Molecular evolution of drought tolerance and wood strength related candidate genes in loblolly pine (*Pinus taeda* L.)

By T. E. KORALEWSKI[1)], J. E. BROOKS[2)] and K. V. KRUTOVSKY[1),3),4),5),*)]

(Received 1st July 2013)

Abstract

Loblolly pine (*Pinus taeda* L.) is an intensely studied species that has become a model system for conifers. It is one of the most important commercial crops in the southeastern United States and grows across a vast territory. Due to exposure to this current diverse environment and the fluctuating climatic conditions of the past, it has likely accumulated substantial variation in adaptive trait and wood strength related genes. We merged a set of newly collected and previously published genomic DNA sequence data and analyzed them for departures from neutrality in 32 drought tolerance and wood strength related candidate genes using neutrality tests, such as Tajima's *D*, HKA, MK and nonsynonymous to synonymous substitutions ratio (*Z*-test). Three other major Southern pines closely related to *P. taeda* (*Pinus echinata* Mill., *P. elliottii* Engelm., and *P. palustris* Mill.) were used as outgroups in interspecific tests. In three loci (*4-coumarate: CoA ligase, putative cell-wall protein* and *trans-cinnamate 4-hydroxylase 2*) neutrality was rejected by both intra- and interspecific tests, consistent with purifying selection. Neutrality was also rejected in several other loci (*alpha-tubulin, arabinogalactan 4, arabinogalactan 6, cinnamate 4-hydroxylase 1, cinnamoyl CoA reductase, cinnamyl alcohol dehydrogenase, caffeoyl CoA O-methyltransferase 1, early response to drought 3, glycine hydroxymethyltransferase, ABI1 protein phosphatase 2C-like, putative wall-associated protein kinase,* and *unknown gene ug_2-498*); however, these results are difficult to interpret because only one of the tests proved significant. This study contributes to the ongoing discussion about natural selection in putative adaptive genes in loblolly pine. However, unambiguous interpretation of the results often remains problematic.

Key words: loblolly pine, longleaf pine, shortleaf pine, slash pine, *Australes*, drought tolerance, wood strength, neutrality tests, SNP, natural selection.

[1)] Department of Ecosystem Science and Management, Texas A&M University, 2138 TAMU, College Station, TX 77843-2138, USA.

[2)] Division of Natural Science, Blinn College, 2423 Blinn Blvd., Bryan, TX 77805, USA.

[3)] Department of Forest Genetics and Forest Tree Breeding, Büsgen-Institute, Georg-August-University of Göttingen, Büsgenweg 2, D-37077 Göttingen, Germany.

[4)] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow 119333, Russia.

[5)] Genome Research and Education Center, Siberian Federal University, 50a/2 Akademgorodok, Krasnoyarsk 660036, Russia.

[*)] Corresponding author: KONSTANTIN V. KRUTOVSKY. Department of Forest Genetics and Forest Tree Breeding, Büsgen-Institute, Georg-August-University of Göttingen, Büsgenweg 2, D-37077 Göttingen, Germany, Phone: +49-551-393-3537. E-Mail: kkrutov@gwdg.de

Introduction

Loblolly pine (*Pinus taeda* L., section *Trifoliis*, genus *Pinus*) is the most intensely studied Southern pine and a model species for conifers. It is wind-pollinated with continuous, expansive, extensively out-crossing populations with large effective population sizes and usually a high gene flow (KRUTOVSKY et al., 2012). Hence, *P. taeda* has a relatively low population differentiation at neutral markers (AL-RABAB'AH and WILLIAMS, 2002; CHHATRE et al., 2013; ECKERT et al., 2010; GONZALEZ-MARTINEZ et al., 2007; SCHMIDTLING et al., 1999). Three other pines from subsection Australes, slash (*P. elliottii* Engelm.), shortleaf (*P. echinata* Mill.), and longleaf (*P. palustris* Mill.), are closely related to *P. taeda*. In the case of these three species, high level of genetic diversity, low differentiation between populations, and low (or lack of) correlation between genetic distance and geographic distance have also been observed (SCHMIDTLING and HIPKINS, 1998; WILLIAMS et al., 2007; XU et al., 2008).

The current natural range of the Southern pine populations is very broad, greatly overlapping among the species, and stretches from warm-temperate to subtropical climate across 13 southeastern states of the USA, expanding into areas of diverse temperatures and rainfall. Therefore, the species have likely accumulated much variation in genes that determine adaptive traits (SCHMIDTLING, 2001), becoming fit for different habitats, including relatively good adaptation to drought. Considering, furthermore, that the Southern pines are evolutionarily young species, we have a rare opportunity to study adaptive and evolutionary processes "in progress" via comparing their nucleotide variation.

To detect departures from neutrality in nucleotide variation, a number of neutrality tests have been developed (for review see KREITMAN, 2000). The null hypothesis of neutrality is based on the neutral theory of molecular evolution developed by KIMURA (1968), which considers mutation and genetic drift as major factors that affect nucleotide genetic variation and population genetic structure. We used the Tajima's *D* (TAJIMA, 1989), HKA (HUDSON et al., 1987), MK (MCDONALD and KREITMAN, 1991) and nonsynonymous to synonymous nucleotide substitutions ratio (LI et al., 1985; NEI and GOJOBORI, 1986) tests. The HKA and MK tests are interspecific and, apart from the investigated population set data, require an outgroup.

Previous studies on selection in loblolly pine targeted candidate genes for drought tolerance, wood strength, and disease response, but interpretation of the results often proved problematic. BROWN et al. (2004) analyzed nucleotide diversity and linkage disequilibrium (LD) in 19 adaptive trait related genes in 32 individuals sam-

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics

pled from various locations within the species' natural range except Florida. Their results demonstrated high positive values of Tajima's *D* statistic in a few genes, particularly in *caffeoyl CoA O-methyltransferase 1* (*ccoaomt-1; D* = 2.81) and *cinnamyl alcohol dehydrogenase* (*cad; D* = 2.09). However, neutrality was not rejected. Interspecific tests were not used in that study.

Gonzalez-Martinez et al. (2006) studied 18 drought-stress response related genes in 32 loblolly pine megagametophytes. Using Tajima's *D* they identified a possible selective sweep in *early response to drought 3* (*erd3*) gene, but pointed out that genetic hitchhiking or a recent population expansion could produce a similar effect. Although the HKA test rejected neutrality in a pairwise comparison with *water-stress inducible protein 1* (*lp3-1*) gene, the MK test did not. Similarly, no robust conclusion was reached for *ccoaomt-1* despite significant positive Tajima's *D* statistic. Fewer haplotypes than expected were found at this locus, but none of the interspecific tests rejected neutrality. A sliding window approach allowed for identification of a few regions in *putative wall-associated protein kinase* (*ppap12*) and *ug-2_498* genes with statistically significant Tajima's *D*. No selection acting upon amino acid sequences was identified from the ratio of nonsynonymous to synonymous substitutions. *Pinus pinaster* was used as an outgroup in the interspecific tests.

These two studies examined in total 34 various drought tolerance, drought-stress response and wood strength related candidate genes in loblolly pine. Both confirmed a low level of LD and moderate nucleotide diversity, and failed to identify significant population genetic structure. Additionally, a more recent study also addressed the question of selection acting in these genes (Ersoz et al., 2010). Patterns consistent with balancing or diversifying selection were observed in *cad, ccoaomt-1* and *cinnamate 4-hydroxylase 1* (*c4h1b*); recent selective sweep or background selection in *coumarate 3-hydroxylase* (*c3h-f4r6*) and *erd3* (selective sweep could also be in progress in the latter); and possible balancing selection in *putative cell-wall protein* (*lp5*). *Pinus sylvestris* was used as an outgroup.

Although very little research has been devoted to outgroup selection in molecular evolution studies, the process has received considerable attention in the sister field of molecular phylogenetics. It is generally accepted that the outgroup choice plays a pivotal role in phylogeny reconstruction (Lyons-Weiler et al., 1998), and taxa closely related to the ingroup should often be preferred over distant ones (Graham et al., 2002; Smith, 1994). Although those conclusions do not need to be binding for molecular evolution studies, Akashi (1999) noted that departure from neutrality can be detected even in comparisons of sequences from closely related species. *P. taeda* and *P. pinaster* (section *Pinea*) may share common ancestry about 25-48 MYA, and *P. taeda* and *P. sylvestris* (section *Pinus*) – about 14-31 MYA (Gernandt et al., 2008; Willyard et al., 2007). Since the species within the subsection *Australes* are thought to begin to diverge 7-15 MYA (Willyard et al., 2007), they offer a sound alternative outgroup choice for *P. taeda*. Consequently, in this work we employed the three *Australes* species as outgroups for *P. taeda*.

We used the primers designed in the two original studies (Brown et al., 2004; Gonzalez-Martinez et al., 2006) to amplify and sequence the same orthologous genes in the four Southern pines, and studied their sequences in a comparative genomic framework. Our objective was to identify departures from neutrality and infer signatures of selection in the studied genes in loblolly pine. In relation to the three abovementioned studies, we used expanded and/or merged datasets, modified methodology and alternative outgroup species.

Material and Methods

Source of data

The DNA was extracted from haploid megagametophytes of four pine species: loblolly, slash, shortleaf, and longleaf pines. The individual seed samples were provided by Dr. Dana Nelson (US Forest Service, Southern Institute of Forest Genetics, Saucier, MS, USA). PCR primers previously developed by Brown et al. (2004) and Gonzalez-Martinez et al. (2006) were used to amplify and resequence the total of 49 amplicons (32 genes; *Table 1*) following standard PCR procedures. DNA from one or two unrelated megagametophytes of each species was sequenced. To minimize sequencing errors, both forward and reverse strands were sequenced, and a consensus sequence was obtained for each individual megagametophyte using the Sequencher computer program (ver. 4.2, Gene Codes Corporation, Ann Arbor, Michigan, USA, http://www.genecodes.com). Additional data coming from two separate groups of individual trees (Brown et al., 2004; Ersoz et al., 2010; Gonzalez-Martinez et al., 2006) were downloaded from the PopSet database of GenBank (http://www.ncbi.nlm.nih.gov).

Multiple nucleotide alignments and coding regions assignment

Multiple nucleotide alignments were generated using BioEdit (ver. 7.0.9.0; Hall, 1999) and SeaView (ver. 4.0; Galtier et al., 1996) software that implements the MUSCLE algorithm (Edgar, 2004). Genes that were sequenced in multiple segments using several (but often overlapping) amplicons generated by different primer pairs were concatenated and analyzed as a single sequence with the assumption that these segments represent the same gene. Sequences with extended gaps due to missing data (i.e. resulting from poor quality of the sequencing output) were excluded from analysis, if an alternative sequence from the same species was available. The coding regions were assigned following available data in GenBank, primarily using the population sets submitted by Brown et al. (2004) and Gonzalez-Martinez et al. (2006). Additional genomic and EST sequences from other pines and conifers, such as *Picea sitchensis* and *Pseudotsuga menziesii* were used to define exon-intron structure, if needed.

Neutrality tests

The loblolly pine nucleotide sequences newly generated in this study were merged into a single population

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics

Table 1. – Pine species, genes and numbers of individual nucleotide sequences studied.

Gene	Abbr	Amp	This study				Brown et al., 2004	Gonzalez-Martinez et al., 2006	Ersoz et al., 2010			
			Piec	Piel	Pipa	Pita	Pita	Pita	Piel	Pira	Pisy	Pita
4-coumarate:CoA ligase	*4cl*	5	1	1	1	1	32					
4-coumarate:CoA ligase (amp. 4)	*4cl*	1	1	2	2	2	30		2	2	2	32
arabinogalactan 4	*agp-4*	1	2	2	2	2	32					
arabinogalactan 6	*agp-6*	2	2	2	1	1	32					
arabinogalactan-like (amp. 2)	*agp-like*	1	2	2	2	2	32					
alpha-tubulin	*α-tubulin*	2	2	2	2	2	32					
aquaporin, membrane intrinsic protein	*aqua-MIP*	1	2	2	2	2		32				
coumarate 3-hydroxylase	*c3h*	5	2	2	2	2	28					
coumarate 3-hydroxylase (amp. 1)	*c3h*	1	2	2	2	2	26					32
cinnamate 4-hydroxylase 1 (amp. 1-5)	*c4h-1*	5	1	1	1	1	32					
cinnamate 4-hydroxylase 1 (amp. 1, 4, 5)	*c4h-1*	3	2	2	2	2	32					
trans-cinnamate 4-hydroxylase 2	*c4h-2*	1	2	2	2	2	32					
cinnamyl alcohol dehydrogenase	*cad*	1	1	1	1	1	28					
caffeoyl CoA O-methyltransferase 1	*ccoaomt-1* (*ccoaomt*)	1		1	1	1	30	32	2	2	1	
cinnamoyl-CoA reductase	*ccr-1* (*ccr*)	2	2	2	2	2	32					
cellulose synthase	*cesA3*	2	1	2	2	1	32					
caffeate O-methyltransferase	*comt-2*	2	1	1		1	30					30
calcium-dependent protein kinase	*cpk3*	1	2	2	2	2		32				
dehydrin 1	*dhn-1*	1	1			1		32				
dehydrin 2	*dhn-2*	1	2	2	2	2		32	2	2	2	
early response to drought 3	*erd3*	1	1	1	1	1		32	2	2	1	
glycine hydroxymethyltransferase	*glyhmt*	1	1	1	1	1	32					
water-stress inducible protein 3	*lp3-3*	1	1	1	1	2		32				
putative cell-wall protein	*lp5* (*lp5-like*)	1	1					32		2	2	
metallothionein-like	*mt-like*	1	1	1	1	1		32	2	2		
phenylalanine ammonia-lyase 1	*pal-1* (*pal*)	1		1	1	1	30	32	2	1	2	
ABI1 protein phosphatase 2C-like	*pp2c*	1	1	1	1	1		32				
putative wall-associated protein kinase	*ppap12*	1	1	1	1	1		32				
LIM domain protein 1 (LIM transcription factor)	*PtLIM1*	1	1	1	1	1	32					
LIM domain protein 2 (LIM transcription factor)	*PtLIM2*	1		1	1	1	32					
cysteine protease	*rd21A-like*	1		1	1	1		32		2	1	1
s-adenosylmethionine synthetase 1	*sams-1* (*sam-1*)	1	1	1	1	1	32					
s-adenosylmethionine synthetase 2	*sams-2* (*sam-2*)	1	1	1	1	1	30	32	1	2	2	
chloroplast Cu/Zn superoxide dismutase	*sod-chl*	1	1	1	1	1		32	1			
unknown; drought stress responsive (University of Georgia uniscript sequence #2_498)	*ug-2_498*	1	1	1				32				

Note: **Abbr** – abbreviation; **Amp** – number of amplicons; **Piec** – *Pinus echinata*, **Piel** – *P. elliottii*, **Pipa** – *P. palustris*, **Pita** – *P. taeda*, **Pira** – *P. radiata*, and **Pisy** – *P. sylvestris*.

set for each gene together with sequences downloaded from GenBank. Tajima's *D* statistic (Tajima, 1989) was calculated using the DnaSP software (ver. 5.10.01; Librado and Rozas, 2009). The test was run with the sliding window option, where window length and step were 100 bp and 25 bp, respectively. Indels were excluded from the analysis and were not counted in the sliding window length. The HKA test (Hudson et al., 1987) was run in a maximum likelihood framework, as implemented in MLHKA software (Wright and Charlesworth, 2004), with chain length set to 500,000. For each locus, the null hypothesis (neutrality) was compared with the

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics

alternative hypothesis (the locus is potentially under selection). Maximum likelihood ratio test was performed to test for significance. The MK test (KREITMAN, 2000; MCDONALD and KREITMAN, 1991; for review see WRAY et al., 2003) was performed also using DnaSP. Substitutions only in coding regions were considered in this case. Both HKA and MK tests were used to compare the loblolly pine set with three other Southern pine species (*P. echinata*, *P. elliottii* and *P. palustris*), and, in a few cases, also with *Pinus radiata* D. Don and *Pinus sylvestris* L. Additionally, neutrality was tested through analysis of nonsynonymous to synonymous substitutions ratio as implemented in the MEGA software (ver. 5; TAMURA et al., 2011). *Z*-test was run to test the null hypothesis, assuming normal distribution of the test statistic. Sites with gaps or missing data were deleted in pairwise comparisons. Standard error was computed through bootstrap with 1,000 replicates. The Nei-Gojobori's nucleotide substitution model was used to calculate d_S and d_N (JUKES and CANTOR, 1969). In all tests results were considered significant at $\alpha = 0.05$ significance level.

Results

Data for 34 genes were initially collected. Due to poor read quality, the sequence for *lp3-1* and *ferritin* genes could not be determined, and some other sequences were trimmed. There was no available data to define exon-intron structure for *ug-2_498*. Consequently, coding regions in 31 genes were assigned and the data for 32 genes were submitted to NCBI GenBank (http://www.ncbi.nlm.nih.gov; accession numbers KF158811–KF158983).

Tajima's D test

Overall positive Tajima's *D* values were significant in only two genes (*Table 2*): *cad* ($P < 0.05$) and *ccoaomt-1*

Table 2. – Tajima's *D* neutrality test for *Pinus taeda*.

Gene	*D*	D_{coding}	D_{syn}	D_{nonsyn}	D_{silent}	Sliding window, *D* (bp)	
4cl (amp. 4)	1.584	1.258	1.258	n.a.	1.585	**2.138***	(73-172)
						2.418*	(98-197)
						2.649*	(123-222)
						2.369*	(148-247)
agp-4	0.304	1.773#	-0.240	**2.212***	-0.805	1.718#	(125-224)
agp-6	0.222	-0.534	-0.551	-0.413	0.465	-1.684#	(51-150)
						-1.895*	(101-200)
						-1.610#	(126-225)
c4h-1 (amp. 1-5)	-0.583	-0.236	-0.805	0.668	-0.801	**2.126***	(1-100)
						2.126*	(26-125)
c4h-1 (amp. 1, 4, 5)	-0.670	-0.241	-0.798	0.648	-1.241	**2.149***	(1-100)
						2.149*	(26-125)
cad	**2.077***	1.026	0.293	1.595	1.827#	**2.094***	(1-120)
						2.405*	(221-320)
						2.094*	(246-345)
						2.094*	(271-370)
ccoaomt-1	**3.028****	1.471	1.471	n.a.	**3.028****	2.026#	(1-127)
						2.566*	(53-152)
						2.785**	(78-177)
						2.785**	(103-202)
						2.785**	(128-227)
						2.617*	(153-252)
						1.989#	(178-277)
						2.310*	(203-302)
						2.433*	(228-327)
						2.433*	(253-352)
						2.644*	(278-377)
						2.433*	(303-418)
						1.989#	(328-443)
erd3	**-2.103***	-1.502	n.a.	-1.502	**-1.888***	–	
ppap12	0.723	0.723	**2.500***	-0.389	**2.500***	-1.728#	(82-181)
						-1.728#	(107-206)
						2.172*	(232-331)
						2.731**	(257-356)
						2.500*	(282-378)
ug-2_498	-1.224	–	–	–	–	**-2.008***	(295-417)

Note: Only loci with results statistically significant at $\alpha = 0.05$ are presented. Significance level: # $P < 0.10$; * $P < 0.05$; ** $P < 0.01$. In bold are values of *D* where $P < 0.05$.

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics

Table 3. – Interspecific HKA neutrality test results for *Pinus taeda* (*P*-values).

Gene	Outgroup			
	Piec	**Piel**	**Pipa**	**Pisy**
4cl	**0.020**	0.444	n/c	–
4cl (amp. 4)	**0.049**	n/c	0.539	0.460
α-tubulin	**0.016**	0.638	n/c	–
c4h-2	0.858	**0.023**	n/c	–
ccr-1	n/c	n/c	**0.021**	–
glyhmt	**0.039**	0.360	0.111	–
lp5	0.410	–	–	**0.000**
pp2c	**0.003**	0.243	0.128	–

Note: Only loci with results statistically significant at $\alpha = 0.05$ are presented. In bold are *P*-values where $P < 0.05$. **Piec** – *Pinus echinata*, **Piel** – *P. elliottii*, **Pipa** – *P. palustris*, **Pisy** – *P. sylvestris*.

($P < 0.01$, mainly due to silent mutations: $P < 0.01$). The sliding window showed that the evidence came from both coding and noncoding regions ($P < 0.05$).

The overall *D* was not significant in *arabinogalactan 4* (*agp-4*); however, *D* based on nonsynonymous substitutions was positive ($P < 0.05$). Similarly, positive Tajima's *D* values were observed for all synonymous substitutions in *ppap12* ($P < 0.05$) and at the 3' end of the amplicon ($P < 0.05$). The sliding window approach identified positive Tajima's *D* values ($P < 0.05$) in *cinnamate 4-hydroxylase 1* (*c4h-1*; the beginning of the first exon), and in the set of 64 sequences of *4-coumarate: CoA ligase* (*4cl*, amplicon 4; the region stretching from the end of the first exon through the intron to the beginning of the second exon) despite not significant overall *D* in both cases.

Erd3 was the only gene with negative overall *D* ($P < 0.05$), primarily due to silent substitutions. The sliding window showed negative *D* for the first exon in *arabinogalactan 6* (*agp-6*; $P < 0.05$) and for the middle region of *ug-2_498* ($P < 0.05$).

HKA and MK tests

The HKA test rejected neutrality at several loci: *4cl, alpha-tubulin* (*α-tubulin*), *trans-cinnamate 4-hydroxylase 2* (*c4h-2*), *cinnamoyl-CoA reductase* (*ccr-1*), *glycine hydroxymethyltransferase* (*glyhmt*), *putative cell-wall protein* (*lp5*), and *ABI1 protein phosphatase 2C-like* (*pp2c*), but results depended strictly on the outgroup species used (*Table 3*). No significant values were observed in the MK test.

Nonsynonymous to synonymous substitutions ratio test

Highly significant values were observed in *4cl* ($Z = -2.871$; $P = 0.005$; the *P*-value changed to 0.058 when only the fourth amplicon with the expanded population set was analyzed), *c4h-2* ($Z = -2.340$; $P = 0.021$), and *lp5* ($Z = -3.102$; $P = 0.002$). Notably, the HKA test rejected neutrality in these three genes as well.

Discussion

Southern pines are an evolutionarily relatively young and closely related group. It has been hypothesized that during the last glacial period that ended about 15,000 years ago (equivalent to about 600 loblolly pine generations) their range was within the regions of central Florida and the Caribbean (Jackson et al., 2000; Schmidtling et al., 1999; Wells et al., 1991). In addition, Mexico and Southern Texas (including the Lost Pines population in Bastrop County, isolated from the main area) have been proposed as a western refugium of *Pinus taeda* (Al-Rabab'ah and Williams, 2004; Schmidtling et al., 1999; Wells et al., 1991), which implies two historically separated refugia of loblolly pine: east and west of the Mississippi river, respectively (Al-Rabab'ah and Williams, 2002). Therefore, in the case of Southern pines, recent demographic expansion could seriously affect nucleotide variation and complicate the search for signatures of natural selection. Both demographic events and selection can leave similar signatures in the genome that are often very difficult to dissect (Ersoz et al., 2010; Gonzalez-Martinez et al., 2006).

In this study we applied multiple tests to examine the data from various perspectives. To aid dissection of potential selection from recent demographic events, we used interspecific comparisons (Hudson et al., 1987; Kreitman, 2000; McDonald and Kreitman, 1991). The HKA test rejected neutrality in a few genes (*4cl*, *α-tubulin*, *c4h-2*, *ccr-1*, *glyhmt*, *lp5*, and *pp2c*); however, the MK test, which is more robust to a potential bias due to demographic processes, was not significant for any gene. In the latter case, possible slow changes in some genes might have been missed, while they could have been more evident in the multilocus framework of the MLHKA. The *Z*-test rejected neutrality in three of these genes (*4cl*, *c4h-2* and *lp5*), implying an excess of synonymous substitutions ($Z < 0$), consistent with purifying selection. *4cl* was the only locus in this group with significant Tajima's *D* (sliding window, $D > 0$).

Previously, Ersoz et al. (2010) concluded lower mutation rate at *pp2c*; our observations are consistent with that. However, their study attributed the variation in *lp5* to potential balancing selection, primarily due to ancestral alleles present in both *P. taeda* and *P. sylvestris*. Given the highly significant *Z*-test score showing underrepresentation of nonsynonymous substitutions ($Z < 0$), and the robustness of this test, the case of potential purifying selection acting upon this locus seems to be well supported, too.

The overall Tajima's *D* was negative only in *erd3* ($P < 0.05$). Other tests failed to reject neutrality at this locus, although the HKA test showed a low, but not significant, *P*-value ($P = 0.069$) in comparison *P. taeda* vs. *P. sylvestris*. Ersoz et al. (2010) hypothesized a possible selective sweep at *erd3*; however, the silent sites were the only data partition here with significant excess of rare alleles ($P < 0.05$), which does not seem to support a selective sweep alone. A plausible explanation could be that the sweep has been completed and that currently purifying selection is acting upon the locus. When slid-

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics

ing window was applied, patterns consistent with purifying selection were found also in *ug_2-498*, reported earlier by GONZALEZ-MARTINEZ et al. (2006), and *agp-6* (coding region). Possible purifying selective pressure in these two cases might have been too weak to be detected by other tests, including overall Tajima's *D*.

Tajima's *D* test was significant in several other loci. Not supported by the other tests, these results are difficult to interpret, as they may likely reflect a combination of demographic and adaptive processes; dissection of the two factors remains challenging. The overall Tajima's *D* was positive in *ccoaomt-1* and *cad* ($P < 0.01$ and $P < 0.05$, respectively), consistent with balancing or positive selection. High positive Tajima's *D* value in both loci was reported by BROWN et al. (2004) and also in *ccoaomt-1* by GONZALEZ-MARTINEZ et al. (2006). In both studies the test result was attributed to dimorphism in haplotype lineages. ERSOZ et al. (2010) associated the observed patterns at these two loci with potential balancing or diversifying selection. In our study the expanded *P. taeda* sample sets did not affect the statistical significance of the Tajima's *D* test, and remains consistent with the previous studies. Considering, however, that in both genes most of Tajima's *D* signal came from silent substitutions ($P < 0.1$ in *cad*, and $P < 0.01$ in *ccoaomt-1*), potential balancing selection may be acting primarily upon introns, possibly in order to lower the cost of transcription and to maintain correct splicing and the appropriate level of affinity for regulatory factors (CARVALHO and CLARK, 1999; CASTILLO-DAVIS et al., 2002; COLLINS, 1988). Alternatively, selection and demographic effects acting simultaneously may be confounding, which could result in such localized non-neutral variation.

In some cases the sliding window approach provided additional insights. *Agp-4* showed positive Tajima's *D* due to nonsynonymous substitutions ($P < 0.05$) primarily in the area at the end of the coding region. Tajima's *D* was positive ($P < 0.05$) also in a short region of *c4h-1* spanning part of the 5'UTR and the beginning of the first exon. In *ppap12* Tajima's *D* was significant for the synonymous substitutions ($P < 0.05$), and sliding window detected a region with $P < 0.01$. Interestingly, in *ppap12* apart from regions showing positive *D* values, a section of the coding region was identified with a negative *D* ($P < 0.10$). Such pattern may indicate that different factors, possibly opposite forms of selection, can affect various regions of the same gene. Therefore, overall *D* statistic at the locus level and lack of significance can be misleading, possibly due to averaging of opposite effects.

The evidence of positive selection remains elusive despite various approaches applied to multiple loci. *Erd3*, the only locus in which possible selective sweep has been detected (ERSOZ et al., 2010; GONZALEZ-MARTINEZ et al., 2006), requires further investigation. This fact, however, does not imply a lack of adaptation in this group of genes. Instead, it could mean that the molecular evolution mechanisms and the adaptive processes slowed down or concluded sometime in the past, and balancing and purifying selection maintains the adaptive variants. Development of new analytical tools, including ones that would consider a functional group of genes as a whole, or expanded datasets could help to advance this inquiry.

Some of the genes studied here have been considered candidates for drought-stress response (CHANG et al., 1996; GONZALEZ-MARTINEZ et al., 2006; GONZALEZ-MARTINEZ et al., 2008). Notably, both intra- and interspecific tests rejected neutrality in *lp5*, but non-neutral patterns were also observed in *ccoaomt-1*, *erd3*, *ppap12* and *ug-2_498* (Tajima's *D*), and in *pp2c* (HKA). Other genes associated with xylem cell-wall biosynthesis, including lignin biosynthesis, play important roles in water transportation within the plant (SPERRY, 2003) – in particular *4cl* and *c4h-2* (BROWN et al., 2003; PETER and NEALE, 2004) where both HKA and *Z*-test rejected neutrality. Recurrent periods of drought, as shown by paleoclimatic records (SEAGER et al., 2009), are a strong evidence for historical cyclic selective pressures, agreeing with both intra- and interspecific test results and shaping drought adaptations in loblolly pine. Although there is much discourse regarding the causes, amplitude and dynamics of the present and future change of climate, there is evidence that we are experiencing climate change on the global scale and modeling efforts continue to target the question of future trends (ADAMS et al., 1990; DESER et al., 2012; HOEGH-GULDBERG and BRUNO, 2010; HUANG et al., 2011; THUILLER et al., 2011). Intensified focus upon these loci may aid in shaping strategies for breeding trees better prepared for the forthcoming climate challenges (KRUTOVSKY et al., 2013).

Perhaps the biggest difficulty in molecular evolution studies is to discriminate between natural selection and demographic events. Although a negative Tajima's *D* value could indicate a negative selective pressure as well as a recent population expansion, when coupled with a significant value of the *Z*-test, they could be a strong indication of purifying selection. Similarly, positive Tajima's *D* values alone could denote a fast expansion following a bottleneck, but when accompanied by significant *Z*-test values they can be a strong signature of balancing selection.

Some of the robust evidence for significant Tajima's *D* values came from silent substitutions and often was strongly localized as detected via sliding window. Such pattern could be an indication of selection rather than a result of recent population expansion. GONZALEZ-MARTINEZ et al. (2006) found no evidence for population structure in the studied area using 21 unlinked nuclear microsatellite markers representing most of loblolly pine linkage groups. F_{ST} was also low among the three studied regions. They observed, however, skewed genome-wide Tajima's *D* distribution toward negative values, noting its significance as a signature of population growth. AL-RABAB'AH and WILLIAMS (2002) found only slight population differentiation across the loblolly pine range, but identified genetic differentiation between the two main parts of the loblolly pine range, i.e. east and west of the Mississippi river, respectively. If we assume that the observed pattern is due to recent post-glacial expansion (a demographic event), then this genome-wide effect of skewedness would be expected in most, if

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics

not all, studied loci, while selection usually affects only a group of genes, and its effect could be very different depending on the form of selection.

The wide spectrum of results from the neutrality tests may be interpreted as evidence of various factors influencing the pine genome. Evidently, the effects of the processes shaping variation on the molecular level are not homogeneous. Highly significant positive Tajima's *D* scores in certain loci and highly significant negative *D* in others cannot be easily explained by population expansion alone, especially when neutrality is also rejected by other tests. Very likely, both population expansion and selective pressures have been acting simultaneously. Further studies focusing not only on *P. taeda* but also on other Southern pines and outgroup species that face various environmental challenges may help to better understand the effects of these processes.

Acknowledgements

We are grateful to Dr. Alan E. Pepper, Dr. Clare A. Gill, Dr. Mariana Mateos and Dr. Ruzong Fan (Texas A&M University) for valuable discussions, suggestions and comments. We thank anonymous reviewers for their comments that helped improve the manuscript. TEK would like to thank Dr. Thomas D. Byram for his support during this study. The project was supported by the United States Department of Agriculture (USDA) Cooperative State Research, Education, and Extension Service (CSREES) and Texas Agricultural Experiment Station (TAES) McIntire-Stennis Project (TEX09122-0210381). The Pine Integrated Network: Education, Mitigation, and Adaptation Project (PINEMAP), a Coordinated Agricultural Project funded by the USDA National Institute of Food and Agriculture, Award #2011-68002-30185, provided support during preparation of this manuscript.

References

Adams, R. M., C. Rosenzweig, R. M. Peart, J. T. Ritchie, B. A. Mccarl, J. D. Glyer, R. B. Curry, J. W. Jones, K. J. Boote and L. H. Allen (1990): Global climate change and US agriculture. Nature **345**(6272): 219–224.

Akashi, H. (1999): Within- and between-species DNA sequence variation and the 'footprint' of natural selection. Gene **238**(1): 39–51.

Al-Rabab'ah, M. A. and C. G. Williams (2002): Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. Forest Ecol Manag **163**(1–3): 263–271.

Al-Rabab'ah, M. A. and C. G. Williams (2004): An ancient bottleneck in the Lost Pines of central Texas. Mol Ecol **13**(5): 1075–1084.

Brown, G. R., D. L. Bassoni, G. P. Gill, J. R. Fontana, N. C. Wheeler, R. A. Megraw, M. F. Davis, M. M. Sewell, G. A. Tuskan and D. B. Neale (2003): Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.). III. QTL verification and candidate gene mapping. Genetics **164**(4): 1537–1546.

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley and D. B. Neale (2004): Nucleotide diversity and linkage disequilibrium in loblolly pine. P Natl Acad Sci USA **101**(42): 15255–15260.

Carvalho, A. B. and A. G. Clark (1999): Intron size and natural selection. Nature **401**(6751): 344–344.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov (2002): Selection for short introns in highly expressed genes. Nat Genet **31**(4): 415–418.

Chang, S. J., J. D. Puryear, M. A. D. L. Dias, E. A. Funkhouser, R. J. Newton and J. Cairney (1996): Gene expression under water deficit in loblolly pine (*Pinus taeda*): Isolation and characterization of cDNA clones. Physiol Plantarum **97**(1): 139–148.

Chhatre, V. E., T. D. Byram, D. B. Neale, J. L. Wegrzyn and K. V. Krutovsky (2013): Genetic structure and association mapping of adaptive and selective traits in the east Texas loblolly pine (*Pinus taeda* L.) breeding populations. Tree Genet Genomes **9**(5): 1161–1178.

Collins, R. A. (1988): Evidence of natural selection to maintain a functional domain outside of the 'core' in a large subclass of Group I introns. Nucleic Acids Res **16**(6): 2705–2715.

Deser, C., A. Phillips, V. Bourdette and H. Y. Teng (2012): Uncertainty in climate change projections: the role of internal variability. Clim Dynam **38**(3–4): 527–546.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra, S. C. Gonzalez-Martinez and D. B. Neale (2010): Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics **185**(3): 969–982.

Edgar, R. C. (2004): MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**(5): 1792–1797.

Ersoz, E. S., M. H. Wright, S. C. Gonzalez-Martinez, C. H. Langley and D. B. Neale (2010): Evolution of disease response genes in loblolly pine: Insights from candidate genes. PLoS One **5**(12): e14234.

Galtier, N., M. Gouy and C. Gautier (1996): SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci **12**(6): 543–548.

Gernandt, D. S., S. Magallon, G. G. Lopez, O. Z. Flores, A. Willyard and A. Liston (2008): Use of simultaneous analyses to guide fossil-based calibrations of Pinaceae phylogeny. Int J Plant Sci **169**(8): 1086–1099.

Gonzalez-Martinez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler and D. B. Neale (2006): DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics **172**(3): 1915–1926.

Gonzalez-Martinez, S. C., D. Huber, E. Ersoz, J. M. Davis and D. B. Neale (2008): Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. Heredity **101**(1): 19–26.

Gonzalez-Martinez, S. C., N. C. Wheeler, E. Ersoz, C. D. Nelson and D. B. Neale (2007): Association genetics in *Pinus taeda* L. I. Wood property traits. Genetics **175**(1): 399–409.

Graham, S. W., R. G. Olmstead and S. C. H. Barrett (2002): Rooting phylogenetic trees with distant outgroups: A case study from the commelinoid monocots. Mol Biol Evol **19**(10): 1769–1781.

Hall, T. A. (1999): BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series **41**: 95–98.

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics

HOEGH-GULDBERG, O. and J. F. BRUNO (2010): The impact of climate change on the world's marine ecosystems. Science **328**(5985): 1523–1528.

HUANG, J., B. ABT, G. KINDERMANN and S. GHOSH (2011): Empirical analysis of climate change impact on loblolly pine plantations in the southern United States. Nat Resour Model **24**(4): 445–476.

HUDSON, R. R., M. KREITMAN and M. AGUADE (1987): A test of neutral molecular evolution based on nucleotide data. Genetics **116**(1): 153–159.

JACKSON, S. T., R. S. WEBB, K. H. ANDERSON, J. T. OVERPECK, T. WEBB, J. W. WILLIAMS and B. C. S. HANSEN (2000): Vegetation and environment in Eastern North America during the Last Glacial Maximum. Quaternary Sci Rev **19**(6): 489–508.

JUKES, T. H. and C. R. CANTOR (1969): Evolution of protein molecules, pp. 21–132. *In:* Mammalian protein metabolism, edited by M. N. MUNRO. Academic Press, New York.

KIMURA, M. (1968): Evolutionary rate at molecular level. Nature **217**(5129): 624–626.

KREITMAN, M. (2000): Methods to detect selection in populations with applications to the human. Annu Rev Genom Hum G **1**: 539–559.

KRUTOVSKY, K., T. BYRAM, R. WHETTEN, N. WHEELER, D. NEALE, M. LU, T. KORALEWSKI and C. LOOPSTRA (2013): PINEMAP + PineRefSeq = Future Forests. PINEMAP (Pine Integrated Network: Education, Mitigation, and Adaptation Project) Year 2 Annual Report | March 2012-February 2013 "Mapping the future of southern pine management in a changing world" (http://www.pinemap.org/reports/annual-reports/PINEMAP_Year_2_Annual_Report_FINAL.pdf): 26-27.

KRUTOVSKY, K. V., J. BURCZYK, I. CHYBICKI, R. FINKELDEY, T. PYHÄJÄRVI and J. J. ROBLEDO-ARNUNCIO (2012): Gene flow, spatial structure, local adaptation, and assisted migration in trees, pp. 71–116. *In:* Genomics of Tree Crops, edited by R. J. SCHNELL and P. M. PRIYADARSHAN. Springer.

LI, W. H., C. I. WU and C. C. LUO (1985): A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol **2**(2): 150–174.

LIBRADO, P. and J. ROZAS (2009): DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25**(11): 1451–1452.

LYONS-WEILER, J., G. A. HOELZER and R. J. TAUSCH (1998): Optimal outgroup analysis. Biol J Linn Soc **64**(4): 493–511.

MCDONALD, J. H. and M. KREITMAN (1991): Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351**(6328): 652–654.

NEI, M. and T. GOJOBORI (1986): Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3**(5): 418–426.

PETER, G. and D. NEALE (2004): Molecular basis for the evolution of xylem lignification. Curr Opin Plant Biol **7**(6): 737–742.

SCHMIDTLING, R. C. (2001): Southern pine seed sources. USDA, Forest Service, Southern Research Station. GTR SRS-44.

SCHMIDTLING, R. C., E. CARROLL and T. LAFARGE (1999): Allozyme diversity of selected and natural loblolly pine populations. Silvae Genet **48**(1): 35–45.

SCHMIDTLING, R. C. and V. HIPKINS (1998): Genetic diversity in longleaf pine (*Pinus palustris*): influence of historical and prehistorical events. Can J Forest Res **28**(8): 1135–1145.

SEAGER, R., A. TZANOVA and J. NAKAMURA (2009): Drought in the southeastern United States: Causes, variability over the last millennium, and the potential for future hydroclimate change. J Climate **22**(19): 5021–5045.

SMITH, A. B. (1994): Rooting molecular trees: problems and strategies. Biol J Linn Soc **51**(3): 279–292.

SPERRY, J. S. (2003): Evolution of water transport and xylem structure. Int J Plant Sci **164**(3): S115–S127.

TAJIMA, F. (1989): Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**(3): 585–595.

TAMURA, K., D. PETERSON, N. PETERSON, G. STECHER, M. NEI and S. KUMAR (2011): MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28**(10): 2731–2739.

THUILLER, W., S. LAVERGNE, C. ROQUET, I. BOULANGEAT, B. LAFOURCADE and M. B. ARAUJO (2011): Consequences of climate change on the tree of life in Europe. Nature **470**(7335): 531–534.

WELLS, O. O., G. L. SWITZER and R. C. SCHMIDTLING (1991): Geographic variation in Mississippi loblolly pine and sweetgum. Silvae Genet **40**(3–4): 105–119.

WILLIAMS, D. A., Y. Q. WANG, M. BORCHETTA and M. S. GAINES (2007): Genetic diversity and spatial structure of a keystone species in fragmented pine rockland habitat. Biol Conserv **138**(1–2): 256–268.

WILLYARD, A., J. SYRING, D. S. GERNANDT, A. LISTON and R. CRONN (2007): Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol Biol Evol **24**(1): 90–101.

WRAY, G. A., M. W. HAHN, E. ABOUHEIF, J. P. BALHOFF, M. PIZER, M. V. ROCKMAN and L. A. ROMANO (2003): The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol **20**(9): 1377–1419.

WRIGHT, S. I. and B. CHARLESWORTH (2004): The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. Genetics **168**(2): 1071–1076.

XU, S. Q., C. G. TAUER and C. D. NELSON (2008): Genetic diversity within and among populations of shortleaf pine (*Pinus echinata* Mill.) and loblolly pine (*Pinus taeda* L.). Tree Genet Genomes **4**(4): 859–868.

DOI:10.1515/sg-2014-0009
edited by Thünen Institute of Forest Genetics