MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 724–738 doi: 10.1111/j.1365-294X.2012.05667.x

Recurrent gene deletions and the evolution of adaptive cyanogenesis polymorphisms in white clover (*Trifolium repens* L.)

KENNETH M. OLSEN, NICHOLAS J. KOOYERS, and LINDA L. SMALL
Department of Biology, Campus Box 1137, Washington University in St. Louis, St. Louis, MO 63130-4899 USA

Abstract

Understanding the molecular evolution of genes that underlie intraspecific polymorphisms can provide insights into the process of adaptive evolution. For adaptive polymorphisms characterized by gene presence/absence (P/A) variation, underlying loci commonly show signatures of long-term balancing selection, with gene-presence and gene-absence alleles maintained as two divergent lineages. We examined the molecular evolution of two unlinked P/A polymorphisms that underlie a well-documented adaptive polymorphism for cyanogenesis (hydrogen cyanide release with tissue damage) in white clover. Both cyanogenic and acyanogenic plants occur in this species, and the ecological forces that maintain this chemical defence polymorphism have been studied for several decades. Using a sample of 65 plants, we investigated the molecular evolution of sequences flanking the two underlying cyanogenesis genes: *Ac/ac* (controlling the presence/absence of cyanogenic glucosides) and *Li/li* (controlling the presence/absence of their hydrolysing enzyme, linamarase). A combination of genome walking, PCR assays, DNA sequence analysis and Southern blotting was used to test whether these adaptive P/A polymorphisms show evidence of long-term balancing selection, or whether gene-absence alleles have evolved repeatedly through independent deletion events. For both loci, we detect no signatures of balancing selection in the closest flanking genomic sequences. Instead, we find evidence for variation in the size of the deletions characterizing gene-absence alleles. These observations strongly suggest that both of these polymorphisms have been evolving through recurrent gene deletions over time. We discuss the genetic mechanisms that could account for this surprising pattern and the implications of these findings for mechanisms of rapid adaptive evolution in white clover.

Keywords: balancing selection, copy number variation (CNV), cyanogenesis, parallel evolution, recurrent gene deletions, *Trifolium repens*

Received 14 February 2012; revision received 12 April 2012; accepted 2 May 2012

Introduction

Understanding the molecular basis of adaptation is a central goal of evolutionary biology. Effective research in this area requires study systems that are not only genetically tractable but also ecologically well characterized, so that the adaptive significance of molecular variation and corresponding phenotypes can be understood in their natural context (Mitchell-Olds *et al.* 2007). With advances in genomic approaches, it is becoming increasingly possible to genetically characterize ‘ecological model species’, where the adaptive significance of natural phenotypic variation has already been well documented. We have taken this strategy in studying the genetic basis of an adaptive polymorphism for cyanogenesis (cyanide release with tissue damage) in white clover (*Trifolium repens* L.) (Olsen *et al.* 2007, 2008; Olsen & Ungerer 2008; Kooyers & Olsen 2012; see also Olson & Levsen 2012). The ecological forces that maintain this chemical defence polymorphism have been

Correspondence: Kenneth M. Olsen, Fax. 1-314-935-4432;
Email: kolsen@wustl.edu

studied for several decades, providing a well-established context for understanding its adaptive significance (reviewed below). In previous work, we have determined the molecular genetic basis of this adaptive polymorphism (Olsen *et al.* 2007, 2008). In this study, we examine the molecular evolutionary mechanisms that are responsible for the long-term evolution and selective maintenance of cyanogenesis genetic variation in white clover.

The white clover cyanogenesis polymorphism

Trifolium repens is a creeping legume and a common component of lawns, roadsides and pastures in temperate regions worldwide. This species is polymorphic for cyanogenesis, with both cyanogenic and acyanogenic plants occurring in natural populations (Armstrong *et al.* 1913; Ware 1925). Cyanogenesis occurs in more than 2500 plant species and is generally thought to function at least in part as a chemical defence against herbivory (Hughes 1991; Møller 2010). Consistent with this hypothesis, cyanogenic white clover plants are differentially protected from small generalist herbivores, including slugs, snails, insects and voles (e.g. Dirzo & Harper 1982a; Pederson & Brink 1998; Saucy *et al.* 1999; Viette *et al.* 2000). At the same time, white clover populations show climate-associated clinal variation in cyanogenesis, with acyanogenic plants predominating at higher latitudes and elevations (e.g. Daday 1954a,b; Till-Bottraud *et al.* 1988; Kooyers & Olsen 2012). The apparent selective advantage of acyanogenic clover in colder climates has been proposed to reflect fitness trade-offs between energetic investment in defence vs. growth/reproduction in regions of high and low herbivore pressure (e.g. Kakes 1989; see also Pennings & Silliman 2005). Alternatively, abiotic factors may be a determining factor; for example, cyanogenic plants may be directly selected against in colder climates, as frost-induced tissue damage may lead to cyanide autotoxicity (Daday 1965; Dirzo & Harper 1982b; Brighton & Horne 1977; but see also Olsen & Ungerer 2008). Whether primarily reflecting biotic or abiotic factors, the fact that cyanogenesis clines have evolved repeatedly, both in the native Eurasian species range (e.g. Daday 1954a,b; De Araujo 1976; Till-Bottraud *et al.* 1988; Pederson *et al.* 1996; Majumdar *et al.* 2004) and in introduced populations worldwide (e.g. Daday 1958; Ganders 1990; Kooyers & Olsen 2012), suggests that the selective factors maintaining this adaptive polymorphism are strong and geographically pervasive.

The cyanogenic phenotype in clover requires two biochemical components that are separated in intact tissue and brought into contact with cell rupture: cyanogenic glucosides, which are stored in the vacuoles of photosynthetic tissue; and their hydrolytic enzyme, linamarase, which is stored in the cell wall (reviewed by Hughes 1991). Acyanogenic clover plants may lack cyanogenic glucosides, linamarase or both components. Inheritance of the two cyanogenic components is controlled by two independently segregating Mendelian genes, where the dominant (functional) allele confers the presence of the component (Coop 1940; Melville & Doak 1940; Corkill 1942). The *Ac* gene controls the presence/absence of cyanogenic glucosides, and the unlinked *Li* gene controls the presence/absence of linamarase. Thus, plants that possess at least one dominant allele at both genes (*Ac_*, *Li_*) are cyanogenic, while homozygous recessive genotypes at either or both genes (*acac*, *lili*) lead to the absence of one or more of the required components. The presence or absence of each component can be determined for individual plants through colorimetric HCN assays (e.g. Feigl & Anger 1966) and exogenously added cyanogenic components (method described in Olsen *et al.* 2007). Among plants producing cyanogenic components, there is wide quantitative variation in the levels of the compounds present. This variation is partly attributable to the *Ac/ac* and *Li/li* genes themselves (heterozygotes produce intermediate levels of the compounds), as well as to environmental effects and uncharacterized genetic factors (Corkill 1942; Vickery *et al.* 1987; Hughes 1991; K. Olsen, unpublished observations).

In recent studies, we have documented the molecular genetic basis of the *Ac/ac* and *Li/li* biochemical polymorphisms. The nonfunctional *ac* and *li* alleles correspond, respectively, to gene deletions at two unlinked loci: *CYP79D15*, which encodes the cytochrome P450 protein catalysing the first dedicated step in cyanogenic glucoside biosynthesis (Olsen *et al.* 2008); and *Li*, which encodes the linamarase protein (Olsen *et al.* 2007). Thus, the *Ac/ac* and *Li/li* biochemical polymorphisms in white clover correspond to two independently segregating gene presence/absence (P/A) polymorphisms.

Evolution of adaptive gene presence/absence polymorphisms

Adaptive P/A polymorphisms have been best studied in plant resistance genes (*R*-genes), where the presence or absence of specific loci is associated with adaptive variation for resistance to co-evolving pathogens (e.g. Grant *et al.* 1998; Stahl *et al.* 1999; Shen *et al.* 2006). There are two different molecular evolutionary mechanisms that can potentially underlie the origin and maintenance of this type of polymorphism. One possibility is that the P/A polymorphism has arisen in the distant evolutionary past, with long-term balancing selection maintaining the gene-presence and gene-absence allele classes up to the present day. This mechanism is expected to create two evolutionarily divergent haplotype lineages corresponding to the two allele classes. Examination of DNA sequences immediately flanking a gene P/A polymorphism (*i.e.* sequences that are present in all individuals and that are linked to the P/A polymorphism) can be used to test for this haplotype structure as well as for other molecular signatures predicted under balancing selection (e.g. Stahl *et al.* 1999; reviewed by Charlesworth 2006; see Fig. 1). This pattern of long-term balancing selection has been documented at several *R*-gene P/A polymorphisms in *Arabidopsis thaliana* (Stahl *et al.* 1999; Tian *et al.* 2002; Shen *et al.* 2006).

Alternatively, an adaptive P/A polymorphism could be evolving through recurrent gene deletion events over time, with the repeated origin of gene-absence alleles from the gene-presence allele class. If gene deletion alleles have evolved repeatedly, a signature of long-term balancing selection is not expected. There is at least one instance where recurrent gene deletions have been documented in *R*-genes: gene-absence alleles of *RPM1* have evolved at least twice independently, in *Arabidopsis thaliana* and *Brassica napus* (Grant *et al.* 1998). More generally, recurrent deletions have been identified as playing a role in several other instances of parallel evolution (e.g. Chan *et al.* 2010; McGrath *et al.* 2011); in these cases, there is not selection to maintain an adaptive polymorphism within populations, but rather directional selection to fix an adaptive trait in separate populations that occur in a specific environment. For the cyanogenesis polymorphisms in white clover, if the *Ac/ac* and *Li/li* genes were evolving through recurrent gene deletion events, one would not expect to find molecular signatures of long-term balancing selection in flanking genomic sequences. Moreover, one might expect to observe variation in the size of the genomic deletions occurring among gene-absence alleles, a pattern consistent with multiple, independent deletion events.

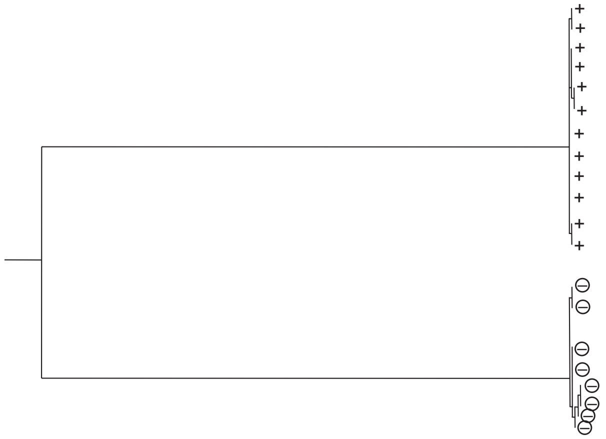

Fig. 1 Balancing selection at an adaptive gene presence/absence (*P/A*) locus. Haplotype tree constructed from downstream sequences flanking the gene deletion boundary of the *Arabidopsis thaliana R*-gene *RPS5*. Polymorphism data were obtained from Fig. 2b of Tian *et al.* (2002), and tree construction followed the protocol employed for the white clover cyanogenesis gene regions (see Materials and methods). Branch lengths are approximately proportional to nucleotide divergence among haplotypes. Plus and minus symbols correspond, respectively, to accessions that have or that lack the *RPS5* gene. A recombinant haplotype present in two accessions was excluded from the analysis.

While the white clover *Li* and *Ac* genes have been sequenced (Olsen *et al.* 2007, 2008), their genomic locations are currently unknown (genetic mapping is in progress; K. Olsen, unpublished data), and their surrounding genomic sequences have not been previously determined. In this study, we have undertaken genome walking to identify and characterize these flanking sequences for assessing the molecular evolution of the cyanogenesis P/A polymorphisms. Using sequences extending up to 13 kb away from the cyanogenesis genes, we examine whether the molecular evolution of these adaptive polymorphisms reflects long-term balancing selection, recurrent gene deletions or some combination of these two mechanisms. Our analyses reveal evidence that, unlike most well-characterized adaptive P/A polymorphisms, the long-term evolution of both *Ac/ac* and *Li/li* has likely been shaped by recurrent gene deletion events.

Materials and methods

Study system and sampling

Trifolium repens is a self-incompatible, perennial species that reproduces by seed and spreads vegetatively by stolons. A native of Eurasia, it has been introduced into temperate and cool tropical regions worldwide as a forage crop and lawn plant, and it is widely naturalized around areas of human activity. The species shows little evidence of population structure, either locally (Ennos 1982; Kooyers & Olsen 2012), on a continental scale (e.g. Kooyers & Olsen 2012), or globally (George *et al.* 2006; Olsen *et al.* 2007). White clover is an allotetraploid species; however, all classical and molecular genetic studies indicate that the two cyanogenesis genes are present in only one of the two parental genomes (Williams & Williamson 2001; Badr *et al.* 2002; Olsen *et al.* 2007, 2008). This pattern suggests that the species may have originated through ancient hybridization between a cyanogenic and an acyanogenic *Trifolium* species.

Plants used in this study were obtained either through the USDA worldwide clover germplasm collection

(51 accessions) or from naturalized North American populations (14 accessions) (Table 1; see also Table S1, Supporting Information). All plants were grown under standard greenhouse conditions at Washington University. Cyanogenesis phenotyping assays were performed using fresh leaf tissue following the methods described by Olsen *et al.* (2007). Previous PCR assays and Southern blotting using a worldwide sample of clover plants (>500 accessions) have indicated a near-perfect correlation between presence/absence of cyanogenic components and cyanogenesis gene P/A variation (Olsen *et al.* 2007, 2008; Kooyers & Olsen 2012; K. Olsen, unpublished observations).

Identification of DNA sequences flanking the cyanogenesis genes

Genomic sequences flanking the *CYP79D15* and *Li* loci were determined through successive rounds of genome walking and DNA sequencing using a BD Clontech Universal Genomewalker™ kit (TaKaRa, Madison, WI). Briefly, this method relies on the creation of adaptor-ligated genomic libraries, derived from several different restriction enzymes, which are used in a series of nested PCR amplifications with adaptor-specific and gene-specific primers (reviewed by Leoni *et al.* 2011). USDA accession PI 239977 (originating from Aguas de Moura, Portugal; Table 1) was used for all genome walking; this accession possesses both cyanogenesis genes. Genomic DNA was extracted from leaf tissue using a modified CTAB protocol (described by Olsen *et al.* 2007). Three restriction enzymes, *Eco*R V, *Dra* I and *Ssp* I, were used to create genomic libraries, which, following adaptor ligation (*ExTaq* polymerase, TaKaRa), were used in successive rounds of PCR, cloning and sequencing, to determine sequences of overlapping contigs at increasing distances from the cyanogenesis loci. DNA sequencing was performed using BigDye reactions run on an ABI 3130 sequencer in the Biology Department of Washington University. A minimum of three clones was sequenced per PCR product; singletons observed in a single clone were treated as polymerase artefacts and eliminated. This yielded one unambiguous sequence haplotype per PCR product. Primers used in genome walking (136 in total) are listed in Tables S2 and S3 (Supporting information). Creation of contigs, sequence alignments and editing were performed using BioLign 4.0.6 (Hall 2001).

Analysis of proximal flanking sequences

As flanking genomic sequences were determined, primer pairs were designed for PCR assays to test for the presence/absence of portions of flanking sequences at increasing distances from the cyanogenesis loci (Fig. 2; Table S3). Ten to twenty accessions, representing genotypes with and without a given cyanogenesis gene, were used in initial PCR assays. PCR employed GoTaq polymerase (Promega) in 20 μl reactions with standard reaction conditions; annealing temperatures were adjusted for primer combinations (Table S3). Negative PCR results were repeated at least three times and under a range of optimizing conditions to confirm the absence of a product. Successful amplification of a targeted PCR product in most or all samples was taken as evidence that the flanking sequence was located beyond the boundary of a cyanogenesis gene deletion. The identity of PCR products was confirmed by sequencing 1–3 cloned PCR products in a subset of accessions.

Because upstream (5′) genome walking was hindered at both *CYP79D15* and *Li* by the occurrence of tandemly repeated genomic sequences (see Results), flanking sequence analyses focused primarily on downstream (3′) regions for both loci. Based on the predominant cyanogenesis gene deletion boundary detected at each locus by PCR screening, the downstream sequences closest to the deletion junction were targeted for PCR amplification in the entire white clover sample set (65 accessions), representing both gene-presence and gene-absence genotypes for both cyanogenesis genes (Table 1); a subset of these accessions was used in DNA sequencing (Tables 1 and S1, Supporting information). For *CYP79D15*, a 1.14-kb region located approximately 2.34 kb downstream of the gene's stop codon was targeted (referred to hereafter as *3CYP-2.34*; Fig. 2a; Table S3, Supporting information); this region was sequenced in 40 accessions, representing 23 accessions possessing *CYP79D15* and 17 accessions without it. For *Li*, a 0.9-kb region located approximately 6.65 kb downstream of the stop codon was targeted (referred to hereafter as *3Li-6.65*; Fig. 2b; Table S3); it was sequenced in 39 accessions, representing 23 accessions with the *Li* gene and 16 accessions without it. As with DNA sequencing for genome walking, at least three clones per PCR product were sequenced to eliminate artefacts of polymerase error.

Sequence alignments were exported as Nexus files to DnaSP 5.0 (Librado & Rozas 2009) for most population genetic analyses. Previously published data sets from three unlinked, neutrally evolving gene regions (*ACO1*, *ALDP* and *ZIP*; Olsen *et al.* 2007) were used as reference comparisons for the cyanogenesis gene regions. Nucleotide diversity was estimated as π (Nei 1987) and θ_W (Watterson 1975), and Hudson's (1987) recombination parameter (*R*) was calculated between adjacent sites using the formula in DnaSP. Frequency spectrum-based tests of selection were performed to test for positive deviations from neutral equilibrium, which would

Table 1 White clover accessions used in analyses. Accessions with prefix PI were obtained from the USDA germplasm collection; accessions with other prefixes were collected from naturalized North American populations. Plants producing both cyanogenic glucosides and linamarase are cyanogenic.

Accession	Origin	Cyanogenic phenotype	*Ac/ac* phenotype	*Li/li* phenotype
PI 100247[a]	New Zealand	No	*Ac+*	*li–**
PI 195534[ab]	Italy	No	*ac–*	*li–*
PI 200372[b]	Israel	Yes	*Ac+*	*Li+*
PI 204930	Turkey	Yes	*Ac+*	*Li+*
PI 205062[c]	Turkey	No	*Ac+*	*li–**
PI 208730[abc]	Italy	No	*ac–*	*li–*
PI 214207[ab]	Israel	Yes	*Ac+*	*Li+*
PI 217444	Italy	Yes	*Ac+*	*Li+*
PI 221961[ab]	Afghanistan	No	*Ac+*	*li–*
PI 226996[ab]	Uruguay	Yes	*Ac+*	*Li+*
PI 230183	Argentina	Yes	*Ac+*	*Li+*
PI 232109[a]	Germany	No	*ac–*	*li–**
PI 234678	France	Yes	*Ac+*	*Li+*
PI 239977[abc]	Portugal	Yes	*Ac+*	*Li+*
PI 246751	Spain	Yes	*Ac+*	*Li+*
PI 251053[abc]	Macedonia	No	*ac–*	*li–*
PI 251190[ac]	Serbia	No	*ac–*	*li–**
PI 251191[abc]	Montenegro	No	*ac–*	*li–*
PI 251197[ab]	Bosnia-Herz.	No	*ac–*	*li–*
PI 253323[a]	Slovenia	No	*ac–*	*Li+*
PI 260646[c]	Greece	Yes	*Ac+*	*Li+*
PI 282378[a]	Italy	No	*ac–*	*li–**
PI 291828[b]	Chile	Yes	*Ac+*	*Li+*
PI 294546[c]	France	Yes	*Ac+*	*Li+*
PI 298485[b]	Israel	Yes	*Ac+*	*Li+*
PI 302441[ab]	Australia	No	*Ac+*	*li–*
PI 311490[a]	Spain	Yes	*Ac+*	*Li+*
PI 311494[bc]	Spain	Yes	*Ac+*	*Li+*
PI 315542[ab]	Russia	No	*Ac+*	*li–*
PI 345529[a]	Australia	Yes	*Ac+*	*Li+*
PI 350706	Australia	Yes	*Ac+*	*Li+*
PI 384699[bc]	Morocco	Yes	*Ac+*	*Li+*
PI 418905	Italy	Yes	*Ac+*	*Li+*
PI 418911	Italy	Yes	*Ac+*	*Li+*
PI 419314[ab]	Greece	No	*ac–*	*li–*
PI 419316[abc]	Greece	Yes	*Ac+*	*Li+*
PI 419401[ab]	Greece	Yes	*Ac+*	*Li+*
PI 420001[b]	Japan	Yes	*Ac+*	*Li+*
PI 440745[c]	Russia	No	*Ac+*	*li–**
PI 440746[b]	Russia	No	*Ac+*	*li–*
PI 494747[c]	Kazakhstan	No	*ac–**	*li–**
PI 499685[a]	China	No	*Ac+*	*li–**
PI 499688[bc]	China	No	*Ac+*	*li–*
PI 516411[bc]	Romania	No	*ac–**	*li–*
PI 517126	Morocco	Yes	*Ac+*	*Li+*
PI 517515[b]	Ethiopia	Yes	*Ac+*	*Li+*
PI 542904[ab]	Croatia	No	*ac–*	*li–*
PI 542905	Croatia	No	*ac–*	*Li+*
PI 542915[a]	Bosnia-Herz.	No	*ac–*	*li–*
PI 556991[ab]	USA	No	*ac–*	*li–*
PI 597530[ab]	Lithuania	No	*Ac+*	*li–*
LA_0410[ab]	USA (Louisiana)	No	*ac–*	*Li+*
MSG_0307[ab]	USA (Mississippi)	No	*ac–*	*li–*
MSG_0507[ab]	USA (Mississippi)	Yes	*Ac+*	*Li+*
MSG_0601[ab]	USA (Mississippi)	Yes	*Ac+*	*Li+*
MSG_1009[a]	USA (Mississippi)	No	*Ac+*	*li–**
MSG_1106[ab]	USA (Mississippi)	No	*ac–*	*Li+*
MSG_1108[ab]	USA (Mississippi)	Yes	*Ac+*	*Li+*
MSJ_0710[ab]	USA (Mississippi)	Yes	*Ac+*	*Li+*
STL_1904[ab]	USA (Missouri)	Yes	*Ac+*	*Li+*
STL_2406[a]	USA (Missouri)	No	*ac–*	*li–**
STL_2604[ab]	USA (Missouri)	Yes	*Ac+*	*Li+*
STL_HG_ F101[ab]	USA (Missouri)	Yes	*Ac+*	*Li+*
TN_0506[ab]	USA (Tennessee)	Yes	*Ac+*	*Li+*
WI_1007[ab]	USA (Wisconsin)	Yes	*Ac+*	*Li+*

Asterisks indicate *ac–* or *li–* accessions that did not PCR-amplify *3CYP-2.34* and *3Li-6.65*, respectively; all other *ac–* and *li–* accessions are inferred to have gene deletion junctions immediately upstream of *3CYP-2.34* and *3Li-6.65*, respectively (see Fig. 2). Superscript letters indicate accessions used in DNA sequencing and Southern hybridizations.

[a] Accessions used in sequencing *3CYP-2.34*.

[b] Accessions used in sequencing *3Li-6.65*.

[c] Accessions used in *3Li-9.16* Southern hybridizations.

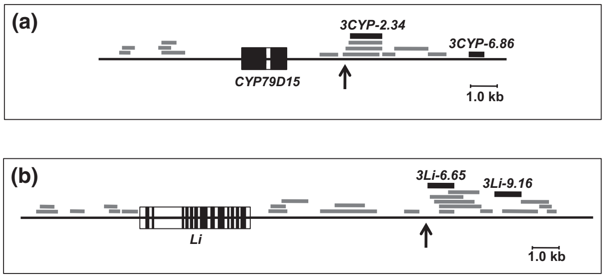

Fig. 2 Characterization of genomic regions around the white clover cyanogenesis genes. Diagrams are approximately to scale. Black- and white-boxed regions correspond to exons and introns, respectively, of: (a) the *Ac* (*CYP79D15*) gene and (b) the *Li* gene. Gray bars indicate approximate locations and sizes of upstream and downstream flanking regions targeted by PCR to determine boundaries of the cyanogenesis gene deletions; each gray bar corresponds to a primer combination indicated in Table S3 (Supporting information). All PCR-targeted upstream regions could only be amplified in plants that possess the neighbouring cyanogenesis gene. For downstream regions, black bars labelled *3CYP-2.34* and *3Li-6.65* indicate the respective locations of the closest regions that could be successfully PCR-amplified in *ac–* and *li–* plants. Arrows indicate the corresponding inferred deletion junction boundaries; these characterize most *ac–* and *li–* accessions (90% and 63%, respectively; see Results). The black bar labelled *3CYP-6.86* indicates the location of the closest downstream region that could be amplified in *ac–* plants that do not amplify the *3CYP-2.34* region; the black bar labelled *3Li-9.16* indicates the location of the region targeted for PCR and Southern blotting to confirm that this genomic region is absent in *li–* plants that do not amplify *3Li-6.65*.

be consistent with balancing selection (Tajima 1989; Fu & Li 1993; Fay & Wu 2000). Potential signatures of balancing selection were also examined by testing for elevated linkage disequilibrium (LD) between pairs of segregating sites, using the B and Q statistics of Wall (1999). Statistical significance for all of these tests was assessed through coalescent simulations in DnaSP (1000 runs), with θ_W and R values in the simulations estimated empirically from the data. For tests of selection requiring an outgroup, sequences from *Trifolium isthmocarpum* were employed (USDA accessions PI 203664 and PI 535571); this is a closely related species that possesses both cyanogenesis genes. Because high recombination and associated LD decay can lead to the erosion of selection signatures in genomic regions around targets of selection (see Charlesworth 2006), we also repeated these tests of selection using only the 5′ half of the *3CYP-2.34* and *3Li-6.65* loci (i.e. sequences most closely linked to the gene deletion junctions). To directly assess the rates at which LD would be expected to decline in these flanking genomic regions, we calculated between-site LD across each gene using Lewontin's (1964) D' measure in DnaSP.

As an additional test for deviations from neutrality, we used Hey's *HKA* program (http://genfaculty.rutgers.edu/hey/software) to perform a multilocus HKA test (Hudson *et al.* 1987) with *3CYP-2.34*, *3Li-6.65* and the three reference genes. In this test, loci evolving under balancing selection would be expected to show an excess of within-species polymorphism relative to fixed differences between species. Within-species polymorphism was measured using the same *T. repens* accessions as used in other tests of selection, and between-species differences were calculated using *T. isthmocarpum*. Statistical significance was assessed using the X^2 statistic of Hudson *et al.* (1987), which approximates a chi-square distribution; values for the chi-square distribution were generated through 1000 coalescent simulations of the data in *HKA*.

To assess whether genealogical relationships between haplotypes were consistent with long-term evolutionary divergence between gene-presence and gene-absence allele classes, haplotype trees were inferred using maximum-likelihood (ML) analyses, with the best-fit model of nucleotide substitution selected in jModelTest 0.1.1 (Guindon & Gascuel 2003; Posada 2008) based on likelihood scores for 88 different models and the Akaike Information Criterion. The GTR model of molecular evolution with rate variation among sites was employed for both of the sequence data sets based on jModelTest results. ML trees were generated in PhyML 3.0 (Guindon *et al.* 2010) via the ATGC web platform (http://www.atgc-montpellier.fr/phyml/) with default settings for tree searching and bootstrap analysis. Trees were rooted using *T. isthmocarpum* sequences.

Extended analysis of neighbouring genomic sequences

Because the closest targeted flanking regions (*3CYP-2.34* and *3Li-6.65*) did not universally amplify in all plants, we extended the PCR surveys further downstream to test for evidence of larger genomic deletions in some *ac*− and *li*− accessions (i.e. accessions lacking the cyanogenic components; see Table S3 for primer combinations and targeted regions). For the *CYP79D15* region, PCR screening was extended as far as a region 6.86 kb downstream of the stop codon (Table S3), a distance at which a ~600-bp PCR product could be successfully amplified in all surveyed accessions. For a subset of 13 accessions (seven *Ac*+ plants, six *ac*− plants), this PCR product (hereafter *3CYP-6.86*) was cloned and sequenced to confirm sequence identity across all individuals.

For *Li*, PCR screening was extended as far as 11.08 kb downstream, a distance at which the *li*− accessions that did not amplify *3Li-6.65* still did not yield a PCR product (see Results). PCR screening results at 11.08 kb downstream were the same as those targeting a 950-bp region located 9.16 kb downstream of *Li*; PCR products for the more proximal region (hereafter referred to as *3Li-9.16*; Fig. 2b) were sequenced in one or more clones of 33 accessions to confirm the sequence identity (Table S1). To further investigate whether *3Li-9.16* PCR results were reflecting the presence/absence of this genomic region, we probed for the region using Southern hybridizations of genomic DNA. For 15 accessions, genomic DNA was digested with *Afl* III, a restriction enzyme that is not predicted to cut within the targeted genomic region. A DIG-labelled probe (Roche, Indianapolis) was generated using the *3Li-9.16* PCR primers and a plasmid containing the *3Li-9.16* sequence (derived from accession PI 239977). Protocols for probe generation and the *3Li-9.16* Southern hybridization followed those used previously with the *Li* and *CYP79D15* loci (Olsen *et al.* 2007, 2008). On a Southern blot, the absence of a band corresponding to the probed *3Li-9.16* region would provide confirmation that the gene sequence is not present in a plant's genome.

Results

Determination of sequences flanking the cyanogenesis genes

Genome walking is a PCR-based approach that can be used to determine the DNA sequences that flank a gene of interest (see Materials and methods; reviewed by Leoni *et al.* 2011). For molecular ecology studies in non-model species, this technique can be especially valuable for identifying sequences in genomic regions that may show signatures of selection associated with adaptive variation. In this study, successive rounds of genome walking, using three adaptor-ligated genomic DNA restriction libraries, were employed to determine DNA sequences upstream and downstream of the *CYP79D15* and *Li* genes. For *CYP79D15*, two rounds of upstream genome walking and five rounds of downstream genome walking yielded approximately 3.4 and 7.4 kb of unambiguous upstream and downstream sequences, respectively (Fig. 2a). For *Li*, six rounds of upstream genome walking and eight rounds of downstream genome walking yielded approximately 2.9 and 13.5 kb of upstream and downstream sequences, respectively (Fig. 2b). For both genes, further upstream genome walking was impeded by the occurrence of genomic sequences that were identical or nearly identical to those already determined, which prevented the design of new, unique PCR primers for further genome walking. From the sequences obtained, we could infer that the *CYP79D15* upstream repeat (starting approximately 3.4 kb upstream of the gene) has a minimum length of 1.2 kb; the *Li* upstream repeat (starting approximately 2.9 kb upstream of the gene) has a minimum length of 0.8 kb. GenBank BLASTs of the repeated sequences did not reveal any similarity to known genes in any organism. While the nature of the genome walking approach is such that we cannot fully characterize the span of these repeats from the available sequences, we can nonetheless conclude that both cyanogenesis gene upstream regions are characterized by large tandem sequence repeats.

Localization of gene deletion junctions

To determine the boundaries of gene deletions in plants lacking cyanogenesis genes, we performed PCR screening assays in a sample of 65 white clover accessions, targeting portions of flanking sequences at increasing distances from the cyanogenesis loci (Fig. 2; Tables S1 and S3, Supporting information). For the upstream sequences of both *CYP79D15* and *Li*, the presence/absence of amplicons in all PCR-targeted regions corresponded to the cyanogenesis gene P/A variation, indicating that our attempts to genome walk beyond the upstream boundaries of the deletions were unsuccessful for both loci.

In contrast, for downstream regions, the inferred boundaries of the gene deletions could be identified for most accessions. For *CYP79D15*, a 1.14-kb region starting 2.34 kb downstream of the stop codon (*3CYP-2.34*; Fig. 2a) could be amplified in all but two *ac*− accessions of 20 surveyed (PI 494747 and PI 516411; Table 1). Based on additional PCR screening in the vicinity of *3CYP-2.34* (see Fig. 2a and Table S3, Supporting information), we were able to narrow down the location of the deletion junction to within a 60-bp window immediately upstream of *3CYP-2.34*. Thus, the *3CYP-2.34* region falls within 60 bp of the gene deletion boundary observed in 90% of *ac*− accessions examined. For the two *ac*− accessions that did not amplify the *3CYP-2.34* region, extension of targeted PCR screening progressively further downstream yielded successful amplifications once we had reached a distance of 6.86 kb from the gene (fragment *3CYP_6.86*; Fig. 2a, Table S3, Supporting information). DNA sequencing of cloned *3CYP_6.86* PCR products in a sample of 13 accessions (Table S1, Supporting information) confirmed the identity of these downstream flanking sequences. Together, these patterns suggest that the downstream boundary of the *Ac* gene deletion occurs approximately 2.34 kb away from the *CYP79D15* locus in most *ac*− plants, but with larger deletions apparently occurring in some accessions.

For *Li*, a 0.9-kb region located 6.65 downstream of the stop codon (*3Li-6.65)* could be amplified and sequenced in 17 of 27 *li-* accessions surveyed, whereas the remaining 10 *li-* accessions (37%) failed to yield a PCR product in this region (Table 1). For the *li-* accessions for which *3Li-6.65* could be amplified, additional PCR screening around this region localized the deletion junction to within a 325-bp window adjacent to *3Li-6.65*. Because of difficulties in amplifying PCR products in this 325-bp region (in both *Li*+ and *li*− plants), we were unable to localize the gene deletion boundary any further within this window. On the basis of these observations, we could conclude that many, but not all, *li*− alleles are characterized by a gene deletion junction occurring within 325 bp upstream of *3Li-6.65.*

For those *li*− accessions that did not amplify *3Li-6.65,* PCR surveys further downstream revealed that these accessions also did not amplify any products within the limits of the known downstream sequence (up to 11.1 kb downstream; see Table S3, Supporting information). To test whether the absence of PCR products in these more distal regions was reflecting a genomic deletion, we performed Southern hybridizations where we probed for a 0.95-kb sequence located 9.16 kb downstream of the *Li* stop codon (the *3Li-9.16* region; Fig. 2). The Southern blotting indicates that a genomic fragment is present in all plants for which *3Li-9.16* can be amplified and is absent in plants with no PCR product (Fig. 3). This pattern confirms that the absence of *3Li-9.16* PCR products corresponds to a deletion of this genomic region. Taken together with the *3Li-6.65* PCR

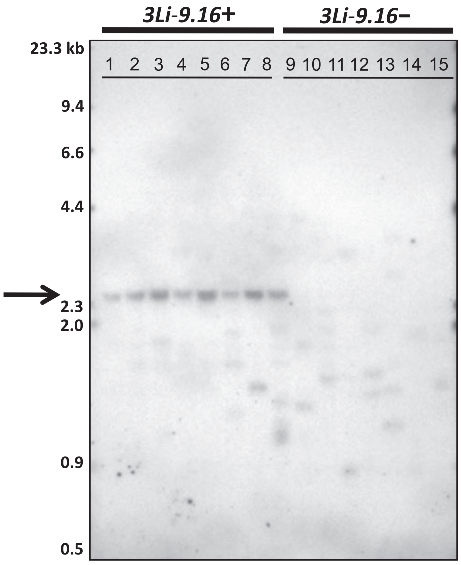

Fig. 3 Presence/absence of the *3Li-9.16* genomic region corresponds to PCR amplification results. Southern hybridizations using the *3Li-9.16* genomic sequence as a probe in white clover accessions where the *3Li-9.16* sequence can be successfully PCR-amplified (lanes 1–8) and where it cannot (lanes 9–15). The arrow indicates the location of a genomic fragment detected only in plants that amplify the *3Li-9.16* sequence. Accessions (lanes 1–15): PI 239977 (*Li*+), PI 419316 (*Li*+), PI 384699 (*Li*+), PI 499688 (*li*−), PI 298485 (*Li*+), PI 208730 (*li*−), PI 251191 (*li*−), PI 260646 (*Li*+), PI 311494 (*Li*+), PI 205062 (*li*−), PI 440745 (*li*−), PI 494747 (*li*−), PI 251053 (*li*−), PI 251190 (*li*−), PI 516441 (*li*−).

screening and sequencing results, these observations suggest that there is variability in the size of the *Li* genomic deletions, with some accessions characterized by one or more deletions that apparently extend beyond the limits of the genomic region surveyed. Unexpectedly, we also observed that the *3Li-9.16* region is apparently absent in some plants that possess the *3Li-6.65* region (e.g. accession PI 311494, Fig. 3; see also Table S1, Supporting information); this potentially suggests a more complex pattern of genomic deletions in the *Li* downstream region than would be predicted by a simple P/A polymorphism at the cyanogenesis locus.

No signatures of balancing selection

A hallmark of long-term balancing selection is a pattern of elevated nucleotide diversity at the selected locus in comparison with unlinked, neutrally evolving genes. For P/A polymorphisms under balancing selection, this pattern will be detectable in the genomic region adjacent to the boundary of the P/A polymorphism, to the extent that this flanking region is in linkage disequilibrium with the target of selection (Stahl *et al.* 1999; Tian *et al.* 2002; Shen *et al.* 2006). For both *3CYP-2.34* and *3Li-6.65*, which immediately flank the common *ac*− and *li*− gene deletion junctions, no pattern of elevated nucleotide diversity is evident in comparisons with three unlinked, neutral gene regions (Table 2). In addition, no significantly positive deviations from neutral equilibrium were detected using frequency spectrum-based tests of selection (Tajima 1989; Fu & Li 1993; Fay & Wu 2000). Similarly, tests for elevated between-site LD revealed no significant signatures of balancing selection (Wall's B and Q values, Table 2).

To examine the possibility that signatures of selection might be evident only in the portions of the flanking regions most tightly linked to the P/A polymorphisms, we repeated all of these tests of selection using only the 5′ half of each flanking gene region; however, the results are unchanged (Table 2). Estimates of LD across each of the two flanking regions provide further evidence that the absence of signatures of selection at these loci is not an artefact of LD decay. For *3CYP-2.34, D′* values decline by approximately 5% per kb (Fig. S1a, Supporting Information); for *3Li-6.65*, the estimated rate of LD decay, while higher (approximately 20% per kb; Fig. S1b, Supporting Information), is still not so great that a signal of balancing selection at the P/A polymorphism would be eroded at this adjacent locus (which occurs within 325 bp of the *Li* deletion junction). Deviations from neutral evolution were also not detected in the more distal *3CYP-6.68* and *3Li-9.16* loci, regions for which smaller subsets of accessions were unambiguously sequenced (i.e. with ≥3 clones/accession) (data not shown).

Consistent with other observations, no evidence of non-neutral evolution was detected in a multilocus HKA test comparing *3CYP-2.34*, *3Li-6.65* and the three neutral genes ($\chi^2 = 3.759$; d.f. = 4; $P = 0.44$). One caveat with this analysis is that we assumed equal effective population sizes (Ne) for all five of the genes. However, because more than one-third of the *li*− plants surveyed were found to lack the *3Li-6.65* sequence, it is possible that Ne at this locus is substantially smaller than that of the other loci. To test whether this violation could be obscuring a selection signal in our data, we repeated the multilocus HKA test, conservatively assuming Ne of *3Li-6.65* to be one-half that of the other loci; the result of the test is unchanged, with no signature of selection detected ($\chi^2 = 3.976$, d.f. = 4, $P = 0.41$).

In addition to creating genomic islands of elevated within-species nucleotide diversity and LD, long-term balancing selection on a P/A polymorphism is also expected to generate haplotype structure in flanking regions, with gene-presence and gene-absence alleles forming two evolutionarily divergent haplotype lineages (e.g. Stahl *et al.* 1999; Tian *et al.* 2002; see Fig. 1). In contrast to previously characterized *R*-gene adaptive

Table 2 Nucleotide variation and tests of selection in genomic regions flanking the cyanogenesis genes and in three unlinked neutral genes. All tests of selection are nonsignificant ($P > 0.1$)

Genes	Approx. Length (kb)	No. sequences	No. Segregating sites	π	θ_W	R[*]	Tajima's D[†]	Fu & Li's D[‡]	Fay & Wu's H[§]	Wall's B[¶]	Wall's Q[¶]
3CYP-2.34	1.14	40	46 (18)	0.009 (0.007)	0.010 (0.008)	0.003	−0.388 (−0.418)	−0.490 (0.513)	−1.36 (0.200)	0.089 (0.177)	0.152 (0.278)
3Li-6.65	0.9	39	15 (5)	0.006 (0.004)	0.004 (0.003)	0.009	1.226 (0.811)	0.731 (0.176)	−2.76 (−0.464)	0.071 (0.000)	0.133 (0.000)
ACO1	0.56	40	36	0.019	0.016	0.012	0.696	1.12	4.12	0.177	0.257
ALDP	0.57	40	17	0.016	0.013	0.007	0.846	−0.244	0.877	0.067	0.125
ZIP	0.53	36	8	0.007	0.005	0.018	1.115	−0.473	−0.308	0.000	0.000

Numbers in parentheses are values for the 5′ half of the sequenced regions, which immediately flank the deletion junctions. Data for the three reference genes are from Olsen *et al.* (2007), with nucleotide diversity calculations based on silent sites.

[*] Hudson's (1987) recombination parameter *R*, between sites.

[†] Tajima (1989).

[‡] Fu & Li (1993).

[§] Fay & Wu (2000).

[¶] Wall (1999).

P/A polymorphisms, sequences flanking the cyanogenesis genes show no evidence of divergence between the two allele classes (Fig. 4a,b). For both *3CYP-2.34* and *3Li-6.65*, haplotype clades within the ML trees are generally well resolved, with many nodes supported by >90% bootstrap values; however, these haplotype groups show no clear correspondence to the gene-presence and gene-absence allele classes. In addition, there are many instances where identical or nearly identical haplotypes are shared among plants that have and that lack the neighbouring cyanogenesis gene (e.g. *3CYP-2.34* haplotype sharing among accessions MSJ-0710, PI 100247, PI 542904 and LA-0410; Fig. 4a); this is again inconsistent with a model of deep evolutionary divergence between the two allele classes. Phylogenetic analysis using only the 5′ half of each sequenced region yielded trees with less phylogenetic resolution but no clearer correspondence to cyanogenesis gene presence/absence phenotypes (data not shown). Together, these findings confirm the results of the statistical tests of selection, indicating that the cyanogenesis gene flanking regions are not evolving as expected under a model of long-term balancing selection.

Discussion

In genomic model organisms, molecular studies of adaptive polymorphism often start out with the identification of a particular locus that shows a signature of balancing selection. This is then followed by efforts to identify associated phenotypes and to characterize their potential adaptive significance in nature. A complementary approach is possible with the white clover cyanogenesis polymorphism. Because the ecology of this chemical defence polymorphism has been studied for more than half a century, there is a wealth of accumulated evidence that the *Ac/ac* and *Li/li* polymorphisms are truly adaptive in natural populations. This ecological foundation has allowed us to investigate how the selective maintenance of this biochemical variation is manifested at the molecular level. Interestingly, when we examine whether the underlying loci show signatures of balancing selection, we find no such signals. Sequences immediately flanking the two P/A polymorphisms show no elevated nucleotide diversity, no elevated between-site LD and no haplotype structure or other signatures that would suggest balancing selection (Table 2; Fig. 4) (Hudson *et al.* 1987; Tajima 1989; Fu & Li 1993; Wall 1999; Fay & Wu 2000). In addition, we find evidence for variation in the size of the genomic deletions that characterize *ac*− and *li*− alleles (e.g. Figs 2 and 3), a pattern that suggests that the gene-absence alleles have arisen through multiple gene deletion events. Together, these observations strongly

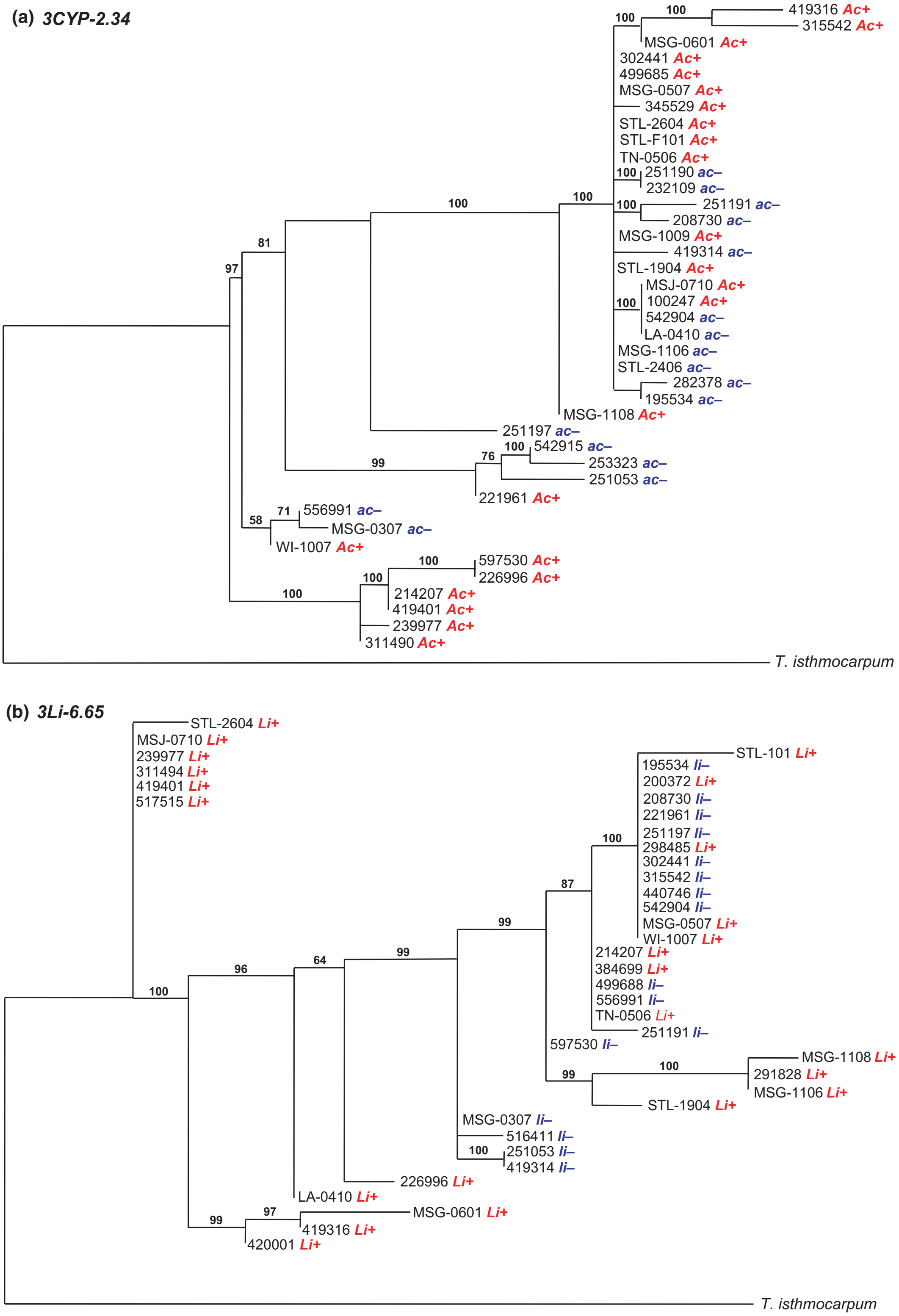

Fig. 4 Maximum-likelihood haplotype trees for sequences flanking the cyanogenesis genes. Haplotype labels correspond to accessions in Table 1. Red and blue labels indicate whether an accession has or lacks the neighbouring cyanogenesis gene. Numbers along branches indicate percentage bootstrap support; only values >50% are shown. (a) *3CYP-2.34*, located approximately 2.34 kb downstream of *CYP79D15* (*Ac*); (b) *3Li-6.65*, located approximately 6.65 kb downstream of *Li*.

suggest that the *Ac/ac* and *Li/li* biochemical polymorphisms have been evolving through recurrent gene deletions over time. Thus, while two alternative phenotypes are being actively maintained by selection at each gene (e.g. Hughes 1991; Kooyers & Olsen 2012), this is apparently occurring at the molecular level through repeated mutational events.

These findings are unexpected for at least two reasons. First, to our knowledge, all other intraspecific adaptive P/A polymorphisms where flanking sequences have been examined show signatures of long-term balancing selection (Stahl *et al.* 1999; Tian *et al.* 2002; Shen *et al.* 2006; Chen *et al.* 2010; see Fig. 1). Second, because the *Ac/ac* and *Li/li* polymorphisms are unlinked, our results suggest that there are not one but two separate regions within the *T. repens* genome where adaptive genomic deletions have apparently occurred repeatedly. This pattern suggests a surprising degree of lability in the white clover genome. Below, we discuss these findings in the context of what is currently known about the evolution of adaptive P/A polymorphisms, potential confounding effects in detecting balancing selection and the evolution of white clover cyanogenesis variation.

Molecular evolution of P/A polymorphisms

Molecular evolutionary analyses of P/A polymorphisms have focused on plant *R*-genes, specifically those genes encoding proteins characterized by nucleotide binding sites and leucine-rich repeats (NB-LRR proteins). Nearly all such studies have examined the model plant *Arabidopsis thaliana* or its relatives within the Brassicaceae. In *A. thaliana*, nine P/A polymorphisms have been examined in detail; these include *RPM1* (Stahl *et al.* 1999), *RPS5* (Tian *et al.* 2002) and seven *R*-genes identified by Shen *et al.* (2006). For all of these genes, flanking sequences are characterized by elevated nucleotide diversity, positive Tajima's *D* values and other signatures of balancing selection (reviewed by Shen *et al.* 2006; Chen *et al.* 2010; Fig. 1). Beyond *R*-genes, other loci have also been identified in *Arabidopsis* where P/A variation corresponds to natural phenotypic variation (e.g. *FLM* and flowering time variation; Werner *et al.* 2005); however, to our knowledge, the molecular evolution of the underlying allelic variation has not been studied.

It is important to note that, unlike white clover, *A. thaliana* is a self-fertilizing species characterized by high homozygosity and very low effective recombination. This mating system is expected to generate elevated LD across the genome, which would maximize the chances of detecting a signature of balancing selection in sequences near a P/A polymorphism (reviewed by Charlesworth 2006). In contrast, for a self-incompatible, widely dispersing species such as white clover, the effective recombination rate is expected to be higher; our empirical estimates of Hudson's (1987) recombination parameter (*R*) between sites are in the range of 0.003–0.018 for the five genes we examined (Table 2). Increased effective recombination could thus confine any signature of balancing selection to a relatively small genomic region most tightly linked to the P/A polymorphism, possibly on the order of several hundred base pairs from the deletion junction (see Hudson & Kaplan 1988). To explicitly test for this possibility, we repeated the statistical tests of selection using only the portions of flanking sequences most closely linked to the cyanogenesis P/A polymorphisms (the 5′ halves of *3CYP-2.34* and *3Li-6.65*, which are, respectively, located a maximum of 60 and 325 bp away from the common deletion junctions). Even these most closely linked sequences show no signatures of balancing selection or trends in that direction (Table 2). Taken together with our estimates of modest rates of LD decay in these flanking genes (Fig. S1, Supporting Information), these results suggest that the absence of signatures of balancing selection at the white clover cyanogenesis loci is not an artefact of low detection ability in this outcrossing species.

Recurrent deletions and adaptive polymorphism

While balancing selection clearly plays a key role in the evolution of many adaptive P/A polymorphisms (as well as many other forms of adaptive polymorphism; see, e.g. Charlesworth 2006), there is also some evidence of a role for recurrent deletions in P/A evolution. In particular, for the *R*-gene *RPM1*, gene-absence alleles have evolved independently in *A. thaliana* and in another mustard, *Brassica napus* (Grant *et al.* 1998). Whether recurrent deletion plays any role at these loci on a shorter evolutionary timescale (i.e. at the intraspecific level) remains unclear. For *RPM1*, no P/A polymorphism has been detected within *A. lyrata,* the sister species of *A. thaliana* (Chen *et al.* 2010). This is consistent with the findings of a genome-wide comparison of the two species, which indicates that the occurrence of a given *R*-gene P/A polymorphism in *A. thaliana* is not clearly correlated with the presence or absence of its ortholog in *A. lyrata* (Guo *et al.* 2011).

Beyond *R*-genes, recent studies on the genetic basis of adaptation have revealed other instances where recurrent gene deletion has played a role in adaptive evolution. This is most clearly documented in systems where directional selection has favoured parallel evolutionary change, either among different isolated populations of one species or in two closely related species. For example, in threespine stickleback fish (*Gasterosteus aculeatus*), the parallel evolution of freshwater-adapted populations from a marine ancestor has been accompanied by parallel deletions of the regulatory region of the *Pitx1* homeobox gene, leading to parallel reductions in pelvis development (Chan *et al.* 2010). In nematodes of the genus *Caenorhabditis*, laboratory selection experiments for adaptation to high population density have led to parallel deletions of *srg* chemoreceptor genes used in pheromone signalling; these parallel gene losses have occurred not only in separate strains of *C. elegans*, but also in the related species *C. briggsae* (McGrath *et al.* 2011). More generally, to the extent that P/A variation can be considered a form of gene copy number variation (CNV), gene deletion events represent a category of genetic variation that is increasingly being recognized as playing a fundamental role in shaping natural phenotypic variation and the process of adaptive evolution (reviewed by Schrider & Hahn 2010).

Potential mechanisms of adaptive gene deletion in white clover

For *T. repens*, the occurrence of two unlinked loci that have apparently undergone recurrent deletions suggests that there may be some feature of this species' genome that facilitates its instability. One possibility is white clover's allotetraploid evolutionary origin. A characteristic of polyploidization is the deletion of portions of the parental species' genomes; this occurs as the new genome becomes stabilized in the generations following the genome-doubling event (e.g. Feldman *et al.* 1997; reviewed by Ma & Gustafson 2005). Thus, if white clover arose through multiple polyploidization events over time, there could conceivably be multiple origins of gene-presence and gene-absence alleles derived from the independent origins of the species. At least two factors suggest that this explanation is improbable. First, only one of white clover's two diploid progenitors is known to be extant (*T. occidentale*; Hand *et al.* 2008), and this species is restricted to the Atlantic coastal cliffs of Western Europe. This would make it extremely unlikely that white clover has originated repeatedly anytime in the recent evolutionary past. In addition, *T. occidentale* as well as several other diploid relatives of white clover are themselves polymorphic for cyanogenic components (Gibson *et al.* 1972; Kakes & Chardonnens 2000), and preliminary analyses indicate that these polymorphisms also reflect P/A variation (K. Olsen, unpublished observations). This pattern suggests that polyploidy *per se* is not an inherent requirement for the evolution of the cyanogenesis gene P/A polymorphisms. The longer-term evolutionary origin(s) of cyanogenesis and cyanogenesis polymorphisms in white clover and its diploid relatives remains to be determined.

An alternative explanation for the *Ac/ac* and *Li/li* recurrent gene deletions is that the particular genomic regions where these genes occur are especially prone to deletion events. Genes located in subtelomeric regions or other genomic regions of tandemly repeated DNA are known to be especially likely to undergo repeated deletions (e.g. Kuo *et al.* 2006). Our current genetic map data do not indicate that either *Ac* or *Li* is located near a telomere (K. Olsen, unpublished observations). On the other hand, the upstream genome walking results in the present study do suggest that both genes are partially bounded by tandemly repeated DNA sequences. Tandemly repeated genomic regions are often associated with unequal crossing over events, and these can lead to deletions and duplications of the genomic regions. Thus, if the *Ac* and *Li* genes were present in tandemly repeated copies (e.g. as small gene families with copies undergoing concerted evolution), and if there were copy number variation among plants, then unequal crossing over could potentially generate gene deletion alleles from the gene-presence allele class (as well as additional copy number variation in the gene-presence allele class). If this model is correct, these gene deletion events are apparently occurring infrequently enough that they have remained undetected in the numerous crossing studies conducted over the last several decades (e.g. Corkill 1942; Hughes 1991; K. Olsen, unpublished observations).

It is also possible that more complex patterns of genomic deletion are at play in the cyanogenesis gene regions. For example, our PCR screening and Southern blotting results for the *Li* downstream region indicate that some plants lack the *3Li-9.16* genomic sequence, while possessing the more proximal *3Li-6.65* sequence (Fig. 3; Table S1, Supporting information). Genomic characterizations of pedigree populations are currently in progress to further clarify the structure of the genomic regions containing the cyanogenesis genes.

The role of cyanogenesis gene deletions in local adaptation

Ultimately, one of the greatest advantages of the white clover cyanogenesis system is its tractability for examining adaptive genetic variation in natural populations. A striking feature of cyanogenesis in this species is the pattern of climate-associated clines that have evolved repeatedly in both native and introduced populations around the world. For example, in North America, where white clover was introduced with European colonization, populations in the central United States show a steady latitudinal decline in cyanogenesis frequencies, from 86% cyanogenic plants in southern Louisiana to 11% in northern Wisconsin (Kooyers & Olsen 2012; see also Olson & Levsen 2012). The emergence of similar climate-associated clines around the world strongly suggests that this species is capable of undergoing a rapid response to selection following introduction to new environments. Given the mechanism of *Ac/ac* and *Li/li* molecular evolution suggested by the present study, the question arises as to what role recurrent gene deletions may play in this rapid evolutionary response. Work is now under way to examine whether adaptive cyanogenesis cline evolution in introduced populations occurs primarily through the sorting of introduced *Ac/ac* and *Li/li* allelic variation, or whether recurrent gene deletions may be a key factor in this adaptive evolutionary process.

Acknowledgements

The authors sincerely thank members of the Olsen laboratory group, Allison Miller (St Louis University) and Richard Moore (Miami University of Ohio) for helpful discussions and comments on the manuscript. We also thank Mike Dyer and other staff of the Washington University greenhouse for maintaining plants used in the study. Funding for this research was provided through a National Science Foundation CAREER grant to KMO (DEB-0845497) and a National Science Foundation Doctoral Dissertation Improvement Grant to NJK (DEB-1110588).

References

Armstrong HE, Armstrong EF, Horton E (1913) Herbage Studies. II.-Variation in *Lotus corniculatus* and *Trifolium repens*: (cyanophoric plants). *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **86**, 262–269.

Badr A, Sayed-Ahmed H, El-Shanshouri A, Watson LE (2002) Ancestors of white clover (*Trifolium repens* L.), as revealed by isozyme polymorphisms. *Theoretical and Applied Genetics*, **106**, 143–148.

Brighton F, Horne MT (1977) Influence of temperature on cyanogenic polymorphisms. *Nature*, **265**, 437–438.

Chan YF, Marks ME, Jones FC *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science*, **327**, 302–305.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, **2**, 379–384.

Chen QH, Han ZX, Jiang HY, Tian DC, Yang SH (2010) Strong positive selection drives rapid diversification of *R*-genes in Arabidopsis relatives. *Journal of Molecular Evolution*, **70**, 137–148.

Coop IE (1940) Cyanogenesis in white clover III. Study of Linamarase. *New Zealand Journal of Science and Technology*, **B22–23**, 71–83.

Corkill L (1942) Cyanogenesis in white clover (*Trifolium repens*) V. The inheritance of cyanogenesis. *New Zealand Journal of Science and Technology*, **23**, 178–193.

Daday H (1954a) Gene frequencies in wild populations of *Trifolium repens*. I. Distribution by latitude. *Heredity*, **8**, 61–78.

Daday H (1954b) Gene frequencies in wild populations of *Trifolium repens*. II. Distribution by altitude. *Heredity*, **8**, 377–384.

Daday H (1958) Gene frequencies in wild populations of *Trifolium repens*. III. World distribution. *Heredity*, **12**, 169–184.

Daday H (1965) Gene frequencies in wild populations of *Trifolium repens*. IV. Mechanisms of natural selection. *Heredity*, **20**, 355–365.

De Araujo AM (1976) The relationship between altitude and cyanogenesis in white clover (*Trifolium repens* L.). *Heredity*, **37**, 291–293.

Dirzo R, Harper JL (1982a) Experimental studies on slug–plant interactions. III. Differences in the acceptability of individual plants of *Trifolium repens* to slugs and snails. *Journal of Ecology*, **70**, 101–117.

Dirzo R, Harper JL (1982b) Experimental studies on slug–plant interactions. IV. The performance of cyanogenic and acyanogenic morphs of *Trifolium repens* in the field. *Journal of Ecology*, **70**, 119–138.

Ennos RA (1982) Association of the cyanogenic loci in white clover. *Genetical Research Cambridge*, **40**, 65–72.

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

Feigl F, Anger V (1966) Replacement of benzidine by copper ethylacetoacetate and tetra base as spot-test reagent for hydrogen cyanide and cyanogen. *Analyst*, **91**, 282–284.

Feldman M, Liu B, Segal G, Abbo S, Levy AA *et al.* (1997) Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics*, **147**, 1381–1387.

Fu Y-X, Li W-H (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Ganders FR (1990) Altitudinal clines for cyanogenesis in introduced populations of white clover near Vancouver, Canada. *Heredity*, **64**, 387–390.

George J, Dobrowolski MP, de Jong EV *et al.* (2006) Assessment of genetic diversity in cultivars of white clover (*Trifolium repens* L.) detected by SSR polymorphisms. *Genome*, **49**, 919–930.

Gibson PB, Barnett OW, Gillingham JT (1972) Cyanoglucoside and hydrolyzing enzyme in species related to *Trifolium repens*. *Crop Science*, **12**, 708–709.

Grant MR, McDowell JM, Sharpe AG *et al.* (1998) Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. *Proceedings of the National Academy of Sciences USA*, **95**, 15843–15848.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

Guo YL, Fitz J, Schneeberger K, Ossowski S, Cao J, Weigel D (2011) Genome-wide comparison of nucleotide-binding site-leucine-rich repeat-encoding genes in Arabidopsis. *Plant Physiology*, **157**, 757–769.

Hall T (2001) BioLign alignment and multiple contig editor. Available from http://en.bio-soft.net/dna/BioLign.html.

Hand ML, Ponting RC, Drayton MC, Lawless KA, Cogan NOI *et al.* (2008) Identification of homologous, homeologous and paralogous sequence variants in an outbreeding allopolyploid species based on comparison with progenitor taxa. *Molecular Genetics and Genomics*, **280**, 293–304.

Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. *Genetical Research (Cambridge)*, **50**, 245–250.

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.

Hughes MA (1991) The cyanogenic polymorphism in *Trifolium repens* L. (white clover). *Heredity*, **66**, 105–115.

Kakes P (1989) An analysis of the costs and benefits of the cyanogenic system in *Trifolium repens* L. *Theoretical and Applied Genetics*, **77**, 111–118.

Kakes P, Chardonnens AN (2000) Cyanotypic frequencies in adjacent and mixed populations of *Trifolium occidentale* Coombe and *Trifolium repens* L. are regulated by different mechanisms. *Biochemical Systematics and Ecology*, **28**, 633–649.

Kooyers NJ, Olsen KM (2012) Rapid evolution of an adaptive cyanogenesis cline in introduced North American white clover (*Trifolium repens* L.). *Molecular Ecology*, **21**, 2455–2468.

Kuo H-F, Olsen KM, Richards EJ (2006) Natural variation in a subtelomeric region of Arabidopsis: implications for the genomic dynamics of a chromosome end. *Genetics*, **173**, 401–417.

Leoni C, Volpicella M, De Leo F, Gallerani R, Ceci LR (2011) Genome walking in eukaryotes. *FEBS Journal*, **278**, 3953–3977.

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics*, **49**, 49–67.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Ma X-F, Gustafson JP (2005) Genome evolution of allopolyploids: a process of cytological and genetic diploidization. *Cytogenetic and Genome Research*, **109**, 236–249.

Majumdar S, De KK, Banjeree S (2004) Influence of two selective forces on cyanogenesis polymorphism of *Trifolium repen*s L. in Darjeeling Himalaya. *Journal of Plant Biology*, **47**, 124–128.

McGrath PT, Xu Y, Ailion M, Garrison JL, Butcher RA, Bargmann CI (2011) Parallel evolution of domesticated *Caenorhabditis* species targets pheromone receptor genes. *Nature*, **477**, 321–325.

Melville J, Doak BW (1940) Cyanogenesis in white clover II. Isolation of glucoside constituents. *New Zealand Journal of Science and Technology*, **22**, 67–70.

Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, **8**, 845–856.

Møller BL (2010) Functional diversifications of cyanogenic glucosides. *Current Opinion in Plant Biology*, **13**, 338–347.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York.

Olsen KM, Ungerer MC (2008) Freezing tolerance and cyanogenesis in white clover (*Trifolium repens* L., Fabaceae). *International Journal of Plant Sciences*, **169**, 1141–1147.

Olsen KM, Sutherland BL, Small LL (2007) Molecular evolution of the *Li/li* chemical defence polymorphism in white clover (*Trifolium repens* L.). *Molecular Ecology*, **16**, 4180–4193.

Olsen KM, Hsu S-C, Small LL (2008) Evidence on the molecular basis of the *Ac/ac* adaptive cyanogenesis polymorphism in white clover (*Trifolium repens* L.). *Genetics*, **179**, 517–526.

Olson MS, Levsen N (2012) Classic clover cline clues. *Molecular Ecology*, **21**, 2315–2317.

Pederson GA, Brink GE (1998) Cyanogenesis effect on insect damage to seedling white clover in a bermudagrass sod. *Agronomy Journal*, **90**, 208–210.

Pederson GA, Fairbrother TE, Greene SL (1996) Cyanogenesis and climatic relationships in U.S. white clover germplasm collection and core subset. *Crop Science*, **36**, 427–433.

Pennings SC, Silliman BR (2005) Linking biogeography and community ecology: latitudinal variation in plant-herbivore interaction strength. *Ecology*, **9**, 2310–2319.

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, **25**, 1253–1256.

Saucy F, Studer J, Aerni V, Schneiter B (1999) Preference for acyanogenic white clover (*Trifolium repens*) in the vole *Arvicola terrestris*. I. Experiments with two varieties. *Journal of Chemical Ecology*, **25**, 1441–1454.

Schrider DR, Hahn MW (2010) Gene copy-number polymorphism in nature. *Proceedings of the Royal Society, Series B, Biological Sciences*, **277**, 3213–3221.

Shen JD, Araki H, Chen LL *et al.* (2006) Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in *Arabidopsis thaliana*. *Genetics*, **172**, 1243–1250.

Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J (1999) Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. *Nature*, **400**, 667–671.

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

Tian DC, Araki H, Stahl E, Bergelson J, Kreitman M (2002) Signature of balancing selection in *Arabidopsis*. *Proceedings of the National Academy of Sciences USA*, **99**, 11525–11530.

Till-Bottraud I, Kakes P, Dommée B (1988) Variable phenotypes and stable distribution of the cyanotypes of *Trifolium repens* L. in Southern France. *Acta Oecologica*, **9**, 393–404.

Vickery PJ, Wheeler JL, Mulcahy C (1987) Factors affecting the hydrogen cyanide potential of white clover (*Trifolium repens* L.). *Australian Journal of Agricultural Research*, **38**, 1053–1059.

Viette M, Tettamanti C, Saucy F (2000) Preference for acyanogenic white clover (*Trifolium repens*) in the vole *Arvicola terrestris.* II. Generalization and further investigations. *Journal of Chemical Ecology*, **26**, 101–122.

Wall JD (1999) Recombination and the power of statistical tests of neutrality. *Genetics Research Cambridge*, **74**, 65–69.

Ware WM (1925) Experiments and observations on forms and strains of *Trifolium repens*. *Journal of Agricultural Science*, **15**, 47–67.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Werner JD, Borevitz JO, Warthmann N, Trainer GT, Ecker JR *et al.* (2005) Quantitative trait locus mapping and DNA array hybridization identify an *FLM* deletion as a cause for natural flowering-time variation. *Proceedings of the National Academy of Sciences USA*, **102**, 2460–2465.

Williams WM, Williamson ML (2001) Genetic polymorphism for cyanogenesis and linkage at the linamarase locus in

Trifolium nigrescens Viv. subsp. *nigrescens*. *Theoretical and Applied Genetics*, **103**, 1211–1215.

The authors' research interests lie in plant evolutionary biology. K.O. is an associate professor in the Biology department of Washington University in St. Louis. His laboratory group studies the genetic basis of adaptation and the genetics of crop domestication. N.K. is a PhD student at Washington University with interests in the genetics of adaptation and ecological trade-offs that maintain intraspecific polymorphisms across space and time. L.S. is a research technician in the Olsen laboratory.

Data accessibility

DNA sequences: GenBank accessions JQ920491-JQ920575; DNA sequence information and alignments: DRYAD entry doi:10.5061/dryad.ks6g0.

Supporting information

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

Fig. S1 Linkage disequilibrium decay across the *3CYP-2.34* and *3Li-6.65 loci*.

Table S1 Details on white clover accessions used in the study.

Table S2 Primers used in genome walking and DNA sequencing.

Table S3 Primer pairs used in PCR assays for presence/absence of flanking genomic regions.

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 corresponding author for the article.