Dual-species origin of an adaptive chemical defense polymorphism

Kenneth M. Olsen¹, David M. Goad¹*, Sara J. Wright¹,²*, Maya L. Dutta¹, Samantha R. Myers¹, Linda L. Small¹ and Lin-Feng Li³

¹Department of Biology, Washington University in St. Louis, 1 Brookings Dr., St Louis, MO 63130, USA; ²Biological Sciences Department, Rowan University, Glassboro, NJ 08028, USA; ³Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China

Summary

- Allopolyploid speciation and chemical defense diversification are two of the most characteristic features of plant evolution; although the former has likely shaped the latter, this has rarely been documented. Here we document allopolyploidy-mediated chemical defense evolution in the origin of cyanogenesis (HCN release upon tissue damage) in white clover (Trifolium repens).
- We combined linkage mapping of the loci that control cyanogenesis (Ac, controlling production of cyanogenic glucosides; and Li, controlling production of their hydrolyzing enzyme linamarase) with genome sequence comparisons between white clover, a recently evolved allotetraploid, and its diploid progenitors (Trifolium pallescens, Trifolium occidentale).
- The Ac locus (a three-gene cluster comprising the cyanogenic glucoside pathway) is derived from T. occidentale; it maps to linkage group 2O (occidentale subgenome) and is orthologous to a highly similar cluster in the T. occidentale reference genome. By contrast, Li maps to linkage group 4P (pallescens subgenome), indicating an origin in the other progenitor species.
- These results indicate that cyanogenesis evolved in white clover as a product of the interspecific hybridization that created the species. This allopolyploidization-derived chemical defense, together with subsequent selection on intraspecific cyanogenesis variation, appears to have contributed to white clover’s ecological success as a globally distributed weed species.

Introduction

A hallmark of plant species diversification has been the evolution of diverse specialized metabolites, many of which function as antitherbivore chemical defenses (Pichersky & Lewinsohn, 2011; Mithöfer & Boland, 2012). In recent decades, advances in genomics, combined with biochemical and physiological studies, have provided important insights into the molecular bases of these defense metabolite pathways and their macroevolutionary origins (e.g. Zhuang et al., 2012; Edger et al., 2015; Xu et al., 2017; Lai et al., 2020). On a microevolutionary timescale, species that are polymorphic for chemical defense production have served as especially valuable systems for studying the molecular basis of adaptive variation and the roles of natural selection and other forces in maintaining it (Brachi et al., 2015; Wager & Li, 2018; Keith & Mitchell-Olds, 2019; Lowry et al., 2019).

Like chemical defense evolution, another defining feature of plant diversification has been allopolyploid speciation, whereby interspecific hybridization and retention of both parental species’ genomes creates a new polyploid daughter species. This allopolyploidization process has the potential to generate novel adaptive variation for diverse traits including chemical defense metabolites (Soltis et al., 2014; Van de Peer et al., 2017). Experiments with synthetically created allopolyploids have demonstrated, for example, that distinct chemical defense profiles can arise through remodeling of parental species’ gene expression networks and post-transcriptional processing (Anssour & Baldwin, 2010). However, studies are largely lacking on the role of allopolyploidization for chemical defense evolution in natural wild species; in particular, there is little empirical work on the relationship between chemical defense traits in allopolyploids and the genetic contributions of their diploid progenitors. Given the likely role of defense metabolite evolution in angiosperm diversification (Pichersky & Lewinsohn, 2011; Mithöfer & Boland, 2012), and the prevalence of allopolyploid speciation in plants (Van de Peer et al., 2017), characterizing this relationship is an important step towards understanding the mechanisms of adaptive diversification in the plant kingdom.

A long-standing model system for plant chemical defense evolution is cyanogenesis (HCN release following tissue damage) and the polymorphism for this trait that occurs in white clover (Trifolium repens L.). First described more than a century ago (Armstrong et al., 1913), white clover plants vary in the cyanogenic response, with both cyanogenic and acyanogenic
genotypes found in wild populations. Latitudinal and elevational clines in cyanogenesis have evolved in white clover throughout the native species range in Europe and in introduced populations worldwide; cyanogenic plants predominate in warmer and drier locations, but are largely absent from populations at higher latitudes and elevations (Daday, 1954, 1958; Kooyers & Olsen, 2012, 2013). Since the 1950s, numerous studies have attempted to identify the selective factors that maintain this polymorphism and favor the recurrent evolution of climate-associated cyanogenesis clines. Key factors likely include herbivore deterrence, energetic costs of producing the defense and abiotic stress adaptation (reviewed by Hughes, 1991; Olsen et al., 2013; Kooyers et al., 2018). More recently, the molecular basis of the white clover cyanogenesis polymorphism also has been characterized (Olsen et al., 2007, 2008; Olsen & Small, 2018). This chemical defense polymorphism has come to serve as a textbook example of intraspecific variation maintained by geographically heterogeneous selection (Briggs & Walters, 2016; Futuyma & Kirkpatrick, 2017), and recently it also has been adopted as a model for understanding adaptation to urban environments (Thompson et al., 2016; Santangelo et al., 2020). White clover is a common weed species worldwide; the fact that climate-associated cyanogenesis clines have evolved repeatedly throughout the native and introduced species range suggests that selection on this chemical defense polymorphism has facilitated its global expansion and widespread ecological success (Daday, 1958; Kooyers & Olsen, 2013; Santangelo et al., 2020).

White clover is an allotetraploid species (2n = 4x = 32), having arisen through the hybridization of two diploid (2n = 2x = 16) Trifolium species (Williams et al., 2012; Griffiths et al., 2019), and a critical unanswered question for the cyanogenesis polymorphism is how the evolution of this chemical defense relates to the species’ hybrid origin. The cyanogenic response occurs through the interaction of two chemical precursors that are separated in intact tissue and brought together upon tissue damage: cyanogenic glucosides (CNGLcs), which in leaf tissue are likely compartmentalized in the vacuole, and their hydrolyzing enzyme linamarase, which is present in the apoplast (Kakes, 1985; Gleadow & Möller, 2014). Inheritance of the polymorphism is determined by two unlinked simple Mendelian polymorphisms that control the production of each precursor: Aclac controls the presence/absence of CNGLcs, and the independently segregating Lili gene controls the presence/absence of linamarase (Hughes, 1991). For both polymorphisms, homozygote recessive genotypes (acac or lili) lack the corresponding chemical precursor. Thus, plants that carry at least one dominant allele at both genes have both required precursors and are cyanogenic (referred to as the AcLi ‘cyanotype’), whereas plants that are homozygous recessive at either or both genes are acyanogenic (acl, AcLi and acl ‘cyanotypes’).

At the molecular level, the Aclac polymorphism corresponds to a genomic presence/absence polymorphism for the three-gene metabolic cluster that constitutes the CNGLc biosynthesis pathway (Olsen & Small, 2018). In parallel, the unlinked Lili polymorphism corresponds to a genomic presence/absence polymorphism for the gene encoding the linamarase protein (Olsen et al., 2007). For both Aclac and Lili, the recessive (gene-deletion) alleles have arisen multiple times independently in white clover (Olsen et al., 2013), and gene deletions at these same loci have evolved in parallel in several related Trifolium species (Olsen et al., 2014; Olsen & Small, 2018). Because inheritance of both the Aclac and Lili polymorphisms follows a single gene, two-allele model, both of the cyanogenesis loci have been presumed to be present within only one of white clover’s two parental subgenomes (Kakes & Halkvoort, 1994; Badr et al., 2002). However, their genomic locations have remained unresolved; this has left open the questions of whether both genes are derived from the same diploid progenitor, and which progenitor(s) contributed them.

The recent publication of a white clover reference genome, together with conclusive identification of its diploid progenitor species (Williams et al., 2012; Griffiths et al., 2019), has opened up the possibility of tracing the evolutionary origins of the cyanogenesis loci. T. repens is estimated to have originated in the very recent evolutionary past (~15–28 000 yr ago) by multiple hybridizations between Trifolium pallescens, a species now distributed in alpine regions of central Europe, and Trifolium occidentale, a species currently restricted to Atlantic coastal margins of Western Europe (Griffiths et al., 2019). Curiously, the present-day occurrence of cyanogenesis in these diploid progenitors provides only limited insight for inferring the trait’s origins. T. occidentale carries the Aclac polymorphism (Kakes & Chardonnens, 2000), and, as in white clover, ac alleles correspond to deletions of the Ac gene cluster (Olsen et al., 2014; Olsen & Small, 2018). However, extensive surveys of wild populations have not detected linamarase (or the Li gene) in this species (Kakes & Chardonnens, 2000). For T. pallescens, neither of the cyanogenic precursors (or their underlying genes) have been detected in population sampling to date (Gibson et al., 1972; Olsen et al., 2014) (see also Supporting Information Table S1).

Because gene sequences of the two diploid progenitor species are highly similar to each other (≥97% identity on average; Table S2) and also are highly conserved with their respective descendant subgenomes in T. repens (Griffiths et al., 2019), most white clover sequence contigs cannot be definitively assigned to one subgenome or the other in the currently available shotgun assembly reference genome (A. Griffiths and R. Moraga, pers. comm.; D. M. Goad, unpublished). Thus, the Ac and Li gene sequences alone are insufficient to determine their subgenome locations and diploid contributors. This problem can potentially be overcome, however, by combining sequence analyses with genetic mapping of the Aclac and Lili polymorphisms and integration of linkage map data into the genome assembly. The availability of draft genome assemblies for each diploid progenitor species (Griffiths et al., 2019) further facilitates this genetic mapping process.

In this study we generated high-density linkage maps from two white clover F₂ mapping populations and combined genetic mapping of the Aclac and Lili cyanogenesis polymorphisms with gene and genome sequence analyses to assess the evolutionary origins of the cyanogenesis loci. We addressed the following questions: (1) Given that the Ac gene cluster is known to exist in T. occidentale, do genetic map data and genome sequence analyses...
support the hypothesis that the white clover Ac locus is derived from that diploid species? (2) Is the Li locus located in the same subgenome as Ac, indicating that both cyanogenic components were contributed by the same diploid progenitor? Our results reveal that this textbook example of a plant chemical defense has a dual-species origin, with the required cyanogenesis genes contributed through the interspecific hybridization that created this widely adapted allopolyploid species.

Materials and Methods

Generation of mapping populations

Trifolium repens is a native species of Europe that has been introduced into mesic, temperate regions worldwide; it is grown for forage and is widely naturalized in lawns, pastures and other disturbed habitats. For this study, two T. repens F2 mapping populations were generated from biparental crosses using three wild-collected North American plants. The parental genotypes were selected such that the presence/absence of cyanogenic glucosides (CNgels) (Acac) and of linamarase (Liλl) polymorphisms would segregate to create all four cyanotypes in the F2 generation: DMN_010 (acac, lili; originating from Duluth, MN); STL_0701 (acac, lili; from St Louis, MO) (Kooyers & Olsen, 2012); and GFL_007 (AcAc, LiLi; from Gainesville, FL). GFL_007 served as a parent in both mapping populations, yielding the DMN_010 × GFL_007 (hereafter DG) and STL_0701 × GFL_007 (hereafter SG) populations. For both populations, hand crosses were performed in the greenhouse between parents in both directions to generate 50–100 F1 genotypes; the species has gametophytic self-incompatibility (Casey et al., 2010), so no self-fertilization occurred. Within each F1 population, random crosspollinations were performed to generate c. 500 F2 genotypes per population (Tables S3, S4). For the DG population, solitary bees (Crown Bees, Woodinville, WA) were used to cross-pollinate F1 plants in enclosed cages in the greenhouse. For the SG population, random crosses were performed by hand to produce F2 seed.

The three parental genotypes, along with F1 and F2 genotypes, were grown in the Washington University glasshouse facilities under standard conditions (16 h:8 h, light:dark photoperiod with 140 µmol supplemental lighting as needed; 18–23°C; c. 50% relative humidity). Seeds were scarified using fine grit sandpaper and placed on mist benches in the greenhouse for germination. Seedlings were transferred to individual pots for establishment after producing at least one true leaf.

Genotyping

For all plants of all generations in both mapping populations, young fresh leaf tissue (120–150 mg) was ground to powder in liquid N2 using a mortar and pestle, and high-purity genomic DNA (gDNA) was extracted using a protocol modified from Whitlock et al. (2008). DNA was run on 0.8% agarose gels to check for quality and integrity. Samples were quantified and standardized to 17–23 ng µl−1 as measured by a Qubit® 2.0 Fluorometer (Invitrogen).

Ac and Li polymorphisms Parental, F1 and F2 plants were genotyped for the Acac and LiLi gene presence/absence polymorphisms using previously described PCR genotyping protocols (Olsen et al., 2007, 2008) (Table S5). Negative PCR results were repeated at least once to confirm recessive (gene-absence) genotypes. Genotypes were further confirmed by cyanogenesis phenotyping using leaf tissue and Feigl–Anger HCN tests (Feigl & Anger, 1966), as described previously (Olsen et al., 2007). For acyanogenic genotypes, the presence of individual cyanogenic components was assessed by exogenously adding complementary reagents (either linamarase or cyanogenic glucosides) individually to leaf tissue and re-testing with Feigl–Anger paper following the previously described protocol. All F1 genotypes in both populations were confirmed to be cyanogenic (AcLi cyanotype), consistent with the expectation that they would be heterozygous at both cyanogenesis genes (Acac, Liλl).

Genotyping-by-sequencing Standardized quantities of gDNA samples from the parent plants and all F1 and F2 plants were prepared in randomized libraries, each containing 95 samples and one negative control sample (water), using a genotyping-by-sequencing (GBS) protocol modified from (Huang et al., 2014) with 100 ng gDNA and ApeiKl for digestion (Elshire et al., 2011). Common and adapter barcode sequences were created using an online GBS barcode generator (Deena Bioinformatics, Thomas van Gurp, Nijmegen, the Netherlands) for design of oligonucleotide barcode sequences (IDT, Coralville, IA, USA) (Table S6). Following enzyme digestion and adapter ligation, cleanup and size-selection steps were performed using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA). Samples then were pooled and PCR-amplified using primers from Elshire et al. (2011) (Table S5).

Paired-end sequencing (150-bp reads) was performed with each library in a unique lane using the Illumina HiSeq 2500 platform (Novogene Corp., Chula Vista, CA, USA). GBS libraries (13 in total) were run individually on sequencing lanes with a 15% PhiX Control spike-in (Illumina, San Diego, CA, USA). The three parents of the mapping populations were included as standards in all 13 libraries to increase read depth and reliability of heterozygous single nucleotide polymorphism (SNP) calls for optimal selection of parental markers for genetic mapping.

Raw GBS reads were analyzed using an in-house version of the FAST-GBS pipeline (Torkamaneh et al., 2017) that was modified to optimize handling of paired-end data. Briefly, reads were demultiplexed with SABRE (https://github.com/najoshi/sabre) and trimmed with CUTFADAPT (Martin, 2011). Reads then were mapped back to the white clover reference genome (Griffiths et al., 2019) using BWA (Li & Durbin, 2009) on default paired-end settings. Parental and F1 SNP variants were called with Platypus (Rimmer et al., 2014), which filtered reads with unmapped or distant mates.

Linkage map construction

In order to generate a set of markers that were most likely to segregate in a 1:2:1 ratio in the F2 populations, SNPs were filtered...
to the subset that were homozygous in both parents and had >0.7 heterozygosity in the F₁ population. Allele counts for each F₂ SNP were passed to GUSMAP v.1.0 (Bilton et al., 2018), where additional SNP filtering was performed before linkage map construction. Any F₂ lines with <15× average read depth were excluded (Tables S2, S3). Filtering also removed any markers that did not meet the following criteria: minor allele frequency (MAF) >0.05, missing data <0.1, average read depth >5×, or a P-value <0.01 on a genotype frequency test (indicating deviations from H₂/O₁ segregation in the F₂ populations). Additionally, SNPs were pruned such that only one SNP was present from any given 2000-bp region to represent a single GBS fragment.

The createLG function in GUSMAP was used to calculate pairwise r² and logarithm of odds (LOD) scores; SNPs were assigned to a linkage group if their LOD score in that group exceeded 45. Because GUSMAP v.1.0 was designed for data from an F₁ pseudotestcross population, this function was modified slightly to use only SNPs that were informative in both parents rather than those that are only maternally or paternally informative. The SNPs then were ordered using the orderLG function with default parameters. Markers that appeared to be incorrectly placed upon manual inspection of recombination rates were removed, and linkage maps were created using the computeMap function with the Kosambi method for calculating distances. Markers that introduced large gaps were then removed from the map and distances were recalculated.

In order to assign labels to linkage groups that were consistent with previously published white clover linkage maps (Barrett et al., 2004; Griffiths et al., 2013), we generated a 200-bp sequence for each marker based on the reference genome sequence 100 bp upstream and downstream of each SNP; these sequences were then used as BLAST queries against the reference genomes of the two progenitor species and Medicago truncatula. Linkage groups were assigned to either the O or P subgenome based on which progenitor species genome the most markers from a given linkage group matched with highest-scoring BLAST hits (Table S7). Within a subgenome, linkage groups were labeled to match previous published linkage maps based on shared synteny to M. truncatula chromosomes (Fig. S1) (Griffiths et al., 2013). The marker order for a given linkage group was inverted where necessary to match its direction in the previously published maps.

Genetic mapping

The Aelac and Lilli polymorphisms were mapped separately in each F₂ population to identify the locations of the corresponding loci in the DG and SG linkage maps. F₂ genotypes were not used in the genetic mapping analysis if they were missing >75% genotype data; this left 423 and 500 genotypes in the DG and SG populations, respectively (Tables S3, S4). Mapping analysis was performed with the R/qtl package (Broman et al., 2003). The scanone function was used to map quantitative trait loci (QTLs) for Aelac, Lilli and cyanogenesis. QTLs were considered significant if their LOD score was higher than a P=0.05 confidence threshold, generated for each phenotype from 1000 permutations of the data. For each QTL, LOD scores and the percentage variance explained (PVE) were calculated with the fitqtl function, and 1.5-LOD drop intervals were calculated with the lodint function. For cyanogenesis, which had two significant peaks, we fit a two-QTL model with refined positions from the refineqtl function.

Cyanogenesis gene sequence analysis

The previously published 138-kb Ac gene sequence (GenBank accession MH059954, comprising the three-gene CNglc metabolic cluster) was used in BLAST analyses on the NCBI platform, querying whole genome shotgun contigs (WGS database) of each diploid progenitor species to identify sequences with highest identity to the white clover sequence. Alignment to the best hit was visualized using the CoGe web platform (https://nomevolution.org/coge/G Evo.pl) with default parameter settings. Likewise, the 3.9-kb Li gene sequence was used as a query in BLAST analyses of the T. occidentale and T. pallescens genomes. To assess potential similarity in the genomic sequence flanking the Li locus to progenitor genome sequences, BLAST analyses were also performed using a larger 77.6-kb genomic sequence containing the Li gene that was derived from a white clover BAC library (Olsen & Small, 2018). To determine whether the Ac and Li loci are present in the T. repens shotgun genome assembly, both loci also were used as queries in BLAST analysis against that reference genome in the NCBI portal.

Results

White clover linkage maps

Full-sib crosses among F₁ plants of each parental cross (DMN_010 × GFL_007 and STL_0701 × GFL_007) yielded 502 and 500 F₂ individuals for the DG and SG populations, respectively. After filtering for the subset of SNPs that could be mapped with paired-end sequence reads and that were homozygous in both parents of a mapping population but heterozygous in F₁ plants (see Materials and Methods), the DG mapping population was characterized by 8095 polymorphic SNP markers and the SG population was characterized by 7665 polymorphic SNPs. Filtering out of any F₂ genotypes with <15× coverage yielded 252 lines in the DG population and 191 lines in the SG population that were used in linkage map construction (Tables S3, S4). For both datasets, linkage maps were composed of 16 linkage groups, consistent with white clover’s haploid chromosome number (n = 2x = 16), with eight linkage groups corresponding to the T. occidentale subgenome (chromosomes 1O to 8O, following the nomenclature of Griffiths et al., 2019) and the other eight corresponding to the T. pallescens subgenome (chromosomes 1P to 8P) (Fig. 1a,b).

The linkage map for the DG F₂ population included 2575 SNPs with a total length of 5057.6 cm and average marker interval of 2.0 cm (Table S8). Of these markers, 1104 mapped to the T. occidentale subgenome and 1471 mapped to the T. pallescens...
The linkage map for the SG F2 population included 2437 SNPs with a total length of 5815.6 cM and average interval of 2.4 cM (Table S9). Of these markers, 1045 were mapped to chromosomes belonging to the \textit{T. occidentale} subgenome and 1392 mapped to the \textit{T. pallescens} subgenome. The SNP markers in both mapping populations thus provide roughly equal representation of white clover’s two parental subgenomes.

**QTL mapping of cyanogenesis loci**

In both mapping populations, the \textit{Ac} gene mapped with high confidence to a single, highly localized genomic region on the top half of chromosome 2O (LOD = 85.2 and PVE = 54.2\% in DG; LOD = 120.7, PVE = 67.1\% in SG) (Figs 1a,b, S2, S3; Table 1). Likewise, \textit{Li} mapped to a single, highly localized region on the bottom of chromosome 4P in both populations, again with high confidence (LOD = 90.8 and PVE = 56.5\% in DG; LOD = 103.2, PVE = 61.3\% in SG) (Figs S4, S5). The identification of a single QTL peak for each cyanogenesis gene is consistent with the well-established simple Mendelian inheritance of the \textit{Acac} and \textit{Lili} polymorphisms. Correspondingly, the cyanogenesis phenotype mapped to the locations of the \textit{Ac} and \textit{Li} loci, and only to those two locations, with high confidence in both populations (> 70 LOD) (Figs S6, S7). Notably, the location of the \textit{Ac} locus within the O subgenome and the \textit{Li} locus within the P subgenome strongly suggests that white clover’s two cyanogenesis components were inherited from different diploid progenitors, with CNglc production derived from \textit{T. occidentale} and linamarase production inherited from \textit{T. pallescens}.

**Genome sequence analyses**

\textit{Ac locus} Sequence comparisons of the white clover \textit{Ac} locus to shotgun genome assemblies of the two progenitor species were

---

**Fig. 1** Linkage maps and cyanogenesis quantitative trait loci (QTL) locations in the DG (a) and SG (b) mapping populations in white clover (\textit{Trifolium repens}). Linkage groups are identified by subgenome chromosome number, where O and P designations correspond (respectively) to the \textit{Trifolium occidentale} and \textit{Trifolium pallescens} subgenomes. Horizontal lines indicate genetic map locations of genotyping-by-sequencing markers. Colored bars indicate locations and approximate span of significant QTLs for cyanogenic glucoside (CNglc) production (\textit{Ac}), linamarase production (\textit{Li}) and cyanogenesis.
consistent with genetic mapping results in suggesting inheritance from \( T. \) \( \text{occidentale} \). \text{BLAST} analysis revealed that much of the 138-\( kb \) \( Ac \) gene cluster, including both coding and noncoding sequence, was nearly identical to a single 169.2-\( kb \) region within the \( T. \) \( \text{occidentale} \) genome (≥97% sequence identity across 78.2% of the region) (Fig. 2; Table S10). All three genes of the CNglc biosynthetic pathway fell within highly conserved regions (97.5%, 99.2% and 99.6% sequence identity for \( \text{CYP79D15} \), \( \text{CYP736A187} \) and \( \text{UGT85K17} \) respectively), and synteny also was preserved across the entire gene cluster. No significant matches were detected for any of the CNglc synthesis genes outside of this 169.2-\( kb \) sequence in the \( T. \) \( \text{occidentale} \) genome. These results suggest that the white clover \( Ac \) locus is likely orthologous to, and thus descended from, the highly similar gene cluster in the \( T. \) \( \text{occidentale} \) genome.

By contrast with the high similarity detected in \( T. \) \( \text{occidentale} \), the white clover \( Ac \) sequence showed only negligible similarity to \( T. \) \( \text{pallescens} \) genomic sequences. Across the entire 138-\( kb \) \( Ac \) region, the longest \( T. \) \( \text{pallescens} \) sequence with ≥97% identity was a 692-bp fragment that corresponds to a previously described \( Tyl1/copia \)-like retrotransposon that occurs within the gene cluster (Olsen & Small, 2018) (Table S10). Among the three CNglc synthesis genes, the closest match was for a 173-bp portion of the 2.0-\( kb \) \( \text{CYP736A187} \) gene which showed 82.1% identity to a \( T. \) \( \text{pallescens} \) sequence. Taken together with the genetic mapping data and the \( T. \) \( \text{occidentale} \) sequence comparisons, these results strongly suggest that the white clover \( Ac \) locus is derived from \( T. \) \( \text{occidentale} \) and not from \( T. \) \( \text{pallescens} \).

**Li locus** Unlike inferences for \( Ac \), genome sequence analyses were inconclusive in identifying the ancestral sequence of the white clover \( Li \) gene. \text{BLAST} analysis of the 3.8-\( kb \) \( Li \) gene against the \( T. \) \( \text{occidentale} \) genome yielded a single best hit with 92.1% sequence identity. However, this \( T. \) \( \text{occidentale} \) sequence shows greater similarity (97.1% identity) to a previously described \( Li \)-paralog in the white clover genome that does not function in cyanogenesis (Olsen et al., 2007) (Table S10); the gene is, thus, more likely to be the \( T. \) \( \text{occidentale} \) ortholog of the non-cyanogenic \( Li \)-paralog than the \( Li \) gene. Likewise, \text{BLAST} queries of the \( T. \) \( \text{pallescens} \) genome with \( Li \) revealed a single best hit; as with \( T. \) \( \text{occidentale} \), the sequence showed higher nucleotide similarity to the white clover noncyanogenic \( Li \)-paralog than to \( Li \) (96.0% vs 94.6% identity, respectively; Table S10). \text{BLAST} searches using a larger 77.6-\( kb \) sequence containing the \( Li \) gene also did not yield any obvious close matches within either diploid genome, with best hits for >1-\( kb \) fragments showing ≤86% identity to both species. Thus, although genetic mapping of the white clover \( Lilli \) polymorphism indicates its placement in the \( T. \) \( \text{pallescens} \) subgenome, the \( Li \) gene sequence does not appear to be present in the published reference genome of that diploid progenitor species. This result suggests that present-day \( T. \) \( \text{pallescens} \) either lacks the \( Li \) gene altogether or that the species is polymorphic for \( Lilli \), with the published genome derived from a \( lilli \) (gene-deletion homozygote) genotype. Consistent with both of these possibilities, cyanogenesis genotyping and phenotyping of the \( T. \) \( \text{pallescens} \) accession used to create the reference genome (AZ 1895) indicated that the \( Li \) gene is not present in that genotype (Table S1).

**Trifolium repens reference genome comparisons** \text{BLAST} analysis of the \( Ac \) locus against the \( T. \) \( \text{repens} \) reference genome confirmed the presence of the gene cluster in the accession used for genome sequencing (100% identity for a 139.1-\( kb \) region; Table S10). The \( T. \) \( \text{repens} \) contig containing the \( Ac \) locus currently is assigned to \( T. \) \( \text{repens} \) chromosome 1P; given the lack of congruence with genetic map data, where \( Ac \) is placed unambiguously at the top of 2O (Fig. 1; Table 1), this designation likely is an artifact of contig misassignment in the current shotgun-assembly draft genome. The \( Li \) locus also was confirmed to be present in the white clover reference genome (100% identity for the 3.8-\( kb \) gene; Table S10). In this case, the \( Li \)-containing region appears to be assigned to the correct chromosome in the draft assembly; it is annotated as a chromosome 4P sequence, consistent with QTL mapping results (Fig. 1; Table 1).

**Discussion**

Two of the most characteristic features of plant species diversification are speciation via allopolyploidy and the evolution of diverse chemical defense metabolites. Here we have examined
how the former phenomenon gave rise to a textbook example of the latter, namely cyanogenesis in the allotetraploid species white clover (Trifolium repens). Through the integration of linkage maps and genetic mapping of the white clover presence/absence of cyanogenic glucosides (CNglcs) (Ac/ac) and of linamarase (Li/li) polymorphisms with genome sequence analyses in its two diploid progenitor species, we have documented a dual-species origin for the chemical defense in this allotetraploid species. Specifically, the synthesis of CNglcs, which requires the three-gene metabolic gene cluster comprising the Ac locus, has been contributed by T. occidentale (Fig. 1a,b; Table 1); sequence comparisons between T. repens and T. occidentale reveal a high degree of conservation across the entire gene cluster (Fig. 2). For the hydrolyzing enzyme, linamarase, which is required to liberate HCN from CNglcs, genetic mapping indicates that the underlying Li gene has an origin in the T. pallescens genome, although the gene is absent in the genotype used to generate the published reference genome (Table S1). Our confidence in these evolutionary inferences is bolstered by the fact that white clover’s two subgenomes show very high genome integrity with their respective subgenomes and very limited interhomeologous recombination since polyploidization (Griffiths et al., 2019); this high degree of subgenome conservation suggests that the linkage map data and genome sequence comparisons are both accurately reflecting the cyanogenesis genes’ evolutionary origins. Support for these conclusions is further strengthened by the very high degree of similarity between our Trifolium–Medicago synteny plots (Fig. S1) and those reported in previous analyses (e.g. Griffiths et al., 2019). Below we discuss these findings in the context of the white clover cyanogenesis polymorphism, Trifolium evolution, and allopolyploid adaptation more broadly.

Cyanogenic glucosides and the Ac locus

As a chemical defense that requires physical interaction between two normally separated biochemical precursors, cyanogenesis has been described as a ‘cyanide bomb’ (Morant et al., 2008) that is triggered by chewing herbivores. By this analogy, the ‘explosives’ component of the white clover cyanide bomb (i.e. CNglcs) was contributed by the diploid progenitor T. occidentale. This coastal western European species is not cyanogenic but does carry the Ac/ac polymorphism for CNglc production (Kakes & Chardonnens, 2000; Olsen et al., 2014). T. occidentale is not atypical among white clover’s close congeners in this respect; several other Trifolium species also have been found to be polymorphic for CNglc production while lacking the capacity to produce linamarase (e.g. T. ambiguum, T. montanum, T. suffucatum, T. uniflorum) (Gibson et al., 1972; Olsen et al., 2014). The most likely explanation for this pattern is that CNglcs in these species serve functions in primary metabolism that are unrelated to cyanide release and herbivore deterrence. CNglcs are produced in >3000 plant species across >130 families, and research over the last two decades has established important roles for them in physiological processes that do not involve cyanide release – most notably as storage and transport molecules for nitrogen (N) and glucose (Gleadow & Møller, 2014; Pičmanová et al., 2015). For white clover evolution, it thus appears that the antiherbivore cyanogenic response evolved in part through co-option of CNglc production for a new adaptive role as a chemical defense.

When functioning as storage and transport molecules, CNglcs provide readily metabolized reserves of reduced-state N; these N reserves can be especially beneficial for plants growing in drought-prone or other stressful environments (Gleadow & Møller, 2014; Kooyers et al., 2014). In the case of white clover, where climate-associated Ac/ac clines are well-documented throughout the species range (Daday, 1954, 1958; Kooyers & Olsen, 2012, 2013; Olsen et al., 2013), both observational and experimental evidence suggest that some of this adaptive variation probably reflects selective pressures unrelated to cyanogenesis – particularly, plant fitness in drought-prone environments (Kooyers et al., 2014). Thus, it seems likely that the presumed ancestral function of CNglcs in primary metabolism was maintained in white clover even after they gained their well-documented adaptive function in herbivore deterrence.

Genome sequence comparisons between white clover and its diploid progenitors suggest that this is a very young polyploid species, having emerged around the time of the last major European glaciation c. 20 000 yr ago, when the progenitor species are expected to have had regions of range overlap in glacial refugia (Griffiths et al., 2019). Consistent with this scenario, the white clover Ac locus shows very high sequence conservation with its...
T. occidentale ortholog; the three CNglc synthesis genes and most intervening noncoding sequence show ≥ 97% identity on average between species (Fig. 2). These similarity estimates would likely be even higher with a more complete version of the T. occidentale reference genome, as the current shotgun assembly has multiple regions of undetermined sequence within the Ac locus (see cross-hatched zones in Fig. 2).

Genes such as Ac that comprise metabolic clusters generally are believed to under strong selection to remain in close physical proximity within the genome to ensure inheritance of the complete biosynthetic pathway, thereby minimizing risk of accumulating toxic pathway intermediates (Takos & Rook, 2012; Gleadow & Møller, 2014). In a previous study of the white clover Ac cluster, we found that the three component genes were orthologous to and highly conserved with those of a CNglc cluster in the cyanogenic species Lotus japonicus (birdfoot trefoil), which last shared an ancestor with Trifolium c. 50 Myr ago (Olsen & Small, 2018). Notably, on the timescale of that much older evolutionary divergence, there was essentially no conservation of the cluster outside the three highly conserved CNglc genes; gene order, transcript orientation and intervening sequences were all completely diverged. This pattern suggests that there is little overall selective constraint on noncoding portions of the CNglc metabolic cluster. In this light, the very high similarity we observe for the entire gene cluster between T. repens and T. occidentale (Fig. 2) further supports our conclusion that the white clover Ac locus is directly and recently descended from that progenitor species.

White clover is characterized by high genetic diversity despite its recent hybrid origin, and coalescent modeling has suggested that the species likely evolved through multiple independent T. pallescens × T. occidentale hybridization events (Griffiths et al., 2019). Given that T. occidentale carries the Aclac polymorphism, independent hybridizations involving Ac- and acac genotypes of that species could easily account for the origin of the white clover Aclac polymorphism. If true, this would mean that the gene-presence and -absence alleles that characterize the well-studied Aclac polymorphism in white clover pre-date the origin of the species, making it a case of trans-specific adaptive polymorphism. However, molecular evolutionary analyses of the Ac locus and flanking sequences indicate that the genomic deletions that characterize ac alleles have evolved repeatedly, both in white clover and in its congeners that carry the Aclac polymorphism (Olsen et al., 2013, 2014). Thus, genomic instability around the Ac locus may lend itself to repeated evolution of species-specific ac alleles even if the original white clover Aclac polymorphism was inherited directly from T. occidentale.

Linamarase and the Li locus

Although genetic mapping places the white clover Li locus unambiguously in the T. pallescens subgenome (linkage group 4P; Fig. 1; Table 1), the absence of the gene from the T. pallescens reference genome prevents sequence comparisons between the species. The gene’s absence can most easily be explained in one of two ways. T. pallescens may carry the Lili gene presence/absence polymorphism, as occurs in white clover and other closely related species (e.g. T. isthmocarpum, T. nigrescens, Olsen et al., 2014), with the reference genome representing a lili genotype. Alternatively, the gene may have been present at the time that T. pallescens hybridized with T. occidentale to create T. repens, but later lost in the T. pallescens lineage. Several factors favor the second explanation. First, neither linamarase nor the underlying Li gene have ever been identified in T. pallescens in surveys to date (Gibson et al., 1972; Olsen et al., 2014) (Table S1). Equally notably, CNglcs have not been detected in the species either, and linamarase has no known function other than in CNglc hydrolysis to release HCN; it thus would be implausible for a species to produce the CNglc-specific hydrolytic enzyme without CNglcs also being produced in at least some individuals. The restricted distribution of T. pallescens within cool alpine habitats further argues against the gene’s present-day occurrence in the species. Cyanogenic Trifolium species (or cyanogenic populations of polymorphic species) occur primarily in warm temperate and hot Mediterranean climates (Olsen et al., 2013, 2014); T. pallescens would thus be a unique exception in the genus if it were an alpine-adapted cyanogenic species. Based on all of these factors, the most likely scenario may be that T. pallescens was once cyanogenic (or polymorphic for cyanogenesis), and that this trait was lost during the species’ post-Pleistocene range contraction into its present-day alpine environment.

It should be noted that a third potential explanation also exists for the occurrence of the Li gene in T. repens which we cannot exclude with our present data – namely, adaptive introgression from some other Trifolium species that carries this gene. Despite differences in ploidy between white clover and the two other Trifolium species that are known to produce linamarase (T. nigrescens and T. isthmocarpum, both 2n = 2x = 16), each of these species can be hybridized with T. repens under controlled conditions, and crosses with T. nigrescens can produce fertile hybrids (Kazimierski & Kazimierska, 1972; Abberton, 2007). Thus, although it is not directly suggested by our genetic map data, an adaptive introgression scenario for the origin of Li in white clover also remains within the realm of possibility.

Allopolyploidy and cyanogenesis evolution in white clover

Speciation via allopolyploidy is widely recognized as a potential source of evolutionary novelty, with the combination of two or more parental species’ genomes providing opportunities for dynamic remodeling of genome organization, gene content, gene expression networks and transposon activity (Soltis et al., 2014). These changes provide a novel genomic toolkit that can potentially facilitate adaptation and ecological niche expansion (reviewed in Van de Peer et al., 2017). White clover may well be an illustration of this phenomenon (Griffiths et al., 2019), as the species occurs abundantly worldwide, in climates ranging from boreal to subtropical, despite its parental species’ narrow ecological and geographical distributions. Although the role of cyanogenesis in white clover’s ability to expand into such a wide environmental range is unknown, the fact that climate-associated cyanogenesis clines have evolved repeatedly throughout the
species range (reviewed by Kooyers & Olsen, 2012; Olsen et al., 2013) suggests that selection on this trait plays an important role in the species’ wide ecological amplitude. At the same time, it should be noted that cyanogenesis is unlikely to have been a novel trait that evolved uniquely in *Trifolium* with the origin of *T. repens*; as discussed above, several congeners also are cyanogenic. In this context, the dual-species origin of white clover cyanogenesis that we have documented in this study may best be viewed as one feature of the newly evolved species that, in combination with other emergent features of its allopolyploid toolkit, facilitated its global expansion and ecological success.

**Acknowledgements**

The authors thank Mike Dyer and other staff of the Washington University greenhouse facility for providing plant care; Andrew Griffiths and Roger Moraga (AgResearch NZ) for reference genome sequence information; Rachel Tan (AgResearch NZ) and Jennifer Barrett (Danforth Plant Science Center, St Louis) for guidance on genotyping-by-sequencing protocols; Marshall Wedger (Washington University) for assistance with QTL mapping scripts; and members of the Olsen lab group for helpful comments on the manuscript. Funding for the work was provided by NSF grants IOS-1557770 to KMO and DEB-1601641 to SJW, and by NSF graduate research fellowship DGE-1143954 to SJW. The authors declare no competing interests.

**Author contributions**

KMO conceived the project, analyzed sequence data, interpreted the results and wrote the manuscript; SJW created the mapping populations, generated GBS genotype data and contributed to writing; L-FL and DMG performed bioinformatics; DMG also constructed linkage maps and performed QTL mapping and subgenome data interpretation; and MLD, SRM and LLS performed cyanogenesis genotyping and phenotyping. All authors contributed comments on earlier drafts of the manuscript. DMG and SJW contributed equally to this work.

**ORCID**

David M. Goad https://orcid.org/0000-0001-8658-6660
Kenneth M. Olsen https://orcid.org/0000-0002-8338-3638

**Data availability**

Genotyping-by-sequencing reads have been submitted to NCBI SRA as aligned BAM files. The STL × GFL population is available under project ID PRJNA740324 (accessions SAMN19839204–SAMN19839729). The DMN × GFL population is under project ID PRJNA740343 (accessions SAMN19841524–SAMN19842038). Marker information and other genetic map data are provided in the online supplementary materials (Tables S8, S9).

**References**


Table S2 Percentage sequence identity between a sample of *Tri-
folium occidentale* and *Trifolium pallescens* orthologs.

Table S3 Parentage and cyanogenesis genotyping for the DG
mapping population.

Table S4 Parentage and cyanogenesis genotyping for the SG
mapping population.

Table S5 Primer combinations for *Ac* and *Li* genotyping, and
GBS PCR primers for final library amplification and Illumina
sequencing.

Table S6 White clover GBS common and barcode adapters.

Table S7 Number of markers on each linkage group with best
BLAST hit to either the *Trifolium occidentale* or *Trifolium pallescens*
reference genomes.

Table S8 GBS marker data for the DG population.

Table S9 GBS marker data for the SG population.

Table S10 GenBank accessions of best BLAST hits for white clover
cyanogenesis loci in the *Trifolium occidentale*, *Trifolium pallescens*
and *Trifolium repens* draft genomes.

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.