GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis

Sylvain Glémin[1], Yves Clément[1,2], Jacques David[2], and Adrienne Ressayre[3]

[1] Institut des Sciences de l'Evolution de Montpellier, Unité Mixte de Recherche 5554, Centre National de la Recherche Scientifique, UMR 5554 CNRS, Université Montpellier 2, F-34095 Montpellier, France
[2] Montpellier SupAgro, Unité Mixte de Recherche 1334 Amélioration Génétique et Adaptation des Plantes Méditerranéennes et Tropicales, F-34398 Montpellier, France
[3] INRA, UMR de Génétique Végétale, INRA/CNRS/Univ Paris-Sud/AgroParistech, Ferme du Moulon, F-91190 Gif sur Yvette, France

In angiosperms (as in other species), GC content varies along and between genes, within a genome, and between genomes of different species, but the reason for this distribution is still an open question. Grass genomes are particularly intriguing because they exhibit a strong bimodal distribution of genic GC content and a sharp 5′–3′ decreasing GC content gradient along most genes. Here, we propose a unifying model to explain the main patterns of GC content variation at the gene and genome scale. We argue that GC content patterns could be mainly determined by the interactions between gene structure, recombination patterns, and GC-biased gene conversion. Recent studies on fine-scale recombination maps in angiosperms support this hypothesis and previous results also fit this model. We propose that our model could be used as a null hypothesis to search for additional forces that affect GC content in angiosperms.

GC content patterns and dynamics in coding and noncoding regions

GC content variation along genomes is a key feature of genomic organization and strongly varies between species. Much work has focused on the so-called 'isochore structure' (see Glossary) of mammalian genomes (i.e., the patchwork of GC-rich and GC-poor regions alternating along the genome) because it is associated with fundamental elements of genome organization [1]. For instance, GC-rich regions exhibit higher gene density and compactness [2], earlier replication timing [3], and higher recombination rates [4] relative to GC-poor regions. In mammals, GC content heterogeneity is found both in coding and noncoding regions, and the GC content of a gene is well correlated with the GC content of its surrounding regions [1] (Box 1). In flowering plants, genic GC content is also heterogeneous and has been found to be associated with some gene characteristics, similar to what has been observed in mammals (e.g., GC-rich genes tend to be more compact than GC-poor ones) [5]. However, angiosperm genomes also show unique features not shared with mammalian genomes.

Contrary to mammals, angiosperm genomes do no exhibit a clear isochore structure. Although GC content in exons and introns is positively correlated for individual genes [6,7], genic GC content is uncorrelated or at best weakly correlated with flanking noncoding GC content [7] (Box 1). This could be explained by the highly dynamic nature of angiosperm genomes, which can vary rapidly and widely in size and structure [8]. For instance, in the *Oryza* genus, the *ADH* region was massively shuffled by the replacement of an intergenic space, gene disruption, and movements mediated by transposable elements approximately 15 million years [9]. Loss of synteny in intergenic regions can even occur within a species, as in maize, with recombination concentrating within or near genes where homology is conserved [10]. Thus, large-scale genome rearrangements can erase a potential isochore structure.

As a consequence, the relation between GC content and recombination is not apparent in noncoding regions: it is

Glossary

Deamination of methyl CpG: chemical reaction that changes the methylated cytosine of a CpG dinucleotide into a thymine, thus causing a C to T mutation. Cytosine deamination typically produces a uracil base that is easily recognized by the repair machinery of the cell. CpG-associated C to T mutations occur about ten times more frequently than regular C to T mutations.
Isochores (or isochore structure): large regions of a genome (few 10 kb to 1 Mb) exhibiting local similarities in GC content. Isochores were initially found in mammals and birds, but also occur in other eukaryotes. In species with isochores, GC content in genic regions (especially GC3) is usually highly correlated with GC content in flanking sequences (at least at few kilobases; Box 1).
GC3: GC content at the third codon position in coding sequences. Given that most mutations occurring at third codon positions are synonymous, GC3 is much less constrained than GC1 and GC2 (GC content at first and second codon positions, respectively) and better reflects forces, other than selection, affecting base composition.
Meiotic distortion: biased segregation of alleles during meiosis, leading to an under or over-representation of the different alleles in gametes, compared with the expected Mendelian ratio (1:1). gBGC is a kind of meiotic distortion caused by the bias in repair mechanism associated with recombination (Box 2). Meiotic distortion can also be caused directly by a 'selfish' sequence that favors its own transmission at meiosis.
Synteny: conservation and co-localization of a group of two sequences, or more, in related taxa or in duplicated blocks within a genome. Genome rearrangements tend to disrupt synteny.

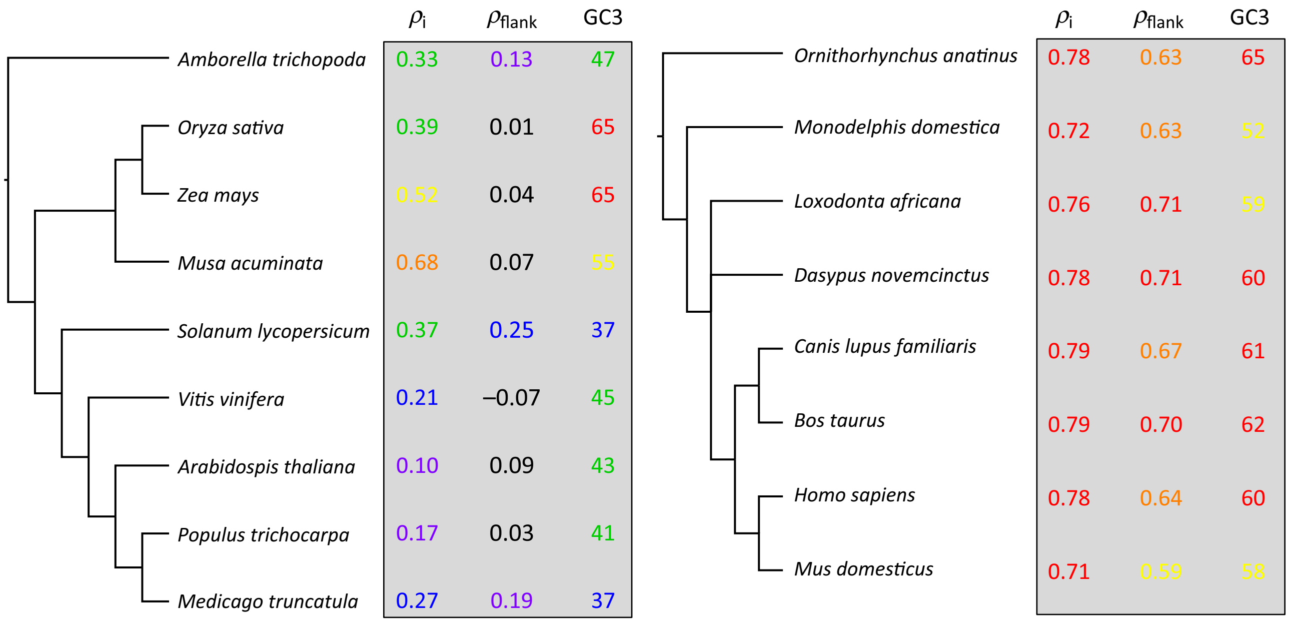
Corresponding author: Glémin, S. (sylvain.glemin@univ-montp2.fr).
Keywords: GC biased gene conversion; GC content; intron/exon structure; methylation; angiosperm genomes; recombination.

0168-9525/
http://dx.doi.org/10.1016/j.tig.2014.05.002

Box 1. Genomic patterns of GC content: angiosperms versus mammals

Mammalian genomes exhibit a strong heterogeneity in GC content at the scale of few tens of kilobases to one megabase, the so-called 'isochore structure' [1]. Some angiosperms (such as grasses) also exhibit strong heterogeneity in GC content. In coding regions, stronger variations between species were found in angiosperms [mean GC3 ranging from 36% to 68% for expressed sequence tag (EST) data] [14] than in mammals (mean GC3 ranging from 44% to 58% for a set of >1000 orthologous genes) [16] (Figure I). However, angiosperm genomes are thought to be less structured (e.g., [7]). To illustrate this point, we computed the correlation coefficients between GC3 and GC content in introns (GC_i) and flanking regions (GC_{flank}) for all protein-coding genes in nine angiosperms and eight mammals covering the phylogeny of the two groups (Figure I). In all mammals, strong, positive, and highly significant correlations between GC3 and GC_i (between 0.71 and 0.79) and between GC3 and GC_{flank} (between 0.59 and 0.71) were observed. In angiosperms, positive and significant correlations between GC3 and GC_i were also observed but they were lower than in mammals in general (between 0.11 and 0.68). However, correlation coefficients between GC3 and flanking GC content were low, especially in commelinids (Banana + grasses, here), which contrasts with the strong correlation between GC3 and GC_i observed in these species. This suggests that, unlike mammals, angiosperms are devoid of a clear isochore structure.

Similarly, the correlation between GC content and meiotic recombination is different between mammals and land plants. In human [4], mouse [22], and dog [58], recombination is strongly correlated with total GC content (mainly corresponding to noncoding regions). In angiosperms, no significant correlation was found between recombination and total GC content in *Arabidopsis thaliana* [24,59], *Oryza sativa* [12,24], *Populus trichocarpa* [24], *Sorghum bicolor* [24], and *Vitis vinifera* [24], and only a weak positive correlation in *Zea mays* [11] and even a negative correlation in *Medicago truncatula* [13]. However, in grasses, where gBGC is supposed to be strong, significant correlations were found in *Brachypodium distachyon*, *O. sativa*, and *Z. mays* when GC3 was used instead of total GC content [14].



	ρ_i	ρ_{flank}	GC3
Amborella trichopoda	0.33	0.13	47
Oryza sativa	0.39	0.01	65
Zea mays	0.52	0.04	65
Musa acuminata	0.68	0.07	55
Solanum lycopersicum	0.37	0.25	37
Vitis vinifera	0.21	–0.07	45
Arabidospis thaliana	0.10	0.09	43
Populus trichocarpa	0.17	0.03	41
Medicago truncatula	0.27	0.19	37

	ρ_i	ρ_{flank}	GC3
Ornithorhynchus anatinus	0.78	0.63	65
Monodelphis domestica	0.72	0.63	52
Loxodonta africana	0.76	0.71	59
Dasypus novemcinctus	0.78	0.71	60
Canis lupus familiaris	0.79	0.67	61
Bos taurus	0.79	0.70	62
Homo sapiens	0.78	0.64	60
Mus domesticus	0.71	0.59	58

TRENDS in Genetics

Figure I. Mean GC3 and Pearson correlation coefficients, ρ, between GC3 and GC_i (ρ_i) and GC_{flank} (ρ_{flank}). The transcript giving the longest coding DNA sequence was used for each gene. For each species, we computed for each gene GC3, the GC content of all its introns, and the GC content of flanking positions, taking 5 kb upstream and downstream of the complete transcript. Data for all mammals were downloaded from the Ensembl database [60]. Data for *Amborella trichopoda* were downloaded from the Plants Ensembl database [61], whereas data for other plants were downloaded from the Gramene database [62]. All data were downloaded using the BioMart interface [63]. Given the very large data sets, all correlations are significant. Colors for correlation coefficients: black: ρ <0.1, purple: 0.1 $\leq \rho$ <0. 2, blue: 0.2 $\leq \rho$ <0.3, green: 0.3 $\leq \rho$ <0.4, yellow: 0.5 $\leq \rho$ <0.6, orange 0.6 $\leq \rho$ <0.7, red: 0.7 $\leq \rho$. Colors for mean GC3: blue: ρ <40%, green: 40% $\leq \rho$ <50%, yellow: 50% $\leq \rho$ <60%, red: 60% $\leq \rho$.

very weak in maize [11], nonsignificant in rice [12], and nonsignificant or even negative (depending on the genomic scale) in *Medicago truncatula* [13], contrary to that observed in mammals [4]. However, as in mammals, a strong positive correlation between recombination and GC3 was found in maize, rice, and *Brachypodium distachyon* [14]. Likewise, in the grass genus *Festuca*, GC content is positively correlated with genome size, which is mainly driven by the invasion of GC-rich transposable elements [15]. This is the reverse of what has been found in mammals, where large genomes with low recombination rates per base pair are less GC rich compared with small genomes [16]. Thus, to study the mechanisms shaping variation in GC content at the nucleotide level in angiosperms, we need to focus on genic GC content to avoid the noisy signals arising from frequent reshuffling of noncoding DNA.

These coding regions also exhibit specific features that remain poorly understood. Much attention has been paid to the peculiar characteristics of grass genomes. In contrast to the eudicot genomes studied so far, GC content in coding regions of grass genomes is highly heterogeneous and GC3 exhibits a bimodal distribution [5,7,17]. In addition, grass genes exhibit a strong 5′ to 3′ negative GC content gradient, both in exons, where the GC3 gradient is the strongest, and in introns [18]. These features seem to set grasses apart from other species, both in the plant and animal kingdoms. However, a large survey across the seed plant phylogeny recently showed that many intermediates exist

Box 2. Mechanism and consequences of GC-biased gene conversion

gBGC is a process associated with recombination in several eukaryotes that favors the transmission of G and C alleles at meiosis (Figure I). This transmission bias was directly observed and estimated to be on the order of a few percent in yeast [26]. During meiosis, a double-strand break (DSB) initiates recombination and is followed by the invasion of the broken strand by a strand of the homologous chromosome. Thus, the strand experiencing the DSB is converted. In yeast, these conversion events are independent of GC content [27]. Then, strand invasion progresses on the flanking regions of the DSB point, forming heteroduplexes and creating mismatches at heterozygous sites. It is the mismatch repair system (MMR), which is thought to be biased towards G and C alleles [27], that favors their transmission.

Thus, gBGC is equivalent to meiotic distortion and it mimics selection in favor of G and C alleles [64]: it increases their frequency in populations and eventually leads to their fixation. Thus, highly recombining regions, especially around recombination hotspots, are enriched in G and C (e.g., [4,22]). At equilibrium, the expected GC content, *GC**, depends on the balance between mutational bias towards A and T alleles, λ, gBGC, *b*, and genetic drift determined by effective population size, N_e: $GC^* = 1/(1 + \lambda \times e^{-4N_e b})$. Even weak gBGC is sufficient to raise GC content much above the expected mutational equilibrium. For instance, using $\lambda = 2$ [65] and $4N_e b = 1.3$ as the average observed in the top 20% recombining regions in human [66] raises the GC content from 33% (mutational equilibrium, $b = 0$) to 65%. Given that gBGC mimics selection, a short episode of strong gBGC (as in a recombination hotspots) can quickly increase GC content, whereas, when a recombination hotspot disappears and gBGC stops, the return to equilibrium takes much longer (Figure II).

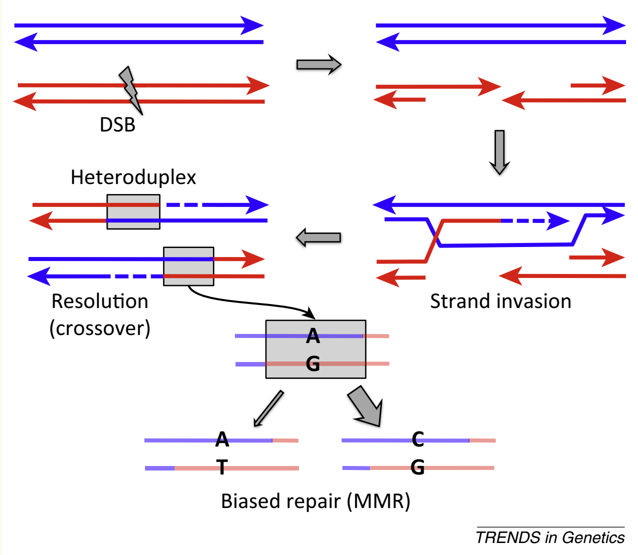


Figure I. Simplified model of recombination (in *Saccharomyces cerevisiae*). Only two homologous DNA double strands are represented (blue and red). Recombination initiates with a double-strand break (DSB) followed by strand invasion, which eventually leads to heteroduplex formation. For simplicity, noncrossover output is not represented. In the heteroduplex, mismatches occur at heterozygous sites. Repair of these mismatches is biased in favor of the bases G and C. The mismatch repair system (MMR) is thought to be the biased mechanism.

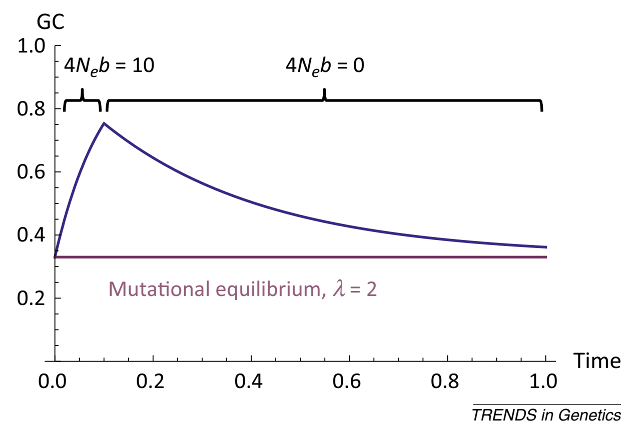


Figure II. Evolution of GC content after a short episode of strong GC-biased gene conversion (gBGC) ($4N_eb = 10$). Mutational bias towards AT is $\lambda = 2$, such that the mutational equilibrium GC content is 33%. Time is scaled by mutation rate.

between GC-rich and highly heterogeneous grass-like genomes with strong 5′–3′ gradients and GC-poor and homogeneous genomes with very weak gradients (such as in *Arabidopsis thaliana*) [14].

Forces affecting GC content evolution: the emerging role of GC-biased gene conversion

Over the past 10 years, GC-biased gene conversion (gBGC) has been established as the main process affecting GC content evolution in mammals (reviewed in [19]). This is a recombination-associated process that favors the fixation of GC alleles over AT ones because of biased mismatch repair following heteroduplex formation at meiosis [20] (Box 2). Within mammalian genomes, there is a strong positive relation between recombination landscapes, GC content, and GC content evolution, both at large megabase (Mb) scales [4,21] and at the scale of recombination hotspots [22,23]. Among mammals, GC content distribution is also correlated with karyotype structure and life-history traits, such as body mass [16]: small genomes, experiencing higher recombination rates per Mb, are more GC rich, as are genomes of small species, probably because their effective population sizes are higher and, hence, gBGC is more efficient (Box 2). There is also increasing evidence for a role of gBGC in many other groups of eukaryotes in addition to mammals [24,25], and gBGC has been demonstrated experimentally in yeast [26,27].

The causes of GC content variation in angiosperm genomes are less clearly established than in mammals, and the role of gBGC has only been investigated more recently [14,28–30]. To explain the peculiar characteristics of grass genomes, several studies have suggested a functional role for the distribution of GC content, linking GC content with regulation of gene expression or gene ontology, hence assuming direct selection on GC content (e.g., [7,31–33]). It was argued that differences in function and amino acid composition between GC-rich and GC-poor genes support a selection hypothesis [31]. However, this was also observed in mammalian genomes [34], whose GC content is mainly shaped by gBGC, where it was shown that gBGC can also affect amino acids [35,36]. Selection on codon usage for translational efficiency could shape GC3 patterns [37]. Differences in selection intensity among genes could explain within-genome variations in GC content, and variations in codon usage preferences could explain differences in GC content between species. However, although there is evidence for selection on codon usage in grasses [28], it does not appear to contribute significantly to shaping GC3 patterns because both lowly and highly expressed genes exhibit bimodal distributions [32]. It has also been suggested that GC content is selected for regulating gene expression, leading to two distinct classes of gene, at least in grasses [7,33]. However, the underlying mechanism behind this hypothesis is not known, and it is not clear why GC content in introns should also be selected for. Thus, we think that selective hypotheses are not clearly established and are currently insufficient to explain all the data adequately.

Alternatively, there is evidence that gBGC affects GC content in grasses [28,38,39] as well as in other groups of plants (e.g., [29,30]). In grasses, both GC content and GC content evolution correlate positively with recombination, both at third positions and in introns [14,28]. A fixation bias favoring GC alleles was also observed in grasses [28] and in *A. thaliana* [30]. Furthermore, a genome-wide analysis of recombination events in *A. thaliana* provides direct evidence of gBGC [40]: in gene conversion tracts, a significant excess of AT→GC over GC→AT conversion events was observed. The consequences of gBGC on GC content will depend on the frequency of conversion events at meiosis, the rate of which is debated [40,41]. However, even infrequent conversion events and tiny biases can have strong consequences in large populations because the impact of gBGC depends on the product of gBGC intensity and effective population size (Box 2).

One problem when trying to determine the causes of observed GC content patterns is the difficulty in untangling causes and consequences from the many correlations reported between genomic variables, such as gene length, expression level, methylation, recombination rate, and GC content. Moreover, recombination data, which are critical for assessing gBGC, were scarce, especially at small genomic scales. Recent descriptions of fine-scale recombination maps in *A. thaliana* [42] and in *Mimulus guttatus* [43], especially at the gene scale, bridge the gap between fine-scale GC content and recombination patterns, which are congruent with the hypothesis that recombination has a major role in shaping nucleotide landscapes in angiosperms. Here, we propose a unifying hypothesis to explain genic GC content variation within and between angiosperm genomes. We propose that gene structure and recombination patterns could be the main determinant of GC content variation through the effect of gBGC, both at the genic (5′–3′ gradient) and genome scale (distribution among genes).

Local patterns of recombination could explain 5′–3′ GC content gradients

One of the peculiar features initially observed in grass genomes is the strong negative 5′–3′ GC content gradient along genes [18], which is also found in most other angiosperms [14]. This gradient is most visible at third codon positions, but can also be found at other positions as well as in introns, although GC content is on average much lower in introns than in exons [6,18]. This strong 5′–3′ decreasing pattern is observed both along exonic and intronic positions in genes [6,18], except for short mono-exonic very GC-rich genes, for which GC3 remains very high (by definition) with a slight increase at the beginning of the gene [7].

In yeast, similar albeit less marked gradients have been documented, and they correlate with the 5′–3′ recombination gradient that has also been observed along genes [26]. This in agreement with the hypothesis that recombination gradients affect variations in base composition and codon usage along genes through the effect of gBGC [44]. Moreover, GC gradients in yeast genes are well fitted by a simple model that only takes into account the localization of meiotic double-strand breaks (initiation sites of recombination), the length of conversion tracts, and the intensity of gBGC [45]. In yeast, recombination patterns, gene structure, and gBGC could be the main determinant of variation in GC content along and between genes, especially GC3.

Extending this simple model to angiosperms has until recently been hampered by the lack of information about recombination patterns in angiosperms. However, two recent publications of detailed recombination maps at the gene scale for *A. thaliana* [42] and *M. guttatus* [43] strikingly support the possible involvement of recombination in shaping GC content along angiosperm genes. Indeed, both studies showed that, as in yeast, crossovers are associated with genes, and recombination hotspots are mainly localized around transcription start sites (TSSs). In *A. thaliana*, crossover frequency also increases around transcription termination sites (TTSs) but the effect is weaker than it is for TSSs. Overall, this generates a 5′–3′ recombination gradient along genes, as observed in yeast (Figure 1). Interestingly, *M. guttatus* also clearly shows a 5′–3′ GC content gradient [14], which parallels the recombination gradient. In addition, both studies found that recombination is on average lower in introns than in exons. This could partly explain the observation that intron gradients are usually less steep than GC3 gradients and the other observation that intronic GC content is usually lower than GC3, although this contrast could also be linked to exon and intron definitions ([46,47] and see below). In *M. guttatus*, recombination rates were estimated separately for each position of exon and intron (first, second, third. . .) [43]. The two staggered 5′–3′ recombination gradients that were found fit the patterns of GC content in introns and exons along genes that were observed in grasses [6]. Recombination data with similar resolution are still lacking in grasses, but if this recombination pattern is ancestral in eukaryotes, as suggested by these two studies and similar results in yeast [26,48], we can speculate that it also holds for grasses and other angiosperms. Given that GC content gradients with different slopes were observed in many species [14], we propose that variation in recombination gradients and/or gBGC intensity could be one of the main determinants of GC patterns along genes in angiosperms. Here, introns likely have a key role. Obviously, they keep exons away from the TSS, further reducing recombination in the middle and the end of genes. In

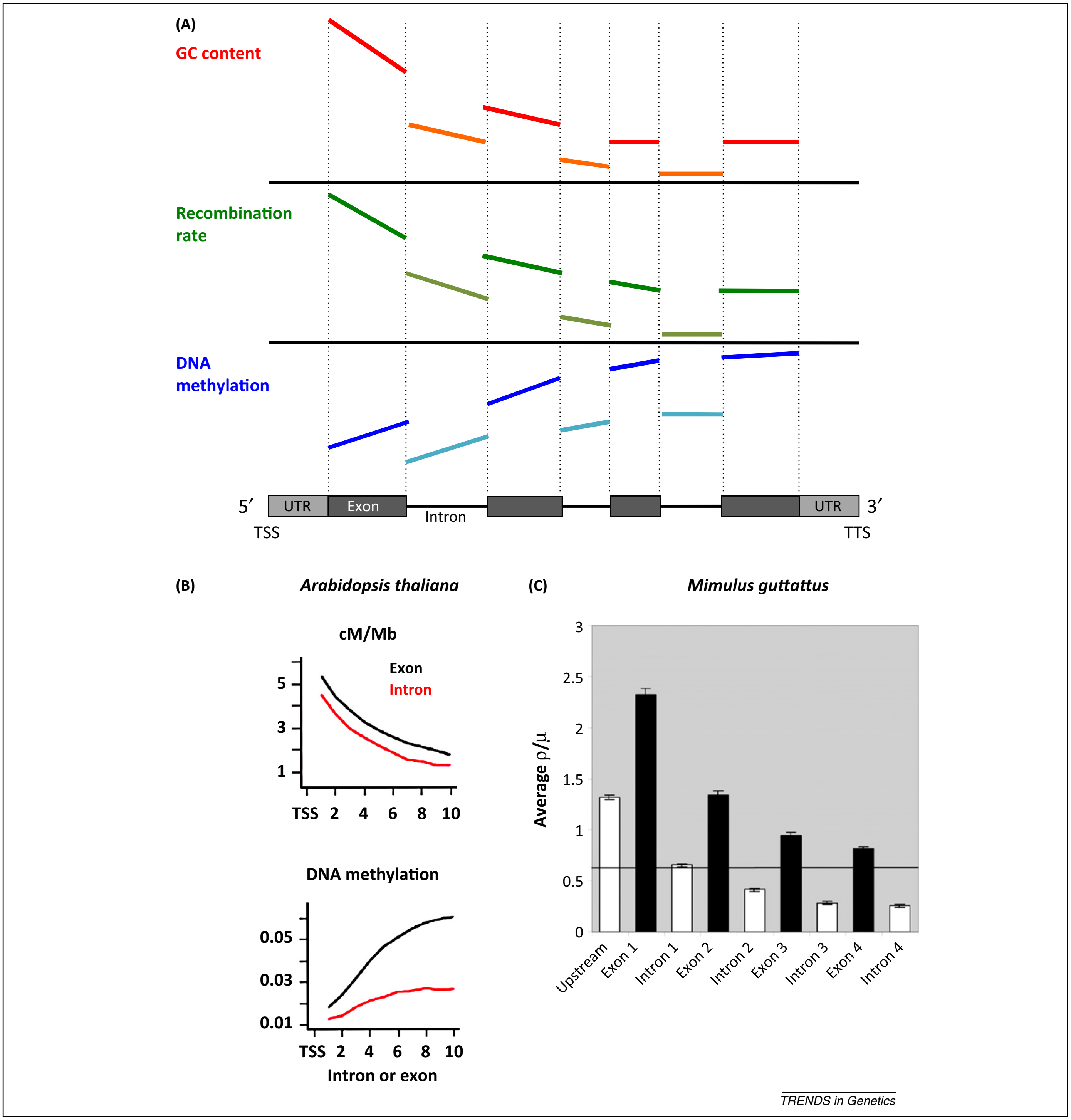


Figure 1. 5′–3′ gradients along genes in angiosperms. **(A)** Schematic and average representation of the different 5′–3′ gradients that can be observed along genes in angiosperms [6,18,42,43]. Patterns are parallel between exons and introns but there are clear quantitative differences between them. **(B)** Recombination and methylation gradients observed in *Arabidopsis thaliana* [42]. **(C)** Recombination rates (relative to mutation rates) in exons and introns along genes in *Mimulus guttatus* [43]. Abbreviations: TSS, transcription start site; TTS, transcription termination site; UTR, untranslated region.

addition, given that they affect gene structure and reduce recombination [42,43], they could have a direct effect beyond their passive role on gene length. Note that the recombination and gBGC gradient hypothesis can also explain the gradients of nucleotide and amino acid substitution rates observed along genes (higher substitution rates at the 5′ end) [6], because gBGC is expected to increase the substitution process (Box 2) [35,36].

GC content gradient and gene structure could explain genome-wide distributions of GC content

A straightforward consequence of the 5′–3′ GC content gradient is that gene structure should affect GC content. Short genes, especially genes with few introns, should be more GC rich than long genes, simply because the GC content of a gene is an average over the gradient. As predicted, there is a strong association between gene structure, gene length, and GC content in flowering plants [6,49,50]. Consequently, at the genome scale, we would expect GC content distributions to be controlled by the distribution of gene structure and the intensity of the recombination and gBGC gradient. As a hypothetical example, let us consider a species with a strong recombination and gBGC gradient and with only two kinds of gene, short mono-exonic genes and long genes with two exons (Figure 2). The resulting GC content distribution would be clearly bimodal. Now, consider the same gene composition but with a weaker gradient. This would result in a unimodal distribution with a lower mean. This example shows that a simple change in the shape of the gradient or a change in the distribution of gene structure (e.g., number of exons or length of introns) can strongly affect overall GC content distribution. The sharpness of the gradient can easily evolve with the intensity of recombination and/or effective population size (Box 2), without any change in the basic molecular machinery of recombination. In agreement with this prediction, there is a strong correlation across angiosperms between the slope of the 5′–3′ GC content gradient and the mean and variance of the GC content distribution: GC-rich and heterogeneous genomes also exhibit the strongest 5′–3′ GC content gradients [14]. As a result, provided that in flowering plants recombination is involved in gradient genesis, gBGC and gene structure could explain genome-wide distributions of GC content.

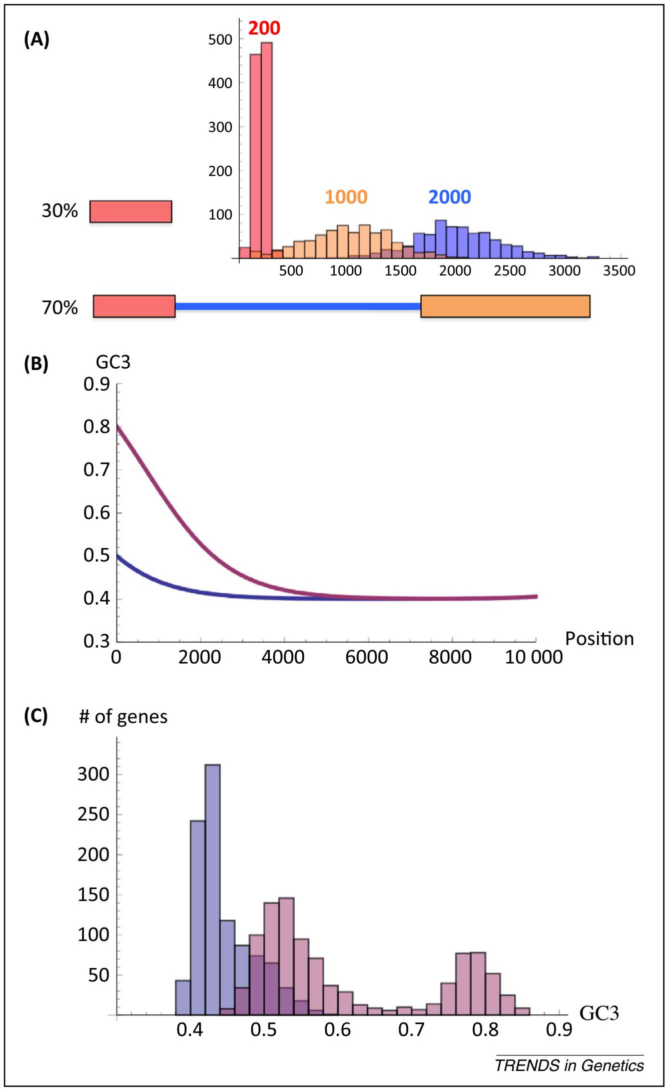


Figure 2. A hypothetical example of the relation between GC content gradient, gene structure, and GC content distribution. **(A)** The hypothetical genome is assumed to comprise two classes of gene: 30% short mono-exonic genes with an average length of 200 codons and 70% long genes with two exons with an average length of 200 and 1000 codons, respectively, separated by an intron with an average length of 2000 nucleotides. A distribution of length is assumed for each exon and for the intron. We only consider two classes of gene to simplify the example, but we assume a larger variance for the introns and the second exon to simulate large variation in total gene length. For simplicity, we assume that introns only have an effect on gene length and no additional specific effect. **(B)** At each third position, a GC or an AT base is drawn from a Bernoulli distribution with a mean given by the 5′–3′ GC content gradient. The gradient function is given by Equation 5 in [45]: $GC(x) = \frac{A(1-p)^x + B(1-p)^{L-x} + C}{D\left(A(1-p)^x + B(1-p)^{L-x}\right) + 1}$. The six parameters are complex combinations of the mechanistic parameters taken into account in [45]. Approximately, L is the average length of a gene, $1/p$ is the average length of conversion tracts, A is proportional to the intensity of gBGC in 5′, B to the intensity of gBGC in 3′, and C and D are functions of the substitution processes, independent and dependent of recombination. Here, the parameters have been adjusted to give a strong gradient (purple: $A = 2$, $B = 0.01$, $C = 0.4$, $D = 1$, $p = 0.001$, $L = 10\,000$) and a weak one (blue: same parameters except $A = 0.2$). The difference between the two gradients corresponds only to a difference in the frequency of recombination initiated in the 5′ region. **(C)** The resulting GC3 distribution is clearly bimodal with high average GC content (0.61) for the strong gradient (purple) and unimodal with lower average GC content (0.44) for the weak gradient (blue).

GC content is also correlated with patterns of expression in grasses. For instance, GC3-rich genes exhibit more variable expression than GC3-poor genes, and it was proposed that high GC3 content could be selected for regulation of gene expression [7,33]. Differences in GC content among genes would result from different classes of gene expression. Alternatively, the relation between GC content and expression patterns could be indirect and driven by gene structure. Indeed, gene structure is associated with the regulation of gene expression. In both *A. thaliana* and rice, highly expressed genes have more and longer introns than lowly expressed genes [51], and in *A. thaliana* (as well as yeast and mice), rapidly regulated genes are more compact with fewer introns [52]. Therefore, by affecting both GC content and patterns of expression, gene structure could underlie the observed correlation between these two variables.

Gene methylation and nucleosome positioning could impact GC content

Recently, patterns of methylation were shown to be strongly associated with GC content distribution in grasses [33,49] and well conserved between rice and *B. distachyon* [49]. In both species, the 20% body-methylated genes are especially GC poor compared with unmethylated genes. Moreover, methylation is linked to gene structure: short genes are hypomethylated and GC rich and mainly correspond to the second mode of the GC content distribution (as in Figure 2C). How could this striking feature be related to the model proposed above? The association between methylated regions and low GC content could be directly explained by the high rates of deamination of methylated cytosines [33,49]. Given that methylation increases along the gene from 5′ to 3′ [42], deamination could also contribute to the decreasing GC content gradient. However, because mutation seems to be generally biased towards AT, the absence of deamination is not sufficient to explain very high GC content.

Alternatively, and not exclusively, the relation between GC content and methylation could be mediated by recombination and gBGC. Interestingly, both *A. thaliana* and *M. guttatus* recombination studies clearly showed a negative association between methylation and recombination [42,43]. Moreover, in *A. thaliana*, loss of methylation increases recombination in euchromatin regions [53,54], and the 5′–3′ methylation gradient correlates negatively with the recombination gradient [42]. As a consequence, short genes are expected to be less methylated than long genes, as observed in rice and *B. distachyon* [49]. Similarly, in *M. guttatus*, recombination hotspots are associated with unmethylated CpG islands [43]. If methylation controls recombination patterns, it could in turn control both local patterns and global distributions of GC content.

Finally, complex interactions between nucleosome occupancy, GC content, intron–exon architecture, and methylation have been described [46,47]. In genes, nucleosome occupancy is observed principally in exons whereas introns are depleted for nucleosomes [47]. Nucleosome positioning is tightly associated with high GC content and low methylation levels, and both could be involved in exon–intron definitions [46,47]. It has been suggested that a similar mechanism could also stall the recombination machinery in exons [43], resulting in the different levels of recombination observed between exons and introns. Thus, gBGC could help maintain differences in GC content between exons and introns, inducing a positive feedback loop between GC content and exon definition. In addition, preferential nucleosome fixation in exons could prevent GC3 reaching levels as low as in introns by protecting cytosines from deamination [55]. Thus, the interplay between methylation, recombination, and possibly nucleosome positioning needs to be clarified to gain a fuller understanding of GC content distribution in plants.

Concluding remarks

The origin of the diversity of nucleotide landscapes in plants is a puzzling question. Given the association between GC content and many other genomic features, it is tempting to look for functional or adaptive explanations (e.g., [7,31,33]). However, Lynch [56] argued that mechanistic molecular constraints should be taken into account to interpret genomic patterns, whereas Galtier and Duret [35] pleaded for 'extending the null hypothesis of molecular evolution' to include the possible effects of gBGC to interpret classical genomic signatures of selection. Following these lines, we also argue that we should first evaluate the main nonadaptive mechanistic processes to explain genomic patterns of base compositions before looking for functional or adaptive explanations. This does not mean that selection does not have a role, but we propose that gene structure and recombination patterns could be the major determinants of base composition. Other molecular mechanisms, such as methylation and nucleosome positioning, could also affect GC content.

Additional research is needed to put our hypothesis to the test. Detailed recombination and methylation maps at the gene scale are lacking in rice and other grasses, which are essential to confirm or disprove the hypothesis that 5′–3′ GC content gradients are caused by recombination gradients. So far, gBGC has been clearly identified in only a few plant species. Thus, tests for the occurrence of gBGC in many other groups are needed. This can be achieved both by direct sequencing of meiosis products (e.g., [40]) or by indirect population genomic approaches (e.g., [28]). Finally, measuring variation in recombination patterns and gene structure along the angiosperm phylogeny should help interpret the wide diversity of nucleotide landscapes.

In turn, GC content patterns could also provide information about the recombination process. Given that it was proposed that the location of recombination hotspots at TSSs could be ancestral to most eukaryotes [42,43], our model could indeed apply to many other groups outside angiosperms, except those that secondarily evolved other mechanisms of hotspot determination, such as mammals [57]. Thus, GC content patterns along genes could give some clues about the molecular control of recombination in nonmodel species without other information. We suggest that the occurrence of a 5′–3′ GC gradient is indicative of recombination initiation at TSSs. Moreover, characterizing a 5′–3′ GC gradient could be a first, rough but easy, step to determine genomic variation in recombination patterns within genomes and between species.

Acknowledgments

We thank Nicolas Galtier for helpful discussions and comments on the manuscript. We also thank three anonymous reviewers for their very helpful and constructive comments on a previous version of the manuscript. This work was supported by Agropolis Fondation under the ARCAD project No 0900-001 (http://www.arcad-project.org/), which funded Y.C. This is ISEM publication number ISEM 2014-060.

References

1 Eyre-Walker, A. and Hurst, L.D. (2001) The evolution of isochores. *Nat. Rev. Genet.* 2, 549–555

2 Mouchiroud, D. *et al.* (1991) The distribution of genes in the human genome. *Gene* 100, 181–187

3 Costantini, M. and Bernardi, G. (2008) Replication timing, chromosomal bands, and isochores. *PNAS* 105, 3433–3437

4 Duret, L. and Arndt, P.F. (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.* 4, e1000071

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

6 Zhu, L. *et al.* (2009) Patterns of exon–intron architecture variation of genes in eukaryotic genomes. *BMC Genomics* 10, 47

7 Tatarinova, T.V. *et al.* (2010) GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics* 11, 308

8 Kejnovsky, E. *et al.* (2009) Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends Ecol. Evol.* 24, 572–582

9 Ammiraju, J.S. *et al.* (2008) Dynamic evolution of oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell* 20, 3191–3209

10 Rafalski, A. and Morgante, M. (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet.* 20, 103–111

11 Gore, M.A. *et al.* (2009) A first-generation haplotype map of maize. *Science* 326, 1115–1117

12 Tian, Z. *et al.* (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? *Genome Res.* 19, 2221–2230

13 Paape, T. *et al.* (2012) Fine-scale population recombination rates, hotspots, and correlates of recombination in the *Medicago truncatula* genome. *Genome Biol. Evol.* 4, 726–737

14 Serres-Giardi, L. *et al.* (2012) Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell* 24, 1379–1397

15 Smarda, P. *et al.* (2008) Genome size and GC content evolution of *Festuca*: ancestral expansion and subsequent reduction. *Ann. Bot.* 101, 421–433

16 Romiguier, J. *et al.* (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20, 1001–1009

17 Wang, H.C. *et al.* (2004) Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* 21, 90–96

18 Wong, G.K. *et al.* (2002) Compositional gradients in Gramineae genes. *Genome Res.* 12, 851–856

19 Duret, L. and Galtier, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* 10, 285–311

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

21 Meunier, J. and Duret, L. (2004) Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* 21, 984–990

22 Clement, Y. and Arndt, P.F. (2013) Meiotic recombination strongly influences GC–content evolution in short regions in the mouse genome. *Mol. Biol. Evol.* 30, 2612–2616

23 Katzman, S. *et al.* (2011) Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hotspots. *Genome Biol. Evol.* 3, 614–626

24 Pessia, E. *et al.* (2012) Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.* 4, 675–682

25 Escobar, J.S. *et al.* (2011) GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Mol. Biol. Evol.* 28, 2561–2575

26 Mancera, E. *et al.* (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454, 479–485

27 Lesecque, Y. *et al.* (2013) GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Mol. Biol. Evol.* 30, 1409–1419

28 Muyle, A. *et al.* (2011) GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Mol. Biol. Evol.* 28, 2695–2706

29 Hazzouri, K.M. *et al.* (2013) Comparative population genomics in *Collinsia* sister species reveals evidence for reduced effective population size, relaxed selection, and evolution of biased gene conversion with an ongoing mating system shift. *Evolution* 67, 1263–1278

30 Gunther, T. *et al.* (2013) Mutational bias and gene conversion affect the intraspecific nitrogen stoichiometry of the *Arabidopsis thaliana* transcriptome. *Mol. Biol. Evol.* 30, 561–568

31 Shi, X.L. *et al.* (2007) Evidence that natural selection is the primary cause of the guanine-cytosine content variation in rice genes. *J. Integr. Plant Biol.* 49, 1393–1399

32 Mukhopadhyay, P. *et al.* (2007) Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes. *Gene* 400, 71–81

33 Tatarinova, T. *et al.* (2013) Cross-species analysis of genic GC3 content and DNA methylation patterns. *Genome Biol. Evol.* 5, 1443–1456

34 D'Onofrio, G. *et al.* (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.* 32, 504–510

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

36 Galtier, N. *et al.* (2009) GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25, 1–5

37 Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 12, 640–649

38 Escobar, J.S. *et al.* (2010) An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). *Evolution* 64, 2855–2872

39 Haudry, A. *et al.* (2008) Mating system and recombination affect molecular evolution in four Triticeae species. *Genet. Res.* 90, 97–109

40 Yang, S. *et al.* (2012) Great majority of recombination events in *Arabidopsis* are gene conversion events. *PNAS* 109, 20992–20997

41 Lu, P. *et al.* (2012) Analysis of *Arabidopsis* genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg *erecta* and all four products of a single meiosis. *Genome Res.* 22, 508–518

42 Choi, K. *et al.* (2013) *Arabidopsis* meiotic crossover hot spots overlap with H2A.Z. nucleosomes at gene promoters. *Nat. Genet.* 45, 1327–1336

43 Hellsten, U. *et al.* (2013) Fine–scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 48, 19478–19482

44 Stoletzki, N. (2011) The surprising negative correlation of gene length and optimal codon use: disentangling translational selection from GC-biased gene conversion in yeast. *BMC Evol. Biol.* 11, 93

45 Marsolier-Kergoat, M.C. (2011) A simple model for the influence of meiotic conversion tracts on GC content. *PLoS ONE* 6, e16109

46 Gelfman, S. *et al.* (2013) DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res.* 23, 789–799

47 Chodavarapu, R.K. *et al.* (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466, 388–392

48 Pan, J. *et al.* (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* 144, 719–731

49 Takuno, S. and Gaut, B.S. (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1797–1802

50 Guo, X. *et al.* (2007) Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes. *FEBS Lett.* 581, 1015–1021

51 Ren, X.Y. *et al.* (2006) In plants, highly expressed genes are the least compact. *Trends Genet.* 22, 528–532

52 Jeffares, D.C. *et al.* (2008) Rapidly regulated genes are intron poor. *Trends Genet.* 24, 375–378

53 Mirouze, M. *et al.* (2012) Loss of DNA methylation affects the recombination landscape in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 5880–5885

54 Melamed-Bessudo, C. and Levy, A.A. (2012) Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 109, E981–E988

55 Chen, X. *et al.* (2012) Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* 335, 1235–1238

56 Lynch, M. (2007) *The Origin of Genome Architecture*, Sinauer

57 Baudat, F. *et al.* (2013) Meiotic recombination in mammals: localization and regulation. *Nat. Rev. Genet.* 14, 794–806

58 Axelsson, E. *et al.* (2012) Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res.* 22, 51–63

59 Giraut, L. *et al.* (2011) Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet.* 7, e1002354

60 Flicek, P. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.* 42, D749–D755

61 Kersey, P.J. *et al.* (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.* 42, D546–D552

62 Monaco, M.K. *et al.* (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.* 42, D1193–D1199

63 Kinsella, R.J. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011, bar030

64 Nagylaki, T. (1983) Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci. U.S.A.* 80, 6278–6281

65 Kong, A. *et al.* (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488, 471–475

66 Spencer, C.C. *et al.* (2006) The influence of recombination on human genetic diversity. *PLoS Genet.* 2, e148