

1 **Title: Legume genome structures and histories inferred from *Cercis canadensis* and**  
2 ***Chamaecrista fasciculata* genomes**

3  
4 **Authors:**

5 Hyun-oh Lee<sup>1</sup>, Jacob S Stai<sup>1</sup>, Qiaoji Xu<sup>2</sup>, Thulani Hewavithana<sup>3</sup>, Rabnoor Batra<sup>3</sup>, Alex Liu<sup>4</sup>,  
6 Brandon D Jordan<sup>5</sup>, Rachel Walstead<sup>6</sup>, Jerry Jenkins<sup>6</sup>, Melissa Williams<sup>6</sup>, Jenell Webber<sup>6</sup>, Jane  
7 Grimwood<sup>6</sup>, John T Lovell<sup>6,7</sup>, Tomáš Brůna<sup>7</sup>, Shengqiang Shu<sup>7</sup>, Keykhosrow Keymanesh<sup>7</sup>,  
8 Joanne Eichenberger<sup>7</sup>, Jeremy Schmutz<sup>6,7</sup>, David M Goodstein<sup>7</sup>, Kerrie Barry<sup>7</sup>, David Sankoff<sup>2</sup>,  
9 Lingling Jin<sup>3</sup>, James H Leebens-Mack<sup>8</sup>, Steven B Cannon<sup>5</sup>

10

11 **Affiliations:**

12 <sup>1</sup>ORISE Fellow, USDA-ARS, Corn Insects and Crop Genetics Research Unit, 819 Wallace Rd, Ames, IA 50011,  
13 USA

14 <sup>2</sup>Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

15 <sup>3</sup>Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5C9, Canada

16 <sup>4</sup>School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

17 <sup>5</sup>USDA-ARS, Corn Insects and Crop Genetics Research Unit, 819 Wallace Rd, Ames, Iowa, USA

18 <sup>7</sup>US Department of Energy Joint Genome Institute, Berkeley, California 94720, USA

19 <sup>6</sup>Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA

20 <sup>8</sup>Department of Biology, University of Georgia, Athens, Georgia 30602, USA

21

22 \*Corresponding authors: [steven.cannon@usda.gov](mailto:steven.cannon@usda.gov), [jleebensmack@uga.edu](mailto:jleebensmack@uga.edu),

23 [lingling.jin@cs.usask.ca](mailto:lingling.jin@cs.usask.ca)

24

25

26 **Summary**

- 27       • The legume family originated ca. 70 million years ago and soon diversified into at least  
28       six lineages (now extant subfamilies). The signal of whole genome duplications (WGD)  
29       is apparent in species sampled from all six subfamilies. The early diversification has  
30       posed difficulties for resolving the legume backbone structure and the timing of WGDs.  
31       • In this study, we report the genome sequences and annotations for *Cercis canadensis*  
32       (Cercidoideae) and *Chamaecrista fasciculata* (Caesalpinoideae) to help resolve the  
33       relative taxonomic placements along the legume backbone, the timings of WGDs relative  
34       to subfamily origins, and the ancestral legume karyotype.  
35       • Analyses of genome assemblies from four subfamilies within Fabaceae show that the last  
36       common ancestor of all legumes likely had seven chromosomes, with a genome structure  
37       similar to the extant *Cercis* genome. Our analysis supports an allopolyploid origin of the  
38       subfamily Caesalpinoideae, with progenitors involving lineages along the backbone of  
39       the legume phylogeny.  
40       • A probable allopolyploid origin of Caesalpinoideae subfamily provides a partial  
41       explanation for the difficulty in resolving the structure of the legume backbone. The  
42       retained karyotype structure and lack of a WGD in the last 100+ Mya, underscore the  
43       utility of the *Cercis* genome as an ancestral reference for the legume family.

44

45 **Keywords:**

46 allopolyploidy, Caesalpinoideae, Cercidoideae, *Cercis canadensis* (redbud), *Chamaecrista*  
47 *fasciculata* (partridge pea), Fabaceae, legumes, whole genome duplication

48

49

50

## 51 **Introduction**

52 The legume family (Leguminosae or Fabaceae) is the third largest family of flowering plants,  
53 comprising about 770 genera and 19,500 species (Lewis *et al.*, 2013; Azani *et al.*, 2017). Species  
54 such as peanuts, cowpeas, soybeans, and fava beans are an important source of protein for a large  
55 proportion of the global human population, and play an important agricultural role in fixing  
56 nitrogen (N<sub>2</sub>) from the atmosphere and converting it to reduced forms that are used by plants to  
57 produce amino acids and other biomolecules (Martins *et al.*, 2003; Chu *et al.*, 2004; Herridge *et*  
58 *al.*, 2008; Salvagiotti *et al.*, 2009; Köpke & Nemecek, 2010).

59

60 The legume family originated within the Rosid II angiosperm clade toward the end of the  
61 Mesozoic era, around 70 Million Years Ago (MYA) (Li *et al.*, 2019). Within a span of roughly  
62 15 MY from its origin, the family had radiated giving rise to six lineages that are recognized as  
63 subfamilies: Cercidoideae, Detarioideae, Dialioideae, Duparquetioideae, Caesalpinioideae, and  
64 Papilionoideae (Azani *et al.*, 2017). Most agronomically important species (and approximately  
65 two thirds of species in the family) fall within the Papilionoid clade, but the other subfamilies  
66 contain many species of great importance in term of ecological service and economic value,  
67 including timber species (e.g. *Gleditsia*, *Robinia*), forage (e.g. *Acacia*, *Caesalpinia*), and human  
68 consumption (e.g. *Tamarindus* [tamarind] *Detarium* [sweet detar], *Ceratonia* [carob], *Tylosema*  
69 [marama bean]) (Cannon *et al.*, 2011).

70

71 Rapid bursts of diversification within the family have made resolution of the legume phylogeny  
72 difficult. Further complicating the understanding of the genomic evolution in the family is the  
73 presence of multiple, apparently independent whole genome duplications within most  
74 subfamilies (Cannon *et al.*, 2015; Zhao *et al.*, 2021). Chromosome numbers in the family vary  
75 widely, though there are clear modal counts for each subfamily: n=12 for Detarioideae, 14 for  
76 Cercidoideae, 14 for Dialoideae, and 14 for Caesalpinioideae. In Papilionoideae, the range of  
77 chromosome numbers is broad, but the majority of species in this subfamily have counts of n=7-  
78 11 (Ren *et al.*, 2019).

79

80 Here, we describe high-quality genome assemblies and annotations for *Cercis canadensis* L. and  
81 *Chamaecrista fasciculata* (Michx.) Greene. These species represent two of the six generally-

82 recognized legume subfamilies, and thus are well placed for helping to infer early events in the  
83 evolution of the legumes. We test the extent to which the *C. canadensis* genome, representing an  
84 exceptional legume lineage that has not experienced a WGD over its evolutionary history  
85 subsequent to the ~135 Mya gamma triplication (Jiao *et al.*, 2012), retains pre-legume ancestral  
86 genome structure including the ancestral monoploid karyotype count for all members of the  
87 family.

88

89 For analyses of chromosome structural and gene family evolution, we make comparisons among  
90 six sequenced legume genomes that represent the four largest legume subfamilies; and also  
91 against four nonlegume outgroup species. Those outgroup species, in order from youngest to  
92 oldest shared ancestry with the legumes, are *Quillaja saponaria* (Quillajaceae, the closest family  
93 to the legumes), *Prunus persica* (Rosaceae, in the Fabideae with the legumes), *Arabidopsis*  
94 *thaliana* (Malvaceae, sister to the Fabideae and within the Rosid clade), and *Vitis vinifera*,  
95 (Vitaceae, Vitales, sister to clade with all other orders in the Rosid clade).

96

97 Of the newly sequenced genomes reported here, *C. canadensis*, also known as the eastern  
98 redbud, is a deciduous ornamental tree native to eastern North America. The *Cercis* genus is  
99 eponymous for the Cercidoideae subfamily. There are approximately 10 species within the genus  
100 - three native to North America, six native to China and south-central Asia, and one native to the  
101 Mediterranean region (Davis *et al.*, 2002). We compare the assembly of *C. canadensis* to a  
102 genome assembly of *C. chinensis* (Li *et al.*, 2023), and evaluate the *C. canadensis* assembly and  
103 annotations in the context of other legume species.

104

105 *Chamaecrista fasciculata*, commonly known as partridge pea, is in the Caesalpinioideae legume  
106 subfamily. *C. fasciculata* is an annual plant common in prairies of eastern and central North  
107 America (Fenster, 1991; Govaerts *et al.*, 2021). *C. fasciculata* exhibits symbiotic nitrogen  
108 fixation (SNF), hosting nitrogen-fixing Bradyrhizobium bacteria in specialized structures  
109 (nodules) on roots. *C. fasciculata* has been used as a model for research on the ecology,  
110 physiology, and evolution of symbiotic nitrogen-fixation (Singer *et al.*, 2009). SNF is found in  
111 most genera in the Papilionoideae and in several clades within the Caesalpinioideae, but in none  
112 of the other legume subfamilies. SNF presence and absence among legume subfamilies has

113 informed inference of SNF evolution with the Rosid Nitrogen Fixing Clade (Sprent *et al.*, 2017;  
114 Griesmann *et al.*, 2018; Zhao *et al.*, 2021).

115

## 116 **Materials and Methods**

### 117 ***Cercis canadensis* and *Chamaecrista fasciculata* genome assembly and annotation**

118 Genome assembly and annotation methods for *Cercis canadensis* and *Chamaecrista fasciculata*  
119 are described in Supporting Information Methods S1.

120

### 121 **Phylogenomic analyses**

122 For phylogenomic analyses (Figs 4, 5), legume-focused gene families were first constructed  
123 using the Pandagma gene family workflow (<https://github.com/legumeinfo/pandagma>; Cannon *et al.*, 2024), using the CDS and protein sequences of 36 legume species in 21 genera (Table S1)  
124 and four nonlegume outgroup genera. Inputs for the initial base families included 15 individual  
125 legume species (Table S1) and exemplar sequences for pangene sets for six legume genera for  
126 which multiple annotations and species are available (*Arachis*, *Cicer*, *Glycine*, *Medicago*,  
127 *Phaseolus*, and *Vigna*). For example, the *Vigna* pangene set was calculated based on seven  
128 accessions of *V. unguiculata*, two of *V. radiata*, and one of *V. angularis*. The non-legume  
129 outgroup species were *Quillaja saponaria*, *Prunus persica*, *Arabidopsis thaliana*, and *Vitis*  
130 *vinifera*. There were 39,981 gene families generated using these inputs, and 25,271 families  
131 containing at least 4 sequences and at least 2 distinct genera. These base gene families are  
132 available at

133 <https://data.legumeinfo.org/LEGUMES/Fabaceae/genefamilies/legume.fam3.VLMQ/>.

134

135  
136 Using on those initial “base” families, the 15 annotation sets described in this manuscript (11  
137 legume genera and 4 outgroup genera) were placed into the base gene families above by  
138 homology, using the “pandagma fsup” workflow, parameterized with identity  $\geq 30\%$  and  
139 coverage  $\geq 40\%$ . Protein sequences from each family were aligned using famsa (Deorowicz *et al.*,  
140 2016). The alignments were modeled using hmmbuild from the hmmer package (Finn *et al.*,  
141 2011) to generate hidden Markov models (HMMs). The original gene family sequences were  
142 then realigned to the HMM for each family, and then trimmed to the match-states of the HMM.

143 Phylogenetic trees were then calculated using FastTree (v. 2.1)(Price *et al.*, 2010). Selected gene  
144 trees (e.g. Figs. 4b-e) were also calculated using RAxML, with 1000 bootstraps.

145  
146 To calculate a consensus gene phylogeny (Fig. 4a), 50 gene families were selected at random  
147 from “complete” families -i.e., families containing the expected number of genes from each  
148 species under the assumption of retention following known WGDs and no additional WGDs.  
149 Homoeologous genes from ancestrally tetraploid species in each of the 50 selected phylogenies  
150 were labeled A or B based on their position in the gene trees; for example, *Medicago.A* and  
151 *Phaseolus.A* were identified in one clade and *Medicago.B* and *Phaseolus.B* in another clade. The  
152 A and B labeling was applied top to bottom in rooted trees that had been ordered relative to the  
153 outgroup, by the Order function in the Archaeopteryx tree viewer (Han & Zmasek, 2009). This  
154 places sister clades with more nodes above those with fewer nodes, which generally results in the  
155 better-represented Papilionoid clade placed at the top, and Cercidoid clade near the bottom. The  
156 effect should otherwise be neutral with respect to other aspects of clade arrangements. Given  
157 these labeled gene families, a supermatrix alignment was generated by concatenating alignments  
158 from all 50 gene families, with genes and paralogs placed in consistent order: *Lotus.A* first,  
159 *Medicago.A* second, etc. The alignment was sampled at modulo 5 (taking every fifth amino acid)  
160 to make phylogenetic calculations tractable. The resulting alignment matrix had 9483 sites. The  
161 consensus phylogeny was then calculated using RAxML-NG (Kozlov *et al.*, 2019), using the  
162 `raxml-ng --all, with model LG+G8+F.`

163

#### 164 **Synonymous-site (Ks) calculations and divergence time estimation**

165 The Ks values (Fig. 5) were calculated as part of the pandagma gene family workflow (Cannon  
166 *et al.*, 2024). For a given species pair, gene pairs were identified first based on homology using  
167 `mmseqs2` (Steinegger & Söding, 2017), then filtered based on inclusion in synteny blocks  
168 identified by DAGChainer (Haas *et al.*, 2004). Given those gene pairs, Ks values were calculated  
169 using PAML (Yang, 2007). A modal value was then calculated for all genes in a synteny block.  
170 The modal values for the genes in the block were applied to all genes in that block, which were  
171 in turn used to calculate genome-wide Ks-value frequency distributions (Fig. 5). Modal Ks  
172 values for each species pair (both speciation and WGD peaks) were used as parameters in a  
173 system of equations to solve for each branch length in the consensus gene tree described above

174 (Fig. 5d). Divergence times for the tree in Fig. 6a were approximated on the consensus tree using  
175 calibrations from TimeTree database (<http://TimeTree.org>)(Kumar *et al.*, 2017). The median  
176 value used for *Vitis* and *Medicago* was taken as 117 Mya (Wikström *et al.*, 2001).

177

### 178 **Analysis of gene family expansions and contractions and gene ontology enrichment**

179 Homologous gene clustering was performed with the OrthoFinder (Emms & Kelly, 2019)  
180 clustering algorithm and default options (e-value 1e-2, inflation value 1.5) by the Orthovenn3  
181 web server (Sun *et al.*, 2023). Gene family contraction and expansion analysis was performed  
182 using CAFE5 (Mendes *et al.*, 2020). Gene ontology (GO) terms for biological process, molecular  
183 function, and cellular component categories and enrichment of the expanded and contracted  
184 genes were then obtained. The enriched horizontal bar plot was drawn by using SRplot (Tang *et*  
185 *al.*, 2023).

186

### 187 **Analysis of duplicated subgenomes resulting from polyploid events**

188 *Chamaecriasta*, *Senna*, and *Bauhinia* all exhibit duplicated syntenic regions compared to *Cercis*  
189 due to their independent WGD events after diverging from *Cercis*. The SyntenyLink algorithm  
190 (Hewavithana *et al.*, 2023) was used to identify two subgenomes derived from ancient  
191 allotetraploidy events in the ancestry of *Chamaecriasta*, *Senna*, and *Bauhinia*. SyntenyLink  
192 assesses differences in substitution and fractionation patterns in synteny blocks as  
193 homoeologous gene duplicates (syntelogs) diverge. The results are visualized using SynVisio  
194 (Bandi & Gutwin, 2020) showing syntenic blocks linking regions of the *Cercis* genome to  
195 homoeologous regions of the in *Bauhinia* (Fig. 1b), *Chamaecriasta* (Fig. 1c), and *Senna* (Fig.  
196 1d) genomes.

197

### 198 **Ancestral genome reconstructions and rearrangement distances support seven** 199 **protochromosomes**

200 We used the RACCROCHE (Xu *et al.*, 2020) procedure to infer gene content and order for  
201 hypothesized ancestral chromosomes in the genomes in a phylogeny. First, we identify adjacent  
202 genes (Yang & Sankoff, 2011), allowing up to a specified number of spacer genes between genes  
203 scored as adjacent, across the chromosomes in the input genomes. Considering the divergence  
204 times among our focal genomes, we allow 2 spacer genes to score gene adjacencies. Then, for

205 each ancestral node in the species phylogeny, we infer adjacencies by generating graphs with all  
206 phylogenetically informative adjacencies. In the graph, vertices are adjacencies, and edges join  
207 any two adjacencies that each contain one of the 5' and 3' ends of the same gene. The graph is  
208 analyzed using the Maximum Weight Matching algorithm, which produces linear ancestral  
209 “contigs” as output, with each contig containing a collection of genes found in proximity in more  
210 than one of the input genomes. To avoid biases due to widely disparate contig lengths, we cut  
211 each contig into “20-mers” of length at most 20 genes, creating a larger set of more comparable  
212 contigs.

213

214 To group the reduced contigs (20-mers) into collections representing inferred ancestral  
215 chromosomes, we match them against the chromosomes of the input genomes and count the  
216 number of times any two contigs match the same chromosome. Contig ordering is taken into  
217 account, and multiple matches within a genome are permitted (to permit modeling of WGDs).  
218 This scoring produces a co-occurrence matrix, which is then clustered using a complete-linkage  
219 clustering of the contigs. The output can be interpreted as the inferred gene content of each  
220 ancestral chromosome. The reconstructed ancestors each contain at most one member of each  
221 gene family from ortholog groups calculated from the descendant genomes. We then performed  
222 “g $\square$ mer” analysis (Xu *et al.*, 2023) to estimate the ancestral number of chromosomes,  $x$ .

223

## 224 **DCJ distance inference of species relationships**

225 The Double Cut and Join (DCJ) distance (Yancopoulos *et al.*, 2005) is used to quantify the  
226 structural differences between two genomes. Smaller DCJ values indicate fewer rearrangements  
227 between the two genomes. We calculated the total number of DCJ distances between all pairs of  
228 ancestors using the UniMOG tool (Hilker *et al.*, 2012).

229

## 230 **Results**

### 231 **Genome assembly and annotation assessment**

232 The genomes of both *C. canadensis* and *C. fasciculata* were initially assembled based on PacBio  
233 ccs data, and then the contigs were oriented and ordered using Hi-C data. Heterozygous snp/indel  
234 phasing errors were corrected using PacBio and Illumina data. The 99.46-99.58% of the  
235 assembled sequences were assigned to chromosomes. Finally, *C. canadensis* was assembled with



236 7 pseudochromosomes and *C. fasciculata* with 8 pseudochromosomes. Two haplotype  
237 assemblies and annotations were resolved for each species. The final chromosome-level  
238 assemblies for *C. canadensis* and *C. fasciculata*, exhibited length variation between haplotypes  
239 (Table 1). To assess the completeness of the genome, we performed BUSCO (v. 5.4.5) analysis  
240 in gene mode using the fabales\_odb10 dataset and found completeness percentages of 96.4 and  
241 93.6% (single copy percentages of 92.3 and 79.1%). The missing rates were 3.4% and 6.1%  
242 (Table 1). The relatively high “missing” BUSCO rates are likely due to the fact that  
243 fabales\_odb10 consisted of only 10 species in the Papilionidae family, and when we expanded  
244 the BUSCO DB to eudicots\_odb10, *C. canadensis* and *C. fasciculata* were found to have 99.4%  
245 and 99.5% complete BUSCOs, respectively.

246

#### 247 **Comparisons of the *Cercis canadensis* and *Chamaecrista fasciculata* genomes with resources** 248 **from related species**

249 The *C. canadensis* genome assembly described here is the second near-complete assembly  
250 published from this genus. Comparisons with the assembly for *C. chinensis* (Li *et al.*, 2023)  
251 show the two assemblies to be similar in size and structure, but with large rearrangements on  
252 chromosomes 3 and 5 (Fig. S2), and having an average identity in alignable regions of only  
253 93.76%. The assembly sizes for the two species are similar: 352.8 and 342.0 Mbp for the total  
254 assembly sizes in *C. chinensis* and *C. canadensis*; and 331.8 and 340.3 Mbp for the  
255 chromosomally anchored sequences in *C. chinensis* and *C. canadensis*.

256

257 The *C. fasciculata* genome assembly described here is the second assembly of this species – the  
258 first being a contig-level assembly (Griesmann *et al.*, 2018), for isolate NF-2018-5 (derived from  
259 line MN87), GenBank accession GCA\_003254925.1, with scaffold N50 of 56.6 kb and total  
260 assembly length of 429.1 Mb. The chromosome-scale, haplotype-resolved assembly described  
261 here, of isolate ISC494698, has scaffold N50 of 757.1 kb and total assembly length of 580.4 Mb.

262

263

264 **Table 1.** Genome and gene statistics for *Cercis canadensis* and *Chamaecrista fasciculata*

265

Species	<i>Cercis canadensis</i>		<i>Chamaecrista fasciculata</i>	
<b>Haplotype</b>	1	2	1	2
<b>Pseudochromosome number</b>	7	7	8	8
<b>Total scaffold length (bp)</b>	342,014,377	315,499,427	580,459,023	564,372,522
<b>No. of scaffolds</b>	43	7	24	17
<b>No. of contigs</b>	56	20	110	107
<b>N50 of scaffolds (bp)</b>	48,286,772	43,568,431	75,709,764	71,910,870
<b>N50 of contigs (bp)</b>	27,822,107	26,270,282	11,464,191	9,553,364
<b>GC Ratio (%)</b>	36.31	35.75	35.35	35.22
<b>No. of gene</b>	27,440	26,713	29,074	28,859
<b>No. of cds</b>	51,848	51,022	49,343	49,167
<b>No. of exon</b>	360,595	359,144	329,134	329,851
<b>Avg. exons per cds</b>	6.4	6.4	6.2	6.2
<b>Avg. gene length (bp)</b>	4,047	4,094	4,257	4,281
<b>Avg. cds length (bp)</b>	1,546	1,553	1,465	1,471
<b>Avg. exon length (bp)</b>	324	322	315	314
<b>Total gene length (bp)</b>	111,055,993	109,386,083	123,788,266	123,559,672
<b>Total cds length (bp)</b>	80,207,802	79,275,015	72,295,815	72,355,461
<b>Longest cds (bp)</b>	16,539	16,539	16,455	16,461
<b>Shortest cds (bp)</b>	93	75	93	96
<b>Complete BUSCOs (C)</b>	5175 / 96.4%	5106 / 95.2%	5023 / 93.6%	5018 / 93.5%
<b>Complete and single-copy BUSCOs (S)</b>	4954 / 92.3%	4943 / 92.1%	4242 / 79.1%	4243 / 79.1%
<b>Complete and duplicated BUSCOs (D)</b>	221 / 4.1%	163 / 3.0%	781 / 14.6%	775 / 14.4%
<b>Fragmented BUSCOs (F)</b>	10 / 0.2%	23 / 0.4%	15 / 0.3%	17 / 0.3%
<b>Missing BUSCOs (M)</b>	181 / 3.4%	237 / 4.4%	328 / 6.1%	331 / 6.2%
<b>Total BUSCO groups searched</b>	5,366	5366	5,366	5,366

266

267

268 **Synteny relationships show independent WGDs early in four legume subfamilies, but**  
 269 **excluding *Cercis***

270 Synteny plots (Figs 1, 2) show a general 1::2 pattern of chromosomal duplication between *Cercis*

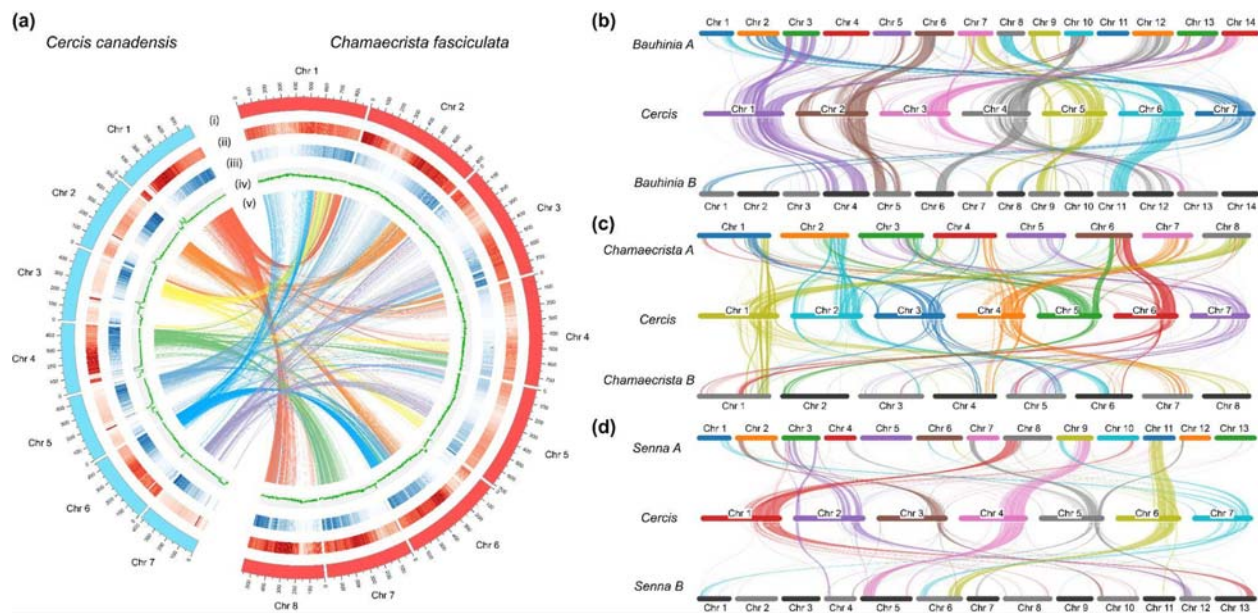
271 and other legume and close outgroup species. This pattern can be seen, for example, in *Cercis*

272 chromosome 1 matching *Bauhinia* chromosomes 3 and 4 (Fig. 1b), *Chamaecrista* chromosomes  
 273 1 and 8, (Figs 1a, c), and *Senna* chromosomes 8 and 13 (Fig. 1c). These patterns are consistent  
 274 with WGDs occurring independently in the two subfamilies (*Cercis* and *Bauhinia* in the  
 275 Cercidoideae subfamily, and *Chamaecrista* and *Senna* in the Caesalpinioideae subfamily), but  
 276 after divergence of *Cercis* from the remaining species in the Cercidoideae (though see discussion  
 277 below regarding inference of allopolyploidy in both subfamilies).

278  
 279 Comparisons of *Cercis* to itself (Fig. S1) identifies only small synteny blocks and no evidence of  
 280 a recent whole genome duplication. The fragmentary duplications in the *Cercis* self-comparison,  
 281 together with *Ks* results (presented below) are consistent with a WGD in the timeframe of the  
 282 ~135 Mya gamma triplication (Jiao *et al.*, 2012).

283

284



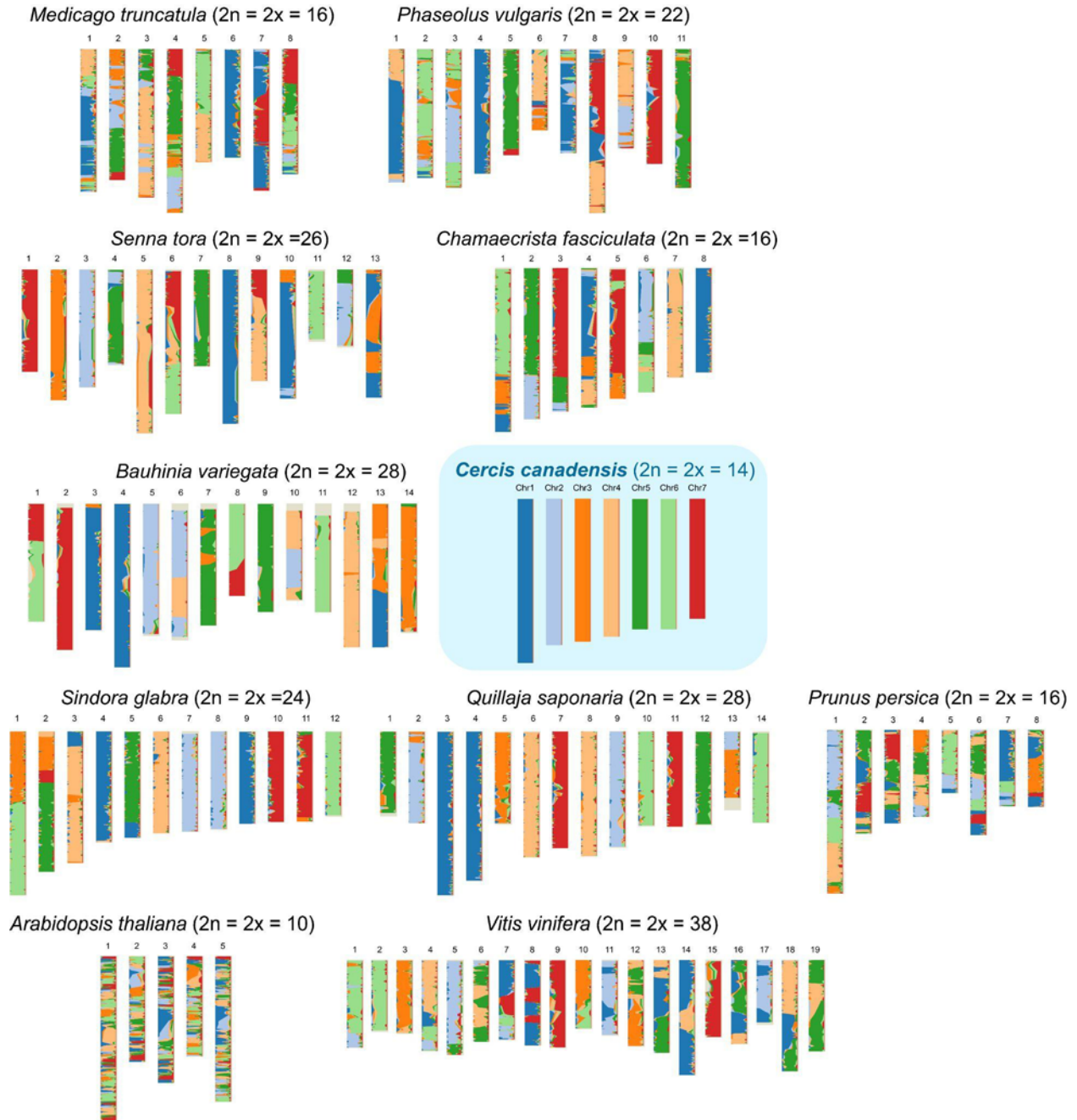
285

286 **Fig. 1.** Plots of genomic features and syntenic relationships.

287 (a) Circos plots representing features of the *C. canadensis* and *C. fasciculata* genomes. Circos track (i):  
 288 chromosome length (Mbp); track (ii): repeat density, track (iii): gene density, track (iv): GC plot, inner  
 289 connecting lines (v): synteny regions, identified with MCScanX. (b-d) Synteny plots shown in  
 290 SyntenyLink (Bandi & Gutwin, 2020; Hewavithana *et al.*, 2023) images, showing two subgenomes in  
 291 *Bauhinia* (b), *Chamaecrista* (c), and *Senna* (d) relative to *Cercis* at the center of each panel and the  
 292 least/most fractionated subgenome above/beneath *Cercis*.

293

294



295

296 **Fig. 2.** Synteny map of *Cercis canadensis* and comparison against legume and outgroup species.

297 Chromosome-level synteny maps were generated with the PanSyn program using *Cercis* as a reference

298 relative to each indicated comparison species.

299

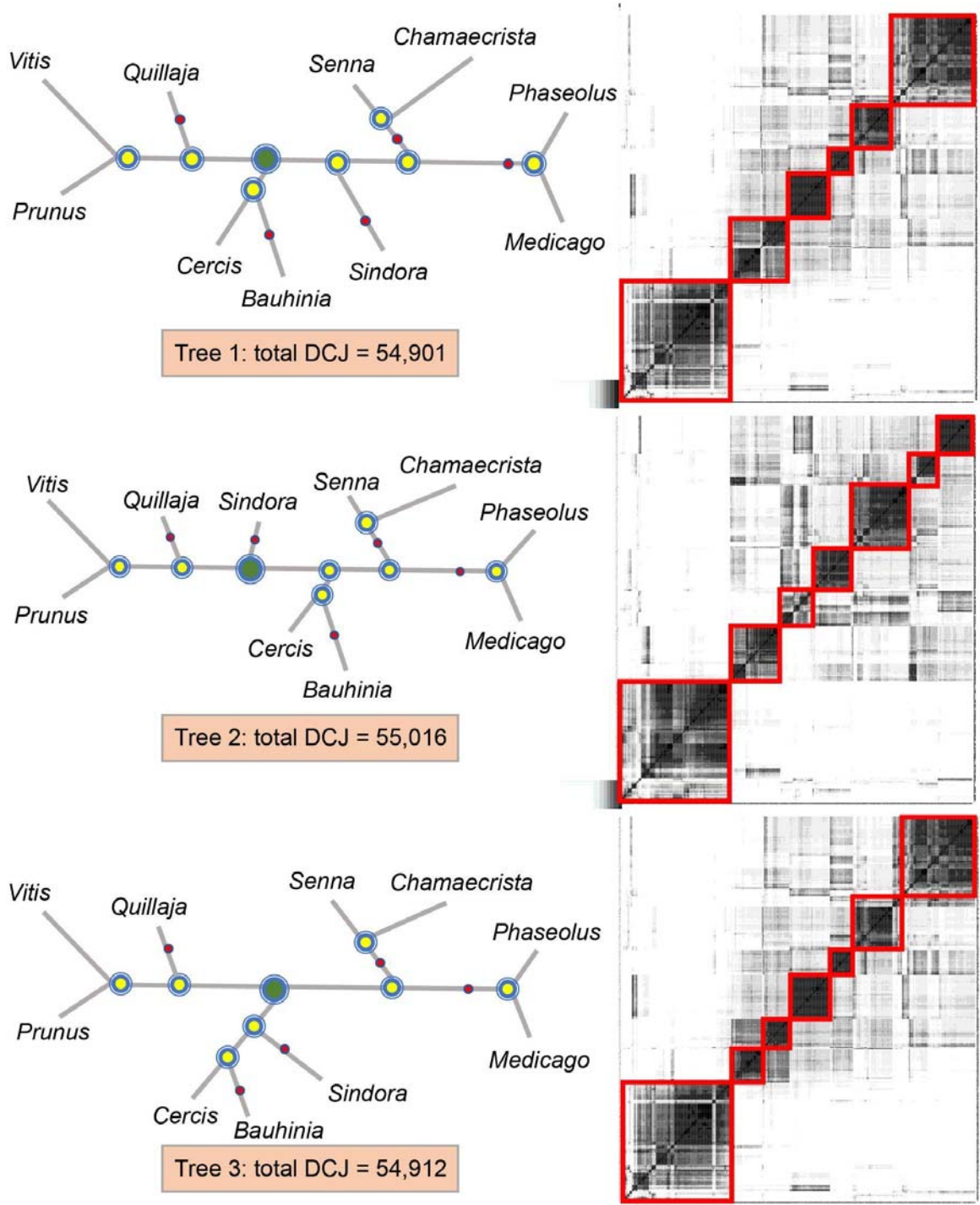
300 **Inference of a seven-chromosome legume progenitor with genome structure similar to that**  
301 **of *Cercis***

302 Synteny and shared-contig analyses suggest that the *Cercis* genome structure is similar to the  
303 ancestral legume karyotype, which we infer was also likely to have had seven chromosomes.  
304 The *Quillaja*, *Sindora*, *Bauhinia*, and *Senna* chromosome structures can all be represented in  
305 terms of doublings of the seven-chromosome *Cercis* genome, with a small number of  
306 rearrangements in each lineage. In *Quillaja* (a close outgroup to the legumes), the chromosomal  
307 correspondence with *Cercis* is a simple 1::2 match, with *Cercis* chromosomes generally  
308 matching two *Quillaja* chromosomes – often with chromosomal-scale synteny across both  
309 species. The taxa requiring the fewest rearrangements relative to a simple doubling of *Cercis*  
310 ( $1n=7$ ) are *Quillaja* ( $1n=14$ ), requiring approximately 3 rearrangements; *Sindora* ( $1n=12$ ),  
311 requiring approximately 6 rearrangements; and *Senna* ( $1n=13$ ), requiring approximately 10  
312 rearrangements. *Medicago* ( $1n=8$ ) and *Phaseolus* ( $1n=11$ ) show more complex restructuring -  
313 which would be consistent with the reduced chromosome counts from a hypothesized  
314 Papilionoid progenitor with  $1n=14$  chromosomes.

315  
316 Analysis of contig co-occurrence among reconstructed ancestral genomes (Fig. 3) supports an  
317 ancestral legume karyotype with 7 chromosomes, with little rearrangement relative to current  
318 *Cercis* chromosomes. In the heat maps in Fig. 3, syntenic contigs identified across the 10  
319 indicated species (7 legumes and 3 outgroups) are clustered by proximity, following three  
320 hypothetical phylogenetic topologies. The clusters indicate probable ancestral chromosomes, as  
321 they show syntenic contig groups that are found in proximity, at the indicated phylogenetic node,  
322 as assessed within the included species and ancestors. The heat maps correspond with the legume  
323 ancestors, supporting the estimated 7-chromosome karyotype.

324  
325 Fig. 3 shows seven clusters of the inferred legume ancestor formed from the co-occurrence  
326 matrix. Given the hypothesized contig content of each chromosome, the contigs are ordered on a  
327 chromosome using a Linear Ordering Problem routine. We find that the monoploid number of  
328 the legume family is likely to be  $x = 7$  (Fig. 3), and that Tree 1 is the most parsimonious  
329 hypothetical phylogeny.

330



331

332 **Fig. 3.** Hypothetical phylogenetic relationships of 10 selected legume and dicot outgroup species, and  
 333 heat maps of contig adjacencies representing inferred ancestral chromosome content. Left: three  
 334 hypothesized phylogenetic relationships. The legume ancestor node is highlighted in green while the other  
 335 ancestors are in yellow. The total Double Cut and Join (DCJ) distances between ancestral nodes suggest  
 336 Tree 1 as the most likely of these phylogenies. Right: Heat maps of the legume ancestors of the



337 corresponding phylogenetic relationships (left side), showing the clusters of reconstructed  
338 contigs likely making up either six or seven ancestral chromosomes.

339

340

341 **Phylogenomic analyses and a consensus phylogenetic model of speciation and whole**  
342 **genome duplications**

343 The duplication and speciation histories for the legumes in this study are evident in gene families  
344 constructed from the proteomes from each species. Using eleven representative genome  
345 sequences and annotations from four legume subfamilies as well as selected nonlegume  
346 outgroups, we calculated gene families for all genes. From 50 gene families with no sequence  
347 losses or gains relative to known WGD events, we calculated a consensus gene family based on a  
348 concatenated alignment from those families (Fig. 4a), which shows relative placements of  
349 speciation and WGD events. The basic species topology is congruent with the legume backbone  
350 topology that has been reported previously (Azani *et al.*, 2017; Ferreira, 2024). Separate WGDs  
351 are apparent early in each of four legume subfamilies included in this study, but complexities in  
352 the Cercidoideae and Caesalpinioideae require discussion and interpretation.

353

354 In the Cercidoideae, WGDs clearly affect *Bauhinia* and *Phanera* but not *Cercis*. However, in the  
355 consensus gene phylogeny (as well as in many individual gene phylogenies examined), *Cercis*  
356 groups with one of the *Bauhinia/Phanera* post-duplication lineages rather than outside (and prior  
357 to) the inferred WGD. A plausible explanation for this topology is that the WGD in the  
358 Cercidoideae was allopolyploid in nature, with a progenitor of *Cercis* contributing one  
359 subgenome and some other species early in the origin of the family contributing the other  
360 subgenome in the allopolyploid event.

361

362 In the Caesalpinioideae, a WGD is clearly evident, resulting in paralogs in each species  
363 examined in this subfamily; yet confusingly, one WGD-derived lineage groups more frequently  
364 with Papiloinoid species than with the Caesalpinioid paralogs. As with the Cercidoideae,  
365 allopolyploidy is a plausible explanation for this pattern. Specifically, the observed consensus  
366 gene family topology is consistent with a speciation along the legume backbone followed by a  
367 significant period of divergence (several million years), followed by an allopolyploid merger that

368 resulted in the Caesalpinoid subfamily – each member of which would have two (divergent)  
369 subgenomes. One of the diploid lineages would then have then gone on to become the progenitor  
370 of the Papilionoideae - which experienced its own WGD prior to diversification within that  
371 subfamily.

372

373 The hypothesized allopolyploid origins of the Caesalpinioideae and the Cercidoideae can be seen  
374 in the schematic in Fig. 4b. Here, an early speciation in the lineage leading to the  
375 Caesalpinioideae and Papilionoideae is represented as a red circle. After significant divergence,  
376 merger of two descendant species would have given rise to the 1n=14 Caesalpinioideae. In any  
377 species within that subfamily, two subgenomes are present, and paralogs from those two  
378 subgenomes may have distinct evolutionary histories that reflect the different histories of the two  
379 species that contributed to the fundamentally allopolyploid subfamily. In the schematic, those  
380 distinct gene histories can be seen as distinct dotted and solid paths leading to the speciation  
381 origin (red circle).

382

383 Similar dual paths can be seen in the Fig. 4b schematic of allopolyploidy in the Cercidoideae. If  
384 a *Cercis* progenitor contributed one of the subgenomes to the allopolyploid Cercidoideae lineage,  
385 then a *Cercis* gene should have greater affinity with one of the two WGD-derived paralogs in,  
386 for example, *Bauhinia*. Indeed, this is seen in the consensus species/WGD phylogeny (Fig. 4a).

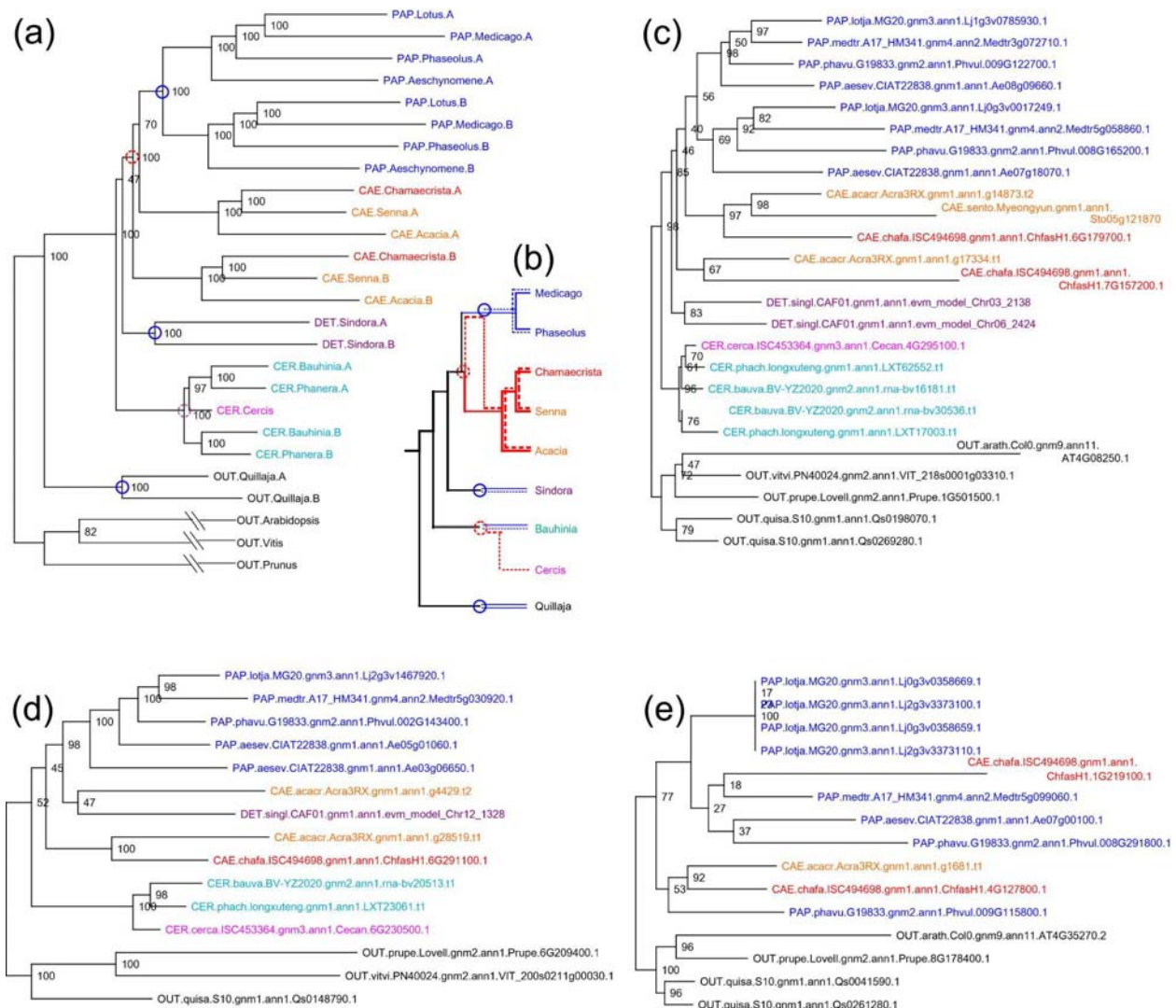
387

388

389

390





391

392 **Fig. 4.** Consensus and example gene family trees.

393 (a) Consensus tree showing inferred whole genome duplications (WGDs), calculated from 50 gene  
 394 families with near-complete representation of species and WGD-derived gene duplications. Red circles  
 395 indicate locations of inferred WGD events. Circles: inferred polyploidy events (blue auto-, red allo-  
 396 polyploidy).

397 (b) Schematic of gene, singl, duplication paths, showing inferred allo- and auto-polyploid events.

398 Individual lines represent evolutionary paths between genes, blue and red indicating inferred auto- and  
 399 allo-polyploidy respectively. Dotted and solid lines show hypothetical alternate paths for genes from  
 400 different subgenomes; for example, the path between two *Chamaecrista* paralogs would consist of a  
 401 dotted path, with greater similarity to the Papilionoideae and a solid-red path with greater similarity to an  
 402 older progenitor.

403 (c) Gene family phylogeny for the family of MtNSP2 (Nodulation Signaling Pathway 2), showing the  
404 same structure as the consensus tree.

405 (d) Gene family phylogeny of the family of LjSYMRK/MtDMI2 (SYMBiotic Receptor Kinase / Doesn't  
406 Make Infections 2), showing loss of orthologs from non-nodulating Senna and presence in the included  
407 nodulating species.

408 (e) Gene family phylogeny of the family of LjNIN (Nodule Inception), showing presence in all included  
409 nodulating species and absence from all non-nodulating species.

410

### 411 **Exemplar gene families and utility for studies of evolution of symbiotic nitrogen fixation in** 412 **the legumes**

413 The consensus topology in Fig. 4a is also seen in individual gene families – as, for example, in  
414 Fig. 4c, which happens to be a gene family that has a key role in symbiotic nitrogen fixation  
415 (SNF). This family contains MtNSP2, named for the *M. truncatula* “Nodulation Signaling  
416 Pathway 2” gene.

417

418 The other two gene families shown in Fig. 4 deviate from the consensus topology, but in  
419 interesting ways. Fig. 4d contains a gene critical for SNF, identified and named independently in  
420 studies using *Lotus japonicus* and *Medicago truncatula*: LjSYMRK/MtDMI2 (SYMBiotic  
421 Receptor Kinase / Doesn't Make Infections 2). In this family, orthologs are present in two  
422 nodulating species in the Caesalpinioideae, *C. fasciculata* and *A. crassicarpa*; but not in the  
423 (non-nodulating) *S. tora*. Significance of this presence-absence pattern needs to be tempered,  
424 however, by the existence of orthologs from the other non-nodulating species in the family:  
425 *Bauhinia*, *Phanera*, *Cercis*, and *Sindora*.

426

427 Fig. 4e contains LjNIN (Nodule Inception), which has been shown to be critical to SNF  
428 (Schäuser *et al.*, 1999; Marsh *et al.*, 2007; Griesmann *et al.*, 2018). In this gene family, the only  
429 legume species present in the family are those that nodulate; and all other non-nodulating species  
430 included in this study are absent. Specifically, the nodulating species present in this gene family  
431 are *L. japonicus*, *M. truncatula*, *P. vulgaris*, *A. evenia* (Papilionoideae) and *A. crassicarpa* and  
432 *C. fasciculata* (Caesalpinioideae). The non-nodulating species that are all absent from this gene  
433 family are *S. glauca* (Detarioideae) and *C. canadensis*, *P. championii*, and *B. variegata*  
434 (Cercidoideae).

435

436 **Ks and phylogenomic analyses indicate independent WGD events in at least four legume**  
437 **subfamilies**

438 Analyses of silent-site mutations between species pairs can be used to examine relative  
439 evolutionary rates associated with both speciations and WGDs. The plots in Fig. 5 show Ks  
440 peaks for selected species pairs. The resolution in these plots is higher than in many Ks analyses  
441 because the Ks values are taken from the modal Ks value per synteny block in the indicated  
442 species comparison, rather than from individual gene pairs.

443

444 In Fig. 5a, showing Ks values between *Cercis* and the indicated species (including *Cercis*  
445 compared with itself), the amplitude of *Cercis-Cercis* plot is near zero in the Ks range shown.  
446 The peak for the *Cercis-Cercis* comparison is near Ks=2.0 (not shown), consistent with the only  
447 duplication in *Cercis* being much older than the origin of the legumes. The Ks peak with the  
448 smallest value is with *Bauhinia*, which is consistent with *Cercis* and *Bauhinia* being relatively  
449 close sister taxa within the Cercidoideae. All other peaks in Fig. 5a are in the range 0.5-0.85,  
450 reflecting the substantial divergence with the other selected species, all of which are in other  
451 subfamilies (and another plant family all together in the case of *Quillaja*).

452

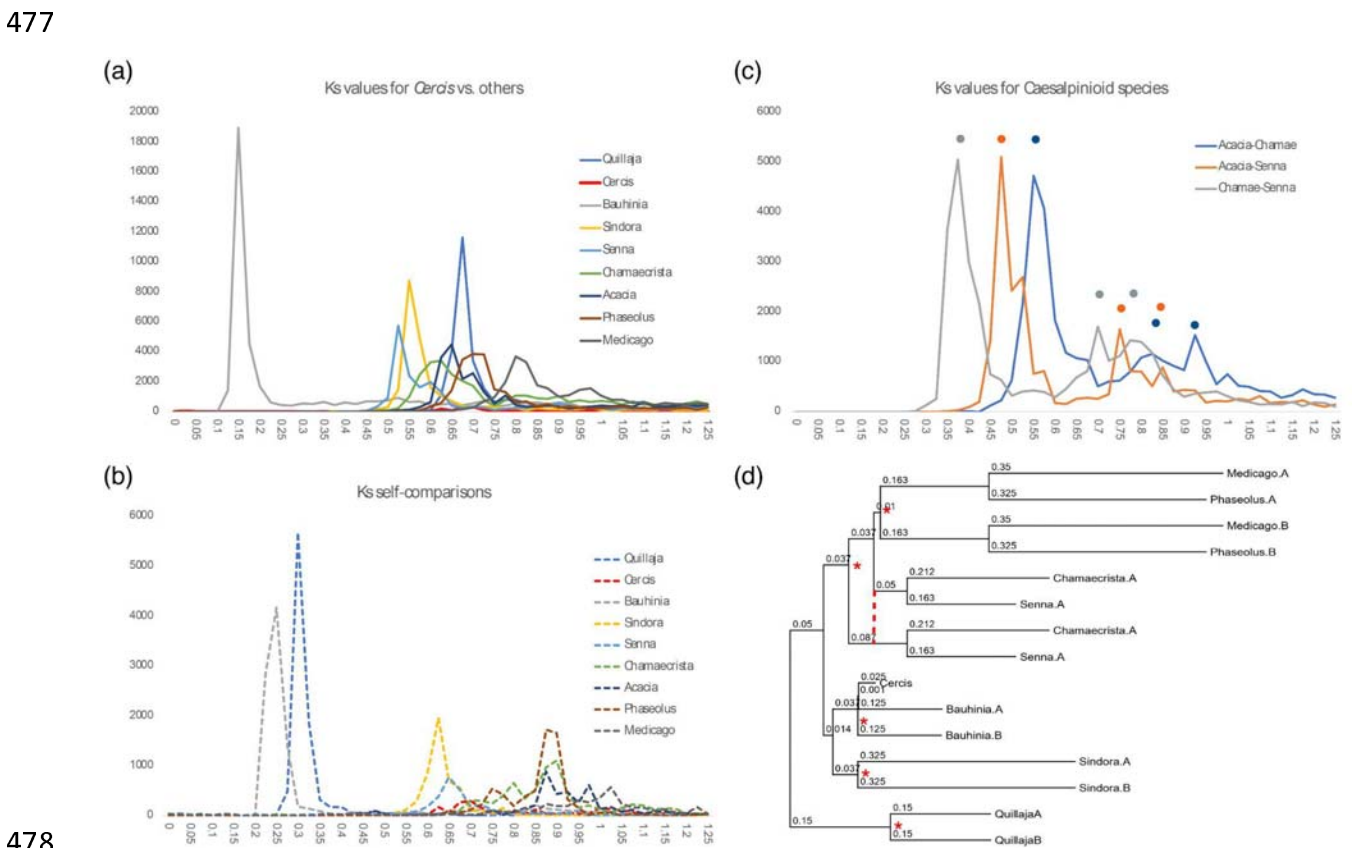
453 In Fig. 5b, showing Ks peaks for self-comparisons for each species, *Cercis* again is notable in its  
454 near-absence (several-fold lower in amplitude than for the other species self-comparisons). The  
455 Ks peaks for *Bauhinia* and *Quillaja* are both in small Ks bins (0.25 and 0.325 respectively),  
456 reflecting the relatively recent WGDs in those two taxa (WGDs that must be independent, since  
457 they are in different families).

458

459 In Fig. 5c, showing comparisons among the three Caesalpinioideae species included in this  
460 study, there is a strong primary peak reflecting speciations (*Acacia-Chamaecrista*, *Acacia-Senna*,  
461 and *Chamaecrista-Senna*); and an intriguing doubled (bimodal) peak in each comparison at more  
462 distant (older) Ks bins: at 0.7 and 0.8 for *Chamaecrista-Senna*, 0.75 and 0.85 for *Acacia-Senna*,  
463 and 0.825 and 0.925 for *Acacia-Chamaecrista*. For each species pair, these secondary double  
464 peaks are separated by 0.1 Ks units. A secondary peak is expected to represent a WGD near the

465 base of the Caesalpinioideae; but a doubled secondary peak may represent alternate evolutionary  
 466 paths associated with allopolyploidy, as depicted in the schematic in Fig. 4b.

467  
 468 Because rates of silent-site mutations may differ in different lineages, the values need to be  
 469 considered in a phylogenetic context. Projecting the Ks values from the species pairs (Table S3)  
 470 onto the consensus topology from Fig. 4a and resolving the branch lengths algebraically gives  
 471 the approximate branch lengths in Ks units. Both the phylogenetic and Ks analyses support  
 472 independent WGDs in each of the examined subfamilies. Some lineages have apparently been  
 473 evolving much faster than others (by the metric of silent-site mutations), with *Medicago* having  
 474 accumulated changes at nearly twice the pace of *Cercis* since their divergence from the common  
 475 legume ancestor (Ks distances to the common ancestor of ~0.5 for *Medicago* and 0.225 for  
 476 *Cercis*).

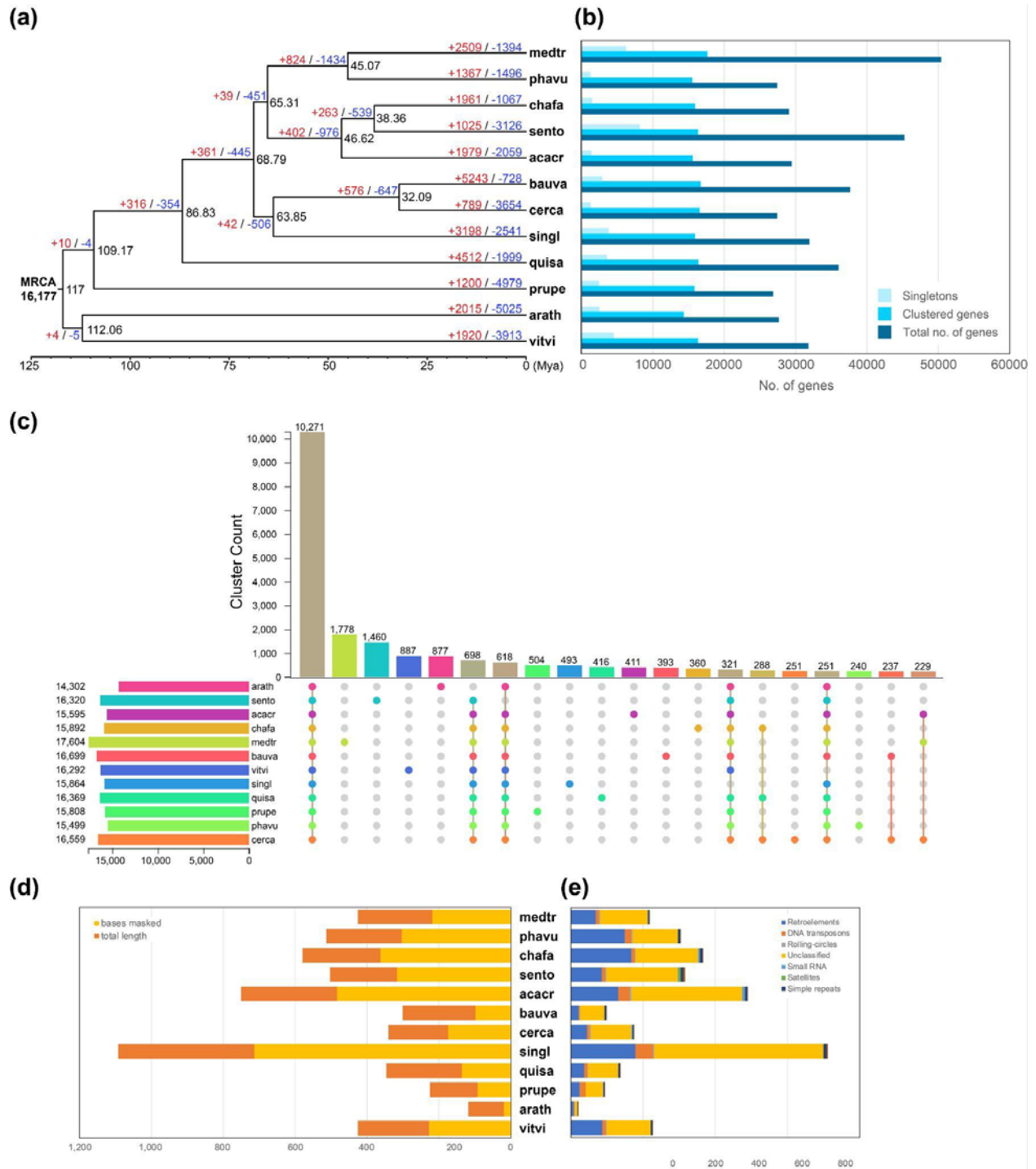


478  
 479 **Fig. 5.** Ks distributions and phylogeny with Ks-derived branch lengths. In panels (a) and (b), modal Ks  
 480 values were calculated for synteny blocks based on gene-pair matches, and those modal values used in  
 481 place of the individual gene-pair values. (a) Comparisons between *Cercis* and other species (peaks  
 482 representing speciations). (b) Ks values for species self-comparisons (peaks representing WGD events).

483 (c) Ks peaks for comparisons among species in the Caesalpinioideae: *Acacia crassicarpa*, *Chamaecrista*  
484 *fasciculata*, *Senna tora*. Dots show locations of Ks peaks – on the left indicating speciations, and on the  
485 right indicating a whole genome duplication. (d) Phylogeny derived from combined gene families (see  
486 Fig. 3 for details), with branch lengths determined algebraically from Ks peaks.

487

488



489

490

**Fig. 6.** Genomic comparison of *C. canadensis* and *C. fasciculata* with related species. (a) Phylogenetic tree with related species. Estimated divergence times are shown on the node, which indicates the number of expansions (red) and contractions (blue) relative to the gene family with the most recent common ancestor (MRCA). The divergence time was calculated based on 117 Mya for *Medicago* and *Vitis*.

493

494

(b) Bar graph comparing the total number of genes, clustered genes (orthogroup), and singleton numbers for the

495 species used in the phylogenetic analysis. (c) Upset plot of clustered genes. The number clustered in all  
496 species is 10,271. (d) Bar graph showing the proportion of repetitive sequences to genome size. (e)  
497 Cumulative bar graph showing the proportion of repetitive sequence content for each species.

498

### 499 **Gene expansion and contraction by comparing to representative species**

500 Analyses of gene expansion and contraction at the level of subfamily showed the largest number  
501 of changes in Papilionoideae (represented by *Medicago* and *Phaseolus*), with an increase of 632  
502 gene families and a decrease of 592 (Fig. 6a). At the species level, *Bauhinia* (Cercidoideae),  
503 showed the highest number of gene family changes, with 5,243 increases and 728 decreases;  
504 while *Cercis* had the lowest number of gene families, with 789 increases and 3,654 decreases.  
505 This is consistent with other indications of slower evolution in *Cercis* than other legumes –  
506 including *Bauhinia* within the same subfamily.

507

508 In the GO enrichment analysis of these increased and decreased gene families, *Cercis* was  
509 compared to *Bauhinia*, and *Chamaecrista* was compared to *Senna*. When comparing *Cercis* and  
510 *Bauhinia*, there was an enrichment of terms for defense response, signal transduction, and pollen  
511 recognition. Although both species are bisexual flowers, the fact that signal transduction and  
512 pollen recognition are most expanded in *Cercis* suggests that they may have unique recognition  
513 systems (Figs S3, S5).

514

515 In comparison to *Chamaecrista* and *Senna*, the top 30 expanded clusters are significantly enriched  
516 in signal transduction, pollen recognition, and auxin-activated signaling pathway functions (Figs  
517 S4). Interestingly, the auxin-activated signaling pathway is consistent with what we expected  
518 since *Chamaecrista* produces nodules unlike *Senna*. It is expected that the flavonoids produced  
519 by *Chamaecrista* act as auxin transport inhibitors during the pre-nodule infection stage, affecting  
520 the promotion of nodule primordial cell division at the nodule site (Figs S4, S6).

521 Orthologous cluster analysis revealed that 10,271 clusters were common to all species, 251  
522 clusters were unique to *Cercis* and 360 clusters were unique to *Chamaecrista* (Fig. 6c).

523

### 524 **Divergence time estimation**

525 Based on a divergence time of 117 Mya between *Medicago* and *Vitis* (Wikström *et al.*, 2001), we  
526 estimated the divergence of the other lineages based on phylogenetic topology and relative  
527 branch lengths. The following four legume subfamilies diverged at 68.8 Mya. This is similar to  
528 the radiometrically dated Cretaceous-Paleozoic (K-Pg) boundary of 66 Mya (Renne *et al.*, 2013).  
529 In our calculations, the estimated divergence time of the Papilionoideae and Caesalpinioideae  
530 was at 65.3 Mya, followed by Detarioideae and Cercidoideae at 63.9 Mya. *Cercis* and *Bauhinia*  
531 (in the same subfamily) diverged at 32.1 Mya (Fig. 6a).

532

### 533 **Centromeric arrays in *Cercis* and *Chamaecrista* suggest divergent evolutionary histories**

534 Tandem repeats in *C. canadensis* and *C. fasciculata* were identified using ULTRA (Olson &  
535 Wheeler, 2018). The most abundant repeats with length greater than 90 bases were evaluated for  
536 chromosomal position and array size. Based on these characteristics, putative centromeric  
537 repeats were selected and clustered in order to identify consensus centromeric repeat sequences  
538 (Data S1).

539

540 The arrays of probable centromeric repeats are strikingly large in *C. canadensis*, extending up to  
541 12 Mb and averaging 6 Mb. In total, they comprise approximately 12.6% of the total genome  
542 size. In contrast, the largest centromeric arrays in *C. fasciculata* comprise approximately 6.2% of  
543 the total genome size. The centromeric arrays generally occur in different (non-syntenic)  
544 locations – the arrays in *Cercis* disrupting synteny with *Chamaecrista* and vice versa (Figs S7-8).  
545 The pattern of synteny disruption suggests that centromeric arrays have originated (or moved)  
546 subsequent to the divergence of the respective lineages.

547

### 548 **Discussion**

549 The taxa selected as the focus of this work were chosen in order to help answer questions about  
550 the early diversification and evolution of the legume family, and to provide resources for further  
551 study of the evolution of symbiotic nitrogen fixation.

552

553 *Cercis*, as the earliest diverging lineage in the Cercidoideae (which itself is early-diverging  
554 within the legumes), is of particular interest due to its lack of WGD, its slow evolutionary rate,  
555 and its similarity to the inferred chromosome structure in the legume progenitor. As the only



556 legume established to this point to be without a WGD in the span of legume evolution, *Cercis*  
557 offers a unique model for the study of evolution of this large and diverse family.

558

559 Synteny, phylogenomic, and Ks analyses confirm that *C. canadensis* does not have a recent  
560 WGD within the timeframe of the legume family. Limited evidence of older duplications are  
561 evident, consistent with the ~135 Mya gamma triplication (Jiao *et al.*, 2012). In addition,  
562 independent WGD events are evident in the other four legume subfamilies that we examined in  
563 this study: the Papilionoideae, Caesalpinioideae, Detarioideae, and Cercidioideae. The estimated  
564 divergence time of each subfamily in the 63.9-68.8 Mya (Fig. 6a).

565

566 We also infer a probable karyotype of seven chromosomes in the legume ancestor, structurally  
567 similar to the current *Cercis* chromosomes, albeit with some rearrangements. Although some  
568 other examined species are found to have complex rearrangements (particularly in the  
569 Papilionoideae), the others are generally well approximated as a doubling of the *Cercis* genome,  
570 followed by a small number of splits, fusions, and inversions, which can be crucial for  
571 reconstructing the genomic history of the legume family.

572

573 We speculate that the unusually large centromeric arrays in *C. canadensis* (comprising roughly  
574 12% of the genome sequence), may be related to the stability of the chromosome structure over  
575 the ~70 million year history of legume evolution - as both cause and effect. In particular,  
576 centromeric arrays may tend to grow if undisturbed by rearrangements; and large centromeres  
577 may also aid in maintaining chromosome structure by providing “unmissable” mitotic  
578 attachment points.

579

580 Analysis of the expansion and contraction of duplicated genes and gene families confirms that  
581 the retention of duplicated genes has not been random, with some families generally retaining  
582 post-WGD duplicates (including in lineages with independent WGDs), and other families  
583 tending to fall back to single-copy status. Stochastic gene loss may have been notably important  
584 regarding SNF, where the pattern of presence and absence of this trait across the nitrogen-fixing  
585 clade has been modeled as a small number of independent gains and numerous losses  
586 (Griesmann *et al.*, 2018; Kates *et al.*, 2024).

587

588 *Chamaecrista* is of particular interest due to its capacity for SNF -- in contrast with many other  
589 lineages in the Caesalpinioideae subfamily -- including genera such as *Senna*, which is a sister  
590 genus within the Cassiinae tribe (and within the broader Mimosoideae–Caesalpinieae–Cassiinae  
591 (MCC) clade in the Caesalpinioideae). *Chamaecrista* has long been proposed as a model for  
592 examination of SNF within the Caesalpinioideae, due in part to characteristics that make  
593 *Chamaecrista* suitable for experimentation (Singer *et al.*, 2009). In particular, *C. fasciculata* is  
594 physically small, with short generation time, it can be outcrossed or selfed, and it exhibits  
595 considerable diversity across the North American habitats in which it is found.

596

597 The genome assemblies and annotations, together with other resources representing the four  
598 largest subfamilies in the legumes, permit construction of gene families that robustly capture the  
599 core genic complement of the family. These gene families support allopolyploid genome  
600 duplication events in both the Cercidoideae and Caesalpinioideae. This result helps explain why  
601 it has been so difficult to resolve the backbone topology for the legume phylogeny. Considerable  
602 discordance is seen in the placement of genes from Caesalpinioideae, such that WGD-derived  
603 paralogs from species such as *Chamaecrista* often do not resolve as sister to one another, but  
604 rather as alternately sister to genes from other subfamilies. An allopolyploid model is consistent  
605 with this observed pattern of discordance in gene families for Caesalpinioideae sequences.

606

607 An important general corollary of allopolyploidy within a taxonomy is that it may not be  
608 possible to faithfully represent the history of species relationships with a bifurcating  
609 phylogenetic model. Rather, a reticulate model is needed. Furthermore, particular genes may  
610 have followed different evolutionary histories due to effects such as incomplete lineage sorting,  
611 gene conversion, and segmental or chromosomal replacement following allopolyploidy.

612

613 Although the work presented here did not focus primarily on nodulation, the availability of a  
614 high-quality genome assembly and annotations for *C. fasciculata* are expected to be of use in  
615 such studies in the future. Examination of key nodulation-related gene families such as SYMRK  
616 and NIN show the utility of high-quality genic sequence from *Chamaecrista* and *Senna* -- the  
617 SYMRK gene family showing retention of both WGD-derived *Chamaecrista* genes but loss of

618 both from *Senna*; and NIN showing presence of orthologs in all examined nodulating species and  
619 absence from all non-nodulating species.

620

621 Within the Caesalpinioideae, nodulation is present in approximately nine lineages and absent in a  
622 comparable number (Sprent *et al.*, 2017; Kates *et al.*, 2024). Nodulation is present in most genera  
623 in the Papilionoideae, and absent in the other four legume subfamilies. The pattern of  
624 taxonomically scattered presence and absence of the trait in the Caesalpinioideae has been  
625 modeled as due to repeated, scattered losses of SNF (Kates *et al.*, 2024). We speculate that this  
626 pattern of loss could have been facilitated by the likely allopolyploid history of this subfamily.  
627 For example, if one of the progenitor species lacked SNF (either due to loss or non-gain) and the  
628 other progenitor had SNF capacity, then the allopolyploid merger might have produced a new  
629 polyploid species that had capacity for diversification, but that was also vulnerable to stochastic  
630 loss of genes crucial to SNF.

631

632 In the SYMRK/DMI2 family (Fig. 4d) (SYMRK identified in *Lotus japonicus* and the ortholog  
633 DMI2 identified in *Medicago truncatula*), two WGD-derived paralogs are present in *Acacia*  
634 (within the Mimosoid group, where nodulation predominates among most taxa). In  
635 *Chamaecrista*, one of two paralogs has evidently been lost. Both paralogs have been lost from  
636 (non-nodulating) *Senna*. While we can't establish the exact timing of the losses or the precise  
637 historical functions of these orthologs given data from this project, this pattern of presence and  
638 loss is consistent with inheritance of progenitor SYMRK/DMI2 genes from two species that  
639 merged to form the early allopolyploid founder of the Caesalpinioideae. In this model, those two  
640 genes may already have acquired differing functions. Both were evidently retained in *Acacia* and  
641 only one in *Chamaecrista*. Both were lost in *Senna*.

642

643 A similar story may apply for the NIN gene family (Fig. 4e). In that case, at least one functioning  
644 progenitor gene must have been present prior to the origin and radiation of the Papilionoideae  
645 and Caesalpinioideae. One WGD paralog was evidently lost from the *Acacia* (mimosid) lineage,  
646 but both WGD-derived paralogs have been retained in each of the Papilionoideae and in  
647 *Chamaecrista*. Intriguingly, one of the *Chamaecrista* genes (ChafaH1.1G219100) resolves sister  
648 to the described NIN gene, Medtr5G099060.

649

## 650 **Conclusions**

651 The work here describes the high-quality genomes and annotations for *Cercis canadensis*,  
652 eponymous for the *Cercidoideae* legume subfamily; and for *Chamaecrista fasciculata* in the  
653 Caesalpinoideae subfamily. These species are well placed taxonomically to aid inferences about  
654 key features of legume evolution, including the legume ancestral karyotype and the respective  
655 timing of subfamily origins and WGDs early in the legume radiation. Both *Cercis* and  
656 *Chamaecrista* show evidence of allopolyploidy – with the progenitor of *Cercis* likely  
657 contributing one subgenome to an allopolyploid event that gave rise to the remaining species in  
658 the *Cercidoideae*. In the *Caesalpinoideae*, the preponderance of gene families and Ks analyses  
659 suggest merger of two species that diverged along the taxonomic grade leading to the  
660 *Papilionoideae*, and then merged to give rise to the allopolyploid *Caesalpinoideae*. Such an  
661 allopolyploid merger early in the evolution of SNF may help to explain the very uneven pattern  
662 of SNF presence and absence across the diverse *Caesalpinoideae*. Finally, a finding of  
663 allopolyploidy during the origin of the legumes provides an important example of diversification  
664 that is not modeled sufficiently with a standard bifurcating phylogeny.

665

## 666 **Acknowledgements**

667 This work (proposal: 10.46936/10.25585/60001405) conducted by the U.S. Department of  
668 Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User  
669 Facility, is supported by the Office of Science of the U.S. Department of Energy operated under  
670 Contract No. DE-AC02-05CH11231. The work was also supported by the United States  
671 Department of Agriculture, Agricultural Research Service (USDA-ARS) CRIS Project 5030-  
672 21000-071-000D. This research used resources provided by the SCINet project and/or the AI  
673 Center of Excellence of the USDA Agricultural Research Service, ARS project numbers 0201-  
674 88888-003-000D and 0201-88888-002-000D. This research used resources of the National  
675 Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported  
676 by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-  
677 05CH11231 using NERSC award BER-ERCAP0027438. The USDA is an equal opportunity  
678 provider and employer. Mention of trade names or commercial products in this article is solely

679 for the purpose of providing specific information and does not imply recommendation or  
680 endorsement by the U.S. Department of Agriculture.

681

## 682 **Competing Interests**

683 None declared.

684

## 685 **Author Contributions**

686 HL and SBC conducted primary analyses, drafted the review, and managed project data. HL,  
687 JSS, and SBC conducted gene family, phylogenomic and Ks analyses. SBC and JSS collected  
688 plant tissue for genome sequencing. BDJ contributed analysis of genomic repeats. JG, JS, JJ, and  
689 RW conducted the genome sequencing and assembly. MW, JW, KK, and JE conducted lab work  
690 for genome sequencing and annotation. TB and SS generated genome annotations. DG and KB  
691 managed genome assembly and annotation work. QX, TH, RB, AL, DS, and LJ carried out  
692 synteny analyses and contributed software for genomic analysis. JL-M and LTL reviewed the  
693 analyses and edited the manuscript. SBC and JL-M conceptualized and designed the research.  
694 DMG, KB, and JL-M provided project management and funding. All authors approved the final  
695 draft.

696

## 697 **Data Availability**

698 BioProject for *Chamaecrista fasciculata* var. ISC494698:

699 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1137390>

700 BioProject for *Cercis canadensis* ISC453364:

701 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1137384>

702 Genome assemblies and annotations for *Chamaecrista fasciculata* var. ISC494698 and *Cercis*  
703 *canadensis* ISC453364: <https://phytozome-next.jgi.doe.gov>

704 Legume gene families and associated phylogenomic analyses:

705 <https://data.legumeinfo.org/LEGUMES/Fabaceae/genefamilies/legume.fam3.VLMQ/>

706 Supporting Information Methods S1 and Results S2 (manuscript supplement)

707 Supporting Information Figures and Tables S3 (manuscript supplement)

708

709 **References**

- 710 **Azani N, Babineau M, Bailey CD, Banks H, Barbosa AR, Pinto RB, Boatwright JS, Borges LM,**  
711 **Brown GK, Bruneau A. 2017.** A new subfamily classification of the Leguminosae based on a  
712 taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). *taxon*  
713 **66:** 44–77.
- 714 **Bandi V, Gutwin C. 2020.** Interactive exploration of genomic conservation.
- 715 **Cannon SB, Lee H-O, Weeks NT, Berendzen J. 2024.** Pandagma: A tool for identifying pan-gene  
716 sets and gene families at desired evolutionary depths and accommodating whole genome  
717 duplications. *Bioinformatics:* btae526.
- 718 **Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B,**  
719 **Stewart Jr CN, Rolf M. 2015.** Multiple polyploidy events in the early radiation of nodulating and  
720 nonnodulating legumes. *Molecular biology and evolution* **32:** 193–210.
- 721 **Cannon SB, Sato S, SatoshiTabata ND, May GD. 2011.** 22 Legumes as a Model Plant Family.  
722 *Biology and breeding of food legumes:* 348.
- 723 **Chu GX, Shen QR, Cao JL. 2004.** Nitrogen fixation and N transfer from peanut to rice cultivated  
724 in aerobic soil in an intercropping system and its effect on soil N fertility. *Plant and soil* **263:** 17–  
725 27.
- 726 **Davis CC, Fritsch PW, Li J, Donoghue MJ. 2002.** Phylogeny and biogeography of Cercis  
727 (Fabaceae): evidence from nuclear ribosomal ITS and chloroplast ndhF sequence data.  
728 *Systematic Botany* **27:** 289–302.
- 729 **Deorowicz S, Debudaj-Grabysz A, Gudyś A. 2016.** FAMSA: Fast and accurate multiple sequence  
730 alignment of huge protein families. *Scientific reports* **6:** 33964.
- 731 **Emms DM, Kelly S. 2019.** OrthoFinder: phylogenetic orthology inference for comparative  
732 genomics. *Genome biology* **20:** 1–14.
- 733 **Fenster CB. 1991.** Gene flow in *Chamaecrista fasciculata* (Leguminosae) I. Gene dispersal.  
734 *Evolution* **45:** 398–409.
- 735 **Ferreira PDL. 2024.** Phylogenomics and the rise of the angiosperms. *Nature.*
- 736 **Finn RD, Clements J, Eddy SR. 2011.** HMMER web server: interactive sequence similarity  
737 searching. *Nucleic acids research* **39:** W29–W37.
- 738 **Govaerts R, Nic Lughadha E, Black N, Turner R, Paton A. 2021.** The World Checklist of Vascular  
739 Plants, a continuously updated resource for exploring global plant diversity. *Scientific data* **8:**  
740 215.
- 741 **Griesmann M, Chang Y, Liu X, Song Y, Haberer G, Crook MB, Billault-Penneteau B,**  
742 **Lauressergues D, Keller J, Imanishi L. 2018.** Phylogenomics reveals multiple losses of nitrogen-  
743 fixing root nodule symbiosis. *Science* **361:** eaat1743.
- 744 **Haas BJ, Delcher AL, Wortman JR, Salzberg SL. 2004.** DAGchainer: a tool for mining segmental  
745 genome duplications and synteny. *Bioinformatics* **20:** 3643–3646.
- 746 **Han MV, Zmasek CM. 2009.** phyloXML: XML for evolutionary biology and comparative  
747 genomics. *BMC bioinformatics* **10:** 1–6.
- 748 **Herridge DF, Peoples MB, Boddey RM. 2008.** Global inputs of biological nitrogen fixation in  
749 agricultural systems. *Plant and soil* **311:** 1–18.
- 750 **Hewavithana T, Koh CS, Kaur A, Spiteri R, Parkin I, Jin L. 2023.** Inference of subgenomes  
751 resulting from polyploid events using synteny based dynamic linking and maximum

752 neighbourhood. In: IEEE, 104–109.

753 **Hilker R, Sickinger C, Pedersen CN, Stoye J. 2012.** UniMoG—a unifying framework for genomic  
754 distance calculation and sorting based on DCJ. *Bioinformatics* **28**: 2509–2511.

755 **Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka  
756 DR, Wafula E, Wickett NJ, et al. 2012.** A genome triplication associated with early diversification  
757 of the core eudicots. *Genome Biology* **13**: R3.

758 **Kates HR, O’Meara BC, LaFrance R, Stull GW, James EK, Liu S-Y, Tian Q, Yi T-S, Conde D, Kirst  
759 M. 2024.** Shifts in evolutionary lability underlie independent gains and losses of root-nodule  
760 symbiosis in a single clade of plants. *Nature Communications* **15**: 4262.

761 **Köpke U, Nemecek T. 2010.** Ecological services of faba bean. *Field crops research* **115**: 217–233.

762 **Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019.** RAxML-NG: a fast, scalable and  
763 user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**: 4453–  
764 4455.

765 **Kumar S, Stecher G, Suleski M, Hedges SB. 2017.** TimeTree: a resource for timelines, timetrees,  
766 and divergence times. *Molecular biology and evolution* **34**: 1812–1819.

767 **Lewis GP, Schrire BD, Mackinder BA, Rico L, Clark R. 2013.** A 2013 linear sequence of legume  
768 genera set in a phylogenetic context—a tool for collections management and taxon sampling.  
769 *South African Journal of Botany* **89**: 76–84.

770 **Li J, Shen J, Wang R, Chen Y, Zhang T, Wang H, Guo C, Qi J. 2023.** The nearly complete  
771 assembly of the *Cercis chinensis* genome and Fabaceae phylogenomic studies provide insights  
772 into new gene evolution. *Plant Communications* **4**.

773 **Li H-T, Yi T-S, Gao L-M, Ma P-F, Zhang T, Yang J-B, Gitzendanner MA, Fritsch PW, Cai J, Luo Y.  
774 2019.** Origin of angiosperms and the puzzle of the Jurassic gap. *Nature plants* **5**: 461–470.

775 **Marsh JF, Rakocevic A, Mitra RM, Brocard L, Sun J, Eschstruth A, Long SR, Schultze M, Ratet P,  
776 Oldroyd GE. 2007.** *Medicago truncatula* NIN is essential for rhizobial-independent nodule  
777 organogenesis induced by autoactive calcium/calmodulin-dependent protein kinase. *Plant  
778 physiology* **144**: 324–335.

779 **Martins LMV, Xavier GR, Rangel FW, Ribeiro JRA, Neves MCP, Morgado LB, Rumjanek NG.  
780 2003.** Contribution of biological nitrogen fixation to cowpea: a strategy for improving grain yield  
781 in the semi-arid region of Brazil. *Biology and fertility of soils* **38**: 333–339.

782 **Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2020.** CAFE 5 models variation in evolutionary  
783 rates among gene families. *Bioinformatics* **36**: 5516–5518.

784 **Olson D, Wheeler T. 2018.** ULTRA: a model based tool to detect tandem repeats. In: 37–46.

785 **Price MN, Dehal PS, Arkin AP. 2010.** FastTree 2—approximately maximum-likelihood trees for  
786 large alignments. *PLoS one* **5**: e9490.

787 **Ren L, Huang W, Cannon SB. 2019.** Reconstruction of ancestral genome reveals chromosome  
788 evolution history for selected legume species. *New Phytologist* **223**: 2090–2103.

789 **Renne PR, Deino AL, Hilgen FJ, Kuiper KF, Mark DF, Mitchell III WS, Morgan LE, Mundil R, Smit  
790 J. 2013.** Time scales of critical events around the Cretaceous-Paleogene boundary. *Science* **339**:  
791 684–687.

792 **Salvagiotti F, Specht JE, Cassman KG, Walters DT, Weiss A, Dobermann A. 2009.** Growth and  
793 nitrogen fixation in high-yielding soybean: Impact of nitrogen fertilization. *Agronomy Journal*  
794 **101**: 958–970.

795 **Schauser L, Roussis A, Stiller J, Stougaard J. 1999.** A plant regulator controlling development of

796 symbiotic root nodules. *Nature* **402**: 191–195.

797 **Singer SR, Maki SL, Farmer AD, Ilut D, May GD, Cannon SB, Doyle JJ. 2009.** Venturing beyond  
798 beans and peas: what can we learn from Chamaecrista? *Plant Physiology* **151**: 1041–1047.

799 **Sprent JI, Ardley J, James EK. 2017.** Biogeography of nodulated legumes and their  
800 nitrogen-fixing symbionts. *New Phytologist* **215**: 40–56.

801 **Steinegger M, Söding J. 2017.** MMseqs2 enables sensitive protein sequence searching for the  
802 analysis of massive data sets. *Nature biotechnology* **35**: 1026–1028.

803 **Sun J, Lu F, Luo Y, Bie L, Xu L, Wang Y. 2023.** OrthoVenn3: an integrated platform for exploring  
804 and visualizing orthologous data across genomes. *Nucleic Acids Research* **51**: W397–W403.

805 **Tang D, Chen M, Huang X, Zhang G, Zeng L, Zhang G, Wu S, Wang Y. 2023.** SRplot: A free online  
806 platform for data visualization and graphing. *PLoS One* **18**: e0294236.

807 **Wikström N, Savolainen V, Chase MW. 2001.** Evolution of the angiosperms: calibrating the  
808 family tree. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**: 2211–  
809 2220.

810 **Xu Q, Jin L, Zheng C, Leebens Mack JH, Sankoff D. 2020.** Raccroche: ancestral flowering plant  
811 chromosomes and gene orders based on generalized adjacencies and chromosomal gene co-  
812 occurrences. In: Springer, 97–115.

813 **Xu Q, Jin L, Zheng C, Zhang X, Leebens-Mack J, Sankoff D. 2023.** From comparative gene  
814 content and gene order to ancestral contigs, chromosomes and karyotypes. *Scientific Reports*  
815 **13**: 6095.

816 **Yancopoulos S, Attie O, Friedberg R. 2005.** Efficient sorting of genomic permutations by  
817 translocation, inversion and block interchange. *Bioinformatics* **21**: 3340–3346.

818 **Yang Z. 2007.** PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and*  
819 *evolution* **24**: 1586–1591.

820 **Yang Z, Sankoff D. 2011.** Generalized adjacency in genetic networks and the conservation of  
821 functional gene clusters. In: IEEE, 173–178.

822 **Zhao Y, Zhang R, Jiang K-W, Qi J, Hu Y, Guo J, Zhu R, Zhang T, Egan AN, Yi T-S. 2021.** Nuclear  
823 phylotranscriptomics and phylogenomics support numerous polyploidization events and  
824 hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Molecular Plant*  
825 **14**: 748–773.