UNIVERSITY of York

This is a repository copy of Genome structural evolution in Brassica crops.

White Rose Research Online URL for this paper: <u>https://eprints.whiterose.ac.uk/173736/</u>

Version: Accepted Version

Article:

He, Zhesi orcid.org/0000-0001-8335-9876, Ji, Ruiqin, Havlickova, Lenka orcid.org/0000-0002-5874-8615 et al. (17 more authors) (2021) Genome structural evolution in Brassica crops. Nature Plants. pp. 757-765. ISSN 2055-026X

https://doi.org/10.1038/s41477-021-00928-8

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk https://eprints.whiterose.ac.uk/ 1 Genome structural evolution in *Brassica* crops

3	Zhesi He ¹ , Ruiqin Ji ^{1†} , Lenka Havlickova ¹ , Lihong Wang ¹ , Yi Li ¹ , Huey Tyng Lee ² , Jiaming Song ³ ,
4	Chushin Koh ⁴ , Jinghua Yang ⁵ , Mingfang Zhang ⁵ , Isobel A.P. Parkin ⁶ , Xiaowu Wang ⁷ , David
5	Edwards ⁸ , Graham J King ⁹ , Jun Zou ³ , Kede Liu ³ , Rod J Snowdon ² , Surinder S. Banga ¹⁰ , Ivana
6	Machackova ¹¹ and Ian Bancroft ^{1*}
7	
8	¹ Department of Biology, University of York, Heslington, York, YO10 5DD, UK
9	² Department of Plant Breeding, Justus Liebig University of Giessen, 35392 Giessen, Germany
10	³ National Key Laboratory of Crop Genetic Improvement, College of Plant Science & Technology,
11	Huazhong Agricultural University, Wuhan, China
12	⁴ Global Institute for Food Security (GIFS), 110 Gymnasium Place, University of Saskatchewan,
13	Saskatoon, SK S7N 0W9 Canada
14	⁵ Department of Horticulture, College of Agriculture & Biotechnology, Zhejiang University, China,
15	310058
16	⁶ Agriculture and Agri-Food Canada, 107 Science Place Saskatoon, SK, S7N OX2
17	⁷ Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences (IVF, CAAS),
18	Beijing, China
19	⁸ School of Biological Sciences and the Institute of Agriculture, Faculty of Science, The University
20	of Western Australia, Crawley, WA, Australia
21	⁹ Southern Cross Plant Science, Southern Cross University, Lismore, NSW 2480, Australia
22	¹⁰ Department of Plant Breeding and Genetics, Punjab Agricultural University, Ludhiana, India
23	¹¹ Selgen, a.s., Plant breeding station, Chlumec nad Cidlinou, 503 51, Czech Republic
24	[†] Current address: Department of Horticulture, Shenyang Agricultural University, Shenyang
25	110866, China
26	* Corresponding author Prof Ian Bancroft: ian.bancroft@york.ac.uk
27	
28	
29	

30 Abstract

The cultivated *Brassica* species include numerous vegetable and oil crops of global importance. 31 32 Three genomes (designated A, B and C) share mesohexapolyploid ancestry and occur both singly 33 and in each pairwise combination to define the *Brassica* species. With organisational errors (such 34 as misplaced genome segments) corrected, we showed that the fundamental structure of each of 35 the genomes is the same, irrespective of the species in which it occurs. This enabled us to clarify 36 genome evolutionary pathways, including updating the Ancestral Crucifer Karyotype (ACK) block 37 organisation and providing support for the *Brassica* mesohexaploidy occurring via a two-step 38 process. We then constructed genus-wide pan-genomes, drawing from genes present in any 39 species in which the respective genome occurs, which enabled us to provide a global gene 40 nomenclature system for the cultivated Brassica species and develop methodology to costeffectively elucidate the genomic impacts of alien introgressions. Our advances not only underpin 41 42 knowledge-based approaches to more efficient breeding of *Brassica* crops, it also provides an exemplar for the study of other polyploids. 43

44

45 The cultivated Brassica species include vegetable (e.g. cabbage, cauliflower, broccoli, pak choi, 46 mustard greens, turnip and rutabaga), and condiment (e.g. mustard rape) crops, as well as the 47 third largest source of vegetable oil globally (oilseed rape)¹. Like Arabidopsis thaliana, which has been widely used to study fundamental plant science, the Brassica species are members of the 48 49 Brassicaceae (Cruciferae or mustard) family. Ancestral genome reconstruction, based on genome sequences, provides insights into genome evolution². Three genomes and pairwise combinations 50 51 thereof distinguish the *Brassica* species. The A genome (AA; n=10) occurs in *B. rapa*, the B genome (BB; n=8) in *B. nigra* and the C genome (CC; n=9) in *B. oleracea*. These diploid genomes 52 53 also occur in each pairwise combination to form the amphidiploid allotetraploid species B. napus (AACC; n=19), *B. juncea* (AABB; n=18) and *B. carinata* (BBCC; n=17)³. Genome sequences are 54 available for most of the *Brassica* species⁴⁻⁸ and are consistent with earlier studies that showed the 55 *Brassica* "diploid" genomes to have mesopolyploid ancestry⁹⁻¹². Assessment of the gene content 56 57 shows a characteristic pattern of one segment showing greater gene retention (least fractionated,

LF) than the other two (which are more fractionated, MF1 and MF2). There are two hypothetical mechanisms for this biased fractionation: (1) a single step hexaploidization followed by expression dominance and loss of silenced genes, and (2) a two-step process, initially involving the formation of a tetraploid by alloploidization and fractionation, before hybridization to introduce a third genome, which shows less fractionation as it has been in a polyploid context for a shorter time.

64 Genetic studies indicated conserved organisation of the *Brassica* genomes, irrespective of the species in which they occur¹³, with extensive collinearity between protein-coding genes^{14,15}. 65 However, studies based on whole genome sequencing suggested considerable variability⁸. 66 67 Adaptation of methods used for high density genic single nucleotide polymorphism (SNP) markers^{16,17} enabled the development of a quality assurance process based on scoring SNP 68 69 markers tagging genes and the identification of inconsistencies in the purported organisation of the genome sequences¹⁸. Such corrections to the organisation of genome sequences are important for 70 71 avoiding misleading results in, for example, the assessment of genome collinearity or positional 72 cloning of genes.

73

74 Reference genome sequences do not represent the entire gene repertoire of a species. The 75 concept of the pan-genome overcomes this limitation by representing the complete genic makeup of a species¹⁹. A pan-genome has been constructed for *B. oleracea*, using a reference-guided 76 assembly approach²⁰. However, that resource represents only the *Brassica* C genome as it occurs 77 78 in *B. oleracea* only, i.e. it does not include the genes present in the C genome of *B. napus* and *B.* carinata. The reference-quided *B. napus* pan-genome²¹ is similarly limited. Cross-species pan-79 genomes would better represent the gene pools from which genetic variation can readily be 80 81 accessed for *Brassica* crop improvement.

82

Gene flow between species by horizontal transfer provides a source of genetic variation and
enables adaptation, particularly in plants^{22,23}. Widely used as a traditional plant breeding method
for broadening genetic diversity, this "alien introgression" approach has enormous scope for future
crop improvement²⁴. However, the technical difficulty of assessing the extent of introduced or

exchanged genome segments at the molecular level impairs a deeper understanding of genome
changes. An example is the fertility restoration locus for the Ogura cytoplasmic male sterility
system (*Rfo*), widely used in oilseed rape, which involves a large segment of the radish genome²⁵.
The gene responsible, a pentatricopeptide repeat (PPR) gene, is known²⁶, and genetically linked
molecular markers are available²⁷. However, a lack of genetic recombination within a very large
introgressed chromosomal segment had made the extent of linked radish sequences and the
substituted *Brassica* chromosome segment difficult to define²⁸.

94

In this study, we corrected errors in the organisation of *Brassica* genome sequences and used the
improved resource to elucidate the ancestry and evolution of the *Brassica* A, B and C genomes.
Using this knowledge, we defined cross-species pan-genomes to underpin a systematic gene
nomenclature system for use by the *Brassica* research community. We also used the new resource
to assess the complex genome impacts of alien introgression.

100

101

102 **Results**

103 <u>1. Establishing fundamental genome configurations for the *Brassica* genus</u>

104 Numerous draft genome sequences have been produced for *Brassica* species, with comparative 105 analyses suggesting substantial differences in genome organisation both between and within 106 species. However, these genome assemblies were not well controlled for misassembly that can 107 result either from repetitive sequences contributing to chimeric scaffolds, or from anomalous 108 placement of scaffolds based on sparse genetic linkage maps. Smaller scaffolds can be 109 particularly difficult to position as the linkage maps used to support the original genome sequence 110 assembly will include regions with limited resolution due to lack of polymorphism between the 111 parents or low recombination rates. In polyploids, polymorphisms between homoeologous regions 112 of genomes can be difficult to differentiate from allelic variation and can lead to mis-assignment of 113 scaffolds to incorrect genomes. To address these shortcomings, the availability of a high density of 114 polymorphic markers is important to assess the reliability of mapping by enabling the discrimination 115 of occasional mis-mapping markers from multiple markers correctly mapping a scaffold. Increasing

116 resolution, by using a larger mapping population to produce more recombination events, adds 117 relatively little if there are additional approaches, such as collinearity with other genomes that can 118 be drawn upon. Advances in high resolution linkage mapping using data from transcriptome re-119 sequencing¹⁷ enables the quality of assemblies to be revisited. For example, analysis of the first genome assembly for *B. juncea⁸* using genome-ordered graphical genotypes (GOGGs)¹⁸ enabled 120 rapid correction into a much-improved version²⁹. This demonstrated the utility of using multiple 121 122 linkage mapping populations to provide more comprehensive resolution across a genome by 123 complementing gaps in polymorphism and/or recombination, and that these do not necessarily 124 need to be made using the same species as that from which the genome sequence is derived, 125 where the same genome is represented in multiple species. In the present study, we visualized 126 genome assemblies for each diploid species using GOGGs that were generated using SNP 127 markers produced from transcriptome or genome re-sequencing of double haploid (DH) linkage 128 mapping populations. For each diploid genome, two high-density genetic maps were used that 129 represented the two corresponding allotetraploid species in which the respective diploid genome is 130 found (*i.e. B. napus* and *B. juncea* for the *B. rapa* Chiifu v3 A genome, *B. juncea* and *B. carinata* 131 for the *B. nigra* early draft NI100 B genome, *B. napus* and *B. carinata* for the *B. oleracea* TO1000 132 C genome). Many anomalies were detected, whereby blocks of adjacent markers clearly mapped 133 genetically to positions discordant with their physical placement in the genome assemblies. The 134 markers used for generating GOGGs can all be assigned to gene models, enabling comparative 135 analysis. A distinctive feature of such misplaced genome sequence segments is their lack of 136 collinearity to orthologous genes in both A. thaliana and T. parvula, whereas they show perfect 137 collinearity to positions indicated by linkage mapping. Thus, such non-collinearity is a good 138 indicator of problematic regions in genome assemblies and precise re-positioning can be achieved 139 using collinearity as a fine-scale guide. We therefore worked manually through the GOGG data 140 (visualized in MS Excel spreadsheets) for all three diploid *Brassica* genome sequences and moved 141 segments manually to achieve congruous linkage mapping and collinearity with A. thaliana and T. 142 parvula genomes, using the appropriate pairs of allotetraploid linkage mapping datasets. Each 143 potential rearrangement was cross-checked with each of the relevant allotetraploid linkage maps 144 and, remarkably, no mapping conflicts were identified. Our interpretation is that the fundamental

145 organisation of the Brassica A, B and C genomes is conserved across species. The resulting 146 GOGGs are illustrated in Figure 1 while the GOGG data, including details of the markers and gene 147 models, are provided in Supplementary Data 1, 2 and 3. The reorganised genomes, as 148 represented by ordered gene models, are provided in Supplementary Data 4. Due to their 149 polyploid ancestry and strong artificial selection by breeders, some individual accessions of 150 cultivated *Brassica* species are expected to have inherited common structural genome rearrangements, particularly in the allopolyploid species³⁰. Genome structural rearrangements are 151 expected to inhibit recombination by blocking pairing in hererozygotes³¹, and therefore cannot be 152 153 elucidated by linkage mapping. Consequently, in chromosome segments where no recombination 154 data useful for linkage mapping was available, the organisation of the corresponding diploid 155 genome sequence in this region was accepted by default.

156

157 <u>2. Defining collinearity relationships of the *Brassica* genomes</u>

158 Shared ancestry with a mesohexaploid progenitor results in extensive collinearity between the 159 Brassica A, B and C genomes. Most effort has focused on comparative analysis of the A and C 160 genomes. Linkage mapping with very high densities of markers confirmed large collinearity blocks 161 (extending to the scale of some whole chromosomes, for example A1 and C1) but with disruption of other parts of the genome into small, apparently non-collinear segments³². This lack of 162 163 uniformity in conserved synteny across the genomes may be due partly to noise in the analyses, 164 for example from genome assemblies containing chimeric scaffolds or by errors in linkage mapping 165 with molecular markers, the scoring of which can be complex in polyploids with closely related genomes, such as *B. napus*¹⁶. In order to develop a clear understanding of the structural genome 166 167 evolution leading to the organisation of the extant *Brassica* genomes from their most recent 168 common ancestor, we aimed first to identify the most reliable orthology relationships. To do this, 169 we undertook a 3-way BLAST similarity analysis between all gene models in each of the re-170 assembled *Brassica* A, B and C genomes. We considered as putative orthologous triplets sets of 171 genes that gave reciprocal top BLAST hits with each other. This conservative approach is 172 designed to minimise noise caused by spurious sequence similarities and resulted in the 173 identification of 22,691 triplets. The set of gene models was curated manually, resulting in a set of

174 21,328 orthologous triplets present in collinear segments across all three genomes, as shown in 175 Supplementary Data 5. Pairwise comparisons of the organisation of the genomes are illustrated in 176 Figure 2. Assessment of molecular markers (as used for the development of GOGGs) could be 177 used to confirm the positioning of small collinear segments by linkage mapping. These results 178 confirm that synteny can be conserved on the chromosome scale, but also that collinearity can 179 break down into relatively small genome segments when rearrangement commences.

180

181 <u>3. Identification of conserved ancestral genome blocks</u>

182 The Brassicaceae provide an excellent model for studying genome structural evolution. Pioneering 183 studies led to the development of a proposed "Ancestral Crucifer Karyotype" (ACK) comprising 24 collinearity blocks (labelled A to X)³³, with greater understanding developing as more genome 184 sequences emerged for the family³⁴. Our improved organisation of the *Brassica* genomes was 185 186 developed without reference to pre-existing knowledge of the ACK collinearity blocks, so a re-187 evaluation of them provides both the opportunity for new insights and a further quality check on our 188 resource. The genome triplication represented in the *Brassica* species occurred close to the time of 189 separation between the *Brassica* and *T. parvula* lineages, making the *T. parvula* genome a good representative of the pre-triplication genome inherited by *Brassica* species³⁵. We therefore used 190 191 the top T. parvula gene model BLAST hits with each of the 21,328 orthologous Brassica gene 192 triplets (Supplementary Data 5) to define collinearity blocks that we could subsequently align with the organisation with the previously defined ACK blocks². As shown in Table 1, the results 193 194 corresponded very well. Our analysis is entirely consistent with the previously-called ancestral 195 blocks, apart from an indication that two small additional blocks can be defined. We label these 196 new blocks as V2 and W2, with the blocks corresponding to the original V and W re-named V1 and 197 W1, respectively, in Table 1. All of our 136 *Brassica* ACK genome blocks could be identified in both 198 T. parvula and A. thaliana genomes apart from two blocks (which we refer to as Tp blocks 27.5 and 199 29.5; see Supplementary Data 6) that are not represented in the *T. parvula* genome sequence. 200 This set of 136 ACK genome blocks can be arranged to represent any of the Brassica A, B or C 201 genomes or, with 3-fold redundancy, the diploid Brassicaceae (i.e. A. thaliana or T. parvula) 202 genomes.

203

204 <u>4. Inference of the origins of paralogous sub-genome segments</u>

205 One of the unresolved questions in the evolution of *Brassica* genomes is the mechanism by which 206 one sub-genome exhibits less gene loss (fractionation) than the other two, when compared with an 207 orthologous Brassicaceae genome such as A. thaliana or T. parvula. We used our improved 208 organisation of the *Brassica* genomes to address this, inferring, at least partially, the chromosomal 209 organisation of their common ancestor prior to speciation. Assuming the rate of structural 210 rearrangement of the subgenomes is approximately constant over time, an approximately equal 211 rearrangement would be expected in each sub-genome if they had come together simultaneously 212 to form a hexaploid, which then preserved one less fractionated sub-genome by a mechanism 213 such as methylation to substantively silence the other two. Alternatively, a two-stage process, in 214 which two genomes fractionated for an extended period of time in a tetraploid before addition of a 215 third genome formed the hexaploid, would result in less rearrangement of the genome that joined 216 most recently, i.e. the less fractionated sub-genome. The approach we used to distinguish between 217 these two scenarios was to manually assess the 134 ACK collinearity blocks with collinearity to T. 218 parvula for fusions represented in the *Brassica* A, B and C genomes, which represent their 219 configurations in the ancestors of the *Brassica* genomes. These could be sorted, based on 220 complementarity of break positions between blocks, into larger blocks and a least fractionated 221 block identified based on gene content in *Brassica* as a proportion of all orthologues in the 222 corresponding region of the *T. parvula* genome (Supplementary Data 7). The 136 ACK collinearity 223 blocks (including the two with no T. parvula orthologues) were then assembled into their putative 224 configuration in the ancestors of the three *Brassica* genomes, based on adjacencies in the extant 225 Brassica genomes (Figure 3, Supplementary Data 8). The least fractionated blocks almost 226 perfectly represent 7 ancestral chromosomes that exhibit excellent collinearity with the T. parvula 227 genome, with only one small block (containing orthologues of Tp6g07740-Tp6g09710, positioned 228 in chromosome 6) having been classified as more fractionated. Ten additional putative ancestral 229 chromosomes were identified that comprise more fractionated blocks, with only one exception: one 230 small block contains orthologues of Tp7g11790-Tp7g11970 (positioned in chromosome 14) that 231 had been classified as least fractionated. These chromosomes exhibited numerous fusions and

rearrangements, indicating extensive rearrangement since formation of the polyploid. Our findings therefore support the two-stage process, with the hexaploid being formed by the fusion of a third genome, very similar in organisation to that of *T. parvula*, with a tetraploid that had already been undergoing fractionation and rearrangement.

236

237 <u>5. Construction of cross-species pan-genomes</u>

238 Having developed a robust understanding of the organisation and evolution of the Brassica A, B 239 and C genomes, we next aimed to optimise our knowledge of the gene complement of the genus. 240 Because we have shown that the fundamental organisation of the genomes is the same 241 irrespective of the species, and each genome is shared across three different species of the 242 genus, we aimed to develop the first genus-wide pan-genomes. Doing this in a robust and 243 comprehensive manner required the development of a new approach to pan-genome construction. 244 The selection of the most appropriate underpinning representative of each genome is important for 245 minimising errors and to provide a resilient gene nomenclature for the genus. First, we evaluated 246 genome sequence resources for completeness and fidelity of organisation, using GOGGs. We 247 concluded, on the basis of least errors in organisation, that the most reliable were the genome sequences of *B. rapa* Z1 for the A genome³⁶, *B. nigra* Ni100 for the B genome³⁷ and *B. oleracea* 248 HDEM for the C genome³⁶. We integrated unanchored scaffolds as listed in Supplementary Data 9 249 250 then anchored (based on BLAST similarity) or interpolated (based on collinearity of flanking genes) 251 gene models from 12 additional genome sequence resources, as listed in Table 2. These genome 252 sequences were released previously to our analysis or in parallel with them; we draw upon their 253 gene content. The gene models comprising the pan-genome were given systematic pan-genome 254 names, as approved by the Steering Committee of the Multinational Brassica Genome Project. 255 These are shown in Supplementary Data 10 (for the A genome), Supplementary Data 11 (for the B 256 genome) and Supplementary Data 12 (for the C genome), along with other details of their source 257 genomes. Of the 197,465 gene models in the *Brassica* pan-genomes, 165,698 (83.9%) were 258 already represented in the underpinning genomes. Of these, 104,339 (63.0%) showed significant 259 similarity to orthologues in *T. parvula* and 93,361 (56.3%) showed significant similarity to 260 orthologues in A. thaliana. Lower proportions of genes showed significant similarity for the B and C

genomes (58.8% and 57.1%, respectively) than did gene models from the A genome (68.3%). A
substantially lower proportion of the 31,767 gene models integrated from additional genome
sequences to form the pan-genomes showed significant similarity to orthologues in *T. parvula*(12.0%, 12.5% and 10.1% for A, B and C genomes, respectively).

265

266 <u>6. Using the pan-genomes to elucidate the genomic impacts of alien trait introgression</u>

267 Alien introgression is an important approach to broadening the genetic diversity of crops by 268 exchanging genetic material between related species. The genomic impacts have been difficult to 269 assess, restricting the application of the system to targets of high commercial value, such as the 270 introduction of components of F₁ hybrid production systems. However, there is growing demand for 271 novel genetic resistances to both biotic and abiotic stresses, and strong allopolyploidization and 272 breeding bottlenecks necessitate the introduction of novel genetic diversity for such traits from 273 secondary or tertiary gene pools. Having generated a comprehensive platform for describing 274 genomic variation in the cultivated *Brassica* species, we tested whether it could be used to support 275 genome analysis of alien introgression lines. As a first example, we assessed the widely-used 276 Ogura fertility restorer system. The restorer gene carried by the *Rfo* locus (the orthologue of which 277 in our pan-genome is C09p019070.1 BoIHDE) was introgressed into *B. napus* (oilseed rape) from 278 the closely-related species Raphanus sativus (radish), however the extent of the Brassica genome 279 which was replaced has not been described at the gene level. As a second example, we screened 280 a population of lines with putative introgressions into *B. juncea* from *Brassica fruticulosa*. Because 281 well-assembled genome sequences are seldom available for wild relatives, we implemented a 282 method termed "curing" to edit a Brassica pan-genome to more closely represent that of a donor species³⁸. For the introgression into *B. napus* (which contains the *Brassica* A and C genomes) we 283 284 cured the *Brassica* B pan-genome with *R. sativus* genome sequence reads and re-named this 285 reference genome R. For the introgression into *B. juncea* (which contains the *Brassica* A and B 286 genomes) we cured the Brassica C pan-genome with B. fruticulosa genome sequence reads and 287 re-named this reference genome F. We then mapped genome sequence reads from recipient and 288 donor parents, and introgression lines, to the appropriate triple reference sequence (ACR or ABF) 289 and processed as described previously for the visualization of homoeologous exchanges in

polyploid species³⁰. The resulting Genome Display Tile Plots (GDTPs) are shown in Figure 4, while
a quantification of reads mapping to each gene in the reference is provided in Supplementary Data
13 for the radish intogression and Supplementary Data 14 for the *B. fruticulosa* intogression. This
approach clearly delineated the extent of the *R. sativus* introgression into *B. napus* and a *B.
<i>fruticulosa* introgression into *B. juncea*. Other putative *B. fruticulosa* introgression lines showed no
evidence of capturing genomic DNA from the donor species, but did show segmental losses from
the recipient genome.

297

As there is a genome sequence available for *R. sativus*, we repeated the analysis for the *Rfo* introgression using the radish genome sequence as the R genome reference. This yielded the same result as the cured reference sequence (Extended Data Figure 1), but with the advantage that visualization was also possible based on the genome order of the radish genes. Testing for the radish introgression, similar results could also be obtained using mRNAseq reads instead of gDNA reads (Extended Data Figure 2), making the approach cost-effective even for species with very large genomes where genome re-sequencing would be prohibitively expensive.

305

306

307 Discussion

308 Rigorous assessment of genome sequences available for *Brassica* species enabled us to show 309 conclusively that that the fundamental organisation of the *Brassica* A, B and C genomes is 310 conserved across the multiple species in which they occur. By identifying triplets of orthologous 311 genes across all three genomes, we were able to define collinearity relationships between the 312 genomes, demonstrating a wide range of sizes for blocks of collinearity, ranging from a few genes 313 to whole chromosomes. Using the resource, in particular the collinearity observed with the genome 314 of T. parvula (as the best extant representative of the genome structure of the ancestor of Brassica 315 species prior to genome triplication), we were able to re-visit analyses of ancestral genome 316 evolution based on collinear genome blocks traceable to the ACK. We identified two additional 317 blocks (V2 and W2) and showed the existence of blocks in the *Brassica* genomes for which no 318 orthologous segments exist in the T. parvula genome sequence, providing 136 genome blocks that

319 can be rearranged to represent the structure of any of the Brassica genomes. Using these 320 structures, we were able to infer the structure of the least fractionated *Brassica* sub-genome, 321 showing that prior to rearrangement it comprised 7 chromosomes of similar structure to those 322 present in *T. parvula*. In contrast, the more fractionated sub-genomes appear to be derived from 323 two genomes that had undergone previous rearrangement, supporting the hypothesised 2-step 324 derivation of the mesohexaploid structure, initially involving the formation of a tetraploid and a 325 period of rearrangement and fractionation, before further hybridization to introduce a third genome, 326 which shows less fractionation as it has been in a polyploid context for a shorter time.

327

328 In addition to understanding more of the evolutionary process by which the Brassica genomes 329 evolved, we developed resources to underpin future research and breeding in the many 330 commercially important *Brassica* crops. First, we established the first genus-wide pan-genomes for 331 any species. Of the 197,465 gene models in the combined *Brassica* A, B, and C pan-genomes, the 332 majority (165,698) were already represented in the underpinning genomes and most of these 333 (>55%) showed significant similarity to orthologues in *T. parvula*. A notably lower proportion of the 334 31,767 gene models integrated from additional genome sequences to form the pan-genomes 335 showed such similarity to orthologues in *T. parvula* (<13%), which may indicate that these contain a greater proportion of spurious gene models. This genus-wide pan genome resource provides a 336 337 framework for describing *Brassica* gene content, via a systematic new nomenclature approved by 338 the Multinational *Brassica* Genome Project. We demonstrated the practical utility of this resource to 339 support crop breeding to introduce traits from related species by developing a novel approach for 340 the first cost-effective analysis of alien introgressions at the level of gene sequences, providing 341 examples of a radish introgression into *B. napus* and a *B. fruticulosa* introgression into *B. juncea*. 342

343

344 **References**

USDA-FAS. USDA Oilseeds: World markets and trade, 2020).
 Murat, F. et al. Understanding Brassicaceae evolution through ancestral genome reconstruction.
 Genome biology 16, 262 (2015).

348 240	3	Nagaharu, U. Genome analysis in Brassica with special reference to the experimental formation of Reparts and posuliar mode of fortilization. <i>Inn L Pot</i> 7 , 289, 452 (1925)
343	л	B. hapus and peculial mode of refinization. <i>Jph J Bol</i> 7, 365-452 (1955).
350 351	4	1035 (2011).
352	5	Liu, S. et al. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid
353		genomes. Nature communications 5, 3930, doi:10.1038/ncomms4930 (2014).
354	6	Parkin, I. et al. Transcriptome and methylome profiling reveals relics of genome dominance in the
355		mesopolyploid Brassica oleracea. <i>Genome Biology</i> 15 , R77 (2014).
356	7	Chalhoub, B. et al. Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed
357		genome. <i>Science</i> 345 , 950-953, doi:10.1126/science.1253435 (2014).
358	8	Yang, J. <i>et al.</i> The genome sequence of allopolyploid Brassica juncea and analysis of differential
359		homoeolog gene expression influencing selection. <i>Nature genetics</i> 48 , 1225-1232.
360		doi:10.1038/ng.3657 (2016).
361	9	Lagercrantz, U. & Lydiate, D. J. Comparative genome mapping in Brassica. <i>Genetics</i> 144 , 1903-1910
362	-	(1996).
363	10	O'neill, C. M. & Bancroft, I. Comparative physical mapping of segments of the genome of Brassica
364		oleracea var. alboglabra that are homoeologous to sequenced regions of chromosomes 4 and 5 of
365		Arabidopsis thaliana. The plant journal 23 , 233-243 (2000).
366	11	Yang, TJ. et al. Sequence-level analysis of the diploidization process in the triplicated FLOWERING
367		LOCUS C region of Brassica rapa. The Plant cell 18, 1339-1347 (2006).
368	12	Town, C. D. et al. Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene
369		loss, fragmentation, and dispersal after polyploidy. The Plant cell 18, 1348-1359 (2006).
370	13	Parkin, I. A., Sharpe, A. G., Keith, D. J. & Lydiate, D. J. Identification of the A and C genomes of
371		amphidiploid Brassica napus (oilseed rape). Genome / National Research Council Canada = Genome
372		/ Conseil national de recherches Canada 38 , 1122-1131, doi:10.1139/g95-149 (1995).
373	14	Rana, D. et al. Conservation of the microstructure of genome segments in Brassica napus and its
374		diploid relatives. <i>Plant Journal</i> 40 , 725-733, doi:10.1111/j.1365-313X.2004.02244.x (2004).
375	15	Cheung, F. et al. Comparative analysis between homoeologous genome segments of Brassica napus
376		and its progenitor species reveals extensive sequence-level divergence. The Plant cell 21, 1912 -
377		1928 (2009).
378	16	Trick, M., Long, Y., Meng, J. & Bancroft, I. Single nucleotide polymorphism (SNP) discovery in the
379		polyploid Brassica napus using Solexa transcriptome sequencing. <i>Plant biotechnology journal</i> 7,
380		334-346, doi:10.1111/j.1467-7652.2008.00396.x (2009).
381	17	Bancroft, I. et al. Dissecting the genome of the polyploid crop oilseed rape by transcriptome
382		sequencing. Nature biotechnology 29, 762-766, doi:10.1038/nbt.1926 (2011).
383	18	He, Z. & Bancroft, I. Organization of the genome sequence of the polyploid crop species Brassica
384		juncea. <i>Nature genetics</i> 50 , 1496-1497 (2018).
385	19	Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. <i>Current</i>
386		opinion in microbiology 23 , 148-154 (2015).
387	20	Golicz, A. A. <i>et al.</i> The pangenome of an agronomically important crop plant Brassica oleracea.
388		Nature communications 7, 13390 (2016).
389	21	Dolatabadian, A. et al. Characterization of disease resistance genes in the Brassica napus
390		pangenome reveals significant structural variation. <i>Plant biotechnology journal</i> (2019).
391	22	Mallet, J. Hybridization as an invasion of the genome. <i>Trends in ecology & evolution</i> 20 , 229-237
392		(2005).
393	23	Arnold, M. L. Transfer and origin of adaptations through natural hybridization: were Anderson and
394		Stebbins right? <i>The Plant cell</i> 16 , 562-570 (2004).
395	24	Zamir, D. Improving plant breeding with exotic genetic libraries. <i>Nature reviews. Genetics</i> 2, 983-
396		989, doi:10.1038/35103589 (2001).
397	25	Delourme, R., Horvais, R., Vallée, P. & Renard, M. Double low restored F1 hybrids can be produced
398		with the Ogu-INRA CMS in rapeseed. in Proc. 10th Int. Rapeseed Congress.

399	26	Brown, G. G. et al. The radish Rfo restorer gene of Ogura cytoplasmic male sterility encodes a
400		protein with multiple pentatricopeptide repeats. The Plant journal : for cell and molecular biology
401		35 , 262-272, doi:10.1046/j.1365-313X.2003.01799.x (2003).
402	27	Hu, X. et al. Mapping of the Ogura fertility restorer gene Rfo and development of Rfo allele-specific
403		markers in canola (Brassica napus L.). Molecular breeding 22, 663-674 (2008).
404	28	Feng, J. et al. Physical localization and genetic mapping of the fertility restoration gene Rfo in
405		canola (Brassica napus L.). Genome / National Research Council Canada = Genome / Conseil
406		national de recherches Canada 52 , 401-407 (2009).
407	29	Yang, J., Ji, C., Liu, D., Wang, X. & Zhang, M. Reply to: 'Organization of the genome sequence of the
408	-	polyploid crop species Brassica juncea', <i>Nature genetics</i> 50 , 1497-1498, doi:10.1038/s41588-018-
409		0240-7 (2018).
410	30	He 7 <i>et al.</i> Extensive homoeologous genome exchanges in allopolyploid crops revealed by
411		mRNAseq-based visualization <i>Plant histechnology journal</i> 15 594-604 doi:doi:10.1111/nbi.12657
412		(2017)
413	31	Crown K N Miller D F Sekelsky I & Hawley R S Local inversion heterozygosity alters
414	51	recombination throughout the genome <i>Current Biology</i> 28 , 2984-2990, e2983 (2018)
717 //15	37	Bancroft I Fraser F. Morgan C & Trick M Collinearity analysis of Brassica A and C genomes
416	52	hased on an undated inferred unigene order. Data in brief 3 , 51-55 (2015)
410 //17	22	Schranz M. Lysak M. & Mitchell-Olds T. The ABC's of comparative genomics in the Brassicaceae:
417 //18	55	building blocks of crucifer genomes Trends Plant Sci 11 535 - 542 (2006)
410 //10	3/	Lycak M A Mandáková T & Schranz M E Comparative naleogenomics of crucifers: ancestral
410	54	genomic blocks revisited <i>Current Oninion in Plant Biology</i> 20 , 108-115 (2016)
420	25	Change E at al. Deciphering the diploid apportant biology 50 , 108-115 (2010).
421	22	Dignt coll 25 15/1-155/ (2013)
422	36	Relser C et al Chromosome-scale assemblies of plant genomes using panonore long reads and
425	50	ontical mans. Nature plants 4, 879 (2018)
424	37	Derumal S et al. A high-contiguity Brassica nigra genome localizes active centromeres and defines
425	57	the ancestral Brassica genome Nature Plants 6, 929-941 (2020)
420	20	Higgins I. Maguein A. Trick M. Erssor E & Paneroft I in <i>BMC ganomics</i> Vol 12, 247 (2012)
427	30 20	Camacho C et al. BLAST+: architecture and applications. BMC bioinformatics 10, 421 (2002).
420	10	Li H & Durbin R East and accurate short read alignment with Burrows-Wheeler transform
425	40	bioinformatics 25 , 1754, 1760 (2000)
430	/11	Li H et al. The sequence alignment/man format and SAMtools <i>Bioinformatics</i> 25 , 2078-2079
431	41	
432	10	(2003). Zhang Last al Improved Brassica rang reference geneme by single melocule sequencing and
455	42	chromosomo conformation capture technologies. Herticulture recearch E 1 11 (2019)
434	13	Zou L et al Genome-wide selection footnrints and deleterious variations in young Asian
435	45	allototraploid raposood. <i>Plant biotechnology journal</i> (2010)
430	4.4	Song L M of all Fight high quality gamping rayoal nan gamping architecture and ecotype
437	44	differentiation of Practice names. Nature Plants 1 12 (2020)
430	15	Lee H et al Chromosome-scale assembly of winter oilseed rane Brassica papus Frontiers in Plant
439	43	Science 11 406 (2020)
440		<i>Science</i> 11 , 450 (2020).
441		
442		
443	Ackno	owledgments
444	This w	ork was supported by: UK Biotechnology and Biological Sciences Research Council

- BB/L002124/1 and BB/R019819/1 to IB, the China Agriculture Research System CARS-25-A-03
- and the Natural Science Foundation of Liaoning Province, China 2013020071 to RJ, grants

447	031B0890A from BMBF and SN14/22-1 from DFG to RJS and HTL, Australia Research Council
448	Project LP160100030 to DE, National Natural Science Foundation of China 31970564 to JZ, and
449	Indian Council of Agricultural Research F.No.27(5)/2007-HRD and Department of Biotechnology
450	and Government of India BT/01/CEIB/12/I/03 to SSB.
451	
452	
453	Author contributions
454	IB conceived the work. ZH, RJ, LH, IM and IB designed the experiments. ZH, RJ, LH, LW, YL,
455	HYL, JS, CK, JY, MZ, IAPP, XW, DE, GJK, JZ, KL, RJS, SSB acquired, analysed and/or
456	interpreted the data. ZH and IB drafted the manuscript.
457	
458	
459	Competing interests statement
460	The authors declare no competing interests.
461	
462	
463	Figure legends
464	
465	Figure 1. Genome-ordered graphical genotypes for the Brassica A, B and C genomes as
466	represented in allotetraploid species. Graphical genotypes are shown for transcriptome or genome
467	SNP markers scored across three doubled haploid (DH) linkage mapping populations: (1) 119 lines
468	of the Varuna x Heera (VHDH) mapping population for A genome <i>B. juncea</i> and B genome <i>B.</i>
469	juncea (Heera alleles in coral, Varuna alleles in blue and missing scores in white). (2) 45 lines of
470	the Tapidor x Ningyou 7 (TNDH) mapping population for A genome <i>B. napus</i> and C genome <i>B.</i>
471	napus (Ningyou 7 alleles in coral, Tapidor alleles in blue and missing scores in white). (3) 93 lines
472	of the Yellowcross x Whiteban (YWDH) mapping population for B genome B. carinata and C
473	genome B. carinata (Whiteban alleles in coral, Yellowcross alleles in blue and missing scores in
474	white). The multi-coloured bars are colour-coded by the top BLAST sequence similarity match to
475	the chromosomes in <i>Arabidopsis thaliana</i> (left bar) and <i>Thellungiella parvula</i> (right bar) of the 15

476	Brassica gene model in which each respective SNP is scored (light blue = chromosome 1, orange
477	= chromosome 2, dark blue = chromosome 3, green = chromosome 4, red = chromosome 5,
478	salmon = chromosome 6, yellow = chromosome 7, light grey = no BLAST hit with E-value < $1e^{-30}$).
479	
480	
481	Figure 2. Collinearity of Brassica A, B and C genomes. 21,328 triplets of orthologous genes in the
482	Brassica A, B and C genomes are plotted by their order in the respective genomes: (a) Brassica A
483	and C genomes, (b) Brassica A and B genomes and (c) Brassica C and B genomes.
484	
485	
486	Figure 3. Inferred structure of chromosomes in the nascent triplicated ancestral genome. Partial
487	hypothesised genome structure is depicted in relation to genome blocks collinear between
488	Brassica and T. parvula, with the range of T. parvula orthologues indicated for each block and
489	colour-coded by chromosome (Tp1 = light blue; Tp 2 = orange; Tp 3 = dark blue, Tp 4 = green, Tp
490	5 = red, Tp 6 = yellow, Tp 7 = beige). The genome segments identified in <i>Brassica</i> for which there
491	is no identified <i>T. parvula</i> orthologue are shown as grey blocks. Hypothesised ancestral
492	chromosomes 1 to 7 comprise Brassica blocks identified as least fractionated and 8 to 17 comprise
493	Brassica blocks identified as more fractionated, except those illustrated with T. parvula orthologue
494	names in red font. Arrows indicate the relative orientations of the inferred contiguous blocks.
495	Chromosome numbers are assigned arbitrarily, except 1 to 7, which are assigned numbers to
496	match the numbering of the orthologous <i>T. parvula</i> chromosomes.
497	
498	
499	
500	Figure 4. Visualization of genomic impacts of alien introgression into allotetraploid Brassica
501	species. Genome Display Tile Plots were generated based on the relative abundance of genome
502	sequence reads mapping to three reference genomes, two of which are genomes of the
503	introgression recipient species and the third a cured genome representing the introgression donor
504	species. Quantification is represented in CMYK colour space for orthologous gene triplets. (a) 16

505 Rapeseed Ogura hybrid system with a radish (Raphanus sativus) introgression in Brassica napus. 506 The cyan component represents abundance of the *Brassica* A genome orthologue, the yellow 507 component that of the *Brassica* C genome orthologue and the magenta component that of the 508 radish (R) genome orthologue, produced by curing. The triplets are plotted in Brassica C genome 509 order, along with controls comprising parental species and *in silico* combinations to render a 510 diagnostic colour key. Three plants representing the male sterile (CMS) plants of the hybrid system 511 (no introgression) and three plants containing the radish introgression harbouring the restorer (Rfo) 512 gene are illustrated. (b) Mustard rape (Brassica juncea) lines with Brassica fruticulosa 513 introgression. The cyan component represents abundance of the *Brassica* A genome orthologue, 514 the yellow component that of the *Brassica* B genome orthologue and the magenta component that 515 of the *B. fruticulosa* (F) genome orthologue, produced by curing. The triplets are plotted in *Brassica* 516 A genome order, with only A1 shown, along with controls comprising parental species and *in silico* 517 combinations to render a diagnostic colour key. Four putative introgression lines are illustrated, 518 including one with no introgression in this chromosome (AD-19-003), one that has inherited a large 519 heterozygous deletion of the orthologous part of the *Brassica* B genome (AD-19-014), one that has 520 inherited a small homozygous deletion of the *Brassica* A genome (AD-19-027) and one confirmed 521 introgression line which has inherited a substitution of part of the Brassica A genome by B. 522 fruticulosa sequences (AD-19-037).

523

524

525 Tables

	-	I		·
ACK block	At start	Tp start	At end	Tp end
A	AT1G01030	Tp1g00050	AT1G19530	Tp1g17400
В	AT1G19570	Tp1g17410	AT1G42990	Tp1g32200
С	AT1G43130	Tp1g32290	AT1G56180	Tp1g41870
D	AT1G64670	Tp2g00040	AT1G56230	Tp2g06750
E	AT1G64960	Tp5g19220	AT1G80950	Tp5g36020
F	AT3G01040	Tp3g00010	AT3G25520	Tp3g23020

526 <u>Table 1. Ancestral Crucifer Karyotype genome blocks identified across the *Brassica* genomes.</u>

G	AT2G04050	Tp3g23100	AT2G06200	Tp3g24630
Н	AT2G11520	Tp3g26330	AT2G20900	Tp3g34170
1	AT2G20920	Tp4g00010	AT2G29710	Tp4g12510
J	AT2G29980	Tp4g12540	AT2G48140	Tp4g30080
К	AT2G01070	Tp2g12540	AT2G04038	Tp2g14790
L	AT3G25545	Tp2g14810	AT3G30975	Tp2g18310
М	AT3G42640	Tp5g18280	AT3G49730	Tp5g12350
N	AT3G49790	Tp5g12330	AT3G63530	Tp5g00010
0	AT4G00080	Tp6g00060	AT4G05420	Tp6g04790
Р	AT4G12600	Tp6g04910	AT4G08320	Tp6g09620
Q	AT5G29560	Tp2g19970	AT5G23090	Tp2g24020
R	AT5G23000	Tp6g22390	AT5G01010	Tp6g41230
S	AT5G42100	Tp7g00010	AT5G35220	Tp7g07090
Т	AT4G12650	Tp7g09040	AT4G16630	Tp7g15250
U	AT4G16765	Tp7g15470	AT4G40080	Tp7g37560
V1	AT5G42220	Tp7g08390	AT5G42420	Tp7g07940
V2	AT5G42490	Tp2g06900	AT5G47780	Tp2g12520
W1	AT5G47820	Tp2g18320	AT5G49620	Tp2g19930
W2	AT5G49660	Tp6g10770	AT5G60800	Tp6g22380
Х	AT5G60820	Tp2g23880	AT5G67630	Tp2g30710
At start = BL	AST hits of Bras	<i>ssica</i> CDS gene	e models to Aral	bidopsis
thaliana CDS	S gene models:	lowest gene mo	odel number	
At end = BL	AST hits of <i>Bras</i>	<i>sica</i> CDS gene	models to Arab	idopsis
thaliana CDS	S gene models:	highest gene m	nodel number	
Tp start = Bl	_AST hits of Bra	<i>ssica</i> CDS gen	e models to The	ellungiella
parvula CDS	6 gene models: I	owest gene mo	odel number	
Tp end = BL	AST hits of Bras	<i>ssica</i> CDS gene	e models to The	llungiella
parvula CDS	S gene models: I	highest gene m	odel number	

529 <u>Table 2. Sources of 197,465 gene models in cross-species *Brassica* pan-genomes.</u>

Source	Reference	Code	Order	A ge	nome	B ge	nome	C gei	nome
			added	gene r	nodels	gene r	nodels	gene r	nodels
				Total	Tp hits	Total	Tp hits	Total	Tp hits
B. rapa Z1	Belser et al ³⁶	BraZAA	1	45819	31322				
	2018								
<i>B. nigra</i> Ni100	Perumal et al ³⁷	BniNIA	1			59422	34940		
	2020								
<i>B. oleracea</i> HDEM	Belser et al ³⁶	BolHDE	1					60457	34539
	2018								
<i>B. rapa</i> Chiifu v3	Zhang et al42	BraCHC	2	2863	592				
	2018								
<i>B. nigra</i> YZ12151	Yang et al ⁸	BniYZA	2			598	69		
	2016								
<i>B. oleracea</i> Pangenome	Golicz et al ²⁰	BolBOP	2					1888	581
	2016								
<i>B. rapa</i> R-o-18	King et al	BraROA	3	3387	172				
	unpublished								
<i>B. oleracea</i> 02-12	Liu et al ⁵ 2014	BolBOA	3					594	146
<i>B. napus</i> Pangenome	Dolatabadian	BnaBNP	4	2409	201			7724	314
	et al ²¹ 2019								
<i>B. napus</i> Ningyou 7	Zou et al ⁴³	BnaNYA	5	1241	111			2950	161
	2019								
<i>B. napus</i> ZS11	Song et al44	BnaZSA	6	500	108			775	125
	2020								
<i>B. napus</i> Express 617	Snowdon et	BnaEXA	7	111	33			184	41
	al ⁴⁵								

<i>B. juncea</i> T84v2	Yang et al ⁸	BjuTUA	8	858	121	1630	222		
	2016								
<i>B. juncea</i> AU213	Yang et al	BjuAUA	9	370	73	661	80		
	unpublished								
<i>B. carinata</i> 080798EM	Parkin et al	BcaBCA	10			1319	157	1705	231
	unpublished								
Underpinning genome				45819	31322	59422	34940	60457	34539
Added				11739	1411	4208	528	15820	1599
Total				57558	32733	63630	35468	76277	36138
Tp hits = BLAST hits of I	<i>Brassica</i> CDS ger	ne models	to Thell	lungiella	parvula	CDS ge	ne mod	els	

530

531

532 Methods

533 <u>Reorganising genome resources</u>

534 GOGGs analysis was performed on each of the A, B, C genome resources (B. rapa Chiifu v3 for the A genome, *B. nigra* Ni100 for the B genome and *B. oleracea* TO1000 for the C genome), 535 essentially as described previously¹⁸. The MS Excel (version 2016) spreadsheets representing the 536 537 GOGGs were displayed with gene models ordered by their coordinate in the genome sequence 538 resource, at 1-pixel row height and screen shots compiled into MS PowerPoint slides for 539 visualization. These were scanned manually for discontinuities in the orderly transition within each 540 line between alleles originating from each of the two parents. When such anomalies (which 541 resemble what would be observed in linkage mapping as apparent double recombinants at the 542 same point in many individual mapping lines) were detected, we manually generated a 543 standardized formatted table for correcting misplaced segments of genomes, with correct 544 placement based on where in the genome the pattern of alleles present in each line of the mapping 545 population best matches. Fine placement (*i.e.*, between precisely which two gene models the segment should be inserted) was determined to preserve collinearity with the A. thaliana genome, 546

- 547 based on top BLAST similarity to *A. thaliana* CDS gene models. These tables and the original
- 548 genome resource files were then taken into an automated R script,
- 549 Genome_Sequence_Reorganise (deposited on GitHub
- 550 https://github.com/hezhesi/Genome_Sequence_Reorganise), to generate the revised genome
- sequence and annotation files. This improvement was undertaken as an iterative process to refine
- genome organisation. As an example, the final iteration of construction is illustrated in Extended
- 553 Data Figure 3 to show mis-placed chromosome segments and unanchored sequence scaffolds,
- with the editing file for construction of the final pseudomolecules presented as Supplementary Data
- 555 15.
- 556

557 <u>Gene anchoring and interpolation</u>

- 558 R (version 3.6) scripts were used to anchor or interpolate gene models from additional genome
- sequence resources with the following steps. Based on the quality and completeness of the
- 560 genome resources, the scripts were run for genome resources, in the order shown in Table 2. The
- 561 process was:
- 562 i. Align gene models onto the pseudomolecules, align CDS gene models (separate by
- 563 chromosomes) onto the previous round of globalABC CDS models on each chromosome using
- 564 BLASTN³⁹ (Version: 2.6.0+). Only top hits and Evalue < 1e-30 were included.
- ii. Sort them by their locations. Also using AT and TP models as a guidance to determine if they
- are anchored correctly.
- 567 iii. Add all extra CDS models and integrate a new list
- 568 iv. Delete overlapped models on the existing gene space.
- v. Generate final ordered CDS model list for next round of processing.
- 570

571 Reference genome curing with donor species

- 572 "Curing" was performed on gene models from the *Brassica* pan-genome with DNA re-sequencing
- 573 data from donor species. This method was first developed and described by Higgins et al^{38} .
- Instead of running it on RNA-seq data for a transcriptome reference, DNA-seq data were used and

a cured reference set of gene models was created. The 150 base reads were split into three files,
each containing a set of 50 base reads using the Perl script illumina_split_read.pl. Using other Perl
scripts from Higgins et al³⁸, with default parameters, iterative mapping and comparing with
consensus was performed over six cycles after which there was no significant gain in alignment
efficiency. This process resulted in a cured *Raphanus* sativus L. (R) derived from the *Brassica* B
genome and *B. fruticulosa* (F) derived from the *Brassica* C genome. Then the combined ACR or
ABF genomes were used as a reference for read mapping.

582

583 <u>Genome quantification with cured reference</u>

584 Genome sequence reads were 150-base paired-end Illumina HiSegX reads obtained from DNA purified from leaves. BWA sequence-alignment program⁴⁰ (Version: 0.7.17-r1188) was used for 585 586 mapping genomic reads, using the appropriate 3-genome DNA gene model reference combination. 587 SAMtools⁴¹ (Version: 1.10) was used to index mapping results and score counts of each gene 588 model for each sample, then R was used to calculate the normalised as reads per kb per million 589 aligned reads (RPKM) values. In silico reads were simulated from gene models seguence file using 590 simulator program wasim version 1.6 (https://github.com/lh3/wgsim) with number of read pairs 591 being 1000000 and read length being 150.

592

593 Analysis of introgressions by genome re-sequencing

Genome representation was analysed using DNA purified from leaves, as described above. The visualization approach based on Genome Display Tile Plots (GDTPs) is essentially the same as that used for TDTPs described in He et al 2017^{30} , except that DNA gene models (i.e. including introns and UTRs) are used as the reference sequences and genome sequence reads are mapped. Only genes with significant signals (mean RPKM across the set of plants analysed in the experiment > 0.01) were used for further analysis. Tile plots were used to visualize genome redundancy data using quantitative representation of DNA gene models.

601

602 Data availability

- Raw sequence reads of *R. sativus* introgression samples can be found under NCBI BioProject
- accession ID PRJNA507350. Raw sequence reads of *B. fruticulosa* introgression samples can be
- found under NCBI BioProject accession ID PRJNA673122. Raw genome re-sequencing reads for
- the *B. carinata* mapping population YWDH can be found under NCBI BioProject accession ID
- 607 PRJNA722822. R-o-18 genome assembly information can be found under NCBI BioProject ID
- 608 PRJNA649364.
- 609
- 610 Code availability
- 611 The R script Genome_Sequence_Reorganise has been deposited on GitHub
- 612 (https://github.com/hezhesi/Genome_Sequence_Reorganise).





B genome B. juncea







C genome B. carinata





1	Tp1g00060	Tp1g02080
	Tp1g02100	Tp1g06340
	Tp1g06370	Tp1g06720
	Tp1g06760	Tp1g10420
	Tp1g10520	Tp1g10850
	Tp1g10880	Tp1g11870
	Tp1g11930	Tp1g15230
	Tp1g15260	Tp1g19420
	Tp1g19450	Tp1g27540
	Tp1g27600	Tp1g28010
	Tp1g28650	Tp1g30260
	Tp1g34890	Tp1g41430
	Tp1g32510	Tp1g34820
2	Tp2g00040	Tp2g06750
0,000		
	Tp2g06890	Tp2g07820
	Tp2g07870	Tp2g20060
	Tp2g20080	Tp2g23870
	Tp2g23880	Tp2g24950
	Tp2g24970	Tp2g29200
	Tp2g29210	Tp2g30530
3	Tp3g00420	Tp3g02230
	Tp3g02250	Tp3g20770
	Tp3g20810	Tp3g24570
	Tp3g26140	Tp3g34060
4	Tp4g00120	Tp4g08120
	Tp4g08210	Tp4g09040
	Tp4g09740	Tp4g11380
	Tp4g11490	Tp4g11900
	Tp4g11930	Tp4g16510
	Tp4g16530	Tp4g22740
	Tp4g22770	Tp4g28140
	Tp4g28200	Tp4g30030
5	Tp5g00110	Tp5g12320
	Tp5g12350	Tp5g13990
	Tp5g14010	Tp5g16130
	Tp5g16200	Tp5g17110
	Tp5g17130	Tp5g18280
	Tp5g19220	Tp5g33970
	Tp5g34000	Tp5g35900
~		
6	Tp6g00060	Tp6g03600
	Tp6g03660	Tp6g07730

8	Tp1g00050	Tp1g13310	
-	Tp1g13360	Tp1g17330	
	Tp1g17640	Tp1g27210	
	Tp1g27440	Tp1g31540	ł
9	Tp2g06900	Tp2g30710	ł
, - · · ,	Tp6g00880	Tp6g07500	Î
10	Tp3g00010	Tp3g22270	ł
	Tp2g02650	Tp2g06750	ł
			t
	Tp5g11300	Tp5g16450	Î
	Tp5g10450	Tp5g11270	Î
11	Tp4g00080	Tp4g30070	Î
	Tp7g00910	Tp7g01970	
	Tp7g02330	Tp7g04930	¥
-			
12	Tp6g26950	Tp6g41140	Î
	Tp6g14530	Tp6g26910	
	Tp6g11200	Tp6g14300	
	Tp5g19300	Tp5g35860	•
12			1
13	Tp5g00050	Tp5g10430	
	Tp5g12430	Tp5g13210	
	Tp5g13240	Tp5g17840	
	Tp5g22160	Tp5g29240	
	Tp5g29380	Tp5g36020	
	Tp5g00010	Tp5g09560	ľ
	Tp7g00010	Tp7g04200	
	Tp7g04300	Tp7g07090	
	Tp7g09040	Tp7g11450	
	Tp7g11490	Tp7g12550	L
	Tp5g09800	Tp5g12200	
	Tn7g13430	Tn7g15270	
	Tn7g15490	Tn7g17370	
	Tp7g17430	Tp7g20000	
	Tp7g20100	Tp7g36000	
	- Broros		
	Tp7g36130	Tp7g37550	T

14	Tp1g01280	Tp1g06880	
	Tp1g07060	Tp1g18390	
	Tp1g19310	Tp1g22020	
	Tp1g22110	Tp1g26320	¥
	Tp7g12030	Tp7g12420	1
	Tp7g34020	Tp7g37420	
	Tp7g21540	Tp7g22030	I
	Tp7g33130	Tp7g33820	
	Tp7g22200	Tp7g26990	
	Tp7g27110	Tp7g31380	
	Tp7g31510	Tp7g33020	¥
	Tp7g18040	Tp7g21470	↑
	Tp7g17410	Tp7g17860	
	Tp7g15470	Tp7g15750	
	Tp7g15870	Tp7g17350	¥
	Tp7g12490	Tp7g15250	1
	Tp1g26420	Tp1g27430	ł
	Tp7g11790	Tp7g11970	1
	Tp7g05910	Tp7g07670	↑
	Tp1g27620	Tp1g32200	1
	Tp7g09420	Tp7g11750	ŧ
	Tp1g32290	Tp1g32670	
	Tp1g32260	Tp1g36750	Ĩ
	Tp1g36760	Tp1g41870	•
	Tp1g36760	Tp1g41870	+
15	Tp1g36760 Tp3g28090	Tp1g41870 Tp3g33960	+ +
15	Tp1g36760 Tp3g28090 Tp2g00090	Tp1g41870 Tp3g33960 Tp2g01220	+
15	Tp1g36760 Tp3g28090 Tp2g00090 Tp2g02230	Tp1g41870 Tp3g33960 Tp2g01220 Tp2g03320	+
15	Tp1g36760 Tp3g28090 Tp2g00090 Tp2g02230 Tp2g04510	Tp1g41870 Tp3g33960 Tp2g01220 Tp2g03320 Tp2g05220	+
15	Tp1g36760 Tp3g28090 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950	Tp1g41870 Tp3g33960 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340	+
15	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g19860	+
15	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g19860 Tp2g20180	+
15	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g19860 Tp2g20180 Tp2g30770	+
15	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g19860 Tp2g20180 Tp2g30770	*
15 16	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g19860 Tp2g20180 Tp2g30770	*
15 16	Tp1g36760 Tp2g0090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240 Tp2g20240	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g20180 Tp2g30770 Tp2g30770	*
15	Tp1g36760 Tp2g0090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240 Tp2g20240	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g20180 Tp2g30770 Tp3g05080 Tp3g06990 Tp3g07720	
15	Tp1g36760 Tp2g0090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240 Tp2g20240	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g20180 Tp2g30770 Tp3g05080 Tp3g06990 Tp3g07720 Tp3g08460	
15	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240 Tp2g20240 Tp3g05110 Tp3g07100 Tp3g07780 Tp3g08600	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g20180 Tp2g30770 Tp3g05080 Tp3g06990 Tp3g07720 Tp3g08460 Tp3g020410	
15	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g15660 Tp2g19980 Tp2g20240 Tp3g05110 Tp3g07100 Tp3g07780 Tp3g08600 Tp3g20540	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g20180 Tp2g30770 Tp3g05080 Tp3g06990 Tp3g07720 Tp3g08460 Tp3g20410 Tp3g22540	
15	Tp1g36760 Tp2g0090 Tp2g02230 Tp2g04510 Tp2g04510 Tp2g15660 Tp2g15660 Tp2g19980 Tp2g20240 Tp3g05110 Tp3g07100 Tp3g07780 Tp3g08600 Tp3g20540 Tp3g22570	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g19860 Tp2g20180 Tp2g30770 Tp3g05080 Tp3g05080 Tp3g07720 Tp3g08460 Tp3g20410 Tp3g22540 Tp3g30400	
15	Tp1g36760 Tp2g0090 Tp2g02230 Tp2g02230 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g15660 Tp2g19980 Tp2g20240 Tp3g07100 Tp3g07100 Tp3g07780 Tp3g08600 Tp3g20540 Tp3g20540 Tp3g20540	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g19860 Tp2g20180 Tp2g30770 Tp3g05080 Tp3g06990 Tp3g07720 Tp3g08460 Tp3g08460 Tp3g20410 Tp3g20410 Tp3g22540	
15	Tp1g36760 Tp2g0090 Tp2g02230 Tp2g04510 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240 Tp3g07100 Tp3g07100 Tp3g07780 Tp3g07780 Tp3g08600 Tp3g20540 Tp3g20540 Tp3g20540 Tp3g20540	Tp1g41870 Tp2g01220 Tp2g03320 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g19860 Tp2g20180 Tp2g30770 Tp3g05080 Tp3g06990 Tp3g06990 Tp3g07720 Tp3g08460 Tp3g20410 Tp3g20410 Tp3g30400 Tp3g32510 Tp3g34170	
15	Tp1g36760 Tp2g0090 Tp2g02230 Tp2g04510 Tp2g04510 Tp2g06950 Tp2g15660 Tp2g19980 Tp2g20240 Tp3g07100 Tp3g07100 Tp3g07780 Tp3g07780 Tp3g08600 Tp3g20540 Tp3g20540 Tp3g20540 Tp3g20540	Tp1g41870 Tp3g33960 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g19860 Tp2g30770 Tp3g05080 Tp3g05080 Tp3g07720 Tp3g08460 Tp3g20410 Tp3g30400 Tp3g32510 Tp3g34170	
15 16 17	Tp1g36760 Tp2g00090 Tp2g02230 Tp2g04510 Tp2g04510 Tp2g04510 Tp2g04510 Tp2g0540 Tp2g15660 Tp2g15060 Tp2g19980 Tp2g20240 Tp3g00110 Tp3g07100 Tp3g07780 Tp3g02540 Tp3g30410 Tp3g30410 Tp3g32920	Tp1g41870 Tp3g33960 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g19860 Tp2g30770 Tp3g05080 Tp3g05080 Tp3g06990 Tp3g08460 Tp3g30400 Tp3g32510 Tp3g34170	
15 16 17	Tp1g36760 Tp3g28090 Tp2g0090 Tp2g02230 Tp2g04510 Tp2g04510 Tp2g04510 Tp2g05950 Tp2g15660 Tp2g15660 Tp2g02240 Tp3g00110 Tp3g07100 Tp3g07780 Tp3g08600 Tp3g20540 Tp3g30410 Tp3g30410 Tp3g32920	Tp1g41870 Tp3g33960 Tp2g01220 Tp2g03320 Tp2g05220 Tp2g15340 Tp2g15340 Tp2g15340 Tp2g30770 Tp3g05080 Tp3g05080 Tp3g07720 Tp3g08460 Tp3g22540 Tp3g30400 Tp3g32510 Tp3g34170	

Tp7g00050	Tp7g00360
Tp7g01440	Tp7g06850
Tp7g07940	Tp7g08390
Tp7g12590	Tp7g15360
Tp7g16710	Tp7g17380
Tp7g17390	Tp7g26340
Tp7g15380	Tp7g16670
Tp7g26370	Tp7g37520

7

Tp6g07740Tp6g09710Tp6g10770Tp6g35180Tp6g35190Tp6g35900Tp6g35920Tp6g40960

¥

 Tp4g00010
 Tp4g12510

 Tp6g10970
 Tp6g12640

 Tp6g00190
 Tp6g10610









A genome	A genome	B genome	B genome	C genome	C genome
B. napus	B. juncea	B. juncea	B. carinata	B. carinata	B. napus

