Chromosome-scale assembly of the highly heterozygous genome of red clover (*Trifolium pratense* L.), an allogamous forage crop species

Derek M. Bickhart[1]*, Lisa M. Koch[1], Timothy P.L. Smith[2], Heathcliffe Riday[1], Michael L. Sullivan[1]*

[1]US Dairy Forage Research Center, USDA-ARS, Madison, WI

[2]US Meat Animal Research Center, USDA-ARS, Clay Center, NE

*Correspondence

Michael L. Sullivan, michael.sullivan@usda.gov, ORCID 0000-0002-8517-4493

Derek M. Bickhart, derek.bickhart@usda.gov, ORCID 0000-0003-2223-9285

Lisa M. Koch, lisa.koch@usda.gov, ORCID 0000-0002-4297-9531

Timothy P.L. Smith, tim.smitj2@usda.gov, ORCID 0000-0003-1611-6828

Heathcliffe Riday, heathcliffe.riday@usda.gov, ORCID 0000-0002-8322-6691

Abstract

Red clover (*Trifolium pratense* L.) is used as a forage crop due to a variety of favorable traits relative to other crops. Improved varieties have been developed through conventional breeding approaches, but progress could be accelerated and gene discovery facilitated using modern genomic methods. Existing short-read based genome assemblies of the ~420 Megabase (Mb) genome are fragmented into >135,000 contigs with numerous errors in order and orientation within scaffolds, likely due to the biology of the plant which displays gametophytic self-incompatibility resulting in inherent high heterozygosity. A high-quality long-read based assembly of red clover is presented that reduces the number of contigs by more than 500-fold, improves the per-base quality, and increases the contig N50 statistic by three orders of magnitude. The 413.5 Mb assembly is nearly 20% longer than the 350 Mb short read assembly, closer to the predicted genome size. Quality measures are presented and full-length isoform sequence of RNA transcripts reported for use in assessing accuracy and for future annotation of the genome. The assembly accurately represents the seven main linkage groups present in the genome of an allogamous (outcrossing), highly heterozygous plant species.

Research Area: Genetics and Genomics

Classifications: Bioinformatics, Plant Genetics

Context

The species *Trifolium pratense* L. (red clover) is an important legume forage crop grown on approximately 4 million hectares worldwide [1]. Red clover is an extremely versatile crop grown as an animal feed and/or as a green manure in pure and mixed stands for hay, haylage, silage, and grazing. Red clover is known for its ease of establishment and shade tolerance, as well as its ability to grow in poorly drained and low pH soils. The reduced need for exogenous nitrogen application due to its ability to fix nitrogen and the relatively high protein content of this plant compared to other forage crops provide potential for reducing the environmental footprint of livestock production. Compared to alfalfa, another common legume forage crop, red clover varieties have higher forage yields, provide a better source of magnesium to avoid grass tetany in grazing cattle, and may have improved post-harvest protein preservation [2] and bypass protein content in ruminant production systems [3]. The improved protein storage and utilization of this forage appears to be due to the post-harvest oxidation of *o*-diphenolic compounds by an endogenous polyphenol oxidase [4], although condensed tannins could also play a role [5]. Red clover tissues accumulate polyphenol oxidizable phenolics (mainly caffeic acid derivatives), condensed tannins, and a variety of specialized metabolites including flavonoid compounds [6, 7]. Such compounds have the potential to influence animal and rumen physiology in both negative [8] and positive ways [9]. Specialized metabolites from red clover have potential medicinal or nutraceutical value as well (see for example [10]). Improved varieties of red clover have been developed, especially with respect to persistence, disease resistance, and yield, but further improvements could be made in these and other traits affecting quality and nutritional value [1]. Genetic progress and greater understanding of the physiology and biochemistry of agronomic and quality traits could be accelerated using genomic tools based on the production of a high-quality reference genome for the species. Such a genome would also facilitate gene discovery efforts.

Red clover is a hermaphroditic allogamous (outcrossing) diploid ($2n = 2x = 14$) with a homomorphic gametophytic self-incompatibility (GSI) system [11] whereby a pistil expressed S-RNase mediates the degradation of pollen tubes from "self" pollen [12]. The GSI locus has been mapped to linkage group one in red clover. The GSI system in red clover appears to be especially effective [13], making red clover an obligate out-crossing species with a high degree of heterozygosity. This high degree of heterozygosity has made genome assembly with short read sequencing data difficult. Two previous short read genome assemblies [14, 15] have been reported with limited contiguity (>135,000 contigs), completeness, and accuracy. We report a long-read based assembly consisting of 258 contigs that provides a much improved reference genome to enhance genome-enabled red clover improvement.

Methods

Sample information

The individual used for sequencing in this study is HEN17-A07, a red clover plant selected out of the U.S. Dairy Forage Research Center (Madison, WI, USA) breeding program representing elite North American red clover germplasm. This individual was derived from 30 years of selection and breeding for red clover grazing tolerance, persistence, biomass yield, and *Fusarium oxysporum* Schlect resistance [16, 17]. Source varieties and germplasm for HEN17-A07 include: red clover varieties 'Dominion' [18] and 'Redlangraze' (ABI Alfalfa Inc., now part of Land O'Lakes, Inc. Arden, MN, USA); and experimental populations C452, C11, and C827 out of the U.S. Dairy Forage Research Center red clover breeding program. Plant material used for all nucleic acid isolations was clonally propagated from the original selected plant and maintained in a growth chamber at 22°C with 18 h days and light intensities of approximately 400 μmol m^{-2} s^{-1}.

DNA and RNA extraction and sequencing

Approximately 0.8 g of frozen unexpanded leaf tissue from the red clover individual Hen17-A07 (hereafter referred to as "red clover") was ground in a mortar and pestle under liquid nitrogen. High molecular weight DNA was extracted using the NucleoBond HMW DNA extraction kit as directed by the manufacturer (Macherey Nagel, Allentown, PA, USA). The DNA pellet was resuspended in 150 µL of 5mM Tris-Cl pH 8.5 (kit buffer HE) by standing at 4ºC overnight, with integrity estimated by fluorescence measurement (Qubit, Qiagen, Germantown, MD, USA), optical absorption spectra (DS-11, DeNovix), and size profile (Fragment Analyzer, Thermo Fisher, Waltham, MA, USA).

The Ligation Sequencing Kit (SQK-LSK109) was used to prepare libraries for nanopore sequencing from the extracted DNA as directed by the manufacturer (Oxford Nanopore Technologies, Oxford, UK). The libraries were sequenced in 14 R9.4 MinION flowcells on a GridION x5 instrument. The Guppy version 3.3 basecaller was used to call sequence bases producing 60 gigabase pairs (Gbp) of nanopore sequence in 4.5 million pass_filter reads having average read length of 13.6 kilobase (kb).

The DNA for HiFi sequencing was sheared (Hydroshear, Diagenode, Denville, NJ, USA) using a speed code setting of 13 to achieve a size distribution with peak at approximately 23 kb. Smaller fragments were removed by size selection for >12 kb fragments (BluePippin, Sage Science, Beverly, MA, USA). Size-selected DNA was used to prepare a SMRTbell library using the SMRTbell Express Template Prep Kit 2.0 as recommended by the manufacturer (Pacific Biosciences, Menlo Park, CA, USA). The library was sequenced in two SMRT Cell 8M cells on a Sequel II instrument using Sequel Sequencing Kit 3.0, producing 23.2 Gbp of HiFi sequence in 1.22 million CCS reads having average length 18.9 kb.

Approximately 200ug of DNA was fragmented to approximately 550bp on a Covaris M220 (Covaris, Woburn, MA, USA) by the University of Wisconsin-Madison Biotechnology Center (Madison, WI, USA) for short read sequencing as specified in the TruSeq DNA PCR-Free Reference Guide (Oct 2017, Illumina, San Diego, CA, USA). A library was prepared using a TruSeq DNA PCR-Free library preparation kit according to manufacturer guidance and was sequenced on a NextSeq500 instrument

(Illumina) with a NextSeq High Output v2 300 cycle kit, generating 198 million 2x150 paired end reads. This resulted in 30.0 Gbp of short read data which was used for error-correction and assembly validation.

The Omni-C library was prepared from unexpanded leaf tissue collected from plants grown in the dark for three days, and ground in liquid nitrogen with mortar and pestle. The pulverized material was processed into a proximity ligation library using the Omni-C Proximity Ligation Assay Protocol of the Omni-C Kit as directed by the manufacturer (Dovetail Genomics, Scotts Valley, CA, USA). The library was sequenced on a NextSeq500 instrument (Illumina) with 2x150 paired end reads, generating 60 million paired end Hi-C reads.

RNA was prepared for Iso-Seq using the Sigma Spectrum Plant Total RNA Kit including On-Column DNAse I Digestion (both products Sigma-Aldrich, St. Louis, MO, USA). One Hen17-A07 plant was sectioned into three parts (roots, leaves/crown, stem/flower) which were ground separately in liquid nitrogen in a mortar and pestle. RNA was prepared from 100 mg of each of the three tissues and pooled in equal proportions to avoid overrepresentation of one portion of the plant in the Iso-Seq reads. The pooled RNA was processed into an Iso-Seq library using the “Iso-Seq Express Template Preparation for Sequel and Sequel II Systems” protocol from the manufacturer (Pacific Biosciences) using the “standard” workflow of the protocol which includes a selection for polyadenylated transcripts. The library was sequenced in four SMRT cells on a Sequel II instrument, producing a total of 49 million sub reads with an average length of 2.9 kilobase pairs (kbp).

Genome assembly and scaffolding

HiFi reads (23.2 Gbp total; approximately 55-60x predicted coverage) were assembled using the PacBio IPA HiFi assembler (https://github.com/PacificBiosciences/pbipa) version 1.3.0 using default settings. This resulted in a primary haplotype assembly of 419.1 megabase pairs (Mbp) in 283 contigs, with a contig N50 of 4.3 Mbp, and an alternate haplotype assembly of 353.6 Mbp in 1,555 contigs. The

relatively large size of the alternate haplotype assembly likely reflects the obligate heterozygosity of red clover, since high heterozygosity supports more complete separation of parental haplotypes during HiFi-based assembly. The primary haplotype assembly was retained for use in downstream polishing and assembly quality assessment. Residual haplotype sequence was removed from the assembled contigs using purge_dups v1.2.5 [19]. Depth of coverage cutoff values for the purge_dups workflow were estimated from minimap2 [20] alignments of HiFi reads to the contigs. A total of 5.6 Mbp (1.4% of the original bases) in 34 contigs were identified as remnant haplotypes in the primary contig assembly and removed. Of the 34 contigs, 25 were entirely composed of remnant haplotype sequence and were completely removed from the purged assembly. The final set of purged contigs (hereafter referred to as "HiFi Contigs") had an identical contig N50 (4.3 Mbp) to the first primary IPA assembly because of the small size of the contigs that were removed, but had 258 contigs and a reduction in size of 5.6 Mbp.

Scaffolds were created from the HiFi Contigs using the SALSA v2 scaffolding workflow [21]. Omni-C reads were aligned to the purged contig assembly using BWA MEM [22] with the '-SP5' flag to disable paired-end read recovery. Resulting BAM files were converted to a bed file format using the Bedtools2 [23] tool, "bamToBed." SALSA was subsequently run without misassembly detection to avoid unnecessary contig breaks and the "DNASE" setting due to the use of OmniC reads for scaffolding. This placed the 258 contigs into 143 scaffolds with a scaffold N50 of 15.6 Mbp (Table 1). This intermediary dataset is referred to as the "Omni-C scaffolds" for convenience. The contiguity as summarized by the contig and scaffold N50 values compared favorably with legume assemblies that had the benefit of extensive polishing, such as the *Medicago truncatula* reference, MedTr 4.0 [24].

Scaffold placement using linkage data

Previously published expressed sequence tag (EST) [25], bacterial artificial chromosome (BAC) end [14], and Oxford Nanopore read overlaps were used to generate super-scaffolds representing the best approximation of red clover linkage group chromosomes. EST and BAC reads were converted to fasta

format and aligned against Hi-C scaffolds using BWA MEM. A custom script (https://github.com/njdbickhart/perl_toolchain/blob/master/assembly_scripts/alignAndOrderSnpProbes.pl) was used to order and orient EST and BAC information into a tabular, bipartite graph format for comparison. Oxford Nanopore reads were aligned to the Omni-C scaffolds with minimap2 [20] and overhanging reads were identified using custom perl scripts (https://github.com/njdbickhart/RumenLongReadASM/blob/master/viralAnalysisScripts/filterOverhangAlignments.pl). Overlapping reads from two different contigs were combined into bipartite graphs as evidence of connection.

The BAC, EST, and Oxford Nanopore datasets were analyzed using the Python NetworkX (https://networkx.org/) module to determine concordance among all three for final scaffold formation. The Oxford Nanopore read overlaps showed substantial overlap with the underlying EST dataset, but the BAC end sequence did not display any substantial overlap with the other datasets. The final linkage group super-scaffolds were generated by assigning Omni-C scaffolds to linkage groups and ordering them according to their placement in the EST alignment dataset. Scaffolds that did not have EST mappings but were identified via Nanopore overlaps (four scaffolds in total) were incorporated into the final super-scaffolds on the side of the scaffold indicated by overlapping read data. The final set of super-scaffolds were generated using the ‘agp2fasta’ utility of the “CombineFasta” Java tool (https://github.com/njdbickhart/CombineFasta). The final set of super-scaffolds is referred to as “ARS_RCv1.1” in the text.

Iso-Seq transcript identification

Iso-Seq sequence data was processed for isoform identification using the Iso-Seq Analysis pipeline in smrtlink v9.0.0.92188 including the option to map putative isoforms to the assembly scaffolds imported as a reference genome. A total of 9.2 million HiFi reads were generated from the 49 million sub-reads, of which 8,899,606 (97%) were classified as Full-Length Non-Concatemer reads (FLNC) with a mean

length of 3.2 kbp. These FLNC reads collapsed to 437,586 predicted unique polished high-quality isoforms, of which 308,804 (70%) mapped to 24,955 unique gene loci in the assembly, consistent with approximately 12 isoforms per unique loci. These gene loci are provided as BED coordinate files for future annotation efforts.

Data validation and quality control

Assembly error-rate assessment

Genome quality was tested using a composite of k-mer and read mapping quality statistics as implemented in the Themis-ASM workflow [26]. All references to short-read WGS data refer to the short-reads generated from the HEN17-A07 individual sequenced and assembled in this study unless otherwise mentioned. The completeness and quality of the assembly was first assessed using Merqury [27] k-mer analysis and freebayes [28] variant analysis. Merqury estimated the overall quality of the assembly at a Phred-based [29] quality value (QV) score of 49 which corresponds to an error every 129,000 bases (Table 2). Comparison of k-mer profiles between the HiFi contigs and the previously published TGACv2 red clover assembly [14] (accession GCA_900292005.1) using the upset python module (https://github.com/ImSoErgodic/py-upset) (Figure 1) indicated that only 55.2% of all k-mers were shared between the two assemblies. This surprisingly low shared content could be the result of real differences in the genomes of the different varieties of this highly heterozygous species (the earlier assembly used an individual from the Milvus variety versus the Hen17-A07 individual used here), or the higher degree of completeness of the current assembly (the previous assembly was comprised of 135,023 contigs and was 68 Mb smaller total size), or assembly and haplotype switching errors in the short read assembly, or a combination of these factors. The Themis-ASM analysis of TGACv2 estimated an error every 142 bases, indicating that the ARS-RCv1.1 assembly has a three orders of magnitude improvement in k-mer based QV estimates. Indeed, the count of erroneous, singleton k-mers identified in the TGACv2 assembly was over 40 million, compared to less than 10,000 in the ARS_RCv1.1 assembly (Figure 3).

198 This represents a substantial improvement in assembly accuracy enabled by the use of improved
199 sequencing technologies.

200 Freebayes QV values were similar to those generated via Merqury analysis, but with a six point decrease
201 in relative QV between the two assemblies. This QV estimate was originally developed to compare the
202 qualities of uniquely mappable regions of assemblies [30], so it is more robust when comparing datasets
203 derived from different breeds or varieties to separate assemblies. The appreciable difference in Freebayes
204 QV between the two assemblies still points towards a higher error rate in the TGACv2 reference, and
205 suggests that the ARS_RCv1.1 assembly is more suitable as a reference for short-read WGS alignment in
206 the red clover species.

207 The MedTr4 assembly represents a high quality reference for most legume species, and has been used in
208 several whole genome comparisons to indicate assembly quality [31, 32]. This includes the original
209 release of the TGACv2 reference, where synteny was identified between MedTr4 and the TGACv2
210 assembly [14]. However, Merqury-estimated error rate of one out of every ten bases when mapping red
211 clover WGS reads suggests that MedTr4 is unsuitable as a reference for red clover WGS alignment. This
212 conclusion is supported by the observation that over 60% of the HEN17-A07 individual WGS reads were
213 unmapped when aligned to the MedTr4 reference. This suggests that more distantly related legume
214 species require a high quality reference genome assembly for satisfactory alignment quality metrics. The
215 approach described here provides a method to develop these reference assemblies for highly heterozygous
216 allogamous species, such as red clover, without the requirement for extensive post-hoc polishing.

217 **Structural variant assessment and comparative alignments**

218 The structural accuracy of the super-scaffolds was assessed using a variety of comparative metrics native
219 to the Themis-ASM workflow [26]. The short-read WGS data alignments were used as a basis for
220 FRC_align [33] quality metrics which identified a relatively low number of regions with predicted inter-

scaffold alignments in ARS_RCv1.1 (Table 3). This was matched by a relatively low count of complex structural variants (SV) in ARS_RCv1.1 compared to TGACv2 as identified by Lumpy [34] analysis, suggesting that small-scale misassemblies that are detectable using short-read alignments were minimized in the ARS_RCv1.1 assembly.

Comparisons of the large scale synteny of our assembly to the TGACv2 reference revealed a substantial number of discrepancies. Alignment of the scaffolds from the TGACv2 reference to the ARS_RCv1.1 assembly was performed with minimap2 [20] using the "-x asm10" preset. A circos plot (http://circos.ca/) derived from these alignments revealed numerous differences in sequence attribution to linkage group super-scaffolds (Figure 4a). Furthermore, these whole-scaffold alignments revealed several structural variants that represented potential expansions of the TGACv2 reference compared to ARS_RCv1.1 (Figure 4b). The accuracy of ARS_RCv1.1 super-scaffold placement on a macro-scale was examined by alignment of previously generated BAC end sequence data from the Milvus B individual [14] to both assemblies with minimap2 using the "-x sr" preset. Resulting PAF files were analyzed with custom scripts to identify three distinct categories of BAC paired-end alignments: 1) if both pairs aligned to the same scaffold, 2) if both pairs aligned to different scaffolds or 3) if both pairs were unmapped (Table 3). The same 483 BAC paired-ends were unmapped to both assemblies, suggesting contamination in the creation of the original BAC library. However, the ARS_RCv1.1 assembly had two-fold more BAC paired-ends that aligned to the same super-scaffold than the TGACv2 reference. This gives greater confidence to the linkage-group assignment on the ARS_RCv1.1 assembly, and suggests that observed structural expansions of the TGACv2 reference are due to misassemblies (Table 2) or other smaller errors (Figure 3).

Re-use potential and conclusions

We report the creation of a new reference assembly for red clover using a combination of HiFi and nanopore-based long read sequencing, with Omni-C and BAC-end sequence scaffolding to produce

chromosome-scale superscaffolds. The quality of the assembly demonstrates that low-error rate long reads are suitable for resolving issues in assembling allogamous heterozygous (> 50%) diploid plant genomes and generating continuous scaffolds. The addition of Omni-C read linkage data supported generation of an assembly with only 143 scaffolds. These scaffolds were then combined into seven final linkage-group super-scaffolds, which better reflected the haploid structure of red clover chromosomes. Compared to a previous reference for the species, ARS_RCv1.1 contains 20% more assembled sequence and has an error rate that is lower by three orders of magnitude. Comparative mapping statistics to other legume genome assemblies suggest that this assembly will enable better alignment of red clover short-read WGS data, will improve the prediction of gene models, and will facilitate transcriptomic studies and gene discover efforts based on both marker-phenotype association and sequence identity. Previous assemblies of red clover were limited by the error-rates or length of reads used to construct them. We demonstrate that recent improvements in DNA sequencing technologies are finally capable of generating a suitable assembly for this highly heterozygous species, and that these methods can be applied to other similar species without the need for expert curation.

Availability of source code and requirements

Project name: Themis-ASM.

Project Home page: https://github.com/njdbickhart/Themis-ASM.

Operating systems: Unix, Linux.

Programming language: Snakemake v3.4+, Python 3.6+, Perl 5.10+

Other requirements: miniconda v3.6+ or Anaconda 3+

License: GNU GPL

Data availability

All sequence data used in the assembly, scaffolding and analysis of ARS_RCv1.1 can be found on the NCBI's SRA under Bioproject accession number PRJNA754186. Genome Accession for the ARS_RCv1.1 assembly is GCA_020283565.1. IsoSeq reads can be found under the NCBI's SRA run accession number SRR15433788. IsoSeq transcripts will be provided via GigaDB accession after peer-review.

List of abbreviations

BAC, bacterial artificial chromosome; EST, expressed sequence tag; FLNC, Full-Length Non-Concatemer; GB, gigabase; Gbp, gigabase pairs; GSI, gametophytic self-incompatibility; kb, kilobase; kbp, kilobase pairs; MB, megabase; Mbp, megabase pairs; QV, quality value; SV, structural varient

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by USDA-ARS Projects 5090-31000-026-00D (DMB), 5090-21000-071-00D (MLS), 5090-21000-001-00D (HR), 3040-31000-100-00D (TPLS).

Permissions

To our knowledge, there are no local, national or international guidelines or legislation governing the study presented in this manuscript and no permissions and/or license required for the study.

Author's contributions

LMK, TPLS, and MLS were responsible for genome WGS, Omni-C, and transcriptome sequencing data generation. DMB and TPLS assembled the genome and DMB ran scaffolding analysis. DMB and LMK ran the analysis of the assembly. All authors read and contributed to the manuscript.

Acknowledgements

We thank Dr. Kristen Kuhn, Kelsey McClure, and Dr. Jennifer McClure for technical assistance.

The USDA does not endorse any products or services. Mentioning of trade names is for information purposes only. The USDA is an equal opportunity employer.

References

1. Riday H. Progress made in improving red clover (*Trifolium pratense* L.) through breeding. Int J Plant Breed. 2010;4:22-9.

2. Albrecht KA and Muck RE. Proteolysis in ensiled forge legumes that vary in tannin concentration. Crop Sci. 1991;31:464-9.

3. Broderick GA. Utilization of protein in red clover and alfalfa silages by lactating dairy cows and growing lambs. J Dairy Sci. 2018;101:1190-205. doi:10.3168/jds.2017-13690.

4. Sullivan ML and Hatfield RD. Polyphenol oxidase and *o*-diphenols inhibit postharvest proteolysis in red clover and alfalfa. Crop Sci. 2006;46:662-70.

5. Berard NC, Wang Y, Wittenberg KM, Krause DO, Coulman BE, McAllister TA, et al. Condensed tannin concentrations found in vegetative and mature forage legumes grown in western Canada. Can J Plant Sci. 2011;91:669-75. doi:10.4141/cjps10153.

6. Saviranta NMM, Julkunen-Tiitto R, Oksanen E and Karjalainen RO. Leaf phenolic compounds in red clovder (*Trifolium pratens*e L.) indcued by exposure to moderately elevated ozone. Environ Polut. 2010;158:440-6.

7. Oleszek W, Stochmal A and Janda B. Concentration of isoflavones and other phenolics in the aerial parts of *Trifolium* species. J Agr Food Chem. 2007;55:8095-100. doi: 10.1021/Jf072024w.

8. Kramer R, Keogh RG and McDonald MF. The accumulation and clearance of equol in the blood of ewes grazed on either high or low formononetin red clovers. Proc N Z Soc Anim Prod. 1996;56:373-7.

9. Harlow BE, Flythe MD, Kagan IA, Goodman JP, Klotz JL and Aiken GE. Isoflavone supplementation, via red clover hay, alters the rumen microbial community and promotes weight gain of steers grazing mixed grass pastures. PLoS One. 2020;15 3:e0229200. doi:10.1371/journal.pone.0229200.

10. Kolodziejczyk-Czepas J, Krzyzanowska-Kowalczyk J, Sieradzka M, Nowak P and Stochmal A. Clovamide and clovamide-rich extracts of three *Trifolium* species as antioxidants and moderate antiplatelet agents in vitro. Phytochemistry. 2017;143:54-63. doi:10.1016/j.phytochem.2017.07.011.

11. Townsend CE and Taylor NL. Incompatibility and plant breeding. In: Taylor NL, editor. Clover Science and Technology. Madison, WI, USA: ASA-CSSA-SSSA; 1985. p. 365-81.

12. Riday H and Krohn AL. Genetic map-based location of the red clover (*Trifolium pratense* L.) gametophytic self-incompatibility locus. Theor Appl Genet. 2010;121:761-7. doi:10.1007/s00122-010-1347-0.

13. Williams RD and Silow RA. Genetics of red clover (*Trifolium pratense* L.): Compatibility I. J Genet. 1933;27:341-62.

14. De Vega JJ, Ayling S, Hegarty M, Kudrna D, Goicoechea JL, Ergon Å, et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Sci Rep. 2015;5:17394. doi:10.1038/srep17394.

15. Istvanek J, Jaros M, Krenek A and Repkova J. Genome assembly and annotation for red clover (*Trifolium pratense*; *Fabaceae*). Am J Bot. 2014;101:327-37. doi:10.3732/ajb.1300340.

16. Riday H, Casler MD, Crooks A and Wood T. Persistence of grazed red clover varieties. Grass Clippings (University of Wisconsin Extension, Center for Integrated Agricultural Systems, and College of Agriculture and Live Sciences). 2007;2:3-8.

17. Venuto BC. Reaction of red clover (*Trifolium pratense* L.) to *Fusarium oxysporum* Schlect. and selection for and inheritance of resistance in red clover. (Thesis) University of Wisconsin-Madison, Madison, WI, USA, 1993.

18. Association of Official Seed Certifying Agencies. Dominion red clover. 2005. Moline, IL, USA: Association of Official Seed Certifying Agencies.

19. Guan D, McCarthy SA, Wood J, Howe K, Wang Y and Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896-8. doi:10.1093/bioinformatics/btaa025.

20. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. 2016;32:2103-10. doi:10.1093/bioinformatics/btw152.

21. Ghurye J, Pop M, Koren S, Bickhart D and Chin C-S. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 2017;18 1:527. doi:10.1186/s12864-017-3879-z.

22. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013.

23. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-2. doi:10.1093/bioinformatics/btq033.

24. Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics. 2014;15:312. doi:10.1186/1471-2164-15-312.

25. Isobe S, Kölliker R, Hisano H, Sasamoto S, Wada T, Klimenko I, et al. Construction of a consensus linkage map for red clover (*Trifolium pratense* L.). BMC Plant Biol. 2009;9:57. doi:10.1186/1471-2229-9-57.

26. Heaton MP, Smith TPL, Bickhart DM, Vander Ley BL, Kuehn LA, Oppenheimer J, et al. A Reference Genome Assembly of Simmental Cattle, *Bos taurus taurus*. J Hered. 2021;112:184-91. doi:10.1093/jhered/esab002.

27. Rhie A, Walenz BP, Koren S and Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245. doi:10.1186/s13059-020-02134-9.

28. Garrison E and Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:12073907 [q-bio]. 2012.

29. Ewing B and Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8:186-94.

30. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643-50. doi:10.1038/ng.3802.

31. Ha J, Satyawan D, Jeong H, Lee E, Cho K-H, Kim MY, et al. A near-complete genome sequence of mungbean (*Vigna radiata* L.) provides key insights into the modern breeding program. Plant Genome. 2021:e20121. doi:10.1002/tpg2.20121.

32. Moghaddam SM, Oladzad A, Koh C, Ramsay L, Hart JP, Mamidi S, et al. The tepary bean genome provides insight into evolution and domestication under heat stress. Nat Commun. 2021;12:2638. doi:10.1038/s41467-021-22858-x.

33. Vezzi F, Narzisi G and Mishra B. Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. PLoS One. 2012;7 12:e52210. doi:10.1371/journal.pone.0052210.

34. Layer RM, Chiang C, Quinlan AR and Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15 6:R84. doi:10.1186/gb-2014-15-6-r84.

Figure Captions

Figure 1: Comparison of unique k-mer counts among the TGACv2 assembly and our HiFi Contigs. Unique k-mers were counted using meryl and compared between both assemblies using exact match comparisons. The top histogram shows the proportion of all unique k-mers shared among each set, with set membership shown in the bottom right dot plot. The leftmost histogram shows the total count of unique k-mers distinct to each assembly, with percentages indicating the amount of k-mers from the combined total dataset.

Figure 2: Comparative assembly statistics. (A) The total percentages of Eudicot lineage single-copy orthologous genes identified by the BUSCO tool are represented by stacked histograms for each assembly. Values larger than 10% are displayed on the histograms for convenience. (B) NG values against an estimated genome size of 420 MB are shown as solid lines on the plot. The NG50 value is distinguished by a vertical dashed bar for each assembly.

Figure 3: Merqury stacked histogram charts of k-mer multiplicity between the ARS_RCv1.1 (A) assembly and the TGACv2 (B) reference. In each case, the k-mers derived from the assembly are colored light red, and the k-mers unique to the short-read WGS data (from the HEN17-07A individual of *T.*

pretense) are dark grey. The farthest left red bar indicates the total number of singleton k-mers for each assembly, which are considered indicators of misassemblies or errors. The bimodal distribution of each plot indicates the heterozygous (left-most) and homozygous (right-most) k-mer values. The prevalence of any area under the "read-only" plot indicates that the assembly does not contain k-mers present in the short-read WGS data.

Figure 4: Structural variation comparison between the TGACv2 and ARS_RCv1.1 reference assemblies. (A) A circos plot constructed from whole-genome alignments of TGACv2 (labelled TGACv2_LG1-7) to ARS-RCv1.1 (labelled LG1-7) is color coded based on originating ARS_RCv1.1 linkage-group information. Only alignment blocks larger than 10 kbp in length are displayed on the plot as ribbons that connect between each assembly. Presence of more than one colored alignment ribbon link to the TGACv2 scaffolds indicates a discrepancy between the two assemblies. (B) Whole-genome alignments also revealed additional structural variant discrepancies between the two assemblies. Given the relative nature of duplications and deletions detected on comparative alignments, arrows that indicate potential expansion of sequence in one assembly compared to another are indicated at the bottom of the plot. For example, tandem contractions of sequence in ARS_RCv1.1 could be considered expansions of genome sequence in TGACv2, and vice versa.

Table 1: Assembly Size Statistics

Statistic	TGACv2	HiFi Contigs	Omni-C Scaffolds	MedTr4
Assembly Length (Mbp)	346.0	413.5	413.5	412.8
Contig / Scaffold count	39,051	258	143	2,186
Scaffold N50 (Mbp)	22.7	4.4	15.6	49.2
Largest Contig / Scaffold (Mbp)	32.6	13.4	34.2	56.6

Table 2: Assembly quality statistics

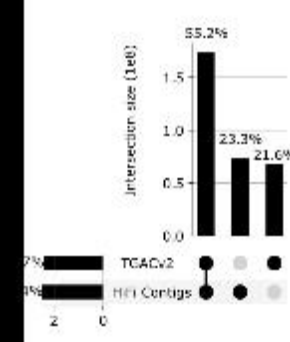
Category	TGACv2	ARS_RCv1.1	MedTr4	Description
Merqury QV	21.5304	48.9101	9.74458	kmer-based Quality
Merqury ErrorRate	0.007	1.29 x 10-5	0.106	kmer-based error rate
Merqury Completeness (%)	61.7428	77.7322	3.86382	Percentage of complete assembly based on kmers
Freebayes QV	20.03	41.71	12.22	SNP and INDEL Quality value
Unmapped reads (%)	3.65	2.37	60.92	Percentage of short-reads unmapped
COMPLETE Single copy (%)	87.5	87.6	92.9	Percent of complete, single-copy BUSCOs
COMPLETE Duplicated (%)	3.2	10.4	4.8	Percent of complete, duplicated BUSCOs
FRAGMENTED (%)	4.9	1.1	0.7	Percent of fragmented BUSCOs
MISSING (%)	4.4	0.9	1.6	Percent of missing BUSCOs

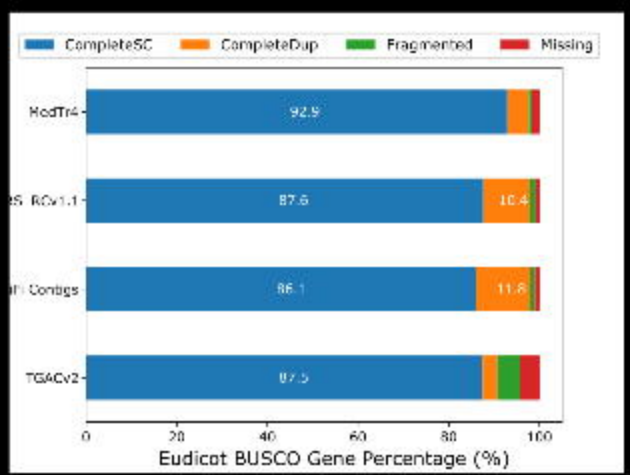
427 Table 3: Structural variant analysis

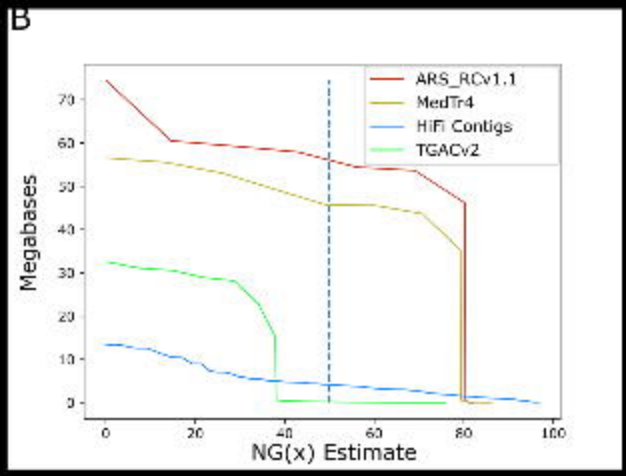
Category	TGACv2	ARS_RCv1.1	Description
HIGH_SPAN_PE	65,254	2,052	FRC_align identified regions with high numbers of inter-contig paired-end read mappings
Lumpy Deletions	20,727	20,945	Number of identified structural variant deletions
Lumpy Duplications	6,554	3,823	Number of identified structural variant duplications
Lumpy Complex	387,898	60,130	Number of complex (multiple tandem deletions or duplications) structural variants
BAC ends to same scaffold	7,357	15,795	BAC end pairs that were best mapped to the same scaffold
BAC ends to different scaffold	21,228	12,791	BAC end pairs with best alignments to different scaffolds
BAC ends unmapped	484	483	Unmapped BAC end pairs

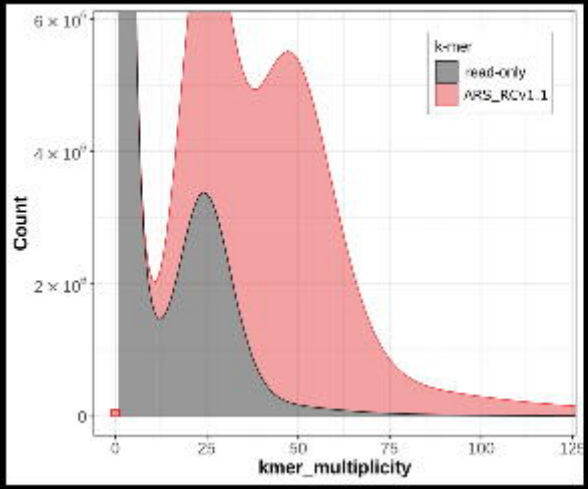
428

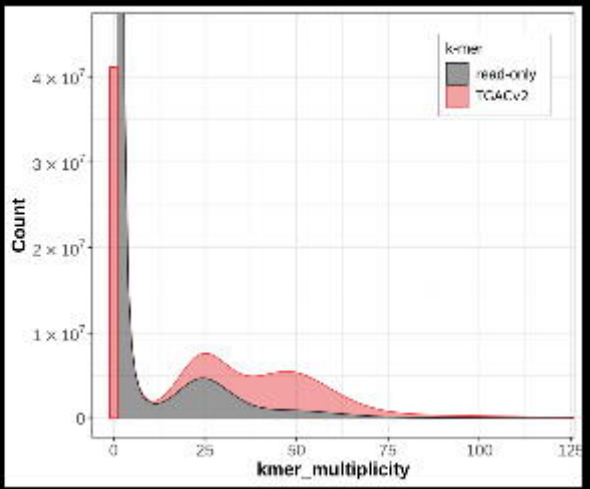
429

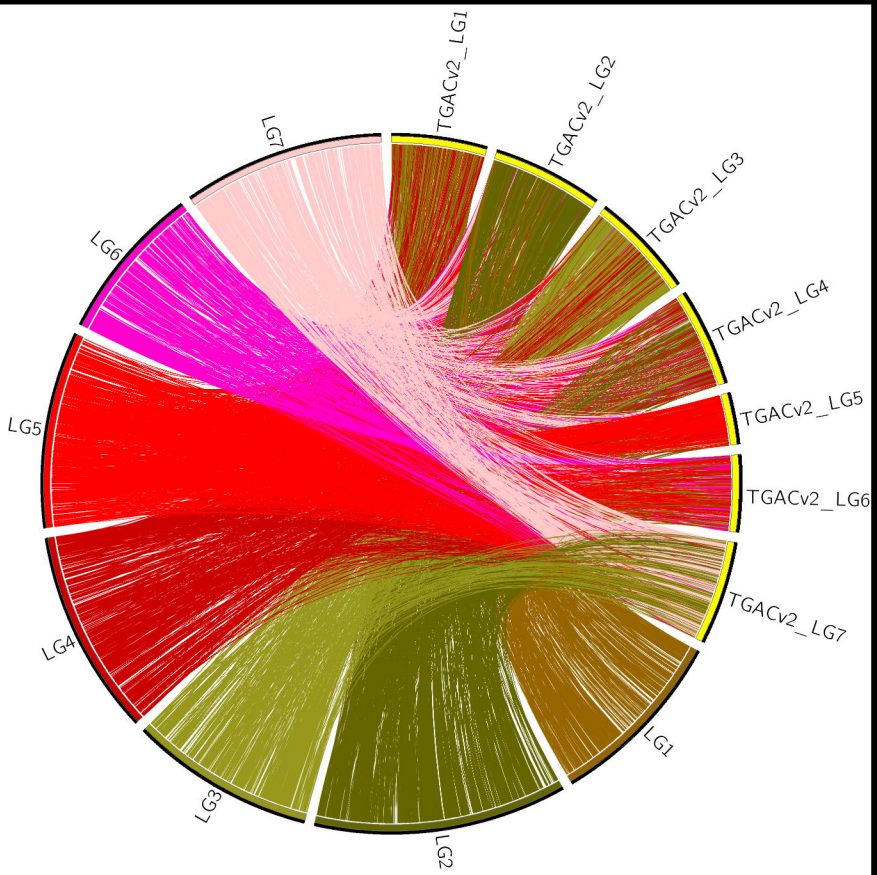

55.2%
23.3%
21.6%
Intersection size (1e6)
1.5
1.0
0.5
0.0
TCACv2
HiFi Contigs
2
0

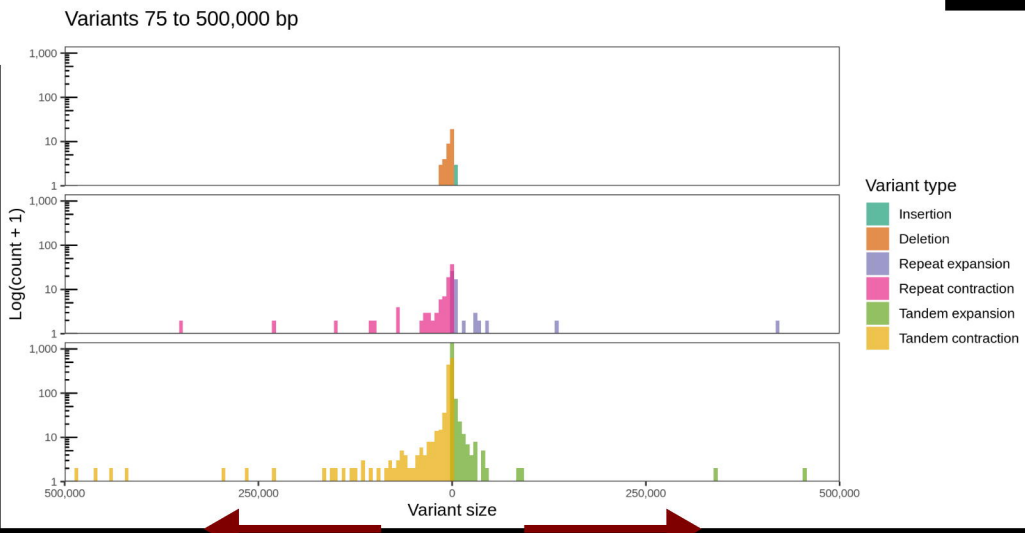

CompleteSC
CompleteDup
Fragmented
Missing
MedTr4
92.9
S_RCv1.1
87.6
10.4
Contigs
86.1
11.8
TGACv2
87.5
0
20
40
60
80
100
Eudicot BUSCO Gene Percentage (%)


B
ARS_RCv1.1
MedTr4
HiFi Contigs
TGACv2
Megabases
70
60
50
40
30
20
10
0
0
20
40
60
80
100
NG(x) Estimate


k-mer
read-only
ARS_RCv1.1
Count
kmer_multiplicity


k-mer
read-only
TGACv2
Count
kmer_multiplicity


TGACv2_LG1
TGACv2_LG2
TGACv2_LG3
TGACv2_LG4
TGACv2_LG5
TGACv2_LG6
TGACv2_LG7
LG1
LG2
LG3
LG4
LG5
LG6
LG7


Variants 75 to 500,000 bp
Log(count + 1)
1,000
100
10
1
500,000
250,000
0
250,000
500,000
Variant size
Variant type
Insertion
Deletion
Repeat expansion
Repeat contraction
Tandem expansion
Tandem contraction