Investigation of transposable element derived polymorphisms in the Scots pine (*Pinus sylvestris* L.) genome

Angelika Voronova & Dainis Rungis

Genetic Resource Centre, Latvian State Forest Research Institute "Silava", 111 Rigas street, Salaspils, Latvia, LV-2169

e-mail: angelika.voronova@silava.lv

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, and contain large inter-gene regions and a high proportion of repetitive sequences. Up to 62% of the sequenced loblolly pine genome (*Pinus taeda*) consists of retrotransposon (RE) sequences and 70% of these are Long Terminal Repeat REs (Neale *et al*., 2014). Conifers have genes with large introns (up to 100 kb), that contain various types of mobile genetic elements. 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 sequences co-expressed with stress associated genes as part of their introns. It has been reported that transposable element 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). RE transposition in stress conditions could lead to novel mutations and formation of novel genetic pathways. The aim of future study is identification of Scots pine genes that are associated with transposable element derived structural variations in the genome; investigation of transposable element derived polymorphism within pine breeding germplasm; and their influence on expression of stress-induced genes.

In our previous studies transcripts of various RE-like sequences were detected in Scots pine during infestation with pine woolly aphids and abiotic stress (Voronova *et al.* 2014). Composition and copy number variation of each detected RE were estimated for Scots pine genome and compared with conifer genomes with available sequences (Voronova *et al*., 2017). Expression of these REs in response to infection with pathogenic fungi (*Heterobasidion annosum*) was analysed in two *P.sylvestris* seedling families. First results displayed strong correlation of expression of 8 studied REs families in each seedling, later correlation analyses between data for 11 seedling families confirmed that results (Figure 2). Transcription level of each RE family couldn't be explained with family abundance of particular RE in the *P.sylvestris* genome (Table 1). There were difficulties to detect relative expression for the low copy number family «*Silava*» as expression in the control was undetectable. Higher expression of distributed RE family «*Conagree*» was found among all samples, however family with the highest copy numbers «*IFG*» (in average 30,000 more copies in the *P.sylvestris* genome) displays even lower levels of transcription. RE family «*Cumberland*» with only 165 ±16,7 copies in average displays similar levels of transcription after inoculation with pathogen. Therefore disproportionately higher variation in the distribution of each RE family exists in the genome of *P.sylvestis* and transcript abundance could not be explained by co-expression of random loci in the genome. Additionally, RE families composition difference in various pine species were observed (Fig.1.), as well as individual variation of RE families in *P.sylvestis* genome.

Table 1. Comparison of RE copy number variation in *Pinus sylvestris* and *Pinus taeda* relative to diploid genome size.

RE name	RE size, bp	*Pinus sylvestris* 2C=44949 Mbp			*Pinus taeda*- 2C=43228 Mbp		
		CN ± SD	Occup., Mbp	%	CN ±SD	Occup., Mbp	%
Conagree	*15 550*	52 682 ±10633,6	819,21	1,8225	63 576 ±2138	988,60	2,2869
Angelina	*14 992*	30 805 ±6674,1	461,83	1,0274	9 ±0,30	0,13	0,0003
IFG-7a	*4 322*	83 985 ±7439,3	362,99	0,8076	84 422 ±1787,7	364,87	0,8441
Appalachian	*5 652*	22 885 ±3157,8	129,35	0,2878	7 284 ±384,7	41,17	0,0952
R4	*10 227*	3 916 ±685,9	40,05	0,0891	11 ±1,2	0,11	0,0003
Pineywoods	*5 253*	7 171 ±1039,9	37,67	0,0838	5 779 ±521,9	30,36	0,0702
Talladega	*15 394*	1 155 ±147,2	17,78	0,0395	2 013 ±52,8	30,98	0,0717
Copia-17	*5 904*	2 532 ±317,4	14,95	0,0333	1 767 ±79,1	10,43	0,0241
Cumberland	*8 951*	165 ±16,7	1,48	0,0033	104 ±5,6	0,93	0,0021
Gypsy-78	*6 592*	116 ± 20	0,76	0,0017	95 ± 3.70	0,63	0,0014
Silava	*7 692*	4 ±0,47	0,03	0,0001	3 ±0,24	0,02	0,0001

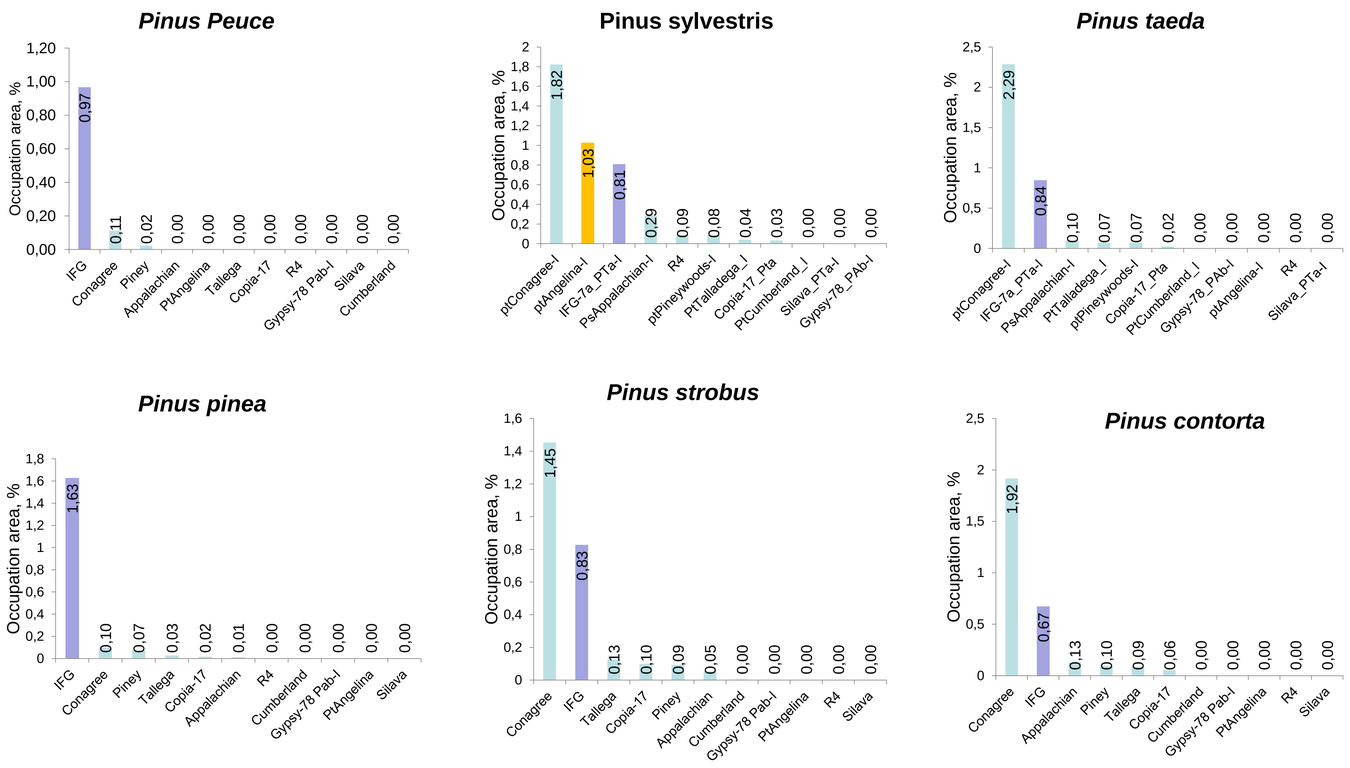


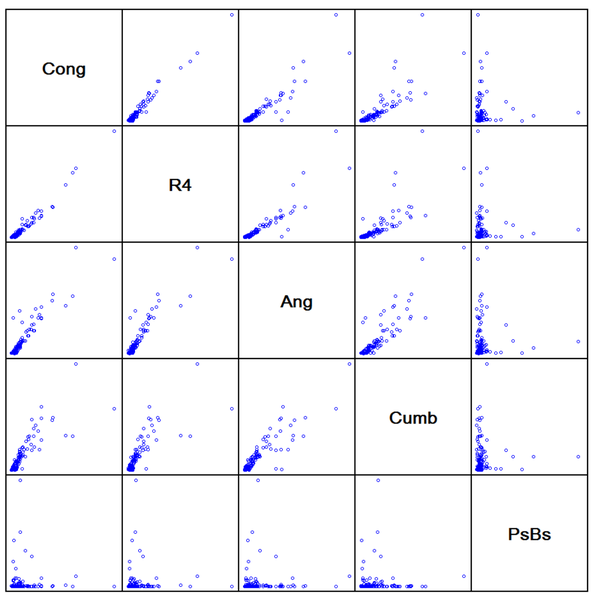
Figure 1. Occupation area (%) of RE families relative to species average genome size.
It was assumed that each estimated copy represents full-length element. Estimation of the copy number of eleven REs was performed using Real-time PCR absolute quantification with Maxima SYBR Green/ROX qPCR Master Mix (*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.

Table 2. Summary statistics for MRPP of experimental groups.
T-test statistic; A- chance-corrected within-group agreement; p- probability of a smaller or equal delta.

Group	Tissue	δ under null hypothesis				T	p	A
		Observed δ	Expected δ	Variance	Skewness			
Infection	root	0.2325	0.4599	0.68810841E-05	-2.42	-86.67	0.00000000	0.4943
	needle	0.2703	0.4599	0.70307401E-05	-2.24	-71.48	0.00000000	0.4121
Treatment	root	0.2073	0.4599	0.29774968E-04	-1.09	-46.29	0.00000000	0.5492
	needle	0.2303	0.4599	0.30422541E-04	-1.00	-41.62	0.00000000	0.4991
Seedling family	root	0.4418	0.4599	0.74323444E-04	-0.69	-2.09	0.03293225	0.0392
	needle	0.4317	0.4599	0.75939897E-04	-0.64	-3.22	0.00450801	0.0611
Damage	root	0.4143	0.4599	0.69030618E-05	-2.4	-17.36	0.00000009	0.0991
	needle	0.4271	0.4599	0.70531958E-05	-2.24	-12.33	0.00000380	0.0712
Damage & inf	root	0.2236	0.4599	0.13804904E-04	-1.70	-63.60	0.00000000	0.5138
	needle	0.2523	0.4599	0.14105146E-04	-1.58	-55.25	0.00000000	0.4512
Family & inf	root	0.3476	0.4599	0.17884583E-03	-0.39	-8.39	0.00000000	0.2441
	needle	0.3678	0.4599	0.18273552E-03	-0.35	-6.80	0.00000010	0.2001

For expression analysis of 11 pine seedling families after inoculation with *H.annosum* only four RE families were selected, additionally one stress responsive gene *PsB* was included in the study. MRPP(Multi-Response Permutation Procedures) provided by PC-ORD v.5. statistical package (McCune *et al.* 2006) was used to investigate if there are significant differences between expression response in different experimental groups (Table 2). Heterogeneity within following groups was tested: infected vs controls; treatment (control, 7dpi; 14dpi; 21dpi), family of seedlings (11), damage of primary needles, damage of secondary needles, damage of the stem, length of the secondary needles, number of secondary roots etc. Expression data were relativized before analyses. Distance matrix was rank-transformed. Sorensen (Bray-Curtis) distance measure could be applied to quantitative data and it retains sensitivity in more heterogeneous data sets compared to Euclidean distance (McCune *et al.* 2002). Significant differences were observed between control and infected sample groups. Multiple pairwise comparisons of groups (data not shown) by treatment time also indicates significant difference between control group and each of sampling points after infection, while heterogeneity between groups taken after 7 dpi, 14 dpi, 21 dpi were similar to expected by chance. Pooled group of plants with visually observed damages didn't differ significantly by RE expression from group of infected plants without noticeable damages. However separation of samples by seedlings family and infection reveal differences in RE response. Six pine families had significant expression changes in roots after inoculation (A=0.20-0.265), while two seedlings families (Sm4, M236) displays no significant changes (A<0.1); M347, M259, M248 small change (A=0.1-0.16).

Figure 2. Scatterplot matrix for RE expression data in needles. Non-metric Multidimensional Scaling (NMS) procedure with correlation as distance measure was used. RE families (Cong, R4, Ang, Cumb); PsBs (*Pinosylvine synthetase* gene).

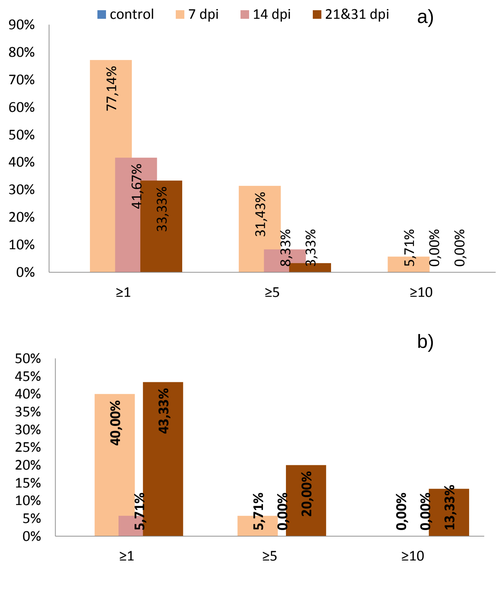


Pearson correlation was calculated and analysed for needles and roots RE expression with traits observed for each pine seedling during experiment. In all-to–all correlation analysis strong correlation (0.998) was found between four RE loci expression (*Conagree, R4, Cumberland, Angelina*) in each tissue type (Fig.2.). RE expression in needles wealky correlates with infection (0.30-0.42), but RE expression in roots weakly correlates with treatment time (0.31). PsB expression in roots weakly correlates with observed damage (0,3409), while PsB expression in needles didn't correlate to any damage traits observed. Moderate correlation (0,4845-0,5490) observed between RE expression in roots and presence of dark ring in the base of needles. Weak correlation (0.332-0,3629) observed between RE expression in roots and stem damage.

Correlation with all traits by seedling family was carried on to investigate role of family structure in RE response. Several differences were observed between 11 pine seedlings families. In five (Sm4, M236, M223, M248, M347) pine seedling families expression between RE in needles and roots correlated strongly (0.5355-0.7299), in two families (Sm12 and Bal303) moderate (0.4495-0.5794) to weak (0,3216- 0,5025), but in M259, M110, M241 and M242 families no correlation was found. In each seedling family moderate correlation between inoculation and observed damage was observed, however RE expression in the roots and damage correlates strongly only in one pine family- M259; in three families (M242, M241 and M110) moderate to weak correlation with stem damage was observed; no correlation found in other pine families. RE expression rate in needles could not to be correlated to any damage observed.

Expression of gene PsB also responds differently in different pine families- expression of PsB in roots and needles correlates strongly in M259, M223, M248, M347; in Sm4 and M236 PsB expression in roots and needles were not associated, in other families weak correlation was observed. In M259, Sm4, M236 PsB expression positively correlates with RE expression in opposite tissue at moderate to weak levels while other families displays no such correlation.

Figure 3. Percentage of seedlings displaying elevated RE expression in needles (a) and in roots (b) after *H.annosum* inoculation.



In each treatment point other seedlings were sampled for RNA analysis therefore we could not track the expression change in one plant during the experiment. Frequency of seedlings with elevated RE expression level in particular treatment point could give an insight into main tendencies of responses (Fig.3.). Data were separated by sampling points (7, 14, 21&31) and frequency of expression level more then 1, 5 & 10 was calculated.

In the needles all four RE loci displays same tendency, that plants with elevated RE responses are more often found after 7dpi with 74-75 % of seedlings displays expression higher than one and only about 30 % higher than 5, from them 3-10 % displays expression higher than 10. Proportion of plants with elevated RE expression drops gradually after 14 days post inoculation and later stage. In 21&31dpi 26-36% of plants display expression levels higher than one and about 3 % higher than 5 and no plants display expression higher than 10. In the roots RE expression changes were different in that 32-40% of plants after 7dpi display expression levels higher than one, but after 14 dpi percentage of plants with elevated RE expression drops to 3-6% and no plants were found with expression level higher than one. After 21&31dpi percentage of plants with elevated RE expression in the roots grows and about 43% displays expression higher than one, but about 13 % of plants exhibits expression higher than 10.

Contrary, stress sensitive gene expression PsB grows with pathogen spread (Fig.4.). PsB gene expression was elevated in all seedlings roots after 21&31dpi similar as in needles, but high PsB expression levels were only in 33% of plants roots after 21&31dpi while in needles 53% of plants had expression levels higher than 100 in the same treatment point.

Figure 4. Percentage of seedlings displaying elevated PsB gene expression in needles (a) and in roots (b) after inoculation with *H.annosum*.