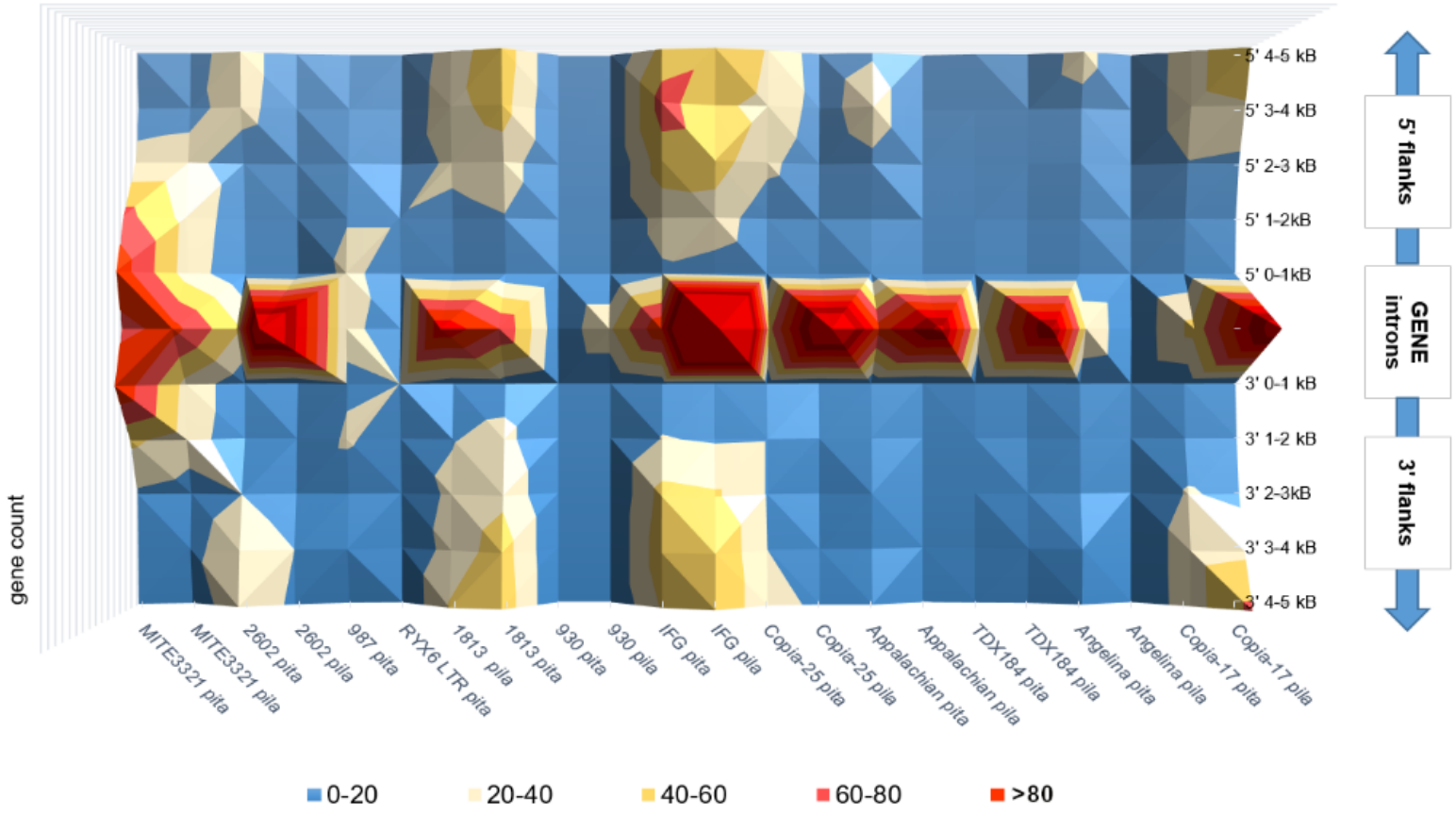
Analysis of flanking gene regions



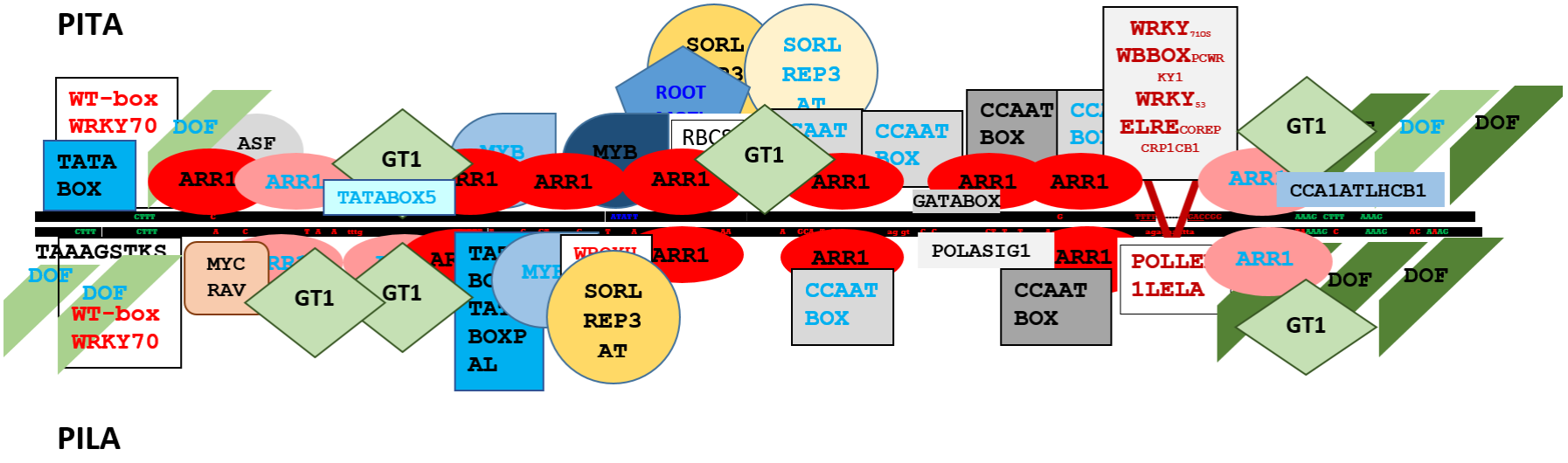
How LTRs are distributed regarding proximity to genes?

Genome and gene set		genes Flanking region from the gene start/end coordinates									
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										
	>100										
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										
	>100										
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										
	>100										
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										
	>100										
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										
	>100										
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										
	>100										

How LTRs are distributed regarding proximity to genes?

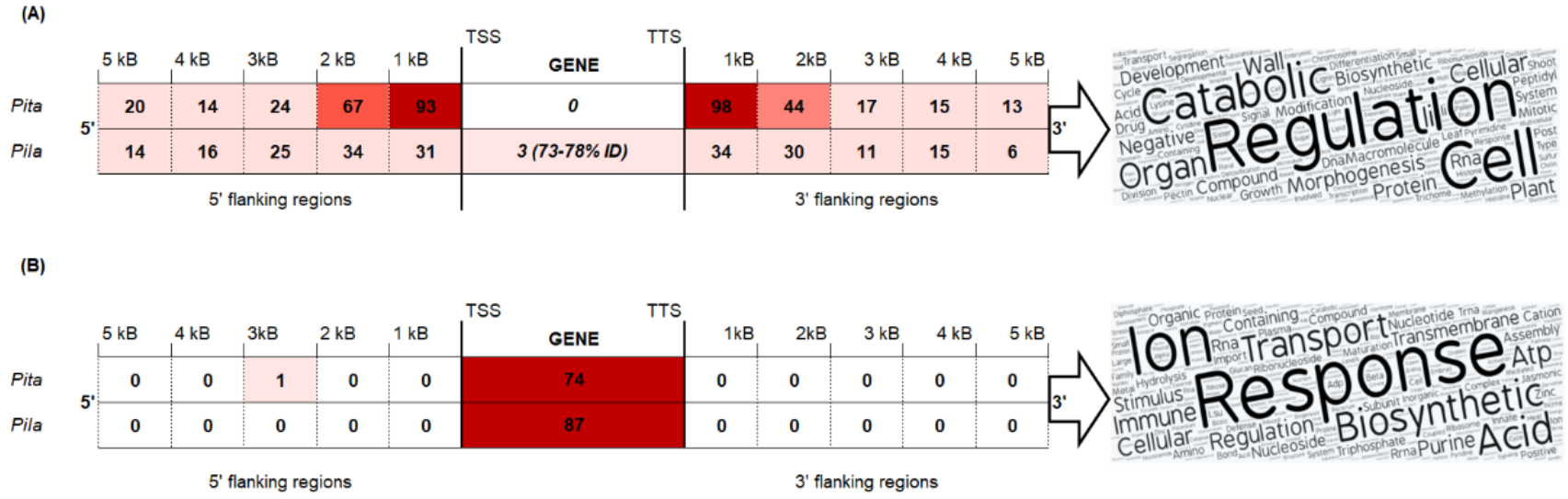


Alignment of *P.taeda* (PITA) and *P.lambertiana* (PILA) consensus sequences with predicted plant cis-acting regulatory elements




```
ARR1AT-pita-10; pila-7;
CAATBOX1-4; 2;
DOFCOREZM-4; 5;
GT1CONSENSUS -3; 3
```

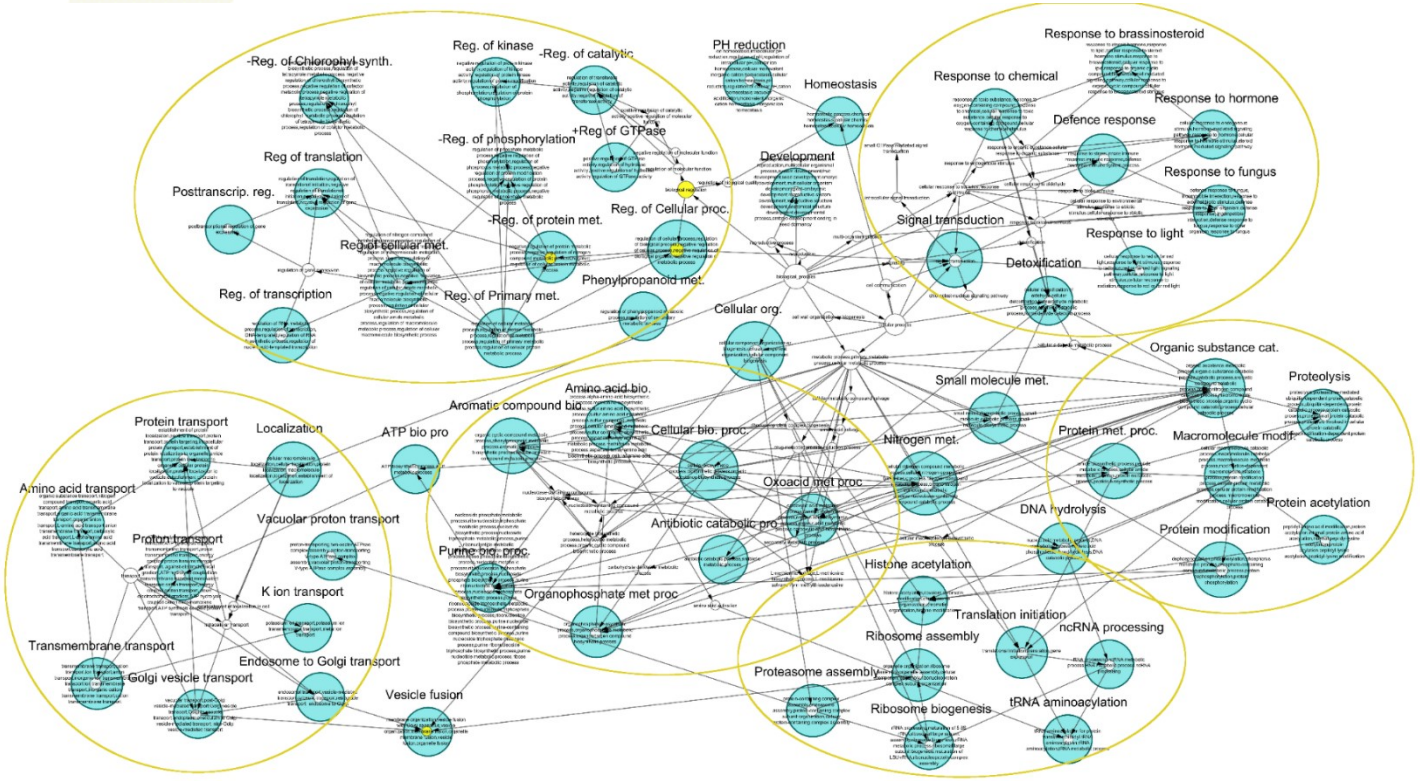
MITE3321 family distribution among gene regions



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

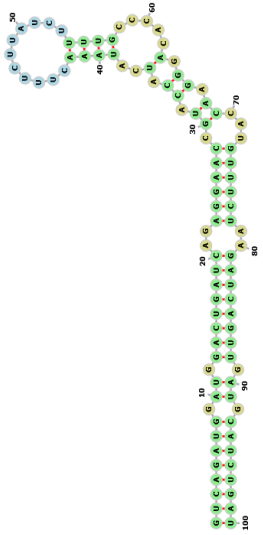
Species	Genes ID with multiple *3321MITEs*	Insertion count	Description
P.taeda v.2.0.	PITA_12742	7	uncharacterized protein with domain of phosphoglucosamine mutase family protein
	PITA_21987	4	subtilisin-like protease SBT5.3
	PITA_00114	3	metal tolerance protein 11
	PITA_24114	2	probable xyloglucan endotransglucosylase/hydrolase 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

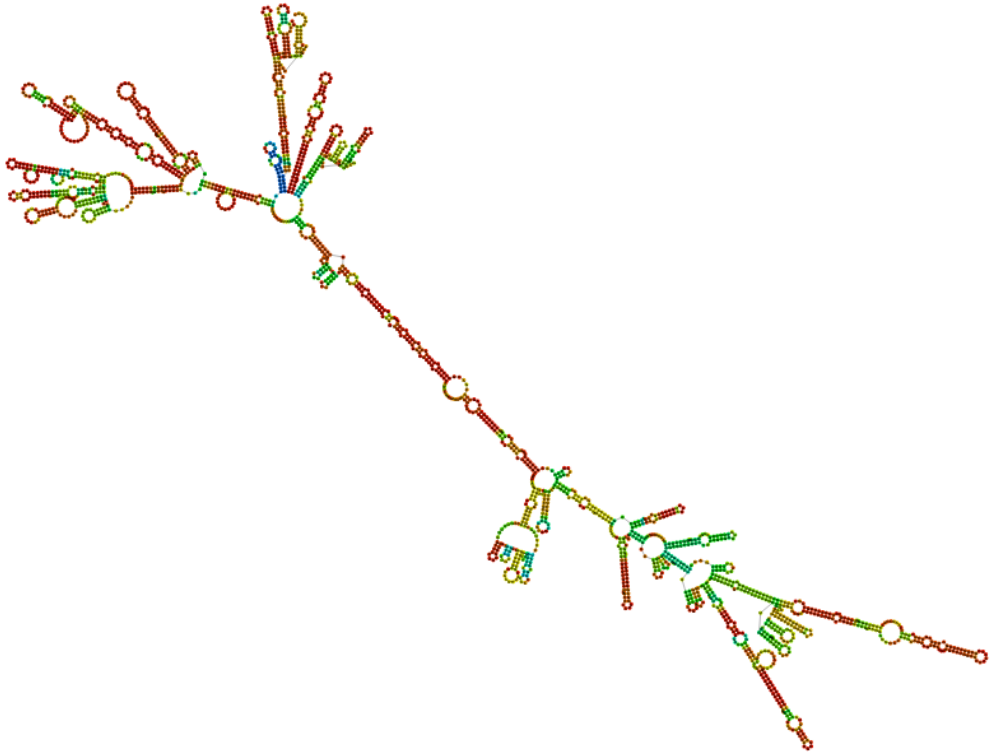
DNA transposon *184DTX found in important stress-responsive gene introns*



NPR1 (Nonexpresser of Pathogenesis-related proteins-1);
histone-binding PHD1 finger protein ALFIN-like 4 coding gene;
COPII-coated ER to Golgi transport vehicle SNARE-like 13 gene
eukaryotic translation initiation complex 2B
PSMD4, a 26S proteasome non-ATPase regulatory subunit gene

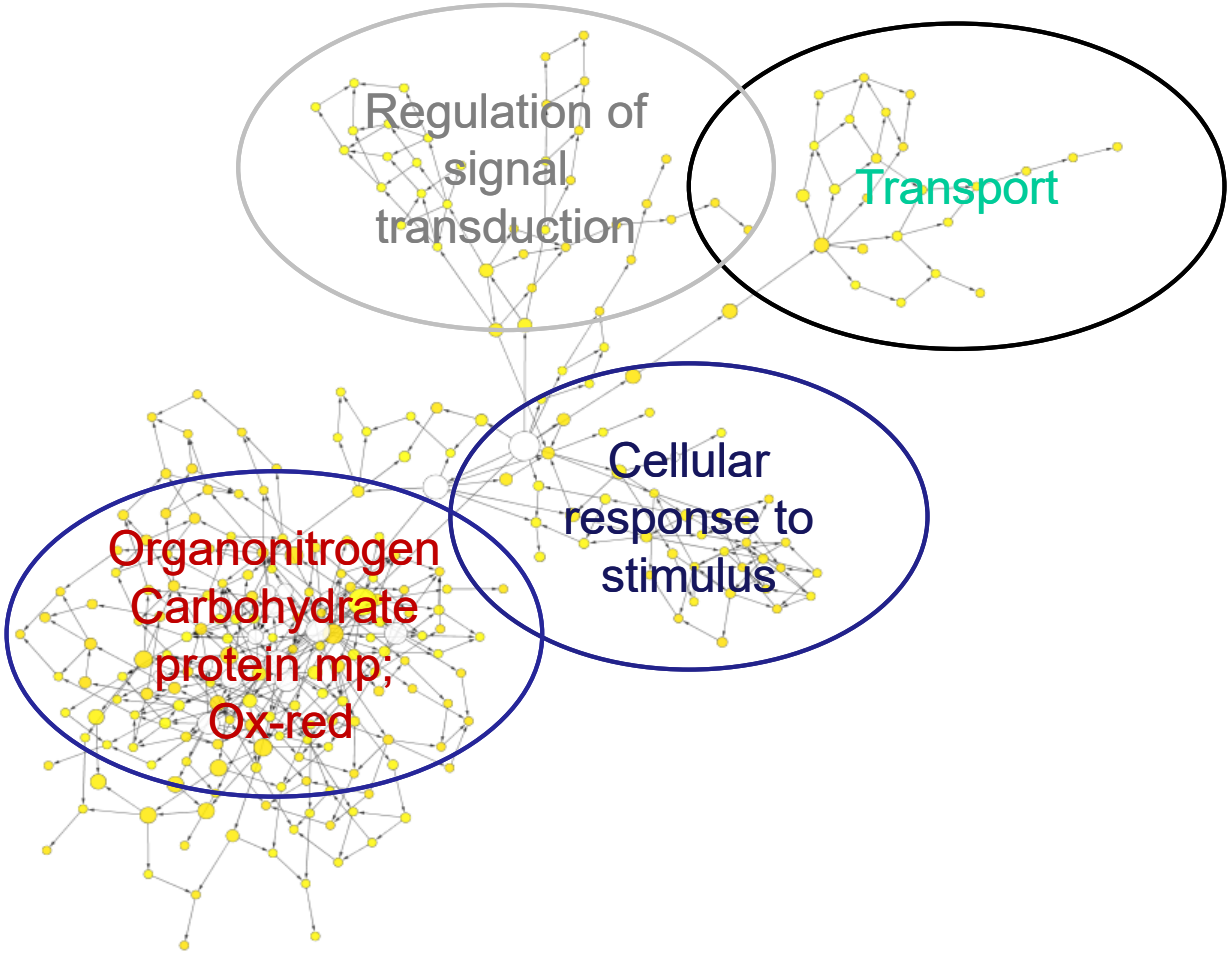
DNA transposon *184DTX* could form mature microRNA and contain microRNA target site



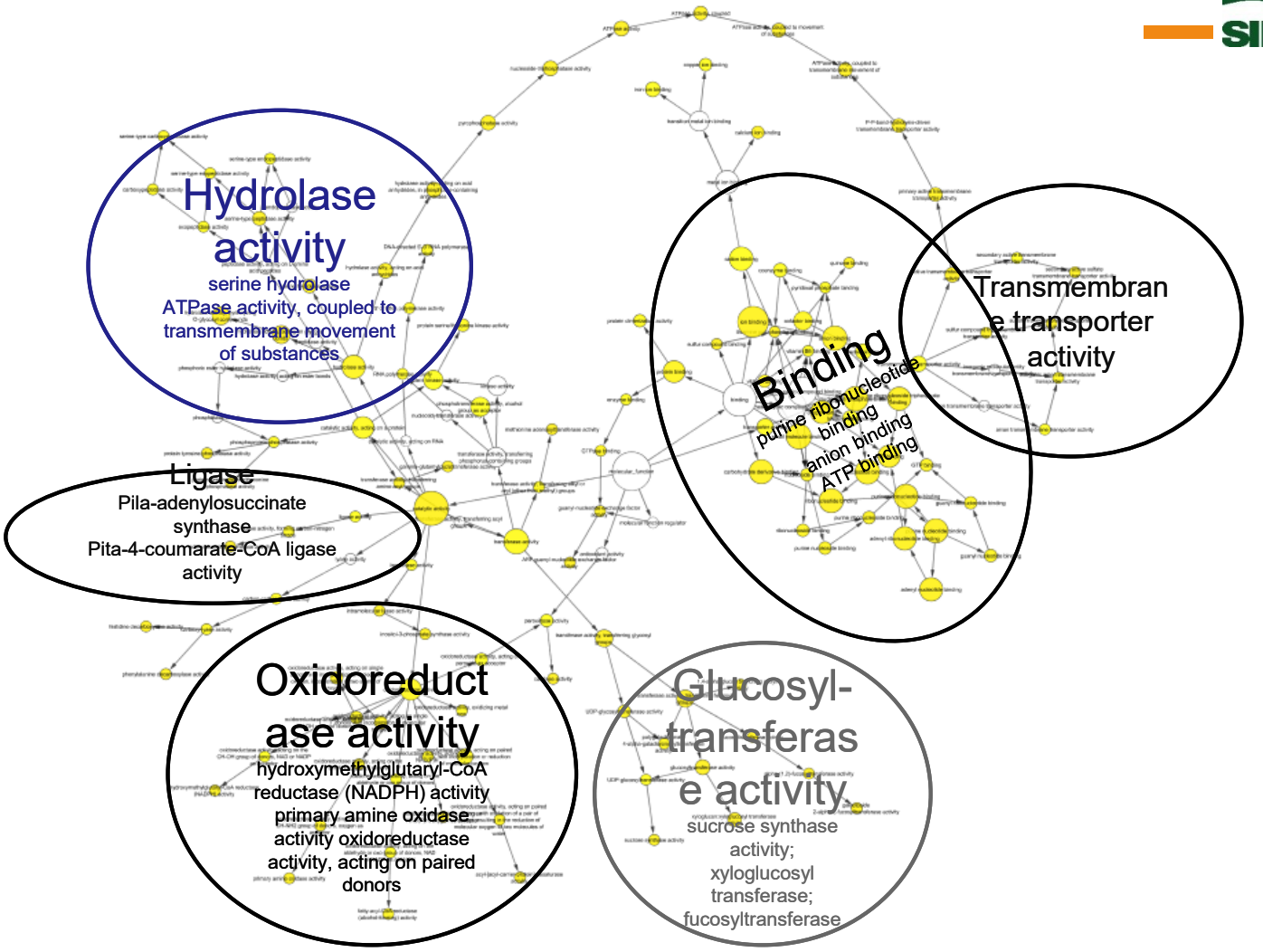


- Start Position : 48
- End Position : 147
- Sequence Size : 100 nucleotides
- Minimum Free Energy : -37 kcal/mol

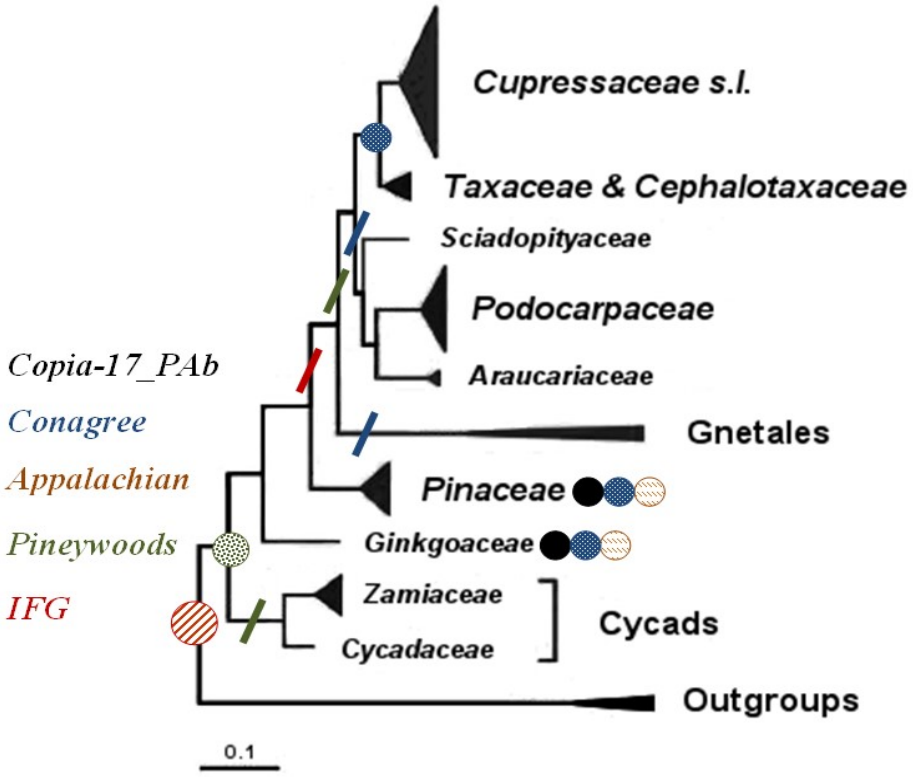
Gene network dependant on insertion of one TE



Gene network dependant on insertion of one TE



IFG retrotransposon



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.

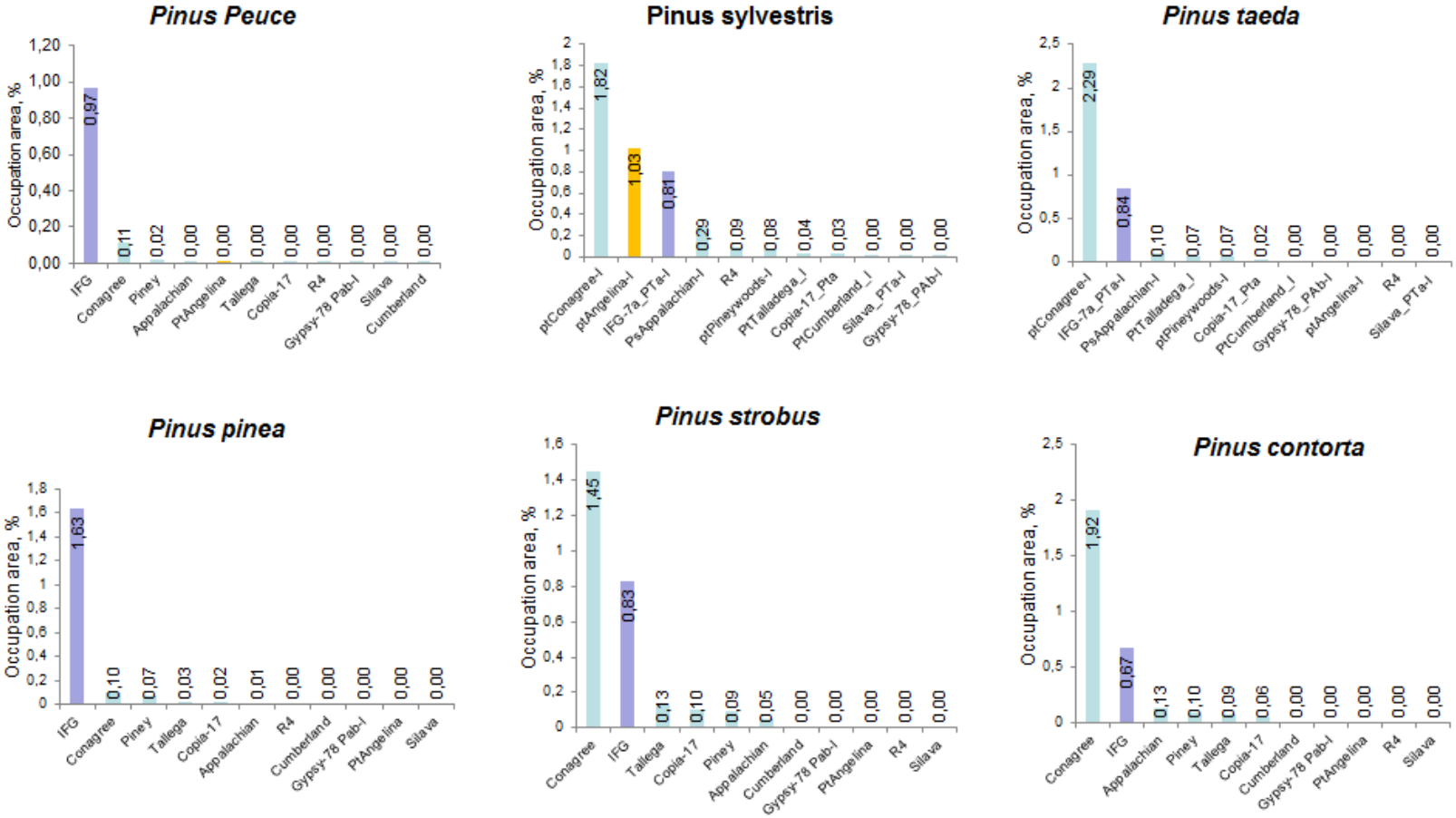


Figure 1. Occupation area (%) of RE families relative to species average genome size.

It was assumed that each estimated copy represents full-length element. Estimation of the copy number of eleven REs was performed using Real-time PCR absolute quantification with Maxima SYBR Green/ROX qPCR Master Mix (*Thermo Scientific*) reagents and StepOne software v.2.2.2 (*Applied Biosystems*). Plasmids with cloned RE sequences were used for standard curves (6 dilutions 1:10; 3 replicates), for plasmid with a known insert sequence, molecular weight was calculated using the Sequence Manipulation Suite: DNA Molecular Weight (Stothard, 2000). Plasmid copy number was calculated using the formula: copy nb.= (amount, ng) * Avogadro nb. (6.022*10^23)/ 1*10^9*(mol weight, Da). Copy number of each RE was calculated relative to the amount of DNA analysed and the genome size (2C) of the various species.

Copia-1813 RLX resides gene introns and flanks

The *P. taeda* LTR contained the following two AG-rich tracts:

$(AGNN)_3$**$(AG)_3$**$(NNAG)_2$ and $(AGNN)_2(AGN)_4$.

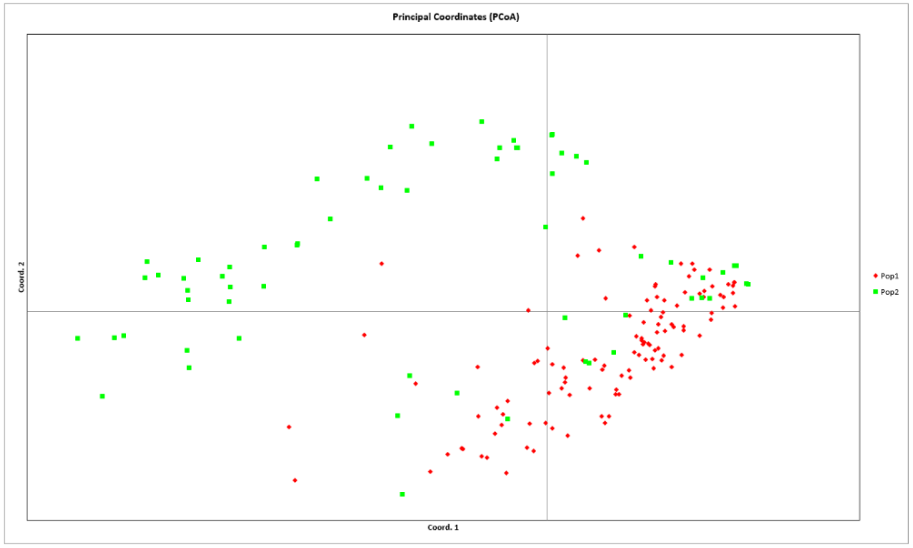
The *P. lambertiana* LTR also contained polypurine-rich motifs 25 bp apart: $AA(AGG)_2A_3(AGG)_2GA_3AGG$ and $GAG(AGG)_3AGA(AG)_3$.

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 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.

Average GC content for introns considering 1-kb hits was 39% for *P. taeda* and 41% for *P. lambertiana*, respectively.

Copia-1813 RLX-network TE patterns embedded in gene introns



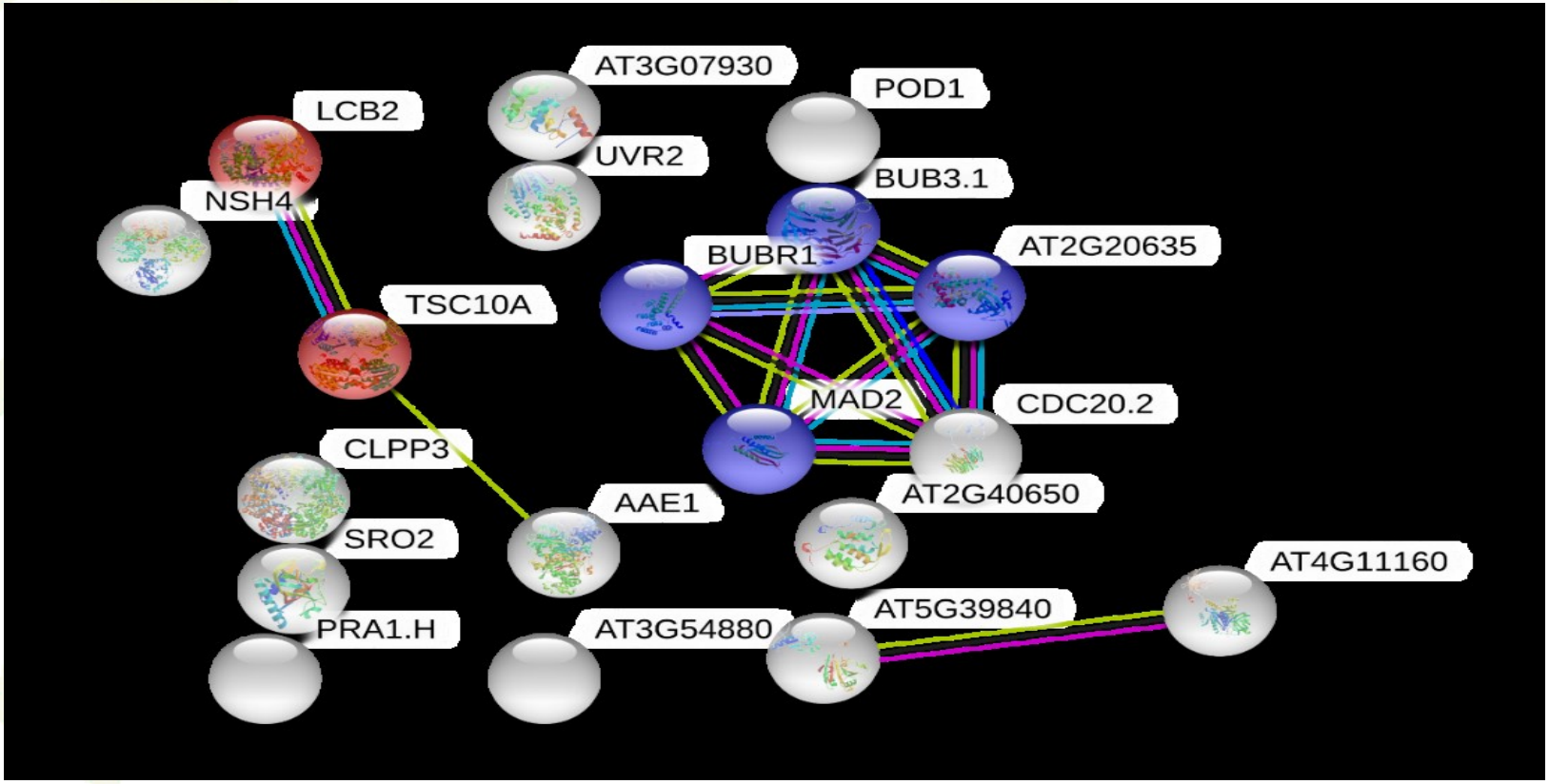
***Copia-1813* RLX + *DTX184* TE** was found within introns of seven *P. lambertiana* genes. Products of these genes were found in different cell compartments and are involved in protein folding in ER (oxidation), positive regulation of RNA export from the nucleus, protein heterodimerization, and SYM-1 stress responsive protein from yeast; the function of this protein is not yet described in plants.

***Copia-1813+2602+Copia*-25** Tes are involved in pH regulation in Golgi, tethering of vesicles to Golgi membranes, nuclear protein import, and intracellular protein transport.

Copia-1813+Copia-2602+Copia25+IFG was found in the following two genes: insulinase (involved in protein targeting to mitochondrion) and histone deacetylase 15 (tag for epigenetic repression).

Genotype ("C") match for the three *P. lambertiana* genes: two of them were annotated as splicing factor 3A subunit 3 genes and one as a cleavage and polyadenylation specificity factor subunit 5-like coding genes. Products of these genes are involved in pre-mRNA maturation and splicing according to the UniProt Knowledgebase.

Copia-1813 RLX-network TE patterns embedded in gene introns



STRING build gene network from recognized gene names (ref. Arabidopsis thaliana) from P.lambertiana genes containing repeats of single *Copia-1813* family. Edges connect genes that are coexpressed, found interacting and mentioned together in other publications.

Repeat rich node genes in *Pinus taeda*.

pita-v2-Ids	LTRsh Nb.	Best hit NCBI, Database Name-refseq_protein; Description- NCBI Protein Reference Sequences; Program-BLASTX 2.9.0+ Citation	hit accession	Conserved domains name	Domain Accession	Domain description	GO annotations	annotations count	co-accuring terms (Based on Entire Annotation set)
PITA_00338	65	two-pore potassium channel 3-like isoform X2	XP_010275702.1	Ion_trans_2	pfam07885	ion channel; This family includes the two membrane helix type ion channels found in bacteria.	GO:0005267	122030	659
						Enables the facilitated diffusion of a potassium ion (by an energy-independent process) involving passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism.			
PITA_00504	55	1,4-alpha-glucan-branching enzyme 2-2, chloroplastic/amyloplastic isoform X1	XP_007204282.1	PLN02447	PLN02447	1,4-alpha-glucan-branching enzyme	GO:0003844	777713	123
PITA_01345	43	GTPase Der	XP_008449721.1	PRK00093; P-loop_NTPase super family	PRK00093; cl21455	GTP-binding protein Der;; P-loop containing Nucleoside Triphosphate Hydrolases; Members of the P-loop NTPase domain superfamily are characterized by a conserved nucleotide phosphate-binding motif, also referred to as the Walker A motif (GxxxxGK[S/T], where x is any residue), and the Walker B motif (hhhh[D/E], where h is a hydrophobic residue). The Walker A and B motifs bind the beta-gamma phosphate moiety of the bound nucleotide (typically ATP or GTP) and the Mg2+ cation, respectively. The P-loop NTPases are involved in diverse cellular functions, and they can be divided into two major structural classes: the KG (kinase-GTPase) class which includes Ras-like GTPases and its circularly permutated YlqF-like; and the ASCE (additional strand catalytic E) class which includes ATPase Binding Cassette (ABC), DExD/H-like helicases, 4Fe-4S iron sulfur cluster binding proteins of NifH family, RecA-like F1-ATPases, and ATPases Associated with a wide variety of Activities (AAA). Also included are a diverse set of nucleotide/nucleoside kinase families.	GO:0005525		4224
PITA_00128	41	S-formylglutathione hydrolase isoform X1	XP_006843471.1	PLN02442	PLN02442	S-formylglutathione hydrolase	GO:0018738		41
PITA_01309	41	cytochrome P450	XP_012078717.1	p450 super family	cl12078	Cytochrome P450; Cytochrome P450s are haem-thiolate proteins involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyze regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures.	GO:0005490; GO:0004497		646
PITA_01333	37	B3 domain-containing transcription repressor VAL2 isoform X1	XP_006841783.1	B3_DNA;	cd10017	Plant-specific B3-DNA binding domain; The plant-specific B3 DNA binding domain superfamily includes the well-characterized auxin response factor (ARF) and the LAV (Leafy cotyledon2 [LEC2]-Abscisic acid insensitive3 [ABI3]-VAL) families, as well as the RAV (Related to ABI3 and VP1) and REM (REproductive Meristem) families. LEC2 and ABI3 have been shown to be involved in seed development, while other members of the LAV family seem to have a more general role, being expressed in many organs during plant development. Members of the ARF family bind to the auxin response element and depending on presence of an activation or repression domain, they activate or repress transcription. RAV and REM families are less studied B3 protein famillies.	GO:0016564; GO:0001227		2,526
				zf-CW	pfam07496	CW-type Zinc Finger; This domain appears to be a zinc finger. The alignment shows four conserved cysteine residues and a conserved tryptophan. It was first identified by, and is predicted to be a "highly specialized mononuclear four-cysteine zinc finger...that plays a role in DNA binding and/or promoting protein-protein interactions in complicated eukaryotic processes including...chromatin methylation status and early embryonic development." Weak homology to pfam00628 further evidences these predictions (personal obs: C Yeats). Twelve different CW-domain-containing protein subfamilies are described, with different subfamilies being characteristic of vertebrates, higher plants and other animals in which these domain is found.			
PITA_00372	33	serine/threonine-protein kinase GRIK1	XP_011627097.1	STKc_LKB1_CaMKK	cd14008	Catalytic domain of the Serine/Threonine kinases, Liver Kinase B1, Calmodulin Dependent Protein Kinase Kinase, and similar proteins; STKs catalyze the transfer of the gamma-phosphoryl group from ATP to serine/threonine residues on protein substrates. Both LKB1 and CaMKKs can phosphorylate and activate AMP-activated protein kinase (AMPK). LKB1, also called STK11, serves as a master upstream kinase that activates AMPK and most AMPK-like kinases. LKB1 and AMPK are part of an energy-sensing pathway that links cell energy to metabolism and cell growth. They play critical roles in the establishment and maintenance of cell polarity, cell proliferation, cytoskeletal organization, as well as T-cell metabolism, including T-cell development, homeostasis, and effector function. CaMKKs are upstream kinases of the CaM kinase cascade that phosphorylate and activate CaMKI and CamKIV. They may also phosphorylate other substrates including PKB and AMPK. Vertebrates contain two CaMKKs, CaMKK1 (or alpha) and CaMKK2 (or beta). CaMKK1 is involved in the regulation of glucose uptake in skeletal muscles. CaMKK2 is involved in	GO:0004674		4,418

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

LTR Nb. *pita*	**LTR Nb. *pila***	**Description**	**Accession,** [h]-homologous genes	**Conserved domain name**	**Accession**	**GO terms**
24	19	plastidial pyruvate kinase 2	XP_006843356.1[h]	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	SrmB	COG0513	RNA secondary structure unwinding
21	21	phospholipid:diacylglycerol acyltransferase 1	XP_006849611.1[h]	PLN02517	PLN02517	acylglycerol biosynthetic process
18	20	nuclear pore complex protein NUP62-like/GPCR-type G protein 1 isoform X2	XP_024396806.1	SMC_prok_B super family	cl37069	RNA export from nucleus; protein import/export into/from nucleus; nucleocytoplasmic transport, localization
13	23	WD repeat-containing protein WRAP73	XP_008798782.1	WD40 super family	COG2319	-
	19	protein RAE1	XP_028076289.1		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.

Conclusions I

- 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.

- Only several homologous genes were revealed in the most studied networks between pine species, indicating that most transposition events occurred after separation of the species. 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.

Conclusions II

- 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.

- 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 highly correlated expression levels of RLX families within individuals found previously.

- 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 gene networks 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).

Conclusions III

- 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.

- DNA TE 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).

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

- No genes with several MITE3321 insertions in its different non-coding regions (flanks and introns) were identified. Genes containing MITE3321 insertions in different regions were associated with different biological proceses. This non-random distribution suggests formation of differentially regulated gene sub-networks, depending on the location of MITE insertions.

- 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

Conclusions IV

- 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; 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.

- 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.


Thank you for your attention!