Evolutionary forces affecting synonymous variations in plant genomes

Yves Clément, Sarah Gallien, Yan Holtz, Félix Homa, Stéphanie Pointet, Sandy Contreras, Benoit Nabholz, François Sabot, Laure Saune, Morgane Ardisson, et al.

PLOS | GENETICS

RESEARCH ARTICLE

Evolutionary forces affecting synonymous variations in plant genomes

Yves Clément[1,2,3]*, Gautier Sarah[4,5], Yan Holtz[1], Felix Homa[5,6], Stéphanie Pointet[5,7,8], Sandy Contreras[5,9], Benoit Nabholz[2], François Sabot[5,10], Laure Sauné[11], Morgane Ardisson[4], Roberto Bacilieri[4], Guillaume Besnard[12], Angélique Berger[7], Céline Cardi[7], Fabien De Bellis[7], Olivier Fouet[7], Cyril Jourda[7,13], Bouchaib Khadari[4], Claire Lanaud[7], Thierry Leroy[7], David Pot[7], Christopher Sauvage[14], Nora Scarcelli[10], James Tregear[10], Yves Vigouroux[10], Nabila Yahiaoui[7], Manuel Ruiz[5,7], Sylvain Santoni[4], Jean-Pierre Labouisse[7], Jean-Louis Pham[10], Jacques David[1], Sylvain Glémin[2,15]*

1 Montpellier SupAgro, UMR AGAP, Montpellier, France, **2** UMR 5554 ISEM (Université de Montpellier-CNRS-IRD-EPHE), Montpellier, France, **3** Ecole Normale Supérieure, PSL Research University, CNRS, Inserm, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris, France, **4** INRA, UMR AGAP, Montpellier, **5** SouthGreen Platform, Montpellier, France, **6** Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden, **7** CIRAD, UMR AGAP, Montpellier, France, **8** ALCEDIAG/CNRS Sys2Diag FRE3690, Biological Complex System Modelling and Engineering for Diagnosis, Cap delta, Montpellier, France, **9** GenoScreen, Lille, France, **10** IRD, UMR DIADE, Montpellier, France, **11** INRA, UMR1062 CBGP, Montferrier-sur-Lez, France, **12** UMR 5174 EDB (CNRS/ENSFEA/IRD/ Université Toulouse III), Toulouse, France, **13** CIRAD, UMR PVBMT, Saint-Pierre, La Réunion, France, **14** UR1052 GAFL (INRA), Montfavet, France, **15** Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

* yclement@biologie.ens.fr (YC); sylvain.glemin@ebc.uu.se (SG)

OPEN ACCESS

Citation: Clément Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, et al. (2017) Evolutionary forces affecting synonymous variations in plant genomes. PLoS Genet 13(5): e1006799. https://doi.org/10.1371/journal.pgen.1006799

Editor: Gregory P. Copenhaver, The University of North Carolina at Chapel Hill, UNITED STATES

Received: December 8, 2016

Accepted: May 4, 2017

Published: May 22, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. All reads have been deposited in the NCBI BioProjet under the accession number 326055 (https://www.ncbi.nlm.nih.gov/bioproject/?term=326055)

Funding: This work was supported by Agropolis Foundation in the framework of the ARCAD project N° 0900-001 (www.arcad-project.org). he funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Base composition is highly variable among and within plant genomes, especially at third codon positions, ranging from GC-poor and homogeneous species to GC-rich and highly heterogeneous ones (particularly Monocots). Consequently, synonymous codon usage is biased in most species, even when base composition is relatively homogeneous. The causes of these variations are still under debate, with three main forces being possibly involved: mutational bias, selection and GC-biased gene conversion (gBGC). So far, both selection and gBGC have been detected in some species but how their relative strength varies among and within species remains unclear. Population genetics approaches allow to jointly estimating the intensity of selection, gBGC and mutational bias. We extended a recently developed method and applied it to a large population genomic dataset based on transcriptome sequencing of 11 angiosperm species spread across the phylogeny. We found that at synonymous positions, base composition is far from mutation-drift equilibrium in most genomes and that gBGC is a widespread and stronger process than selection. gBGC could strongly contribute to base composition variation among plant species, implying that it should be taken into account in plant genome analyses, especially for GC-rich ones.

Competing interests: The authors have declared that no competing interests exist.

Author summary

In protein coding genes, base composition strongly varies within and among plant genomes, especially at positions where changes do not alter the coded protein (synonymous variations). Some species, such as the model plant *Arabidopsis thaliana*, are relatively GC-poor and homogeneous while others, such as grasses, are highly heterogeneous and GC-rich. The causes of these variations are still debated: are they mainly due to selective or neutral processes? Answering to this question is important to correctly infer whether variations in base composition may have functional roles or not. We extended a population genetics method to jointly estimate the different forces that may affect synonymous variations and applied it to genomic datasets in 11 flowering plant species. We found that GC-biased gene conversion, a neutral process associated with recombination that mimics selection by favouring G and C bases, is a widespread and stronger process than selection and that it could explain the large variation in base composition observed in plant genomes. Our results bear implications for analysing plant genomes and for correctly interpreting what could be functional or not.

Introduction

Base composition strongly varies across and within plant genomes [1]. This is especially striking at the coding sequence level for synonymous sites where highly contrasted patterns are observed. Most Gymnosperms, basal Angiosperms and Eudicots have relatively GC-poor and homogeneous genomes. In contrast, Monocot species present a much wider range of variation from GC-poor species to GC-rich and highly heterogeneous ones, some with bimodal GC content distribution among genes, these differences being mainly driven by GC content at third codon position (GC3) [1]. Commelinids (a group containing palm trees, banana and grasses, among others) have particularly GC-rich and heterogeneous genomes but GC-richness and bimodality have been showed to be ancestral to Monocots, suggesting erosion of GC content in some lineages and maintenance in others [2]. As a consequence, in most species, synonymous codons are not used in equal frequency with some codons more frequently used than others, a feature that is called the codon usage bias [reviewed in 3]. This is true even in relatively homogeneous genomes such as in *Arabidopsis thaliana* [e.g. 4].

Which forces drive the evolution of genome base composition and codon usage is still under debate. Mutational processes can contribute to observed variations between species and within genomes [e.g. 5]. However, mutation can hardly explain a strong bias towards G and C bases, as it is biased towards A and T in most organisms studied so far [Chapter 6 in 6]. Selection on codon usage (SCU) has thus appeared as one of the key forces shaping codon usage as it has been demonstrated in many organisms both in prokaryotes and eukaryotes [reviewed in 3]. Codon bias can thus result from the balance between mutation, natural selection and genetic drift [7]. The main cause for SCU is likely that preferred codons increase the accuracy and/or the efficiency of translation but other mechanisms involving mRNA stability, protein folding, splicing regulation and robustness to translational errors could also play a role [3,8,9]. In some species, SCU appears to be very weak or inexistent, typically when effective sizes are small [10], as typically assumed for mammals [but see 8]. However, mammalian genomes exhibit strong variations in base composition, the so-called isochore structure [11], which are mainly driven by GC-biased gene conversion (gBGC) [12]. gBGC is a neutral recombination-associated process favouring the fixation of G and C (hereafter S for strong) over A and T (hereafter W for weak) alleles because of biased mismatch repair following heteroduplex

formation during meiosis [13]. Although gBGC is a neutral process–*i.e*. the fate of S vs. W alleles is not driven by their effect on fitness—gBGC induces a transmission dynamic during reproduction identical to natural selection for population genetics [14]. Therefore, we here refer to it as a "selective-like" process as opposed to mutation and drift. gBGC has been experimentally demonstrated in yeast [15,16], humans [17,18], birds [19] and rice [20]. Many indirect genomic evidences also supported gBGC in eukaryotes [21,22] and even recently in some prokaryotes [23], although it seems to be weak or absent in some species as *Drosophila* [24] where selection on codon usage predominates [25,26,27,28].

In plants, both SCU [4,29,30] and gBGC [21,31,32] have been documented, but how their magnitudes and relative strength vary among species remains unclear. Recently, it has been proposed that the wide variations in genic GC content distribution observed in Angiosperms could be explained by the interaction between gene structure, recombination pattern and gBGC [33]. Increasing evidence suggests that in various organisms, including plants, recombination occurs preferentially in promoter regions of genes, or near transcription initiation sites [34,35,36]. This generates a 5'-3' recombination gradient, and consequently a gBGC gradient, which could explain the 5'-3' GC content gradient observed in GC-rich species, such as Commelinids [1,2]. A mechanistic consequence is that short genes, especially with no or few introns, are on average GC-richer [37]. A stronger gBGC gradient and/or a higher proportion of short genes would increase the average GC content and simple changes in the gBGC gradient can explain a wide range of GC content distribution from unimodal to bimodal ones [33].

So far, the magnitude of gBGC and SCU has been quantified only in a handful of plant species [29,30,32,38]. As in other species studied, weak SCU and gBGC intensities were estimated. The population-scale coefficients, $4N_e s$ or $4N_e b$, are usually of the order of 1, where N_e is the effective population size and s and b the intensity of SCU and gBGC respectively [26,29,30,32,38,39]. However, high gBGC values ($4N_e b > 10$) have been estimated in the close vicinity of recombination hotspots in mammals [38,40] and across the entire honeybee genome [41]. Differences in population-scale intensities can be due to variation in N_e and/or in s or b. For gBGC, b is the product of the recombination rate r and the basal conversion rate per recombination event, b_0. Within a genome, variations in gBGC intensities are mainly due to variation in recombination rate [e.g. 38]. Among species, b_0 can also vary. For instance, b was estimated to be 2.5 times lower in honeybees than in humans but recombination rate is more than 18 times higher [41], suggesting that b_0 could be 45 times lower in honeybees than in humans. The very intense population-scale gBGC in honeybees is thus explained by the combination of a large N_e and extremely high recombination rates [41].

Several methods have been developed to estimate the intensity of SCU and gBGC, either from polymorphism data alone, or from the combination of polymorphism and divergence data [e.g. 26,27,38]. These methods rely on the fact that preferred codons (for SCU) or GC alleles (for gBGC) are expected to segregate with higher frequency than neutral and un-preferred or AT alleles, fitting a population genetics model with selection or gBGC to the different site frequency spectra (SFS). As demography affects SFS, it must be taken into account in the model. Moreover, mutations must be polarized, *i.e*. the ancestral or derived state of mutations must be determined using one or several outgroup species. Otherwise, selection or gBGC can be estimated from the shape of the folded SFS by assuming equilibrium base composition [42] or allowing only recent change in base composition [e.g. 25,26,27], which is not the case in mammals [43] and some Monocots [2], for example. As errors in the polarization of mutations can lead to spurious signatures of selection or gBGC [44], this issue must also be taken into account.

We specifically address the following questions: (i) do neutral or selective forces mainly affect base composition? (ii) if active, what are the intensities of gBGC and SCU and how do

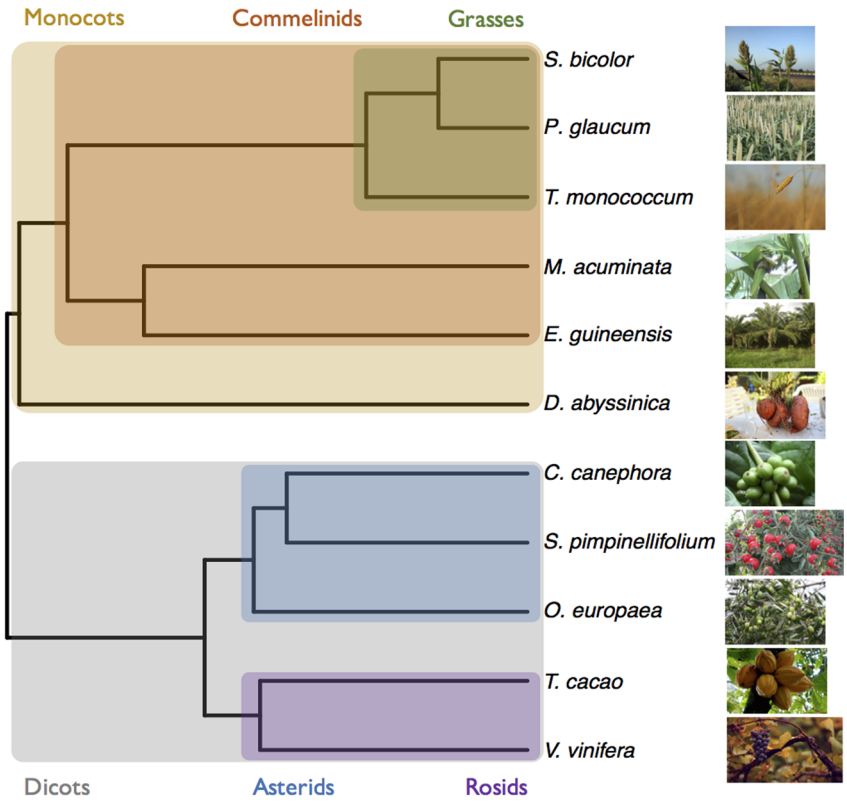

Fig 1. Phylogeny of the species used in this study. Phylogenetic relationship of the species used in this study. The phylogeny was computed with PhyML [75] on a set of 33 1–1 orthologous protein clusters obtained with SiLiX [76] and the resulting tree was made ultrametric (see untransformed trees in S5 and S6 Figs). Images for *S. bicolor*, *T. monococcum*, *D. abyssinica* and *O. europaea* come from the pixabay website. Images for *S. pimpinellifolium* and *M. acuminata* are provided by the authors. All other images come from the Wikimedia website.

https://doi.org/10.1371/journal.pgen.1006799.g001

they vary across species? (iii) are the average gBGC and the 5'-3' gBGC gradient stronger in GC-rich genomes? To do so we used and extended the recent method developed by Glémin et al. [38] that controls for both demography and polarization errors. We applied it to a large population genomic dataset of 11 species spread across the Angiosperm phylogeny to detect and quantify the forces affecting synonymous positions. Our results show that base composition is far from mutation-drift equilibrium in most studied genomes, that gBGC is a widespread process being the major force acting on synonymous sites, overwhelming the effect of SCU and contributing to explain the difference between GC-rich (Commelinids, here) and GC-poor genomes (Eudicots and yam, here).

Results

Building a large dataset of sequence polymorphism and divergence in 11 plant species

We focused our analyses on 11 plant species spread across the Angiosperm phylogeny with contrasted base composition and mating systems (Fig 1 and Table 1). To survey the wide variation observed in Monocots, and in line with the sampling of a previous study [2], we sampled one basal Monocots (*Dioscorea abyssinica*, yam), two non-grass Commelinids (*Musa acuminata*, banana and *Elaeis guineensis*, palm tree) and three grasses with contrasted mating system

Table 1. List of studied species and datasets characteristics.

Species	Name	Group	Mating system	Outgroup 1	Outgroup 2	Reference	# of individuals
Sorghum bicolor	Sorghum	Monocot—Commelinid	Mixed	*Sorghum brachypodum*	*Zea mays*	Genome	9
Pennisetum glaucum	Pearl millet	Monocot—Commelinid	Outcrossing	*Pennisetum polystachion*	*Pennisetum alopecuroides*	Transcriptome	10
Triticum monococcum	Einkorn wheat	Monocot—Commelinid	Selfing	*Taeniatherum caput-medusae*	*Eremopyrum bonaepartis*	Transcriptome	10
Musa acuminata	Banana	Monocot—Commelinid	Outcrossing	*Musa balbisiana*	*Musa becarii*	Transcriptome	10
Elaeis guineensis	Oil palm tree	Monocot—Commelinid	Outcrossing	*Phoenix dactylifera*	*Mauritia flexuosa*	Transcriptome	10
Dioscorea abyssinica	Yam	Monocot—Basal	Outcrossing	*Dioscorea praheensilis*	*Dioscorea trifida*	Transcriptome	5
Coffea canephora	Coffee tree	Eudicot—Asterid	Outcrossing	*Empogona ruandensis*	*Coffea pseudozanguebariae*	Transcriptome	12
Solanum pimpinellifolium	Tomato	Eudicot—Asterid	Mixed	*Solanum melongena*	*Capsicum annuum*	Genome	10
*Olea europaea subsp. europaea**	Olive tree	Eudicot—Asterid	Outcrossing	*Olea europaea subsp. cuspidata*	*Phillyrea angustifloia*	Transcriptome	10
Theobroma cacao	Cocoa	Eudicot—Rosid	Outcrossing	*Herrania nititda*	*Theobroma speciosa*	Genome	10
Vitis vinifera	Grape vine	Eudicot—Rosid	Outcrossing	*Vitis romaneti*	*Vitis riparia*	Genome	12

* Simply noted *Olea europaea* in the rest of the article

https://doi.org/10.1371/journal.pgen.1006799.t001

(*Pennisetum glaucum*, pearl millet, *Sorghum bicolor*, sorghum and *Triticum monococcum*, einkorn wheat). In Eudicots, both Rosids (*Theobroma cacao*, cacao and *Vitis vinifera*, grapevine) and Asterids (*Coffea canephora*, coffee tree, *Olea europaea*, olive tree and *Solanum pimpinellifolium*, tomato) are represented. For practical reasons cultivated species have been chosen but we only sampled wild individuals over the species range, except for palm tree for which cultivated individuals were sampled (See S1 Table for sampling details). In this species cultivation is very recent without real domestication process (19th century [45]). For each species, we used RNA-seq techniques to sequence the transcriptome of about ten individuals plus two individuals from two outgroup species, giving a total of 130 individual transcriptomes. Using transcriptomes has been shown to be a useful approach for comparative population genomics with no or minor bias for genome wide comparison [46,47]. When a well-annotated reference genome was available (see Material and methods), we used it as a reference for read mapping. Otherwise we used a *de novo* transcriptome assembly already obtained for these species (focal + outgroups) [48] (Table 1 and S2 Table). After quality trimming and mapping of the raw reads, we kept contigs with at least one read mapped for every individual, giving between more than 24,000 (*P. glaucum*) and 45,000 (in *O. europaea*) contigs per species (Table 1). This initial dataset was used for gene expression analyses (see below). Genotype calling and filtering of paralogous sequences were performed using the *read2snp* software [47] for each species separately, and coding sequence regions were extracted (see Material and methods). The resulting datasets were used to compute nucleotide diversity statistics that did not require any outgroup information. The number of identified SNPs varies from 4,409 in *T. monococcum* (which suffered from the lowest depth of sequencing) to 115,483 in *C. canephora*. Variations in the numbers of SNPs also revealed the large variation in polymorphism levels with π_S ranging from 0.17% in *E. guineensis* to 1.22% in *M. acuminata*. The level of constraints on proteins, as measured by the π_N/π_S ratio, varies between 0.122 in *T. monococcum* and 0.261 in *E. guineensis* (Table 2).

Table 2. Global statistics for each dataset.

Species	# of contigs			Total length	# of SNPS		Base composition					Polymorphism		
	Total	Genotyped	With outgroup		Total	Polarized	GC	GC3	Average ENC	Codon Preference[a]	Cor(GC3, Expression)[b]	π_S (in %)	π_N (in %)	π_N/π_S
Sorghum bicolor	29448	18518	3884	25849393	77703	12201	0.52	0.56	40.33	15 / 7	0.30	0.407	0.065	0.161
Pennisetum glaucum	24618	12443	9616	8870196	95068	78360	0.48	0.53	39.75	13 / 10	0.27	0.710	0.121	0.170
Triticum monococcum	33381	3766	1319	1758789	4409	3522	0.46	0.48	40.06	26 / 2	0.38	0.272	0.033	0.122
Musa acuminata	36115	14366	10546	6796494	113585	89793	0.49	0.52	39.42	28 / 1	0.31	1.223	0.237	0.194
Elaeis guineensis	26791	14970	9144	10623105	28097	27514	0.47	0.47	39.33	28 / 4	0.28	0.175	0.046	0.261
Dioscorea abyssinica	30551	18497	11544	16125630	84961	49552	0.46	0.46	41.10	26 / 12	0.17	0.417	0.085	0.205
Coffea canephora	28975	13290	9064	11180913	115483	78519	0.45	0.42	40.68	27 / 6	0.22	0.593	0.145	0.245
Solanum pimpinellifolium	34727	12357	1074	9438177	25392	3253	0.43	0.38	42.79	22 / 8	0.18	0.213	0.051	0.238
Olea europaea	45389	12816	8512	6718947	90397	68299	0.44	0.42	39.09	28 / 6	0.23	1.070	0.231	0.216
Theobroma cacao	28798	9918	7901	5510955	37455	32674	0.45	0.42	44.06	27 / 8	0.31	0.484	0.124	0.257
Vitis vinifera	29971	12398	9325	12513219	101351	68315	0.46	0.45	44.30	27 / 8	0.21	0.744	0.147	0.197

GC and GC3 have been computed on the total number of contigs

[a] # of preferred codons ending in G or C / ending in A or T

[b] correlation between GC at third codon positions and gene expression (log10(RPKM))

ENC: effective number of codons (computed with method X)

π_S: nucleotide diversity at synonymous sites

π_N: nucleotide diversity at non-synonymous sites

https://doi.org/10.1371/journal.pgen.1006799.t002

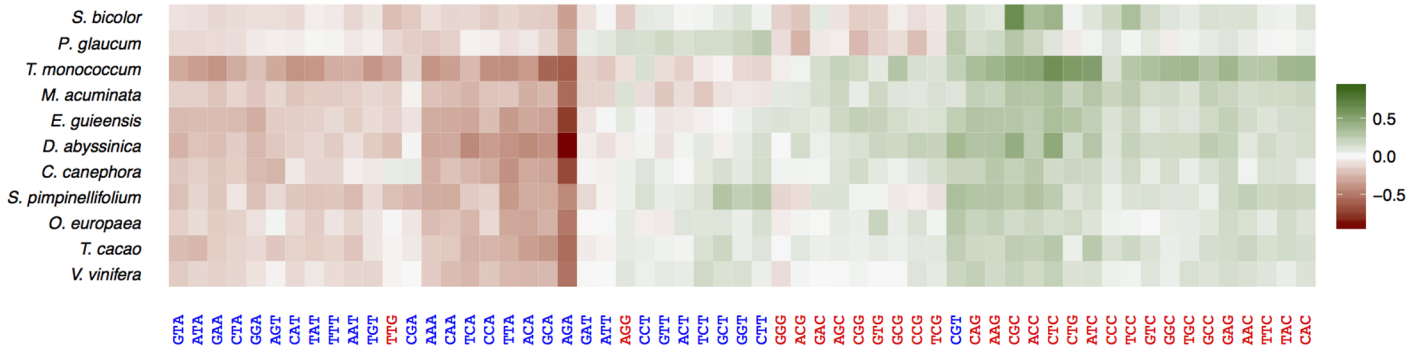

Fig 2. Patterns of codon preference among the 11 studied species. The colour scale indicates the magnitude of Δ RSCU, the difference in the Relative Synonymous Codon Usage between highly and lowly expressed genes. The greenest codons are the most preferred and the reddest the least preferred. Codons ending in G or C are in red and those ending in A or T in blue.

https://doi.org/10.1371/journal.pgen.1006799.g002

For the analyses requiring polarized SNPs, we also added orthologous sequences from two outgroups to each sequence alignment of the focal species individuals (see Material and methods). The number of polarized SNPs ranged from 3,253 in *S. pimpinellifolium* to 89,793 in *M. acuminata*. Other details about the datasets are given in Table 2. Overall, although the dataset does not represent the full transcriptome of each species it allows large-scale comparative analyses.

Base composition, patterns of codon usage and codon preferences vary across species

We first looked at base composition: GC3 varies from 0.38 to 0.44 in Eudicots and from 0.46 to 0.56 in Monocots (Table 2). As observed in previous studies [2,43], these values tend to be lower than genome wide averages (when available) but the relative differences in base composition among species were conserved, notably the GC-poorness of Eudicots compared to Monocots. Grass species exhibited a bimodal GC3 distribution except *T. monococcum* where bimodality was not apparent (S1 Fig). This is likely because the sequencing depth was lower for this species so that GC-rich genes (most likely short ones [37]) have been under sampled. We also characterized codon usage in each species by computing the Relative Synonymous Codon Usage (RSCU) for every codon as the frequency of a particular codon normalised by the frequency of the amino acid it codes for (S3 Table, S2 Fig). Patterns of RSCU were relatively consistent between species but reflected differences of GC content between them, notably a higher usage of G or C-ending codons in GC-rich species.

In order to evaluate the possible effect of selection on codon usage, we defined the sets of preferred (P) and un-preferred (U) codons for each species. The fitness consequences of using optimal or suboptimal codons should be higher in highly expressed genes, causing the usage of optimal codons to increase with gene expression (and that of non-optimal ones to decrease). Thus, we defined preferred (or un-preferred) codons as codons for which RSCU increases (or decreases) with gene expression as in [49] (see Materials & methods for more details). S3 Table shows detailed results for each species. In contrast with genome-wide codon usage, nearly all species showed a bias towards preferred codons ending in G or C (Table 2, Fig 2 and S3 Table), only *P. glaucum* and *S. bicolor* showing a more balanced AT/GC sharing of codon preference. Preferences for two-fold degenerated codons were highly conserved across species, with only GC-ending preferred codon except for aspartic acid and tyrosine in *P. glaucum* (Fig 2, S3 Table). Preferences for other amino acids were slightly more labile but there were always one preferred GC-ending and one un-preferred AT-ending codon common to all species.

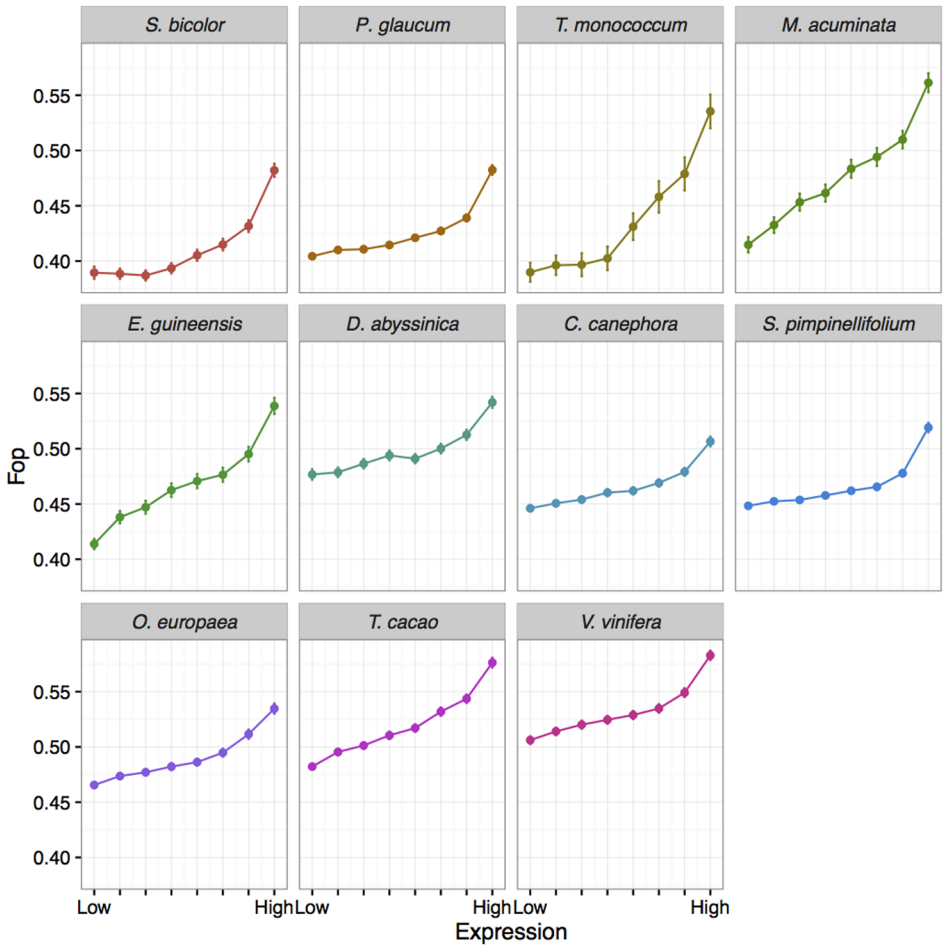

Fig 3. Relationship between the frequency of optimal codons (FOP) and expression in the 11 studied species. For each species, genes have been split into eight categories of expression (based on RPKM) of same size and the mean FOP for each category is plotted with its 95% confidence interval.

https://doi.org/10.1371/journal.pgen.1006799.g003

Frequency of optimal codons of a gene (Fop, *i.e.* the frequency of preferred codons [50]), increased with expression as expected but the difference in Fop between the most highly and most lowly expressed genes was weak to moderate (from ~5% in *C. canephora* to 15% in *T. monococcum* and *M. acuminata*) and tended to be higher in Commelinid species (Fig 3). Because most preferred codons ended with G or C, GC3 and expression were also positively correlated in all species.

Selective-like evolutionary forces affect base composition

To determine which forces affect variation in base composition and codon usage among species, we first evaluated whether base composition at synonymous sites was at mutation-drift equilibrium. Glémin et al. [38] showed that the asymmetry of the distribution of non-polarized GC allele frequencies (measured by the skewness coefficient of the distribution) was a robust test of this equilibrium. This statistic is not affected by possible polarization errors (see later for more on polarization errors). A skewness coefficient equal to 0 is expected under equilibrium whereas negative (or positive) values mean higher (or lower) GC content than expected under mutation-drift equilibrium. The same rationale can be applied to codon frequencies.

Table 3. Skewness, neutrality index (NI) and direction of selection (DoS) statistics for GC content and codon usage.

Species	GC content						Codon usage					
	Mean allele frequency of GC alleles	Skewness	p-value[a]	NI	DoS	p-value[b]	Mean frequency of Pref alleles	Skewness	p-value[a]	NI	DoS	p-value[b]
Sorghum bicolor	0.576	-0.351	**<10E-16**	0.834	0.043	**7.50E-07**	0.535	-0.164	**5.45E-06**	0.94	0.02	0.256
Pennisetum glaucum	0.562	-0.294	**<10E-16**	0.963	0.009	**0.007**	0.534	-0.158	**<10E-16**	0.87	0.03	**3.72E-15**
Triticum monococcum	0.547	-0.222	**1.81E-05**	0.728	0.078	**8.70E-11**	0.550	-0.236	**1.16E-05**	0.71	0.08	**3.84E-11**
Musa acuminata	0.570	-0.343	**<10E-16**	0.827	0.047	**<10E-16**	0.570	-0.344	**<10E-16**	0.83	0.05	**7.01E-15**
Elaeis guineensis	0.540	-0.201	**<10E-16**	0.819	0.050	**3.30E-09**	0.535	-0.170	**3.06E-13**	0.82	0.05	**1.79E-08**
Dioscorea abyssinica	0.554	-0.277	**<10E-16**	0.856	0.037	**0.035**	0.549	-0.252	**<10E-16**	0.87	0.03	0.112
Coffea canephora	0.450	0.234	**<10E-16**	0.913	0.022	**3.13E-05**	0.458	0.199	**<10E-16**	0.92	0.02	**5.47E-04**
Solanum pimpinellifolium	0.534	-0.152	**0.019**	1.132	-0.031	0.051	0.539	-0.174	**0.016**	0.73	0.08	**1.04E-06**
Olea europaea	0.509	-0.047	**0.001**	0.884	0.031	**0.003**	0.510	-0.051	**0.001**	0.89	0.03	**0.017**
Theobroma cacao	0.515	-0.071	**4.66E-04**	0.838	0.044	**7.14E-14**	0.510	-0.045	0.053	0.88	0.03	**5.38E-06**
Vitis vinifera	0.550	-0.229	**<10E-16**	0.737	0.075	**<10E-16**	0.538	-0.172	**<10E-16**	0.66	0.10	**3.80E-88**

[a] Null hypothesis: skewness = 0

[b] Null hypothesis: NI = 1 / DoS = 0 (equivalent test done on the same contingency table).

https://doi.org/10.1371/journal.pgen.1006799.t003

We found that GC content and the frequency of preferred codons were significantly higher than predicted by mutational effects in all species, with the exception of coffee, which interestingly showed a lower GC content than expected under mutation-drift balance (Table 3).

As base composition equilibrates slowly under mutation pressure [33], non-equilibrium conditions could be due to long-term changes in mutational patterns. To test further whether selective-like forces can explain the excess of GC and preferred codons, we developed a modified MacDonald Kreitman test [51] comparing W→S (or U→P) to S→W (or P→U) polymorphic and divergent sites (Material & Methods and S1 Text). SNPs and fixed mutations (substitutions) were polarized by parsimony using two outgroup taxa for each focal species. We built contingency tables by counting the number of polymorphic or divergent sites for each of the two mutational categories. From these contingency tables, we computed neutrality, NI, [52] and direction of selection, DoS, [53] indices. In the case of selective-like forces favouring the fixation of W→S or U→P mutation, NI values are expected to be lower than 1 and DoS values to be positive. P-values were computed from a Chi-squared test on the contingency tables. NI was lower than 1 and DoS positive in all species except *S. pimpinellifolium* (Table 3), indicating that selective-like forces drove the fixation of GC and preferred codon alleles. In *P. glaucum*, although significant, the departure from the neutral expectation for GC content is minute, which can be explained by very weak gBGC but also by a recent increase in its intensity (see Results below and S1 Text). Overall, this analysis showed that in most species selective-like forces tended to drive base and codon composition away from

their mutational equilibrium. Selection and gBGC are the two known alternatives whose effects have to be distinguished.

Disentangling gBGC and SCU?

Although they may have different mechanistic causes and biological consequences, selection and gBGC leave similar evolutionary footprints and are not easy to disentangle, especially in species where most preferred codons end in G or C (Table 2). We first applied correlative approaches to try to disentangle both processes. Then we tried to quantify their respective intensities.

Under the SCU hypothesis, departure from neutrality should be stronger for highly expressed genes and/or genes with strongly biased codon composition. Under the gBGC hypothesis, departure from neutrality should increase with recombination rates. However, recombination data was not available in our datasets. As gBGC leads to an increase in GC content, departure from neutrality should thus also increases with GC content. We split synonymous SNPs and substitutions into eight groups of same size according to their GC3 or their gene expression level (measured by the mean RPKM values across all individuals of a given population), and computed the NI and DoS indices for each category based on W/S or U/P changes. For all species except *D. abyssinica* and *S. bicolor*, we found a strong positive (or negative) correlation between GC3 and DoS (or NI), indicating a stronger bias in favour of S alleles in GC-rich genes (Fig 4). In contrast, the relationship between expression level and DoS or NI measured on codon usage was weaker, with more variable and on average lower correlation coefficients (Fig 4). These results tend to point out gBGC as a stronger force than SCU affecting synonymous variations in our datasets.

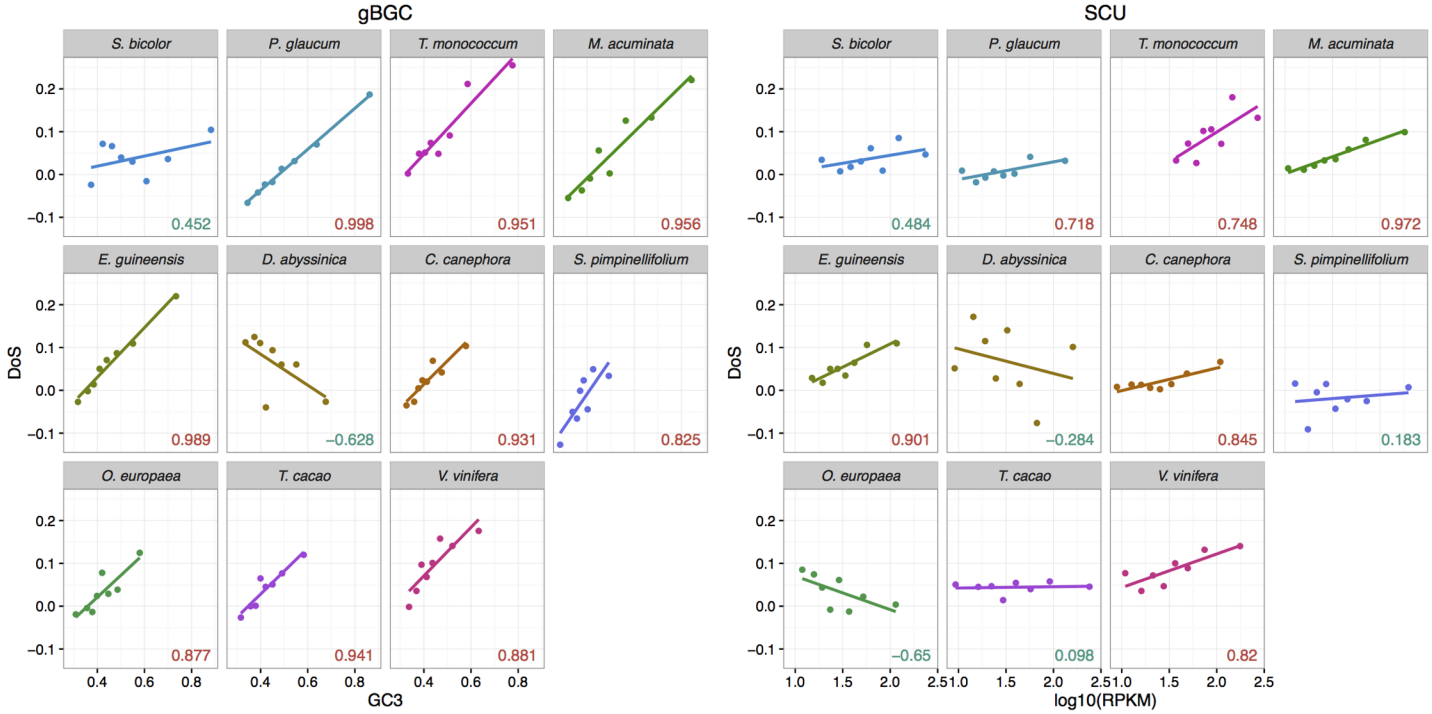

Fig 4. DoS statistics as a function of GC3 and expression level. Correlation between GC3 and DoS computed on WS changes (left panel) or between expression level (measured through RPKM) and DoS computed on UP changes (right). Pearson correlation coefficients are given for each species (red: significant at the 5% level, blue non-significant).

https://doi.org/10.1371/journal.pgen.1006799.g004

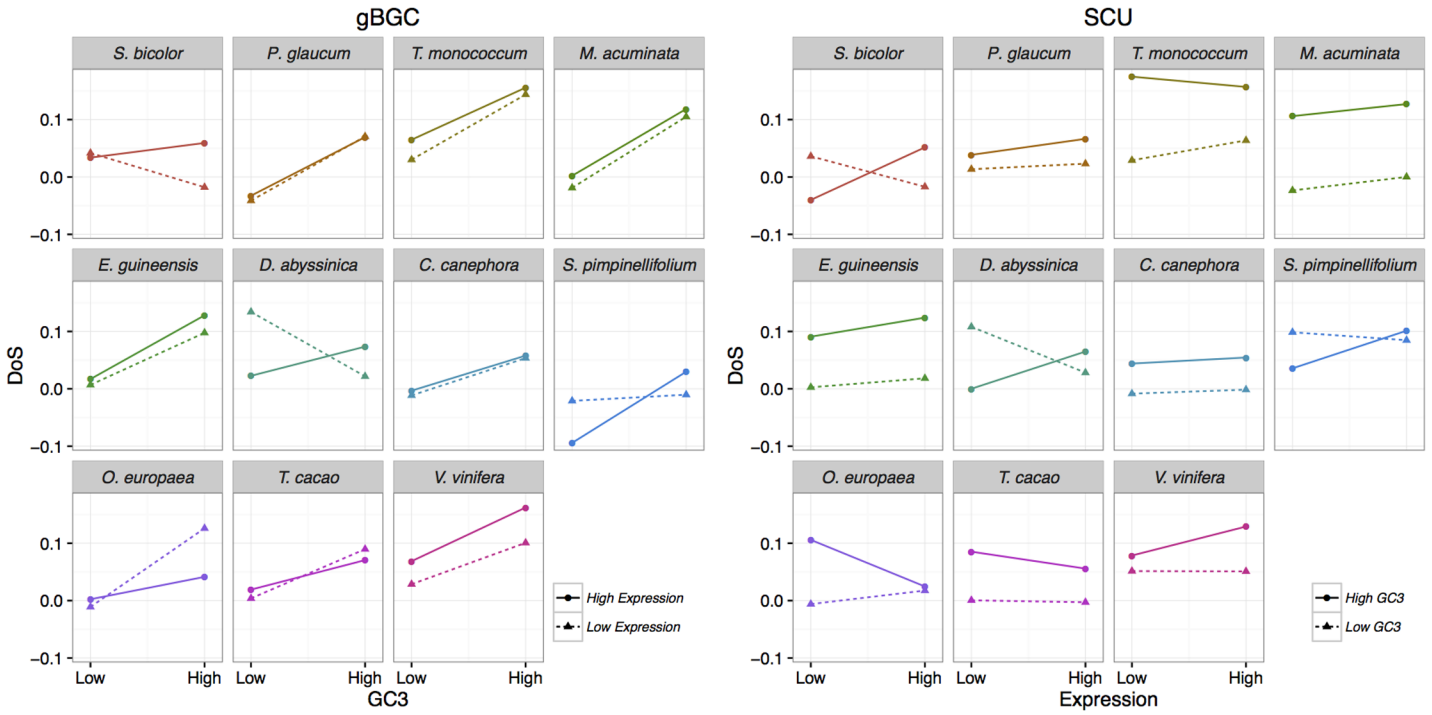

Fig 5. Combined effect of GC3 and expression level on DoS statistics. The DoS statistics was computed on W/S (gBGC) or U/P (SCU) changes for four gene categories: GC-rich and highly expressed, GC-rich and lowly expressed, GC-poor and highly expressed, GC-poor and lowly expressed.

https://doi.org/10.1371/journal.pgen.1006799.g005

We then split our datasets into four independent categories based on two GC3 groups crossed by two expression level groups to test which factor has the strongest effect on the bias towards S or P alleles. The rationale is that SCU should make the bias towards P alleles increase with gene expression independently of GC3. On the other hand, gBGC should increase the bias towards S alleles with GC3 independently of gene expression. We found that DoS clearly increased with GC3 in all species for both lowly and highly expressed genes, with the exception of *D. abyssinica* and *S. bicolor* where it decreased for lowly expressed genes, and *S. pimpinellifolium* where there was little change for lowly expressed genes. On the other hand, the effect of expression on DoS was inconsistent or only weak in most species (Fig 5). These results confirm that the effect of gBGC appears stronger than the effect of SCU.

Estimation of gBGC/SCU intensity and mutational bias

To evaluate further the forces affecting base composition we estimated the intensity of selection ($S = 4N_e s$) and gBGC ($B = 4N_e b$) from site frequency spectra (SFS). SFS for all species are represented in S3 Fig. We used the method recently developed by Glémin et al. [38] that takes SNP polarization errors into account, which avoids observing spurious signature of selection or gBGC. As mentioned above, the observed pattern in *P. glaucum* (excess of GC content but almost no departure from neutrality according to the NI and DoS indices, see Table 3) suggests a recent change in the intensity of selection and/or gBGC. Also, transition to selfing, which usually can be very recent in plants [54], could have effectively shut down gBGC in the recent past due to a deficit in heterozygous positions. To capture these possible changes of fixation bias through time, we extended the model of [38] by combining frequency spectra and divergence estimates as summarized on Fig 6 (and see S2 Text for full details). Divergence is determined by both mutation and selection/gBGC so it is not possible to disentangle these two

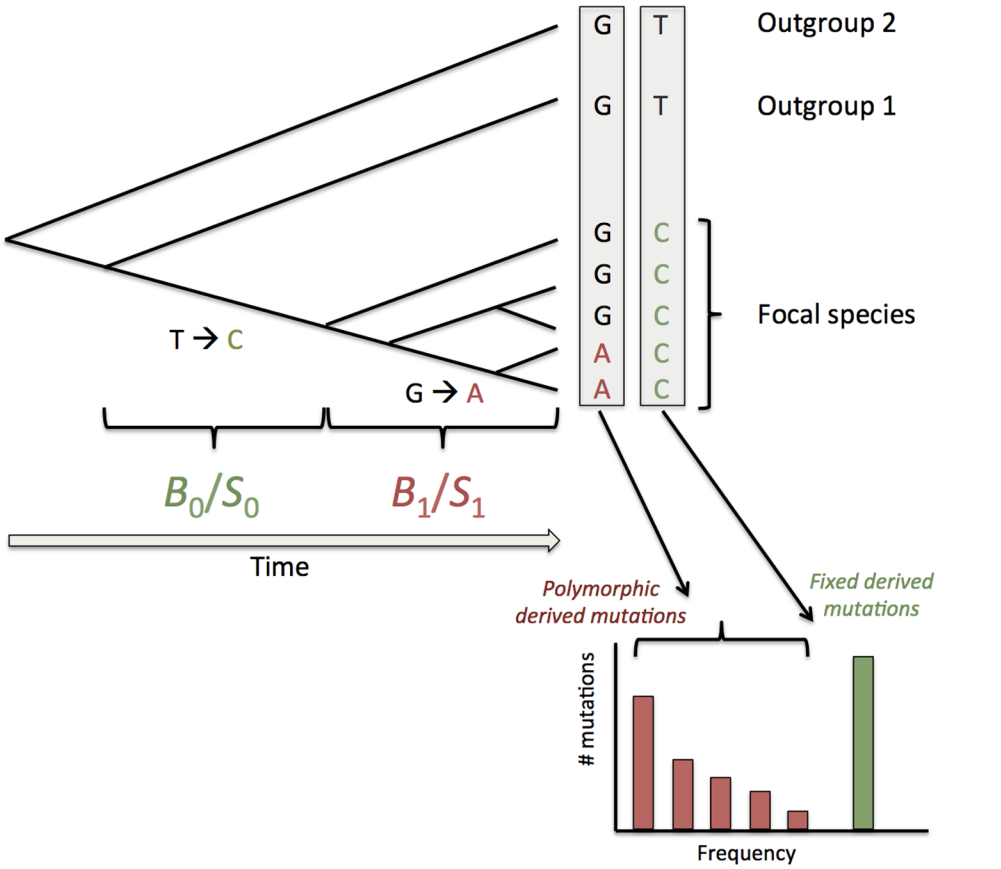

Fig 6. Schematic presentation of the method to estimate recent and ancestral gBGC or SCU. In addition to polymorphic derived mutations used to infer recent gBGC or selection (B_1/S_1) as in [38] we also consider substitutions (*i.e.* fixed derived mutations) on the branch leading to the focal species. Each box corresponds to a site position in a sequence alignment. Both kinds of mutations are polarized with the two same outgroups and are thus sensitive to the same probability of polarization error. We assume that gBGC and selection may have change so that fixed mutations may have undergo a different intensity. Note that these two *B* or *S* values correspond to average of potentially more complex variations over the two periods.

https://doi.org/10.1371/journal.pgen.1006799.g006

factors from the divergence data alone. However, if we assume constant and identical mutation bias at the polymorphism and the divergence level, this leave enough degrees of freedom to fit an additional *S* or *B* parameter. Thus, we assumed a single mutation bias but two different selection/gBGC intensities, one fitted on polymorphism data and the other on divergence. We evaluated the statistical significance of the shift in intensity by a likelihood ratio test with the model where the two intensities were equal (*i.e.* no change over time). Simulations showed that not taking polarization errors into account can bias selection/gBGC estimates as already shown in [38] and also leads to spurious detection of changes in selection/gBGC intensities (S2 Text). Simulations also showed that the estimated differences between the two intensities were often underestimated. This is expected as *B* values estimated in the model correspond to averages over the conditions that mutations have experienced during their lifetime (drift and gBGC/selection intensities), so it depends on when changes occurred. However, the method accurately retrieved the appropriate weighted averages for B_0 and B_1 and efficiently accommodates for demographic variations (see S2 Text). Overall, the test of heterogeneity of selection/gBGC is a conservative approach. If we relax the assumption of constant mutational bias, changes in both bias and selection/gBGC are no more identifiable. Recent *S*/*B* estimates are

not affected but ancestral estimates are underestimated (resp. overestimated) when mutation bias decreases (resp. increases). However, the method is still powerful to detect departure from a constant regime of selection/mutation/drift equilibrium (S2 Text).

We applied the method to the total frequency spectra, either for W/S or U/P polymorphisms and substitutions. In all species, significant (at the 5% level) gBGC or SCU were detected but at low intensity (B or $S < 1$, Table 4). In four species (*P. glaucum*, *E. guineensis*, *D. abyssinica* and *V. vinifera*) we found significant differences between ancestral and recent intensities for gBGC and/or SCU. In particular, the recent significant increase in gBGC in *P. glaucum* (from 0.224 to 0.524, Table 4) can explain why NI is very close to one (or DoS close to zero) (see above and S1 Text). On average, Monocots, especially Commelinids species tended to exhibit stronger gBGC than Eudicots and B tended to increase with mean GC3, but no relationship is significant with only 11 species when either B_0 or B_1 are used. However, using the constant B estimates (S4 Table), weakly significant relationships were found for the difference between Commelinids and other species (Wilcoxon test: p-value = 0.0519) and the correlation between B and GC3 ($\rho_{Spearman} = 0.691$, p-value = 0.023). No significant relationship was found for SCU. No significant relationship between B or S and π_S was found either.

As the two processes are entangled, it is difficult to properly and separately estimate their respective intensities. To do so, we developed a second extension of the method of [38]. Combining the two processes, nine kinds of mutations can occur (see S2 Text). By assuming that selection and gBGC act additively, it is in theory possible to estimate separately the two effects. We fit a general model to the nine SFS and the nine substitution counts, with a constant mutation bias, two B and two S values. The details of the model are reported in S2 Text. Simulations showed that the method could efficiently estimate both gBGC and SCU but tended to slightly underestimate recent gBGC and overestimate recent SCU (S2 Text). When the distributions of SNPs and substitutions are highly unbalanced (typically S/P and W/U states are confounded and there are very few WS-PU and SW-UP mutations), it is more difficult to detect both effects with a significant level (S2 Text). Finally, if assignation of codon preference is not perfect, typically for four-fold and six-fold degenerated codons, this could also underestimate SCU and reduce the power to detect it, especially for highly unbalanced dataset for which it is anyway inherently difficult to distinguish gBGC and SCU (see S2 Text). For both selection and gBGC and both ancestral and recent periods, we either fixed the value to 0 or let it be freely estimated, leading to 16 different models. For each species, the best model according to AIC criteria (see Methods) is given in Table 5 while all results are given in S5 Table. In six species the model with only gBGC was the best one, this could also include *M. acuminata* where it was not possible to disentangle between gBGC and SCU. For three species, the best model included both gBGC and SCU and only *S. pimpinellifolium* appeared to be affected by SCU but not gBGC. If codon preferences were perfectly determined, this result is expected to be robust and conservative because simulations suggest that SCU is slightly more easily detected than gBGC. If there were some errors in codon preference identification, this can partly explain that SCU was less often detected. However, the species for which SCU was not detected did not present the most unbalanced codon preference (see Table 2) and identification error rate should have been rather high (>20% see S2 Text) to strongly bias results. Overall, this confirms that synonymous sites are widely affected by gBGC in the studied plant species and that SCU either only plays a minor role or is partly masked by the effect of gBGC.

This method also allowed us to estimate mutation bias. As already observed in most species, mutation was biased towards AT alleles, with a bias slightly ranging from 1.6 to 2.2 (Table 4), which is of the same order as what was found in humans [38,55]. Interestingly, *C. canephora* was again an exception with almost no mutational bias ($\lambda = 1.05$).

Table 4. Separated estimations of recent and ancestral gBGC ($B = 4N_e b$) and SCU ($S = 4N_e s$).

Species	gBGC					
	lambda	$4N_e b$ ancestral	$4N_e b$ recent	p-value ancestral = 0	p-value recent = 0	p-value recent = ancestral
Sorghum bicolor	1.61 [1.51–2.69]	0.378 [0.290–0.516]	0.078 [-0.492–0.739]	**2.73E-14**	0.758	0.189
Pennisetum glaucum	1.73 [1.69–1.83]	0.224 [0.189–0.261]	0.524 [0.383–0.661]	**<10E-16**	**1.15E-13**	**2.18E-06**
Triticum monococcum	1.99 [1.67–2.25]	0.448 [0.269–0.613]	-0.008 [-0.824–0.691]	**1.39E-05**	0.985	0.164
Musa acuminata	1.71 [1.66–1.80]	0.313 [0.253–0.370]	0.397 [0.234–0.546]	**<10E-16**	**2.68E-06**	0.343
Elaeis guineensis	1.84 [1.77–1.93]	0.328 [0.267–0.400]	0.516 [0.328–0.702]	**<10E-16**	**1.76E-07**	**0.034**
Dioscorea abyssinica	2.20 [2.10–2.47]	1.171 [0.127–4.067]	0.008 [-0.221–0.264]	**0.032**	0.949	0.072
Coffea canephora	1.05 [1.02–1.10]	0.154 [0.110–0.202]	0.243 [0.113–0.366]	**9.47E-11**	**3.77E-04**	0.171
Solanum pimpinellifolium	2.05 [1.74–2.63]	0.114 [-0.057–0.392]	0.759 [-0.491–3.785]	0.215	0.153	0.193
Olea europaea	1.58 [1.53–1.64]	0.167 [0.080–0.268]	0.031 [-0.127–0.168]	**<10E-16**	0.687	0.132
Theobroma cacao	1.67 [1.59–1.74]	0.316 [0.258–0.377]	0.465 [0.222–0.683]	**<10E-16**	**6.54E-05**	0.135
Vitis vinifera	2.15 [2.08–2.22]	0.360 [0.318–0.413]	0.024 [-0.101–0.153]	**<10E-16**	0.71	**1.55E-08**
Species	**SCU**					
	lambda	$4N_e s$ ancestral	$4N_e s$ recent	p-value ancestral = 0	p-value recent = 0	p-value recent = ancestral
Sorghum bicolor	2.04 [1.70–2.47]	0.139 [0.023–0.260]	0.439 [-0.251–1.083]	**0.010**	0.143	0.341
Pennisetum glaucum	1.76 [1.70–1.87]	0.181 [0.137–0.226]	0.126 [-0.062–0.289]	**2.33E-15**	0.165	0.484
Triticum monococcum	2.84 [2.33–3.31]	0.534 [0.353–0.718]	0.236 [-0.610–1.029]	**1.14E-06**	0.581	0.409
Musa acuminata	2.02 [1.96–2.15]	0.315 [0.256–0.362]	0.392 [0.221–0.553]	**<10E-16**	**5.21E-06**	0.394
Elaeis guineensis	1.58 [1.50–1.66]	0.324 [0.233–0.396]	0.512 [0.322–0.704]	**3.00E-15**	**6.51E-07**	**0.043**
Dioscorea abyssinica	1.68 [1.39–1.74]	1.909 [0.306–9.994]	-0.101 [-0.311–0.135]	**0.023**	0.470	**0.037**

(*Continued*)

Table 4. (*Continued*)

Coffea canephora	0.89 [0.86–0.95]	0.148 [0.079–0.197]	0.196 [0.039–0.330]	**5.91E-08**	**0.012**	0.515
Solanum pimpinellifolium	1.56 [1.32–2.05]	0.465 [0.270–0.857]	0.566 [-0.567–3.900]	**3.39E-06**	0.285	0.834
Olea europaea	1.18 [1.13–1.22]	0.148 [0.040–0.241]	0.025 [-0.162–0.186]	**0.004**	0.772	0.214
Theobroma cacao	1.09 [1.02–1.16]	0.245 [0.167–0.339]	0.397 [0.107–0.673]	**2.85E-11**	**3.00E-03**	0.185
Vitis vinifera	1.26 [1.22–1.32]	0.470 [0.421–0.525]	0.118 [-0.028–0.258]	**<10E-16**	0.103	**7.09E-08**

https://doi.org/10.1371/journal.pgen.1006799.t004

Variation along genes

So far, we considered either global effects at the transcriptome scale or variations among genes belonging to different categories. However, most plant species exhibit a more or less pronounced gradient in base composition from 5' to 3' [1], which is strongly linked to exon-intron structure [37]. In particular, in some species the first exon is much GC-richer than other exons. Moreover, it has been proposed that this gradient could be due to a gBGC gradient

Table 5. Best model for the joined estimations of recent and ancestral gBGC ($B = 4N_eb$) and SCU ($S = 4N_es$).

Species	$4N_eb$ ancestral	$4N_eb$ recent	$4N_es$ ancestral	$4N_es$ recent
Sorghum bicolor	0.439 [0.334–0.525]	0	0	0
Pennisetum glaucum	0.218 [0.182–0.253]	0.561 [0.393–0.689]	0.139 [0.106–0.175]	0
Triticum monococcum	0.264 [0.042–0.443]	0	0.247 [0.027–0.468]	0
Musa acuminata 1	0.312 [0.281–0.395]	0.394 [0.241–0.580]	0	0
Musa acuminata 2	0	0	0.317 [0.284–0.400]	0.398 [0.176–0.540]
Elaeis guineensis	0.329 [0.241–0.383]	0.517 [0.234–0.744]	0	0
Dioscorea abyssinica	1.256 [0.564–2.202]	0	0	0
Coffea canephora	0.154 [0.119–0.227]	0.244 [0.070–0.361]	0	0
Solanum pimpinellifolium	0	0	0.459 [0.311–0.603]	0
Olea europaea	0.168 [0.074–0.250]	0	0	0
Theobroma cacao	0.318 [0.241–0.383]	0.474 [0.234–0.744]	0	0
Vitis vinifera	0.256 [0.216–0.295]	0	0.380 [0.323–0.439]	0

For *Musa acuminata* the two best models with very close AIC values are given.

https://doi.org/10.1371/journal.pgen.1006799.t005

associated with a recombination gradient [33]. To quantitatively test this hypothesis, we separated SNPs and fixed derived mutations as a function of their position along genes. The best choice would have been to split them according to exon ranking [37]. However, as exon annotation was lacking (or imprecise) for most species in our datasets, we split contigs into two sets: the first 252 base pairs, corresponding to the median length of the first exon in *Arabidopsis*, banana and rice (Gramene database [56]), used as a proxy for the first exon, and the rest of the contig. We then estimated *B* on these two sets of contigs. Some imprecision in the "first exon" definition and variation in transcript length among species reduced the power of this analysis and results should be interpreted with caution. However, we did not expect that it could create artifactual *B* gradient as the use of a stringent criterion reinforced the observed patterns despite reducing datasets (see below).

For all species except *D. abyssinica* and *S. pimpinellifolium*, the ancestral *B* was higher in the first part than in the rest of contigs. The signature was less clear for recent *B* as far less values were significant. Ancestral and recent *B* were not significantly different in most species (S6 Table). To illustrate the global pattern, Fig 7 shows average gBGC gradients for all species, *i.e.* assuming the same ancestral and recent *B* values. Interestingly, while there was no clear taxonomic effect on global gBGC estimates (Table 4), there was a sharp difference between Commelinid species and the others for the first part of contigs (Wilcoxon test p-value = 0.030, Fig 7C), in agreement with the strong 5'– 3' GC gradient observed in these species [1,2]. *B* values and GC3 tended to be positively correlated on the first part of contigs ($\rho_{Spearman}$ = 0.591, p-value = 0.061) but not significantly in the rest ($\rho_{Spearman}$ = 0.382, p-value = 0.248). These analyses were performed on all contigs but some of them do not start by a start codon. We restricted the analyses to the subset of contigs starting by a start codon and we found very similar results with stronger statistical support: in the first exon, *B* was significantly higher in Commelinids than in other species (Wilcoxon test p-value = 0.0043) and *B* values and GC3 were significantly and positively correlated both on the first part of contigs ($\rho_{Spearman}$ = 0.80, p-value = 0.0052) and in the rest of contigs ($\rho_{Spearman}$ = 0.70, p-value = 0.0208) (S6 Table and S4 Fig). In line with previous results showing that first exons contribute to most of the variation in GC content among species [2,33,37], these results show that species also mostly differ in their gBGC intensities in the first part of genes.

Discussion

Selective-like evolution of synonymous variations in plant genomes

It has already been shown that base composition in grass genomes is not at mutation-drift equilibrium with both gBGC and selection increasing GC content despite mutational bias toward A/T [31]. Our results demonstrate that even in GC-poor genomes base composition is not at mutation-drift equilibrium, implying that selective-like forces are widespread in all the 11 plant species we studied. In all species, either the skewness and/or the DoS/NI statistics show evidence of departure from equilibrium and purely neutral evolution (Table 3). All species except *C. canephora* have higher GC content than predicted by mutational effect alone, which could be explained by a mutation/gBGC (or selection)/drift balance.

The case of *C. canephora* remains intriguing. Mutation seems not to be biased towards AT as observed in all mutation accumulation experiments [reviewed in 57] and through indirect methods [58]. So far, GC biased mutation has only been observed in the bacteria *Burkholderia cenocepacia* [57]. However, despite no apparent or very weak AT mutational bias and evidence of both recent and ancestral gBGC (Table 4), GC content is rather low (GC3 = 0.42, Table 2) and lower than expected under mutation pressure alone ($1/(1+\lambda)$ = 0.49) as revealed by the positive skewness statistics (Table 3). Preferred codons mostly end in G or C (Table 2) so that

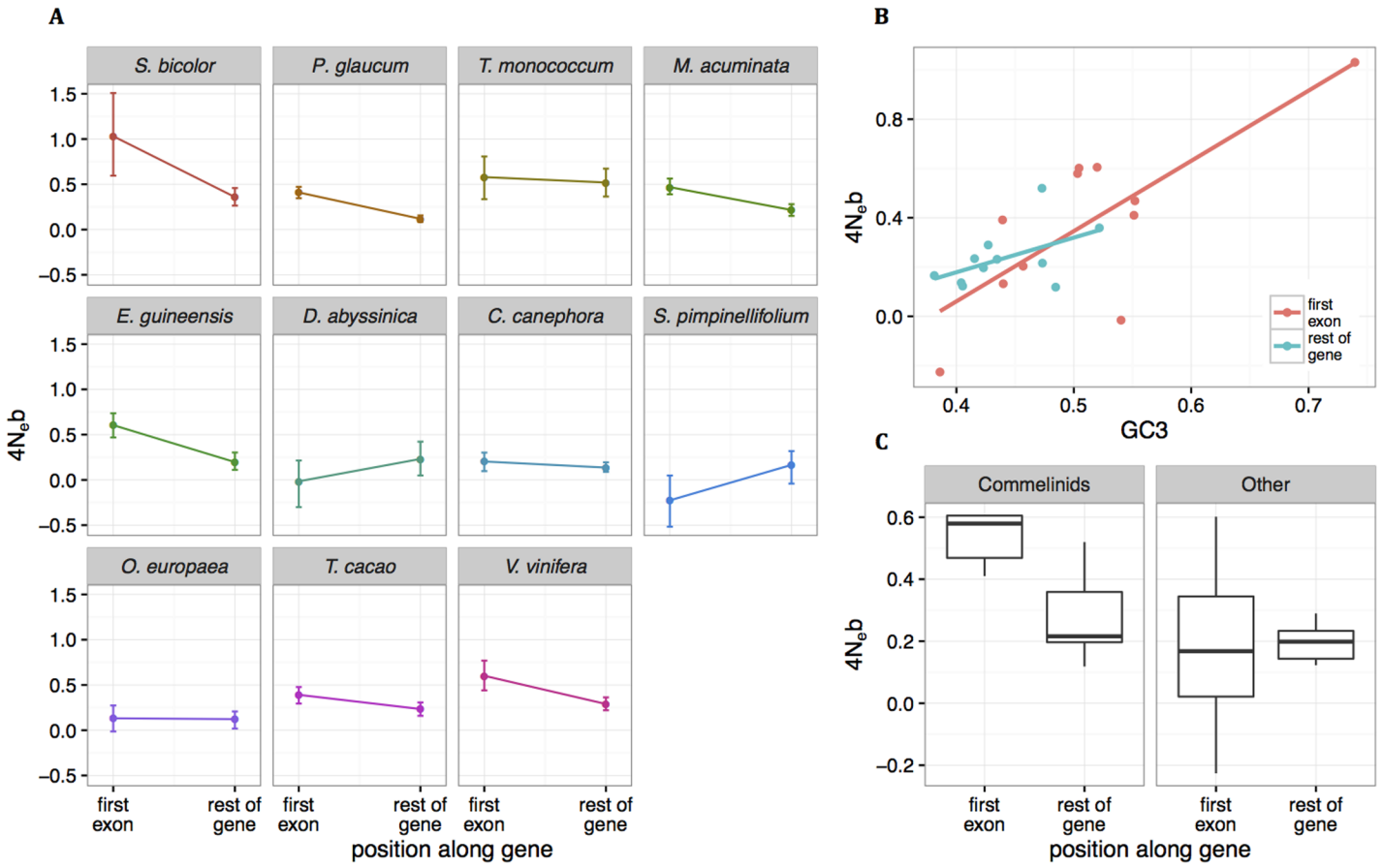

Fig 7. GC3 and gBGC gradients along genes. A: gBGC strength estimations ($4N_eb$) for first exons (252 first bp of contigs) and rest of gene. Error bars indicate the 95% confidence intervals. With the exception of *D. abyssinica* and *S. pimpinellifolium*, all species exhibit stronger gBGC in the first exons compared to the rest of genes. B. Correlations between GC3 and gBGC strength in first exons (red) and rest of genes (blue). Each dot corresponds to one species. GC3 and $4N_eb$ tend to be positively correlated in both regions: $\rho_{Spearman} = 0.591$, p-value = 0.061 for first exons and $\rho_{Spearman} = 0.382$, p-value = 0.248 for the rest of genes. C. Comparison of 4Neb estimates between first exons and rest of genes for Commelinids (all Monocots with the exception of *D. abyssinica*, left panel) and other species (right panel). $4N_eb$ values are higher in first exons compared to rest of genes in Commelinids species, while other species exhibit no differences between first exons and rest of genes.

https://doi.org/10.1371/journal.pgen.1006799.g007

SCU is not a possible explanation for this low GC content. Rather, a recent change in mutation bias is a more probable explanation. Using $B_0 = 0.154$ or $B_1 = 0.243$ (Table 4), a mutational bias of 1.61 or 1.76 would be necessary to reach the observed GC3 (= 0.42). Such values are in the same range as observed for the other species. *D. abyssinica* is another intriguing case where DoS decreases with GC content, contrary to other species (Fig 4). We currently have no clear hypothesis to explain this pattern and it should be viewed with caution because DoS is estimated with few substitutions in this species but it would be compatible with an increase in AT mutation bias with GC content. Further investigation of mutational patterns in these species would be useful to understand better these two intriguing cases.

Beyond departure from equilibrium, comparison of ancestral and recent gBGC or selection also reveals the dynamic nature of forces affecting base composition. At least four species (*P. glaucum*, *E. guineensis*, *D. abyssinica* and *V. vinifera*) show evidence of significant change in gBGC and/or SCU intensity over time (Table 4). If we consider the first part of genes only, changes also occurred in *M. acuminata* and *T. cacao* (S6 Table). Moreover, our method is conservative (see S2 Text) so we may have missed variations in other species. Changes occurred in

both directions. In the three selfing or mixed mating species (*S. pimpinellifolium*, *T. monococcum*, and *S. bicolor*) the ancestral gBGC or SCU intensity is significantly positive but the recent one is null. This is supported by the rather recent evolution of selfing in these species, which nullifies the effect of gBGC through the increase in homozygosity levels and reduces the efficacy of selection [59]. In other species, gBGC or SCU have increased (e.g. *P. glaucum*) or decreased (e.g. *V. vinifera*). Recalling that $B = 4N_e r b_0$ (see Introduction), this could be explained by changes in effective population size (N_e) recombination rate (r), gBGC intensity per recombination event (b_0) and also conversion tract length, which might also vary among species [60]. To date, we do not know anything about the stability of b_0 across generations and how fast it can evolve. In some species, such as mammals, recombination can evolve very rapidly, at least at the hotspot scale [61] but it can be more stable in other species like in birds [62], yeast [63] or maize [64]. Moreover, we average gBGC over the whole transcriptome so recent genome-scale changes in recombination should be necessary to explain changes in B. Although recent changes in r and b_0 are possible, changes in effective population size over time appears to be the most likely explanation.

Selective-like evolution and non-equilibrium conditions can have practical impacts on several genomic analyses. First, gBGC can lead to spurious signatures of positive selection [65], significantly increasing the rate of false positive in genome scan approaches in mammals [66]. This problem should also be taken into account in plant genomes, even in GC-poor ones. Second, SCU/gBGC and non-stationary evolution, due for instance to changes in population size, can strongly affect the estimation of the rate of adaptive evolution through McDonald-Kreitman approaches, especially at high GC content [67]. In species far from equilibrium such as Commelinids, it should be an issue to consider.

gBGC, SCU or both?

Technical issues. We found clear evidences that base composition evolution is not driven only by mutation. However, it was more difficult to distinguish gBGC from SCU because we only used coding regions in our study. Unfortunately, we were not able to use 5' or 3' flanking regions to compare them with synonymous coding positions. These flanking regions were too short and of lower sequencing coverage and quality: they were not frequently sequenced and corresponded to sequence ends. Comparison with introns or non-coding regions would be helpful in the future to confirm our findings, as it was done in rice [31] or maize [32]. To bypass this problem, we developed a new method that jointly estimates gBGC and SCU and allows testing which processes are significant. However, the two processes are especially difficult to distinguish in species where most preferred codons end in G or C, such as *M. acuminata* and *T. monococcum* (Tables 2 and 5 and S2 Text) and when the power is limited by the number of SNPs (*S. pimpinellifolium* and *T. monococcum*). An additional problem is that codon preferences can be imperfectly characterized (whereas there is no ambiguity to define W and S positions). When codon preference are correctly identified, simulations suggest that weaker SCU than gBGC could be estimated even for a highly unbalanced dataset (at least ancestral SCU, see S2 Text). However, it becomes more problematic for unbalanced dataset when some preferences are incorrectly identified, reducing the power to detect SCU (S2 Text). Finally, correlative approaches with GC content and expression can also help distinguishing the two processes. Overall, although each individual result (species-specific and or approach-specific) can be insufficiently conclusive, they collectively point towards the general conclusion of a major contribution of gBGC and a lower contribution of SCU, or a contribution partly masked by gBGC, to explain synonymous variation in the studied plant species.

Predominant signature of gBGC. The combination of our different results suggests that gBGC prevails over SCU in the studied plants. While signatures of gBGC were detected in all species but *S. pimpinellifolium*, SCU was detected only in four or five species (Table 5). However, in these species, the change in NI/DoS with expression is consistent with SCU only in *P. glaucum* (Fig 4). These poorly supported results do not necessarily mean that SCU is not active. Indeed, we were able to defined preferred codons in all our species, and Fop increases with expression level in all of them (Fig 2). However, changes in Fop with expression are moderate to low (15% to 5%) and on average lower to what was observed in *Drosophila* (15%) or *Caenorhabditis* (25%), but slightly higher than *Arabidopsis* (5%) [49]. Thus, SCU is likely active but at a level too low to be detected by our methodology in some species, especially because gBGC masks its effect. In some species such as maize, recombination and gene expression levels are positively correlated as they mainly occurred in open chromatin regions of the genome [32]. This could affect the ability to identify preferred codons because S alleles would increase with expression (and be considered as preferred) because of gBGC, not SCU. Beyond the potential methodological artefact, it also means that gBGC would counteract (for W preferred codons) or reinforce (for S preferred codons) the action of SCU, with a global reduction of SCU on average [68]. A larger dataset (increasing both the number of SNPs and of individuals) would probably be necessary to properly estimate SCU in the presence of gBGC, especially when the most preferred codons end with G or C. It should be noted that in *P. glaucum*, one of the species where SCU was quite confidently detected, a high number of SNPs and a rather equilibrated patterns of codon preference were identified. Finally, in *Drosophila*, it was shown that SCU varies among codons [27], while we only assumed a constant selection coefficient. Generalization of our model by including the approach of [27] is likely a promising avenue to dissect the interaction between gBGC and SCU.

Coevolution between GC and codon usage?. The difficulty in distinguishing gBGC and SCU also raises the question of the interaction between these two processes. The predominance of GC ending preferred codons has also been observed in many bacteria [69]. The bias towards GC ending preferred codons increases with genomic GC content, with species having a GC content higher than 40% being strongly biased towards GC preference [69]. The classical Bulmer's model of coevolution between preferred codons and tRNA predicts a match between the frequency of tRNAs and preferred codons with two equivalent stable states (either AT or GC preference), and so does not explain the observed bias in preference [70]. However, our results are compatible with a modified version of this model in which an external force on base composition is introduced [71]. We propose that gBGC could act as such a force. By increasing GC content, gBGC could disrupt the co-evolutionary equilibrium between preferred codons and tRNAs abundance towards a higher level of GC preference. This would in turn leads to the confounding effects of gBGC and SCU.

GC content gradient and the gBGC hypothesis

We detected gBGC in all but one species but its intensity is rather weak (Tables 4 and 5 and S4 and S5 Tables), of the same order to what was estimated in humans [38] but lower than in other mammals [39], maize [72], and particularly honey bee [41]. Low values can be explained by the fact that we only estimated average *B* values. In many plants studied so far, recombination was found to be heterogeneous along chromosomes [e.g. 36] and locally occurring in hotspots [e.g. 34,35,64], so that gBGC can be locally much higher than average estimates. However, we did not apply the hotspot model proposed by [38] because it behaves poorly when not constrained by additional information on hotspot structure, which we lack in the species studied here. In addition, recombination hotspots are preferentially located outside

genes, especially in 5' upstream regions (and 3' downstream regions to a lesser extent) [34,35,36]. As we estimated gBGC intensities within coding regions, this can also explain why we only estimated rather weak *B* values.

A consequence of this specific recombination hotspot location is the induction of a 5'– 3' recombination gradient along genes (or more generally an exterior to interior gradient if also considering downstream location) [34,35]. Recently, it has been proposed that this recombination gradient could explain the 5'– 3' gradient observed in grasses and more generally in many plant species [33]. We tested this model by looking at signatures of gBGC along contigs in our datasets. In agreement with this model, we found stronger gBGC signatures at the 5' end of contigs compared to the rest of contigs in most of our species (Fig 7). The fact that we observed this gBGC gradient in both Eudicots and Monocots suggests that all these species share the same meiotic recombination structure with preferential location of recombination in upstream regions of gene, which was hypothesized to be the ancestral mode of recombination location in Eukaryotes [73].

Glémin et al. [33] also proposed that changes in the steepness of the recombination/gBGC gradient could explain variation in GC content distributions among species, from unimodal GC-poor to bimodal GC-rich distributions. Alternatively, if gradients are stable among species, changes in gene structure, especially the number of short mono-exonic genes and the distribution of length of first introns, could also generate variations in GC content distribution [33,37]. Here we found that, in the first part of genes, gBGC is the highest in Commelinid species, which exhibit the richest and most heterogeneous GC content distributions (Fig 7). This result parallels the sharp difference in GC content in first exons between rice and *Arabidopsis* whereas the centres of genes have a very similar base composition [37]. Our results support the hypothesis that genic base composition in GC-rich and heterogeneous genomes has been driven by high gBGC/recombination gradients. As GC content bimodality is likely ancestral to monocot species and has been lost several times later [2], our results suggest that an increase in gBGC and or recombination rates occurred at the origin of the Monocot lineage.

Conclusion

Overall, we show that selection on codon usage only plays a minor role in shaping base composition evolution at synonymous sites in plant genomes and that gBGC is the main driving force. Our study comes along an increasing number of results showing that gBGC is at work in many organisms. Plants are no exception. If, as we suggest, gBGC is the main contributor to base composition variation among plant species, it shifts the question towards understanding why gBGC may vary between species and more generally why gBGC evolved. Our results also imply that gBGC should be taken into account when analysing plant genomes, especially GC-rich ones. Typically, claims of adaptive significance of variation in GC content should be viewed with caution and properly tested against the "extended null hypothesis" of molecular evolution including the possible effect of gBGC [65].

Materials & methods

Dataset

We focused our study of synonymous variations in 11 species spread across the Angiosperm phylogeny with contrasted base composition and mating systems, *Coffea canephora*, *Olea europaea*, *Solanum pimpinellifolium*, *Theobroma cacao*, *Vitis vinifera*, *Dioscorea abyssinica*, *Elaeis guineensis*, *Musa acuminata*, *Pennisetum glaucum*, *Sorghum bicolor* and *Triticum monococcum*. A phylogeny of these species is shown in Fig 1. For practical reasons, we chose diploid cultivated species but focused our analysis on wild populations except in *Elaeis guineensis* where

domestication is very recent and limited (19th century [45]). Using the same methodology as [48], we sequenced for each species the transcriptome of ten individuals (12 in the case of *C. canephora* and *V. vinifera*, nine in the case of *S. bicolor* and five in the case of *D. abyssinica*) plus two individuals coming from two outgroup species, using RNA-seq (see S3 Text for details). After quality cleaning, reads were either mapped on the transcriptome extracted from the reference genome (when available, see Table 1) or on the de novo transcriptome of each species (including outgroups) obtained from [48]. For *C. canephora* and its outgroups, no transcriptome was available. We thus applied the same methodology and pipeline as in [48] to assemble and annotate contigs. For banana, *M. acuminata*, Robusta coffee tree, *C. canephora*, and for the outgroup *Phoenix dactylifera*, genome sequences were available but the quality of mapping was not optimal because of problems of definition of exon/intron boundaries. We thus preferred assembling a new transcriptome from our data using the same protocol. Details of the assemblies for all species are given in S2 Table. Details of data processing are provided in S4 Text. Only contigs with at least one mapped read for each individual was kept for further analysis. Expression levels for each individual in each contig were computed as RPKM values (*i.e.* the number of Reads per Kilobase per Millions mapped reads). We called genotypes and filtered out paralogs for each species individually using the *read2snp* software [47] (see S4 Text for details). Genotypes were called when the coverage was at least 10x and the posterior probability of the genotype higher than 0.95. Otherwise, the genotype of the individual was considered as missing data. Orthology between focal and outgroup individuals was determined by best reciprocal blast hit. Finally, we aligned orthologous contigs (focal and outgroup individuals) sequences using MACSE [74].

SNPs detection and polarization

We scanned contig alignments in each focal species for polymorphic sites. We only considered gapless sites for which all focal individuals were genotyped. Only bi-allelic SNPs were considered. In the highly selfing *T. monococcum*, the deficit in heterozygous sites can lead to abnormal site frequency spectra that are difficult to analyse. We thus used an allele sampling procedure that effectively divides the number of chromosomes by two by merging together homologous chromosomes in each individual. For heterozygous sites, one allele was randomly chosen. For the mixed mating *S. bicolor* and *S. pimpinellifolium*, we used the full SFSs.

SNPs were polarized using parsimony by comparing alleles in focal individuals to orthologous positions in outgroups. For each polymorphic site, the ancestral allele was inferred to be the one identical to both outgroup species, while the other allele was inferred to be derived. All polarized SNPs are marked ancestral → derived for the remainder of the paper. A and T bases were grouped together as W (for weak) while G and C bases were grouped together as S (for strong). We thus classified mutations as W→S, S→W or neutral with respect to gBGC (S←→S or W←→W).

SNPs and preferred codons

In each species, preferred (P) and un-preferred (U) codons were defined using the ΔRSCU method [49]. In each contig, we computed for each codon its RSCU value, or relative frequency (*i.e.* its frequency in a contig normalized by the frequency of its amino-acid in the same contig). Contigs were divided into eight groups of identical size based on their expression levels (RPKM values averaged over all individuals). For each codon, we compared its RSCU in the first (least expressed) and last (most expressed) class using a Mann-Whitney U test. Codons were annotated as preferred (or un-preferred) if their RSCU increased (or decreased) significantly with gene expression levels. All other codons were marked as non-significant. All

synonymous SNPs for which an ancestral allele is unambiguously identified were annotated with regards to codon preference: mutations increasing codon preference (from un-preferred to either non-significant or preferred, or from non-significant to preferred) were annotated U→P while mutations decreasing codon preference (from preferred to either un-preferred or non-significant, or from non-significant to un-preferred) were annotated P→U. Mutations not affecting preference were considered as neutral with respect to SCU.

Substitutions

Using the three species alignments (Focal + two outgroups), we also counted and polarized substitutions specific to the focal species lineage. Divergent sites were determined as sites that were fixed in the focal population and different from both outgroups. Only sites identical in both outgroups were considered. As described above for SNPs, substitutions were classified as W→S, S→W or neutral, and U→P, P→U and neutral.

Modified MK-test, neutrality and direction of selection indices

We performed a modified McDonald-Kreitman (MK) test [51], comparing W→S to S→W polymorphic and divergent sites on one hand (gBGC set) and U→P to P→U polymorphic and divergent sites on the other (SCU set). The underlying theory is detailed in S1 Text. For each category, the total number of synonymous polymorphic and divergent sites was computed following the criteria detailed above. We performed a Chi-squared test for each set. Significant tests indicate that sequences do not evolve only under mutation pressure: selection and/or gBGC must be at work. Furthermore, we computed for each set a neutrality [52] and a direction of selection [53] indices as follows:

$$NI = \frac{P_{WS}/P_{SW}}{D_{WS}/D_{SW}}$$

$$DoS = \frac{D_{WS}}{D_{WS} + D_{SW}} - \frac{P_{WS}}{P_{WS} + P_{SW}}$$

Where P_{WS} and P_{SW} are the number of W→S and S→W SNPs and D_{WS} and D_{SW} the number of W→S and S→W substitutions respectively. Assuming constant mutational bias, NI values lower than 1 or positive DoS values indicate SCU and/or gBGC of similar or stronger intensity at the divergence than at the polymorphism level. Respectively, NI values higher than 1, or negative DoS values indicate stronger selection and/or gBGC at the polymorphism than at the divergence level (see S1 Text).

Because these statistics rely on polarized polymorphisms and substitutions, they are potentially sensitive to polarization errors, which could lead to spurious signature of selection/gBGC [38,44]. Importantly, we showed in S1 Text that the sign of both statistics is insensitive to polarization errors (as far as they are lower than 50%) and that polarization errors decrease the magnitude of the statistics, which makes our tests conservative to polarization errors.

Estimation of gBGC and SCU

To estimate gBGC and SCU we extended the method of Glémin et al. [38] as detailed in S2 Text. The rationale of the approach is to fit population genetic models to the three derived SFS including fixed mutations (W→S or U→P, S→W or P→U, and neutral). Parameters estimated are ancestral (B_0 or S_0) and recent (B_1 or S_1) gBGC or selection, mutational bias (λ), as well as other parameters (see S2 Text for details). We ran a series of nested models where B_0

and B_1 (or S_0 and S_1) are either fixed to zero or freely estimated, plus one model where they are set to be equal. Models were compared by the appropriate likelihood ratio tests (LRT). To jointly estimate gBGC and selection, we also extended the model by fitting nine SFS corresponding to the combination of the three basic SFS (e.g. W→S and P→U see S2.1 Table in S2 Text for the complete list). We tested all combinations of models where each parameter can be either null or freely estimated, so from the null neutral model, $B_0 = B_1 = S_0 = S_1 = 0$, to the model with the four parameters being freely estimated. As all models are not nested, we then chose the best model using the Akaike Information Criterion (AIC). When AICs were very close we chose the model with the lowest number of free parameters.

Supporting information

S1 Text. Neutrality and direction of selection indices under gBGC or SCU.
(PDF)

S2 Text. Estimation of gBGC and selection intensities—Extension of Glémin et al. (2015).
(PDF)

S3 Text. Data preparation.
(PDF)

S4 Text. Data processing.
(PDF)

S1 Table. List of sampled species and individuals.
(XLSX)

S2 Table. Summary of assemblies' characteristics.
(XLSX)

S3 Table. Codon preferences for the eleven species.
(XLSX)

S4 Table. Detailed results of gBGC and SCU estimates.
(XLSX)

S5 Table. Results of all gBGC/SCU nested models.
(XLSX)

S6 Table. Results of all models in the first part and rest of genes.
(XLSX)

S1 Fig. Distribution of GC3 content in the transcriptome of the 11 species.
(PDF)

S2 Fig. RSCU (Relative synonymous codon usage) in the 11 species. Codons are grouped by amino acids. Codons ending with A or T are in blue, those ending with G or C in red. Blue colour corresponds to the most frequent codons and yellow to the least frequent.
(PDF)

S3 Fig. SFSs in the eleven species. Site-frequency spectra for synonymous gBGC SNPs, *i.e.* W→S, S→W or S→S and W→W SNPs grouped together as "neutral."
(PDF)

S4 Fig. GC3 and gBGC gradients along genes starting with a start codon. In the first exon, B is significantly higher in Commelinid than in other species (Wilcoxon test p-value = 0.0043).

B values and GC3 are significantly and positively correlated both on the first part of contigs ($\rho_{Spearman}$ = 0.80, p-value = 0.0052) and in the rest of contigs ($\rho_{Spearman}$ = 0.70, p-value = 0.0208).
(PDF)

S5 Fig. Phylogenetic relationship between species used in this study. Top-left panel: phylogeny of the species used in the study. The phylogeny was computed with PhyML [75] on a set of 33 1–1 orthologous protein clusters obtained with SiLiX [76]. Top-right and bottom-left panels: dN and dS values between species used in this study. We used the branch model of codeml [77] to infer dN and dS values independently in each branch of the phylogeny. We used the topology inferred from PhyML.
(PDF)

S6 Fig. Phylogeny with detailed branch lengths. Phylogeny of the species used in this study (see S5 Fig for Method) with detailed branch lengths for each individual branches. Only the branch between *D*. *abyssinica* and the other monocot species shows a bootstrap support lower than 0.98 (namely 0.71).
(PDF)

S1 File. This contains: 1) the mathematica script used to jointly estimate gBGC and SCU from SFS and divergence data 2) the R script used to simulate SFS under various demographic scenarios 3) Processed site frequency spectra used in this analysis.
(ZIP)

Acknowledgments

We thank Nicolas Galtier for numerous discussions and for sharing scripts during the course of the project, Aurélien Bernard for help with bioinformatics, Carina Mugal and Laurent Duret for helpful comments on the manuscript. We thank the following colleagues and institutions for providing plant material: Dr Hyacinthe Legnate and the CNRA (Ivory Coast), Pr Adrien Kalondji and the University of Kinshasa (DRC), Bernard Perthuis (CIRAD French-Guiana) for Coffee-tree, Michel Boccara and the Cocoa Research Center (Trinidad) for Cocoa, Pierre-Oliver Cheptou, the staff of the experimental field of the Plateforme des Terrains d'Experience du LabEx CeMEB (Montpellier, France) for *Phillyrea* and *Olea* species, the Domaine de Vassal grapevine seed bank (INRA, Marseillan-Plage, France) for *Vitis* accessions, Frédérique Aberlenc for palm tree (Montpellier, France), CRB Plantes Tropicales Guadeloupe and Collection Musacées CARBAP Cameroun for banana. This is publication ISEM 2017–091.

Author Contributions

Conceptualization: SG JD.

Data curation: GS YH FH SP SC BN RB CJ FS MR.

Formal analysis: YC SG.

Funding acquisition: JLP JD SG.

Investigation: YC SG.

Methodology: SG.

Project administration: JLP JPL.

Resources: RB AB GB CC JD FDB OF BK CL TL DP CS NS JT YV NY LS MA SS.

Supervision: SG JD.

Visualization: YC SG.

Writing – original draft: YC SG.

Writing – review & editing: YC SG.

References

1. Serres-Giardi L, Belkhir K, David J, Glémin S (2012) Patterns and evolution of nucleotide landscapes in seed plants. The Plant Cell 24: 1379–1397. https://doi.org/10.1105/tpc.111.093674 PMID: 22492812
2. Clement Y, Fustier MA, Nabholz B, Glémin S (2015) The bimodal distribution of genic GC content is ancestral to monocot species. Genome Biology and Evolution 7: 336–348.
3. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nature Review Genetics 12: 32–42.
4. Wright SI, Yau CB, Looseley M, Meyers BC (2004) Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. Molecular Biology & Evolution 21: 1719–1726.
5. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2004) Codon usage between genomes is constrained by genome-wide mutational processes. Proceeding of the National Academy of Science USA 101: 3480–3485.
6. Lynch M (2007) The origin of genome architecture. Sunderland: Sinauer.
7. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129: 897–907. PMID: 1752426
8. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Review Genetics 7: 98–108.
9. Cusack BP, Arndt PF, Duret L, Roest Crollius H (2011) Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. PLoS Genetics 7: e1002276. https://doi.org/10.1371/journal.pgen.1002276 PMID: 22022272
10. Subramanian S (2008) Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics 178: 2429–2432. https://doi.org/10.1534/genetics.107.086405 PMID: 18430960
11. Eyre-Walker A, Hurst LD (2001) The evolution of isochores. Nature Review Genetics 2: 549–555.
12. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annual Review of Genomics and Human Genetics 10: 285–311. https://doi.org/10.1146/annurev-genom-082908-150001 PMID: 19630562
13. Marais G (2003) Biased gene conversion: implications for genome and sex evolution. Trends in Genetics 19: 330–338. https://doi.org/10.1016/S0168-9525(03)00116-1 PMID: 12801726
14. Nagylaki T (1983) Evolution of a finite population under gene conversion. Proceedings of the National Academy of Sciences of the USA 80: 6278–6281. PMID: 6578508
15. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454: 479–485. https://doi.org/10.1038/nature07135 PMID: 18615017
16. Lesecque Y, Mouchiroud D, Duret L (2013) GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Molecular Biology and Evolution 30: 1409–1419. https://doi.org/10.1093/molbev/mst056 PMID: 23505044
17. Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I (2015) Crossovers are associated with mutation and biased gene conversion at recombination hotspots. Proceeding of the National Academy of Science USA 112: 2109–2114.
18. Williams AL, Genovese G, Dyer T, Altemose N, Truax K, et al. (2015) Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. eLife 4.
19. Smeds L, Mugal CF, Qvarnstrom A, Ellegren H (2016) High-Resolution Mapping of Crossover and Non-crossover Recombination Events by Whole-Genome Re-sequencing of an Avian Pedigree. PLoS Genetics 12: e1006044. https://doi.org/10.1371/journal.pgen.1006044 PMID: 27219623
20. Si W, Yuan Y, Huang J, Zhang X, Zhang Y, et al. (2015) Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. The New Phytologist 206: 1491–1502. https://doi.org/10.1111/nph.13319 PMID: 25664766

21. Escobar JS, Glémin S, Galtier N (2011) GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms, and Other Eukaryotes. Molecular Biology and Evolution 28: 2561–2575. https://doi.org/10.1093/molbev/msr079 PMID: 21444650

22. Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, et al. (2012) Evidence for widespread GC-biased gene conversion in eukaryotes. Genome Biology and Evolution 4: 675–682. https://doi.org/10.1093/gbe/evs052 PMID: 22628461

23. Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, et al. (2015) GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genetics 11: e1004941. https://doi.org/10.1371/journal.pgen.1004941 PMID: 25659072

24. Robinson MC, Stone EA, Singh ND (2014) Population genomic analysis reveals no evidence for GC-biased gene conversion in Drosophila melanogaster. Molecular Biology & Evolution 31: 425–433.

25. Zeng K, Charlesworth B (2010) Studying patterns of recent evolution at synonymous sites and intronic sites in Drosophila melanogaster. Journal of Molecular Evolution 70: 116–128. https://doi.org/10.1007/s00239-009-9314-6 PMID: 20041239

26. Zeng K, Charlesworth B (2009) Estimating selection intensity on synonymous codon usage in a non-equilibrium population. Genetics 183: 651–662, 651SI–623SI. https://doi.org/10.1534/genetics.109.101782 PMID: 19620398

27. Zeng K (2010) A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. Molecular Biology & Evolution 27: 1327–1337.

28. Jackson BC, Campos JL, Haddrill PR, Charlesworth B, Zeng K (2017) Variation in the intensity of selection on codon bias over time causes contrasting patterns of base composition evolution in Drosophila. Genome Biol Evol.

29. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D (2011) Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. Genome Biol Evol 3: 868–880. https://doi.org/10.1093/gbe/evr085 PMID: 21856647

30. Qiu S, Bergero R, Zeng K, Charlesworth D (2011) Patterns of codon usage bias in Silene latifolia. Molecular Biology & Evolution 28: 771–780.

31. Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S (2011) GC-biased gene conversion and selection affect GC content in the Oryza genus (rice). Molecular Biology and Evolution 28: 2695–2706. https://doi.org/10.1093/molbev/msr104 PMID: 21504892

32. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES (2016) Open chromatin reveals the functional maize genome. Proceeding of the National Academy of Science USA.

33. Glémin S, Clement Y, David J, Ressayre A (2014) GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends in Genetics 30: 263–270. https://doi.org/10.1016/j.tig.2014.05.002 PMID: 24916172

34. Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, et al. (2013) Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nature Genetics 45: 1327–1336. https://doi.org/10.1038/ng.2766 PMID: 24056716

35. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, et al. (2013) Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proceedings of the National Academy of Sciences of the USA.

36. Li X, Li L, Yan J (2015) Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. Nature Communications 6: 6648. https://doi.org/10.1038/ncomms7648 PMID: 25800954

37. Ressayre A, Glémin S, Montalent P, Serre-Giardi L, Dillmann C, et al. (2015) Introns Structure Patterns of Variation in Nucleotide Composition in *Arabidopsis thaliana* and Rice Protein-Coding Genes. Genome Biol Evol 7: 2913–2928. https://doi.org/10.1093/gbe/evv189 PMID: 26450849

38. Glémin S, Arndt PF, Messer PW, Petrov D, Galtier N, et al. (2015) Quantification of GC-biased gene conversion in the human genome. Genome Research 25: 1215–1228. https://doi.org/10.1101/gr.185488.114 PMID: 25995268

39. Lartillot N (2013) Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. Molecular Biology & Evolution 30: 489–502.

40. Clement Y, Arndt PF (2013) Meiotic Recombination Strongly Influences GC-Content Evolution in Short Regions in the Mouse Genome. Molecular Biology and Evolution.

41. Wallberg A, Glémin S, Webster MT (2015) Extreme recombination frequencies shape genome variation and evolution in the honeybee, Apis mellifera. PLoS Genetics 11: e1005189. https://doi.org/10.1371/journal.pgen.1005189 PMID: 25902173

42. Smith NG, Eyre-Walker A (2001) Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Molecular Biology & Evolution 18: 982–986.

43. Romiguier J, Ranwez V, Douzery EJ, Galtier N (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Research 20: 1001–1009. https://doi.org/10.1101/gr.104372.109 PMID: 20530252

44. Hernandez RD, Williamson SH, Zhu L, Bustamante CD (2007) Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Molecular Biology & Evolution 24: 2196–2202.

45. Zeven AC (1972) The Partial and Complete Domestication of the Oil Palm (*Elaeis guineensis*). Economic Botany 26: 274–279.

46. Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, et al. (2014) Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature.

47. Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, et al. (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genetics 9: e1003457. https://doi.org/10.1371/journal.pgen.1003457 PMID: 23593039

48. Sarah G, Homa F, Pointet S, Contreras S, Sabot F, et al. (2016) A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. Molecular Ecology Resources.

49. Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, *Arabidopsis*. Proceedings of the National Academy of Sciences of the USA 96: 4482–4487. PMID: 10200288

50. Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. Journal of Molecular Biology 151: 389–409. PMID: 6175758

51. McDonald JH, Kreitman M (1991) Adaptative protein evolution at the ADH locus in Drosophila. Nature 351: 652–654. https://doi.org/10.1038/351652a0 PMID: 1904993

52. Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Molecular Biology & Evolution 13: 735–748.

53. Stoletzki N, Eyre-Walker A (2011) Estimation of the neutrality index. Molecular Biology & Evolution 28: 63–70.

54. Igic B, Busch JW (2013) Is self-fertilization an evolutionary dead end? New Phytologist 198: 386–397. https://doi.org/10.1111/nph.12182 PMID: 23421594

55. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475. https://doi.org/10.1038/nature11396 PMID: 22914163

56. Tello-Ruiz MK, Stein J, Wei S, Preece J, Olson A, et al. (2016) Gramene 2016: comparative plant genomics and pathway resources. Nucleic Acids Research 44: D1133–1140. https://doi.org/10.1093/nar/gkv1179 PMID: 26553803

57. Dillon MM, Sung W, Lynch M, Cooper VS (2015) The Rate and Molecular Spectrum of Spontaneous Mutations in the GC-Rich Multichromosome Genome of Burkholderia cenocepacia. Genetics 200: 935–946. https://doi.org/10.1534/genetics.115.176834 PMID: 25971664

58. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genetics 6.

59. Marais G, Charlesworth B, Wright SI (2004) Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biology 5: R45. https://doi.org/10.1186/gb-2004-5-7-r45 PMID: 15239830

60. Mugal CF, Weber CC, Ellegren H (2015) GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. Bioessays 37: 1317–1326. https://doi.org/10.1002/bies.201500058 PMID: 26445215

61. Lesecque Y, Glémin S, Lartillot N, Mouchiroud D, Duret L (2014) The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. PLoS Genet 10: e1004790. https://doi.org/10.1371/journal.pgen.1004790 PMID: 25393762

62. Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, et al. (2015) Stable recombination hotspots in birds. Science 350: 928–932. https://doi.org/10.1126/science.aad0843 PMID: 26586757

63. Tsai IJ, Burt A, Koufopanou V (2010) Conservation of recombination hotspots in yeast. Proceeding of the National Academy of Science USA 107: 7847–7852.

64. Rodgers-Melnick E, Bradbury PJ, Elshire RJ, Glaubitz JC, Acharya CB, et al. (2015) Recombination in diverse maize is stable, predictable, and associated with genetic load. Proceeding of the National Academy of Science USA 112: 3823–3828.

65. Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in Genetics 23: 273–277. https://doi.org/10.1016/j.tig.2007.03.011 PMID: 17418442

66. Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, et al. (2010) Detecting positive selection within genomes: the problem of biased gene conversion. Philosophical Transactions of the Royal Society B-Biological Sciences 365: 2571–2580.

67. Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H (2016) Codon Usage Selection Can Bias Estimation of the Fraction of Adaptive Amino Acid Fixations. Molecular Biology & Evolution.

68. Glémin S (2010) Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics 185: 939–959. https://doi.org/10.1534/genetics.110.116368 PMID: 20421602

69. Hershberg R, Petrov DA (2009) General rules for optimal codon choice. PLoS Genetics 5: e1000556. https://doi.org/10.1371/journal.pgen.1000556 PMID: 19593368

70. Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. Nature 325: 728–730. https://doi.org/10.1038/325728a0 PMID: 2434856

71. Higgs PG, Ran W (2008) Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Molecular Biology & Evolution 25: 2279–2291.

72. Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES (2016) Open Chromatin Reveals the Functional Maize Genome. Proceeding of the National Academy of Science USA in press.

73. Choi K, Henderson IR (2015) Meiotic recombination hotspots—a comparative view. The Plant Journal 83: 52–61. https://doi.org/10.1111/tpj.12870 PMID: 25925869

74. Ranwez V, Harispe S, Delsuc F, Douzery EJ (2011) MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS ONE 6: e22594. https://doi.org/10.1371/journal.pone.0022594 PMID: 21949676

75. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59: 307–321. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638

76. Miele V, Penel S, Duret L (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12: 116. https://doi.org/10.1186/1471-2105-12-116 PMID: 21513511

77. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591. https://doi.org/10.1093/molbev/msm088 PMID: 17483113