Chromosome-scale reference genome and RAD-based genetic map of yellow starthistle (*Centaurea solstitialis*) reveal putative structural variation and QTLs associated with invader traits

Running Title: Reference genome, genetic map, and invasion QTL in *Centaurea solstitialis*

Authors and affiliations:

Bryan Reatini[1,3]*, F. Alice Cang[1], Qiuyu Jiang[1], Michael T. W. McKibben[1,4], Michael S. Barker[1,5], Loren H. Rieseberg[2,6], Katrina M. Dlugosch[1,7]

[1]Department of Ecology and Evolutionary Biology, University of Arizona

[2]Department of Botany, University of British Columbia

[3]ORCID:0000-0001-8643-4296

[4]ORCID:0000-0002-1342-0085

[5]ORCID:0000-0001-7173-1319

[6]ORCID:0000-0002-2712-2417

[7]ORCID:0000-0002-7302-6637

* *To whom correspondence should be addressed:* reati1bs@gmail.com

Abstract: Yellow starthistle (*Centaurea solstitialis*) is a highly invasive species that has been a model system for the potential contribution of evolution to invader traits. Here, we report the construction of a chromosome-scale reference genome for *C. solstitialis* using a combination of PacBio HiFi and Dovetail Omni-C technologies, and functional gene annotation with RNAseq. We validate the reference genome using a restriction site-associated DNA (RAD)-based genetic map from an F2 mapping population. We find that syntenic comparisons to other taxa in the Asteraceae reveal a chromosomal fusion in the lineage of *C. solstitialis*, and widespread fission in globe artichoke (*Cynara cardunculus*). Using a QTL analysis from the mapping population (derived from a cross between native and invading parents) we identify 13 QTL underpinning size traits that are associated with adaptation in the invaded range, including a putative large-scale chromosomal inversion that has a pleiotropic and overdominant effect on key invader traits.

Keywords: reference genome, QTL, genetic map, invasive species, Centaurea solstitialis

Introduction

Yellow starthistle (*Centaurea solstitialis* L., Asteraceae) is a highly invasive plant on at least four continents and occupies a broad distribution throughout its native range in Eurasia (Maddox et al., 1985). Due to its ability to occupy a wide range of environments, and its profoundly negative ecological and economic impact in the invaded range, *C. solstitialis* has become a model system for understanding how invasive species might exploit open niches and dominate the communities in which they invade (e.g. Barker et al., 2017; Dlugosch et al., 2015a; Dukes et al., 2011; Graebner et al., 2012; Hierro et al., 2006; Zavaleta and Hulvey, 2004). This species is also well-suited to genomic study of the genetic basis of its success, as an annual, diploid (1N = 8), species of modest genome size (860 Mbp) (Bancheva & Greilhuber, 2006; Heiser Jr. & Whitaker, 1948; Irimia et al., 2017). Here, we present a chromosome-scale reference genome and associated analyses to facilitate genomic investigations of this model invader.

Centaurea solstitialis was introduced to South America in the 1600s and then to North America in the 1800s, and has since spread aggressively throughout the Western US, Chile, and Argentina (Barker et al., 2017; Gerlach, 1997). For invasive populations in California in particular, severe invasions are associated with the evolution of increased biomass and reproduction (Barker et al., 2017; Widmer et al., 2007). Within this invasion, plant size and biomass show evidence of adaptive clines along local environmental gradients (Dlugosch, Cang, et al., 2015)**.** Larger plant size is associated with increased reproductive fitness (Dlugosch, Cang, et al., 2015) and potentially an increased competitive ability (Montesinos et al., 2019;

56 Montesinos & Callaway, 2017). The former is predicted to lead to higher population growth
57 rates for invader genotypes relative to native genotypes (Dlugosch, Cang, et al., 2015).
58 Together, these observations suggest that the evolution of increased plant size in
59 Californian *C. solstitialis* populations may have directly contributed to the invasiveness of this
60 species, which makes it an attractive system to investigate how local adaptation can contribute
61 to biological invasions. Understanding the genetic basis of putative invasion traits is a core
62 objective of invasion biology (Bock et al., 2015; Dlugosch, Anderson, et al., 2015). However,
63 genomic resources are currently limited for many invasive species, which has impeded
64 investigations of the genetic basis of local adaptation for important invasion traits (Bock et al.,
65 2015; Dlugosch, Anderson, et al., 2015).
66 Here, we present a chromosome-scale reference genome for *C. solstitialis*, with
67 annotations of repetitive elements and functional genes. We validate the reference genome
68 assembly by comparing it with a genetic map constructed from an F2 mapping population, and
69 by carrying out gene synteny comparisons between starthistle and other representatives of
70 *Asteraceae*. We then demonstrate the utility of these genomic resources by investigating the
71 regions of the genome underpinning increased plant size in Californian populations via QTL
72 analysis, identifying candidate genes within those QTL regions.
73
74 **Materials and Methods**
75 *Genomic reference sample collection and sequencing*
76 Plant tissue used for the reference genome was grown from seed collected in August
77 2018 from the native range near Canales, Spain (site code 'CAN'; Lat: 41.00033, Long: -4.89718).

78 A reference voucher from the same locality is archived at the University of Arizona herbarium
79 (ARIZ #425375). The plant was reared to the early bolting stage in a greenhouse at the
80 University of Arizona (Tucson, Arizona, USA) under ambient light conditions from January to
81 July 2020, and placed in the dark for 24 hours before harvesting leaves directly into liquid
82 nitrogen for DNA extraction.

83 Leaf tissue was sent to Dovetail Genomics (Scotts Valley, California, USA) for reference
84 genome sequencing and assembly using a combination of PacBio HiFi and Dovetail Omni-C
85 approaches. Specifically, PacBio HiFi reads were generated by first constructing a ~20kb library
86 using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA, USA) using the
87 manufacturer's recommended protocol. The library was then bound to polymerase using the
88 Sequel II Binding Kit 2.0 (PacBio) and sequenced on PacBio Sequel II 8M SMRT cells, generating
89 45 Gb of data. Dovetail Omni-C libraries were prepped by fixing chromatin with formaldehyde,
90 then digesting chromatin with DNAse I. Proximity ligation was then performed, crosslinks were
91 reversed, and DNA was purified. Sequencing libraries were generated using NEBNext Ultra
92 enzymes and Illumina-compatible adapters, followed by PCR enrichment. Libraries were
93 sequenced on an Illumina HiSeqX platform, at ~30X coverage.

94

95 *Genome assembly*

96 The initial HiFi assembly was constructed *de novo* from PacBio reads using Wtdbg2
97 (Ruan & Li, 2020). Potential contamination in the initial assembly was identified and removed
98 based on BLAST v2.9 results against the NT database using Blobtools v1.1.1 (Laetsch & Blaxter,
99 2017). Haplotypic duplications were then filtered from the assembly using purge_dups v1.1.2

100 (Guan et al., 2020). The filtered initial assembly was used as input along with the Dovetail
101 Omni-C reads for scaffolding using the HiRise pipeline (Putnam et al., 2016). Briefly, Omni-C
102 reads were aligned to the draft assembly, a likelihood model for genomic distance between
103 read pairs was produced, and that model was then used to make joins and break misjoins in the
104 input draft assembly. The completeness of the final HiRise assembly was evaluated using the
105 255 Benchmarking Universal Single-Copy Orthologs (BUSCOs) from the Eukaryota dataset
106 (eukaryote_odb10) using BUSCO v4.0.5 (Manni et al., 2021).
107
108 *Genome annotation*
109 Repeat family identification and classification were performed on the final genome
110 assembly using RepeatModeler v2.0.1 (http://www.repeatmasker.org/RepeatModeler), relying
111 on RECON v1.08 (Bao & Eddy, 2002) and RepeatScout v1.0.6 (Price et al., 2005) for *de novo*
112 identification of repeats. Custom repeat libraries produced by RepeatModeler were then used
113 to identify and mask repeats in the final assembly using RepeatMasker v4.1.0
114 (https://www.repeatmasker.org/RepeatMasker/). High-confidence transposable elements were
115 then identified and classified by running RepeatMasker v4.1.0 with the plant TE database
116 nrTEplantsApril2020 (Contreras-Moreira et al., 2021) as the input library. In brief, this library
117 combines multiple plant TE databases and prunes the resulting library to remove redundant
118 sequences and minimize overlap with protein-coding domains in nucleotide-binding, leucine-
119 rich repeat (NLR) genes (Contreras-Moreira et al., 2021). Repetitive elements were then soft-
120 masked in the reference.

121 For gene identification, transcripts from six *C. solstitialis* individuals were sequenced
122 using RNAseq. Tissues included leaves from the same individual used for the reference genome,
123 whole shoots of two additional small seedlings from the same population ('CAN'), and three
124 individuals (again including two seedlings and a mature bolting individual) from seed collected
125 in September 2016 near Gilroy, California, USA in the invaded range (site code 'GIL'; Lat:
126 37.03373, Long: -121.53674; reference voucher ARIZ #425113). All seeds were grown in early
127 2020 in greenhouses at the University of Arizona under ambient light conditions. Tissue was
128 flash-frozen in liquid nitrogen and sent to Genewiz (Azenta Life Sciences, South Plainfield, NJew
129 Jersey, USA) for RNA extraction and sequencing. RNA libraries were prepared via rRNA-
130 depletion, and paired-end 2x150 reads were generated on the Illumina HiSeq platform.
131 Gene annotation was performed on the final repeat masked genome using a
132 combination of AUGUSTUS v2.5.5 (Stanke et al., 2006) and SNAP v2006-07-28 (Korf, 2004).
133 Specifically, coding sequences from *Cynara cardunculus* (Acquadro et al., 2020), *Lactuca sativa*
134 (Reyes-Chin-Wo et al., 2017), and *Helianthus annuus* (Badouin et al., 2017) were used to train
135 two independent *ab initio* models for *C. solstitialis* using AUGUSTUS and SNAP. RNAseq data
136 were then aligned to the final genome assembly using the STAR aligner software v2.7 (Dobin et
137 al., 2013) and AUGUSTUS was used to generate intron hints using the *bam2hints* tool. Gene
138 predictions were generated using MAKER (Holt & Yandell, 2011), SNAP, and AUGUSTUS, using
139 Swiss-Prot peptide sequences from the UniProt database to guide prediction and generate
140 peptide evidence in the MAKER pipeline. The final set of genes was filtered to contain only
141 genes predicted by both AUGUSTUS and SNAP. Putative gene function was assessed by
142 performing a BLAST search of the peptide sequences against the UniProt database, and tRNA

143 predictions were generated using tRNAscan-SE v2.05 (Chan et al., 2021). The quality of the
144 annotation was evaluated by plotting annotation edit distance from MAKER2 and gene synteny
145 was investigated between *C. solstitialis*, *Cynara cardunculus,* and *Lactuca sativa* using MCScanX
146 via TBTools v1.098696 (Chen et al., 2020) and visualized using SynVisio (Bandi & Gutwin, 2020).
147 The genome assembly and annotation of globe artichoke were downloaded from
148 http://www.artichokegenome.unito.it and the lettuce assembly (Lsat_Salinas_v7) and
149 annotation were downloaded from NCBI (BioProject: PRJNA173551).
150 Putative centromeric regions were identified using a sliding window analysis of gene and
151 repeat density using custom Python scripts (*YST_genome.ipynb*; zenodo citation). Specifically,
152 gene density from the gene annotation, unique repetitive element density from the repeat
153 annotation, and Gypsy/Copia long terminal repeat retrotransposon (LTR-RT) density from the
154 curated TE database were all calculated within 5Mbp sized windows with a step size of 100,000
155 bp for each chromosome. The midpoint of each element was used to define its location in the
156 genome. Counts for each of these four categories were normalized by dividing the count for the
157 window by the maximum count for the category within the chromosome, yielding a
158 measurement of the relative density of each category across the chromosome. Given that each
159 of the categories differs in their predicted association with centromere proximity (i.e. gene
160 density is predicted to be low near centromeres but LTR-RT density is often predicted to be
161 high), the direction of the effect was rescaled such that higher values always corresponded with
162 putative association with the centromere for all four categories. The average of the four
163 categories was then calculated for each window in order to assign a score ranging from 0-1 to
164 the window, with a score of 1 indicating a perfect score across all four categories (minimum

165 gene density, minimum unique repeat density, and maximum LTR-RT density for both Copia
166 and Gypsy elements). This score was then plotted across each chromosome to estimate the
167 location of centromeric regions.
168
169 *Mapping population collection and sequencing*
170 An F2 mapping population was created using an initial cross between a single native
171 range maternal parent and a single invaded range paternal parent. The native parent was
172 grown from seeds collected near Kirklareli, Turkey (site code 'TK23'; Lat: 41.751233, Long:
173 27.247883) in September 2008. The invader parent was grown from seeds collected near
174 Mariposa, California, USA (site code 'TRI'; Lat: 37.46178, Long: -119.79218) in 2008. Vouchers
175 from each of these populations were deposited in the ARIZ herbarium (#425116-425117).
176 Parental plants were grown in a greenhouse at the University of British Columbia
177 (Vancouver, British Columbia, Canada) and hand pollinated to produce F1 seed. Crosses were
178 performed by covering flowering heads (capitula) with white organza bags before the heads
179 opened to prevent pollination, then clipping heads presenting pollen and using these to brush
180 against heads presenting a large fraction of receptive stigmas, before re-covering receiving
181 heads until seed maturation. Of the resulting F1 progeny, four were selected at random and
182 reared to a large size in 11.4 L pots of commercial potting soil with 1 mL / L of 13:13:13
183 Osmocote fertilizer (Scotts Miracle-Gro, Marysville, Ohio, USA). The four F1 plants were
184 reciprocally crossed to one another in two pairs, and samples of their leaves stored at -80 C for
185 later genotyping.

186 Resulting F2 seeds were germinated on moist potting soil in an environmental room set
187 to 12 hr days and 16-18 C days / 14 C nights, and misted by hand every other day. A total of
188 2901 F2 seedlings were transplanted to 410 ml Deepots (Steuwe & Sons, Tangent, Oregon, USA)
189 in a 50:50 mix of silica sand and potting soil, with 2 mL of 13:13:13 Osmocote fertilizer. Deepots
190 were watered daily from below on a flood table in a greenhouse at the University of British
191 Columbia, under ambient lighting. Of the 2901 F2 plants, 300 were selected for genotyping
192 based on measurements of maximum leaf length at the first time point (see trait measurements
193 below), including 150 plants with the longest maximum leaf length and 150 plants with the
194 shortest maximum leaf length from the F2 generation.
195 Leaves from the 300 selected F2s, all four F1s, and six siblings of each parent (from field-
196 collected seeds) were used for genotyping. DNA was extracted using a modified CTAB/PVP
197 extraction protocol (Webb & Knapp, 1990) and sent to Floragenex (Beaverton, Oregon, USA) for
198 single digest Restriction Site Associated DNA (RAD) sequencing (Miller et al., 2007). RAD library
199 preparation was performed with the CpNpG 5-methylcytosine sensitive enzyme, *PstI*, and 1x80
200 bp reads were sequenced on the Illumina Genome Analyzer II platform, yielding an average of
201 29.8X coverage per sample.
202
203 *Reference-guided genetic map*
204 Raw single-end reads from the F2 mapping population were trimmed with the
205 *process_radtags* function in STACKS v2.60 (Rochette et al., 2019) using the *c*, *q*, and *r* flags and
206 specifying the *Pst*I restriction enzyme used in RAD library preparation. Trimmed reads were
207 aligned to the largest eight scaffolds of the final reference assembly (reference chromosomes,

208 see Results) using the *mem* function of the Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009)
209 and the resulting alignments were coordinate-sorted using samtools v1.10 (Li et al., 2009). RAD
210 loci were built and genotyped using the reference-aligned pipeline of STACKS via the *ref_map.pl*
211 wrapper. The *populations* function of STACKS was then used to filter the resulting genotype
212 data, requiring that a locus be present in each generation (P1, F1, and F2) and with a minimum
213 of 50% individuals represented at genotyped sites in order to be kept. Filtered data were
214 exported in variant call format, and parental genotypes for genetic mapping were called for
215 each RAD marker using an allele frequency-based approach. A custom python script
216 (*genmap_preprocessvcf.py*; zenodo citation) used the genotypes of six siblings of each parent to
217 polarize the alleles present in F1s as derived from either the native or invading population,
218 using the following criteria: for each allele present in the F1s at each RAD locus, the allele
219 frequency among native (TK) sibs and invader (TRI) sibs was compared and if freq(TRI)>freq(TK)
220 the allele was identified as predominantly invaded range (denoted 'A'), whereas if
221 freq(TK)>freq(TRI) the allele was identified as predominantly native range (denoted 'B'). For a
222 marker to be retained, all alleles in F1s needed to be successfully polarized, and there could be
223 no more than two missing F1 genotypes or one missing parental population genotype. Filtered
224 alleles were collapsed into parental classes (A and B), and F2 genotypes were exported to R/qtl
225 format using these genotype calls.
226 Importantly, although the reference genome was used to call genotypes at each RAD
227 marker, it was not used to group markers into linkage groups for genetic map construction.
228 Instead, initial linkage groups for the genetic map were built in R/qtl v1.50 (Broman et al., 2003)
229 using the recombination fractions observed among F2 genotypes. Specifically, a maximum

230 recombination fraction of 0.35 and minimum LOD score of 15 were used to group markers into
231 initial linkage groups. Markers showing evidence of significant segregation distortion were
232 removed using a significance cutoff of p=1E-5. The eight largest linkage groups were used for
233 the genetic map, assuming these linkage groups corresponded with the eight chromosomes of
234 *C. solstitialis*. The genetic distances between markers in the genetic map were then estimated
235 using R/qtl2 v0.28 (Broman et al., 2019). Specifically, genetic distances were first estimated
236 using the *est_map* function of R/qtl2 under a range of genotype error rates [0.001, 0.01, 0.025,
237 0.05, 0.075, 0.1]. The most likely error rate (0.075) was estimated using the $\log_{10}$ likelihood of
238 each estimated map and this rate was used to estimate distances in the final genetic map.
239 To validate the linear sequence of the chromosomes in the reference genome, we
240 plotted the collinearity between the genetic map and the physical reference using custom
241 Python scripts (*YST_genome.ipynb*; zenodo citation). Abberant markers in the genetic map -
242 those that fell on different chromosomes than the rest of the linkage group - were quantified to
243 determine the consistency between the reference genome and genetic map, and these
244 abberant markers where ultimately removed from downstream analyses. Collinearity patterns
245 were then compared with centromere genome scans in order to evaluate whether patterns of
246 recombination were consistent with the estimated centromere positions, again using custom
247 Python scripts (*YST_genome.ipynb*; zenodo citation).
248
249 *Trait measurements and QTL analyses*
250 The length of the longest leaf and the total number of leaves were recorded for all 2901
251 plants at 3.5 weeks and 5 weeks of age. Potential confounding factors during trait

252 measurement were accounted for by fitting linear models for each trait measurement using the
253 *aov* function in R, with fixed effects of the day measurements were collected, the individual
254 collecting the data, the spatial location in the greenhouse (block), the location of the plant
255 within the block, and the identity of the specific F1 mother of each F2. Residuals of the model
256 were then extracted and added to the grand mean to obtain corrected values of maximum leaf
257 length and total number of leaves at both time points for use in QTL analysis.

258 Additional representatives of both parental populations and F1 crosses between these
259 populations were included in the experiment for phenotypic comparison to the F2s. Parental
260 population plants were grown from field-collected seed including 28 'TRI' plants from the
261 invasion and 29 'TK23' plants from the native range. F1 seeds came from additional controlled
262 crosses, and included 17 plants with 'TRI' genotypes as the maternal parents and 24 plants with
263 'TK23' genotypes as the maternal parents. Germination and rearing were concurrent with the
264 F2 populations, under the same conditions.

265 Genome scans to identify QTL were performed using both Haley-Knott regression (Haley
266 & Knott, 1992) and the leave one chromosome out (LOCO; Yang et al., 2014) approach in R/qtl2
267 using the final genetic map and corrected phenotypic data as input. Both models were used to
268 qualitatively gauge support for QTL by comparing overlap between them. Given that the LOCO
269 method accounts for kinship when crosses are generated with parents from divergent source
270 populations – as was the case for the present mapping population – the LOCO model was used
271 as the primary model for candidate gene identification. For each method, two thresholds were
272 used to define QTL peaks using a permutation-based approach in R/qtl2: a standard significance
273 threshold of 0.05 and a suggestive threshold of 0.1. Separate peaks on the same chromosome

274 were identified using a minimum LOD decline of 1 between peaks. Candidate genes within QTL
275 peaks were identified by extracting all genes within a decline of 1 LOD on either side of the
276 peak.
277
278 **Results**
279 *Genome assembly and annotation*
280 The final chromosome-scale reference genome contained 8 primary scaffolds
281 corresponding with the haploid chromosome number of *C. solstitialis* (n=8) totaling ~725 Mbp.
282 The 8 primary scaffolds composed 94.8% of the total sequence content of the reference (L90=8,
283 Figure 1A), and 84.4% of the average genome size of *C. solstitialis* (860 Mbp) estimated by flow
284 cytometry (Irimia et al., 2017). An additional 1072 unplaced scaffolds totaled ~39 Mbp. Of the
285 255 BUSCOs in the eukaryote_odb10 dataset, 237 (92.9%) complete BUSCOs were found in the
286 8 primary scaffolds: 188 (73.7%) were complete and single-copy and 49 (19.2%) were complete
287 and duplicated (Table 1). None of the BUSCOs that were missing or fragmented in the 8 primary
288 scaffolds were located in the trailing scaffolds, but 4 of the BUSCOs that were found in single-
289 copy in the 8 primary scaffolds were duplicated in the trailing scaffolds. Given these results, we
290 hereafter refer to the 8 primary scaffolds as chromosomes, numbered by size.
291 Roughly 63.3% of the reference was repetitive DNA, with 29.04% identified as class I
292 transposable elements (retrotransposons) and 2.07% identified as class II transposable
293 elements (DNA transposons) (Table 1). A total of 481,954 retroelements - composing 12% of
294 the genome - were successfully classified using the high-confidence nrTEplants database
295 including 122,163 Ty1/Copia and 172,604 Gypsy LTR-RTs.

296 The complete reference annotation included 34,323 predicted gene models
297 encompassing a total of 59.8 Mbp of the reference sequence (Table 1). Of the 32,431 genes
298 that fell on the 8 putative chromosomes, 24,011 (74%) were able to be functionally annotated
299 using the UniProt database. A total of 616 tRNA prediction models were identified on the 8
300 putative chromosomes using tRNAscan-SE, which is similar to the 639 identified in *Arabidopsis*
301 *thaliana* (Chan et al., 2021). Using the cumulative frequency of annotation edit distance (from 0
302 being perfect support to 1 being no support) as a metric of annotation quality, the majority of
303 genes in the complete annotation are well supported by overlapping aligned RNA-seq and
304 protein homology data (Supplemental Figure 1).
305 Gene synteny analyses between the *C. solstitialis* genome and that of globe artichoke
306 (*Cynara cardunculus*) – the most closely related thistle (subfamily *Carduoideae*) whose genome
307 has been assembled to the chromosome-scale – using MCScanX revealed large blocks of
308 syntenic genes across the genome (Figure 2). Conserved synteny was also evident in
309 comparisons between the *C. solstitialis* genome and that of *Lactuca sativa* (subfamily
310 *Cichorioideae*), albeit with more evidence for interchromosomal rearrangements (Figure 2).
311 Three-way comparisons between *C. solstitialis*, *Cynara cardunculus,* and *Lactuca sativa* revealed
312 a complex history of chromosome evolution within the thistle subfamily *Carduoideae* (Figure X).
313 In particular, by using *Lactuca sativa* as an outgroup, multiple chromosome fissions are evident
314 within the *Cynara cardunculus* lineage (Figure 2C), whereas chromosome 3 of *C. solstitialis*
315 shows evidence of being derived from chromosome fusion (Figure 2D), and chromosome 7 is
316 highly conserved across all three taxa (Figure 2E).

317 For the sliding window analysis to detect putative centromeric regions, the raw counts
318 for each of the individual categories varied in their ability to predict centromere location for
319 each of the 8 chromosomes (Supplemental Figure 2). However, the average score of the four
320 categories revealed a clear signal indicating the location of putative pericentromeric regions for
321 each of the 8 chromosomes (Figure 3). Sliding window analysis of gene density alone revealed
322 roughly 2X higher gene density at the ends of chromosome arms relative to putative
323 centromeric regions for the 8 presumed chromosomes of the *C. solstitialis* genome
324 (Supplemental Figure 2).
325
326 *Reference-guided genetic map*
327 A total of 1,765 RAD markers across the 300 F2 individuals were successfully polarized
328 and were used as input for initial linkage group formation in the genetic map. After filtering for
329 segregation distortion, 1,064 markers remained in the 8 linkage groups of the genetic map.
330 Each of the 8 linkage groups corresponded with one of the 8 chromosomes in the *C. solstitialis*
331 reference genome (Figure 4B), with only 5 aberrant markers from a single linkage group falling
332 on a different chromosome than the rest of the markers from their linkage group. These
333 aberrant markers were pruned in the final map. Strong linkage and low estimated
334 recombination were evident between markers within each chromosome, whereas linkage was
335 weak and recombination was estimated to be frequent between markers on different
336 chromosomes (Figure 4A).
337 Collinearity analyses between the genetic map and physical map revealed megabase-
338 scale regions of low recombination in the center of each chromosome, with chromosomes 1, 5,

339 7, and 8 showing low recombination at one of the two extremes of the chromosome as well
340 (Figure 3A). To evaluate whether these regions of low recombination correspond with
341 pericentromeric regions, we compared our collinearity analyses with our centromere genome
342 scans (Figure 3A,B). For all chromosomes except 1 and 7, regions of low recombination in the
343 genetic map corresponded closely with estimated pericentromeric regions (Figure 3A,B).
344 Moreover, these same regions corresponded with lower RAD marker density (Figure 3A;
345 Supplemental Figure 2). Chromosomes 1 and 7 displayed low recombination across the majority
346 of the chromosome - in putative pericentromeric regions and on one of the distal arms of each
347 chromosome (Figure 3A,B).
348
349 *Trait differences and QTL analyses*
350 Invader parental genotypes had greater total number of leaves and greater maximum
351 leaf length than native parental genotypes at both time points (Supplemental Figure 3), which is
352 consistent with previous studies of invasive and native *C. solstitialis* populations (Dlugosch,
353 Cang, et al., 2015). F1 individuals displayed heterosis for maximum leaf length at both time
354 points, and for total number of leaves at the second time point (Supplemental Figure 3),
355 whereas F2 individuals showed the greatest range in trait values (Supplemental Figure 3). Given
356 that the 300 genotyped F2s were selected from the ends of the distribution of longest leaf
357 length at time point 1, they showed an expected bimodal distribution for this trait
358 (Supplemental Figure 4). However, total number of leaves followed a roughly normal
359 distribution for both time points within the F2 mapping population used for downstream QTL
360 analyses (Supplemental Figure 4).

361 Haley-Knott regression and LOCO genome scans revealed a total of 13 distinct QTL
362 peaks with varying degrees of support (summarized in Table 2). A total of seven peaks were
363 associated with maximum leaf length alone, four were associated with total number of leaves
364 alone, and two peaks were associated with both traits (Table 2). For all four peaks associated
365 with total number of leaves, the dominance pattern appeared to be additive, and the invader
366 alleles at each QTL were associated with a greater number of leaves (Table 2). Of the seven
367 peaks associated with maximum leaf length, mixed patterns of dominance (additive, complete,
368 and overdominance) were observed and the parental allele driving larger leaves varied
369 between loci (Table 2). One of the two pleiotropic QTL displayed complete dominance favoring
370 the native allele, and the other pleiotropic QTL displayed overdominance. The latter QTL
371 corresponded with the large region of low recombination outside of the putative centromeric
372 region on chromosome 1 (Figure 3). Similarly, one of the suggestive QTL for total number of
373 leaves corresponded with the other large region of low recombination outside of the
374 pericentromeric region on chromosome 7 (Figure 3). The number of candidate genes under QTL
375 peaks depended on the size of the QTL peak, varying from 16 genes within 0.2 Mbp to 3532
376 genes within the 88.9 Mbp region of low recombination on chromosome 1 (Table 2).
377

378 **Discussion**

379 We constructed a chromosome-scale genome assembly, gene and repeat annotations,
380 and genetic map for yellow starthistle (*Centaurea solstitialis*), and we validated these genomic
381 resources using a combination of assembly statistics, collinearity analyses, and gene synteny
382 analyses. This is the first non-cultivated species to be assembled to chromosome-scale in the

383 thistle subfamily (*Carduoideae*) of *Asteraceae*, with the only other representative genome of
384 *Carduoideae* being cultivated globe artichoke.

385 Following genome assembly using PacBio HiFi reads and scaffolding using Dovetail
386 Omni-C data, the 8 pseudohaploid chromosomes of the assembly made up 94.8% of the total
387 sequence content and contained 92.9% complete BUSCO genes (Table 1). The lack of any
388 additional BUSCO genes within the 1062 trailing scaffolds - in addition to these scaffolds only
389 comprising 5.2% of the total sequence content of the genome (Figure 1A) - suggest that these
390 trailing scaffolds are likely composed of contamination or mis-assemblies due to heterozygosity
391 in the outbred wild genotype used for genome assembly. The Omni-C contact map of the 8
392 pseudohaploid chromosomes of the assembly (Figure 1B) also demonstrates two expected
393 patterns of chromosome-scale assemblies: 1) link density is higher within chromosomes than
394 between, and 2) link density decays with physical distance within a chromosome (Lajoie et al.,
395 2015). Each of the 8 linkage groups in the genetic map also corresponded with one of the 8
396 chromosomes of the assembly (Figure 4), and the recombination frequencies between markers
397 in the genetic map were consistent with the linear sequence of each chromosome (Figure 3A).
398 Together, these results suggest that the 8 pseudohaploid chromosomes of the assembly
399 represent the complete set of chromosomes of *C. solstitialis*.

400 Estimated recombination was generally highest at the ends of each chromosome and
401 decreased towards putative pericentromeric regions located roughly in the center of each
402 chromosome (Figure 3A,B) - although chromosomes 5 and 8 had highly asymmetric centromere
403 positions. The general locations of these putative pericentromeric regions were estimated using
404 genome scans of gene density, unique repeat density, and LTR-RT density (Figure 3B).

405 Pericentromeric regions also corresponded with low RAD marker density in the genetic map
406 (Figure 3A).
407 Interestingly, chromosomes 1 and 7 displayed recombination patterns that did not
408 appear to be explained by the location of centromeres alone. For both of these chromosomes,
409 no recombination was observed across one of the chromosome arms, despite the central
410 location of the centromeres (Figure 3A,B). Similar collinearity patterns have been explained by
411 chromosomal inversions in a range of plant and animal taxa (Huang et al., 2020; Kirubakaran et
412 al., 2016; Lee et al., 2017; Tong et al., 2016), suggesting that large-scale chromosomal
413 rearrangements may exist between native and invasive genotypes of *C. solstitialis* on these
414 chromosomes.
415 The *C. solstitialis* assembly is the second chromosome-scale reference genome for a
416 member of the *Carduoideae*, which provides new opportunities to examine genome structural
417 evolution in the Asteraceae. Gene synteny analyses between *C. solstitialis*, *Cynara cardunculus,*
418 and *Lactuca sativa* revealed large regions of conserved synteny between all three genomes
419 (Figure 2). The ancestral haploid chromosome number of *Asteraceae* is estimated to be n=9
420 (Mota et al., 2016), which is retained in *Lactuca sativa* (subfamily *Chicoriae*; n=9) but not *C.*
421 *solstitialis* (1N = 8) or *Cynara cardunculus* (1N = 17). By using *Lactuca sativa* as an outgroup, we
422 attempted to assess the most parsimonious history of chromosomal evolution within the *C.*
423 *solstitialis* and *Cynara cardunculus* ancestral lineages. *Cynara cardunculus* has over double the
424 chromosome number as *C. solstitialis*, yet Scaglione et al. (2016) found that no whole-genome
425 duplication appears to have occurred after speciation between *Chicoriae* and *Carduoideae*,
426 using the distribution of Ks values between *Cynara cardunculus* and *Lactuca*. Consistent with

427 Scaglione et al. (2016), our gene synteny analyses revealed evidence for multiple chromosome
428 fissions in *Cynara cardunculus* rather than whole-genome duplication (Figure 2A; exemplified in
429 Figure 2C). Karyotype analyses of *Cynara cardunculus* have revealed that the chromosomes
430 display an unusual size distribution with discrete size categories of large, medium, and small
431 chromosomes (Falistocco, 2016) - perhaps due to multiple fission events of larger ancestral
432 chromosomes into smaller chromosomes as our results suggest. Our gene synteny analyses also
433 suggest that the reduction in the chromosome number of *C. solstitialis* (n=8) relative to the
434 base number for *Asteraceae* (n=9) may be due to the fusion of two ancestral chromosomes
435 within the lineage leading up to *C. solstitialis* (Figure 2D) - although there appear to be many
436 other interchromosomal rearrangements between these genomes (Figure 2B).

437 Invasive *C. solstitialis* populations show evidence for adaptive evolution of increased size
438 and reproduction, genetic changes which are thought to have aided in its invasion success
439 (Barker et al., 2017; Dlugosch, Cang, et al., 2015). Using the reference genome and genetic
440 mapping population described here, we identified 13 QTL associated with two phenotypic
441 components of increased plant size in this species. We identified four QTL peaks associated
442 with the total number of leaves, and all of these loci displayed greater leaf number for invader
443 genotypes relative to the native genotypes (Table 2) - a pattern consistent with differences
444 observed between parental phenotypes (Supplementary Figure 3). We identified seven QTL
445 associated with maximum leaf length, for which the dominance pattern and favored allele
446 depended on the locus, suggesting that both native and invasive populations have segregating
447 variation for larger leaf length at different loci.

448 Two pleiotropic QTL were identified by our QTL analyses, one displaying overdominance
449 and the other favoring the native genotype. Interestingly, the overdominant pleiotropic QTL on
450 chromosome 1 corresponded with the putative inversion identified by collinearity analyses
451 (Figure 3). Given that heterosis is observed in the F1 generation for both traits (Supplemental
452 Figure 3), it is plausible that overdominance (and pleiotropy) at this QTL could be driven by an
453 inversion (Faria et al., 2019). Specifically, a large-scale inversion would prevent recombination
454 across a large portion of the chromosome, as appears to be the case in this region of
455 chromosome 1 (Figure 3A), which would "fix" heterozygous loci affecting both traits across its
456 length in the F2 generation. The inverted region - and all phenotypically relevant genes within it
457 - would act as a single heterotic and pleiotropic locus. Balancing selection on overdominant
458 inversions is a classic form of local adaptation in general (Faria et al., 2019), making this QTL a
459 particularly intriguing candidate for local adaptation in this system. Alternatively,
460 overdominance at this QTL could be explained by heterozygous loci across the inversion
461 masking deleterious alleles (Connallon & Olito, 2022; Jay et al., 2022), which are likely to
462 accumulate during biological invasions due to strong genetic drift (i.e. expansion load: Gilbert et
463 al., 2017; Peischl et al., 2015). In this case, the maintenance of the inversion polymorphism
464 within invasive populations would not need to be driven by local adaptation *per se*, but it could
465 still be an important variant contributing to invasion success if invader genotypes that are
466 heterozygous for the inversion have higher fitness than homozygotes.
467 The putative inversion on chromosome 7 (identified by collinearity analyses) was also
468 identified as a suggestive QTL using Haley-Knott regression, though with additive effects
469 (favoring more leaves in the invader). It is perhaps not surprising that QTL are associated with

470 both of the large inversions that we inferred from our genetic map and reference genome,
471 given that any trait loci in these chromosomal regions will be linked to the entire inversion.
472 Nevertheless, loci that have large phenotypic effects - such as inversions containing many
473 tightly linked genes affecting a trait - are predicted to be favored by selection during invasions
474 due to the relative strength of genetic drift (Bock et al., 2015; Dlugosch, Anderson, et al., 2015),
475 particularly when gene flow is involved (Reatini & Vision, 2020). Inversions are increasingly
476 being identified by genomic studies as playing important roles in adaptation (Huang and
477 Rieseberg, 2020) and have recently been shown to contribute to the success of a plant invasion
478 (Battlay et al., 2022). Our results suggest that these structural variants may play a role in
479 phenotypic evolution in invasive populations of *C. solstitialis* and merit further investigation.
480 Investigating the genetic basis of invasion traits remains a central objective of invasion
481 biology, and here we advance that goal in the *Centaurea solstitialis* system. Together, the
482 reference genome assembly, annotation, genetic map, and QTL candidate regions presented
483 here provide essential genomic resources for studies of *C. solstitialis* invasion. Moreover, these
484 genomic resources will be useful for investigating the history of genome evolution across the
485 agriculturally and evolutionarily important Asteraceae family.
486

487 **References**

488 Acquadro, A., Portis, E., Valentino, D., Barchi, L., & Lanteri, S. (2020). "Mind the Gap": Hi-C Technology
489 Boosts Contiguity of the Globe Artichoke Genome in Low-Recombination Regions. *G3:*
490 *Genes|Genomes|Genetics*, *10*(10), 3557–3564. https://doi.org/10.1534/g3.120.401446
491 Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Brière, C., Owens, G. L.,

Carrère, S., Mayjonade, B., Legrand, L., Gill, N., Kane, N. C., Bowers, J. E., Hubner, S., Bellec, A., Bérard, A., Bergès, H., Blanchet, N., … Langlade, N. B. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, *546*(7656), 148–152. https://doi.org/10.1038/nature22380

Bancheva, S., & Greilhuber, J. (2006). Genome size in Bulgarian Centaurea s.l. (Asteraceae). *Plant Systematics and Evolution*, *257*(1), 95–117. https://doi.org/10.1007/s00606-005-0384-7

Bandi, V., & Gutwin, C. (2020). Interactive Exploration of Genomic Conservation. *Graphics Interface*. https://doi.org/10.20380/GI2020.09

Bao, Z., & Eddy, S. R. (2002). Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Research*, *12*(8), 1269–1276. https://doi.org/10.1101/gr.88502

Barker, B. S., Andonian, K., Swope, S. M., Luster, D. G., & Dlugosch, K. M. (2017). Population genomic analyses reveal a history of range expansion and trait evolution across the native and invaded range of yellow starthistle (Centaurea solstitialis). *Molecular Ecology*, *26*(4), 1131–1147. https://doi.org/10.1111/mec.13998

Battlay, P., Wilson, J., Bieker, V. C., Lee, C., Prapas, D., Petersen, B., Craig, S., Boheemen, L. van, Scalone, R., Silva, N. P. de, Sharma, A., Konstantinović, B., Nurkowski, K. A., Rieseberg, L., Connallon, T., Martin, M. D., & Hodgins, K. A. (2022). *Large haploblocks underlie rapid adaptation in an invasive weed* (p. 2022.03.02.482376). bioRxiv. https://doi.org/10.1101/2022.03.02.482376

Bock, D. G., Caseys, C., Cousens, R. D., Hahn, M. A., Heredia, S. M., Hübner, S., Turner, K. G., Whitney, K. D., & Rieseberg, L. H. (2015). What we still don't know about invasion genetics. *Molecular Ecology*, *24*(9), 2277–2297. https://doi.org/10.1111/mec.13032

Broman, K. W., Gatti, D. M., Simecek, P., Furlotte, N. A., Prins, P., Sen, Ś., Yandell, B. S., & Churchill, G. A. (2019). R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. *Genetics*, *211*(2), 495–502.

516 https://doi.org/10.1534/genetics.118.301595

517 Broman, K. W., Wu, H., Sen, Ś., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses.
518 *Bioinformatics*, *19*(7), 889–890. https://doi.org/10.1093/bioinformatics/btg112

519 Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: Improved detection and
520 functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096.
521 https://doi.org/10.1093/nar/gkab688

522 Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., & Xia, R. (2020). TBtools: An Integrative
523 Toolkit Developed for Interactive Analyses of Big Biological Data. *Molecular Plant*, *13*(8), 1194–
524 1202. https://doi.org/10.1016/j.molp.2020.06.009

525 Connallon, T., & Olito, C. (2022). Natural selection and the distribution of chromosomal inversion
526 lengths. *Molecular Ecology*, *31*(13), 3627–3641. https://doi.org/10.1111/mec.16091

527 Contreras-Moreira, B., Filippi, C. V., Naamati, G., García Girón, C., Allen, J. E., & Flicek, P. (2021). K-mer
528 counting and curated libraries drive efficient annotation of repeats in plant genomes. *The Plant*
529 *Genome*, *14*(3), e20143. https://doi.org/10.1002/tpg2.20143

530 Dlugosch, K. M., Anderson, S. R., Braasch, J., Cang, F. A., & Gillette, H. D. (2015). The devil is in the
531 details: Genetic variation in introduced populations and its contributions to invasion. *Molecular*
532 *Ecology*, *24*(9), 2095–2111. https://doi.org/10.1111/mec.13183

533 Dlugosch, K. M., Cang, F. A., Barker, B. S., Andonian, K., Swope, S. M., & Rieseberg, L. H. (2015).
534 Evolution of invasiveness through increased resource use in a vacant niche. *Nature Plants*, *1*.
535 https://doi.org/10.1038/nplants.2015.66

536 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras,
537 T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.
538 https://doi.org/10.1093/bioinformatics/bts635

539 Dukes, J. S., Chiariello, N. R., Loarie, S. R., & Field, C. B. (2011). Strong response of an invasive plant

species (Centaurea solstitialis L.) to global environmental changes. *Ecological Applications*, *21*(6), 1887–1894. https://doi.org/10.1890/11-0111.1

Falistocco, E. (2016). Cytogenetic characterization of cultivated globe artichoke (Cynara cardunculus var. Scolymus) and cardoon (C. cardunculus var. Altilis). *Caryologia*, *69*(1), 1–4. https://doi.org/10.1080/00087114.2015.1109935

Faria, R., Johannesson, K., Butlin, R. K., & Westram, A. M. (2019). Evolving Inversions. *Trends in Ecology & Evolution*, *34*(3), 239–248. https://doi.org/10.1016/j.tree.2018.12.005

Gerlach, J. D. (1997). How the West Was Lost: Reconstructing the Invasion Dynamics of Yellow Starthistle and Other Plant Invaders of Western Rangelands and Natural Areas. *Calif. Exot. Pest Plant Counc. Symp. Proc.*, 6.

Gilbert, K. J., Sharp, N. P., Angert, A. L., Conte, G. L., Draghi, J. A., Guillaume, F., Hargreaves, A. L., Matthey-Doret, R., & Whitlock, M. C. (2017). Local Adaptation Interacts with Expansion Load during Range Expansion: Maladaptation Reduces Expansion Load. *The American Naturalist*, *189*(4), 368–380. https://doi.org/10.1086/690673

Graebner, R. C., Callaway, R. M., & Montesinos, D. (2012). Invasive species grows faster, competes better, and shows greater evolution toward increased seed size and growth than exotic non-invasive congeners. *Plant Ecology*, *213*(4), 545–553. https://doi.org/10.1007/s11258-012-0020-x

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Haley, C. S., & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, *69*(4), 315–324. https://doi.org/10.1038/hdy.1992.131

Heiser Jr., C. B., & Whitaker, T. W. (1948). Chromosome Number, Polyploidy, and Growth Habit in California Weeds. *American Journal of Botany*, *35*(3), 179–186. https://doi.org/10.1002/j.1537-

564 2197.1948.tb05204.x

565 Hierro, J. L., Villarreal, D., Eren, Ö., Graham, J. M., & Callaway, R. M. (2006). Disturbance Facilitates
566 Invasion: The Effects Are Stronger Abroad than at Home. *The American Naturalist*, *168*(2), 144–
567 156. https://doi.org/10.1086/505767

568 Holt, C., & Yandell, M. (2011). MAKER2: An annotation pipeline and genome-database management tool
569 for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491.
570 https://doi.org/10.1186/1471-2105-12-491

571 Huang, K., Andrew, R. L., Owens, G. L., Ostevik, K. L., & Rieseberg, L. H. (2020). Multiple chromosomal
572 inversions contribute to adaptive divergence of a dune sunflower ecotype. *Molecular Ecology*,
573 *29*(14), 2535–2549. https://doi.org/10.1111/mec.15428

574 Huang, K., & Rieseberg, L. H. (2020). Frequency, Origins, and Evolutionary Role of Chromosomal
575 Inversions in Plants. *Frontiers in Plant Science*, *11*.
576 https://www.frontiersin.org/articles/10.3389/fpls.2020.00296

577 Irimia, R.-E., Montesinos, D., Eren, Ö., Lortie, C. J., French, K., Cavieres, L. A., Sotes, G. J., Hierro, J. L.,
578 Jorge, A., & Loureiro, J. (2017). Extensive analysis of native and non-native Centaurea solstitialis
579 L. populations across the world shows no traces of polyploidization. *PeerJ*, *5*, e3531.
580 https://doi.org/10.7717/peerj.3531

581 Jay, P., Tezenas, E., Véber, A., & Giraud, T. (2022). Sheltering of deleterious mutations explains the
582 stepwise extension of recombination suppression on sex chromosomes and other supergenes.
583 *PLOS Biology*, *20*(7), e3001698. https://doi.org/10.1371/journal.pbio.3001698

584 Kirubakaran, T. G., Grove, H., Kent, M. P., Sandve, S. R., Baranski, M., Nome, T., De Rosa, M. C., Righino,
585 B., Johansen, T., Otterå, H., Sonesson, A., Lien, S., & Andersen, Ø. (2016). Two adjacent
586 inversions maintain genomic differentiation between migratory and stationary ecotypes of
587 Atlantic cod. *Molecular Ecology*, *25*(10), 2130–2143. https://doi.org/10.1111/mec.13592

588 Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*(1), 59.
589 https://doi.org/10.1186/1471-2105-5-59
590 Laetsch, D. R., & Blaxter, M. L. (2017). *BlobTools: Interrogation of genome assemblies* (6:1287).
591 F1000Research. https://doi.org/10.12688/f1000research.12232.1
592 Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines.
593 *Methods (San Diego, Calif.)*, *72*, 65–75. https://doi.org/10.1016/j.ymeth.2014.10.031
594 Lee, C.-R., Wang, B., Mojica, J. P., Mandáková, T., Prasad, K. V. S. K., Goicoechea, J. L., Perera, N.,
595 Hellsten, U., Hundley, H. N., Johnson, J., Grimwood, J., Barry, K., Fairclough, S., Jenkins, J. W., Yu,
596 Y., Kudrna, D., Zhang, J., Talag, J., Golser, W., … Mitchell-Olds, T. (2017). Young inversion with
597 multiple linked QTLs under selection in a hybrid zone. *Nature Ecology & Evolution*, *1*(5), 1–13.
598 https://doi.org/10.1038/s41559-017-0119
599 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform.
600 *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324
601 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R.
602 (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.
603 https://doi.org/10.1093/bioinformatics/btp352
604 Maddox, D. M., Mayfield, A., & Poritz, N. H. (1985). Distribution of Yellow Starthistle (Centaurea
605 solstitialis) and Russian Knapweed (Centaurea repens). *Weed Science*, *33*(3), 315–327.
606 Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and
607 Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of
608 Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–
609 4654. https://doi.org/10.1093/molbev/msab199
610 Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective
611 polymorphism identification and genotyping using restriction site associated DNA (RAD)

612 markers. *Genome Research*, *17*(2), 240–248. https://doi.org/10.1101/gr.5681207

613 Montesinos, D., & Callaway, R. M. (2017). Inter-regional hybrids of native and invasive Centaurea

614 solstitialis display intermediate competitive ability. *Ecography*, *40*(7), 801–802.

615 https://doi.org/10.1111/ecog.02653

616 Montesinos, D., Graebner, R. C., & Callaway, R. M. (2019). Evidence for evolution of increased

617 competitive ability for invasive Centaurea solstitialis, but not for naturalized C. calcitrapa.

618 *Biological Invasions*, *21*(1), 99–110. https://doi.org/10.1007/s10530-018-1807-z

619 Mota, L., Torices, R., & Loureiro, J. (2016). The Evolution of Haploid Chromosome Numbers in the

620 Sunflower Family. *Genome Biology and Evolution*, *8*(11), 3516–3528.

621 https://doi.org/10.1093/gbe/evw251

622 Peischl, S., Kirkpatrick, M., & Excoffier, L. (2015). Expansion Load and the Evolutionary Dynamics of a

623 Species Range. *The American Naturalist*, *185*(4), E81–E93. https://doi.org/10.1086/680220

624 Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large

625 genomes. *Bioinformatics (Oxford, England)*, *21 Suppl 1*, i351-358.

626 https://doi.org/10.1093/bioinformatics/bti1018

627 Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A.,

628 Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S., & Green, R. E. (2016). Chromosome-

629 scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*,

630 *26*(3), 342–350. https://doi.org/10.1101/gr.193474.115

631 Reatini, B., & Vision, T. J. (2020). Genetic architecture influences when and how hybridization

632 contributes to colonization. *Evolution*, *74*(8), 1590–1602. https://doi.org/10.1111/evo.13972

633 Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., Xia, L., Froenicke, L., Lavelle, D. O.,

634 Truco, M.-J., Xia, R., Zhu, S., Xu, C., Xu, H., Xu, X., Cox, K., Korf, I., Meyers, B. C., & Michelmore, R.

635 W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication

636 in lettuce. *Nature Communications*, *8*(1), 14953. https://doi.org/10.1038/ncomms14953

637 Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end
638 sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–
639 4754. https://doi.org/10.1111/mec.15253

640 Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*(2),
641 155–158. https://doi.org/10.1038/s41592-019-0669-3

642 Scaglione, D., Reyes-Chin-Wo, S., Acquadro, A., Froenicke, L., Portis, E., Beitel, C., Tirone, M., Mauro, R.,
643 Lo Monaco, A., Mauromicale, G., Faccioli, P., Cattivelli, L., Rieseberg, L., Michelmore, R., &
644 Lanteri, S. (2016). The genome sequence of the outbreeding globe artichoke constructed de
645 novo incorporating a phase-aware low-pass sequencing strategy of F1 progeny. *Scientific*
646 *Reports*, *6*(1), 19427. https://doi.org/10.1038/srep19427

647 Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio
648 prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435–W439.
649 https://doi.org/10.1093/nar/gkl200

650 Tong, C., Li, H., Wang, Y., Li, X., Ou, J., Wang, D., Xu, H., Ma, C., Lang, X., Liu, G., Zhang, B., & Shi, J.
651 (2016). Construction of High-Density Linkage Maps of Populus deltoides × P. simonii Using
652 Restriction-Site Associated DNA Sequencing. *PLOS ONE*, *11*(3), e0150692.
653 https://doi.org/10.1371/journal.pone.0150692

654 Webb, D. M., & Knapp, S. J. (1990). DNA extraction from a previously recalcitrant plant genus. *Plant*
655 *Molecular Biology Reporter*, *8*(3), 180–185. https://doi.org/10.1007/BF02669514

656 Widmer, T. L., Guermache, F., Dolgovskaia, M. Y., & Reznik, S. Ya. (2007). Enhanced Growth and Seed
657 Properties in Introduced vs. Native Populations of Yellow Starthistle (Centaurea solstitialis).
658 *Weed Science*, *55*(5), 465–473.

659 Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in

660 the application of mixed-model association methods. *Nature Genetics*, *46*(2), 100–106.
661 https://doi.org/10.1038/ng.2876
662 Zavaleta, E. S., & Hulvey, K. B. (2004). Realistic Species Losses Disproportionately Reduce Grassland
663 Resistance to Biological Invaders. *Science*, *306*(5699), 1175–1177.
664 https://doi.org/10.1126/science.1102643

665 **Acknowledgements**

666 We thank B.M. Anderson, S. Lin and K. Nurkowski for help with data collection and D.
667 Kaplan for assistance with greenhouse logistics. This work was supported by funding from a
668 Natural Sciences and Engineering Research Council of Canada grant #353026 to L.H.R., United
669 States National Science Foundation grants #1750280 to KMD and #2109625 to BR.

670

671 **Author Contributions**

672 BR, KD, MB, and LR conceived the work. BR, KD, AC, QJ, and MM conducted the research
673 and analyzed the resulting data. BR and KD drafted the manuscript, and all authors provided
674 revisions and feedback.

675

676 **Tables and Figures**

677 **Tables**

678 Table 1. Statistics of the *Centaurea solstitialis* genome assembly and annotation.

Metric	Initial PacBio assembly	Final Omni-C Assembly
Contiguity		
Total length	764,896,586	765,086,668
Total number of scaffolds	2,969	1,080
N50	1,356,249	100,696,929
L50	127	4
L90	873	8

Complete BUSCOs	231 (90.98%)	237 (92.9%)
Complete and single copy BUSCOs	185 (72.55%)	184 (72.1%)
Annotation		
Total repetitive DNA	--	63.31%
Transposable elements	--	31.1%
Simple repeats	--	1.16%
Total number of genes	--	34,323
Total coding region (bp)	--	59,845,476
Average gene length (bp)		1,743

679

680 Table 2. Summary of QTL peaks associated with maximum leaf length (len) and total number of leaves
681 (num) at time points 1 and 2 (e.g. len1 and len2, respectively, or just len if the QTL is associated with the
682 trait at both time points). The identification method, location in the reference, size of the peak, favored
683 allele (i.e. TRI for invader, TK for native, and H for overdominance), dominance pattern at the locus, and
684 number of candidate genes within 1 LOD drop of the peak are all given for each peak.

PeakID	method	trait	chr	start	end	size (Mbp)	favored_allele	dominance_pattern	#candidates
1	LOCO	len_num	Chr_1	74748	88979819	88.90507	H	overdominant	3532
2	HK	len	Chr_1	106564625	106770853	0.206228	H	overdominant	16
3	LOCO	num	Chr_2	4353261	12526728	8.173467	TRI	additive	495
4	LOCO	num2	Chr_2	80999070	99091674	18.0926	TRI	additive	880
5	LOCO	num1	Chr_2	95334365	102637104	7.302739	TRI	additive	370
6	LOCO	len	Chr_4	1797495	7470976	5.673481	TK	additive	407
7	LOCO	len	Chr_4	20690837	75543191	54.85235	TK	additive	2198
8	LOCO	len1	Chr_5	60237273	68394149	8.156876	TK	complete	391
9	LOCO	len	Chr_6	75934695	81496578	5.561883	TK	complete	364
10	LOCO	len	Chr_7	6148370	9845385	3.697015	TRI	complete	282
11	HK	num2	Chr_7	19245630	74419950	55.17432	TRI	additive	2102

12	LOCO	len	Chr_8	1027923	2687378	1.659455	TK	complete	147
13	LOCO	len_num	Chr_8	10425376	16906147	6.480771	TK	complete	353

Figures

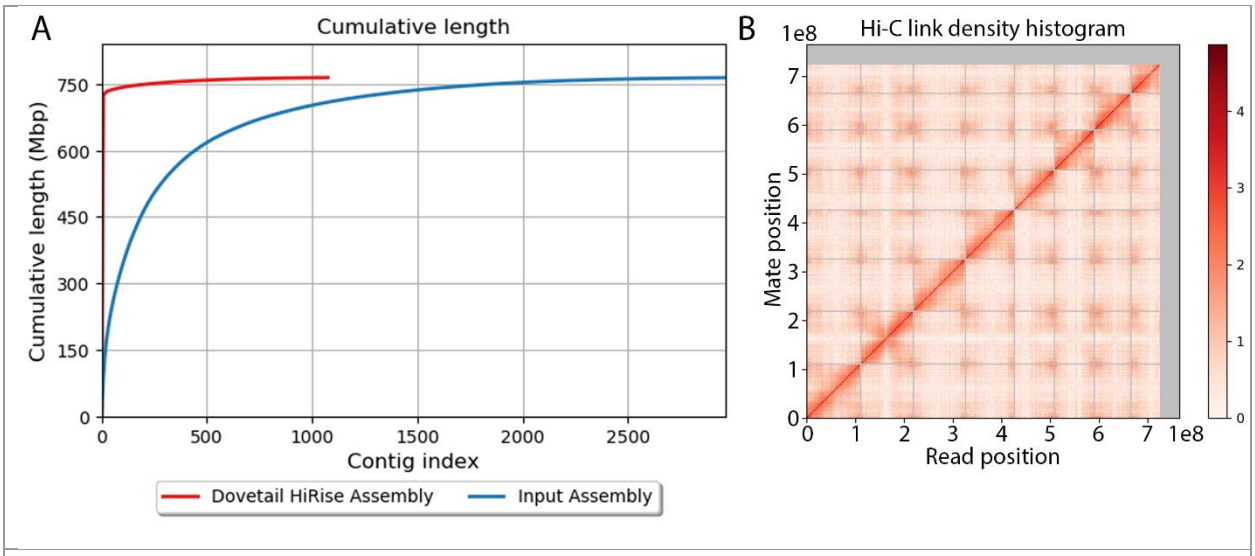


Figure 1. Scaffolding to the chromosome-level of the *C. solstitialis* genome using Omni-C data via the dovetail HiRise pipeline. A) Cumulative length of scaffolds in the final HiRise assembly versus the initial PacBio HiFi assembly. B) Link density of read pairs from the Dovetail Omni-C data mapped to the final *C. solstitialis* assembly. X and Y-axes represent the mapping positions of the first and second read in each read pair, respectively. Positions are binned, and link density within bins are represented as a heat map. Grey bars denote boundaries between scaffolds.

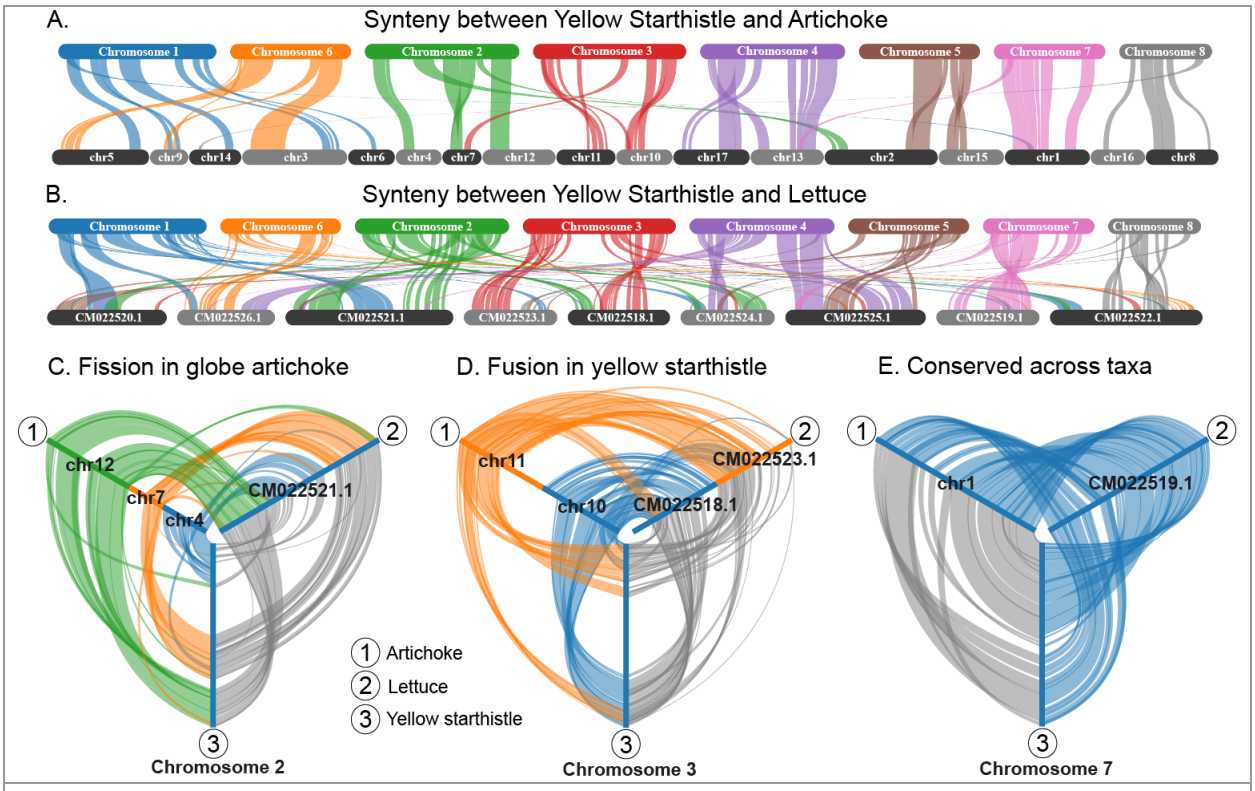


Figure 2. Gene synteny between yellow starthistle (*C. solstitialis*; n=8), globe artichoke (*Cynara cardunculus*; n=17), and lettuce (*Lactuca sativa*; n=9) genomes reveal chromosome evolution within the thistle subfamily (*Carduoideae*). A) Genome-wide synteny between yellow starthistle and artichoke displaying synteny blocks with match score of at least 2866. B) Genome-wide synteny between yellow starthistle and lettuce displaying synteny blocks with match score of at least 981. C) Putative fission of an ancestral chromosome into three separate chromosomes in globle artichoke revealed by three-way comparisons of gene synteny. D) Putative chromosome fusion between two ancestral chromosomes into one chromosome in yellow starthistle revealed by three-way comparisons of gene synteny. E) Example of a highly conserved chromosome between all three taxa revealed by three-way comparisons of gene synteny.

690

691

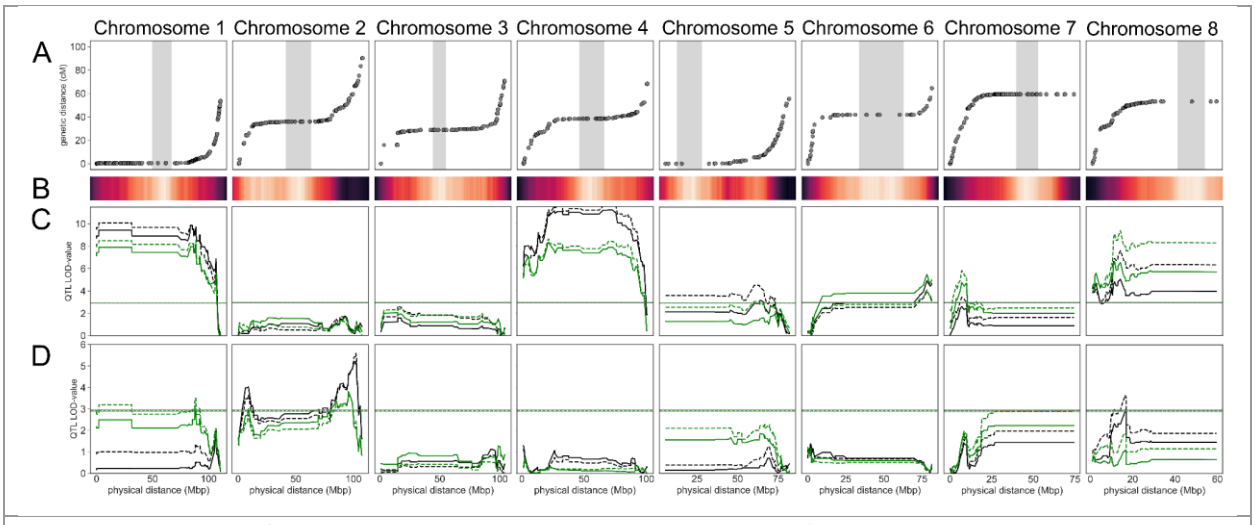


Figure 3. Summary of genome structure and QTL analyses. Panel A) Collinearity between the physical map on the x-axis and genetic map on the y-axis reveals suppressed recombination in putative pericentromeric regions (vertical grey bars) for all chromosomes and putative structural variation on the arms of chromosomes 1 and 7. Panel B) centromere genome scan score heatmap of the average gene density, unique repetitive element density, Gypsy LTR-RT density, and Copia LTR-RT density across each chromosome. Panel C) QTL LOD scores for maximum leaf length at time point 1 (black lines) and time point 2 (green lines) using the HK method (dashed lines) and LOCO method (solid lines). Panel D) QTL LOD scores for total number of leaves at time point 1 (black lines) and time point 2 (green lines) using the HK method (dashed lines) and LOCO method (solid lines). Horizontal lines represent the suggestive QTL threshold of 0.1 for each time point and method.

692

693

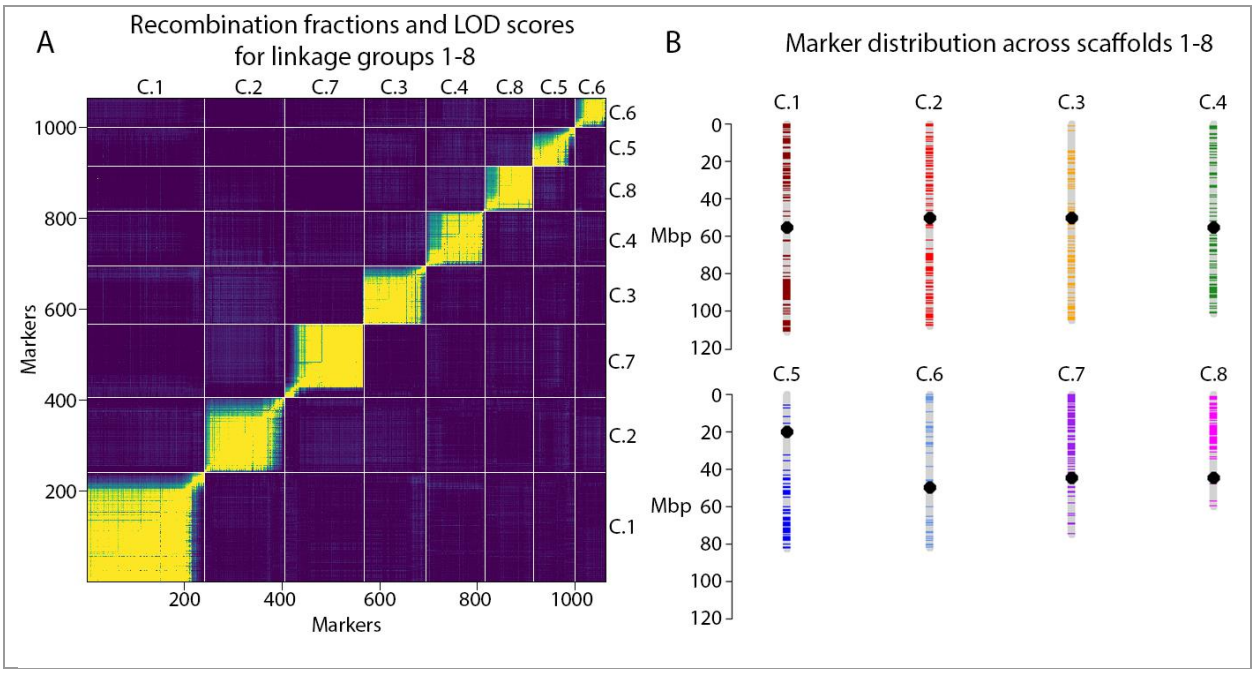
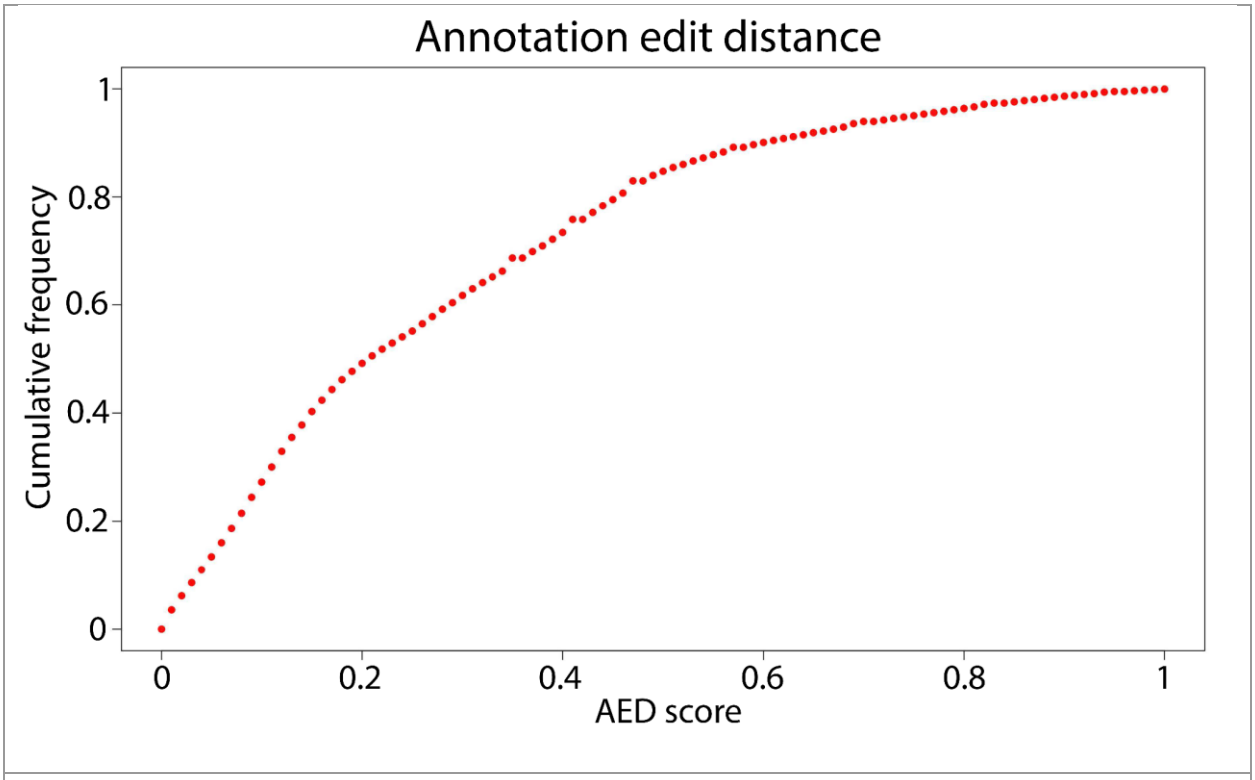


Figure 4. Visualization of the reference-aligned genetic map. A) Heatmap of recombination fractions (RF) in the top left of the diagonal and logarithm of odds (LOD) scores in the bottom right of the diagonal for ordered markers in linkage groups 1-8 reveal strong linkage and low recombination for markers within linkage groups and weak linkage with high recombination for markers on different linkage groups. Each linkage group is labeled with the corresponding chromosome number from the reference genome. B) Physical distribution of markers from the genetic map across chromosomes 1-8 of the reference genome, with each linkage group displayed as a different color. Putative centromere locations from genome scans are indicated with black dots.

714 **Supplemental Material**

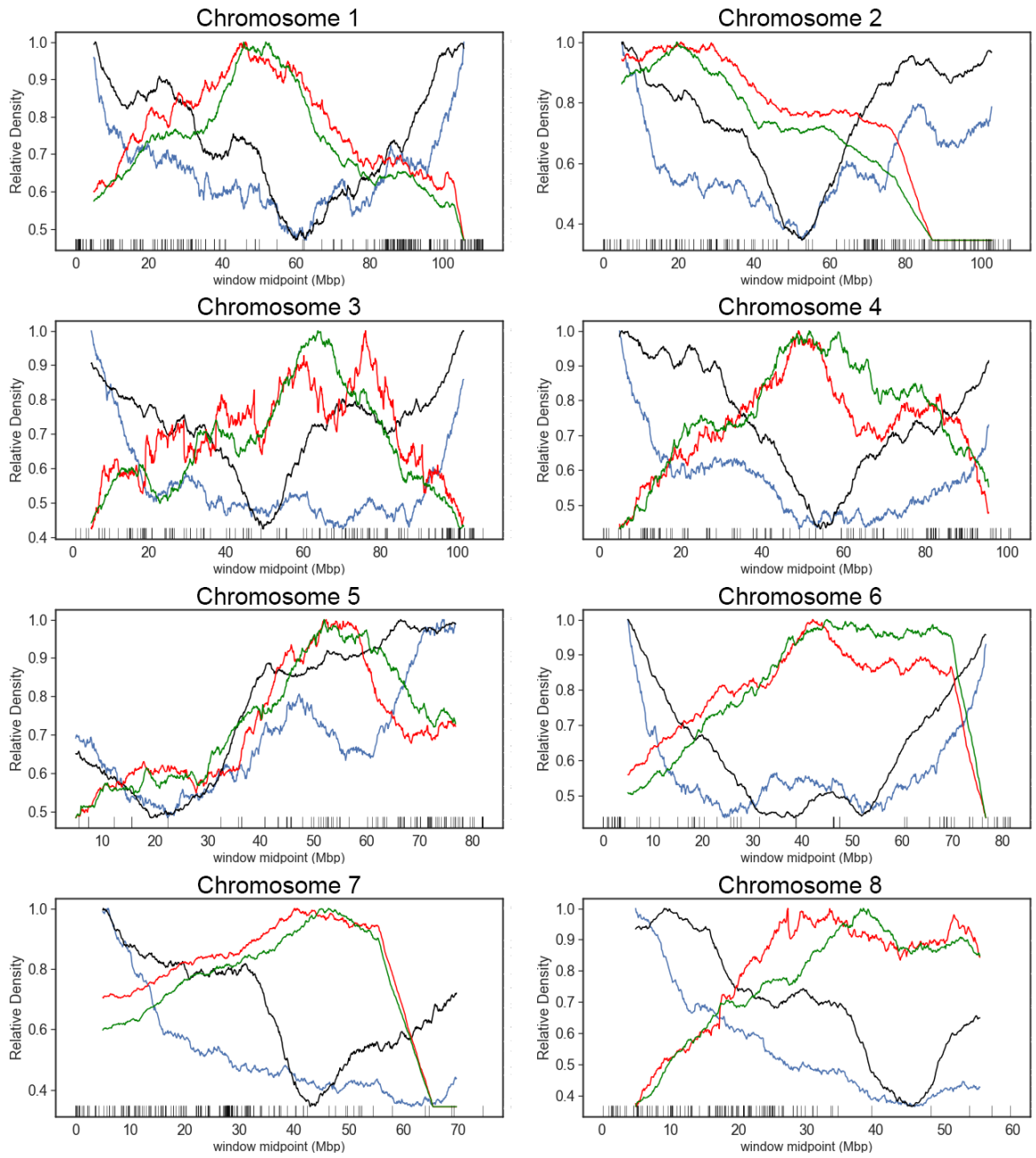
715



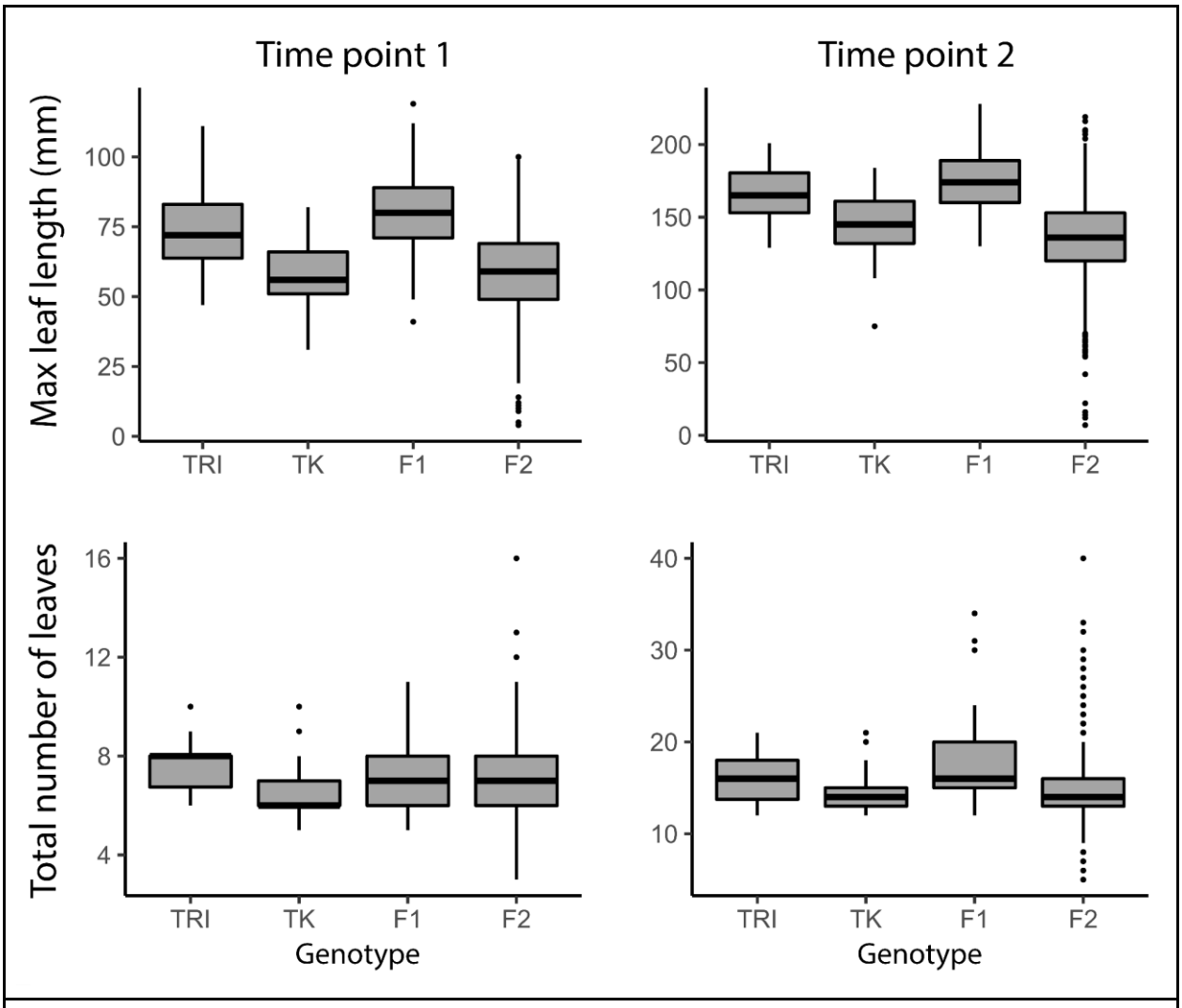
Supplemental Figure 1. Annotation Edit Distance (AED) for the *C. solstitialis* genome annotation.

716

717



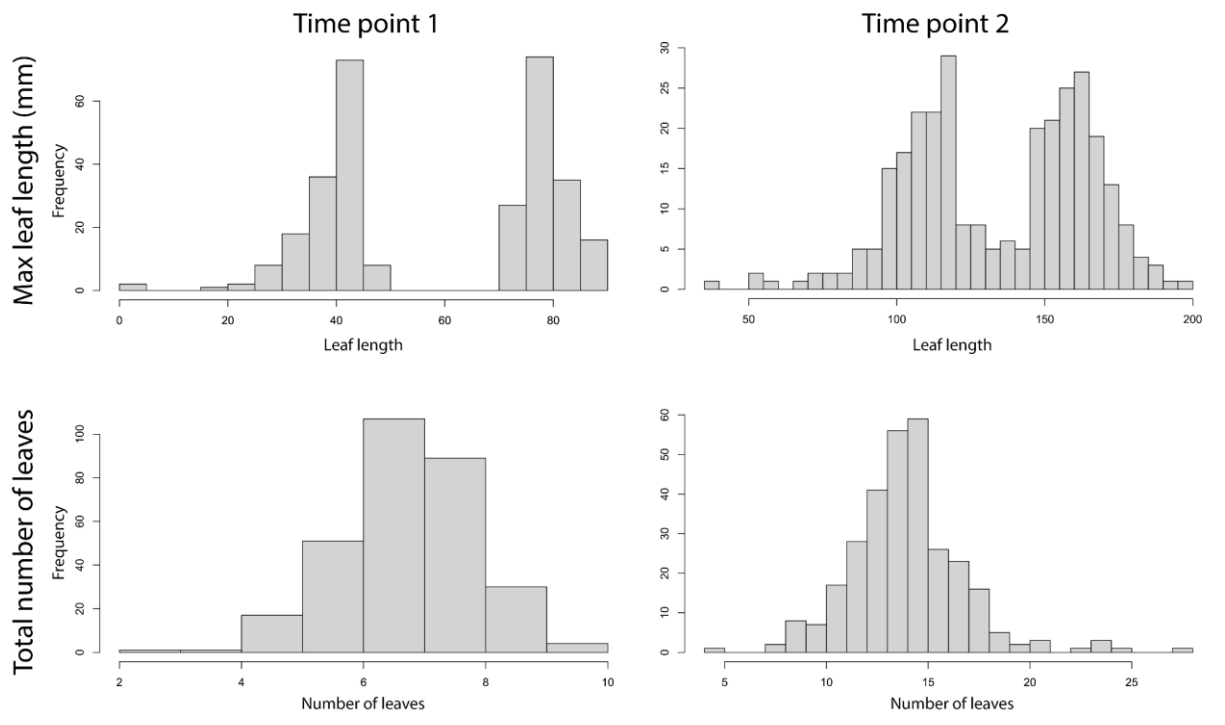
Supplemental Figure 2. Normalized density of genes, unique repetitive elements, Gypsy LTR-RT elements, and Copia LTR-RT elements across the genome.



Supplemental Figure 3. Phenotypic values for invader parental genotype (TRI), native parental genotype (TK), F1, and F2 genotypes of the mapping population.

720

721



Supplemental Figure 4. Distribution of corrected values for maximum leaf length and total number of leaves at both time points for the 300 genotyped F2s used as input for QTL analysis.