This is a repository copy of Assembly and comparison of two closely related Brassica napus genomes.

White Rose Research Online URL for this paper: http://eprints.whiterose.ac.uk/114855/

Version: Published Version

Article:
Bayer, Philippe E, Hurgobin, Bhavna, Golicz, Agnieszka A et al. (13 more authors) (2017) Assembly and comparison of two closely related Brassica napus genomes. Plant biotechnology journal. pp. 1602-1610. ISSN 1467-7644

https://doi.org/10.1111/pbi.12742

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

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.
Assembly and comparison of two closely related \textit{Brassica napus} genomes

Philipp E. Bayer$^1$, Bhavna Hurgobin$^{1,2}$, Agnieszka A. Golicz$^3$, Chon-Kit Kenneth Chan$^1$, Yuxuan Yuan$^1$, HueyTyng Lee$^{1,2}$, Michael Renton$^{1,4}$, Jinling Meng$^5$, Ruiyuan Li$^5$, Yan Long$^5$, Jun Zou$^5$, Ian Bancroft$^6$, Boulos Chalhoub$^{7,8}$, Graham J. King$^{5,9}$, Jacqueline Batley$^1$ and David Edwards$^{1,*}$

$^1$School of Biological Sciences, The University of Western Australia, Crawley, WA, Australia
$^2$School of Agriculture and Food Sciences, University of Queensland, St. Lucia, Qld, Australia
$^3$Plant Molecular Biology and Biotechnology Laboratory, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, Melbourne, Vic., Australia
$^4$School of Agriculture and Environment, The University of Western Australia, Crawley, WA, Australia
$^5$National Key Laboratory of Crop Genetic Improvement, Key Laboratory of Rapeseed Genetic Improvement, Ministry of Agriculture P. R. China, Huazhong Agricultural University, Wuhan, China
$^6$Department of Biology, University of York, York, UK
$^7$Organization and Evolution of Complex Genomes (OECG), Institut National de la Recherche agronomique (INRA), Université d’Evry Val d’Essonne (UEVE), Evry, France
$^8$Institute of System and Synthetic Biology, Genopole, Centre National de la Recherche Scientifique, Université d’Evry Val d’Essonne, Université Paris-Saclay, Evry, France
$^9$Southern Cross Plant Science, Southern Cross University, Lismore, NSW, Australia

Received 16 November 2016; revised 6 February 2017; accepted 9 April 2017.
*Correspondence (Tel +61 (0)8 6488 2415; fax +61 (8) 6488 1108; email dave.edwards@uwa.edu.au)

Summary

As an increasing number of plant genome sequences become available, it is clear that gene content varies between individuals, and the challenge arises to predict the gene content of a species. However, genome comparison is often confounded by variation in assembly and annotation. Differentiating between true gene absence and variation in assembly or annotation is essential for the accurate identification of conserved and variable genes in a species. Here, we present the \textit{de novo} assembly of the \textit{B. napus} cultivar Tapidor and comparison with an improved assembly of the \textit{Brassica napus} cultivar Darmor-bzh. Both cultivars were annotated using the same method to allow comparison of gene content. We identified genes unique to each cultivar and differentiate these from artefacts due to variation in the assembly and annotation. We demonstrate that using a common annotation pipeline can result in different gene predictions, even for closely related cultivars, and repeat regions which collapse during assembly impact whole genome comparison. After accounting for differences in assembly and annotation, we demonstrate that the genome of Darmor-bzh contains a greater number of genes than the genome of Tapidor. Our results are the first steps towards comparison of the true differences between \textit{B. napus} genomes and highlight the potential sources of error in future production of a \textit{B. napus} pangenome.

Keywords: genome assembly, whole genome comparison, genotyping by sequencing, genome assembly improvement, \textit{Brassica napus}, Tapidor, contigPlacer.

Introduction

\textit{Brassica} is a genus which contains economically important crop species, including condiments such as mustards; vegetable crops including broccoli, Brussels sprouts, cabbage, cauliflower, and commercially important oilseeds which are used as both a food and biofuel. The genomes of the \textit{Brassica} family are related as described in the triangle of U (Nagaharu, 1935). There are three diploid \textit{Brassica} species, \textit{B. rapa} (AA, \(n = 10\)), \textit{B. nigra} (BB, \(n = 8\)), and \textit{B. oleracea} (CC, \(n = 9\)), and these genomes evolved from a common ancestor which underwent at least one whole genome triplication (Lysak et al., 2005). Combinations of these genomes result in three amphidiploid \textit{Brassica} species: \textit{B. juncea} (AABB, \(n = 18\)), \textit{B. carinata} (BBCC, \(n = 17\)) and \textit{B. napus} (AACC, \(n = 19\)).\footnote{For a detailed account of \textit{Brassica} genome organization see Lysak et al., 2005.}\footnote{For a detailed account of \textit{Brassica} genome organization see Lysak et al., 2005.}\footnote{For a detailed account of \textit{Brassica} genome organization see Lysak et al., 2005.}\footnote{For a detailed account of \textit{Brassica} genome organization see Lysak et al., 2005.} Several \textit{Brassica} genome assemblies have now been published, including the A genome of the \textit{B. rapa} cultivar Chiifu (Wang et al., 2011), the C genome of the \textit{B. oleracea} cultivars O2-12 (Liu et al., 2014) and TO1000 (Parkin et al., 2014), the AACC genome of the \textit{B. napus} cultivar Darmor-bzh (Chalhoub et al., 2014) and the BB and AABBB genome of \textit{B. nigra} and \textit{B. juncea} (Yang et al., 2016).

Early indications of significant differences in genome content between individuals were apparent from data obtained from RFLP and SSR analysis in \textit{Brassica} species, where a relatively large number of null alleles were detected with co-dominant molecular assays (Udall et al., 2005). With the increased application of next generation DNA sequencing technology, these genomic differences are becoming more apparent and there is a growing understanding that a single genome does not reflect the gene content of a species. Following the production of reference genome assemblies, researchers are moving towards the characterization of pangenomes, representing the gene content of a species, including core genes which are found in all individuals and variable genes which are only present in some individuals.
(Golicz et al. 2016a). Variable genes can be split into two groups: copy number variations (CNVs), in which the number of copies of a gene differs between individuals, and the presence/absence variations (PAVs), an extreme form of CNV in which a gene is present in some individuals but absent in others (Golicz et al., 2016a; Saxena et al., 2014).

Gene presence/absence variation is important for crop improvement as some variable genes have been shown to be associated with agronomic traits. In rice, comparison of three divergent lines led to the identification of several variable genes including the submergence tolerance gene Sub1A which is absent in submergence intolerant lines (Schatz et al., 2014). A recent study investigated the pangenome for 1483 rice accessions and found 1913 high confidence dispensable genes, of which 1489 were expressed (Yao et al., 2015). A genome-wide association study (GWAS) with SNPs for those rice accessions found that 23.5% of metabolic traits had higher association signals with SNPs located on dispensable genes than with SNPs located on the core reference genome. Similarly, comparison of three Brassica rapa morphological variations (turnip, rapid cycling and Chinese cabbage) revealed around 1224 unique genes in each of the three genomes (Lin et al., 2014), while another recent study produced low coverage sequence data for 199 B. rapa and 119 B. oleracea accessions to identify SNPs and trace parallel selection signals in the two subgenomes (Cheng et al., 2016). Regions undergoing positive selection could be identified (25 in B. rapa and 58 in B. oleracea), of which nine were shared between the genomes.

In chickpea, the re-sequencing of 16 lines identified up to 32 genes absent in each line (Thudi et al., 2016), while a comparison of 503 inbred maize lines demonstrated that only 16.4% of 8,681 representative transcripts were expressed in all lines (Hirsch et al., 2014). Recently, Golicz et al. (2016a,b) identified 61,379 genes in a Brassica oleracea pangeneome study, of which 18.7% demonstrated the presence/absence variation among individuals, and with the variable genes being enriched for annotations associated with important agronomic traits.

The repetitive content of plant genomes makes genome assembly and validation a challenge (Edwards et al., 2013). Assemblers often cannot accurately assign sequence reads from repetitive regions to their correct genomic location and repeat sequences in the genome frequently collapse into a single copy in the assembly. In addition, different annotation results may confound comparative analyses as some genes may not be predicted in some cultivars even if they share significant sequence identity with genes predicted in other cultivars. These factors impact the direct comparison of whole genome assemblies for complex crop genomes and make the assessment of species pan-genomes a major challenge.

Here, we describe the de novo assembly and annotation of the amphidiploid B. napus cultivar Tapidor and comparison with an improved and reannotated assembly of the published B. napus cultivar Darmor-bzh (Chalhoub et al., 2014). Comparison of the gene content between Darmor-bzh and Tapidor identified genes that are conserved or unique to each cultivar. In addition, we identified genes which appear to be absent in one or other cultivar due to variation in assembly or annotation. Our results present a detailed assessment and comparison of assemblies for complex B. napus genomes, the first step towards establishing a B. napus pangeneome.
To assess and compare the completeness of the two genome assemblies, CEGMA (Parra et al., 2007) was used to identify core eukaryotic genes (CEGs), and BUSCO (Simão et al., 2015) was used with the plants profile to identify Single-Copy Orthologs (SCOs). All 248 CEGs aligned completely with the Darmor-bzh assembly, while only 246 CEGs aligned completely with the Tapidor assembly, with the remaining two partially identified in both assemblies. Based on genomic read mapping, three Tapidor genes were predicted to be absent in Tapidor -bzh genes, with 59 099 genes in Tapidor reads, while 70 162 predicted genes for Tapidor (31 969 on the A and 39 334 on the C pseudomolecules, 3228 on unplaced contigs), 37 220 genes for Darmor -bzh (37 220 on the A and 39 334 on the C pseudomolecules, 3228 on unplaced contigs) (Table 1). The genes predicted to be absent in Tapidor -bzh, Tapidor, Ningyou-7 and Tapidor x Ningyou-7 RNA-Seq data. Of the 73 genes, 21 (29%) showed no expression in both cultivars to each of the assemblies. Based on genomic read mapping, three Tapidor genes were predicted to be absent in Tapidor -bzh, while 73 Darmor-bzh genes were predicted to be absent in Tapidor (Table 1). The genes predicted to be absent were compared with Darmor-bzh, Tapidor, Ningyou-7 and Tapidor x Ningyou-7 RNA-Seq data. Of the 73 genes, 21 (29%) showed no expression in

| Name | Darmor-bzh SNPs | Darmor-bzh Length (Mbp) | Darmor-bzh SNPs/Mbp | Darmor-bzh predicted filtered genes | Tapidor predicted filtered genes | Tapidor SNPs/Mbp | Tapidor genes with no reads | Darmor-bzh genes with no reads |
|------|----------------|------------------------|---------------------|-----------------------------------|---------------------------------|-----------------|---------------------------|-------------------------------|
| chrA01 | 45 870 | 31.16 | 1472 | 37 875 | 23.9 | 1586 | 3687 | 3050 | 0 | 3 |
| chrA02 | 45 229 | 31.34 | 1443 | 46 599 | 27.9 | 1672 | 3528 | 3315 | 1 | 0 |
| chrA03 | 59 180 | 39.49 | 1499 | 55 319 | 32.1 | 1723 | 5429 | 4408 | 0 | 4 |
| chrA04 | 57 270 | 23.31 | 2457 | 49 357 | 21 | 2351 | 2676 | 2638 | 0 | 0 |
| chrA05 | 54 784 | 28.6 | 1916 | 36 548 | 20.1 | 1818 | 3506 | 2634 | 1 | 1 |
| chrA06 | 72 163 | 31.9 | 2262 | 57 613 | 29.1 | 1977 | 3926 | 3713 | 0 | 0 |
| chrA07 | 45 458 | 28.9 | 1573 | 34 101 | 22.7 | 1505 | 3555 | 2976 | 0 | 0 |
| chrA08 | 24 807 | 21.74 | 1141 | 22 096 | 16.4 | 1346 | 2827 | 2147 | 0 | 0 |
| chrA09 | 79 586 | 46.72 | 1704 | 53 008 | 30.8 | 1721 | 5286 | 3913 | 0 | 3 |
| chrA10 | 27 723 | 19.96 | 1389 | 35 378 | 22.5 | 1575 | 2800 | 3175 | 0 | 1 |
| Total A | 512 070 | 303 | 427 894 | 247 | 37 220 | 31 969 | 2 | 12 |
| chrC01 | 92 316 | 47.95 | 1925 | 63 462 | 31.9 | 1990 | 3739 | 3083 | 0 | 5 |
| chrC02 | 49 665 | 58.66 | 847 | 47 872 | 40.5 | 1181 | 4182 | 3572 | 1 | 23 |
| chrC03 | 64 224 | 71.85 | 894 | 72 229 | 55.2 | 1307 | 6448 | 5800 | 0 | 9 |
| chrC04 | 95 329 | 61.04 | 1562 | 79 006 | 45.6 | 1733 | 4658 | 4356 | 0 | 2 |
| chrC05 | 25 226 | 52.72 | 479 | 43 752 | 45.6 | 959 | 4452 | 4839 | 0 | 4 |
| chrC06 | 41 036 | 44.61 | 920 | 41 252 | 34.1 | 1211 | 3543 | 3230 | 0 | 6 |
| chrC07 | 34 089 | 52.5 | 649 | 41 001 | 37.3 | 1099 | 4165 | 3523 | 0 | 1 |
| chrC08 | 43 857 | 46.29 | 947 | 48 079 | 40.2 | 1197 | 4140 | 4295 | 0 | 0 |
| chrC09 | 46 660 | 60.21 | 775 | 57 479 | 51 | 1127 | 4607 | 4864 | 0 | 2 |
| Total C | 492 402 | 3496 | 494 132 | 381 | 39 934 | 37 562 | 1 | 52 |
| Unplaced contigs | 2513 | 51.33 | 49 | 23 380 | 8.5 | 2752 | 3228 | 631 | 0 | 9 |
| Total | 1 006 985 | 850 | 945 406 | 636 | 80 382 | 70 162 | 3 | 73 |
Darmor-bzh or Tapidor, 19 (26%) were expressed in Darmor-bzh but neither Tapidor nor Ningyou-7, 27 (37%) were expressed in Darmor-bzh and Ningyou-7 but not in Tapidor, and expression segregated in the Tapidor × Ningyou-7 DH population, 2 (2.7%) were expressed in Tapidor and Ningyou-7 but not in Darmor-bzh, while 4 (5.5%) were expressed in all tissues in cases (Table S3). The three genes predicted to be absent in Darmor-bzh but not in Tapidor were not expressed in Darmor-bzh, Ningyou-7 or Tapidor (Table S4).

Proteins encoded by genes predicted to be absent in one of the two cultivars were compared with Swiss-Prot and checked for enriched GO terms. In genes present in Darmor-bzh but not in Tapidor, terms such as ‘RNA splicing’ and ‘floral meristem growth’ were enriched (P < 0.05, Table S5). The three genes predicted to be present in Tapidor but not in Darmor-bzh had no Swiss-Prot hits, and therefore, transfer of GO annotation was not possible.

Assessment of repetitive and collapsed assembly regions

Using the differential comparative read mapping pipeline CoReFinder, we identified 30,200 and 26,812 collapsed regions longer than 50 bp, with a total length of 12,495,844 bp and 10,522,089 bp and average sizes of 413 bp and 392 bp in the Darmor-bzh and Tapidor assemblies, respectively (Figure 1, Table S6). We also identified 43,775 and 2,191 repetitive regions, where the sequence is represented more than once in the assembly, totalling 31,807,543 bp and 972,605 bp, with an average of 720 bp and 437 bp in Darmor-bzh and Tapidor, respectively (Figure 2, Table S7).

Genes were identified in both the collapsed and repetitive regions of the Darmor-bzh and Tapidor assemblies. In Darmor-bzh, 2,455 genes (3.1%) were located in collapsed regions and 5,703 (7.1%) were located in repetitive regions, while in Tapidor, 2,651 genes (3.8%) were located in collapsed regions and 2,464 (0.4%) were located in repetitive regions (Table S8). In Tapidor, chromosomes A2 and C2 carried the largest number of collapsed genes (410, 15% and 245, 9.2%) while in Darmor-bzh, chromosomes A3 and C3 carried the largest number of collapsed genes (212, 8.6% and 222, 9.0%). In Darmor-bzh, the chromosomes A3 and C3 carried the largest number of repetitive genes (588, 10.3% and 470, 8.2%) while chromosomes A1 and C1 carried the most repetitive genes in Tapidor (26, 10.5% and 28, 11.4%).

Pfam protein domains and UniProtKB/Swiss-Prot hits were identified for genes located in repetitive and collapsed regions in both assemblies. Both sets of results were compared between Darmor-bzh and Tapidor in order to identify any common patterns that may lead to problems in genome assemblies. The largest number of shared Pfam domains and shared Swiss-Prot hits was between genes located in repetitive regions in Darmor-bzh and genes located in collapsed regions in Tapidor. In repetitive genes in Darmor-bzh, 10% of Pfam domains overlapped with collapsed genes in Tapidor and Pfam domains collapsed in Darmor-bzh, while only 0.5% of genes located in collapsed and repetitive regions in Tapidor shared Pfam domains (Figure 3).

The repetitive genes in Darmor-bzh and collapsed genes in Tapidor contained repetitive domains such as ‘Myb-like DNA-
binding pattern was exhibited in shared Swiss-Prot hits where the greatest overlap in terms of Swiss-Prot hits was between the set of proteins encoded by Darmor-bzh genes located in collapsed regions and Tapidor genes located in repetitive regions (Figure S3). As with the Pfam domain overlaps, 10% of Swiss-Prot hits were shared between repetitive genes in Darmor-bzh and collapsed genes in Tapidor. The Swiss-Prot hits contained proteins such as ‘Myosin-6’ or ‘Proline-rich receptor-like protein kinase’ (Table S10).

**Discussion**

*Brassica napus* canola is an important global crop and a major source of vegetable oil for human consumption. While genetic markers have been used for crop improvement for several years (Hayward et al., 2012), the application of genomics is only now leading to improved traits in the field. A major limitation of genomics based crop improvement is the lack of available reference genome assemblies. The first public *Brassica napus de novo* whole genome assembly was published in 2014, supported by a large multinational endeavour (Chalhoub et al., 2014). Since this achievement, DNA sequencing and bioinformatics technologies have advanced, and it has become increasingly accepted from studies in a wide range of crop species that a single reference is insufficient to describe the gene content of a species. To start to address this issue, we have produced an improved genome assembly for Darmor-bzh and produced a second *de novo* whole genome assembly for a closely related cultivar, *B. napus* Tapidor. Tapidor and Darmor-bzh both share Jet Neuf as an ancestor, but Darmor-bzh additionally has the cultivar Bronovski in its direct ancestry (Foisset et al., 1995), while Tapidor contains Bienvenu in its ancestry (Sharpe and Lydiate, 2003).

To permit a direct comparison between the Tapidor and Darmor-bzh genome assemblies, we improved the published Darmor-bzh genome using high-resolution skim genotyping by sequencing (Bayer et al., 2015) to position previously unplaced contigs within pseudomolecules. Both assemblies were then annotated using the same data and methods to prevent confounding effects of different annotation pipelines.

The assembly of the cultivar Tapidor was smaller (635 Mbp) compared with an assembly size of 850 Mbp for Darmor-bzh (Table 1). Using k-mers to predict both the assembly size and the genome size, we found that the estimated genome size was similar for both Darmor-bzh (1345 Mbp) and Tapidor (1335 Mbp). In contrast, the predicted assembly size was similar to, although smaller than the actual assembly size with an estimated 808 Mbp for Darmor-bzh and 697 Mbp for Tapidor. The predicted genome size estimate accounts for coverage, sequencing depth and unique k-mers, while the genome assembly estimates account only for unique k-mers. The Tapidor sequence data contains fewer unique k-mers than the Darmor-bzh sequencing data, which suggests that the Tapidor genome is actually smaller than the Darmor-bzh genome. Both estimated genome sizes here are larger than the flow-cytometry based estimations for different *B. napus* (AAC) cultivars of 1129–1235 Mbp (Arunuganathan and Earle, 1991). While the predicted genome size difference probably reflects a real difference in genome size, the discrepancy between the assembly size of Tapidor and Darmor-bzh is probably due to the different sequencing technologies that were used for assembly. Tapidor was assembled with short paired end sequences, while Darmor-bzh was assembled using a combination of paired end, 454 and Sanger sequencing (Chalhoub et al., 2014). The consistent differences between the assembly sizes and predicted genome sizes are common for genome assemblies and reflect the challenges of assembling the most repetitive regions of the genome.

While there appears to be a real difference in the genome size of these two homoygous individuals, we hypothesize that the different assembly methods cause the majority of observed differences between the assemblies. The CEGMA results suggest that both assemblies contain the majority of the gene content and that both assemblies are of high quality. Both assemblies contain all core eukaryotic genes (CEGs), although two CEGs were only partially present in Tapidor, which suggests that this assembly may be more fragmented. Similarly, the BUSCO results show that both assemblies carry the majority of gene content, with the Darmor-bzh assembly having more gene copies than Tapidor. This suggests that some gene copies have collapsed in the Tapidor assembly which have not collapsed in the Darmor-bzh assembly. The increased collapsing of regions in the Tapidor assembly compared to the Darmor-bzh assembly is probably due to the use of long read 454 and Sanger sequence data in the production of the Darmor-bzh assembly.

Both the genomes were annotated using the same methods and supporting data. Gene prediction is dependent on a wide range of factors and the number of predicted genes is influenced by the stringency of gene calling. To enable the most robust comparison of predicted genes between these assemblies, we applied a relatively strict gene calling approach. The number of predicted genes for Darmor-bzh is lower in our study than the 101 040 gene models originally predicted by Chalhoub et al. (2014), and this difference is likely due to the more stringent methods of repeat masking employed, including the removal of transposon-related and short genes.

As the Tapidor contigs were scaffolded based on the Darmor-bzh assembly, and both Tapidor and Darmor-bzh have shared ancestry, we expected to see extensive collinearity between the genome assemblies. We assessed this by comparison of genes present in collinear blocks along the pseudomolecules. We also examined whether genes which were predicted to be missing from collinear blocks were in fact missing or just remained unannotated. In Darmor-bzh, 11 688 genes (14.5%) and in Tapidor 10 903 genes (15.5%) were not in collinear blocks and were not identified in their expected regions, suggesting that they are either misplaced in the assembly or are nonsyntenic in the genome. Wicker et al. (2011) estimated that 17% of Triticaceae genes are nonsyntenic with model grasses, which is similar to our findings in *B. napus*. Extensive homeologous recombination has been observed in *B. napus* (Cai et al., 2014; Chalhoub et al., 2014; Nicolas et al., 2007) which may also explain some of the nonsyntenic regions observed here.

As differences in genome assembly and annotation may suggest differences in gene content which are not reflected in the actual genome, we investigated the number of genes predicted to be absent in each cultivar based on raw data and independent of gene assembly or annotation (Table 1). Mapping genomic sequence reads between references suggests that 73 genes are unique to Darmor-bzh and absent in Tapidor, while three genes are unique to Tapidor and absent in Darmor-bzh. Of the 73 genes predicted to be absent, 27 (37%) were expressed in Darmor-bzh and Ningyou-7 but not in Tapidor, with clear segregation in the Tapidor × Ningyou-7 DH population, suggesting that these genes have been truly lost in Tapidor but are
present in Ningyou-7 (Table S3). The 19 (26%) genes expressed in Darmor-bzh but not in Tapidor or Ningyou-7 may also be absent in Ningyou-7, or their expression may not be captured in the Darmor-bzh RNA-Seq libraries. Some error remains, as two genes (2.7%) predicted to be absent in Tapidor and present in Darmor-bzh appear to be expressed in Tapidor but not in Darmor-bzh, and four genes (5.5%) appear to be expressed in all cultivars. These results may be due to the more relaxed mapping parameters for the single RNA-Seq reads compared to the paired genomic sequence reads, leading to erroneous read mapping and false calling of gene presence in the RNA-Seq analysis. Genes which showed no expression in any of the cultivars assessed may be pseudogenes, misannotated or only expressed in certain conditions or tissues which were not sampled during RNA-Seq library preparation (Table S4).

The relatively small difference in the presence of unique genes compares with 10 220 annotated Darmor-bzh genes which are not found in the Tapidor assembly. To test whether this difference is due to duplicate gene copies that have collapsed into single copies in the Tapidor assembly, we investigated the number and size of collapsed and repetitive regions in both assemblies (Table S6, Table S7, Figures 1 and 2). The number and the size of repetitive regions were greater in the Darmor-bzh assembly by a factor of 19 and 32, respectively. These numbers are in line with the above gene number differences as well as the differences of assembly size and estimated assembly size, and suggests that the longer 454 and Sanger reads used in the Darmor-bzh assembly lead to more repetitive regions being assembled and less collapsing of genomic regions in Darmor-bzh.

Homeologous exchanges (HEs) are frequently observed in polyploid species, and the number of genes located in repetitive and collapsed regions can be used for the detection of HE between the A and the C genome (Chalhoub et al., 2014). Here, the largest and second largest number of collapsed and repetitive genes in both assemblies is always between a chromosome and its homeologous partner chromosome in the same cultivar, such as A2 and C2 for the number of collapsed genes in Darmor-bzh. The pseudomolecule with the greatest size of collapsed region is C2 in Darmor-bzh which is also consistent with prior reports (Chalhoub et al., 2014). When comparing both cultivars, chromosome C2 shows the greatest difference, with 23 (31%) of genes not present in Tapidor in the first 10 Mbp, and 245 (9.2%) of collapsed genes in Tapidor being located on C2 in the first 10 Mbp. The number of collapsed genes suggests homeologous exchange of the first 10 Mbp between A2 and C2 with subsequent loss of the original C2 arm in Tapidor (a homeologous nonreciprocal translocation, HNRT), and that this genome structure is not present in Darmor-bzh. The genes in this region are enriched for the GO-term ‘floral meristem growth’ (P < 0.05), and chromosome C2 contains several breeding related QTLs such as seed yield, seed weight and heterosis (Zhao et al., 2016). This suggests that selection for agronomic traits may be responsible for the observed differences between Darmor-bzh and Tapidor on this chromosome.

The inability to assemble repetitive sequences can lead to the underestimation of gene content and the incorrect calling of presence/absence variation between assemblies. The annotation of genes located in repetitive and collapsed regions helps to identify which protein domains may be associated with poor assembly and subject to incorrectly called the presence/absence variation. We observed an overlap in shared domains and Swiss-Prot hits between genes located in repetitive regions in Darmor-bzh and collapsed regions in Tapidor (Figure 3, Figure S3). Shared domains include known repetitive domains such as WD repeats or pentatricopeptide repeats (PPRs), while shared Swiss-Prot hits included leucine-rich repeat receptor kinases, as well as somatic embryogenesis receptor kinases which contain leucine repeats (Hecht et al., 2001). This suggests that genes which have repetitive domains under-assemble using only short-read technology but at least some of these may be correctly assembled when longer sequence reads are included in the assembly process. However, the inclusion of 454 reads does not correctly assemble all repetitive genes, as many collapse-related Pfam and Swiss-Prot domains are located in collapsed regions in Darmor-bzh, indicating that there is still missing gene content in the high quality Darmor-bzh assembly (Figure 3). The potential collapsing of genomic regions containing important agronomic genes is important for both future genome assembly projects and future pangeneome analysis as the capturing of all gene content which may play a role in agronomic performance is important for the application of genomics assisted breeding.

The differential assembly of genes with repetitive domains along with the presence of homeologous exchanges is likely to confound the assessment of gene presence variation between individuals and the translation of this information for the agronomic improvement of crop species. For example, there are several reports of variation in NBS LRR gene content between individuals and that this may play an important role in the resistance or susceptibility of crops to disease (Tollenare et al., 2012; Wu et al., 2014). While the evidence suggests that there is likely to be some true variation in the presence of these genes, it is important to differentiate between true variation and assembly or annotation issues when interpreting the biological function of these genes as assembly and annotation variation can confound gene trait associations.

Conclusions

We have de novo assembled and annotated a second genome for the important oilseed crop species, B. napus canola, and compared this with an improved and reannotated assembly of the published cultivar Darmor-bzh (Chalhoub et al., 2014). We demonstrate that comparison of gene content between individuals needs to account for differences in assembly and annotation to avoid misinterpretation. Between the two assemblies, only three and 73 predicted genes represent real gene loss between these closely related cultivars, 524 and 701 (0.7%–0.8%) are due to misannotation (i.e. the gene annotation process failed to identify a gene in one assembly but predicted its presence in the other assembly), while 10 903 and 11 688 (14%–15%) of the differences were due to either misplaced contigs or real gene movement. Our results suggest that researchers comparing different genome assemblies should not rely solely on the results of annotation pipelines but should also compare the assemblies and the unassembled read data to differentiate between real differences and artefacts. In particular, genes with repetitive domains may collapse leading to an underestimation of gene copy number. By accounting for these potential errors, it is possible for future studies to establish an accurate pangeneome for this important oilseed crop.

Methods

Darmor-bzh and Tapidor genome assemblies

All Tapidor reads were cleaned using sickle (Joshi and Fass, 2011) and normalized using khmer v1.0 normalize_by_median.py (Brown et al., 2012; Crusoe et al., 2015). VelvetOptimiser
Gene models were produced using Tophat v2.1.0 and cufflinks (Gladman and Seemann, 2012) together with Repbase v were used to mask repeats. RepeatModeler v1.0.8 and RepeatMasker v4.0.6 (http://repeatmasker.org) together with Repbase v1.0.8 and RepeatMasker v4.0.6 (http://repeatmasker.org) to assemble contigs. Contigs greater than 1 Kbp were sorted by comparison with the Darmor-bzh assembly using blastn (Altschul et al., 1990) with an e-value of 1e-6. Contigs with two highest-scoring alignments within 1% were unplaced. Contigs were then sorted using LASTZ (Harris, 2007) and a custom script (LASTZSorter.py, available at http://appliedbioinformatics.com.au/index.php/Darmor_Tapidor). MASTMap (Wu et al., 2008) was used to calculate the genetic map using the unimputed SNPs of the DH population (distance_function: kosambi, cut_off_p_value: 0.0000001, no_map_size 2, missing_threshold 0.1).

SGSautoSNP (Lorenz et al., 2012) was used with the SkimGBS pipeline (Bayer et al., 2015) to call SNPs and genotypes using either the Tapidor or Darmor-bzh-genome as reference (Chalhoub et al., 2014). ContigPlacer was used to place unplaced contigs in the Darmor-bzh and Tapidor assemblies and is available at https://github.com/phillipnbayer/contigPlacer.

Gene-level comparison between Darmor

RepeatModeler v1.0.8 and RepeatMasker v4.0.6 (http://repeatmasker.org) together with Repbase v were used to mask repeats. Gene models were produced using Tophat v2.1.0 and cufflinks v2.2.0 (Kim et al., 2013), and annotation was performed using MAKER v2.21 (Cantarel et al., 2008). RNA-Seq data for the cultivars Tapidor, Ningyou and TN DH (Higgins et al., 2012) were used together with Darmor-bzh RNA-Seq reads (Chalhoub et al., 2014).

CEGMA v2.5 (Parra et al., 2007) was used to annotate the Brassica oleracea C genome (Parkin et al., 2014) with core gene models (COGs). The annotated core gene models were used as input for AUGUSTUS v3.0.2 (Keller et al., 2011). A set of ESTs from the 95k microarray (Trick et al., 2009), Brassicaceae proteins from RefSeq, and B. rapa, B. oleracea and B. napus unigenes from UniGene (NCBI Resource Coordinators 2013) were used as evidence. Predicted genes with an AED-score of 1, shorter than 100 bp, or carrying transposase-related domains were removed from subsequent analysis. Both assemblies were assessed using CEGMA v2.5 (Parra et al., 2007) and BUSCO v1.1.b1 (Simão et al., 2015) with the early release plants dataset. Predicted transcripts and proteins were renamed according to the standardized Brassica nomenclature (Østergaard et al., 2008).

Gene-level comparison between Darmor-bzh and Tapidor

MCScanX (Wang et al., 2012) was used to analyse collinearity between the Darmor-bzh and Tapidor genes and to assign predicted genes to blocks of at least five collinear genes. Custom scripts were used to parse the MCScanX output. Genes were compared with the Tapidor pseudomolecules using blastp with an e-value cut-off of 1e-10. The scripts are available at http://appliedbioinformatics.com.au/index.php/Darmor_Tapidor under ‘Collinearity analysis’.

Disclosure declaration

The authors declare no competing interests.

References

Alexa, A. and Rahnenfuhrer, J. (2010) topGO: enrichment analysis for gene ontology. R package version, 2.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Arumuganathan, K. and Earle, E.D. (1991) Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9, 211–215.

Bayer, P.E., Ruperao, P., Mason, A.S., Stiller, J., Chan, C.-K.K., Hayashi, S., Long, Y. et al. (2015) High-resolution skin genotyping by sequencing reveals
the distribution of crossovers and gene conversions in Cicer anetinum and Brassica napus. Theor. Appl. Genet. 128, 1039–1047.
Brown, C.T., Howe, A., Zhang, Q., Pyrkoz, A.B. and Brom, T.H. (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. http://arxiv.org/abs/1203.4802.
Cai, G., Yang, Q., Yi, B., Fan, C., Edwards, D., Batley, J., Zhou, Y. et al. (2014) A complex recombination pattern in the genome of allotetraploid Brassica napus as revealed by a high-density genetic Map ed. H. Candela. PLoS One, 9, e109910.
Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C. et al. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196.
Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, W., Wang, X., Chiquet, J. et al. (2014) Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science, 345, 950–953.
Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, Y., Liu, B. et al. (2016) Subgenomic parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. Nat. Genet. 1–10.
Cruise, M.R., Alamedin, H.F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Chalmers, A. et al. (2015) The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research, 4, 900.
Edwards, D., Batley, J. and Snowdon, R.J. (2013) Accessing complex crop genomes with next-generation sequencing. Theor. Appl. Genet. 126, 1–11.
Foisset, N., Delourme, R., Barret, P. and Renard, M. (1995) Molecular tagging of the dwf dwarf BZH gene in Brassica napus. Theor. Appl. Genet. 91, 756–761.
Gladman, S. and Seemann, T. (2012) Velvet Optimiser (Version 2.2.4) [Software]. http://www.bioinformatics.utas.edu.au/software.velvetoptimiser.shtml.
Golicz, A.A., Martinez, P.A., Zander, M., Patel, D.A., Van De Wouw, A.P., Visendi, P., Fitzgerald, T.L. et al. (2014) Gene loss in the fungal canola pathogen Leptosphaeria maculans. Funct. Integr. Genomics, 15, 189–196.
Golicz, A.A., Batley, J. and Edwards, D. (2016a) Towards plant pangenomics. Plant Biotechnol. J. 14, 1099–1105.
Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Chan, C.K.K. et al. (2016b) The panome of an agronomically important crop plant Brassica oleracea. Nat. Commun. 7, 13390.
Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.
Hayward, A., Mason, A.S., Dalton-Morgan, J., Zander, M., Edwards, D. and Batley, J. (2012) SNP discovery and application in Brassica napus. J. Plant Physiol. 39, 49–61.
Hecht, V., Vieille-Calzada, J.P., Hartog, M.V., Schmidt, E.D., Boutilier, K., Grossniklaus, U. and de Vries, S.C. (2001) The Arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR KINASE 1 gene is expressed in developing ovules and embryos and enhances embryogenic competence in culture. Plant Physiol. 127, 803–816.
Higgins, J., Magusin, A., Trick, M., Fraser, F. and Bancroft, I. (2012) Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species Brassica napus. BMC Genom. 13, 247.
Hirsch, C.N., Foerster, J.M., Johnson, J.M., Skehon, R.S., Muttoni, G., Vaillancourt, B., Perigagacino, F. et al. (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell, 26, 121–135.
Joshi, N.A. and Fass, J.N. (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. https://github.com/nvr1/najoshi/sickle.
Keller, O., Kollmar, M., Stanke, M. and Waack, S. (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics, 27, 757–763.
Kim, D., Park, S., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36.
Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357–359.
Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient short of DNA sequences to the human genome. Genome Biol. 10, R25.
Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y. et al. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. 43, 1035–1039.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T. et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, e49.

Wicker, T., Mayer, K.F.X., Gundlach, H., Martis, M., Steuernagel, B., Scholz, U., Simkova, H. et al. (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23, 1706–1718.

Wu, Y., Bhat, P.R., Close, T.J. and Lonardi, S. (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. ed. L. Kruglyak. PloS Genet. 4, e1000212.

Wu, P., Shao, Z.-Q., Wu, X.-Z., Wang, Q., Wang, B., Chen, J.-Q., Hang, Y.-Y. et al. (2014) Loss/retention and evolution of NBS-encoding genes upon whole genome triplication of *Brassica rapa*. Gene, 540, 54–61.

Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., Hu, Z. et al. (2016) The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homeolog gene expression influencing selection. Nat. Genet. 1546–1718.

Yao, W., Li, G., Zhao, H., Wang, G., Lian, X. and Xie, W. (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol. 16, 187.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 17, 847–848.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

Zhao, X., Li, B., Zhang, K., Hu, K., Yi, B., Wen, J., Chaozhi, M. et al. (2016) Breeding signature of combining ability improvement revealed by a genomic variation map from recurrent selection population in *Brassica napus*. Sci. Rep. 6, 29553.

**Supporting information**

Additional Supporting Information may be found online in the supporting information tab for this article:

**Figure S1** Dotplot comparison of the Darmor-bzh assembly (y-axis) and the Tapidor assembly (x-axis) pseudomolecules.

**Figure S2** Comparison of ranks between 19 linkage maps from MSTMap and physical placement using the Darmor-bzh genome assembly as reference.

**Figure S3** Number of shared Swiss-Prot hits between genes located in collapsed and repetitive regions between Darmor-bzh and Tapidor.

**Table S1** Overlap between collections of previously unplaced contigs and contigs placed in the improved Darmor assembly.

**Table S2** Repeats in Darmor and Tapidor.

**Table S3** Gene expression values for 73 genes present in Darmor and predicted to be absent in Tapidor using RNA-Seq data from Darmor, Tapidor, Ningyou-7 and the Tapidor x Ningyou-7 population.

**Table S4** Gene expression values for 3 genes present in Tapidor and predicted to be absent in Darmor using RNA-Seq data from Darmor, Tapidor, Ningyou-7 and the Tapidor x Ningyou-7 population.

**Table S5** TopGO enriched terms for genes absent in Tapidor but present in Darmor.

**Table S6** Number and total sizes of collapsed regions in Darmor.

**Table S7** Number and total sizes of repetitive regions in Darmor.

**Table S8** Number of genes in predicted repetitive and collapsed regions in Darmor and Tapidor.

**Table S9** Shared domains between Darmor genes located in repetitive regions and Tapidor genes located in collapsed regions (only ten most common domains).

**Table S10** Ten most common Swiss-Prot hits for repetitive genes in Darmor and collapsed genes in Tapidor with an e-value cutoff of 1e-5.