RESEARCH ARTICLE

Genome-wide analysis of codon usage bias in four sequenced cotton species

Liyuan Wang[1]☯, Huixian Xing[1]☯, Yanchao Yuan[1], Xianlin Wang[1], Muhammad Saeed[2], Jincai Tao[1], Wei Feng[1], Guihua Zhang[3], Xianliang Song[1]*, Xuezhen Sun[1]*

1 State Key Laboratory of Crop Biology/Agronomy College, Shandong Agricultural University, Taian, China, **2** Department of Botany, Government College University, Faisalabad, Pakistan, **3** Heze Academy of Agricultural Sciences, Heze, China

☯ These authors contributed equally to this work.
* songxl999@163.com (XSO); sunxz@sdau.edu.cn (XSU)

Abstract

Codon usage bias (CUB) is an important evolutionary feature in a genome which provides important information for studying organism evolution, gene function and exogenous gene expression. The CUB and its shaping factors in the nuclear genomes of four sequenced cotton species, *G. arboreum* (A_2), *G. raimondii* (D_5), *G. hirsutum* (AD_1) and *G. barbadense* (AD_2) were analyzed in the present study. The effective number of codons (ENC) analysis showed the CUB was weak in these four species and the four subgenomes of the two tetraploids. Codon composition analysis revealed these four species preferred to use pyrimidine-rich codons more frequently than purine-rich codons. Correlation analysis indicated that the base content at the third position of codons affect the degree of codon preference. PR2-bias plot and ENC-plot analyses revealed that the CUB patterns in these genomes and subgenomes were influenced by combined effects of translational selection, directional mutation and other factors. The translational selection (P2) analysis results, together with the non-significant correlation between GC12 and GC3, further revealed that translational selection played the dominant role over mutation pressure in the codon usage bias. Through relative synonymous codon usage (RSCU) analysis, we detected 25 high frequency codons preferred to end with T or A, and 31 low frequency codons inclined to end with C or G in these four species and four subgenomes. Finally, 19 to 26 optimal codons with 19 common ones were determined for each species and subgenomes, which preferred to end with A or T. We concluded that the codon usage bias was weak and the translation selection was the main shaping factor in nuclear genes of these four cotton genomes and four subgenomes.

OPEN ACCESS

Citation: Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, et al. (2018) Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS ONE 13(3): e0194372. https://doi.org/10.1371/journal.pone.0194372

Editor: David D Fang, USDA-ARS Southern Regional Research Center, UNITED STATES

Received: November 28, 2017

Accepted: March 1, 2018

Published: March 27, 2018

Data Availability Statement: The third-party data underlying this study were obtained from the CottonGen database. These data sets are publicly available through the following links: G. raimondii (ftp://ftp.bioinfo.wsu.edu/species/Gossypium_raimondii/JGI_221_G.raimondii_Dgenome/assembly/G.raimondii_JGI_221_v2.0.psuedomolecules.fasta.gz); G. arboreum (ftp://ftp.bioinfo.wsu.edu/species/Gossypium_arboreum/CGP-BGI_G.arboreum_Agenome/assembly/G.arboreum_BGI-A2_chromosomes_only.fas.gz); G. hirsutum (ftp://ftp.bioinfo.wsu.edu/species/Gossypium_hirsutum/NAU-NBI_G.hirsutum_

Introduction

Genetic information is transmitted from DNA to mRNA, then from mRNA to protein. In the latter process, information is transmitted in the form of codons. Codon is an important link in the output of nucleic acid information. Genetic code has degenerate feature that a single amino acid, except methionine (Met) and tryptophan (Trp), is encoded by more than one

AD1genome/assembly/AD1-NBI_v1.1_chromosomes_only.fas.gz); G. barbadense (ftp://ftp.bioinfo.wsu.edu/species/Gossypium_barbadense/HAU-NBI_v1.0_AD2genome/assembly/AD2-NBI_v1.0_chromosomes_only.fas.gz).

Funding: This work was supported by the Major Projects for Transgenic Breeding of China (2017ZX08005-004-006); The System of Modern Agriculture Industrial Technology (SDAIT-03-03/05) and the Natural Science foundation (ZR2017MC057). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

codon known as synonymous codons. The synonymous codons are not used at equal frequencies in coding sequences in many organisms [1]. This phenomenon called 'synonymous codon usage bias (SCUB)' reflects non-uniform usage of synonymous codons encoding the same amino acid during the translation of genes to proteins [2,3].

The degree of SCUB divergence varies greatly among different species and genes [4–7]. SCUB was affected by many factors, including directional mutation, neutral selection [8], GC content, synonymous substitution rate [9], tRNA abundance [10], selection for efficient translation initiation [11], codon hydropathy and DNA replication initiation site [12], gene length [13] and expression level [14], etc. Of these factors, directional mutation and neutral selection are the two main ones, with varying relative importance in different species [15–17]. The codon usage patterns always represent balance between the directional mutation and the neutral selection that leads to translational efficiency of genes [8–11]. SCUB patterns were also associated with phylogenetic relationship among given species. Distant phylogenetic species usually has greater variations in codon usage bias [18,19].

Information on the SCUB patterns can provide significant insights pertaining to the prediction, classification, and molecular evolution of genes, design of highly expressed genes and cloning vectors and reveal about the host-pathogen coevolution and adaptation of pathogens to specific hosts [20,21]. Clustering results based on relative synonymous codon usage (RSCU) values could provide useful reference for phylogenetic relationship analysis [18,19,22].

Cotton (*Gossypium* spp.) is the main source of renewable textile fibers and is also grown to produce vegetable oil and high protein meals for humans and livestock [23]. The genus *Gossypium* includes around 46 diploid (2n = 2x = 26) and 6 tetraploid (2n = 4x = 52) and 1 purported species [24–26], including four commercial ones, *G. arboreum* (A_2), *G. herbaceum* (A_1), *G. hirsutum* (AD_1) and *G. barbadense* (AD_2). It has been proposed that all diploid cotton species may have evolved from a common ancestor and allopolyploid cotton may have appeared through hybridization and subsequent polyploidization events between the A- and D-subgenome progenitors. The D-genome species *G. raimondii* (D_5) and the A-genome species much like modern *G. arboreum* (A_2) and *G. herbaceum* (A_2) are the donor species for the D and A chromosome groups of the tetraploid cotton species, respectively [27–28]. Recently, the nuclear genome sequences of *G. arboreum*, *G. raimondii*, *G. hirsutum* and *G. barbadense*, and chloroplast and mitochondrial genome sequences of *G. hirsutum* have successively been released [29–38], which has advanced the understanding of cotton genomics and genetics, and made it possible to investigate SCUB patterns in cotton nuclear and organelle genomes. But until now, SCUB analysis has been performed only in chloroplast genome of *G. hirsutum* [36]. The purpose of this study was to analyze the SCUB patterns of the four sequenced cotton species and explore the key factors influencing codon choice.

Materials and methods

Coding sequence data

All the CDS sequences of *G. raimondii* [30], *G. arboreum* [31], *G. hirsutum* [33] and *G. barbadense* [34] were downloaded from the CottonGen database (https://www.cottongen.org/). We got the Excel files containing all genes of *G. hirsutum* and *G. barbadense* from the Cotton Research Institute (CRI) of Nanjing Agricultural University (http://mascotton.njau.edu.cn), and then separated the two allotetraploids into *At* and *Dt* subgenomes using Seqkit [39] (https://github.com/shenwei356/seqkit), named At_1 and Dt_1 in *G. hirsutum*, and At_2 and Dt_2 in *G. barbadense* respectively. Detailed information of CDSs and codons used in this work was listed in Table 1.

Table 1. The number of CDSs and codons of 4 cotton species and 4 subgenomes used in this study.

Species or subgenomes	Genome	Number of CDSs	Number of Codons
G. arboreum	A_2	40134	14538888
G. raimondii	D_5	77267	32995450
G. hirsutum	$(AD)_1$	66434	27096230
G. barbadense	$(AD)_2$	77358	29259373
At_1	$(AD)_1$	32032	13188672
Dt_1	$(AD)_1$	34402	13907558
At_2	$(AD)_2$	39568	14786710
Dt_2	$(AD)_2$	37790	14472663

https://doi.org/10.1371/journal.pone.0194372.t001

Statistical analyses

Codon usage bias indices. The nuclear genome CDSs of each cotton species were firstly analyzed as a whole to clarify the codon usage features. Methionine (Met) and tryptophan (Trp), each having a single codon, were excluded from further analysis. Stop codons (UAG, UAA, and UGA) were also excluded from the analysis because each stop codon can only occur once in a single CDS sequence.

Using CodonW 1.4.2 software (http://codonw.sourceforge.net/), a number of indices of codon usage bias including the relative synonymous codon usage (RSCU), effective number of codons (ENC), and the frequency of the nucleotides G+C at the third position (GC3s) were calculated. Several codon composition indices including GC contents of the entire gene (GC), the content or frequency of each individual base A, T, G, and C at the third position of codons (A3s, T3s, G3s, C3s) were also counted. The G+C content at the first, second positions of codons (GC1, GC2) and the average GC content of the first and second positions (GC12) were determined by the online Cusp program from Galaxy (https://usegalaxy.org/). The correlations between nucleotide contents were calculated with the statistical software SPSS V21.0 (http://www.spss.com.cn/).

Multiple comparisons. Multiple comparisons can be used to infer any significant differences between the effects of the factors. Then the T3s, G3s, GC and ENC of each gene were calculated by CodonW 1.4.2. The average value of all the genes was expressed in T3s(av), G3s(av), GC(av) and ENC(av), and multiple comparisons across the cotton species and subgenomes were made by SPSS V21.0. The box and whisker plots were drawn through OmicShare (www.omicshare.com/tools).

PR2-bias plot. Parity Rule 2 (PR2) is a rule of DNA composition. When there is no deviation between mutation and selection pressure of two DNA chains, the fractional content of the four bases follows A = T and G = C (where A + T + G + C = 1) [40]. PR2-bias plots are particularly informative when PR2 biases at the third codon position are plotted. The center of the plot, where both coordinates are 0.5, is the place where A = T and G = C (PR2). The degree of deviation from PR2 allows us to estimate the chain bias affected by mutation, selection, or both [41]. If genes are evenly distributed across the plan view, that is, if the codon usage frequency of A + T is the same as that of G + C at the third position, then the codon usage preference is likely to be entirely caused by mutation [40]. The A3s/(A3s+T3s) and G3s/(G3s+C3s) of each gene were calculated and used as the ordinate and the abscissa to show the relationship between the two base contents of genes, namely purine (A and G) and pyrimidine (T and C) at the third codon position. The PR2-plots were drawn by Matlab R2016a (https://www.mathworks.com/).

ENC-plot (ENC versus GC3s). The effective codon number (ENC) determines the degree of preference for the unbalanced use of codons. ENC value ranges from 20 (only one codon is used for each amino acid) to 61 (when all synonymous codons are used for each amino acid) and is negatively correlated with codon usage bias. The codon usage pattern across genes was examined by the ENC-plot drawn by Matlab R2016a, which is a plot of ENC versus GC3s. ENC-plot is one of the most widely used measures for judging whether or not organism codon usage is biased through exploring the use of codon bias, and detecting the effect of base content on CUB [36]. The expected ENC values from GC3s (denoted by '*S*') were calculated according to the equation given by Wright [42] and Novembre [43]:

$$\mathrm{ENC}_{\mathrm{expected}} = 2 + S + \frac{29}{S^2 + (1-S)^2}$$

The genes would be distributed along the standard curve or near the standard curve when codon bias is only affected by mutation, while they would fall below the standard curve if codon bias is influenced by selection and other factors.

Translational selection (P2). The translation option (P2) measures the efficiency of codon-anticodon interactions and provides an indication of translation efficiency as long as information for preferred codon sets is not available. The results of PR2-bias plot can reflect the relationship between purine (A and G) and pyrimidine (T and C) in codon composition. P2 value > 0.5 shows preference for translational selection.

P2 was calculated according to the following equation, where W = A or U, S = C or G, and Y = C or U [2,44–46]:

$$\mathrm{P2} = \frac{\mathrm{WWC} + \mathrm{SSU}}{\mathrm{WWY} + \mathrm{SSY}}$$

Determination of putative optimal codons. Optimal codon is the preferred codon, which is determined by calculation and sequencing of the ENC values of all genes. Generally speaking, highly expressed genes have a large degree of codon preference and therefore a small ENC value. Low expression genes contain more rare codons and have a larger ENC value. Therefore, the relative level of gene expression is currently generally determined by comparing ENCs. The smaller the ENC value is, the higher the corresponding gene is often expressed. 5% of the sequence data were taken from the upper and lower limit regions of the ordered data set, to establish two high- and low-bias gene datasets. To define optimal codons, we used a T-test to examine the significance of codon usage difference between the two datasets [47]. The RSCU values of the codons from the two databases were compared. If the difference (ΔRSCU) is equal to or greater than 0.08, and codons with a frequency of usage that was significantly higher ($P < 0.01$) in high-bias genes than that in genes with low bias were defined as the optimal codons [22,48]. SPSS V21.0 was implemented for statistical analysis.

RSCU-based cluster analysis. The RSCU value is the ratio between the actual observed values of the codon and the theoretical expectations. It reflects the relative usage preference for the specific composition of codons encoding the same amino acid [49]. If RSCU = 1, codon usage is unbiased; if RSCU > 1, specific codon frequency is higher than other synonymous codons, otherwise, the frequency is low [49].

In the cluster analysis, 4 cotton species and 4 *At*- and *Dt*- subgenomes were clustered according to their RSCU values using the Hierarchical Cluster Analysis (HCA) tool from OmicShare. In the clustering process, each cotton species was used as an object, and the relative use of codon was taken as variable.

Table 2. The composition parameters values of codon usage in 4 cotton species and 4 subgenomes.

Species and subgenomes	T3s	A3s	G3s	C3s	GC3s	GC	ENC
G. arboreum	0.425	0.341	0.262	0.230	0.381	0.437	54.08
G. raimondii	0.437	0.344	0.259	0.220	0.370	0.434	53.39
G. hirsutum	0.425	0.341	0.262	0.232	0.382	0.437	54.11
G. barbadense	0.423	0.342	0.263	0.232	0.383	0.437	54.26
At_1	0.425	0.340	0.262	0.231	0.382	0.438	54.11
Dt_1	0.425	0.341	0.262	0.232	0.382	0.437	54.10
At_2	0.423	0.342	0.264	0.233	0.384	0.437	54.41
Dt_2	0.424	0.342	0.263	0.230	0.382	0.436	54.11

https://doi.org/10.1371/journal.pone.0194372.t002

Results and discussion

Codon base composition and multiple comparisons

Firstly, several codon usage parameters were calculated for each cotton species and subgenome taking all their CDSs as a whole and shown in Table 2 and Fig 1. From Table 2 and Fig 1, we can see little difference among all genomes and subgeomes studied except for *G. raimondii*, suggesting there are similar codon base compositions at genome and subgenome level. Briefly, the base composition at the third codon position conforms to T > A > G > C. Both the GC3s and GC content were less than 0.5, illustrating that these four cotton species tend to use pyrimidine-rich codons more frequently than purine-rich codons. In *G. raimondii*, the T3s was even

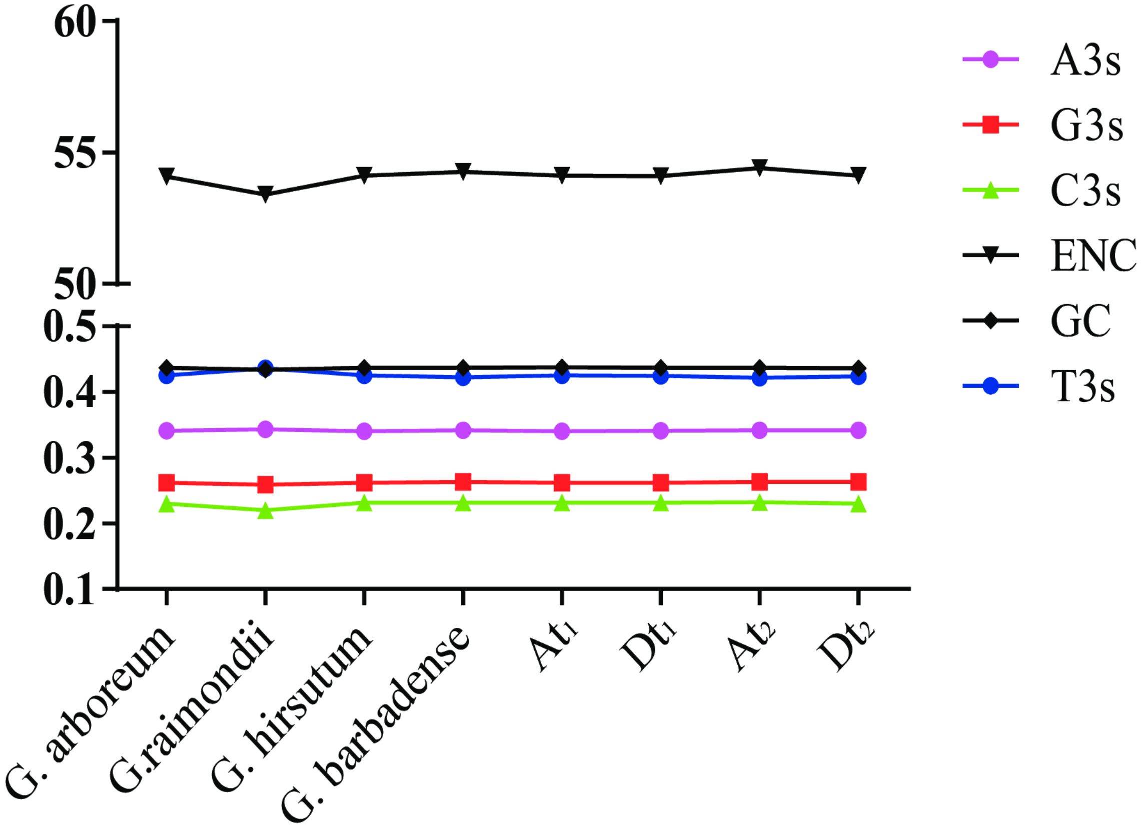

Fig 1. The composition parameters values of codon usage in 4 cotton species and 4 subgenomes.

https://doi.org/10.1371/journal.pone.0194372.g001

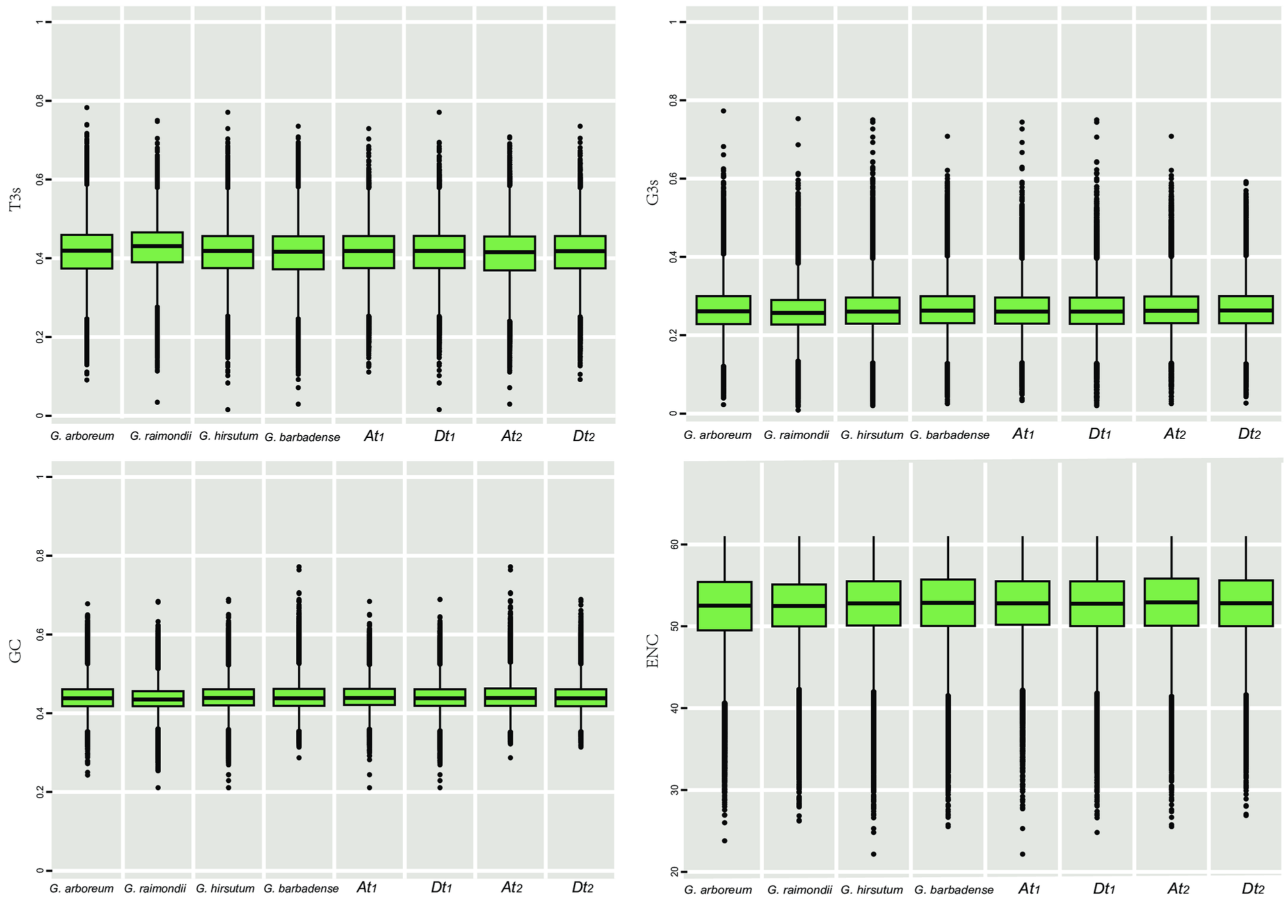

Fig 2. The distribution of T3s, GC3s, GC and ENC of genes in 4 cotton species and 4 subgenomes.

https://doi.org/10.1371/journal.pone.0194372.g002

greater than GC content. The average GC contents are higher than $GC3_S$. Such codon base composition preference was also previously reported in upland cotton chloroplast genes whose GC3s (27.38%) and GC (37.89%) were even lower [36]. Totally, these results were consistent with the lower nuclear GC contents in such cotton species [50–53].

The effective codon number (ENC) values revealed the degree of CUB. ENC is negatively correlated with CUB. According to previous studies [19,42,54], ENC values less than 35 mean high codon preference and ENC values more than 50 reveal general random codon usage. Herein, the average ENC values ranged from 53.39 in *G. raimondii* to 54.41 in At_2 subgenome (Table 2). The distribution of ENC values of the total genes in each genome or subgenome (Fig 2) revealed that less than 0.5%, and more than 70% of genes had low (< 35) and high (> 50) ENC values respectively in all genomes and subgenomes, indicating weak codon usage bias.

In order to further explore their differences in codon base compositions among these genomes and subgenomes, several indices were selected and calculated for each gene, and distribution analysis (Fig 2) and multiple comparisons (Table 3) were performed. In Fig 2, we got roughly similar distribution patterns of the four SCUB indices among the genomes and subgenomes. A majority of genes distributed relatively concentrated, close to the mean value, and there were some genes with extreme values distributed far away from the mean values of T3s, GC, GC3s, and ENC in two directions.

Table 3. Values of T3s(av), G3s(av), GC(av) and ENC(av) of genes in 4 species and 4 subgenomes and their multiple comparisons.

Genomes and subgenomes	T3s$_{(av)}$ * $\bar{x} \pm SD$	G3s$_{(av)}$ * $\bar{x} \pm SD$	GC$_{(av)}$ * $\bar{x} \pm SD$	ENC$_{(av)}$ * $\bar{x} \pm SD$
G. arboretum	0.4147±0.0677^{b}	0.2670±0.0624^{b}	0.4413±0.0378^{c}	52.06±0.0267^{e}
G. raimondii	0.4252±0.0594^{a}	0.2607±0.0540^{d}	0.4378±0.0334^{d}	52.35±0.0160^{d}
G. hirsutum	0.4133±0.0632^{c}	0.2655±0.0592^{c}	0.4422±0.0358^{bc}	52.47±0.0190^{c}
G. barbadense	0.4116±0.0651^{d}	0.2683±0.0580^{a}	0.4427±0.0368^{b}	52.64±0.0171^{ab}
At$_{1}$	0.4130±0.0627^{c}	0.2660±0.0584^{bc}	0.4432±0.0349^{ab}	52.55±0.0263^{bc}
Dt$_{1}$	0.4136±0.0636^{bc}	0.2650±0.0599^{c}	0.4412±0.0366^{c}	52.40±0.0272^{cd}
At$_{2}$	0.4098±0.0663^{e}	0.2682±0.0579^{ab}	0.4435±0.0378^{a}	52.69±0.0241^{a}
Dt$_{2}$	0.4134±0.0638^{bc}	0.2684±0.0582^{a}	0.4417±0.0358^{c}	52.58±0.0242^{b}

*: The "(av)" represents the average of all genes. The multiple comparisons were performed by Duncan's Multiple Range Method.

The various lowercase letters following the data in the same column indicate significant differences at 0.05 level.

https://doi.org/10.1371/journal.pone.0194372.t003

However, many significant differences were detected by multiple comparisons (Table 3). The four species genomes showed significant differences in T3s$_{(av)}$, G3s$_{(av)}$ and ENC$_{(av)}$. As for GC$_{(av)}$, significant differences were found between *G. raimondii* and other three species, between *G. arboreum* and *G. barbadense*, while no significant difference was detected between the two tetraploids, and between *G. aboreum* and *G. hirsutum*.

Comparisons were also performed between *At* or *Dt* subgenomes and their putative progenitor genome respectively. For T3s$_{(av)}$, significant differences were detected among the two *At* subgenomes and *G. arboreum* (A_2), but not among the two *Dt* subgenomes and *G. raimondii* (D_5). With respect to G3s$_{(av)}$, significant differences were detected among the two *Dt* subgenomes and *G. raimondii* (D_5), but not among the two *At* subgenomes and *G. arboreum* (A_2). As for gene average GC content GC$_{(av)}$, the putative donor genome *of G. arboreum* (A_2) had significant lower value than *At*$_{1}$ and *At*$_{2}$ subgenomes, while *G. raimondii* (D_5) had significant higher value than *Dt*$_{1}$ and *Dt*$_{2}$ subgenomes. No significant differences were found between *At* and *Dt* subgenomes either in *G. hirsutum* or *G. barbadense*. For ENC$_{(av)}$, significant differences were detected among the two *At* subgenomes and *G. arboreum* (A_2), and between *G. raimondii* and *Dt*$_{1}$, but not between *Dt*$_{1}$ and *Dt*$_{2}$.

We also compared the *At* and *Dt* subenomes in the same genome of the two tetraploids. In *G. hirsutum*, significant difference was detected only in GC$_{(av)}$ between *At*$_{1}$ and *Dt*$_{1}$ subgenomes. Conversely, significant differences were not detected only in GC$_{(av)}$ between *At*$_{2}$ and *Dt*$_{2}$ subgenomes.

Totally, these four cotton species and the four subgenomes exhibited weak codon usage bias to use pyrimidine-rich codons more frequently than purine-rich codons and there are significant difference in codon base composition and codon usage preference.

Correlation analysis between codon usage bias indices

Studies have shown that the higher the gene expression level is, the stronger is the preferred use of codon [1–3,6,15,16,49,55–60]. In our study, the codon usage bias results of all cotton species were shown in S1 Table. The correlation between the parameters of 4 cotton species and 4 subgenomes had the same rule, except the correlation between G3s and C3s. There was a negative correlation between G3s and C3s among *G. arboreum*, *G. barbadense*, *At*$_{2}$ and *Dt*$_{2}$ subgenomes; and a positive correlation among *G. raimondii*, *G. hirsutum*, *At*$_{1}$ and *Dt*$_{1}$ subgenomes.

Table 4. The correlation coefficients between GC12 and GC3 in 4 cotton species and 4 subgenomes.

	GC1	GC2	GC12
GC2	.827/.969*		
GC12	.984*/.992**	.914/992**	
GC3	-.929/-.283	-.610/-.401	-.865/-.344

The digits before and after backslash represent correlation coefficient among 4 cotton species and 4 subgenomes, respectively.

* Correlation was significant at the 0.05 level (2-tailed).

** Correlation was significant at the 0.01 level (2-tailed).

https://doi.org/10.1371/journal.pone.0194372.t004

The results indicated that there was significant negative correlation between the ENC value and T3s, meanwhile, the ENC value was positively correlated with G3s, C3s and GC3s ($P < 0.01$); in addition, T3s had positive correlation with A3s, and negative correlation with G3s, C3s and GC3s. These correlation results indicated that the base content at the third position of the synonymous codons directly affects the degree of codon usage preference. It could be concluded that genes with stronger codon usage bias (with lower ENC value) would have lower G3s, C3s and higher T3s values. These results indicated that the genes of these species and subgenomes preferred to use high expression codons ending with pyrimidines (T/A).

GC12 represents the average GC content of the first and second positions of the codons. A significant correlation between GC12 and GC3 values means that mutational stress is superior to translation selection in the formation of codon usage bias while non-significant correlation between them reveal that translation selection plays dominant role in codon usage preference [55,61–63]. In our study, firstly we took the nuclear genome CDSs of each cotton species and subgenome as a whole and calculated one GC12/GC3 value per cotton species to analyze the correlation coefficients between GC12 and GC3 in 4 cotton species and 4 subgenomes. From the results in Table 4, there was no significant correlation between GC12 and GC3, implying that codon usage bias was influenced primarily by translation selection in these 4 cotton species and 4 subgenomes.

PR2-bias plot analysis

The PR2-bias plots of the four cotton species and four subgenomes were shown in Fig 3. From Fig 3, it can be seen that along the ordinate, all species genomes and subgenomes presented similar distribution that a majority of genes distributed on the lower left area or the lower right area. However, along the abscissa, there were two types of distributions. A slightly more number of genes of *G*. *arboreum*, *G*. *barbadense* and its two subgenomes distributed on the G > C side than the G < C side while nearly equal amount of genes of *G*. *raimondii*, *G*. *hirsutum* and its two subgenomes distributed on both sides. These results revealed a codon usage imbalance between A + T and G + C at the third base position and indicated that not only the mutation, but also the selection and other factors determined the codon usage patterns in these four cotton species and four subgenomes, and the degree of the third codon position preferences in *G*. *arboreum* and *G*. *barbadense* are slightly different from *G*. *raimondii* and *G*. *hirsutum*. And this was similar to the codon usage bias in chloroplast genome of *G*. *hirsutum* that SCUB was formed under effect of both mutation and selection [36].

ENC and GC3s scatter plot (ENC-plot)

Since ENC is constrained by G+C content of the gene, it is often plotted against GC3s of the gene to investigate patterns of codon usage [42,64]. The ENC-plot of CDSs of four cotton

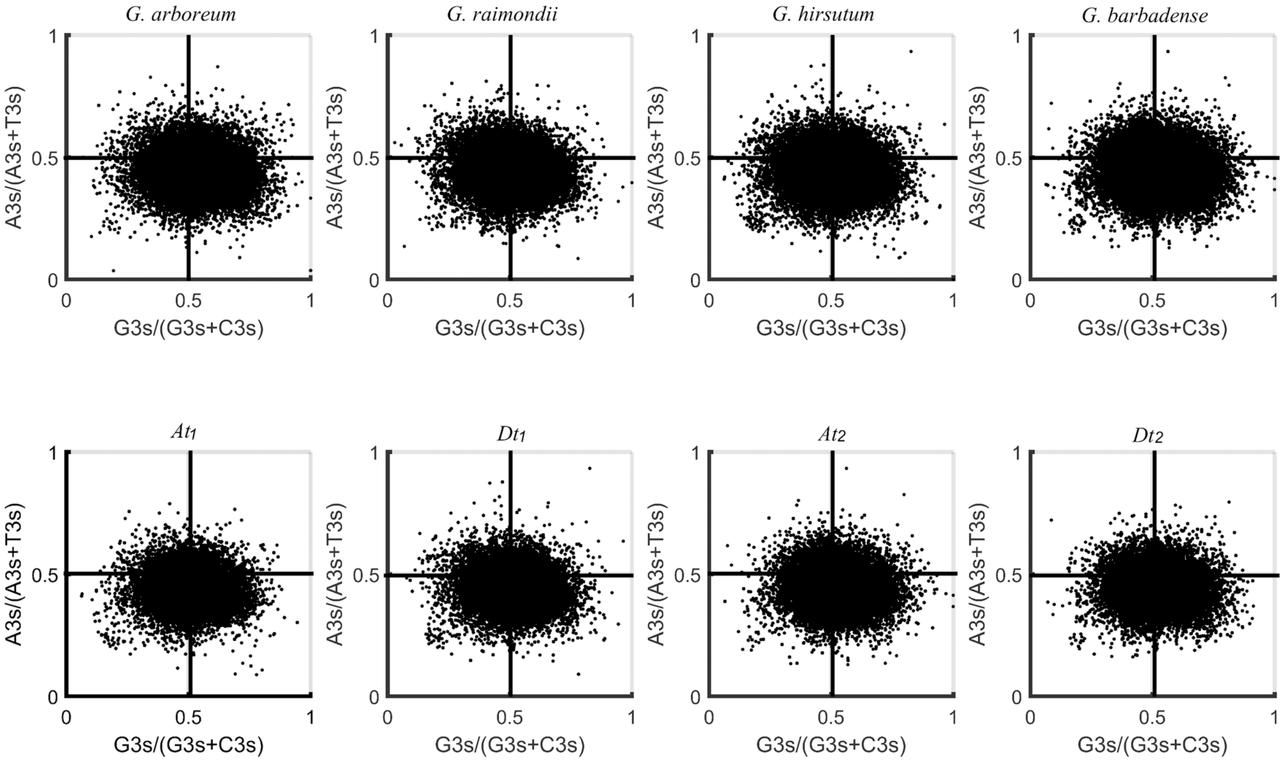

Fig 3. The PR2-bias plots of 4 cotton species and 4 subgenomes.

https://doi.org/10.1371/journal.pone.0194372.g003

species was presented in Fig 4. The solid curve represented the expected position of CDSs whose codon usage was only shaped by the GC3s. Similar ENC-plots are found among all the four cotton species genomes and the four subgenomes of the two tetraploid species (Fig 4). CDSs appeared to cluster around the expected ENC of 30–60. Although a small number of genes distributed in the vicinity of the expected curve, indicated that compositional constraint was the only determinant factor shaping the codon usage pattern. A majority of genes with low ENC values deviated well below the expected curve, indicating that GC3s value was a major determinant of codon usage bias and that other factors independent of nucleotide composition shaped codon usage as well [65]. Consequently, the codon usage pattern of genes in the four cotton genomes and the four subgenomes might be shaped by the combined effects of directional mutation and neutral selection.

RSCU values analysis and determination of putative optimal codons

The RSCU values of CDSs in 4 cotton species and 4 subgenomes were calculated and shown in S2 Table. The RSCU values of 25 codons were greater than 1 and 31 codons were smaller than 1 in all the cotton species, except AAG (Gly), GUG (Gly), and AAA (Lys). Among them, AAG is a low frequency codon in *G. raimondii*, on the contrary, it appeared with high frequency or no preference in other cotton species. In addition, these high frequency codons mostly ended with T (15 of 25) or A (8 of 25) (except UUG and AGG). *At* the same time, most of the low frequency codons ended with C (16 of 31) or G (9 of 31). The above results showed that the RSCU values and such preference of all codons maintained a high degree of unity among all the cotton species.

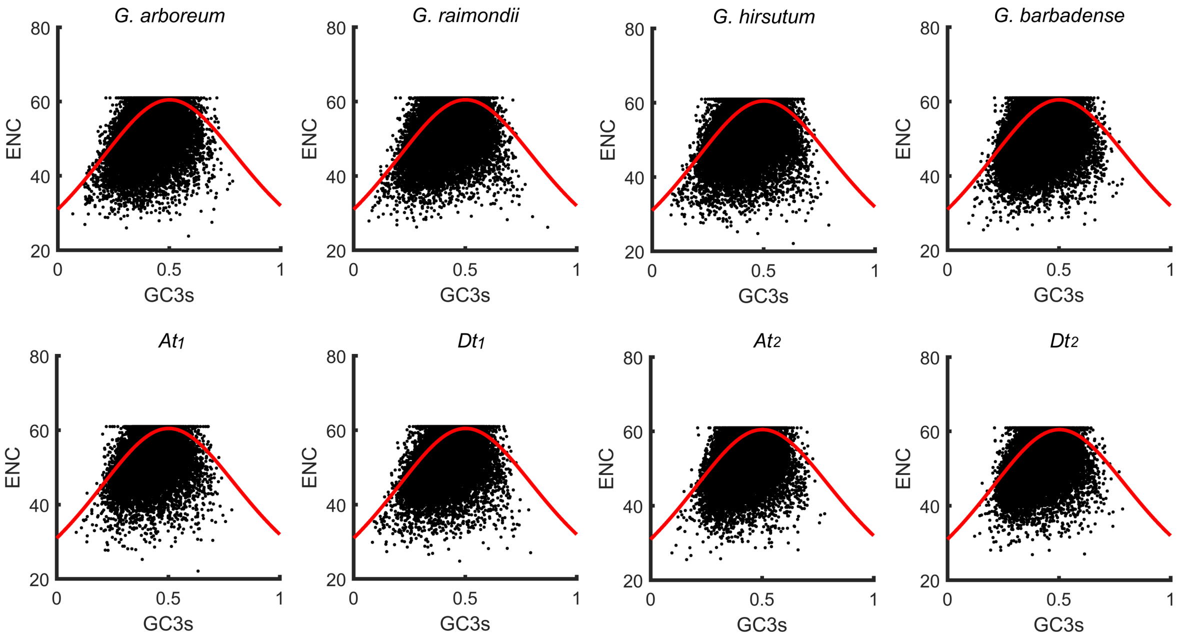

Fig 4. The ENC-plot of 4 cotton species and 4 subgenomes.

https://doi.org/10.1371/journal.pone.0194372.g004

By comparing the RSCU values from two bias libraries of each cotton species, 19 to 26 optimal codons were determined for each cotton species (Table 5), 15 of them ending with T, 9 ending with A, 2 ending with G. However, although most optimal codons ended in either T or A, codons with T at the third position were detected more frequently. There were 19 common optimal codons in these four cotton species.

Compared with the chloroplast genome in *G. hirsutum* [36], the optimal codons of the host nuclear genome in the present study were quite different. The number of optimal codons detected in chloroplast genome [36] and nuclear genome in *G. hirsutum* herein were 23 and 26 respectively, with 12 shared optimal codons (UUG, AUU, UCU, ACU, GCU, CAA, UGU, AGA, GGU, UUA, CCU and GAA). Most of the rest 11 optimal codons specifically determined in chloroplast genome ended with C (7 of 11) while most of the rest 14 optimal codons specifically determined in nuclear genome end with U (8 of 14).

Analysis of translational selection (P2) and choice between pyrimidines in the third position of codon

Grosjean et al first noted in the *MS2 phage* genome that there was a bias in the choice between C and U bases in the third position of codon [44,66]. They found that nucleotides at degenerate positions consistently produced moderate-strength codon-anticodon binding energy. If the first two bases of a codon are both A or U, the C at the third position will give a closer to average codon binding energy than U. Similarly, if the first two bases are either C or G, the third base of "right choice" is U because C gives a strong binding energy. Therefore, we can characterize the translational efficiency of a gene by the frequency of "correct selection", called P2 index, between the pyrimidines in codons starting with AA, AU, UA, UU, CC, CG, GC or GG.

First, the values of WWC, SSC, WWU and SSU were calculated according to the RSCU values of the corresponding codons (Table 6). From Table 6, it was seen that both SSU and WWU were higher than SSC and WWC in all species and subgenomes, especially in *G.*

Table 5. The optimal codons of 4 cotton species and 4 subgenomes.

Codon	*G. arboreum*		*G. raimondii*		*G. hirsutum*		*G. barbadense*		At_1		Dt_1		At_2		Dt_2	
	High	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**	**High**	**Low**
UUU	1.24	1****	1.24	1****	1.21	0.99****	1.22	0.98****	1.18	0.98****	1.23	1****	1.21	0.97****	1.22	1****
UUA	1.04	0.88***	1.1	0.88****	1.07	0.87****	0.99	0.9****	0.98	0.86****	1.16	0.89****			1	0.9****
UUG	1.7	1.41****	1.67	1.4****	1.69	1.41****	1.75	1.3****	1.71	1.41****	1.66	1.41****	1.7	1.23****	1.79	1.38****
CUU	1.46	1.27****	1.54	1.32****	1.49	1.29****					1.49	1.29****				
AUU	1.55	1.18****	1.53	1.22****	1.48	1.2****	1.53	1.18****	1.49	1.2****	1.47	1.21****	1.54	1.18****	1.53	1.19****
GUU	1.79	1.38****	1.91	1.42****	1.86	1.39****	1.86	1.3****	1.88	1.39****	1.85	1.4****	1.84	1.27****	1.88	1.34****
UCU	1.58	1.15****	1.65	1.23****	1.61	1.16****	1.5	1.15****	1.6	1.16****	1.64	1.17****	1.5	1.14****	1.51	1.18****
UCA	1.36	1.15****	1.49	1.17****	1.52	1.14****	1.43	1.16****	1.5	1.14****	1.53	1.14****	1.45	1.16****	1.41	1.15****
CCU	1.61	1.26****	1.63	1.28****	1.53	1.25****					1.53	1.25****				
CCA	1.54	1.21****	1.52	1.18****	1.61	1.19****	1.55	1.22****	1.59	1.17****	1.62	1.2****	1.57	1.25****	1.53	1.19****
ACU	1.59	1.18****	1.59	1.19****	1.53	1.16****	1.54	1.14****	1.55	1.15****	1.51	1.16****	1.53	1.12****	1.56	1.17****
ACA	1.2	1.05***	1.25	1.08****	1.21	1.05****	1.21	1.11****	1.17	1.05****	1.25	1.05****			1.21	1.07****
GCU	1.9	1.38****	1.96	1.42****	1.91	1.4****	1.87	1.34****	1.9	1.39****	1.91	1.39****	1.87	1.29****	1.88	1.38****
GCA					1.16	1.07****			1.15	1.06****					1.17	1.08****
UAU	1.31	1.03****	1.31	1.06****	1.3	1.02****	1.25	1.02****	1.27	1.02****	1.32	1.03****	1.23	1.02****	1.27	1.03****
CAU	1.44	1.14****	1.37	1.15****	1.38	1.14****	1.41	1.12****	1.38	1.13****	1.38	1.16****	1.41	1.1****	1.42	1.16****
CAA	1.36	1.16****	1.32	1.16****	1.37	1.16****	1.33	1.18****	1.34	1.16****	1.4	1.17****	1.31	1.18****	1.35	1.18****
AAU	1.29	1.04****	1.27	1.06****	1.25	1.03****	1.26	1.06****	1.22	1.02****	1.28	1.03****	1.25	1.07****	1.26	1.04****
GAU	1.52	1.27****	1.53	1.28****	1.52	1.26****	1.51	1.25****	1.51	1.25****	1.52	1.26****	1.49	1.25****	1.52	1.25****
GAA			1.19	1.09****	1.19	1.09****					1.21	1.09****				
UGU	1.23	0.98***	1.21	0.99****	1.19	0.96****	1.22	0.99****	1.18	0.96****	1.2	0.97****	1.21	1****	1.23	0.99****
AGU	1.22	0.94****	1.12	0.91****	1.08	0.89****	1.2	0.96****	1.03	0.89****	1.1	0.88****	1.17	0.96****	1.22	0.95****
AGA	2.24	1.36****	2.38	1.4****	2.32	1.32****	2.25	1.34****	2.23	1.31****	2.39	1.33****	2.19	1.32****	2.31	1.36****
AGG	2.01	1.39****	1.83	1.34****	1.87	1.35****	2.06	1.33****	1.9	1.32****	1.84	1.36****	2	1.31****	2.06	1.35****
GGU	1.49	1.14****	1.49	1.13****	1.51	1.14****	1.45	1.14****	1.5	1.13****	1.51	1.15****	1	1.13****	1.43	1.14****
GGA	1.25	1.12****	1.28	1.19****	1.26	1.13****	1.21	1.13****	1.24	1.13****	1.28	1.13****				

*** Correlation is significant at the 0.005 level.

**** Correlation is significant at the 0.001 level.

https://doi.org/10.1371/journal.pone.0194372.t005

raimondii with the largest SSU and WWU and the lowest SSC and WWC. That is, the choice between two pyrimidines (U and C) in the third position of codon tends to U. Then the P2 index of species and subgenomes was calculated (Table 6). All species and subgenomes had P2 values more than 0.5, which revealed that translational selection played the dominant role over mutation pressure in the codons' usage.

Table 6. The values of WWC, SSC, WWU, SSU and P2 in 4 cotton species and 4 subgenomes.

Genomes and subgenomes	SSU	WWU	SSC	WWC	P2
G. arboreum	5.30	4.94	2.40	3.28	0.5389
G. raimondii	5.44	5.09	2.33	3.11	0.5354
G. hirsutum	5.31	4.94	2.40	3.29	0.5395
G. barbadense	5.32	4.94	2.40	3.29	0.5398
At_1	5.29	4.94	2.41	3.29	0.5386
Dt_1	5.26	4.94	2.40	3.29	0.5381
At_2	5.25	4.93	2.43	3.30	0.5374
Dt_2	5.26	4.94	2.39	3.28	0.5381

https://doi.org/10.1371/journal.pone.0194372.t006

RSCU-based cluster analysis

The results of RSCU-based cluster analysis were shown in Fig 5. In Fig 5, the two tetraploids *G. hirsutum* and *G. barbadense* were grouped into two different clusters. The two *Dt* subgenomes

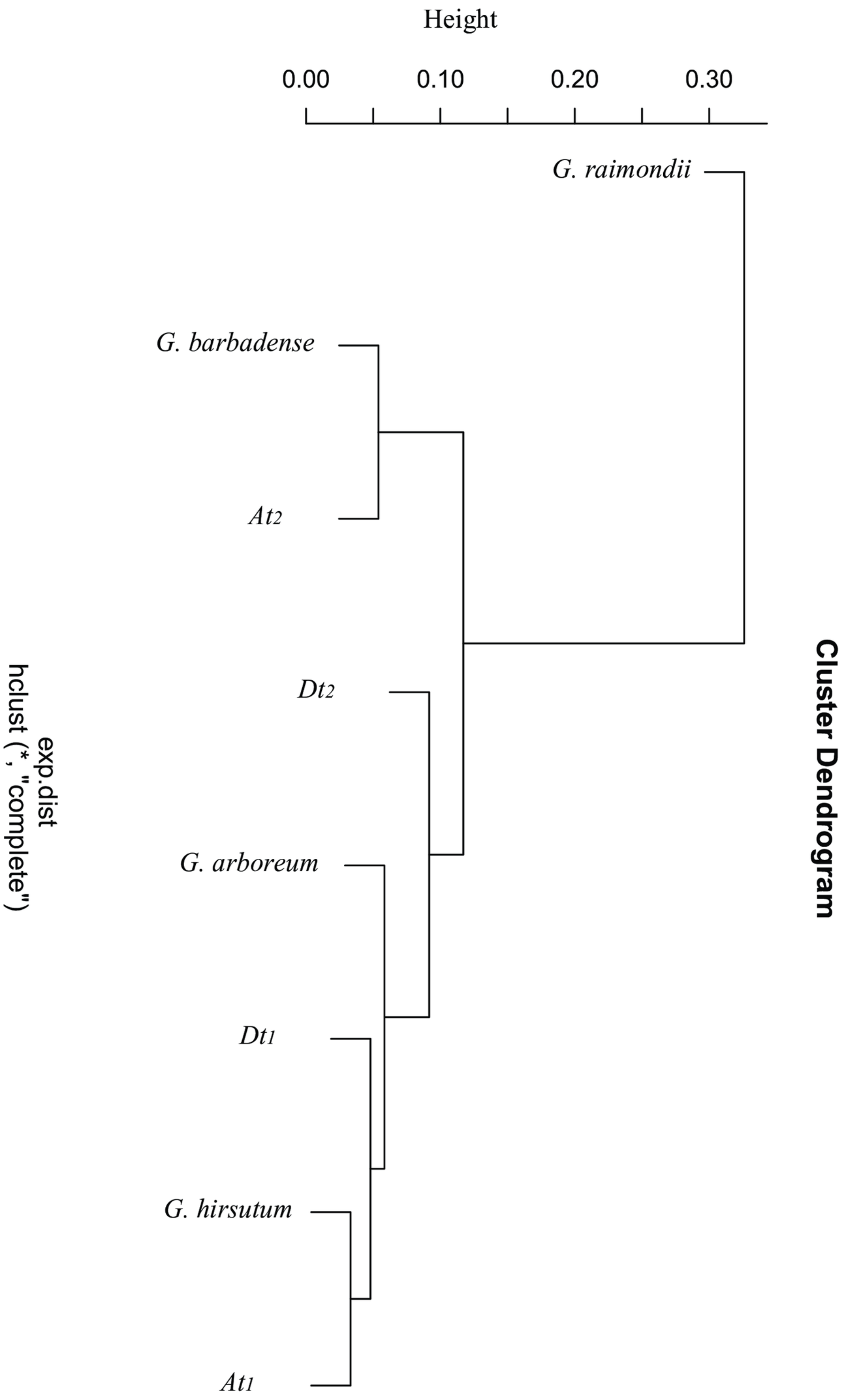

Fig 5. Cluster tree based on the RSCU values of 4 cotton species and 4 subgenomes.

https://doi.org/10.1371/journal.pone.0194372.g005

and the D_5 genome (*G. raimondii*) were grouped into different clusters, so as to the two *At* subgenomes and the A_2 genome (*G. arboreum*). These results are quite inconsistent with the widely accepted taxonomic and phylogenetic relationship of these four cotton species [24,26,28,67–70]. The allopolyploid cotton species may have appeared through hybridization and subsequent polyploidization events between the A- and D-subgenome progenitors. The D-genome species *G. raimondii* (D_5) and the A-genome species much like modern *G. arboreum* (A_2) and *G. herbaceum* (A_1) have been supported by molecular methods and other evidence [28,68,71–72] to be the donor species for the *Dt* and *At* subgenome of the tetraploid cotton species, respectively. The monophyly of polyploid Gossypium species was also studied through cluster analysis based on sequences of a 2.8-kb intergenic region from all diploid species belonging to the genome groups from which the polyploid originates [28], in which all the *Dt* subgenomes and D_5 genome were grouped into one cluster, and all the *At* subgenomes and the A_1 and A_2 genomes were grouped into another cluster. The results of the present study indicated that the evolutionary relationship among these four cotton species could not be well reflected by RSCU-based cluster analysis.

Conclusions

In the present study, codon usage bias patterns and the shaping factors in the four sequenced cotton genomes of *G. arboreum* (A_2), *G. raimondii* (D_5), *G. hirsutum* (AD_1) and *G. barbadense* (AD_2), and the four subgenomes (At_1, Dt_1, At_2, and Dt_2) of these two tetraploids were addressed and compared. All these genomes and subgenomes exhibited similar weaker codon usage bias revealed by the results of less ($< 0.5\%$) genes with low (< 35) ENC and more genes ($> 70\%$) with high ENC. Codon composition analysis revealed these species and subgenomes had low GC and GC3, tended to use pyrimidine-rich codons more frequently than purine-rich codons at the third positions of codons and follow the $T > A > G > C$ trend, although there was significant difference in codon composition and codon usage preference among them. Correlation analysis indicated that the base content at the third position of codons affected the degree of codon preference. PR2-bias plot and ENC-plot revealed that not only translation selection but also directional mutation and other factors shaped the CUB. The P2 analysis results, with the non-significant correlation between GC12 and GC3, further revealed that translation selection was the main factor influencing the CUB pattern herein. Through RSCU analysis, 25 high frequency codons preferentially ended with T or A, and 31 low frequency codons preferentially ended with C or G common in these genomes and subgenomes were determined. And 19 to 26 optimal codons were determined, including 19 common ones, for each species and subgenome. The optimal codons preferred to end with A or T. Finally, we concluded that these four cotton genomes had weak CUB, translation selection played dominant role over mutation pressure in codon usage preference in these four cotton species, and *At* and *Dt* subgenomes had similar codon usage patterns with their A- and D-genome progenitors.

Supporting information

S1 Table. The correlation analysis between codon usage bias indices in 4 cotton species and 4 subgenomes.
(XLSX)

S2 Table. The RSCU values of CDSs in 4 cotton species and 4 subgenomes.
(XLSX)

Author Contributions

Data curation: Liyuan Wang, Yanchao Yuan, Xianlin Wang, Jincai Tao, Wei Feng.

Formal analysis: Liyuan Wang, Xianliang Song, Xuezhen Sun.

Methodology: Liyuan Wang.

Project administration: Guihua Zhang, Xianliang Song, Xuezhen Sun.

Visualization: Xianliang Song.

Writing – original draft: Liyuan Wang.

Writing – review & editing: Liyuan Wang, Huixian Xing, Muhammad Saeed, Xianliang Song, Xuezhen Sun.

References

1. Gu W, Zhou T, Ma J, Sun X, Lu Z. Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the *Nidovirales*. Virus Research. 2004; 101(2):155–161. https://doi.org/10.1016/j.virusres.2004.01.006 PMID: 15041183.
2. Chakraborty S, Nag D, Mazumder TH, Uddin A. Codon usage pattern and prediction of gene expression level in *Bungarus* species. Gene. 2017; 604:48–60. https://doi.org/10.1016/j.gene.2016.11.023 PMID: 27845207.
3. Behura SK, Severson DW. Comparative analysis of codon usage bias and codon context patterns between *Dipteran* and *Hymenopteran* sequenced genomes. PLoS One. 2012; 7(8):e43111. https://doi.org/10.1371/journal.pone.0043111 PMID: 22912801.
4. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. Genome Biology and Evolution. 2011; 3:868–80. https://doi.org/10.1093/gbe/evr085 PMID: 21856647.
5. Salim HMW, Cavalcanti ARO. Factors influencing codon usage bias in genomes. Journal of the Brazilian Chemical Society. 2008; 19(2):257–62. https://doi.org/10.1590/S0103-50532008000200008
6. Baeza M, Alcaino J, Barahona S, Sepulveda D, Cifuentes V. Codon usage and codon context bias in *Xanthophyllomyces dendrorhous*. BMC Genomics. 2015; 16(1):293. https://doi.org/10.1186/s12864-015-1493-5 PMID: 25887493
7. Dohra H, Fujishima M, Suzuki H. Analysis of amino acid and codon usage in *Paramecium bursaria*. FEBS Letters. 2015; 589(20 Pt B):3113–3118. https://doi.org/10.1016/j.febslet.2015.08.033 PMID: 26341535.
8. Marais G, Mouchiroud D, Duret L. Neutral effect of recombination on base composition in *Drosophila*. Genetical Research. 2003; 81(2):79–87. PMID: 12872909.
9. Sharp PM, Li WH. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Molecular Biology and Evolution. 1987; 4(3):222–230. https://doi.org/10.1093/oxfordjournals.molbev.a040443 PMID: 3328816.
10. Olejniczak M, Uhlenbeck OC. tRNA residues that have coevolved with their anticodon to ensure uniform and accurate codon recognition. Biochimie, 2006, 88(8):943–950. https://doi.org/10.1016/j.biochi.2006.06.005 PMID: 16828219.
11. Zalucki YM, Power PM, Jennings MP. Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins. Nucleic Acids Research, 2007, 35(17):5748. https://doi.org/10.1093/nar/gkm577 PMID: 17717002.
12. Huang Y, Koonin EV, Lipman DJ, Przytycka TM. Selection for minimization of translational frame shifting errors as a factor in the evolution of codon usage. Nucleic Acids Research, 2009, 37(20):6799–6810. https://doi.org/10.1093/nar/gkp712 PMID: 19745054
13. Sun Z, Ma L, Murphy R, Zhang XS, Huang DW. Analysis of codon usage on *Wolbachia pipientis* wMel genome. Science in China Series C: Life Sciences, 2009, 39(10):948–953.
14. Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y. Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe*. Genes to Cells, 2009, 14:499–509. https://doi.org/10.1111/j.1365-2443.2009.01284.x PMID: 19335619.
15. Prabha R, Singh DP, Sinha S, Ahmad K, Rai A. Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes. Marine Genomics. 2017; 32:31–9. https://doi.org/10.1016/j.margen.2016.10.001 PMID: 27733306.

16. Vicario S, Moriyama EN, Powell JR. Codon usage in twelve species of *Drosophila*. BMC Evolutionary Biology. 2007; 7:226. https://doi.org/10.1186/1471-2148-7-226 PMID: 18005411

17. Subramanian S. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics. 2008; 178(4):2429–32. https://doi.org/10.1534/genetics.107.086405 PMID: 18430960.

18. Wu XM, Wu SF, Ren DM, Zhu YP, He FC. The analysis method and progress in the study of codon bias. Hereditas (Beijing). 2007; 29(4):420–6. CSCD:2726594.

19. Zhao Y, Zheng H, Xu A, Yan D, Jiang Z, Qi Q, et al. Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution. BMC Genomics. 2016; 17:677. https://doi.org/10.1186/s12864-016-3021-7 PMID: 27558469.

20. Liu XS, Zhang YG, Fang YZ, Wang YL. Patterns and influencing factor of synonymous codon usage in porcine circovirus. Virology Journal. 2012; 9:68. https://doi.org/10.1186/1743-422X-9-68 PMID: 22416942

21. Pandit A, Sinha S. Differential trends in the codon usage patterns in *HIV-1* genes. PLoS One. 2011; 6(12):28889. https://doi.org/10.1371/journal.pone.0028889 PMID: 22216135.

22. Liu H, He R, Zhang H, Huang Y, Tian M, Zhang J. Analysis of synonymous codon usage in *Zea mays*. Molecular Biology Reports. 2010; 37(2):677–84. https://doi.org/10.1007/s11033-009-9521-7 PMID: 19330534

23. Chen ZJ, Scheffler BE, Dennis E, Triplett BA, Zhang TZ, Guo WZ, et al. Toward sequencing cotton (*Gossypium*) genomes. Plant Physiology. 2007; 145(4):1303–1310. https://doi.org/10.1104/pp.107.107672 PMID: 18056866.

24. Wendel J, Albert VA. Phylogenetics of the cotton genus (*Gossypium*): character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. Systematic Botany. 1992; 17:115–143. https://doi.org/10.2307/2419069

25. Krapovickas A, Seijo G. *Gossypium ekmanianum* (Malvaceae), algodon Silvestre de la Republica Dominicana. Bonplandia. 2008; 17:55–63.

26. Gallagher JP, Grover CE, Rex K, Moran M, Wendel JF. A new species of cotton from Wake *At*oll, *Gossypium stephensii* (Malvaceae). Systematic Botany. 2017; 42(1):115–123. https://doi.org/10.1600/036364417X694593

27. Wendel JF, Brubaker C, Alvarez I, Cronn R, Stewart JM, Lubbers EL, et al. Genetics and genomics of cotton. Genetics & Genomics of Cotton, 2009, 3.

28. Grover CE, Grupp KK, Wanzek RJ, Wendel JF. Assessing the monophyly of polyploid *Gossypium* species. Plant Systematics and Evolution. 2012; 298(6):1177–1183. https://doi.org/10.1007/s00606-012-0615-7

29. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, et al. The draft genome of a diploid cotton *Gossypium raimondii*. Nature Genetics. 2012; 44(10):1098–103. https://doi.org/10.1038/ng.2371 PMID: 22922876.

30. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibers. Nature. 2012; 492(7429):423–428. https://doi.org/10.1038/nature11798 PMID: 23257886.

31. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, et al. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nature Genetics. 2014; 46(6):567–72. https://doi.org/10.1038/ng.2987 PMID: 24836287.

32. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, et al. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nature Biotechnology. 2015; 33 (5):524–530. https://doi.org/10.1038/nbt.3208 PMID: 25893780.

33. Zhang T, Hu Y, Jiang W, Fang L, Guan X, Chen J, et al. Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nature Biotechnology. 2015; 33 (5):531–7. https://doi.org/10.1038/nbt.3207 PMID: 25893781.

34. Yuan D, Tang Z, Wang M, Gao W, Tu L, Jin X, et al. The genome sequence of Sea-Island cotton (*Gossypium barbadense*) provides insights into the allopolyploidization and development of superior spinnable fibres. Scientific Reports. 2015; 5:17662. https://doi.org/10.1038/srep17662 PMID: 26634818.

35. Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, Yang CQ, et al. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Scientific Reports. 2015; 5:14139. https://doi.org/10.1038/srep14139 PMID: 26420475.

36. Shang M, Liu F, Hua J, Wang K. Analysis on codon usage of chloroplast genome of *Gossypium hirsutum*. Scientia Agricultura Sinica. 2011; 44(2):245–53. https://doi.org/10.3864/j.issn.0578-1752.2011.02.003 CABI:20113088541.

37. Lee SB, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, et al. The complete chloroplast genome sequence of *Gossypium hirustum*: Organization and phylogenetic relationships to other angiosperms. BMC Genomics. 2006; 7:61–72. https://doi.org/10.1186/1471-2164-7-61 PMID: 16553962.

38. Liu G, Cao D, Li S, Su A, Geng J, Grover CE, et al. The Complete mitochondrial genome of *Gossypium hirsutum* and evolutionary analysis of higher plant mitochondrial genomes. PLoS One. 2013; 8(8): e69476. https://doi.org/10.1371/journal.pone.0069476 PMID: 23940520.

39. Wei S, Shuai L, Yan L, Hu FQ. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016, 11(10):e0163962. https://doi.org/10.1371/journal.pone.0163962 PMID: 27706213.

40. Sueoka N. Intrastrand parity rules of DNA base composition and usage of synonymous codons. Journal of Molecular Evolution. 1995; 40(3):318–25. https://doi.org/10.1007/bf00163236 PMID: 7723058.

41. Sueoka N. Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses. Journal of Molecular Evolution. 2001; 53(4–5):469–76. https://doi.org/10.1007/s002390010237 PMID: 11675607

42. Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87(1):23–9. https://doi.org/10.1016/0378-1119(90)90491-9 PMID: 2110097.

43. Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. Molecular Biology and Evolution. 2002; 19(8):1390–4. https://doi.org/10.1093/oxfordjournals.molbev.a004201 PMID: 12140252.

44. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Research. 1982; 10(22):7055. https://doi.org/10.1093/nar/10.22.7055 PMID: 6760125.

45. Mcewan NR, Gatherer D. Codon indices as a predictor of gene functionality in a *Frankia* operon. Canadian Journal of Botany. 1999; 77(9):1287–92. https://doi.org/10.1139/b99-068

46. Gatherer D, Mcewan NR. Small regions of preferential codon usage and their effect on overall codon bias-The case of the *plp* gene. Biochemistry and Molecular Biology International. 1997, 43(1):107–114. https://doi.org/10.1080/15216549700203871 PMID: 9315288.

47. Liu Q, Xue Q. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. Journal of Genetics. 2005, 84(1):55–62. https://doi.org/10.1007/BF02715890 PMID: 15876584.

48. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc Natl Acad Sci U S A. 1999; 96(8):4482–4487. https://doi.org/10.1073/pnas.96.8.4482 PMID: 10200288.

49. Wang SF, Su MW, Tseng SP, Li MC, Tsao CH, Huang SW, et al. Analysis of codon usage preference in hemagglutinin genes of the swine-origin influenza A (*H1N1*) virus. Journal of Microbiology and Immunology Infection. 2016; 49(4):477–86. https://doi.org/10.1016/j.jmii.2014.08.011 PMID: 25442859.

50. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter. 1991; 9(3):208–18. https://doi.org/10.1007/BF02672069

51. Han Z, Wang C, Song X, Guo W, Gou J, Li C, et al. Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theoretical Applied Genetics. 2006; 112 (3):430–9. https://doi.org/10.1007/s00122-005-0142-9 PMID: 16341684.

52. Tao T, Zhao L, Lv Y, Chen J, Hu Y, Zhang T, et al. Transcriptome sequencing and differential gene expression analysis of delayed gland morphogenesis in *Gossypium australe* during seed germination. PLoS One. 2013; 8(9):e75323. https://doi.org/10.1371/journal.pone.0075323 PMID: 24073262.

53. Hui G, Wang X, Gundlach H, Mayer KF, Peterson DG, Scheffler BE, et al. Extensive and biased intergenomic nonreciprocal DNA exchanges shaped a nascent polyploid genome, *Gossypium* (cotton). Genetics. 2014; 197(4):1153–63. https://doi.org/10.1534/genetics.114.166124 PMID: 24907262.

54. Jiang Y, Deng F, Wang HL, Hu ZH. An extensive analysis on the global codon usage pattern of *baculoviruses*. Archives of Virology. 2008; 153(12):2273–2282. https://doi.org/10.1007/s00705-008-0260-1 PMID: 19030954.

55. Wang H, Liu S, Zhang B, Wei W. Analysis of synonymous codon usage bias of *Zika Virus* and its adaption to the hosts. PLoS One. 2016; 11(11):e0166260. https://doi.org/10.1371/journal.pone.0166260 PMID: 27893824.

56. Ma YP, Zhou ZW, Liu ZX, Hao L, Ma JY, Feng GQ, et al. Codon usage bias of the phosphoprotein gene of spring viraemia of carp virus and high codon adaptation to the host. Archives of Virology. 2014; 159(7):1841–7. https://doi.org/10.1007/s00705-014-2000-z PMID: 24519460.

57. Deb S, Basak S. Comparative study of codon usage pattern and compositional distribution between whole genome and virulence gene set of *Vibrio cholerae* N16961. Computational Molecular Biology. 2015. https://doi.org/10.5376/cmb.2015.05.0006

58. Kattoor JJ, Malik YS, Sasidharan A, Rajan VM, Dhama K, Ghosh S, et al. Analysis of codon usage pattern evolution in avian rotaviruses and their preferred host. Infection, Genetics and Evolution. 2015; 34:17–25. https://doi.org/10.1016/j.meegid.2015.06.018 PMID: 26086995.

59. Suzuki H, Morton BR. Codon adaptation of plastid genes. PLoS One. 2016; 11(5):e0154306. https://doi.org/10.1371/journal.pone.0154306 PMID: 27196606.

60. Li N, Sun MH, Jiang ZS, Shu HR, Zhang SZ. Genome-wide analysis of the synonymous codon usage patterns in apple. Journal of Integrative Agriculture. 2016; 15(5):983–91. https://doi.org/10.1016/s2095-3119(16)61333-3

61. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Research. 2003; 92(1):1–7. https://doi.org/10.1016/s0168-1702(02)00309-x PMID: 12606071.

62. Wang M, Liu YS, Zhou JH, Chen HT, Ma LN, Ding YZ, et al. Analysis of codon usage in *Newcastle* disease virus. Virus Genes. 2011; 42(2):245–53. https://doi.org/10.1007/s11262-011-0574-z PMID: 21249440

63. Hussain S, Rasool ST. Analysis of synonymous codon usage in *Zika* virus. Acta Tropica. 2017; 173:136–46. https://doi.org/10.1016/j.actatropica.2017.06.006 PMID: 28606821.

64. Das S, Paul S, Dutta C. Synonymous codon usage in adenoviruses: Influence of mutation, selection and protein hydropathy. Virus Research. 2005; 117(2):227–36. https://doi.org/10.1016/j.virusres.2005.10.007 PMID: 16307819.

65. Liu H, Huang Y, Du X, Chen Z, Zeng X, Chen Y, et al. Patterns of synonymous codon usage bias in the model grass *Brachypodium distachyon*. Genetics and Molecular Research. 2012; 11(4):4695–4706. https://doi.org/10.4238/2012.October.17.3 PMID: 23096921.

66. Grosjean H, Sankoff D, Jou WM, Fiers W, Cedergren RJ. Bacteriophage *MS2* RNA: a correlation between the stability of the codon: anticodon interaction and the choice of code words. Journal of Molecular Evolution. 1978; 12(2):113. https://doi.org/10.1007/BF01733262 PMID: 368346.

67. Galau GA, Wilkins TA. Alloplasmic male sterility in AD allotetraploid *Gossypium hirsutum* upon replacement of its resident A cytoplasm with that of D species *G*. *harknessii*. Theoretical and Applied Genetics. 1989; 78(1):23–30. https://doi.org/10.1007/BF00299748 PMID: 24227025.

68. Wendel JF. New World tetraploid cottons contain Old World cytoplasm. Proc Natl Acad Sci U S A. 1989; 86(11):4132–4136. PMID: 16594050.

69. Wendel JF, Cronn RC. Polyploidy and the evolutionary history of cotton. Advances in Agronomy. 2003; 78:139–86. https://doi.org/10.1016/s0065-2113(02)78004-8

70. Wu YX, Chen JH, He QL, Zhu SJ. Parental origin and genomic evolution of tetraploid *Gossypium* species by molecular marker and GISH analyses. Caryologia. 2013, 66(4):368–374. https://doi.org/10.1080/00087114.2013.857830

71. Endrizzi JE, Turcotte EL, Kohel RJ. Genetics, cytology, and evolution of *Gossypium*. Advances in Genetics. 1985, 23:271–375.

72. Wendel JF, Brubaker CL, Seelanan T. The origin and evolution of *Gossypium*. Physiology of cotton. 2010:1–18. https://doi.org/10.1007/978-90-481-3195-2_1