Genetics: Published Articles Ahead of Print, published on February 2, 2012 as 10.1534/genetics.111.137349

Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy

Haibao Tang[*, †, 1], Margaret R. Woodhouse[*, 1], Feng Cheng[‡], James C. Schnable[*], Brent S. Pedersen[*], Gavin Conant[§, ††], Xiaowu Wang[‡], Michael Freeling[*] and J. Chris Pires[**, ††]

[*] Department of Plant and Microbial Biology, University of California, Berkeley, California, USA

[†] J. Craig Venter Institute, 9704 Medical Center Dr., Rockville, Maryland, USA

[‡] Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China

[§] Division of Animal Sciences, University of Missouri, Columbia, Missouri, USA

[**] Division of Biological Sciences, University of Missouri, Columbia, Missouri, USA

[††] Informatics Institute, University of Missouri, Columbia, Missouri, USA

[1] **These authors contributed equally**

Running Title

Patterns of fractionation in Brassica subgenomes

Keywords:

Brassica rapa, genome duplication, paleohexaploidy, fractionation, genome dominance

Corresponding author:

Haibao Tang, Ph. D.

J. Craig Venter Institute,

9704 Medical Center Drive,

Rockville, MD, 20850

Tel: 706-248-3895

E-mail: tanghaibao@gmail.com

ABSTRACT

The genome sequence of the paleohexaploid *Brassica rapa* shows that fractionation is biased among the three subgenomes, and that the least fractionated subgenome has approximately twice as many orthologs as its close (and relatively unduplicated) relative Arabidopsis than had either of the other two subgenomes. One evolutionary scenario is that the two subgenomes with heavy gene losses (I and II) were in the same nucleus for a longer period of time than the third subgenome (III) with the fewest gene losses. This "two-step" hypothesis is essentially the same as that proposed previously for the eudicot paleohexaploidy; however, the more recent nature of the *Brassica rapa* paleohexaploidy makes this model more testable. We found that subgenome II suffered recent small deletions within exons more frequently than exons in subgenome I, as would be expected if the genes in subgenome I had already been near-maximally fractionated before subgenome III was introduced. We observed that some sequences, before these deletions, were flanked by short direct repeats, a unique signature of intra-chromosomal illegitimate recombination. We also found, through simulations, that short – single or two-gene – deletions appear to dominate the fractionation patterns in *B. rapa*. We conclude that the observed patterns of the triplicated regions in the *Brassica* genome are best explained by a two-step fractionation model. The triplication and subsequent mode of fractionation could impact the potential to generate morphological diversity – a hallmark of the *Brassica* genus.

INTRODUCTION

Ancient polyploidies are prevalent in most eukaryotic lineages, including plants (JIAO *et al.* 2011; PROOST *et al.* 2011; VAN DE PEER *et al.* 2009), fungi (KELLIS *et al.* 2004) and animals (JAILLON *et al.* 2004; AURY *et al.* 2006). Much progress has been made in dating these evolutionary events, and quantifying the retention and loss of gene duplicates after them. Gene content impacts the potential for diversification and specialization of biological functions (FORCE *et al.* 1999) and the potential for increases in morphological complexity (THOMAS *et al.* 2006). In both the eudicot and monocot clades of flowering plants, there have been multiple rounds of polyploidy followed by selective gene losses (TANG *et al.* 2010; TANG *et al.* 2008), leaving the gene repertoire of many angiosperm species greatly expanded from an estimated ancestral (i.e. in the last common ancestor) gene number of 12000-14000 loci (STERCK *et al.* 2007; TANG *et al.* 2008).

Despite the initial expansion of gene numbers immediately following genome duplications, most lineages have since experienced drastic gene loss, genome downsizing (BENNETT and LEITCH, 2005; LEITCH and LEITCH, 2008) and ultimately genetic 'diploidization' at many loci (WOLFE, 2001). A number of mechanisms could lead to the diploidization, among which the "fractionation" of duplicate genes is a major force (LANGHAM *et al.* 2004; THOMAS *et al.* 2006). During the fractionation process, many gene copies with redundant functions, and with product levels not under stringent control ("gene dosage" theory (BIRCHLER and VEITIA 2010)) tend to be lost, resulting in a reduction of gene complement that offsets the initial expansion from genome mergers. In the paleotetraploid maize, using the sorghum genome as outgroup, the fractionation mechanism was shown to be predominantly short deletions, probably via intra-

chromosomal recombination, and is certainly not randomization by nucleotide substitutions (WOODHOUSE *et al.* 2010). By whatever mechanism, the initially identical subgenomes generated by whole genome duplication events do not fractionate equally – one subgenome consistently has more genes retained on it than the other; this holds true for eukaryotes ranging from Paramecium to flowering plants and fish (SANKOFF *et al.* 2010). This phenomenon, called "fractionation bias," was first described in the Arabidopsis genome (THOMAS *et al.* 2006) and later generalized throughout major eukaryote lineages with paleopolyploidies (SANKOFF *et al.* 2010).

In plants, if not all eukaryotes, when two genomes find themselves in the same nucleus, one subgenome – as defined by fractionation bias – expresses its genes to a higher mRNA level than does the other subgenome. This is the phenomenon of genome dominance (SCHNABLE *et al.* 2011b). Since the 12 million year old maize paleotetraploid displays substantial genome dominance (SCHNABLE *et al.* 2011b), the result that *Brassica rapa* subgenome III expresses its genes to a higher level than does either subgenomes I or II was more affirming than surprising (WANG *et al.* 2011). In another study, the 23-40 million year old Arabidopsis tetraploidy (known as the alpha event; BOWERS *et al.* 2003) and the ~70 million year old pre-grass tetraploidy were shown to express genome dominance today (SCHNABLE *et al.* 2011a). However, genome dominance is most evident when tetraploidy was recent: in synthetic and natural hybrids and allotetraploids of cotton (FLAGEL and WENDEL 2010), wide hybrids of *Arabidopsis* species (CHANG *et al.* 2010; WANG *et al.* 2006), allotetraploids of *Tragopogon*

species (BUGGS *et al.* 2010a; BUGGS *et al.* 2010b), and synthesized *Brassica* lines (GAETA *et al.* 2007; XIONG *et al.* 2011). It is not yet fully understood why genome dominance persists until today after tens of millions of years of evolution.

The diploid *Brassica* species were first hypothesized to have been triplicated based on comparative mapping studies (LAGERCRANTZ 1998; LAGERCRANTZ and LYDIATE 1996; PARKIN *et al.* 2005; PARKIN *et al.* 2003). There was some skepticism based on the observation that most loci were not triplicated; however, subsequent BAC-FISH (LYSAK *et al.* 2005) and comparative BAC sequencing studies (YANG *et al.* 2006) further supported the triplication hypothesis. The recent sequencing of *Brassica rapa* has confirmed the genome triplication event that occurred in the common ancestor of all *Brassica* species (WANG *et al.* 2011).

It was demonstrated that *Brassica rapa* underwent biased fractionation — subgenome III has retained almost two thirds of *A. thaliana* orthologous genes, while subgenomes I and II have retained significantly fewer genes (WANG *et al.* 2011). Based on biased fractionation results much like those in *B. rapa*, the eudicot paleohexaploidy, known as the gamma event, was proposed to have happened by a two-step fractionation process (LYONS *et al.* 2008). Fortunately, the relatively recent paleohexaploidy in *B. rapa* and the position of the Arabidopsis genome as outgroup provide a phylogenetic system with superior analytical power. The two-step fractionation hypothesis was suggested for *Brassica*'s biased fractionation, to explain the fact that subgenome I is the most fractionated genome and subgenome III the least fractionated genome (WANG *et al.* 2011). However, this hypothesis was not formally tested.

Herein, we test the “two-step fractionation” hypothesis by examining short, exonic deletions in retained *Brassica* genes, using Arabidopsis as the outgroup. Such deletions were associated with recent, ongoing biased fractionation in maize (WOODHOUSE *et al.* 2010). We found that subgenome II had more deletions than subgenome I or subgenome III, suggesting that a two-step process of genome fractionation did indeed occur. We also show that deletions tend to accumulate in multi-copy retained genes rather than in genes retained as a single copy, a phenonomenon best explained by relaxed selection in duplicate genes.

METHODS

Partitioning of subgenomes according to number of retained genes

The identification of orthologous regions and partitioning into subgenomes follows the method described in the Supplementary Information in the *Brassica rapa* release (WANG *et al.* 2011). Briefly, multiple chromosomal segments in *B. rapa* that are orthologous to the same *A. thaliana* segment are numbered accordingly, using the established "A to X" numbering system (WANG *et al.* 2011). All *B. rapa* segments that match to the same *A. thaliana* segment are partitioned into three subgenomes (for example, segments matching *A. thaliana* segment *R* are partitioned into *R-I*, *R-II* and *R-III*) (**Figure 1A**). We exhaustively enumerated all partitions and evaluated each partition based on heuristic rules that were detailed in (WANG *et al.* 2011). After the partitioning, we counted the number of syntenic orthologs within each sub-genome. According to the number of retained orthologous genes in each subgenome, each segment was classified and named them as I, II, and III for "Most fractionated", "Moderately fractionated" and "Least fractionated", respectively, for each of to A to X segment (**Figure 1B**). We then examined the number of orthologous genes in each block; nearly all blocks showed a significant difference (with *P*-value cutoff 0.01) in gene numbers between the three subgenomes, with the only exception being block T. Finally, we concatenated each set of most fractionated (*A-I*, *B-I*… to *X-I*), moderately fractionated (*A-II*, *B-II*… to *X-II*) and least fractionated blocks (*A-III*, *B-III*… to *X-III*) respectively for downstream analyses.

Determining the sequence divergence among *B. rapa* homeologs

For paired genes inferred from syntenic alignments, we aligned the protein sequences using CLUSTALW (LARKIN *et al.* 2007) and used the protein alignments to guide coding sequence alignments by PAL2NAL (SUYAMA *et al.* 2006). To calculate *Ks*, we used the Nei-Gojobori method implemented in the *yn00* program in the PAML package (YANG 2007). A Python script was used to create a pipeline for all the calculations and is available at http://github.com/tanghaibao/bio-pipeline/tree/master/synonymous_calculation/. The actual distribution of *Ks* values is modeled and fitted as a log-transformed normal distribution (TANG *et al.* 2008).

Automated cataloging of internal deletion sites within the *B. rapa* genes

The sites of deletions were identified by an automated pipeline, illustrated in **Figure 2**. Using the *A. thaliana* orthologs as reference, we aligned one, two, or three homeologs in *B. rapa*. For each of 3648 (*A. thaliana*, *B. rapa*) pairs, we detected deletions of various sizes in the *B. rapa* gene compared to the *A. thaliana* gene. The DNA sequences of the complete genes (containing all exons and introns) were extracted for the BLASTN comparisons. For each gene pair, we used BLASTN with parameters favoring short, strong sequence matches (word size 7, spike length 15 bp, low-complexity filter off). We identified all collinear high scoring segment pairs (HSP) through "heaviest increasing subsequence" algorithm (KURTZ *et al.* 2004). There are un-matching sequences (gaps) between adjacent HSPs. For each gap pair, we noted the size in *A. thaliana*, as well as in *B. rapa*, and identified all the sites that were smaller in the *B. rapa* genes. Links to the GEvo (LYONS and FREELING 2008) URL were configured in the spreadsheet to assist manual proofing.

The changes in the sizes of the sequences are documented in the following notation: "**Bite (A=>B)"**, which means there are A bases in *A. thaliana*, but B bases in bases in *B. rapa* (**Figure 2**). For example, "Bite (81=>0)" means 81 *A. thaliana* bases were removed in the *B. rapa* gene. A very useful effect of this notation is that is also possible for "B" to be negative. For example, "Bite (82=>-7)" means 82 *A. thaliana* bases were removed and adjacent HSPs overlap by 7 bases. This is an indication of 7 bases of flanking direct repeats (as proposed in (WOODHOUSE *et al.* 2010)). We applied cutoffs of A > 30 bp and B < 10 bp, in order to select DNA chunks that decreased in size from *A. thaliana* to *B. rapa*. Exonic deletions were further identified for the deletion locations that intersect *A. thaliana* exon locations, using the tool INTERSECTBED (QUINLAN and HALL 2010).

The full catalog containing a total of 4539 deletion sites along with their locations, gene identifiers, deleted bases and GEvo links are available in the **Supporting File S1**.

Simulation of deletions of homeologs and likelihood ratio test for model selection

Based on the initial hypothesis of a deletion mechanism that independently eliminates one gene at a time, a simulation of gene loss was carried out. Starting with a length equal to the number of all genes, genes were deleted at random until the simulated number of deletions was equal to the true observed number. The distribution of apparent deletion lengths for the run was then saved, and the preceding steps were repeated 1,000 times. This gives a distribution of deletion lengths.

A genetic algorithm (GA) using 20 character states, each representing a deletion-length of various lengths was used to determine, given the region length and the distribution of observed deletion lengths, the most likely deletion model to achieve the best match between simulated and observed data. The fitness values of solutions in the genetic algorithm were scored after each step with the fittest solutions being those where the simulated number of deletion runs was least different from the observed number of runs. The components for our deletion size simulation include the following:

- Simulate under "model 1 (with only deletion size of 1 gene)" and then report counts for various deletion sizes
- Simulate under "model 1+2 (with deletion size up to 2 genes)" and then report counts for various deletion sizes
- Continue the simulation. Add one more deletion size for each new model.
- Likelihood ratio test to see which model gives the best likelihood while keeping the model as simple as possible (based on Occam's Razor). The likelihood function is defined as: $\ln L = \sum_i C_i \ln p_i$, where C_i is the simulated count and p_i is the actual frequency of deletion size i.

Scripts that perform the simulations and likelihood calculation are available at http://github.com/tanghaibao/bio-pipeline/blob/master/gap_simulations.

RESULTS

One subgenome has retained significantly more genes than the other two

As had been noted by WANG *et al.* 2011, subgenomes I, II, and III have retained 5966, 7679, and 11536 genes, respectively (ignoring genes that do not show conserved synteny with *A. thaliana*, e.g. those that are unique to *B. rapa* or have transposed) (**Table 1**). We find a similar trend in number of nucleotides per subgenome and number of genes per subgenome (both retained and non-retained) (**Table 1**). The difference in size among all three subgenomes (subgenome I the smallest, subgenome III the largest) is primarily due to the level of biased fractionation among the three subgenomes. Conversely, whole-gene deletions are 2.1 times more frequent in I than in III (10423 deletions in subgenome I versus 4853 deletions in subgenome III). There are also significant differences in numbers of singletons (no WGD duplicates) in the three subgenomes. The genes that exist only on I, II, and III are 1592, 2449, and 5211, respectively, which suggests that most single-copy genes are retained in the least fractionated subgenome III. In contrast, the differences in gene densities of the three subgenomes are less dramatic than the sheer counts (**Table 1**). We conclude that: 1) the observed gene retention bias cannot be explained by uneven gene density (for example, varied level of heterochromatic versus euchromatic sequences); and 2) the sequence removal mechanism that has shaped the retention bias did not exclusively target gene-rich regions.

Genetic distances to *A. thaliana* orthologs cannot distinguish among *B. rapa* subgenomes

The median *Ks* value between *A. thaliana – B. rapa* orthologs is 0.48, while the median *Ks* value between *B. rapa – B. rapa* homeologs is 0.37 (**Figure 3**), supporting the conclusion that *Brassica* hexaploidy occurred *after* its divergence from *Arabidopsis* (WANG *et al.* 2011). Both *A. thaliana – B. rapa* gene pairs and the *B. rapa – B. rapa* gene pairs show unimodal peak in the *Ks* distribution (**Figure 3**).

We collected *A. thaliana* genes that were represented in *B. rapa* by two or three orthologs. For each of these orthologs, we noted their subgenome assignment and determined their *Ks* value in comparison with their single *A. thaliana* ortholog. The *A. thaliana - B. rapa Ks* values were compared in a pairwise fashion with the "winner" subgenome inferred (**Table 2**). The distance between *A. thaliana* – subgenome III appears to be slightly larger than the distance between *A. thaliana* – subgenome II (χ^2-test P=0.004), while the other two pairwise comparisons are not significant at α=0.01 level. This suggests that although *Ks* is able to clearly differentiate between the time of Arabidopsis-Brassica divergence and the hexaploidy, it fails to differentiate the three subgenomes within the hexaploidy.

Additionally, we employed a tree-based method to attempt to differentiate the *B. rapa* triplets. We used PhyML (GUINDON and GASCUEL 2003) to construct the phylogeny of the 3 *B. rapa* genes using the single *A. thaliana* gene as outgroup. We evaluated a total of 1655 trees with *B. rapa* triplets. A total of 952 (58%) trees had poor branch support (aLRT value <=0.8), suggesting that in most cases, the relationship among the *B. rapa* triplets are poorly resolved. Even among the trees that have good resolution on the splitting of the triplets, 243

trees have the "(I, II), III" topology, 203 trees have the "(I, III), II" topology and 257 trees show the "(II, III), I" topology. These counts do not favor a dominant topology (P=0.04, χ^2 – test, significance level 0.01).

In general, our findings are in agreement with previous results (WANG *et al.* 2011). On the sequence level, all three subgenomes appear equally diverged from *Arabidopsis thaliana*, as would be expected if the divergence between the Arabidopsis and Brassica genomes predated the triplication.

Deletions in *B. rapa* genes through comparison to the *A. thaliana* ortholog

In order to understand the mechanism underlying biased whole-gene removals in the three subgenomes, we asked whether there are differences in the rate of sequence removals within the gene sequences. We catalogued a list of sequence removal events based on pairwise comparisons between each *B. rapa* gene and its *A. thaliana* orthologs through an automated pipeline. Briefly, we listed the intervening gaps between the adjacent matching regions (HSPs) and checked whether the corresponding gap in *B. rapa* was substantially smaller than the corresponding gap in *A. thaliana* (**see Methods**) (**Figure 2**). In this study, we only focused on the deletions that are longer than 30 bases. Shorter deletions are likely affected by the artifacts of sequence alignments, so this arbitrary cutoff is a result of our favoring accuracy over sensitivity.

Using our automated deletion discovery pipeline, we identified a total of 4539 deletion sites of 3648 *B. rapa* genes examined (14.5% of *B. rapa* genes inspected in this study) (all deletion sites identified are available in **Supporting File S1**). Some *B. rapa* genes have experienced more than one deletion. Gap sizes ranged from 31 bases (just

above the computational cutoff) to 1363 bases, with the size distribution shown in **Figure 4**. There is an apparent excess of deletion sizes between 70-80 bases, in addition to the peak at smaller deletion size ranges.

Different parts of the genes have experienced different rates of deletion. Deletions in the intronic sequences were ~8 times more likely than in exonic deletions (1.63% of total intronic bases versus 0.24% of exonic bases; **Table 3**). 5-` or 3-` untranslated regions (UTRs) have incurred the fewest deletions, even fewer than in exons (0.04% of UTRs versus 0.24% of exonic bases), suggesting that some UTRs have functional roles and are under strong purifying selection. Inferred deletions that fall within sequences corresponding to Arabidopsis exons are likely to be the most reliable, and therefore are the types of deletions we used to investigate the mechanism of biased gene deletion.

Distribution of deletion sites among *B. rapa* homeologs

We anticipated that exonic deletions would be rarest in genes within subgenome III, and that genes in subgenome I would be the most likely to have gaps. Unexpectedly, we found that a higher proportion of genes in subgenome II had deletions versus subgenome I (7.9% versus 7.1%; **Table 4**). This is true whether we count the number of deletion sites or count the number of deleted bases. These data track observation of whole-gene fractionation bias among the three subgenomes, in that subgenomes I and II had more numbers of genes with deletions than subgenome III. However, subgenome II still had more genes with exonic deletions than expected, given the overall genome fractionation bias as discussed earlier.

We also observed differences of deletion frequencies in singlet, doublet, or triplet genes in *B. rapa*. Singlet genes contain significantly fewer deletions than doublets or triplets. A total of 6.4% versus 7.1% and 7.3% of the singlet, doublet, and triplet genes contain deletions, respectively (**Table 4**). This is consistent with different selection regimes on single-copy genes relative to genes with duplicate copies. Single-copy genes are expected to be under stronger purifying selection than genes with duplicate copies that can potentially buffer their functions. We further note that the real differences of the strength of purifying selection on singlet and duplicate genes might be larger than we have observed. The deletions we have counted include both the selectively neutral as well as deleterious deletions. Indeed, there is a background rate for neutral deletions, which are expected to be the same between singlet and duplicate genes. This background component in our deletion counts dilutes the signal reflecting only the purifying selection.

Direct repeats flanking the removed sequences

In maize, sequence deletions flanked by direct repeats (WOODHOUSE *et al.* 2010) are associated with the biased fractionation among homeologous regions. In *B. rapa*, about one third of exonic deletions were flanked by direct repeats, with length up to 19 bp (**see Methods**). Only one copy of the two original repeat units remained at the deletion site, probably as a direct result of the deletion mechanism (**Figure 5**). These data suggest that fractionation via small deletions occurs in *B. rapa* as it does in maize (WOODHOUSE *et al.* 2010), and may be a phenomenon general to plants.

Several instances of the flanking repeats are given in **Table 5**. Some repeats are low-complexity simple sequence repeats (SSRs), e.g. tri-nucleotide repeats $(GAT)_n$, $(TTC)_n$ (**Table 5**). SSRs have been shown to have high potential for illegitimate recombination of genes (ROCHA *et al.* 2002). Other repeat instances with higher nucleotide complexity are also present. Direct repeats are known to be hotspots of homologous recombination between the repeat units, making the intervening sequences more easily removed (**Figure 5C**).

Distribution of transposable-element related sequences

Any sort of mechanism that removes DNA in the genome could potentially be "induced" by a transposon bloom. Since we are testing the two-step model for paleohexaploid fractionation, a past transposon bloom could affect each subgenome in different ways. The deletion mechanism in plants has been hypothesized as an adaptation to fight "genetic obesity" (DEVOS *et al.* 2002).

Identification of the *Brassica rapa* interspersed repetitive elements followed published methods (WANG *et al.* 2011). Elements were categorized into classes, with LINEs, SINEs, LTRs, and DNA transposons being the largest classes (**Table 6**). Among all major classes of TEs, the distributions are not biased towards any subgenome, suggesting that the “background” insertion and removal rates are equal across the three subgenomes, at least when viewed as they exist today. None of the most abundant TE families with the large counts (>500 copies) across the genome showed any preference in a single subgenome (*P*-value cutoff of 0.01, χ^2-test).

DISCUSSION

Two-step genome merger model could explain the retention bias

The two-step model for paleohexaploidy formation and fractionation (LYONS *et al.* 2008) suggests that two of the genomes came together first, and then the third genome was added some time later (**Figure 6**). The common way to form a hexaploid is between a diploid (2N) and tetraploid (4N) cross resulting in a triploid which on doubling produces a hexaploid. If this were the case, two subgenomes could be in the same nucleus for a longer period of time (as a viable tetraploid) than the third, which is then relatively less fractionated than the first two. Additional support comes from the gene loss pattern between subgenome I and II, where low density regions of one of the two more fractionated genomes is compensated by less loss on the other, which indicates that the tetraploid genome (I+II) could be viable since most genes tend to have at least one copy (WANG *et al.* 2011).

We devised an experiment to test for the *B. rapa* genes that recently underwent fractionation. We reasoned that, if subgenomes I and II had been "at war" for a long time, then perhaps the non-dominant genome, I, had already lost nearly all the genes it could lose, so that, when III entered the fray, subgenome II would still have removable genes to be deleted. This is indeed what we observe – subgenome II is the one that has incurred the most exonic deletions, rather than subgenome I.

While the two-step model is the general expectation for the *formation* of a hexaploid, our model does require *fractionation* to also have occurred in two distinct steps. There are other alternatives that we have to rule out. For example, the two genome mergers might have occurred sequentially so there was no time to fractionate

after the first step. While this alternative is technically two-step formation, in terms of fractionation it is one-step. We rendered this hypothesis unlikely.

Rates of recent deletions are actively influenced by selection on gene functions

We found that deletions in the intronic sequences have occurred ~8 times more frequently than deletions in the exonic sequences (**Table 3**). Intuitively, corresponding exons are usually much more conserved in genome comparisons than the sequences of the introns, when exons are under strong selective pressure to retain coding capacity. Although the majority of the deletions occurred in the intronic sequences, there are still deletions in the exonic sequences that are tolerated by the *B. rapa* genes in our comparisons. Both functionally and in terms of technical reliability, exonic deletions are the more relevant to our study.

We consider the deletions in exonic sequences to be recent events, following the assumption that once a gene undergoes short deletions within exons, it is increasingly likely to be rendered nonfunctional, and more deletions will eventually accumulate until the gene sequence is no longer identifiable. Therefore, the presence of one or a few small deletions in the exonic sequences of any gene suggests that the deletion event was relatively recent.

On the whole-gene level, we found that singly retained genes contain fewer sequence removals than the doublets or triplets. In general, genes with duplicate copies show a higher likelihood of functional compensation than single-copy genes (Gu *et al.* 2003). However, there are many exceptions to this rule. Many

duplicate genes that have survived the sequence removal processes following polyploidy have diversified in their regulatory roles (TANG *et al.* 2010) and have acted as the hub or bottleneck enzymes of metabolic pathways (WU and QI 2010). One well-studied example is the FLC gene, for which all three copies are retained in the *B. rapa* genome (SCHRANZ *et al.* 2002). All three *B. rapa* homeologs are additive, in a dosage-dependent manner. We failed to find exonic deletions in any one of the three copies of the FLC homeologs in *B. rapa*, suggesting that the structures of the three copies are relatively intact. There is likely selective advantage in retaining multiple copies of those genes critical to the adaptation to the environment, e.g. flowering time, so that the actual dosage of the protein can be finely modulated.

Recent deletions in *B. rapa* homeologs are affected by a complex interplay of genome dominance

Our initial expectation was that the ratios of small deletion events per subgenome should match the gene fractionation ratio per subgenome. In maize, the deletion bias matched the fractionation bias between homeologs (WOODHOUSE *et al.* 2010). However, in *B. rapa*, the ongoing deletions do not closely reflect the pattern of gene loss: among the three subgenomes of *B. rapa*, subgenome II appears to have undergone more deletions than either subgenome I or subgenome III (**Table 4**).

Our proposed two-step fractionation hypothesis is capable of explaining this unexpected deviation of subgenome II from the general trend of fractionations. During the first genome merger, fractionation bias favored the retention of genes from one subgenome versus another, and then the more fractionated genome may have reached

its level of saturation for fractionation (subgenome I). When a new genome was introduced, the earlier, more retained genome (subgenome II) may have more fractionation in comparison to the new genome as well as the older, greatly fractionated genome (**Figure 6**). In this case, one would expect to see more deletions in the genome undergoing more recent fractionation. We would then hypothesize that the genome now undergoing the most fractionation is subgenome II. The two-step process involves a shifting role for subgenome II: after the first step, it is the dominant genome, but only until the second step, when subgenome III dominates both other subgenomes (I and II veteran subgenomes) (**Figure 6**).

Deletions, not point mutations, are the major mechanism for gene inactivation

We failed to observe consistent relationships among the three subgenomes using both the synonymous substitutions (*Ks*) and the gene tree method, indicating that base substitutions do not account for the fractionation biases or the three parental species are approximately equally diverged. In contrast, the deletion rates are distinctly different among the three subgenomes, suggesting that the deletions contribute more to the fractionation process than point mutations.

In mammals, the mode of gene inactivation appears to be a pseudogene pathway, which mostly involves point mutations that over time accumulate to render gene products non-functional (SCHRIDER *et al.* 2009). In the case of *B. rapa*, we favor the sequence removal model as we have observed the presence of

small deletions within exons that track the fractionation biases (more exonic deletions in subgenomes I and II than subgenome III), while in comparison, frequencies of point mutations among the three subgenomes are similar.

Sequence removals are likely facilitated by illegitimate recombination via direct repeats

We argue that fractionation in *B. rapa* appears to be due in part to a short deletion mechanism via illegitimate recombination, similar to previous observations in maize (WOODHOUSE *et al.* 2010). Direct repeats exhibit high levels of recombination intensity (ROCHA 2003). Bzymek and Lovett (2001) (BZYMEK and LOVETT 2001) proposed three major mechanisms for illegitimate recombination: simple replication slippage, sister-chromosome exchange-associated slippage, and single-strand annealing. Presence of the genomic repeats in proximity, e.g. simple di- and tri-nucleotide repeats, increases the likelihood of illegitimate recombination (**Table 5**).

Illegitimate recombination is not the only avenue for sequence removals. Interspersed repeats also tend to carry and transpose DNA segments. We could not find such a bias based on our scan for major types of repeat elements (**Table 6**), suggesting that the sequence removals – especially the "biased" removal patterns – are unlikely to be caused by interspersed repetitive elements, e.g. transposons or retrotransposons.

Cumulative small deletions set the ground for whole-gene removals

We simulated the process of generating the observed gap patterns. Our likelihood ratio test combined with simulations (**see Methods**) suggested that small deletion sizes

consisting of mostly single-gene and a small number of two-gene deletions are sufficient to generate the observed patterns. The pattern of predominantly one- or two-gene removals at a time is in concordance with the observation of small, internal deletion of genes, in which small chunks of sequences are removed per deletion event. However, when enough sequence removals have accumulated and/or removal of sequences containing critical functional domains have taken place, the entire gene function is compromised and thus more mutations will follow, since there is no purifying selection in place to protect against deleterious mutations. We therefore view the short deletion mechanisms as cumulative mutations that eventually resulted in whole-gene removals that have shaped the gene loss patterns we observe.

Alternative hypotheses that might explain genome dominance in *B. rapa*

Although we have support for the "two-stage" scenario through the observation that subgenome II contains the most exonic deletions, we do not exclude other possibilities that might also contribute to the current gene loss pattern among *B. rapa* subgenomes. For example, paleohexaploidy could have quickly taken place with no initial differences in fractionation, but the three genomes may instead have acquired different epigenetic marks, and these epigenetic differences included the fractionation differences observed in *B. rapa* today. "Genomic dominance" of subgenome D over subgenome A is clear in the allotetraploid cotton genome (FLAGEL and WENDEL 2010). Differential epigenetic marks resulting in the differential expression strengths of duplicate genes might correlate with gene retention favoring the homeolog with higher expression levels. Work in *Arabidopsis suecica* showed that homeologous gene loss is certainly correlated with

levels of expression and perhaps histone modifications as well (CHANG *et al.* 2010). Genomic and epigenomic changes directed toward one parental genome have also been observed in *Brassica napus* (in which the C-subgenome tended to be preferentially modified) (GAETA *et al.* 2007) and in Triticale polyploids (in which the rye subgenome tended to be preferentially modified) (MA and GUSTAFSON 2006). Biases in the strength and patterns of epigenetic modifications can lead to different selective constraints on the subgenomes. Such biases may be still ongoing in the modern-day Arabidopsis and maize genomes, and the bias is associated with differential epigenetic marks that result in differential expression levels between homeologs. Biased gene loss is the result of selection against the loss of the homeolog copy that has a higher expression value (which is more likely to compromise the biological function) (SCHNABLE and FREELING 2011).

In any case, there is nothing equal about the behavior of the three different genomes in *B. rapa*. The competing model which involves differential epigenetic marking might have impact on the subgenome differences we observe, which we hope to evaluate by studying the patterns of gene expression or histone modifications in *B. rapa* using high-throughput RNA-seq or CHIP-seq data.

ACKNOWLEDGEMENT

We appreciate financial support from the US National Science Foundation (MCB-0820821 to M.F.).

TABLES

Table 1. The number of genes and retained genes in each of the three subgenomes in *B. rapa* when compared to *A. thaliana*. The "Number of retained genes" in each subgenome is taken from Wang et al. (2011).

	Subgenome I	Subgenome II	Subgenome III	Arabidopsis
Genome span (Mb)	56.9	78.1	104.6	119.1
Number of genes	8890	11957	16838	27134
Gene density (genes/Mb)	156.3	153.0	160.9	227.8
Number of retained genes	5966	7679	11536	16423
Retained gene density (genes/Mb)	104.9	98.3	110.3	137.8
Percentage of genes retained (compared to Arabidopsis)	36%	46%	70%	100%

Table 2. “Horse race” *Ks* comparisons. The distances of two *B. rapa* genes to the *A. thaliana* reference gene can be compared to each other. For example, 1191 of “I - At > II - At” means that there are among all the I, II comparisons, 1191 of them showed higher *Ks* value of the gene in subgenome I than the gene in subgenome II. Asterisk (*) marks the *P*-value significant at $\alpha < 0.01$.

“Horse race” *Ks* comparisons	Counts	*P*-value (χ^2-test)
I - At > II - At	1191	0.489
I - At < II - At	1225	
II - At > III - At	2256	0.004 (*)
II - At < III - At	2067	
I - At > III - At	1678	0.027
I - At < III - At	1809	

Table 3. The locations of the deletion sites in *A. thaliana* genes identified through *A. thaliana* – *B. rapa* comparisons. Note that it is possible for some deletions to be situated across exon-intron boundaries.

	Number of deletions	Number of bases	Number of bases in *A. thaliana* genome	Percentage (%)
CDS	1863	78125	33050356	0.24
Introns	3345	319220	19590057	1.63
5`-UTR	27	1563	2610978	0.06
3`-UTR	49	2604	4442021	0.06

Table 4. Number of deletions and exonic deletions within *B. rapa* genes, grouped based on their subgenome assignments and copy numbers.

	Number of exonic deletions	Number of deleted exonic bases	Number of genes	Percentage of genes with exonic deletions (%)
Subgenome I	423	18155	5966	7.1
Subgenome II	609	27096	7679	7.9
Subgenome III	714	32874	11537	6.2
Singlet	594	28131	9252	6.4
Doublet	799	33752	10962	7.3
Triplet	353	16242	4968	7.1

Table 5. Partial list of instances of internal deletions within *B. rapa* genes that are flanked by direct repeats. This list only shows the flanking sequences that are 19 bases in length (an arbitrary number). See **Methods** for the notation of deletion identifiers.

Deletion ID	Left Flank	Right Flank
AT1G10570_Bra019923_Bite(32=>-19)	ggagaatttagtgtattga	agagaatttagtgtattga
AT1G47970_Bra018696_Bite(119=>-19)	gacgatgacgatgatgatg	gacgatgatgatgaggatg
AT1G52870_Bra018995_Bite(104=>-19)	tttgatgttgcttgagtga	tttgatgttgcttgagtga
AT2G21560_Bra030293_Bite(44=>-19)	cttttcttcagtgctctgt	cttatcttcaatgctctgt
AT2G44160_Bra004810_Bite(44=>-19)	ttaatgtagataccagctg	ttaatgtagataccagttg
AT3G03590_Bra031985_Bite(298=>-19)	tgtaaagactctaagcaaa	tgtgaagactctaaacaaa
AT3G49140_Bra018005_Bite(81=>-19)	aacctcagtcattctcttt	aacctcagtcattctcttt
AT3G51260_Bra036824_Bite(43=>-19)	catgttctataactaaacc	aatgttctataactaaacc
AT4G02880_Bra018525_Bite(50=>-19)	tgaaaatagtgatgccgag	tgaaaatggtgatccagag
AT4G18430_Bra012597_Bite(99=>-19)	atctagtcaaatattatat	atctagttaatattatatt
AT5G40120_Bra025619_Bite(89=>-19)	gttaatgcagcaggagctt	gttgatgcagcatgaactt
AT5G46740_Bra025000_Bite(62=>-19)	cagcaaatggcttctcaga	cagcaaatggtttctcaga
AT5G61150_Bra029332_Bite(101=>-19)	ttcctcttcttcatcttca	ttcttcctcttcttcttcc

Table 6. The number of major classes of transposable elements in three subgenomes.

Type	Total counts	Frequency in subgenomes (counts per Mb)		
		I	II	III
LTR	62510	239	248	241
LINE/L1	28215	109	112	108
LTR/Copia	24809	94	100	95
LTR/Gypsy	17004	69	70	61
DNA/hAT-Ac	13715	54	54	52
DNA/En-Spm	10438	41	42	39
SINE	10232	40	39	40
DNA/MuDR	8950	35	35	34
DNA/Harbinger	4990	18	21	19
DNA/TcMar-Pogo	4983	18	20	19
DNA	4701	18	19	18
DNA/hAT	4542	18	18	17
RC/Helitron	3190	11	13	12
LINE/Penelope	3096	11	12	12
DNA/TcMar-Stowaway	2674	11	10	10
DNA/hAT-Tag1	2262	8	10	9

FIGURE LEGENDS

Figure 1. (**A**) Dot plot between *B. rapa* and *A. thaliana*, with *B. rapa* segments that are derived from the same *A. thaliana* origin grouped together, to illustrate the partitioning and test of non-random fractionation among *B. rap*a triplicated regions. This is only showing one of the 24 sets of blocks (block *R*). The table under the dot plot contains the counts of *A. thaliana – B. rapa* orthologs in the respective subgenomes. Gene losses are not equally distributed in most of the duplicated blocks, as tested by χ^2-test ($P=1 \times 10^{-28}$ in the case of block *R*). (**B**) The partitioning of *B. rapa* chromosomes into three inferred subgenomes following the partitioning algorithm in (WANG *et al.* 2011).

Figure 2. The pipeline for automated deletion discovery. We first compared between *A. thaliana* and *B. rapa* orthologous genes using BLASTN. From the initial BLASTN HSPs, we computed a set of collinear HSPs. The un-matching regions in *A. thaliana* and *B. rapa* are compaired in a pairwise fashion, recording the sizes of the corresponding gaps in *A. thaliana* and *B. rapa*, in a notation of "Bite (A=>B)". We only selected the deletions that have A > 30 bp and B < 10 bp, to screen for substantial downsizing in the *B. rapa* sequence. As examples, the bites in black color are selected on the basis of these criteria whereas the gray ones are ignored.

Figure 3. *Ks* distribution between *A. thaliana – B. rapa* orthologs and *B. rapa – B. rapa* homeologs. Solid lines are the observed distribution, dashed lines are the fitted distribution based on log-normal distribution (TANG *et al.* 2008).

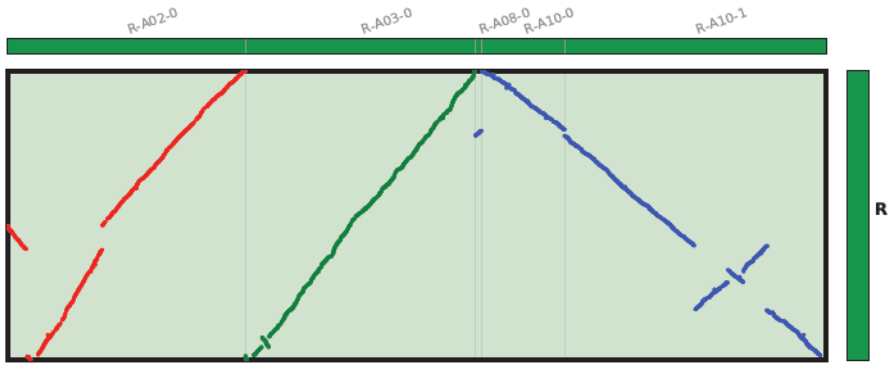
Figure 4. Size distribution of the deletions in *B. rapa* genes (sequences present in *A. thaliana* but removed in *B. rapa*) that we catalogued in this study. The distribution stops at 30 since we only focus on the deletions that are > 30 bases (a computational cutoff).

Figure 5. (**A**) A GEvo graphic of BLASTN output between orthologous genes in *A. thaliana* and three *B. rapa* homeologs. The first panel is a region in *A. thaliana* and used as the reference, the following three panels are three *B. rapa* regions that were derived from the recent hexaploidy event. Arrows represent gene models and colored rectangles show the extents of BLASTN matches (HSPs). The colored rectangles (pink, tan, brown) represent high-scoring sequence pairs (HSPs), or regions with high sequence similarity to each other. *A. thaliana* is the reference sequence (top panel). As can be seen, a gap is evident (blue arrow) when comparing the HSPs of *AT1G68590* and *Bra038364* (bottom panel). The deleted sequence (circled in blue) is evident in comparison to other *B. rapa* homeologs. An overlap between the HSP blocks that flank the deletion can be seen (blue circle); this indicates that the pre-deleted sequence was flanked by direct repeats. To reproduce this analysis, go to (http://genomevolution.org/r/rmi). (**B**) ClustalW alignment of the *A. thaliana* and the three *B. rapa* sequences from (A). The sequences in the blue box in the whole *B. rapa* homeologs indicate the locations of the direct repeat sequence that originally flanked the deletion in the homeolog containing the deletion (*Bra038364*). (**C**) Proposed mechanism for the within-gene deletion via intra-chromosomal illegitimate recombination.

Figure 6. The proposed “two-step” model of genome mergers. First genome I and genome II form a tetraploid, subsequent addition of genome III formed the hexaploid. Such step-wise genome additions involve shifting roles of the “dominant” genome: in the formation of tetraploid, subgenome II was the dominant genome, whereas in the hexaploid, subgenome III became the new dominant genome.

Figure 1

A



χ^2 −test of the number of orthologs in each sub−genome (P =1e−28)

	R−I	R−II	R−III
Observed	685	687	1083
Expected	818	818	818

B.

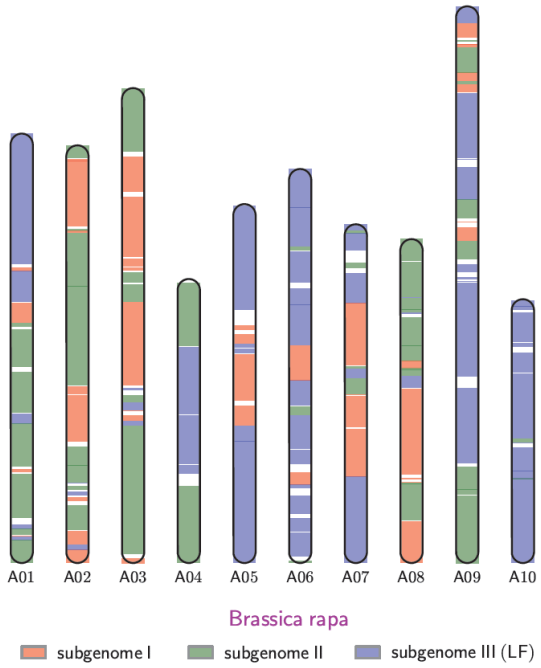


Figure 2.

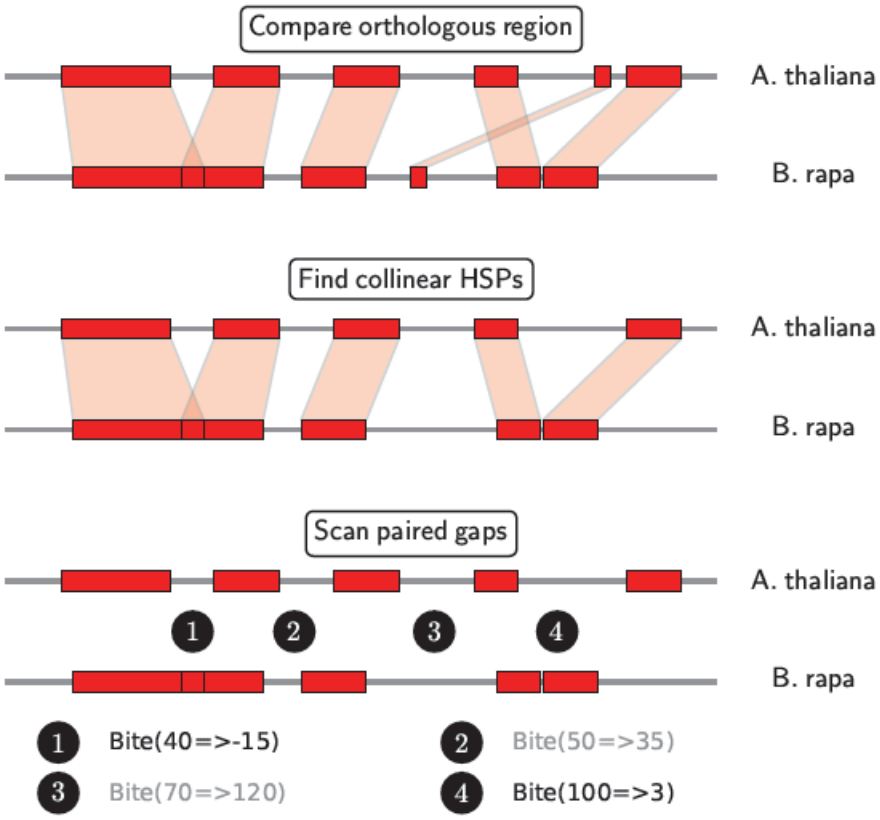

Compare orthologous region
A. thaliana
B. rapa
Find collinear HSPs
A. thaliana
B. rapa
Scan paired gaps
A. thaliana
1
2
3
4
B. rapa
1 Bite(40=>-15)
2 Bite(50=>35)
3 Bite(70=>120)
4 Bite(100=>3)

Figure 3.

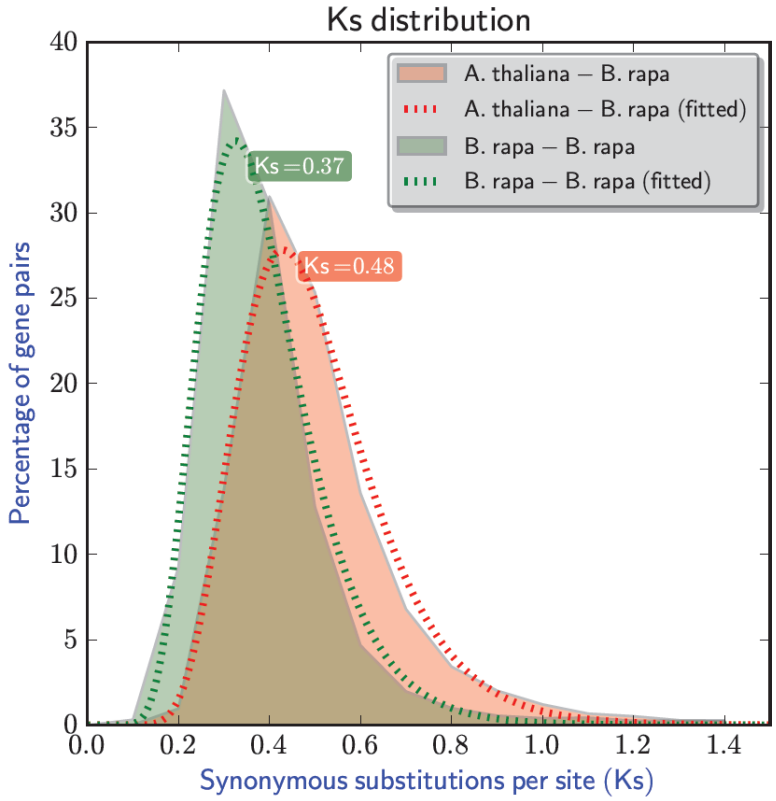

Ks distribution
A. thaliana – B. rapa
A. thaliana – B. rapa (fitted)
B. rapa – B. rapa
B. rapa – B. rapa (fitted)
Ks = 0.37
Ks = 0.48
Percentage of gene pairs
Synonymous substitutions per site (Ks)
0
5
10
15
20
25
30
35
40
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Figure 4.

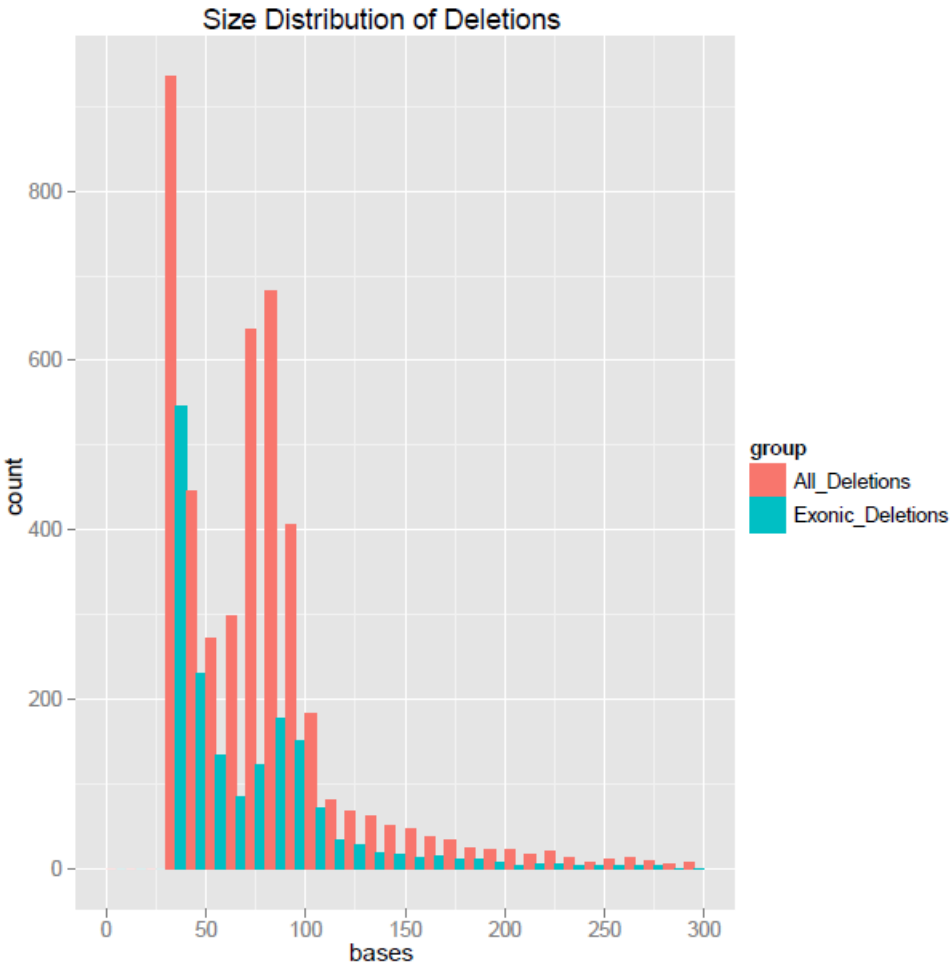
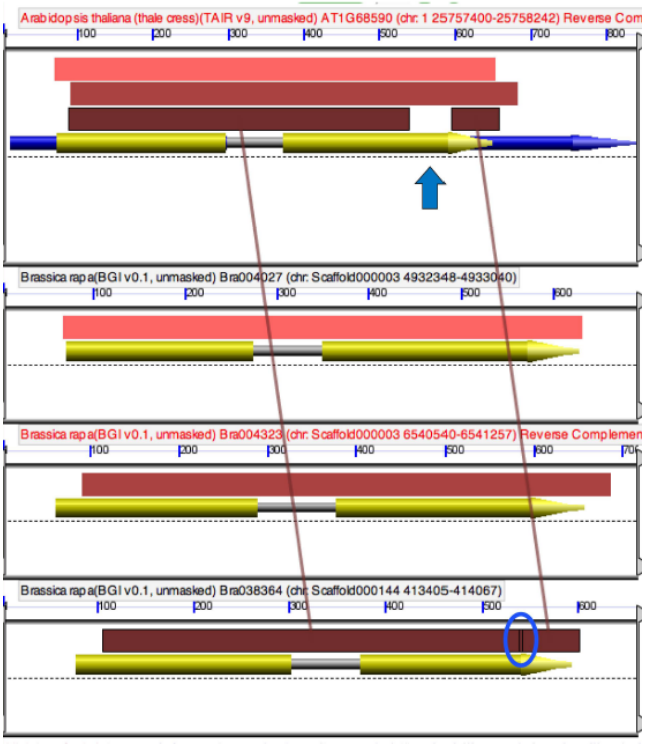
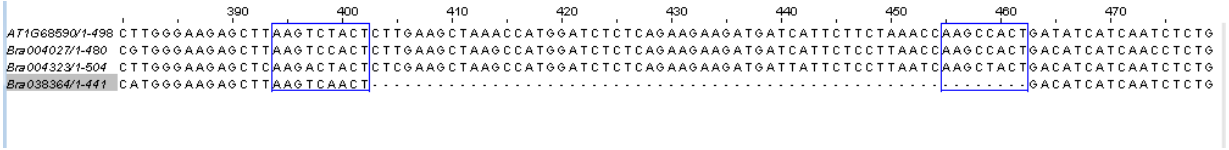

Size Distribution of Deletions
count
800
600
400
200
0
0
50
100
150
200
250
300
bases
group
All_Deletions
Exonic_Deletions

Figure 5

A


Arabidopsis thaliana (thale cress)(TAIR v9, unmasked) AT1G68590 (chr: 1 25757400-25758242) Reverse Com
Brassica rapa(BGI v0.1, unmasked) Bra004027 (chr: Scaffold000003 4932348-4933040)
Brassica rapa(BGI v0.1, unmasked) Bra004323 (chr: Scaffold000003 6540540-6541257) Reverse Complemen
Brassica rapa(BGI v0.1, unmasked) Bra038364 (chr: Scaffold000144 413405-414067)

B


390 400 410 420 430 440 450 460 470
AT1G68590/1-498 CTTGGGAAGAGCTTAAGTCTACTCTTGAAGCTAAACCATGGATCTCTCAGAAGAAGATGATCATTCTTCTAAACCAAGCCACTGATATCATCAATCTCTG
Bra004027/1-480 CGTGGGAAGAGCTTAAGTCCACTCTTGAAGCTAAGCCATGGATCTCTCAGAAGAAGATGATCATTCTCCTTAACCAAGCCACTGACATCATCAACCTCTG
Bra004323/1-504 CTTGGGAAGAGCTCAAGACTACTCTCGAAGCTAAGCCATGGATCTCTCAGAAGAAGATGATTATTCTCCTTAATCAAGCTACTGACATCATCAATCTCTG
Bra038364/1-441 CATGGGAAGAGCTTAAGTCAACT----------GACATCATCAATCTCTG

C

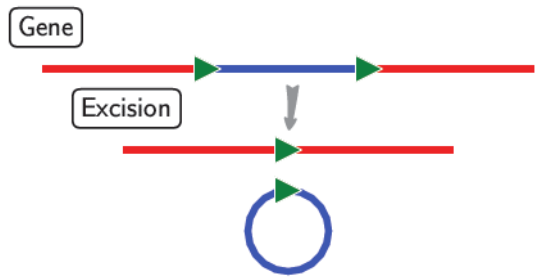
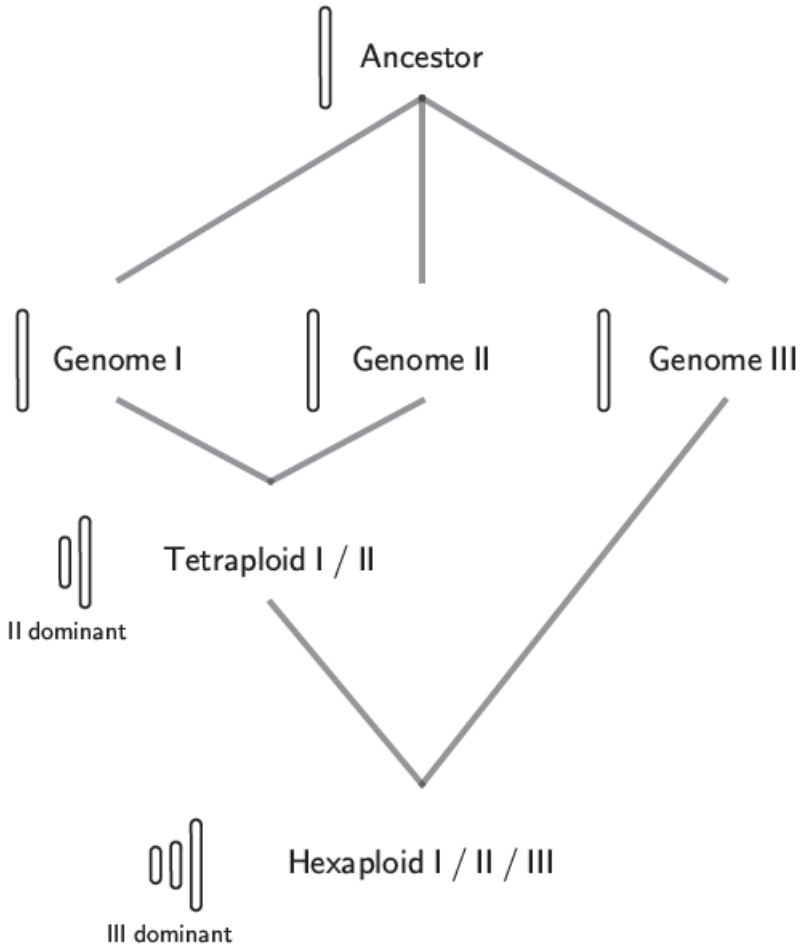

Gene
Excision

Figure 6.


Ancestor
Genome I
Genome II
Genome III
Tetraploid I / II
II dominant
Hexaploid I / II / III
III dominant

LITERATURE CITED

AURY, J. M., O. JAILLON, L. DURET, B. NOEL, C. JUBIN *et al.*, 2006 Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature **444:** 171-178.

BENNETT, M. D., and I. J. LEITCH, 2005 Nuclear DNA amounts in angiosperms: progress, problems and prospects. Ann Bot **95:** 45-90. BIRCHLER, J. A., and R. A. VEITIA, 2010 The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol **186:** 54-62.

BOWERS, J. E., B. A. CHAPMAN, J. RONG and A. H. PATERSON, 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438.

BUGGS, R. J., S. CHAMALA, W. WU, L. GAO, G. D. MAY *et al.*, 2010a Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. Mol Ecol **19 Suppl 1:** 132-146.

BUGGS, R. J., N. M. ELLIOTT, L. ZHANG, J. KOH, L. F. VICCINI *et al.*, 2010b Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. New Phytol **186:** 175-183.

BZYMEK, M., and S. T. LOVETT, 2001 Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc Natl Acad Sci U S A **98:** 8319-8325.

CHANG, P. L., B. P. DILKES, M. MCMAHON, L. COMAI and S. V. NUZHDIN, 2010 Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. Genome Biol **11:** R125.

DEVOS, K. M., J. K. BROWN and J. L. BENNETZEN, 2002 Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res **12:** 1075-1079.

FLAGEL, L. E., and J. F. WENDEL, 2010 Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. New Phytol **186:** 184-193.

FORCE, A., M. LYNCH, F. B. PICKETT, A. AMORES, Y. L. YAN *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531-1545.

GAETA, R. T., J. C. PIRES, F. INIGUEZ-LUY, E. LEON and T. C. OSBORN, 2007 Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell **19:** 3403-3417.

GU, Z. L., L. M. STEINMETZ, X. GU, C. SCHARFE, R. W. DAVIS *et al.*, 2003 Role of duplicate genes in genetic robustness against null mutations. Nature **421:** 63-66.

GUINDON, S., and O. GASCUEL, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:** 696-704.

JAILLON, O., J. M. AURY, F. BRUNET, J. L. PETIT, N. STANGE-THOMANN *et al.*, 2004 Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature **431:** 946-957.

JIAO, Y., N. J. WICKETT, S. AYYAMPALAYAM, A. S. CHANDERBALI, L. LANDHERR *et al.*, 2011 Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97-100.

KELLIS, M., B. W. BIRREN and E. S. LANDER, 2004 Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428:** 617-624.

KURTZ, S., A. PHILLIPPY, A. L. DELCHER, M. SMOOT, M. SHUMWAY *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol **5:** R12.

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.

LAGERCRANTZ, U., and D. J. LYDIATE, 1996 Comparative genome mapping in Brassica. Genetics **144:** 1903-1910.

LANGHAM, R. J., J. WALSH, M. DUNN, C. KO, S. A. GOFF *et al.*, 2004 Genomic duplication, fractionation and the origin of regulatory novelty. Genetics **166:** 935-945.

LARKIN, M. A., G. BLACKSHIELDS, N. P. BROWN, R. CHENNA, P. A. MCGETTIGAN *et al.*, 2007 Clustal W and Clustal X version 2.0. Bioinformatics **23:** 2947-2948.

LEITCH, A. R., and I. J. LEITCH, 2008 Genomic plasticity and the diversity of polyploid plants. Science **320:** 481-483.

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

LYONS, E., B. PEDERSEN, J. KANE and M. FREELING, 2008 The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. Tropical Plant Biology **1:** 181-190.

LYSAK, M. A., M. A. KOCH, A. PECINKA and I. SCHUBERT, 2005 Chromosome triplication found across the tribe Brassiceae. Genome Res **15:** 516-525.

MA, X. F., and J. P. GUSTAFSON, 2006 Timing and rate of genome variation in triticale following allopolyploidization. Genome **49:** 950-958.

PARKIN, I. A., S. M. GULDEN, A. G. SHARPE, L. LUKENS, M. TRICK *et al.*, 2005 Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. Genetics **171:** 765-781.

PARKIN, I. A., A. G. SHARPE and D. J. LYDIATE, 2003 Patterns of genome duplication within the Brassica napus genome. Genome **46:** 291-303.

PROOST, S., P. PATTYN, T. GERATS and Y. VAN DE PEER, 2011 Journey through the past: 150 million years of plant genome evolution. Plant J **66:** 58-65.

QUINLAN, A. R., and I. M. HALL, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26:** 841-842.

ROCHA, E. P., 2003 An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res **13:** 1123-1132.

ROCHA, E. P., I. MATIC and F. TADDEI, 2002 Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? Nucleic Acids Res **30:** 1886-1894.

SANKOFF, D., C. ZHENG and Q. ZHU, 2010 The collapse of gene complement following whole genome duplication. BMC Genomics **11:** 313.

SCHNABLE, J. C., and M. FREELING, 2011 Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. PLoS One **6:** e17855.

SCHNABLE, J. C., Y. KIM, S. SUBRAMANIAM, H. TANG, G. TURCO *et al.*, 2011a Genome dominance in plants and gene regulatory consequences of whole genome duplication. Submitted.

SCHNABLE, J. C., N. M. SPRINGER and M. FREELING, 2011b Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A **108:** 4069-4074.

SCHRANZ, M. E., P. QUIJADA, S. B. SUNG, L. LUKENS, R. AMASINO *et al.*, 2002 Characterization and effects of the replicated flowering time gene FLC in Brassica rapa. Genetics **162:** 1457-1468.

SCHRIDER, D. R., J. C. COSTELLO and M. W. HAHN, 2009 All human-specific gene losses are present in the genome as pseudogenes. J Comput Biol **16:** 1419-1427.

STERCK, L., S. ROMBAUTS, K. VANDEPOELE, P. ROUZÉ and Y. VAN DE PEER, 2007 How many genes are there in plants (... and why are they there)? Current Opinion in Plant Biology **10:** 199-203.

SUYAMA, M., D. TORRENTS and P. BORK, 2006 PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res **34:** W609-612.

TANG, H., J. E. BOWERS, X. WANG and A. H. PATERSON, 2010 Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci U S A **107:** 472-477.

TANG, H., X. WANG, J. E. BOWERS, R. MING, M. ALAM *et al.*, 2008 Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res **18:** 1944-1954.

THOMAS, B. C., B. PEDERSEN and M. FREELING, 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., J. A. FAWCETT, S. PROOST, L. STERCK and K. VANDEPOELE, 2009 The flowering world: a tale of duplications. Trends Plant Sci **14:** 680-688.

WANG, J., L. TIAN, H. S. LEE, N. E. WEI, H. JIANG *et al.*, 2006 Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. Genetics **172:** 507-517.

WANG, X., H. WANG, J. WANG, R. SUN, J. WU *et al.*, 2011 The genome of the mesopolyploid crop species Brassica rapa. Nat Genet.

WOLFE, K. H., 2001 Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet **2:** 333-341.

WOODHOUSE, M. R., J. C. SCHNABLE, B. S. PEDERSEN, E. LYONS, D. LISCH *et al.*, 2010 Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol **8:** e1000409.

WU, X., and X. QI, 2010 Genes encoding hub and bottleneck enzymes of the Arabidopsis metabolic network preferentially retain homeologs through whole genome duplication. BMC Evol Biol **10:** 145.

XIONG, Z., R. T. GAETA and J. C. PIRES, 2011 Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. Proc Natl Acad Sci U S A **108:** 7908-7913.

YANG, T. J., J. S. KIM, S. J. KWON, K. B. LIM, B. S. CHOI *et al.*, 2006 Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of Brassica rapa. Plant Cell **18:** 1339-1347.

YANG, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24:** 1586-1591.