MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 2369–2379 doi: 10.1111/mec.12189

FROM THE COVER

# Are long-lived trees poised for evolutionary change? Single locus effects in the evolution of gene expression networks in spruce

JUKKA-PEKKA VERTA,*†[1] CHRISTIAN R. LANDRY, †‡ and JOHN J. MACKAY*†

*Département des Sciences du Bois et de la Forêt & Centre d'Étude de la Forêt, Université Laval, Québec, QC, Canada, G1V 0A6, †Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, QC, Canada, G1V 0A6, ‡Département de Biologie & PROTEO, Université Laval, Québec, QC, Canada, G1V 0A6

## Abstract

**Genetic variation in gene expression traits contributes to phenotypic diversity and may facilitate adaptation following environmental change. This is especially important in long-lived organisms where adaptation to rapid changes in the environment must rely on standing variation within populations. However, the extent of expression variation in most wild species remains to be investigated. We address this question by measuring the segregation of expression levels in white spruce [*Picea glauca* (Moench), Voss] in a transcriptome-wide manner and examining the underlying evolutionary and biological processes. We applied a novel approach for the genetic analysis of expression variation by measuring its segregation in haploid meiotic seed tissue. We identified over 800 transcripts whose abundances are most likely controlled by variants in single loci. Cosegregation analysis of allelic expression levels was used to construct regulatory associations between genes and define regulatory networks. The majority (67%) of segregating transcripts were under linkage. Regulatory associations were typically among small groups of genes (2–3 transcripts), indicating that most segregating expression levels can evolve independently from one another. One notable exception was a large putative *trans* effect that altered the expression of 180 genes that includes key regulators of protein metabolism, highlighting a regulatory cascade affected by variation in a single locus in this conserved metabolic pathway. Overall, segregating expression variation was associated with stress response- and duplicated genes, whose evolution may be linked to functional innovations. These observations indicate that expression variation might be important in facilitating diversity of molecular responses to environmental stresses in wild trees.**

*Keywords*: adaptation, gymnosperms, molecular evolution, transcriptomics

*Received 16 November 2012; accepted 26 November 2012*

## Introduction

Long-lived and sessile organisms such as trees require long-term resistance and environmental adaptability to survive in a context where environmental changes can outpace their generation time. White spruce is a keystone species of the North American taiga whose populations are affected by global environmental change (Peng *et al.* 2011). Their potential for adaptation is likely to have wide-sweeping impacts on boreal animal and plant communities and the ecosystem services they provide. The ranges of spruce trees and other conifers cover large climatic gradients while subpopulations can be adapted to their local environments (Aitken *et al.* 2008; Mimura & Aitken 2010; Savolainen *et al.* 2011). These populations may draw upon alternative molecular solutions to respond to local environmental

[1]Correspondence: Jukka-Pekka Verta, 1030, Avenue de la Médecine, Pavillon Charles-Eugene-Marchand, office 2225, Québec, QC, Canada G1V 0A6, Fax: +1 418 656 7493; E-mail: jukka-pekka.verta.1@ulaval.ca, jp.verta@gmail.com

conditions (Prunier *et al.* 2011, 2012), a phenomenon also seen in annual plants (Fournier-Level *et al.* 2011). Variation in gene expression contributes to phenotypic diversity in cellular and physiological processes, including responses to environmental stresses and, therefore, may sustain the adaptability of conifer populations. Despite the fact that it could help predict their responses to a changing environment (Aitken *et al.* 2008), the extent of expression variation in wild conifers remains to be investigated.

Scans covering whole transcriptomes have identified hundreds of heritable gene expression changes in organisms ranging from budding yeast to primates (Brem *et al.* 2002; Schadt *et al.* 2003; Hubner *et al.* 2005; Kirst *et al.* 2005; Whitehead & Crawford 2006; West *et al.* 2007; Drost *et al.* 2010; Romero *et al.* 2012). A standard method to study the genetic architecture of expression variation (the number of loci involved, size of the genetic effects and degree of linkage) in model organisms has been to map gene expression variation to genetic loci (Gilad *et al.* 2008; Kliebenstein 2009). Evolutionary inferences can then be drawn based on, for instance, the identification of molecular pathways and gene features contributing to expression variation (Landry *et al.* 2006, 2007; Gan *et al.* 2011) or adaptive expression differences (Fraser *et al.* 2010). Furthermore, the genetic description of expression variation provides information about the interconnectedness of variable expression traits, which is a key in evaluating the evolutionary potential of populations. If expression variation in different genes is controlled by independent mutations, their evolution can take different paths and thus provide alternative adaptive solutions. If, however, expression variation in most genes is due to a few genetic variants with wide-ranging effects, the diversity in independent evolutionary paths is reduced. In addition, the linkage between beneficial and detrimental expression variation would likely not permit positive selection on favourable variants. The examination of the genetic architecture of expression variation requires the definition of the level of cosegregation between expression traits (Keurentjes *et al.* 2007; Ayroles *et al.* 2009; Drost *et al.* 2010). In both model and nonmodel organisms, this involves segregating populations that allow the tracking of genetic effects on expression variation. Novel approaches need to be developed in nonmodel organisms, such as forest trees, where it is unfeasible to generate the inbred lines or controlled crossed progeny that facilitate traditional analysis (Gilad *et al.* 2008).

The conserved seed morphology of the gymnosperms may provide an unexplored approach for the genetic tracking of gene expression variation. The gymnosperm seed contains a haploid tissue, the megagametophyte (Fig. 1c), which is derived from one of the four megaspores produced during meiosis. Megagametophytes inherit maternal alleles in a 1:1 ratio (Fig. 1b), allowing analysis of each allele in distinct haploid meiotic products (O'Malley *et al.* 1996). Heritable expression variation that is linked to a single major causal variant can be identified by screening for expression levels that segregate in the haploid progeny. No breeding is required because expression variation is detected between segregating maternal alleles. The method can thus be applied to any wild individual. No information on segregant genotypes is needed, given that the presence of two allelic forms can be inferred from expression differences that segregate in Mendelian ratios. Furthermore, megagametophytes were reported to have the largest number of expressed genes in a survey of white spruce tissues (Raherison *et al.* 2012), indicating that the megagametophytes allow expression variation to be analysed in the majority of genes. The megagametophyte is also relevant for studying adaptation to a changing environment because it serves as an important life-transition stage that affects embryo viability and thus contributes to fitness.

Here, we report the first study that applies the megagametophyte approach to gene expression analysis in a transcriptome-wide scale. Our objectives were to first determine the extent of segregating gene expression variation using the megagametophyte system. This would allow a determination of the feasibility of the approach and a characterization of gene expression variation most directly affected by genetic variants. Second, we aimed to study the linkage between expression traits, which would allow to determine the level of connectedness between variable expression traits and subsequent evolutionary prospects. Third, along with functional categories, we investigated the evolutionary events that could be associated with segregating expression variation. One prominent source of differentially expressed gene variants is gene duplication (Gu *et al.* 2004; Landry *et al.* 2007), and expression differences due to gene duplications can have significant contribution to adaptation both in short and in long timescales (Kondrashov 2012). Gene duplication rates (Lipinski *et al.* 2011), duplicate gene half-lifes (Lynch & Conery 2003) and evolutionary paths of duplicate genes (Carreterro-Paulet & Fares 2012) seem to vary significantly between organisms, and it is unclear whether gene duplication is a significant source of expression variation in conifers.

## Materials and methods

We focused our study on two undomesticated and field-grown mother trees representing the same East Canadian population. The wild mother trees (tree A and B) served as references in genetic mapping studies

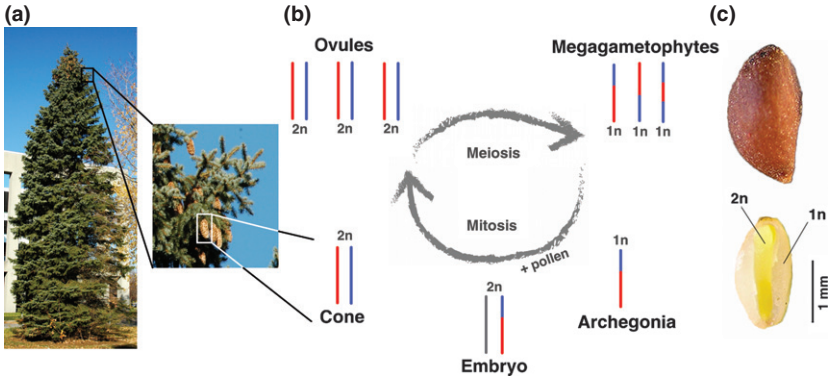


**Fig. 1** A gymnosperm system to study expression variation. (a) A spruce tree and cones. Produced each year, the cones may contain hundreds of seeds. (b) Replication, segregation and recombination of chromosomes during gymnosperm seed development (red and blue represent homologous chromosomes, grey represents paternal chromosome of the embryo). (c) Intact and dissected white spruce seed, showing the haploid and diploid tissues.

and correspond to study trees 77 111 and 80 112 in Pelgas *et al.* (2011). The trees were grown in even-aged plantations near Cap Tourmente (tree A, +47° 4′ 1.28″, −70° 49′ 4.03″) and Aubin (tree B, +46° 40′ 15.27″, −71° 29′ 34.54″), Québec, Canada. We obtained controlled crossed seed material from the Canadian Forest Service seed bank in Québec City, QC, Canada (Dr J. Beaulieu, Laurentian Forestry Centre). Seed lots were collected in a single year (tree A: 1994, tree B: 1996) from a single tree and stored at −20 °C until use for the study. Surface-sterilized seeds were stratified by submersion in sterile water at 4 °C for 24 h, followed by storage at 4 °C under 100% RH for 28 days and then dark incubation at 26 °C for 4 h to start germination. The megagametophyte was removed from the seed under a dissecting microscope, separated into two halves of equal size, flash-frozen on liquid nitrogen, and stored at −80 °C until used for RNA extraction.

### *Microarray analysis*

We randomly selected 18 megagametophytes from each parent tree for transcriptome profiling. We performed independent replication of sample preparation, microarray hybridization and fluorescent dyes for 16 and 18 of the samples from trees A and B, respectively. mRNA was independently extracted from the separate megagametophyte halves using magnetic oligo-d(T) beads, transcribed *in vitro*, labelled using two different dyes (one for each replicate sample set) and randomly hybridized on microarrays comprising oligonucleotide gene-specific probes matching 23 804 unique white spruce transcripts (Beaulieu *et al.* 2011; Raherison *et al.* 2012). The two mRNA samples from each megagametophyte were hybridized onto two independent microarrays using two different dyes. The microarray data were deposited in the GEO database (accession GSE35337) with detailed methods and protocols.

### *Segregation analysis*

Analyses were carried out separately for the two microarray dyes. Fluorescence data were background corrected and normalized using the normexp and quantile methods of the Limma package (Smyth 2005) in R (Ihaka & Gentleman 1996). Other normalization methods were investigated, but were rejected due to their reduced power in identifying segregating expression variation (Data S1). Information from spots that had abnormal morphology, that were overlaid by dust particles, or that were saturated in more than four samples were ignored in further analyses.

We defined transcripts with heritable expression variation that segregated according to Mendel's first law as Mendelian Expression Traits (METs). Mendelian Expression Traits were predicted to correspond to bimodal expression distributions in which the two modes were observed in a 1:1 frequency (falling within the 95% IC of 1:1 segregation, i.e. between five and 13 observations per mode when $N = 18$), based on differential transcript abundance in haploid progeny (Fig. 2).

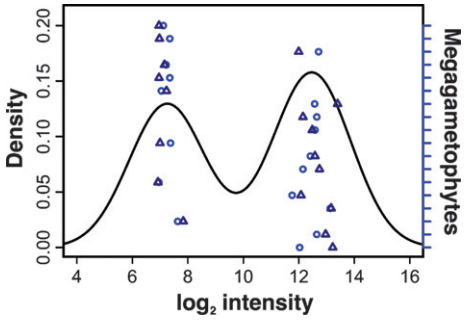


**Fig. 2** Segregation of intensity levels of a Mendelian Expression Trait (MET) transcript. Transcript abundance ($\log_2$ intensities from microarray hybridizations) of a gene exhibiting MET. Open blue circles and triangles represent independent replicates of each megagametophyte (right *y*-axis). Density estimate of expression distribution (solid line, left *y*-axis).

The number of modes in fluorescence distributions of each microarray spot was determined with a combined expectation-maximization clustering and mixture modelling approach as implemented in the Mclust function (Fraley & Raftery 2006; Hsieh *et al.* 2007) in R, allowing a maximum of two modes. Spots with bimodal fluorescence distributions were divided into those with among sibling observed frequencies of distribution modes falling between five and 13 and to those whose frequencies fell outside this range and thus could be exhibiting the contribution of more than one locus.

Transcripts meeting the 1:1 criterion were designated as displaying a MET, given that an identical segregation pattern was observed in both replicates. The replicated measurements were used to determine the stringency of our MET identification criteria. Because noise in microarray data may cause replicated samples to be incorrectly assigned to expression distribution modes, three mismatches were allowed between the mode annotations of the replicate datasets (Fig. S1). Data for each gene were independently randomized between the 18 samples for each replicate dataset while conserving information on spot quality. The mean number of METs that could be identified in 100 randomized datasets was compared to the number of METs that had been obtained by running the analysis on nonrandomized data, thereby allowing a given number of mismatches in each case. Allowing three mismatches gave an empirical 'false discovery rate' < 5%. The number of analysed (expressed) genes was defined as the number of genes whose mean intensity over all microarrays was higher or equal than the mean intensity of the lowest identified MET transcript.

### *Quantitative RT-PCR validation of METs*

We selected 23 genes exhibiting METs in tree A and whose expression differences covered a large part of the observed range (Table S1). We measured the expression of these genes in 15 previously analysed megagametophytes with quantitative RT-PCR. 250-ng of aaRNA was reverse transcribed using SuperScript II reagents (Invitrogen, Carlsbad, CA, USA), and random hexamer primers and transcripts were quantified using the QuantiTect SybrGreen reagents (Qiagen, Hilden, Germany) on a LightCycler 480 qPCR instrument (Roche, Indianapolis, IN, USA). Fluorescence data were analysed using the linear regression of efficiency method (Rutledge & Stewart 2008), and target gene expression levels were normalized to an endogenous housekeeping gene (ribosomal protein L3, GenBank: BT115036) by calculating a $\log_2$ ratio between the target and housekeeping gene. Low amplification efficiency (<80%) was detected for one of the genes, which was not considered for further analysis. The expression levels of each gene were attributed to one of two clusters according to the microarray-determined expression modes for given samples, and the clusters were tested for difference in mean relative expression level using the Welch two-sample *t*-test (corrected for multiple testing with the Benjamini–Hochberg false discovery rate) (Benjamini & Hochberg 1995) (Fig. S2). METs were considered confirmed if the segregating expression clusters determined by microarrays differed in mean expression level when measured by quantitative RT-PCR with a *P*-value threshold of 0.05 (Table S1).

For independent validation, mRNA levels of (i) 18 of the genes analysed previously and that validated the microarray results; (ii) two genes analysed previously that did not validate; and (iii) two other genes for which METs were not detected were directly measured starting from independently stratified, germinated and dissected megagametophytes of the same mother tree. mRNA was extracted as described previously, and 10-ng of mRNA was reverse transcribed using oligo-d (T) primers. Expression levels were measured and normalized as explained previously. Evidence for METs was tested with the Mclust function allowing two cluster centres.

### *Cosegregation analysis*

Coregulated transcripts were defined using a simple cosegregation analysis in the haploid segregants. The segregation patterns of each gene's expression modes were compared to the patterns of all other MET genes using a custom R script. Two mismatches were allowed between segregation patterns. This was determined following a similar approach that was used to identify METs; segregation patterns were randomized in each gene, and the number of cosegregation patterns that were observed by chance was compared to the number of cosegregation calls that were obtained using the nonrandomized data. Because of intense computational requirement, spot quality information was ignored in the analysis. A mean of 100 data permutations was used to calculate the final 'false discovery rate'.

A control for cosegregation calling was performed. Each METs segregation pattern had to be within two mismatches from all other segregation patterns included in a cosegregating group. Some cosegregating partners of a focal MET can be left of a group if they are not within two mismatches from all other group members. As a consequence, the number and size of cosegregating groups may depend on the order in which the METs segregation patterns were compared. We randomized the order in which MET segregation patterns were compared and counted the number and median of

groups. A mean of 100 permutations was reported for the number of cosegregating groups and for the median number of genes included in cosegregating groups.

### *GO analysis*

We analysed functional gene annotations of the genes exhibiting METs to identify gene classes more or less prone to expression variation. *Arabidopsis thaliana* gene annotations were assigned to each identified MET gene based on the white spruce gene catalogue (Rigault *et al.* 2011). GOslim biological process annotations for all expressed genes (see above) in each tree and the MET genes were then downloaded from the TAIR database (www.arabidopsis.org), and the number of expected versus observed annotation counts was compared using the hypergeometric test. The *P*-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate correction (Benjamini & Hochberg 1995).

### *Paralog definition and dS*

We tested whether the MET genes were associated with gene duplications. EST cluster sequences corresponding to MET genes and their predicted protein sequences were downloaded from the white spruce gene catalogue (Rigault *et al.* 2011). These sequences were searched against the whole white spruce transcript sequence catalogue with BLASTp (Altschul *et al.* 1997) using a word size of five and default parameters. Sequences were identified as paralogs if their protein sequences aligned for more than 30 amino acids with more than 80% alignment identity. The protein sequences of paralog pairs were aligned using ClustalW2 (Larkin *et al.* 2007) and default parameters. Codon sequences were then forced to fit the protein alignment using PAL2NAL (Suyama *et al.* 2006), and the number of synonymous substitutions per synonymous site (dS) was calculated using the codeml function using the Phylogenetic Analysis by Maximum Likelihood (PAML) -software (Yang 1997) and the substitution model of Yang *et al.* (1998). Only dS values less than 1.5 were considered in statistical testing. The difference in proportions of duplicated METs and duplicated white spruce genes was tested using the two-proportion *z*-test.

### *Analysis of budding yeast data*

Normalized *Saccharomyces cerevisiae* data were downloaded from the GEO website (GDS1115 & GDS1116) (Brem & Kruglyak 2005). An analysis protocol of identical steps as with the white spruce data was run on 18 segregants, each replicated on two channels and two arrays. A comparative analysis was performed on a compiled dataset obtained personally from the authors of the original study (Brem & Kruglyak 2005). This was a compiled dataset of two replicate channels, and thus the analysis steps were run on two sets of 18 different segregants. Positive segregation of two alleles was called if both sets exhibited bimodal expression patterns consistent with segregation of two alleles in one locus.

## Results

### *Megagametophyte analysis permits the measurement of segregating expression variation*

We identified 841 and 807 transcripts (30 and 22 expected by chance) whose expression levels segregated in a 1:1 ratio in the haploid progeny and thus corresponded to METs in white spruce trees A and B, respectively. A total of 111 METs were shared between the two trees. The differences between expression distribution modes ranged from 1.04- to 64.8-fold with a median of 2.1. We estimated that roughly 20 000 genes are expressed in the megagametophytes, based on lowest detected MET expression (Materials and methods). Segregation of transcript abundances in genes exhibiting METs was confirmed by quantitative RT-PCR analysis in 82% of the transcripts ($N = 22$) when assayed in the transcriptome-profiled samples of tree A and in 80% ($N = 20$) when measured in an independent set of megagametophytes from the same parent (Table S1). Two negative control genes showed no segregation of expression levels in the additional megagametophyte set. In addition to the METs, we identified 420 and 451 segregating transcripts that did not follow a 1:1 ratio. This pattern is consistent with the possible contribution of two or more loci. Finally, when applied to published budding yeast data (Brem & Kruglyak 2005), our procedure identified METs in 1.8% of assayed genes (63 transcripts, two expected by chance), consistent with the 1.86% found by eQTL analysis (Brem & Kruglyak 2005). This result suggests that our analysis procedure is able to detect Mendelian expression traits in a manner comparable with eQTL analyses.

### *Cosegregation of METs reveals distinct transcription networks*

Haploid segregant analysis allowed us to study the cosegregation among METs. Cosegregating METs represent allelic expression levels in different genes that were always found to be correlated between individual megagametophytes and are thus expected to be affected by the same segregating genetic factors. These factors can be either a common *trans* effect, such as a

transcription factor or other upstream regulator, or a *cis* effect affecting all loci, such as shared regulatory elements or multigenic copy number variation (Fig. 3).

Most of the METs (65% & 67%, trees A & B, respectively) cosegregated with at least one other Mendelian expression trait. The METs formed 147 coregulated groups in tree A (eight groups were expected by chance) and 111 groups in tree B (eight expected by chance). In tree A, the standard deviation in the number of cosegregating groups and the median number of METs included in groups was 2.9 and 0.22, respectively, which indicates that the order in which segregation patterns were compared did not greatly affect the number or size of cosegregating groups. The coregulated groups included members with opposite gene expression patterns in 63% and 54% of the cases, respectively, indicating that most genetic effects were not associated with specifically higher or lower expression levels. The median numbers of METs in a given group were 3 and 2, indicating that most single-gene *trans* or common *cis* effects act on a limited number of genes. A notable exception was observed in tree B, which contained a coregulated group that accounted for more than 22% of all discovered METs (180 transcripts, Fig. 3). Consistent with expectations for a *trans* hotspot (Brown *et al.* 2008; Wu *et al.* 2008), the group is associated with similar functional annotations, overrepresented in protein metabolism (34 observed vs. 24 expected, $P = 4.5 \times 10^{-2}$) and underrepresented in stress response (nine observed vs. 23 expected, $P = 5.6 \times 10^{-3}$).

We analysed cosegregation between the 111 shared METs to determine shared regulatory groups between white spruce individuals. In total, 22% (24/111) of the shared METs cosegregated with at least one other shared MET in both trees, forming seven coregulated groups. Gene membership in only two of these groups was totally shared in trees A and B. In addition, in two cases, shared METs could be coexpressed in either the same or the opposite directions, depending upon the individual. These results suggest that while the same genes may exhibit expression variation in two individuals, the underlying genetic variants may be different or have opposite effects. This may be the case if the genes exhibiting a switch in effect directions are affected by different *trans* variation in the two individuals, for example.

### *METs are associated with distinct biological processes and gene duplications*

We analysed the representation of Gene Ontology (Ashburner *et al.* 2000) categories in the METs to define gene groups whose expression is more prone to vary due to direct Mendelian effects. GO annotations were assigned to 55% and 58% of the METs in trees A and B, respectively, based on their sequence similarity to

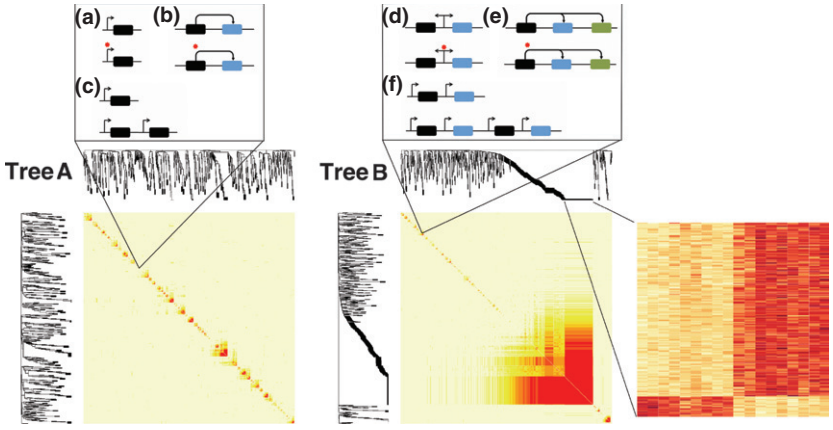


**Fig. 3** White spruce individuals show distinct putative *cis* and *trans* effect sizes. A topological overlap matrix of MET segregation patterns in trees A and B shows the differential landscapes of Mendelian expression variation in the two studied trees (tree A & tree B). For illustration purposes, MET segregation patterns are compared to each other using hierarchical clustering. The pattern of each MET is represented as a clustering tree branch. Cosegregating METs correspond to adjacent branches on the clustering tree. The levels of correlation between MET segregation patterns are colour coded, so that cosegregating METs correspond to red clusters in the matrix. Putative heterozygous genotypes of mother trees are illustrated in boxes at the top of the figure (a-f; genes represented as boxes, promoter *cis* elements as open arrows and regulatory associations as connected arrows). METs that do not cosegregate can be caused by, but not restricted to, *cis* variation (a), single-gene *trans* variation (b) or copy number variation (c). Cosegregation can be caused by, but not restricted to, variation in shared *cis* elements (d), multigenic *trans* variation (e) or multigenic copy number variation (f). Bottom right: hierarchical clustering of expression levels of 180 cosegregating METs in tree B (rows, genes; columns, megagametophytes).

*Arabidopsis thaliana* (Rigault *et al.* 2011). Overrepresentation was observed in stress response [tree A: 123 observed vs. 71 expected, $P = 8 \times 10^{-9}$; tree B (excluding large cosegregating group): 88 observed vs. 61 expected, $P = 1.5 \times 10^{-3}$], while genes associated with developmental processes were underrepresented (tree A: 48 observed vs. 70 expected, $P = 6.3 \times 10^{-3}$).

We investigated whether genes exhibiting METs could be associated with gene duplications. The proportions of duplicated genes in the MET datasets were significantly higher than in the white spruce gene catalogue in general ($p_{genome} = 0.108$ vs. tree A: $p_{MET} = 0.187$, $P = 3 \times 10^{-12}$; vs. tree B: $p_{MET} = 0.142$, $P = 3 \times 10^{-3}$, two-sample *z*-test of proportions), indicating an association between gene duplication and Mendelian expression variation in white spruce. We also observed an overrepresentation in gene functions associated with stress response in the duplicated MET genes in tree A (40 observed vs. 21 expected, $P = 2.8 \times 10^{-4}$), pointing towards an interplay of Mendelian expression variation, gene duplication and stress response genes.

Next, we investigated the timescale of gene duplication events that contribute to Mendelian expression variation. We calculated the relative time of duplication of MET genes based on the silent-site divergence from their most probable protein pairs. The ratio of synonymous substitutions to synonymous sites was significantly smaller in duplicated MET genes of tree A compared to expressed duplicate genes in the white spruce genome on average [$\text{mean}(dS_{<1.5})_{MET} = 0.21$ vs. $\text{mean}(dS_{<1.5})_{genome} = 0.28$, $P = 8 \times 10^{-4}$, Wilcoxon–Mann–Whitney *U*-test, Fig. 4a]. This indicates that METs are preferentially found in younger duplicated genes. The duplicated gene pairs in which METs were observed for at least one paralog were divided into two groups; those in which we observed METs for both paralogous copies and those in which just one gene of a duplicate pair exhibited METs. The majority (88%) of duplicate pairs exhibited METs in only one gene paralog. This indicates that the evolutionary paths of most gene duplicates diverge; while expression levels are nonvariable in one copy, they are polymorphic in the other. Asymmetric expression evolution where expression variation is constrained to only one copy of a duplicate pair is often interpreted as evidence for unequal selection on the two copies (Innan & Kondrashov 2010). Those paralog pairs in which both genes did exhibit METs were very young, suggesting that symmetric expression evolution, which is consistent with relaxed selection on both copies (Innan & Kondrashov 2010), is restricted to young duplicates (Fig. 4b).

## Discussion

We developed a method for tracking segregating gene expression variation in haploid meiotic products of wild gymnosperm individuals. We identified expression variation associated with variants in single loci (METs) using microarray data from megagametophytes of two wild white spruce trees. METs were for most part specific to one or the other genotype, consistent with an abundance of segregating alleles in the population. As we concentrated on two unrelated individuals, more in-depth studies on the diversity of METs in white spruce would be needed to determine the level to which METs are shared within and among populations. We also applied our procedure for the identification of segregating expression variation to gene expression data from three model species (*Arabidopsis thaliana*, *Saccharomyces cerevisiae* & *Rattus norvegicus*) to gain an interspecific perspective of the frequencies of segregating expression traits (Data S1). METs were twice as frequent in white spruce as in the other species, which could be accounted by differences in expression network connectedness, for example (Figure S4). Our megagametophyte analysis allowed a robust identification of METs, supported by technical replication and quantitative RT-PCR validation. However, future studies using megagametophytes might benefit from the analysis of a larger number of segregant genotypes, as it may allow the tracking of more complex inheritance patterns of expression traits.

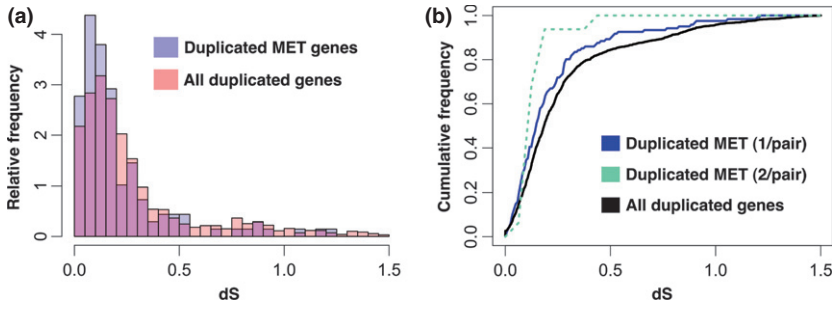


**Fig. 4** Duplicated genes that exhibit METs tend to be young paralogs. (a) A histogram of the relative frequency of duplicated genes for each bin of ratio between synonymous substitutions and synonymous sites (dS) in tree A. Bin size 0.05. (b) Cumulative frequencies of different duplicated sequences over relative age (dS) in tree A.

To understand the effect of single Mendelian alleles on whole gene networks, we investigated whether METs were associated with many independently segregating mutations or with few mutations having widespread effects. Most METs (67% and 65%) cosegregated with at least one other MET in both studied trees. Combining the number of cosegregating groups and the number of independently segregating METs, over 400 independent mutations causing segregating gene expression levels could be inferred in each of the trees. The numbers of METs that were included in coexpressed regulatory groups were roughly similar to observed *trans* proportions in eQTL studies on *Zea mays* (Swanson-Wagner *et al.* 2009), *A. thaliana* (West *et al.* 2007) and budding yeast (Brem *et al.* 2002), yet most of the groups did not exhibit directional biases on gene expression levels predicted for *trans* effects (West *et al.* 2007). Most coregulated groups were relatively small (2–3 members), which indicated a trend towards generally short *trans* cascades and, to a large extent, independent segregation of Mendelian expression variation. This aspect of MET cosegregation is important for their possible evolutionary implications. Given that most regulatory associations were between small numbers of METs, natural selection can, to some extent, act on their phenotypes in an independent manner. It is therefore possible that the diversity in METs might contribute to molecular evolution by providing alternative phenotypes. Again, a more fine-grained dissection of these cosegregation patterns will benefit from the analysis of a larger number of progeny. Furthermore, assessing whether METs may have evolutionary impacts will require elucidation of their effects on fitness in haploid and diploid tissues as well as their possible association with adaptative traits. We also note that the predominant effect sizes of regulatory variants within the white spruce population remain to be determined, considering that we observed widely varying effects of single regulatory variants, affecting chiefly fewer than four genes in both trees but up to 180 genes in tree B.

Many genes that are associated with transcription and translation were included in the group of 180 genes segregating in tree B, such as putative homologs of *A. thaliana* histone acetyltransferase *GCN5* and three eukaryotic translation initiation factors (*eIF5A*, *eIF2B* & *eIF4A*). Their null mutants exhibit striking phenotypic effects on growth and development in *A. thaliana* (Vlachonasios *et al.* 2003; Feng *et al.* 2007). The putative eIF factors were particularly interesting as *trans* controlling genes, because the highly conserved eIF proteins are essential for proper mRNA 5′ capping and translation in eukaryotes (Hernández *et al.* 2010). In budding yeast, at least *eIF2B* and *eIF5A* have been directly associated with a transcriptional/translational cascade that includes *GCN5* through *GCN4* (Georgakopoulos & Thireos 1992; Hinnebusch 1997; Kuo *et al.* 2000; Yamamoto *et al.* 2005). GCN4, which is a master regulator of amino acid biosynthesis, recruits GCN5 to specific gene promoters, leading to histone hyperacetylation and transcriptional activation of amino acid biosynthesis genes (Kuo *et al.* 2000). We observed that high allelic expression levels of *GCN5* and *eIF* genes cosegregated with high expression levels of many genes that are involved in protein metabolism. They also cosegregated with low levels of expression of genes involved in protein degradation and associated with the vacuole, which is indicative of a generally lower level of protein turnover. Our results suggested that a conserved regulatory link exists between *eIF*'s and amino acid metabolism genes in gymnosperms. No plant *GCN4* has yet been characterized (Hey *et al.* 2010).

A large *trans* effect on protein metabolism has also been observed in budding yeast and was caused by a single nucleotide frameshift (Brown *et al.* 2008), indicating that such strong *trans* effects can have very simple molecular bases. Our findings illustrate the power of our approach to identifying putative Mendelian *trans* regulators, an ability that has a wide range of applications in studying gymnosperm populations, phenotypes and gene networks. Notably, because our approach identifies gene expression variation due to simply inherited Mendelian variants, the results of MET analysis can be readily combined with genetic mapping or marker analysis, along with physiological responses. The approach offers simple and direct ways to study the frequencies of Mendelian variants in wild individuals and their possible association with locally adapted populations.

We identified functional and evolutionary gene categories that were associated with Mendelian expression traits. We found that METs were preferentially associated with genes that were involved in stress responses and relatively rare in genes involved in developmental processes. Nearly half of the identified MET transcripts had no sequence similarity to *A. thaliana*, indicating that diversity is also present in genes without close *A. thaliana* homologs. Our results indicate that METs might contribute to the diversity of stress responses in white spruce. Moreover, genes exhibiting METs were associated with gene duplications, with recent duplication events being more likely to contribute to Mendelian expression variation than old ones. We investigated the distribution of Mendelian variation between duplicated gene pairs and observed that MET genes generally followed an asymmetric evolution in which one gene of a duplicate pair exhibits expression variation. Our results supported the view that a change in evolutionary paths as a function of age is common in duplicated genes (Gu *et al.* 2005; He & Zhang 2005; Zou *et al.* 2009) and

suggested that the accumulation of Mendelian expression differences might be coupled with relaxation of selection pressure through single-gene duplication events. The retention of paralogs with asymmetric expression has been associated with functional innovation in the copy whose expression is under lower constraint (Zou *et al.* 2009). Our observations included METs in asymmetrically expressed paralogs that had high dS values and that may have been retained through this mechanism.

Taken together, our results agreed with observations of a link between expression variation and gene duplication in selfing annual plants (Zou *et al.* 2009) and model organisms (Gu *et al.* 2004; Landry *et al.* 2006) and have highlighted the role of gene duplication as a prominent source of evolutionary novelty that facilitates gene diversity especially in stress responses. This mechanism might be an important source of diversity on a local scale, which could play a role in local adaptation through alternative genes and alleles, as suggested by recent studies in annual plants (Fournier-Level *et al.* 2011) as well as spruce trees (Prunier *et al.* 2011, 2012).

In summary, we showed that gymnosperm megagametophytes facilitate the tracking of segregating expression traits within single individuals and that their haploid nature can provide a powerful system for estimating the total number of loci that are associated with such traits. Megagametophytes can facilitate direct analyses of any number of unrelated individuals or meiotic products, allowing the study of gene network evolution also in long-lived, out bred species with large genomes. The relatively high frequency of segregating expression traits, their diversity between individuals and their tendency to be associated with stress response genes could provide the means for phenotypic diversity and facilitate adaptation in white spruce.

The impracticality of standard approaches is a common problem that hampers genetic analyses in non-model species, yet expanding our scope to wild species is necessary to advance the field of ecological and evolutionary genomics (Pavey *et al.* 2012). The megagametophyte approach is a step towards enabling such analyses in a large number of wild plant species.

## Acknowledgements

Funding to JM was provided by Genome Canada and Genome Quebec for the Arborea II and projects as well as NSERC (Natural Sciences and Engineering Research Council of Canada). CRL is a CIHR New Investigator and his research in ecological genomics is funded by a NSERC Discovery Grant. The authors thank R. Brem for providing a *S. cerevisiae* dataset, members of J. MacKay laboratory for technical assistance and R. Sederoff, N. Aubin-Horth and W. Parsons for helpful comments on the manuscript.

## References

Aitken SN, Yeaman S, Holliday JA, Wang T, Curtis-McLane S (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evolutionary Applications*, **1**, 95–111.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Ashburner M, Ball C, Blake J *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Ayroles JF, Carbone MA, Stone EA *et al.* (2009) Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genetics*, **41**, 299–307.

Beaulieu J, Doerksen T, Boyle B *et al.* (2011) Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. *Genetics*, **188**, 197–214.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a pratical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B - Methodological*, **57**, 289–300.

Brem R, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 1572–1577.

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Brown KM, Landry CR, Hartl DL, Cavalieri D (2008) Cascading transcriptional effects of a naturally occurring frameshift mutation in *Saccharomyces cerevisiae*. *Molecular Ecology*, **17**, 2985–2997.

Carreterro-Paulet L, Fares MA (2012) Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Molecular Biology and Evolution*, **29**, 3541–3551.

Drost DR, Benedict CI, Berg A, Zhang J, Yang X, Zuo J (2010) Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of Populus. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 8492–8497.

Feng H, Chen Q, Feng J, Zhang J, Yang X, Zuo J (2007) Functional characterization of the Arabidopsis eukaryotic translation initiation factor 5A–2 that plays a crucial role in plant growth and development by regulating cell division, cell growth, and cell death. *Plant Physiology*, **144**, 1531–1545.

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.

Fraley C, Raftery A (2006) *MCLUST version 3 for R: Normal Mixture Modeling and Model-Based Clustering*. Technical Report 504, Department of Statistics, University of Washington, Seattle, WA, USA.

Fraser HB, Moses AM, Schadt EE (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 2977–2982.

Gan X, Stegle O, Behr J *et al.* (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, **477**, 419–423.

Georgakopoulos T, Thireos G (1992) Two distinct yeast transcriptional activators require the function of the GCN5 protein to promote normal levels of transcription. *EMBO Journal*, **11**, 4145–4152.

Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, **24**, 408–415.

Gu Z, Rifkin SA, White KP, Li W (2004) Duplicate genes increase gene expression diversity within and between species. *Nature Genetics*, **36**, 577–579.

Gu X, Zhang Z, Huang W (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 707–712.

He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, **169**, 1157–1164.

Hernández G, Altmann M, Lasko P (2010) Origins and evolution of the mechanisms regulating translation initiation in eukaryotes. *Trends in Biochemical Sciences*, **35**, 63–73.

Hey SJ, Byrne E, Halford NG (2010) The interface between metabolic and stress signalling. *Annals of Botany*, **105**, 197–203.

Hinnebusch AG (1997) Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome. *Journal of Biological Chemistry*, **272**, 21661–21664.

Hsieh W, Passador-Gurgel G, Stone E, Gibson G (2007) Mixture modeling of transcript abundance classes in natural populations. *Genome Biology*, **8**, R98.

Hubner N, Wallace CA, Zimdahl H *et al.* (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, **37**, 243–253.

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Innan H, Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, **11**, 97–108.

Keurentjes JJB, Fu J, Terpstra IR *et al.* (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 1708–1713.

Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR (2005) Genetic architecture of transcript-level variation in differentiating xylem of a Eucalyptus hybrid. *Genetics*, **169**, 2295–2303.

Kliebenstein D (2009) Quantitative genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTLs. *Annual Review of Plant Biology*, **60**, 93–114.

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 5048–5057.

Kuo MH, vom Baur E, Struhl K, Allis CD (2000) Gcn4 activator targets Gcn5 histone acetyltransferase to specific promoters independently of transcription. *Molecular Cell*, **6**, 1309–1320.

Landry CR, Oh J, Hartl DL, Cavalieri D (2006) Genome-wide scan reveals that genetic variation for transcriptional plasticity in yeast is biased towards multi-copy and dispensable genes. *Gene*, **366**, 343–351.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL (2007) Genetic properties influencing the evolvability of gene expression. *Science*, **317**, 118–121.

Larkin MA, Blackshields G, Brown NP *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U (2011) High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology*, **21**, 306–310.

Lynch M, Conery JS (2003) The origins of genome complexity. *Science*, **302**, 1401–1404.

Mimura M, Aitken SN (2010) Local adaptation at the range peripheries of Sitka spruce. *Journal of Evolutionary Biology*, **23**, 249–258.

O'Malley DM, Grattapaglia D, Chaparro JX *et al.* (1996) Molecular markers, forest genetics and tree breeding In:*Genomes of Plants and Animals*, 21st Stadler Genetics Symposium, (eds Gustafson JP, Flavell RB), pp. 87–102, Plenum Press, New York.

Pavey SA, Bernatchez L, Aubin-Horth N, Landry C (2012) What is needed for next-generation ecological and evolutionary genomics? *Trends in Ecology and Evolution*, **27**, 673–678.

Pelgas B, Bousquet J, Meirmans PG, Ritland K, Isabel N (2011) QTL mapping in white spruce: gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. *BMC Genomics*, **12**, 145.

Peng C, Ma Z, Lei X, Zhu Q *et al.* (2011) A drought-induced pervasive increase in tree mortality across Canada's boreal forests. *Nature Climate Change*, **1**, 467–471.

Prunier J, Laroche J, Beaulieu J, Bousquet J (2011) Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Molecular Ecology*, **20**, 1702–1716.

Prunier J, Gérardi S, Laroche J, Beaulieu J, Bousquet J (2012) Parallel and lineage-specific molecular adaptation to climate in boreal black spruce. *Molecular Ecology*, **21**, 4270–4286.

Raherison E, Rigault P, Caron S *et al.* (2012) Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression. *BMC Genomics*, **13**, 434.

Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, MacKay JJ (2011) A white spruce gene catalog for conifer genome analyses. *Plant Physiology*, **157**, 14–28.

Romero IG, Ruvinsky I, Gilad Y (2012) Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, **13**, 505–516.

Rutledge RG, Stewart D (2008) A kinetic-based sigmoidal model for the polymerase chain reaction and its application to high-capacity absolute quantitative real-time PCR. *BMC Biotechnology*, **8**, 47.

Savolainen O, Kujala ST, Sokol C *et al.* (2011) Adaptive potential of northernmost tree populations to climate change, with emphasis on Scots pine (*Pinus sylvestris* L.). *Journal of Heredity*, **102**, 526–536.

Schadt EE, Monks SA, Drake TA *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.

Smyth GK (2005) Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*(eds Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W), pp. 397–420. Springer, New York.

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**, W609–612.

Swanson-Wagner RA, DeCook R, Jia Y *et al.* (2009) Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. *Science*, **326**, 1118–1120.

Vlachonasios KE, Thomashow MF, Triezenberg SJ (2003) Disruption mutations of ADA2b and GCN5 transcriptional adaptor genes dramatically affect Arabidopsis growth, development, and gene expression. *Plant Cell*, **15**, 626–638.

West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, St Clair DA (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics*, **175**, 1441–1450.

Whitehead A, Crawford DL (2006) Neutral and adaptive variation in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 5425–5430.

Wu C, Delano DL, Mitro N *et al.* (2008) Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genetics*, **4**, e1000070.

Yamamoto Y, Singh CR, Marintchev A, Hall NS, Hannig EM, Wagner G, Asano K (2005) The eukaryotic initiation factor (eIF) 5 HEAT domain mediates multifactor assembly and scanning with distinct interfaces to eIF1, eIF2, eIF3, and eIF4G. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 16164–16169.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computational and Applied Biosciences*, **13**, 555–556.

Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, **15**, 1600–1611.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu S (2009) Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genetics*, **5**, e1000581.

---

J-P.V. designed and performed the experiments, analyzed the data and drafted the manuscript. This study is part of his PhD research towards understanding heritable expression variation in conifer trees. J-P.V. is broadly interested in evolutionary and ecological genomics in model and non-model species. C.R.L. and J.J.M. provided advice and guidance on the design of experiments, data analysis and interpretation of results, and contributed to the writing of the manuscript. C.R.L. is interested in evolutionary systems biology and ecological genomics. J.J.M. is interested in forest genomics and investigates transcriptional regulation as a component of adaptive diversity and metabolic plasticity.

---

## Data accessibility

Gene expression data: Gene Expression Omnibus accession GSE35337.

R scripts: included in online supporting information.

## Supporting information

Additional supporting information may be found in the online version of this article.

**Table S1** Quantitative RT-PCR validation of METs in two independent megagametophyte sets of tree A.

**Fig. S1** Expected and observed numbers of Mendelian expression traits when allowing 1 to 3 mismatches (MM) between Mclust -mode annotations of replicate datapoints in tree A.

**Fig. S2** Quantitative RT-PCR validation of segregation calls.

**Fig. S3** Relative tissue-preferential expression of a co-segregating MET group.

**Fig. S4** METs are twice as frequent in white spruce megagametophytes as in model species.

**Data S1** Supplementary analysis, materials and methods: comparison of MET frequencies across species, between-tree differences in gene expression, alternative normalization and cosegregation analysis.