New Phytologist

Selective histories of poplar protease inhibitors: elevated polymorphism, purifying selection, and positive selection driving divergence of recent duplicates

Maurine Neiman[1], Matthew S. Olson[2] and Peter Tiffin[1]

[1]Department of Plant Biology, 250 Biosciences, University of Minnesota, Saint Paul, MN 55105, USA; [2]Institute of Arctic Biology, 311 Irving 1, University of Alaska Fairbanks, Fairbanks, AK 99775, USA

Author for correspondence:
Peter Tiffin
Tel: +1 612 624 7406
Email: ptiffin@umn.edu

Received: *12 February 2009*
Accepted: *21 May 2009*

New Phytologist (2009) **183**: 740–750
doi: 10.1111/j.1469-8137.2009.02936.x

Key words: coevolution, herbivores, Kunitz trypsin inhibitor, plant defense, Poplar, population genetics, protease inhibitors.

Summary

- To further our understanding of plant defense evolution and the consistency of selection at the nucleotide level we analysed polymorphism data from five protease inhibitor (PI) genes in *Populus balsamifera*.
- We compared diversity at the five PI genes to diversity at nondefense loci in both range-wide samples as well as in two subpopulations, one from the northern edge of the species range and one from the southern edge of the range. We also compared our data with previously reported diversity in *Populus tremula*, a European species with similar ecology to North American *P. balsamifera*.
- The PIs show diverse histories, including repeated bouts of positive selection and excess diversity. These genes also exhibit diverse histories in *P. tremula* but the signatures of selection acting at the specific loci differed between the species. One locus, *KTI3*, segregates several recent duplicates that show evidence of either positive selection or relaxed selective constraints.
- The patterns of diversity at the PIs varied within *P. balsamifera* and between two closely related species. The lack of consistent patterns suggests that evolution of host defense genes, including adaptations to enemy-imposed selection, may often be lineage- and gene-specific.

Introduction

In recent years, analyses of whole genome sequences and nucleotide diversity at specific loci have provided insight into the relative effects of selective versus neutral forces on patterns of genetic diversity. Such analyses have revealed that some of the strongest evidence of both positive and balancing selection is found at genes associated with biotic interactions, including those that encode reproductive proteins (Clark *et al.*, 2006) and proteins that defend hosts against pathogens and parasites (Ford, 2002). Evidence for strong selection on plant defense genes is consistent with observations in contemporary populations in which genotypes that are better defended against herbivores and pathogens often have higher fitness than less well-defended genotypes (reviewed in Rausher, 2001; deMeaux & Mitchell-Olds, 2003). Population genetic analyses complement these studies by examining the evolutionary history of individual loci over longer time-scales and by providing a powerful tool for integrating the effects of selection from the phenotype to the genotype to the individual gene.

Although strong evidence for selection is found at many defense genes, not all putative defense genes show clear evidence of a selective history that differs from that of other genes in the genome (Moeller & Tiffin, 2005, 2008; Bakker *et al.*, 2006). Moreover, when population genetic analyses reveal evidence for selection, the nature of selection appears to be highly variable and includes recent selective sweeps, frequency-dependent or balancing selection, purifying selection and selective neutrality (reviewed in Tiffin & Moeller, 2006). Open questions remain as to the proportion of putative defense genes that harbor signatures of recent selection and whether we can identify general patterns of selection that have acted on defense genes. Alternatively, the evolutionary response to enemy-imposed selection may be gene and lineage-specific as a result of temporal and spatial variation in selection. Plant protease inhibitors (PIs) are an interesting class of defense genes in which to

investigate these questions because they represent a class of defense genes with similar biochemical function for which population genetic analyses have detected the full range of selective histories described for plant defense genes (Tiffin & Gaut, 2001; Clauss & Mitchell-Olds, 2003, 2004; Ingvarsson, 2005; Talyzina & Ingvarsson, 2006; Moeller & Tiffin, 2008).

Protease inhibitors are defense proteins that function by binding to protease enzymes and inhibiting proteolytic activity (reviewed in Ryan, 1990; Haq *et al.*, 2004). Several lines of evidence support a role for PIs in plant defense, including their induction following insect feeding and their ability to inhibit insect gut proteases and slow insect development (reviewed in Lopes *et al.*, 2004). Moreover, many insect herbivores have apparent adaptations that lessen the potentially negative impact of PIs contained in their diets either by expressing proteases that are unaffected by the PIs (Mazumdar-Leighton & Broadway, 2001) or through proteolytic inactivation of PI function (Giri *et al.*, 1998), suggesting a coevolutionary history between host PIs and insect proteases. Nevertheless, some PIs are likely involved in nondefense functions such as regulation of endogenous proteases (Lopes *et al.*, 2004).

The efficacy of PI inhibition depends on the three-dimensional fit between proteases and inhibitors (Bode & Huber, 1992). The implications are that the ability of PIs to defend against herbivore attack will depend upon the specific structure of the protease, and that any PI may show activity against only a limited range of enemy proteases. This specificity sets the stage for gene-for-gene coevolution; a mutation in an enemy protease that weakens inhibitor binding will be selectively favored in the enemy, which in turn will favor alleles in the host that encode inhibitors that bind better to the new protease (Lopes *et al.*, 2004). The importance of inhibitor–protease interaction has been invoked to explain the presence of highly diverse PI genes found in the genomes of many species (reviewed in Christeller, 2005), just as avirulence–virulence protein interactions are viewed as the main force driving the evolutionary divergence and variability of plant resistance genes (reviewed in Meyers *et al.*, 2005).

Our primary objective in this study was to further our understanding of the selection acting on host-defense genes, and PIs in particular, by examining the nucleotide diversity of four Kunitz trypsin inhibitors (*KTIs*) and one cysteine PI gene in balsam poplar, *Populus balsamifera*. These PIs are among the most strongly induced genes in poplar following mechanical damage and herbivore feeding (Bradshaw *et al.*, 1990; Christopher *et al.*, 2004; Major & Constabel, 2006). We characterize the role selection has played in the evolution of these genes by comparing patterns of diversity in PIs to patterns of diversity in 28 nondefense, reference, genes. In order to test for evidence of selection favoring different alleles in different parts of the species range, we sampled and analysed diversity for three samples: a range-wide sample and two geographically defined subpopulations. Finally, we took advantage of an earlier study of PI nucleotide diversity in the European poplar *P. tremula* (Ingvarsson, 2005) to assess whether these genes harbor similar signatures of selection in two geographically distinct but ecologically similar species.

Materials and Methods

Populus balsamifera ssp. *balsamifera* L. is a short-lived woody tree distributed across northern North America from Newfoundland to northwest Alaska (Fig. 1). It is among the most rapidly-growing trees in the boreal forest, and is an early successional colonizer of floodplains and disturbed sites (Bonan, 2002). *Populus balsamifera* spp. *balsamifera* is closely related to *Populus balsamifera* ssp. *trichocarpa* (Torr. & A. Gray) Hulten, for which a genome has been sequenced and partly annotated (Tuskan *et al.*, 2006). Although the subspecific status of the relationship among these taxa is recognized in the forthcoming Flora of North America, forestry practitioners and previous publications usually recognized these taxa as separate species, *P. balsamifera* and *P. trichocarpa*, a precedent we follow here.

We extracted DNA from leaves from 46 *P. balsamifera* trees collected from across the species range. These genotypes represent a portion of the Agriculture and Agri-Food Canada Balsam Poplar (AgCaBaP) collection growing in a common garden in Indian Head, Canada. The individuals comprise three geographically distinct samples: a range-wide sample composed of a single individual from each of 16 locations from throughout the species range (Fig. 1), a southern population represented by 15 individuals sampled from near Edmonton, Alberta, Canada, and a northern population represented by 16 individuals (13 sampled from near Galena, AK, USA, and 3 from near Nome, AK, USA; preliminary analyses revealed no qualitative differences between samples from Galena and Nome). We hereafter refer to these as the range-wide, southern and northern samples, respectively.

We used PCR to amplify each of 5 PI and 12 reference loci from each of the 46 sampled individuals as well an additional 16 reference loci from the range-wide sample only. For these additional reference loci we also sampled a sequence from *P. tremula* (clone 61, population 7 of the SWASP collection; Luquez *et al.*, 2008). The PIs included four Kunitz trypsin inhibitors, *KTI3–5*, *Gwin3*, which is homologous to *P. tremula KTI1* and *KTI2* (there appears to have been a duplication in the lineage leading to *P. tremula*), and one cysteine inhibitor, *CI1* (nomenclature following Bradshaw *et al.*, 1990; Ingvarsson, 2005). The sequenced regions of the PIs ranged from 540–715 bases and included most of the entire coding region of each gene. The 28 reference loci ranged from 380 to 714 bp and contained 0–647 bp of coding sequence. The reference loci were haphazardly selected from around 600 loci that are part of an ongoing study of genome-wide nucleotide variation in *P. balsamifera* (M. S. Olson *et al.*, unpublished). Primer sequences, PCR conditions, and genomic locations for all loci are in provided in the Supporting Information, Table S1.

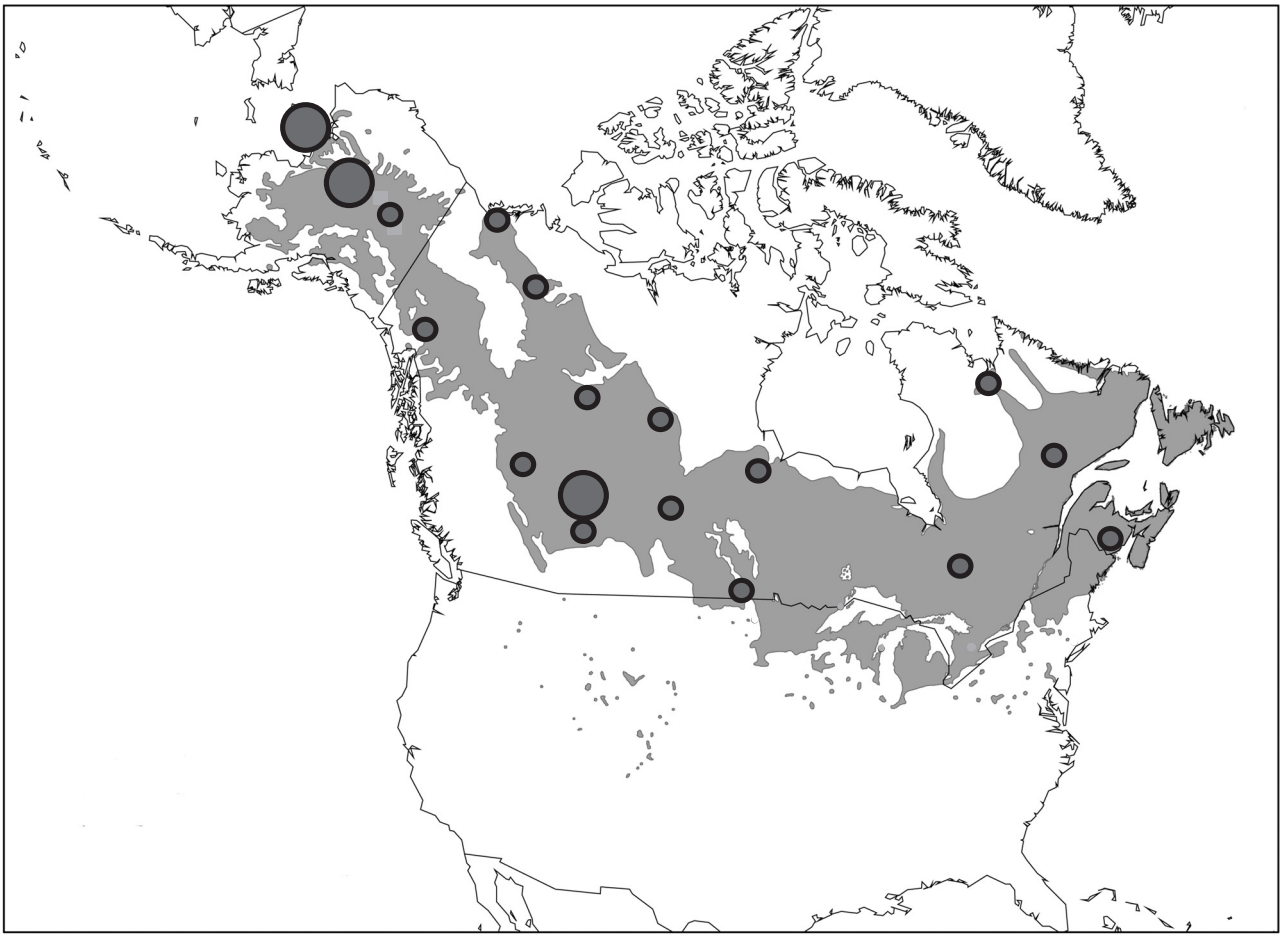

Fig. 1 Present-day distribution of *Populus balsamifera* (tinted region) with the location of samples taken for range-wide (small circles), northern and southern (large circles) samples across northern North America.

All PCR products were sequenced after treating with ExoSAP-IT (USB Corp., Cleveland, OH, USA). One PI (*KTI3*) and six of the reference loci that were sampled from all three subpopulations (*177996*, *178400*, *178844*, *179330*, *548079*, *751366*) segregated indels that prevented us from resolving allelic sequences for some individuals. In these cases, we cloned PCR products into pGem-TA vectors (Promega); for each reference locus we sequenced three clones from one PCR product, whereas for *KTI3* we sequenced 8–10 clones from each of two PCR reactions (see later). Singleton variants found in clones but not verified in the direct sequences or a second cycle of PCR and cloning were not included in the analyses. Sequence data were aligned using BIOEDIT v. 7.0.9 (Hall, 1999). We used PHASE (Stephens *et al.*, 2001; Stephens & Scheet, 2005) to resolve haplotypes for sequences from PCR products (all posterior probabilities > 0.8). All sequences have been deposited in GenBank (accession numbers provided in Table S1).

Data analysis

For each of the three data sets (range-wide, northern and southern) we estimated standard descriptors of nucleotide diversity including pairwise differences per site between sequences π (Nei, 1987), the average number of segregating sites per site, θ_w (Watterson, 1975), haplotype diversity, *Hd* (Nei, 1987), and Tajima's D (Tajima, 1989). Values of θ_w, π and D were calculated for all sites as well as for silent sites only. For the PI loci we also estimated Fay and Wu's H (Fay & Wu, 2000) to test for significant excess of high-frequency derived variants, and used McDonald-Kreitman (MK) tests (McDonald & Kreitman, 1991) to determine whether there were significant differences in the ratio of replacement per synonymous polymorphism vs replacement per synonymous substitution for each locus. Finally, we estimated dN : dS using the modified Nei-Gojobori method as implemented in MEGA4 (Zhang *et al.*, 1998; Tamura *et al.*, 2007). We tested whether H values deviated significantly from expectations under a standard neutral model using 10 000 coalescent simulations conditioned on the number of segregating sites and assuming a stable population size and no recombination. Similar coalescent simulations were used to obtain confidence intervals for θ around each of the reference loci. For Fay and Wu's H and MK tests we used sequence from homologous loci in *P. tremula* as an outgroup. We used *Fst* and *Snn* (Hudson,

2000) to estimate the extent of genetic differentiation between the southern and northern populations. Summary statistics and coalescent simulations for obtaining confidence intervals around θ for individual genes and testing the significance of H-statistics were conducted in DNAsp (Rozas *et al.*, 2003).

We used three approaches to compare the patterns of diversity at PIs with those at reference loci. First, we compared θ at the PI loci to the distributions of θ generated from 1000 coalescent simulations of a single panmictic, exponentially growing population generated using Serial SimCoal (Anderson *et al.*, 2005) a modified version of SIMCOAL (Excoffier *et al.*, 2000); PIs with values that fall in the tails of the distribution of the coalescent-generated values are candidates for selection (Tenaillon *et al.*, 2004). For these simulations we assumed a present-day effective population size of 60 000, a mutation rate of 1×10^{-9} mutations per site per year, growing exponentially at a rate of 0.0015% per generation. For each of the coalescent simulations we sampled 30 sequences to calculate nucleotide diversity (θ, π and D). The coalescent parameters were chosen because they produced a mean Tajima's D and a mean number of segregating sites (S) for a 588 bp region of DNA (mean length of reference loci, mean silent sites = 336) very similar to the mean observed values from the reference loci (mean $S_{observed} = 4.9$, $S_{simulations} = 5.1$; mean $D_{observed} = -0.45$, $D_{simulations} = -0.44$). Because selection may affect diversity and thus bias estimates of demographic history we also conditioned these coalescent simulations using data from silent (synonymous and noncoding) sites only. For these simulations we used similar conditions as above except the sequence length was 336 bases and the population grew exponentially at a rate of 0.00087% per generation. These parameters produced a mean number of silent segregating sites and Tajima's D values calculated only on silent sites that were similar to those estimated from our data (mean $S_{observed} = 3.3$, $S_{simulations} = 3.2$; mean $D_{observed} = -0.23$, $D_{simulations} = -0.26$). While the demographic models used for these simulations should provide reasonable distributions of θ for comparing the PI with the reference loci, we caution that a simple growth model might not accurately capture the demographic history of *P. balsamifera*. Second, we used nonparametric Kruskal–Wallis tests to determine if the frequency distribution of polymorphic sites, as reflected by Tajima's D, differed significantly between PI and reference genes. Finally we used the 16 reference loci for which we had a sequence from *P. tremula* in maximum-likelihood HKA tests as implemented in MLHKA (Wright & Charlesworth, 2004, http://www.yorku.ca/stephenw/programs.html) to test for deviations from the neutral expectation of a positive correlation between intraspecific polymorphism and interspecific divergence. Because of limited numbers of silent sites at our PI loci we conducted these analyses using total diversity.

For the HKA tests, we initially estimated the goodness of fit to the data for two models: a neutral model and a model where selection is operating on PIs but not the reference loci. Under the neutral model, the selection parameter *k* was fixed across all loci, whereas our selection model allowed for a separate *k* value for the PI loci. We ran 1 million chains and fitted each model with three independent runs initiated with a different random number seed. For each model we used the mean likelihood scores from the three runs for likelihood ratio testing.

KTI4, *Gwin3* and *KTI3*

Preliminary analyses of the PCR-generated sequence data from *KTI4* and *Gwin3* revealed unexpected patterns of diversity that were caused by the presence of two highly diverged alleles at each locus. The presence of these alleles was confirmed by cloning and sequencing. We think these highly diverged alleles are found in *P. balsamifera* as a consequence of recent hybridization (see the Results section) and thus excluded these sequences from all analyses. The situation for *KTI3* was more complicated. Sequences from two independent rounds of PCR from both the northern and southern samples revealed that many individuals were segregating three nucleotides at multiple sites. To determine if this was caused by sequencing error or amplification of multiple gene copies, we cloned products from multiple PCRs and sequenced 8–10 clones per individual. The sequenced clones revealed 2–10 distinct alleles from each individual, consistent with recent duplication events. Neighbor-joining genealogies showed that the *KTI3* sequences fell into three clades; sequences within each clade had 98–99% similarity and sequences in different clades showed *c.* 95–97% similarity. Despite weak bootstrap support (< 65%, Fig. 3) we separated the *KTI3* data into these three clades for subsequent analyses, hereafter referred to as *KTI3-A*, *KTI3-B*, and *KTI3-C*. The existence of apparently duplicated copies of *KTI3* meant that we were unable to definitively assign ortholog status; we thus excluded *KTI3* from all comparisons of PIs vs reference loci.

Results

Nucleotide diversity and divergence

We sequenced an average of 2958 bp (2608 coding) from the PI loci and 6830 bp (3438 coding) from reference loci from each of 47 individuals – 16 from the range-wide sample, 15 from the northern sample and 16 from the southern sample (Fig. 1) – as well as 9410 bp (5391 coding) from the 16 reference loci that were sampled from the range-wide sample only. In the range-wide sample reference loci harbored 0–14 polymorphic sites per locus; π ranged from 0 to 0.008, π_{silent} from 0 to 0.008, θ_{Wtotal} from 0 to 0.0055, $\theta_{Wsilent}$ from 0 to 0.0123, and *Hd* from 0 to 0.81. In the same sample, the protease inhibitors, excluding *KTI3*, harbored 4–12 polymorphic sites per locus; π ranged from 0.0019 to 0.007, π_{silent} from 0.002 to 0.006, θ_{Wtotal} from 0.0019 to 0.046, $\theta_{Wsilent}$ from 0.0039 to 0.0077 and *Hd* from 0.33 to 0.86 (Table 1, Table S2). There was no significant difference in the ratio of

Table 1 Summary statistics for each of the protease inhibitors and the mean, minimum, and maximum of the 28 (range-wide) or 12 (northern, southern samples) reference loci

Sample		Length	*N*	*Hd*	θ_{Wtotal}	$\theta_{Wsilent}$	Tajima's D	*Fst*
Range-wide								
	CI1	715	28	0.82	0.0046	0.0077	–0.69	
	KTI4	590	28	0.86	0.0043	0.0040	–0.49	
	KTI5	639	32	0.33	0.0018	0.0054	–0.29	
	Gwin3	603	20	0.63	0.0019	0.0044	–0.15	
Reference	Mean	569	30	0.42	0.0015	0.0032	–0.37	
	Minimum			0	0	0	–1.55	
	Maximum			0.81	0.0045	0.0123	0.85	
Southern								
	CI1	715	26	0.78	0.0051	–	–1.59	–0.003
	KTI4	590	30	0.86	0.0037	–	0.65	–0.007
	KTI5	639	24	0.56	0.0015	–	1.64	–0.003
	Gwin3	603	20	0.68	0.0023	–	–0.67	–0.011
Reference	Mean			0.42	0.0016		–0.59	0.000
	Minimum			0	0.00		–1.52	–0.022
	Maximum			0.76	0.00		1.58	0.027
Northern								
	CI1	715	30	0.78	0.00379	–	–0.22	
	KTI4	590	32	0.94	0.00458	–	0.86	
	KTI5	639	26	0.43	0.00142	–	1.13	
	Gwin3	603	26	0.51	0.00174	–	–0.54	
Reference	Mean			0.38	0.00148		–0.32	
	Minimum			0.063	0.00041		–1.506	
	Maximum			0.69	0.00239		0.826	

Data for individual reference loci are provided in the Supporting Information Tables S2 and S3.
N is the number of alleles. Tajima's D values did not differ from neutral expectations.
Fst compares the northern to southern samples (only 12 reference loci used for *Fst*).
θ_{total} and (b) θ_{silent}.

replacement to synonymous diversity in the PIs vs reference loci ($\pi_{rep} : \pi_{syn}$), (Kruskal–Wallis, $\chi^2 = 0.545$, $P = 0.46$); neither were any of the $\pi_{rep} : \pi_{syn} > 1$.

On average, the PI loci harbored greater diversity and have evolved more rapidly than the reference loci. The greater diversity was largely caused by two genes, *CI1* and *KTI4*; the estimates of both *CI1* and *KTI4* θ_{Wtotal} were greater than all but 3 of the 28 reference locus estimates and fell within the upper 5% of the values obtained from the coalescent simulations (Fig. 2a). These PIs also showed higher diversity than reference loci when only silent sites were examined (Fig. 2b). The estimates of haplotype diversity at *CI1* and *KTI4* were greater than the estimates for any of the reference loci, with estimates that fell outside of the coalescent-based confidence intervals of the estimates for 23 of the 28 reference loci (Table S2). The divergence of sequences from *P. balsamifera* and *P. tremula* revealed that the PIs had evolved significantly faster than the reference loci; there were an average of 26.3 fixed differences at the PIs compared with 13.4 at the reference loci (*t*-test, $P = 0.022$).

Neutrality tests

The frequency distribution of segregating sites, as reflected by Tajima's D, did not deviate significantly from expectations under a neutral-equilibrium model for any locus; neither did Tajima's D values for PI loci differ from values estimated from reference loci (Kruskal–Wallis $P > 0.5$, Table 1, Table S2). Similarly, the mlHKA tests did not detect a significant difference in the fit of a model with selection on the PIs vs a neutral model ($\chi^2 = 5.234$, df = 5, $P > 0.5$), and thus revealed no evidence that PIs had experienced consistently different selective histories than the reference loci. Fay and Wu's H values for the reference loci, which ranged from −2.1 to 0.873, were not significantly different from 0 and fell within the range of the reference loci H values. For three of the PIs, the ratios of replacement to synonymous fixed differences vs polymorphisms did not differ significantly from the neutral expectations. For *Gwin3*, however, the MK test revealed a significant departure from neutral expectations for the north sample ($G = 4.421$, $P = 0.035$). However, this test did not remain significant after correcting for multiple comparisons.

Comparison of *Gwin3* sequences with homologous sequences from related species revealed evidence that *Gwin3* recently has evolved more rapidly at replacement than synonymous sites. Specifically, sequences from *P. balsamifera* differed from a sequence from a *P. trichocarpa* × *P. deltoides hybrid* at 0–2 synonymous sites and 9–10 replacement sites, with a minimum dN : dS = 1.4; 10 of 66 of the pairwise dN : dS values were significantly > 1 ($P < 0.005$), which is consistent

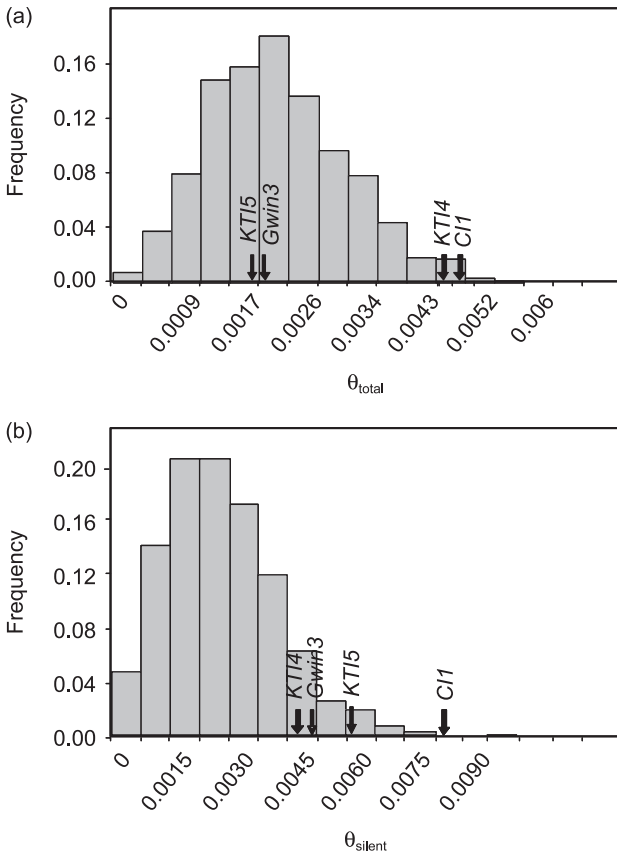

Fig. 2 Distribution of (a) θ_{total} and (b) θ_{silent} from coalescent simulations with population growth. Arrows designate the range-wide estimates of θ for each of the protease inhibitors. The mean values of θ and Tajima's D from the simulations closely matched those calculated from the 28 reference loci (see the Materials and Methods section).

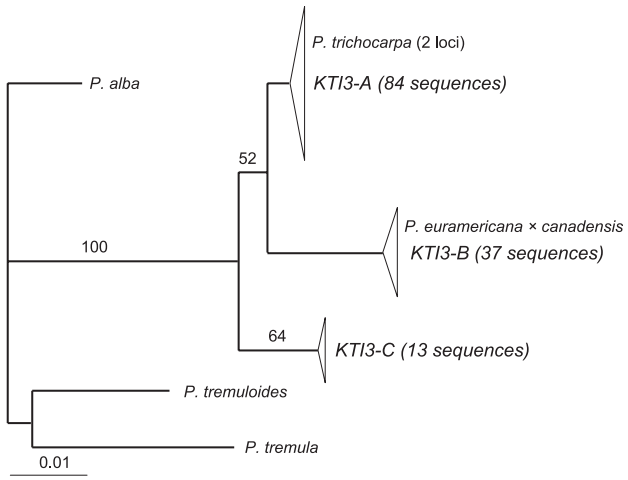

Fig. 3 Neighbor-joining tree of *KTI3* sequences showing relationships among *Populus balsamifera KTI3-A*, *KTI3-B*, and *KTI3-C* alleles and those from other poplar species. Numbers above branches are bootstrap values > 50.

with positive selection having driven the evolution of at least some of the alleles. *Populus balsamifera* differed from the more similar of two *Populus nigra Gwin3*-like sequences by one to three synonymous sites and seven to eight replacement sites, with a mean dN : dS > 1 (minimum dN : dS = 0.74); few of these values were significantly greater than 1, although most were significantly greater than 0.

Geographic variation in diversity

The northern and southern samples showed patterns similar to those from the range-wide sample; diversity at PI loci as a class did not differ significantly from that at reference loci, and Tajima's D and Fay and Wu's H values were similar for all three samples (Table S2). *KTI4* and *CI1* harbored greater diversity than the reference loci in both the northern and southern samples, although patterns in each of these samples fell within the distribution of values generated from coalescent simulations based on the range-wide sample (Table S2). Tests of genetic differentiation also revealed little evidence for population structure: estimates of *Fst* were near zero and estimates of *Snn* were near 0.5 (Table S3). In summary, these results provide no evidence for geographically variable selection acting at the PI loci or that population structure has played an important role in shaping diversity within *P. balsamifera*.

Divergence of recent duplicates

Analyses of *KTI3* data were complicated by the amplification of multiple, highly similar sequences, presumably owing to recent duplications. Since we were unable to assign sequences to specific loci, we identified three haplotype groups on the basis of genealogical relationships. Multiple PCRs from each individual revealed that some individuals contained 10 distinct alleles with as many as seven sequences from the same allelic class. By contrast, only two distinct sequences were sampled from other individuals and only one allelic class was sampled from some individuals. Together, the data suggest that *P. balsamifera* is polymorphic for the number of *KTI3* loci in the genome, with a minimum of five copies in some individuals. By comparison, in the sequenced genome of the closely related *P. trichocarpa*, three full-length *KTI3*-like sequences are identified by BLAST searches. Two of these were > 97% identical to the *P. balsamifera* sequences; the third was only 86–90% identical. Therefore, it appears that *P. trichocarpa* contains two members of the *KTI3* gene family we sampled with our PCR, suggesting either that duplications have occurred since *P. balsamifera* and *P. trichocarpa* diverged or that *P. trichocarpa* is also polymorphic for *KTI3* copy number.

Estimating diversity among these *KTI3* sequences is difficult because diversity estimators are sensitive to the length of the sequenced region, and with an unknown number of loci it is impossible for us to know the total length of the sequenced region included in our sample. Comparison of divergence among the *KTI3* allelic classes does, however, suggest that

these loci have evolved under relaxed selection or that positive selection has contributed to their divergence. The dN : dS values from allelic comparisons within each class were almost all < 1 (only 23 of 3775 comparisons were > 1). Similarly, pairwise comparisons within alleles of the least-frequent class and with these alleles vs the other two classes had dN : dS > 1 in only 10 of 1573 comparisons . By contrast, alleles from the two most common allelic classes differed from members of the other class by 0–4 synonymous sites and 4–12 replacement sites (mean dN : dS = 1.9, range = 0.27–6.9), and over 25% of pairwise comparisons with other members of other allelic classes had dN : dS > 1 (excluding 24 sequences for which dN : dS is not defined because dS = 0). Despite the high dN : dS estimates, few of these values were significantly > 1, likely because of the limited power associated with the number of sites that differentiate the sequences.

Introgression of *KTI4* and *Gwin3* alleles

Two individuals in the range-wide sample harbored highly divergent *KTI4* alleles and were heterozygous at > 30 sites, whereas other individuals were heterozygous at no more than six sites for *KTI4*. Similarly, two individuals from the southern *Gwin3* sample were heterozygous at 10–12 sites whereas other individuals were heterozygous at no more than three sites. Such highly diverged alleles segregating at these loci is highly unlikely in the absence of either strong selection maintaining the polymorphism or recent introgression. Reproductive barriers between many poplar species are relatively weak and hybridization among North American poplars is common (Stettler *et al.*, 1996; Hamzeh *et al.*, 2007). As such, the potential for introgression, rather than long-term balancing selection appears likely. In fact, this appears to be the case; the low-frequency highly diverged alleles we sampled were much more similar to sequences from other *Populus* species than they were to the more common alleles we sampled. For *KTI4*, the rare allele differed from the other *P. balsamifera KTI4* sequences by 30–36 sites and a maximum of one replacement site, but differed from a *P. deltoides* allele in GenBank at only four bases. Similarly, the rare *Gwin3* alleles differed from other sampled sequences at 9–11 sites (9–10 replacement sites), but were identical (or nearly so) to a sequence from a *P. trichocarpa* × *P. deltoides* hybrid (Fig. S1).

Discussion

We used a survey of nucleotide diversity at five protease inhibitor genes and 28 reference loci from three populations of *P. balsamifera* to gain insight into the evolution of plant defense genes and the consistency of molecular evolutionary patterns at homologous loci in closely related lineages. The five PI genes showed varied patterns of diversity. For the four loci for which we were able to estimate intraspecific diversity, *Gwin3* and *KTI5* harbored diversity similar to that of reference loci. Although diversity similar to reference loci is consistent with a recent evolutionary history shaped primarily by purifying selection, the ratio of replacement to synonymous site divergence in *Gwin3* suggests that this locus has experienced repeated bouts of positive selection. Moreover, the PI loci as a class and *Gwin3* in particular, have diverged more rapidly then reference loci, which may also reflect nonneutral evolution. This apparent inconsistency between intraspecific diversity and interspecific divergence at *Gwin3* may reflect differences in the statistical power of intraspecific and interspecific tests of selection (Zhai *et al.*, 2009). Alternatively, and not necessarily independent of the issues of the powers of these statistical tests, the apparent inconsistency may result from episodic selection, perhaps reflecting the time between favorable mutations or intermittent selection associated with fluctuations in herbivore loads (Tiffin & Gaut, 2001).

The mean diversity of the PIs was greater than the mean diversity of the reference loci and two of the PIs, *KTI4* and *CI1*, harbor significantly greater total diversity than coalescence-based expectations as well as 25 of the 28 reference loci. All four PIs also show high diversity when only silent sites are considered. There are at least four potential explanations for the elevated diversity we find at the PI genes. One possibility is that PIs are found at more rapidly evolving regions of the genome, and the elevated diversity reflects higher mutation rates or weaker background selection. The more rapid divergence of PIs than reference loci (greater numbers of fixed differences between *P. balsamifera* and *P. tremula*) is consistent with this hypothesis.

A second possibility is that the PI genes have experienced weaker purifying selection, or conditional neutrality (perhaps because plants have not been attacked recently by enemies against which the defense genes are effective). The hypothesis of conditional neutrality for plant *R*-genes was motivated by finding a high frequency of premature stop codons and frame-shift mutations segregating at NBS-LRR genes in *Arabidopsis thaliana* (Gos & Wright, 2008). Three lines of evidence suggest that conditional neutrality is not responsible for the elevated diversity we find segregating at *KTI4* and *CI1*: high diversity was found when only silent sites were considered, we detected no premature stop codons and biochemical assays have established that at least some of the PIs we examined encode proteins that are active against current enemies of poplar (Major & Constabel, 2008).

If selection is responsible for the excess diversity in range-wide samples it may reflect selection that favors different alleles in different parts of a species range. Geographically variable selection is consistent with evidence that herbivore- and pathogen-imposed selection in contemporary populations is often highly variable (Thompson, 2005). We, however, found similar patterns of diversity in the range-wide, northern, and southern samples, and thus no evidence that excess diversity in our sample results from geographically variable selection. Although this conclusion should be viewed with caution given

that we sampled only two subpopulations, more extensive surveys of defense-gene diversity within and among populations of other plant species have also detected little evidence for local adaptation (Bakker *et al.*, 2006; Moeller & Tiffin, 2008). Thus, although there is extensive evidence that population structure affects coevolutionary interactions (Thompson, 2005), local adaptation does not appear to have had a strong effect on the nucleotide diversity of most defense genes. The different apparent signals from experiments in contemporary populations compared with population genetic analyses suggests that selection may be temporally variable such that strong selection measured in contemporary populations may not be stable enough to strongly affect nucleotide diversity. Alternatively, selection in contemporary populations is often measured on polygenic traits, which may leave little evidence of selection at the individual loci that contribute to their variation (Kelly, 2006).

A final potential explanation for the apparent excess of diversity at defense genes is they have evolved in response to frequency-dependent selection (Rose *et al.*, 2004) resulting in partial sweeps of rare alleles. Both the excess diversity and rapid evolution of the PIs appear to be consistent with a history of frequency-dependent selection. Nevertheless, claims that frequency-dependent selection is responsible for elevated levels of polymorphism in defense genes are made largely on the basis of verbal arguments. Formal models investigating the effects that ecologically realistic frequency-dependent or temporally variable selection are expected to have on nucleotide diversity and divergence would be valuable.

Evidence for diverse evolutionary histories of PI loci within *P. balsamifera* is similar to the conclusions from studies of PIs in other plant lineages (Tiffin & Gaut, 2001; Clauss & Mitchell-Olds, 2003, 2004; Moeller & Tiffin, 2005, 2008; Ingvarsson, 2005) as well as studies of other plant defense loci such as R genes (Bakker *et al.*, 2006; Dodds *et al.*, 2006), chitinases (Tiffin, 2004), and glucosinolate synthesis enzymes (Kroymann *et al.*, 2003; Stranger & Mitchell-Olds, 2005). This diversity raises an obvious question of whether the different selective histories (purifying selection driving the evolution of some PIs, positive selection driving the evolution of others, and an excess of diversity at other loci) correspond to different functional roles. Our picture of adaptation of host defenses would be quite different if the PIs we examined are all involved in defense compared with the PIs that show signatures of positive selection or excess diversity are involved in defense whereas the others are involved in regulating endogenous enzymes. Similarly, there has been speculation that defense gene diversity may depend on whether those defenses are primarily targeted against generalists or specialist enemies (Clauss & Mitchell-Olds, 2004; Tiffin *et al.*, 2004). While we are not aware of assays establishing the ecological function of the protease inhibitors we studied, there is evidence that they are induced following herbivore and/or mechanical damage, and at least some are highly effective inhibitors of exogenous proteases (Bradshaw *et al.*, 1990; Christopher *et al.*, 2004; Major & Constabel, 2006, 2008). Although expression and biochemical function do not allow for direct inferences of ecological function, induced genes with potential roles in protecting the host against herbivores and pathogens are often assumed to be part of the host's defense response. If this assumption is sound, then our study provides no evidence that the different patterns of diversity we found at the PI loci are caused by differences in the function of these genes.

Evolutionary history of PIs in *P. balsamifera* and *P. tremula*

We compared the evolution of *P. balsamifera* PIs with those previously characterized in *P. tremula* (Ingvarsson, 2005) to provide some insight into the consistency of molecular evolution at the nucleotide level. In *P. tremula*, the strongest evidence for recent selection on PIs is found at *KTI1* and *KTI2*, which harbor significantly greater replacement than synonymous site diversity ($\pi_R : \pi_S > 1$) and two distinct alleles segregating at *KTI1* (Ingvarsson, 2005). By contrast, diversity at *Gwin3*, the apparent *P. balsamifera* homolog of *P. tremula KTI1* and *KTI2*, does not differ from that at the reference loci although dN : dS values indicate that positive selection has driven the between-species divergence of *Gwin3*. In both species, *CI1* is a candidate for recent selection, but the signatures of selection differ. *Populus tremula CI1* harbors an excess of rare polymorphisms (significantly negative Tajima's D; Ingvarsson, 2005), as expected during recovery from a selective sweep, whereas *P. balsamifera CI1* harbors excess diversity with no strong skew towards rare variants. *KTI3*, *KTI4* and *KTI5* harbor diversity in *P. tremula* that is consistent with reference loci (Ingvarsson, 2005). By comparison, in *P. balsamifera KTI3* has a complicated history in which positive selection probably plays a part (see below), *KTI4* and KTI5 show some evidence of excess diversity. In summary, our data do not provide a strong case that the evolutionary responses to enemy-imposed selection are consistent in related lineages. The general pattern found at PIs – varied selective histories with little evidence for consistent response to selection in closely-related species – is similar to that found for *Zea* chitinase genes (Tiffin, 2004), which encode proteins involved in pathogen defense. This variation in the apparent targets of selection at the nucleotide level may be caused by differences in the composition of the enemies that attack plants – if enemies harbor different proteases then, presumably, the PIs that are effective against those enemies will differ. Alternatively, the selection imposed on plant defense may be similar in the two lineages but temporal variation in selection, differences in genetic background or the complexity of potential adaptive responses may mean that it is unlikely that selection will similarly affect the evolution of the same genes in different lineages (Cohan & Hoffmann, 1989; Simões *et al.*, 2008).

Rapid evolution of *KTI3* sequences

Analysis of the *KTI3* data was complicated because the sequences clearly did not come from a single locus and the mixing of orthologous and paralogous sequences precluded us from using diversity-based statistics to infer past selection. Nevertheless, pair-wise dN : dS values suggest that either relaxed or positive selection has acted on recent duplicates. If positive selection is responsible then this result may reflect a selective advantage for the high diversity of *KTI3* proteins, possibly because higher diversity increases the range of herbivores against which a host genotype is defended (Duda & Palumbi, 1999). Interestingly, only one of the amino acid substitutions that differentiate the three allelic groups we identified is in the reactive-site loop, which is the region of the molecule that has strongest effect on the efficacy and specificity of inhibitory activity (Bode & Huber, 1992). Similarly, the selection that has acted during the divergence of *KTI* paralogs within and among poplar species appears to target regions outside of the reactive-site loop (Talyzina & Ingvarsson, 2006). Of the 11 amino acids that Talyzina & Ingvarsson (2005) identified as likely targets of selection during the divergence of poplar KTIs, one of them, which is located in a loop connecting β-sheets (Major & Constabel, 2008; Fig. S2), segregates three amino acids each differing in side-chain properties (uncharged polar, nonpolar, basic). This amino acid is within two amino acids of several other substitutions and indels that differ among the three allelic *KTI3* classes. This region of the molecule thus appears to be a strong candidate for selection. Because this region is distant from the active site it is not likely to have a strong effect on the efficacy of protease inhibition. One possibility is that this is a region of the molecule that enemy proteases target to deactivate host PIs and thereby circumvent defense.

One of the complicating factors in the *KTI3* analyses is that *P. balsamifera* appears to be segregating copy number variants. Copy number polymorphism has also been documented in some plant *R*-gene clusters (Sun *et al.*, 2001), antiherbivore defense genes (Kroymann *et al.*, 2003) and antimicrobial defensin genes in humans (e.g. Aldred *et al.*, 2005). While polymorphism in the size of defense gene families may be common, the evolutionary forces acting on copy number and potential costs associated with increases in the size of these families remains largely unexplored. Regardless of the evolutionary forces acting, our *KTI3* data raise a note of caution for population genetic surveys that identify excess diversity. We suspected that we were amplifying multiple copies only when we detected several sites at which three bases were segregating. Without those sites we may have easily assumed that *KTI3* was a single locus, or with information from the *P. trichocarpa* genome, at most, duplicated. Only repeated rounds of amplification and the sequencing of multiple clones provided insight into the potential complexity of the *KTI3*-gene family. Given that polymorphism for copy number may be common (Borevitz *et al.*, 2003, 2007; Ossowski *et al.*, 2008), the possibility of sampling paralogs should be considered when excess nucleotide diversity in PCR-based resequencing data is detected.

Conclusion

Surveys of nucleotide diversity provide an opportunity to investigate whether the genes that underlie responses to selection are similar in closely related lineages and whether selection acting on the phenotype results in similar adaptive responses at the nucleotide level. Our analyses of nucleotide diversity segregating at protease inhibitors in *P. balsamifera* reveal that the adaptive history of genes encoding proteins with similar biochemical and putative ecological function varies both within a species and between closely related species. Whether these diverse histories are the result of differences in the agents of selection, variation in the time and type of mutations that entered populations or stochastic responses of translating selection on a complex phenotype to adaptive response at the nucleotide level remain open questions. The coupling of comparative population genetic analyses, functional characterization of gene function and characterization of genetic architecture underlying variation in complex traits should help to evaluate the relative importance of these factors in determining the repeatability and nature of adaptation.

Acknowledgements

We thank W. D. Shroeder and Salim Silim of the Agroforestry Division of Agriculture and Agri-Food Canada in Indian Head, Saskatchewan, for generously providing access to the balsam poplar genotypes analysed in this report, Brian Arnold, Molly Peterson, Riya Jay and Jennifer Reese for helping in the collection of sequence data, Peter Constabel and Ian Major for discussion about protease inhibitors, P. K. Ingvarsson for providing PCR primers for PI loci and the anonymous reviewers for comments and criticisms that improved the paper. Funding for this work was provided by the University of Minnesota Institute for Renewable Energy and the Environment and the US National Science Foundation (NSF DBI-0701911 awarded to M. S. Olson and P. Tiffin).

References

Aldred PM, Hollox EJ, Armour JA. 2005. Copy number polymorphism and expression level variation of the human alpha-defensin genes *DEFA1* and *DEFA3*. *Human Molecular Genetics* **15**: 2045–52.

Anderson CNK, Ramakrishnan U, Chan YL, Hadly EA. 2005. Serial SimCoal: a population genetic model for data from multiple populations and points in time. *Bioinformatics* **21**: 1733–1734.

Bakker EG, Toomajian C, Kreitman M, Bergelson J. 2006. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell* **18**: 1803–1818.

Bode W, Huber R. 1992. Natural protein proteinase inhibitors and their interaction with proteinases. *European Journal of Biochemistry* **204**: 433–451.

Bonan GB. 2002. *Ecological climatology: concepts and applications*. Cambridge, UK: Cambridge University Press.

Borevitz J, Liang D, Plouffe D, Chang H, Zhu T, Weigel D, Berry C, Winzeler E, Chory J. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research* **13**: 513–523.

Borevitz JO, Hazen SP, Michael TP, Morris GP, Baxter IR, Hu TT, Chen H, Werner JD, Nordborg M, Salt DE *et al.* 2007. Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **104**: 12057–12062.

Bradshaw HD, Hollick JB, Parsons TJ, Clarke HR, Gordon MP. 1990. Systemically wound-responsive genes in poplar trees encode proteins similar to sweet potato sporamins and legume Kunitz trypsin inhibitors. *Plant Molecular Biology* **14**: 51–59.

Christeller JT. 2005. Evolutionary mechanisms acting on proteinase inhibitor variability. *FEBS Journal* **272**: 5710–5722.

Christopher ME, Miranda M, Major IT, Constabel CP. 2004. Gene expression profiling of systemically wound-induced defenses in hybrid poplar. *Planta* **219**: 936–947.

Clark NL, Aagaard JE, Swanson WJ. 2006. Evolution of reproductive proteins from animals and plants. *Reproduction* **131**: 11–22.

Clauss MJ, Mitchell-Olds T. 2003. Population genetics of tandem trypsin inhibitor genes in *Arabidopsis* species with contrasting ecology and life history. *Molecular Ecology* **12**: 1287–1299.

Clauss MJ, Mitchell-Olds T. 2004. Functional divergence in tandemly duplicated *Arabidopsis thaliana* trypsin inhibitor genes. *Genetics* **166**: 1419–1436.

Cohan FM, Hoffmann AA. 1989. Uniform selection as a diversifying force in evolution: evidence from *Drosophila*. *American Naturalist* **134**: 613–637.

DeMeaux J, Mitchell-Olds TM. 2003. Evolution of plant resistance at the molecular level: ecological context of species interactions. *Heredity* **94**: 343–352.

Dodds PN, Lawrence GJ, Catanzariti AM, Teh T, Wang CI, Ayliffe MA, Kobe B, Ellis JG. 2006. Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proceedings of the National Academy of Sciences, USA* **103**: 8888–8893.

Duda TF, Palumbi SR. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proceedings of the National Academy of Sciences, USA* **96**: 6820–6823.

Excoffier L, Novembre J, Schneider S. 2000. SIMCOAL: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. *Journal of Heredity* **91**: 506–509.

Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.

Ford MJ. 2002. Applications of selective neutrality tests to molecular ecology. *Molecular Ecology* **11**: 1245–1262.

Giri AP, Harsulkar AM, Deshpande VV, Sainani MN, Gupta VS, Ranjekar PK. 1998. Chickpea defensive proteinase inhibitors can be inactivated by podborer gut proteinases. *Plant Physiology* **116**: 393–401.

Gos G, Wright S. 2008. Conditional neutrality at two adjacent NBS-LRR disease resistance loci in natural populations of *Arabidopsis lyrata*. *Molecular Ecology* **17**: 4953–4962.

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

Hamzeh M, Sawchyn P, Périnet P, Dayanandan S. 2007. Asymmetrical natural hybridization between *Populus deltoides* and *P. balsamifera* (Salicaceae). *Canadian Journal of Botany* **85**: 1227–1232.

Haq SK, Atif SM, Khan RH. 2004. Protein proteinase inhibitors in combat against insects, pests and pathogens: natural and engineered phytoprotection. *Archives of Biochemistry and Biophysics* **431**: 145–159.

Hudson RR. 2000. A new statistic for detecting genetic differentiation. *Genetics* **155**: 2011–2014.

Ingvarsson PK. 2005. Molecular population genetics of herbivore-induced protease inhibitor genes in European aspen (*Populus tremula* L. Salicaceae). *Molecular Biology and Evolution* **22**: 1802–1812.

Kelly JK. 2006. Geographical variation in selection, from phenotypes to molecules. *American Naturalist* **167**: 481–495.

Kroymann J, Donnerhacke S, Schnabelrauch D, Mitchell-Olds T. 2003. Evolutionary dynamics of an *Arabidopsis* insect resistance quantitative trait locus. *Proceedings of the National Academy of Sciences, USA* **100**: 14587–14592.

Lopes AR, Juliano MA, Juliano L, Terra WR. 2004. Coevolution of insect trypsins and inhibitors. *Archives of Insect Biochemistry and Physiology* **55**: 140–152.

Luquez V, Hall D, Albrectsen BR, Karlsson J, Ingvarsson P, Jansson S. 2008. Natural phenological variation in aspen (*Populus tremula*): the SwAsp collection. *Tree Genetics & Genomes* **4**: 279–292.

Major IT, Constabel CP. 2006. Molecular analysis of poplar defense against herbivory: comparison of wound- and insect elicitor-induced gene expression. *New Phytologist* **172**: 617–635.

Major IT, Constabel CP. 2008. Functional analysis of the Kunitz trypsin inhibitor family in poplar reveals biochemical diversity and multiplicity in defense against herbivores. *Plant Physiology* **446**: 888–993.

Mazumdar-Leighton S, Broadway RM. 2001. Transcriptional induction of diverse midgut trypsins in larval *Agrostis ipsilon* and *Helicoverpa zea* feeding on the soybean trypsin inhibitor. *Insect Biochemistry and Molecular Biology* **31**: 645–657.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–655.

Meyers BC, Kaushik K, Nandety RS. 2005. Evolving disease resistance genes. *Current Opinion in Plant Biology* **8**: 129–134.

Moeller DA, Tiffin P. 2005. Genetic diversity and the evolutionary history of plant immunity genes in two species of Zea. *Molecular Biology and Evolution* **22**: 2480–2490.

Moeller DA, Tiffin P. 2008. Geographic variation in adaptation at the molecular level: a case study of plant immunity genes. *Evolution* **62**: 3069–3081.

Nei M. 1987. *Molecular evolutionary genetics*. New York, NY, USA: Columbia University Press.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Research* **18**: 2024–2033.

Rausher MD. 2001. Co-evolution and plant resistance to natural enemies. *Nature* **411**: 857–864.

Rose LE, Bittner-Eddy PD, Langley CH, Holub EB, Michelmore RW, Beynon JL. 2004. The maintenance of extreme amino acid diversity at the disease resistance gene, *RPP13*, in *Arabidopsis thaliana*. *Genetics* **166**: 1517–1527.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DNASP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.

Ryan CA. 1990. Protease inhibitors in plants: genes for improving defenses against insects and pathogens. *Annual Review of Phytopathology* **28**: 425–449.

Simões P, Santos J, Fragata I, Mueller LD, Rose MR, Matos M. 2008. How repeatable is adaptive evolution? The role of geographical origin and founder effects in laboratory adaptation. *Evolution* **62**: 1817–1829.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics* **76**: 449–462.

Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* **68**: 978–989.

Stettler RF, Zsuffa L, Wu R. 1996. The role of hybridization in the genetic manipulation of *Populus*. In: Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM, eds. *Biology of Populus and its implications for management and conservation*. Ottawa, Canada: NRC Research Press, 87–112.

Stranger B, Mitchell-Olds T. 2005. Nucleotide variation at the myrosinase-encoding locus,*TGG1*, and quantitative myrosinase enzyme activity variation in *Arabidopsis thaliana*. *Molecular Ecology* **14**: 295–309.

Sun Q, Collins N, Ayliffe M, Smith S, Drake J, Pryor T, Hulbert S. 2001. Recombination between paralogues at the *rp1* rust resistance locus in maize. *Genetics* **158**: 423–438.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Talyzina NM, Ingvarsson PK. 2006. Molecular evolution of a small gene family of wound inducible Kunitz trypsin inhibitors in *Populus*. *Journal of Molecular Evolution* **63**: 108–119.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**: 1596–1599.

Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. *Molecular Biology and Evolution* **21**: 1214–1225.

Thompson JN. 2005. *The geographic mosaic of coevolution*. Chicago, IL, USA: University of Chicago Press.

Tiffin P. 2004. Comparative evolutionary histories of chitinase genes in the genus *Zea* and family Poaceae. *Genetics* **167**: 1331–1340.

Tiffin P, Gaut BS. 2001. Molecular evolution of the wound-induced serine protease inhibitor *wip1* in *Zea* and related genera. *Molecular Biology and Evolution* **18**: 2092–2101.

Tiffin P, Moeller DA. 2006. The molecular evolution of plant immune system genes. *Trends in Genetics* **22**: 662–670.

Tiffin P, Hacker R, Gaut BS. 2004. Population genetic evidence for rapid changes in intraspecific diversity and allelic cycling of a specialist defense gene in Zea. *Genetics* **168**: 425–434.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7: 188–193.

Wright SI, Charlesworth B. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. *Genetics* **168**: 1071–1076.

Zhai W, Nielsen R, Slatkin M. 2009. An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution* **26**: 273–283.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences, USA* **95**: 3708–3713.

Supporting Information

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

Table S1 Length, primers, GenBank accessions, outgroup sequence, and genomic location for PI and reference loci

Table S2 A summary of basic descriptive and diversity statistics for each of the loci in our study

Table S3 Estimates of *Fst* and *Snn* values for the PI and reference loci

Fig. S1 Neighbor-joining genealogies showing relationships of *P. balsamifera KTI4* and *Gwin3* alleles to sequences from other poplar species.

Fig. S2 Amino acid sequence of *KTI3* showing structural features and amino acids previously identified as having diverged among Poplar *KTI* gene family members.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.