Genet. Res., Camb. (2008), **90**, *pp.* 97–109. © 2008 Cambridge University Press
doi:10.1017/S0016672307009032 Printed in the United Kingdom

Mating system and recombination affect molecular evolution in four *Triticeae* species

A. HAUDRY[1,2], A. CENCI[1], C. GUILHAUMON[1], E. PAUX[3], S. POIRIER[1], S. SANTONI[1], J. DAVID[1] AND S. GLÉMIN[2]*

[1] *UMR Diversité et Adaptation des Plantes Cultivées, Montpellier SupAgro – INRA – IRD – UMII, 2, Place Pierre Viala, 34060 Montpellier Cedex 1, France*
[2] *Institut des Sciences de l'Evolution, Université Montpellier 2, place Eugène Bataillon, Montpellier, France*
[3] *UMR ASP 1095, Université Clermont Ferrand, INRA, F-63100 Clermont Ferrand, France*

(*Received 26 August 2007 and in revised form 20 October 2007*)

Summary

Mating systems and recombination are thought to have a deep impact on the organization and evolution of genomes. Because of the decline in effective population size and the interference between linked loci, the efficacy of selection is expected to be reduced in regions with low recombination rates and in the whole genome of self-fertilizing species. At the molecular level, relaxed selection is expected to result in changes in the rate of protein evolution and the pattern of codon bias. It is increasingly recognized that recombination also affects non-selective processes such as the biased gene conversion towards GC alleles (bGC). Like selection, this kind of meiotic drive in favour of GC over AT alleles is expected to be reduced in weakly recombining regions and genomes. Here, we investigated the effect of mating system and recombination on molecular evolution in four *Triticeae* species: two outcrossers (*Secale cereale* and *Aegilops speltoides*) and two selfers (*Triticum urartu* and *Triticum monococcum*). We found that GC content, possibly driven by bGC, is affected by mating system and recombination as theoretically predicted. Selection efficacy, however, is only weakly affected by mating system and recombination. We investigated the possible reasons for this discrepancy. A surprising one is that, in outcrossing lineages, selection efficacy could be reduced because of high substitution rates in favour of GC alleles. Outcrossers, but not selfers, would thus suffer from a 'GC-induced' genetic load. This result sheds new light on the evolution of mating systems.

1. Introduction

Mating systems are thought to play a major role in genome evolution (Charlesworth & Wright, 2001). They affect the effective population size, N_e, and the efficacy of recombination, which both play a crucial role in molecular evolution by controlling patterns of polymorphism, the efficacy of selection, and specific processes such as biased gene conversion towards GC (Marais *et al.*, 2004). Inbreeding reduces N_e because of non-independent gamete sampling (corresponding to a 50% reduction under complete selfing) (Pollak, 1987) and the efficacy of recombination (Nordborg, 2000), which in turn reduces N_e further through hitchhiking effects due to the recurrent elimination of deleterious alleles, (background selection; Charlesworth *et al.*, 1993), or the spread of advantageous mutations (selective sweeps; Maynard-Smith & Haigh, 1974). Such a reduction in N_e is also expected in regions of low recombination in outcrossing species (Charlesworth *et al.*, 1993). Finally, self-fertilizing species are usually more prone to recurrent bottlenecks (Schoen & Brown, 1991) thanks to their capacity for founding new populations with few seeds or even only one seed (Baker, 1955). In many cases, extinction-recolonization dynamics also reduces local and species-wide N_e (Ingvarsson, 2002).

Because the efficacy of selection mainly depends on the product $N_e s$, where s is the selection coefficient,

* Corresponding author. Telephone: (+33) 4 67 14 48 18. Fax: (+33) 4 67 14 36 10. e-mail: glemin@univ-montp2.fr

highly self-fertilizing species should be less efficient than outcrossers at purging slightly deleterious alleles or fixing new advantageous mutations. At the molecular level, we would expect to observe signatures of relaxed selection in selfers, such as an elevated ratio of non-synonymous over synonymous substitution rates (d_N/d_S ratio) due to the fixation of slightly deleterious mutations, and la ow level of codon bias due to the inefficacy of selection for preferred codons (Akashi, 1995).

While the effect of selfing on polymorphism is now well documented (Glémin *et al*., 2006; Hamrick & Godt, 1996; Nybom, 2004), its impact on selection efficacy has been only weakly supported. In a meta-analysis on a wide set of plants, Glémin *et al*. (2006) found a weak signal of relaxed selection both against slightly deleterious alleles and in favour of new advantageous mutations in self-fertilizing species, compared with outcrossing ones. Wright *et al*. (2002), however, did not find any difference between the selfer *Arabidopsis thaliana* and its outcrossing close relative *Arabidopsis lyrata*, either in the rate of protein evolution or in codon bias. Patterns of selection, especially patterns of codon bias, can be obscured by biased gene conversion towards GC (bGC), which mimics selective effects (Marais, 2003). Biased gene conversion is a kind of meiotic drive, in which GC alleles are favoured over AT alleles. Increasing evidence suggests that it might affect genome evolution and that it should be taken into account in genomic studies (Marais, 2003; Meunier & Duret, 2004; Galtier & Duret, 2007). It occurs at heterozygote sites involved in the Holliday junction during recombination, so that it is expected to be rare or absent in selfers and in regions of low recombination (Marais, 2003; Marais *et al*., 2004). Recently, Wright *et al*. (2007) compared codon usage and base composition between the outcrossers *A. lyrata* and *Brassica oleracea* and the selfer *A. thaliana*. Because most preferred codons ended in G or C, the higher GC content at synonymous sites in outcrossers can be the result of more efficient selection for codon usage or the result of stronger bGC in outcrossers. Because the shift in base composition is independent of gene expression level, Wright *et al*. (2007) concluded that base composition is more probably caused by bGC (or change in mutation bias) rather than by a reduction in the efficacy of selection in *A. thaliana*. Until now, studies using species other than *A. thaliana* have been very scarce, and the effect of mating systems on protein and base composition evolution remains unclear. Testing the effect of mating systems in other groups of species appears timely.

In this study we investigated the effect of mating systems on molecular evolution in four diploid *Triticeae* species: two outcrossers, (i) rye (*Secale cereale*) (Lundqvist, 1954) and (ii) *Aegilops speltoides* (Dvorak *et al*., 1998); and two self-fertilizing sister species, (iii) *Triticum urartu* and (iv) *Triticum monococcum* (Dvorak *et al*., 1993; Yamane & Kawahara, 2005). *Ae. speltoides*, *T. urartu* and *T. monococcum* are wild diploid relatives of durum wheat (*T. turgidum* subsp. *durum*). The phylogenetic relationships and timeline of the evolution of these lineages were estimated by Huang *et al*. (2002) (Fig. 1). Under the parsimony hypothesis, selfing should have evolved sometime after the last common ancestor of *T. urartu* and *T. monococcum* with *Ae. speltoides*.

In durum wheat, a representative of polyploid *Triticeae*, a steep recombination gradient along chromosome arms has been found using C-banding polymorphism among chromosomes: about 80% of recombination occurs in the most distal 20% of the chromosomes (Lukaszewski, 1992; Lukaszewski & Curtis, 1993). The distribution of recombination differs also between physically short and long chromosome arms in wheat, with presumably higher recombination rates in short than in long arms (Lukaszewski & Curtis, 1993). This pattern of recombination is found in both wheat and barley (Kunzel *et al*., 2000). These two species are quite divergent among *Triticeae* so that this pattern is probably the norm in this tribe. In *Aegilops* species, including *Ae. speltoides*, Dvorak *et al*. (1998) showed that these recombination gradients affect levels of diversity. RFLP polymorphism is higher in telomeric regions than in centromeric regions, the ratio being much higher in some self-fertilizing (*Ae. searsi*: 24·0) than in outcrossing species (*Ae. speltoides*: 1·5). These results suggest that recombination should strongly affect genome evolution, even in selfers. *Triticeae* species thus offer a good opportunity to test for the effect of both mating systems and recombination on patterns of molecular evolution. It is also worth noting that bGC may occur in *Triticeae* species. Glémin *et al*. (2006) found highly significant differences in GC content between self-fertilizing and outcrossing *Gramineae* species, in both coding and non-coding regions. They suggested that bGC could be higher in *Gramineae* than in other flowering plants, which is in agreement with the high GC content and heterogeneity in GC found in numerous *Gramineae* genomes (Barakat *et al*., 1997; Carels & Bernardi, 2000; Wong *et al*., 2002).

Using the wheat, rye and barley genomic resources available in the publicdomain, we sequenced 52 genes in the four *Triticeae* species and used *Hordeum spontaneum* as an outgroup (or *H. vulgare* when *H. spontaneum* was not available). Most genes were chosen to belong to the same chromosome in order to cover the recombination gradient between centromeric and telomeric regions. To locate genes, we used the genomic resources of the 3B chromosome of wheat, which is the most conserved chromosome

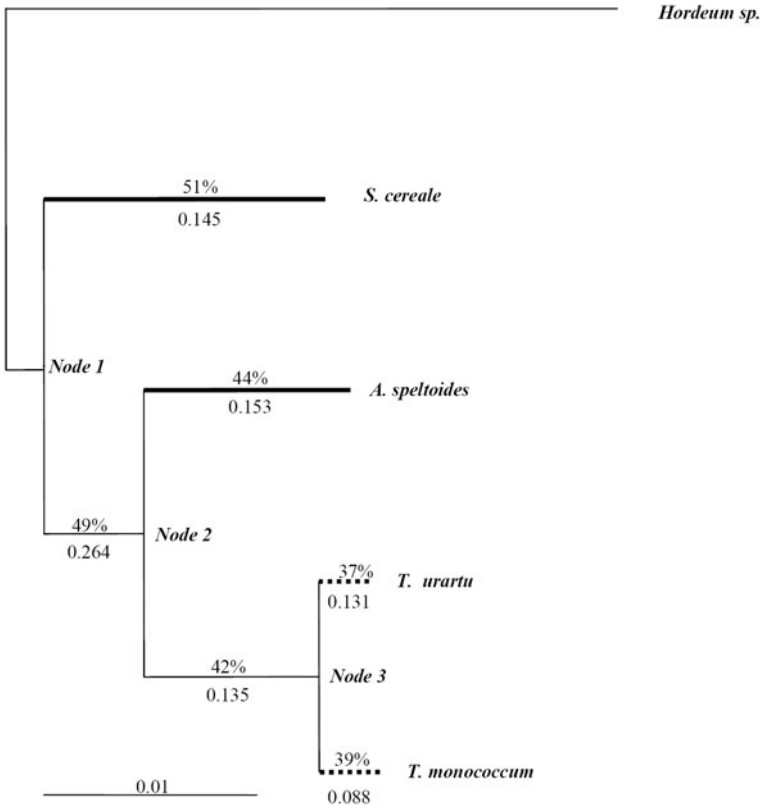

Fig. 1. Phylogenetic relationships of the studied species. Barley diverged between 11 and 12 MYA from the diploid *Triticum* and *Aegilops* species, whereas rye diverged more recently, about 7 MYA. *Aegilops speltoides* diverged between 2·5 and 6 MYA from the two *Triticum species*, which separated more recently (around 1 MYA) (Huang *et al*., 2002). Dashed lines, self-fertilizing lineages leading to *T. monococcum* and *T. urartu*; bold lines, outcrossing lineages leading to *S. cereale* and to *Ae. speltoides*. GC* are shown above each branch. Ratios of non-synonymous to synonymous substitutions (ω) shown below each branch were calculated using CODEML (free-ratio model, Mb-8; see text).

between wheat and rice (chromosome 1; Munkvold *et al*., 2004; Sorrells *et al*., 2003), and for which the relationship between genetic and physical maps has been intensively investigated (Paux *et al*., 2006).

We studied the effect of mating systems in regions of low and high recombination on the efficacy of selection through the rate of protein evolution (d_N/d_S) and the pattern of codon bias. We also investigated GC content and GC evolution between mating systems and chromosomal regions. We thus tested whether d_N/d_S ratios are higher, and codon bias and GC content lower, in self-fertilizing than in outcrossing species. Similarly, we tested the same predictions between regions of high and low recombination.

2. Materials and methods

(i) *Plant materials*

Among *Triticeae*, it is difficult to find several pairs of sister species differing by their mating systems because of phylogenetic and mating system uncertainties. The phylogenetic relationships among the four species we chose are well supported even if gene flow among the *Triticum*/*Aegilops* group cannot be excluded. This point is discussed below. We also chose species for which the mating system is well known, and to avoid incorrect assignation of ancestral mating systems, we assumed that mating systems are known for the terminal branches only. We used five species: two selfers, *T. monococcum* (accession DV-92-4-1, collection, NSF Jan Dvorak) and *T. urartu* (DV-1792-3, collection NSF Jan Dvorak); and two outcrossers, *Ae. speltoides* (S1 collection INRA Rennes, J. Jahier and *S. cereale* (PI 561793, USDA). *H. spontaneum* (PI 282585, USDA) was used as an outgroup. When sequences from *H. spontaneum* were not available, we used sequences of *H. vulgare*, the domesticated form of *H. spontaneum*, extracted from the Expressed Sequence Tag (EST) database of GenBank (National Center for Biotechnology Information, NCBI; http://www.ncbi.nlm.nih.gov).

(ii) *Sampled locus*

We sequenced a set of 46 genes expected to cover the homoeologous group 3 chromosome, to allow separation into hypothesized regions differing in recombination rate. PCR primers for gene sequencing

https://doi.org/10.1017/S0016672307009032 Published online by Cambridge University Press

were designed using either a 19 400 BAC End Sequences dataset (Paux *et al*., 2006) generated from a chromosome 3B-specific BAC (Safar *et al*., 2004), or with barley and wheat ESTs matching rice chromosome 1. We used the location of rice orthologues (on chromosome 1) as a proxy for their chromosomal position. Although synteny between rice chromosome 1 and wheat chromosome 3 is sometimes broken, gene repertoire and order is mostly conserved (Munkvold *et al*., 2004; Rota & Sorrells, 2004; Sorrells *et al*., 2003). Because we used two raw recombination categories only, the expected degree of synteny between the two chromosomes seems to be sufficient for our analyses. The relative distance to the centromere was computed for each chromosome arm, assuming that the centromere is located around 17 000 000 bp from the telomere of the short arm in rice (Sasaki *et al*., 2002). This defined the short arm versus long arm groups of genes according to their position relative to the centromere. Genes located further than 75% of the arm length from the centromere were grouped into the so-called telomeric group, while other genes were grouped into the centromeric group. We thus assumed that gross recombination patterns remain unchanged between species. Correspondence between rice and wheat chromosome locations (assumed to reflect the position in the four species studied here) was partly through EST physical mapping in wheat nullisomic-tetrasomic, ditelosomic and deletion lines (Sears, 1954; Sears & Sears, 1978; Endo & Gill, 1996), either by direct homology of the studied locus with a mapped EST (Qi *et al*., 2004; http://wheat.pw.usda.gov/NSF/progress_mapping.html), or by mapping of molecular markers from the same BAC or BAC contig (Paux *et al*., 2006).

We added five nuclear genes (*Crtiso*, *Eif4e*, *Eifiso4e*, *Gsp-1*, *Psy2*) from different genome locations and one chloroplastic gene (*matK*). No specific location was attributed to these genes in our analysis. Primers, PCR and sequencing conditions are described in Supplementary Table S2. Sequences were aligned manually with the Staden Package (Staden *et al*., 2001). Coding regions were identified by comparison with EST data and the annotated rice genome (http://www.gramene.org). In what follows, we analysed only coding regions.

(iii) *Protein evolution*

We used the maximum likelihood method of the CODEML program in the PAML package (Yang, 1997) to test for variation in the d_N/d_S ratio (hereafter ω) under different scenarios (see below). For such analyses, we used the best phylogenetic tree found by maximum likelihood reconstruction (model GTR + Γ) using the PHYML software (Guindon & Gascuel, 2003) on all loci concatenated (Fig. 1). This tree is consistent with existing literature (for example see Huang *et al*., 2002).

To test the effect of mating system on ω, we considered several nested models of sequence evolution (Fig. 2). Because we investigated selection patterns on the four focal species, we used *Hordeum* as an outgroup to distinguish internal and external branches. When this distinction was not necessary, we excluded *Hordeum* sequences because we wanted to detect selection events on the four focal species only. First, we tested for differences in ω between branches according to their mating system. Here, we thus included *Hordeum* sequences and we ran the following models: (i) the null model (M0), which assumes the same ω ratio for all branches; (ii) a second model (Mb-2), which assumes two ratios, one for internal branches plus the outgroup *Hordeum*, ω_0, and one for the external branches leading to the four focal species, ω_1; (iii) a third model (model Mb-3), which assumes three ratios, one for internal branches plus the outgroup *Hordeum*, ω_0, and two different ratios for the external branches, ω_{out} for outcrossing species and ω_{self} for self-fertilizing ones; (iv) a free ratio model (Mb-8), in which each branch has its own ω ratio (8 ω). We used ω_0 in Mb-2 and Mb-3 because the mating systems of internal branches are unknown. To test for the effect of mating systems, Mb-3 was compared with Mb-2. These analyses were performed on every gene independently, on all genes concatenated, and on groups of genes concatenated according to their putative chromosome location (long vs short arm; telomeric vs centromeric region). To test for differences in ω between genomic regions (long vs short chromosome arm, telomeric vs centromeric region), we used the codon models for multiple gene categories as described by Yang & Swanson (2002) (Fig. 2): we compared a model Mc-0, assuming the same ω (ω_0) for both categories (and for all branches), with a model Mc-1, assuming one ω (for all branches) for each category (ω_{high}, ω_{low}). In this analysis we excluded the *Hordeum* outgroup.

Under purifying selection ($\omega < 1$), ω increases when selection is relaxed, but the reverse is expected under adaptive evolution ($\omega > 1$). Few adaptive substitutions can thus lead to patterns with higher ω values in outcrossing than in selfing lineages, which is misleading relative to our assumption of accumulating more deleterious mutations in selfers. We thus tested for the possible occurrence of positive selection in the dataset (Fig. 2). First, we ran the nested site models M7 and M8 of CODEML (Yang, 1998). Second, we ran the clade model that allows the d_N/d_S ratio to vary among sites and among lineages to detect lineage-specific variation in selection pressures and compared it with the null nearly neutral model (M1a) (Bielawski & Yang, 2004). In these tests, we excluded the *Hordeum* sequences to avoid detecting

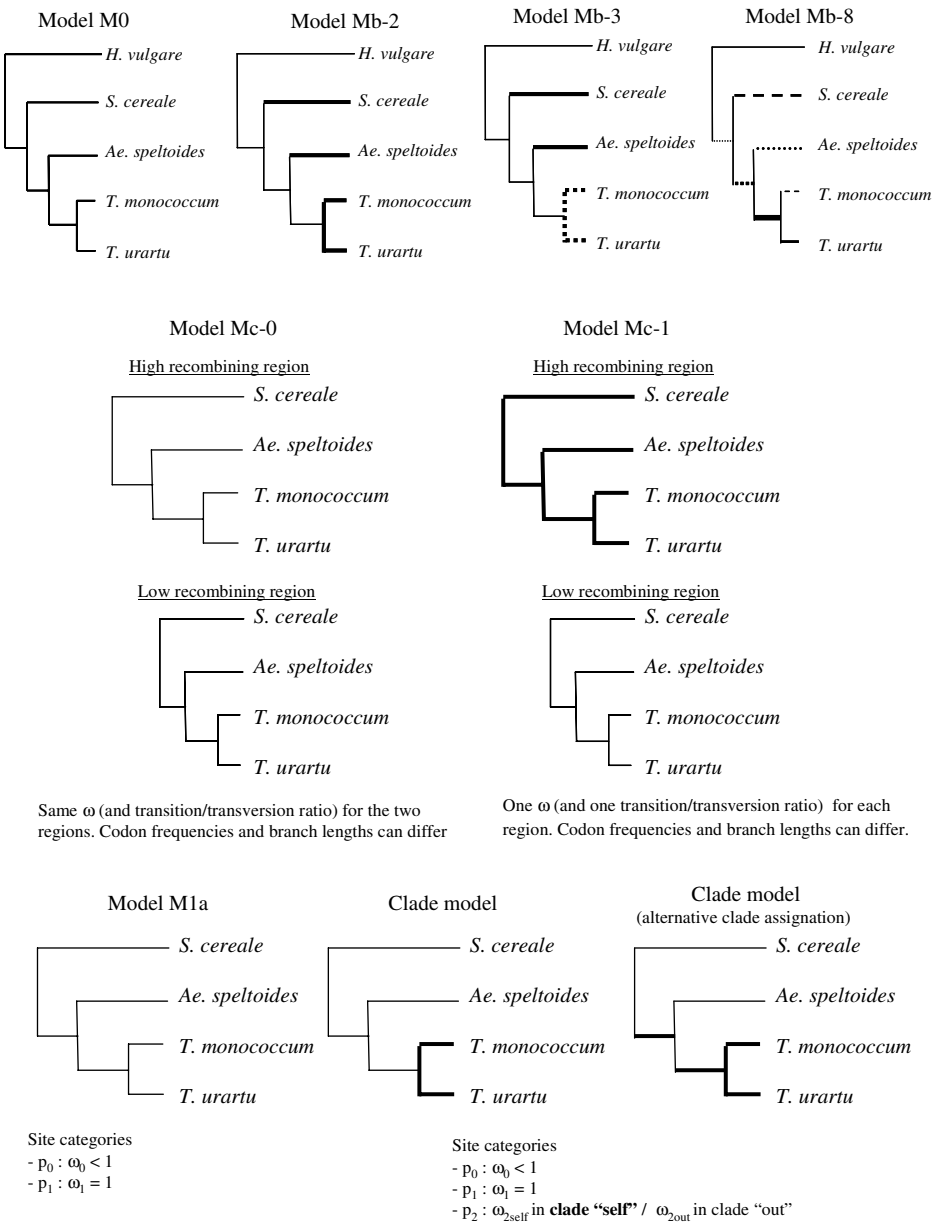

Fig. 2. Schematic representation of the different branch and branch-site models tested with codeml. Branches with identical thickness lines have the same ω value (M0, Mb-2, Mb-3, Mb-8, Mc-0 and Mc-1), or the same distribution of ω categories (M1a, *clade*). In models Mc-0 and Mc-1, the two gene categories are allowed to have their own branch lengths as drawn. Model Mb-3 is tested against model Mb-2 (*d.f.* = 1); Model Mc-1 is tested against model Mc-0 (*d.f.* = 2); the clade model is tested against model M1a (*d.f.* = 3).

a signature of positive selection occurring in the *Hordeum* branch.

The maximum likelihoods of the models were computed and their significance was tested by likelihood ratio tests (LRT) with the appropriate degrees of freedom (Yang, 1998; Yang & Nielsen, 1998).

(iv) *Base composition and codon usage evolution*

For every gene except the chloroplastic gene *matk*, the GC contents of the total, first, second and third codon positions were computed (GC, GC1, GC2 and GC3, respectively). We also computed the equilibrium GC content a sequence would eventually reach if the substitution pattern occurring in a given lineage remained constant (denoted GC* hereafter). Given the rate of the two classes of substitutions, AT→GC (u) and GC→AT (v), the GC content should converge to GC* = $u/(u+v)$ (Sueoka, 1962). We used the phylogenetic tree of Fig. 1 and estimated the two categories of substitutions on every branch by parsimony using the DNAPARS procedure of the PHYLIP package (Felsenstein, 1989). To achieve sufficient power we pooled the substitutions occurring on external

https://doi.org/10.1017/S0016672307009032 Published online by Cambridge University Press

branches leading to *S. cereale* plus *Ae. speltoides* (bold lines in Fig. 1) and *T. monococcum* plus *T. urartu* (dashed lines in Fig. 1). We computed GC* for all sites, and separately for the first (GC1*), second (GC2*) and third (GC3*) codon positions. Substitutions were pooled for all genes or by genomic region (long vs short arm, telomeric vs centromeric region). Significance of differences in GC* between lineages or between genomic regions were assessed by chi-square tests.

We also examined the pattern of codon usage in the four studied species. The list of preferred codons was determined on both *T. aestivum* and *H. vulgare* (Kawabe & Miyashita, 2003; Liu & Xue, 2005; Wang & Roossinck, 2006). As the preferred codons are almost identical in these two species, we assumed that the four species studied here share the same preferred codons as *T. aestivum*. We estimated the frequency of optimal codons (*Fop*; Ikemura, 1985) for every gene. Letting U be the substitution rate from preferred to non-preferred codons and P the reverse substitution rate, we estimated the stationary *Fop* values as $Fop^* = P/(P+U)$. P and U were estimated by parsimony as described for GC and AT substitutions (see above). To distinguish between the effect of selection efficacy on codon usage and the effect of biased gene conversion, we used the fact that in four-fold and six-fold degenerated codons (used for some amino acids), C is preferred over G (Kawabe & Miyashita, 2003; Liu & Xue, 2005; Wang & Roossinck, 2006). For these amino acids, we computed C* and G* values as described above. Substitutions were also pooled for all genes or by genomic regions and chi-square tests were performed.

3. Results

We obtained the complete sequences of 46 gene fragments expected to belong to chromosome 3 (17 designed on BES, 29 based on rice homology), and for 6 other gene fragments located elsewhere in the genome (Supplementary Table S1). Fragments have a minimum length of 700 bp. A total of 31 218 bp of coding regions was available for analyses. Twelve genes (6531 bp) were assigned to the telomeric region versus 34 to the centromeric region (22 437 bp); 35 were assigned to the long arm (19 161 bp) versus 11 to the short arm (8592 bp). The location of seven genes is uncertain. Genes are unevenly located but the average synonymous divergences in each genomic region are very similar (telomeric region, 0·116; centromeric region, 0·109; short arm, 0·117; long arm, 0·105).

(i) *Rate of protein evolution*

The average ω ratio computed on all concatenated genes is 0·160 (model M0) and varied between 0·088 and 0·264 in the free ratio model (Mb-8). None of the 52 single gene fragments, when analysed separately, revealed a significant difference between selfing and outcrossing lineages: ω_{self} is higher than ω_{out} in no more than about half the genes (Fig. 3). In the concatenated gene analyses, contrary to expectations, ω_{self} (from 0·077 to 0·115) is lower than ω_{out} (from 0·114 to 0·159), but the difference is significant only for centromeric genes ($P = 0{\cdot}048$; Table 1).

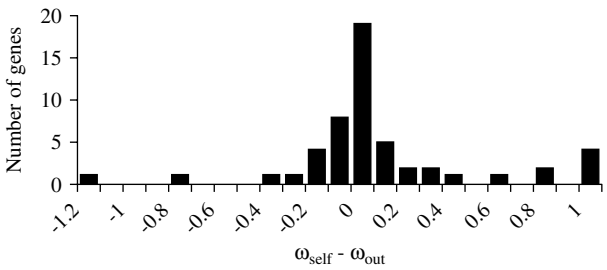

Fig. 3. Distribution of the differences $\omega_{self} - \omega_{out}$ estimated in model Mb-3 for the 52 genes fragments sequenced. This model has three ω ratios: one for the internal branches + *Hordeum*; one for the external outcrossing branches, ω_{out}; and one for the external selfing branches, ω_{self}.

On the whole concatenated sequence dataset, we detected a proportion of 2·5% of sites under possible positive selection with $\omega = 2{\cdot}88$ on average ($P = 0{\cdot}031$) (M8 vs M7; Table 2). Among these sites, none has a posterior probability of being under positive selection higher than 0·95 (Yang *et al.*, 2005). Excluding these sites from the dataset did not change previous conclusions (data not shown). Under the clade model we detected a significant difference between outcrossing and self-fertilizing lineages ($P = 0{\cdot}021$): a proportion $P_2 = 3\%$ of sites are under positive selection in outcrossers ($\omega_{2out} = 3{\cdot}02$), while these sites evolve almost neutrally in selfers ($\omega_{2self} = 1{\cdot}01$) (Table 2). In summary, we found no clear evidence that the two self-fertilizing species have fixed more slightly deleterious alleles than outcrossers, but positive selection seems to be more efficient in outcrossers than in selfers. This might obscure the putative difference in fixation rates of slightly deleterious alleles.

Contrasts in substitution patterns in short versus long arms of chromosome 3 and telomeric versus centromeric regions were used to estimate the impact of variation in local recombination rates on rates of protein evolution. We found no difference between short arm and long arm regions, but the ω ratio is higher in centromeric than in telomeric regions both in selfing (0·104 and 0·077, respectively; Table 1) and in outcrossing lineages (0·169 and 0·125, respectively; Table 1). However, in both cases, model Mc-1 is not significantly better than model Mc-0 (Table 1).

https://doi.org/10.1017/S0016672307009032 Published online by Cambridge University Press

Table 1. *Summary of maximum likelihood estimates of the ratio of non-synonymous to synonymous substitutions (ω) in different models contrasting selfing and outcrossing lineages (model Mb-3 against model Mb-2, 1 d.f.) and highly and weakly recombining regions (model Mc-1 against model Mc-2, 2 d.f.)*

Sequence set		All genes	Chr. 3 telomeric	Chr. 3 centromeric	Chr. 3 short arm	Chr. 3 long arm
No. of genes		52	12	34	11	35
Size (bp)		31 218	6531	22 425	8592	19 161
Model M0	ω	0·160	0·136	0·161	0·159	0·158
	lnL	−59042·70	−11043·52	−38485·80	−14757·85	−32805·47
Model Mb-2	ω_0	0·175	0·159	0·169	0·171	0·167
	ω_1	0·140	0·110	0·150	0·144	0·146
	lnL	−59040·22	−11042·40	−38485·33	−14757·47	−32804·95
Model Mb-3	ω_0	0·174	0·158	0·168	0·170	0·167
	ω_{self}	0·105	0·074	0·101	0·099	0·113
	ω_{out}	0·150	0·120	0·164	0·155	0·156
	lnL	−59038·56	−11041·90	−38483·36	−14756·85	−32804·14
LRT		3·33	1·01	3·94	1·22	1·63
P value		0·068	0·316	**0·047**	0·269	0·202
Model Mc-0	ω_0		0·155		0·159	
	lnL		−46002·94		−44164·85	
Model Mc-1	$\omega_{high}/\omega_{low}$		0·136/0·161		0·150/0·164	
	lnL		−46002·42		−44164·59	
LRT			1·03		0·53	
P value			0·309		0·468	

Table 2. *Summary of maximum likelihood of site models and branch-site models. Model M8 is tested against model M7 (1 d.f.). The clade model is tested against model M1a (3 d.f.). In models M7 and M8 the ω ratio of sites under purifying selections follows a beta distribution with parameters p and q (column 2, Beta: p/q)*

Model M7	Beta: p/q	0·013/0·070
	lnL	−54007·67
Model M8	Beta: p/q	9·72/99·00
	$p_1/\omega_1 > 1$	2·5 %/2·88
	lnL	−54005·34
LRT		4·66
P value		**0·031**
Model M1a	$p_0/\omega_0 < 1$	84 %/0·00
	$p_1/\omega_1 = 1$	16 %
	lnL	−54006·58
Clade model	$p_0/\omega_0 < 1$	97 %/0·085
	$p_1/\omega_1 = 1$	0
	$p_2/\omega_{2out}/\omega_{2self}$	3 %/ω_{2out} = 3·01/ω_{2self} = 1·01
	lnL	−54001·73
LRT		9·69
P value		**0·021**
Clade model (alternative assignation)	p_0/ω_0	86 %/0·003
	$p_1/\omega_1 = 1$	11·80 %
	$p_2/\omega_{2out}/\omega_{2self}$	2·1 %/ω_{2out} = 2·73/ω_{2self} = 0·00
	lnL	−54002·10
LRT		8·95
P value		**0·030**

(ii) *Base composition and codon usage*

Current GC and GC3 values are very similar in all species, being only slightly higher in the two outcrossers than in the two selfers (Table 3). GC* values for each branch estimated on all concatenated genes are presented in Fig. 1. For each codon position (GC1*, GC2* and GC3*) and for each genomic region, outcrossing lineages have higher GC* values than selfing lineages, but in numerous cases the number of variable sites is not large enough to result in a significant difference (Table 2). Highly recombining regions (telomeric region, short arm) also show higher GC* and GC3* values than weakly recombining ones (centromeric region, long arm), in both selfing and outcrossing lineages (Table 2). When pooling substitutions from all branches, GC3* is significantly higher ($P = 0{\cdot}02$) on the short arm of chromosome 3 (0·53) than on the long arm (0·43). The same trend is observed between telomeric (GC3* = 0·53) and centromeric regions (GC3* = 0·45), but the difference is not significant ($P = 0{\cdot}097$). Overall, these results suggest that GC enrichment is positively correlated with the efficacy of recombination, both among (outcrossing vs selfing) and within genomes (recombination gradient).

Because preferred codons in *Triticeae* species typically end in G or C (Kawabe & Miyashita, 2003; Liu & Xue, 2005; Wang & Roossinck, 2006), results on GC3* can be due to higher codon bias in highly recombining genomes and regions. Accordingly, *Fop* is

https://doi.org/10.1017/S0016672307009032 Published online by Cambridge University Press

Table 3. *GC content and GC evolution* (*GC**) *in the two self-fertilizing* (T. monococcum *and* T. urartu) *and the two outcrossing species* (S. cereale *and* Ae. speltoides)

Sequence	MS lineages	GC1	GC2	GC3	GC	GC1*	GC2*	GC3*		GC*	
All nuclear genes (29 673 bp)	Outcrossing	0·535	0·403	0·480	0·473	0·454 (81)	0·375 (62)	0·480 (352)		0·453 (495)	
	Self-fertilizing	0·535	0·402	0·475	0·471	0·362 (18)	0·249 (15)	0·366 (109)	$p=0·06$	0·353 (142)	$\mathbf{p=0·03}$
Chr. 3 centromeric (22 437 bp)	Outcrossing	0·548	0·384	0·487	0·473	0·426 (66)	0·384 (50)	0·477 (253)		0·443 (369)	
	Self-fertilizing	0·548	0·383	0·487	0·473	0·310 (11)	0·210 (10)	0·358 (78)	$p=0·08$	0·326 (99)	$\mathbf{p=0·04}$
Chr. 3 telomeric (6531 bp)	Outcrossing	0·542	0·413	0·488	0·481	0·611 (14)	0·375 (13)	0·528 (82)		0·511 (109)	
	Self-fertilizing	0·543	0·413	0·482	0·479	0·283 (4)	0·588 (3)	0·463 (25)	$p=0·62$	0·449 (32)	$p=0·53$
Chr. 3 long arm (19 161 bp)	Outcrossing	0·542	0·408	0·494	0·481	0·461 (52)	0·418 (35)	0·464 (213)		0·452 (300)	
	Self-fertilizing	0·541	0·407	0·491	0·480	0·336 (10)	0·369 (13)	0·352 (72)	$p=0·10$	0·352 (95)	$p=0·08$
Chr. 3 short arm (8592 bp)	Outcrossing	0·554	0·395	0·478	0·476	0·473 (24)	0·376 (25)	0·529 (112)		0·496 (161)	
	Self-fertilizing	0·553	0·395	0·475	0·474	0·202 (6)	0·000 (2)	0·455 (29)	$p=0·50$	0·385 (37)	$p=0·20$

Values within parentheses indicate the number of substitutions used to compute GC*. Chi-square tests were performed using contingency tables of the number of AT→GC and GC→AT substitutions in the two categories tested.

Table 4. *Stationary optimal codon frequency* (*Fop**) *in the two self-fertilizing* (T. monococcum *and* T. urartu) *and the two outcrossing species* (S. cereale *and* Ae. speltoides)

	*Fop**			
Sequence set	All branches	Outcrossing	Self-fertilizing	
All genes		0·339 (242)	0·235 (77)	$p=0·057$
Chr. 3 centromeric	0·300 (501)	0·336 (163)	0·202 (50)	$\mathbf{p=0·041}$
Chr. 3 telomeric	0·384 (149)	0·375 (55)	0·377 (17)	$p=0·988$
	$\mathbf{p=0·044}$	$p=0·585$	$p=0·123$	
Chr. 3 long arm	0·301 (439)	0·324 (138)	0·218 (44)	$p=0·130$
Chr. 3 short arm	0·359 (210)	0·385 (80)	0·313 (24)	$p=0·496$
	$p=0·118$	$p=0·344$	$p=0·341$	

Values within parentheses indicate the number of substitutions used to compute *Fop**. Chi-square tests were performed using contingency tables of the number of preferred→unpreferred and unpreferred→preferred substitutions in the two categories tested.

slightly higher in outcrossing species (35·4%) than in self-fertilizing ones (35·1%). As for GC content, these values are very similar, but *Fop** values are much more contrasted (Table 4). As theoretically predicted, *Fop** is higher in outcrossing than in selfing lineages but the difference is only significant for the centromeric genes ($P=0·04$), and marginally significant when all genes are concatenated ($P=0·057$). *Fop** is also higher in highly than in weakly recombining regions. Pooling substitutions from all the branches, the difference is significant between centromeric and telomeric regions ($P=0·04$), but not significant between the long and short arm ($P=0·118$) (Table 4).

https://doi.org/10.1017/S0016672307009032 Published online by Cambridge University Press

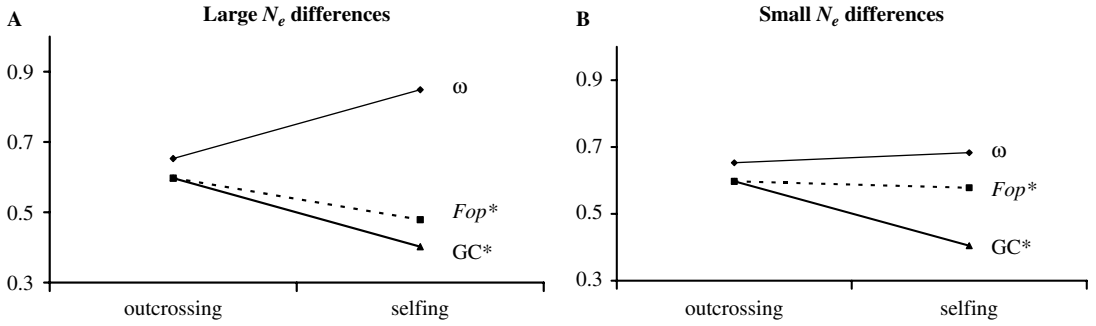

Fig. 4. Theoretical ω ratio for slightly deleterious alleles, *Fop** and GC*, computed in an outcrossing species and in a highly self-fertilizing species (selfing rate: $S=95\%$) with different effective population sizes. Effective population sizes are $N_e=N_{out}$ and $N_e=N_{self}(1-S/2)$ for outcrossing and self-fertilizing species, respectively. N_{self} can be lower than N_{out} if factors other than the automatic effect of selfing $(1-S/2)$ reduce N_e (bottlenecks, hitch-hiking effects). For co-dominant mutations and $N_{self}=N_{out}$, selection is independent of S (see Charlesworth, 1992), and $\omega=\frac{4Ns}{e^{4Ns}-1}$ (Charlesworth, 1992), and $Fop^*=\frac{e^{4Ns}}{u/v+e^{4Ns}}$ (Bulmer, 1991); $GC^*=\frac{e^{2N\gamma(1-S)(2-S)}}{u/v+e^{2N\gamma(1-S)(2-S)}}$ (Marais *et al.*, 2004). N is N_{out} in outcrossers and N_{self} in selfers; s is the selection coefficient against deleterious alleles (ω) or in favour of preferred codons (*Fop**); γ is the intensity of bGC; u is the mutation rate towards unpreferred codons (*Fop**) or towards AT alleles (GC*); and v is the reverse mutation rate. For the two graphs, $S=0{\cdot}95$, $s=\gamma=0{\cdot}0002$, and $u/v=1{\cdot}5$. (*A*) $N_{out}=1000$ and $N_{self}=400$, that is $N_e(\text{self})\approx 200$. (*B*) $N_{out}=1000$ and $N_{self}=900$, that is $N_e(\text{self})\approx 450$.

Given the strong correlation between *Fop* and GC, these results could be due to the sole effect of biased gene conversion (or biased mutation) rather than selection on codon usage. If selection on codon usage occurs, we would predict that C*, but not G*, should vary with mating system and recombination in four-fold and six-fold degenerated codons because C is the preferred base at the third position in those codons. Using all concatenated genes, G* is almost constant ($P=0{\cdot}87$) in outcrossing (G*$=0{\cdot}29$) and in selfing lineages (G*$=0{\cdot}28$). C* is more variable ($0{\cdot}20$ and $0{\cdot}13$, respectively), but the difference is not significant ($P=0{\cdot}20$). G* does not differ significantly either between short and long arm (G*$=0{\cdot}24$; G*$=0{\cdot}25$; $P=0{\cdot}87$) or between telomeric and centromeric regions (G*$=0{\cdot}32$; G*$=0{\cdot}23$; $P=0{\cdot}46$). C* differs significantly between short and long arm (C*$=0{\cdot}25$; C*$=0{\cdot}16$; $P=0{\cdot}02$), but not between telomeric and centromeric regions (C*$=0{\cdot}23$; C*$=0{\cdot}17$; $P=0{\cdot}15$). These results weakly support selection being less efficient in weakly recombining than in highly recombining genomes or regions.

(iii) *Protein and GC evolution*

GC1* and GC2* shared a pattern similar to GC3*, suggesting that bGC should be effective and could affect protein evolution, partially masking selective effects. To test this hypothesis, we computed GC* on the whole tree for every gene. We then separated the genes into two groups: genes having lower GC* than the median value (45%) and genes having higher GC* than the median. The two groups had similar size (16 725 and 13 776 bp, respectively). First, we re-ran models Mc-0 and Mc-1 to compare these two gene categories, substituting high and low GC* categories for high and low recombination categories. ω is significantly ($P=0{\cdot}018$) lower in the low GC* group ($\omega=0{\cdot}127$) than in the high GC* group ($\omega=0{\cdot}172$). Second, we re-ran the *branch* model (model 2) analysis on the two GC* groups. In the low GC* group, there is no significant difference ($P=0{\cdot}664$) between $\omega_{self}=0{\cdot}111$ and $\omega_{out}=0{\cdot}126$. On the contrary, in the high GC* group, $\omega_{self}=0{\cdot}099$ and $\omega_{out}=0{\cdot}180$ are significantly different ($P=0{\cdot}032$). These results suggest that GC evolution affects protein evolution, at least in outcrossers.

4. Discussion

We studied the impact of selfing and recombination on molecular evolution patterns in four *Triticeae* species. We aimed to test whether selfing and low recombination reduce the efficacy of selection and the strength of bGC. Selection effects scale in N_es, where s is the selection coefficient, while bGC effects scale in $N_e\gamma(1-S)$, where γ is the intensity of bGC and S the selfing rate (Marais *et al.*, 2004). Selfing on the whole-genome scale and recombination at a more local scale affect selection efficacy through N_e variation only, but they affect bGC through both N_e and the effective intensity of bGC, $\gamma(1-S)$. BGC depends on the efficacy of recombination (Marais, 2003) and is likely to vary dramatically from γ to 0 depending on the mating system of the species. Fig. 4 illustrates how ω, *Fop** and GC* are affected by mating systems when reduction in N_e due to selfing is large (Fig. 3*A*: $N_e(\text{self})/N_e(\text{out})=0{\cdot}2$), or small (Fig. 3*B*: $N_e(\text{self})/N_e(\text{out})=0{\cdot}45$, that is just below the $\frac{1}{2}$ threshold). We thus expect a stronger effect of mating system and recombination on bGC than on selection efficacy.

https://doi.org/10.1017/S0016672307009032 Published online by Cambridge University Press

(i) *Mating systems and recombination affect base composition*

Results on GC content and codon usage fit well with the theoretical predictions on mating system and recombination effects. GC* and *Fop** are higher in outcrossing than in self-fertilizing species, and in highly than in weakly recombining regions. However, differences were not all significant because of the lack of statistical power for some comparisons (Tables 3 and 4). Patterns of GC3* and *Fop** can be explained either by selection on codon usage or by bGC (or both). Because GC1* and GC2* patterns fit theoretical predictions as well as GC3*, bGC probably drives, at least partly, the evolution of GC content in outcrossing species and in genomic regions with high recombination rates. We also found some evidence for selection on codon usage by contrasting G* and C* in four-fold and six-fold codons. However, this effect is weak: we found no significant mating system effect on preferred codon selection, and a significant but weak effect of recombination only when contrasting centromeric versus telomeric regions. Finally, we cannot rule out mutational bias to explain our results. However, such a bias should be mating system dependent, which has not been documented. Overall, our results are better explained by bGC than by other factors, but polymorphism data in non-coding regions would be necessary to disentangle the different hypotheses.

The bGC interpretation is consistent with the highly significant differences in GC content found between self-fertilizing and outcrossing species detected among *Gramineae* in both coding and non-coding regions (Glémin *et al*., 2006) and among *Triticeae* species in EST (Akhunov *et al*., 2003*a*). It is surprising, however, to have found a recombination effect even in self-fertilizing species (Table 3). Marais *et al*. (2004) showed that GC* should almost not vary with recombination in highly self-fertilizing species, which could explain the lack of variation in GC content observed along the *A. thaliana* genome. We hypothesize that the *Triticum* species studied here exhibit a lower selfing rate than *A. thaliana*, and that very strong recombination gradients, such as those observed in several *Triticeae* species (Lukaszewski, 1992; Lukaszewski & Curtis, 1993), leave a genomic signature even in self-fertilizing species. For example, RFLP polymorphism is 12 to 25 times higher in telomeric than in centromeric regions in some self-fertilizing *Aegilops* species (Dvorak *et al*., 1998).

(ii) *Mating systems and recombination do not clearly affect protein evolution*

Contrary to expectations, we found no evidence that the two self-fertilizing species *T. monococcum* and *T. urartu* have fixed more slightly deleterious mutations than the two outcrossing species *S. cereale* and *Ae. speltoides*. Neither gene-by-gene analysis (Fig. 3) nor analyses of concatenated genes (Table 1) showed a significantly higher ω_{self} than ω_{out}, and the reverse tendency is even observed. Several reasons can be invoked to explain the lack of evidence of reduced selection efficacy in the self-fertilizing species studied. First, the four *Triticeae* species studied here diverged recently. Polymorphism for deleterious alleles, especially in outcrossing species, can therefore account for a part of the divergence estimates, since only one sequence per gene was produced for each species. In such cases we would expect higher ω ratios in terminal than in internal branches because π_N/π_S ratios are expected to be higher than d_N/d_S ratios under purifying selection. This could especially be the case for rye because of recent increased drift due to domestication. However, we found the reverse pattern: rye has a lower ω than *Ae. speltoides* (see Fig. 1). Intraspecific polymorphism alone cannot explain our results. The recent origin of selfing was also invoked to explain similar results in the comparison of *A. thaliana* and *A. lyrata* (Wright *et al*., 2002). We cannot rule out this hypothesis for the two *Triticum* species. As selfing can rapidly and recurrently evolve, transition from outcrossing to selfing may have occurred independently and recently in the two species, although this scenario is not the most parsimonious. Similarly, the differences in branch lengths between selfing and outcrossing lineages can bias the results. However, if selfing was that recent (and correspondingly the selfing branch lengths too short), it should not have affected base composition dynamics. Gene flow among these species complex may also have obscured the expected pattern, but once again GC content dynamics should also have been obscured as well. Together, results on protein and base composition evolution suggest that Fig. 4*B* matches our dataset better than Fig. 4*A*. Mating systems apparently weakly affect the efficacy of selection whereas GC dynamics is more strongly affected, at least partly through bGC effects.

Finally, we suggest that other processes can obscure the predictions. First, bGC could affect protein evolution (see below). Second, we also found an excess of non-synonymous substitutions in outcrossing lineages ($p_2 = 3\%$ with $\omega_{2out} = 3{\cdot}02$ in outcrossers vs $\omega_{2self} = 1{\cdot}01$ in selfers), suggesting that positive selection is more efficient for these species than in the two self-fertilizing species. Short branches of self-fertilizing lineages could explain why we did not detect positive selection in these lineages. However, if we consider all internal branches as selfers (Fig. 2) we still detect positive selection in the outcrossing lineages only ($p_2 = 2{\cdot}1\%$ with $\omega_{2out} = 2{\cdot}73$ vs $\omega_{2self} = 0$ in selfers, p value $= 0{\cdot}030$). This suggests that

https://doi.org/10.1017/S0016672307009032 Published online by Cambridge University Press

outcrossing and self-fertilizing lineages actually have different selection patterns. Few adaptive substitution events would explain why ω_{out} is slightly higher than ω_{self}. Given these values, we can roughly estimate a mean corrected ω ratio, ω'_{tot}, by excluding sites under putative positive selection. We can use $\omega_{tot} = p_2\ \omega_2 + (1 - p_2)\ \omega'_{tot}$ with ω_{tot} standing for ω_{self} and ω_{out} of model Mb-2 (Table 1). It gives, $\omega'_{out} = 0{\cdot}060$ instead of $\omega_{out} = 0{\cdot}150$, and $\omega'_{self} = 0{\cdot}078$ instead of $\omega_{self} = 0{\cdot}105$, which matches theoretical predictions slightly better.

We also tried to detect higher rates of protein evolution in low recombining chromosomic regions than in higher recombining ones by contrasting both long versus short arm and proximal versus distal regions. The ω ratio is very similar between short and long arms, while the difference is higher, although non-significant, between telomeric and centromeric regions. Few studies have demonstrated a positive relation between recombination and the efficacy of selection through d_N/d_S measures (see for instance Haddrill *et al*., 2007; Pal *et al*., 2001). In *Drosophila*, genomic regions with no crossing-over experience a severe reduction in the efficacy of selection, but even a low frequency of crossing over in other regions seems to be enough to maintain the efficacy of selection (Haddrill *et al*., 2007). In *Triticeae* species, recombination gradients are very steep with virtually no recombination occurring around the centromeres (Lukaszewski, 1992; Lukaszewski & Curtis, 1993). This would explain why selection seems to be somewhat weaker in centromeric than in telomeric regions, but not between short and long chromosome arms, between which differences in average recombination rates are lower. These results must be viewed with caution because of possible assignation errors. Synteny between rice chromosome 1 and wheat chromosome 3 is sometimes broken and physical assignation of our locus using wheat deletion lines revealed a few discrepancies between rice and wheat locations, especially in the telomeric region of the short arm (Supplementary Table S1). This synteny erosion in the genome of wheat ancestors appears to be linked to recombination intensity and mating system (Akhunov *et al*., 2003*b*). Another limitation of this study is the under-representation of telomeric genes. Extensive gene sequencing along the same chromosome would allow a more robust analysis of the effect of recombination. Despite these limitations, it is worth noting that, contrary to protein evolution, GC content differences between genomic regions match well the theoretical predictions, suggesting that gene location assignation should be mainly correct.

(iii) *Does bGC affect protein evolution?*

We found that mating system and recombination affect GC content evolution even on first and second codon positions, and thus protein evolution. Accordingly, ω ratios are higher in regions of high than low GC*, especially in outcrossing lineages. Protein evolution could affect GC content but it is difficult to explain why changes in amino acid would preferentially result in AT→GC substitutions. We thus favour the reverse explanation that GC evolution drives protein evolution. Fast-evolving genomic regions exhibiting mostly AT→GC changes have also been observed in highly recombining regions in humans (Pollard *et al*., 2006) and in mice (Galtier & Duret, 2007). Positive selection is mostly invoked to explain such a lineage-specific acceleration of substitution rates in comparison with close relatives. Recently, Galtier & Duret (2007) proposed a neutral alternative explanation. The bGC molecular drive could overcome purifying selection, and lead to the fixation of G and C weakly deleterious mutations. BGC can thus increase the mutation load, as previously suggested by Bengtsson (1990). If high GC* is a consequence of bGC hotspots rather than selection on codon usage, our results are consistent with a bGC-associated mutation load in the two outcrossing species *S. cereale* and *Ae. speltoides*. According to Galtier & Duret (2007), outcrossing species, but not self-fertilizing ones, would suffer from a 'genomic Achille's heel'. BGC could also lead to spurious detection of sites under positive selection in outcrossing lineages. This unexpected effect contributes to obscuring the effect of mating system and recombination on patterns of selection. Including non-coding regions in such molecular studies would help to characterize parameters, such as the strength of bGC, to be included in the null model to test more accurately the effect of selection.

(iv) *Genomic approaches on mating system evolution*

Such genomic approaches shed light on the evolution of mating systems. Self-fertilization has been suggested to be an evolutionary dead end because of the accumulation of slightly deleterious mutations, and because low genetic diversity may preclude adaptation (see review in Takebayashi & Morrell, 2001). Very few studies have provided support for the premise of this hypothesis, namely reduced selection efficacy in selfers. Our results gave no clear support for an increased drift load in selfers, partly because other processes may interfere. We suggest that, surprisingly, outcrossers may also suffer from an increase mutation load due to bGC, not to drift. We also found that positive selection could be more efficient in outcrossing than in self-fertilizing species, in agreement with the dead end theory. Wright *et al*. (2002) did not find any evidence for an increased load in the selfer *A. thaliana* compared with the outcrosser *A. lyrata*, but they did not test for difference

in adaptive evolution nor for interference with bGC. Similar studies in other species are needed to evaluate the generality of these results. Larger genomic data and better knowledge of the distribution of local recombination rates would also help in estimating the genetic load associated with each mating system. The dead end theory also posits that self-fertilizing species should be of recent origin, and that transitions from selfing to outcrossing are rare (Takebayashi & Morrell, 2001). Inferring shifts in mating systems and their timing would thus be useful. If GC* values are well correlated with effective recombination rates (Meunier & Duret, 2004) and thus to mating systems, we hypothesize that GC3* or GC* computed on non-coding sequences could be useful statistics to infer mating system evolution along phylogenies.

We thank A. Tsitrone and J. Ronfort for helpful discussion, and N. Galtier, B. Mable, and two anonymous reviewers for comments on the manuscript. S.G. also thanks D. Charlesworth for her help on his first study on molecular evolution. This work was supported by the French Agence Nationale de la Recherche (ANR 'Exegese-Blé' project) and by the French Institut National de la Recherche Agronomique (INRA 'Tritipol' project). This is publication ISE-M 2007-152 of the Institut des Sciences de l'Evolution de Montpellier.

References

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. *Genetics* **139**, 1067–1076.

Akhunov, E. D., David, J. L., Chao, S., Lazo, G., Anderson, O. D., *et al.* (2003*a*). GC composition and codon usage in genes of inbreeding and outcrossing *Triticeae* species. In *Tenth International Wheat Genetics Symposium*, pp. 203–206. Italy: Paestum.

Akhunov, E. D., Goodyear, A. W., Geng, S., Qi, L. L., Echalier, B., Gill, B. S., Miftahudin, T., Gustafson, J. P., Lazo, G., Chao, S., *et al.* (2003*b*). The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Research* **13**, 753–763.

Baker, H. G. (1955). Self-compatibility and establishment after 'long-distance' dispersal. *Evolution* **9**, 347–348.

Barakat, A., Carels, N. & Bernardi, G. (1997). The distribution of genes in the genome of gramineae. *Proceedings of the National Academy of Sciences of the USA* **94**, 6857–6861.

Bengtsson, B. O. (1990). The effect of biased conversion on the mutation load. *Genetical Research* **55**, 183–187.

Bielawski, J. P. & Yang, Z. (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *Journal of Molecular Evolution* **59**, 121–132.

Bulmer, M. (1991). The selection–mutation–drift theory of synonymous codon usage. *Genetics* **129**, 897–907.

Carels, N. & Bernardi, G. (2000). Two classes of genes in plants. *Genetics* **154**, 1819–1825.

Charlesworth, B. (1992). Evolutionary rates in partially self-fertilizing species. *The American Naturalist* **140**, 126–148.

Charlesworth, B., Morgan, M. T. & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.

Charlesworth, D. & Wright, S. I. (2001). Breeding systems and genome evolution. *Current Opinion in Genetics & Development* **11**, 685–690.

Dvorak, J., Diterlizzi, P., Zhang, H. B. & Resta, P. (1993). The evolution of polyploid wheats: identification of the A-genome donor species. *Genome* **36**, 21–31.

Dvorak, J., Luo, M. C. & Yang, Z. L. (1998). Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. *Genetics* **148**, 423–434.

Endo, T. R. & Gill, B. S. (1996). The deletion stocks of common wheat. *Journal of Heredity* **87**, 295–307.

Felsenstein, J. (1989). PHYLIP: phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.

Galtier, N. & Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics* **23**, 273–277.

Glémin, S., Bazin, E. & Charlesworth, D. (2006). Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proceedings of the Royal Society of London, Series B* **273**, 3011–3019.

Guindon, S. E. P. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696–704.

Haddrill, P. R., Halligan, D. L., Tomaras, D. & Charlesworth, B. (2007). Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biology* **8**, R18.

Hamrick, J. L. & Godt, M. J. W. (1996). Effects of life history traits on genetic diversity in plants species. *Philosophical Transactions of the Royal Society of London, Series B* **351**, 1291–1298.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R. & Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences of the USA* **99**, 8133–8138.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution* **2**, 13–35.

Ingvarsson, P. K. (2002). A metapopulation perspective on genetic diversity and differentiation in partially self-fertilizing plants. *Evolution* **56**, 2368–2373.

Kawabe, A. & Miyashita, N. T. (2003). Patterns of codon usage bias in three dicot and four monocot plant species. *Genes and Genetic Systems* **78**, 343–352.

Kunzel, G., Korzun, L. & Meister, A. (2000). Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* **154**, 397–412.

Liu, Q. & Xue, Q. (2005). Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *Journal of Genetics* **84**, 55–62.

Lukaszewski, A. J. (1992). A comparision of physical distribution of recombination in chromosome 1R in diploid rye and in hexaploid triticale. *Theoretical and Applied Genetics* **83**, 1048–1053.

Lukaszewski, A. J. & Curtis, C. A. (1993). Physical distribution of recombination in B-genome chromosomes of tetraploid wheat. *Theoretical and Applied Genetics* **86**, 121–127.

Lundqvist, A. (1954). Studies on self-sterility in rye, *Secale cereale* L. *Hereditas* **40**, 278–294.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics* **19**, 330–338.

Marais, G., Charlesworth, B. & Wright, S. I. (2004). Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. *Genome Biology* **5**, R45.

Maynard-Smith, J. & Haigh, D. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35.

Meunier, J. & Duret, L. (2004). Recombination drives the evolution of GC-content in the human genome. *Molecular Biology and Evolution* **21**, 984–990.

Munkvold, J. D., Greene, R. A., Bertmudez-Kandianis, C. E., La Rota, C. M., Edwards, H., Sorrells, S. F., Dake, T., Benscher, D., Kantety, R., Linkiewicz, A. M., *et al*. (2004). Group 3 chromosome bin maps of wheat and their relationship to rice chromosome 1. *Genetics* **168**, 639–650.

Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929.

Nybom, H. (2004). Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. *Molecular Ecology* **13**, 1143–1155.

Pal, C., Papp, B. & Hurst, L. D. (2001). Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Molecular Biology and Evolution* **18**, 2323–2326.

Paux, E., Roger, D., Badaeva, E., Gay, G., Bernard, M., Sourdille, P. & Feuillet, C. (2006). Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *The Plant Journal* **48**, 463–474.

Pollak, E. (1987). On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics* **117**, 353–360.

Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J. S., Bejerano, G., Baertsch, R., *et al*. (2006). Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* **2**, e168.

Qi, L. L., Echalier, B., Chao, S., Lazo, G. R., Butler, G. E., Anderson, O. D., Akhunov, E. D., Dvorak, J., Linkiewicz, A. M., Ratnasiri, A., *et al*. (2004). A chromosome bin map of 16 000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**, 701–712.

Rota, M. & Sorrells, M. E. (2004). Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Functional and Integrative Genomics* **4**, 34–46.

Safar, J., Bartos, J., Janda, J., Bellec, A., Kubalakova, M., Valarik, M., Pateyron, S., Weiserova, J., Tuskova, R., Cihalikova, J., *et al*. (2004). Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *The Plant Journal* **39**, 960–968.

Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, J., Niimura, Y., Cheng, Z., Nagamura, Y., *et al*. (2002). The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312–316.

Schoen, D. J. & Brown, A. H. D. (1991). Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proceedings of the National Academy of Sciences of the USA* **88**, 4494–4497.

Sears, E. R. (1954). The aneuploid of common wheat. *Missouri Agricultural Experiment Station Research Bulletin* **572**, 1–58.

Sears, E. R. & Sears, L. (1978). The telocentric chromosomes of common wheat. In *Proceedings of the Fifth International Wheat Genetics Symposium* (ed. S. Ramanujams), pp. 389–407. New Delhi, India: Indian Agricultural Research Institute.

Sorrells, M. E., La Rota, M., Bermudez-Kandianis, C. E., Greene, R. A., Kantety, R., Munkvold, J. D., Miftahudin, T., Mahmoud, A., Ma, X. F., Gustafson, P. J., *et al*. (2003). Comparative DNA sequence analysis of wheat and rice genomes. *Genome Research* **13**, 1818–1827.

Staden, R., Judge, D. P. & Bonfield, J. K. (2001). Sequence assembly and finishing methods. *Methods of Biochemical Analysis* **43**, 303–322.

Sueoka, N. (1962). On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the USA* **48**, 582–592.

Takebayashi, N. & Morrell, P. L. (2001). Is self-fertilization an evolutionary dead end? Revisiting an old hypothesis with genetic theories and a macroevolutionary approach. *American Journal of Botany* **88**, 1143–1150.

Wang, L. J. & Roossinck, M. J. (2006). Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants. *Plant Molecular Biology* **61**, 699–710.

Wong, G. K., Wang, J., Tao, L., Tan, J., Zhang, J., Passey, D. A. & Yu, J. (2002). Compositional gradients in *Gramineae* genes. *Genome Research* **12**, 851–856.

Wright, S. I., Lauga, B. & Charlesworth, D. (2002). Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*. *Molecular Biology and Evolution* **19**, 1407–1420.

Wright, S. I., Iorgovan, G., Misra, S. & Mokhtari, M. (2007). Neutral evolution of synonymous base composition in the *Brassicaceae*. *Journal of Molecular Evolution* **64**, 136–141.

Yamane, K. & Kawahara, T. (2005). Intra- and interspecific phylogenetic relationships among diploid *Triticum-Aegilops* species (*Poaceae*) based on base-pair substitutions, indels, and microsatellites in chloroplast noncoding sequences. *American Journal of Botany* **92**, 1887–1898.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**, 568–573.

Yang, Z. & Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**, 409–418.

Yang, Z. H. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555–556.

Yang, Z. & Swanson, W. J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular Biology and Evolution* **19**, 49–57.

Yang, Z., Wong, W. S. & Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* **22**, 1107–1118.