Université de Rennes 1

Master 1 IMABEE

Internship Report

Effects of genic GC gradients and gene structure through GC-biased gene conversion on the variation of coding GC-content in angiosperm genomes

---

Maya SCHRÖDL

Hosting structure: ECOBIO
Tutor: Sylvain Glémin

Internship coordinator: Maxime Hervé

Internship period: 01/04/2019 - 28/06/2019
Defended: 13/05/2019

CONTENTS

1. INTRODUCTION

Nucleotide landscapes, which are the way base composition is distributed along a genome, strongly vary between organisms. In eukaryotes, mean genomic GC-content (proportion G+C versus A+T) ranges from 20% to 60% (Lynch, 2007). At first sight, this is surprising because a naive expectation would be that GC-content should be close to 50%.

GC content also varies within a genome. Vertebrates, and especially mammals, exhibit a very heterogeneous genomic base composition. Their genomes generally consist of alternating GC-rich and GC-poor regions. These large regions with similar GC-contents are called "isochores" (Eyre-Walker and Hurst, 2001). There, inner genic GC-content correlates with flanking regions (Clay et al., 1996). Inside these isochores, GC-content of coding sequences (CDS) is highly variable at different levels: the first, second, and the third codon positions (GC1, GC2, GC3). GC1 and GC2 are most constrained, since any change in these compositions often causes protein changes, whereas changes in GC3 are mostly synonymous (silent) (Glémin et al., 2014).

Which evolutionary forces cause these isochore structures, and therefore heterogeneous nucleotide landscapes in general, is still much debated (Duret and Galtier, 2009; Serres-Giardi et al., 2012; Clément et al., 2017). During the last two decades, the GC-biased gene conversion (bGC) hypothesis has increasingly won attention and acceptance as being the major force determining GC-content evolution, particularly in mammals (Marais, 2003; Duret and Galtier, 2009). bGC is a mechanism taking place during meiotic recombination. After double-strand break and invasion of the homologous sequence by a single-stranded DNA, mismatches occur when the parental alleles differ. The repair of these mismatches is often biased towards GC (Marais, 2003). bGC affects equally GC-content on both - synonymous sites (introns and GC3) and non-synonymous sites (GC1 and GC2) (Glémin et al., 2014). bGC is a neutral process, not depending on the fitness effect of individuals carrying the alleles (Muyle et al., 2011). Nevertheless, bGC behaves in some aspects like selection, since it increases the probability of fixation of GC alleles (Nagylaki, 1983). Thus, it can be considered as a kind of meiotic drive at the nucleotide level. The bGC mechanism is proposed to counterbalance a possible mutational bias towards AT (Serres-Giardi et al., 2012).

bGC has been demonstrated to be widespread over the whole eukaryotic phylogeny (Pessia et al., 2012). There is direct evidence for bGC in yeast (Mancera et al., 2008) and humans (Williams et al., 2015; Arbeithuber et al., 2015), and much indirect evidence in other mammals (Duret and Galtier, 2009) and birds (Nabholz et al., 2011). Plants have been

less studied but some evidence exists in different plant species, especially in grasses (Muyle et al., 2011; Clément et al., 2017). Nevertheless, very little is known on whether bGC exists in other plant species (Serres-Giardi et al., 2012).

As opposed to mammals, flowering plants do not reveal isochore structures (Tatarinova et al., 2010). It has been suggested that genome-wide rearrangements erase potential isochore structures (Glémin et al., 2014). Notwithstanding, very heterogeneous and GC-rich genomes like commelinid monocots, have been documented with a bimodal GC distribution among genes within a genome (Serres-Giardi et al., 2012). On the contrary, GC-poor genomes, like some eudicots, are more homogeneous with a unimodal distribution (Serres-Giardi et al., 2012).

In flowering plants, GC3 does not correlate to flanking regions like in isochores, but correlates to GC1 and GC2 (Tatarinova et al., 2010). This suggest that GC-content is highly structured at a local scale around genes. In Glémin et al. (2014) it is suggested that the variation of GC-content between genes could explain the different patterns of variation observed amongst angiosperm genomes.

A possible explanation for the variation in CDS GC-content between genes could be the presence of a decreasing $5' - 3'$ GC gradient along the genes over exons, which is the strongest for GC3 (Glémin et al., 2014) and can also be found in introns. GC-poor and homogeneous species tend to have flat mean CDS GC3-gradients over all genes. On the contrary, GC-rich, heterogeneous species display steep gradients (Serres-Giardi et al., 2012; Wong et al., 2002; Clément et al., 2015). A mechanistic consequence of this gradient is that the shorter a gene is, i.e. the less exons it has, the higher its GC-content is (Zhu et al., 2009). This variation in gene GC-content with the gene structure is simply the case because the GC-content of a gene is the average over its gradient (Glémin et al., 2014; Ressayre et al., 2015). Therefore not only the gradient strength but also the proportion of short genes might have an impact on the mean GC-content and the GC heterogeneity of a species (Glémin et al., 2014).

It has been proposed that this decreasing GC gradient is due to a higher recombination rate at transcription start sites (TSS), i.e. at the beginning of genes ($5'$-end). This might lead to a recombination gradient over the gene. This recombination gradient, in interaction with exon-intron structure and through bGC, would lead to the observed GC gradient (Glémin et al., 2014). Indeed, in yeast, similar GC gradients as in angiosperms have been documented, and they correlate with a decreasing $5' - 3'$ recombination gradient (Mancera et al., 2008). Furthermore, in two studies on the two angiosperm

species *Arabidopsis thaliana* (Choi et al., 2013) and *Mimulus guttatus* (Hellsten et al., 2013), it has been shown that recombination hotspots are mainly situated around the TSS. For *A. thaliana*, the recombination rate also slightly increases at the $3'$-end, at the transcription termination site (TTS). This fits in line with the bGC hypothesis.

CDS GC gradients, as an averaged slope over all genes regardless their length, have already been established for a large amount of angiosperm species (Serres-Giardi et al., 2012), using expressed sequence tags (EST). EST correspond to pools of mRNA fragments, so that exon/intron structure cannot be defined, and it is a biased sample towards highly expressed genes (Serres-Giardi et al., 2012). These analyses have not yet been conducted on whole genomes, which is now made possible through the large amount of recently sequenced genomes (Goodstein et al., 2012). Furthermore, there is a lack of knowledge when it comes to describing a species' gradients distinguished by gene structure. This has only been done on *Arabidopsis thaliana* and *Oryza sativa* (Ressayre et al., 2015), which are respectively a GC-poor and a GC-rich species. Ressayre et al. (2015) state that it is crucial to take into account gene structure and not simply construct a mean gradient over all genes.

Moreover, it is not known whether the gradient amplitude or the global gene structure distribution prevail when explaining the species' mean GC-content and variation. Lastly, Ressayre et al. (2015) recognise that the GC gradients are always U-shaped, and a more precise description of the gradient than the overall slope is needed in order to determine which evolutionary forces are at. bGC has been put forward to be the main evolutionary force determining the gradient strength, but there is still a lack of evidence for angiosperms (Glémin et al., 2014).

Glémin et al. (2014) suggested that the GC-content variation along and between genes, as well as between species, can be mainly explained by different recombination patterns, and therefore different gradient amplitudes, and/or gene structure (number of exons/introns per gene) with the major evolutionary force being bGC.

We will test if the variation in GC-content among angiosperm species is due to the different amplitude of the $5' - 3'$ genic GC gradient and/or gene structure, as suggested in Glémin et al. (2014). Then, we will tempt to explain whether the gradient is mainly caused by bGC.

To achieve these two goals, a large comparative study was undertaken on 52 angiosperm genomes from which all coding sequences have been extracted and analysed. GC patterns were analysed at the genome level (distribution) and the gene level (gradient) and compared to different recombination proxies.

We expect that the difference in GC-content between angiosperm species is mainly due to a U-shaped genic $5' - 3'$ gradient. We expect that GC-rich heterogeneous species present a strong gradient which could be explained by a higher intensity in bGC than GC-poor species (Serres-Giardi et al., 2012).

Our results show that the difference in gradient strength between species, especially at the beginning of the gene, is the main factor creating variation in mean GC-content between species. Our results strongly suggest that the leading mechanism is bGC, but further studies are needed.

2. Material and methods

2.1. Genomic data

The analyses were conducted on a representative but unbalanced sample of 52 angiosperm species, of which the annotated chromosomal genomes were available on Phytozome v12 (Goodstein et al., 2012): twelve commelinid monocots, two alismatid monocots, and twenty-nine rosids eudicots, four asterid eudicots, four other eudicots, and one basal dicot. We only focused on the genic coding parts to avoid the noisy signals arising from frequent reshuffling of noncoding DNA (Glémin et al., 2014). The genes containing one or more introns in their $3'$ and/or $5'$ untranslated regions (UTR) were taken out from the total dataset (preliminary analyses: between 0% and 52% of the total gene number, depending on the species) in order to exclude the effect introns in the UTR-regions might have on the coding sequence GC-content (Ressayre et al., 2015). To avoid confusion with the UTR of the first and last exons, the term CDS was used for the coding exon parts.

2.2. Compositional analyses

For the compositional analyses Python v3.6.1 (van Rossum) was used with the package Biopython v1.73 (Cock et al., 2009). Firstly, the whole genome size was computed for each species. Then, for the 35 species for which the karyotype was available, the chromosome length was computed. For each gene, its length and the number of exons were calculated. The CDSs of each gene were numbered according to their rank from the $5'$-end toward the $3'$-end. For each CDS, the total mean GC-contents at the first, second, and third codon position (GC1, GC2, and GC3) were computed as in Serres-Giardi et al. (2012). Compositional data were obtained by a previous student (Zeballos, 2018).

2.3. Intraspecies general statistics

For all the following analyses, R v3.5.0 (R Core Team, 2018) was used. For each species, the mean GC3 ($GC3_{mean}$) and its standard deviation ($GC3_{sd}$) were computed to describe respectively the GC-richness and the GC heterogeneity of each species. For better illustration, two representative species were chosen: a GC-rich one, *Oryza sativa*, and a GC-poor one, *Arabidopsis thaliana*.

2.4. 5'-3' CDS GC3 GRADIENTS & DISCRETE GRADIENT MODEL

For each species, genes were grouped according to their exon number. Let i be the exon number per gene ($i \in (1, ..., 14)$) and j the ranking index of a given CDS ($j \in (1, ..., i)$). The median GC3-contents were calculated as: $GC3^{obs}_{i,j} = median(GC3_{i,j})$. We limited our analyses to genes with $i \leq 14$, because only few genes had more exons. The exon and intron lengths were not taken into account.

To quantify the gradient, we chose to describe it by the following discrete phenomenological model, which can capture several shapes (from flat to U-shaped):

$$GC3^{pred}_{i,j} = (A-e)c^{j-1}+(B-e)d^{i-j}+e \quad (1)$$

where $GC3^{pred}_{i,j}$ is the predicted median GC3-content, i the exon number per gene, and j the CDS rank. A corresponds approximately to the GC3-content at the $5'$-end (j = 1), B approximately to the GC3-content at the $3'$-end ($j = i$), e corresponds approximately to the GC3-content at the lowest point of the gradient. c and d ($0 \leq c, d \leq 1$) are coefficients that determine the curvature of the gradient at the $5'$- and the $3'$-end respectively. The i- and j-independent parameters $A, B, e, c,$ and d where determined for each species by fitting the model to the observed $GC3^{obs}_{i,j}$ data with the *nlsLM()* function from the R package minpack.lm (Elzhov et al., 2016).

As a next step, the gradient at the beginning of a gene ($5'$ $gradient$) was approximated by $grad_{5'} = A - e$. The gradient at the end of a gene ($3'$ $gradient$) was likewise approximated by $grad_{3'} = B - e$.

2.4.1. *Goodness of fit*

It was tested whether the gradient model (equation 1) predicts well the GC3-content of a gene with a certain exon number i. The mean GC3-content of the genes with i exons, which was observed/predicted by the model, was estimated weighting by the mean CDS length for each CDS rank j, using the following equation:

$$GC3^{pred,obs}_{i} = \frac{\sum_{j=1}^{i}(L_j\ GC3^{pred,obs}_{i,j})}{\sum_{j=1}^{i} L_j} \quad (2)$$

where L_j the mean length of the CDS with a rank j.

Then the difference between the $GC3^{pred}_{i}$ and the $GC3^{obs}_{i}$ was computed for each i.

This difference for each i for each species was summarised by taking the mean difference for each i of all species, which was compared to zero difference for each i using a wilcoxon-test.

2.5. RECOMBINATION PROXIES

The recombination rate was approximated by the crossover rate per Megabase (Mb). The crossover rate itself was estimated at different levels. Firstly, it was approximated by taking the whole genome size, since crossovers tend to be more frequent in smaller genomes

(Hartl and Jones, 2005). For the 35 species, for which the annotated karyotype was available, the crossover rate was additionally approximated by the chromosome length, since the number of crossovers per Mb tends to be higher in smaller chromosomes (Kaback et al., 1999). In order to have a more precise idea of whether bGC plays a role, local recombination rates were computed through the comparison of genetic and physical maps using the R package MareyMaps (Siberchicot et al., 2017). The three GC-rich grass species *Oryza sativa, Brachypodium distachyon, Zea mays* and the GC-poor species *Arabidopsis thaliana* (Serres-Giardi et al., 2012), for which genetic maps were available and physically mapped on the genome, were analysed. The 1202 markers for *O. sativa*, the 558 markers for *B. distachyon* and the 1366 markers for *Z. mays* were taken from Serres-Giardi et al. (2012). The 400 markers for *A. thaliana* were taken from the example maps of MareyMaps (Siberchicot et al., 2017). As in Serres-Giardi et al. (2012), a loess regression was applied to each chromosome's genetic-physical map comparison and windows containing 20% of the total number of markers were taken. After obtaining the fitted function, the recombination rate for each gene for each of the four species was estimated, taking the centre coordinate of each gene as a reference position.

2.6. Statistical analyses

2.6.1. *General interspecific correlations*

Firstly, the Spearman correlation between the $GC3_{mean}$ and the $GC3_{sd}$ was computed, expecting to observe a positive correlation as in Serres-Giardi et al. (2012).

Following the results of Ressayre et al. (2015) on *A. thaliana* and *O. sativa*, we expect to observe also U-shaped $5' - 3'$ GC3 gradients for each gene class with a certain exon number for each of the 52 species. The Spearman correlations between $GC3_{mean}$ per species and respectively the slope of the gradients $grad_{5'}$ and $grad_{3'}$ were tested. The confidence intervals for each gradient's slope were computed by bootstrap ($n_{rep} = 1000$). This correlation is expected to be positive (Serres-Giardi et al., 2012). In order to test whether the GC-content richness at the beginning of the gradient and at the end are due to a similar mechanism, the Spearman correlation between the $grad_{5'}$ and the $grad_{3'}$ was computed.

We presume that the more exons a gene contains, the lower its GC3-content, as proposed by Glémin et al. (2014). Therefore, the Spearman correlation between exon number and mean GC3-content per gene was computed ($i \sim GC3_i$). Furthermore, we suppose that this last - presumably negative - correlation gets stronger, the stronger the gradient is (Ressayre et al., 2015). This would be a direct mechanistic consequence of the U-shaped gradient.

To test this, the Spearman correlation between the correlation coefficient of $i \sim GC3$ and the $grad_{5'}$ was computed.

2.6.2. *Model selection*

As a next step, it was tested whether the gradient strength and/or the distribution of gene size contribute more or equally to the difference in the species' mean GC-contents. To determine which variable prevails when it comes to the species' mean GC-content ($GC3_{mean}$), a stepwise, AIC based model selection was undertaken, based on the following linear model:

$$GC3_{mean} \sim grad_{5'} + grad_{3'} + monoex + i_{mean} \quad (3)$$

where the proportion of monoexonic genes $monoex$ and the mean exon number i_{mean} are approximations of the gene size distribution. We excluded the possible interactions between the different variables.

2.6.3. *Mechanisms explaining GC-gradients*

In order to determine whether the mechanism causing the different gradient amplitudes has a different effect on GC-content regarding codon position, the Spearman correlation coefficients between synonymous (GC3) and non-synonymous (GC1, GC2) sites: $GC3 \sim (GC1 + GC2)$, was computed as in Ressayre et al. (2015) for each species. The Spearman correlation between these last coefficients and the $GC3_{mean}$ were computed. Under the bGC-hypothesis, we suppose that the correlation $GC3 \sim (GC1 + GC2)$ is strictly positive for all species, and that it is even stronger for GC-rich species, as observed in Ressayre et al. (2015). When it comes to the explanation for the gradients' presence, we suppose that it can be mainly caused by bGC (Glémin et al., 2014). Serres-Giardi et al. (2012) propose that the positive correlation between mean GC3-content and gradient is due to a stronger intensity of bGC in GC-rich genomes. Therefore, we expect that the GC-richer a species is, the stronger is the correlation between the recombination rate and the GC3-content at the first exon. Only the GC3-content of the first CDS ($GC3_{j=1}$) was taken into account under the hypothesis that the recombination rate is the highest at the beginning of the gene.

This correlation was tested at different organisational levels. We expect larger genomes, which generally have a lower recombination rate per base pair, to be less GC-rich than smaller genomes (Glémin et al., 2014). Thus, the Spearman correlation between the whole genome length and the mean genomic $GC3_{j=1}$ was tested. Moreover, recombination rate per base pair is generally lower in long chromosomes (Pessia et al., 2012). Consequently, the Spearman correlation between the chromosome length and the mean chromosome $GC3_{j=1}$ was tested for each species. We expect a negative correlation, which would be stronger for GC-rich genomes, since it has

been proposed that GC-richer species experience higher bGC (Ressayre et al., 2015). Lastly, a strong negative correlation between local recombination rates and GC3-content of the first CDS is expected for GC-rich species. For each one of the four species, the genes were grouped into twenty bins according to their recombination rate estimation (Serres-Giardi et al., 2012). The correlation between the local recombination rate (r) and mean $GC3_{j=1}$ per gene bin ($r \sim GC3_{j=1}$) was established.

3. Results

3.1. GC-heterogeneity increases with GC-richness

The GC-richer the species is ($GC3_{mean}$), the more heterogeneous it is ($GC3_{sd}$) (fig. 1). The representative example species *Oryza sativa* is relatively GC-rich and *Arabidopsis thaliana* relatively GC-poor (fig. 1).

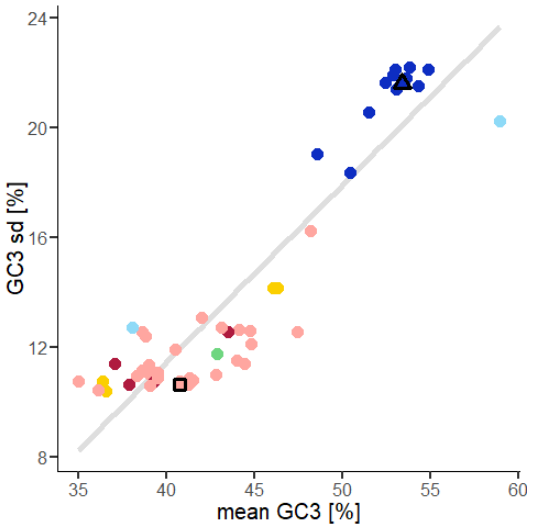

Figure 1: Standard deviation of the mean coding sequence GC content at the third codon position (GC3) as a function of the mean coding sequence GC3 content per species' genome. GC richness (mean GC3) and GC heterogeneity (GC3 sd) is positively correlated ($Spearman's\ \rho = 0.812$, $P < 10^{-16}$, n = 52) (grey line). Each point corresponds to a species. Dark blue: commelinid monocots (12), light blue: alismatid monocots (2), pink: rosid eudicots (29), red: asterid eudicots (4), yellow: other eudicots (4), green: basal dicot (1). The example species *Oryza sativa* and *Arabidopsis thaliana* are represented respectively by a triangle and a square.

3.2. 5'-3' U-shaped GC3 gradients & gradient model

The two species *O. sativa* and *A. thaliana* show both a very clear $5' - 3'$-gradient with a similar U-shape for each exon number class i (fig. 2). The longer a gene, ie. the higher i, the more the gradient of this gene is stretched in the middle.

Both gradients show a very different shape: the GC-rich species *O. sativa* shows a strong $5'$- and $3'$-gradient (respectively: 42.7; 9.4), whereas the GC-poor species *A. thaliana* has two weak gradients (respectively: 7.8; 5.7).

The five coefficients A, B, c, d, e of the fitted discrete model (eq. 1) were significant ($P < 10^{-3}$) for all but one species, which was excluded in the following analyses. Almost all the species showed a U-shaped gradient (fig. 3), 98% (± 0.02 (CI)) of the species having a positive $5'$-gradient ($grad_{5'}$) and at the same time a positive $3'$-gradient ($grad_{3'}$) (fig. 4).

The $grad_{5'}$ and $grad_{3'}$ are positively correlated when taking all species ($Spearman's\ \rho = 0.85, P < 10^{-16}$). The $grad_{5'}$ is generally stronger than $grad_{3'}$ (not overlapping bootstrap confidence intervals, fig. 4).

The higher the GC_{mean} of a species, the stronger the $grad_{5'}$ and $grad_{3'}$ (fig. 4).

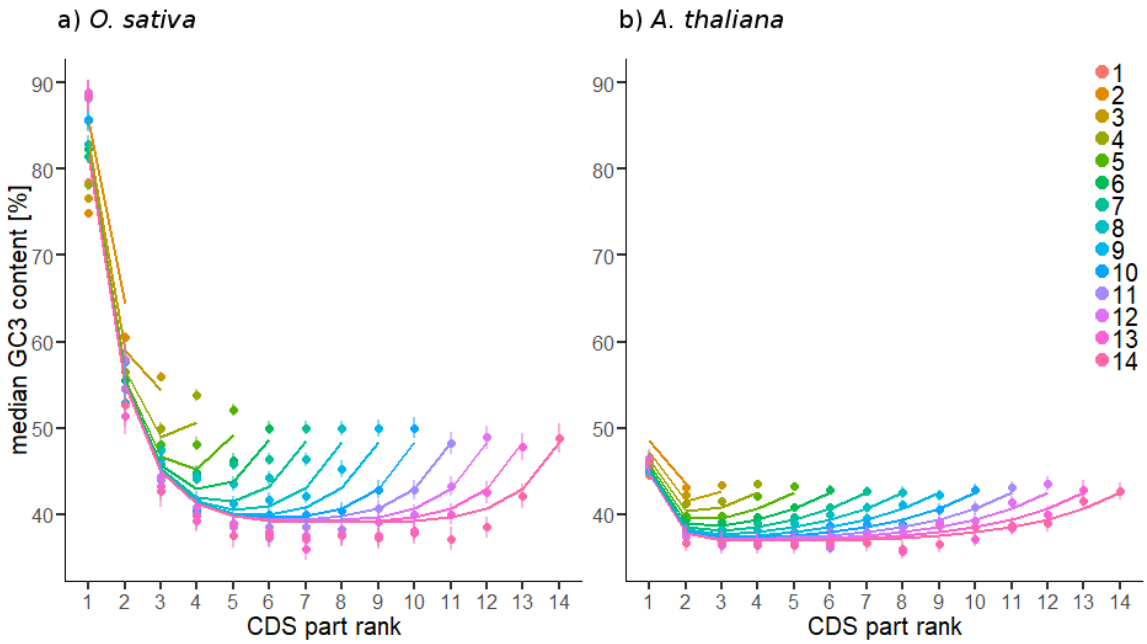

Figure 2: Genic $5' - 3'$ GC3 gradients and fitted discrete gradient models according to coding sequence (CDS) position along the gene ($j \in (1, ..., i)$) within each gene class with a certain exon number ($i \in (0, ..., 14)$). Two representative example species: a GC-rich one, *Oryza sativa* (a), and a GC-poor one, *Arabidopsis thaliana* (b). The gradients are represented as the median GC3 content [%] as a function of CDS part rank j. Each colour corresponds to a gene class, having a certain exon number i (legend shown at the top right). Each point corresponds to a calculated median GC3 content for i, j. Bars on dots represent the 95% confidence interval. Each line corresponds to the predicted values for each i, j by the fitted discrete gradient model.

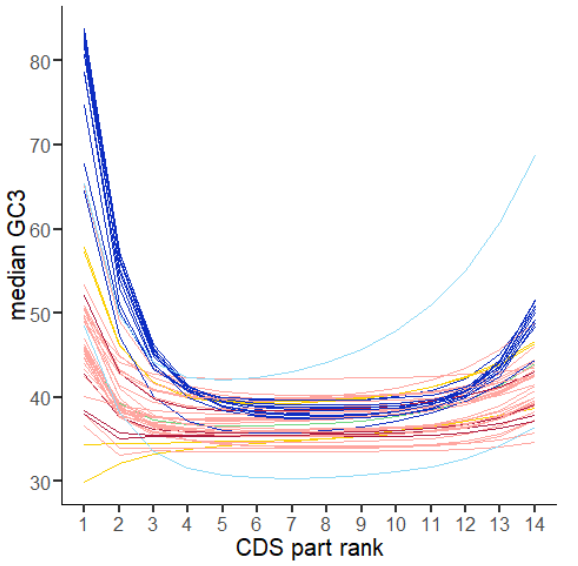

Figure 3: Summary of the genic $5' - 3'$ GC3 gradients for each species for genes with 14 exons ($GC3^{pred}_{14,j}$), computed by the gene-length-independent model. Each gradient is the median coding sequence GC3-content as a function of the coding sequence rank (CDS rank) from the $5'$-end to the $3'$-end. Each line represents a species' gradient. Dark blue: commelinid monocots (12), light blue: alismatid monocots (2), pink: rosid eudicots (29), red: asterid eudicots (4), yellow: other eudicots (3), green: basal dicot (1).

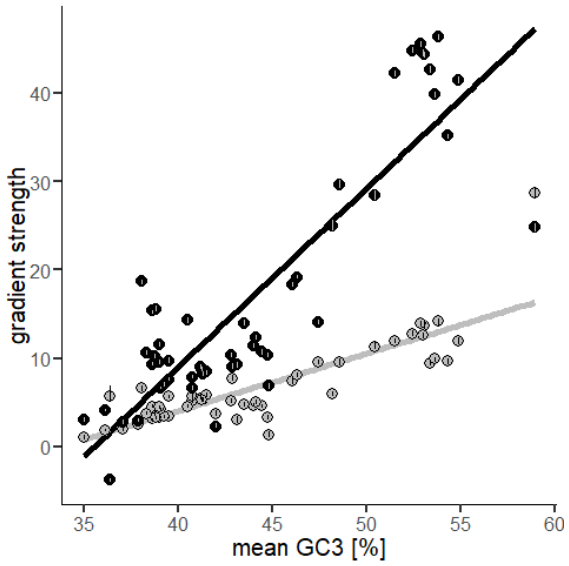

Figure 4: Genic $5' - 3'$ GC3 gradient amplitudes as a function of mean GC3 per species. Black: gradient at the beginning of the gene ($5'$-gradient); white: gradient at the end of the gene ($3'$-gradient). The mean GC3 content and the gradient strength are positively correlated for the $5'$-gradient ($Spearman's\ \rho = 0.74, P < 10^{-16}$) as well as for the $3'$-gradient ($Spearman's\ \rho = 0.76, P < 10^{-16}$). Each point per colour corresponds to a species ($n = 51$). Bars on dots represent the 95% bootstrap-generated confidence intervals ($n_{rep} = 1000$).

3.3. Discrete gradient model overestimates GC-content of short genes

When weighting by exon length (eq. 2), there is no difference between the predicted GC-content and the observed GC-content of genes for long genes ($i > 2$). For short genes ($i \leq 2$), the model overestimates the genic GC-content (fig. 5).

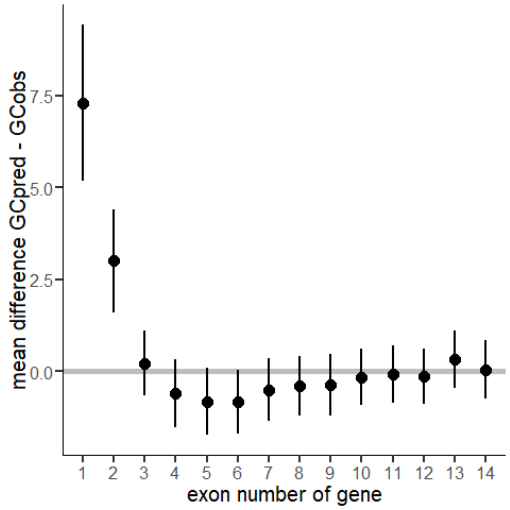

Figure 5: Difference between the observed GC3 and the GC3 per gene predicted by the discrete gradient model, as a mean over all species ($n = 52$) for each gene class with a certain exon number. Each point represents the mean difference for a certain gene class. Bars on dots represent the 95% confidence intervals. There is no difference between the predicted GC-content and the observed GC-content of genes for long genes ($i > 2$) (difference compared to zero over all species, for each i: wilcoxon-test, $p > 0.06$) (fig. 5). For short genes ($i \leq 2$), the model overestimates the genic GC-content (wilcoxon-test, $w > 2499, P < 10^{-19}$).

3.4. The stronger the GC gradient, the more the GC-content varies with gene size

The longer a gene (higher exon number), the lower the mean GC3-content of the gene ($Spearman's\ \rho < 0$, $P < 10^{-16}$, for each species). This association between exon number and mean GC3-content was stronger, the stronger the $5'$-gradient was ($Spearman's\ \rho = 0.925, P < 10^{-16}$).

3.5. Most variation in mean GC-content among species can be explained by different gradient amplitudes

After the stepwise model comparison based on the model in equation 3, the chosen model to describe the variation in mean GC3-content between species was the following:

$$GC3_{mean} \sim grad_{5'} + grad_3 + i_{mean} \quad (4)$$

The gradient strength at the beginning and the end ($grad_{5'}, grad_{3'}$) contributed positively to the $GC3_{mean}$ of a species, whereas the mean exon number (i_{mean}) lowered the $GC3_{mean}$ (table 1). Over 75% of the GC3-content variation was explained by the $grad_{5'}$. About 10% of the variation was explained by the $grad_{3'}$, and only a very small part (1.4%) by the i_{mean} (table 1).

Table 1: Summary of the chosen model explaining the variation of the mean GC3 content. The coefficient estimate, the Anova results (F, P), and the explained part, calculated through the sum of squares, are represented for each variable. The estimate of the intercept ($\pm se$) is 44.2 ± 3.6. $r^2 = 0.87$; adjusted $r^2 = 0.86$. The corresponding model is: $GC3_{mean} \sim grad_{5'} + grad_{3'} + i_{mean}$. $grad_{5'}$ and $grad_{3'}$ correspond respectively to the slope of the GC3 gene gradient at the beginning and the end of the gene. i_{mean} is the mean exon number per gene.

	coefficient estimate ($\pm se$)	F	P	estimated part [%]
$grad_{5'}$	0.23 ± 0.03	278.9	$< 10^{-16}$	76.3
$grad_{3'}$	0.59 ± 0.1	34.3	$4 * 10^{-7}$	9.4
i_{mean}	-1.97 ± 0.86	5.3	0.026	1.4
$residuals$	-	-	-	12.9

3.6. GC-CONTENT AT SYNONYMOUS AND NON-SYNONYMOUS POSITIONS

There was a negative correlation between the GC-content at synonymous (GC3) and non-synonymous (GC1+GC2) positions for GC-poor species and a positive correlation for GC-rich species (fig. 6). The GC-richer the species, the higher the correlation coefficient.

3.7. CORRELATION BETWEEN RECOMBINATION PROXIES AND GC3

There was no significant correlation between the genome size and the mean GC3-content of the first CDS ($GC3_{j=1}$) ($Spearman's\ \rho = 0.063, P = 0.66$).

For 34 of the 35 analysed species, there was no significant correlation between the mean chromosome $GC3_{j=1}$ and the chromosome length ($P > 0.05$). Only one species had a negative slightly significant correlation ($Spearman's\ \rho = -0.6, P = 0.044$). A positive correlation between the mean $GC3_{j=1}$ per gene and the local recombination rate was observed for the three GC-rich species (fig. 7). For the GC-poor species *A. thaliana*, there was no such correlation ($Pearson's\ \rho = -0.24, P = 0.3$).

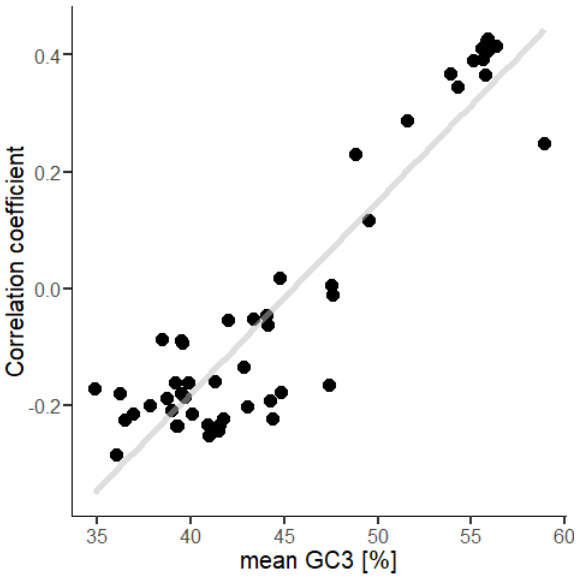

Figure 6: The correlation coefficient ($Spearman's\ \rho$) of the correlation between the GC content at synonymous (GC3) and non-synonymous (GC1 + GC2) codon positions ($GC3 \sim (GC1 + GC2)$) as a function of mean GC3-content. The correlations $GC3 \sim (GC1 + GC2)$ were different from 0 ($P < 10^{-16}$) for each species. The higher the mean GC3-content, the higher the correlation coefficient of the latter correlation ($Spearman's\ \rho = 0.92, P < 10^{-16}$). Each point represents a species.

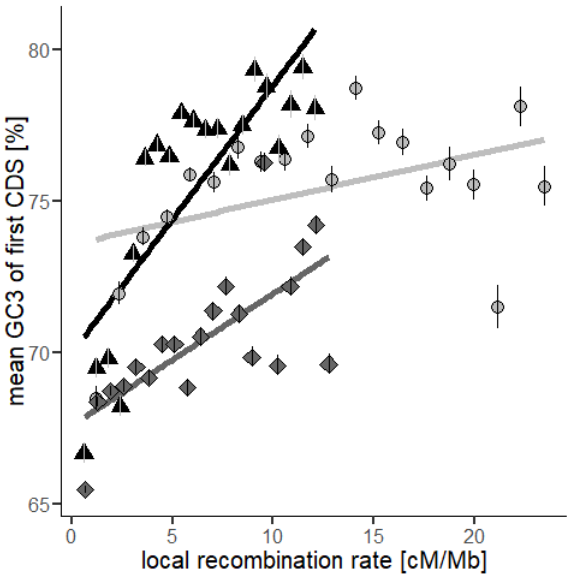

Figure 7: Relationship between mean GC3 of the coding part of the first exon (CDS) and the local recombination rate in three grass species. There is a positive correlation for *Brachypodium distachyon* ($Pearson's\ \rho = 0.53, P = 0.04$), *Oryza sativa* ($Pearson's\ \rho = 0.81, P = 10^{-5}$) and *Zea mays* ($Pearson's\ \rho = 0.69, P = 7*10^{-4}$). Genes have been grouped into twenty bins according to their local recombination rate estimation. Dots correspond to the mean GC3 of each bin and bars to the SEs. Black triangles: *O. sativa*; light grey points: *B. distachyon*; dark grey diamonds: *Z. mays*.

4. DISCUSSION

The reason for variation in GC-content among angiosperm species remains unknown. Glémin et al. (2014) suggested that this variation could be explained by different amplitudes of the $5' - 3'$ genic GC gradient, and/or gene structure (number of exons/introns per gene) with the major evolutionary force being GC-biased gene conversion (bGC). According to our results, this variation is mainly due to the variation of amplitude in the beginning ($5'$-end) of the $5' - 3'$ genic GC gradient. Our results suggest that this gradient might be caused by bGC.

4.1. GENERAL GC-RICHNESS

Firstly, we observe with our analysis on the fully sequenced genomes of 52 species, the same relation as Serres-Giardi et al. (2012) with EST sequences: the GC-richer a species is, the more heterogeneous it is (fig. 1). Since the analysed species samples were distributed all over the angiosperm phylogeny, it is most probable that this characteristics is true for most other angiosperm species.

4.2. GENIC 5'-3' GRADIENT AMPLITUDE DETERMINES GC-RICHNESS

Even though the presence of many small or many long genes (gene size distribution) can explain a small part of the GC-content variation between species, the $5' - 3'$ genic GC gradient amplitude plays a major role when it comes to the determination of mean GC-content of a species (table 1). Therefore, the evolutionary cause(s) (like bGC) leading to the differences in gradient amplitude between species, might have a major impact on the overall GC-richness of a species.

4.3. ANGIOSPERM SPECIES PRESENT A GENIC U-SHAPED GC GRADIENT FOR EACH GENE CLASS

As predicted, almost all species represent a U-shaped gradient (fig. 3). Having taken many distant angiosperm species, we can hypothesise as Ressayre et al. (2015), that the U-shaped gradient is ancestral to Angiosperms.

When we look at the extreme examples *O. sativa* (GC-rich) and *A. thaliana* (GC-poor) (fig. 2), we observe as Ressayre et al. (2015), that the U-shaped gradient stretches in the middle of the gene for genes with more exons. As a mechanistic consequence of this stretched gradient in the middle for long genes, we found that gene structure has a direct effect on the GC-richness of a species (section 3.4): the more coding sequences (CDS) a gene contains, the lower the mean GC-content of a gene, as observed in Zhu et al. (2009); Takuno and Gaut (2013) and Guo et al. (2007). Moreover, the stronger the genic GC gradient, the more the GC-content varies with the approximate gene length, which confirms the theoretical model of Glémin et al. (2014). This is simply a mechanistic consequence, since the mean GC-content of a gene is the average over the whole gradient. Thus, taking into account exon number

when characterising a species' GC-gradient is a crucial step. Indeed, not considering this variation in GC-content between gene class, may lead to misconclusions. For example, for GC-poor species like *A. thaliana*, no gradient was detected when pooling all genes (Wong et al., 2002), whereas our results suggest that there is a gradient when taking into account exon number. We therefore encourage, that future analyses on GC-gradients will not be done without considering exon (or intron) number.

4.4. A NEW CONTINUOUS MODEL TAKING INTO ACCOUNT EXON AND INTRON LENGTH IS NEEDED

Our discrete model, which takes into account exon number, predicts the median GC-content of a gene well for long genes (fig. 5). We can thereby conclude that at least long genes have a very similar gradient pattern, as we proposed when constructing the model. Yet, for genes with only one or two exons, this model strongly overestimates their GC-content. An hypothesis for this phenomenon is that genes with few exons might have longer exons than genes with many exons. This would result in lower GC-content than predicted by the gradient model. To test this hypothesis, we computed subsequently the correlation between the number of exon per gene and the mean exon length. Indeed, for all species, the less exons a gene had, the longer its exons were (mean over all species: $Spearman's\ \rho = -0.33\ (\pm\ 0.05\ sd)$; for all species: $P < 10^{-6}$). To avoid this bias through exon length, we propose to create a new gradient model. A different approach would be a continuous model based on gene length in terms of base pairs. This model would take into account exon and intron size.

4.5. EXPLANATORY HYPOTHESES FOR GC-CONTENT VARIATIONS AMONG AND WITHIN GENOMES

4.5.1. *GC-biased gene conversion*

Our results suggest that the genic GC gradient, and therefore the variation in GC-content between species, is mainly caused by bGC.

First of all, we observe that the beginning and the end of the gradient are positively correlated (section 3.2). We therefore suggest that the higher GC-content at the beginning and the end of the gene are due to a same mechanism that increases GC-content in the external regions. It has been suggested that recombination hotspots occur in the external regions of a gene (Choi et al., 2013; Hellsten et al., 2013), and linked to bGC, this could explain the latter correlation.

Moreover, we observe that the $5'$-gradient is always stronger than the $3'$-gradient. This fits in line with the bGC hypothesis, under the hypothesis that recombination is higher at the beginning of the gene, as shown for *A. thaliana* (Choi et al., 2013). Indeed, it was suggested, that recombination often initiates within gene promotors (Baudat and Nicolas, 1997; Mancera et al., 2008).

Furthermore, our results show that GC-rich species have stronger $5'$- and $3'$-gradients (fig. 4). A possible explanation for this phenomenon might be a higher intensity in bGC for GC-rich species, as suggested in (Serres-Giardi et al., 2012). Another result that suggests much bGC in GC-rich species is the observed positive correlation between GC-content at synonymous and non-synonymous coding positions for GC-rich species (6). This correlation is stronger, the GC-richer the species is. Since bGC affects all codon-positions equally (Serres-Giardi et al., 2012), this last result may suggest that bGC is stronger for GC-richer genomes than for GC-poor ones.

Another result that suggests the presence of bGC in GC-rich species is the observed positive correlation between local recombination rate and GC-content of the first exon in three GC-rich grass species (fig. 7), as expected under bGC. A full conclusion on the presence of bGC in GC-rich species cannot be made, since our local recombination analyses were only done on few species. To test whether bGC is present in other angiosperm species, more recombination maps will be needed.

4.5.2. *Genome and chromosome size are no suitable recombination proxies for plants*

We did not observe any correlation between GC-content and genome size. Additionally, there was almost no correlation between GC-content and chromosome length (section 3.7). One could conclude, that there is no bGC for angiosperm species at all. But this seems improbable, since bGC has already been suggested many times to occur in angiosperms (Glémin et al., 2014; Serres-Giardi et al., 2012; Clément et al., 2017; Ressayre et al., 2015). Furthermore, our results on the correlation between local recombination rate and GC-content in few species indicate that bGC is likely a present mechanism at least for some angiosperm species.

The more probable explanation for the lack of the latter correlations is that the approximation of recombination rate by genome size and chromosome size is probably not precise enough. Angiosperm genomes show a highly dynamic nature: their genomes can vary rapidly and to a large extent in size and structure (Kejnovsky et al., 2009). The recombination proxy by chromosome length works well for other eukaryotes (Pessia et al., 2012), but it can be misleading for plants, because of rapid chromosomal rearrangements. Transposable elements for example can enlarge chromosomes and genomes at an instant, without affecting the long-term recombination rate (Serres-Giardi et al., 2012).

4.5.3. *Further possible explanations*

We did not detect any correlation between the GC-content and the local recombination rate for the GC-poor species *A. thaliana*. This result suggests, by the example of one species,

that bGC might not occur in GC-poor species. Yet, we observe a genic GC gradient for GC-poor species.

A possible explanation for this contradiction might be that bGC is weaker in these species and that our recombination proxy through recombination maps is too imprecise to highlight weak bGC intensities. This explanation is supported by the observed weaker GC gradients for GC-poor species.

A stronger hypothesis explaining this contradiction is that the GC gradient might be caused by a different mechanism than bGC, at least for GC-poor species. For GC-poor species, the correlation between GC-content at synonymous and non-synonymous coding positions is negative (fig. 6). This has already been observed for *A. thaliana* (Ressayre et al., 2015). The only possible phenomenon able to produce such negative correlations is, to our knowledge, stabilising selection. GC-content in introns is always lower than in exons (Glémin et al., 2014), and stabilising selection could keep this intron-exon difference balanced (Goodall and Filipowicz, 1991). This difference could be regulated by the synonymous part (GC3) of the exons. The cause of this potential stabilising selection is unknown, but it could be that the difference between intron and exon GC-content is needed for splicing initiation (Ressayre et al., 2015).

We therefore propose that the GC gradients might be caused by a stabilising selection in interaction with different intensities in bGC. A simple way to have a look at this hypothesis would be to test the correlation between synonymous and non-synonymous positions for each CDS for each gene having a certain exon number.

5. Conclusion

Our results shed light on the possible causes of variation in nucleotide landscapes amongst angiosperms. GC-biased gene conversion (bGC) might be the major evolutionary force shaping the GC-content of angiosperm species through $5' - 3'$ genic GC gradients. We did not directly prove the role of bGC, but our results are most compatible with this hypothesis. More recombination maps are needed to confirm the presence and intensity of bGC. Finally, we strongly suggest taking bGC into account in future genomic analyses.

ACKNOWLEDGEMENTS

This project would not have been possible without Sylvain Glémin, whom I would like to thank for introducing me to the "GC world" and giving much helpful advice. Many thanks to my dearest friends Colin Guétemme and Monique Van Dorssen who proofread my report and provided me with constructive comments (and coffee). I would also like to thank Cécile Carpentier for her psychological support. Thanks to Nathalie Zeballos for the prepared data set. This internship was financed by the CNRS.

REFERENCES

Arbeithuber, B., Betancourt, A. J., Ebner, T., and Tiemann-Boege, I. (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences*, 112(7):2109–2114.

Baudat, F. and Nicolas, A. (1997). Clustering of meiotic double-strand breaks on yeast chromosome III. *Proceedings of the National Academy of Sciences of the United States of America*, 94(10):5213–5218.

Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C. H., McVean, G., and Henderson, I. R. (2013). Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature Genetics*, 45(11):1327–1336.

Clay, O., Cacciò, S., Zoubak, S., Mouchiroud, D., and Bernardi, G. (1996). Human coding and noncoding DNA: Compositional correlations. *Molecular Phylogenetics and Evolution*, 5(1):2–12.

Clément, Y., Fustier, M.-A., Nabholz, B., and Glémin, S. (2015). The Bimodal Distribution of Genic GC Content Is Ancestral to Monocot Species. *Genome Biology and Evolution*, 7(1):336–348.

Clément, Y., Sarah, G., Holtz, Y., Homa, F., Pointet, S., Contreras, S., Nabholz, B., Sabot, F., Sauné, L., Ardisson, M., Bacilieri, R., Besnard, G., Berger, A., Cardi, C., De Bellis, F., Fouet, O., Jourda, C., Khadari, B., Lanaud, C., Leroy, T., Pot, D., Sauvage, C., Scarcelli, N., Tregear, J., Vigouroux, Y., Yahiaoui, N., Ruiz, M., Santoni, S., Labouisse, J.-P., Pham, J.-L., David, J., and Glémin, S. (2017). Evolutionary forces affecting synonymous variations in plant genomes. *PLOS Genetics*, 13(5):e1006799.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423.

Duret, L. and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*, 10(1):285–311.

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., and Bolker, B. (2016). *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds*. R package version 1.2-1.

Eyre-Walker, A. and Hurst, L. D. (2001). The evolution of isochores. *Nature Reviews. Genetics*, 2(7):549–555.

Glémin, S., Clément, Y., David, J., and Ressayre, A. (2014). GC content evolution in coding regions of angiosperm genomes: A unifying hypothesis. *Trends in Genetics*, 30(7):263–270.

Goodall, G. J. and Filipowicz, W. (1991). Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. *The EMBO Journal*, 10(9):2635–2644.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, 40(Database issue):D1178–D1186.

Guo, X., Bao, J., and Fan, L. (2007). Evidence of selectively driven codon usage in rice: Implications for GC content evolution of Gramineae genes. *FEBS letters*, 581(5):1015–1021.

Hartl, D. L. and Jones, E. W. (2005). *Genetics: Analysis of Genes and Genomes*. Jones & Bartlett Learning.

Hellsten, U., Wright, K. M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S. R., Schmutz, J., Willis, J. H., and Rokhsar, D. S. (2013). Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proceedings of the National Academy of Sciences*, 110(48):19478–19482.

Kaback, D. B., Barber, D., Mahon, J., Lamb, J., and You, J. (1999). Chromosome Size-Dependent Control of Meiotic Reciprocal Recombination in Saccharomyces cerevisiae: The Role of Crossover

Interference. *Genetics*, 152(4):1475–1486.

Kejnovsky, E., Leitch, I. J., and Leitch, A. R. (2009). Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends in Ecology & Evolution*, 24(10):572–582.

Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, Mass. OCLC: 682103807.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, 454(7203):479–485.

Marais, G. (2003). Biased gene conversion: Implications for genome and sex evolution. *Trends in Genetics*, 19(6):330–338.

Muyle, A., Serres-Giardi, L., Ressayre, A., Escobar, J., and Glémin, S. (2011). GC-Biased Gene Conversion and Selection Affect GC Content in the Oryza Genus (rice). *Molecular Biology and Evolution*, 28(9):2695–2706.

Nabholz, B., Künstner, A., Wang, R., Jarvis, E. D., and Ellegren, H. (2011). Dynamic Evolution of Base Composition: Causes and Consequences in Avian Phylogenomics. *Molecular Biology and Evolution*, 28(8):2197–2210.

Nagylaki, T. (1983). Evolution of a finite population under gene conversion. *Proceedings of the National Academy of Sciences of the United States of America*, 80(20):6278–6281.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. B. (2012). Evidence for Widespread GC-biased Gene Conversion in Eukaryotes. *Genome Biology and Evolution*, 4(7):675–682.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ressayre, A., Glémin, S., Montalent, P., Serre-Giardi, L., Dillmann, C., and Joets, J. (2015). Introns Structure Patterns of Variation in Nucleotide Composition in *Arabidopsis thaliana* and Rice Protein-Coding Genes. *Genome Biology and Evolution*, 7(10):2913–2928.

Serres-Giardi, L., Belkhir, K., David, J., and Glémin, S. (2012). Patterns and Evolution of Nucleotide Landscapes in Seed Plants. *The Plant Cell*, 24(4):1379–1397.

Siberchicot, A., Rezvoy, C., Charif, D., Gueguen, L., and Marais, G. (2017). *MareyMap: Estimation of Meiotic Recombination Rates Using Marey Maps*. R package version 1.3.4.

Takuno, S. and Gaut, B. S. (2013). Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences*, 110(5):1797–1802.

Tatarinova, T. V., Alexandrov, N. N., Bouck, J. B., and Feldmann, K. A. (2010). GC3 biology in corn, rice, sorghum and other grasses. *BMC genomics*, 11:308.

van Rossum, G. The Python Language Reference. page 171.

Williams, A. L., Genovese, G., Dyer, T., Altemose, N., Truax, K., Jun, G., Patterson, N., Myers, S. R., Curran, J. E., Duggirala, R., Blangero, J., Reich, D., and Przeworski, M. (2015). Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *eLife*, 4:e04637.

Wong, G. K.-S., Wang, J., Tao, L., Tan, J., Zhang, J., Passey, D. A., and Yu, J. (2002). Compositional Gradients in Gramineae Genes. *Genome Research*, 12(6):851–856.

Zeballos, N. (2018). *GC Content Variation in Coding Regions of Angiosperm Genomes (Internship report, unpublished)*.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J.-Q., and Tian, D. (2009). Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC genomics*, 10:47.

Effects of genic GC gradients and gene structure through GC-biased gene conversion on the variation of coding GC-content in angiosperm genomes

Abstract: The coding parts of angiosperm genomes present a large range of variation in GC-content. GC-rich species have more heterogeneous genomes, while GC-poor species are homogeneous. The reason for these patterns is still unknown. GC-biased gene conversion (bGC), a recombination-associated process favouring GC over AT, has won a lot of attention in the last decades. Recently, it has been suggested that bGC is the major factor creating U-shaped $5' - 3'$ genic GC gradients, under the hypothesis that recombination is strongest at the beginning and the end of the genes. By action of these GC gradients, shorter genes are generally GC-richer than longer genes. Alongside gene structure, amplitude variation of these gradients might be the origin of GC-content variation among species.

To tackle these issues, we analysed fully sequenced genomes from over fifty angiosperm species on their genomic and genic GC patterns, while taking into account gene structure characterised as exon number per gene. Through recombination maps of a few species we also tested whether bGC plays a role.

Almost all analysed species presented such a U-shaped genic GC gradient and GC-richer species had a stronger gradient. This gradient was the major force impacting GC-content variation among species. Our results suggest that bGC is most present in GC-rich species, making them through stronger gradients to GC-rich species.

Overall, our results suggest that variation in bGC through its effect on gene GC-gradient, is likely the main determinant of GC-content variation within and among angiosperm species.

Key-words: nucleotide landscapes - plant evolution - recombination - coding regions

Les effets de gradients géniques de GC et de la structure génique, à travers de la conversion génique biaisée vers GC, sur la variation du taux de GC dans les régions codantes des génomes d'Angiospermes

Résumé : Les régions codantes des génomes d'Angiospermes recouvrent un grand spectre de variation du taux de GC. Les espèces GC-riches ont des génomes plus hétérogènes en GC, tandis que les espèces GC-pauvres sont plus homogènes. Les causes évolutives de ces tendances sont encore méconnues. Une cause possible est la conversion génique biaisé vers GC (bGC) qui est un processus lié à la recombinaison favorisant les allèles GC sur AT. Récemment, il a été proposé que la bGC est la force principale génératrice des gradients géniques de GC ($5'$ vers $3'$), sous l'hypothèse que le taux de recombinaison est le plus fort aux extrémités des gènes. A travers de ces gradients de GC, les gènes courts sont généralement les plus GC-riches. La variation de l'amplitude de ces gradients, avec la structure des gènes, pourrait ainsi être l'origine de la variation de GC entre espèces.

Pour tester ces hypothèses nous avons analysé les génomes entiers d'une cinquantaine d'espèces d'Angiospermes pour leurs caractéristiques génomiques et géniques de GC. Nous avons pris en considération la structure des gènes, caractérisé par le nombre d'exons par gène. Nous avons également testé sur quelques espèces par moyen de cartes de recombinaison si la bGC joue un rôle dans la formation de gradient.

Presque toutes les espèces présentaient un tel gradient de GC en forme de U et les espèces GC-riches avaient un gradient plus fort. Ce gradient semble être la force principale influençant la variation de taux de GC entre espèces. Nos résultats suggèrent que la bGC est plus présente dans les espèces GC-riches à cause des gradients de GC plus forts.

Nos résultats suggèrent que la variation en bGC via son effet sur les gradients géniques de GC, est probablement le déterminant principal de la variation de taux de GC au sein des espèces Angiospermes et entre elles.

Mots-clés : paysages nucléotidiques - évolution des plantes - recombinaison - régions codantes