









Haplotype-Resolved, Chromosome-Level Assembly of White Clover (*Trifolium repens* L., Fabaceae)

James S. Santangelo ^{1,*}, Paul Battlay ², Brandon T. Hendrickson³, Wen-Hsi Kuo ⁴, Kenneth M. Olsen ⁴, Nicholas J. Kooyers ³, Marc T.J. Johnson ¹, Kathryn A. Hodgins ², and Rob. W. Ness ¹

¹Department of Biology, University of Toronto Mississauga, Mississauga, Ontario, Canada

²School of Biological Sciences, Monash University, Melbourne, Victoria, Australia

³Department of Biology, University of Louisiana, Lafayette, Louisiana, USA

⁴Department of Biology, Washington University in St. Louis, St. Louis, Missouri, USA

*Corresponding author: E-mail: james.santangelo37@gmail.com.

Accepted: 29 July 2023

Abstract

White clover (*Trifolium repens* L.; Fabaceae) is an important forage and cover crop in agricultural pastures around the world and is increasingly used in evolutionary ecology and genetics to understand the genetic basis of adaptation. Historically, improvements in white clover breeding practices and assessments of genetic variation in nature have been hampered by a lack of high-quality genomic resources for this species, owing in part to its high heterozygosity and allotetraploid hybrid origin. Here, we use PacBio HiFi and chromosome conformation capture (Omni-C) technologies to generate a chromosome-level, haplotype-resolved genome assembly for white clover totaling 998 Mbp (scaffold N50 = 59.3 Mbp) and 1 Gbp (scaffold N50 = 58.6 Mbp) for haplotypes 1 and 2, respectively, with each haplotype arranged into 16 chromosomes (8 per subgenome). We additionally provide a functionally annotated haploid mapping assembly (968 Mbp, scaffold N50 = 59.9 Mbp), which drastically improves on the existing reference assembly in both contiguity and assembly accuracy. We annotated 78,174 protein-coding genes, resulting in protein BUSCO completeness scores of 99.6% and 99.3% against the embryophyta_odb10 and fabales_odb10 lineage datasets, respectively.

Key words: allotetraploid, genome assembly, haplotype-resolved, legume, polyploidy.

Significance

We provide two white clover genome assemblies as part of this project: 1) a haplotype-resolved, chromosome-level assembly and 2) a functionally annotated haploid mapping assembly. These assemblies place white clover among the best sequenced legumes to date, and one of the best assemblies for a plant of recent polyploid origins. This work will facilitate marker-assisted breeding programs for traits of agronomic importance and provide increased resolution and ability to identify the genomic basis of adaptation in this increasingly used model in evolutionary ecology and genetics.

Introduction

White clover (*Trifolium repens* L., Fabaceae) is a prostrate, herbaceous perennial that spreads via stolons, forming large clonal patches up to 1 meter across (Burdon 1983). It originated as an allotetraploid in the Mediterranean 15–28 Ka resulting from the hybridization of its diploid

progenitors, *T. occidentale* and *T. palleescens* (Williams et al. 2012; Griffiths et al. 2019). Because of its rapid growth and symbiosis with nitrogen-fixing bacteria, white clover is an important forage crop in agricultural pastures, and it has become naturalized in diverse climates around the world over the last several hundred years (Burdon

© The Author(s) 2023. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

1983; Kjærsgaard 2003). Today, there are large efforts to improve production and survival in variable environments, including traits such as yield and biomass production (Barrett et al. 2009; Moeskjær et al. 2022), salt tolerance (Wang et al. 2010), drought tolerance (Annicchiarico and Piano 2004; Jiang et al. 2010), frost tolerance (Inostroza et al. 2018; Zhang et al. 2022), and disease resistance (Panter et al. 2012). Currently, most white clover breeding relies on phenotypic selection, although marker-assisted breeding designs are increasingly common and would be greatly facilitated by a well-annotated, chromosome-level reference genome assembly (Faville et al. 2012; Moeskjær et al. 2022).

In addition to its use in agricultural mixed-grass pastures and breeding programs, white clover has become a model in evolutionary ecology and genetics for understanding adaptation to environmental gradients and agents of selection in nature. Early work documented latitudinal and altitudinal clines in the frequency of cyanogenesis, the production of hydrogen cyanide in response to tissue damage—an anti-herbivore defense whose metabolic components can also affect tolerance to abiotic stressors (e.g., drought, frost) (Daday 1954a, 1954b, 1965, 1958; Hughes 1991). More recent work has corroborated these widespread continental clines (Innes et al. 2022; Kooyers and Olsen 2012, 2013), uncovered clines on smaller spatial scales across urban–rural gradients (Santangelo et al. 2022), identified the molecular mechanisms underlying genetic variation in cyanogenesis (Olsen et al. 2007, 2008; Olsen and Small 2018; Olsen et al. 2021), and experimentally tested the ecological factors maintaining the cyanogenesis polymorphism (Kooyers et al. 2014, 2018; Albano and Johnson 2023; Emad Fadoul et al. 2023) and its evolutionary consequences (Thompson and Johnson 2016; Santangelo et al. 2018). Although much of this work has focused on the cyanogenesis polymorphism—a trait with well-characterized inheritance attributable to two epistatically-interacting Mendelian loci—ongoing and future work will leverage white clover’s rich history in evolutionary ecology to examine the genetic basis of adaptation at various spatial scales for which a high-quality reference assembly will be essential. In particular, a chromosome-level, haplotype-resolved assembly would facilitate identifying structural variants involved in adaptation (Battlay et al. 2023) and improve our understanding of the evolutionary consequences of polyploidization in this ecologically and agronomically important allotetraploid.

Owing to the inherently repetitive nature of polyploid genomes, chromosome-level and haplotype-resolved genome assemblies have been challenging for these taxa. However, new technologies allow us to span difficult repetitive elements and offer the ability to greatly improve and expand earlier genome assemblies. Here, we present a chromosome-level, haplotype-resolved genome assembly of the model legume white clover using PacBio HiFi and

chromosome conformation capture (Dovetail Omni-C) technologies. We present two genomes as part of this project: 1) an unannotated, haplotype-resolved assembly and 2) a functionally annotated haploid mapping assembly, which we compare with the previous reference assembly (Griffiths et al. 2019) using two recently generated linkage maps for the species (Olsen et al. 2021).

Results and Discussion

Genome Assemblies

Our final haplotype-resolved assembly totaled 998,247,995 bp for haplotype 1 ($N = 693$ scaffolds total) and 1,009,398,733 bp for haplotype 2 ($N = 1,022$ scaffolds total) (table 1), slightly shorter than the ~1.1 Gbp previously estimated genome size for *T. repens* species (Griffiths et al. 2019). Haplotype 1 had genome BUSCO completeness

Table 1

Assembly and Annotation Statistics for the Haploid Reference (16 Chromosomes + 2 Organelles) and the Haplotype-Resolved Diploid Reference Assemblies

Statistic	Reference		
	Haploid	Diploid_Hap1	Diploid_Hap2
Assembly statistics			
Number of contigs	18	693	1022
Largest contig	67,442,382	67,442,382	66,872,020
Total length	968,215,888	998,247,995	1,009,398,733
N50	59,895,999	59,287,556	58,622,306
N90	54,808,019	54,808,019	47,359,824
L50	8	8	9
L90	15	15	16
GC %	33.70	33.97	34.05
Ns per 100 kbp	0.80	0.92	0.87
Annotation statistics			
# genes	78,174	—	—
# genes with common names	15,871	—	—
# transcripts (mRNA)	87,929	—	—
Transcript level			
Mean gene length	2,872	—	—
Total CDS exons	448,332	—	—
Single exon transcripts	18,658	—	—
Mean exon length	234	—	—
% of genome comprised of genes	23.2	—	—
Functional			
GO terms	54,159	—	—
Interproscan	67,338	—	—
Eggnog	77,327	—	—
Pfam	55,891	—	—
Cazyme	3,313	—	—
Merops	3,039	—	—
Busco	3,408	—	—

“—” means statistic was not estimated for those assemblies.

scores of 99.6% and 99.5% against the embryophyta ($N = 1,614$ genes total) and fabales ($N = 5,340$ genes total) lineage datasets, respectively (supplementary table S1, Supplementary Material online). Similarly, haplotype 2 had BUSCO completeness scores of 99.5% against both databases.

Our haploid mapping assembly totaled 968,215,888 bp assembled into 18 chromosomes (16 chromosomes + 2 organelles), with unplaced scaffolds removed from this assembly because our goal was to focus on assembled, contiguous chromosomes for this assembly. This resulted in the exclusion of 1,697 scaffolds from across both haplotypes, representing ~30 Mbp of sequence (~3% of genome). A total of 93.8% ($N = 2,186$) of the 2,330 filtered linkage markers for the “DG” mapping population mapped to the correct chromosome in our new assembly, which was a dramatic improvement compared with the 39.4% ($N = 919$) of the correctly mapped markers from the previous assembly (fig. 1A). Similar results were obtained for the “SG” mapping population (supplementary fig. S2, Supplementary Material online). The haploid mapping assembly showed complete BUSCO scores of 99.6% and 99.5% against the embryophyta and fabales lineage datasets, respectively, with most of these genes occurring in duplicate copy (supplementary table S1, Supplementary Material online).

Annotation

We softmasked 59.4% (~576.5 Mbp) of the haploid reference assembly (fig. 1B) to improve gene model prediction during annotation. Of the classified repetitive elements, most (27.2%) were Long Terminal Repeats (i.e., LTRs) elements, with Ty1/Copia (13.5%) and Gypsy/DIRS1 (9.7%) elements making up the majority (supplementary table S2, Supplementary Material online). We annotated 78,174 genes consisting of 87,929 messenger RNA (mRNA) transcripts that together account for 23.2% of the genome (table 1 and fig. 1B). Thirty-nine thousand four hundred twenty-five of our annotated genes occur on the *T. occidentale* subgenome, with the remaining 38,749 on the *T. pallescens* subgenome, consistent with the number of genes of closely related diploid *Trifolium* species (*T. pratense*: 43,682; *T. subterraneum*: 42,704). Synteny between the subgenomes is largely preserved, except for three translocations between nonhomoeologous chromosomes and six inversions between homoeologous chromosomes (fig. 1C). Of the 78,174 genes, 4,868 (6.2%) are completely overlapped by repeats and likely represent transposable element protein-coding sequences. Most mRNAs (~87%; $N = 77,043$) had at least one functional annotation (table 1), with 15,871 genes containing common names. Our final annotated protein set had complete protein BUSCO scores of 99.6% and 99.3% against

the embryophyta and fabales lineage datasets, respectively (supplementary table S1, Supplementary Material online).

Conclusion

We have provided a chromosome-level, haplotype-resolved genome assembly of the allotetraploid white clover (*T. repens*), and a functionally annotated haploid mapping assembly that shows substantial improvements over the existing reference genome for the species. These assemblies will facilitate marker-assisted breeding programs for traits of agronomic importance and provide increased resolution and ability to identify the genomic basis of adaptation in this increasingly used model in evolutionary ecology and genetics. Together with an alternative and upcoming chromosome-level assembly (Wang et al. 2023) and other high-quality reference genomes in the genus (Dluhošová et al. 2018; Bickhart et al. 2022; Shirasawa et al. 2023), our haplotype-resolved assembly will be particularly useful for identifying structural variation and facilitate the development of pangenomic references (Eizenga et al. 2020) for which haplotype-resolved assemblies are an asset (Garg et al. 2022).

Materials and Methods

Plant Sample

We sequenced an F4 *T. repens* genotype that was generated as part of a separate experiment. As a diploidized allotetraploid, *T. repens* exhibits disomic inheritance with chromosomes from both subgenomes segregating independently, and plants are obligately outcrossing due to a gametophytic self-incompatibility. The sequenced plant originated from an F0 crosses between a plant from Ontario, Canada, and a plant from Louisiana, USA, followed by three generations of random crossing among the F1s, F2s, and finally the F3s. The sequenced plant was maintained in a 1 L pot in potting soil (Pro-Mix LP15; Premier Tech, Rivière-du-Loup, Canada) in a growth chamber set to 25 °C on a 12 h light:12 h dark cycle, though the plant was maintained in the dark for 48 h prior to sampling to reduce polysaccharide content. The plant was nondestructively harvested on March 28, 2022, by sampling approximately 2.5 g of leaf tissue, immediately flash-freezing tissue in liquid nitrogen, and storing it at -80 °C prior to shipping on dry ice to Dovetail Genomics for DNA extraction, library preparation, and sequencing.

Sequencing

DNA samples were quantified using a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The PacBio SMRTbell library (~20 kbp mean insert length) for PacBio Sequel was constructed using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA) using the manufacturer’s recommended protocol. The library was bound to polymerase using the Sequel II Binding Kit 2.0

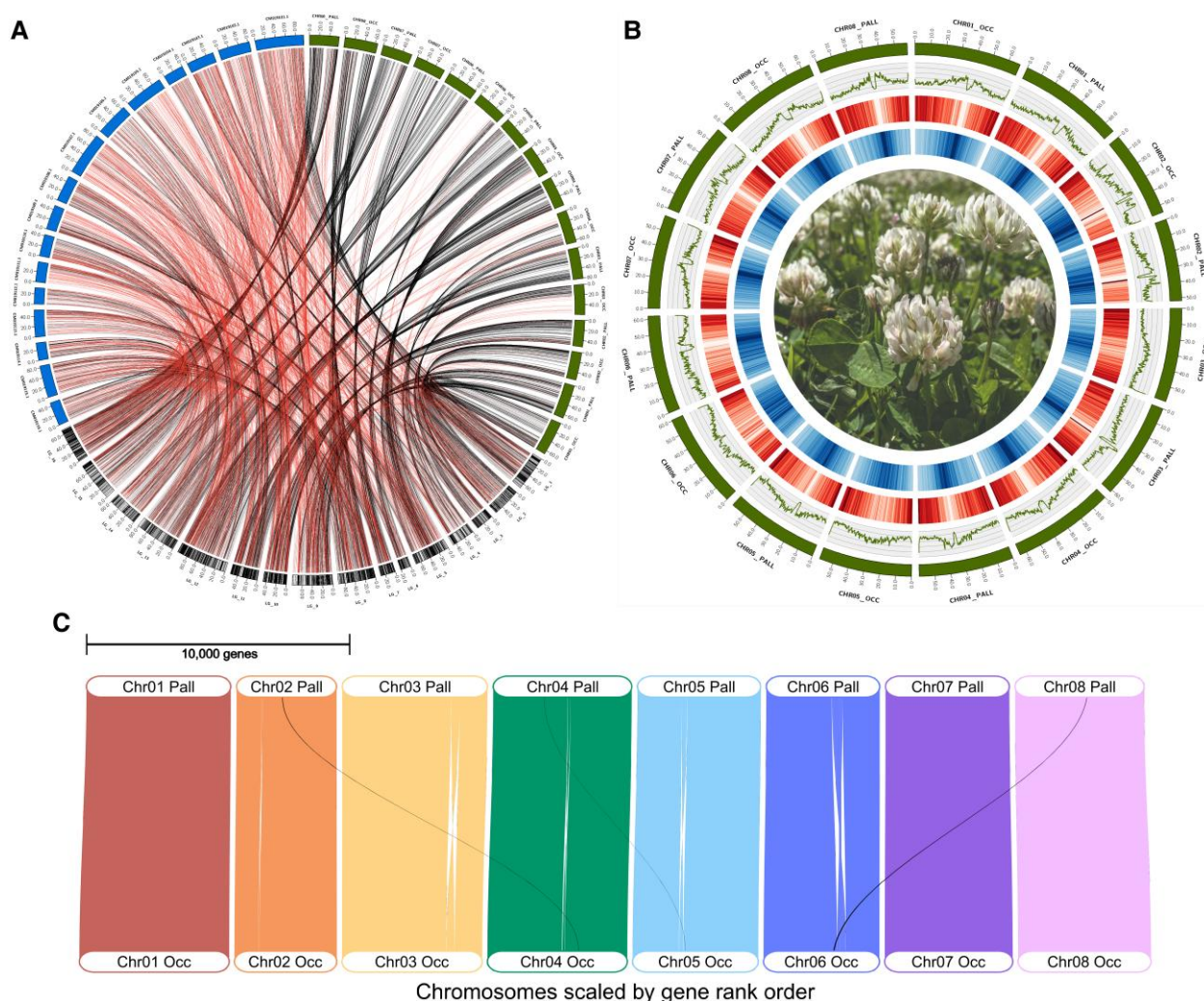


Fig. 1—(A) Linkage map from the “DG” mapping population (Olsen et al. 2021, bottom, converted to physical positions in Mbp) with markers (vertical black lines in ideogram) connected to their physical positions in both the previous reference assembly (Griffiths et al. 2019, blue, upper left) and the current haploid assembly (green, upper right). Lines connecting markers to their physical position are colored red if they map to the wrong chromosome based on the linkage data or black if they map to the correct chromosome. (B) Circos plot of haploid mapping assembly consisting of 16 chromosomes. From outside to inside: chromosomes (green ideograms), GC%, gene density (red), repeat proportion (blue), and a photo of flowering *Trifolium repens* (credit: James Santangelo). GC%, gene density, and repeat density were estimated in 500 Kb windows with a 100 Kb step. (C) GENESPACE Riparian plot showing synteny between the *T. occidentale* (bottom) and *T. pallescens* (top) subgenomes of *T. repens*. Black lines show inferred translocations between nonhomoeologous chromosomes ($N = 3$), whereas white gaps within homoeologous chromosomes show inversions ($N = 6$; Chr_06 contains two nested inversions).

(PacBio) and loaded onto PacBio Sequel II. Sequencing was performed on PacBio Sequel II 8 M SMRT cells generating 58 Gbp of data. These PacBio Circular Consensus Sequencing (i.e., CCS) reads were used as an input to “hifiasm” v0.16.1-r375 (Cheng et al. 2022, 2021) (see Scaffolding below).

For each Dovetail Omni-C library, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with DNase I; chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed,

and the DNA was purified. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before polymerase chain reaction enrichment of each library. The library was sequenced on an Illumina HiSeqX platform (Illumina, San Diego, California, USA) to produce $\sim 30\times$ sequence coverage. The PacBio CCS reads and Omni-C reads ($MQ > 50$) were then used as input for “hifiasm” to produce two haplotype-resolved assemblies (hap1 and hap2) using default parameters.

Scaffolding

We first produced an initial assembly of all PacBio HiFi data with "hifiasm" in the "primary" mode. This resulted in two sets of contigs: primary and alternative. We then combined primary and alternative contigs into a single set of all contigs, containing 1,384,338,092 bp of sequence in 6,189 contigs with N50 size of 15,304,949 bp, which we call the "unresolved" contig set below. Next, to determine which contig was derived from which subgenome, we used Illumina reads for the diploid parental species *Trifolium occidentale* (SRR8593471) and *Trifolium pallescens* (SRR8617466) downloaded from NCBI's Sequence Read Archive (SRA). We mapped the Illumina reads to the unresolved contig set with "bwa mem" (Li and Durbin 2009), used the best alignment for each Illumina read, counted the number of reads from each parental species mapped to each contig, and divided it by the total number of Illumina reads in each parental set. Based on the weighted number of alignments of the parental reads, we labeled the contigs in the unresolved set with "Pal" and "Occ" labels corresponding to the two subgenomes, resulting in a "labeled" set of contigs.

We then aligned the PacBio HiFi reads and the Omni-C reads to the labeled contigs with the "minimap2" (Li 2021) and "bwa mem" aligners, computed the best alignment for each read, and split the HiFi and Omni-C reads into subsets for each subgenome. We required that both Omni-C reads have the best alignment to the same subgenome to be assigned to that subgenome. Next, we assembled the two subsets of HiFi/Omni-C reads separately with "hifiasm" *Hi-C* in haplotype resolved mode. This yielded haplotype-resolved assemblies for the two subgenomes. We then scaffolded the assemblies with "HiRise" scaffolder (Putnam et al. 2016) and closed gaps in the scaffolds with "SAMBA" scaffolder (Zimin and Salzberg 2022). The final step was to remove redundant haplotype contigs that "hifiasm" sometimes keeps in the assembly. We did this by aligning all contigs shorter than 1 Mbp to the assembly for each of the haplotypes with "nucmer" (Marçais et al. 2018) and excluding the contigs that mapped to the interior of other bigger contigs with better than 95% similarity over at least 75% of their length. This resulted in the final set of assembled haplotypes (2 subgenomes \times 2 haplotypes each = 4 haplotypes).

Assembly

All analyses from here forward are implemented in an open and reproducible Snakemake v7.16 pipeline (Mölder et al. 2021). The pipeline begins with input of the Dovetail haplotype assemblies, associated AGP (i.e., "A Golden Path") files and linkage map data from (Olsen et al. 2021), and ends with the generation of the phased diploid assembly in FASTA format (NCBI BioProjects PRJNA957817 and PRJNA957816), the annotated haploid mapping assembly in FASTA, NCBI Sequin, and GFF3 formats (BioProject

PRJNA951196), and manuscript figures. See Data Accessibility for links to data and code.

Before assembling the reference genomes, the assembled haplotypes required manual curation to correct minor misassemblies and fill gaps generated during scaffolding. First, we used BLAST v2.12.0 (Altschul et al. 1990) to align two previously generated linkage maps for the species (Olsen et al. 2021) to each of the four haplotypes and to the previous *T. repens* reference genome (Griffiths et al. 2019). We removed alignments that were less than 175 bp in length of the 200 bp total length for each linkage mapping marker sequence and had less than 95% identity, and retained only the best alignment (i.e., lowest E-value) for each marker. These alignments were used to identify misassembled scaffolds and to assess correspondence between the scaffolds in the newly assembled haplotypes and the chromosomes in the previous reference genome.

Second, we used "Minimap2" v2.24 (Li 2021) to generate pairwise alignments between all four haplotypes. Together with the linkage map alignments above, these alignments enabled us to fill in three gaps (likely spanning the centromere) and one telomere with unplaced scaffolds (supplementary fig. S1, Supplementary Material online). In addition, the scaffolding generated a double telomere at the end of one of the chromosomes in the *T. pallescens* subgenome; this extra telomere was removed and added to its correct location at the end of the homoeologous chromosome in the *T. occidentale* subgenome (supplementary fig. S1, Supplementary Material online). All manual fixes were implemented in BioPython v1.8 (Cock et al. 2009).

We used the revised haplotypes to generate two separate reference genome assemblies: a haplotype-resolved assembly and a collapsed haploid mapping assembly. As a diploidized allotetraploid (see above), *T. repens'* four haplotypes can be collapsed into two haplotypes, each containing eight chromosomes from each subgenome (i.e., $N = 16$) resulting in a phased "diploid" assembly (i.e., $2N = 32$). We therefore present this assembly as two FASTA files, with one for each of these two haplotypes. These FASTA files additionally include all unplaced scaffolds for each of the haplotypes. We additionally created a haploid mapping assembly, generated by taking the longer chromosome of each of the two haplotypes for each linkage group. This haploid mapping assembly was used for the structural and functional annotation described below. Both the diploid and haploid assemblies were checked for annotation completeness by running BUSCO v5.4.6 (Seppey et al. 2019) in "genome" mode against the embryophyta_odb10 and fabales_odb10 lineage datasets.

Structural Annotation

To improve gene-model predictions, we softmasked repeats prior to proceeding with annotation. First, we used

"RepeatModeler" v2.0.3 (Flynn et al. 2020) to generate a repeat library using the haploid mapping reference as input. This database was then merged with RepBase (v. RedBaseRepeatMaskerEdition-20181026), and the combined repeat library was used to softmask repeats using "RepeatMasker" v4.1.3.

We predicted gene models and generated a structural annotation of the haploid mapping assembly by combining evidence from proteins in related plant species and RNA-Seq evidence in *T. repens*. First, we ran "BRAKER" v3.0.0 (Brůna et al. 2021) in "protein mode" using proteins from all green plants (i.e., Viridiplantae) as input, supplemented with proteins from all legumes (family: Fabaceae) from the UniProtKB database ($N=1,233,771$ proteins). Next, we downloaded a subset of all RNA-Seq data from *T. repens* available from four published sources (Nagy et al. 2013; Griffiths et al. 2019; Zhou et al. 2021; Zhang et al. 2022), selected to represent diverse tissue types and library preparation protocols ($N=21$ RNAseq libraries total, [supplementary table S3, Supplementary Material](#) online). We mapped the raw RNA-Seq reads to the haploid mapping reference using "STAR" v2.7.0b (Dobin et al. 2013) in "two-pass mode" and merged the resulting BAM files using SAMtools v1.16.1 (Li et al. 2009). We used this merged BAM file as input to "BRAKER" in "RNAseq" mode. Next, we combined evidence from "protein" and "RNAseq" modes using "TSEBRA" (Gabriel et al. 2021) before using the "agat_convert_gxf2gxf.pl" script from "AGAT" v1.0.0-pl5321hdfd78af_0 (Dainat et al. 2022) to convert the BRAKER-generated GTF to GFF3 format and proceeding to functional annotation.

Functional Annotation

We added functional annotations by querying extracted proteins against numerous databases prior to merging and formatting annotations for uploading to NCBI. First, we retrieved functional annotations using "InterProScan" v5.61.93-0 (Jones et al. 2014) and "Eggnog-mapper" v2.1.10 (Cantalapiedra et al. 2021). The resulting outputs were passed as input to "funannotate" v1.8.14 (Palmer and Stajich 2019), which combined the annotations and queried some additional databases. For any protein annotated as "hypothetical protein" and containing a fully resolved enzyme commission (i.e., EC) number (i.e., resolved to four digits), we replaced the "hypothetical protein" annotation with the EC number's product in the ExPASSY Enzyme database (Bairoch 2000). If the EC number was only resolved to three digits or fewer, we kept the "hypothetical protein" annotation and removed the EC number. In the end, the following databases were queried for annotations: InterPro v93.0 (Blum et al. 2021), EggNog v5.0 (Huerta-Cepas et al. 2019), MEROPS v12.0 (Rawlings et al. 2018), Uniprot v2023_01 (UniProt

Consortium 2023), dbCan v11.0 (Yin et al. 2012), Pfam v35.0 (Mistry et al. 2021), GO v2023-03-06 (Ashburner et al. 2000; Gene Ontology Consortium 2021), and MiBig v1.4 (Terlouw et al. 2023). We queried our final annotation against the embryophyta_odb10 and fabales_odb10 BUSCO databases using BUSCO v5.4.6 (Seppey et al. 2019) in "protein" mode and assessed synteny between the *T. occidentale* and *T. pallescens* subgenomes using GENESPACE with default parameters except "useHOGs" and "orthofinderInBlk" were set to TRUE (Lovell et al. 2022).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

Acknowledgments

We thank Rory Craig for thoughtful discussions that greatly improved the genome and manuscript. Plant lines used for sequencing were originally created by L. Albano and subsequently maintained by H. Fargo, K. Bhachu, and I. Arif. DNA extraction, genome sequencing, and scaffolding were provided by Dovetail genomics.

Author Contributions

J.S.S.: conceptualization, sample preparation, software, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, and visualization. P.B.: software, formal analysis, and writing—review and editing. B.T.H.: software, formal analysis, and data curation—review and editing. W.-H.K.: validation and writing—review and editing. K.M.O.: writing—review and editing. N.J.K.: conceptualization, writing—review and editing, and funding acquisition. M.T.J.J.: conceptualization, sample preparation, interpretation, reviewing and editing, and funding acquisition. K.A.H.: conceptualization, writing—review and editing, and funding acquisition. R.W.N.: conceptualization, writing—review and editing, and funding acquisition.

Funding

J.S.S. was supported by an NSERC PDF. B.T.H. and N.J.K. were supported by NSF OIA-1920858. K.M.O. and W.-H.K. were supported by NSF IOS-1557770. M.T.J.J. and R.W.N. were supported by independent NSERC Discovery grants, and M.T.J.J. was further supported by a Canada Research Chair and E.W.R. Steacie Fellowship. K.A.H. was supported by ARC DP220102362 and HSPF RGP0001/2019.

Data Availability

All code to reproduce this manuscript's results can be found on JSS's GitHub (see <https://github.com/James-S-Santangelo/dcg>) and is archived on Zenodo (<https://zenodo.org/record/8,180,534>). In addition to code, the Zenodo repository contains the raw haplotypes assembled by Dovetail and the AGP files required as input to the pipeline. All other data are contained within the GitHub/Zenodo repository, except for proprietary databases (e.g., RepBase) that could not be included (see GitHub README). The raw data used in the assemblies have been deposited on NCBI (BioProject [PRJNA979795](https://ncbi.nlm.nih.gov/bioproject/PRJNA979795)) along with the annotated haploid mapping assembly (BioProject [PRJNA951196](https://ncbi.nlm.nih.gov/bioproject/PRJNA951196)) and individual haplotype assemblies (BioProjects [PRJNA957816](https://ncbi.nlm.nih.gov/bioproject/PRJNA957816) and [PRJNA957817](https://ncbi.nlm.nih.gov/bioproject/PRJNA957817)).

Literature Cited

- Albano LJ, Johnson MTJ. 2023. Interactions between environmental factors drive selection on cyanogenesis in *Trifolium repens*. *Oikos* 23:e09629.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.
- Annicchiarico P, Piano E. 2004. Indirect selection for root development of white clover and implications for drought tolerance. *J Agron Crop Sci.* 190:28–34.
- Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet.* 25:25–29.
- Bairoch A. 2000. The ENZYME database in 2000. *Nucleic Acids Res.* 28: 304–305.
- Barrett B, Baird I, Woodfield D. 2009. White clover seed yield: a case study in marker-assisted selection. In: Yamada T and Spangenberg G, editors. *Molecular breeding of forage and turf*. New York (NY): Springer. p. 241–250.
- Battlay P, et al. 2023. Large haploblocks underlie rapid adaptation in the invasive weed *Ambrosia artemisiifolia*. *Nat Commun.* 14:1717.
- Bickhart DM, Koch LM, Smith TPL, Riday H, Sullivan ML. 2022. Chromosome-scale assembly of the highly heterozygous genome of red clover (*Trifolium pratense* L.), an allogamous forage crop species. *GigaByte* 2022:1–13.
- Blum M, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49:D344–D354.
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform.* 3:lqaa108.
- Burdon JJ. 1983. Biological flora of the British isles: *trifolium repens* L. *J Ecol.* 71:307–330.
- Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol.* 38:5825–5829.
- Cheng H, et al. 2022. Haplotype-resolved assembly of diploid genomes without parental data. *Nat Biotechnol.* 40:1332–1335.
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 18:170–175.
- Cock PJA, et al. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.
- Daday H. 1954a. Gene frequencies in wild populations of *Trifolium repens* I. Distribution by latitude. *Heredity (Edinb).* 8:61–78.
- Daday H. 1954b. Gene frequencies in wild populations of *Trifolium repens* II. Distribution by altitude. *Heredity (Edinb).* 8:377–384.
- Daday H. 1958. Gene frequencies in wild populations of *Trifolium repens* L III. World distribution. *Heredity (Edinb).* 12:169–184.
- Daday H. 1965. Gene frequencies in wild populations of *Trifolium repens* L IV. Mechanism of natural selection. *Heredity (Edinb).* 20: 355–365.
- Dainat J et al. 2022. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format (Version v1.0.0). Zenodo. <https://www.doi.org/10.5281/zenodo.3552717>.
- Dluhošová J, Ištváněk J, Nedělník J, Řepková J. 2018. Red clover (*Trifolium pratense*) and zigzag clover (*T. medium*)—a picture of genomic similarities and differences. *Front Plant Sci.* 9:724.
- Dobin A, et al. 2013. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* 29:15–21.
- Eizenga JM, et al. 2020. Pangenome graphs. *Annu Rev Genomics Hum Genet.* 21:139–162.
- Fadoul H E, Albano LJ, Bergman ME, Phillips MA, Johnson MTJ. 2023. Assessing the benefits and costs of the hydrogen cyanide antiherbivore defense in *Trifolium repens*. *Plants* 12:1213.
- Faville MJ, Griffiths AG, Jahufer MZZ, Barrett BA. 2012. Progress towards marker-assisted selection in forages. *ProNzG* 74:189–194.
- Flynn JM, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 117: 9451–9457.
- Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. 2021. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics.* 22:566.
- Garg S, Balboa R, Kuja J. 2022. Chromosome-scale haplotype-resolved pangenomics. *Trends Genet.* 38:1103–1107.
- Gene Ontology Consortium. 2021. The gene ontology resource: enriching a GOLD mine. *Nucleic Acids Res.* 49:D325–D334.
- Griffiths AG, et al. 2019. Breaking free: the genomics of allopolyploidy-facilitated niche expansion in white clover. *Plant Cell.* 31:1466–1487.
- Huerta-Cepas J, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47:D309–D314.
- Hughes MA. 1991. The cyanogenic polymorphism in *Trifolium repens* L (white clover). *Heredity (Edinb).* 66:105–115.
- Innes SG, Santangelo JS, Kooyers NJ, Olsen KM, Johnson MTJ. 2022. Evolution in response to climate in the native and introduced ranges of a globally distributed plant. *Evolution* 76:1495–1511.
- Inostroza L, et al. 2018. Understanding the complexity of cold tolerance in white clover using temperature gradient locations and a GWAS approach. *Plant Genome.* 11:170096.
- Jiang Q, et al. 2010. Improvement of drought tolerance in white clover (*Trifolium repens*) by transgenic expression of a transcription factor gene WXP1. *Funct Plant Biol.* 37:157–165.
- Jones P, et al. 2014. Interproscan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240.
- Kjærsgaard T. 2003. A plant that changed the world: the rise and fall of clover 1000–2000. *Landscape Res.* 28:41–49.
- Kooyers NJ, Gage LR, Al-Lozi A, Olsen KM. 2014. Aridity shapes cyanogenesis cline evolution in white clover (*Trifolium repens* L. *Mol Ecol.* 23:1053–1070.
- Kooyers NJ, Hartman Bakken B, Ungerer MC, Olsen KM. 2018. Freeze-induced cyanide toxicity does not maintain the cyanogenesis polymorphism in white clover (*Trifolium repens*). *Am J Bot.* 105:1224–1231.
- Kooyers NJ, Olsen KM. 2012. Rapid evolution of an adaptive cyanogenesis cline in introduced north American white clover (*Trifolium repens* L.). *Mol Ecol.* 21:2455–2468.

- Kooyers NJ, Olsen KM. 2013. Searching for the bull's Eye: agents and targets of selection vary among geographically disparate cyanogenesis clines in white clover (*Trifolium repens* L.). *Heredity* (Edinb). 111:495–504.
- Li H, et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.
- Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 37:4572–4574.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Lovell JT, et al. 2022. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* 11:e78526.
- Marçais G, et al. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 14:e1005944.
- Mistry J, et al. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res*. 49:D412–D419.
- Moeskjær S, et al. 2022. Major effect loci for plant size before onset of nitrogen fixation allow accurate prediction of yield in white clover. *Theor Appl Genet*. 135:125–143.
- Mölder F, et al. 2021. Sustainable data analysis with snakemake. *F1000Res* 10:33.
- Nagy I, Barth S, Mehenni-Ciz J, Abberton MT, Milbourne D. 2013. A hybrid next generation transcript sequencing-based approach to identify allelic and homeolog-specific single nucleotide polymorphisms in allotetraploid white clover. *BMC Genomics*. 14:100.
- Olsen KM, et al. 2021. Dual-species origin of an adaptive chemical defense polymorphism. *New Phytol*. 232:1477–1487.
- Olsen KM, Hsu S-C, Small LL. 2008. Evidence on the molecular basis of the ac/ac adaptive cyanogenesis polymorphism in white clover (*Trifolium repens* L.). *Genetics* 179:517–526.
- Olsen KM, Small LL. 2018. Micro- and macroevolutionary adaptation through repeated loss of a complete metabolic pathway. *New Phytol*. 219:757–766.
- Olsen KM, Sutherland BL, Small LL. 2007. Molecular evolution of the li/li chemical defence polymorphism in white clover (*Trifolium repens* L.). *Mol Ecol*. 16:4180–4193.
- Palmer J, Stajich J. 2019. *nextgenusfs/funcannotate* (Version 1.8.14). <https://zenodo.org/record/2604804>.
- Panter S, et al. 2012. Molecular breeding of transgenic white clover (*Trifolium repens* L.) with field resistance to *Alfalfa* mosaic virus through the expression of its coat protein gene. *Transgenic Res*. 21:619–632.
- Putnam NH, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res*. 26:342–350.
- Rawlings ND, et al. 2018. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 46:D624–D632.
- Santangelo JS, et al. 2022. Global urban environmental change drives adaptation in white clover. *Science* 375:1275–1281.
- Santangelo JS, Thompson KA, Johnson MTJ. 2018. Herbivores and plant defences affect selection on plant reproductive traits more strongly than pollinators. *J Evol Biol*. 32:4–18.
- Seppy M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. *Methods Mol Biol*. 1962:227–245.
- Shirasawa K, et al. 2023. An improved reference genome for *Trifolium subterraneum* L. Provides insight into molecular diversity and intra-specific phylogeny. *Front Plant Sci*. 14:1103857.
- Terlouw BR, et al. 2023. MIBig 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res*. 51:D603–D610.
- Thompson KA, Johnson MTJ. 2016. Antiherbivore defenses alter natural selection on plant reproductive traits. *Evolution* 70:796–810.
- UniProt Consortium. 2023. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res*. 51:D523–D531.
- Wang J, et al. 2010. Identification of genetic factors influencing salt stress tolerance in white clover (*Trifolium repens* L.) by QTL analysis. *Theor Appl Genet*. 120:607–619.
- Wang H, et al. 2023. High-quality chromosome-level de novo assembly of the *Trifolium repens*. *BMC Genomics*. 24:326.
- Williams WM, Ellison NW, Ansari HA, Verry IM, Hussain SW. 2012. Experimental evidence for the ancestry of allotetraploid *Trifolium repens* and creation of synthetic forms with value for plant breeding. *BMC Plant Biol*. 12:55.
- Yin Y, et al. 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 40:W445–W451.
- Zhang X, et al. 2022. Time-course RNA-Seq analysis provides an improved understanding of genetic regulation in response to cold stress from white clover (*Trifolium repens* L.). *Biotechnol Biotechnol Equip*. 36:1–8.
- Zhou L, Lu Q-W, Yang B-F, Zagorchev L, Li J-M. 2021. Integrated small RNA, mRNA, and degradome sequencing reveals the important role of miRNAs in the interactions between parasitic plant *Cuscuta australis* and its host *Trifolium repens*. *Sci Hortic*. 289:110458.
- Zimin AV, Salzberg SL. 2022. The SAMBA tool uses long reads to improve the contiguity of genome assemblies. *PLoS Comput Biol*. 18:e1009860.

Associate editor: Vincent Castric