The Contributions of the Allopolyploid Parents of the Mesopolyploid Brassiceae are Evolutionarily Distinct but Functionally Compatible

Yue Hao[1], Makenzie E. Mabry[2], Patrick P. Edger[3,4], Michael Freeling[5], Chunfang Zheng[6], Lingling Jin[7], Robert VanBuren[3,8], Marivi Colle[3], Hong An[2], R. Shawn Abrahams[2], Xinshuai Qi[9], Kerrie Barry[10], Christopher Daum[10], Shengqiang Shu[10], Jeremy Schmutz[10,11], David Sankoff[6], Michael S. Barker[9], Eric Lyons[12], J. Chris Pires[2,13] and Gavin C. Conant[1,14,15,16]

1. Bioinformatics Research Center, North Carolina State University, Raleigh, NC
2. Division of Biological Sciences, University of Missouri - Columbia, Columbia, MO
3. Department of Horticulture, Michigan State University, East Lansing, MI
4. Ecology, Evolutionary Biology and Behavior, Michigan State University, East Lansing, MI
5. Department of Plant and Microbial Biology, University of California, Berkeley, CA
6. Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada
7. Department of Computing Science, Thompson Rivers University, Kamloops, BC, Canada
8. Plant Resilience Institute, Michigan State University, East Lansing, MI
9. Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ
10. Department of Energy Joint Genome Institute, Walnut Creek, CA
11. HudsonAlpha Institute for Biotechnology, Huntsville, AL
12. School of Plant Sciences, University of Arizona, Tucson, AZ
13. Informatics Institute, University of Missouri – Columbia, Columbia, MO
14. Program in Genetics, North Carolina State University, Raleigh, NC
15. Department of Biological Sciences, North Carolina State University, Raleigh, NC
16. Division of Animal Sciences, University of Missouri – Columbia, Columbia, MO

Correspondence: gconant@ncsu.edu

Running title: Subgenome differentiation after Brassiceae hexaploidy

Keywords: *Brassica*, polyploidy, biased fractionation, subgenomes, *Crambe hispanica*

Abstract

The members of the tribe Brassiceae share an ancient whole genome triplication (WGT), and plants in this tribe display extraordinarily high within-species morphological diversity. One proposed model for the formation of these hexaploid Brassiceae is that they result from a "two-step" pair of hybridizations. However, direct evidence supporting this model of formation has been lacking; meanwhile, the evolutionary and functional constraints that drove evolution after the hexaploidy are even less understood. Here we report a new genome sequence of *Crambe hispanica*, a species sister to most sequenced Brassiceae. After adding this new genome to three others that are also descended from the ancient hexaploidy, we traced the history of gene loss after the WGT using a phylogenomic pipeline called POInT (the Polyploidy Orthology Inference Tool). This approach allowed us to confirm the two-step model of hexaploidy formation and to assign statistical confidence to our parental "subgenome" assignments for >90,000 individual genes. We show that each subgenome has a statistically distinguishable rate of homeolog losses. Moreover, our modeling allowed us to infer that there was a significant temporal gap between the two allopolyploidizations, with about one third of the total shared gene losses between the four analyzed Brassiceae species in the first two subgenomes prior to the arrival of the third subgenome. There is little indication of functional distinction between the three subgenomes: the individual subgenomes show no patterns of functional enrichment, no excess of shared protein-protein or metabolic interactions between their members, and no biases in their likelihood of having experienced a recent selective sweep. We propose a "mix and match" model of allopolyploidy, where subgenome origin drives homeolog loss propensities but where genes from different subgenomes function together without difficulty.

Introduction

Fifty years ago, Ohno (Ohno 1970) published his famous, and forceful, opus on the role of gene duplication, and in particular of *genome* duplication (aka polyploidy), in evolutionary innovation. Since then, evidence both of polyploidy's ubiquity (Wolfe and Shields 1997; Soltis and Soltis 2012; Van de Peer et al. 2009, 2017) and of its role in evolutionary innovations such as yeast aerobic glucose fermentation, the organization of the retinae of teleost fishes and in plant defensive compounds has continued to accumulate (Conant and Wolfe 2007; Merico et al. 2007; van Hoek and Hogeweg 2009; Edger et al. 2015; Sukeena et al. 2016). Preeminent among the polyploid lineages are the flowering plants, where over 180 ancient polyploidies are known (One Thousand Plant Transcriptomes Initiative 2019).

The tribe Brassiceae experienced hexaploidy between 5 and 9 MYA and after its divergence from *Arabidopsis thaliana* (Wang et al. 2011). This polyploidy was originally inferred by comparative linkage mapping studies between *Brassica* species and *A. thaliana* (Lagercrantz 1998; Lukens et al. 2004; Parkin et al. 2005; Schranz et al. 2006) and confirmed by chromosome painting (Lysak et al. 2005; Lysak 2009). The Brassiceae are also the most morphologically diverse tribe in the family *Brassica*ceae, a condition that is believed to be partly due to this hexaploidy (Cheng et al. 2014). The Brassiceae also contain many important vegetable and oilseed crops, such as broccoli, cabbage, kale, mustard and canola. Among these crops, species in the genus *Crambe* are not only an important industrial oilseed source because of their high erucic acid content (Lazzeri et al. 1997; Warwick and Gugel 2003; Carlsson et al. 2007) but also could serve as resources for *Brassica* crop breeding and development (Rudloff and Wang 2011).

Among allopolyploidies (formed by the merging of similar but not identical genomes), the preferential retention of gene copies from one of the parental subgenomes has been observed in yeast, maize, cotton, monkeyflower, *Arabidopsis*, *Brassica*, and nematodes (Thomas et al. 2006; Conant and Wolfe 2008a; Cheng et al. 2012; Parkin et al. 2014; Renny-Byfield et al. 2015; Edger et al. 2017; Emery et al. 2018; Schoonmaker et al. 2020). Allopolyploids also show a tendency for genes from one of the subgenomes to be more highly expressed: genes from the remaining subgenomes have a correspondingly

greater chance of being silenced or even lost completely, suggesting one mechanism that might drive preferential retention. This phenomenon has been called “subgenome dominance”, and the resulting pattern of gene retention is known as “biased fractionation” (Thomas et al. 2006; Schnable et al. 2011; Yoo et al. 2014). A number of hypothesis have been proposed to explain these differences in homeolog expression, including differences in transposon silencing across subgenomes (Freeling et al. 2012; Woodhouse et al. 2014; Zhao et al. 2017; Alger and Edger 2020), disruption of the organelle-nucleus communication in the more fractionated subgenomes (Sharbrough et al. 2017; Costello et al. 2019), and epigenetic changes induced by the genomic shock of merging genetically distinct subgenomes (McClintock 1984; Wendel et al. 2018; Bird et al. 2018). In addition, we and others have speculated that allopolyploids might bring together coevolved and conflicting copies of multi-protein complexes (Codoñer and Fares 2008; Gong et al. 2012; Scienski et al. 2015; Emery et al. 2018): early random gene losses from one subgenome that partly resolved these conflicts might then set the polyploidy down a path favoring losses from that subgenome. A related proposal was made by Makino and McLysaght (2012), namely that selection to maintain dosage balance among interacting gene neighbors could produce local and eventually global biases in fractionation.

Not all homeologs are equally likely to revert to single-copy after a polyploidy, regardless of the level of biased fractionation. Duplicated genes coding for transcription factors, ribosomal proteins and kinases are over-retained after independent polyploidies in flowering plants, yeasts, ciliates and vertebrates (Seoighe and Wolfe 1998; Blanc and Wolfe 2004; Maere et al. 2005; Aury et al. 2006; Makino and McLysaght 2010). These patterns are best explained by a need to maintain dosage balance among highly interacting genes (Birchler et al. 2005; Hakes et al. 2007; Birchler and Veitia 2012, 2014; Conant et al. 2014). Curiously, there are also genes that prefer *not* to be duplicated: genes for DNA repair and those targeted to organelles have returned to single-copy rapidly after genome duplication (De Smet et al. 2013; Conant 2014).

The *Brassica* WGT is a particularly interesting system to study all of these phenomena because the nature of the triplication allows us to explore each in unusual detail. The patterns of biased

fractionation observed in the genome of *Brassica rapa* suggested that the triplication "event" was actually two separate allopolyploid hybridizations involving three distinct diploid progenitor species, with the merger of the two currently highly fractionated ancestral genomes occurring first, followed by the subsequent addition of a third subgenome, which currently possesses the most retained genes (Tang et al. 2012; Cheng et al. 2012).

Although this scheme of subgenome additions and fractionation is very reasonable, there are alternatives. Moreover, the very distinctness of these three subgenomes allows for a more complete analysis of the consequences of these types of genomic mergers and of the proposed mechanisms underlying biased fractionation. With the benefit of a new genome sequence from *Crambe hispanica*, presented here, we analyzed the *Brassica* WGT with our tool for modeling post-polyploidy genome evolution: POInT (the Polyploidy Orthology Inference Tool) (Conant and Wolfe 2008a). POInT, which we recently extended to allow the analysis of WGTs (Schoonmaker et al. 2020), probabilistically infers the orthologous regions surviving from ancient duplications/triplications; it can be used to test for biases in fractionation without *ad hoc* assumptions. Moreover, POInT's models operate on complete genomes along a phylogeny (Dunn et al. 2018; Smith et al. 2020), helping to mask rearrangements in single genomes and giving larger synteny blocks (Emery et al. 2018).

We find strong support for the two-step WGT formation model, with evidence for a significant gap in time between the two events. Strikingly, we find no evidence that co-evolved functional modules have driven gene losses: while the three subgenomes are clearly distinguishable in their loss patterns, there are no indications that members of the same subgenome share more functional associations than do genes from differing subgenomes.

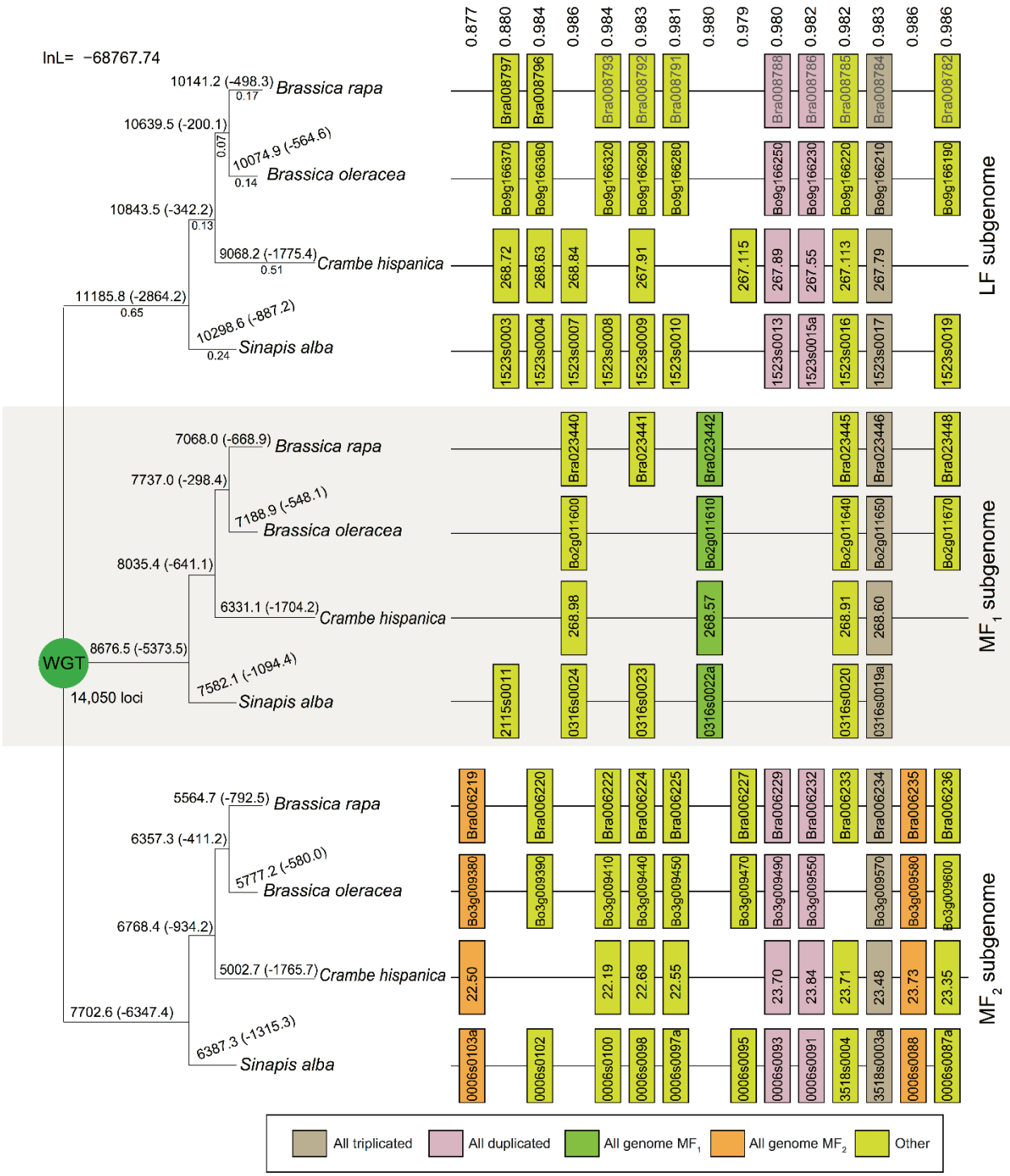
Results

A well-assembled, annotated genome of Crambe hispanica

Our assembly of the *Crambe hispanica* genome contains 1,019 contigs with 11 terminal telomeres and has a total size of 480MB; its contig N50 size is 4.4 Mb. The assembly graph showed low heterozygosity and few assembly artefacts, with the exception of one mega-cluster consisting of a high copy number LTR across 500 contigs and spanning ~30 Mb. The resulting annotated genome is of high quality: we compared its gene set against the Benchmarking Universal Single-Copy Orthologs (BUSCO v.2) (Simão et al. 2015) plant dataset (embryophyta_odb9), finding that 95.8% of these expected genes were present in our annotation.

Inferring blocks of triple-conserved synteny in four triplicated Brassiceae genomes and estimating an ancestral gene order

Using the genome of *Arabidopsis thaliana* as an unduplicated reference, we inferred blocks of triple conserved synteny (TCS) for each of four genomes sharing the *Brassica* hexaploidy. We then merged these blocks across all of the species: we refer to each such locus as a "pillar." Each pillar consists of between 1 and 3 surviving genes in each of the four genomes. As described in the *Methods*, we used both a set of TCS blocks inferred with POInT containing 14,050 pillars ($P_{pillars}$) and a separate ancestral genome reconstruction that estimates the gene order that existed just prior to the WGT. The latter contains five reconstructed ancestral chromosomes involving 89 scaffolds with a total of 10,868 ancestral genes. When we match these genes to the TCS blocks computed with POInT, the result is 7,993 ancestrally-ordered pillars (A_{pillars}).


lnL= −68767.74
WGT
14,050 loci
11185.8 (-2864.2)
0.65
10141.2 (-498.3)
0.17
Brassica rapa
10639.5 (-200.1)
0.07
10074.9 (-564.6)
0.14
Brassica oleracea
10843.5 (-342.2)
0.13
9068.2 (-1775.4)
0.51
Crambe hispanica
10298.6 (-887.2)
0.24
Sinapis alba
8676.5 (-5373.5)
7068.0 (-668.9)
Brassica rapa
7737.0 (-298.4)
7188.9 (-548.1)
Brassica oleracea
8035.4 (-641.1)
6331.1 (-1704.2)
Crambe hispanica
7582.1 (-1094.4)
Sinapis alba
7702.6 (-6347.4)
5564.7 (-792.5)
Brassica rapa
6357.3 (-411.2)
5777.2 (-580.0)
Brassica oleracea
6768.4 (-934.2)
5002.7 (-1765.7)
Crambe hispanica
6387.3 (-1315.3)
Sinapis alba
0.877 0.880 0.984 0.986 0.984 0.983 0.981 0.980 0.979 0.980 0.982 0.982 0.983 0.986 0.986
Bra008797 Bra008796 Bra008793 Bra008792 Bra008791 Bra008788 Bra008786 Bra008785 Bra008784 Bra008782
Bo9g166370 Bo9g166360 Bo9g166320 Bo9g166290 Bo9g166280 Bo9g166250 Bo9g166230 Bo9g166220 Bo9g166210 Bo9g166190
268.72 268.63 268.84 267.91 267.115 267.89 267.55 267.113 267.79
1523s0003 1523s0004 1523s0007 1523s0008 1523s0009 1523s0010 1523s0013 1523s0015a 1523s0016 1523s0017 1523s0019
LF subgenome
Bra023440 Bra023441 Bra023442 Bra023445 Bra023446 Bra023448
Bo2g011600 Bo2g011610 Bo2g011640 Bo2g011650 Bo2g011670
268.98 268.57 268.91 268.60
2115s0011 0316s0024 0316s0023 0316s0022a 0316s0020 0316s0019a
MF1 subgenome
Bra006219 Bra006220 Bra006222 Bra006224 Bra006225 Bra006227 Bra006229 Bra006232 Bra006233 Bra006234 Bra006235 Bra006236
Bo3g009380 Bo3g009390 Bo3g009410 Bo3g009440 Bo3g009450 Bo3g009470 Bo3g009490 Bo3g009550 Bo3g009570 Bo3g009580 Bo3g009600
22.50 22.19 22.68 22.55 23.70 23.84 23.71 23.48 23.73 23.35
0006s0103a 0006s0102 0006s0100 0006s0098 0006s0097a 0006s0095 0006s0093 0006s0091 3518s0004 3518s0003a 0006s0088 0006s0087a
MF2 subgenome
All triplicated
All duplicated
All genome MF1
All genome MF2
Other

1

Figure 1. Subgenome assignment and inference of gene loss after the shared WGT in four species. After the WGT, each ancestral locus could potentially expand to three gene copies, but because of the biased gene loss, the number of genes left in each subgenome is not even. One subgenome is less fractionated (LF), while the other two are more fractionated (MF1 and MF2). Shown here is a window of 15 post-WGT loci (out of the total 14,050 loci) in four species that shared the WGT, *Brassica rapa*, *Brassica oleracea*, *Crambe hispanica* and *Sinapis alba*. Each pillar corresponds to an ancestral locus, and the boxes represent genes connected by synteny (straight line). The numbers on top of each pillar are the posterior probabilities assigned to this combination of orthology relationships relative to the other $(3!)^4-1=1295$ possible orthology states. The numbers above each branch given the number of genes in each subgenome surviving to that point, with the number of gene losses in parentheses. The numbers below the branches in the first subtree are POInT's branch length estimates (αt).

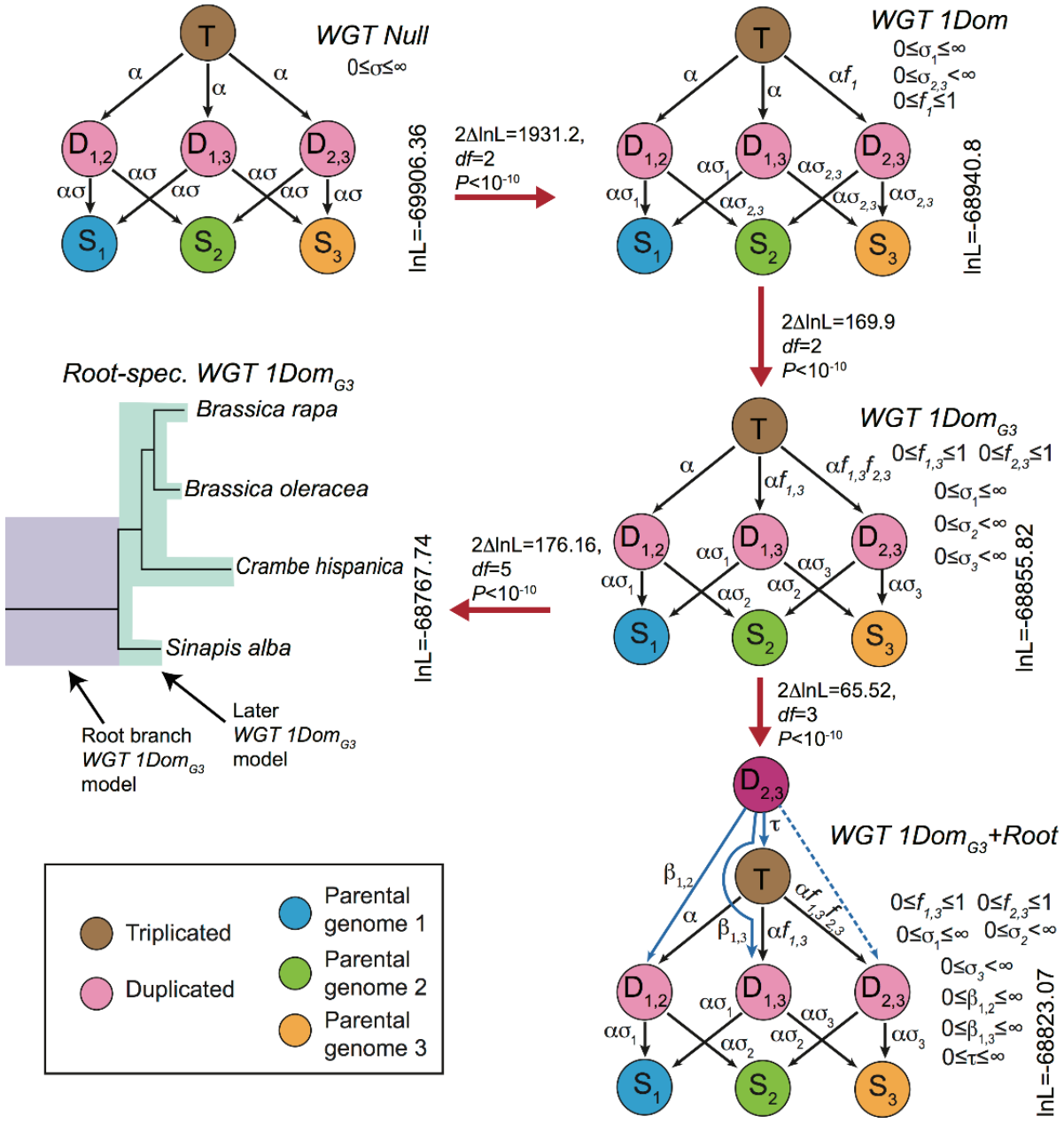
Inferring the relationship of the four Brassiceae genomes based on gene loss patterns

We fit models of WGT evolution (see below) to both several different orderings of the 14,050 pillars in the $P_{pillars}$ set and to the $A_{pillars}$ (Supplementary Table S1). These orderings of the $P_{pillars}$ differed in their number synteny breaks: we used the ordering with the highest likelihood under the WGT 3rate G1Dom model for our remaining analyses (see below). Similarly, we compared the fit of three possible phylogenetic topologies to the pillars under this model: the remainder of our analyses use the topology shown in Figure 1, which has the highest likelihood. Curiously, one of the other two topologies, while having a lower likelihood under POInT's models (Supplementary Figure S1), is the phylogeny estimated using the plastid genome (Arias and Pires 2012). Because the $A_{pillars}$ give similar parameter estimates but comprise a smaller dataset, we will discuss our results in terms of the $P_{pillars}$.

The three subgenomes differ in their propensity for ohnolog copy loss

We used POInT to assign genes from the four triplicated genomes to three subgenomes with high resolution: 75% of the pillars have subgenome assignments with posterior probabilities > 0.84 (Supplementary Figure S3). We observe clear signals of biased fractionation: while we estimate that 2,864 genes were lost from the LF subgenome along the root branch, the corresponding figures for MF1 and MF2 are 5,373 and 6,347 respectively (Figure 1). These values are in qualitative agreement with previous findings (Xie et al. 2019; Liu et al. 2014; Cheng et al. 2014, 2012).

We assessed the statistical support for these estimated differences in the subgenomes' rates of ohnolog loss using a set of nested models of post-WGT gene loss. We started with a model (WGT Null) that did not differentiate between the subgenomes, meaning that the shared base transition rate from **T** to $\mathbf{D_{1,2}}$, $\mathbf{D_{1,3}}$ or $\mathbf{D_{2,3}}$ was defined to be α ($0 \leq \alpha < \infty$, Figure 2). The transition rate from $\mathbf{D_{1,2}}$, $\mathbf{D_{1,3}}$ or $\mathbf{D_{2,3}}$ to $\mathbf{S_1}$, $\mathbf{S_2}$ or $\mathbf{S_3}$ is scaled by σ: e.g., occurs at rate $\alpha \bullet \sigma$. We compared this model to a more complex one that allowed losses of both triplicated and duplicated genes to be less frequent from a posited *less-fractionated* subgenome LF (WGT 1Dom, Figure 2). This model introduces a fractionation parameter f_1 ($0 \leq f_1 \leq 1$), which potentially makes the transitions between T and $D_{2,3}$ rarer than the other T-to-D rates ($\alpha \bullet f_1$; see Figure 2). The WGT 1Dom model fits the pillar data significantly better than does WGT Null (Figure 2; $P<10^{-10}$, likelihood ratio test with two degrees of freedom). We next compared the WGT 1Dom model to a WGT $1Dom_{G3}$ model that gives MF1 and MF2 separate loss rates. Again, this model gives a better fit to the pillar data than did WGT 1Dom ($P<10^{-10}$, likelihood ratio test with two degrees of freedom, Figure 2). We hence confirm the presence of three subgenomes, distinguishable by their patterns of homeolog loss. It is important to recall here that our approach does not require the identification of these three subgenomes *a priori*: the probabilistic assignment of genes to subgenomes is an integral part of the POInT orthology computation: as a result, the inherent uncertainty in these assignments is accounted for in estimating the various biased fractionation parameters.


WGT Null
0≤σ≤∞
lnL=-69906.36
2ΔlnL=1931.2,
df=2
P<10-10
WGT 1Dom
0≤σ1≤∞
0≤σ2,3<∞
0≤f1≤1
lnL=-68940.8
2ΔlnL=169.9
df=2
P<10-10
WGT 1DomG3
0≤f1,3≤1 0≤f2,3≤1
0≤σ1≤∞
0≤σ2<∞
0≤σ3<∞
lnL=-68855.82
2ΔlnL=176.16,
df=5
P<10-10
lnL=-68767.74
2ΔlnL=65.52,
df=3
P<10-10
WGT 1DomG3+Root
0≤f1,3≤1 0≤f2,3≤1
0≤σ1≤∞ 0≤σ2<∞
0≤σ3<∞
0≤β1,2≤∞
0≤β1,3≤∞
0≤τ≤∞
lnL=-68823.07
Root-spec. WGT 1DomG3
Brassica rapa
Brassica oleracea
Crambe hispanica
Sinapis alba
Root branch
WGT 1DomG3
model
Later
WGT 1DomG3
model
Triplicated
Duplicated
Parental genome 1
Parental genome 2
Parental genome 3

1

Figure 2. POInT's models for inferring WGT. Five different models of post-WGT evolution and their ln-likelihoods are shown. In each model, the colored circles represent different states. the brown circle represents triplicated state (**T**); the pink circles are duplicated states (**$D_{1,2}$**, **$D_{1,3}$** and **$D_{2,3}$**); the blue, green and yellow circles are three single-copy states (**S_1** for LF subgenome, **S_2** for MF1 subgenome and **S_3** for MF2 subgenome). The transition rates between states are shown above the arrows. α: transition rate from triplicated state to duplicated states; ασ: transition rates from duplicated states to single copy states; f: fractionation parameters; β and τ: root model parameters. Red arrows point to the models tested using likelihood ratio tests. See *Methods*. *WGT Null model*: transition rates are the same across three subgenomes, modeling the scenario of no biased fractionation. *WGT 1Dom model*: with the biased fractionation parameter f_1 ($0 \leq f_1 \leq 1$), the MF1 and MF2 subgenomes are more fractionated than LF subgenome. *WGT 1Dom$_{G3}$ model*: two fractionation parameters $f_{1,3}$ and $f_{2,3}$ were introduced, distinguishing three subgenomes: MF2 is more fractionated than MF1, and MF1 is more fractionated than LF. *Root-spec. WGT 1Dom$_{G3}$ model:* similar to the previous model, but with two sets of parameters, one set for the root branch and the other for the rest of the branches. *WGT 1Dom$_{G3}$ + Root model*: an intermediate state $D_{2,3}$ was added, representing the merging of MF1 and MF2 subgenomes, as the "first step" of the hexaploidy formation. The other different states in the model represent the merging of LF subgenome, and gene loss process before or after the joining of LF.

Patterns of post-WGT gene loss support the "two-step" model of hexaploidy

To test the hypothesis that the WGT proceeded in two steps (Cheng et al. 2012; Tang et al. 2012), we used two approaches. First, we applied an extended version of the WGT 1Dom$_{G3}$ model where each model parameter was allowed to take on distinct values on the root branch and on the remaining branches (Root-spec. WGT 1Dom$_{G3}$ in Figure 2). This extended model fits the pillar data significantly better than does the original WGT 1Dom$_{G3}$ model ($P<10^{-10}$, likelihood ratio test with five degrees of freedom, Figure 2). The biased fractionation parameters for the root branch differ from those of the remaining branches:

the value of $f_{1,3}$ on the root is smaller than on later branches (0.6445 versus 0.7368) while $f_{2,3}$ is larger (0.6766 versus 0.4078). These values are consistent with a two-step hypothesis: prior to the arrival of LF, there would have been a number of losses from MF1 and MF2, meaning that the relative preference for LF would be higher (smaller $f_{1,3}$).

In our second approach, we developed a specific model of the two-step hexaploidy (WGT $1Dom_{G3}$+$Root_{LF}$ in Figure 2). This model describes the transition from a genome *duplication* to a triplication: all pillars start in state $\mathbf{D_{2,3}}$: e.g., the MF1 and MF2 genes are present but not the LF ones. We then model the addition of LF as transitions to either the **T**, $\mathbf{D_{1,2}}$ or the $\mathbf{D_{1,3}}$ states (with rates τ, $\beta_{1,2}$ or $\beta_{1,3}$, respectively). State **T** is seen when no losses occurred prior to the arrival of LF, the other states occur when either MF1 or MF2 experienced a loss prior to the arrival of LF. Any pillars that remain in $\mathbf{D_{2,3}}$ had no corresponding gene arrive from LF. Of course, at the level of the individual pillar, we have insufficient data to make such inferences: the utility of this model is to give global estimates of the degree of fractionation seen in MF1 and MF2 prior to the arrival of LF. This model offers a significantly improved fit over WGT $1Dom_{G3}$ ($P<10^{-10}$, likelihood ratio test with three degrees of freedom, Figure 2). More importantly, we can propose other versions of this model where either MF1 or MF2 is the last arriving subgenome: when we do so, the model fit is much worse than seen with WGT $1Dom_{G3}$+$Root_{LF}$ model (Supplementary Table S1). Hence, we can conclude that subgenomes MF1 and MF2 had already begun a process of (biased) fractionation prior to the addition of the LF subgenome.

A gap between the two allopolyploidies

This root-specific model also allows us to estimate the state of MF1 and MF2 immediately before the arrival of LF. In particular, we can estimate the percentage of pillars that had already experienced losses prior to LF's arrival. About 28% of all of the MF1 homeologs inferred to have been lost on the root branch were lost prior to the arrival of LF, with the equivalent number of MF2 losses being 38%. A negligible 0.3% of pillars do not appear to have received a copy of the LF homeolog.

LF copy in retained triplets are under strong selective constraints

In our dataset there 218 loci that have retained triplicates in all four genomes with confidence ≥ 95%. For each loci, we calculated the selective constraints the group of 12 genes using codeml (Yang 2007), allowing the genes from each subgenome to have a different dN/dS value. On average, among the retained triplets, genes from the LF subgenome show slightly greater selective constraint (smaller dN/dS; Supplementary Figure S4).

Single copy genes from multiple subgenomes are enriched in genes functioning in DNA repair

GO overrepresentation tests were performed with the *Arabidopsis* orthologs of genes returned to single copy by the end of the root branch from each subgenome. Similar to previous findings (De Smet et al. 2013), we found that single copy genes are enriched in biological processes such as DNA repair and DNA metabolism (Supplementary Figure S5). More specifically, single copy genes from the LF subgenome are enriched in base-excision repair, while MF1 single copy genes are enriched in nucleotide-excision repair, non-recombinational repair and double-strand break repair (Supplementary Figure S5a). Intriguingly, single copy genes from both LF and MF1 show overrepresented molecular functions in endo and exodeoxyribonuclease activities (Supplementary Figure S5b). LF single copy genes are also enriched in RNA interference processes, suggesting that such interference, targeted to the MF1 and MF2 subgenomes, could be one mechanism by which biased fractionation was driven.

Genes from the same subgenome are not overly likely to physically or metabolically interact

For genes with high subgenome assignment confidence (≥ 95%), we mapped those assignments (LF, MF1 or MF2) and the duplication status at the end of the root branch onto the nodes (gene products) of the *A. thaliana* protein-protein interaction (PPI) network (*Methods)*. For comparative purposes, we also produced a mapping of an extant network, based on the gene presence/absence data and subgenome assignments in *B. rapa*. Not surprisingly, in the "ancient" network inferred at the end of the common root

branch, there are a relatively large number of nodes (1,952) associated with surviving triplicated loci: these nodes were connected by a total of 2,384 triplet-to-triplet edges. The *B. rapa*-specific network contains fewer nodes with retained triplets (662): there were 263 edges connecting these nodes (Figure 3a).

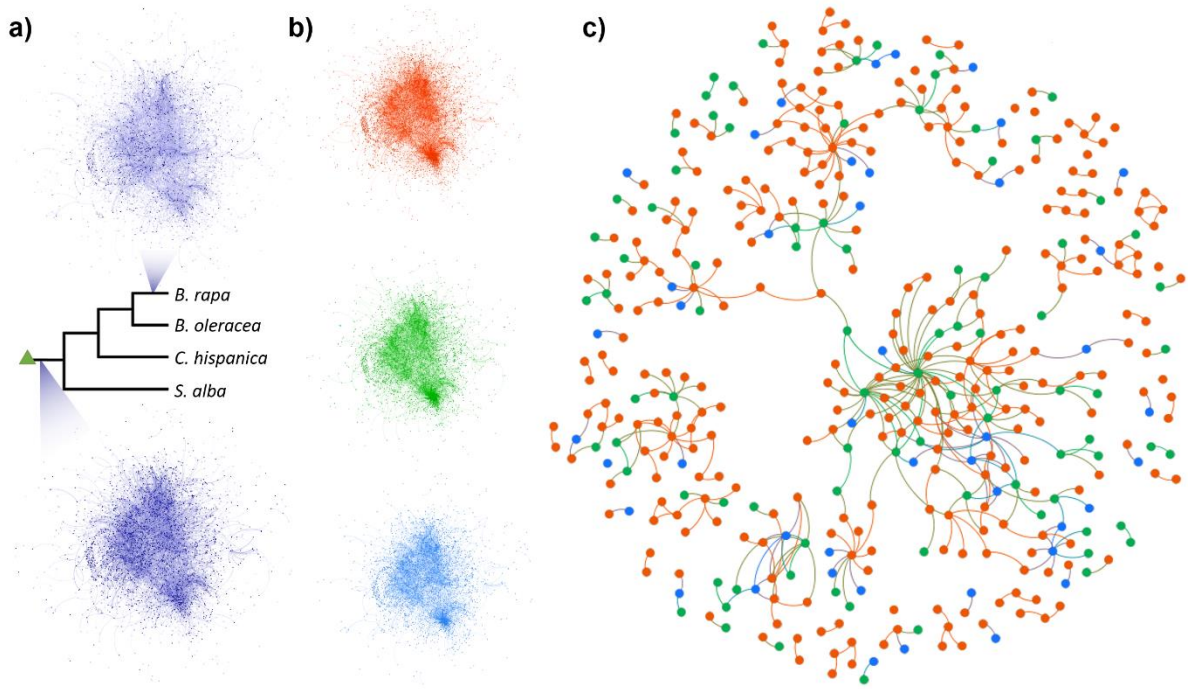


Figure 3. Protein-protein interaction networks after the WGT. a) The *Arabidopsis* PPI network at the root branch (bottom) and the same PPI network colored by the *Brassica* rapa gene retention status (top). The dark purple colored nodes represent retained triplets. b) PPI network partitioned by subgenome assignment at the root branch. LF subgenome: red - 4,249 nodes and 8,454 edges. MF1 subgenome: green - 3,379 nodes and 6,442 edges. MF2 subgenome: blue - 3,073 nodes and 4,961 edges. c) A subset of the PPI network where only nodes encoded by single copies genes and connected with other single copy nodes are shown. Red nodes are from the LF subgenome, green nodes are from the MF1 subgenome and blue nodes are from the MF2 subgenome.

The dosage constraints that affect surviving gene copies post-polyploidy will tend to result in the retention of genes involved in multiunit complexes or in the same signaling pathways (Birchler and Veitia 2007, 2012; Conant et al. 2014). Thus, we expected to see that the retained triplets showed higher network connectivity. And indeed, our permutation tests reveal that the retained triplets on the root branch are significantly over-connected to each other in the PPI network (P = 0.018, Supplementary Figure S6). We also hypothesized that proteins coded for from the same subgenome would be more likely to be connected due to preferential retention of genes from a single complex from the same subgenome. To test this idea, we partitioned the gene products based on their subgenome of origin. The LF subgenome contains more genes and thus more exclusive connections (Figure 3b). When considering only genes that had returned to single-copy by the end of the root, we identified 188 LF-LF edges among 886 single copy LF genes, with fewer edges exclusive to MF1 and MF2 genes (30 and 3, respectively). We used randomization methods to test whether the numbers of such subgenome-specific edges differed from what would be expected by chance. When considering the network as a whole, we found that there were significantly *fewer* LF-LF edges than expected (P = 0.022; Supplementary Figure S6). However, when we considered only the single-copy genes in the network, the number of subgenome-specific edges did not differ from that seen in random networks for any of the three subgenomes (P = 0.286 for LF-LF edges, see Supplementary Figure S6), suggesting that the original dearth of such edges was a statistical artifact resulting from the excess of triplet-to-triplet edges.

We also explored the association of between genes' role in metabolism and their pattern of post-hexaploidy evolution using the *A. thaliana* metabolic network (*Methods).* However, again considering the state of each pillar at the end of the root branch, we did not find an excess of shared metabolic interactions between triplicated or single-copy genes in this network (Supplementary Figure S6).

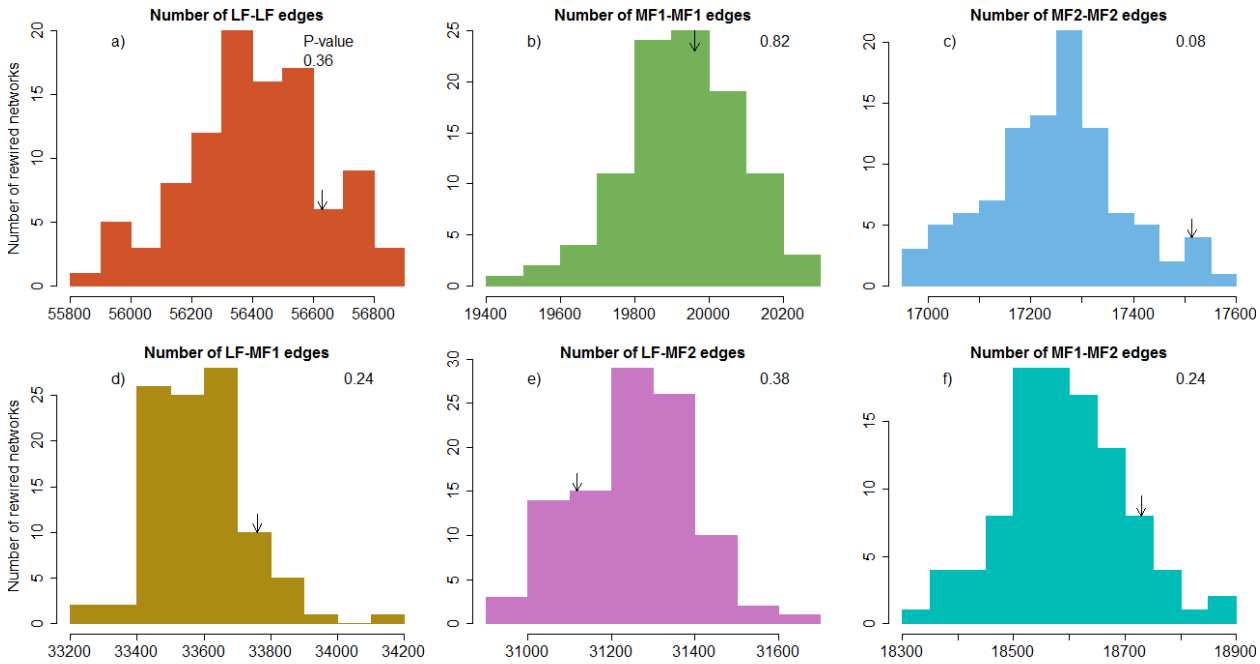


Figure 4. Subgenome-specific edge counts for 100 rewired *Brassica rapa* co-expression networks compared to those from the actual network. a) Distribution of the number of edges connecting pairs of *B. rapa* genes both from the LF subgenome in 100 rewired networks. b) Distribution of the number of edges connecting pairs of genes both from the MF1 subgenome. c) Distribution of the number of edges connecting pairs of genes both from the MF2 subgenome. d) Distribution of the number of edges connecting LF genes to MF1 genes. e) Distribution of the number of edges connecting LF genes to MF2 genes. f) Distribution of the number of edges connecting MF1 and MF2 genes. In each panel, the black arrows show the number of edges with that set of subgenomes assignments for the true network.

Finally, we asked whether genes from the same subgenome are more likely to be co-expressed. We constructed a *B. rapa* co-expression network from the RNASeq data described in the *Methods* section. In this network edges connect pairs of genes that are highly correlated in their expression (*Methods*). The inferred co-expression network contains 3,933 nodes (e.g., genes) from the LF subgenome, 2,310 nodes from MF1 and 1,982 from MF2. We then counted the number of edges

connecting pairs of nodes from the same subgenome. To assess whether there was an excess of such shared subgenome co-expression relationships, we randomly rewired the network 100 times and compared the edge count distributions from these randomized networks to those of the real network. We found that the real network did not show a significant excess of shared edges between genes from the same subgenome when compared to the randomized networks (LF-LF P=0.36, MF1-MF1 P=0.82, MF2-MF2 P=0.08, Figure 4).

Subgenome of origin does not affect the propensity to experience a selective sweep

We tested for associations between genes' subgenome of origin and their propensity to experience recent selective sweeps. Data on these sweeps was taken from a recent scan in *B. rapa* by Qi et al. (Qi et al. 2019). No subgenome had either an excess or a deficit of observed sweeps relative to the other two (Supplementary Figure S7). Genes from the MF1 subgenome showed slightly negative association with selective sweeps (P = 0.0089, chi-square test).

Discussion

The combination of the new genome sequence of *Crambe hispanica* and our modeling of the post-WGT evolution of the four Brassiceae genomes using POInT allowed us to draw a number of conclusions regarding the *Brassica* WGT. We confirmed previous work (Tang et al. 2012; Cheng et al. 2012) arguing that these genomes derive from a pair of ancient allopolyploidies: more subtly, we also show that, as had also been proposed, the least fractionated subgenome (e.g., the one with the most retained genes) is very likely the genome that was added last. To these proposals, we add the new observation that these hybridization events were likely not particularly closely spaced in time: our model predicts that on the order of 1/3 of the gene losses from subgenomes MF1 and MF2 that occurred on the root branch occurred *before* the arrival of the LF subgenome. In the future, it will be interesting to further refine the timing of these events: the problem, however, is a challenging one because the allopolyploid

nature of the events means that molecular clock approaches will tend to estimate speciation times for the allopolyploid ancestors rather than hybridization times.

Many forces shape genome evolution after polyploidy. A tendency for genes that operate in multiunit complexes or involved in signaling cascades to remain duplicated post-polyploidy is best explained by the presence of dosage constraints driven by a need to maintain the stoichiometry and kinetics of assembly for such functional units (Birchler et al. 2005; Birchler and Veitia 2007, 2012; Conant et al. 2014; Birchler et al. 2016). On the other hand, genes involved in functions such as DNA repair very often return rapidly to singleton status after duplication (Freeling 2009; De Smet et al. 2013). Our results illustrate the importance of these dosage effects, with genes whose products interact with many other gene products in *A. thaliana* being overly likely to be retained in triplicate in these *Brassica*e genomes. Notably, this pattern is not observed for metabolic genes, a result we interpret as illustrating metabolism's dynamic robustness to gene dosage changes (Kacser and Burns 1981).

We had previously argued that one force driving the biased fractionation that distinguishes the LF, MF1 and MF2 subgenomes might be selection to maintain coadapted complexes from a single parental genome (Emery et al. 2018). That such coadapted complexes exist and respond to polyploidy is suggested by the gene conversions seen after the yeast polyploidy among the duplicated ribosomal and histone proteins (Evangelisti and Conant 2010; Scienski et al. 2015). However, these examples may be exceptions rather than the rule, meaning that pressure to maintain coadapted complexes is not a significant driver of biases in fractionation. We found that although there was some degree of functional distinction for single copy genes from the LF subgenome (e.g., enrichment in biological processes such as DNA repair and RNA interference), more generally speaking, there was no significant evidence of functionally incompatibility of single-copy genes from different subgenomes. Thus, genes from the same subgenome were not more likely to interact with each other physically, nor were the genes returned to single copy on the common root branch functionally subdivided among the subgenomes. And even the DNA repair enzymes that rapidly returned to single-copy appear to derive from at least two of the three subgenomes. It hence appears that De Smet et al.'s (De Smet et al. 2013) original hypothesis that these

genes may be prone to dominant negative interactions may best explain their preference for a single-copy state.

It remains to be seen if the "mix and match" pattern of subgenome retention observed here represents the dominant mode of evolution for most allopolyploidies. Of course, whether or not subgenome conflicts exist may be partly a question of the preexisting differences between the parental species, and a more general survey of allopolyploidies that includes estimates of the subgenome divergence prior to the polyploidy events would be most enlightening. If the pattern holds, however, the implications could be significant: hybridization represents a potentially important means of adaption (Paterson 2005; Hollister 2015; Alix et al. 2017; Blanc-Mathieu et al. 2017; Smukowski Heil et al. 2017), and if we combine this factor with the propensity for polyploidies to generate evolutionary innovations (Edger et al. 2015) and the value of holding dosage sensitive genes in duplicate long enough to allow such innovations (Blanc and Wolfe 2004; Conant and Wolfe 2008b; Conant et al. 2014; Zhao et al. 2017; Liang and Schnable 2018; Qiu et al. 2020), we can find ourselves in a position to strongly support Ohno's arguments on the power of polyploidy.

Methods

Crambe hispanica (PI 388853) sample preparation, genome sequencing

Leaf tissue was harvested from 36 dark treated inbred plants (selfed for nine generations; PI 388853). Dark treatment was performed to reduce chloroplast abundance and involved leaving the plants in a dark room for 3-4 days. After treatment, 5g of tissue was collected across 36 plants. This process was repeated three times, allowing us to obtain a total of 15g of tissue. This tissue was then sent to the University of Delaware Sequencing and Genotyping Center at the Delaware Biotechnology Institute (Newark, DE, USA) for high molecular weight DNA isolation and library preparation prior to PacBio (Pacific Biosciences, Menlo Park, CA, USA) and Illumina (San Diego, CA) sequencing. Libraries were prepared using standard SMRTbell procedures, followed by sequencing of 11 PacBio SMRT cells on a PacBio sequel and one PacBio SMRT cell of RSII sequencing. Paired-end 150 bp reads were generated

on an Illumina HiSeq 2500 system. For Hi-C scaffolding, 0.5g tissue sample was sent to Phase Genomics (Phase Genomics Inc. Seattle, WA, USA).

Crambe hispanica v1.1 genome assembly and annotation

The assembly of the *Crambe hispanica* v1.1 genome was performed using Canu v1.6 (Koren et al. 2017). In total, 3.9 million raw PacBio reads spanning 48 Gb were used as input for Canu. The following parameters were modified for assembly: minReadLength=1000, GenomeSize=500Mb, corOutCoverage=200 "batOptions=-dg 3 -db 3 – dr 1 -ca 500 -cp 50". All other parameters were left as default. The draft assembly output had a contig N50 of 4.4 Mb across 1,019 contigs with a total assembly size of 480 Mb. The assembly graph was visualized using Bandage (Wick et al. 2015) to assess for ambiguities in the graph related to repetitive elements and heterozygosity. The draft Canu assembly was polished reiteratively using high-coverage Illumina paired-end data (82 million reads) with Pilon v1.22 (Walker et al. 2014). Quality filtered Illumina reads were aligned to the genome using bowtie2 v2.3.0 (Langmead and Salzberg 2012) under default parameters and the resulting bam file was used as input for Pilon with the following parameters: --flank 7, --K 49, and --mindepth 8. Pilon was run recursively three times using the updated reference each time to correct the maximum number of residual errors.

A Proximo Hi-C library was prepared as described (Phase Genomics Inc. Seattle, WA, USA) and sequenced on an Illumina HiSeq 2500 system with paired-end 150 bp reads. The *de novo* genome assembly of Hi-C library reads were used as input data for the Phase Genomics Proximo Hi-C genome scaffolding platform. This resulted in 18 scaffolds that includes 99.5% of the original assembly length with a scaffold N50 of 32.6 Mbp and scaffold N90 of 30.1 Mbp.

The genome was annotated using MAKER (Campbell et al. 2014), using evidence of protein sequences downloaded from the Araport 11 and Phytozome 12 plant databases (Cheng et al. 2017; Goodstein et al. 2012), and *C. hispanica* transcriptome data assembled with StringTie (Pertea et al. 2015). Repetitive regions in the genome were masked using a custom repeat library and Repbase Update (Bao et al. 2015) through Repeatmasker Open-4.0 (Smit et al. 2015). *Ab initio* gene prediction was performed

using SNAP (Korf 2004) and Augustus (Stanke and Waack 2003). The resulting MAKER gene set was filtered to select gene models with Pfam domain and annotation edit distance (AED) < 1.0. Then, the amino acid sequences of predicted genes were searched against a transposase database using BLASTP and an E-value cutoff of 10^{-10} (Campbell et al. 2014). If more than 30% of a given gene aligned to transposases after the removal of low complexity regions, that gene was removed from the gene set.

Triple-conserved Synteny reconstruction

Based on their phylogenetic placement and assembly quality, we selected and retrieved from CoGe (Lyons and Freeling 2008; Lyons et al. 2008a) three additional mesohexaploid genomes for our analyses: those of *Brassica rapa* (version 1.5, CoGe id 24668) (Wang et al. 2011), *Brassica oleracea* (TO1000 version 2.1, CoGe id 26018) (Liu et al. 2014; Parkin et al. 2014) and *Sinapis alba* (version 1.1, CoGe id 33284). We have developed a three-step pipeline for inferring the conserved synteny blocks created by polyploidy (Emery et al. 2018). For the first step of this pipeline, we used *Arabidopsis thaliana* Col-0 version 10.29 (CoGe genome id 20342) as a nonhexaploid outgroup and identified homologous genes between it and each of the four hexaploid genomes using GenomeHistory (Conant and Wagner 2002). Genes were defined as homologous if their translated products shared 70% percent amino acid sequence identity, and the shorter sequence was at least 80% percent of the length of the longer. In the second step, we sought to place genes from each of the hexaploid genomes into blocks of triple-conserved synteny (TCS) relative to their *A. thaliana* homologs. To do so, we inferred a set of "pillars," each of which contains a single gene (or group of tandem duplicates) from *A. thaliana* and between 1 and 3 genes from the hexaploidy genome. Using simulated annealing (Kirkpatrick et al. 1983; Conant and Wolfe 2006), we sought a combination of pillar gene assignments and relative pillar order that maximized the TCS. In the third and final step, we merged the pillars across the four hexaploid genomes, using their *A. thaliana* homologs as indices. We then sought a global pillar order that minimized the number of synteny breaks across all of the hexaploid genomes. These three steps resulted in a set of 14,050 ordered pillars, each with at least one surviving gene from each of the four genomes (Figure 1) and a corresponding

"ancestral" gene from *A. thaliana*. Supplementary Table S1 shows that POInT's model inferences are consistent across a number of such estimated ancestral orders.

An ancestral genome order reconstruction

As a verification of our POInT pipeline, we also sought an independent inference of the order of the genes in the parental subgenomes just prior to the first step of the *Brassica* triplication. First, we used CoGe's SynMap (Lyons et al. 2008b) to identify homologs between the *A. thaliana* and *Arabidopsis lyrata* genomes and those between *B. rapa* and *B. oleracea.* The SynMap algorithm was applied with a chaining distance of 50 genes and a minimum of five aligned gene pairs to identify likely orthologous genes in all pairwise-comparisons of the four genomes. Paralogs were identified by self-comparisons of each of the two *Brassica* genomes with SynMap. Then these orthologs and paralogs were grouped into 24,011 homology sets with the 'OMG!' program (Zheng et al. 2011). Every homology set consists of 1-3 *Brassica* paralogs from each of the three *Brassica* genomes and a single *Arabidopsis* gene from each of the two *Arabidopsis* genomes, representing one "candidate gene" in the reconstructed ancestral genome. Among these, 2,178 homology sets contained the maximum of eight genes (one each from the two *Arabidopsis* genomes and three each from the two *Brassica* genomes).

The homology sets were used to retrieve the ancestral gene order from adjacency graph using an efficient algorithm called Maximum Weight Matching (MWM) (Zheng et al. 2013). We identified all the gene adjacencies in the four genomes, considering only the genes in the homology sets. Each adjacency was then weighted according to how many of the 8 possible adjacencies were actually observed. The MWM produced an optimal set of 10,944 linear contigs containing all 24,001 putative ancestral genes from the homology sets that included 13,057 of 45,982 total adjacencies in the data set, with the remaining adjacencies being inconsistent with this optimal set. We used the contigs in the output of the MWM to reconstruct each of the 5 ancestral chromosomes. There were 34 contigs containing large proportions of genes originating in two or more of the ancient chromosomes that were discarded, as were any contigs containing four or fewer genes from a *Brassica* genome. While the 9,712 contigs so omitted

represent 89% of all contigs, they represent only 55% of the genes, leaving a small group of large contigs with strong synteny relations in our ancestral reconstruction. We next identified adjacencies among the contigs themselves and applied a second iteration of MWM on them, giving the optimal ordering of those contigs. Combining these orders with the existing gene order information within each contig yields the position of all the genes on each ancestral chromosome. This order was mapped to our set of pillars of TCS, giving a subset of those pillars ordered by this ancestral order estimate.

Modeling WGT using POInT

POInT employs user-defined Markov models of gene loss after WGT. These models have seven states (Figure 2): the triplicated state **T** in which all three copies after WGT are still present; the "duplicated" states $\mathbf{D_{1,2}}$, $\mathbf{D_{1,3}}$, $\mathbf{D_{2,3}}$ where one out of the three gene copies has been lost, and the three single-copy state $\mathbf{S_1}$, $\mathbf{S_2}$, $\mathbf{S_3}$. Previous work suggested that the three subgenomes that formed these hexaploids are distinct in their patterns of gene preservation (Tang et al. 2012; Cheng et al. 2012), consisting of a "less fractionated" genome (LF), a subgenome with intermediate levels of gene loss (more fractionated 1 or MF1) and an even more fractionated subgenome (MF2). We hence defined state $\mathbf{S_1}$ to correspond to LF and $\mathbf{S_2}$ and $\mathbf{S_3}$ to MF1 and MF2, respectively.

*The phylogenetic relationships of the triplicated members of the Brassica*ceae

POInT fits these models to the pillar data under an assumed phylogenetic topology using maximum likelihood, allowing us to use that likelihood statistic to compare different phylogenetic relationships among these four hexaploid taxa. POInT's computational demands were too great to allow testing all 15 rooted topologies of 4 species (POInT's models are not time reversible). However, by making the reasonable assumption that *B. rapa* and *B. oleracea* are sister to each other, we were able to test the three potential relationships of *C. hispanica* and *S. alba* to the two *Brassicas.* Figure 1 gives the maximum likelihood topology: the two alterative topologies and their likelihoods are given in Supplementary Figure S1.

Selective constraints of the retained triplets

We identified 218 pillars that retained triplicated genes across all four genomes and where the confidence in their subgenome assignments was ≥ 95%. For each such pillar, the 12 sequences were aligned using T-coffee (Notredame et al. 2000). The cladogram for each 12 genes consists of three subtrees grouping four sequences that belong to same subgenome in the same sister group (Supplementary Figure S4). Then, using codeml in PAML (Yang 2007) with CodonFreq set to F3X4, we inferred three distinct dN/dS ratios: one for each of the three subtrees deriving from the three parental subgenomes.

Functional analysis of single-copy genes from different subgenomes

We performed functional analysis for genes where we have high (≥ 95%) confidence that they returned to single copy along the common root branch. Using the corresponding "ancestral" locus from *A. thaliana*, we performed individual Gene Ontology analyses with PANTHER (Mi et al. 2019) Overrepresentation Tests (release date 20190711) for genes from each subgenome. The background list used in all cases was the loci that remained duplicated or triplicated at the end of the root branch.

Protein-protein interaction and metabolic network analysis

The *A. thaliana* protein-protein interaction (PPI) network was downloaded from BioGRID (Stark et al. 2011; Arabidopsis Interactome Mapping Consortium 2011). The root-branch post-WGT subgenome assignments for each "ancestral" locus represented by an *Arabidopsis* gene were mapped onto the nodes (gene products) of the PPI network, so long as our confidence in those subgenome assignments was ≥ 95%. Similarly, for the extant *B. rapa*, we took loci with high subgenome assignment confidence ≥ 95% and mapped their *A. thaliana* orthologs onto network nodes. The resulting PPI network (Figure 3) was visualized using Gephi 0.9.2 (Bastian et al. 2009) with the Fruchterman Reingold and Yifan Hu layout algorithms (Fruchterman and Reingold 1991; Hu 2006). To test whether gene products from the same

subgenome are over-connected in this network, we permutated the subgenome assignments 1,000 times, holding the network topology unchanged. We then compared the actual number of edges connecting single copy genes from the same subgenome with the distribution of this value seen in the randomized networks (Supplementary Figure S6). We also asked whether the ancestral genes corresponding to retained triplets showed an excess of connections amongst themselves. Because the number of edges between retained triplets and between single-copy genes are not independent, we performed an additional set of permutations, in which we held all the triplet rows constant and only shuffled the subgenome assignments of the remaining nodes.

We performed similar analyses using the AraGEM v1.2 metabolic network from *A. thaliana* (de Oliveira Dal'Molin et al. 2010; Bekaert et al. 2012). In this network, each node represents a biochemical reaction, and pairs of nodes are connected by edges if their respective reactions share a metabolite. For each *A. thaliana* gene encoding an enzyme catalyzing one such reaction, we mapped the root-branch subgenome assignments (again requiring ≥ 95% confidence), assigning to that gene three presence/absence variables (one per subgenome). Then, for each subgenome, we counted the number of edges between pairs of nodes with at least one pair of single-copy genes from a common subgenome. We assessed significance by holding the network topology and *Arabidopsis* gene assignments constant and randomizing the subgenome assignments 1,000 times. We then compared the distributions of the single-subgenome edge counts from the simulations with the actual values (Supplementary Figure S6).

Brassica rapa co-expression network analysis

We generated a gene expression dataset for *Brassica rapa* spanning diverse experimental conditions, including following: a cold treatment in leaves (4hrs and 28hrs post), methyl jasmonate treatment in leaves (4hrs and 28hrs post), anaerobic treatment in leaves (4 and 8hrs post), salt treatment in roots (4hrs and 28hrs post) and a diurnal time course in leaves (every 4hrs, 6 timepoints) in standard light-dark conditions but also in complete dark and complete light conditions. Total RNA was extracted from above organs using the Invitrogen Purelink RNA Mini Kit (Carlsbad, CA, USA), converted into a library

using the Illumina TruSeq RNA kit (San Diego, CA, USA), and paired-end 100bp reads were sequenced on the HiSeq-2000 instrument at the VJC Genomics Sequencing Laboratory at the University of California, Berkeley. The NextGENe V2.17 (SoftGenetics, State College, PA, USA) software package was used to remove low-quality Illumina data, map reads to the *B. rapa* FPsc (v1.0, CoGe id 20101) reference genome, and calculate normalized RPKM (reads per kilobase of transcript per million) values for all genes.

We filtered the dataset to only include genes that were missing a measured expression value for at most one of the 32 RNAseq libraries, leaving 24,907 *B. rapa* genes in it. The gene identifiers used for the expression dataset were from the *B. rapa* FPsc (v1.0, CoGe id 20101) reference genome, so we translated these identifiers those from *B. rapa* Chiifu (v1.5, id 24668) using CoGe SynMap (Lyons et al. 2008b). In so doing, we only used *B. rapa* genes with one-to-one matches between the two references. For any pair of genes in the expression dataset, we calculated the Spearman correlation coefficient of their RPKM values. A co-expression network was then constructed using highly correlated gene pairs, e.g., pairs having Spearman correlation coefficients ≥ 0.9 (positive correlations), or ≤ -0.9 (negative correlations). Thus, the nodes of this co-expression network are *B. rapa* genes, and the edges represent correlation in expression. The co-expression network was randomized 100 times by rewiring the edges, while holding the nodes and their subgenome assignments unchanged. In other words, all edges were broken and randomly reconnecting, preserving the degree of every node (Pérez-Bercoff et al. 2011). The distributions of inter-subgenome and intra-subgenome edge counts are shown in Figure 4.

Association between recent selective sweeps in B. rapa and subgenomes

B. rapa genes were divided into those in the regions of selective sweeps detected by SweeD (Pavlidis et al. 2013) in either turnip, toria, Indian sarson, pak choi or Chinese cabbage (vegetable types of *B. rapa*), and those showing no such signatures (Qi et al. 2017, 2019). We tested whether particular subgenomes (posterior probability ≥ 0.95) were unusually likely or unlikely to have experienced a

selective sweep using chi-square test. The association plot as shown in Supplementary Figure S7 was visualized using the vcd package version 1.4-4 in R 3.6.0 (Meyer et al. 2006; Zeileis et al. 2007).

Data Access

The assembly of the *Crambe hispanica* v1.1 genome will be available under NCBI BioProject PRJNA631330. The annotated *Crambe hispanica* v1.1 genome will also be available to download from CoGe (id58014). POInT input files are available on figshare (https://figshare.com/s/05d8f68dd2b8f465d013).

Acknowledgements

This project was supported by U.S. National Science Foundation grant NSF-IOS-1339156 (Y.H., M.M., P.P.E., H.A., R.S.A., X.Q., M.B., E.L., J.C.P., and G.C.C.). We would like to thank the U.S. Department of Energy Joint Genome Institute and the Brassicaceae Map Alignment Project (BMAP) consortium for allowing us access to the *Sinapis alba* genome. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.We would like to thank A. Platts for assistance with our analyses and J. A. Birchler, R. Roberts, J. Thorne and H. Ashrafi for helpful discussions.

References

Alger EI, Edger PP. 2020. One subgenome to rule them all: underlying mechanisms of subgenome dominance. *Curr Opin Plant Biol* **54**: 108–113.

Alix K, Gérard PR, Schwarzacher T, Heslop-Harrison JSP. 2017. Polyploidy and interspecific hybridization: Partners for adaptation, speciation and evolution in plants. *Ann Bot* **120**: 183–194.

Arabidopsis Interactome Mapping Consortium. 2011. Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science* **333**: 988–993.

Arias T, Pires JC. 2012. A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. *Taxon* **61**: 980–988.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Ségurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**: 171–178.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: An open source software for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social. *Int AAAI Conf Weblogs Soc Media* 361–362.

Bekaert M, Edger PP, Hudson CM, Pires JC, Conant GC. 2012. Metabolic and evolutionary costs of herbivory defense: Systems biology of glucosinolate synthesis. *New Phytol* **196**: 596–605.

Birchler JA, Johnson AF, Veitia RA. 2016. Kinetics genetics: Incorporating the concept of genomic balance into an understanding of quantitative traits. *Plant Sci* **245**: 128–134.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends Genet* **21**: 219–226.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci U S A* **109**: 14746–14753.

Birchler JA, Veitia RA. 2014. The gene balance hypothesis: dosage effects in plants. In *Plant Epigenetics and Epigenomics: Methods and Protocols*, pp. 25–32, Humana Press, Totowa, NJ.

Birchler JA, Veitia RA. 2007. The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell* **19**: 395–402.

Bird KA, VanBuren R, Puzey JR, Edger PP. 2018. The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol* **220**: 87–93.

Blanc-Mathieu R, Perfus-Barbeoch L, Aury JM, Da Rocha M, Gouzy J, Sallet E, Martin-Jimenez C, Bailly-Bechet M, Castagnone-Sereno P, Flot JF, et al. 2017. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. *PLoS Genet* **13**: 1–36.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**: 1679–1691.

Campbell MS, Law MY, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ, et al. 2014. MAKER-P: A Tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol* **164**: 513–524.

Carlsson AS, Clayton D, Salentiijn E, Toonen M. 2007. *Oil crop platforms for industrial uses*. www.cplbookshop.com.

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J* **89**: 789–804.

Cheng F, Wu J, Fang L, Sun S, Liu B, Lin K, Bonnema G, Wang X. 2012. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS One* **7**: e36442.

Cheng F, Wu J, Liang J, Wang X. 2014. Genome triplication drove the diversification of Brassica plants. *Hortic Res* **1**: 14024.

Codoñer FM, Fares MA. 2008. Why should we care about molecular coevolution? *Evol Bioinforma* **2008**: 237–246.

Conant GC. 2014. Comparative genomics as a time machine: How relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. *Mol Biol Evol* **31**: 3184–3193.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol* **19**: 91–98.

Conant GC, Wagner A. 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res* **30**: 3378–3386.

Conant GC, Wolfe KH. 2006. Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol* **4**: e109.

Conant GC, Wolfe KH. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol* **3**: 129.

Conant GC, Wolfe KH. 2008a. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* **179**: 1681–1692.

Conant GC, Wolfe KH. 2008b. Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* **9**: 938–950.

Costello R, Emms DM, Kelly S. 2019. Gene Duplication Accelerates the Pace of Protein Gain and Loss from Plant Organelles. *Mol Biol Evol* **37**: 969–981.

de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK. 2010. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol* **152**: 579–589.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent

gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A* **110**: 2898–2903.

Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc Natl Acad Sci U S A* **115**: E409–E417.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, et al. 2015. The butterfly plant arms-race escalated by gene and genome duplications. *Proc Natl Acad Sci* **112**: 8362–8366.

Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-W, Bewick AJ, Ji L, Platts AE, Bowman MJ, et al. 2017. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* **29**: 2150–2167.

Emery M, Willis MMS, Hao Y, Barry K, Oakgrove K, Peng Y, Schmutz J, Lyons E, Pires JC, Edger PP, et al. 2018. Preferential retention of genes from one parental genome after polyploidy illustrates the nature and scope of the genomic conflicts induced by hybridization. *PLoS Genet* **14**: e1007267.

Evangelisti AM, Conant GC. 2010. Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biol Evol* **2**: 826–834.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol* **60**: 433–453.

Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC. 2012. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr Opin Plant Biol* **15**: 131–139.

Fruchterman TMJ, Reingold EM. 1991. Graph Drawing by Force-directed Placement. *Softw Pract Exp* **21**: 1129–1164.

Gong L, Salmon A, Yoo MJ, Grupp KK, Wang Z, Paterson AH, Wendel JF. 2012. The cytonuclear dimension of allopolyploid evolution: An example from cotton using rubisco. *Mol Biol Evol* **29**: 3023–3036.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res* **40**: 1178–1186.

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biol* **8**: R209.

Hollister JD. 2015. Polyploidy: Adaptation to the genomic environment. *New Phytol* **205**: 1034–1039.

Hu Y. 2006. Efficient, High-Quality Force-Directed Graph Drawing. *Math J* **10**: 37–71.

Kacser H, Burns JA. 1981. The molecular basis of dominance. *Genetics* **97**: 639–666.

Kirkpatrick S, Gelatt CDJ, Vecchi MP. 1983. Optimization by simulated annealing. *Science* **220**: 671–680.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**: 722–736.

Korf I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**: 59.

Lagercrantz U. 1998. Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. *Genetics* **150**: 1217–1228.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lazzeri L, De Mattei F, Bucelli F, Palmieri S. 1997. Crambe oil - A potential new hydraulic oil and quenchant. *Ind Lubr Tribol* **49**: 71–77.

Liang Z, Schnable JC. 2018. Functional divergence between subgenomes and gene pairs after whole genome duplications. *Mol Plant* **11**: 388–397.

Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, Zhao M, Ma J, Yu J, Huang S, et al. 2014. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun* **5**: 3930.

Lukens LN, Quijada PA, Udall J, Pires JC, Schranz ME, Osborn TC. 2004. Genome redundancy and plasticity within ancient and recent Brassica crop species. *Biol J Linn Soc* **82**: 665–674.

Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* **53**: 661–673.

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, et al. 2008a. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol* **148**: 1772–1781.

Lyons E, Pedersen B, Kane J, Freeling M. 2008b. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. *Trop Plant Biol* **1**: 181–190.

Lysak MA. 2009. Comparative cytogenetics of wild crucifers (Brassicaceae). In *Biology and Breeding of Crucifers* (ed. S.K. Gupta), pp. 177–205, CRC Press Taylor & Francis Group, Boca Raton London New York.

Lysak MA, Koch MA, Pecinka A, Schubert I. 2005. Chromosome triplication found across the tribe Brassiceae. *Genome Res* **15**: 516–525.

Maere S, Bodt S De, Raes J, Casneuf T, Montagu M Van, Kuiper M, Peer Y Van de. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci* **102**: 5454–5459.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci* **107**: 9270–9274.

Makino T, McLysaght A. 2012. Positionally biased gene loss after whole genome duplication: Evidence from human, yeast, and plant. *Genome Res* **22**: 2427–2435.

McClintock B. 1984. The significance of responses of the genome to challenge. *Science* **226**: 792–801.

Merico A, Sulo P, Piškur J, Compagno C. 2007. Fermentative lifestyle in yeasts belonging to the Saccharomyces complex. *FEBS J* **274**: 976–989.

Meyer D, Zeileis A, Hornik K. 2006. The strucplot framework: Visualizing multi-way contingency tables with vcd. *J Stat Softw* **17**: 1–48.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* **47**: D419–D426.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205–217.

Ohno S. 1970. *Evolution by Gene Duplication*. Springer, Verlag Berlin Heidelberg.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**.

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ. 2005. Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. *Genetics* **171**: 765–781.

Parkin IAP, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, et al. 2014. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biol* **15**: R77.

Paterson AH. 2005. Polyploidy, evolutionary opportunity, and crop adaptation. In *Genetics of Adaptation* (ed. R. Mauricio), pp. 191–196, Springer Netherlands, Dordrecht.

Pavlidis P, Živković D, Stamatakis A, Alachiotis N. 2013. SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol* **30**: 2224–2234.

Pérez-Bercoff Å, McLysaght A, Conant GC. 2011. Patterns of indirect protein interactions suggest a spatial organization to metabolism. *Mol BioSyst* **7**.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295.

Qi X, An H, Hall TE, Di C, Blischak PD, Pires JC, Barker MS. 2019. Genes derived from ancient polyploidy have higher genetic diversity and are associated with domestication in Brassica rapa. *bioRxiv*.

Qi X, An H, Ragsdale AP, Hall TE, Gutenkunst RN, Pires JC, Barker MS. 2017. Genomic inferences of domestication events are corroborated by written records in Brassica rapa. *Mol Ecol* **26**: 3373–3388.

Qiu Y, Tay Y Van, Ruan Y, Adams KL. 2020. Divergence of duplicated genes by repeated partitioning of splice forms and subcellular localization. *New Phytol* **225**: 1011–1022.

Renny-Byfield S, Gong L, Gallagher JP, Wendel JF. 2015. Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Mol Biol Evol* **32**: 1063–1071.

Rudloff E, Wang Y. 2011. Crambe. In *Wild Crop Relatives: Genomic and Breeding Resources: Oilseeds*

(ed. C. Kole), pp. 97–116, Springer, Heidelberg Dordrecht London New York.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci* **108**: 4069–4074.

Schoonmaker A, Hao Y, McK. Bird DM, Conant GC. 2020. A single, shared triploidy in three species of parasitic nematodes. *G3-Genes Genom Genet* **10**: 225–233.

Schranz ME, Lysak MA, Mitchell-Olds T. 2006. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends Plant Sci* **11**: 535–542.

Scienski K, Fay JC, Conant GC. 2015. Patterns of gene conversion in duplicated yeast histones suggest strong selection on a coadapted macromolecular complex. *Genome Biol Evol* **7**: 3249–3258.

Seoighe C, Wolfe KH. 1998. Extent of genomic rearrangement after genome duplication in yeast. *Proc Natl Acad Sci U S A* **95**: 4447–4452.

Sharbrough J, Conover JL, Tate JA, Wendel JF, Sloan DB. 2017. Cytonuclear responses to genome doubling. *Am J Bot* **104**: 1277–1280.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Smit A, Hubley R, Green P. 2015. RepeatMasker. http://www.repeatmasker.org/.

Smith SD, Pennell MW, Dunn CW, Edwards S V. 2020. Phylogenetics is the New Genetics (for Most of Biodiversity). *Trends Ecol Evol* 1–11.

Smukowski Heil CS, DeSevo CG, Pai DA, Tucker CM, Hoang ML, Dunham MJ. 2017. Loss of Heterozygosity Drives Adaptation in Hybrid Yeast. *Mol Biol Evol* **34**: 1596–1612.

Soltis PS, Soltis DE. 2012. *Polyploidy and Genome Evolution*. Springer-Verlag, Berlin Heidelberg.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**: 215–225.

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, et al. 2011. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**: 698–704.

Sukeena JM, Galicia CA, Wilson JD, McGinn T, Boughman JW, Robison BD, Postlethwait JH, Braasch I, Stenkamp DL, Fuerst PG. 2016. Characterization and Evolution of the Spotted Gar Retina. *J Exp Zool Part B Mol Dev Evol* **326**: 403–421.

Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC. 2012. Altered patterns of fractionation and exon deletions in Brassica rapa support a two-step model of paleohexaploidy. *Genetics* **190**: 1563–1574.

Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* **16**: 934–946.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet* **10**: 725–732.

Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet* **18**: 411–424.

van Hoek MJA, Hogeweg P. 2009. Metabolic adaptation after whole genome duplication. *Mol Biol Evol* **26**: 2441–2453.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**: e112963.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species Brassica rapa. *Nat Genet* **43**: 1035–1039.

Warwick SI, Gugel RK. 2003. Genetic variation in the Crambe abyssinica - C. hispanica - C. glabrata complex. *Genet Resour Crop Evol* **50**: 291–305.

Wendel JF, Lisch D, Hu G, Mason AS. 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr Opin Genet Dev* **49**: 1–7.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* **31**: 3350–3352.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708–713.

Woodhouse MR, Cheng F, Pires JC, Lisch D, Freeling M, Wang X. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc Natl Acad Sci U S A* **111**: 5283–5288.

Xie T, Zhang F-G, Zhang H-Y, Wang X-T, Hu J-H, Wu X-M. 2019. Biased gene retention during diploidization in Brassica linked to three-dimensional genome organization. *Nat Plants* **5**: 822–832.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Yoo M-J, Liu X, Pires JC, Soltis PS, Soltis DE. 2014. Nonadditive Gene Expression in Polyploids. *Annu Rev Genet* **48**: 485–517.

Zeileis A, Meyer D, Hornik K. 2007. Residual-based shadings for visualizing (conditional) independence. *J Comput Graph Stat* **16**: 507–525.

Zhao M, Zhang B, Lisch D, Ma J. 2017. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell* **29**: 2974–2994.

Zheng C, Chen E, Albert VA, Lyons E, Sankoff D. 2013. Ancient eudicot hexaploidy meets ancestral eurosid gene order. *BMC Genomics* **14**: S3.

Zheng C, Swenson K, Lyons E, Sankoff D. 2011. OMG! Orthologs in Multiple Genomes – Competing Graph-Theoretical Formulations. In *International Workshop on Algorithms in Bioinformatics*, pp. 364–375.

1