bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 **Genome sequencing and analysis of two early-flowering cherry (*Cerasus* × *kanzakura*) varieties,**
2 **'Kawazu-zakura' and 'Atami-zakura'**
3
4 Kenta Shirasawa[1,*], Akihiro Itai[2], Sachiko Isobe[1]
5
6 [1]Kazusa DNA Research Institute, Chiba 292-0818, Japan
7 [2]Kyoto Prefectural University, Kyoto 606-8522, Japan
8
9 [*]To whom correspondence should be addressed:
10 Kenta Shirasawa
11 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan
12 Tel.: +81 438 52 3935
13 Fax: +81 438 52 3934
14 E-mail: shirasaw@kazusa.or.jp
15
16 **Running title:** Genomes of early-flowering cherry varieties
17

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18 **Abstract**

19 To gain genetic insights into the early-flowering phenotype of ornamental cherry, also known as sakura, we
20 determined the genome sequences of two early-flowering cherry (*Cerasus* × *kanzakura*) varieties,
21 'Kawazu-zakura' and 'Atami-zakura'. Since the two varieties are interspecific hybrids, likely derived from
22 crosses between *Cerasus campanulata* (early-flowering species) and *Cerasus speciosa*, we employed the
23 haplotype-resolved sequence assembly strategy. Genome sequence reads obtained from each variety by
24 single molecule real-time sequencing (SMRT) were split into two subsets, based on the genome sequence
25 information of the two probable ancestors, and assembled to obtain haplotype-phased genome sequences.
26 The resultant genome assembly of 'Kawazu-zakura' spanned 519.8 Mb with 1,544 contigs and an N50
27 value of 1,220.5 kb, while that of 'Atami-zakura' totaled 509.6 Mb with 2,180 contigs and an N50 value of
28 709.1 kb. A total of 72,702 and 72,528 potential protein-coding genes were predicted in the genome
29 assemblies of 'Kawazu-zakura' and 'Atami-zakura', respectively. Gene clustering analysis identified 2,634
30 clusters uniquely presented in the *C. campanulata* haplotype sequences, which might contribute to its
31 early-flowering phenotype. Genome sequences determined in this study provide fundamental information
32 for elucidating the molecular and genetic mechanisms underlying the early-flowering phenotype of
33 ornamental cherry tree varieties and their relatives.

34

35 **Keywords:** early-flowering, genome assembly, haplotype-phased genome sequence, long-read sequencing,
36 sakura

37

38 **Introduction**

39 Flowering cherry, called sakura in Japanese, is an ornamental plant popular worldwide. A major *Cerasus* ×
40 *yedoensis* cultivar ‘Somei-Yoshino’, which is an interspecific hybrid of *Cerasus spachiana* and *Cerasus*
41 *speciosa*[1], usually blooms from March to April in Japan. In addition, early-flowering sakura species, such
42 as *Cerasus campanulata*, usually bloom 1–2 months earlier than ‘Somei-Yoshino’, and its interspecific
43 hybrids such as *Cerasus* × *kanzakura* also exhibit early flowering. *C.* × *kanzakura* is considered a hybrid

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

44 between *C. campanulata* and *Cerasus speciosa* and/or *Cerasus jamasakura*[2], but its origin is still debated.
45 Two *C.* × *kanzakura* cultivars, 'Kawazu-zakura' and 'Atami-zakura', also bloom early (January and
46 February, respectively); however, the molecular mechanisms underlying their early-flowering phenotype
47 remain unknown. Although the mechanisms of early flowering in Rosaceae family members, Japanese
48 plum (*Prunus mume*) and peach (*Prunus persica*), which flower in February and March, respectively, are
49 well known[3], it remains unclear whether these mechanisms are common between *Cerasus* and *Prunus*.

50 Genome sequence analysis provides information on nucleotide polymorphisms and gene copy
51 number variation, which can lead to phenotypic differences among individuals and cultivars[4].
52 Pan-genomics, which involves de novo genome sequencing of multiple lines within a species, is conducted
53 to obtain information on variation in all genes within a species to understand the origin of the organism
54 under study[5,6]. In 'Somei-Yoshino', haplotype-phased genome sequences have been reported, and
55 comprehensive changes in gene expression during floral bud development that contribute toward flowering
56 have been revealed by time-course transcriptome analysis[7]. Therefore, comparative genomics of multiple
57 lines of flowering cherry varieties, such as 'Kawazu-zakura', 'Atami-zakura', and 'Somei-Yoshino', could
58 provide genetic insights into their early-flowering phenotypes.

59 A trio-binning strategy[8], previously used in a bovine F1 hybrid to resolve two haplotype-phased
60 genome sequences, was recently applied to 'Somei-Yoshino'[7]. Genes associated with the early-flowering
61 phenotype of 'Kawazu-zakura' and 'Atami-zakura' were assumed to be encoded by the *C. campanulata*
62 haplotype sequences. Therefore, in this study, we used the trio-binning strategy to determine the
63 haplotype-phased sequences of 'Kawazu-zakura' and 'Atami-zakura'. Comparative analysis of three
64 sakura genomes ('Kawazu-zakura', 'Atami-zakura', and 'Somei-Yoshino') facilitated the identification of
65 genes unique to the *C. campanulata* haplotype sequences of 'Kawazu-zakura' and 'Atami-zakura' as
66 candidates responsible for the early-flowering phenotype of these varieties.

67

68 **Materials and methods**

69 *Plant materials and DNA extraction*

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

70 Two early-flowering cherry (*Cerasus* × *kanzakura*) varieties, 'Kawazu-zakura' and 'Atami-zakura', were
71 used in this study. Both varieties were planted at the orchard of Kyoto Prefectural University (Kyoto,
72 Japan). Genome DNA was extracted from young leaves by a modified sodium dodecyl sulfate (SDS)
73 method[9].
74
75 *Genome size estimation*
76 Software tools used for data analyses are listed in Supplementary Table S1. Genome libraries for short-read
77 sequencing were prepared with the TruSeq DNA PCR-Free Sample Prep Kit (Illumina, San Diego, CA,
78 USA), and sequenced on the NextSeq 500 platform (Illumina, San Diego, CA, USA) in paired-end, 150 bp
79 mode. The genome size was estimated with Jellyfish.
80
81 *De novo genome sequence assembly and reference-guided contig ordering and orientation*
82 Genomes of the two cherry varieties were sequenced using the single molecule real-time (SMRT)
83 sequencing technology. Long-read DNA libraries were constructed using the SMRTbell Express Template
84 Prep Kit 2.0 (PacBio, Menlo Park, CA, USA) and sequenced on SMRT cells (1M v3 LR) in a PacBio
85 Sequel system (PacBio). Raw sequence reads of each variety were divided into two subsets with the
86 trio-binning strategy[8] using the short-read data of *C. campanulata* ('Kanhi-zakura') and *C. speciosa*
87 ('Ohshima-zakura') (DDBJ sequence archive accession no.: DRA008096)[7]. The sequence read subsets
88 were assembled separately with Falcon or Canu to build haplotype-phased diploid genome sequences.
89 Sequence errors in the contigs were corrected twice using long reads with ARROW. Potential
90 contaminating sequence reads from organelle genomes were identified by alignments with the chloroplast
91 and mitochondrial genome sequences of *Prunus avium* (GenBank accession no.: MK622380 and
92 MK816392) with Minimap2, and then removed from the final assemblies. Haplotype-phased sequences,
93 based on binning with *C. campanulata* and *C. speciosa*, were aligned against the *C. spachiana* and *C.*
94 *speciosa* haplotype sequences, respectively, of the 'Somei-Yoshino' genome using Ragoo to build
95 pseudomolecule sequences. Genome sequences were compared with D-Genies.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

96

97 *Gene prediction and repetitive sequence analysis*

98 Potential protein-coding genes were predicted with the MAKER pipeline, which was based on peptide
99 sequences predicted from the genome sequences of sweet cherry (PAV_r1.0)[10], peach (v2.0.a1)[11], and
100 Japanese plum[12]. Short genes (<300 bp) as well as genes predicted with an annotation edit distance (AED)
101 greater than 0.5, which is proposed as a threshold for good annotations in the MAKER protocol, were
102 removed to facilitate the selection of high-confidence (HC) genes. Functional annotation of the predicted
103 genes was performed with Hayai-Annotation Plants. Gene clustering was performed with OrthoFinder.

104 Repetitive sequences in the pseudomolecules were identified with RepeatMasker using repeat
105 sequences registered in Repbase and a *de novo* repeat library built with RepeatModeler. The identified
106 repetitive sequences were classified into nine types, in accordance with RepeatMasker: short interspersed
107 nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeat (LTR)
108 elements, DNA elements, small RNAs, satellites, simple repeats, low complexity repeats, and unclassified.

109

110 **Results and data description**

111 *De novo assembly of 'Kawazu-zakura' and 'Atami-zakura' genomes*

112 Short reads amounting to 64.0 and 127.7 Gb were obtained for 'Kawazu-zakura' and 'Atami-zakura',
113 respectively. The genome sizes of 'Kawazu-zakura' and 'Atami-zakura' were estimated at 672.7 and 675.2
114 Mb, respectively (Figure 1). Since 'Kawazu-zakura' and 'Atami-zakura' are interspecific hybrids, we used
115 a trio-binning strategy to establish haplotype-resolved genome assemblies representing each parental
116 genome sequence.

117 Long-read data (34.9 Gb) of 'Kawazu-zakura' obtained from two SMRT cells were divided into two
118 subsets (17.2 and 17.6 Gb), in accordance with the short-read data of potential parental species, *C.*
119 *campanulata* and *C. speciosa*[7], respectively. Reads in each subset were independently assembled with
120 Falcon to construct contigs representing the two haplotype sequences. Potential errors in the haplotype
121 sequences were corrected with long reads, and sequences of organelle genomes were removed to obtain the

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

122 final assembly of the diploid genome of 'Kawazu-zakura'. The resulting assemblies consisted of *C.*
123 *campanulata* (262.2 Mb, N50 = 1.4 Mb) and *C. speciosa* (257.6 Mb, N50 = 1.1 Mb) haplotypes (Table 1),
124 and were designated as KWZcam_r1.0 and KWZspe_r1.0, respectively. Although the total assembly size
125 was shorter than the estimated size, the complete BUSCO scores of KWZcam_r1.0 and KWZspe_r1.0
126 were 93.1% and 96.7%, respectively, indicating that the assemblies were complete (Table 1). The two
127 assemblies were merged to generate KWZ_r1.0, with a complete BUSCO score of 98.0%.

128 The 'Atami-zakura' genome was sequenced in parallel with the 'Kawazu-zakura' genome.
129 Long-read data of 'Atami-zakura' (14.3 Gb) were obtained from two SMRT cells, and divided into two
130 subsets (7.4 and 6.8 Gb) using the short-read data of *C. campanulata* and *C. speciosa*[7], respectively. The
131 reads were assembled with Falcon to generate two haplotype contig sequences. However, the sizes of the
132 two assemblies (139.9 and 110.0 Mb) were much smaller than the estimated sizes. Therefore, we used
133 Canu to obtain long assemblies. This was followed by potential sequence error correction and organelle
134 genome sequence removal. The sizes of the resultant assemblies were improved to 267.4 Mb (N50 = 853.5
135 kb) and 242.2 Mb (N50 = 569.4 Mb) for the *C. campanulata* and *C. speciosa* haplotypes, respectively
136 (Table 1), and the assemblies were designated as ATMcam_r1.0 and ATMspe_r1.0, respectively. The
137 complete BUSCO scores were 93.4% and 93.5% for ATMcam_r1.0 and ATMspe_r1.0, respectively (Table
138 1), and 98.2% for the merged assembly (ATM_r1.0).

139

140 *Reference-guided pseudomolecule sequence construction*

141 Since the genome structures are well conserved across the *Cerasus* and *Prunus* species[7], we used the two
142 haplotype pseudomolecule sequences of the 'Somei-Yoshino' genome, CYEspachiana_r3.1 and
143 CYEspeciosa_r3.1, as references to establish the pseudomolecule sequences of 'Kawazu-zakura' and
144 'Atami-zakura'. A total of 777 and 746 contigs of KWZcam_r1.0 and KWZspe_r1.0, respectively, were
145 aligned against CYEspachiana_r3.1 and CYEspeciosa_r3.1 sequences, respectively. The lengths of the
146 resultant 'Kawazu-zakura' pseudomolecule sequences were 256.7 Mb (KWZcam_r1.0) and 246.5 Mb
147 (KWZspe_r1.0) (Table 2). On the other hand, 1,110 ATMcam_r1.0 and 1,041 ATMspe_r1.0 contigs were

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

148 aligned with the CYEspachiana_r3.1 and CYEspeciosa_r3.1 sequences, respectively, and the lengths of the
149 'Atami-zakura' pseudomolecule sequences obtained were 261.5 Mb (ATMcam_r1.0) and 238.9 Mb
150 (ATMspe_r1.0) (Table 2). The pseudomolecule sequences of 'Kawazu-zakura' and 'Atami-zakura'
151 genomes covered the entire genome sequence of 'Somei-Yoshino' (Figure 2).
152
153 *Gene and repetitive sequence predictions*
154 A total of 36,264 and 36,264 HC protein-coding genes were predicted in KWZcam_r1.0 and KWZspe_r1.0
155 assemblies, respectively (Table 2). The complete BUSCO scores of genes in the KWZcam_r1.0 and
156 KWZspe_r1.0 were 88.3% and 86.6%, respectively, while the BUSCO score of all 72,582 genes was
157 97.0%. Functional gene annotation revealed that 9,430, 17,907, and 12,603 sequences were assigned to
158 Gene Ontology (GO) slim terms in the biological process, cellular component, and molecular function
159 categories, respectively, and 2,264 genes had enzyme commission numbers.
160 On the other hand, 36,281 and 36,421 HC genes were predicted in ATMcam_r1.0 and ATMspe_r1.0
161 assemblies, respectively (Table 2). Complete BUSCOs of genes in ATMcam_r1.0 and ATMspe_r1.0 were
162 88.3% and 86.6%, respectively, while that of all 72,702 genes was 96.8%. According to the functional
163 gene annotation, 9,836, 18,586, and 13,020 sequences were assigned to GO slim terms in the biological
164 process, cellular component, and molecular function categories, respectively, and 2,301 genes had enzyme
165 commission numbers.
166 Repeat sequences occupied varying proportions of the different genome assemblies: 48.0%
167 (KWZcam_r1.0), 45.7% (KWZspe_r1.0), 47.7% (ATMcam_r1.0), and 43.2% (ATMspe_r1.0). LTR
168 elements were the most abundant repetitive sequences (15.1–17.7%), followed by unclassified repeats
169 (12.7–13.7%) and DNA transposons (11.1–13.2%) (Table 3).
170
171 *Clustering analysis of flowering time related genes in cherry varieties*
172 Four sets of genes predicted in the haplotype-phased genomes of 'Kawazu-zakura' and 'Atami-zakura'
173 clustered with two sets of genes in the two haploid sequences of 'Somei-Yoshino'. A total of 35,226

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

174 clusters were obtained, of which 10,702 were common across all six gene sets. The early-flowering
175 phenotype of *C.* × *kanzakura* could be explained by genes uniquely present in the *C. campanulata*
176 haplotype sequences. In the *C. campanulata* haplotype sequences of 'Kawazu-zakura' and 'Atami-zakura'
177 genomes, a total of 2,634 clusters were found to include 3,113 and 3,123 genes, respectively, suggesting
178 that these genes might be associated with the early-flowering phenotype of 'Kawazu-zakura' and
179 'Atami-zakura'.
180

181 **Conclusion and future perspectives**

182 Here, we report haplotype-phased genome assemblies of two early-flowering cherry (*C.* × *kanzakura*)
183 cultivars, 'Kawazu-zakura' and 'Atami-zakura', both of which are interspecific hybrids derived from *C.*
184 *campanulata* and *C. speciosa*. Although the origin of *C.* × *kanzakura* remains unclear, *C. campanulata* and
185 *C. speciosa* and/or *C. jamasakura* are considered as its potential parents[2]. Another possibility is that
186 'Atami-zakura' originated from *C. jamasakura* and *C. campanulata*[13]. This is supported by the fact that our
187 attempt to divide the long reads of 'Atami-zakura' into two subsets using short-read data of *C. serrulata*
188 (closely related to *C. jamasakura*[7]) and *C. campanulata* failed (data not shown). Therefore, we used short
189 reads of *C. campanulata* and *C. speciosa* for both 'Kawazu-zakura' and 'Atami-zakura'. This result
190 suggests that both 'Kawazu-zakura' and 'Atami-zakura' are closely related to *C. campanulata* and *C.*
191 *speciosa*.

192 Clustering analysis of genes predicted in the genomes of 'Kawazu-zakura' and 'Atami-zakura'
193 together with those of 'Somei-Yoshino' revealed that 2,634 gene clusters were uniquely present in the
194 genome of *C. campanulata* but absent from the genomes of *C. spachiana* and *C. speciosa*. Such copy
195 number variation (or presence/absence variation) of genes could explain the early-flowering phenotype of
196 'Kawazu-zakura' and 'Atami-zakura'. Previously, we performed a time-course transcriptome analysis of
197 the floral buds and flowers of 'Somei-Yoshino' to clarify gene expression patterns during flowering[7]. A
198 similar time-course transcriptome analysis could be applied to 'Kawazu-zakura' and 'Atami-zakura'.
199 Comparative transcriptome analysis of three cultivars could identify the genes responsible for the

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

200 early-flowering phenotype of sakura. Furthermore, comparative transcriptome analysis of Japanese apricot
201 and peach[3] could reveal the genetic mechanisms controlling flowering time across all *Prunus* and *Cerasus*
202 species.

203 Although several flowering cherry cultivars are known to bloom in late-spring, fall, and winter
204 seasons[7], genome sequences of only a few of these cultivars are publicly available[7,14,15]. Comparative
205 genomics and transcriptomics, also known as pan-genomics[4-6], of sakura would provide insights into the
206 origins of these cultivars and their flowering mechanisms, which could facilitate the development of new
207 cultivars with attractive flower characteristics and provide us with the ability to forecast the date of sakura
208 blooming.
209

210 **Acknowledgments**

211 We thank Y. Kishida, C. Minami, H. Tsuruoka, and A. Watanabe (Kazusa DNA Research Institute) for
212 technical assistance.
213

214 **Data availability**

215 Sequence reads are available from the DNA Data Bank of Japan (DDBJ) Sequence Read Archive (DRA)
216 database (accession no.: DRA012553). The DDBJ accession numbers of assembled sequences are
217 BPUM01000001–BPUM01000783 (KWZcam_r1.0), BPUM01000784–BPUM01001544 (KWZspe_r1.0),
218 BPUL01000001–BPUL01001124 (ATMcam_r1.0), and BPUL01001125–BPUL01002180 (ATMspe_r1.0).
219 The genome sequence information generated in this study is available at Plant GARDEN
220 (https://plantgarden.jp).
221

222 **Funding**

223 This work was supported by the Kazusa DNA Research Institute Foundation.
224

225 **Conflict of interest**

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

226 None declared.

227

228 **Supporting information**

229 **Supplementary Table S1** Software tools used for genome assembly and gene prediction.

230

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

231 **References**

232 1. Takenaka, Y. 1963, The Origin of the Yoshino Cherry Tree. *Journal of Heredity*, **54**,
233 207-211.
234 2. Kato, S., Matsumoto, A., Yoshimura, K., et al. 2014, Origins of Japanese flowering cherry
235 (Prunus subgenus Cerasus) cultivars revealed using nuclear SSR markers. *Tree Genetics*
236 *& Genomes*, **10**, 477-487.
237 3. Yamane, H. 2014, Regulation of Bud Dormancy and Bud Break in Japanese Apricot (Prunus
238 mume Siebold ^|^amp; Zucc.) and Peach [Prunus persica (L.) Batsch]: A Summary of Recent
239 Studies. *Journal of the Japanese Society for Horticultural Science*, **83**, 187-202.
240 4. Chen, F., Song, Y., Li, X., et al. 2019, Genome sequences of horticultural plants: past,
241 present, and future. *Hortic Res*, **6**, 112.
242 5. Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B. and Hirsch, C. N. 2021, How the
243 pan-genome is changing crop genomics and improvement. *Genome Biol*, **22**, 3.
244 6. Tao, Y., Zhao, X., Mace, E., Henry, R. and Jordan, D. 2019, Exploring and Exploiting
245 Pan-genomics for Crop Improvement. *Mol Plant*, **12**, 156-169.
246 7. Shirasawa, K., Esumi, T., Hirakawa, H., et al. 2019, Phased genome sequence of an
247 interspecific hybrid flowering cherry, 'Somei-Yoshino' (Cerasus x yedoensis). *DNA Res*,
248 **26**, 379-389.
249 8. Koren, S., Rhie, A., Walenz, B. P., et al. 2018, De novo assembly of haplotype-resolved
250 genomes with trio binning. *Nat Biotechnol*, **36**, 1174-1182.
251 9. Teramoto, S., Kano-Murakami, Y., Hori, M. and Kamiyama, K. 1994, DNA Finger-Printing'
252 to Distinguish Cultivar and Parental Relation of Japanese Pear. *Journal of the Japanese*
253 *Society for Horticultural Science*, **63**, 17-21.
254 10. Shirasawa, K., Isuzugawa, K., Ikenaga, M., et al. 2017, The genome sequence of sweet
255 cherry (Prunus avium) for use in genomics-assisted breeding. *DNA Res*, **24**, 499-508.
256 11. Verde, I., Jenkins, J., Dondini, L., et al. 2017, The Peach v2.0 release: high-resolution

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

257 linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity.
258 *BMC Genomics*, **18**, 225.
259 12. Zhang, Q., Chen, W., Sun, L., et al. 2012, The genome of Prunus mume. *Nat Commun*, **3**,
260 1318.
261 13. Ogawa, T., Kameyama, Y., Kanazawa, Y., Suzuki, K. and Somego, M. 2012, Origins of
262 early-flowering cherry cultivars, Prunus×kanzakura cv. Atami-zakura and
263 Prunus×kanzakura cv. Kawazu-zakura, revealed by experimental crosses and AFLP analysis.
264 *Scientia Horticulturae*, **140**, 140-148.
265 14. Baek, S., Choi, K., Kim, G. B., et al. 2018, Draft genome sequence of wild Prunus yedoensis
266 reveals massive inter-specific hybridization between sympatric flowering cherries.
267 *Genome Biol*, **19**, 127.
268 15. Yi, X. G., Yu, X. Q., Chen, J., et al. 2020, The genome of Chinese flowering cherry
269 (Cerasus serrulata) provides new insights into Cerasus species. *Hortic Res*, **7**, 165.
270

271

Table 1 Statistics of the contig sequences of two flowering cherry (*Cerasus* × *kanzakura*) cultivars, 'Kawazu-zakura' and 'Atami-zakura'

	KWZ_r1.0	KWZcam_r1.0	KWZspe_r1.0	**ATM_r1.0**	ATMcam_r1.0	ATMspe_r1.0
Total contig size (bases)	**519,843,677**	262,196,010	257,647,667	**509,633,549**	267,393,285	242,240,264
Number of contigs	**1,544**	783	761	**2,180**	1,124	1,056
Contig N50 length (bases)	**1,220,495**	1,445,144	1,108,133	**709,113**	853,547	569,444
Longest contig size (bases)	**8,019,066**	5,955,677	8,019,066	**5,799,312**	5,799,312	3,381,444
Gap (bases)	**0**	0	0	**0**	0	0
Complete BUSCOs	**98.2%**	93.1%	96.7%	**98.0%**	93.4%	93.5%
Single-copy BUSCOs	**7.5%**	86.7%	89.0%	**16.0%**	86.5%	87.8%
Duplicated BUSCOs	**90.7%**	6.4%	7.7%	**82.0%**	6.9%	5.7%
Fragmented BUSCOs	**0.3%**	0.7%	0.4%	**0.4%**	0.7%	1.6%
Missing BUSCOs	**1.5%**	6.2%	2.9%	**1.6%**	5.9%	4.9%
#Genes	**72,702**	36,281	36,421	**72,528**	36,264	36,264

272

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

273 **Table 2** Statistics of the pseudomolecule sequences of flowering cherry (*C.* × *kanzakura*) cultivars, 'Kawazu-zakura' and 'Atami-zakura'

		'Kawazu-zakura'						'Atami-zakura'					
	Chrom.	Total length	%	Number of contigs	%	Number of genes	%	Total length	%	Number of contigs	%	Number of genes	%
C. campanulata haplotype	1	38,834,322	14.8	86	11.0	5,173	14.3	37,943,013	14.2	123	10.9	5,180	14.3
	2	45,216,402	17.2	216	27.6	6,881	19.0	49,279,926	18.4	289	25.7	6,710	18.5
	3	28,286,294	10.8	74	9.5	3,837	10.6	29,043,603	10.9	108	9.6	3,885	10.7
	4	31,150,796	11.9	110	14.0	3,747	10.3	33,258,136	12.4	157	14.0	4,442	12.2
	5	28,296,805	10.8	82	10.5	3,852	10.6	26,412,617	9.9	83	7.4	3,625	10.0
	6	33,630,106	12.8	83	10.6	4,908	13.5	34,947,855	13.1	132	11.7	4,871	13.4
	7	20,603,721	7.9	61	7.8	2,702	7.4	21,052,878	7.9	92	8.2	3,003	8.3
	8	30,708,646	11.7	65	8.3	4,559	12.6	29,538,870	11.0	126	11.2	3,745	10.3
	Unassigned	5,546,318	2.1	6	0.8	622	1.7	6,027,887	2.3	14	1.2	803	2.2
	Total	**262,273,410**	**100.0**	**783**	**100.0**	**36,281**	**100.0**	**267,504,785**	**100.0**	**1,124**	**100.0**	**36,264**	**100.0**
C. speciosa haplotype	1	42,661,824	16.6	82	10.8	5,912	16.2	38,473,692	15.9	136	12.9	5,644	17.0
	2	33,947,397	13.2	125	16.4	4,804	13.2	32,131,565	13.3	179	17.0	4,495	13.5
	3	29,126,079	11.3	81	10.6	4,367	12.0	30,781,260	12.7	111	10.5	4,269	12.8
	4	34,989,196	13.6	162	21.3	4,668	12.8	35,466,104	14.6	214	20.3	4,532	13.6
	5	25,691,272	10.0	99	13.0	3,729	10.2	23,383,776	9.6	100	9.5	3,062	9.2
	6	29,111,560	11.3	56	7.4	4,240	11.6	32,639,480	13.5	120	11.4	4,532	13.6
	7	16,872,426	6.5	20	2.6	2,328	6.4	16,033,168	6.6	40	3.8	2,174	6.5
	8	34,141,902	13.2	121	15.9	5,125	14.1	30,022,323	12.4	141	13.4	4,075	12.3
	Unassigned	11,181,211	4.3	15	2.0	1,248	3.4	3,413,596	1.4	15	1.4	481	1.4
	Total	**257,722,867**	**100.0**	**761**	**100.0**	**36,421**	**100.0**	**242,344,964**	**100.0**	**1,056**	**100.0**	**33,264**	**100.0**

274

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

275 **Table 3** Repetitive sequences in two flowering cherry (*C.* × *kanzakura*) cultivars, 'Kawazu-zakura' and 'Atami-zakura'

Repeat type	'Kawazu-zakura'						'Atami-zakura'					
	C. campanulata haplotype			*C. speciosa* haplotype			*C. campanulata* haplotype			*C. speciosa* haplotype		
	Number of elements	Length occupied (bp)	%	Number of elements	Length occupied (bp)	%	Number of elements	Length occupied (bp)	%	Number of elements	Length occupied (bp)	%
SINEs	5,278	495,207	0.2	7,013	665,451	0.3	8,832	896,223	0.3	6,537	608,541	0.3
LINEs	9,358	3,548,357	1.4	9,980	3,635,040	1.4	9,242	3,175,048	1.2	9,285	3,460,432	1.4
LTR elements	63,025	45,423,275	17.3	57,175	42,749,594	16.6	61,503	47,443,444	17.7	52,221	36,517,551	15.1
DNA transposons	85,647	33,936,563	12.9	84,151	30,984,015	12.0	88,999	35,176,601	13.2	77,636	26,829,236	11.1
Unclassified	131,199	36,041,455	13.7	116,201	32,825,941	12.7	130,209	34,407,194	12.9	112,663	31,370,494	12.9
Small RNA	5,384	657,326	0.3	7,211	1,536,541	0.6	6,949	828,201	0.3	2,598	503,911	0.2
Satellites	1,072	277,222	0.1	297	53,425	0.0	1,083	399,307	0.2	342	75,860	0.0
Simple repeats	75,567	3,104,558	1.2	77,082	3,144,046	1.2	77,750	3,266,196	1.2	74,414	3,048,442	1.3
Low complexity	14,265	706,271	0.3	14,754	717,629	0.3	14,352	695,481	0.3	14,137	693,339	0.3

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

276 **Figure legends**

277 **Figure 1** Estimation of the genome size of two flowering cherry (*Cerasus* × *kanzakura*) varieties,

278 ‘Kawazu-zakura’ and ‘Atami-zakura’, based on *k*-mer analysis ($k =$ □17), with the given multiplicity

279 values.

280

281 **Figure 2** Comparative analysis of the genome sequence and structure of flowering cherry varieties,

282 ‘Atami-zakura’, ‘Kawazu-zakura’, and ‘Somei-Yoshino’.

283 Chromosome numbers are indicated above the x-axis and on the right side of the y-axis. Genome

284 sizes (Mb) are below the x-axis and on the left side of the y-axis.

285

'Kawazu–zakura'

Number of distinct k–mers (million)

3.0
2.5
2.0
1.5
1.0
0.5
0.0

0 100 200 300 400

k–mer multiplicity

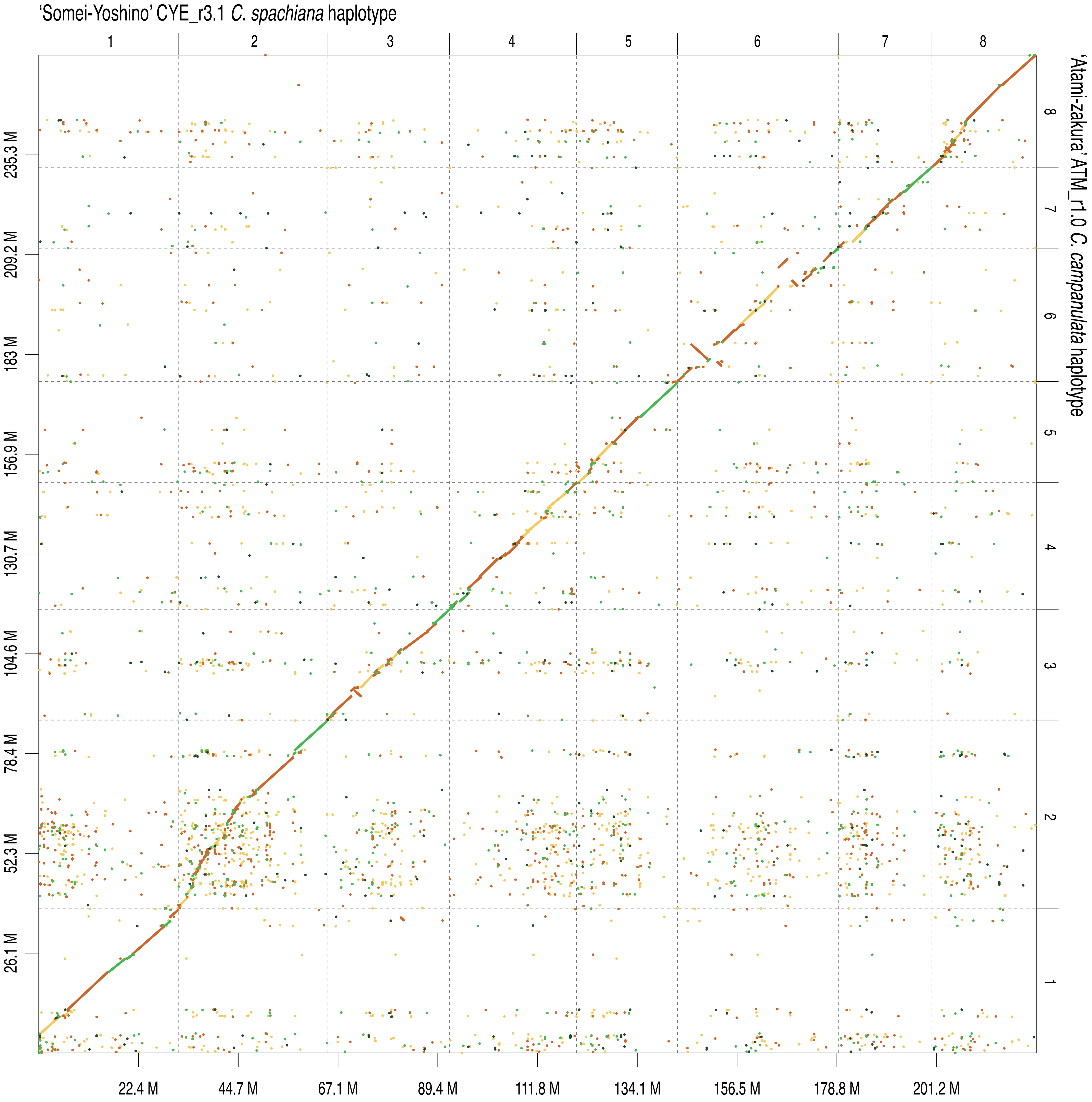
'Atami–zakura'

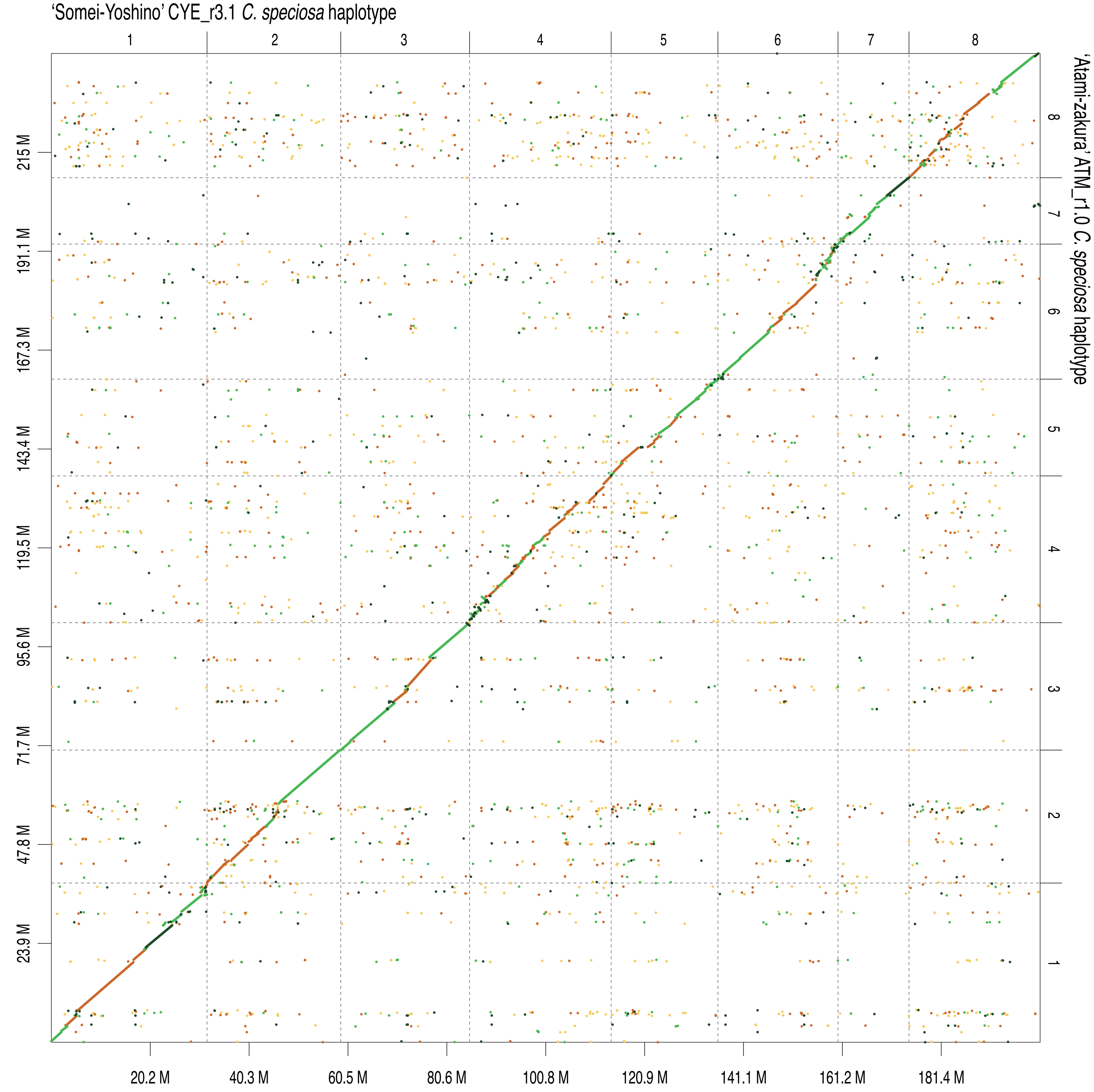
Number of distinct k–mers (million)

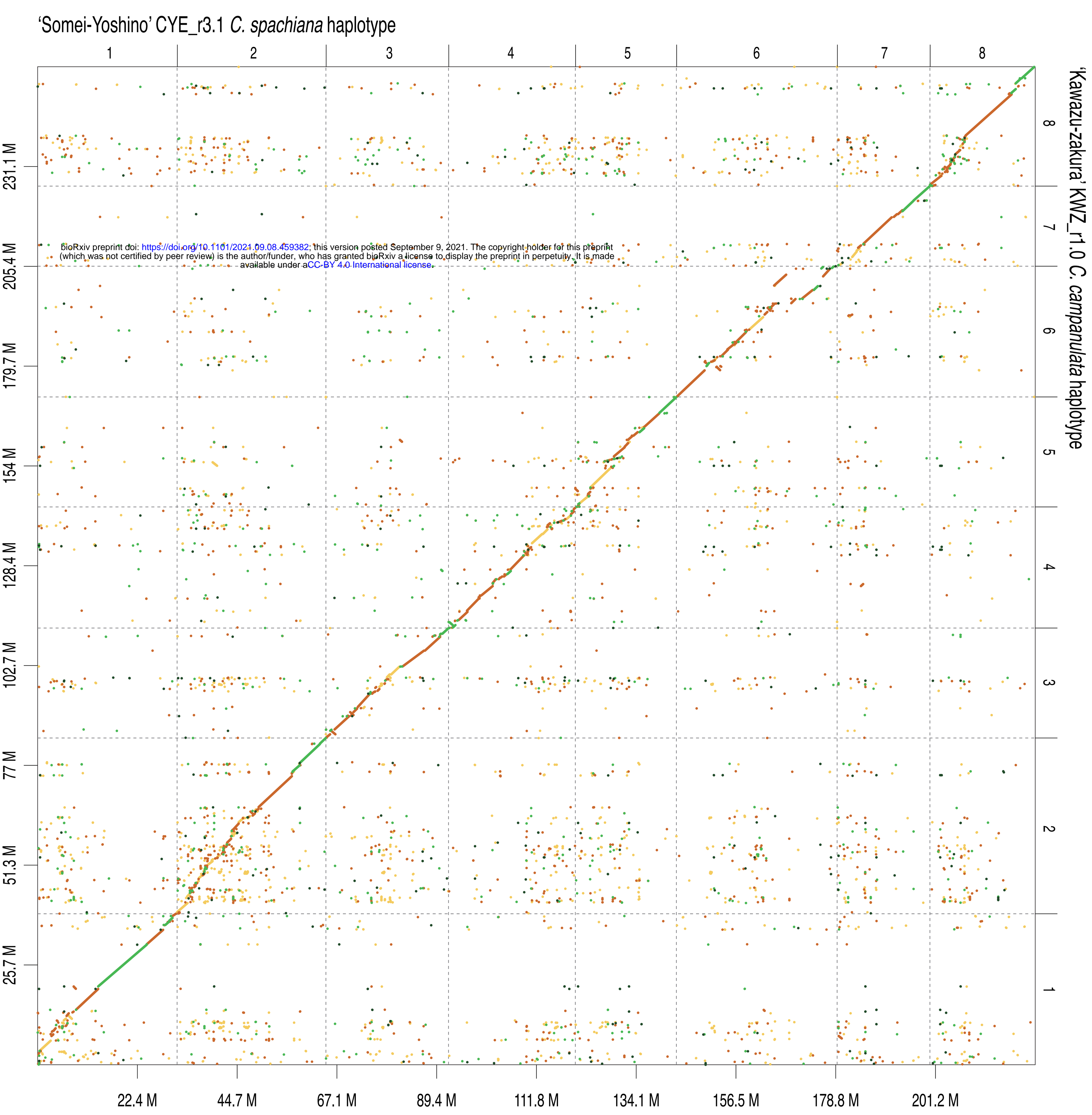
2.0
1.5
1.0
0.5
0.0

0 200 400 600 800

k–mer multiplicity







bioRxiv preprint doi: https://doi.org/10.1101/2021.09.08.459382; this version posted September 9, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.