BMC Biology

RESEARCH ARTICLE **Open Access**

De novo genome assembly of white clover (*Trifolium repens* L.) reveals the role of copy number variation in rapid environmental adaptation

Wen-Hsi Kuo[1], Sara J. Wright[1,2], Linda L. Small[1] and Kenneth M. Olsen[1*]

Abstract

Background White clover (*Trifolium repens*) is a globally important perennial forage legume. This species also serves as an eco-evolutionary model system for studying within-species chemical defense variation; it features a well-studied polymorphism for cyanogenesis (HCN release following tissue damage), with higher frequencies of cyanogenic plants favored in warmer locations worldwide. Using a newly generated haplotype-resolved genome and two other long-read assemblies, we tested the hypothesis that copy number variants (CNVs) at cyanogenesis genes play a role in the ability of white clover to rapidly adapt to local environments. We also examined questions on subgenome evolution in this recently evolved allotetraploid species and on chromosomal rearrangements in the broader IRLC legume clade.

Results Integration of PacBio HiFi, Omni-C, Illumina, and linkage map data yielded a completely de novo genome assembly for white clover (created without a priori sequence assignment to subgenomes). We find that white clover has undergone extensive transposon diversification since its origin but otherwise shows highly conserved genome organization and composition with its diploid progenitors. Unlike some other clover species, its chromosomal structure is conserved with other IRLC legumes. We further find extensive evidence of CNVs at the major cyanogenesis loci; these contribute to quantitative variation in the cyanogenic phenotype and to local adaptation across wild North American populations.

Conclusions This work provides a case study documenting the role of CNVs in local adaptation in a plant species, and it highlights the value of pan-genome data for identifying contributions of structural variants to adaptation in nature.

Keywords Allopolyploid, Copy number variation (CNV), Cyanogenesis, Haplotype-resolved genome, Inverted repeat-lacking clade (IRLC) legumes, Karyotype, Local adaptation, Pangenome, Subgenome, White clover (*Trifolium repens*)

Background

White clover (*Trifolium repens* L., 2n = 4x = 32) is the most widely grown temperate forage legume worldwide due to its superior nitrogen-fixing ability, perenniality, and resilience in frequently disturbed habitats [1–3]. Prior to the twentieth century invention of synthetic nitrogen fertilizer, this species was so important as a source of soil nitrogen that it was considered the

*Correspondence:
Kenneth M. Olsen
kolsen@wustl.edu
[1] Department of Biology, Washington University in St. Louis, St. Louis, MO 63130, USA
[2] Present address: Department of Biological and Biomedical Sciences, Rowan University, Glassboro, NJ 08028, USA

agricultural equivalent of coal in fueling agricultural expansion during the Industrial Revolution [4]. In addition to its longstanding agricultural importance, white clover is also a well-known eco-evolutionary model system for studying how selection maintains within-species chemical defense variation. The species is characterized by a genetically well characterized polymorphism for cyanogenesis (HCN release following tissue damage); multiple studies over the last 80 years have established that climate-associated clines in cyanogenesis evolve rapidly across latitudinal and other environmental gradients worldwide, with cyanogenic plants generally predominating in warmer locations [5–8]. While native to Europe, white clover has become widely naturalized in mesic environments worldwide (from subtropical to boreal climates), and recent genomic research has linked its ability to become so widely adapted to its allopolyploid origin from two ecologically distinct diploid progenitor species (*T. occidentale* and *T. pallescens*) [9].

White clover's importance both in agriculture and as an eco-evolutionary model system has spurred recent efforts to develop high-quality genetic and genomic resources for the species [9–13]. Most recently, two independent genome sequencing projects have utilized long-read sequencing technologies to overcome the scaffolding difficulties associated with polyploid genome assembly [11, 12]. While these two high-quality genomes are a major step in white clover studies, the recent growth of plant pangenome projects (e.g., [14–16]) has demonstrated clearly that multiple high-quality reference genomes are required to properly understand the genomic structure and intraspecific diversity of a species; this is especially true when patterns of genomic structural variation contribute to natural phenotypic variation. In addition, high-quality reference genomes can facilitate investigations of genome evolution on a macroevolutionary scale. For white clover specifically, genome sequence comparisons between this species and related legume genera could fill a current knowledge gap concerning karyotype evolution in the inverted repeat-lacking clade (IRLC) of the legume family, which contains multiple economically important crop species, such as soybean and chickpea, and is characterized by extensive genome rearrangements [17].

Among the many types of structural variation that can occur in a genome, copy number variation (CNV), defined as the variable repetition of specific sequence motifs ranging from 50 bp to several Mbp, is a major contributor to both genetic and phenotypic variability across eukaryotes [18–22]. Despite abundant evidence that CNVs are pervasive in genomes and that they contribute to phenotypic variation, such as in domesticated crops [19, 23, 24], there are remarkably few documented cases where CNVs have been shown to underlie adaptation in wild species [25–27]. To the best of our knowledge, the only potential example in a plant species involves flowering time regulation in common waterhemp, where CNVs of an ATP synthesis pathway locus appear related to natural flowering time variation [25]. Consequently, our knowledge about the role of CNVs in shaping local adaptation in wild populations is lacking, particularly for plant species. Even in the model plant Arabidopsis, studies of CNVs have been restricted to either documenting genome-wide CNV distributions without knowledge of associated phenotypes (e.g., [16]), or to functional characterization of the phenotypic impact of CNVs without data from natural populations to assess their role, if any, in adaptation (e.g., [28]).

Based on our prior knowledge of white clover and the genetic basis of adaptation in this species, we hypothesized that local environmental adaptation in wild populations could arise, in part, through CNVs, particularly at the loci known to control the well-documented cyanogenesis polymorphism. This chemical defense polymorphism is controlled by two independently segregating simple Mendelian genetic polymorphisms that determine the presence or absence of two cyanogenic chemical precursors, both of which must be present for a plant to produce the cyanogenic phenotype: (1) *Ac/ac*, controlling the presence/absence of cyanogenic glucosides, and (2) *Li/li*, controlling the presence/absence of their hydrolyzing cyanogenic β-glucosidase enzyme, linamarase (reviewed by [29]). At the molecular level, *Ac* is a 3-gene metabolic cluster on chromosome 2 that comprises the cyanogenic glucoside biosynthetic pathway, while *Li* is a single gene located on chromosome 12 that encodes linamarase [13, 30–32]. Interestingly, the *Ac* gene is only present in the *T. occidentale* subgenome, and the *Li* gene is only present in the *T. pallescens* subgenome, which suggests that the two genes in white clover were contributed by its two different diploid progenitors [13]. For both loci, the recessive (nonfunctional) alleles are the result of gene deletions, meaning that both the *Ac/ac* and *Li/li* polymorphisms are gene presence/absence variations (PAVs) [30, 32, 33]. However, while the recessive *ac* and *li* alleles have consistently been shown to be gene deletions [34, 35], our observations of *Ac* and *Li* dominant allele inheritance in greenhouse pedigree populations suggested the occurrence of CNVs at both loci (see the "Results" section). Given that variability in the cyanogenesis phenotype is known to contribute to white clover's adaptation across climatic gradients [5, 6, 29, 36], we therefore hypothesized that CNVs at the *Ac* and *Li* loci could play an important role in this adaptive response.

In this study, we tested the hypothesis that CNVs at the *Ac* and *Li* cyanogenesis loci contribute to natural adaptive variation in white clover. We used PacBio HiFi, Omni-C, and two linkage maps to generate and de novo assemble a new chromosome-scale and haplotype-resolved genome. We then compared structural variation at the cyanogenesis loci among three high-quality white clover genomes (this study; [11, 12]). Next, we assessed *Ac* and *Li* CNV occurrence and distributions across wild populations spanning much of the North American species range (419 accessions across 43 locations). In a complementary analysis, we examined the contribution of CNVs to natural phenotypic variation by assessing the relationship between *Ac* CNVs, gene expression, and cyanogenic glucoside content. We further tested whether CNVs at the cyanogenesis loci contribute to adaptation in nature by conducting association analyses between the CNVs and the local environments of the sampled wild populations. Finally, we leveraged our high-quality genome to address questions concerning karyotype evolution in the clover genus *Trifolium* and related genera in the “IRLC clade” of the legume family (Fabaceae). To our knowledge, this study is among the first to document the contribution of CNVs to local environmental adaptation in a wild plant species, and it also provides new insights into the history of genome rearrangements in an economically important clade of legume species.

Results

Genome assembly statistics

We isolated DNA from a wild North American white clover accession (GFL_007) to assemble a haplotype-resolved white clover genome using a combination of PacBio HiFi, Omni-C, and linkage map technologies. The total length of the primary haplotig was 995,597,458 bp and contained 1,033 contigs ($N_{50}=17{,}751{,}673$ bp). After Omni-C and linkage map scaffolding, 94.529% of the haplotig sequences were placed into 16 linkage groups, corresponding to the 16 chromosomes in white clover ($N_{50}=57{,}173{,}913$ bp). The total length of the alternative haplotig was of similar size (1,002,960,897 bp) and contained 584 contigs ($N_{50}=15{,}995{,}289$ bp); after scaffolding (performed independently of the primary haplotig), 95.546% of the alternative haplotig sequences were placed into 16 chromosome scale linkage groups ($N_{50}=60{,}046{,}767$ bp). An independent Illumina-sequenced dataset (2×150 bp, 185 million read pairs; SRR27541471) of the same plant accession was used for genome size estimation and quality assessment. The estimated short-read haploid genome size was similar to the PacBio HiFi assembly (1,025,512,829 bp), with 1.46% heterozygous sites (k-mer length = 21) (Additional file 4: Fig. S1). The consensus quality (QV) of the diploid genome was 51.3697, indicating > 99.99% accuracy in the consensus base calls. The k-mer completeness was 98.4765%.

Approximately 60% of the genome assembly was made up of repeated elements; half of these (30% of the genome) were retroelements, which closely matches retroelement proportions in the two recently published long-read assemblies [11, 12]. The retroelements did not show any biased distribution between the two white clover subgenomes. Furthermore, our phylogenetic analyses revealed several major, recent retroelement diversification events in the white clover genome (Additional file 4: Fig. S2, Fig. S3). These patterns are consistent with transposable element (TE) proliferation following the allopolyploid speciation origin of white clover, a phenomenon that has been reported in other allopolyploid systems [37]. Unfortunately, due to the lack of high-quality genomes of the two diploid progenitor species (*T. occidentale* and *T. pallescens*), a direct test of this TE proliferation hypothesis is currently unfeasible.

BUSCO assessments indicated that 98% of the single-copy orthologous genes of the fabales_odb10 database (5366 genes) were covered in our assembly. Of these, 82% were found to be duplicated, which is consistent with the allotetraploid origin of white clover. More detailed genome assembly statistics are provided in Additional file 1: Table S1. The quality of the separation of the two haplotigs was assessed by comparing the k-mer distributions between the raw reads, the primary haplotig and the alternative haplotig (Fig. 1A, B). The k-mer multiplicity distributions of the primary and alternative haplotigs show nearly complete overlap (indicated by purple shading in Fig. 1A), suggesting a highly accurate separation of the genome’s two haplotigs. The k-mer distribution of the diploid genome indicates that most k-mers were found only once (Fig. 1B), consistent with the high heterozygosity of the genome. The much lower peaks of the 3- or 4-hit distributions compared to 1- or 2-hit distributions indicate that there is divergence between homeologous sequences of the two subgenomes; this suggests that while the genome is tetraploid, it likely behaves meiotically as a diploid species.

Omni-C and linkage map scaffolding and subgenome characterization

The two haplotigs of the contiguous genome were first scaffolded by Omni-C linked reads and then scaffolded by our two previously published linkage maps (DG and GS F_2 mapping populations; 6,173 unique GBS SNP markers after combining the two mapping populations) [13]. Where there were conflictions between the Omni-C and linkage map scaffolding results, we prioritized the results from the linkage maps (see the “Discussion” section). The read count distribution of the chromatin contact map of the scaffolded genome supports our finding that our 16 scaffolds (pseudomolecules) are high-quality

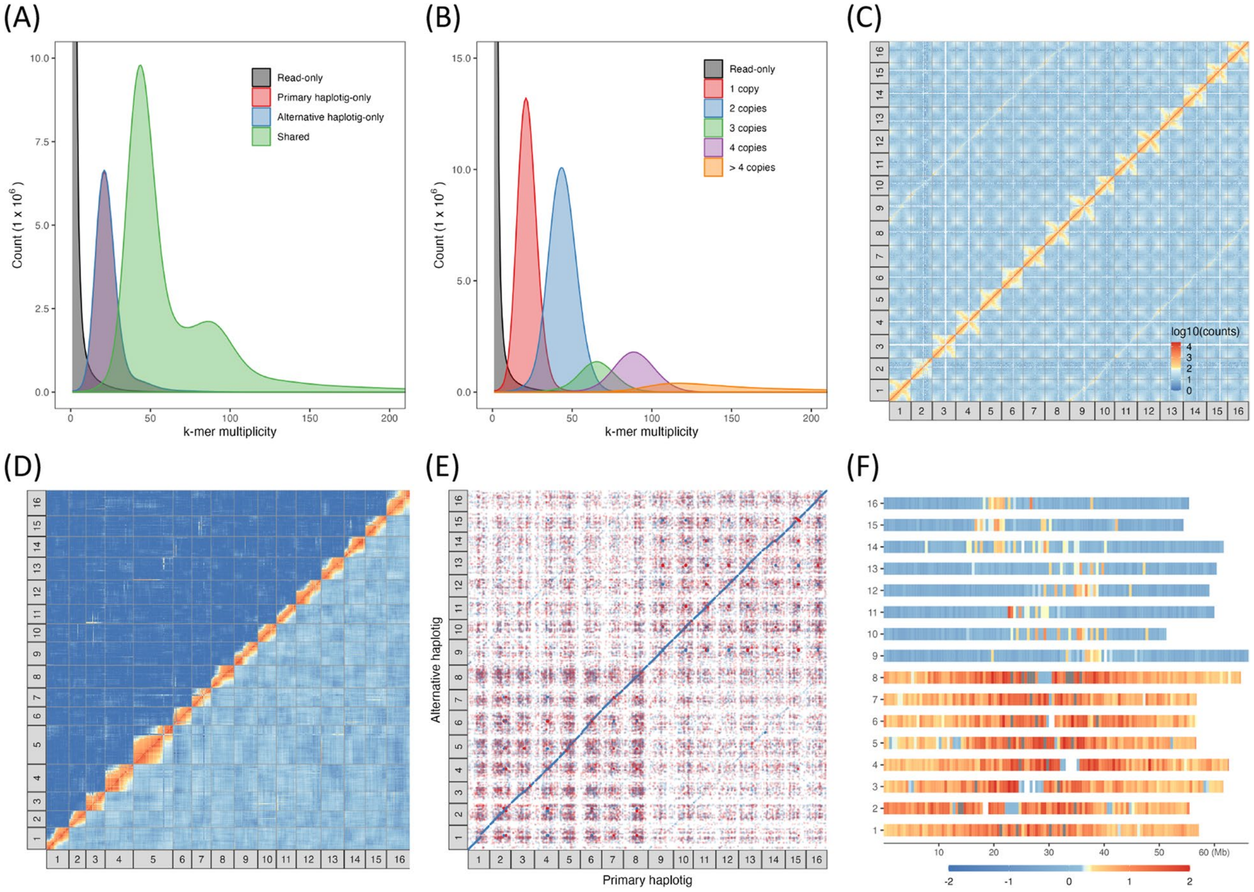

Fig. 1 Genome quality assessment and subgenome identity characterization. **A** k-mer multiplicity distribution between the primary and alternative haplotigs. Purple color indicates the overlap between the primary and alternative haplotigs, reflecting their nearly identical distributions. **B** The number of the times of the k-mers are found in the diploid genome. **C** Omni-C heatmap visualized by KR normalization method with window size at 1 Mbp. **D** Linkage map visualized by calculating the LOD (upper triangle) and the recombination fraction (RF, lower triangle) of the genetic markers of the DG F_2 mapping population at their physical location. Only the primary haplotig is presented here. **E** Whole genome nucleotide alignment of the primary and alternative haplotigs (identity > 95%, query length > 10,000 bp, hit length > 10,000 bp). **F** Chromosomes categorized into *T. occidentale* (To) or *T. pallescens* (Tp) subgenomes. The gradient color is the mapping rate to either diploid progenitor, where − 1 is ideally completely *T. pallescens*-like and + 1 is ideally completely *T. occidentale*-like; values exceeding that range indicate mapping to multiple locations in the diploid progenitors' genomes

representatives of the 16 white clover chromosomes (Fig. 1C). Although we do see weak signals between the homeologous chromosomes (e.g., between chr 1 and chr 9), we believe this result can be attributed to the highly similar gene sequences (see BUSCO analyses in Additional file 1: Table S1) and inaccurate read-mapping of Omni-C short-reads (2 × 150 bp). In addition, the inter-homeologous signals are not evident in the recombination map (Fig. 1D), as most of the GBS markers used in linkage map construction are in the non-genic regions, which are more divergent between the homeologous chromosomes.

The subgenome identities of the 16 scaffolds/chromosomes were assessed using BLAST against the two diploid progenitors' genomes. Across the board, there were more hits to the *T. occidentale* genome than to the *T. pallescens* genome (Fig. 1F). This is consistent with previous evidence that present-day individuals of *T. occidentale* are genomically highly similar to the ancestral species that contributed one of white clover's subgenomes (chr 1–chr 8), whereas present-day representatives of *T. pallescens* are somewhat diverged from the ancestral species that contributed the other subgenome (chr 9–16) [38]. When setting the mapping threshold at 0.45 (where − 1 is completely *T. pallescens*-like and + 1 is completely *T. occidentale*-like; see the "Methods" section), we were able to categorize the 16 chromosomes into eight of *T. occidentale* origin (chr 1–chr 8) and eight of *T. pallescens* origin (chr 9–chr 16) (Fig. 1F and Additional file 4: Fig. S4). We detected no obvious inter-chromosomal translocation signal, except for at the center of the chromosomes where the signals were ambiguous; this is likely because these regions include centromeric long tandem repeats, which can interfere with the mapping process. However, we cannot validate this conclusion due to the limited resolution of the diploid progenitors' genomes [9].

As predicted, our assessments of haplotig similarity using our whole genome nucleotide alignment revealed clear alignment hits between the primary and alternative haplotigs (indicated by the major blue diagonal line in Fig. 1E). In contrast, hits between homeologous loci are much less evident (the two weak blue diagonal lines in the top-left and bottom-right quadrants of Fig. 1E). We also found that there are generally more intra-subgenome hits than inter-subgenome hits (Fig. 1E and Additional file 1: Table S2). Together, these observations suggest that inter-subgenome recombination (i.e., homeologous translocation) was very limited after the polyploidization event that created white clover and that the integrity of the two parental species' subgenomes has been preserved, in agreement with the same conclusion inferred from the k-mer analysis above and previous analysis [9].

Gene synteny and synonymous mutation rate

We annotated 96,293 protein coding genes in the primary haplotig, which is slightly more than double the number of genes in the diploid genome of the related legume *Medicago truncatula* (44,295); this finding is consistent with white clover's tetraploid genome. A total of 61,352 of the annotated genes were successfully mapped by BLAST against the NCBI RefSeq database and functionally annotated with the GO database. The median lengths of the concatenated coding sequence (CDS) and intron sequence per gene were 927 bp and 1693 bp, respectively, which is comparable to closely related legume species (Additional file 4: Fig. S5). The gene synteny between the two subgenomes was especially high (Fig. 2A), suggesting the lack of any large-scale genome rearrangements in white clover or its closest relatives either before or after the polyploidization event that produced the species. The orthologous synonymous rate (K_S) analysis showed a neutral genetic distance to the other species, reflecting the species-level phylogenetic relationships (Fig. 2B). Interestingly, the K_A/K_S distribution for the two white clover subgenomes [*T. repens* (occ) vs. *T. repens* (pall)] was statistically indistinguishable from the K_A/K_S of the two diploid progenitor species [*T. pallescens* vs. *T. occidentale*] (Additional file 4: Fig. S6). This pattern suggests that, following the polyploidization event that created white clover, the homeologous gene copies within the tetraploid genome have exhibited no detectable sign of relaxed purifying selection, despite the high likelihood of functional redundancy among the many gene duplicates.

Interspecific gene synteny analysis indicated that the white clover genome is highly syntenic to the closely related species *Medicago truncatula* (24.53 MYr divergence) and moderately syntenic to the more distantly related *Cicer arietinum* (32.01 MYr divergence) (Fig. 2C). There were only two chromosome-scale translocations between *M. truncatula* and white clover: between *Mt* chr 2 and *Tr* chr 6 and between *Mt* chr 4 and *Tr* chr 8 (Fig. 2C and Additional file 4: Fig. S7E, F). Given the relatively high synteny between these species in different genera, we were surprised to detect multiple genome rearrangements within the genus *Trifolium* when comparing white clover with its two congener species with reference genomes, *T. subterraneum* and *T. pratense* (12.97 MYr divergence). *Trifolium subterraneum* has the same basal chromosome number as white clover ($n=8$); however, this lineage appears to have undergone at least 2 or 3 chromosome-scale translocations in chr 2, chr 3, chr 4, chr 6, chr7, and chr 8 (Fig. 2C and Additional file 4: Fig. S7A, B). *Trifolium pratense*, in comparison, appears to have lost one chromosome after divergence from the *T. subterraneum* lineage, and it shows more complex patterns of genomic rearrangements (Fig. 2C and Additional file 4: Fig. S7C, D). Based on the phylogeny and the gene synteny analyses, we therefore infer that among the three *Trifolium* species with reference genomes, white clover preserves the more ancestral genome structure, and that there have been extensive rearrangements among its diploid congeners.

(See figure on next page.)

Fig. 2 Genome assembly of white clover. **A** Circular view of genomic feature distribution along the 16 chromosomes. The blue, green, and brown tracks stand for the GC content, gene density, and repetitive element density, respectively. The links show the gene synteny between the two subgenomes, where chromosomes 1–8 (light green) and chromosomes 9–16 (light blue) belong to the *T. occidentale* (To) and *T. pallescens* (Tp) subgenomes, respectively. **B** Pairwise synonymous mutation rate (K_S) between the subgenomes and closely related species in the legume family (Fabaceae, subfamily Papilionoideae). **C** Gene synteny plot and consensus gene tree of the inverted repeat-lacking clade (IRLC) within subfamily Papilionoideae. The node numbers show the crown ages in Myr, which are adapted from [39]. For visualization purposes, only the synteny of the chromosome 1–8 was plotted. The plot of the chromosome 9–16 is highly similar

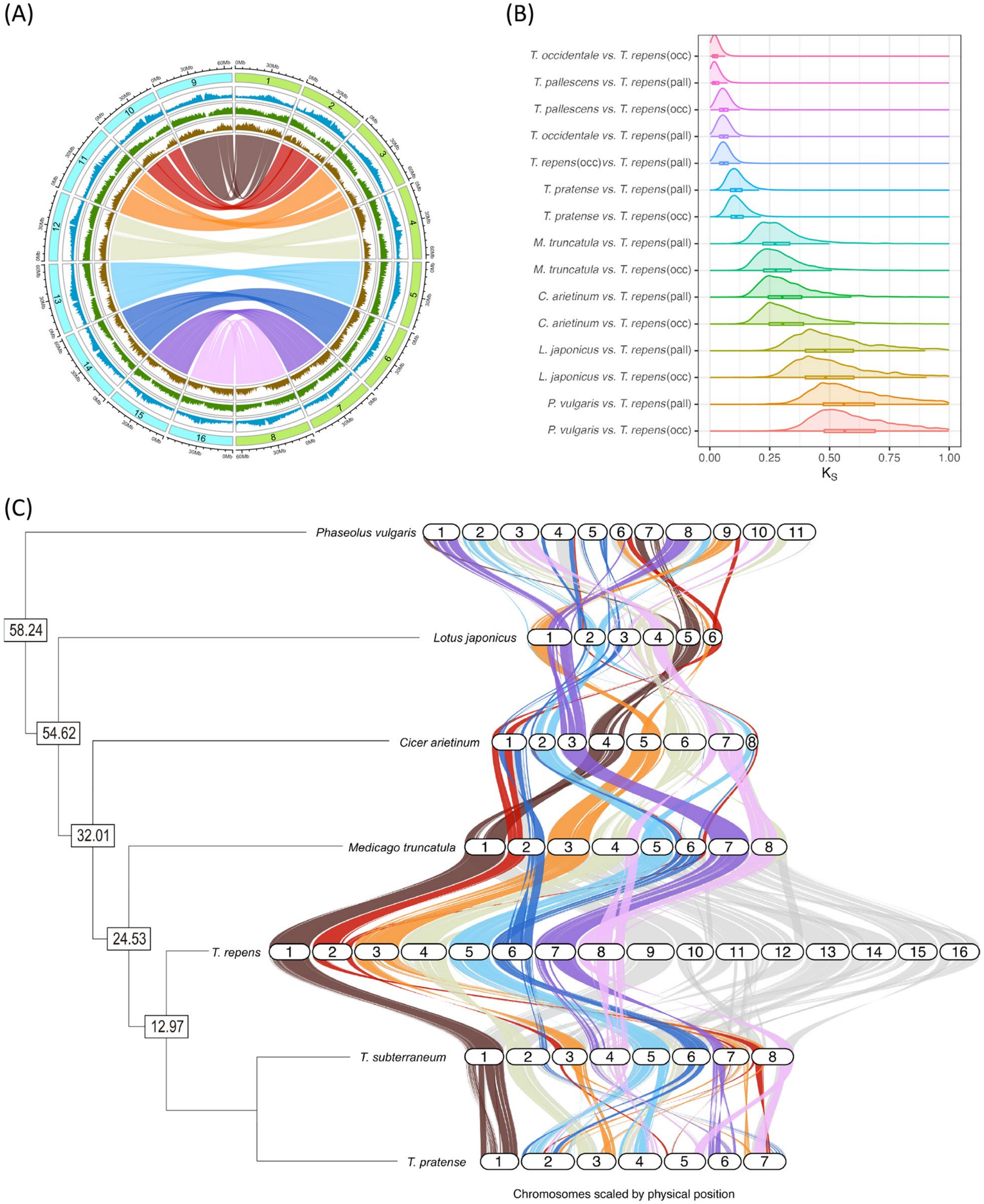

Fig. 2 (See legend on previous page.)

Cyanogenesis genes and copy number variation

Genome sequence analyses

We hypothesized that CNVs at the *Ac* and *Li* loci contribute to the phenotypic variation found in white clover populations across environmental gradients. This hypothesis was initially based on patterns of SNP inheritance in greenhouse pedigree populations, where Sanger sequencing of PCR amplicons revealed that *Acac* and *Lili* hemizygotes could transmit dominant alleles with SNPs that appeared heterozygous (Additional file 2: Table S3). To assess this hypothesis with our whole-genome data, we first tested for evidence of gene copy number variation (CNV) between the primary and alternate haplotigs in our newly generated genome. We found that our long-read based assembly unambiguously confirmed the presence of CNVs at both loci. We aligned the sequences around the *Ac* gene cluster (comprising the three cyanogenic glucoside biosynthesis genes: *CYP79D15*, *CYP736A187*, and *UGT85K17*) and the *Li* gene and performed BLAST analyses between haplotigs (Fig. 3). For *Ac*, there are two complete gene clusters in the primary haplotig and three in the alternative haplotig (Fig. 3A; Additional file 1: Table S4). The gene order inside the *Ac* gene cluster is consistent with previous findings [32] (Additional file 4: Fig. S8). For *Li*, there are two gene copies in both haplotigs (Fig. 3B; Additional file 1: Table S4). To determine the span of the *Ac* and *Li* CNV repeat motif, we plotted the gene synteny in the chromosomal region surrounding each locus (Fig. 3C, D). A single copy of the *Ac* motif is approximately 255–320 Kbp (Additional file 1: Table S5), while a single copy of the *Li* motif is approximately 24 Kbp (Additional file 1: Table S6). Although the CNV motifs are clearly replicated, the non-coding regions between the copies show sequence divergence (Fig. 3C, D). Multiple repetitive sequences are evident in the *Li* gene region, including other glucosidase genes that are predicted to have different substrate specificities than linamarase [40] (Fig. 3D and Additional file 4: Fig. S9). The presence of these repetitive sequences could contribute to genome instability that underlies the structural variation of this genomic region.

In addition to comparing CNVs in our newly generated genome, we performed similar analyses with the two other published white clover genomes. In the haplotig-resolved genome of Santangelo et al. [12], we found two copies of the *Ac* gene cluster in the primary haplotig and a deletion of the entire *Ac* gene cluster in the alternative haplotig (a ~ 500 Kbp deletion compared to the primary haplotig in the present study) (Additional file 4: Fig. S10A, B). For the *Li* gene, we found two copies per haplotig, similar to our new genome assembly (Additional file 4: Fig. S10C, D). In the genome of Wang et al. [11], which does not provide haplotig resolution, we found three copies of the *Ac* gene cluster and two copies of the *Li* gene (Additional file 4: Fig. S11). In none of the three white clover whole genomes did we find a haplotig with a complete deletion of the *Li* locus. However, since *Li* deletions are known to be common in nature [30, 35], we wanted to estimate of the size of the *Li* genomic deletion in plants where it occurs. To do this we mapped the raw reads (2 × 150 bp) of the genome sequences of an accession lacking linamarase activity (DMN_010, *lili* genotype, NCBI SRR27541472) to the primary haplotig of our novel genome. In the *lili* accession, we found a low coverage window of ~ 500 Kbp (Additional file 4: Fig. S12); this result suggests a much larger genomic deletion than the span of the *Li* CNV motif (24 Kbp) alone.

Gene expression and phenotypic data

After demonstrating the existence of both *Ac* and *Li* CNVs in all three white clover whole genomes, we sought to test the hypothesis that CNVs at the *Ac* and *Li* loci directly affect the cyanogenesis phenotype. To do this, we focused on the *Ac* locus because synthesis of cyanogenic glucosides, mediated by the *Ac* gene cluster, is predicted to be directly correlated with the cyanogenic response [41]. We performed qPCR of genomic DNA and mRNA to estimate the correlation between CNV counts and expression in the *Ac* gene cluster (*CYP79D15*, *CYP736A187*, *UGT85K17*). First, we found that the inferred copy numbers of the three genes in the *Ac* cluster consistently occur in a 1:1:1 ratio (Additional file 4: Fig. S13); this is consistent with our findings above that the entire *Ac* gene cluster is duplicated as a unit. In addition, we found that cyanogenic glucoside content is significantly positively correlated with copy number for each of the three genes in the *Ac* cluster (Fig. 4A–C and

(See figure on next page.)

Fig. 3 Nucleotide alignment and gene synteny of the white clover cyanogenesis genes. **A**, **B** Nucleotide alignment of the regions around the *Ac* gene cluster and the *Li* locus. Blue, the pairwise alignment is in the same direction. Red, the pairwise alignment is in the opposite direction. **C**, **D** Gene synteny of the regions around the *Ac* gene cluster (**C**) and the *Li* locus (**D**). Schematic plot on the top-right of each panel shows the overall arrangement of the repeated gene copies that are shown in detail in the main plots. Top sequence, the chromosome in the primary haplotig. Bottom sequence, the corresponding homeologous chromosome in the alternative haplotig. Gray links, nucleotide alignments (same as in **A** and **B**). Purple stacks, the number of nucleotide alignment hits in the homeologous chromosome. Note that Chr. 2 is in the *T. occidentale* subgenome and Chr. 12 is in the *T. pallescens* subgenome

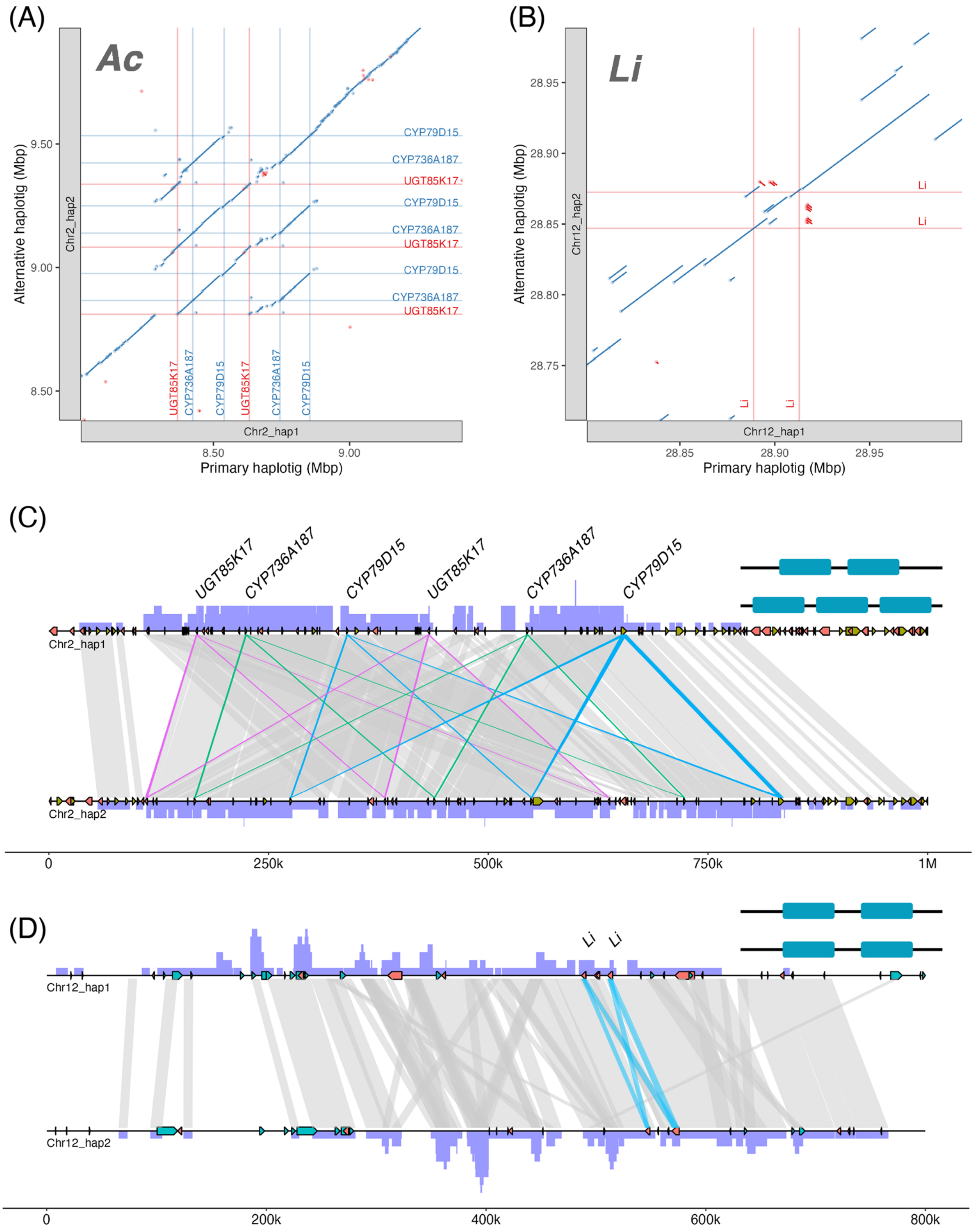

Fig. 3 (See legend on previous page.)

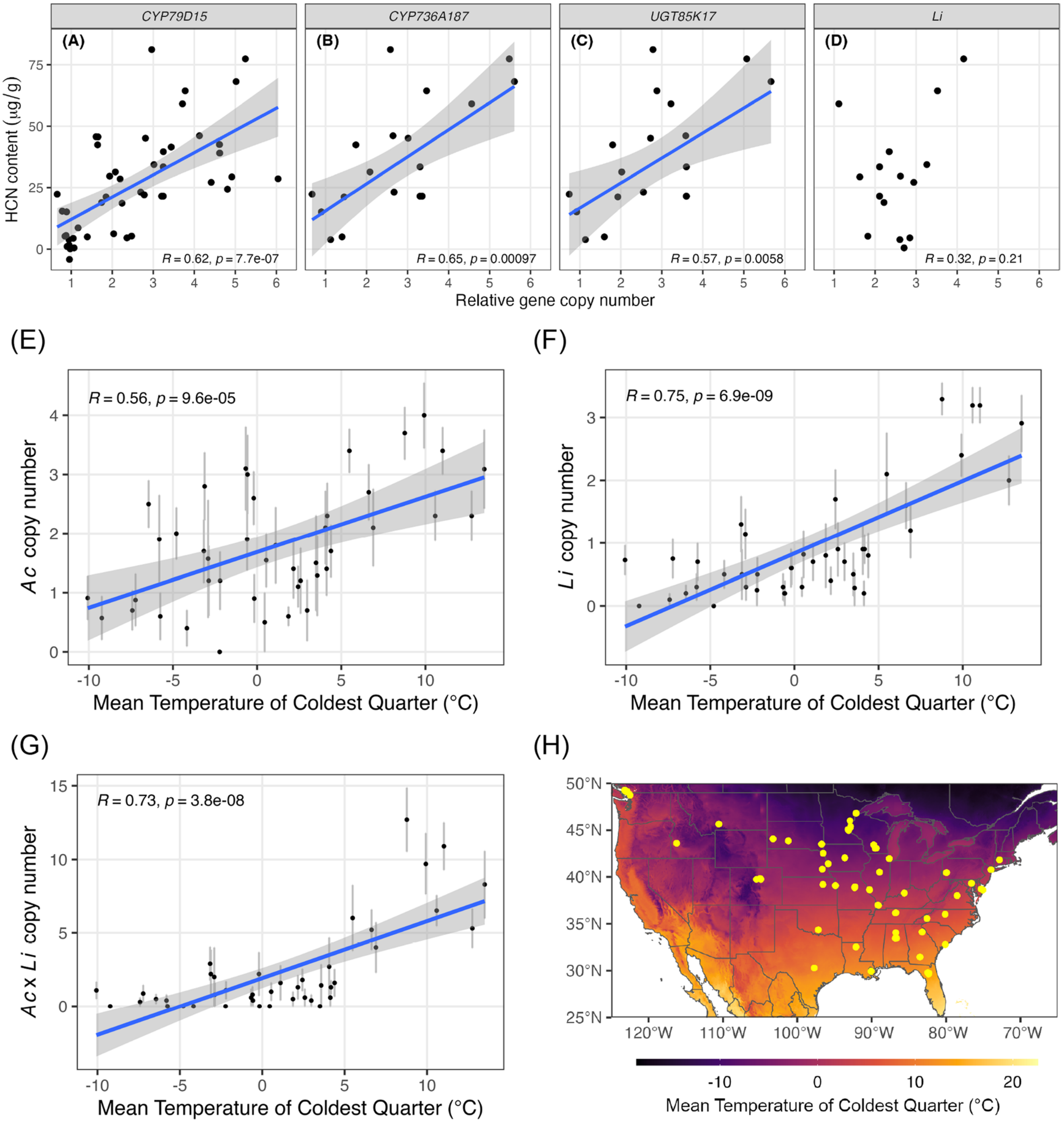

Fig. 4 Cyanogenesis gene copy number variation and associations with cyanogenic potential and climate. **A**–**D** Gene copy number variation and cyanide (HCN) content (per gram fresh leaf tissue). Linear regression line and standard error are plotted if the slope is significantly different from zero ($p < 0.05$). Each point stands for the mean value of 3–4 repeated measurements. **E**–**H** Association between cyanogenesis gene copy number and *Mean Temperature of Coldest Quarter* (MTCQ). **E** *Ac* copy number. **F** *Li* copy number. **G** The product of (*Ac* copy number)*(*Li* copy number) within an accession. **H** Collection locations of 419 sampled accessions (43 populations) across North America. The map is shaded by MTCQ. Pearson correlation coefficient was calculated; if the value was significantly different from zero, the linear model was plotted. Raw data are available at Additional file 1: Table S7 and Additional file 3: Table S8

Additional file 1: Table S7). In comparison, no such correlation was observed for the *Li* gene, whose function is unrelated to cyanogenic glucoside biosynthesis (Fig. 4D and Additional file 1: Table S7). We further detected that cyanogenic glucoside content is significantly positively correlated with the expression of *CYP79D15* and *CYP736A187*, which correspond to the first two steps in the cyanogenic glucoside biosynthetic pathway (Additional file 4: Fig. S14). Surprisingly, however, no significant correlation was observed between the number of gene copies and levels of gene expression (Additional file 4: Fig. S15); this could be an artifact of inadequate time point sampling of tissue, as regulation of cyanogenic glucoside synthesis is known to vary diurnally in other cyanogenic plant species [42].

Wild populations and environmental associations

Finally, to assess whether CNVs at the *Ac* and *Li* cyanogenesis loci contribute to environmental adaptation in wild populations of white clover, we sampled 419 wild accessions from 43 white clover populations distributed across North America, estimated the copy number at the *Ac* and *Li* loci using genomic DNA qPCR, and tested for associations with climatic variables of the collection locales using the WorldClim 2 dataset [43]. CNVs for the wild population samples ranged from 0 to 8 per accession for *Ac* and 0 to 6 for *Li* (Additional file 3: Table S8 and Additional file 4: Fig. S16). Because previous observations of white clover cyanogenesis clines have detected strongest associations between cyanogenesis frequencies and minimum winter temperature [6, 36], we focused on the climatic variable *Mean Temperature of Coldest Quarter* (MTCQ) for CNV association tests. Notably, both *CYP79D15* (the first gene in the *Ac* cluster) and *Li* are significantly correlated with MTCQ (Fig. 4E, F, H). Importantly, these correlations remained significant after removing accessions with zero copies of the genes (corresponding to the *acac* and *lili* recessive genotypes); this indicates that the significant correlations between CNVs and climate are not simply reflecting the previously described PAVs at the *Ac* and *Li* loci, which are well known to contribute to white clover cyanogenesis clines [6, 36] (Additional file 4: Fig. S17, Fig. S18).

Because the cyanogenic phenotype arises through an epistatic interaction between the two cyanogenesis loci, we also examined the cumulative effects of CNVs at both loci by assessing MTCQ associations with the product of *CYP79D15* and *Li* gene copy numbers. As with the individual loci, this assessment revealed a significant positive correlation (Figs. 4G and Additional file 4: Fig. S17C, S18C). This finding further suggests that adaptive clinal variation at the white clover cyanogenesis loci arises through selection acting on the *Ac* and *Li* CNVs.

Discussion

PacBio HiFi, Omni-C, and linkage map bridge the gap to allotetraploid genome assembly

The increased availability of high-quality long-read genomic data has contributed to our understanding of the importance of structural variation, including CNVs, as contributors to phenotypic variation. However, despite the fact that CNVs are prevalent in nature, very few studies to date have documented the distribution of CNVs in the wild and correlated these with known adaptive phenotypes along an environmental gradient. In this study, we bridge this knowledge gap by revisiting the molecular mechanism underlying the well-known cyanogenesis polymorphism in white clover using three newly available high-quality genomes generated with long-read based sequencing technology (this study; 11, 12). The genome assembly presented here is the second haplotig-resolved genome of white clover and was generated by incorporating high-depth (84×) PacBio HiFi sequencing, Omni-C chromatin interaction, and linkage map information. The quality of our new genome assembly highlights the value of high-density linkage map data for complementing Omni-C data in scaffolding, especially when joining two long-range scaffolds in their correct orientation (e.g., across centromeres). The ability to accurately scaffold chromosomes is often limited when using Omni-C data alone (see, e.g., comparisons of the three white clover genomes with and without linkage map assisted scaffolding; Additional file 4: Fig. S19, Fig. S20). Our genome also represents a true de novo genome assembly for white clover — created without first assigning sequences to a subgenome before attempting the assembly. The resulting assembly thereby provides the most unbiased and reliable chromosome-level and haplotig-resolved genome to date for this economically important legume species.

Unexpectedly little genome reorganization in white clover

Polyploidization is characterized by an immediate multiplication in chromosome number through a whole genome duplication event, which can produce a major shock to the genome that may cause a series of unpredictable but discernible downstream changes [44]. These may include but are not restricted to bursts of transposable element (TE) proliferation [37], large-scale genome rearrangements [45], fractionation [46], and the accumulation of deleterious mutations due to duplicated gene redundancy [47]. In our study, we found extensive distribution of TEs in white clover genome (30% of the genome). The short branches of the phylogenetic trees of LTR retrotransposons (LTR-RTs) suggest a recent multiplication of LTR-RT elements within the white clover genome (Additional file 4: Fig. S2, Fig. S3). It is possible that the high number of TEs in the white clover genome has facilitated the diploidization process by accumulating new mutations that contributed to subgenome divergence, thereby repressing recombination between homeologous chromosomes [37]. Further investigation of the TE family expansions, including in the diploid progenitor species, would be useful to test this hypothesis.

The genus *Trifolium* is characterized by extensive chromosome number variation between species and occasional polyploid speciation events [48]. In our genomic comparisons with the closely related legume *Medicago truncatula*, white clover shows the most conserved chromosome structure of the three *Trifolium* species with reference genomes, the other two of which are diploid; *T. pratense* (2n = 14) and *T. subterraneum* (2n = 16) each possess a minimum of 2–3 chromosomal rearrangements in a majority of their chromosomes (Fig. 2C). This finding suggests that polyploidization by itself does not necessarily lead to large-scale genome reorganizations and, conversely, that extensive genome reorganization can occur in the absence of polyploidization.

Beyond the genus *Trifolium*, inclusion of the white clover genome in legume genome sequence comparisons is also pivotal in accurately reconstructing the karyotype evolution of the IRLC legume clade. This economically important group contains the majority of agriculturally important legumes, including chickpea, lentil, pea, vetch, alfalfa, and clovers. Notably, without the white clover genome sequence comparison, it would be impossible to tell that the basal chromosomal organization in *Trifolium* is conserved with related legume genera (Fig. 2C).

Evolutionary theory predicts that species may experience a relaxation of purifying selection after polyploidization because the redundant gene copies (homeologs) are capable of masking the effect of deleterious recessive mutations [47]. We therefore expected to detect more mutations in homeologous gene pairs than in pairwise comparisons of homologous (orthologous) genes of the extant progenitors. However, no significant difference in the ratio of nonsynonymous versus synonymous mutation rate (K_A/K_S) was found (Additional file 4: Fig. S6). Moreover, we did not detect any major genome fractionation in either subgenome (Fig. 2A) (see also [9]). These observations together suggest that white clover has experienced little relaxation of purifying selection after the polyploidization event that created this species. Although this finding is contrary to the classic expectations for homeologous gene copies, it is possible that selection to maintain dosage balance could be preventing any such relaxation of purifying selection [46]. It is also possible that white clover's very recent evolutionary origin (estimated at 15,000–28,000 years ago) [9] is simply too short of a time period to allow for the detectable accumulation of deleterious mutations.

Revisiting the evolution of the adaptive cyanogenesis polymorphism

Understanding the genetic and biochemical bases of the white clover cyanogenesis polymorphism, and the ecological factors that shape its evolution, has been a century-long research endeavor. Initial characterizations of the biochemical components and the polymorphism's inheritance patterns predated the molecular era [49–53]. In the 1950s, a series of classic studies by Daday surveyed cyanogenesis frequencies in wild populations worldwide; this work revealed the widespread occurrence of latitudinal and elevational cyanogenesis clines throughout the native and introduced species range, with cyanogenic plants consistently present at higher frequencies in warmer climates [7, 8, 54]. In subsequent decades, numerous studies have examined the potential environmental factors that maintain the polymorphism and drive the evolution of climate-associated cyanogenesis clines (reviewed by [5, 29, 36, 55, 56]). Key selective factors likely include chemical defense in areas of high herbivore abundance, fitness tradeoffs associated with the energetic costs of producing cyanogenic components, and abiotic stresses including cold and drought. Complementary studies have focused on the molecular basis of the polymorphism and its evolutionary origins. This work has revealed that the *Ac/ac* and *Li/li* polymorphisms arise through two unlinked PAVs [30–32]; that the recessive (deletion) alleles have evolved repeatedly in white clover as well as in related *Trifolium* species [32, 33, 35]; and that the *Ac* and *Li* loci of white clover are derived from different diploid progenitors, with *Ac* derived from *T. occidentale* and *Li* derived from the *T. pallescens* lineage [13]. As the body of white clover cyanogenesis research has progressed, this system has come to be regarded as a textbook example of a balanced polymorphism maintained by geographically heterogeneous selective pressures [57, 58].

By identifying the occurrence of CNVs for the dominant (gene-presence) alleles of the *Ac* and *Li* loci, and by documenting their associations with climatic variation in wild populations across the North American species range, the present study adds new depth to our understanding of the white clover cyanogenesis polymorphism and its role in adaptive evolution. Analysis of the three long-read genome assemblies (this study; 11, 12) revealed CNV counts of 0, 2, and 3 for the *Ac* locus and 2 copies for the *Li* locus in the haploid genomes (Fig. 3, Additional file 4: Fig. S10, Fig. S11). With expanded sampling to include 419 wild North American accessions, estimated CNV counts ranged from 0 to 8 per accession for *Ac* and 0–6 for *Li* (Additional file 3: Table S8 and Additional file 4: Fig. S16). The mechanism responsible for the producing these CNVs remains unknown; however, our genome sequence analyses suggest that it is unlikely to be due to mobile element activity. TEs are not especially abundant in the genomic regions of the *Ac* and *Li* loci, and those elements that are present do not show any detectable patterns associated with CNV counts

(Additional file 4: Fig. S21). This may suggest that other mechanisms, such as nonallelic homologous recombination (NAHR), are responsible for the observed structural variation.

The physical chromosome map produced in the present study corroborates our previous findings, based on QTL mapping of the cyanogenesis phenotype, on the genomic locations of the *Ac* and *Li* loci [13]. Interestingly, unlike most genes in the polyploid white clover genome, no functional homeologs of either *Ac* or *Li* were detected in BLASTs against all gene models (data not shown), despite the highly conserved gene synteny between the two subgenomes overall. These findings lend further support to our prior conclusions that the *Ac* and *Li* genes of white clover are uniquely derived from its two different diploid progenitors [13].

Another key finding of our study is the positive correlation between the number of CNVs at the *Ac* locus and cyanogenic glucoside production (Fig. 4A–C). To our knowledge, this is the first reported evidence in any cyanogenic species of quantitative variation in cyanogenic glucoside production due to CNVs of the underlying biosynthetic pathway. In contrast to white clover, the cyanogenic glucoside content in cassava [59] and *Lotus corniculatus* [60] has been reported to be regulated by cell membrane transporter genes and stress responsive genes, respectively. It thus appears that quantitative control of cyanogenic glucoside accumulation in different plant lineages has evolved independently through different molecular mechanisms. The correlation between *Ac* CNV dosage and cyanogenic glucoside production also illustrates the importance of multiplication of the entire gene cluster. We found that each *Ac* CNV unit includes all three component genes of the biosynthetic pathway (*CYP79D15*, *CYP736A187*, *UGT85K17*) [32] and that the gene order is conserved across CNV copies (Fig. 3A, C; Additional file 4: Fig. S10A, Fig. S11A). If there were multiplication of some but not all of the three genes, the imbalance and resulting accumulation of unstable chemical intermediates could potentially be maladaptive or even toxic [32, 61, 62]. To our knowledge, only one other study has documented the role of variable copies of an entire gene cluster in contributing to phenotypic variation (soybean nematode resistance locus *Rhg1*) [19].

Finally, our results demonstrate that CNVs contribute to local adaptation and the evolution of climate-associated adaptive clines in wild populations of a plant species. As noted above, the white clover cyanogenesis polymorphism has long served as a model for understanding the forces that generate and maintain adaptive polymorphisms within species. Our findings here that CNVs in wild white clover populations are significantly correlated with local climate (Fig. 4E–H), together with the evidence that CNVs directly affect the cyanogenic phenotype (Fig. 4A–C), provide a new level of resolution to our understanding of how natural selection interacts with the genome to shape this adaptive polymorphism. To our knowledge, these results provide the first extensively documented case of CNV-mediated local environmental adaptation in a wild plant species. Equally importantly, they highlight the value of long-read sequencing technologies in accurately capturing the full spectrum of genomic variants that contribute to adaptation. The last decade has witnessed a rapid growth in studies using whole genome sequencing to assess molecular basis of adaptation in plants and many other species [63, 64]. However, reliance on short-read data and alignments to a single reference genome has meant that discoveries are heavily biased towards SNP variants, with little knowledge of the role that CNVs or other structural variants (SVs) may play in adaptation. While population-scale pangenome sequencing is still economically unfeasible for nearly all eukaryotic systems, a hybrid approach such as employed in our study (i.e., using a low-sample pangenome to discover and validate SVs, followed by qPCR to conduct population-scale investigations) could be an affordable way to uncover the role of SVs in adaptation.

Conclusions

Through analyses based on a newly generated, de novo assembled haplotype-resolved genome for white clover, we are able to draw several key inferences about the genomic composition and evolution of this economically important species: (1) the allotetraploid white clover genome, while characterized by recently elevated TE activities, retains highly conserved genome organization and gene composition, with no evidence of relaxed purifying selection in homeologs since its origin; (2) unlike its two diploid congeners with reference genomes (*T. pratense and T. subterraneum*), both of which have undergone extensive chromosomal reorganizations, white clover has preserved a more ancestral genome structure that is shared with other members of the IRLC legume clade; (3) the *Ac* and *Li* cyanogenesis loci are both characterized by CNVs, which contribute to quantitative variation in the cyanogenic phenotype and to local climatic adaptation in wild populations spanning the North American species range. Finally, this study highlights the value of integrating PacBio long-read sequencing, Omni-C chromatin interaction, and linkage map data in polyploid genome assembly and the power of pangenome perspectives in identifying the functional effects of structural variation and its role in local adaptation in nature.

Methods

De novo genome assembly

Young leaf tissue of a wild North American white clover accession “GFL_007” (grown from seed collected in Gainesville, Florida; NCBI SAMN37329216) was dark-treated for 48 h before DNA isolation. High molecular weight (HMW) DNA was isolated by Polar Genomics (Ithaca, NY) using a CsCl gradient to remove mitochondrial and chloroplast genomes. The HMW DNA was then submitted to the Roy J. Carver Biotechnology Center (University of Illinois at Urbana-Champaign, Urbana, IL 61801) for PacBio HiFi sequencing. The HMW DNA was sheared with a Megaruptor 3 to a size of ~ 15 Kbp. Sheared gDNA fragments were converted to a library with the SMRTBell Express Template Prep kit 2.0. The library was sequenced on three SMRTcell 8 M on a PacBio Sequel II using the circular consensus sequencing (CCS) sequencing mode and a 30-h movie time. CCS analysis was performed using SMRTLink V8.0 using the following parameters: ccs –min-length 1000 –max-length 50000 –min-passes 3 –min-rq 0.99. By this approach, PacBio HiFi generated 7.3 million reads with average length 11.6 Kbp. The expected mean coverage is 77 × for a collapsed haploid genome based on the sequencing depth and previously reported genome size [9]. In order to phase the diploid genome and scaffold, Omni-C was used to capture the chromatin proximity interactions. The Omni-C library and sequencing services were also provided by the Roy J. Carver Biotechnology Center. One hundred million read pairs (2 × 150 bp) were sequenced by using NovaSeq 6000 System with SP flow cell (Illumina). The raw reads were demultiplexed and the adaptors were trimmed using the Illumina standard software.

Although the extant white clover genome is likely to have minimal inter-subgenome translocations (e.g., homeologous recombination) [9], we cannot exclude that translocations might have occurred following polyploidization. Therefore, unlike the most recent previous assembly [12], which categorized the PacBio HiFi reads into subgenomes before final assembly, we conducted a complete de novo assembly from the raw PacBio HiFi reads. Hifiasm v0.16.1 was used for the de novo genome assembly with Hi-C integrated mode [65]. The output of the fully phased contigs of the primary and alternative haplotigs were used as the template for Omni-C scaffolding independently. The adaptor-trimmed Omni-C reads were mapped to the two haplotigs independently according to the mapping pipeline provided by Arima Genomics [66]. In short, the pipeline first maps the reads to the reference in single-end mode. Then, the chimeric mapped reads (if present) were processed to retain only the 5′-end of the sequence (on the assumption that the 3′-end was a result of proximal ligation). Finally, the paired reads were joined and sorted in the output.

The genome scaffolding was conducted in two steps. First, the Hi-C scaffolding was performed by SALSA2 with the default settings [67]. Second, two linkage maps created from two F_2 mapping populations (see the “Linkage map construction” section) were used to place the Hi-C scaffolds into pseudochromosomes. The two linkage maps were merged based on the physical locations of the markers in the corresponding contigs, and the genetic distances (cM) were re-estimated using either the mean genetic distance of the two linkage maps or the only available value if the marker was only present in one linkage map. Chromonomer was used to integrate the information of the Hi-C scaffolding and the linkage map and then generate the final genome assembly information file (AGP file) [68]. Then, agptools was used to update the final assembly [69]. The chromosome numbers were labeled by the BLAST against the diploid progenitors’ genome (see subgenome characterization) and the nucleotide alignment to *Medicago truncatula* (see gene synteny and nucleotide alignment). The complete chloroplast and the partial mitochondrial genomes were assembled from the Illumina reads in GetOrganelle [70]. Then, the chloroplast and mitochondrial genome, plus the NCBI prokaryotic RefSeq genomes were used as the reference in BLAST to remove the contaminations from the nuclear genome.

k-mer based genome quality assessment

To estimate the reference-free genome size and heterozygosity before assembly, an independent Illumina short read (2 × 150 bp) dataset of the same accession (GFL_007; SAMN37329216) was used to generate a histogram of k-mer frequencies in jellyfish with default settings (k-mer length of 21) [71]. Then, the estimation was performed by GenomeScope [72]. To assess the assembly quality, the same Illumina dataset was used and the assessment was performed by Merqury [73].

Linkage map construction

Two linkage maps were used to scaffold and assess the scaffolding quality of the new assembly. These were generated from two F_2 mapping populations derived from biparental crosses of GFL_007 and two other white clover genotypes [DG population from a cross of accessions DMN_010 (SAMN34157026) and GFL_007 (SAMN37329216), GS population from accessions GFL_007 (SAMN37329216), and STL_0701 (SAMN34157027)] [13, 74]. Protocols for generation of genotyping-by-sequencing (GBS) marker data used in linkage map construction are described in [13]. For the present study, de novo linkage map construction was performed by mapping raw GBS reads to the two Omni-C scaffolded haplotigs; SNPs were called independently following the GATK best practice workflow [75]. In brief, the reads were mapped to the reference by Bowtie2 (bowtie2 –sensitive –no-mixed –no-discordant –minins 100 –maxins 1000). Then, the SNPs were called by HaplotypeCaller (-ERC GVCF) and combined by GenotypeGVCFs. The output in vcf format was first hard-filtered (bcftools filter -e 'QD < 0.5 || FS > 200.0 || MQ < 20.0 || MQRankSum < -12.5 || MQRankSum > 12.5 || ReadPosRankSum < -8.0 || SOR > 8 || INFO/DP < 2000'); then, individuals with missing genotype (SNP site) > 0.25 were removed and then filtered for minor allele frequency > 0.35, max number of alleles ≤ 2, *p*-value > 1e − 20 on a genotype frequency test (indicating no significant deviation from 1:2:1 segregation in the F_2 populations), minimum mean depth > 15, and missing data < 0.25; finally, only genotypes (SNP sites) that were homozygous in the parents were preserved. The linkage maps were then constructed by GUSMap package with the standard protocol with the minimum LOD threshold that can form 16 linkage groups [76]. The quality of the final phased genome assembly (primary haplotig) was further validated by serving as the reference for mapping the raw GBS reads. All procedures for raw read mapping and follow-up steps were the same as for linkage map construction, except for that all the individuals of the mapping populations were kept. The final vcf files were converted to the "cross" format and imported to the R/qtl package [77]. Then, the heatmaps of the LOD scores and the recombination fractions between the pairwise markers were visualized.

Subgenome categorization

In order to characterize the subgenome identities of either *T. occidentale* or *T. pallescens* origin, the final assembly (primary and alternative haplotigs) was used as the query to BLAST (which allows multiple alignment hits to ambiguous genomic regions) against published genomes of the two diploid progenitors, *T. occidentale* and *T. pallescens* [9]. Only the hits with alignment > 1000 bp and identity > 95% were kept. Then, a customized index was calculated based on: $\sum (To\,alignment\,length)_i - \sum (Tp\,alignment\,length)_k / Bin\,size\,(0.5\,Mb)$, where i and k are the filtered hits to either the *T. occidentale* (To) or *T. pallescens* (Tp) subgenome, respectively. This measure creates an index where a value of -1 is completely *T. pallescens*-like and + 1 is completely *T. occidentale*-like.

Gene prediction and functional annotation

The two haplotigs of the diploid assembly were independently annotated with the BRAKER2 pipeline [78]. The haploid genome was first softmasked for the repeated regions by RepeatModeler v2.0.4 and RepeatMasker v4.1.4 with a de novo constructed and lineage-specific database. Then, the gene model was separately predicted with RNA-seq and protein data. The results were then combined by TSEBRA as suggested [79]. The RNA-seq data were derived from leaf tissue samples of the three accessions used to create the linkage maps (DMN_010, GFL_007, STL_0701); tissue was collected from plants grown in the greenhouse under control and drought treatments (N = 18, including three accessions and three clonal replications of each accession). The RNA-seq reads were mapped to the reference in a relaxed mode (–outFilterScoreMinOverLread 0 –outFilterMatchNminOverLread 0) by STAR v2.7.10a [80]. The protein data were from the translated sequences of the previous published white clover genome [9] and OrthoDB v10 "Fabales" database. For functional annotation, the primary transcripts of each gene model were translated into protein sequences and analyzed using the blastp algorithm against the NCBI RefSeq genomes of Arabidopsis, soybean (*Glycine max*), and *Medicago truncatula*. The BLAST result was then imported in Blast2GO software. After combining the results from InterProScan, the GO terms were mapped and annotated following standard protocols [81].

LTR retrotransposons (LTR-RTs) characterization

LTR-RTs were detected and classified by inpactor2 [82]. The functional components of LTR-RT were then classified by TEsorter [83] with REXdb database [84]. The reverse transcriptase (RT) amino acid sequences were extracted and aligned by MAFFT [85]. Maximum likelihood tree was reconstructed by IQ-TREE2 [86].

Gene synteny and nucleotide alignment

Interspecific gene synteny analysis was conducted for related legume species by GENESPACE v0.9.3 [87], which integrates the results from OrthoFinder [88] and MCscanX [89]. In short, the primary transcripts of *Trifolium pratense* [90], *Trifolium subterraneum* [91], *Medicago truncatula* [92], *Cicer arietinum* [93], *Lotus japonicus* [94], and *Phaseolus vulgaris* [95] were used as the input in OrthoFinder in nucleotide mode. Then, the gene synteny was generated by MCscanX with the settings: blkSize = 20, nGaps = 25. Nucleotide alignment was conducted by Mummer 4.0.0 with the function: nucmer [96]. Only the alignment hits with more than 1000 bp were kept. The gene synteny and nucleotide alignment were then visualized using package gggenomes v0.9.5 in R.

Synonymous mutation rate (K_S) and nonsynonymous mutation rate (K_A)

To assess the divergence among orthologs across species and examine subgenome-specific patterns of molecular evolution within white clover (e.g., relaxed purifying selection), synonymous mutation rate (K_S) and nonsynonymous mutation rate (K_A) were calculated based on the single copy orthologous genes that are universally available in all the taxa in the synteny analysis (except for *Trifolium subterraneum*). After filtering, 3884 genes were kept. The orthologous gene sequences were aligned and guided by their translated protein sequences in MACSE v2 [97]. The low-quality alignment was removed following the OMM_MACSE pipeline. The cleaned alignments were then used for mutation rate calculation in KaKs_Calculator 2.0 with YN method [98].

Cyanogenesis gene copy number variation, gene expression and cyanide quantification

Plant materials used to characterize cyanogenesis variation were collected as seeds or stolons from North American wild populations and cultivated in the greenhouse of Washington University in St. Louis (Additional file 1: Table S7, S8). The cyanogenic phenotype, including the presence/absence of the cyanogenic glucosides and the substrate-specific glucosidase, was assessed biochemically by Feigl-Anger cyanogenesis assay and genetically by PCR-genotyping of *CYP79D15* (the first gene in the *Ac* metabolic cluster) and *Li* [30, 31]. Plants that produced one or both cyanogenic precursors were further examined by qPCR with gDNA to estimate the gene copy number information and with cDNA to estimate the relative gene expression level. The qPCR primers of *CYP79D15*, *CYP736A187*, and *UGT85K17* (together comprising the *Ac* locus) and *Li* were designed based on the sequences in our previous publications [30, 32] and are available in Additional file 1: Table S9. The qPCR was conducted in the CFX96 Real-Time PCR Detection System (BIO-RAD), and the data were analyzed by $\Delta\Delta C_T$ method with efficiency calibration in CFX Maestro™ Software (BIO-RAD) according to the manufacturer's instructions. At least three independent qPCR replicates were conducted for each gene. In each independent replicate, three technical replicates were included. The C_T value was first normalized to *Ef1α* gene and then scaled to the normalized C_T value of the single-copy genotype. To provide a reference genotype for quantification of gene copy numbers, accessions possessing a single copy of each assayed gene were identified in pedigree populations created by crossing cyanogenic and acyanogenic parents (Additional file 2: Table S3).

The cyanogenic glucoside content per unit weight of the fresh leaf tissue was measured by a colorimetric method. Leaf tissue (1 g) was homogenized in liquid nitrogen and resuspended in 2 mL of Na_2HPO_4 buffer (0.067 M, pH 6.0). Then, the sample was centrifuged at 4 °C for 20 min at 13,000 × g twice. After each centrifugation, the supernatant was transferred to a new tube. Then, an excess amount of linamarase (100 μL bulk protein extract from cassava latex, [99]) was added to the sample (50 μL supernatant + 450 μL Na_2HPO_4 buffer) and incubated 30 min at 37 °C to release cyanide from cyanogenic glucosides (linamarin and lotaustralin). After incubation, 600 μL NaOH solution (0.2 M) was added to terminate the reaction. Then, 500 μL of the reaction mix was used for cyanide quantification by Spectroquant® Cyanide Test kit (Millipore cat. no. 1.09701) following the standard protocol. The standard cyanide solution was prepared by the same procedure but replaced the supernatant with the commercial cyanide standard (Millipore cat. no. 119533). The colorimetric reaction had the maximum absorbance at 607 nm. We used the absorbances at 700 nm and 775 nm to estimate the background absorbance at 607 nm by simple linear regression. Finally, the absorbance of 607 nm was subtracted by the background, and then compared to the standard curve for concentration.

Environmental association

To investigate the association between the cyanogenesis gene copy number variation and environmental factors, climatic information at 1-km resolution was downloaded from WorldClim 2 [43]. We extracted the environmental information for collection locations of the sampled wild clover populations as the mean value from the 1000-m-buffered sample location. Then, the gene copy number was correlated to *Mean Temperature of Coldest Quarter* (MTCQ) with Pearson correlation method and visualized in R.

Abbreviations

IRLC	Inverted repeat-lacking clade
CNV	Copy number variation
PAVs	Presence/absence variations
TE	Transposable element
CDS	Coding sequence
MTCQ	*Mean Temperature of Coldest Quarter*
LTR-RTs	LTR retrotransposons
SVs	Structural variants
HMW DNA	High molecular weight DNA
CCS	Circular consensus sequencing
GBS	Genotyping-by-sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-024-01962-6.

Additional file 1: Table S1. Genome assembly summary. Table S2. Total alignment length. Table S4. Genomic locations of cyanogenesis genes. Table S5. Genomic locations of *Ac* gene cluster. Table S6. Genomic locations of *Li* genes. Table S7. Accession information for cyanogenesis CNVs and resultant HCN content variation. Table S9. Primer list.

Additional file 2: Table S3. SNPs detected by direct Sanger sequencing of PCR products of *Ac* and *Li* genes in parents and F$_1$ progeny in greenhouse crossing experiments.

Additional file 3: Table S8. Accession information for characterizing cyanogenesis variation.

Additional file 4: Fig. S1. GenomeScope genome size and heterozygosity assessment. Fig. S2. Phylogeny of the reverse transcriptase amino acid sequences of the Copia super-family in the white clover genome by maximum likelihood method. Fig. S3. Phylogeny of the reverse transcriptase amino acid sequences of the Gypsy super-family in the white clover genome by maximum likelihood method. Fig. S4. Chromosome categorized into *T. occidentale* or *T. pallescens* subgenome. Fig. S5. Coding sequence length and intron length distribution. Fig. S6. Pairwise nonsynonymous mutation rate and synonymous mutation rate between the white clover subgenomes and the diploid progenitors' genomes. Fig. S7. Whole genome nucleotide alignment of the white clover and related species. Fig. S8. *Ac* gene cluster compared between white clover and *Lotus japonica*. Fig S9. Nucleotide alignment of the *Li* gene and other glucosidase genes in white clover. Fig S10. Nucleotide alignment of the cyanogenesis gene copy number variation regions between the white clover reference genome of the present study and that of Santangelo et al. Fig S11. Nucleotide alignment of the cyanogenesis gene copy number variation regions between the white clover reference genome of the present study and Wang et al. Fig. S12. Mapping result of the whole genome re-sequencing of the accessions GFL_007, DMN_010, and PI239977. Fig. S13. Pairwise gene copy number correlation of the genes in *Ac* gene cluster. Fig. S14. Relative gene expression and cyanide content. Fig. S15. Gene copy number variation and relative gene expression. Fig. S16. Estimated *Ac* and *Li* copy number of 419 wild accessions by genomic qPCR. Fig. S17. Association between cyanogenesis gene copy number variationand mean temperature of coldest quarter. Fig. S18. Association between cyanogenesis gene copy number variationand mean temperature of coldest quarter. Fig. S19. Whole genome nucleotide alignment between the primary haplotig and Santangelo et al. Fig. S20. Whole genome nucleotide alignment between the primary haplotig and Wang et al. Fig. S21. Nucleotide alignment and gene synteny of the white clover cyanogenesis genes with LTR-RTs labeled.

Acknowledgements

We thank the high school Biology teachers, plant scientists, and others who participated in the crowd-sourced collection of wild white clover accessions across North America, including Miranda Colletta, Sunita Crittenden, Felix Fritschi, Briana Gross, Nic Kooyers, Lisa Limeri, Mary Lyman-Onkka, Ed McAssey, John McDonald, John McIntire, Maria Monteros, Jim Mullen, Patricio Munoz, Sandra Pelc, Anne Puzzo, Nicole Riddle, Trisha Spanbauer, Daniel Tabor, Rebecca Thomson, Kathryn Turner, Cynthia Vigueira, Kate Waselkov, Marshall Wedger, and Dan Cui Zhou. We especially thank Alexander Mahmoud (undergraduate student at Washington University in St. Louis) for processing wild accessions and maintaining them in greenhouse. We also thank Mike Dyer and other staffs of the Washington University in St. Louis greenhouse facility for providing plant care. Finally, we thank Brock Mashburn, other members of the Olsen lab group, and two anonymous reviewers for valuable comments on earlier versions of the manuscript.

Authors' contributions

W-HK designed the experiments, analyzed all data, interpreted the results, and wrote the manuscript. SJW created the mapping populations, collected and maintained the wild accessions, prepared and sequenced GBS library, and phenotyped the cyanogenesis traits. LLS maintained the plants in greenhouse, prepared GBS library, phenotyped the cyanogenesis traits, and quantified the cyanogenic glucoside content per plant. KMO conducted the greenhouse pedigree crossings, sequenced the cyanogenesis genes, conceived the project, interpreted the results, and wrote the manuscript. All authors read and approved the final manuscript.

Funding

W-HK was funded through the William H. Danforth Plant Science Graduate Research Fellowship in the Division of Biology and Biomedical Sciences at Washington University and by a scholarship from the Taiwan Ministry of Education. Additional funding for the project was provided by US National Science Foundation grants IOS-1557770 to KMO, and DEB-1601641 and DGE-1143954 to SJW.

Availability of data and materials

All data generated or analyzed during this study are included in this published article, its supplementary information files and publicly available repositories. The de novo genome assemby datasets generated during the current study have been deposited with links to BioProject accession number PRJNA953427 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/). It includes the PacBio HiFi raw reads (SRR24107384), the RNA-seq (SRR24147954–SRR24147971), the GBS adapter-trimmed reads of the GS F$_2$ mapping population (SRR24400943–SRR24401442), the GBS adapter-trimmed reads of the DG F$_2$ mapping population (SRR24330286–SRR24330790), Omni-C raw reads (SRR24133794), and the assembled genome and annotation (GCA_032173615.1; JAVQLX000000000). The shotgun resequencing datasets generated during the current study have been deposited to NCBI BioProject database (PRJNA1064563). The raw data for cyanogenesis analyses are available at Additional file 1: Table S7 and Additional file 3: Table S8.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 24 January 2024 Accepted: 24 July 2024
Published online: 07 August 2024

References

1. Zeven AC. Four hundred years of cultivation of Dutch white clover landraces. Euphytica. 1991;54(1):93–9.
2. Taylor NL. A century of clover breeding developments in the united states. Crop Sci. 2008;48(1):1–13.
3. Abberton MT, Marshall AH. White clover. In: Boller B, Posselt UK, Veronesi F, editors. Fodder crops and amenity grasses. New York: NY: Springer New York; 2010. p. 457–76.
4. Kjærgaard T. A plant that changed the world: the rise and fall of clover 1000–2000. Landsc Res. 2003;28(1):41–9.
5. Santangelo JS, Ness RW, Cohan B, Fitzpatrick CR, Innes SG, Koch S, et al. Global urban environmental change drives adaptation in white clover. Science. 2022;375(6586):1275–81.
6. Kooyers NJ, Olsen KM. Searching for the bull's eye: agents and targets of selection vary among geographically disparate cyanogenesis clines in white clover (*Trifolium repens* L.). Heredity. 2013;111(6):495–504.
7. Daday H. Gene frequencies in wild populations of *Trifolium repens* L. Distribution by latitude Heredity. 1954;8(1):61–78.
8. Daday H. Gene frequencies in wild populations of *Trifolium repens*. II Distribution by altitude Heredity. 1954;8(3):377–84.

9. Griffiths AG, Moraga R, Tausen M, Gupta V, Bilton TP, Campbell MA, et al. Breaking free: the genomics of allopolyploidy-facilitated niche expansion in white clover. Plant Cell. 2019;31(7):1466–87.
10. Ravagnani A, Abberton MT, Skøt L. Development of genomic resources in the species of Trifolium L. and its application in forage legume breeding. Agronomy. 2012;2(2):116–31.
11. Wang H, Wu Y, He Y, Li G, Ma L, Li S, et al. High-quality chromosome-level de novo assembly of the *Trifolium repens*. BMC Genomics. 2023;24(1):326.
12. Santangelo JS, Battlay P, Hendrickson BT, Kuo WH, Olsen KM, Kooyers NJ, et al. Haplotype-resolved, chromosome-level assembly of white clover (*Trifolium repens* L., Fabaceae). Genome Biol Evol. 2023;15(8):2023.06.06.543960.
13. Olsen KM, Goad DM, Wright SJ, Dutta ML, Myers SR, Small LL, et al. Dual-species origin of an adaptive chemical defense polymorphism. New Phytol. 2021;232(3):1477–87.
14. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell. 2020;182(1):145–61e23.
15. Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell. 2021;184(13):3542–58.e16.
16. Zmienko A, Marszalek-Zenczak M, Wojciechowski P, Samelak-Czajka A, Luczak M, Kozlowski P, et al. AthCNV: a map of DNA copy number variations in the Arabidopsis genome. Plant Cell. 2020;32(6):1797–819.
17. Zhuang W, Chen H, Yang M, Wang J, Pandey MK, Zhang C, et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nat Genet. 2019;51(5):865–76.
18. Pos O, Radvanszky J, Buglyo G, Pos Z, Rusnakova D, Nagy B, et al. DNA copy number variation: main characteristics, evolutionary significance, and pathological aspects. Biomed J. 2021;44(5):548–59.
19. Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, et al. Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. Science. 2012;338(6111):1206–9.
20. Wang Y, Xiong G, Hu J, Jiang L, Yu H, Xu J, et al. Copy number variation at the *GL7* locus contributes to grain size diversity in rice. Nat Genet. 2015;47(8):944–8.
21. Wright D, Boije H, Meadows JR, Bed'hom B, Gourichon, Vieaud A, et al. Copy number variation in intron 1 of *SOX5* causes the Pea-comb phenotype in chickens. PLoS Genet. 2009;5(6):e1000512.
22. Stalder L, Oggenfuss U, Mohd-Assaad N, Croll D. The population genetics of adaptation through copy number variation in a fungal plant pathogen. Mol Ecol. 2023;32(10):2443–60.
23. Wang X, Li MW, Wong FL, Luk CY, Chung CYL, Yung WS, et al. Increased copy number of gibberellin 2-oxidase 8 genes reduced trailing growth and shoot length during soybean domestication. Plant J. 2021;107(6):1739–55.
24. Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA. Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). PLoS ONE. 2012;7(3):e33234.
25. Kreiner JM, Hnatovska S, Stinchcombe JR, Wright SI. Quantifying the role of genome size and repeat content in adaptive variation and the architecture of flowering time in *Amaranthus tuberculatus*. PLoS Genet. 2023;19(12):e1010865.
26. Ishikawa A, Stuart YE, Bolnick DI, Kitano J. Copy number variation of a fatty acid desaturase gene *Fads2* associated with ecological divergence in freshwater stickleback populations. Biol Lett. 2021;17(8):20210204.
27. Ishikawa A, Kabeya N, Ikeya K, Kakioka R, Cech JN, Osada N, et al. A key metabolic gene for recurrent freshwater colonization and radiation in fishes. Science. 2019;364(6443):886–9.
28. DeBolt S. Copy number variation shapes genome diversity in arabidopsis over immediate family generational scales. Genome Biol Evol. 2010;2:441–53.
29. Hughes MA. The cyanogenic polymorphism in Trifolium repens L. (white clover). Heredity. 1991;66(1):105–15.
30. Olsen KM, Sutherland BL, Small LL. Molecular evolution of the Li/li chemical defence polymorphism in white clover (Trifolium repens L.). Mol Ecol. 2007;16(19):4180–93.
31. Olsen KM, Hsu SC, Small LL. Evidence on the molecular basis of the Ac/ac adaptive cyanogenesis polymorphism in white clover (Trifolium repens L.). Genetics. 2008;179(1):517–26.
32. Olsen KM, Small LL. Micro- and macroevolutionary adaptation through repeated loss of a complete metabolic pathway. New Phytol. 2018;219(2):757–66.
33. Olsen KM, Kooyers NJ, Small LL. Adaptive gains through repeated gene loss: parallel evolution of cyanogenesis polymorphisms in the genus *Trifolium* (Fabaceae). Philos Trans R Soc Lond B Biol Sci. 2014;369:(1648).
34. Kooyers NJ, Olsen KM. Adaptive cyanogenesis clines evolve recurrently through geographical sorting of existing gene deletions. J Evol Biol. 2014;27(11):2554–8.
35. Olsen KM, Kooyers NJ, Small LL. Recurrent gene deletions and the evolution of adaptive cyanogenesis polymorphisms in white clover (Trifolium repens L.). Mol Ecol. 2013;22(3):724–38.
36. Kooyers NJ, Olsen KM. Rapid evolution of an adaptive cyanogenesis cline in introduced North American white clover (Trifolium repens L.). Mol Ecol. 2012;21(10):2455–68.
37. Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, et al. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol. 2010;186(1):37–45.
38. Hand ML, Ponting RC, Drayton MC, Lawless KA, Cogan NO, Charles Brummer E, et al. Identification of homologous, homoeologous and paralogous sequence variants in an outbreeding allopolyploid species based on comparison with progenitor taxa. Mol Genet Genomics. 2008;280(4):293–304.
39. Zhao Y, Zhang R, Jiang K-W, Qi J, Hu Y, Guo J, et al. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. Mol Plant. 2021;14(5):748–73.
40. Kongsaeree PT, Ratananikom K, Choengpanya K, Tongtubtim N, Sujiwattanarat P, Porncharoennop C, et al. Substrate specificity in hydrolysis and transglucosylation by family 1 β-glucosidases from cassava and Thai rosewood. J Mol Catal B Enzym. 2010;67(3):257–65.
41. McMahon J, Sayre R, Zidenga T. Cyanogenesis in cassava and its molecular manipulation for crop improvement. J Exp Bot. 2021;73(7):1853–67.
42. Schmidt FB, Cho SK, Olsen CE, Yang SW, Møller BL, Jørgensen K. Diurnal regulation of cyanogenic glucoside biosynthesis and endogenous turnover in cassava. Plant Direct. 2018;2(2):e00038.
43. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. 2017;37(12):4302–15.
44. McClintock B. The significance of responses of the genome to challenge. Science. 1984;226(4676):792–801.
45. Chen ZJ, Ni Z. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. BioEssays. 2006;28(3):240–52.
46. Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. Gene retention, fractionation and subgenome differences in polyploid plants. Nat Plants. 2018;4(5):258–68.
47. Otto SP. The evolutionary consequences of polyploidy. Cell. 2007;131(3):452–62.
48. Ellison NW, Liston A, Steiner JJ, Williams WM, Taylor NL. Molecular phylogenetics of the clover genus (*Trifolium*—Leguminosae). Mol Phylogenet Evol. 2006;39(3):688–705.
49. Coop I. Cyanogenesis in white clover (Trifolium repens L.) III. A study of linamarase, the enzyme which hydrolyses lotaustralin. N Z J Sci Technol Sect B. 1940;22:71–83.
50. Melville J, Doak B. Cyanogenesis in white clover (Trifolium repens L.) II. isolation of the gluco-sidal constituents. N Z J Sci Technol Sect B. 1940;22:67–71.
51. Corkill L. Cyanogenesis in white clover (Trifolium repens L.) V. The inheritance of cyanogenesis. N Z J Sci Technol Sect B. 1942;23:178–93.
52. Armstrong HE, Armstrong EF, Horton E. Herbage Studies. II-Variation in Lotus corniculatus and Trifolium repens (Cyanophoric plants). Proc Natl Acad Sci U S A. 1913;86(587):262–9.
53. Ware WM. Experiments and observations on forms and strains of *Trifolium repens* L. J Agric Sci. 1925;15(1):47–67.
54. Daday H. Gene frequencies in wild populations of *Trifolium repens* L. III World distribution Heredity. 1958;12(2):169–84.
55. Kooyers NJ, Hartman Bakken B, Ungerer MC, Olsen KM. Freeze-induced cyanide toxicity does not maintain the cyanogenesis polymorphism in white clover (*Trifolium repens*). Am J Bot. 2018;105(7):1224–31.

56. Kooyers NJ, Gage LR, Al-Lozi A, Olsen KM. Aridity shapes cyanogenesis cline evolution in white clover (Trifolium repens L.). Mol Ecol. 2014;23(5):1053–70.
57. Briggs D, Walters SM. Plant variation and evolution. 4th ed. Cambridge: Cambridge University Press; 2016.
58. Futuyma DJ, Kirkpatrick M. Evolution. Fourth edition ed. Sunderland, Massachusetts: Sinauer Associates, Inc., Publishers Sunderland, Massachusetts; 2017.
59. Ogbonna AC, Braatz de Andrade LR, Rabbi IY, Mueller LA, Jorge de Oliveira E, Bauchet GJ. Large-scale genome-wide association study, using historical data, identifies conserved genetic architecture of cyanogenic glucoside content in cassava (*Manihot esculenta* Crantz) root. Plant J. 2021;105(3):754–70.
60. Chen C, Zhang K, Liu F, Wang X, Yao Y, Niu X, et al. Resequencing of global *Lotus corniculatus* accessions reveals population distribution and genetic loci, associated with cyanogenic glycosides accumulation and growth traits. BMC Biol. 2023;21(1):176.
61. Kristensen C, Morant M, Olsen CE, Ekstrom CT, Galbraith DW, Moller BL, et al. Metabolic engineering of dhurrin in transgenic Arabidopsis plants with marginal inadvertent effects on the metabolome and transcriptome. Proc Natl Acad Sci U S A. 2005;102(5):1779–84.
62. Takos AM, Rook F. Why biosynthetic genes for chemical defense compounds cluster. Trends Plant Sci. 2012;17(7):383–8.
63. Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, et al. Adaptation genomics: the next generation. Trends Ecol Evol. 2010;25(12):705–12.
64. Song B, Ning W, Wei D, Jiang M, Zhu K, Wang X, et al. Plant genome resequencing and population genomics: current status and future prospects. Mol Plant. 2023;16(8):1252–68.
65. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170–5.
66. Arima-HiC Mapping Pipeline: Arima Genomics, Inc.; 2019 [Available from: https://github.com/ArimaGenomics/mapping_pipeline.
67. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15(8):e1007273.
68. Catchen J, Amores A, Bassham S. Chromonomer: a tool set for repairing and enhancing assembled genomes through integration of genetic maps and conserved synteny. G3. 2020;10(11):4115–28.
69. Rice E. agptools: tools for working with agp files [Available from: https://warrenlab.github.io/agptools/.
70. Jin J-J, Yu W-B, Yang J-B, Song Y, Depamphilis CW, Yi T-S, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:241. https://doi.org/10.1186/s13059-020-02154-5.
71. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics. 2011;27(6):764–70.
72. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14):2202–4.
73. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245. https://doi.org/10.1186/s13059-020-02134-9.
74. Wright SJ, Goad DM, Gross BL, Muñoz PR, Olsen KM. Genetic trade-offs underlie divergent life history strategies for local adaptation in white clover. Mol Ecol. 2022;31(14):3742–60.
75. Poplin R, Ruano-Rubio V, Depristo MA, Fennell TJ, Carneiro MO, Van Der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017. https://doi.org/10.1101/201178.
76. Bilton TP, Schofield MR, Black MA, Chagné D, Wilcox PL, Dodds KG. Accounting for errors in low coverage high-throughput sequencing data when constructing genetic maps using biparental outcrossed populations. Genetics. 2018;209(1):65–76.
77. Broman KW, Wu H, Sen Ś, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19(7):889–90.
78. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3(1). https://doi.org/10.1093/nargab/lqaa108.
79. Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics. 2021;22:566. https://doi.org/10.1186/s12859-021-04482-0.
80. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29(1):15–21.
81. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674–6.
82. Orozco-Arias S, Humberto Lopez-Murillo L, Candamil-Cortés MS, Arias M, Jaimes PA, Rossi Paschoal A, et al. Inpactor2: a software based on deep learning to identify and classify LTR-retrotransposons in plant genomes. Brief Bioinform. 2023;24(1):bbac511. https://doi.org/10.1093/bib/bbac511.
83. Zhang R-G, Li G-Y, Wang X-L, Dainat J, Wang Z-X, Ou S, et al. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic Res. 2022;9:uhac017. https://doi.org/10.1093/hr/uhac017.
84. Neumann P, Novák P, Hoštáková N, Macas J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob DNA. 2019;10(1):1.
85. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
86. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530–4.
87. Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. eLife. 2022;11:e78526. https://doi.org/10.7554/eLife.78526.
88. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238.
89. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.e.
90. Bickhart DM, Koch LM, Smith TPL, Riday H, Sullivan ML. Chromosome-scale assembly of the highly heterozygous genome of red clover (Trifolium pratense L.), an allogamous forage crop species. Gigabyte. 2022;2022:1–13.
91. Shirasawa K, Moraga R, Ghelfi A, Hirakawa H, Nagasaki H, Ghamkhar K, et al. An improved reference genome for Trifolium subterraneum L. provides insight into molecular diversity and intra-specific phylogeny. Front Plant Sci. 2023;14:1103857.
92. Pecrix Y, Staton SE, Sallet E, Lelandais-Brière C, Moreau S, Carrère S, et al. Whole-genome landscape of *Medicago truncatula* symbiotic genes. Nat Plants. 2018;4(12):1017–25.
93. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol. 2013;31(3):240–6.
94. Li H, Jiang F, Wu P, Wang K, Cao Y. A high-quality genome sequence of model legume *Lotus japonicus* (MG-20) provides insights into the evolution of root nodule symbiosis. Genes. 2020;11(5):483.
95. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46(7):707–13.
96. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018;14(1):e1005944.
97. Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Mol Biol Evol. 2018;35(10):2582–4.
98. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. Kaks_calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics. 2010;8(1):77–80.
99. Haque MR, Bradbury JH. Preparation of linamarase solution from cassava latex for use in the cassava cyanide kit. Food Chem. 1999;67(3):305–9.

Publisher's Note