1 **Considering founding and variable genomes is critical in studying polyploid** 2 **evolution**

3
4 Xueling Ye^{1,2,†}, Haiyan Hu^{1,3,†}, Hong Zhou^{1,2,†}, Yunfeng Jiang^{1,2}, Shang Gao¹, Zhongwei
5 Yuan^{1,2}, Jiri Stiller¹, Chengwei Li³, Guoyue Chen², Yaxi Liu², Yuming Wei², You-Liang
6 Zheng², Chunji Liu^{1,*}

7
8 ¹ CSIRO Agriculture and Food, St Lucia, Queensland 4067, Australia.

9 ² Triticeae Research Institute, Sichuan Agricultural University, Wenjiang, Chengdu 611130,
10 China

11 ³ College of Life Science and Technology, Henan Institute of Science and Technology,
12 Xinxiang, Henan 453003, China

13 [†] These authors contributed equally to this publication

14
15 ***Corresponding author:** E-mail: chunji.liu@csiro.au

17 **Abstract**

18 A wide range of differences between the subgenomes, termed as subgenome asymmetry or SA,
19 has been reported in various polyploids and different species seem to have different responses
20 to polyploidization. We compared subgenome differences in gene ratio and relative diversity
21 between artificial and natural genotypes of several allopolyploid species. Surprisingly,
22 consistent differences in neither gene ratio nor relative diversity between the subgenomes were
23 detected between these two types of polyploid genotypes although they differ in times exposed
24 to evolutionary selection. As expected, the estimated ratio of retained genes between a
25 subgenome and its diploid donor was invariably higher for the artificial allopolyploid

genotypes due likely to the presence of variable genome components (VGC). Clearly, the presence of VGC means that exaggerated differences between a donor and a subgenome in a polyploid are inevitable when random genotypes were used to represent species of either a polyploid or its donors. SA was also detected in genotypes before the completion of the polyploidization events as well as in those which were not formed via polyploidization. Considering that significant changes during and following polyploidization have been detected in previous studies, our results suggest that the influence of VGC needs to be considered in evaluating SA and that diploid donors may define changes in polyploid evolution.

Key words: asymmetric evolution, subgenome asymmetry, donor, polyploids, progenitor

Introduction

Polyploidy, forming new species by enveloping two or more genomes within a single nucleus, is a prominent speciation process. Previous studies show that polyploidization was likely involved in the evolution of all flowering plant species (Bowers et al. 2003; Paterson et al. 2006; Jiao et al. 2011; Van et al. 2017; Ding and Chen, 2018). Understanding why polyploid species have been so pervasive and successful has attracted wide attention in the scientific community worldwide. As a result, possible effects of polyploidization on various aspects of plant biology have been intensively investigated (Feldman et al. 1997; Soltis and Soltis 1999; Li et al. 2014a; Soltis and Soltis 2016; Van et al. 2017; Edger et al. 2018; Ghanbari et al. 2019).

There are two major types of polyploids: autopolyploids and allopolyploids. The former results from whole-genome duplication (WGD) and each of them bears two or more copies of the same genome. The majority of the latter results from inter-specific or inter-generic hybridization followed by chromosome doubling, and each of them bears two or more different

genomes enveloped within a single nucleus. Many important crop species, including wheat, cotton, *Brassica*, potato and banana, are typical allopolyploids.

Results from previous studies show that different species react differently to polyploidization. Wheat (Feldman et al. 1997; Uauy et al. 2017) and *Brassica* (Gaeta et al. 2007; Liu et al. 2018) are among species for which rapid and massive changes in DNA sequences were detected in newly synthesized allopolyploid genotypes. However, such changes were not detected in either cotton (Liu et al. 2001; Page et al. 2016; Wang et al. 2018) or *Arabidopsis* (Wang et al. 2006; Del Pozo et al. 2015). It is also widely believed that different subgenomes of a polyploid species have different responses to internal interactions as well as external stimuli. Such differences, termed as contrasted plasticity, subgenome dominance, or asymmetric evolution (AE), result in subgenome asymmetry (SA) or biased features in gene number, diversity or expression between the constituting subgenomes (Feldman et al. 1997; Flagel and Wendel 2010; Liu et al. 2014; Zhang et al. 2015; Wang et al. 2017; Xu et al. 2018). Possible evolutionary advantages of SA have been proposed in various species (Feldman et al. 1997; Flagel and Wendel 2010; Grover et al. 2012; Pont et al. 2013; Liu et al. 2014; Zhang et al. 2015; Kasianov et al. 2017; Wang et al. 2017; Bao et al. 2018; Xu et al. 2018).

Recent studies reveal that an individual genotype does not possess all genes existing in a species. Genes not shared by different members of a species, termed as the dispensable or variable genome components (VGC), account for about 36% in bread wheat (Liu et al. 2016; Montenegro et al. 2017), 38% in barley (Ma et al. 2019), and as high as 50% in maize (Jin et al. 2016). However, possible influence of VGC in studying polyploid evolution has not been considered. Most of previous studies reporting the existence of SA were based on the analyses of polyploid genotypes only (e.g., Pont et al. 2013; Zhang et al. 2015; Wang et al. 2017). Limited comparisons between subgenomes and their progenitors have been carried out in some studies but used only random genotypes to represent the donor species (e.g., Flagel and Wendel

2010; Li et al. 2014a). Clearly, due to the existence of VGC, using random genotypes to represent either a polyploid or its progenitor species would likely lead to over-estimation of differences between a subgenome and its donor.

Artificial allopolyploid genotypes have been obtained in different studies for several polyploid species. Different from those natural genotypes, the artificial ones all have very short time-frames of existence. In other words, the artificial genotypes have not been exposed to evolutionary selection for long thus parts of the differences between them and those natural genotypes should belong to changes accumulated during evolution in the latter. Further, the exact parents for some of the artificial allopolyploid genotypes are known thus using them to estimate shared features between a given subgenome in a polyploid genotype and its diploid donor could avoid the interference of VGC. We thus decided to use the artificial polyploid genotypes as surrogates to represent the beginning of polyploid evolution in the study reported here. Our results do not only show that considering VGC is important in evaluating SA but also suggest that changes accumulated in polyploid evolution must be defined by their founding genomes.

Material and Methods

Genome and RNA sequences used

Sequence data from *Brassica*, wheat, cotton and *Arabidopsis* were used in this study (Supplementary table 1). All the sequences were downloaded from the National Centre for Biotechnology Information (NCBI) Short Sequence Read Archive (SRA) database (<http://www.ncbi.nlm.nih.gov/sra>).

For *Brassica*, a total of 26 sets of transcriptomic data from 25 genotypes and 15 sets of genomic data from 15 genotypes were used (Supplementary table 1). The transcriptomic

sequences were obtained from natural genotypes representing three allotetraploid species (*B. juncea* with a genome of AABB, *B. napus* with a genome of AACC and *B. carinata* with a genome of BBCC) and their diploid progenitors (*B. rapa* for A subgenome, *B. nigra* for the B subgenome, and *B. alboglabra* for the C genome). Sequences from artificial allotetraploids for each of the three allotetraploid species and their parental genotypes were also used. They included three artificial F1 hybrids with a genome of AC (di-haploids) and an allotetraploid derived from reciprocal hybridization between the diploid A and C donors (genomes CCAA). In addition, genomic sequences from 15 genotypes each representing the three diploid donor species were used to estimate the difference in relative genetic variation among these species.

For wheat, 30 sets of transcriptomic data from 16 different genotypes were collected for this study (Supplementary table 1). They included genotypes representing natural and artificial allotetraploids (genome AABB), natural and artificial allohexaploids (genome AABBDD), their three diploid donors (*T. urartu* accession for A genome, *Ae. longissima* accession for the S genome, *Ae. tauschii* accession for the D genome), and one tetraploid donor (*T. durum* (genome AABB) as the maternal donor to the hexaploid wheat). The wheat genotypes also included an artificial tri-haploid (genome ABD) and a hexaploid (genome AABBDD) derived from it by chromosome doubling. It is of note that the S genome from *Ae. Longissimi* was used as the surrogate for the B subgenome donor in the artificial polyploids. Results from a recent study indicates that the B subgenome donor was likely the product from introgressions of several *Aegilops* species that yet to be identified if not already extinct (El Baidouri et al. 2017).

For cotton, genomic sequences from seven genotypes were used (Supplementary table 1). They included four natural allotetraploids, two belonging to the subspecies *G. hirsutum* and the other two belonging to *G. barbadense*. They all share the same genome structure of AADD (Song et al. 2017). They also include one artificial tetraploid (genome AADD) and its two

diploid parents (an accession of *G. arboreum* as the A genome donor and an accession of *G. raimondii* as the D genome donor).

Eighteen sets of transcriptomic data from *Arabidopsis* were also used in this study (Supplementary table 1). They were derived from one artificial and nine natural allotetraploid genotypes (all with a genome of AATT) with their two diploid subgenome donors (an accession from *A. arenosa* as the A genome donor and an accession from *A. thaliana* as the T genome donor).

Transcriptomic sequence analyses

Based on the initial assessments using the software FastQC (version 0.11.5) (Andrews 2016), transcriptomic sequences were trimmed and filtered using the software SolexaQA++ (Cox et al. 2010). Low quality reads were removed with the criteria of Q<30 and length<50bp. As both the quality and quantity of sequences would likely influence the estimation of the numbers of genes in a genome and the percentage of shares genes between genotypes, we decided to control the quantity of sequences for each of genotype based on its relative genome size. For example, as the B subgenome or its diploid progenitor is bigger than the diploid A genome donor in wheat (Ling et al. 2013; IWGSC 2014; Avni et al. 2017), a larger quantity of the trimmed sequences was used for the former. Further, similar quantities of sequences were used for the artificial and natural genotypes (Supplementary table 1).

Following the quality and quantity controls, the selected transcriptomic sequences were assembled using Trinity (version 2.5.1) (Grabherr et al. 2011) with K-mers=25 and *de novo* assembled with a minimum length of 200bp. Redundant sequences were then removed using the cd-hit package (version 4.6.4) (Li and Godzik 2006) with the sequence similarity threshold of 95%. Coding sequences (CDS) for the assembly sequences were then predicted using TransDecoder (version 5.0.2) (<https://github.com/TransDecoder/TransDecoder/releases>).

In estimating genes shared between the subgenomes with their respective diploid donors, BLAST+ (version 2.7.1) (Camacho et al. 2009) was used to identify syntenic genes with a minimum E-value of 1e-5, a minimum coverage size of 100bp and a minimum identity threshold of 95%. The diploid sequences were first used as a query to BLAST against the polyploid sequences and then a reciprocal analysis was carried out by using the polyploid sequences as a query to BLAST against the diploid sequences. A custom script (Powell et al. 2017) was used to retrieve all those sequences where the polyploids or diploids gave the best hits. These sequences were used to calculate the percentages of genes shared between a given subgenome and its diploid donor.

Genomic sequence analyses

Quality control of the genomic sequences was also carried out using the software FastQC (version 0.11.5) (Andrews 2016) and low-quality sequences were removed using SolexaQA++ (Cox et al. 2010) with $Q < 30$ and $\text{length} < 50\text{bp}$. The selected genomic reads were assembled using the software SOAPdenovo2-r240 (Luo et al. 2012) with the default settings based on k-mer size 63 for cotton, and k-mer size 55 for *Arabidopsis*. The GapCloser (including in SOAPdenovo2 package) was used to reduce gaps among the assembled scaffolds and the cd-hit package (version 4.6.4) (Li and Godzik 2006) was used to remove redundant sequences. For the genomic sequences from cotton, gene prediction was conducted by comparing sequences with known protein structures with DIAMOND blastx (Buchfink et al. 2015) (E value $< 10^{-10}$ and bitscore > 60) based on the SWISS-PROT database (Bairoch and Apweiler 1999). The gene prediction of *Arabidopsis* was conducted by augustus (version 3.2.3) (Stanke and Morgenstern 2005) with the default settings. The genes predicted from *A. thaliana* and *A. arenosa* were used to identify shared genes in the subgenomes of the artificial allotetraploids, because sequences from the parents of the artificial allotetraploid were not available. The same

methods described above for the transcriptomic sequences were used to identify genes in genome or subgenome and to calculate the percentage of genes shared between them.

For estimating the relative diversity among the three diploid donor species in *Brassica*, genome sequences were used to identify SNPs in each of them. Sequence reads were mapped to the reference genome of *B. rapa* (assembly Brapa_1.0), *B. nigra* (assembly ASM168289V1) and *B. oleracea* (assembly BOL) with BWA (version 0.7.12) (Li and Durbin 2009), after viewed and sorted by SAMtools (version 1.5) (Li 2011), duplicates were removed by samtools rmdup and SNPs were identified using SAMtools mpileup by skipping alignments with mapQ smaller than 20, finally filtered SNPs with DP < 4 and SNPs within 3 base pairs of an indel by SAMtools/BCFtools (version 1.5) (Li 2011). Genome diversity for each of the species was calculated by dividing the total numbers of SNPs with the genome size.

Results

Gene ratios between the subgenomes for a given species are highly consistent despite of the wide variation in predicted CDS or gene numbers among different data sets used

As expected, the numbers of predicted CDS or genes do not only vary among genotypes but also for a given subgenome among genotypes. The largest variations were found for the three polyploid species of *Brassica*, where more than two-fold differences in the number of predicted genes were detected. These differences apparently reflect differences in the quantity and quality of sequences from various sources (Supplementary table 2). However, the differences did not seem affected the estimated gene ratios between subgenomes much. In fact, the estimated gene ratios between subgenomes among various genotypes for a given species are highly consistent. It was the same subgenome which gave a lower or higher ratio of genes among different genotypes for a given species, irrespective of either transcriptomic or genomic sequences were used for the analyses. Specifically, the B subgenome is the one giving a higher ratio of genes

in *B. juncea* (genome AABB); C subgenome for *B. carinata* (genome BBCC); C subgenome for *B. napus* (genome AACC); B subgenome for tetraploid (genome AABB) wheat and D subgenome for the tetraploid cotton (genome AADD) (table 1). The highly consistent gene ratios obtained between subgenomes among the different allopolyploid genotypes indicate that the use of large quantity of high-quality sequences is not critical in estimating gene ratios among subgenomes.

Significant and consistent difference in gene ratio among subgenomes was not detected between natural and artificial allopolyploids

Compared with natural ones, the artificial allopolyploid genotypes all have very short times frames of existence. We thus decided to use the artificial genotypes as surrogates to represent the beginning of polyploid evolution, and estimated changes accumulated during evolution. Surprisingly, little difference in gene ratio was detected between the artificial and natural polyploid genotypes for any of the species studied when sequences from the same tissue was used. These included the estimations based on the use of the transcriptomic data for the four artificial and five natural allotetraploid genotypes of *B. napus* (table 1 and fig.1); for the one artificial and two natural genotypes of *B. carinata* (table 1 and fig.1), for the one artificial and two natural polyploid genotypes of *B. juncea* (table 1 and fig.1), for the six natural and two artificial tetraploid genotypes of wheat (table 1 and fig. 1). Similar results were also obtained based on the analysis of genomic sequences on cotton, although the natural genotypes of this species have a much longer history of evolution compared with either the polyploid *Brassica* or wheat (Paterson et al. 2012) (table 1 and fig. 1).

Difference in gene ratios between subgenomes was detected before polyploidization events

As mentioned earlier, the process of forming the majority of allopolyploids can be divided into two steps: inter-specific or inter-generic hybridization and chromosome doubling. Assuming that the two diploid species (one with a genome of XX and the other YY) are involved in forming an allotetraploid genotype, the first step results in the formation of a di-haploid with a genome XY, and the allotetraploid with a genome XXYY will then be formed with the completion of the second step. To examine possible changes in gene ratios among the subgenomes before and after the polyploidization event, we analysed an artificial F1 tri-haploid hybrid (genome ABD) and its corresponding allohexaploid genotype (genome AABBDD) based on the available transcriptome sequences from these materials. This analysis showed that the gene ratios among the subgenomes between these two genotypes were very similar (table 2). As rapid and intensive changes following polyploidisation have been reported in previous studies (Soltis and Soltis 1999; Adams et al. 2005; Soltis et al. 2015; Soltis and Soltis 2016), the results from this analysis can only be explained by the likelihood that the founding genomes define changes caused by polyploidization in the subgenomes of a polyploid.

Gene ratios in allopolyploid genotypes not formed via polyploidization events

Different from those for the majority of allotetraploid species, the allotetraploid species of *Arabidopsis suecica* can be synthesized without the step of polyploidization. As the A subgenome donor is already a tetraploid (with a genome of AAAA), the first step in generating artificial *A. suecica* genotypes is to generate an autotetraploid genotype of *A. thaliana* by doubling its chromosomes. Then an allotetraploid with a genome of AATT can be generated by crossing the two autotetraploid genotypes. Thus, comparing gene ratios between the artificial allotetraploid genotypes with those natural ones of *A. suecica* provides another mean to test the possible effect of polyploidization on SA. Again, little difference in gene ratio

between the subgenomes was detected between these two types of allotetraploid genotypes (table 3 and fig. 1).

Differences in relative gene diversity among the three subgenomes of the various allopolyploid *Brassica* species correlate with those among their diploid donors

To evaluate the possible relationship between a subgenome and its diploid donor, we estimated the gene diversity among the three diploid donors which contributed to the three subgenomes of the polyploid genotypes assessed. As expected, the numbers of predicted SNPs increased with the increase in quantity of sequences used. With a similar amount of sequences used, the relative gene diversity among the three diploid donors was in the order of *B. rapa* (AA) > *B. oleracea* (CC) > *B. nigra* (BB), which is the same as their order as subgenomes in the various allotetraploids (Supplementary table 2). The same order in relative gene diversity between the three subgenomes and the three diploid donors again cannot be a coincidence. Rather, they must indicate changes occurred in the subgenomes must be defined by their founding genomes.

Difference in the ratio of shared genes between a subgenome and its diploid donor between artificial and natural polyploid genotypes

Considering the existence of VGC, the ratios of shared genes between a subgenome and its diploid donor should be higher for an artificial polyploid than for those natural polyploids. This is because that the exact diploid donors for the former are known thus influences from the VGC can be removed when conducting such an analysis. We conducted such a comparison for *Brassica*, wheat, cotton and *Arabidopsis*. As expected, the ratios shared genes between the artificial tetraploid with its two diploid parental genotypes are indeed significantly higher than those shared between the natural genotypes and those ‘random’ diploid genotypes from the progenitor species for each of these species assessed (tables 1, 3). These results reflect the

importance of considering variations among individual genotypes within and between species in investigating the mechanism of polyploid evolution. In other words, the presence of VGC means that using random diploid genotypes would invariably lead to under-estimation of shared gene ratios between a subgenome and its diploid donor.

Discussion

Polyploids are pervasive in flowering plants and SA has been reported in all allopolyploid species investigated to date. Differences in many aspects, including gene number and relative diversity, have been used to assess SA (Akhunov et al. 2003; Qi et al. 2004) and that and epigenetic variation seems to affect subgenome dominance (Hollister and Gaut 2009; Guo and Han 2014; Song et al. 2015; Cheng et al. 2016; Bird et al. 2018; Cheng et al. 2018; Ding and Chen 2018; Xu et al. 2018). Many previous studies on polyploid evolution analysed natural genotypes only and treated differences detected between subgenomes as originated from asymmetric evolution (Flagel et al. 2010; Schnable et al. 2011; Feldman and Levy 2012; Cuadrado et al. 2013; Guo and Han 2014; El Baidouri et al. 2017; Bao et al. 2018; Xu et al. 2018). To investigate possible changes in SA during evolution, we compared artificial and natural genotypes of several allopolyploid species. Different from those natural genotypes, artificial polyploids have hardly been exposed to evolutionary processes and their exact donors are known. With the believe that AE alters SA, we expected to find significantly enhanced levels of differences between the subgenomes in the natural genotypes. Surprisingly, little differences in either gene ratio or relative diversity among subgenomes were detected between these two types of allopolyploid genotypes. By examining di-haploid genotypes, we further showed that polyploidization does not seem to have dramatic influence on SA in allopolyploid genotypes. Rather, our results suggest that the differences observed between subgenomes of a

polyploid seem to simply reflect those between the diploid donors of a polyploid. The less variable subgenome of a polyploid genotype was derived from the less variable diploid donor and they tend to share a higher ratio of genes. These results seem to contradict with those previous studies which have shown unequivocally that extensive changes do occur during and following polyploidization in at least some species (Qi et al. 2004; Cheng et al. 2016; Edger et al. 2017). The contradiction can be explained by the likelihood that founding genomes are critical in defining changes during polyploid evolution. In other words, considering that significant changes in various aspects including gene number (Akhunov et al. 2003; Qi et al. 2004), diversity (Adams et al. 2005) and epigenetic regulation (Wendel et al. 2000; Guo and Han 2014; Song et al. 2015; Cheng et al. 2016; Ding and Chen 2018; Edger et al. 2018) have been detected in polyploid evolution, our results suggest that the levels and directions of the changes during and following polyploidization events could be profoundly influenced by the founding genomes.

It is important to note that assaying whole gene contents for any of the genotypes used in this study is not our intention and that they are not required for estimating either gene ratios or relative diversity between subgenomes. Clearly, only expressed genes can be detected in transcriptomic analysis and the expression of many genes can be tissue-specific (Coram et al. 2008; Mittapalli et al. 2010; Fagerberg et al. 2014). However, with the use of a single set of parameters, comparing subgenomes with the large quantity of expressed genes should likely provide high-quality estimates for both gene ratio and relative diversity. This likelihood is supported by the fact that the obtained results in this study are highly consistent in that significant difference was not detected between artificial and natural polyploid genotypes for any of the species investigated with the use of transcriptomic sequences from different tissues. Considering the large number of combinations of genotypes by tissues used, the highly consistent results cannot be a coincidence. Similarly, the genomic sequences used in this study

may not be adequate for high quality whole genome assembly for any of the genotypes from which they were generated. However, the results obtained from the genomic sequences again showed consistently that significant changes in neither gene ratio nor relative diversity among subgenomes were not detected between artificial and natural polyploid genotypes.

In addition to epigenetic regulations, previous studies in *Brassica* also showed that genomic exchanges including non-reciprocal homoeologous ones also affect SA (Chalhoub et al. 2014; Liu et al. 2014; An et al. 2019; Lu et al. 2019). However, similar to that found in any other species examined in this study, significant changes in SA was not detected between the artificial and natural genotypes and the differences observed between subgenomes seem to simply reflect those between the diploid donors. In fact, the triangle formed among the three diploids and their corresponding subgenomes in *Brassica* provides strong evidence showing the relationship between a subgenome in an allopolyploid and its diploid donors. The A subgenomes in either of the allotetraploids it was involved in (*B. juncea* with a genome of AABB and *B. napus* with a genome of AACC) was the one containing the lower gene ratios; the C subgenome is the one containing a higher gene ratios in both of the allotetraploids it was involved (*B. napus* with a genome of AACC and *B. carinata* with a genome of BBCC); and the B subgenome is the one containing a higher ratio in the combination of AB (*B. juncea*) but it is the one containing a lower ratio of genes in the combination of BC (*B. carinata*).

Different from many other species, the artificial allotetraploids of *Arabidopsis* (genome AATT) were not formed via typical polyploidization. They were generated by hybridizing two autotetraploid genotypes, one with the genome of AAAA and the other TTTT. In other word, no change in ploidy level was involved in the last step in obtaining the allotetraploid genotypes. However, similar to the allopolyploids of *A. suecica*, the gene ratio between the artificial tetraploids of *Arabidopsis* do not differ significantly from those of natural genotypes (table 3 and fig. 1). These results indicate that polyploidization was not be a prerequisite for the

differences between the two subgenomes. This likelihood is supported by that significant difference in gene ratio was not detected between the di-haploids and their allopolyploid derivatives.

Possible relationships between a subgenome and its donors are apparently not limited to gene diversity and gene ratios. For example, results from previous studies also showed that the significant difference in genome size between the subgenomes of polyploid wheats is also present between its progenitor species (IWGSC 2014; Šafář et al. 2010). Similarly, the huge difference in genome size between the two subgenomes (Zhang et al. 2015) also exists between the two progenitor species in cotton (Hendrix and Stewart 2005; Li et al. 2014b). These relationships provide further evidence suggesting that the differences between the subgenomes of the polyploid wheat are likely significantly influenced by its donors. As the exact parents for natural polyploids are unknown and VGC account for large proportions of various genomes (Jin et al. 2016; Liu et al. 2016; Montenegro et al. 2017; Ma et al. 2019), random genotypes from the donor species cannot be used to accurately estimate the relationship between a subgenome and its donor. Considering VGC or differences among individual genotypes in a species is critical in studying SA in polyploid evolution.

Supplementary Material

Supplemental Table 1. Genomic and transcriptomic sequences used in this study

Supplemental Table 2. Numbers of SNPs detected in each of the three diploid donor species of *Brassica*

Acknowledgments

We are grateful to Dr. Donald Gardiner for his constructive suggestions in revising the manuscript. This work was supported by the Commonwealth Scientific and Industrial

Organization (CSIRO), Australia (Project code: R-10191-01). X.Y., H.Z., Y.J. and Z.Y. are grateful to the Sichuan Agricultural University and the China Scholarship Council for funding his visit to CSIRO. H.H. wishes to acknowledge Henan Institute of Science and Technology and CSC for supporting her visit to CSIRO.

Author contributions

C.L., Y.-L.Z., Y.W., G.C., L.C. and Y.L. conceived the study; X.Y., H.Z., H.H., Y.J., Z.Y., S.G. and J.S. collected and analyzed data; X.Y., H.H., H.Z. and C.L. prepared the manuscript with contribution from other authors.

References

- Adams K L, Wendel J F. 2005. Polyploidy and genome evolution in plants. *Curr Opin Plant Biol.* 8(2): 135–141.
- Akhunov E D, et al. 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *PNAS USA.* 100(19): 10836–10841.
- An H, et al. 2019. Transcriptome and organellar sequencing highlights the complex origin and diversification of allotetraploid *Brassica napus*. *Nat Commun.* 10(1):2878.
- Andrews S. 2016. FastQC: a quality control tool for high throughput sequence data. (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>)
- Avni R, et al. 2017. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science.* 357(6346):93–97.
- Bairoch A, Apweiler R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* 27(1):49–54.
- Bao Y, Zhang X, Xu X. 2018. Abundant small genetic alterations after upland cotton domestication. *Biomed Res Int.* 2018.
- Bird KA, VanBuren R, Puzey JR, Edger PP. 2018. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol.* 220(1):87–93.

- Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*. 422:433–438.
- Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 12(1):59.
- Camacho C, et al. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. 10(1):421.
- Chalhoub B, et al. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*. 345(6199):950–953.
- Cheng F, et al. 2016. Epigenetic regulation of subgenome dominance following whole genome triplication in *Brassica rapa*. *New Phytol*. 211(1): 288–299.
- Cheng F, et al. 2018. Gene retention, fractionation and subgenome differences in polyploid plants. *Nat Plants*. 4(5): 258.
- Coram TE, Brown-Guedira G, Chen X. 2008. Using transcriptomics to understand the wheat genome. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*. 3(83):1–9.
- Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*. 11(1):485.
- Cuadrado A, Carmona A, Jouve N. 2013. Chromosomal characterization of the three subgenomes in the polyploids of *Hordeum murinum* L.: New insight into the evolution of this complex. *PLoS One*. 8(12): e81385.
- Del Pozo JC, Ramirez-Parra E. 2015. Whole genome duplications in plants: an overview from *Arabidopsis*. *J Exp Bot*. 66(22):6991–7003.
- Ding M, Chen ZJ. 2018. Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. *Curr Opin Plant Biol*. 42:37–48.
- Edger PP, et al. 2017. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell*. 29(9): 2150–2167.
- Edger PP, McKain MR, Bird KA, VanBuren R. 2018. Subgenome assignment in allopolyploids: challenges and future directions. *Curr Opin Plant Biol*. 42:76–80.
- El Baidouri M, et al. 2017. Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytol*. 213(3):1477–1486.
- Fagerberg L, et al. 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*. 13(2):397–406.

- Feldman M, Levy AA. 2012. Genome evolution due to allopolyploidization in wheat. *Genetics*. 192(3):763–774.
- Feldman M, et al. 1997. Rapid elimination of low-copy DNA sequences in polyploidwheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics*. 147(3):1381–1387.
- Flagel LE, Wendel JF. 2010. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytol*. 186(1):184–193.
- Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. 2007. Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *Plant Cell*. 19(11):3403–3417.
- Ghanbari MA, Jowkar A, Salehi H, Zarei M. 2019. Effects of polyploidization on petal characteristics and optical properties of *Impatiens walleriana* (Hook.). *Plant Cell Tiss Organ Cult*. 2019:1–2.
- Grabherr MG, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol*. 29(7):644.
- Grover CE, et al. 2012. Homoeolog expression bias and expression level dominance in allopolyploid. *New Phytol*. 196(4):966–971.
- Guo X, Han F. 2014. Asymmetric epigenetic modification and elimination of rDNA sequences by polyploidization in wheat. *Plant Cell*. 26(11):4311–4327.
- Hendrix B, Stewart JM. 2005. Estimation of the nuclear DNA content of *Gossypium* species. *Ann Bot*. 95(5):789–797.
- Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res*. 19(8):1419–1428.
- Jiao Y, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*. 473:97–100.
- Jin M, et al. 2016. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. *Sci Rep*. 6:18936.
- Kasianov AS, et al. 2017. High-quality genome assembly of *Capsella bursa-pastoris* reveals asymmetry of regulatory elements at early stages of polyploid genome evolution. *Plant J*. 91(2):278–291.
- Li A, et al. 2014a. mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. *Plant Cell*. 26:1878–1900.
- Li F, et al. 2014b. Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat Genet*. 46(6):567–572.

- Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 27(21):2987–2993.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25(14):1754–1760.
- Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22(13):1658–1659.
- Ling HQ, et al. 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature*. 496(7443):87.
- Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF. 2001. Polyploid formation in cotton is not accompanied by rapid genomic changes. *Genome*. 44(3):321–330.
- Liu M, et al. 2016. Chromosome-specific sequencing reveals an extensive dispensable genome component in wheat. *Sci Rep*. 6:36398.
- Liu S, Snowdon R, Chalhoub B, editors. 2018. The *Brassica Napus* genome. Springer. 2018 Oct 22.
- Liu S, et al. 2014. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun*. 5:3930.
- Lu K, et al. 2019. Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. *Nat Commun*. 10(1):1154.
- Luo R, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 1(1):18.
- Ma Y, Liu M, Stiller J, Liu C. 2019. A pan-transcriptome analysis shows that disease resistance genes have undergone more selection pressure during barley domestication. *BMC Genomics*. 20(1):12.
- Mittapalli O, et al. 2010. Tissue-specific transcriptomics of the exotic invasive insect pest emerald ash borer (*Agrilus planipennis*). *PLoS One*. 5(10):e13708.
- Montenegro JD, et al. 2017. The pangenome of hexaploid bread wheat. *Plant J*. 90(5):1007–1013.
- Page JT, et al. 2016. DNA sequence evolution and rare homoeologous conversion in tetraploid cotton. *PLoS Genet*. 12:e1006012.
- Paterson AH, Bowers JE, Van de Peer Y, Vandepoele K. 2005. Ancient duplication of cereal genomes. *New Phytol*. 165(3):658–661.
- Paterson AH, et al. 2012. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature*. 492(7429):423.

- Pont C, et al. 2013. Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *Plant J.* 76(6):1030–1044.
- Powell JJ, et al. 2017. The defence-associated transcriptome of hexaploid wheat displays homoeolog expression and induction bias. *Plant Biotechnol J.* 15(4):533–543.
- Qi LL, et al. 2004. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics.* 168(2):701–712.
- Šafář J, et al. 2010. Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome Res.* 129(1–3):211–223.
- Schnable J C, Springer N M, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *PNAS USA.* 108(10): 4069–4074.
- Soltis D E, Soltis P S. 1999. Polyploidy: recurrent formation and genome evolution. *Trends Ecol Evol.* 14(9): 348–352.
- Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev.* 35:119–125.
- Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Curr Opin Plant Biol.* 30:159–165.
- Song Q, Chen Z J. 2015. Epigenetic and developmental regulation in plant polyploids. *Curr Opin Plant Biol.* 24: 101–109.
- Song Q, Zhang T, Stelly DM, Chen ZJ. 2017. Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biol.* 18(1):99.
- Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33(suppl_2):W465–W467.
- The International Wheat Genome Sequencing Consortium (IWGSC). 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science.* 345(6194):1251788.
- Uauy C. 2017. Wheat genomics comes of age. *Current opinion in plant biology.* 36: 142-148.
- Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet.* 18(7):411.
- Wang J, et al. 2006. Genome wide nonadditive gene regulation in *Arabidopsis* allotetraploids. *Genetics.* 172:507–517.
- Wang M, et al. 2018. Evolutionary dynamics of 3D genome architecture following polyploidization in cotton. *Nat Plants.* 4(2): 90.

Wang M, et al. 2017. Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat Genet.* 49(4):579–587.

Wendel J F. 2000. Genome evolution in polyploids[M]//Plant molecular evolution. Springer, Dordrecht. 225–249.

Xu P, et al. 2018. The allotetraploid origin and asymmetrical genome evolution of common carp *Cyprinus carpio*. *BioRxiv.* 498774.

Zhang T, et al. 2015. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat Biotechnol.* 33(5):531–537.

Figure Legends

Figure 1. Gene ratios between the subgenomes of various polyploid species.

Although there is a time difference between their formations, consistent difference in gene ratio between the subgenomes do not exist between artificial and natural genotypes for any of the polyploid species. “?” indicates that the time is unknown for the species in concern, and MYs stands for million years.

Table 1. Shared genes between subgenomes and their respective progenitors and gene ratios between the subgenomes of different genotypes in different species

Species	Genotype	Species	Genome	Tissue	Shared genes with Xa		Shared genes with Ya		Gene ratio	
					No.	Percentage	No.	Percentage	X/(X+Y)	Y/(X+Y)
<i>B. napus</i>	Artificial F1 hybrid 1	F1 hybrid	AC	leaf	16031	54.1%	17814	59.0%	47.4%	52.6%
	Artificial F1 hybrid 2	F1 hybrid	AC	leaf	15012	59.0%	16586	61.3%	47.5%	52.5%
	Artificial F1 hybrid 3	F1 hybrid	AC	leaf	16428	60.9%	18746	62.1%	46.7%	53.3%
	Average of artificial F1				15824	58.0%	17715	60.8%	47.2%	52.8%
	Artificial allotetraploid 1	Artificial tetraploid	CCAA	leaf	16940	57.2%	18434	61.1%	47.9%	52.1%
	Average of artificial				16382	57.6%	18074.5	61.0%	47.5%	52.5%
	Natural allotetraploid 1	<i>B. napus</i>	AACC	leaf	15518	52.4%	17122	56.7%	47.5%	52.5%
	Natural allotetraploid 2	<i>B. napus</i>	AACC	leaf	15012	50.7%	17256	57.2%	46.5%	53.5%
	Natural allotetraploid 3	<i>B. napus</i>	AACC	leaf	15861	53.6%	17324	57.4%	47.8%	52.2%
	Natural allotetraploid 4	<i>B. napus</i>	AACC	leaf	15347	51.8%	17498	58.0%	46.7%	53.3%
<i>B. carinata</i>	Natural allotetraploid 5	<i>B. napus</i>	AACC	leaf	15646	52.8%	17331	57.4%	47.4%	52.6%
	Average of natural				15477	52.3%	17306	57.3%	47.2%	52.8%
	Artificial allotetraploid 1	Artificial tetraploid	BBCC	leaf	6070	49.8%	6347	43.7%	48.9%	51.1%
	Natural allotetraploid 1	<i>B. carinata</i>	BBCC	leaf	6009	49.3%	6365	43.8%	48.6%	51.4%
	Natural allotetraploid 2	<i>B. carinata</i>	BBCC	leaf	3297	27.0%	3455	23.8%	48.8%	51.2%
<i>B. juncea</i>	Average of natural				5619	38.2%	5919	33.8%	48.7%	51.3%
	Artificial allotetraploid 1	Artificial tetraploid	AABB	silique	3381	47.9%	4233	51.8%	44.4%	55.6%
	Natural allotetraploid 1	<i>B. juncea</i>	AABB	silique	2967	42.0%	3613	44.2%	45.1%	54.9%
	Natural allotetraploid 2	<i>B. juncea</i>	AABB	silique	2995	42.4%	3734	45.7%	44.5%	55.5%
	Average of natural				2981	42.2%	3674	45.0%	44.8%	55.2%
Wheat	Artificial tetraploid 1	Artificial tetraploid	SSAA	leaf	12115	68.8%	13832	64.8%	46.7%	53.3%
	Natural tetraploid 1	<i>T. dicoccoides</i>	BBAA	leaf	11888	67.6%	12294	57.6%	49.2%	50.8%
	Natural tetraploid 2	<i>T. durum</i>	BBAA	leaf	11168	63.5%	11771	55.2%	48.7%	51.3%
	Natural tetraploid 3	<i>T. dicoccoides</i>	BBAA	leaf	11036	62.7%	11837	55.5%	48.3%	51.8%
	Natural tetraploid 4	<i>T. dicoccoides</i>	BBAA	leaf	10404	59.1%	10927	51.2%	48.7%	51.3%
	Average of natural				11124	63.2%	11707	54.9%	48.7%	51.3%
	Artificial tetraploid 2	Artificial tetraploid	SSAA	spike	20093	57.5%	22034	43.4%	47.7%	52.3%
	Natural tetraploid 5	<i>T. dicoccoides</i>	BBAA	spike	18982	54.3%	21243	41.8%	47.2%	52.8%
	Natural tetraploid 6	<i>T. durum</i>	BBAA	spike	18268	52.3%	18502	36.4%	49.7%	50.3%
	Average of natural				18625	53.3%	19873	39.1%	48.5%	51.6%
Cotton*	Artificial tetraploid	Artificial tetraploid	AADD	leaf	23274	63.3%	24170	61.5%	49.1%	50.9%
	Natural tetraploid 1	<i>G. hirsutum</i>	AADD	leaf	22514	61.2%	22940	58.3%	49.5%	50.5%
	Natural tetraploid 2	<i>G. barbadense</i>	AADD	leaf	21215	57.7%	21811	55.5%	49.3%	50.7%
	Natural tetraploid 3	<i>G. hirsutum</i>	AADD	leaf	20434	55.5%	21084	53.6%	49.2%	50.8%
	Natural tetraploid 4	<i>G. barbadense</i>	AADD	leaf	21937	59.6%	22702	57.7%	49.1%	50.9%
	Average of natural				21525	58.5%	22134	56.3%	49.3%	50.7%

^a, X and Y represent A and B genome or subgenome for *B. juncea*; B and C for *B. carinata*; A and C for *B. napus*; A and B for wheat, and A and D for cotton.

*, Genomic sequences were used for cotton and transcriptomic sequences were used for all other species.

Table 2. Lack of differences in gene ratio between the subgenomes of an artificial F1 hybrid and the hexaploid wheat genotype derived from it*

Genotype (genome)	Tissue	AABB		DD		AB(D+AB)	D/(D+AB)
		No. CDS	percentage	No. CDS	percentage		
Artificial F1 hybrid (BAD)	leaf	25496	48.4%	18467	49.5%	58.0%	42.0%
Artificial hexaploid (BBAADD)	leaf	25494	48.4%	18389	49.3%	58.1%	41.9%

*The triploid F1 hybrid (with a triploid genome of ABD) was obtained by crossing an allotetraploid genotype (producing AB gametes) with a diploid genotype (producing D gametes). The artificial hexaploid (genome AABBDD) was derived from doubling the chromosomes of the triploid F1 hybrid.

Table 3. Shared genes between subgenomes and their respective progenitors and gene ratios between the subgenomes in natural and artificial *A.suecica* based on transcriptomic sequences

Species	Genotype	Genome	Tissue	Shared gene with AA		Shared gene with TT		Gene ratios	
				No.	Percentage	No.	Percentage	AA/(AA+TT)	TT/(AA+TT)
<i>A. suecica</i>	Artificial allotetraploid 1	AATT	leaf	12713	58.1%	15227	65.8%	45.5%	54.5%
<i>A. suecica</i>	Natural allotetraploid 1	TTAA	leaf	12071	55.1%	13464	58.2%	47.3%	52.7%
<i>A. suecica</i>	Natural allotetraploid 2	TTAA	leaf	11241	51.3%	13375	58.2%	45.7%	54.3%
<i>A. suecica</i>	Natural allotetraploid 3	TTAA	leaf	12004	54.8%	13399	57.9%	47.3%	52.7%
<i>A. suecica</i>	Natural allotetraploid 4	TTAA	leaf	12056	55.1%	13198	57.1%	47.7%	52.3%
<i>A. suecica</i>	Natural allotetraploid 5	TTAA	leaf	11389	52.0%	13377	57.8%	46.0%	54.0%
<i>A. suecica</i>	Natural allotetraploid 6	TTAA	leaf	10238	46.8%	13129	56.8%	43.8%	56.2%
<i>A. suecica</i>	Natural allotetraploid 7	TTAA	leaf	12110	55.3%	13092	56.6%	48.1%	51.9%
<i>A. suecica</i>	Natural allotetraploid 8	TTAA	leaf	12052	55.0%	13210	57.1%	47.7%	52.3%
<i>A. suecica</i>	Natural allotetraploid 9	TTAA	leaf	10873	49.7%	12657	54.7%	46.2%	53.8%
Average of natural				11559	52.8%	13211	57.2%	46.7%	53.3%

Figure 1