

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

3

4 Derek M. Bickhart^{1*}, Lisa M. Koch¹, Timothy P.L. Smith², Heathcliffe Riday¹, Michael L. Sullivan^{1*}

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

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

7

8 *Correspondence

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

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

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

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

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

14 **Abstract**

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

28

29 Research Area: Genetics and Genomics

30 Classifications: Bioinformatics, Plant Genetics

31

32

33

34 **Context**

35 The species *Trifolium pratense* L. (red clover) is an important legume forage crop grown on
36 approximately 4 million hectares worldwide [1]. Red clover is an extremely versatile crop grown as an
37 animal feed and/or as a green manure in pure and mixed stands for hay, haylage, silage, and grazing. Red
38 clover is known for its ease of establishment and shade tolerance, as well as its ability to grow in poorly
39 drained and low pH soils. The reduced need for exogenous nitrogen application due to its ability to fix
40 nitrogen and the relatively high protein content of this plant compared to other forage crops provide
41 potential for reducing the environmental footprint of livestock production. Compared to alfalfa, another
42 common legume forage crop, red clover varieties have higher forage yields, provide a better source of
43 magnesium to avoid grass tetany in grazing cattle, and may have improved post-harvest protein
44 preservation [2] and bypass protein content in ruminant production systems [3]. The improved protein
45 storage and utilization of this forage appears to be due to the post-harvest oxidation of *o*-diphenolic
46 compounds by an endogenous polyphenol oxidase [4], although condensed tannins could also play a role
47 [5]. Red clover tissues accumulate polyphenol oxidizable phenolics (mainly caffeic acid derivatives),
48 condensed tannins, and a variety of specialized metabolites including flavonoid compounds [6, 7]. Such
49 compounds have the potential to influence animal and rumen physiology in both negative [8] and positive
50 ways [9]. Specialized metabolites from red clover have potential medicinal or nutraceutical value as well
51 (see for example [10]). Improved varieties of red clover have been developed, especially with respect to
52 persistence, disease resistance, and yield, but further improvements could be made in these and other
53 traits affecting quality and nutritional value [1]. Genetic progress and greater understanding of the
54 physiology and biochemistry of agronomic and quality traits could be accelerated using genomic tools
55 based on the production of a high-quality reference genome for the species. Such a genome would also
56 facilitate gene discovery efforts.

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

66 **Methods**

67 **Sample information**

68 The individual used for sequencing in this study is HEN17-A07, a red clover plant selected out of the
69 U.S. Dairy Forage Research Center (Madison, WI, USA) breeding program representing elite North
70 American red clover germplasm. This individual was derived from 30 years of selection and breeding for
71 red clover grazing tolerance, persistence, biomass yield, and *Fusarium oxysporum* Schlect resistance [16,
72 17]. Source varieties and germplasm for HEN17-A07 include: red clover varieties ‘Dominion’ [18] and
73 ‘Redlangraze’ (ABI Alfalfa Inc., now part of Land O’Lakes, Inc. Arden, MN, USA); and experimental
74 populations C452, C11, and C827 out of the U.S. Dairy Forage Research Center red clover breeding
75 program. Plant material used for all nucleic acid isolations was clonally propagated from the original
76 selected plant and maintained in a growth chamber at 22°C with 18 h days and light intensities of
77 approximately $400 \mu\text{mol m}^{-2} \text{s}^{-1}$.

78 **DNA and RNA extraction and sequencing**

79 Approximately 0.8 g of frozen unexpanded leaf tissue from the red clover individual Hen17-A07
80 (hereafter referred to as “red clover”) was ground in a mortar and pestle under liquid nitrogen. High
81 molecular weight DNA was extracted using the NucleoBond HMW DNA extraction kit as directed by the
82 manufacturer (Macherey Nagel, Allentown, PA, USA). The DNA pellet was resuspended in 150 μ L of
83 5mM Tris-Cl pH 8.5 (kit buffer HE) by standing at 4°C overnight, with integrity estimated by
84 fluorescence measurement (Qubit, Qiagen, Germantown, MD, USA), optical absorption spectra (DS-11,
85 DeNovix), and size profile (Fragment Analyzer, Thermo Fisher, Waltham, MA, USA).

86 The Ligation Sequencing Kit (SQK-LSK109) was used to prepare libraries for nanopore sequencing from
87 the extracted DNA as directed by the manufacturer (Oxford Nanopore Technologies, Oxford, UK). The
88 libraries were sequenced in 14 R9.4 MinION flowcells on a GridION x5 instrument. The Guppy version
89 3.3 basecaller was used to call sequence bases producing 60 gigabase pairs (Gbp) of nanopore sequence
90 in 4.5 million pass_filter reads having average read length of 13.6 kilobase (kb).

91 The DNA for HiFi sequencing was sheared (Hydroshear, Diagenode, Denville, NJ, USA) using a speed
92 code setting of 13 to achieve a size distribution with peak at approximately 23 kb. Smaller fragments were
93 removed by size selection for >12 kb fragments (BluePippin, Sage Science, Beverly, MA, USA). Size-
94 selected DNA was used to prepare a SMRTbell library using the SMRTbell Express Template Prep Kit
95 2.0 as recommended by the manufacturer (Pacific Biosciences, Menlo Park, CA, USA). The library was
96 sequenced in two SMRT Cell 8M cells on a Sequel II instrument using Sequel Sequencing Kit 3.0,
97 producing 23.2 Gbp of HiFi sequence in 1.22 million CCS reads having average length 18.9 kb.

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

103 (Illumina) with a NextSeq High Output v2 300 cycle kit, generating 198 million 2x150 paired end reads.

104 This resulted in 30.0 Gbp of short read data which was used for error-correction and assembly validation.

105 The Omni-C library was prepared from unexpanded leaf tissue collected from plants grown in the dark for

106 three days, and ground in liquid nitrogen with mortar and pestle. The pulverized material was processed

107 into a proximity ligation library using the Omni-C Proximity Ligation Assay Protocol of the Omni-C Kit

108 as directed by the manufacturer (Dovetail Genomics, Scotts Valley, CA, USA). The library was

109 sequenced on a NextSeq500 instrument (Illumina) with 2x150 paired end reads, generating 60 million

110 paired end Hi-C reads.

111 RNA was prepared for Iso-Seq using the Sigma Spectrum Plant Total RNA Kit including On-Column

112 DNase I Digestion (both products Sigma-Aldrich, St. Louis, MO, USA). One Hen17-A07 plant was

113 sectioned into three parts (roots, leaves/crown, stem/flower) which were ground separately in liquid

114 nitrogen in a mortar and pestle. RNA was prepared from 100 mg of each of the three tissues and pooled in

115 equal proportions to avoid overrepresentation of one portion of the plant in the Iso-Seq reads. The pooled

116 RNA was processed into an Iso-Seq library using the “Iso-Seq Express Template Preparation for Sequel

117 and Sequel II Systems” protocol from the manufacturer (Pacific Biosciences) using the “standard”

118 workflow of the protocol which includes a selection for polyadenylated transcripts. The library was

119 sequenced in four SMRT cells on a Sequel II instrument, producing a total of 49 million sub reads with an

120 average length of 2.9 kilobase pairs (kbp).

121 **Genome assembly and scaffolding**

122 HiFi reads (23.2 Gbp total; approximately 55-60x predicted coverage) were assembled using the PacBio

123 IPA HiFi assembler (<https://github.com/PacificBiosciences/pbipa>) version 1.3.0 using default settings.

124 This resulted in a primary haplotype assembly of 419.1 megabase pairs (Mbp) in 283 contigs, with a

125 contig N50 of 4.3 Mbp, and an alternate haplotype assembly of 353.6 Mbp in 1,555 contigs. The

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

137 Scaffolds were created from the HiFi Contigs using the SALSA v2 scaffolding workflow [21]. Omni-C
138 reads were aligned to the purged contig assembly using BWA MEM [22] with the ‘-SP5’ flag to disable
139 paired-end read recovery. Resulting BAM files were converted to a bed file format using the `Bedtools2`
140 [23] tool, “`bamToBed.`” SALSA was subsequently run without misassembly detection to avoid
141 unnecessary contig breaks and the “DNASE” setting due to the use of OmniC reads for scaffolding. This
142 placed the 258 contigs into 143 scaffolds with a scaffold N50 of 15.6 Mbp (Table 1). This intermediary
143 dataset is referred to as the “Omni-C scaffolds” for convenience. The contiguity as summarized by the
144 contig and scaffold N50 values compared favorably with legume assemblies that had the benefit of
145 extensive polishing, such as the *Medicago truncatula* reference, MedTr 4.0 [24].

146 **Scaffold placement using linkage data**

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

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

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

169 **Iso-Seq transcript identification**

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

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

178 **Data validation and quality control**

179 **Assembly error-rate assessment**

180 Genome quality was tested using a composite of k-mer and read mapping quality statistics as
181 implemented in the Themis-ASM workflow [26]. All references to short-read WGS data refer to the
182 short-reads generated from the HEN17-A07 individual sequenced and assembled in this study unless
183 otherwise mentioned. The completeness and quality of the assembly was first assessed using Merqury
184 [27] k-mer analysis and freebayes [28] variant analysis. Merqury estimated the overall quality of the
185 assembly at a Phred-based [29] quality value (QV) score of 49 which corresponds to an error every
186 129,000 bases (Table 2). Comparison of k-mer profiles between the HiFi contigs and the previously
187 published TGACv2 red clover assembly [14] (accession GCA_900292005.1) using the upset python
188 module (<https://github.com/ImSoErgodic/py-upset>) (Figure 1) indicated that only 55.2% of all k-mers
189 were shared between the two assemblies. This surprisingly low shared content could be the result of real
190 differences in the genomes of the different varieties of this highly heterozygous species (the earlier
191 assembly used an individual from the Milvus variety versus the Hen17-A07 individual used here), or the
192 higher degree of completeness of the current assembly (the previous assembly was comprised of 135,023
193 contigs and was 68 Mb smaller total size), or assembly and haplotype switching errors in the short read
194 assembly, or a combination of these factors. The Themis-ASM analysis of TGACv2 estimated an error
195 every 142 bases, indicating that the ARS-RCv1.1 assembly has a three orders of magnitude improvement
196 in k-mer based QV estimates. Indeed, the count of erroneous, singleton k-mers identified in the TGACv2
197 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-

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

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

242 **Re-use potential and conclusions**

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

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

259

260 **Availability of source code and requirements**

261 Project name: Themis-ASM.

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

263 Operating systems: Unix, Linux.

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

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

266 License: GNU GPL

267 **Data availability**

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

273 **List of abbreviations**

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

277 **Competing interests**

278 The authors declare that they have no competing interests

279 **Funding**

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

282 **Permissions**

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

285 **Author's contributions**

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

289 **Acknowledgements**

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

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

293 **References**

- 294 1. Riday H. Progress made in improving red clover (*Trifolium pratense* L.) through breeding. Int J
295 Plant Breed. 2010;4:22-9.
- 296 2. Albrecht KA and Muck RE. Proteolysis in ensiled forage legumes that vary in tannin
297 concentration. Crop Sci. 1991;31:464-9.
- 298 3. Broderick GA. Utilization of protein in red clover and alfalfa silages by lactating dairy cows and
299 growing lambs. J Dairy Sci. 2018;101:1190-205. doi:10.3168/jds.2017-13690.
- 300 4. Sullivan ML and Hatfield RD. Polyphenol oxidase and *o*-diphenols inhibit postharvest
301 proteolysis in red clover and alfalfa. Crop Sci. 2006;46:662-70.
- 302 5. Berard NC, Wang Y, Wittenberg KM, Krause DO, Coulman BE, McAllister TA, et al.
303 Condensed tannin concentrations found in vegetative and mature forage legumes grown in
304 western Canada. Can J Plant Sci. 2011;91:669-75. doi:10.4141/cjps10153.
- 305 6. Saviranta NMM, Julkunen-Tiitto R, Oksanen E and Karjalainen RO. Leaf phenolic compounds in
306 red clover (*Trifolium pratense* L.) induced by exposure to moderately elevated ozone. Environ
307 Polut. 2010;158:440-6.
- 308 7. Oleszek W, Stochmal A and Janda B. Concentration of isoflavones and other phenolics in the
309 aerial parts of *Trifolium* species. J Agr Food Chem. 2007;55:8095-100. doi: 10.1021/Jf072024w.

- 310 8. Kramer R, Keogh RG and McDonald MF. The accumulation and clearance of equol in the blood
311 of ewes grazed on either high or low formononetin red clovers. Proc N Z Soc Anim Prod.
312 1996;56:373-7.
- 313 9. Harlow BE, Flythe MD, Kagan IA, Goodman JP, Klotz JL and Aiken GE. Isoflavone
314 supplementation, via red clover hay, alters the rumen microbial community and promotes weight
315 gain of steers grazing mixed grass pastures. PLoS One. 2020;15 3:e0229200.
316 doi:10.1371/journal.pone.0229200.
- 317 10. Kolodziejczyk-Czepas J, Krzyzanowska-Kowalczyk J, Sieradzka M, Nowak P and Stochmal A.
318 Clovamide and clovamide-rich extracts of three *Trifolium* species as antioxidants and moderate
319 antiplatelet agents in vitro. Phytochemistry. 2017;143:54-63.
320 doi:10.1016/j.phytochem.2017.07.011.
- 321 11. Townsend CE and Taylor NL. Incompatibility and plant breeding. In: Taylor NL, editor. Clover
322 Science and Technology. Madison, WI, USA: ASA-CSSA-SSSA; 1985. p. 365-81.
- 323 12. Riday H and Krohn AL. Genetic map-based location of the red clover (*Trifolium pratense* L.)
324 gametophytic self-incompatibility locus. Theor Appl Genet. 2010;121:761-7.
325 doi:10.1007/s00122-010-1347-0.
- 326 13. Williams RD and Silow RA. Genetics of red clover (*Trifolium pratense* L.): Compatibility I. J
327 Genet. 1933;27:341-62.
- 328 14. De Vega JJ, Ayling S, Hegarty M, Kudrna D, Goicoechea JL, Ergon Å, et al. Red clover
329 (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Sci Rep.
330 2015;5:17394. doi:10.1038/srep17394.

- 331 15. Istvanek J, Jaros M, Krenek A and Repkova J. Genome assembly and annotation for red clover
332 (*Trifolium pratense*; *Fabaceae*). *Am J Bot.* 2014;101:327-37. doi:10.3732/ajb.1300340.
- 333 16. Riday H, Casler MD, Crooks A and Wood T. Persistence of grazed red clover varieties. *Grass*
334 *Clippings* (University of Wisconsin Extension, Center for Integrated Agricultural Systems, and
335 College of Agriculture and Live Sciences). 2007;2:3-8.
- 336 17. Venuto BC. Reaction of red clover (*Trifolium pratense* L.) to *Fusarium oxysporum* Schlect. and
337 selection for and inheritance of resistance in red clover. (Thesis) University of Wisconsin-
338 Madison, Madison, WI, USA, 1993.
- 339 18. Association of Official Seed Certifying Agencies. Dominion red clover. 2005. Moline, IL, USA:
340 Association of Official Seed Certifying Agencies.
- 341 19. Guan D, McCarthy SA, Wood J, Howe K, Wang Y and Durbin R. Identifying and removing
342 haplotypic duplication in primary genome assemblies. *Bioinformatics.* 2020;36:2896-8.
343 doi:10.1093/bioinformatics/btaa025.
- 344 20. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.
345 *Bioinformatics.* 2016;32:2103-10. doi:10.1093/bioinformatics/btw152.
- 346 21. Ghurye J, Pop M, Koren S, Bickhart D and Chin C-S. Scaffolding of long read assemblies using
347 long range contact information. *BMC Genomics.* 2017;18 1:527. doi:10.1186/s12864-017-3879-
348 z.
- 349 22. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
350 arXiv:13033997 [q-bio]. 2013.
- 351 23. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.
352 *Bioinformatics.* 2010;26:841-2. doi:10.1093/bioinformatics/btq033.

- 353 24. Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, et al. An improved genome
354 release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics.
355 2014;15:312. doi:10.1186/1471-2164-15-312.
- 356 25. Isobe S, Kölliker R, Hisano H, Sasamoto S, Wada T, Klimenko I, et al. Construction of a
357 consensus linkage map for red clover (*Trifolium pratense* L.). BMC Plant Biol. 2009;9:57.
358 doi:10.1186/1471-2229-9-57.
- 359 26. Heaton MP, Smith TPL, Bickhart DM, Vander Ley BL, Kuehn LA, Oppenheimer J, et al. A
360 Reference Genome Assembly of Simmental Cattle, *Bos taurus taurus*. J Hered. 2021;112:184-91.
361 doi:10.1093/jhered/esab002.
- 362 27. Rhie A, Walenz BP, Koren S and Phillippy AM. Merqury: reference-free quality, completeness,
363 and phasing assessment for genome assemblies. Genome Biol. 2020;21:245. doi:10.1186/s13059-
364 020-02134-9.
- 365 28. Garrison E and Marth G. Haplotype-based variant detection from short-read sequencing.
366 arXiv:12073907 [q-bio]. 2012.
- 367 29. Ewing B and Green P. Base-calling of automated sequencer traces using phred. II. Error
368 probabilities. Genome Res. 1998;8:186-94.
- 369 30. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule
370 sequencing and chromatin conformation capture enable de novo reference assembly of the
371 domestic goat genome. Nat Genet. 2017;49:643-50. doi:10.1038/ng.3802.
- 372 31. Ha J, Satyawan D, Jeong H, Lee E, Cho K-H, Kim MY, et al. A near-complete genome sequence
373 of mungbean (*Vigna radiata* L.) provides key insights into the modern breeding program. Plant
374 Genome. 2021:e20121. doi:10.1002/tpg2.20121.

- 375 32. Moghaddam SM, Oladzad A, Koh C, Ramsay L, Hart JP, Mamidi S, et al. The tepary bean
376 genome provides insight into evolution and domestication under heat stress. *Nat Commun.*
377 2021;12:2638. doi:10.1038/s41467-021-22858-x.
- 378 33. Vezzi F, Narzisi G and Mishra B. Reevaluating Assembly Evaluations with Feature Response
379 Curves: GAGE and Assemblathon. *PLoS One.* 2012;7 12:e52210.
380 doi:10.1371/journal.pone.0052210.
- 381 34. Layer RM, Chiang C, Quinlan AR and Hall IM. LUMPY: a probabilistic framework for structural
382 variant discovery. *Genome Biol.* 2014;15 6:R84. doi:10.1186/gb-2014-15-6-r84.

383 **Figure Captions**

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

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

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

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

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

414

415

416

417

418

419 Table 1: Assembly Size Statistics

420

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

421

422

423 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

424

425

426

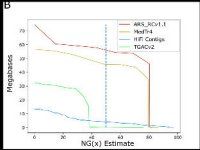
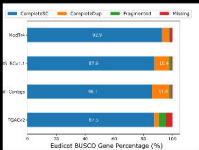
427 Table 3: Structural variant analysis

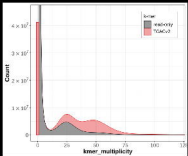
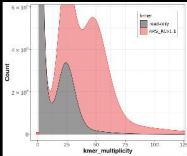
Category	TGACv2	ARS_RCv1.1	Description
HIGH_SPAN_PE	65,254	2,052	FRC_align identified regions with high numbers of inter-contig paired-end read mappings
Lumpy Deletions	20,727	20,945	Number of identified structural variant deletions
Lumpy Duplications	6,554	3,823	Number of identified structural variant 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

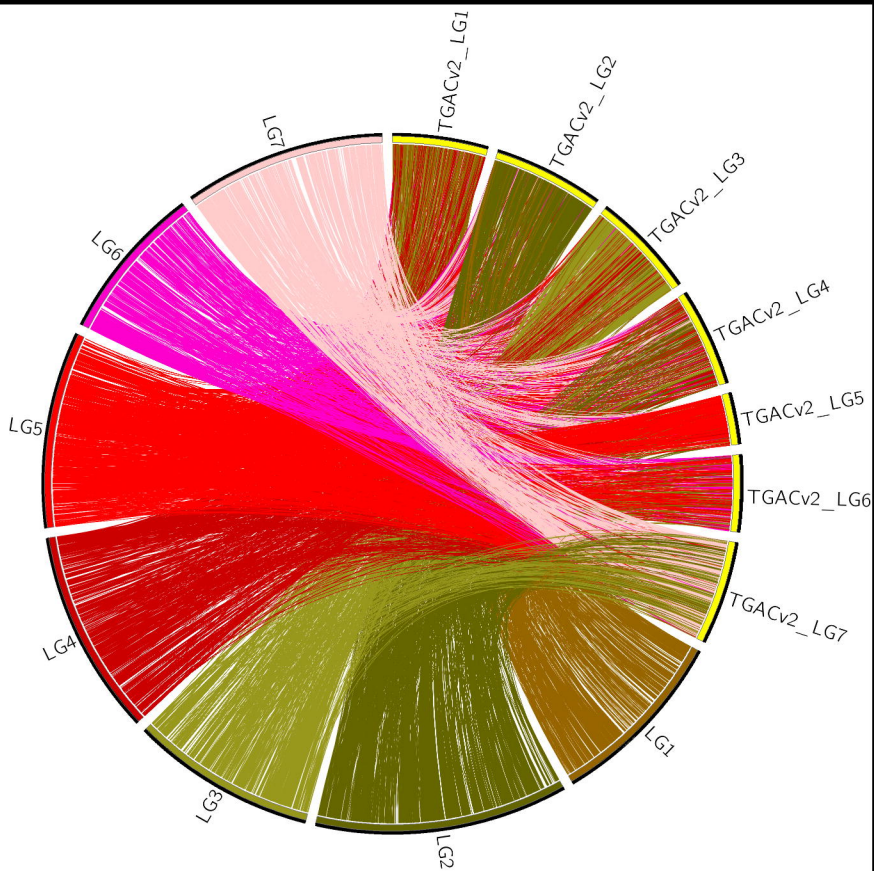
428

429









Variants 75 to 500,000 bp

