New Phytologist

Research

Micro- and macroevolutionary adaptation through repeated loss of a complete metabolic pathway

Kenneth M. Olsen and Linda L. Small

Biology Department, Washington University, Campus Box 1137, St Louis, MO 63130-4899, USA

Author for correspondence:
Kenneth M. Olsen
Tel: +1 314 935 7013
Email: kolsen@wustl.edu

Received: *22 January 2018*
Accepted: *27 March 2018*

New Phytologist (2018) **219:** 757–766
doi: 10.1111/nph.15184

Key words: adaptation, chemical defense metabolites, coadapted gene complex, cyanogenic glucosides, metabolic cluster, Papilionoideae, white clover (*Trifolium repens*).

Summary

- There is growing evidence for the convergent evolution of physically linked gene clusters encoding chemical defense pathways. Metabolic clusters are proposed to evolve because they ensure co-inheritance of all required genes where the defense is favored, and prevent inheritance of toxic partial pathways where it is not. This hypothesis rests on the assumption that clusters evolve in species where selection favors intraspecific polymorphism for the defense; however, they have not been examined in polymorphic species.
- We examined metabolic cluster evolution in relation to an adaptive polymorphism for cyanogenic glucoside (CNglc) production in clover. Using 163 accessions, we performed CNglc assays, BAC sequencing, Southern hybridizations and molecular evolutionary analyses.
- We find that the CNglc pathway forms a 138-kb cluster in white clover, and that the adaptive polymorphism occurs through presence/absence of the complete cluster. Component genes are orthologous to those in the distantly related legume *Lotus japonicus*.
- These findings provide empirical support for the co-inheritance hypothesis, and they indicate that adaptive CNglc variation in white clover evolves through recurrent deletions of the entire pathway. They further indicate that the shared ancestor of many important legume crops was likely cyanogenic and that this defense was lost repeatedly over the last 50 Myr.

Introduction

The location of genes within eukaryotic genomes has historically been viewed as random with respect to gene function. In the last two decades, however, support has grown for the occurrence of operon-like 'metabolic clusters' consisting of linked groups of nonhomologous genes that constitute a single metabolic pathway (Hurst *et al.*, 2004; Osbourn, 2010; Yeaman, 2013; Nützmann *et al.*, 2016; Slot, 2017). In plants, this phenomenon has been increasingly documented for chemical defense pathways (Takos & Rook, 2012; Boycheva *et al.*, 2014). First described for a benzoxazinoid gene cluster in maize (Frey *et al.*, 1997), metabolic clusters have subsequently been detected in diverse angiosperm species for di- and triterpenes (Qi *et al.*, 2004; Field & Osbourn, 2008; Matsuba *et al.*, 2013), steroidal and isoquinoline alkaloids (Winzer *et al.*, 2012; Itkin *et al.*, 2013), polyketides (Schneider *et al.*, 2016), furanocoumarins (Roselli *et al.*, 2017) and cyanogenic glucosides (CNglcs) (Takos *et al.*, 2011; Boycheva *et al.*, 2014).

Given their recent discovery, much remains unknown about the origins and evolution of chemical defense gene clusters. One question concerns the extent to which their occurrence in different species reflects evolutionary convergence vs conservation from shared ancestors. Cross-species comparisons have generally pointed to independent origins of gene clusters in the taxa where they have been detected; this has apparently occurred through recruitment and neofunctionalization of nonorthologous ancestral genes (Field & Osbourn, 2008; Field *et al.*, 2011; Takos *et al.*, 2011). However, with few exceptions (Itkin *et al.*, 2013), available comparisons have been between species from different plant families. The role of evolutionary convergence vs conservation is largely unaddressed for phylogenetically less-diverged species, where gene clusters and/or their constituent genes might be more likely to reflect inheritance from a common ancestor.

A second unresolved question concerns why gene cluster evolution is favored for some chemical defense metabolites. Two nonmutually exclusive adaptive explanations have been proposed for gene clusters in general (Osbourn, 2010): clustering may facilitate coordinated regulation of the pathway genes (e.g. through chromatin modification of the gene cluster region); and it may ensure co-inheritance of all required pathway components. The second 'co-inheritance' explanation has been proposed to be most important for plant chemical defense metabolites (Takos & Rook, 2012). This argument is based on the fact that environmental heterogeneity often acts to maintain within-species adaptive polymorphisms for chemical defense production (reviewed in Moore *et al.*, 2014). In such cases, genomic clustering of the biosynthetic pathway genes would ensure that all required genes are present in the particular environments where the defense is favored. Equally importantly, clustering would safeguard against inheritance of a partial pathway in plants growing where it is not favored, which could lead to maladaptive genotypes that produce

toxic metabolic intermediates (Kristensen *et al.*, 2005; Mylona *et al.*, 2008; McGary *et al.*, 2013; Nützmann *et al.*, 2016). The co-inheritance hypothesis rests on the assumption of within-species adaptive polymorphism for the chemical defense, because physical linkage becomes irrelevant for co-inheritance when the defense and its underlying genes are universally favored in the species.

Although the co-inheritance hypothesis assumes within-species adaptive polymorphism, gene cluster studies to date have not examined any species where such chemical defense polymorphisms are known to exist. As described below, this study examines gene cluster evolution in a species with a well-documented adaptive chemical defense polymorphism, allowing us to test how selection on the polymorphism shapes the evolution of the biosynthetic pathway and its individual genetic components. In addition, the phylogenetic position of our study system permits assessment of the role of evolutionary convergence vs conservation in the origins of the chemical defense pathway in two members of the same plant family, providing a complement to previous cross-family comparisons.

Cyanogenesis and white clover

Cyanogenesis (hydrogen cyanide release with tissue damage) occurs in > 3000 plant species and is considered a classic chemical defense against herbivores (Møller, 2010). Several decades of ecological studies have confirmed this function in white clover (*Trifolium repens*), a species that is naturally polymorphic for the cyanogenic phenotype (Armstrong *et al.*, 1913; reviewed in Hughes, 1991; Olsen *et al.*, 2013). Populations worldwide have evolved climate-associated cyanogenesis clines, with cyanogenic plants predominating in warmer and drier locations (Daday, 1954a,b, 1958; Kooyers & Olsen, 2012, 2013). The selective factors that maintain this adaptive polymorphism are proposed to include climate-associated variation in herbivore abundance and/or abiotic stresses, as well as resource allocation trade-offs associated with the energetic costs of producing the cyanogenic components (Kooyers & Olsen, 2012, 2013; Kooyers *et al.*, 2014; Thompson *et al.*, 2016).

Cyanogenesis in white clover occurs through the interaction of two biochemical components that are separated in intact tissue: cyanogenic glucosides (CNglcs) (stored in the vacuoles of photosynthetic tissue), and their hydrolyzing enzyme linamarase (present in the apoplast) (reviewed in Hughes, 1991; Olsen *et al.*, 2013). The polymorphism reflects two independently segregating Mendelian polymorphisms for the presence/absence of each component (Coop, 1940; Melville & Doak, 1940; Corkill, 1942); the genes *Ac/ac* and *Li/li* control the presence/absence of CNglcs and linamarase, respectively, with the dominant allele of each gene conferring the presence of the component.

In previous work, we determined that at the molecular level, the *Ac/ac* and *Li/li* polymorphisms correspond to two unlinked gene presence/absence polymorphisms (Olsen *et al.*, 2007, 2008). For the *Ac/ac* (CNglc) polymorphism, we observed that the recessive *ac* allele corresponds to a deletion of *CYP79D15*, the cytochrome P450 that catalyzes the first dedicated step in the three-step CNglc biosynthesis pathway (Olsen *et al.*, 2008). Molecular evolutionary studies further indicated that both the *Ac/ac* and *Li/li* adaptive polymorphisms have evolved through selection favoring recurrently evolving gene deletions, and that this has occurred both in white clover (Olsen *et al.*, 2013; Kooyers & Olsen, 2014) and in several related *Trifolium* species (Olsen *et al.*, 2014). However, all molecular studies of the *Ac/ac* (CNglc) polymorphism to date have focused solely on *CYP79D15*, as the genes encoding the second and third metabolic steps have remained unidentified in clover.

Research on other cyanogenic plants indicates that the CNglc pathway is organized as gene clusters in species where the component genes have been identified: sorghum (*Sorghum bicolor*, Poaceae), cassava (*Manihot esculenta*, Euphorbiaceae) and birdsfoot trefoil (*Lotus japonicus,* Fabaceae) (Takos *et al.*, 2011). There is no recognizable conservation among these clusters in nucleotide sequence, gene order within the cluster, or transcript orientation. Thus, CNglc clusters appear to have evolved convergently in the three plant families where they have been detected so far.

Because *L. japonicus* and white clover are both papilionoid legumes, with a most recent common ancestor estimated at *c.* 50 Myr ago (Lavin *et al.*, 2005), they provide an opportunity to examine whether CNglc gene clusters also have evolved convergently within a single plant family. Moreover, the white clover *Ac/ac* polymorphism provides an opportunity to test the co-inheritance hypothesis for gene cluster evolution in a species with a well-documented adaptive polymorphism. In principle, the previously documented *CYP79D15* presence/absence polymorphism should alone be sufficient to control the CNglc polymorphism, because the pathway cannot proceed without this first committed step. Thus, if co-inheritance is *not* a critical factor for the evolution of this chemical defense polymorphism, one would predict that the CNglc polymorphism should involve only the *CYP79D15* locus, as this would be the simplest genetic mechanism. Alternatively, if co-inheritance *is* critical for the evolution of chemical defense polymorphisms, one would predict that all three genes in the pathway should be physically linked and collectively involved in the adaptive polymorphism.

In this study, we have identified and characterized the genes encoding the second and third steps in the white clover CNglc pathway, and we use these data to address the following questions: (1) are the component genes orthologous to those in the legume *L. japonicus*, or have they evolved convergently as with other CNglc pathways? (2) Do the white clover CNglc genes form a metabolic cluster? If so, what role, if any, do genes other than the first gene in the pathway play in the CNglc adaptive polymorphism? Our results reveal that all three genes are orthologous to those in *L. japonicus,* indicating that cyanogenesis was likely present in their common ancestor and has been repeatedly lost in the papilionoid legumes over the last 50 Myr. In addition, we find that the white clover CNglc genes form a metabolic cluster as in other cyanogenic species, and that, consistent with predictions of the co-inheritance hypothesis, the CNglc polymorphism arises through a genomic presence/absence polymorphism for the complete cluster. Considered in the context of

previous ecological studies of the cyanogenesis polymorphism, these results further indicate that climatic adaptation in clover has occurred through selection favoring recurrent loss of the complete CNglc metabolic pathway.

Materials and Methods

Sampling and CNglc phenotyping

Accessions of white clover *Trifolium repens* L. ($N = 81$) and related *Trifolium* species ($N = 82$) were sampled for analyses (Supporting Information Tables S1, S2). For white clover, wild ecotype accessions were obtained either as seed from collections of the USDA Germplasm Resources Information Network (GRIN), or as stolon cuttings sampled from naturalized North American populations as described previously (Kooyers & Olsen, 2012). Other *Trifolium* species accessions were obtained as seed from GRIN or other sources listed in Table S2. Seedlings or stolon cuttings were established on mist benches, and plants were maintained under standard conditions in the Washington University glasshouse. For each accession, cyanogenic glucoside (CNglc) presence/absence was determined using fresh leaf tissue that was frozen to rupture cells and then thawed and incubated with exogenous linamarase in the presence of Feigl–Anger HCN test paper, as described previously (Olsen *et al.*, 2007).

Genetic analyses

Gene presence/absence data for *CYP79D15* were obtained from previous publications (Olsen *et al.*, 2008, 2014). PCR amplification of clover candidate genes for the second and third steps in CNglc synthesis (referred to below as C*YP736A187* and *UGT85K17*, respectively) initially used *Lotus japonicus* primer sequences (Takos *et al.*, 2011), after which clover-specific primers were designed and used in all analyses (Table S3). PCR to determine presence/absence of *CYP736A187* and *UGT85K17* followed the PCR protocol and reaction conditions described previously by Olsen *et al.* (2008) for *CYP79D15*. PCR assays for *CYP736A187* presence/absence relied on the primer combination CYP736A2_03Fb and CYP736A2_03Rb for most accessions; in cases of weak or ambiguous PCR results, alternative primer combinations were used from among those listed in Table S3. Likewise, PCR assays for *UGT85K17* relied on the primer combination UGT85_04Fb and UGT85_04Ra for most accessions, with alternative combinations used as needed.

In order to confirm that PCR genotyping results accurately indicated gene presence/absence variation, genomic digests of a subset of accessions were probed using Southern hybridizations. Protocols for Southern hybridizations of *CYP736A187* and *UGT85K17* followed those described previously for detecting genomic presence/absence of *CYP79D15* (see Olsen *et al.*, 2008, 2014). The following primer combinations were used to amplify DIG-labeled probes for the two genes: CYP736A2_01Fi + CYP736A2_01Ri and UGT85_03Fi + LjUGT85K2_R. Genomic DNA digests followed the restriction enzyme manufacturer's protocols (NEB, Ipswich, MA, USA) and were performed using *Ase*I for most hybridizations; *Afl*II was also used for a subset of white clover accessions (see Table S1).

The genomic sequence of the CNglc cluster region was determined by sequencing pooled clones from a BAC library created from white clover USDA accession PI 239977 (Table S1). BAC library construction, identification of clones that were positive for CNglc genes, and sequence assembly from pooled sequencing were performed by the Clemson University Genomics Institute (CUGI). Screening of the complete BAC library revealed 12 BACs that were positive for one or more of the CNglc genes, and for these clones 250 × 250 Illumina MiSeq mate pair read Nextera libraries were generated with a jump distance of 8 kb. This was followed by a round of linear paired-end 250 × 250 Illumina MiSeq (jump distance of 560 bp), and finally rounds of Sanger sequencing for gap filling in repetitive regions.

Comparison of the white clover and *L. japonicus* CNglc gene clusters was performed using GEvo (Lyons & Freeling, 2008) in the CoGe web platform (https://genomevolution.org/coge/GEvo.pl) with default parameter settings. This was followed by BLAST analysis of regions of interest in each cluster against the NCBI nucleotide database and the Kazusa *L. japonicus* genome browser (http://www.kazusa.or.jp/lotus/). To assess whether levels of nucleotide divergence between white clover and *L. japonicus* for the three CNglc genes are consistent with levels of divergence at other orthologous loci across the genome, we searched for molecularly characterized protein-coding white clover genes in the NCBI nucleotide database and used BLAST analysis against the *L. japonicus* reference genome to identify uniquely matching loci with high sequence similarity. This yielded a set of six reference nuclear genes (*Actin*, *Adh-1*, *EF1-α*, *GAPDH*, *Ser/Thr protein kinase* and *rbcS*) that were used as representative orthologs for the two species. Levels of nucleotide and amino acid divergence at these orthologous loci provided a set of reference values for comparisons of three white clover CNglc genes to their putative orthologs in *L. japonicus*.

For molecular population genetic analyses, Sanger sequencing was performed in the Washington University Biology Department on a subset of *CYP736A187* and *UGT85K17* PCR products from the gene presence/absence genotyping assays (Tables S1, S2). DNA sequencing followed the protocol described previously for *CYP79D15* (Olsen *et al.*, 2008), with a minimum of three cloned PCR products sequenced per gene per accession. Singleton mutations present in individual clones were disregarded as potential PCR artifacts, yielding one unambiguous haplotype per accession. Sequences were edited and aligned using BioLign (Hall, 2001) or GENEIOUS v.8.0 (Kearse *et al.*, 2012). For white clover sequences, nucleotide diversity/divergence analyses and tests for deviations from neutral equilibrium were performed in DNASP v.5 (Librado & Rozas, 2009). To confirm orthology between CNglc gene sequences identified in white clover and genes amplified in related *Trifolium* species, PCR products were cloned and sequenced for *CYP736A187* and *UGT85K17* in 1–6 accessions apiece for seven *Trifolium* species where *CYP79D15* had been sequenced previously; sampling included each of the three evolutionarily diverged subspecies that are recognized within *Trifolium nigrescens* (Williams *et al.*, 2001). All new DNA sequences have

been deposited in the NCBI nucleotide database under GenBank accession numbers MH059811-MH059954.

Results

Identification of the CNglc pathway genes in white clover

CNglc synthesis proceeds in a three-step pathway in all plant species studied to date (Gleadow & Møller, 2014). The first two steps are catalyzed by cytochrome P450s, which convert amino acid precursors into hydroxynitrile intermediates (oximes followed by cyanohydrins). This is followed by a glycosylation step catalyzed by a UDP glucosyltransferase. Along with several other cyanogenic species (including *L. japonicus*, cassava, flax and rubber tree), white clover produces two chemically closely related CNglcs, linamarin and lotaustralin, in the same pathway; these are derived from the amino acids L-valine and L-isoleucine, respectively (Gleadow & Møller, 2014). As the gene encoding the first dedicated step, *CYP79D15* can be considered the signature gene of the pathway (Olsen *et al.*, 2008; Osbourn, 2010; Nützmann *et al.*, 2016). Members of the CYP79 subfamily of P450s also have been found to catalyze this first step in all cyanogenic plant species studied to date (Takos *et al.*, 2011; Gleadow & Møller, 2014). *CYP79D15* shares 92–93% coding-region sequence identity with the *L. japonicus* CNglc signature genes (the paralogs *CYP79D3* and *CYP79D4*) and is considered orthologous to them (Olsen *et al.*, 2008; Takos *et al.*, 2011). Comparison to six unlinked nuclear genes confirms that the divergence between *CYP79D15* and *CYP79D3/CYP79D4* is similar to that of other orthologs (90% average sequence identity; Table 1).

In order to identify potential candidate genes for the second and third steps, we designed PCR primers for conserved regions of the corresponding *L. japonicus* genes, *CYP736A2* and *UGT85K3* (Takos *et al.*, 2011) (Table S3). PCR amplifications consistently yielded a single amplicon per targeted gene in white clover accessions that produce CNglcs (Table S1). DNA sequencing of the PCR products revealed high nucleotide sequence similarity between each white clover gene and its *L. japonicus* counterpart (91% and 94% identity to coding regions of *CYP736A2* and *UGT85K3*, respectively); as with *CYP79D15*, these levels of sequence similarity are very similar to those of reference nuclear genes (Table 1). Following nomenclature committee guidelines for cytochrome P450s (Nelson, 2009) and UDP glucosyltransferases (Mackenzie *et al.*, 1997), the white clover genes are named and recorded in their respective gene family databases as *CYP736A187* and *UGT85K17*. The coding region of *CYP736A187* is 2.03 kb and includes a single intron of 0.53 kb; the predicted amino acid sequence shows a 90% positive match to *L. japonicus CYP736A2*. *UGT85K17* spans 1.64 kb, including a single intron of 0.19 kb, and shows a 93% positive match in predicted amino acid sequence to *L. japonicus UGT85K3*. Together these results suggest that *CYP736A187* and *UGT85K17* are white clover orthologs of the *L. japonicus* genes that encode the second and third steps in CNglc synthesis.

CNglc synthesis genes are clustered in white clover

In order to determine whether the three putative CNglc synthesis genes are clustered in white clover as in other cyanogenic species, we generated a BAC library from a cyanogenic accession (USDA PI 239977; Table S1) and sequenced the subset of 12 clones that showed positive matches to *CYP79D15*, *CYP736A187* and/or *UGT85K17*. BAC sequences reveal that the three white clover genes form a cluster spanning a 138-kb genomic region (Fig. 1). The size of this cluster falls within the range previously observed in other cyanogenic species (83 kb in cassava, 104 kb in sorghum, 166 kb in *L. japonicus*) (Takos *et al.*, 2011).

The overall structure of the two gene clusters shows little evidence of conservation between white clover and *L. japonicus*. The clusters differ with respect to both gene order and transcript orientation, and there is very little identifiable sequence conservation

Table 1 Sequence divergence between white clover (*Trifolium repens*) and *Lotus japonicus* genes for the cyanogenic glucoside (CNglc) biosynthesis pathway and for unlinked nuclear genes

	Trifolium repens	*Lotus japonicus*	Coding sequence length (kb)[b]	Nucleotide identity (exons)	Amino acid positive match
CNglc pathway					
1. *CYP79D15*		*CYP79D3* (*CYP79D4*)[a]	1.72 [1.58]	93% (92%)	93% (89%)
2. *CYP736A187*		*CYP736A2*	2.03 [1.50]	91%	90%
3. *UGT85K17*		*UGT85K3*	1.64 [1.21]	94%	93%
			Mean ± SD	93% ± 0.02 (92% ± 0.02)	92% ± 0.02 (91% ± 0.02)
Reference genes (GenBank accessions)					
Actin	AM419900	EU195536	0.35 [0.23]	96%	100%
Adh-1	X14826	AJ717414	1.27 [1.14]	90%	98%
EF1-α	KC710340	AK33796	0.63 [0.63]	89%	98%
GAPDH	JF968420	BT144712	0.30 [0.30]	88%	97%
Ser/Thr protein kinase	X99100	AB113574	1.30 [1.02]	92%	98%
rbcS	X52293	BT136865	1.26 [0.54]	83%	84%
			Mean ± SD	90% ± 0.03	96% ± 0.06

[a]Values in parentheses are for comparisons between *CYP79D15* and *CYP79D4*, a paralog of *CYP79D3*.
[b]Values in brackets are for exons only; each CNglc gene contains a single intron.

outside of the CNglc genes themselves (Fig. 1). Genomic comparison of the two clusters by GEvo analysis (Lyons & Freeling, 2008) confirmed close matches between coding regions of each white clover CNglc gene and its *L. japonicus* ortholog, as well as similarity between *CYP736A187* and previously reported ψ*CYP736A2* paralogs (Takos *et al.*, 2011). However, the only other region of sequence similarity detected was a *c.* 8 kb region located at *c.* 80 kb within the white clover cluster that showed moderate conservation (72% nucleotide identity) with a region located 20 kb upstream of the *L. japonicus* cluster (Fig. 1). BLAST analysis of this sequence against the entire NCBI nucleotide database revealed no similarity to any annotated genes or to sequences other than clones from the *L. japonicus* reference genome. Sequence conservation outside of the CNglc genes thus appears negligible between the two clusters.

In order to further examine potential sequence conservation between the white clover gene cluster and genes or genomes of other taxa, we performed a BLAST search of the cluster against the entire NCBI nucleotide database. Aside from the *L. japonicus* CNglc genes, the closest match was for a 760 bp sequence located *c.* 17.6 kb upstream of *CYP736A187*; this sequence shows 85–91% nucleotide identity to a *Ty1/copia-like* retrotransposon reported in several other legume genera. No other annotated gene sequences showed matches with > 82% nucleotide identity. We also performed a BLAST search specifically against the genome of the closest relative with a published reference genome, *Medicago truncatula* Gaertn.; this noncyanogenic species co-occurs with *Trifolium* in tribe Trifolieae. The closest hit was for a 4.5-kb genomic sequence sharing 82% nucleotide identity with a region located 2 kb upstream of *CYP79D15*; the sequence does not correspond to any expressed or annotated gene. Together these results indicate that there is little evidence of conserved gene sequences in the white clover cluster other than the three CNglc genes.

The *Ac/ac* polymorphism: presence/absence of the entire CNglc gene cluster

We previously determined that the white clover CNglc adaptive polymorphism corresponds to a presence/absence polymorphism at *CYP79D15* (Olsen *et al.*, 2008). To assess the potential roles of the second and third pathway genes in CNglc variation, we tested for molecular variation at *CYP736A187* and *UGT85K17* using a sample of 81 wild white clover accessions. Plants included 41 USDA accessions from locations worldwide and 40 wild ecotypes collected in the central and southern USA; 52 of the ecotypes produce CNglcs whereas the remaining 29 do not (Table S1). PCR genotyping assays targeting each CNglc gene (*CYP79D15*, *CYP736A187*, *UGT85K17*) revealed a perfect correlation between the presence/absence of PCR products for all three targeted loci and CNglc production (Table S1). To confirm the gene identity of amplified sequences, *CYP736A187* and *UGT85K17* PCR products were purified and sequenced by Sanger sequencing ($N = 29$ for *CYP736A187*, $N = 24$ for *UGT85K17*). All sequences correspond to the targeted genes and have been deposited in GenBank. Analysis of the aligned sequences indicate that patterns and levels of nucleotide diversity are very similar across the three CNglc genes (silent-site π: 0.0013–0.0032; silent-site θ_W: 0.0029–0.0035), with none showing statistically significant deviations from neutral equilibrium in tests of selection (Table S4). These similarities in population genetic parameters suggest similar mutation rates and evolutionary dynamics for these closely linked and functionally related loci.

In order to definitively confirm that the absence of PCR amplifications corresponded to CNglc gene deletions, we performed Southern hybridizations with probes targeting *CYP736A187* and *UGT85K17* using a selection of 50 white clover accessions, 32 of which produce CNglcs, 18 of which do not. Previous analyses of *CYP79D15* revealed strong hybridization to a probe targeting this gene in plants that produce CNglcs and no detectable hybridization in plants that do not (Olsen *et al.*, 2008). We observed same result for probes specific to *CYP736A187* and *UGT85K17*; hybridization was detected in plants that produce CNglcs whereas no corresponding bands were present in plants that do not (Fig. 2a,b; see also Fig. S1). Together with the results of PCR genotyping assays, this finding strongly suggests that the CNglc polymorphism in white clover arises through a presence/absence polymorphism for the complete gene cluster.

In previous work we detected polymorphisms for CNglc production in five congeners of white clover (*T. ambiguum*, *T. isthmocarpum*, *T. montanum*, *T. suffocatum* and *T. uniflorum*, all of which co-occur with white clover in *Trifolium* section *Trifoliastrum*); in all cases the polymorphisms corresponded to *CYP79D15* presence/absence as in white clover (Olsen *et al.*, 2014). To assess whether these CNglc polymorphisms in other *Trifolium* species also occur through presence/absence of the

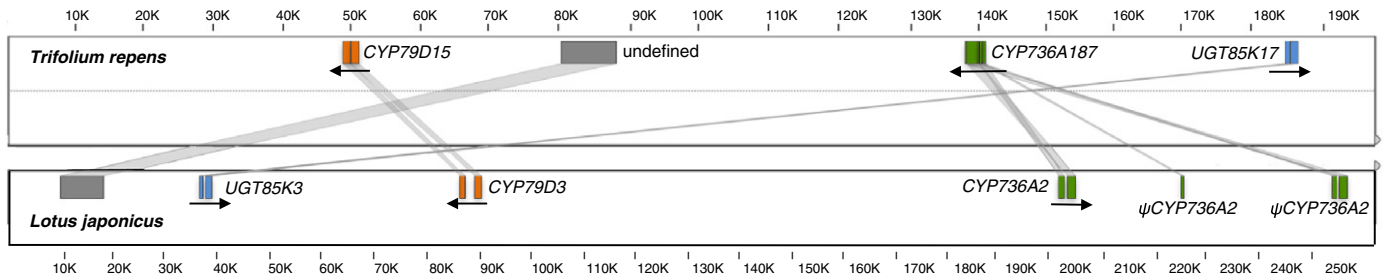

Fig. 1 Genomic comparison of the white clover (*Trifolium repens*) and *Lotus japonicus* cyanogenic glucoside (CNglc) gene cluster regions. Colored or gray shading indicates regions of sequence similarity as identified in GEvo analysis (Lyons & Freeling, 2008). Nonmatching regions within genes correspond to introns. Arrows below genes indicate transcript orientation.

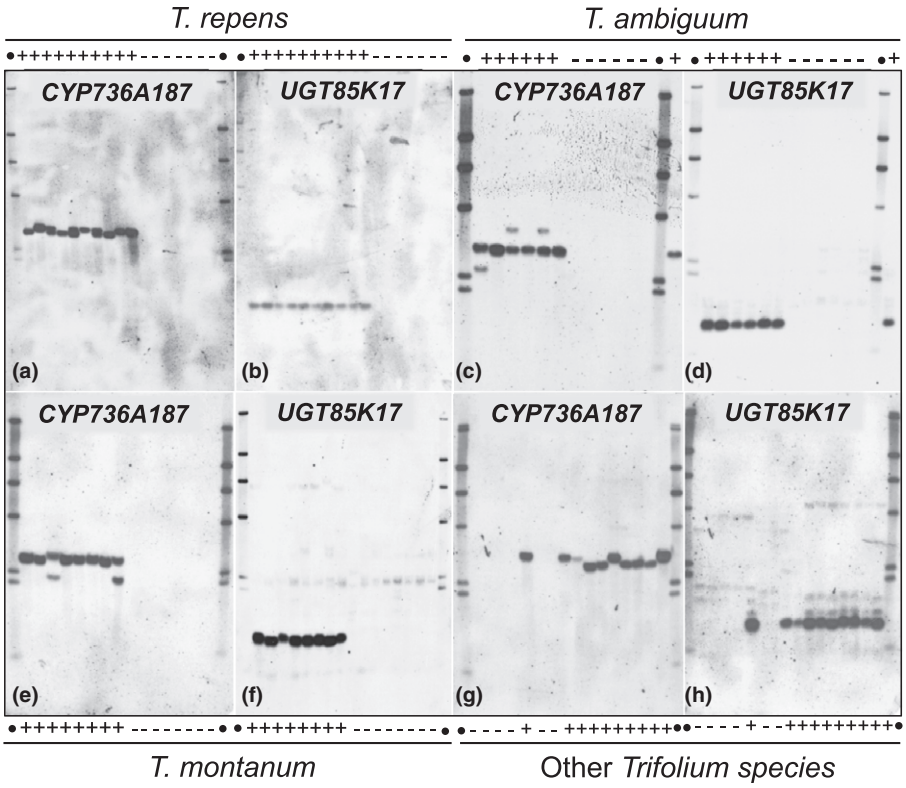

Fig. 2 Southern hybridizations for *CYP736A187* and *UGT85K17* in white clover and other *Trifolium* species. Sampled accessions are from white clover (*Trifolium repens*; a, b), *T. ambiguum* (c, d), *T. montanum* (e, f), and three other species (g, h): *T. suffocatum* (lanes 1–5), *T. uniflorum* (lanes 6–9) and *T. occidentale* (lanes 10–16). Plus and minus symbols correspond, respectively, to accessions with and without detectable concentrations of cyanogenic glucosides (CNglcs) in HCN assays. Solid circles indicate lanes with a ladder of genomic size standards (0.9–23.3 kb). All genomic digests were performed using *Ase*I. See Supporting Information Tables S1 and S2 for accession identifications per lane.

entire CNglc cluster, we performed PCR assays and Southern hybridizations as described above for white clover. Across a sample of 53 accessions spanning the five species, we again found a perfect correlation between CNglc production and the presence of *CYP736A187* and *UGT85K17* as detected by PCR assays (Table S2) and Southern hybridizations (Figs 2c–h, S2). Thus, as in white clover, the *Ac/ac* CNglc polymorphism in related *Trifolium* species occurs through the presence/absence of the complete CNglc gene cluster.

In order to confirm orthology between sequences from these other *Trifolium* species and the white clover CNglc genes, we sequenced PCR amplicons of 1–6 accessions per species for each gene, plus additional accessions of two *Trifoliastrum* species that were monomorphic for the presence of CNglc production in our samples (*T. occidentale*, *T. nigrescens*); 45 accessions were sequenced in total per gene (Table S2). All amplified products showed high sequence similarity to *T. repens CYP736A187* and *UGT85K17* (≥ 96% nucleotide identity in coding regions, ≥ 96% positive amino acid match), confirming that they are orthologous to the white clover and *L. japonicus* CNglc genes.

Discussion

Cyanogenic glucosides (CNglcs) are one of several classes of plant secondary metabolite where the underlying biosynthetic genes have been found to occur in physically linked clusters, and where these clusters have evolved convergently in different plant families (Takos *et al.*, 2011; Nützmann *et al.*, 2016). Here we have leveraged knowledge of the CNglc pathway in the model legume *Lotus japonicus* to identify the three component genes in white clover, all of which are orthologous to those of *L. japonicus* (Table 1). As with the three previously described CNglc pathways (cassava, sorghum, *L. japonicus*; Takos *et al.*, 2011), the white clover genes form a metabolic cluster (Fig. 1). Identification of this pathway in white clover has allowed us to test a key prediction of the co-inheritance hypothesis for metabolic gene cluster evolution: that adaptive polymorphism for the chemical defense involves the entire metabolic pathway. Consistent with this prediction, we find that the *Ac/ac* CNglc polymorphism arises through presence/absence variation for the complete CNglc gene cluster (Figs 2, S1, S2). Below we discuss these findings in the context of metabolic gene cluster evolution and the molecular evolution of cyanogenesis.

Metabolic cluster evolution

All metabolic clusters described to date in plants encode chemical defense compounds (Takos & Rook, 2012). Gene expression analyses have shown that these gene clusters are not universally characterized by tightly coordinated gene regulation (von Rad *et al.*, 2001; Swaminathan *et al.*, 2009; Takos *et al.*, 2011), and this has led to the proposition that assured co-inheritance is the key force driving their evolution (Takos & Rook, 2012). Importantly, an inherent assumption of the co-inheritance hypothesis is that natural selection acts to maintain an intraspecific polymorphism for the chemical defense; the genomic location of the component genes becomes irrelevant for co-inheritance once the chemical defense is universally favored in a species. Although an intriguing proposition, the co-inheritance hypothesis has remained purely speculative in the absence of data from species

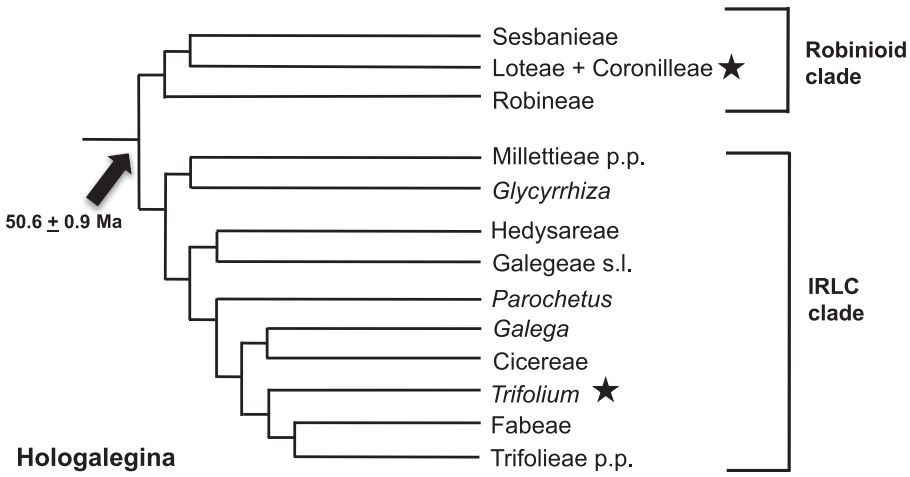

Fig. 3 Phylogenetic relationships within the papilionoid legume clade Hologalegina. Tree topology and crown-clade age estimate are from Lavin *et al.* (2005) and LPWG (2013). Stars indicate positions of branches containing clover (*Trifolium*) and *Lotus japonicus* (Loteae + Coronilleae). Taxonomic abbreviations: s.l., *sensu lato*; p.p., *pro parte*. Ma, Myr ago; IRLC, inverted repeat-lacking clade.

with intraspecific polymorphisms. Here, our findings for the clover CNglc polymorphism provide strong empirical support for the predictions of this hypothesis. Rather than occurring through variation at the *CYP79D15* locus only, which would be the simplest genetic mechanism, we instead find that the CNglc polymorphism corresponds to the presence/absence of the entire gene cluster. Moreover, this pattern holds true not only in white clover but also in several related *Trifolium* species.

For CNglcs specifically and plant chemical defense metabolites more generally, the toxicity of intermediates in the biosynthetic pathway is likely a critical factor driving physical linkage and co-inheritance of the component genes (Kristensen *et al.*, 2005; Mylona *et al.*, 2008; Blomstedt *et al.*, 2016; Nützmann *et al.*, 2016). In the case of CNglcs, the oximes and cyanohydrins that are produced as intermediates during their synthesis are highly unstable and reactive (Gleadow & Møller, 2014). The toxicity of these intermediates has been well-documented in studies that have transformed Arabidopsis and tobacco, two noncyanogenic species, with components of the sorghum CNglc pathway (Bak *et al.*, 2000; Kristensen *et al.*, 2005). In the context of the clover CNglc polymorphism, complete linkage of functional gene copies or nonfunctional gene copies (in this case gene deletions) prevents the possibility of recombination generating a maladaptive genotype with a functional copy of *CYP79D15* but no functional copies of the second and/or third pathway genes. This selection for physical linkage of functionally-related genes is in accordance with R. A. Fisher's model for the evolution of co-adapted gene complexes in heterogeneous environments (Fisher, 1930), which likewise predicts the evolution of nonrecombining sex chromosomes and self-incompatibility supergenes (reviewed in Schwander *et al.*, 2014).

Although the data presented here for white clover strongly support the co-inheritance hypothesis, the question remains as to whether this is the primary explanation for the existence of most plant chemical defense gene clusters in nature. To our knowledge, none of the many other metabolic gene clusters that have been identified (reviewed in Osbourn, 2010; Boycheva *et al.*, 2014; Nützmann *et al.*, 2016) are associated with intraspecific adaptive polymorphisms for the chemical defense. A possibility on one hand is that many of these metabolites do show adaptive polymorphisms that have simply gone undetected. A recent review of intraspecific diversity in plant secondary metabolites suggests that such polymorphisms may indeed be common in nature (Moore *et al.*, 2014). Detailed population surveys in species with known metabolic gene clusters would be useful for testing this hypothesis. On the other hand, some of the best-documented adaptive polymorphisms in plants are for classes of metabolites that show no evidence of gene cluster evolution; this includes both glucosinolates and flavonoids (Nützmann *et al.*, 2016). Thus, co-inheritance is clearly not a requirement for the evolution of all adaptive chemical defense polymorphisms.

It is also likely that coordinated gene regulation rather than co-inheritance is the predominant factor in the evolution of some metabolic clusters. Recent work in Arabidopsis has revealed gene cluster-specific signatures of chromatin modification that are associated with localized, coordinated regulation of the component genes. Specifically, metabolic clusters have been found to be enriched for histone 3 lysine trimethylation (H3K27me3) and histone 2 variant H2A.Z, which are associated with cluster repression and activation, respectively (Nützmann & Osbourn, 2015; Yu *et al.*, 2016). These same chromatin modification signatures also were detected in metabolic gene clusters of maize and oat, suggesting that chromatin-mediated co-regulation may be a general phenomenon for at least some classes of metabolic cluster.

For the specific case of CNglcs, however, coordinated expression does not appear to be an important factor. Expression profiling of the three component genes in *L. japonicus* indicates that they are not strictly co-regulated (Takos *et al.*, 2011). Likewise, expression analyses of the white clover CNglc genes in leaf tissue do not indicate tight co-expression across the three-gene cluster (K. M. Olsen, unpublished observations). In addition, the lack of conservation between the white clover and *L. japonicus* metabolic clusters in overall structure – including gene order, transcript orientation, and intergenic sequences (Fig. 1) – suggests that there is no obvious conservation of secondary structure in these genomic regions as might be predicted if there were strong selection to maintain a pattern of coordinated expression (although it should be noted that similar regulatory processes can occur independent of cluster size and

component genes; e.g. Slot, 2017). Thus, although co-regulation may be a key factor in the evolution of some metabolic clusters, it does not appear to play a critical role for the CNglc pathway.

Microevolutionary adaptation through recurrent deletions of the CNglc pathway

The white clover cyanogenesis polymorphism was first described more than a century ago (Armstrong *et al.*, 1913), and scores of subsequent studies have sought to understand its genetic basis and the ecological factors that maintain it (reviewed in Hughes, 1991; Olsen *et al.*, 2013). Cyanogenic clover plants are differentially protected against a variety of small generalist herbivores (reviewed in Kooyers *et al.*, 2014). At the same time, acyanogenic plants predominate at higher latitudes and elevations, and climate-associated cyanogenesis clines have evolved in native and naturalized populations worldwide (e.g., Daday, 1954a,b, 1958; Kooyers & Olsen, 2012, 2013; Thompson *et al.*, 2016). Glasshouse experiments indicate that there are energetic costs associated with the production of the cyanogenic components (Kakes, 1989; Kooyers *et al.*, 2014); acyanogenic plants may therefore have a competitive advantage in cooler climates if fewer herbivores are present in these locations. CNglcs may also function as nitrogen storage compounds that are particularly beneficial in drought-prone environments (Gleadow & Møller, 2014; Kooyers *et al.*, 2014). Taken together, these observations suggest that a combination of herbivore pressure, energetic costs of CNglc production, and drought-stress adaptation are all likely to be acting to maintain the *Ac/ac* polymorphism.

Previous molecular characterizations of the *Ac/ac* CNglc polymorphism focused solely on the signature gene, *CYP79D15* (Olsen *et al.*, 2008, 2013, 2014). Assessments of the genomic boundaries of the *CYP79D15* deletion, which specifically examined the sequence downstream of the transcript (i.e. upstream of the three-gene cluster as shown in Fig. 1), revealed that *ac* (gene-deletion) alleles have evolved multiple times through independent deletion events of varying size (Olsen *et al.*, 2013). Moreover, these recurrently evolved deletions have contributed to the rapid evolution of adaptive cyanogenesis clines in the introduced species range (Kooyers & Olsen, 2014). In light of the findings of the present study, it is now apparent that these *ac* gene deletions encompass the entire CNglc cluster. To our knowledge, this is the first reported case of microevolutionary adaptation occurring through selection for the repeated loss of a complete biosynthetic pathway. The occurrence of these same CNglc cluster deletions in several related *Trifolium* species (Figs 2, S2; Table S2) suggests that this mechanism of environmental adaptation is not unique to white clover.

Conservation and loss of cyanogenesis in papilionoid legumes

Whereas the genes comprising the CNglc cluster have been found to have independent evolutionary origins in sorghum, cassava and *L. japonicus* (Takos *et al.*, 2011), they are orthologous between *L. japonicus* and white clover (Table 1; Fig. 1). Both species are members of the legume subfamily Papilionoideae; their phylogenetic relationship indicates that they last shared a common ancestor with the origin of the large 'Hologalegina' clade, which has been dated to 50.6 ± 0.9 Myr ago (Lavin *et al.*, 2005) (Fig. 3). Interestingly, the vast majority of the Hologalegina, which includes > 4800 species and many of the world's important legume crops (including pea, alfalfa, chickpea, lentil and broad bean), are *not* cyanogenic. Thus, it appears that CNglc production was present in the common ancestor of this large legume clade and was repeatedly lost as the clade diversified over the last 50 Myr. We can only speculate on the forces that might have driven this process; however, the tremendous diversity of other chemical defense metabolites that can be found in the Hologalegina today (including various alkaloids, phenolics and terpenoids; Wink, 2013) suggests that CNglcs may have been superseded by more complex metabolites in the course of evolutionary arms races with herbivores. It is notable that even within the genus *Trifolium* (*c.* 255 species total), CNglc production is apparently only present in the small *Trifoliastrum* clade comprising white clover and its closest relatives (*c.* 20 species total) (Ellison *et al.*, 2006; Olsen *et al.*, 2014).

Conclusions

The overall picture that emerges for CNglc pathway evolution is one of both flexibility and conservation: independent evolution of CNglc metabolic clusters across the angiosperms (e.g., Takos *et al.*, 2011); conservation of the component genes over a 50-Myr timescale within the legumes (Table 1; Figs 1, 3); and microevolutionary adaptation in white clover through repeated deletions of the entire pathway (Figs 2, S1, S2). Although the CNglc variation examined here is for a discrete presence/absence polymorphism, it seems likely that quantitative CNglc variation, which we suspect plays an equally important role in environmental adaptation, has been shaped by a similar degree of evolutionary complexity at the genetic level.

Acknowledgements

This work was supported by grants from the National Science Foundation (DEB-0845497, IOS-1557770) to K.M.O. The authors thank Mike Dyer and staff of the Washington University glasshouse for assistance in growing plants; Chris Saski (CUGI) for assistance in BAC library construction and sequencing; David Nelson (University of Tennessee) and Michael Court (Washington State University) for guidance on P450 and UGT gene nomenclature; David Goad (Olsen lab, Washington University) for assistance with GEvo analysis; and additional members of the Olsen lab group for helpful comments on the manuscript.

Author contributions

K.M.O. planned and designed the research; L.L.S. performed the experiments and preliminary analyses; and K.M.O. analyzed the data and wrote the manuscript.

ORCID

Kenneth M. Olsen http://orcid.org/0000-0002-8338-3638

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 B* **86**: 262–269.

Bak S, Olsen CE, Halkier BA, Møller BL. 2000. Transgenic tobacco and Arabidopsis plants expressing the two multifunctional sorghum cytochrome P450 enzymes, CYP79A1 and CYP71E1, are cyanogenic and accumulate metabolites derived from intermediates in Dhurrin biosynthesis. *Plant Physiology* **123**: 1437–1448.

Blomstedt CK, O'Donnell NH, Bjarnholt N, Neale AD, Hamill JD, Møller BL, Gleadow RM. 2016. Metabolic consequences of knocking out UGT85B1, the gene encoding the glucosyltransferase required for synthesis of dhurrin in *Sorghum bicolor* (L. Moench). *Plant & Cell Physiology* **57**: 373–386.

Boycheva S, Daviet L, Wolfender JL, Fitzpatrick TB. 2014. The rise of operon-like gene clusters in plants. *Trends in Plant Science* **19**: 447–459.

Coop IE. 1940. Cyanogenesis in white clover (*Trifolium repens* L.). III. A study of Linamarase, the enzyme which hydrolyses lotaustralin. *New Zealand Journal of Science and Technology. A: Agricultural Research Section* **22**: 71B–83B.

Corkill L. 1942. Cyanogenesis in white clover (*Trifolium repens* L.) V. The inheritance of cyanogenesis. *The New Zealand Journal of Science and Technology. B* **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* L. II. Distribution by altitude. *Heredity* **8**: 377–384.

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

Ellison NW, Liston A, Steiner JJ, Williams WM, Taylor NL. 2006. Molecular phylogenetics of the clover genus (*Trifolium*–Leguminosae). *Molecular Phylogenetics and Evolution* **39**: 688–705.

Field B, Fiston-Lavier A-S, Kemen A, Geisler K, Quesneville H, Osbourn AE. 2011. Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proceedings of the National Academy of Sciences, USA* **108**: 16 116–16 121.

Field B, Osbourn AE. 2008. Metabolic diversification – independent assembly of operon-like gene clusters in different plants. *Science* **320**: 543–547.

Fisher RA. 1930. *The genetical theory of natural selection.* Oxford, UK: Clarendon Press.

Frey M, Chomet P, Glawischnig E, Stettner C, Grün S, Winklmair A, Eisenreich W, Bacher A, Meeley RB, Briggs SP *et al.* 1997. Analysis of a chemical plant defense mechanism in grasses. *Science* **277**: 696–699.

Gleadow RM, Møller BL. 2014. Cyanogenic glycosides: synthesis, physiology, and phenotypic plasticity. *Annual Review of Plant Biology* **65**: 155–185.

Hall T. 2001. *BioLign alignment and multiple contig editor.* [WWW document] URL http://en.bio-soft.net/dna/BioLign.html [accessed 4 May 2017].

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

Hurst LD, Pál C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. *Nature Reviews. Genetics* **5**: 299–310.

Itkin M, Heinig U, Tzfadia O, Bhide AJ, Shinde B, Cardenas PD, Bocobza SE, Unger T, Malitsky S, Finkers R *et al.* 2013. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science* **341**: 175–179.

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.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C *et al.* 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–9.

Kooyers NJ, Gage LR, Al-Lozi A, Olsen KM. 2014. Aridity shapes cyanogenesis cline evolution in white clover (*Trifolium repens* L.). *Molecular Ecology* **23**: 1053–1070.

Kooyers NJ, Olsen KM. 2012. Rapid evolution of an adaptive cyanogenesis cline in introduced North American white clover (*Trifolium repens* L.). *Molecular Ecology* **21**: 2455–68.

Kooyers NJ, Olsen KM. 2013. Searching for the bull's eye: agents and targets of selection vary among geographically disparate cyanogenesis clines in white clover (*Trifolium repens* L.). *Heredity* **111**: 495–504.

Kooyers NJ, Olsen KM. 2014. Adaptive cyanogenesis clines evolve recurrently through geographical sorting of existing gene deletions. *Journal of Evolutionary Biology* **27**: 2554–2558.

Kristensen C, Morant M, Olsen CE, Ekstrøm CT, Galbraith DW, Møller BL, Bak S. 2005. Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome. *Proceedings of the National Academy of Sciences, USA* **102**: 1779–1784.

Lavin M, Herendeen PS, Wojciechowski MF. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Systematic Biology* **54**: 575–94.

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

LPWG. 2013. Legume phylogeny and classification in the 21st Century: progress, prospects and lessons for other species-rich clades. *Taxon* **62**: 217–248.

Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant Journal* **53**: 661–673.

Mackenzie PI, Owens IS, Burchell B, Bock KW, Bairoch A, Bélanger A, Fournel-Gigleux S, Green M, Hum DW, Iyanagi T *et al.* 1997. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence. *Pharmacogenetics* **7**: 255–269.

Matsuba Y, Nguyen TTH, Wiegert K, Falara V, Gonzales-Vigil E, Leong B, Schäfer P, Kudrna D, Wing RA, Bolger AM *et al.* 2013. Evolution of a complex locus for terpene biosynthesis in solanum. *Plant Cell* **25**: 2022–2036.

McGary KL, Slot JC, Rokas A. 2013. Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds. *Proceedings of the National Academy of Sciences, USA* **110**: 11481–11486.

Melville J, Doak BW. 1940. Cyanogenesis in white clover (*Trifolium repens* L.). II. Isolation of the glucosidal constituents. *New Zealand Journal of Science and Technology, Section A* **22**: 67B–71B.

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

Moore BD, Andrew RL, Külheim C, Foley WJ. 2014. Explaining intraspecific diversity in plant secondary metabolites in an ecological context. *New Phytologist* **201**: 733–750.

Mylona P, Owatworakit A, Papadopoulou K, Jenner H, Qin B, Findlay K, Hill L, Qi X, Bakht S, Melton R *et al.* 2008. Sad3 and sad4 are required for saponin biosynthesis and root development in oat. *Plant Cell* **20**: 201–12.

Nelson DR. 2009. The cytochrome p450 homepage. *Human Genomics* **4**: 59–65.

Nützmann H-W, Huang A, Osbourn A. 2016. Plant metabolic clusters – from genetics to genomics. *New Phytologist* **211**: 771–789.

Nützmann H-W, Osbourn A. 2015. Regulation of metabolic gene clusters in *Arabidopsis thaliana*. *New Phytologist* **205**: 503–510.

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.

Olsen KM, Kooyers NJ, Small LL. 2013. Recurrent gene deletions and the evolution of adaptive cyanogenesis polymorphisms in white clover (*Trifolium repens* L.). *Molecular Ecology* **22**: 724–738.

Olsen KM, Kooyers NJ, Small LL. 2014. Adaptive gains through repeated gene loss: parallel evolution of cyanogenesis polymorphisms in the genus *Trifolium* (Fabaceae). *Philosophical Transactions of the Royal Society B* **369**: 20130347–20130347.

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.

Osbourn A. 2010. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics* **26**: 449–457.

Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A. 2004. A gene cluster for secondary metabolism in oat: implications for the evolution of metabolic diversity in plants. *Proceedings of the National Academy of Sciences, USA* **101**: 8233–8238.

von Rad U, Hüttl R, Lottspeich F, Gierl A, Frey M. 2001. Two glucosyltransferases are involved in detoxification of benzoxazinoids in maize. *Plant Journal* **28**: 633–642.

Roselli S, Olry A, Vautrin S, Coriton O, Ritchie D, Galati G, Navrot N, Krieger C, Vialart G, Bergès H *et al.* 2017. A bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip. *Plant Journal* **89**: 1119–1132.

Schneider LM, Adamski NM, Christensen CE, Stuart DB, Vautrin S, Hansson M, Uauy C, von Wettstein-Knowles P. 2016. The *Cer-cqu* gene cluster determines three key players in a β-diketone synthase polyketide pathway synthesizing aliphatics in epicuticular waxes. *Journal of Experimental Botany* **67**: 2715–2730.

Schwander T, Libbrecht R, Keller L. 2014. Supergenes and complex phenotypes. *Current Biology* **24**: R288–94.

Slot JC. 2017. Fungal gene cluster diversity and evolution. *Advances in Genetics* **100**: 141–178.

Swaminathan S, Morrone D, Wang Q, Fulton DB, Peters RJ. 2009. CYP76M7 is an ent-cassadiene C11alpha-hydroxylase defining a second multifunctional diterpenoid biosynthetic gene cluster in rice. *Plant Cell* **21**: 3315–3325.

Takos AM, Knudsen C, Lai D, Kannangara R, Mikkelsen L, Motawia MS, Olsen CE, Sato S, Tabata S, Jørgensen K *et al.* 2011. Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. *Plant Journal* **68**: 273–286.

Takos AM, Rook F. 2012. Why biosynthetic genes for chemical defense compounds cluster. *Trends in Plant Science* **17**: 383–388.

Thompson KA, Renaudin M, Johnson MTJ. 2016. Urbanization drives the evolution of parallel clines in plant populations. *Proceedings of the Royal Society B* **283**: 20162180.

Williams W, Ansari HA, Ellison NW, Hussain SW. 2001. Evidence of three subspecies in *Trifolium nigrescens* Viv. *Annals of Botany* **87**: 683–691.

Wink M. 2013. Evolution of secondary metabolites in legumes (Fabaceae). *South African Journal of Botany* **89**: 164–175.

Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, Li Y, Meade F, Teodor R, Vaistij FE *et al.* 2012. A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science* **336**: 1704–1708.

Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences, USA* **110**: 1–9.

Yu N, Nützmann H-W, MacDonald JT, Moore B, Field B, Berriri S, Trick M, Rosser SJ, Kumar SV, Freemont PS *et al.* 2016. Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Research* **44**: 2255–2265.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Southern hybridizations for *CYP736A187* and *UGT85K17* in additional samples of white clover.

Fig. S2 Southern hybridizations for *CYP736A187* and *UGT85K17* in *Trifolium isthmocarpum.*

Table S1 Presence/absence variation for CNglc production in white clover

Table S2 Presence/absence variation for CNglc production in congeners of white clover

Table S3 Primer sequences used for PCR amplifications and DNA sequencing of *Trifolium* CNglc genes

Table S4 Nucleotide diversity, divergence and tests of selection for white clover CNglc genes

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.