MITE insertion-based polymorphism markers from transcriptome of non-model plant (*Pinus sylvestris* L.)

Angelika Voronova[1]*, Dainis Ruņģis[1] [1]Forest Molecular Biology and Genetics group, Latvian State Forest research institute "Silava", Salaspils, Latvia

Introduction

Transposable elements (TEs) and derived sequences could influence gene function by disruption of their structures and regulatory motifs, influencing transcription, gene splicing and heterochromatin formation, play a role in functional non-coding RNA formation and could wire genes into networks. Conifers are ancient outcrossing plants with high adaptability to the environment and short history of artificial selection. Conifer genomes contain an increased TE content and exhibit high genetic diversity. Investigation of TE-associated genes could promote the development of molecular markers linked to variable traits. However, the majority of TE insertions are located outside gene regions, therefore for species with increased TE-content, methods such as TE display in combination with short-read sequencing techniques would result in a vast range of reads with no information on co-localization with coding genes or location in other repeats. Additionally, mapping to a reference genome is necessary in order to reveal the position of the TE insertion. Therefore, these type of approaches are less efficient for non-model plant species with no reference genome available. High quality sequencing of a large conifer genome still requires considerable investments of time and financial resources. In addition, one reference genome will not provide information on TE polymorphisms among individuals, which requires even more high coverage sequencing of many individuals. Long-read sequencing techniques are necessary for the accurate positioning of TEs in genomes, but can be expensive to apply to large numbers of individuals. Targeted sequencing methods are widely utilised techniques, most of which are used in studies of coding genes. Gene-TE associated transcript selection method is suggested, which could simultaneously provide information on transcripts from several individuals, TE families, or conditions. This method was implemented for the investigation of the Scots pine (*Pinus sylvestris*) genome and has the potential to identify TE-associated transcripts in other non-model species. Application design, evaluation, and analysis of the results will be provided and discussed.

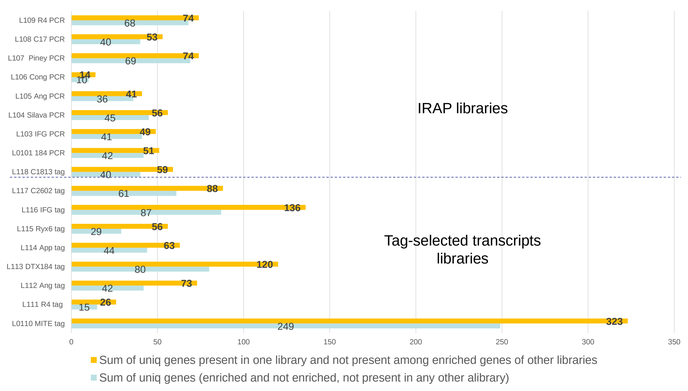

Figure 2. Comparison of uncommon genes among libraries. Same samples and all treatments were used for the selection with different TE-tags, therefore presence of hits to unique genes for particular library suggests eventuated selection.

Targeted TE-transcript selection

Gene-TE associated transcript selection method was inspired by the approach of (Coffin et al., 1974), where DNA-RNA hybrids were selected from tissues using probes for virus isolation. We developed approach to select DNA-dscDNA hybrids using specific biotinylated probes complementary to TEs, with further separation from other transcriptome with streptavidin coated magnetic beads (Fig. 1). Double stranded cDNA was used instead of RNA for a stability reasons and for increasing probability of hybridization to every transcript containing TE insertions in both probable directions. Streptavidin-biotin complex was further cleaved from hybrids using MspI restriction enzyme site introduced in the probe. Separated transcripts were propagated using multiple displacement amplification with *Phi-29* rolling-cycle polymerase (Dean et al., 2002) and sequenced with *IonTorrent Ion GeneStudio S5* (TFS) with read length up to 600bp.

For the implementation of this protocol, information about prevalent TE families in closely related species was required. While similar TE families are distributed among closely related species, the majority of TE insertions occur into non-homologous genes within particular species or varieties (Butelli et al., 2012; Voronova et al., 2020). TE sequences located nearby protein coding genes or within gene introns, are transcribed and these primary transcripts or transcripts from truncated genes were targets for the selection. RNA samples were extracted from stressed plants, in conditions with proven increased expression level of transposable elements (Voronova, 2019).

Libraries obtained with biotin probes contained in average higher diversity of genes and more enriched genes. Libraries generated from IRAP products (Kalendar et al., 2011) were less enriched with genes. To assess the extent of non-specific binding, we compared the lists of matching genes and counted only the unique genes for each library. The number of unique genes was larger than the number of common genes for each library. Some genes were present in several libraries, but were enriched only in one of the libraries (Fig.2). The common genes identified among libraries included photosystem II (chloroplast) and ribosomal proteins coding genes, transcripts which are prevalent in plant cells.

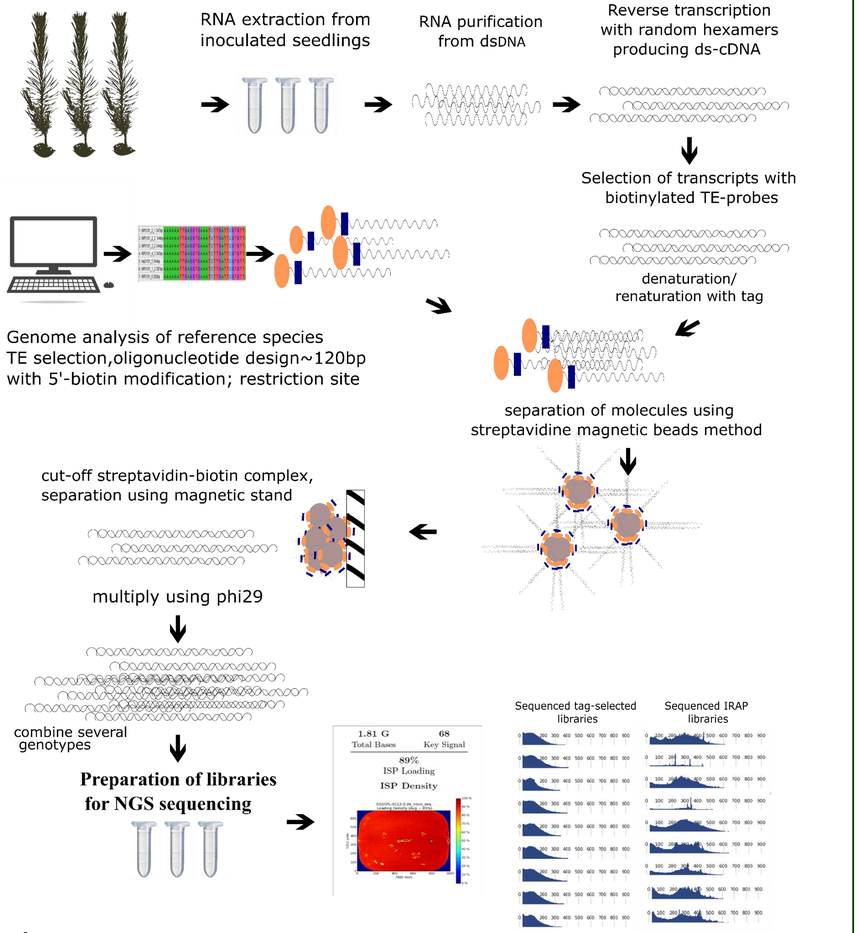

Figure 1. Overall scheme of the TE-transcript selection method used in the study. RNA samples were extracted from Scots pine trees inoculated with needle cast causative pathogen (*Lophodermium seditiosum*), TE-transcription induction after inoculation with this fungus was observed previously (Voronova et al. 2019).

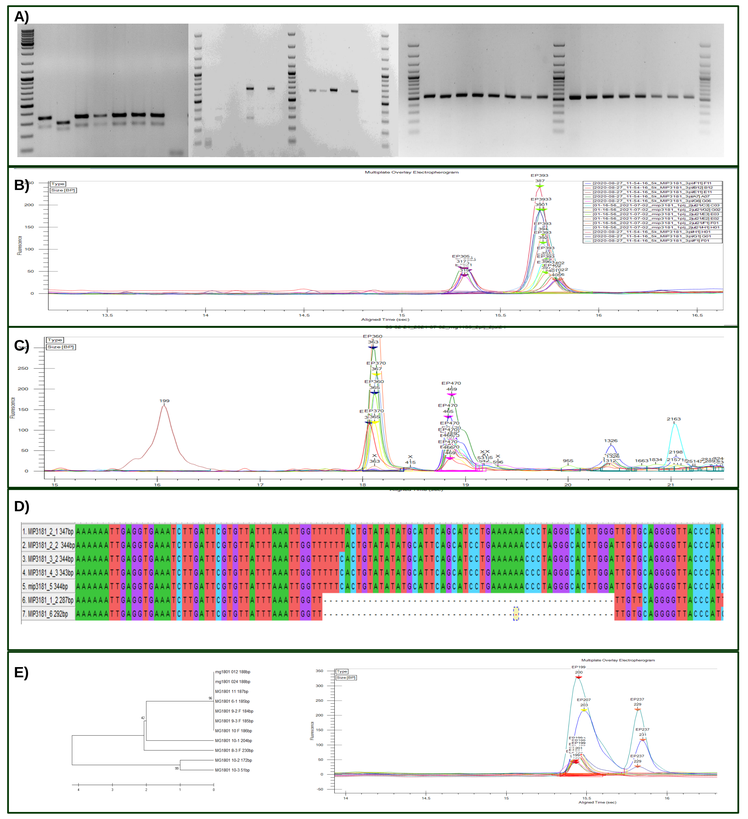

Figure 3. Examples of polymorphism produced from different markers. **A)** Markers amplification tests in 1.7% agarose showing insertion/deletion polymorphism, presence/absence polymorphism and non-polymorphic loci; **B)** MIP-4 marker amplifies specific loci with *Plater* MITE containing 52bp insertion/deletion polymorphism. **C)** MIP-10 chromotogram showing common polymorphoc bands and rare alleles, all containing *Plater* MITE element; **D)** Multiple sequence alignment of MIP-4 loci; **E)** MIP-9 amplified specific 187 bp fragment containing *Plater* MITE, however also rare 230bp alelle was revealed wich contains *Copia* retroelement *Silava* LTR repeat.

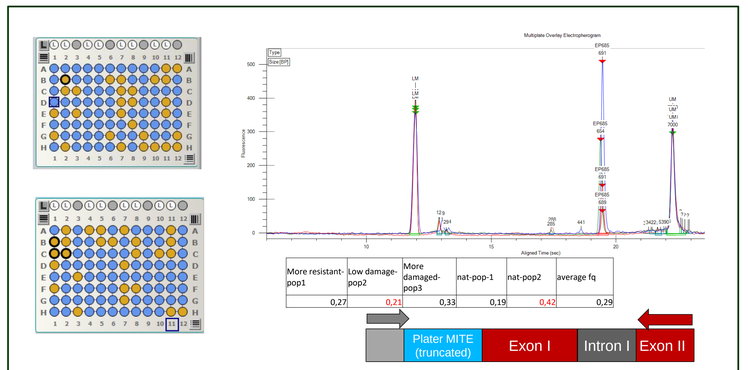

Figure 4. Example of validated gene polymorphism produced from an insertion of *Plater* MITE before *NEC3* gene. This gene contain an ancient insertion in 5' UTR, as element is truncated, therefore specific amplification was obtained.

Development of molecular markers

Plater MITE was previously shown to be distributed in gene regions of pine reference genomes (Voronova et al., 2020). About 30 reads were identified with similarity to *Plater* MITE element and surrounding sequences in one sequenced library. Primers for the MITE insertional polymorphism detection were designed for all these loci. Some of the reads were ended in the repeat, therefore additional amplification with unspecific primers with subsequent cloning and *Sanger* sequencing was applied. Gradient PCR was used for reaction optimisation, amplified fragments were ligated with plasmid, cloned and *Sanger*-sequenced.

Majority of amplification products contained expected MITE element. Some markers revealed a high degree of diversity, while amplified fragments contained the expected intact primer sequences. Empty sites were also observed for each locus. High throughput capillary gel-electrophorese on *LabChip GX-touch* instrument (*PerkinElmer*) was used for genotyping of several pine populations (from 192 to 384 samples with included biological repeats).

Pine samples were selected based on available studies on resistance of their progeny to needle cast causing pathogen (Aris Jansons, Una Neimane, 2008). Additionally, a natural pine stand was analysed to trace influence of abiotic factors. Variable alleles amplified from a larger sample set, were also cloned and sequenced. Size variations of the alleles were caused by presence of several types of TEs, by insertion/deletion polymorphism within the *Plater* element (Fig.4), by amplification of different loci, probably from a similar gene family. Several types of empty alleles were also observed for some markers. Band frequencies were calculated for each population and compared.

Conclusions

- NGS libraries were derived from inoculated Scots pine samples, identical for all targeted selections with biotinylated probes complementary to 10 TE families distributed in pine genome. Obtained libraries contained differing patterns of unique genes, especially if enriched genes were compared. The presence of unique genes in each library as well as differences in enrichment level between libraries indicates that transcript selection was specific for each probe.
- Applied transcript selection method in combination with massively parallel short-read sequencing technology provided data for the development of 20 MITE insertional polymorphism molecular markers based on the only one library of *Plater* element family.
- Transcribed alleles were determined with the suggested method. Evidence of deleterious effect could be suggested as frequencies for most MITE-containing alleles were elevated in the less resistant populations of Scots pine.
- Progeny of the analysed plus trees were studied in regard of resistance to needle cast in large trials, better characterised plant material is needed for the further characterisation of obtained markers.
- Developed targeted gene-TE associated polymorphism detection method could be used for other non-model plant species with high repeat content, for the evaluation of TE-associated mutations in gene regions. Using barcoding and reduced amount of selected data allows to combine several genotypes, as well as several mobile element families.
- Long-read technology would be preferable for the further method development and studies of longer mobile element superfamilies, as well as for more information about evaluated transcripts. Short-read technology results in reads that often ends in the repeats, but the evaluation of specific sequence requires mapping to a species reference genome or an additional approaches.

FLPP FUNDAMENTAL AND APPLIED RESEARCH PROJECTS

This research is funded by the Latvian Council of Science, Open Call of the 2020 Individual Scientist Project Proposals for Fundamental and Applied Research, project No *lzp-2020/2-0151 (2020-2021).*

PostDoc Latvia

Sequencing data developed with the support of the The State Education Development Agency 1.1.1.2. "Post-doctoral Research Aid". Nb. *1.1.1.2/VIAA/1/16/094 (2017-2019).*

References

Aris Jansons, Una Neimane, I. B. (2008). Needlecast resistance of Scots pine and possibilities of its improvement. *Mezzinatne* 18, 3–18. Available at: http://www.silava.lv/mainen/Journals/Mezzinatnemain.aspx.
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.
Coffin, J. M., Parsons, J. T., Rymo, L., Haroz, R. K., and Weissmann, C. (1974). A new approach to the isolation of RNA-DNA hybrids and its application to the quantitative determination of labeled tumor virus RNA. *J. Mol. Biol.* 86, 373–396. doi:10.1016/0022-2836(74)90026-6.
Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5261–5266. doi:10.1073/pnas.082089499.
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.
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., Rendón-Anaya, M., Ingvarsson, P., Kalendar, R., and Ruņģis, D. (2020). Comparative Study of Pine Reference Genomes Reveals Transposable Element Interconnected Gene Networks. *Genes (Basel).* 11, 1216. doi:10.3390/genes11101216.