Molecular Population Genetics of Herbivore-induced Protease Inhibitor Genes in European Aspen (Populus tremula L., Salicaceae)

Pär K. Ingvarsson
Umeå Plant Science Centre, Department of Ecology and Environmental Science, University of Umeå, SE-891 87 Umeå, Sweden

Plants defend themselves against the attack of natural enemies by using an array of both constitutively expressed and induced defenses. Long-lived woody perennials are overrepresented among plant species that show strong induced defense responses, whereas annual plants and crop species are underrepresented. However, most studies of plant defense genes have been performed on annual or short-lived perennial weeds or crop species. Here I use molecular population genetic methods to survey six wound-inducible protease inhibitors (PIs) in a long-lived woody, perennial plant species, the European aspen (Populus tremula), to evaluate the likelihood of either recurrent selective sweeps or balancing selection maintaining amino acid polymorphisms in these genes. The results show that none of the six PI genes have reduced diversities at synonymous sites, as would be expected in the presence of recurrent selective sweeps. However, several genes show some evidence of nonneutral evolution such as enhanced linkage disequilibrium and a large number of high-frequency–derived mutations. A group of at least four Kunitz trypsin inhibitor genes appear to have experienced elevated levels of nonsynonymous substitutions, indicating allelic turnover on an evolutionary timescale. One gene, TII, has enhanced levels of intraspecific polymorphism at nonsynonymous sites and also has an unusual haplotype structure characterized by two divergent haplotypes occurring at roughly equal frequencies in the sample. One haplotype has very low levels of intraallelic nucleotide diversity, whereas the other haplotype has levels of diversity comparable to other genes in P. tremula. Patterns of sequence diversity at TII do not fit a simple model of either balancing selection or recurrent selective sweeps. This suggests that selection at TII is more complex, possibly involving allelic cycling.

Introduction

Plants have evolved numerous adaptations to defend themselves against attack by herbivores, including mechanisms to deter, starve, or poison potential herbivores (Marquis 1992). Herbivores are expected, however, to respond to plant defenses by evolving counteradaptations that render defenses less effective or even useless. Such counterselection can ultimately lead to an “arms race” between plants and herbivores, and theoretical models predict that such coevolutionary interactions will lead to an escalation of traits in both plants and herbivores (Bergelson, Dwyer, and Emerson 2001). At the molecular level, such an arms race is expected to lead to the sequential sweep of new defense and counterdefense alleles in the plant and herbivore populations, respectively (Kniskern and Rausher 2001; DeMeux and Mitchell-Olds 2003). A coevolutionary interaction between plants and herbivores characterized by an arms race model therefore predicts that defense genes should exhibit low levels of standing nucleotide sequence variation and high levels of amino acid differentiation (compared to closely related species), induced by strong directional selection for rapid turnover of alleles (Bergelson et al. 2001; DeMeux and Mitchell-Olds 2003).

Alternatively, natural selection may lead to increased diversity in defense genes, either because frequency-dependent selection favors rare alleles or because variability itself is favorable if different alleles confer resistance to different herbivore genotypes (DeMeux and Mitchell-Olds 2003). Such a model for plant-herbivore coevolution, sometimes termed “the trench warfare model” (Stahl et al. 1999) or the “recycling polymorphism” model (Holub 2001), predicts qualitatively different evolutionary dynamics compared to the arms race model. Because different alleles are maintained over long periods in a polymorphic state, balancing selection is expected to result in enhanced levels of sequence diversity because different allelic lineages accumulate mutations more or less independently. Therefore, a trench warfare model of coevolution predicts that defense genes should show enhanced levels of segregating amino acid polymorphisms within populations that sometimes even exceed levels of silent variation (DeMeux and Mitchell-Olds 2003). Such natural selection, favoring allelic variability, has been implicated in the maintenance of hypervariability in both human major histocompatibility complex genes (Ohta 1998) and pathogen recognition genes in many plants (Bergelson et al. 2001).

A growing number of studies are focusing on the molecular population genetic aspects of plant-herbivore and plant-pathogen interactions, and these studies have shown support for both the arms race and trench warfare mode of coevolution (reviewed in DeMeux and Mitchell-Olds 2003). One conclusion emerging from these studies is that specialist defenses, that is, defenses against a small number of natural enemies, are more likely to show deviations from neutral expectations (Tiffin, Hacker, and Gaut 2004). In particular, specialist defenses appear to be more likely to be under the influence of balancing selection (Tiffin, Hacker, and Gaut 2004). In contrast, many generalist defenses, that is, defenses effective against a broad spectrum of enemies, show patterns of intraspecific sequence diversity that either are consistent with a neutral model (Kawabe and Miyashitya 1999; Tiffin and Gaut 2001; Clauss and Mitchell-Olds 2004; Tiffin, Hacker, and Gaut 2004) or show signs of recent positive selection (i.e.,
Plants defend themselves against the attack of natural enemies by using an array of constitutively expressed defenses, such as thorns, spines, and many toxic secondary metabolites (Karban and Baldwin 1997). Plants also actively defend themselves using inducible defense mechanisms (Karban and Baldwin 1997). The active part of a plant’s defense arsenal involves, for instance, oxidative enzymes, genes involved in strengthening the cell wall, and/or a wide array of protease inhibitors (PIs) (Constabel 1999). PIs, in particular, are a class of wound-inducible defense genes for which function in herbivore defense is well established and where the physiological mechanisms underlying the defense function are well studied (Karban and Baldwin 1997; Koiwa, Bressan, and Hasegawa 1997; Constabel 1999; Haq, Atif, and Khan 2004). Plant PIs are proteins that function as specific substrates for proteolytic enzymes in the digestive tracts of herbivores (Ryan 1990; Constabel 1999; Haq, Atif, and Khan 2004). However, a PI is not cleaved by a protease, like normal substrates are, but rather forms a stable complex that limits or completely inhibits the proteolytic activity of the protease (Ryan 1990; Constabel 1999; Haq, Atif, and Khan 2004). This results in much reduced proteolysis in the digestive tract, a lack of available amino acids, and ultimately lowered growth rates or starvation of the herbivore (Ryan 1990; Constabel 1999). In addition to their antinutritive effects, PIs may also have toxic or directly lethal effects on herbivores (Duffey and Stout 1996). There are several classes of PIs in plants, each effective against a different class of proteases. The two most common classes of PIs are inhibitors of serine and cysteine proteases, and this is thought to reflect the predominance of these proteases in herbivore digestive systems (Koiwa, Bressan, and Hasegawa 1997; Constabel 1999; Haq, Atif, and Khan 2004). Serine proteases are the predominant digestive enzyme in Lepidoptera, whereas both serine and cysteine PIs are commonly found in Coleoptera (Koiwa, Bressan, and Hasegawa 1997; Haq, Atif, and Khan 2004).

Recent studies of transcriptional patterns following wounding by herbivores have shown that a large number of defense-related genes are upregulated following tissue damage in Populus (Christopher et al. 2004). In particular, a wide range of PIs are among the most upregulated genes following damage by herbivores (Haruta et al. 2001; Christopher et al. 2004). Haruta et al. (2001) identified three wound-induced genes from Populus tremuloides, belonging to the Kunitz trypsin inhibitor class of serine PIs. These genes, TI1, TI2, and TI3, belong to a small gene family and show amino acid sequence similarities ranging from 52% to 83% (Haruta et al. 2001). In particular, TI1 and TI2 appear to be the result of a recent gene duplication as these genes are over 90% identical at the nucleotide level. Christopher et al. (2004) identified two additional Kunitz type PIs from a Populus trichocarpa × Populus deltoides hybrid. These two genes, named TI4 and TI5, represent more diverged members of the Kunitz TI class, sharing less than 30% amino acid similarity with TI1, TI2, and TI3. The transcription of all five TI genes is highly upregulated following mechanical wounding or herbivory, and all five genes can also be induced by external application of the defense signal jasmonate (Haruta et al. 2001; Christopher et al. 2004). However, the patterns of induction of the five genes are quite different and distinct, both locally and systemically throughout the plant (Haruta et al. 2001; Christopher et al. 2004). In addition to the TI genes, several other PIs show strong induction following wounding, including highly upregulated cysteine PI, CI1 (Christopher et al. 2004).

In this paper, I use molecular population genetic analyses to study the evolution of wound-inducible PIs in the European aspen (Populus tremula). I do so by analyzing in more detail previously published surveys of natural variation in TI3 and CI1 (Ingvarsson 2005) and by adding data on patterns of nucleotide diversity in four additional PI genes (TI1, TI2, TI4, and TI5) from the same set of individuals. These six genes are surveyed for departures from neutral expectations, with a special emphasis placed on evaluating the likelihood of either recurrent selective sweeps or balancing selection maintaining amino acid polymorphisms.

Materials and Methods

Sequence Collection

The five serine PIs studied (TI1 to TI5) are all members of the Kunitz TI class, although TI1, TI2, and TI3 appear to form a small gene family of more closely related sequences because they share between 65% and 85% amino acid identity (Haruta et al. 2001). TI4 and TI5 represent more diverged members of the Kunitz TI class, sharing 16%–25% amino acid identity with the TI1–TI3 group of sequences (Christopher et al. 2004). Sequences of TI3 and CI1 were obtained from Ingvarsson (2005). New sequences of TI1, TI2, TI4, and TI5 were collected from the same set of P. tremula trees that were included in Ingvarsson (2005). Leaves were collected from four different sites throughout Europe: Besacön in eastern France, Klagenfurt in southern Austria, Färlöv in southeastern Sweden, and Umeå in northern Sweden. TI1 and TI2 were also sequenced from an additional population located close to Halmstad, in southwestern Sweden, about 300 km west of the Färlöv population.

Primers to amplify TI1, TI2, TI4, and TI5 from P. tremula were designed from the publicly available GenBank records for these genes (accession numbers AF349441, AF349442, AY378089, and AY378090, respectively). The entire coding sequence was obtained for all genes, including some 5′- and 3′-flanking sequences for TI1, TI2, and CI1. All genes are intronless, except CI which contains two exons, separated by a short intron. Finally, homologous sequences from P. trichocarpa were obtained for all genes by Blast searches of the assembled genome sequence of P. trichocarpa available at http://genome.jgi-psf.org/.

Polymerase chain reaction products were cloned into the pCR2.1 vector using a TA-cloning kit from Invitrogen (Carlsbad, Calif.). At least five, and often more, different clones of each fragment were sequenced using BigDye chemistry (Applied Biosystems Inc., Foster City, Calif.)
Table 1
Intraspecific Polymorphism in *Populus tremula* Wound-induced Protease Inhibitor Genes

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
<th>Sites</th>
<th>Coding</th>
<th>S</th>
<th>$\theta_W$</th>
<th>$\pi_{nt}$</th>
<th>$\pi_{syn}$</th>
<th>$\pi_{rep}$</th>
<th>$C_{min}$</th>
<th>$C_{rep}$</th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T11$</td>
<td>47</td>
<td>699</td>
<td>663</td>
<td>46</td>
<td>0.01596</td>
<td>0.01913</td>
<td>0.0140</td>
<td>0.0168</td>
<td>0.0003</td>
<td>0.0050</td>
<td>0.059</td>
</tr>
<tr>
<td>$T12$</td>
<td>33</td>
<td>717</td>
<td>684</td>
<td>53</td>
<td>0.01909</td>
<td>0.01855</td>
<td>0.0211</td>
<td>0.0167</td>
<td>0.0022</td>
<td>0.0232</td>
<td>0.171</td>
</tr>
<tr>
<td>$T13$</td>
<td>48</td>
<td>639</td>
<td>606</td>
<td>49</td>
<td>0.01728</td>
<td>0.01436</td>
<td>0.0276</td>
<td>0.0112</td>
<td>0.0030</td>
<td>0.0328</td>
<td>0.040</td>
</tr>
<tr>
<td>$T14$</td>
<td>44</td>
<td>648</td>
<td>645</td>
<td>31</td>
<td>0.01115</td>
<td>0.0106</td>
<td>0.0128</td>
<td>0.0101</td>
<td>0.0011</td>
<td>0.0238</td>
<td>0.108</td>
</tr>
<tr>
<td>$T15$</td>
<td>44</td>
<td>645</td>
<td>642</td>
<td>41</td>
<td>0.01575</td>
<td>0.01393</td>
<td>0.0349</td>
<td>0.0069</td>
<td>0.0026</td>
<td>0.0304</td>
<td>0.214</td>
</tr>
<tr>
<td>$CI1$</td>
<td>46</td>
<td>802</td>
<td>429</td>
<td>62</td>
<td>0.01841</td>
<td>0.01014</td>
<td>0.0057</td>
<td>0.0082</td>
<td>0.0018</td>
<td>0.0226</td>
<td>0.052</td>
</tr>
</tbody>
</table>

Average of PI genes: 0.0164
Average of seven other gene regions: 0.0127

Note.—n, number of sampled haplotypes; sites, length of gene regions surveyed; S, number of segregating sites; $\theta_W$, per site estimate of Watterson’s $\theta = 4N_{e}mu$ assuming no recombination); $\pi_{nt}$ (±SD); $\pi_{syn}$, $\pi_{rep}$, average pairwise diversity per site for total, synonymous, and replacement sites, respectively; $C_{min}$, per site recombination rate consistent with the number of inferred recombination events; $C_{rep}$, estimate of per site recombination rate; 4Nr (Hey and Wakeley 1997). Data from gene regions not involved in herbivore defense are from Ingvarsson (2005) and unpublished data.

Results

Structure of Sequeced Regions

Between 33 and 47 alleles were sequenced from the six PI genes. The length of the sequenced regions ranged from 630 bp for $T13$ to 802 bp for $CI1$. The coding regions of the five $T$ genes range in size from 606 to 684 bp (table 1). The predicted protein sequence of all five $T$ proteins contains a signal peptide (Nielsen et al. 1997), consisting of 25 amino acids, that is cleaved off to form the active PI. The mature $T$ protein consists of 10–12 antiparallel $\beta$ strands that form the $\beta$-trefoil fold characteristic of Kunitz type PIs (Song and Suh 1998) and carries a reactive region situated on an external loop. The $CI1$ gene contains a coding region that is 429 bp long and encodes a protein showing strong similarity to other phytochymoyts. After cleavage of a 27-amino acid signal peptide (Nielsen et al. 1997), the mature protein forms two hairpin loops and a largely unstructured N-terminal protein end (Brown and Dziegielewksa 1997). The first hairpin loop carries the conserved QXXSG motif, thought to be the site active in protein inhibition, and the second hairpin loop carries the conserved PW sequence motif present in all cystatins (Brown and Dziegielewksa 1997).

Intraspecific Patterns of Nucleotide Diversity

There is abundant variation in all genes, with the number of segregating sites ranging from 31 to 62 (table 1, Appendix). Estimates of Watterson’s $\theta = 4N_{e}mu$ range from 0.0112 to 0.0191, which is within the range of earlier estimates of $\theta$ obtained from other *P. tremula* genes (Ingvarsson 2005 and unpublished data). In addition to base pair substitutions, several indels were polymorphic in the sample. Two different indels, each resulting in the deletion of a complete codon, were segregating in $T11$. $T12$ contained four indels that were polymorphic in the sample, three indels resulting in the deletion of a complete codon each and a larger indel causing the deletion of four contiguous codons. $T14$ contained two polymorphic indels, one resulting in the deletion of a complete codon and one causing the
deletion of two contiguous codons. TI5 was also polymorphic for a two-codon deletion. Note that none of these indels disrupted the open reading frame, and all alleles at the five TI genes were predicted to be fully functional based on their amino acid sequence. CI1 only contained indels in the intron, and these were all associated with mononucleotide repeats that differed in length.

If the PI genes were under recurrent directional selection due to continual fixation of new beneficial alleles, silent site diversity is expected to be sharply reduced. There does not, however, appear to be any evidence for a reduction in nucleotide diversity at synonymous sites in the PI genes studied (table 1), compared with other gene regions that are not involved in herbivore defense in *P. tremula* (Ingvarsson 2005, unpublished data). In fact, synonymous site diversity is higher in the six PI genes, (Ingvarsson 2005, unpublished data). In fact, synonymous site diversity is expected to be sharply reduced. There does not, however, appear to be any evidence for a reduction in nucleotide diversity at synonymous sites in the PI genes compared with other gene regions that are not involved in herbivore defense in *P. tremula* (Ingvarsson 2005, unpublished data).

### Table 2

<table>
<thead>
<tr>
<th>Locus</th>
<th>Tajima’s D</th>
<th>P</th>
<th>Cmin</th>
<th>P</th>
<th>Chw</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI1</td>
<td>-0.691</td>
<td>0.781</td>
<td>0.877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TI2</td>
<td>-0.104</td>
<td>0.273</td>
<td>0.269</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TI3</td>
<td>-0.586</td>
<td>0.098</td>
<td>0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TI4</td>
<td>-0.170</td>
<td>0.321</td>
<td>0.207</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TI5</td>
<td>-0.192</td>
<td>0.171</td>
<td>0.086</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CI1</td>
<td>-1.523</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H_{Cmin}</th>
<th>P</th>
<th>Cmin</th>
<th>P</th>
<th>Chw</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI1</td>
<td>12.200</td>
<td>0.051</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>TI2</td>
<td>-6.560</td>
<td>0.085</td>
<td>0.037</td>
<td></td>
</tr>
<tr>
<td>TI3</td>
<td>-0.642</td>
<td>0.377</td>
<td>0.410</td>
<td></td>
</tr>
<tr>
<td>TI4</td>
<td>-3.841</td>
<td>0.114</td>
<td>0.067</td>
<td></td>
</tr>
<tr>
<td>TI5</td>
<td>-0.347</td>
<td>0.450</td>
<td>0.482</td>
<td></td>
</tr>
<tr>
<td>CI1</td>
<td>-2.659</td>
<td>0.234</td>
<td>0.224</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Z_{rep}</th>
<th>P</th>
<th>Cmin</th>
<th>P</th>
<th>Chw</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI1</td>
<td>0.249</td>
<td>0.080</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>TI2</td>
<td>0.135</td>
<td>0.150</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>TI3</td>
<td>0.089</td>
<td>0.210</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>TI4</td>
<td>0.169</td>
<td>0.120</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>TI5</td>
<td>0.084</td>
<td>0.840</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>CI1</td>
<td>0.038</td>
<td>0.540</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

|Note.—Statistical significance of tests were evaluated using coalescent simulation with two different recombination estimates (C_{min} and C_{hw}) and population structure (see text for further details).|

There is evidence for an excess of linkage disequilibrium in all five Kunitz trypsin inhibitor genes if the significance of $Z_{\text{rep}}$ (Kelly 1997) is evaluated using levels of recombination estimated from the data set ($C_{\text{rep}}$), although if the lower bound on the recombination rate ($C_{\text{min}}$) is used, none of the genes have $Z_{\text{rep}}$ that are significant (table 2). However, this approach provides an extremely conservative test of excessive linkage disequilibrium because $C_{\text{min}}$ estimates of the scaled recombination rate are about an order of magnitude lower than those obtained through the method of Hey and Wakeley (1997) (table 2). Enhanced levels of linkage disequilibrium can be generated by directional selection if a favorable mutation has not yet reached fixation or if recombination occurred during a selective sweep, so that variation at linked sites was not completely eradicated during the sweep (Fay and Wu 2000; Przeworski 2002; Kim and Nielsen 2004). An excess of linkage disequilibrium is also expected if natural selection maintains distinct allelic classes at the locus, as is the case with balanced polymorphisms (Charlesworth, Nordborg, and Charlesworth 1997). Finally, demographic scenarios, such as population structure or population growth, can also result in increased linkage disequilibrium compared to neutral expectations (Przeworski 2002). However, effects of demography are expected to be genome wide, and there is no indication of excessive linkage disequilibrium in other *P. tremula* genes (Ingvarsson 2005), suggesting that the presence of enhanced linkage disequilibrium in TI defense genes is the product of natural selection.

Three of the defense genes, TI1, TI2, and TI4, also show an excess of high-frequency–derived variants, as measured by Fay and Wu’s $H$, although this excess is only marginally significant in TI4 ($P < 0.067$, table 2). Under the much more conservative assumption of the low recombination ($C_{\text{min}}$ in table 2), $H$ remains significant only for TI1 (table 2). An excess of high-frequency–derived variants is also a pattern that is likely to be explained by strong directional selection fixing newly arisen mutations because derived mutations occurring at high frequencies are rare under a neutral model (Fay and Wu 2000; Kim and Nielsen 2004). Again, demographic factors, and in particular population structure, may provide an alternative explanation for an excess of high-frequency–derived variants (Przeworski 2002). However, population subdivision is not likely to be a factor in explaining the negative values of $H$ for TI1, TI2, and TI4 for at least two reasons. First of all, the effect of population structure on $H$ becomes more severe with increasing strength of subdivision (Przeworski 2002), and population subdivision is generally low in *P. tremula* (table 1 and Ingvarsson 2005). Second, population structure was explicitly taken into account when assessing the significance of $H$ through coalescent simulations. It is possible that the coalescent simulations did not accurately characterize the population structure of *P. tremula*. However, given the long-distance dispersal of both seeds and pollen seen in *P. tremula*, the island model used in the coalescent simulations should provide a reasonable approximation for the population structure of *P. tremula*, and including data is available in *P. tremula* (table 1, Ingvarsson 2005, unpublished data, range for nondefense genes $2.96 	imes 10^{-3} - 3.44 	imes 10^{-3}$).
a model of population structure in the coalescent simulations, even if approximate, is likely better than completely ignoring it.

Polymorphism and Divergence

Divergence at silent sites in the six defense genes is similar or slightly higher than at other \textit{P. tremula} loci not involved in herbivore defense (fig. 1), whereas nonsynonymous divergence is generally higher at the defense genes than in nondefense genes. In two of the defense genes, \textit{TII} and \textit{TII2}, there are signs of long-term adaptive evolution, as they show greater rates of divergence at nonsynonymous than synonymous sites, as evidenced by $K_s/K_a > 1$ (fig. 1). Using the modified pairwise Nei-Gojobori method with $p$ distances (Nei and Kumar 2000), as implemented in MEGA 2.1 (http://www.megasoftware.net), the $K_s/K_a$ ratios at both \textit{TII} and \textit{TII2} are significantly greater than one ($P < 0.037$ and $P < 0.006$, respectively). A $K_s/K_a$ ratio exceeding unity is a clear sign of adaptive evolution, suggesting that nonsynonymous mutations offer a fitness advantage and are fixed at a higher rate than synonymous mutations (Yang 2001).

Surprisingly, the McDonald and Kreitman (1991) (MK) test, which compare rates of divergence with levels of standing polymorphism within species, show no evidence for adaptive evolution in any of the \textit{TI} genes. However, if polymorphism and divergence data from the five \textit{TI} genes are pooled, these genes appear to have experienced elevated rates of fixation of nonsynonymous mutations (the neutrality index of Rand and Kann [1996], $NI = 0.584$, $P < 0.016$, table 3). Excluding data from \textit{TI4}, which harbors a nonsignificant excess of intraspecific polymorphism at nonsynonymous sites (table 2), this pattern of an excess of fixed nonsynonymous mutations becomes even stronger ($NI = 0.476$, $P < 0.0019$). If the same pooled comparison is made with seven protein-coding genes that are not involved in herbivore defense (Ingvarsson 2005, unpublished data), no deviation from neutral expectations is detected ($NI = 1.433$, $P < 0.155$, table 2).

Contrary to the observations for the five \textit{TI} genes, an MK test indicates a marginally significant ($P < 0.075$) excess of intraspecific polymorphism at replacement sites in \textit{CI1} (table 3). Much of this excess can be explained by the large number of singleton mutations causing amino acid replacements in \textit{CI1}. There are 13 nonsynonymous mutations found only in single alleles and an additional 7 are found in two to four alleles (Appendix). This is also evident from the Tajima’s $D$, which is significantly negative for \textit{CI1} (table 2).

Haplotype Structure at \textit{TII}

\textit{TII} has several features suggesting that it has been influenced by natural selection, including an excess of derived mutations occurring at high frequencies, extensive linkage disequilibrium, and $\pi_d/\pi_s$ and $K_s/K_a$ ratios exceeding one. The genealogy at \textit{TII} shows strong haplotype structure, with alleles falling into two distinct haplotypes (haplotypes A and B in fig. 2, also see Appendix) that differ by 14 fixed substitutions, 13 of which occur at replacement sites. The strong haplotype structure also coincides with both an excess of linkage disequilibrium (LD) and high proportions of high-frequency–derived sites (table 2).

One of the haplotypes (A in fig. 2) shows very little intrahaplotype variation ($S = 5$, $\pi = 0.00095$, Appendix), despite accounting for 25 out of 47 sampled alleles. The remaining sequences, in haplotype B, are more heterogeneous and have a level of polymorphism ($\pi = 0.0117$) that is comparable to other loci in \textit{P. tremula} and about an order of magnitude greater than haplotype A. To test the likelihood of observing two clades with such different levels of intraallelic variation, I used a modified version of the test devised by Hudson et al. (1994). This test evaluates the likelihood, under an equilibrium-neutral model, of observing a haplotype with only five segregating sites making up 25 out of 47 sampled alleles, conditional on the variation observed in the total sample ($S = 46$). I modified the original approach of Hudson et al. (1994) to explicitly take

---

### Table 3

<table>
<thead>
<tr>
<th>Locus</th>
<th>Synonymous Fixed</th>
<th>Synonymous Poly</th>
<th>Replacement Fixed</th>
<th>Replacement Poly</th>
<th>NI</th>
<th>MK ($P$ value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{TII}</td>
<td>17</td>
<td>10</td>
<td>53</td>
<td>30</td>
<td>0.962</td>
<td>0.02 (0.887)</td>
</tr>
<tr>
<td>\textit{TII2}</td>
<td>11</td>
<td>15</td>
<td>42</td>
<td>28</td>
<td>0.489</td>
<td>1.74 (0.187)</td>
</tr>
<tr>
<td>\textit{TII3}</td>
<td>17</td>
<td>23</td>
<td>17</td>
<td>27</td>
<td>0.483</td>
<td>1.27 (0.259)</td>
</tr>
<tr>
<td>\textit{TI4}</td>
<td>13</td>
<td>8</td>
<td>23</td>
<td>24</td>
<td>1.696</td>
<td>0.53 (0.467)</td>
</tr>
<tr>
<td>\textit{TI5}</td>
<td>6</td>
<td>26</td>
<td>7</td>
<td>17</td>
<td>0.560</td>
<td>0.35 (0.552)</td>
</tr>
<tr>
<td>Combined \textit{TI}</td>
<td>54</td>
<td>82</td>
<td>142</td>
<td>126</td>
<td>0.584</td>
<td>5.85 (0.016)</td>
</tr>
<tr>
<td>\textit{CI1}</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>25</td>
<td>3.889</td>
<td>3.97 (0.075)</td>
</tr>
<tr>
<td>Combined nondefense genes</td>
<td>52</td>
<td>133</td>
<td>45</td>
<td>165</td>
<td>1.433</td>
<td>2.02 (0.155)</td>
</tr>
</tbody>
</table>

**Note.**—Divergence calculated from \textit{P. trichocarpa}. NI, neutrality index (Rand and Kann 1996); MK, G-test (McDonald and Kreitman 1991).
have been maintained for long periods of time, through selection. A more likely scenario is that the haplotypes A and B have dispersed between haplotypes A and B, it is unlikely that haplotype A represents a novel haplotype that evolved recently and increased in frequency because of positive selection. A more likely scenario is that the haplotypes A and B have been maintained for long periods of time, through some form of balancing selection. However, patterns of sequence diversity within the two allelic haplotypes are not consistent with a model of balancing selection where the relative frequency of the two haplotypes has been stable over time. Innan and Tajima (1999) showed that if balancing selection maintains two distinct allelic classes at a locus at constant frequencies, the sum of pairwise differences within the two classes is roughly constant and equal to 0, regardless of the strength and pattern of selection. For TI1, the sum of pairwise diversities within haplotypes A and B equals \( \pi = 0.01263 \), which is significantly less \( (P < 0.05) \) than the pairwise diversity observed for the total sample \( (P = 0.0191) \). A model of constant balancing selection thus appears to provide a poor fit to the TI1 data as well.

Discussion

Sequence Diversity and Evidence for Nonneutral Evolution

In this study I have examined patterns of nucleotide variation in six PIs, five serine PI and one cysteine PI, from European aspen (\( P. \) tremula). Although there are differences among the six genes, several signs point to the fact that the evolution of these defense genes is not well characterized by a neutral model. There is no apparent reduction in levels of polymorphism, either at silent or replacement sites in any of the six genes, compared to \( P. \) tremula genes that are not involved in herbivore defense (table 1). In particular, the TI1 and TI2 genes stand out by having \( K_s/K_a \) ratios that exceed one (fig. 1). A \( K_s/K_a \) ratio exceeding unity suggest that nonsynonymous mutations are fixed at a higher rate than synonymous mutations and is a clear sign of adaptive evolution (Yang 2001). Both TI1 and TI2 also have an excess of high-frequency–derived sites (table 2) and excess LD, suggesting that the recent evolutionary histories at these two genes are characterized by one or more selective sweeps that have driven newly arisen mutation to high frequency in the population.

Of the other four genes surveyed, TI3, TI4, and TI5 all show an excess of LD, at least when the significance of the \( Z_{ax} \) test is evaluated using the recombination rates inferred from the sequence data (table 2). Such excess LD could result from either directional or balancing selection, although demographic factors such as population structure could also explain these observations. There is no evidence for excess LD in \( P. \) tremula genes not involved in herbivore defense (Ingvarsson 2005). The observed patterns could thus be a product of natural selection, although patterns of intra-specific sequence diversities at TI3, TI4, and TI5 show little other evidence of having been influenced by natural selection.

There are, however, indications that the TI genes collectively are evolving in a nonneutral manner because they have an elevated rate of fixation of nonsynonymous mutations (table 3). An evolutionary history of increased amino acid substitution but little current evidence for natural selection could be indicative of episodic selection, where favorable mutations appear and rapidly sweep through the population but are then followed by long periods of selective neutrality (Tiffin and Gaut 2001).
CII has an excess of amino acid polymorphisms, much of which can be explained by a large number of singletons. While such an excess of low-frequency amino acid polymorphisms could be an indication of diversifying selection, an excess is also expected under a slightly deleterious model of molecular evolution, where selection is weak enough to allow polymorphism at nonsynonymous sites but where fixation is limited (Ohta 1992).

The different PI genes thus appear to have been influenced by natural selection to different degrees. TII and TI2 show several signs of both historical and current action of natural selection, whereas the remaining TI genes and CII show sporadic signs of deviations from neutral expectations. In Arabidopsis thaliana, Clauss and Mitchell-Olds (2004) showed that patterns of sequence diversity varied 10-fold across six duplicated trypsin inhibitor genes, located in tandem across a 10-kb segment on chromosome II. Interestingly, they also showed that sequence diversity in the coding regions was positively correlated to variation in expression levels of the different TI genes. They interpreted this among-locus variation in both sequence polymorphism and gene expression to be the result of selection for coping with a diverse set of both specialist and generalist herbivores through subfunctionalization of different loci (Clauss and Mitchell-Olds 2004).

Molecular population genetic data from two PIs in maize (Zea mays), wip1 and mpi, did not show any deviations from neutrality (Tiffin and Gaut 2001; Tiffin, Hacker, and Gaut 2004). Relative rate tests, however, suggested significant rate heterogeneity at wip1 with an increased rate evolution in the Zea lineage; these changes were preferentially located in one of the inhibitory loops of wip1 (Tiffin and Gaut 2001). This suggests, even when molecular population genetic analysis in contemporary populations fails to provide evidence for nonneutral evolution, that defense genes can be characterized by an evolutionary history influenced by natural selection.

Tiffin, Hacker, and Gaut (2004) argue that specialist defenses, that is, defenses active against one or a few natural enemies, are more likely to show evidence for nonneutral evolution compared to generalist defenses. Although PIs as a group are expected to target many different natural enemies, it is not clear whether individual PIs should be viewed as specialist or generalist defenses. The large number of TI genes found in P. tremula, each with their own distinct patterns of gene expression following wounding (Christopher et al. 2004), suggests that these genes might have different targets. It is possible that some of the TI genes have more generalized targets whereas others have more specialized functions, but at present there is no data to shed any light on these ideas. The same view was advocated by Clauss and Mitchell-Olds (2004), who argued that the differences in both sequence diversity and expression pattern seen among the six TI genes in their study are most likely the result of selection for subfunctionalization of the different genes. More studies are clearly needed, both of the patterns of expression following attack by different herbivores and also of the inhibitory functions at the biochemical level of the various PI genes against different herbivore digestive enzymes.

Haplotype Structure at TII

The TII gene has both elevated levels of intraspecific amino acid polymorphism and a $K_a/K_s$ ratio that exceeds one (table 1 and fig. 1). Moreover, the genealogy at TII shows a haplotype structure with two divergent haplotypes that have different levels of intrahaplotype variation. The presence of two diverged haplotypes, with one haplotype characterized by a very low level of standing variation and another that carries more normal levels of sequence variation, has been interpreted as a hallmark signature of selective sweeps in a highly recombining genomic region (Parsch, Meiklejohn, and Hartl 2001; Quesada et al. 2003; Kim and Nielsen 2004). However, the two haplotypes at TII differ by 14 fixed mutations, 13 of which are nonsynonymous, although there is no difference between the two haplotypes in the number of derived mutations that they carry. The presence of such diverged haplotypes is therefore probably a result of them having been maintained in the population over evolutionary time, even though the patterns of polymorphism observed in haplotypes A and B do not fit a model of long-term, stable balancing selection (Innan and Tajima 1999). A possible alternative explanation is a recent admixture of two previously isolated populations. This seems unlikely, however, because an admixture should leave a genome-wide signature. First, the population subdivision at TII is low (table 1), and both haplotypes are present in all five populations sampled across Europe, reducing the likelihood of recent admixture. Second, although the excess of LD (table 2) could be interpreted as signs of enhanced haplotype structure also at other TI genes, none of the nondefense genes show such patterns (Ingvarsson 2005). Rather, the enhanced LD could be the result of selection maintaining similar although weaker forms of the haplotype structure seen at TII also in other PI genes (Appendix).

Patterns of polymorphism and haplotype structure therefore suggest that neither the arms race nor trench warfare model alone is adequate for describing the evolutionary history of TII. Rather, the data are suggestive of a model that incorporates both directional selection (arms race) and a selective maintenance of intraspecific amino acid polymorphism (trench warfare). One possible explanation that fits both of these criteria is that different resistance haplotypes are maintained through balancing or negative-frequency–dependent selection, but where frequencies of different resistance haplotypes vary over time (Seger 1992).

Tiffin, Hacker, and Gaut (2004) also showed that neither the arms race nor the trench warfare model was adequate for explaining patterns of polymorphism at the disease-resistance gene hm2 in Z. mays. Tiffin, Hacker, and Gaut (2004) found several diverged haplotypes with very little intrahaplotype variation and proposed a scenario where these haplotypes were maintained by natural selection, but the frequencies of these haplotypes varied over evolutionary time. In A. thaliana, Rose et al. (2004) also found two major allelic haplotypes at the disease-resistance gene RPP13. They interpreted their results as the consequence of negative-frequency–dependent selection, where single RPP13 alleles have not swept to fixation but have
rather been maintained in the population because different alleles encode resistance specificities against different strains of the pathogen (Rose et al. 2004). The RPP13 data is clearly inconsistent with the arms race model and more in line with the predictions made by the trench warfare model. Finally, data from antibacterial peptides in Drosophila melanogaster also do not show clear support for either the arms race or the trench warfare model (Lazzaro and Clark 2003), although these genes show some signs of natural selection, such as relatively high levels of intra-specific amino acid polymorphisms and excesses of derived sites occurring at high frequencies (Lazzaro and Clark 2003).

The scenario proposed above to explain the observed patterns of nucleotide diversity at TII is clearly ad hoc. However, current models that have been proposed to explain coevolutionary interactions of plant-herbivore or plant-pathogen interactions at the molecular level, including the trench warfare and arms race models, are rather heuristic, and it is not clear what type of data are needed to either verify or reject these models (see also Tiffin, Hacker, and Gaut [2004] for a similar argument). The fact that almost all studies that have examined coevolutionary interactions between plant and pathogens or herbivores at the level of individual genes have found patterns of polymorphism that cannot be explained by either the arms race or trench warfare models suggests that more research is needed to specifically examine patterns of nucleotide polymorphism under a wide range of conditions both in these rather simple and in more complex models of plant-parasite coevolution (Seger 1992).

Acknowledgments

I would like to thank Carin Olofsson for obtaining the DNA sequences presented in this paper. John McDonald and three anonymous reviewers provided comments that improved the paper. I am also grateful to Barbara Giles who provided help with linguistic corrections. This research has been funded by a grant from the Swedish Research Council (Vetenskapsrådet, VR).

Appendix

Alignment of polymorphic sites in the six protease inhibitor genes. Residue numbers are indicated above each polymorphic site, and each polymorphic site is indicated as either nonsynonymous (N), synonymous (S), indels (D) or occurring in flanking regions or introns (I). All residues identical to the top sequence are denoted by dots (.) and gaps are given by a dash (–).
Evolution of Defense Genes in *Populus tremula* 1811

**Literature Cited**


John H. McDonald, Associate Editor

Accepted May 17, 2005