The landscape of SNP diversity and relationships with gene function and expression in white spruce




Nathalie Pavy , Astrid Deschênes, John Mackay, and Jean Bousquet

ProCoGen 2nd Training Workshop
University of Alcalá, Spain, February 19th – 21st 2014

UNIVERSITÉ LAVAL

SMarTForests

Canada Research Chair in Forest and Environmental Genomics

SNP discovery in white spruce

Raw material for constructing SNP genotyping arrays for:

- Gene mapping
- Environmental association studies
- Association tests
- Genomic selection

White spruce sequence resources

- 212 individuals from Québec
  - Natural populations
  - Germplasm collections (natural genetic resources collections)

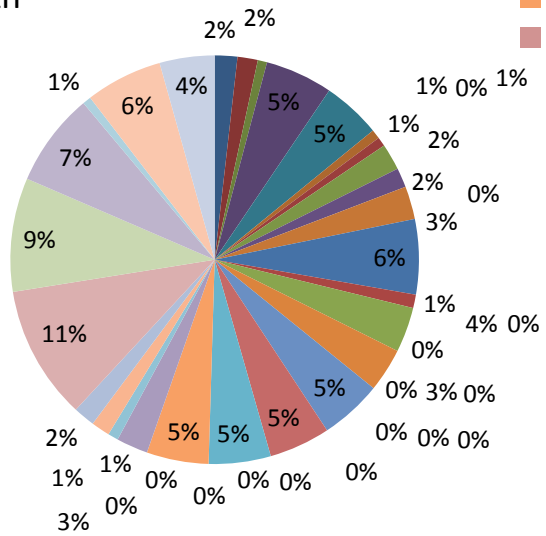
- 48 cDNA libraries
  - To maximize the number of genes discovered
  - 43 cDNA libraries -> Sanger sequencing
  - 5 cDNA libraries ->Next- generation sequencing (Roche 454, Illumina GA II)

- Sequencing effort
  - 64 million quality reads
  - Released by Arborea1 (Pavy et al. 2005 BMC Genomics) and Arborea2 (Rigault et al. 2011 Plant Physiol.)

Sequence resources used

Part 1: 144k quality reads obtained by Sanger sequencing (Pavy et al. 2005)

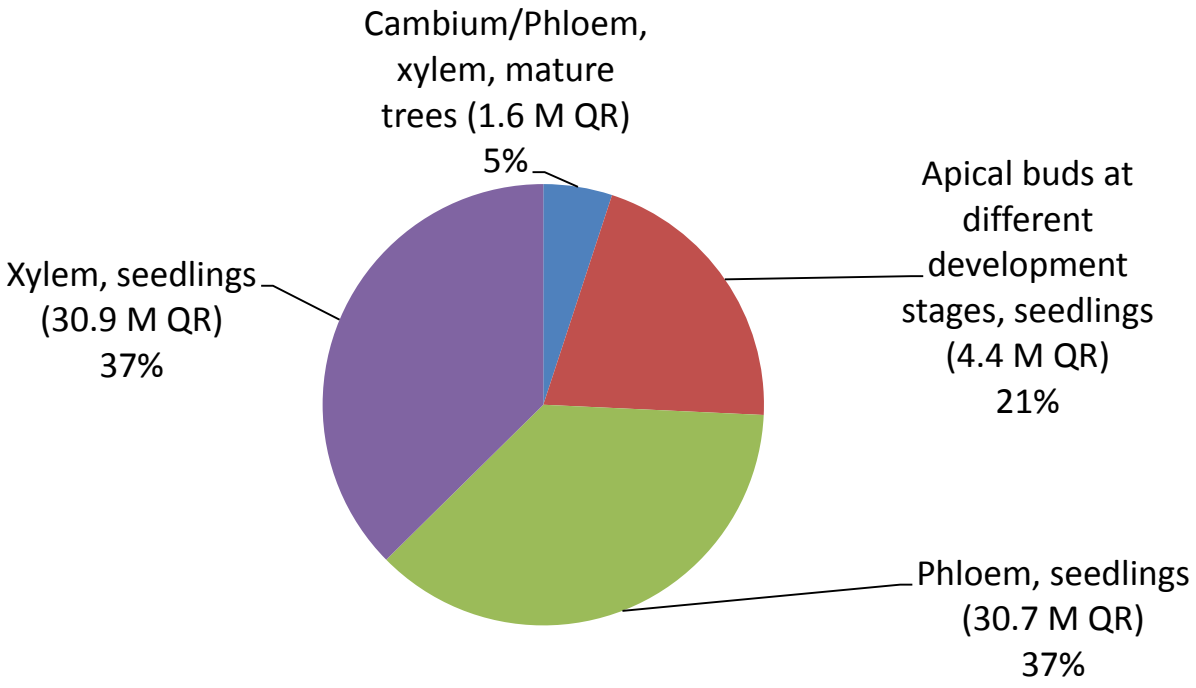
- Male strobili developmental sequence
- Female cones developmental sequence
- Vegetative buds developmental sequence
- Non-lignified secondary xylem from mature trees
- Cambium and phloem from mature tree
- Secondary xylem of girdled saplings
- Cambium, phloem and bark of girdled saplings
- Elongating ROOTS tips - saplings
- Primary, sec. SHOOT - N fertil. Treatments
- Immature somatic EMBRYOS
- Clean ROOTS systems - N treatments
- Clean ROOTS systems - P treatments
- Clean ROOTS systems - Diurnal cycle
- ROOT XYLEM - mature trees
- Annual flush SHOOTS diurnal cycle - trees
- NEEDLES - N fertilization treatments
- Mature Somatic EMBRYOS
- Elongating ROOTS tips - saplings
- Cambium / phloem scrappings (Normalized library)
- Xylem scrapings - AT NITE
- Xylem planings - daytime and seasonal
- Xylem Scrapings (Normalized library)
- Shoot tip (Normalized library)
- Terminal leader (Normalized library)
- Needles - Mid-season
- Needles - End of season
- Shoot tip - Active growth
- Shoot tip - Dormant
- Stem - Active growth
- Stem - Dormant
- Shoot tip - Active
- growth (Normalized)



Sequence resources used

Part 2: 64.3 M of quality reads obtained by next generation sequencing (Rigault et al. 2011)

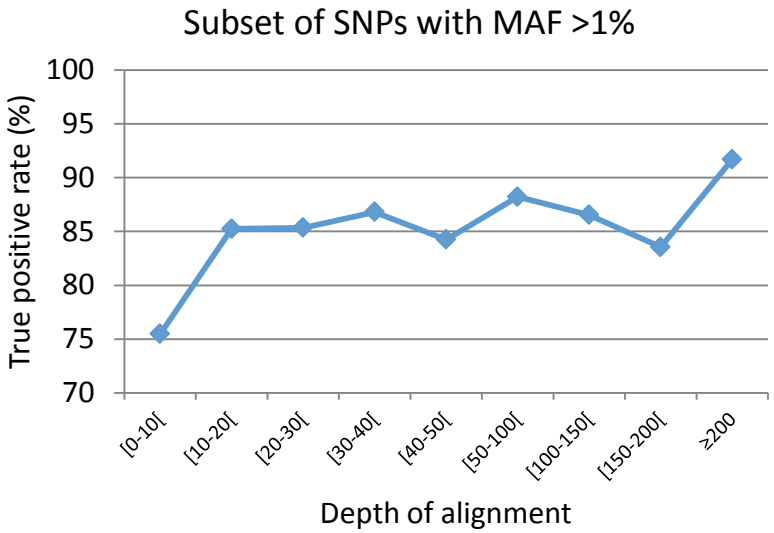
- most of the data used for the SNP discovery is derived from the GAIIx sequencing
  64.3 M of QR in total = 4.9M by GS-FLX (8%) + 59.4M by Illumina GA IIx (92%)
- 5.7 Gb of sequences incorporated in the alignments and subsequent SNP search:

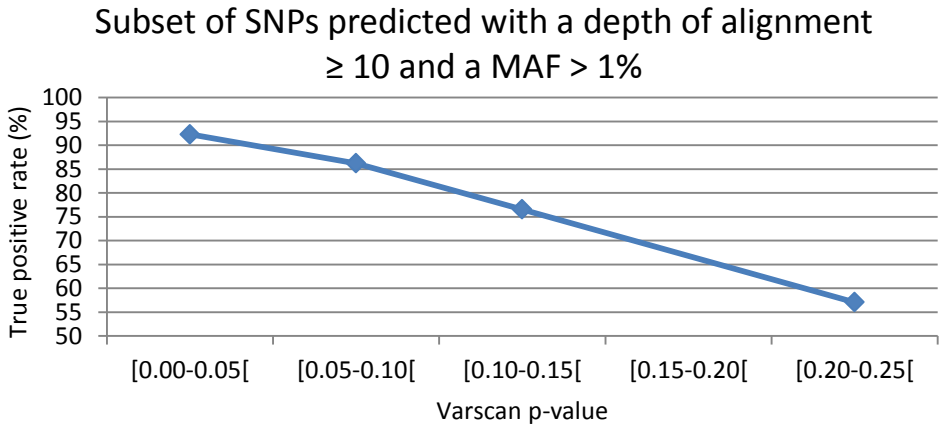


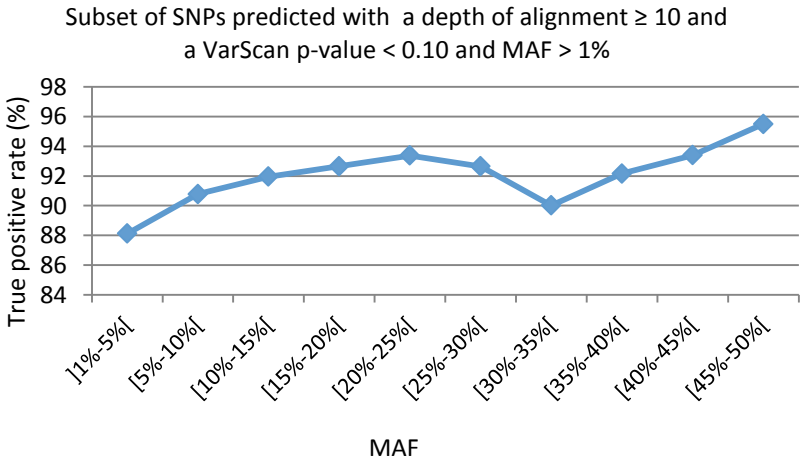
- 80.4% of the length of the reads mapped on the reference database were from individuals from germplasm collections sequenced with GAIIx

Criteria for identifying high-confidence SNPs

Analysis ok 5k SNPs predicted by Varscan and validation with genotyping data (Infinium array)







215k high confidence SNPs minimum 92% validation rate in genotyping (Pavy et al. 2013 M.E.R.)

Building a SNP atlas of >200k high-confidence SNPs

Natural and
association populations
64 million quality reads

White spruce gene catalogue
27,720 gene representatives
(not EST clusters !!!)
(Rigault et al. 2011 Plant Physiol)

33.5 million reads aligned over
27,645 genes

- Stringent alignment parameters
- Reliable reference
  to avoid inducing paralogous variation in the gene assemblies

374k non-singleton SNPs
(80.5% of success rate)

Sample of 5,938 predicted SNPs tested by Infinium genotyping for validation (Pavy et al. 2013 MER)

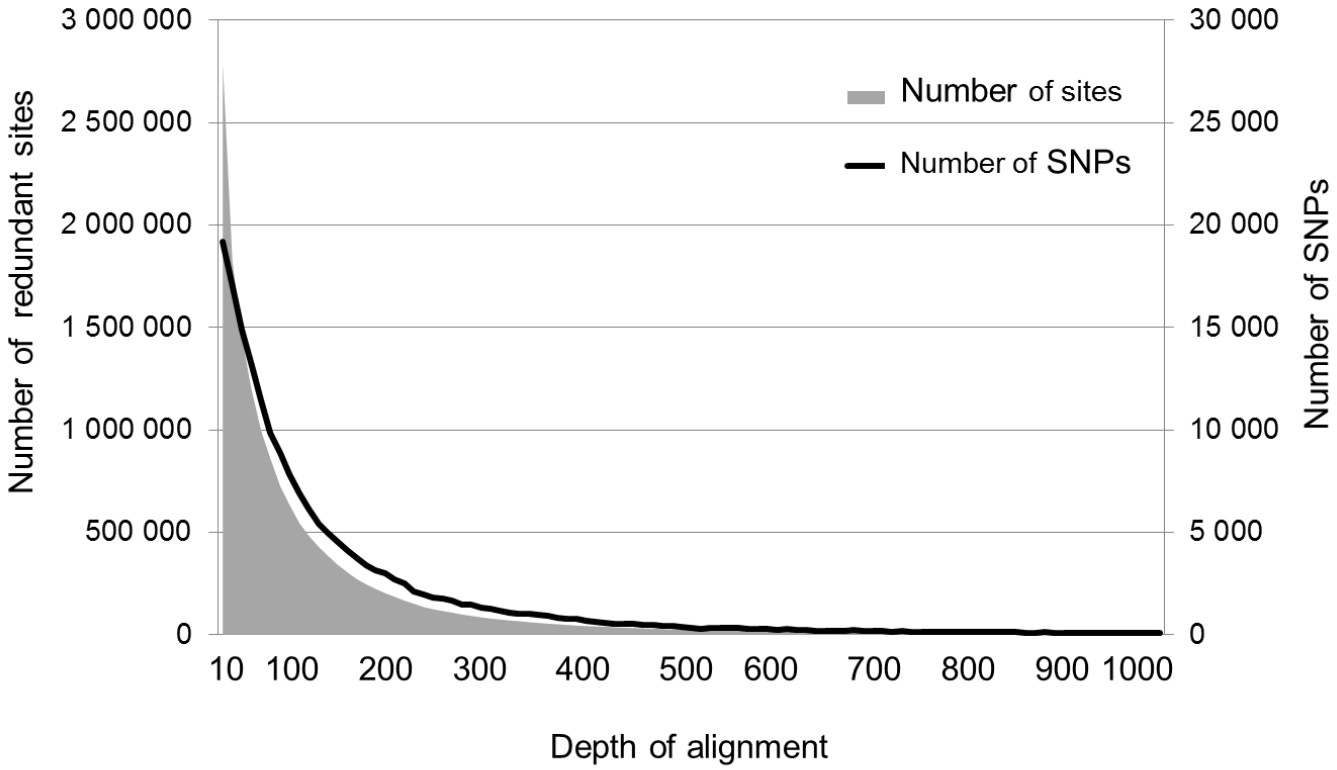
215k high-confidence SNPs
High confidence: 92.1% success rate in Infinium genotyping
Over 17,568 genes represented

Available in dbSNP

195k high-confidence SNPs
over 15,382 snped Full Length Insert Clone sequences

- 76k SNPs over 4,813 confirmed complete coding sequences
- 44k SNPs over 2,973 predicted complete coding sequences
- 76k SNPs over 7,596 other coding sequences
  (partial cds or with not match with a reference protein)

SNP distribution by alignment depth



Because of variable alignment depth, it is impossible to directly estimate unbiased nucleotide diversity

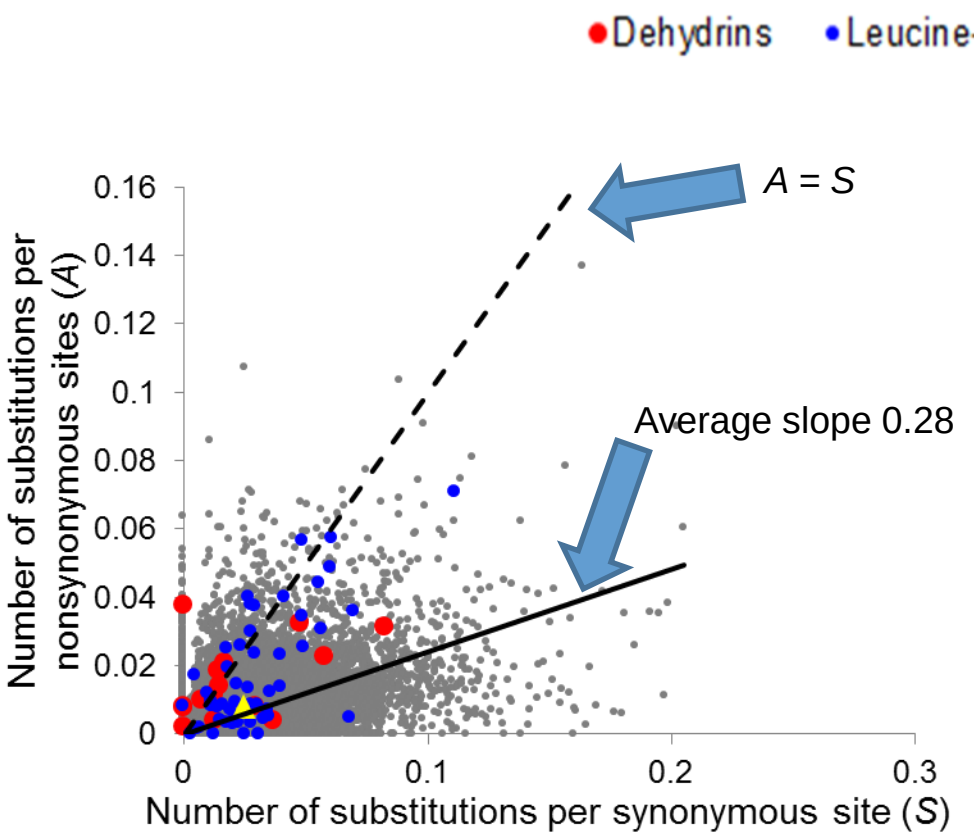
A/S ratio as an indicator of positive selection

- *A* = number of substitutions per nonsynonymous site
- *S* = number of substitutions per synonymous site
- Analogous to interspecific *dN/dS* ratio (Liu et al. 2008 Genome Biol.)
- At the intraspecific level, *A/S* may not be linearly related to *s*, the selection coefficient, so that *A/S* values above 1 may not be strictly indicative of non-neutral evolution (Kryazhimskiy & Plotkin 2008 PloS Genet.)
- *A/S* may be interpreted in relative term: genes with highest *A/S* ratios more likely to be under positive selection than those with lowest *A/S* ratios, more likely to be under purifying selection

SNP abundance within coding sequences

15k cDNAs (gene representatives)

P. glauca



Comparison with angiosperm data

Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*

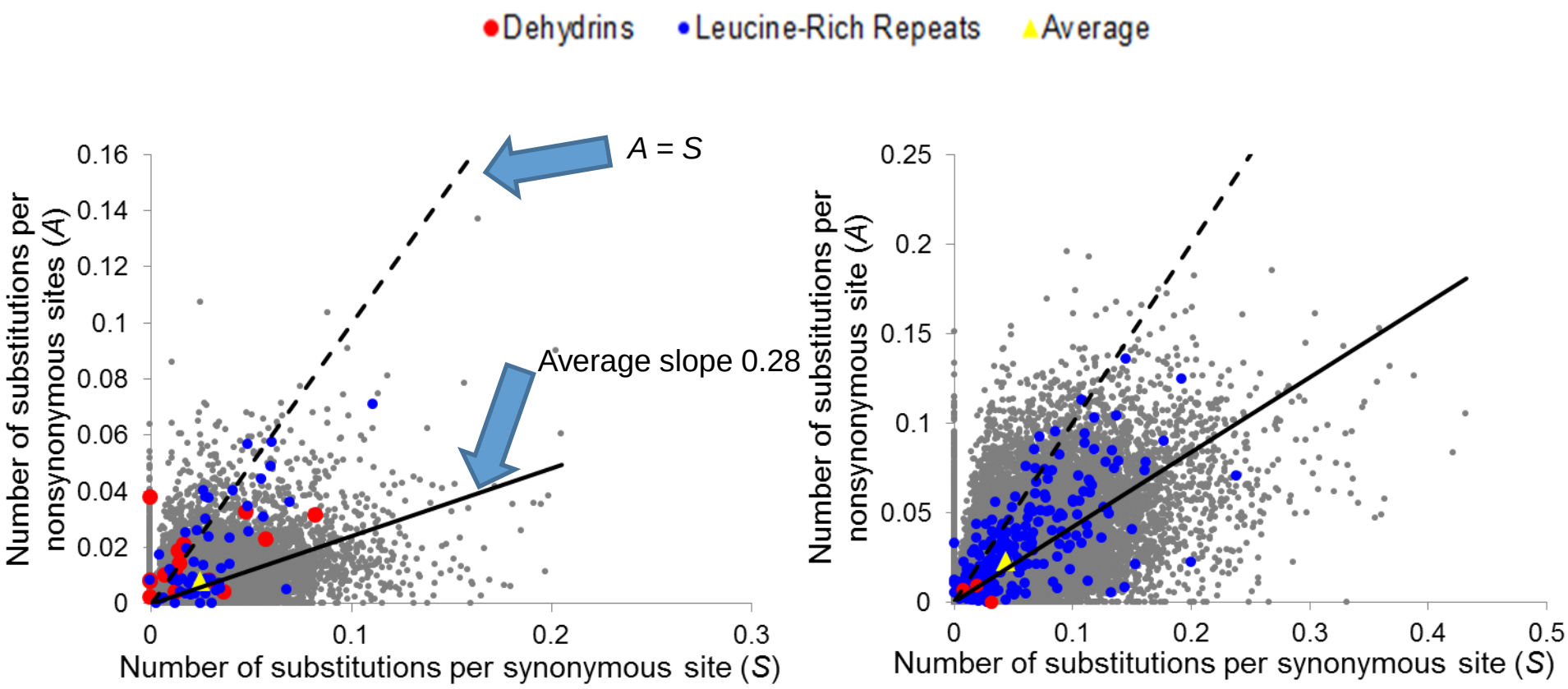
Antoine Branca[a,1], Timothy D. Paape[a], Peng Zhou[b], Roman Briskine[c], Andrew D. Farmer[d], Joann Mudge[d], Arvind K. Bharti[d], Jimmy E. Woodward[d], Gregory D. May[d], Laurent Gentzbittel[e], Cécile Ben[e], Roxanne Denny[b], Michael J. Sadowsky[f], Joëlle Ronfort[g], Thomas Bataillon[h], Nevin D. Young[a,b], and Peter Tiffin[a,2]

[a]Department of Plant Biology, University of Minnesota, Saint Paul, MN 55108; [b]Department of Plant Pathology, University of Minnesota, Saint Paul, MN 55108; [c]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455; [d]National Center for Genome Resources, Sa

SNP abundance within coding sequences: *A/S* ratios

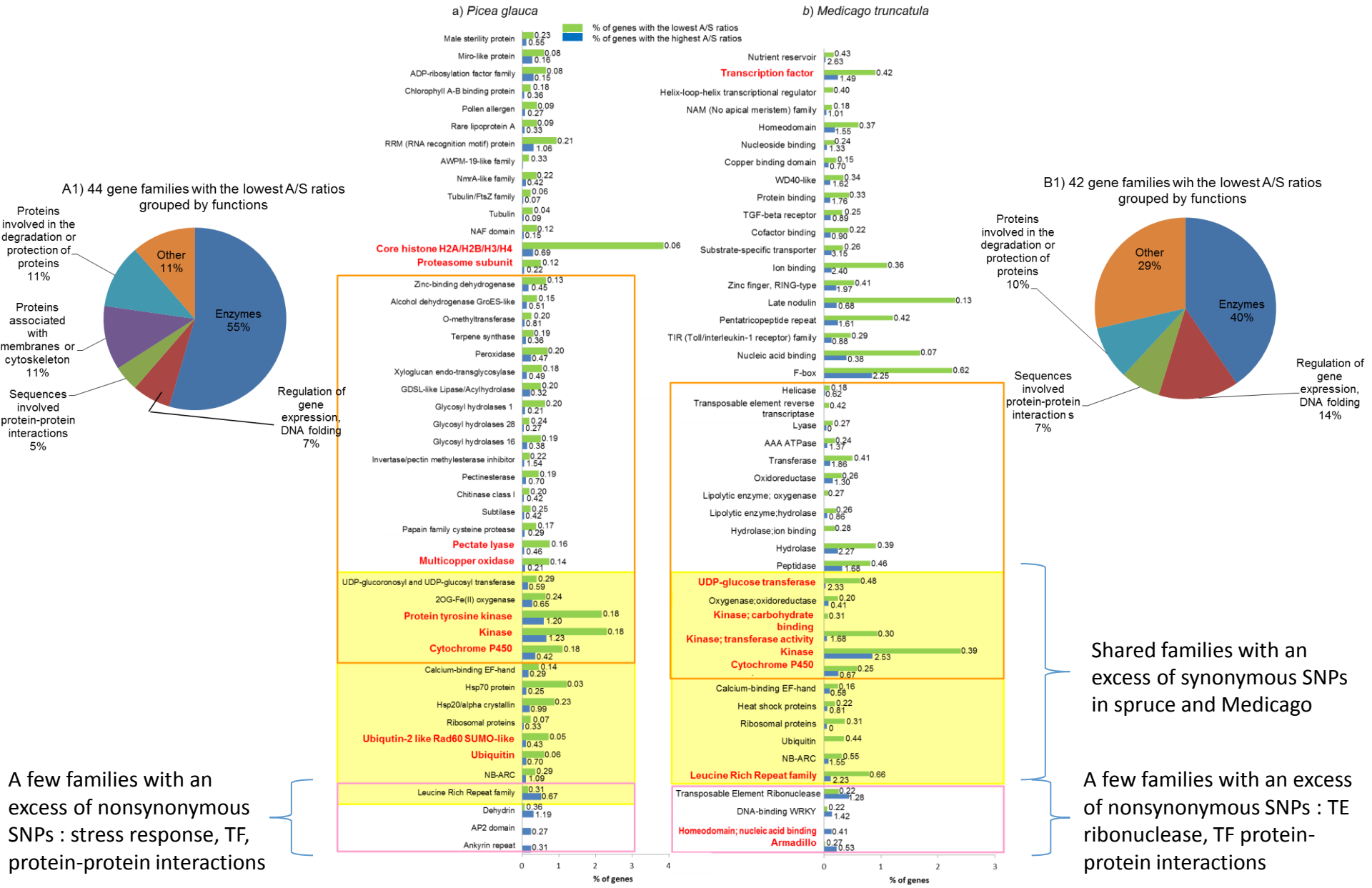
15k cDNAs (gene representatives)
P. glauca

30k cDNAs (gene representatives)
Medicago truncatula (Branca et al. 2011 PNAS)



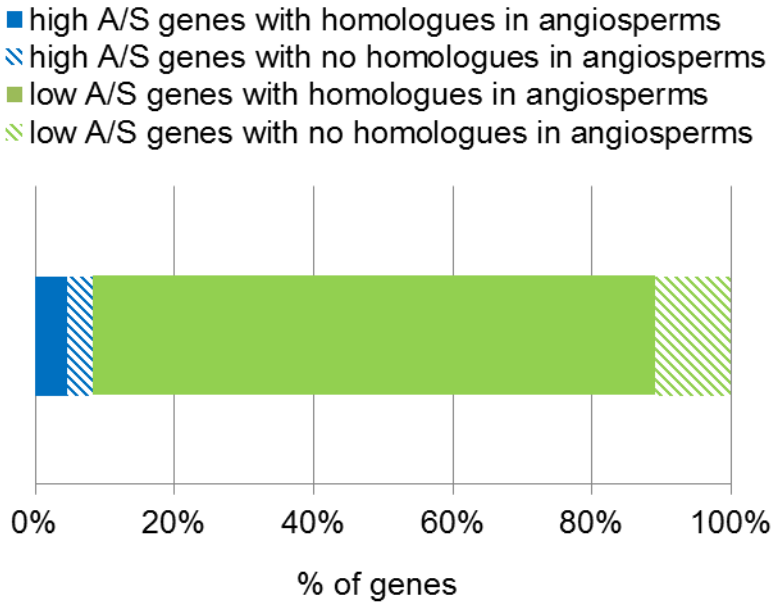
- Lower numbers of substitution in white spruce: generation time effect? bottleneck effect during ice age (Namroud et al. 2010; Pavy et al. 2012)

A/S ratio distribution across 2,500 spruce gene families (7k genes) and comparison with *Medicago*

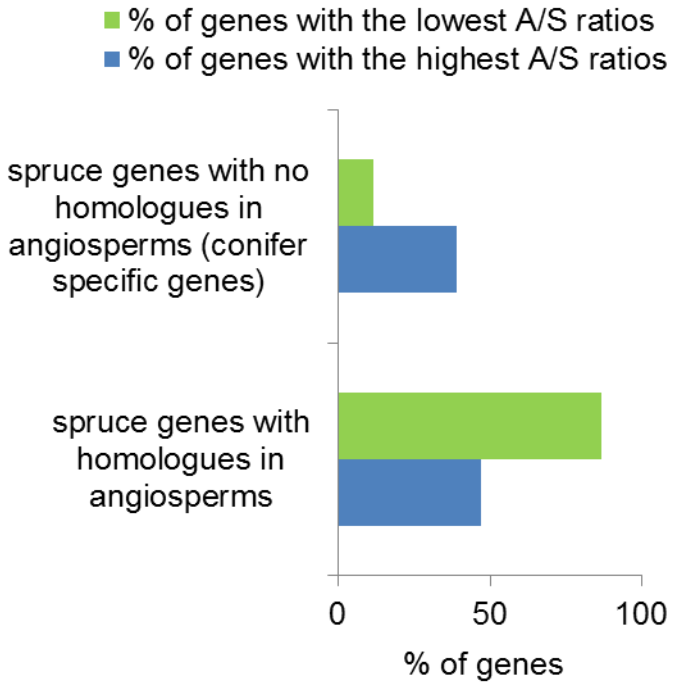


Conifer-specific genes associated to highest *A/S* ratios

➢ **2K conifer-specific genes versus 13K conserved genes with angiosperms**



14% high *A/S* (in blue)
86% low *A/S* (in green)



➢ Conifer-specific genes more abundant among genes with highest *A/S* ratios

Linking *A/S* ratios and expression patterns

Negative correlation between Ka and tissue distribution breadth (Duret & Mouchiroud 2000)

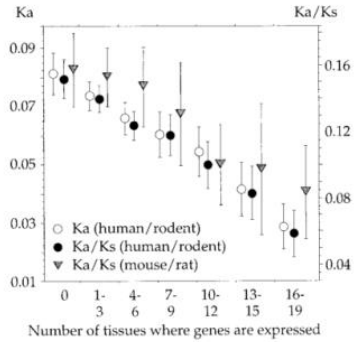


FIG. 1.—Relationship between the gene expression pattern and the nonsynonymous substitution rate (K_a) or the ratio of nonsynonymous/synonymous substitution rates (K_a/K_s). Human/rodent: N = 2,400. Mouse/rat: N = 834. Error bars indicate the 95% confidence interval.

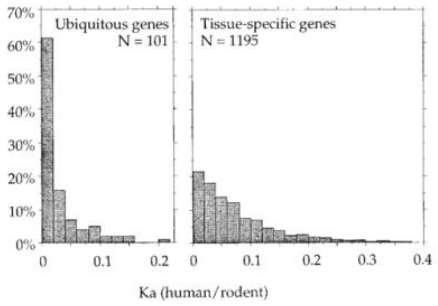


FIG. 2.—Frequency distribution of K_a values in tissue-specific genes (0–3 tissues) and ubiquitous genes (≥16 tissues).

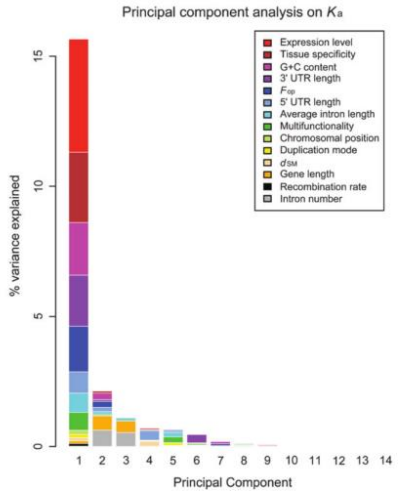
has been suggested that this correlation might be due to neighboring effects (Wolfe and Sharp 1993). Indeed, it is known that the rate of mutation at a given base is

Factors contributing to variation in evolutionary rate among *Arabidopsis* genes (Yang & Gaut 2011)

Table 2. Pairwise Correlations of Evolutionary Rates with Potentially Contributing Factors.

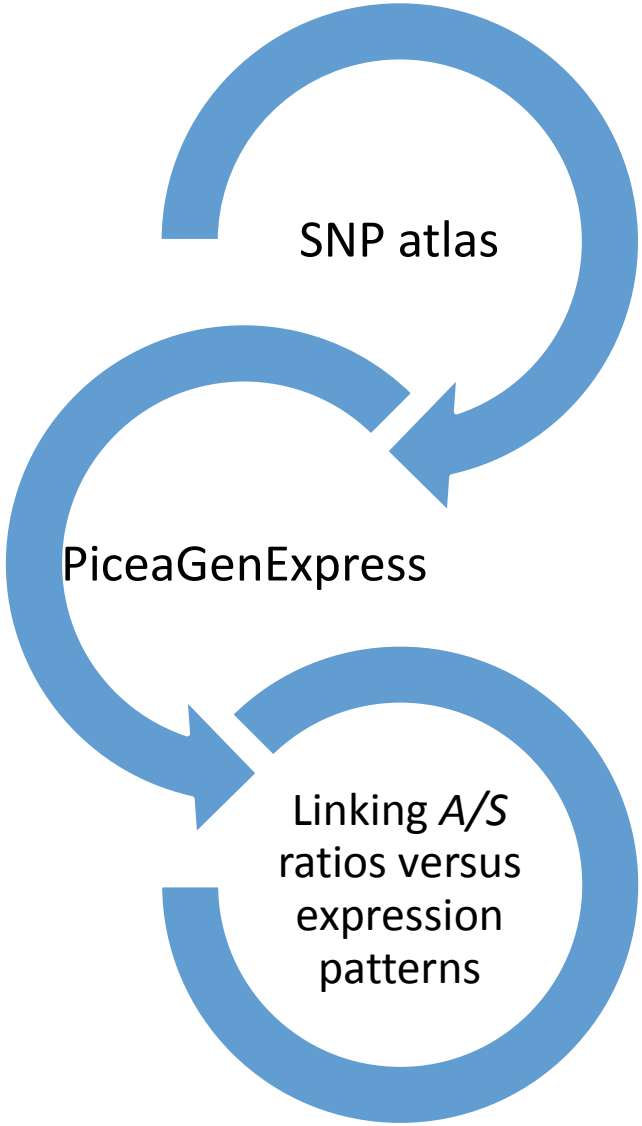
Variable	K_a	K_s	K_a/K_s
Duplication mode	0.13***	0.01	0.13***
Chromosomal position	−0.06***	−0.17***	0.001
Recombination rate	0.04*	0.12***	−0.002
Expression level	−0.42***	−0.09***	−0.39***
Tissue specificity (τ)	0.28***	0.12***	0.24***
a_{SM}	0.18***	0.11***	0.14***
F_{op}	−0.12***	0.04*	−0.14***
Multifunctionality	−0.19***	−0.03#	−0.18***
Gene length	−0.25***	−0.20***	−0.18***
5′ UTR length	−0.20***	−0.13***	−0.15***
3′ UTR length	−0.20***	−0.10***	−0.16***
Intron number	−0.19***	−0.26***	−0.10***
Average intron length	−0.11***	−0.03*	−0.09***
G + C content	−0.20***	−0.01	−0.20***

NOTE.—The coefficients were calculated based on Spearman rank correlation. #$P < 0.05$, *$P < 10^{-3}$, **$P < 10^{-6}$, *** $P < 10^{-9}$.



Gene expression best explains *Ka* variation among genes

Linking *A/S* ratios and expression patterns



Raherison *et al. BMC Genomics* 2012, **13**:434
http://www.biomedcentral.com/1471-2164/13/434

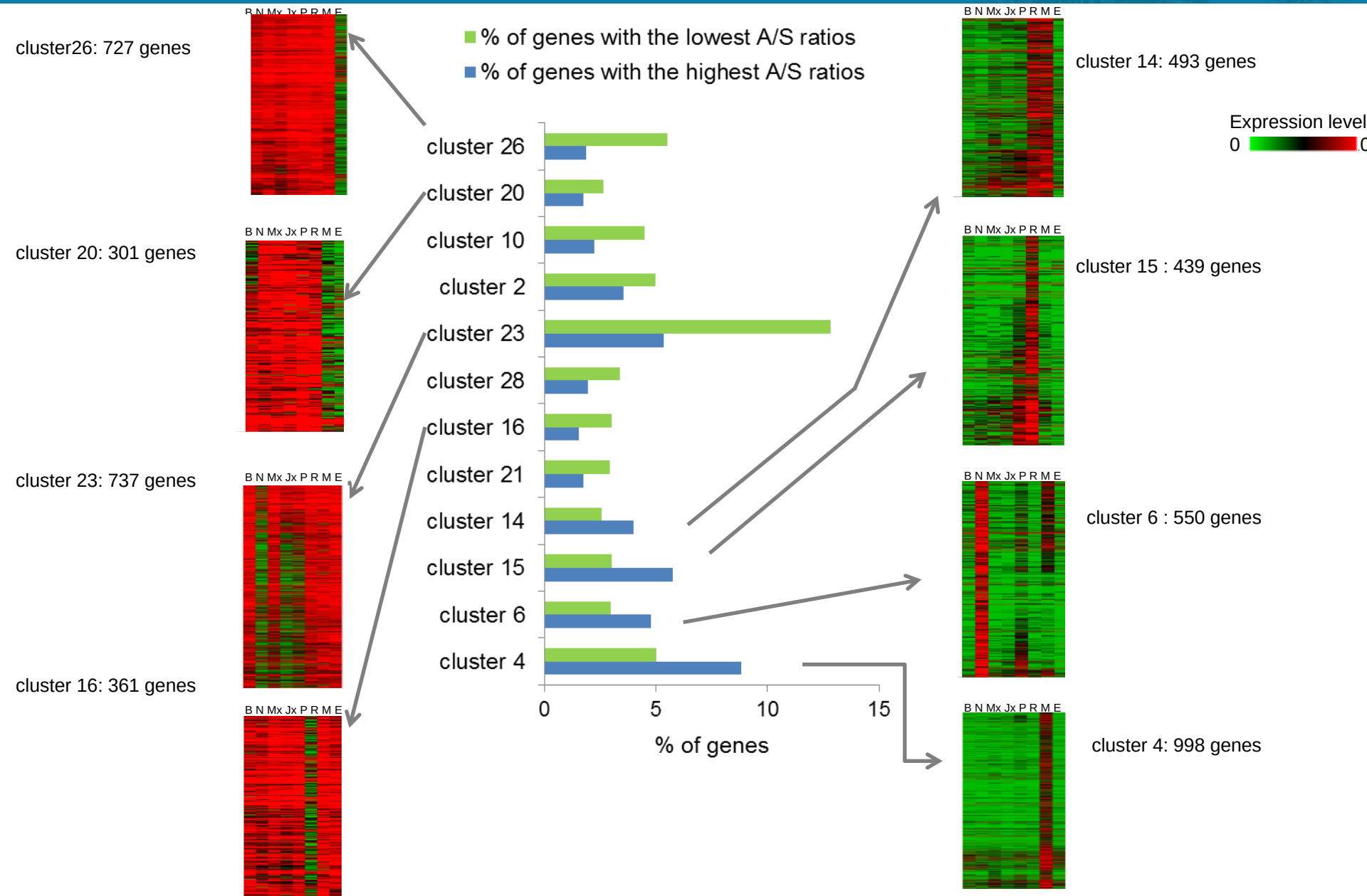
BMC Genomics

RESEARCH ARTICLE Open Access

Transcriptome profiling in conifers and the PiceaGenExpress database show patterns of diversification within gene families and interspecific conservation in vascular gene expression

Elie Raherison[1†], Philippe Rigault[2†], Sébastien Caron[1†], Pier-Luc Poulin[1], Brian Boyle[1], Jukka-Pekka Verta[1], Isabelle Giguère[1], Claude Bomal[1], Jörg Bohlmann[3] and John MacKay[1*]

Linking *A/S* ratios and expression patterns (13k genes): tissue-specific co-expressed gene clusters associated to highest *A/S*



For more information on methods and findings

Genome Biology and Evolution

GBE

The Landscape of Nucleotide Polymorphism among 13,500 Genes of the Conifer *Picea glauca*, Relationships with Functions, and Comparison with *Medicago truncatula*

Nathalie Pavy[1,*,†], Astrid Deschênes[1,†], Sylvie Blais[1], Patricia Lavigne[2], Jean Beaulieu[1,3], Nathalie Isabel[1,2], John Mackay[1], and Jean Bousquet[1]

[1]Canada Research Chair in Forest and Environmental Genomics, Centre for Forest Research and Institute for Systems and Integrative Biology, Université Laval, Québec, Canada

[2]Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, Québec, Canada

[3]Natural Resources Canada, Canadian Wood Fibre Centre, Laurentian Forestry Centre, Québec, Canada

*Corresponding author: E-mail: nathalie.pavy@sbf.ulaval.ca.

†These authors contributed equally to this work.

Accepted: September 15, 2013

Data deposition: This project has been deposited in dbSNP at: http://www.ncbi.nlm.nih.gov/SNP/snp_viewBatch.cgi?sbid=1058878.

Published in Nov. 2013


Thanks !!!
Canada Research Chair in Forest and Environmental Genomics
Canadian Forest Service
GenomeQuébec
GenomeCanada