Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids

Philipp E. Bayer¹, Armin Scheben¹, Agnieszka A. Golicz², Yuxuan Yuan¹, Sebastien Faure³, HueyTyng Lee⁴, Harmeet Singh Chawla⁵, Robyn Anderson¹, Ian Bancroft⁷, Harsh Raman⁶, Yong Pyo Lim⁷, Steven Robbens⁸, Lixi Jiang⁹, Shengyi Liu¹⁰, Michael S. Barker¹¹, M. Eric Schranz¹², Xiaowu Wang¹³, Graham J. King¹⁴, J. Chris Pires¹⁵, Boulos Chalhoub¹⁶, Rod J. Snowdon⁴, Jacqueline Batley¹⁷ and David Edwards¹.¹

¹School of Biological Sciences and the Institute of Agriculture, Faculty of Science, The University of Western Australia, Crawley, WA, Australia
²Plant Molecular Biology and Biotechnology Laboratory, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, VIC, Australia
³Innolea SAS, Mondonville, France
⁴Department of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig University Giessen, Giessen, Germany
⁵Department of Biology, University of York, York, UK
⁶NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, PMB, Wagga Wagga, NSW, Australia
⁷Department of Horticulture, Chungnam National University, Daejeon, South Korea
⁸BASF Innovation Center, Gent, Belgium
⁹Institute of crop science, Department of Agronomy and Plant Breeding, Zhejiang University, Hangzhou, China
¹⁰Chinese Academy of Agricultural Sciences, Oil Crops Research Institute, Wuhan, China
¹¹Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ, USA
¹²Bioysystems Group, Wageningen University and Research Center, Wageningen, The Netherlands
¹³Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences (IVF, CAAS), Beijing, China
¹⁴Southern Cross Plant Science, Southern Cross University, Lismore, NSW, Australia
¹⁵Division of Biological Sciences, Bond Life Sciences Center, University of Missouri, Columbia, Missouri, USA

Received 28 January 2021; revised 11 July 2021; accepted 20 July 2021.

*Correspondence (Tel +61 (08) 6488 2415; fax +61 (08) 6488 1108; email dave.edwards@uwa.edu.au)

**Key words:** *Brassica*, pangenome, XGBoost, gene loss propensity, machine learning, transposable elements.

**Summary**

Plant genomes demonstrate significant presence/absence variation (PAV) within a species; however, the factors that lead to this variation have not been studied systematically in *Brassica* across diploids and polyploids. Here, we developed pangenomes of polyploid *Brassica napus* and its two diploid progenitor genomes *B. rapa* and *B. oleracea* to infer how PAV may differ between diploids and polyploids. Modelling of gene loss suggests that loss propensity is primarily associated with transposable elements in the diploids while in *B. napus*, gene loss propensity is associated with homeologous recombination. We use these results to gain insights into the different causes of gene loss, both in diploids and following polyploidization, and pave the way for the application of machine learning methods to understanding the underlying biological and physical causes of gene presence/absence.

**Introduction**

A single reference genome does not represent the gene content of a species due to gene presence/absence variation (PAV) between individuals. In plants, genome duplication through polyploidization provides an opportunity for differential gene loss and subsequent presence/absence variation between individuals, and species that have experienced relatively recent polyploidy often host a relatively high proportion of dispensable genes. Several studies have examined gene conservation and loss following polyploidization. Neo-functionalization of duplicated genes has been observed in cotton (Adams et al., 2003; Kong et al., 2010; Yang et al., 2017), while in *Brassica napus*, homeologous exchange (HE) between chromosomes is associated with gene loss (Hurgobin et al., 2018) and with the generation of novel chimeric genes (Zhang et al., 2020).

Differential fractionation of genomes has been observed following ancient triplication in the diploid Brassica species *B. rapa* and *B. oleracea* (Cheng et al., 2014), while in octoploid strawberry (*Fragaria ananassa*), the diploid *F. vesca* subgenome dominates the other three subgenomes, having lost the fewest genes (Edger et al., 2019). Differential loss and retention of genes following two rounds of polyploidy has been reported in hexaploid bread wheat (*Triticum aestivum*) (Berkman et al., 2003).

reported in Song et al., (2020), which may be due to different de novo repeat-finding pipelines. In the v9 assembly, the total size of all common repeat classes increased two-fold. For example, Helitron repeat content grew from 153 to 240 Mbp (Tables S2–S4). The difference in the size of C02 is explained by the difference in assembled repetitive elements: in v4.1, C02 contains 26 Mbp of repetitive elements, while in the v9 assembly, C02 contains 91 Mbp of repeats.

Construction of three new pangenomes

Using the iterative mapping and assembly approach (Bayer et al., 2020; Hurgobin and Edwards, 2017), we have assembled pangenomes for B. oleracea, B. rapa and B. napus, representing the C, A and amphidiploid AC subgenomes, using 87, 77 and 79 individuals respectively (Table 2). Compared to the reference assemblies, each pangenome increased in size and gene content. The model of gene numbers converges asymptotically with the addition of each new individual suggesting that we have assembled almost all of the genes for these three species (Figure 1).

Annotation of the pangenomes predicted 58 315 gene-models in B. oleracea, 59 864 gene-models in B. rapa and 108 580 gene-models in B. napus. Out of these, 5963, 13 244 and 5735 gene-models are located on newly assembled pangenome contigs of the three pangenomes. Modelling of the pangenome size resulted in predicted total gene numbers for B. oleracea, B. rapa and B. napus of 58 347 (+/−2), 59 923 (+/−4) and 108 586 (+/−4), with predicted core gene numbers of 46 261 (+/−7), 40 391 (+/−11) and 65 096 (+/−150) respectively. The predicted pangenome size of B. oleracea is lower than the first B. oleracea pangenome which predicted a pangenome size of 63 865 +/−31 (Golicz et al., 2016) perhaps because the first pangenome used a wild relative in the calculations (B. macrocarpa), leading to a higher estimate in the first pangenome, but also used different annotation methods and repeat-masking methods. Similarly, the first B. napus pangenome predicted a pangenome size of 95 730 +/−11 (Hurgobin et al., 2018), lower than what we observe here. When we exclude synthetic lines, the predicted B. napus gene number drops to 108 537 (+/−9), while the core gene number increases to 79 663 (+/−119). Therefore, while the addition of the synthetic lines only increases the predicted total gene number by 49 genes, the proportion of genes that demonstrate presence/absence variation increases from 26 to 38% (Table 3).

Our findings suggest that the synthetics contribute a greater diversity of gene combinations without significantly increasing gene number. The discrepancy in gene content between synthetic and non-synthetic B. napus lines is expected due to differential gene loss between the multiple independent polyploidization events. Natural B. napus is predicted to have derived from a single polyploidy event, while each of the 20 synthetic lines is more recently derived from combinations of 11 female and 14 male parents (Schmutzer et al., 2015). Synthetic lines also demonstrate a greater diversity of homoeologous exchange events followed by subgenome-specific gene loss (Hurgobin et al., 2018).

Brassica rapa and B. oleracea diverged from a common ancestor around 3 MYA (Sun et al., 2019), so they may be expected to share a similar pangenome content. Based on read-alignments out of 58 315 B. oleracea genes, 57 729 (99%) are present in at least one B. rapa individual, and similarly, out of 59 864 B. rapa genes, 57 957 (97%) are present in at least one
B. oleracea individual. Of the 108 580 B. napus genes, 105 149 (97%) and 106 977 (99%) are present in at least one individual of B. oleracea and B. rapa respectively (Figure 2, Table 3). B. rapa has a greater proportion of dispensable genes (33%) than B. oleracea (21%) (Figure S3), suggesting greater genetic diversity in B. rapa, which is in line with a higher genetic diversity observed in the A subgenome of B. napus (Wu et al., 2019). Only 360, 711 and 955 genes were found to be unique in B. oleracea, B. rapa and B. napus respectively. Some of these are likely to be annotation artefacts or genes that have not yet been sampled in the other species, though this result does suggest that there may be genes unique to these species that could be of agronomic interest (Figure 3).

### Table 1
Assembly statistics for the newly assembled B. napus cv. Darmor-bzh v9 compared with v4.1 (Chalhoub et al., 2014)

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Assembly size (Mb)</th>
<th>Anchored chromosome (Mb)</th>
<th>TEs (%)</th>
<th>Number of annotated genes</th>
<th>Completeness (BUSCO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V4.1 (Chalhoub et al., 2014)</td>
<td>850.3</td>
<td>645.4</td>
<td>46.5</td>
<td>101 040</td>
<td>99.5%</td>
</tr>
<tr>
<td>v9</td>
<td>1043.4</td>
<td>933.3</td>
<td>64.5</td>
<td>108 580</td>
<td>99.5%</td>
</tr>
</tbody>
</table>

### Table 2
Pangenome additional contigs assembly statistics

<table>
<thead>
<tr>
<th>Pangenome</th>
<th>Assembly size (Mb)</th>
<th>Assembly N50</th>
<th>Predicted genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brassica oleracea</td>
<td>121.8</td>
<td>3848</td>
<td>6715</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>180.5</td>
<td>2500</td>
<td>19 767</td>
</tr>
<tr>
<td>Brassica napus</td>
<td>87.2</td>
<td>2295</td>
<td>5060</td>
</tr>
</tbody>
</table>

### Gene loss specific to B. Rapa rapid cycling lines
PCA-clustering of B. rapa individuals identified a highly diverged cluster consisting of rapid cycling, self-compatible lines that have undergone intensive selection (FastPlants sc, FPSc). In these individuals, an additional 177 genes were found to be dispensable compared to the non-FPSc B. rapa individuals. Proteins encoded by these 177 genes share sequence identity with stress-response genes including HVA22 (a stress-response gene which regulates vesicular traffic (Brands and Ho, 2002)) and G-type lectin S-receptor-like serine/threonine-protein kinase SRK, a salinity-stress linked regulator (Sun et al., 2013) which is also involved in self-incompatibility (Zhang et al., 2011). The loss of these abiotic stress-related genes may be associated with faster growth of these plants. As the FPSc lines are self-compatible it may be expected that these lines have lost the self-incompatibility-linked genes SLG, SRK and SCR/SP11 within the S-locus (Nasrallah, 1997). However, versions of these three genes are present in all of the FPSc lines, suggesting that self-compatibility in these lines is not caused by gene loss but rather by previously described polymorphisms (Kitashiba and Nasrallah, 2014).
Dispensable genes are commonly associated with abiotic and biotic stress

Dispensable genes are annotated predominantly with GO-terms associated with biotic and abiotic stress response for each of the three Brassica pangenomes (Table S5), with the term ‘defense response’ (GO:0006952) appearing significantly enriched in variable genes of B. oleracea, B. rapa and B. napus. Dispensability of stress response genes has been observed previously in crop pangenomes (Bayer et al., 2019; Golicz et al., 2016). In the B. oleracea pangenome, the GO terms ‘response to salt stress’ and ‘defense to bacterium’ were enriched in dispensable genes (Golicz et al., 2016), while in the wheat pangenome, ‘defense response’ was among the GO terms with the greatest enrichment in dispensable genes (Montenegro et al., 2017). Similar patterns were observed in the pangenomes of rice (Zhao et al., 2018), B. napus (Hurgobin et al., 2019), sunflower (Yu et al., 2018), sesame (Yu et al., 2019), pigeon pea (Zhao et al., 2020), sunflower (Hübner et al., 2019) and soybean (Liu et al., 2020b), where biotic and abiotic stress resistance-related genes were enriched among variable genes.

The strong but variable selection pressure on disease resistance genes associated with the presence or absence of associated pathogens likely impacts their differential conservation and loss between individuals. We found 206, 379 and 445 nucleotide-binding leucine-rich repeat (NLR) genes in B. oleracea, B. rapa and B. napus respectively. The B. oleracea pangenome contained 89 fewer NLR genes than the B. napus C subgenome, while in contrast, the B. rapa A subgenome assembly contained 52 more NLR genes than the B. napus A subgenome. Many of these additional B. rapa NLR genes were not found in the B. rapa reference assembly, highlighting the importance of pangenomes for species comparisons (Figure S4a). This pattern of differential loss was not apparent for two other classes of genes involved in disease resistance, RLP and RLK (Figure S4b), suggesting that the observed differences are not assembly artefacts and that there is a range of R-genes that are only present in the B. rapa gene pool and not in the B. napus gene pool.

Protein-protein interaction networks and the pangenome

Gene conservation and loss are associated with many factors. It has previously been observed that genes associated with protein-protein interaction networks tend to be more resistant to loss following polyploidy than genes outside of such networks. However, this resistance to loss is also affected by selection, with a greater loss of networked genes in new polyploids under strong selection than those under more relaxed selection (Schoenrock et al., 2017). This is exemplified in bread wheat, where the formation of the tetraploid occurred before domestication, while the hexaploid formed post domestication, with greater selection pressure that resulted in a greater loss of networked genes (Berkman et al., 2013).

In our newly assembled B. napus pangenome, excluding synthetic lines, 86% of core genes are predicted to be in networks, while only 72% of dispensable genes are predicted to be in networks (Table S6). There was a statistically significant difference in network retention between the two subgenomes, with 91% and 81% of core genes within networks in the A and the C subgenomes, respectively (X²-test, P < 0.005 in all cases). The retention of networked genes is slightly higher in the diploid species, with 87% and 90% of B. oleracea and B. rapa core genes predicted to be in networks compared with 86% of B. napus core genes in networks (Table S6), while only 68% and 70% of dispensable genes are predicted to be in networks. In the two diploids, as in B. napus, there was a statistically significant association between membership in protein interaction networks and variable genes (X²-test, P < 0.005). The diploid genomes may be under greater pressure to maintain networked genes, as the presence of a duplicate gene set in the polyploid may partially compensate for the loss of genes in functional networks.

Searching for A and C genome ancestors

Several genomic studies have attempted to identify the diploid parents of B. napus (Lu et al., 2019; Song et al., 2020). Here, we compared PAV patterns based on PCA between the two B. napus subgenomes and the B. rapa and B. oleracea individuals. This identified close relatives for the A subgenome (Figure 5a) but not for the C subgenome (Figure 5b), similar to previous observations based on SNPs, suggesting a complex origin for the C subgenome (Song and Osborn, 1992). We hypothesized that there may be different ancestors for different C subgenome chromosomes. We therefore repeated this analysis for each chromosome and observed inconsistencies between chromosomes (Figures S6 and S7). For example, A05 shows very little divergence between

### Table 3  Shared genes between the three pangenomes based on exon-level read alignments. For B. rapa, FPSc (Fast Plants, self-compatible) and non-FPSc lines are compared. For B. napus, non-synthetic and synthetic lines are compared.

<table>
<thead>
<tr>
<th></th>
<th>B. oleracea pangenome</th>
<th>B. rapa pangenome</th>
<th>B. napus pangenome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total genes</td>
<td>58 315</td>
<td>59 864</td>
<td>108 580</td>
</tr>
<tr>
<td>Dispensable genes within the same species</td>
<td>12 354 (21%)</td>
<td>With FPScs 19 912 (33%)</td>
<td>Without FPScs 19 735 (33%)</td>
</tr>
<tr>
<td>Core genes within the same species</td>
<td>45 961 (79%)</td>
<td>With FPScs 39 952 (67%)</td>
<td>Without FPScs 40 129 (67%)</td>
</tr>
<tr>
<td>Present in all three species in at least one individual each</td>
<td>57 717 (99%)</td>
<td>57 941 (97%)</td>
<td>104 465 (96%)</td>
</tr>
<tr>
<td>Present only in...</td>
<td>B. napus and B. oleracea 226 (0.4%) 0</td>
<td>648 (0.6%)</td>
<td>0</td>
</tr>
<tr>
<td>B. napus and B. rapa</td>
<td>0 1198 (2%)</td>
<td>2512 (2.3%)</td>
<td>0</td>
</tr>
<tr>
<td>B. oleracea and B. rapa</td>
<td>12 (0.02%)</td>
<td>16 (0.02%)</td>
<td>0</td>
</tr>
<tr>
<td>B. napus</td>
<td>0 0</td>
<td>955 (0.9%)</td>
<td>0</td>
</tr>
<tr>
<td>B. rapa</td>
<td>0 711 (1.1%)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>B. oleracea</td>
<td>360 (0.6%)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
individuals, which may be due to a previously described low frequency of homoeologous recombination of this chromosome (Pele et al., 2017). C03 and C09 diverged the most, possibly due to elevated crossover frequency. However, we found no chromosome-specific ancestors, suggesting that the C-genome ancestors are not represented by the publicly available *B. oleracea* data.

Comparing transposon content between *B. oleracea*, *B. rapa* and *B. napus*

Many traits of agronomic interest in *B. napus* and its diploid ancestors have been linked with transposon insertions, including an LTR-insertion linked with resistance to pod shattering and silique length (Liu et al., 2020a), and hAT, MITE and LINE insertions linked with flowering time (Song et al., 2020). In *B. napus*, a Helitron insertion within the promoter region of the self-incompatibility gene *BnSP11-1* has been linked with self-fertilization (Gao et al., 2016). This insertion has not been observed in the diploid ancestors, suggesting that it arose after the formation of the polyploid *B. napus*.

Here most classes of transposons show a similar abundance between the A and C subgenomes of *B. napus* and their respective diploid ancestors *B. rapa* and *B. oleracea* (Tables S2–S4, S7–S12). For example, the percentage of hAT (DNA/DTA) elements ranged from 0.8 to 0.9% in *B. oleracea* and 0.5 to 0.8% in the C subgenome of *B. napus*, and a range of 0.7% to 0.9% in *B. rapa* compared with 0.6% to 0.9% in the A subgenome of *B. napus*. However, other classes of transposons

---

**Figure 2** Genes shared across *B. oleracea*, *B. rapa* and *B. napus* in the three assembled pangenomes. (a) *B. oleracea* pangenome (58,315 genes), (b) *B. rapa* pangenome (59,864 genes) and (c) *B. napus* pangenome (108,580 genes).
Factors influencing gene loss propensity in the three pangenomes

We examined factors that may influence gene loss propensity. We built models that used genomic features to predict gene loss propensity in the three pangenomes to ask which genomic features have the largest impact on gene loss. These features include distance from centromeres (Mason et al., 2016), gene size, pseudomolecule size, distance from transposons, and in B. napus, whether a gene is located in a block syntenic with the homoeologous genome (Figure S8), using genes located only on pseudomolecules and ignoring B. napus genes only variable in synthetic lines. This builds on previous observations in B. oleracea showing that dispensable R-genes are closer to transposable elements than expected (Bayer et al., 2019), frequent nonreciprocal homoeologous exchanges between chromosomes in B. napus (Sharpe et al., 1995), and lineage-specific gene loss propensity across eukaryotes (Krylov et al., 2003).

We compared five different statistical and machine learning approaches (Logistic Regression, Gaussian Naïve Bayes, Random Forest, AdaBoost and XGBoost) and settled on gradient boosting models (XGBoost) because this model showed the highest accuracy (0.86) and F1-score (0.23) (Table S13). We built gradient boosting models predicting gene loss propensity while accounting for the strong class imbalance by using different sample weights, balancing of positive and negative weights, stratified test and training data, and a Bayesian hyperparameter search to optimize model parameters. These models achieved an accuracy of 85% (AUC: 0.7, average precision-recall score: 0.2, F1: 0.18) in B. napus, 88% in B. oleracea (AUC: 0.6, average precision-recall score: 0.1, F1: 0.01) and 86% in B. rapa (AUC: 0.6, average precision-recall score: 0.14, F1: 0.02) (Figure S9). Confusion matrices revealed that all models had an almost 99% accuracy in predicting whether a gene is core (98% accuracy in B. napus), but poor accuracy in predicting whether a gene is dispensable (16% accuracy in B. napus) (Table S14). This indicates that the features used in these models do not fully explain gene loss, but explain the extent of gene retention. It is possible that a portion of gene loss in Brassica is truly random, in which case the model has no means to explain gene loss. Another possible reason for the low predictability of variable genes in this model is that there are different types of variable genes that we currently cannot distinguish. Genes that are lost due to homeologous recombination are indistinguishable from novel genes created by Helitrons during every generation of the true Helitron content as the accurate prediction of Helitrons remains challenging (Ou et al., 2019).

Class II DNA transposons of superfamily CACTA (DNA/DTC) make up between 1.9% to 2.4% of the B. oleracea genome and 0.9% to 1.4% of the B. napus genome. We observed an increased number of CACTA transposons in the B. napus C subgenome compared to B. oleracea (2.4% compared with 1.9%). The greater abundance of CACTA elements in B. oleracea compared to B. rapa has been observed before (Alix et al., 2008) and CACTA elements have undergone several rounds of amplification since B. rapa and B. oleracea divergence. Similar CACTA expansions have been observed in amphidiploid cotton compared with its diploid ancestors (Chen et al., 2020), though in our study the difference may be due to repetitive elements collapsing in the B. oleracea assembly, while they were assembled correctly in the more complete B. napus assembly. A recent high-quality genome of B. napus cv. Z511 (Chen et al., 2021) found similar recent repeat expansions compared to the diploid ancestor which supports our findings.
proximity to transposable elements was among 13 and 12 of the top 20 predictors of gene loss propensity in *B. rapa* and *B. oleracea*, respectively; however, in the *B. napus* pan-genome-based model, transposable elements appeared only three times within the top 20 strongest predictors. In *B. napus*, membership in homoeologous blocks and position on different chromosomes were among the strongest ten predictors (Figure 4). This suggests that different mechanisms of gene loss dominate in the diploids and the amphidiploids, with homoeologous exchanges being most strongly linked with gene dispensability in *B. napus*, and transposable elements being most strongly linked with gene dispensability in *B. rapa* and *B. oleracea* (Figure 4).

We examined which rare factors have an impact on the prediction of gene loss propensity using the in-built F-score of XGBoost. In *B. rapa*, the strongest rare predictors of gene loss propensity were the presence of LTR and Helitron repeats, while in *B. oleracea* MITEs and pseudomolecule position were the strongest predictors of gene loss propensity. Interestingly, MITEs were common factors between *B. oleracea* and *B. napus*, suggesting that they play a greater role in the shared C genome.

When plotting the importance of ‘distance to centromere’ for each pseudomolecule separately, the *B. napus* model shows a clear pattern of increasing loss propensity distal to the centromeres, while in the corresponding plots for *B. oleracea* and *B. rapa*, gene loss propensity is distributed across the pseudomolecules (Figure 5). In wheat and *B. napus*, HEs show a similar pattern, with a greater number of HEs towards the telomeres (Zhang et al., 2020), and again indicates the importance of homoeologous recombination in predicting dispensable gene status in *B. napus*.

Subgenome dominance is a well-established phenomenon in polyploids and has previously been observed for specific regions in *B. napus* (Wu et al., 2018; Xie et al., 2019; Zhou et al., 2016). However, studies of subgenome dominance differ in their methodology, with some focusing on differences in gene expression between homoeologous gene-pairs, and others on gene loss. It has been shown that A subgenome regions are more likely to be replaced by C subgenome regions following homoeologous recombination (Bird et al., 2019; Hurgobin et al., 2018) but it is currently unclear if this is related to subgenome expression dominance.

Within *B. napus*, subgenome dominance has usually been observed through differences in gene expression levels between subgenomes (Bird et al., 2019) though it has also been associated with differential gene loss between the subgenomes (Hurgobin et al., 2018). Differential gene loss has also been linked to subgenome dominance in the tetraploid ancestors of *A. thaliana* (Thomas et al., 2006) and maize (Woodhouse et al., 2010). The pseudomolecules C01, C02 and C09 have the strongest association with gene loss propensity among the pseudomolecules tested. This agrees with previous observations showing preferential homoeologous exchange from the A subgenome to the C subgenome in *B. napus* (Hurgobin et al., 2018). Interestingly, these three chromosomes are also the fourth, second and third-longest chromosomes in *B. napus*, suggesting that preferential loss may be associated with longer chromosomes, as previously observed (Chalhoub et al., 2014). However, the longest chromosome, C03, does not appear in the ranking of chromosomes associated with gene loss, suggesting that other mechanisms such as selection may prevent genes on C03 from being lost.

Additional information such as variation in chromosome architecture and behaviour (e.g. crossover frequency) is likely to improve the accuracy of our models, as seen in *B. rapa/B. oleracea* where gene retention is associated with three-dimensional chromosomal organization (Xie et al., 2019).

This study provides insights into the evolution of *Brassica* genomes through a comparative analysis of gene presence/absence variation at the species level. We have shown that gene loss propensity differs between the diploid progenitors of *B. napus* and highlight the genomic differences between synthetic and natural *B. napus* lines. We built models linking the physical location of genes with their gene loss propensity. These models show that the position of a gene on the chromosome is the strongest predictor of gene loss propensity in polyploid *B. napus*, while transposable elements have a greater role in gene loss in the diploids. These results pave the way for the application of machine learning methods to understanding the underlying biological and physical causes of gene presence/absence.

**Methods**

**A new Darmor-Bzh reference genome**

A new *Brassica napus* cv. Darmor-Bzh reference genome assembly was assembled by NRGene using the DeNovoMAGIC™ software platform (NRGene, Nes Ziona, Israel), a proprietary DeBruin graph-based assembler. This assembler used paired-end Illumina reads (450 and 800 bp insert sizes) along with mate-paired Illumina reads (2-4 and 8–10 kb insert sizes) with a total coverage >180×. Scaffolds were joined using 80× of 10× Chromium data and manually corrected using published genetic maps (Chalhoub et al., 2014). The scaffolds were ordered into pseudomolecules using the v4 assembly (Chalhoub et al., 2014) and RaGOO v1.02 (Alonge et al., 2019). Gene space completeness of both assemblies was assessed using BUSCO v5.1.2 database: viridiplantae_odb10 (Simão et al., 2015). The two assemblies were aligned using minimap2 v2.18 and differences were visualized using pafr v0.0.2 (https://github.com/dwinter/paf r). Repeats in the new Darmor-Bzh assembly and the v4 assembly were searched using EDTA v1.9.6 (Ou et al., 2019) and mapped using RepeatMasker v2.0 (Smit and Hubley, 2008).

**Construction of three new pangeneses**

We assembled three pangeneses for *B. napus*, *B. oleracea* and *B. rapa* using the approach of Golicz et al., (2016). We used publicly available paired-end Illumina reads with more than 9× coverage (except the reference cultivar Darmor-Bzh) of 87, 77 and 59 individuals for *B. oleracea*, *B. rapa* and *B. napus* respectively (Tables S15). We sequenced 20 additional *B. napus* individuals using Illumina HiSeq 3000 (PRJNA613532). This number of individuals is sufficient to capture the majority of gene content in the population as in previous pan-genome assemblies, the rate with which novel gene content increases with each added individuals stops growing after 10 to 50 individuals (Gao et al., 2019; Golicz et al., 2016; Hurgobin et al., 2018; Montenegro et al., 2017).

We aligned these three datasets separately to the new *B. napus* assembly, the v2.1 *B. oleracea* assembly (Parkin et al., 2014) and the v3.0 *B. rapa* assembly (Zhang et al., 2018) respectively. Bowtie2 v2.2.9 (Langmead and Salzberg, 2012) was used for all read alignments (options: –end-to-end, –sensitive). The three sets of reads that did not align were assembled using MaSuRCA v3.2.3 (Zimin et al., 2013) into three pangeneses: one
for *B. oleracea* using only *B. oleracea* individuals, one for *B. rapa* using only *B. rapa* individuals, and one for *B. napus* using only *B. napus* individuals. The resulting contigs were aligned with NCBI-NR (accessed 2nd June 2019) using blast + v2.5.0 (Camacho et al., 2009), and contigs with best hits outside the Viridiplantae were considered to be contamination and removed from subsequent steps.

**Gene prediction**

For each species pangenome and the reference genome, all publicly available paired RNASeq data (Table S16) were used in the BRAKER v2.0 (Hoff et al., 2019) gene prediction pipeline after each pangenome was soft-masked using RepeatModeler (Smit and Hubley, 2008) and RepeatMasker (Smit et al., 1996) to avoid removing true genes (Bayer et al., 2018). BRAKER produces AUGUSTUS (Stanke et al., 2006) and GeneMark-EX (Lomsadze et al., 2014) gene predictions. All RNASeq data were aligned using HISAT2 v2.1.0 (Kim et al., 2019) and converted into genome coordinates using StringTie v1.3.4 (Pertea et al., 2015). The RNASeq alignment coordinates were used together with RepeatModeler-based repeat regions, AUGUSTUS and GeneMark-EX predictions, and gene models of the already published *B. oleracea* v2.1 (Parkin et al., 2014), *B. rapa* v3 (Zhang et al., 2018) and *B. napus* v4 (Chalhoub et al., 2014) in the EVidenceModeler v1.1.1 (Haas et al., 2008) pipeline to produce final gene models. Gene models without RNASeq
support and no hits in the previously published gene models were removed from the final annotation. Disease resistance gene analogue (RGA) candidates were predicted using RGaugury (Li et al., 2016).

Gene presence/absence calling

Gene presence/absence variation (PAV) was called using an approach based on SGSGeneLoss (Golicz et al., 2015). For each of the three pangenomes, we aligned all B. oleracea, B. rapa and B. napus reads using Bowtie2 v2.2.9 (Langmead and Salzberg, 2012). Mosdepth v0.2.2 (Pedersen and Quinlan, 2018) and bedtools v 2.27.0 (Quinlan and Hall, 2010) were used to calculate the coverage of all gene exons. Genes where all exon bases were covered by fewer than 2 reads and where the exons’ length was covered by less than 5% of their total length were deemed to be absent. While this may lead to some genes being incorrectly classified as present when they are absent, these parameters provide confidence that absent gene calls are truly absent. We used these results to calculate three PAV tables: one for the B. oleracea pangene containing gene presence information all B. oleracea, B. rapa and B. napus individuals, one for the B. rapa pangene containing gene presence information for all B. oleracea, B. rapa and B. napus individuals, and one for the B. napus pangene containing gene presence information for all B. oleracea, B. rapa and B. napus individuals.

PAV-based PCA modelling of dispensable and core genes and GO-enrichment were performed using R v3.6.3 (R Core Team, 2020) using the packages logisticPCA (Landgraf and Lee, 2015), minpack.nlm (Elzhov et al., 2010) and topGO (Alexa and Rahnenführer, 2009). GO-terms were assigned to all proteins using PANNZER2 (Törönen et al., 2018) (accessed 5.7.2020, database: Viridiplantae). For each possible number of combinations of genomes, 500 000 pairs were chosen for the modelling of pangenome and core gene numbers.

Proteins were compared using DIAMOND v0.9.29.130 with the STRING v11 Arabidopsis database (Szklarczyk et al., 2019) to find proteins within functional networks. Association between network membership and gene status was assessed using the function chisq.test() implemented in R v3.6.3 (R Core Team, 2020). Genes were located within syntenic blocks by self-comparison of the B. napus annotation using MCScanX (Wang et al., 2012).

Assessing gene loss propensity using machine learning

Gene absence was predicted by building three separate feature tables for the three genomes, using genes located on pseudomolecules only, and genes that are lost in at least 2 individuals. The feature tables contained for each gene: which pseudomolecule the gene is located on, GC content, distance to the end of the pseudomolecule, overlap/1kb/2kb/3kb distance to de novo predicted transposon-classes as predicted by EDTA v1.9.6 (Ou et al., 2019), distance to the centromeres as described in (Mason et al., 2016), and, for B. napus, whether a gene was located within a syntenic block. Genes variable only in synthetic individuals were assumed to be core. Accuracy, F1-score and AUC-scores were compared between five machine learning
approaches (logistic regression, Gaussian Naive Bayes, Random Forest, AdaBoost and XGBoost). Three different XGBoost v1.0.2 models (Chen and Guestrin, 2016) were trained using the three PAV feature tables for the B. oleracea, B. rapa and B. napus pangenomes. For this we removed the PAV information of the other species – i.e. the B. oleracea pangenome gene feature table contained only information as to whether a gene was variable of B. oleracea individuals, not B. rapa or B. napus individuals.

Scikit-learn v0.21.3 (Pedregosa et al., 2011) was used to calculate supporting statistics such as F1-score, receiver operating characteristic curves and prediction accuracy. The feature table was split into an 80/20 training/test dataset while stratifying for the gene PAV output using scikit-learn’s train_test_split() function with a random state of 123. Sample weights were computed using the compute_sample_weight function in scikit-learn. The following XGBoost parameters were optimized using scikit-optimize BayesSearchCV: learning_rate (step size shrinkage used in updates to prevent overfitting), min_child_weight (minimum sum of instance weight needed in child, used to decide whether to stop partitioning), max_depth (maximum depth of a tree), max_delta_step (maximum step for each leaf update), subsample (subsample ratio of all training instances), colsample_bytree (subsample ratio of columns when constructing trees), colsample_bylevel (subsample ratio of columns for each level), reg_lambda (L2 regularization term on weights), reg_alpha (L1 regularization term on weights), gamma (minimum loss reduction required to make a further partition), n_estimators (number of trees in the model) and scale_pos_weight (controls level), reg_lambda (L2 regularization term on weights), reg_alpha (L1 regularization term on weights), gamma (minimum loss reduction required to make a further partition), n_estimators (number of trees in the model) and scale_pos_weight (controls the balance of positive and negative weights) (Head et al., 2018). Model metrics were calculated using the scikit-learn functions confusion_matrix, accuracy_score, roc_auc_score and f1_score. Feature importance in the trained models was assessed using TreeExplainer in Shapley Additive Explanations (SHAP) v0.31.0 (Lundberg and Lee, 2017).

Acknowledgements
This work is funded by the Australia Research Council (Projects DP1601004497, LP140100537 and LP130100925), and resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. Dr. Philipp Bayer acknowledges the support of the Forrest Research Foundation. YP Lim was supported by the Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET), Ministry of Agriculture, Food and Rural Affairs (MAFRA), Ministry of Oceans and Fisheries (MOF), Rural Development Administration (RDA) and Korea Forest Service (KFS), South Korea. This work was supported by UK Biotechnology and Biological Sciences Research Council BB/LO02124/1 and BB/R019819/1 to IB.

Conflict of Interests
The authors declare no conflict of interests.

Author Contributions
PEB conceived the research. PEB, AS, AAG, YY and RA carried out the research.
SF, HL, HSC, IB, HR, SR, LJ, SL, MSB, ES, XW, GJK, JCP, BC, WJS and contributed to the genome assembly. YPL contributed additional B. rapa seeds. PEB, JB and DE co-wrote the manuscript. All authors read and contributed to the manuscript.

Data availability
All code generated for this study is available at https://github.com/AppliedBioinformatics/Brassica_oleracea_rapa_napus_code
All data generated for this study is available at BioProject PRJNA613532. The assemblies, annotations, PAV-matrices and other supporting data are available at https://doi.org/10.26182/5f1936836a1c4 and http://brassicagenome.net/databases.php.
JBrowse (Buels et al., 2016) and KnetMiner (Hassani-Pak et al., 2020) instances are available at http://brassicagenome.net/databases.php.

References


10 Philipp E. Bayer et al.
 Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Comparison between the v4.1 assembly (Chalhoub et al., 2014) and the new v9 assembly.

Figure S2 Comparison of repeat content by class in Mb between the two assemblies showing that the six most abundant classes have roughly doubled in size in the v9 assembly.

Figure S3 Number of core and dispensable genes for the A and the C genome, compared between B. napus and the C genome, compared between B. napus and the C genome.

Figure S4 (a) NLR-genes compared between B. napus, B. rapa and B. oleracea, along with additional pangenome contigs.

Figure S5 PCA plots based on PAV patterns of genes located on each chromosome in B. napus split into subgenomes A and C (subfigures A and B respectively) showing strong divergence in PAV patterns between some chromosomes of the B. napus A and the C subgenome, especially C03, C09, A05 and A07.

Figure S6 PCA plot showing divergence of individuals based on gene presence/absence patterns on the A genome. A) chromosome A01, B) A02, C) A03, D) A04, E) A05, F) A06, G) A07, H) A08, I) A09 and J) A10. FPSC: Fast Plants, self-compatible.

Figure S7 PCA plot showing divergence of individuals based on gene presence/absence patterns on the C genome.

Figure S8 Different kinds of reciprocal and non-reciprocal inferences after homoeologous recombination in B. napus.

Figure S9 Receiver-operating curves comparing the three XGBoost models trained on B. oleracea, B. rapa and B. napus data respectively.

Figure S10 Twenty features with the strongest impact on the B. rapa (A), B. oleracea (B) and B. napus (C) models measured by relative quantity as assessed using XGBoost’s inbuilt feature importance methods (‘cover’), showing that in rare feature attributes, the B. oleracea and the B. rapa model focus mostly on retrotransposons in its best-predicting attributes, and in B. napus, the best predictors are pseudomolecule membership.

Table S1 BUSCO results for B. napus v4.1 (Chalhoub et al., 2014) and the new NRGene v9 assembly.

Table S2 Comparison of repeats between B. napus v4.1 and v9 assembly.

Table S3 Comparison of repeats between B. napus v4.1 and v9 assembly.

Table S4 Comparison of repeats between B. napus v4.1 and v9 assembly.

Table S5 Top 15 enriched GO terms in the dispensable genes of B. oleracea, B. rapa, B. napus, and B. napus without synthetic lines.
Table S6 Numbers of core and dispensable genes in STRING functional networks without synthetic lines.
Table S7 Count of transposable elements per pseudomolecule in the *B. oleracea* assembly.
Table S8 Total length (Mbp) of transposable elements per pseudomolecule in the *B. oleracea* assembly.
Table S9 Total length (Mbp) of transposable elements as percentage of total pseudomolecule length in the *B. oleracea* assembly.
Table S10 Count of transposable elements per pseudomolecule in the *B. rapa* assembly.

Table S11 Total length (Mbp) of transposable elements per pseudomolecule in the *B. rapa* assembly.
Table S12 Total length (bp) of transposable elements as percentage of total pseudomolecule length in the *B. rapa* assembly.
Table S13 Comparison of models using the *B. napus* gene loss data.
Table S14 Confusion matrix for the three XGBoost models trained on *B. oleracea*, *B. rapa*, and *B. napus* data.
Table S15 Data used for the assembly of the three pangenomes.
Table S16 RNASeq data used for the annotation of the three pangenomes.