the plant journal
SEB Society for Experimental Biology

The Plant Journal (2017) doi: 10.1111/tpj.13541

Selective modes determine evolutionary rates, gene compactness and expression patterns in *Brassica*

Yue Guo[1], Jing Liu[1], Jiefu Zhang[2], Shengyi Liu[3] and Jianchang Du[1,2,3,*]

[1] *Provincial Key Laboratory of Agrobiology, Institute of Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China,*
[2] *Key Laboratory of Cotton and Rapeseed, Ministry of Agriculture of People's Republic of China, Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China, and*
[3] *Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture of People's Republic of China, Oil Crops Research Institute, Chinese Academy of Agricultural Sciences, Wuhan 430062, China*

Received 28 September 2016; revised 28 February 2017; accepted 15 March 2017.
*For correspondence (e-mail dujianchang@hotmail.com).

SUMMARY

It has been well documented that most nuclear protein-coding genes in organisms can be classified into two categories: positively selected genes (PSGs) and negatively selected genes (NSGs). The characteristics and evolutionary fates of different types of genes, however, have been poorly understood. In this study, the rates of nonsynonymous substitution (K_a) and the rates of synonymous substitution (K_s) were investigated by comparing the orthologs between the two sequenced Brassica species, *Brassica rapa* and *Brassica oleracea*, and the evolutionary rates, gene structures, expression patterns, and codon bias were compared between PSGs and NSGs. The resulting data show that PSGs have higher protein evolutionary rates, lower synonymous substitution rates, shorter gene length, fewer exons, higher functional specificity, lower expression level, higher tissue-specific expression and stronger codon bias than NSGs. Although the quantities and values are different, the relative features of PSGs and NSGs have been largely verified in the model species Arabidopsis. These data suggest that PSGs and NSGs differ not only under selective pressure (K_a/K_s), but also in their evolutionary, structural and functional properties, indicating that selective modes may serve as a determinant factor for measuring evolutionary rates, gene compactness and expression patterns in Brassica.

Keywords: *Brassica*, nonsynonymous substitution, positive selection, purifying selection, synonymous substitution.

INTRODUCTION

It is well established that almost all the protein-coding genes in organisms can be classified into three categories: positively selected genes (PSGs); negatively (or purifying) selected genes (NSGs); and neutral genes (Yang, 2005). Differentiating gene type can be achieved by comparing the rates of nonsynonymous substitution (K_a) with the rates of synonymous substitution (K_s). $K_a/K_s > 1$ (also denoted as ω) is potential evidence for positive selection during gene sequence divergence, and $K_a/K_s < 1$ usually indicates that negative selection has acted (Zhang *et al.*, 2002). Sometimes $K_a/K_s = 1$ has been detected, thus possibly providing evidence of neutral evolution in some functional genes.

The evolutionary forces underlying synonymous and nonsynonymous substitution rates remain unclear. Nevertheless, they have been previously found to be associated with gene type (Nei, 1987), physical location of the gene (Wolfe *et al.*, 1989), guanine-cytosine (GC) content (Ticher and Graur, 1989) and mutation rates (Wolfe *et al.*, 1989; Matassi *et al.*, 1999). Nonsynonymous substitution rates are also found to correlate with expression patterns, indicating that the levels and patterns of expression may be major determinants of K_a (Yang and Gaut, 2011). Our recent comparisons of rate variations between the orthologs of *Glycine max* and its progenitor species *Glycine soja* show that K_a in recombination-suppressed pericentromeric regions is significantly lower than their homologs in chromosomal arms, indicating that chromatin environments can also serve as a determinant factor for K_a (Du *et al.*, 2012).

To estimate the relative gene frequencies under different selection patterns, the rate of divergence between *Drosophila melanogaster* and its close relatives has been performed (Coulthart and Singh, 1988; Civetta and Singh, 1995; Swanson *et al.*, 2001). These efforts have revealed that reproductive proteins evolved much faster than nonreproductive-related proteins (Coulthart and Singh, 1988; Civetta and Singh, 1995), and that >10% of male reproductive proteins have a ratio of $K_a/K_s > 1$ (Swanson *et al.*, 2001), the best explanation for which can only be explained by positive selection (Tsaur and Wu, 1997; Begun *et al.*, 2000). Similar conclusions have also been drawn in primates (Wyckoff *et al.*, 2000). Other fast-evolving genes (i.e. strong candidates for positive selection), include the mammalian olfactory receptor (Buck and Axel, 1991), major histocompatibility complex (MHC) class-I and -II genes (Klein and Figueroa, 1985; Hughes and Nei, 1988) and immune system genes (Hughes, 1999). Generally, the overall PSGs comprise only a small part of the whole gene set in a genome, ranging from 0.5 to 5.3% (Fay and Wu, 2001). This is reasonable because the majority of mutations in nonsynonymous sites are believed to be deleterious to individuals, and therefore most of them will be quickly lost from the population (Li, 1997), leading to a much lower K_a value and therefore lower K_a/K_s ratio. It also at least partially explains why most genes, no matter which genome they belong to, were always being detected to have evolved under purifying/negative selection (Hurst, 2002; Kondrashov *et al.*, 2002; Nielsen *et al.*, 2005; Nei *et al.*, 2010).

Genome-wide analysis of the rates of gene variation in plants, however, is quite limited compared with those in animal studies. This is partially because almost all higher plants have undergone one or more rounds of genome duplication (polyploidization; Jiao *et al.*, 2011), and the true orthologous relationships are difficult to identify (Zhang *et al.*, 2002). To date, several studies regarding evolutionary rates have been performed using the orthologous genes between *Arabidopsis thaliana* and *Arabidopsis lyrata* (Beilstein *et al.*, 2010; Yang and Gaut, 2011), and between *G. max* and *G. soja* (Du *et al.*, 2012). These efforts have facilitated our understanding of the values and the distribution of evolutionary rates, the expression patterns of genes and the underlying mechanisms of rate variation. Nevertheless, the rates, gene features and expression patterns of genes under different selective modes (here, referring to positive selection and negative selection) have not yet been investigated. Therefore, the consequences and evolutionary fates of different types of genes are largely unknown.

The mesopolyploid crop species Brassica belongs to the Brassicaceae family and is one of the most economically important genera, with approximately 100 species, including many important vegetable crops (Labana and Gupta, 1993). Of particular interest, the Brassica genus contains six representative diploid species: *B. rapa* (AA, $2n = 20$), *Brassica nigra* (BB, $2n = 16$), *B. oleracea* (CC, $2n = 18$), and their allotetraploid progeny species *Brassica napus* (AACC, $2n = 38$), *Brassica carinata* (BBCC, $2n = 34$) and *Brassica juncea* (AABB, $2n = 36$) (Beilstein *et al.*, 2006). The origins and relationships of these six species were thoroughly investigated by the 'triangle of U' (Nagaharu, 1935; Wang *et al.*, 2011). It seems clear now that the Brassica lineage diverged from *A. thaliana* about 20 million years ago (20 MYA; Lysak *et al.*, 2005; Town *et al.*, 2006; Yang *et al.*, 2006; Lysak *et al.*, 2007; Mun *et al.*, 2009; Zhao *et al.*, 2013). Another feature of Brassica genomes is that these genomes not only share the three paleo-polyploidy events (Bowers *et al.*, 2003), but most of them have also undergone an additional whole-genome triplication (WGT), which was thought to have occurred 13–17 MYA (Yang *et al.*, 1999; Town *et al.*, 2006; Zhao *et al.*, 2013). Because of their agronomic and evolutionary importance, Brassica species have attracted much attention from many researchers, and often serve as a modeling system for the study of genome evolution (Song *et al.*, 1995; Johnston *et al.*, 2005; Ge *et al.*, 2009). As a result, the *B. rapa* (Wang *et al.*, 2011), *B. oleracea* (Liu *et al.*, 2014) and *B. napus* (Chalhoub *et al.*, 2014) genomes have been sequenced recently. The release of these genome sequences provides an unprecedented opportunity to investigate evolutionary changes between different Brassica species and answer some basic evolutionary questions.

The purpose of this study is to compare the differences of PSGs and NSGs in evolutionary rates, gene compactness, gene retention capability and expression patterns. Thus, the data from this study may provide strong evidence that different selective modes contribute to the outcomes and evolutionary fates for different types of genes. To do this, we initially identified 23 817 orthologous gene pairs between *B. rapa* and *B. oleracea*, and then calculated the K_a, K_s and K_a/K_s for each gene pair. Our data show that PSGs exhibit significantly higher K_a values, lower K_s values, smaller gene size, lower expression level and higher tissue specificity than NSGs. Our data also indicate that PSGs have the tendency to preserve more triplication copies than NSGs, suggesting that PSGs may play an important role in gene retention capability during the early stage after the Brassica WGT event.

RESULTS

The distribution of K_a, K_s and K_a/K_s, and their correlations in Brassica

We first obtained 24 219 orthologous gene pairs between *B. rapa* and *B. oleracea*, and then discarded 402 genes with $K_s > 0.3$, because genes with such a high K_s value may imply either potential sequence saturation or misalignment

(Yang, 2007). After discarding the 402 genes, the data set contained 23 817 orthologous gene pairs, which were further analyzed.

For each gene pair, K_a, K_s and K_a/K_s (ω) were calculated (Figures 1 and S1; Tables S1 and S2). Overall, K_a ranged from 0 to 0.15, with a mean of 0.017 (Figure 1a; Table S2). The K_s estimates ranged from 0 to 0.3, with a mean of 0.085 (Figure 1b; Table S2). The K_a/K_s ratio had an average mean of 0.271, and 90% of K_a/K_s estimates fell within the range of 0.023–0.790 (Figure 1c; Table S2). The K_a and K_s values in Brassica are both significantly lower (one-sided Mann-Whitney U test, $P < 10^{-22}$; Table S2) than in Arabidopsis, but the K_a/K_s in Brassica is significantly higher than in Arabidopsis (one-sided Mann-Whitney U test, $P << 1$; Figure S2; Table S2; Yang and Gaut, 2011). These data indicate that the genes in Brassica may evolve at a lower evolutionary rate and under lower selective pressure than in Arabidopsis. It is a reasonable hypothesis because *B. rapa* and *B. oleracea* are important vegetable crops, and individuals with higher mutation rates are not allowed and will not be preserved by breeders. In contrast, Arabidopsis grows naturally and widely, and therefore it can bear more mutations than Brassica.

The relationships of K_a, K_s and K_a/K_s in Brassica were then established. As shown in Figure 1 and Table S3, K_a increases gradually with the increase of K_s (Spearman's rank correlation $r = 0.14$, $P < 10^{-22}$). This observation is basically consistent with that in Arabidopsis ($r = 0.21$; Yang and Gaut, 2011) and in soybean ($r = 0.22$; Du *et al*., 2012), although the degree of correlation is slightly different,

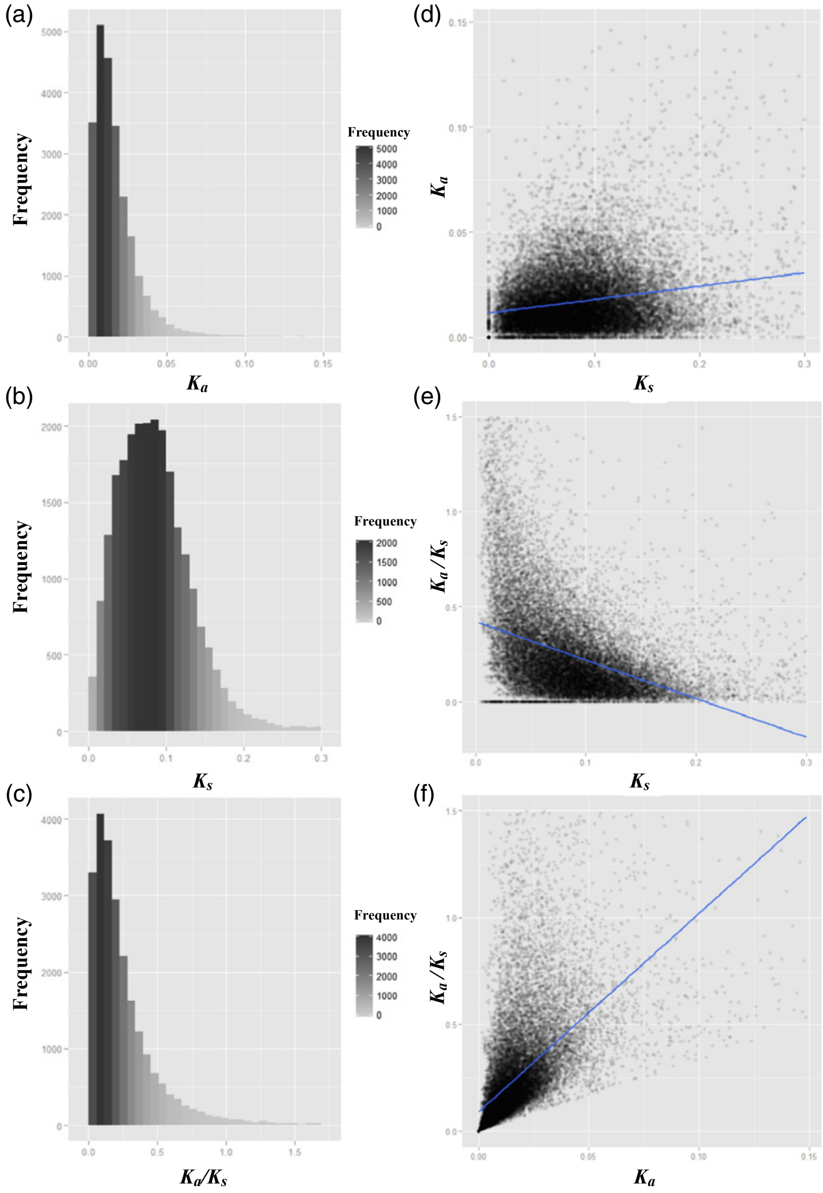

Figure 1. The frequency distributions and correlation analyses of K_a, K_s and K_a/K_s in Brassica. (a–c) The frequency displays of K_a, K_s and K_a/K_s, respectively. (d) The correlation between K_s (*x*-axis) and K_a. (e) The correlation between K_s (*x*-axis) and K_a/K_s. (f) The correlation between K_a (*x*-axis) and K_a/K_s. [Colour figure can be viewed at wileyonlinelibrary.com]

indicating that the mechanisms affecting both synonymous and nonsynonymous sites may be shared in different genomes. As expected, the K_a/K_s ratio was linked to both K_a and K_s (Figure 1; Table S3). The K_a/K_s ratio was highly negatively correlated with K_s ($r = 0.45$, $P < 10^{-22}$), and strongly positively correlated with K_a ($r = 0.76$, $P < 10^{-22}$). This piece of data may indicate that K_a is a determinant factor for K_a/K_s, and K_s can also affect K_a/K_s, even without changes in K_a.

Based on the ratio of K_a to K_s, 23 018 out of 23 817 gene pairs could be classified into three categories: 698 positively selected genes ($K_a/K_s > 1$); 22 309 negatively selected genes; and 11 neutrally evolved genes. The remaining 810 genes could not be classified further based on this criterion because either K_a (633 genes) or K_s (137 genes), or both (29 genes), are zero (Figure S1; Table S1). Therefore, these 810 genes may represent a specific type of gene set in the Brassica genome. Nevertheless, these genes may also evolve under negative selection ($K_a = 0$; $K_s \neq 0$), positive selection ($K_a \neq 0$; $K_s = 0$) or strongly negative selection ($K_a = K_s = 0$), because they are subject to strong constraints.

Higher K_a and lower K_s for PSGs than for NSGs

In order to understand the differences in evolutionary rates between PSGs and NSGs, the K_a and K_s values for each different gene set were calculated separately. As shown in Figure 2 and Table 1, the average value of K_a for PSGs in Brassica is 0.0278, significantly higher than the average value of NSGs of 0.0138 ($P < 10^{-145}$). In contrast, the overall K_s value for PSGs is 0.0201, fourfold lower than the overall value of NSGs of 0.0816 ($P < 10^{-300}$). To further verify the much lower K_s values for PSGs, we first randomly selected five known genes from *B. rapa*, *B. olerocea* and *A. thaliana*, and aligned their orthologous gene sequences (Table S4). The data show that the gene frames and the sequence alignments are largely correct, indicating that the lower K_s values for PSGs may not be artificial (Table S4). We then calculated the K_a and K_s values for two gene sets with and without unknown genes (Table S5). It is obvious that the values of K_a and K_s do not vary greatly, indicating that much lower K_s values for PSGs might also not be caused by annotation mistakes (Table S5). Therefore, it is not surprising to see that the higher K_a and much lower K_s values together make the K_a/K_s of PSGs (1.3472) much higher than that of NSGs (0.1887) (Table 1).

To further compare the evolutionary rates between PSGs and NSGs, the distributions of K_a and K_s were investigated (Figure 3). Although PSGs and NSGs both have a K_a peak, the peak of PSGs is much higher than that of NSGs (0.03 versus 0.01; Figure 3a). In contrast, the majority of the K_s

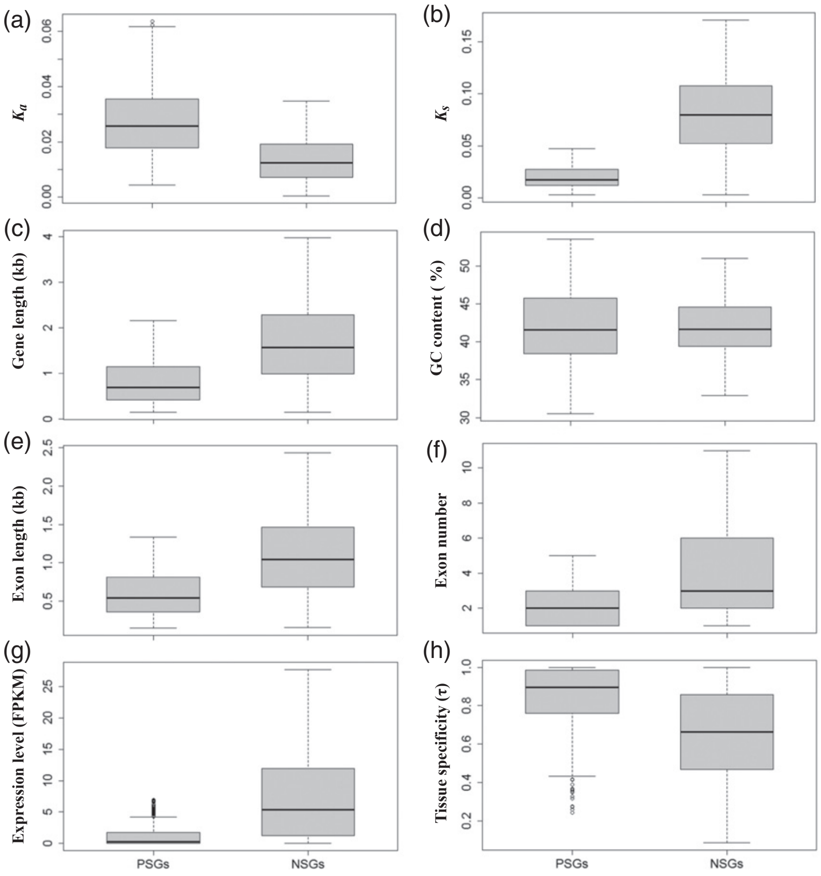

Figure 2. Comparisons of genomic features between positively selected genes (PSGs) and negatively selected genes (NSGs) in Brassica. The line in the box is the median value, and the lines at the bottom and top of each box are the first (lower) and third (higher) quartiles. The ends of the whiskers indicate 1.5 interquartile ranges of the first and third quartiles, respectively. Points outside the range are mild outliers. (a–h) The box plot displays of K_a, K_s, gene length, GC content, exon length, exon number, expression level and tissue specificity between two selections.

Table 1 Comparisons between positively selected genes (PSGs) and negatively selected genes (NSGs) in *Brassica rapa*

Variable	PSGs[a]	NSGs[a]	P[b]
K_a	0.0278 ± 0.0131	0.0138 ± 0.0082	1.430E–146
K_s	0.0201 ± 0.0101	0.0816 ± 0.0373	7.105E–302
K_a/K_s	1.3472 ± 0.2753	0.1887 ± 0.1233	<E–500
Gene length (kb)	0.8306 ± 0.4885	1.6831 ± 0.8837	6.611E–142
Exon length (kb)	0.6067 ± 0.2978	1.1042 ± 0.5202	2.179E–141
Exon number	2.0 ± 1.2	4.0 ± 2.9	1.627E–77
GC content	0.4202 ± 0.049	0.4203 ± 0.038	0.3419
Expression level (FPKM)	1.145 ± 1.674	7.532 ± 7.295	5.845E–139
Tissue specificity (τ)	0.8430 ± 0.1684	0.6558 ± 0.2313	1.842E–83

[a] Mean ± SD.
[b] One-sided Mann-Whitney U test.

values of PSGs are close to 0.01, and most of the K_s values of NSGs are concentrated in the 0.04–0.14 range, and some are even higher (Figure 3b).

The above observations have also been verified in Arabidopsis. Reanalysis of the evolutionary rates between *A. thaliana* and *A. lyrata* orthologs indicate that higher K_a (0.1187 versus 0.0217) and lower K_s (0.0910 versus 0.1431) values were observed for PSGs (Table S6; $P \ll 0.001$; Yang and Gaut, 2011).

Lower exon number and shorter exon and gene lengths for PSGs, compared with NSGs

To understand whether and how evolutionary rates and selective modes affect gene structure, the genetic features for each Brassica ortholog were characterized, including gene length, exon length, exon number and GC content. The data showed that PSGs in *B. rapa* had a significantly shorter gene length (0.831 kb versus 1.683 kb), shorter exon length (0.607 kb versus 1.104 kb) and fewer exons (two versus four) than NSGs (Figures 2c,e,f and 3c; Table 1; $P << 0.001$). No difference in GC content was detected, however (Figures 2d and 3d; Table 1).

A similar analysis was also performed in Arabidopsis using previously published data (Yang and Gaut, 2011). The average gene length, exon length and exon number of Arabidopsis PSGs was 0.896 kb, 0.746 kb and 1.6, respectively, which is significantly less than those of NSGs (Figure S3c,e,f; Table S6; $P << 0.001$). Differing from Brassica, however, the Arabidopsis PSGs have significantly lower GC content than NSGs (42.8% versus 44.5%; Figures 2d and S3d; Tables 1 and S6; $P << 0.001$).

Comparison of functional differentiation between PSGs and NSGs

To discern whether there is functional differentiation between PSGs and NSGs, the orthologous genes in Brassica were annotated using homologous genes in *A. thaliana*. The data showed that a total of 633 PSGs in Brassica were found as protein coding genes in Arabidopsis (Table S7). Further analysis indicated that these PSGs were mainly significantly enriched in nine categories: transcription related (five), DNA binding (two), regulation of RNA metabolic process (one) and endomembrane system (one) (Table S8; $P < 10^{-4}$). Analysis in DAVID also supports that PSGs were mainly related to transcription, DNA binding and response to stimulus (Table S9). This view was further verified by gene family analysis. As illustrated in Figure 4(a), PSGs (compared with NSGs) tend to be transcription factors, resistance genes and auxin genes. In contrast, a large number of NSGs (relative to PSGs) are protein kinase genes, flower-related genes and glucosinolate genes, although many NSGs belong to transcription factors. The data may indicate that functional differentiation has occurred between PSGs and NSGs, and that PSGs may play an important role during Brassica phenotypic diversification and evolution.

Functional differentiation can also be reflected by the analysis of GO term numbers. Most PSGs (relative to NSGs) had only one or two GO terms, and very few had more than six GO terms (Figure 4b). By contrast, NSGs had a tendency to contain three or more GO terms (relative to PSGs), and some had between seven and 10 GO terms (Figure 4b). These data may suggest that PSGs usually perform a specific function, but the NSGs may participate in several (or many) biological pathways or processes.

Lower expression level and higher tissue-specific expression for PSGs than NSGs

Previous studies had indicated that evolutionary rates often strongly correlate with gene expression, including expression level (Pál *et al.*, 2001; Subramanian and Kumar, 2004; Drummond *et al.*, 2006) and expression breadth (number of tissues where a gene is expressed) (Duret and Mouchiroud, 2000; Zhang and Li, 2004). In order to determine whether there is any difference in the expression patterns between PSGs and NSGs, we used the RNA-seq data from *B. rapa* to calculate each gene

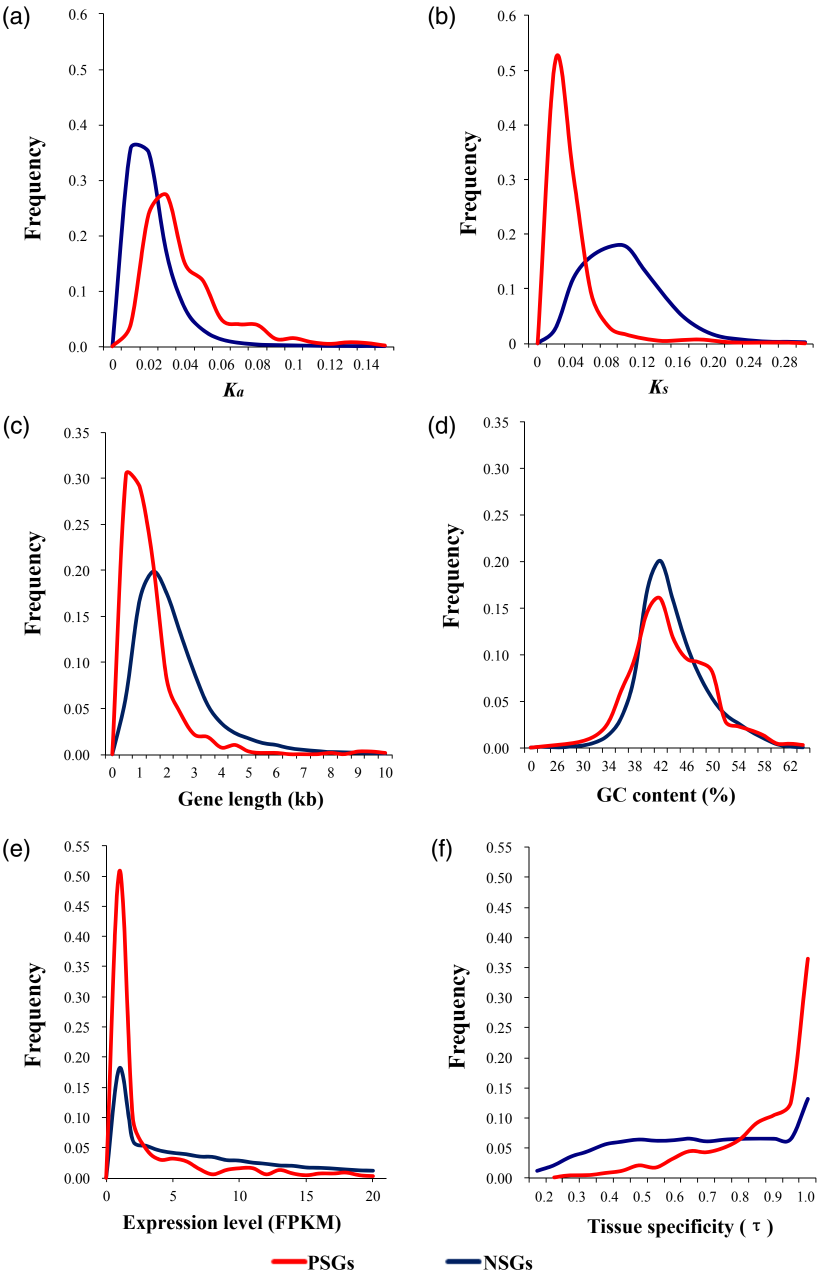

Figure 3. Frequency distributions of K_a (a), K_s (b), gene length (c), GC content (d), expression level (e), and tissue specificity (f) between positively selected genes (PSGs) and negatively selected genes (NSGs). Red and blue lines represent positively and negatively selected genes, respectively ($P < 10^{-50}$ by one-sided Mann-Whitney U test). [Colour figure can be viewed at wileyonlinelibrary.com]

expression level and tissue specificity (Tong *et al*., 2013). Meanwhile, we classified the orthologs into two groups: weakly expressed genes (fragments per kilobase of transcript per million mapped reads, FPKM ≦ 3) and strongly expressed genes (FPKM ≧ 50) (Chen *et al*., 2013). The data showed that PSGs had an overall much lower expression level than NSGs (1.145 versus 7.532; Figures 2g, 3e and S5; Table 1). A total of 313 PSGs (45.3%) were weakly expressed, and only 7.38% of them have a high expression level (Fisher's exact test, $P < 10^{-10}$; Table 2). In contrast, only 26.19% of NSGs were expressed at a very low level, and 10.89% of them were highly expressed (Table 2). To understand the degree of tissue specificity among different genes, we also classified the orthologs into two groups: low tissue specificity (expressed in two or more tissues) and high tissue specificity (expressed in only one tissue). We found that 18.48% of PSGs were expressed with high tissue specificity (Table 2). In contrast, only 4.6% of NSGs were expressed in only one tissue, and the majority (95.4%) could be expressed in multiple tissues (Table 2). The overall degree of tissue specificity (τ) of PSGs is significantly higher than that of NSGs (0.8430 versus 0.6558) (one-sided Fisher's exact test, $P < 10^{-30}$; Figures 2h and 3f; Tables 1 and 2).

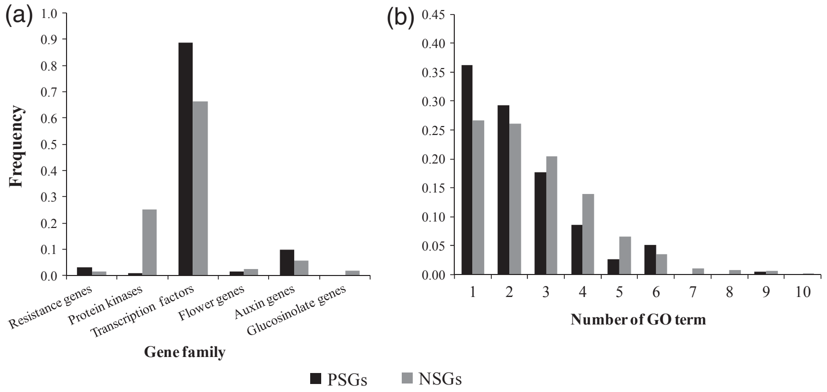

Figure 4. Frequency distributions of gene families (a) and gene ontology (GO) number (b) in Brassica. Black and grey columns represent positively and negatively selective genes, respectively.

Because PSGs and NSGs have different K_a/K_s ratios, we wondered whether the K_a/K_s ratios were correlated with expression patterns. To answer this question, we collected all of the K_a/K_s values from *B. rapa* to *B. oleracea* orthologs and correlated them with expression level and tissue specificity. The data showed that the K_a/K_s ratios were significantly negatively correlated with expression level ($r = -0.2824$, $P < 10^{-320}$) and positively correlated with tissue specificity ($r = 0.1973$, $P < 10^{-178}$) (Table S3). Because K_a and K_s are both associated with K_a/K_s, it is reasonable to deduce that expression patterns are correlated with K_a, K_s, and K_a/K_s.

Retention capability of triplicate genes between PSGs and NSGs

It was found that after the Brassica WGT event, the three subgenomes had undergone extensive genome fractionation, chromosome reduction and block reshuffling (Cheng *et al.*, 2014). During this process, some genes, such as auxin-related genes, had been over-retained (Cheng *et al.*, 2014). To understand whether PSGs and NSGs had been equally preserved after WGT, the frequency of each gene set was calculated on the basis of retention capability, i.e. retaining three copies, retaining two copies (or lost one copy) and retaining only one copy (or lost two copies) (Figure S4; Table 3). It was shown that for the three WGT genes that were retained, PSGs comprised 2.76% of the total genes (i.e. the total number of PSGs and NSGs), which is significantly higher than for when only one WGT gene was retained (1.57%) (one-sided Fisher's exact test, $P = 0.026$; odds ratio, OR = 1.78; Table 3). In statistics, the OR (Cornfield 1951, Edwards 1963, Mosteller 1968) is one of three main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population. This percentage is also higher than the case when only two WGT genes were retained (1.99%), or when one or two WGT genes were retained (1.81%) (Table 3). Although the two data sets do not significantly differ, it seems that the percentage of PSGs increases with the number of WGT genes retained (from 1.57, 1.81 and 1.99 to 2.76%; Table 3). Because the total number of PSGs is low, whether PSGs really have a higher capability of retaining triplicate gene copies needs to be further explored.

Table 2 Comparisons of expression patterns between positively selected genes (PSGs) and negatively selected genes (NSGs) in *Brassica rapa*

	Number of weakly expressed[a]	Number of highly expressed[b]
PSGs	313	51
NSGs	5832	2425
	P = 1.06E–11, OR = 2.55[c]	
	Number of tissue specific[d]	Number of non-tissue specific[e]
PSGs	102	450
NSGs	990	20 509
	P = 1.62E–31, OR = 4.69[c]	

[a] 0 < FPKM ≦ 3 (FPKM, fragments per kilobase of transcript per million mapped reads).
[b] FPKM ≧ 50.
[c] One-sided Fisher's exact test.
[d] Expressed only in one tissue.
[e] Expressed in two or more tissues.

Comparison of codon bias between PSGs and NSGs, and their impacts on gene properties

Codon bias refers to the unequal use of synonymous codons for an amino acid (Hershberg and Petrov, 2008; Larracuente *et al.*, 2008; Plotkin and Kudla, 2011). To understand whether PSGs and NSGs have codon bias, the codon adaptation index (CAI), codon bias index (CBI) and frequency of optimal codons (FOP) were calculated, respectively. As shown in Table S10, for the three codon bias parameters used, PSGs show consistently higher codon bias than NSGs do, and this difference has reached statistical significance for CAI ($P < 10^{-9}$) and FOP ($P = 0.0269$) (Table S10). To further understand the influences of codon bias on gene properties, correlations

Table 3 Comparison of the number of positively selected genes (PSGs) and negatively selected genes (NSGs) under different retention rates after whole-genome triplication (WGT) event in *Brassica rapa*

	Fully retained (three copies)	Partially retained (one copy)
PSGs	18	73
NSGs	633	4573
	$P = 0.026$, OR = 1.78[a]	

	Fully retained (three copies)	Partially retained (one or two copies)
PSGs	18	184
NSGs	633	10 002
	$P = 0.062$, OR = 1.54[a]	

	Fully retained (three copies)	Partially retained (two copies)
PSGs	18	110
NSGs	633	5428
	$P = 0.131$, OR = 1.39[a]	

[a] One-sided Fisher's exact test.

analysis between them has been performed. As shown in Table S11: K_a/K_s is correlated with CBI and CAI; gene length and exon length are negatively correlated with CBI, CAI or FOP; exon number is negatively correlated with CBI and FOP, but is positively correlated with CAI; GC content is positively correlated with CBI and FOP, but is negatively correlated with CAI; expression level is positively correlated with CBI, but is negatively correlated with CAI and FOP; tissue specificity is positively correlated with CBI, CAI and FOP (Table S11).

DISCUSSION

Comparison of evolutionary rates, gene compactness, and expression patterns between PSGs and NSGs: an overview

Several prior studies had performed similar genome-wide analysis of gene evolution by comparing orthologs of closely related relatives (Gaut *et al.*, 2011; Yang and Gaut, 2011; Du *et al.*, 2012). These efforts have revealed evolutionary rate variation, and the factors contributing to this variation (Gaut *et al.*, 2011; Yang and Gaut, 2011; Du *et al.*, 2012). Although positive selection and negative selection were thought to be the two major modes of natural selection (Xu *et al.*, 2014), their influences acting on nuclear genes are still poorly understood. In this study, we used Brassica species as a model system and identified 23 817 orthologs between *B. rapa* and *B. oleracea* (Table S1). The K_a/K_s analyses could classify most of these genes into two types, i.e. PSGs ($K_a/K_s > 1$) and NSGs ($K_a/K_s < 1$). Such classifications may be somewhat artificial, but the comparisons between the two gene sets under different selective modes indeed reveal several interesting observations. These include: (i) PSGs in Brassica show twofold higher K_a and fourfold lower K_s values; (ii) PSGs have twofold shorter gene/exon lengths, and twofold fewer exons; and (iii) PSGs are very weakly expressed and are expressed in a more tissue-specific manner (Table 1). These observations suggest that PSGs and NSGs differ not only in selective pressure (measured as K_a/K_s), but also in evolutionary rates, gene characteristics and expression patterns. This inference has been largely verified by Arabidopsis genes, indicating that such selective patterns may be shared in cruciferous plants.

It is quite interesting to find that no matter in *B. rapa* or in *A. thaliana*, the K_s of PSGs is much lower than NSGs (approximately two- to fourfold). One possibility may be that PSGs have a stronger codon bias (Table S10), are weakly expressed and are expressed in fewer tissues (Table 1). Such properties may be able to help PSGs reduce their synonymous mutation rates, leading to much smaller K_s values. It is also likely that stronger codon usage may increase translation efficiency, because the use of codons that match the most abundant tRNA reduces the time to find and bind the correct tRNA (Duret and Mouchiroud, 1999). Moreover, gene length and exon length were found to be much shorter for PSGs (Table 1). Our data are also consistent with previous large-scale analysis in *Caenorhabditis elegans*, *D. melanogaster* and *A. thaliana* (Duret and Mouchiroud, 1999). In this study, the authors revealed a strong negative correlation between codon usage and protein length (Duret and Mouchiroud, 1999), further supporting the view that selective modes may be an important factor for gene properties.

Can selective modes serve as an alternative indicator of gene compactness?

The relationship between gene structure and gene expression has been frequently investigated, but the debates are still continuing. In humans, highly expressed genes were found to contain fewer and shorter introns/exons, and shorter coding sequences (Castillo-Davis *et al.*, 2002; Eisenberg and Levanon, 2003; Urrutia and Hurst, 2003; Vinogradov, 2004). The compact gene structure was explained by transcriptional efficiency (Castillo-Davis *et al.*, 2002), mutation bias (Urrutia and Hurst, 2003) or genomic design (Vinogradov, 2004). Subsequently, highly expressed genes in dicot species of Arabidopsis and monocot species of rice were reported to contain longer gene transcripts (Ren *et al.*, 2006). The contrasts had been explained by different turn and outcome of selective forces between animals and plants after their split (Ren *et al.*, 2006). This view, however, had been argued immediately in haploid moss *Physcomitrella patens* (Stenøien, 2007). In this study, the author provided data showing that highly expressed genes in moss contain shorter intron and shorter gene lengths (Stenøien, 2007), in contrast with the conclusions of Ren *et al.* (2006), who found that in plants the highly expressed genes are the least compact. Even in a single genome, such as soybean, the expression breadth and exon length were found to be both positively correlated at low and intermediate expression level, and negatively correlated at high expression level (Woody *et al.*, 2010).

Given the fact that correlation between gene structure and gene expression varies within and between different genomes, we here propose that selective modes may serve as an alternative indicator for gene compactness. In this study, PSGs in both *B. rapa* and Arabidopsis consistently show twofold lower intron/exon numbers and shorter gene lengths (Tables 1 and S6). In mammalian genomes, the patterns of positive selection had been analyzed (Kosiol *et al.*, 2008). Although the structural properties of PSGs were not investigated, the expression patterns of PSGs in mammalian genomes were very similar to those in plants, such as being expressed at significantly lower levels and in a more tissue-specific manner (Kosiol *et al.*, 2008). More genomic data are indeed needed to establish and verify the relationships between gene properties and selective modes. We also want to point out that gene compactness may also be affected by other factors, such as evolutionary rates, recombination rates and duplication status. When two or more factors interplay with each other, their independent influences on gene compactness should be cautiously analyzed.

Functional roles of PSGs in the evolution and diversification of *Brassica rapa*

Although the Brassica lineage shared a common ancestor with *A. thaliana* about 20 MYA, only the Brassica lineage underwent a WGT event (Lysak *et al.*, 2005, 2007; Town *et al.*, 2006; Yang *et al.*, 2006; Mun *et al.*, 2009; Zhao *et al.*, 2013). Therefore, the unique WGT event and subsequent genomic rearrangement may have contributed to Brassica speciation and diversification (Cheng *et al.*, 2014). The exact mechanisms underlying Brassica-rich morphotypes, however, remain unclear. In this study, several lines of evidence suggest that PSGs may have contributed to Brassica phenotype diversification after the Brassica–Arabidopsis split. First of all, the number of PSGs in *B. rapa* is much greater than that in *A. thaliana* (698 versus 90; Tables S1 and S6), which may mean that there are more genes to participate in the cell signal pathway and developmental processes. Second, PSGs evolve at much higher protein rates (K_a) but with significantly reduced mutation rates (K_s) (Table 1), which may drive protein diversification without generating more deleterious mutations. Third, PSGs have a more compact gene structure, low expression level and high tissue specificity (Table 1), which may improve transcription/translation efficiency and promote specific tissue development. Fourth, many gene families, including resistance genes, transcription factors and auxin-related genes were proven to be over-retained as PSGs (Figure 4a), indicating that PSGs may play important roles in these gene-related pathways. Last, it seems that PSGs tend to have a higher capability of retaining more triplicate gene copies (Figure S4; Table 3), which may provide more gene resources for further phenotypic diversification. In summary, positive selection may be an important evolutionary force during rediploidization and the diversification process in Brassica.

EXPERIMENTAL PROCEDURES

Data source

The genome sequences for *B. rapa* and *B. olecarea* were downloaded from BRAD 1.2 (http://brassicadb.org; Cheng *et al.*, 2011) and Bolbase (http://www.ocri-genomics.org/bolbase; Liu *et al.*, 2014), respectively. The data for *A. thaliana* were obtained from The Arabidopsis Information Resource (TAIR 9; http://www.arabidopsis.org/index.jsp).

Orthologs, triplicates, sequence alignment and rate estimation

The orthologs between *B. rapa* and *B. olecarea* were determined by InParanoid (Östlund *et al.*, 2009). The alignments were performed using MUSCLE (Edgar, 2004). The K_a and K_s values were estimated using the yn00 module integrated in PAML (Yang, 2007).

The triplicate genes and triplicate relationships of the *B. rapa* genome were determined using previously published data (Wang *et al.*, 2011; Cheng *et al.*, 2012; http://ocri-genomics.org/cgi-bin/bolbase/genewise_on_block.cgi?block=A&page=1).

Functional annotation and gene family classification

The gene annotation for *B. rapa* (v1.2) was downloaded from Phytozome 9.1 (http://phytozome.jgi.doe.gov/pz/portal.html). Gene ontology, functional clustering annotation and tissue expression annotation were performed using the DAVID tool (http://david.ncifcrf.gov/; Huang *et al.*, 2009). The GO enrichment analysis was performed by conducting hypergeometric tests using the 23 817 orthologous gene pairs between *B. rapa* and *B. olerarea* as the background.

The resistance, transcription factors, flower, auxin and glucosinolate related genes were identified based on the classification of BRAD (http://brassicadb.org/brad/browseOverview.php; Wang *et al.*, 2011; Cheng *et al.*, 2012). The protein sequences of *A. thaliana* were used to search the *B. rapa* homologous genes with the lowest *E*-value as the best hit.

Gene expression analysis

The transcriptome data for *B. rapa* were obtained from the six tissues of Chiifu-401-42 generated by Tong *et al.* (2013). Gene expression levels were measured by FPKM using CUFFLINKS 2.1.0 (http://cole-trapnell-lab.github.io/cufflinks/).

Tissue specificity was measured with the index τ (Yanai *et al.*, 2005; Yang and Gaut, 2011). The index τ ranges from 0 to 1. The higher τ values indicate higher specificity. In contrast, if a gene is expressed in only one tissue, τ is close to 1 (Yang and Gaut, 2011).

Statistical tests

We used R 3.0.3 (Field *et al.*, 2012) for statistical analysis and plotting.

ACKNOWLEDGEMENTS

We thank LetPub (http://www.letpub.com) for its linguistic assistance during the preparation of this manuscript. This work was supported by the National Natural Science Foundation of China (31370266), Jiangsu Agricultural Science and Technology Innovation Fund (CX(14)5010), Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture of People's Republic of China (2016002) and Jiangsu Province for Postdoctoral Research Fund (1501143C). The authors declare no conflicts of interest.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Calculations of K_a and K_s among 23 817 orthologs from *Brassica rapa* and *Brassica oleracea*.

Figure S2. Comparisons of K_a/K_s estimates between PSGs and NSGs in Brassica (a) and Arabidopsis (b).

Figure S3. Comparisons of evolutionary rates and gene features between PSGs and NSGs in Arabidopsis.

Figure S4. Frequency distributions of PSGs and NSGs under different retention rates (one, two and three copies) in Brassica.

Figure S5. Frequency distributions of expression levels in different Brassica tissues.

Table S1. Features of K_a/K_s in 23 817 orthologs between *Brassica rapa* and *Brassica oleracea*.

Table S2. Evolutionary rates for 23 817 orthologs between *Brassica rapa* and *Brassica oleracea*.

Table S3. Correlation analyses of evolutionary rates with expression features in Brassica.

Table S4. Sequences and alignments of five randomly selected genes.

Table S5. K_a and K_s values with and without unknown genes in *Brassica rapa*.

Table S6. Comparisons between PSGs and NSGs in *Arabidopsis thaliana.*

Table S7. Functional descriptions of PSGs in Brassica annotated by Arabidopsis.

Table S8. Gene ontology categories enriched for PSGs under a background of all orthologous genes analyzed in DAVID ($P < 0.01$).

Table S9. Gene ontology categories clustered for PSGs under a background of all orthologous genes analyzed by DAVID (enrichment score >2.0).

Table S10. Codon bias comparisons between PSGs and NSGs in *Brassica rapa.*

Table S11. Correlation analysis between codon bias and gene properties in *Brassica rapa.*

REFERENCES

Begun, D.J., Whitley, P., Todd, B.L., Waldrip-Dail, H.M. and Clark, A.G. (2000) Molecular population genetics of male accessory gland proteins in *Drosophila. Genetics*, **156**, 1879–1888.

Beilstein, M.A., Al-Shehbaz, I.A. and Kellogg, E.A. (2006) Brassicaceae phylogeny and trichome evolution. *Am. J. Bot.* **93**, 607–619.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. and Mathews, S. (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana. Proc. Natl Acad. Sci. USA*, **107**, 18724–18728.

Bowers, J.E., Chapman, B.A., Rong, J.K. and Paterson, A.H. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433–438.

Buck, L. and Axel, R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell*, **65**, 175–187.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V. and Kondrashov, F.A. (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* **31**, 415–418.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C. and Samans, B. (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science*, **345**, 950–953.

Chen, X., Zhu, W., Azam, S., Li, H., Zhu, F., Li, H., Hong, Y., Liu, H., Zhang, E. and Wu, H. (2013) Deep sequencing analysis of the transcriptomes of peanut aerial and subterranean young pods identifies candidate genes related to early embryo abortion. *Plant Biotechnol. J.* **11**, 115–127.

Cheng, F., Liu, S.Y., Wu, J., Fang, L., Sun, S.L., Liu, B., Li, P.X., Hua, W. and Wang, X.W. (2011) BRAD, the genetics and genomics database for *Brassica* plants. *BMC Plant Biol.* **11**, 136.

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

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

Civetta, A. and Singh, R.S. (1995) High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J. Mol. Evol.* **41**, 1085–1095.

Cornfield, J. (1951) A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J. Nat. Cancer Inst.* **11**, 1269–1275.

Coulthart, M.B. and Singh, R.S. (1988) High level of divergence of male-reproductive-tract proteins, between *Drosophila melanogaster* and its sibling species, *D. simulans*. *Mol. Biol. Evol.* **5**, 182–191.

Drummond, D.A., Raval, A. and Wilke, C.O. (2006) A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337.

Du, J.C., Tian, Z.X., Sui, Y., Zhao, M.X., Song, Q.J., Cannon, S.B., Cregan, P. and Ma, J.X. (2012) Pericentromeric effects shape the patterns of divergence, retention, and expression of duplicated genes in the paleopolyploid soybean. *Plant Cell*, **24**, 21–32.

Duret, L. and Mouchiroud, D. (1999) Expression pattern and surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl Acad. Sci. USA*, **96**, 4482–4487.

Duret, L. and Mouchiroud, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**, 68–74.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797.

Edwards, A.W.F. (1963) The measure of association in a 2×2 table. *J. Roy. Stat. Soc. Series A (General)*, **126**, 109–114.

Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* **19**, 362–365.

Fay, J.C. and Wu, C.I. (2001) The neutral theory in the genomic era. *Curr. Opin. Genet. Dev.* **11**, 642–646.

Field, A., Miles, J. and Field, Z. (2012) *Discovering Statistics Using R*, 1st edn. Thousand Oaks: SAGE Publications.

Gaut, B., Yang, L., Takuno, S. and Eguiarte, L.E. (2011) The patterns and causes of variation in plant nucleotide substitution rates. *Annu. Rev. Ecol. Evol. Syst.* **42**, 245–266.

Ge, X.H., Wang, J. and Li, Z.Y. (2009) Different genome-specific chromosome stabilities in synthetic *Brassica* allohexaploids revealed by wide crosses with *Orychophragmus*. *Ann. Bot.* **104**, 19–31.

Hershberg, R. and Petrov, D.A. (2008) Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57.

Hughes, A.L. (1999) *Adaptive Evolution of Genes and Genomes*, 1st edn. New York: Oxford University Press.

Hughes, A.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Hurst, L.D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18**, 486–487.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H. and Soltis, P.S. (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.

Johnston, J.S., Pepper, A.E., Hall, A.E., Chen, Z.J., Hodnett, G., Drabek, J., Lopez, R. and Price, H.J. (2005) Evolution of genome size in *Brassicaceae*. *Ann. Bot.* **95**, 229–235.

Klein, J. and Figueroa, F. (1985) Evolution of the major histocompatibility complex. *Crit. Rev. Immunol.* **6**, 295–386.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. (2002) Selection in the evolution of gene duplications. *Genome Biol.* **3**(2), research0008.

Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R. and Siepel, A. (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**(8), e1000144.

Labana, K.S. and Gupta, M.L. (1993) Importance and origin. In *Breeding Oilseed Brassicas* (Labana, K.S., Banga, S.S. and Banga, S.K., eds). Heidelberg: Springer, pp. 1–7.

Larracuente, A.M., Sackton, T.B., Greenberg, A.J., Wong, A., Singh, N.D., Sturgill, D., Zhang, Y., Oliver, B. and Clark, A.G. (2008) Evolution of protein-coding genes in *Drosophila*. *Trends Genet.* **24**, 114–123.

Li, W.H. (1997) *Molecular Evolution*, 1st edn. Sunderland: Sinauer Associates Incorporated.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A.P., Zhao, M., Ma, J., Yu, J. and Huang, S. (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* **5**, 3930.

Lysak, M.A., Koch, M.A., Pecinka, A. and Schubert, I. (2005) Chromosome triplication found across the tribe *Brassiceae*. *Genome Res.* **15**, 516–525.

Lysak, M.A., Cheung, K., Kitschke, M. and Bureš, P. (2007) Ancestral chromosomal blocks are triplicated in *Brassiceae* species with varying chromosome number and genome size. *Plant Physiol.* **145**, 402–410.

Matassi, G., Sharp, P.M. and Gautier, C. (1999) Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.* **9**, 786–791.

Mosteller, F. (1968) Association and estimation in contingency tables. *J. Am. Stat. Assoc.* **63**, 1–28.

Mun, J.H., Kwon, S.J., Yang, T.J., Seol, Y.J., Jin, M., Kim, J.A., Lim, M.H., Kim, J.S., Baek, S. and Choi, B.S. (2009) Genome-wide comparative analysis of the *Brassica rapa* gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biol.* **10**, 1.

Nagaharu, U. (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn. J. Bot.* **7**, 389–452.

Nei, M. (1987) *Molecular Evolutionary Genetics*, 1st edn. New York: Columbia University Press.

Nei, M., Suzuki, Y. and Nozawa, M. (2010) The neutral theory of molecular evolution in the genomic era. *Annu. Rev. Genomics Hum. Genet.* **11**, 265–289.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D. and White, T.J. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170.

Östlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D.N., Roopra, S., Frings, O. and Sonnhammer, E.L.L. (2009) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196–D203.

Pál, C., Papp, B. and Hurst, L.D. (2001) Highly expressed genes in yeast evolve slowly. *Genetics*, **158**, 927–931.

Plotkin, J.B. and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42.

Ren, X.Y., Vorst, O., Fiers, M.W.E.J., Stiekema, W.J. and Nap, J.P. (2006) In plants, highly expressed genes are the least compact. *Trends Genet.* **22**, 528–532.

Song, K.M., Lu, P., Tang, K.L. and Osborn, T.C. (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. *Proc. Natl Acad. Sci. USA*, **92**, 7719–7723.

Stenøien, H.K. (2007) Compact genes are highly expressed in the moss *Physcomitrella patens*. *J. Evol. Biol.* **20**, 1223–1229.

Subramanian, S. and Kumar, S. (2004) Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, **168**, 373–381.

Swanson, W.J., Yang, Z., Wolfner, M.F. and Aquadro, C.F. (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl Acad. Sci. USA*, **98**, 2509–2514.

Ticher, A. and Graur, D. (1989) Nucleic acid composition, codon usage, and the rate of synonymous substitution in protein-coding genes. *J. Mol. Evol.* **28**, 286–298.

Tong, C., Wang, X., Yu, J., Wu, J., Li, W., Huang, J., Dong, C., Hua, W. and Liu, S. (2013) Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in *Brassica rapa*. *BMC Genomics*, **14**, 689.

Town, C.D., Cheung, F., Maiti, R., Crabtree, J., Haas, B.J., Wortman, J.R., Hine, E.E., Althoff, R., Arbogast, T.S. and Tallon, L.J. (2006) Comparative genomics of Brassica oleracea and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell*, **18**, 1348–1359.

Tsaur, S.C. and Wu, C.I. (1997) Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of *Drosophila*. *Mol. Biol. Evol.* **14**, 544–549.

Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* **13**, 2260–2264.

Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* **20**, 248–253.

Wang, X.W., Wang, H.Z., Wang, J., Sun, R.F., Wu, J., Liu, S.Y., Bai, Y.Q., Mun, J.H., Bancroft, I. and Cheng, F. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039.

Wolfe, K.H., Sharp, P.M. and Li, W.H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature*, **337**, 283–285.

Woody, J.L., Severin, A.J., Bolon, Y.T., Joseph, B., Diers, B.W., Farmer, A.D., Weeks, N., Muehlbauer, G.J., Nelson, R.T. and Grant, D. (2010) Gene expression patterns are correlated with genomic and genic structure in soybean. *Genome*, **54**, 10–18.

Wyckoff, G.J., Wang, W. and Wu, C.I. (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature*, **403**, 304–309.

Xu, L., Bickhart, D.M., Cole, J.B., Steven, G.S., Jiuzhou, S., Curtis, P.V.T., Tad, S.S. and George, E.L. (2014) Genomic signatures reveal new evidences for selection of important traits in domestic cattle. *Mol. Biol. Evol.* **32**(3), 711–725.

Yanai, I., Benjamin, H., Shmoish, M. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.

Yang, Z.H. (2005) The power of phylogenetic comparison in revealing protein function. *Proc. Natl Acad. Sci. USA*, **102**, 3179–3180.

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

Yang, L. and Gaut, B.S. (2011) Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol. Biol. Evol.* **28**, 2359–2369.

Yang, Y.W., Lai, K.N., Tai, P.Y. and Li, W.H. (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* **48**, 597–604.

Yang, T.J., Kim, J.S., Kwon, S.J., Lim, K.B., Choi, B.S., Kim, J.A., Jin, M., Park, J.Y., Lim, M.H. and Kim, H.I. (2006) Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region of *Brassica rapa*. *Plant Cell*, **18**, 1339–1347.

Zhang, L. and Li, W.H. (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.* **21**, 236–239.

Zhang, L.Q., Vision, T.J. and Gaut, B.S. (2002) Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol. Biol. Evol.* **19**, 1464–1473.

Zhao, M., Du, J., Lin, F., Tong, C., Yu, J., Huang, S., Wang, X., Liu, S. and Ma, J. (2013) Shifts in the evolutionary rate and intensity of purifying selection between two *Brassica* genomes revealed by analyses of orthologous transposons and relics of a whole genome triplication. *Plant J.* **76**, 211–222.