Comprehensive analysis of codon usage pattern of whole genome in protozoa, Stylonychia lemnae

Ying Wang (yingyingWY2018@outlook.com)
Harbin Normal University https://orcid.org/0000-0001-9263-1034
Lin Yao
Harbin Normal University
Jinfeng Fan
Harbin Normal University
Xueying Zhang
Harbin Normal University
Changhong Guo
Harbin Normal University
Ying Chen
Harbin Normal University

Research article

Keywords: S. lemnae; macronucleus; codon usage; mutation pressure; natural selection

Posted Date: October 16th, 2019

DOI: https://doi.org/10.21203/rs.2.16114/v1

Abstract

Background: Codon usage pattern is an important evolutionary feature in genomes widely observed in many organisms. Stylonychia lemnae is a classical model single-celled eukaryote, and a quintessential ciliate typified by dimorphic nuclei: a germline micronucleus and a vegetative macronucleus. Analysis of codon usage pattern of S. lemnae macronucleus genome helps in understanding evolution at molecular level and acquires significance in mRNA translation, design of transgenic and new gene discovery. Results: The codons of the macronucleus genome sequence of S. lemnae were analyzed and 20,750 coding sequences (CDS) were screened. The overall codon usage of S. lemnae is similar and slightly biased. The value of effective number of codons (ENC) showed that the overall extent of codon usage bias in S. lemnae is relatively high. Nucleotide analysis showed that the overall codon usage is biased toward A- and U-ending codons. The phylogenetic analysis indicated that ciliate is independent evolutionary origins from a common ancestor. The RSCU analysis showed that the codon usage pattern of S. lemnae is more similar to that of Thtrahymana thermophila and Paramecium caudatum . Correlation analysis, ENC-GC 3S plot, and PR2 plot indicated that the codon usage patterns of S. lemnae are not only influenced by mutational pressure but also by natural selection, but neutrality plot analysis showed that the latter plays a major role. C onclusions : Codon usage patterns in eukaryotes are not determined by translational efficiency, but also are determined by the genome. Our study is the first attempt to evaluate the codon usage pattern of S.lemnae macronucleus genome to better understand the evolutionary changes. These results built the base for further research on the molecular evolution of S. lemnae .

Background

The ciliate *S. lemnae* is a classical model single-celled eukaryote, which widespread in ponds, rivers and marshes. *S. lemnae* possesses both a macronucleus (MAC), specialized for gene expression, and a micronucleus (MIC), containing the germline genome that permits recombination and transmission of genetic information across sexual generations [1]. The macronucleus genome of *S. lemnae* is 50.2-Mb, contains 80 ribosomal relate genes [2].

As we all know, the codons encoding the same amino acid are called the synonymous codons, and all amino acids are coded by more than one codon except for methionine and tryptophan[3].The synonymous codons don't occur equally both within and between genomes, and several different code variants have arisen independently, even within a single class[4].

Ciliate is a complex protozoan with two nuclei and its codon usage bias is not universal [5, 6]. With the exception of a few species, most of higher eukaryotes use the standard genetic coding system for protein translation, as do viral and prokaryotic. However, in lower eukaryotic cell, the genetic code expresses its particularity. *Stylonychia* are hypotrichous ciliates with alternative genetic codes. It has reassigned the canonical stop codons TAR to glutamine codons, resulting in genomes with only one stop codon (TGA) [7,8,9]. Whereas in *Oxytricha*, UAA and UAG encode glutamine, TGA is only one stop codon. Furthermore, Non-standard glutamine codons are not used in the four species of Euplotes [10].

The codon usage patterns in *S. lemnae* was analyzed by Martindale D.W in 1989[11], but only two tubulin protein-coding gene sequences were available for analysis. Codon usage has been shown to affect protein structure and function through interfering with translation kinetics, and cotranslational protein folding [12]. Furthermore, synonymous codon usage patterns can be an essential tool in revealing specie evolution [13]. The sequences data were incomplete for *S. lemnae* codon analysis in previous studies, and it is not clear whether the influencing factor on its codon usage with each other, such as mutation pressure [14] or natural selection [15].

Heritable Variations in *Stylonychia sp* is reported firstly in 1915, and causing who the main fact has considered to be cumulative effects of selection [16] .But due to the lack of genomic data, the molecular mechanism of genetic change cannot be elucidated. Basing on the reason, we have downloaded all the complete *S. lemnae* macronuclear genome coding sequences of until June the 21st, 2019 and have analyzed their codon usage patterns.

Results

Phylogenetic analysis of *S. lemnae* based on 18s ribosomal DNA (18s-rDNA)

To determine the phylogenetic relationship of *S. lemnae*, a phylogenetic tree was drawn (Fig.1). The results show that *S. lemnae* clusters with Stylonychia sp. with high support (ML/BI:86/0.76), which is then sister to *Tetrahymena sp.*and *S. cerevisiae* with low bootstrap value (ML/BI: 78/0.76, ML/BI: 60/0.45 respectively) ; which clusters with *C.elegans* and *H.sapiens* in a poorly supported clade (ML/BI: 55/0.43); and also groups with *Paramecium sp.*.

Codon composition analysis

CodonW software was used to analyze the macronucleus genome coding sequence of the *S. lemnae*. The GC content ranged from 22.1% to 58.6% in the gene of *S. lemnae*. The GC content was mainly distributed in 25% ~ 40% with an average of 33.7%, indicating that AT was enriched in the genome (Supplementary Fig. S1).

Relative synonymous codon usage analysis

The patterns of synonymous codon usage in *S. lemnae* coding sequences were assessed by RSCU analysis. Among 26/ preferred codons of corresponding all amino acids (except Methionine and Tryptophan) in *S. lemnae* coding sequences, 26 are A/U-ended (twelve A-ended; fourteen U-ended) and the remaining are C/G ended (one C-ended; one G-ended). Therefore, most of preferentially used codons in *S. lemnae* are A-ended or U-ended codons (Table 1). By analyzing the 26 preferred codons, we can find that the RSCU values of six codons, GGA (G), UCA (S), CCA(P), ACU(T), GCU (A) and AGA(R) are >1.6, whereas the RSCU values of the remaining are also found to be > 1 and< 1.6 (Table 1). Nucleotide composition (A/T-rich) and RSCU analysis (A/U-ended) show that selection of the preferred codons has been influenced by compositional constraints, which indicated that nature selection mostly, shaped its codon pattern.

Based on these data, it was found that in *S. lemnae* coding sequences, 25 are A/U-ended in highly expressed gene, but 21 are C/G end in lowly-expressed gene (Table 2). It verifies A-ended or U-ended codons are preferentially used codons of highly expressed genes.

To determine the potential influences (mutation pressure or natural selection) on the codon usage patterns, the RSCU values of the codons in *S. lemnae* coding sequences were calculated and then were compared with five model organism (*H. sapiens*, *C. elegans, S. cerevisiae, T. thermophila* and *P. caudatum*). We find that preferred codons is eight between *S. lemnae* and *H.Sapiens*; fourteen between *S. lemnae* and *C. elegans*; twenty one between *S. lemnae* and *S. cerevisiae*; twenty one between *S. lemnae* and *T.thermophila*; and nineteen between *S. lemnae* and *P. caudatum* (Table 1). In all, the similarity in codon pattern between *S. lemnae* and *H. Sapiens* is lower than that among *C. elegans* *S. cerevisiae* *T.thermophila* or *P. caudatum*. The codon usage bias of *S. lemnae* differs greatly from that of higher eukaryotes and is similar to that of lower eukaryotes. These results suggest that the selection pressure maybe affect the codon usage pattern of *S. lemnae*.

Correlation analysis

To determine whether the codon usage patterns of *S. lemnae* coding sequences are mainly influenced by mutation pressure or natural selection, we performed a correlation analysis between the nucleotide compositions and the third base of synonymous codons(Table 3). The results show that the A content has a significant positive correlation with the content of A3s and G3s, but has a significant negative correlation with the content of C, T, G, GC, C3s, T3s and GC3s. The C content has a significant positive correlation with the content of G, GC, C3s, GC3s and ENC, but has a significant negative correlation with the content of A, T, A3s T3s and G3s. The T content has a significant positive correlation with T3s contents, but has a significant negative correlation with the content of A, C, G, GC, A3s, C3s, G3S, GC3s and Enc. The G content has a significant positive correlation with the content of G, GC, C3s, G3s, GC3s and ENC, but has a significant negative correlation with A, T, A3s and T3s content. The GC contents has a significant positive correlation with the content of CG, G3s, C3s, GC3s, but has a significant negative correlation with the content of A, T, A3s and T3s. The ENC value has a significant positive correlation with the content of C, G, GC, C3s, G3s and GC3s, but has a significant negative correlation with the content of T, A3s and T3s. These results indicate that compositional constraints under mutation pressure may affect the codon usage pattern for *S. lemnae*.

To study the relative contribution of two major factors, i.e., natural selection and mutational pressure on codon usage, we performed PCA analysis taking RSCU scores to find out major trends of codon usage in *S. lemnae* genes. A plot of PC1 and PC2 showed important features of the codon usage pattern in *S. lemnae* genes (Fig.2a). From this analysis major trends in codon usages were detected in which axis1 (PC1) accounted for 13.4%, whereas axis2 (PC2), axis3 and axis4 accounted for 10.2%, 6.6%, and 4.6% of total variation in *S. lemnae*. Axis1-axis4 explaining 34.8% of the cumulative variances, which indicates that no single factor influence condon usage patteren in in *S. lemnae*.

In order to characterize the codon usage patterns from different types of genes, ribosomal related genes was statistical analysis by PCA (Fig. 2b).It was clearly seen that ribosomal genes of *S. lemnae were* clustered on the right side of PC1, and indicate that compositional constraints are a major factor in CUB, that is n mutation pressure mostly shaped its codon pattern, but other factors are also powerful.

Additationlly, Correlation analysis was also performed to determine the correlations between the first two axes and nucleotide constraints of *S. lemnae* genome (Table 4). The results show that the Axis1 is positively correlated with the A and A3s, whereas it is negatively correlated with the contents of C, G, GC, C3s, GC3s and ENc. Meanwhile, Axis2 is insignificant correlated with the C, T, G, GC, A3s,C3s,T3s, G3s, GC3s and ENC. Overall, these results indicating that mutation pressure has played a major role in shaping the codon usage patterns of *S. lemnae* genomes.

To determine the potential influence of natural selection, correlation analysis was performed between the characters of aminoacid (Gravy values and Aroma) and the codon bias (Axis1, Axis2, ENC, and GC 3s) (Table 5). Our analysis indicates that Axe1 have a significant negative correlation

with AROMA and GRAVY, and Axis 2 has a significant positively correlation with AROMA and GRAVY. Earlier it is found that AROMA and hydropathy of the encoded proteins has a significant correlation with the base composition of third codon positions in some other prokaryotes, several eukaryotes and viral genomes [17,18,19,20,21].However, there is no report of such correlation in any of the ciliate genomes studied so far. To our knowledge, this is for the first time; a correlation has been demonstrated between the synonymous codon usages in genes of *S. lemnae* genomes. All in, the aromaticity and hydrophobicity of amino acid have effect on the codon usage pattern of *S. lemnae*, which reveal that the importance of nature selection.

ENC- GC 3S plot analysis

To determine whether the codon usage patterns of *S. lemnae* coding sequences have been shaped by mutation pressure, natural selection or both, we constructed ENC-GC 3S plot, PR2 plot and neutrality plot analysis.

The degree of codon bias is reflected by the size of ENC value. ENC value ranges from 20 to 61, with the level of base composition bias increasing as the ENC values approach 20. Similarly, genes expressed at low levels contain numerous rare codons, with a higher ENC value approach 61. The convention uses 35 as the criterion for biased bias [22]. The ENC values of the *S. lemnae* genome range from 24.8 to 61, and most of them are more than 35, so the codon bias of the gene is weak. The average GC3 content is 33.7%, GC1 and GC2 are 38.68% and 30.1%, respectively, indicating that the codon base composition is mostly A and U.

The association analysis between ENC - GC3 is shown in Fig. 3. If ENC value of genes will lie on or just below the continuous curve of the expected ENC values, it indicates that the codon bias is only constrained by a G3+ C3 mutational bias [23]. In the figure, more ENC value of genes loci are on the top or far below the curve of the expected ENC values, but a little ENC value of genes lie on or just below the expected curve. It indicates that the codon usage patterns have not only been influenced by mutation pressure, but also mainly influenced by other factors, such as natural selection.

PR2-plot plot analysis.

The relationship between purine(A and G) and primidines (T and C) of partial amino acids of each gene was analyzed by PR2-plot mapping. According to Supplementary Fig. S2, most of the genes are distributed on the high right of the plan, indicating that the frequency of A is higher than T, and the frequency of G is higher than C. If the codon bias of *S. lemnae* is completely affected by random mutation, it shows that A=U and G=C, that is, the use frequency of purine base is equal to that of the pyrimidine base. The use frequency of A differs from that of T, G differ from C indicate that the formation of codon bias is weakly influenced by random mutation, and is strongly influenced by mutation pressure, natural selection, and other factors in *S. lemnae*.

Neutral Plot Analysis

A neutrality plot was constructed to determine the extent of influence between mutation pressure and natural selection by comparing the value of GC 12 and GC 3. When the value of GC 12 is statistically significantly correlated to GC 3 and the slope of the regression line is close to 1 in the neutrality plot, mutation pressure is regarded as the main force forming the codon usage bias. Conversely, if selection is the dominant factor, then the slope of the regression line is close to 0. The analysis show that no correlation is observed between the value of GC 12 and GC 3 ($r = 0.286$, $P > 0.05$) which seemed indicative of mutation pressure playing a little role in codon usage bias of *S. lemnae* genome (Fig. 4).then, after calculating the slope of the regression in the neutrality plot, this was the case. The slope of the regression line was calculated to be 0.2016, highlighting the relative GC 3 (natural selection) is 79.94%, while the relative constraint on neutrality (mutation pressure) is 20.16%. Compared with mutation pressure, natural selection is the dominant factor in shaping the codon usage pattern of *S. lemnae* genes.

Conclusions And Discussion

According to our result, Phylogenetic relationships of *S. lemnae* are similar to *Tetrahymena sp., Paramecium sp.* and *S. cerevisiae*, but are far from *C. elegans* and *H.sapien*. It indicated that the origin and early evolution of many eukaryotes remain unknown or ambiguous, the codon usage pattern is non-universal and influenced by evolutionary processes.

ENC is a simple measure of the degree of codon usage bias. Generally, when the ENC value is less than 35, the codon usage bias is high in a given gene[24].From Supplementary Table S2, the ENC values range from 16.7 to 80.4, with the mean value of 34.21, so the codon usage bias of *S. lemnae* is slightly weak.

Base composition is an important feature of a genome and is the main factor that affects codon usage pattern. The organisms with AT-rich genome, such as *T. thermophila* and *P. tetraurelia*, tend to use A or T at the third position in coding sequence [25,26,27]. Likewise, AT-rich genome in *S. lemnae* preferes A/T as the third condon. However, The The organisms with GC rich are Archea, Bacteria, *Hordeum vulgare*, *Triticum aestivum*, fungi and *Oryza sativa*[28,29,30].In all, mutation pressure is the main factor affecting the codon usage bias of the species[31].

Surprisingly, highly expressed genes with G/C end are significantly more frequent in *Tetrahymena* and *Paramecium*, but it with A/T end is significantly most frequent in *S. lemnae*. Most optimal codons end in T or A, rather than G or C in RSCU analysis (T >A > C >G). PR2 plot shows that T and C are used more frequently than A and G in the third base of synonymous codons in *S. lemnae* complete coding regions. On the contrary, C and T end are preferred in the 3' end of genes in *Tetrahymena* and *Paramecium* [32].In addition, G3S and G 3S are strongly positively correlated with ENC, but are strongly negatively correlated with axis1 and axis2. However, A3S and U 3S are strongly negatively correlated with ENC, but are strongly positively correlated with axis1 and axis2. In general, the high A content and the low G content of *S. lemnae* coding region may be may be the reason of the high A/T content in the third base of synonymous codons, related to the low ENC value, which indicate that the codon bias is relatively high. In another respect, because the A, T, C, G, GC, A3s, T3S, C3S, G3s and GC3s content have a correlation with the two principle axes, which also reveal that mutation pressure from base composition is an important factor shaping the codon usage patterns.

However, according to our PR2 plot, ENC-GC 3S plot, the codon usage patterns of *S. lemnae* have also been influenced by natural selection, such as the hydropathicity and the aromaticity, which plays a major role in shaping the codon usage bias by the result of neutrality plot in *S. lemnae*. From Table 4, AROMO and GRAVY have a negative correlation with Axis1 while which have a positive correlation with Axis1 and Axis2. It firstly indicates the role of hydropathicity and aromaticity forming the codon usage pattern of the cliate.

In short, our analysis revealed that the codon usage bias in *S. lemnae* is low and natural selection such as aromaticity, hydropathicity, is the main factor that affects codon usage variation in *S. lemnae*. Mutation pressure of nucleotide composition is also an important factor influencing codon usage bias. The evolution of *S. lemnae* probably reflects a dynamic process of mutation and natural selection to adapt its codon usage. This study of codon usage patterns in *S. lemnae* is the first reveal information about molecular evolution, and also builds the base for analysis alternative genetic codes of hypotrichous ciliates.

Methods

The complete *S. lemnae* genome is obtained from the *S. lemnae* genome database (http://stylo.ciliate.org/index.php/). A total of 21,061 coding sequences were downloaded. Then we deleted the short sequence whose length is less than 300 bp. Finally, we extracted 20,750 CDSs which had correct start and stop codons with exact multiple of three bases to do data analysis.

Nucleotide composition analysis

The following nucleotide contents of CDS of *S. lemnae* macronucleus genomes were calculated by CodonW software (http://codonw.sourceforge.net), including (i) frequency of occurrence of the nucleotides (A%, C%, U%, G%, GC%); (ii)frequency of each nucleotide at the third position of the synonymous codons (A 3s %, C 3s %, U 3s %, and G 3s %); (iii) frequencies of occurrence of nucleotides G+C at the first (GC 1), second (GC 2), and third base of codon (GC 3); (iv) mean frequencies of nucleotides G+C at the first and second position (GC 1,2).

Relative Synonymous Codon Usage (RSCU) Analysis

RSCU was defined as the ratio of the observed frequency of a specific codon to the expected value, if the each codon of synonymous codons group was used equally[33].The codon with RSCU value is less than 1.0, it means that this codon has relative negative codon usage bias, while the value more than 1.0 has positive codon usage bias; when the RSCU value of codon is close to 1.0, it means that this codon is chosen equally and randomly[34]. In this study, the RSCU values of *S. lemnae* were calculated by CodonW software, the RSCU values for *H.sapiens* were retrieved from Singh R.K.[35], the RSCU values of *C.elegans* were retrieved from Stenico M [36], the RSCU values of *S.cerevisiae* were retrieved from Lloyd A.T [37], and the the RSCU values of *P.caudatum* and *T.thermophila* were retrieved from Barth D and Hannah M.W [38,39], respectively.

Effective number of codons analysis. The ENC was calculated to quantify the codon usage bias of gene and genome level, which is the best estimator of absolute codon usage bias [40,41]. ENC values ranging from 20 to 61 don't require any prior knowledge or a reference set. The value of 20 indicates extreme codon usage bias and the value to 61 indicates no bias. When the ENC value is less than or equal to 35, it is generally believed that the gene has an obvious codon bias.

Determination of optimal codons

Optimal codons definition referred to the analytical method of Liu et al [42]. Gene expression levels were determined by CAI values. The method is as follows: Calculate the CAI of each gene, arrange it of each gene based on CAI values, and take 5% of total genes with the highest CAI valuea to form a high. The RSCU values of each codon in the two data sets is calculated, and then the RSCU values of the high dataset is subtracted that of the low dataset. If difference is greater than 0.08, the corresponding codon is defined as the optimal codon. The software used in this study includes CodonW 1.4.2, SPSS 20 and Excel2016.

ENC-GC_{3S} plot analysis

The ENC- GC_{3S} plot was used to analysis the influence of the GC 3S content on codon usage. The expected ENC value for each GC 3S was calculated using the following formula: $ENC_{expected} = 2 + GC3 + 29/[GC_3^2 + (1 - GC_3)^2]$ [43].

Parity rule 2 analysis

The Parity rule 2 (PR2) plot analysis was used to explore the effects of mutation and natural selection on the codon usage of genes. In this PR2 plot, take the value of AU-bias [A 3s /(A 3s + U 3s)] at the third base of codon as the ordinate and the GC-bias [G 3s / (G 3s+C 3s)] as the abscissa[44].

Neutral Plot Analysis

The neutrality plot was used to examine the extent of the effect of mutation pressure and natural selection on the codon usage patterns by plotting the GC12 values against the GC3 values. In this plot, mutation pressure is assumed to be the main force shaping codon usage when the regression line falls near the diagonal. Alternatively, the regression curve tends to tilt or parallel to the horizontal axis which indicates the dominant role of natural selection on the codon usage bias [21,45].

Principal Component Analysis (PCA).

PCA was used to analyze the major trends in synonymous codon usage of genes mutation. We normalize the data according to the method by Sharp and Li [46]. In each PC, the score of the gth gene (yg) gene was normalized by the mean (m) and standard deviation (SD), expressed as zg = (yg - m) / SD. Highly expressed genes were a zg score greater than 5.17 (theoretically covering only the range of 1.5% of total genes). Then gene expression levels were identified as the major trends in PC score changes in genes. In addition, we analyzed the distribution of PC scores for constitutively overexpressed genes (encoding ribosomal proteins).

Phylogenetic analysis

A phylogenetic tree was constructed based on the nucleotide sequences of the coding regions of *S. lemnae*, using maximum likelihood (ML) method with a bootstrap value of 1000 replicates in CIPRES Science Gateway using RAxML-HPC2 on XSEDE v8.2.10 [47]. Bayesian inference (BI) analysis was carried out with MrBayes v3.2.6 on XSEDE [48] on the CIPRES Science Gateway.

Abbreviations

CUB	Codon usage bias
CDS	coding DNA sequences
RSCU	Relative synonymous codon usage
ENC	effective number of codon
PCA	Principal component analysis
PC	Principal component
SD	standard deviation
CAI	Codon adaptation index
S. lemnae	*Stylonychia lemnae*
H. sapiens	*Homo sapiens*
C. elegans	*Caenorhabditis elegans*
S. cerevisiae	*Saccharomyces cerevisiae*
T. thermophila	*Tetrahymena thermophila*
P. caudatum	*Paramecium caudatum*

Declarations

Ethics approval and consent to participate. Because the organisms studied in this paper are single-celled eukaryotes, they do not involve ethical issues.

Consent for publication. Not applicable" in this section.

Availability of data and materials.

Data source address：http://stylo.ciliate.org/index.php/home/welcome/

Competing interests. The authors declare that they have no competing interests" in this section.

Funding. The workstation used for data processing was financially supported by the National Natural Science Foundation of China (31501893, 31471950) and Natural Science Foundation of Heilongjiang province, China (QC2018021) .

Authors' contributions. YL analyzed and interpreted data and was a major contributor in writing the manuscript. WY and CY carried out previous data processing. JF, XZ and CG interpreted the data . All authors have read and approved the final manuscript.

Acknowledgement

Sequence data was Institute of aquatic biology, Chinese Academy of Sciences provides the transcriptome data of *Stylonychia lemnae*.

References

1. Prescott, D M. "The DNA of ciliated protozoa. " *Microbiol Rev*58.2(1994):233-267.

2. Mu, Weijie , et al. "Epidermal growth factor-induced stimulation of proliferation and gene expression changes in the hypotrichous ciliate, Stylonychia lemnae." *GENE*592.1(2016):186-192.

3. Mehmood, Butt Azeem , et al. "Genome-Wide Analysis of Codon Usage and Influencing Factors in Chikungunya Viruses." *PLoS ONE* 9.3(2014):e90905-.

4. Tourancheau, A. B. , et al. "Genetic code deviations in the ciliates: evidence for multiple and independent events. " *The EMBO Journal* 14.13(1995):3262-3267.

5. Heaphy, Stephen M. , et al. "Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in\r, *C. magnum*." *Molecular Biology and Evolution*(2016):msw166.

6. Chen, Lin , et al. "Analysis of codon usage patterns in Taenia pisiformis through annotated transcriptome data." *Biochemical and Biophysical Research Communications* 430.4(2013):1344-1348.

7. Helftenbein, E. "Nucleotide sequence of a macronuclear DNA molecule coding for alpha-tubulin from the ciliate Stylonychia lemnae. Special codon usage: TAA is not a translation termination codon. " *Nucleic Acids Research* 13.2(1985):415-33.

8. Lekomtsev, S. , et al. "Different modes of stop codon restriction by the Stylonychia and Paramecium eRF1 translation termination factors." *Proceedings of the National Academy of Sciences* 104.26(2007):10824-10829.

9. Lozupone, Catherine A. , R. D. Knight , and L. F. Landweber . "The molecular basis of nuclear genetic code change in ciliates." *Current Biology* 11.2(2001):65-74.

10. Tourancheau, A. B. , et al. "Genetic code deviations in the ciliates: evidence for multiple and independent events. " *The EMBO Journal* 14.13(1995):3262-3267.

11. Martindale, D. W. . "CODON USAGE IN TETRAHYMENA AND OTHER CILIATES." *The Journal of protozoology* 36.1(1989):29-34.

12. Athey, John , et al. "A new and updated resource for codon usage tables." *BMC Bioinformatics* 18.1(2017):391.

13. Wong, Emily Hm , et al. "Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus." *Bmc Evolutionary Biology* 10.1(2010):253-0.

14. Sharp, P. M., et al. "DNA sequence evolution: the sounds of silence. " *Philosophical Transactions of the Royal Society of London*349.1329(1995):241-247.

15. Reis, and M. D. . "Solving the riddle of codon usage preferences: a test for translational selection." *Nucleic Acids Research* 32.17(2004):5036-5044.

16. Middleton, A. R. "Heritable Variations and the Results of Selection in the Fission Rate of Stylonychia pustulata." *Proceedings of the National Academy of Sciences of the United States of America* 19.4(1915).

17. Singh, Niraj K. , et al. "Characterization of codon usage pattern and influencing factors in Japanese Encephalitis Virus." *Virus Research* (2016):S016817021630288X.

18.D'Onofrio, G., Jabbari, K., Musto, H., & Bernardi, G. (1999). The correlation of protein hydropathy with the base composition of coding sequences. *Gene, 238*(1), 3-14.

19. Jabbari, Kamel , et al. "The correlation between GC3 and hydropathy in human genes." *Gene (Amsterdam)*317.none(2003):0-140.

20. Das, Sabyasachi , S. Paul , and C. Dutta . "Synonymous codon usage in adenoviruses: Influence of mutation, selection and protein hydropathy." *Virus Research* 117.2(2006):0-236.

21. Jun Tao , Huipeng Yao..Comprehensive analysis of the codon usage patterns of polyprotein of Zika virus. *Progress in biophysics and molecular biology,*.(2018):05.001.

22. Mochizuki, Kazufumi."High efficiency transformation of Tetrahymena using a codon-optimized neomycin resistance gene." *Gene* 425.1-2(2008):0-83.

23. Mehmood, Butt Azeem , et al. "Genome-Wide Analysis of Codon Usage and Influencing Factors in Chikungunya Viruses." *PLoS ONE* 9.3(2014):e90905.

24. Liu, and Xiong'en. "A more accurate relationship between 'effective number of codons' and GC3s under assumptions of no selection." *Computational Biology and Chemistry*42.Complete(2013):35-39.

25."Macronuclear Genome Sequence of the Ciliate Tetrahymena thermophila, a Model Eukaryote." *PLoS Biology* 4.9(2006):e286.

26. Aury, Jean Marc , et al. "Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia." *NATURE* 444.7116(2006):171-178.

27. Salim, Hannah M. W. , K. L. Ring , and A. R. O. Cavalcanti . "Patterns of Codon Usage in two Ciliates that Reassign the Genetic Code: Tetrahymena thermophila and Paramecium tetraurelia." *Protist* 159.2(2008):283-298.

28. Ruth, Hershberg , D. A. Petrov , and M. W. Nachman . "General Rules for Optimal Codon Choice." *PLoS Genetics*5.7(2009):e1000556.

29.Kawabe, Akira , and N. T. Miyashita . "Patterns of codon usage bias in three dicot and four monocot plant species." *Genes & Genetic Systems* 78.5(2003):343-352.

30. Yang, Xing , X. Luo , and X. Cai . "Analysis of codon usage pattern in Taenia saginata based on a transcriptome dataset." *Parasites & Vectors* 7.1(2014):527.

31. Sharp, P. M. . "An evolutionary perspective on synopnymous codon usage in unicellular organisms." *J. Mol. Evol.* 24(1986).

32. Salim, Hannah M. W. , K. L. Ring , and A. R. O. Cavalcanti . "Patterns of Codon Usage in two Ciliates that Reassign the Genetic Code: Tetrahymena thermophila and Paramecium tetraurelia." *Protist* 159.2(2008):283-298.

33.Sharp, P. M. . "An evolutionary perspective on synopnymous codon usage in unicellular organisms." *J. Mol. Evol.* 24(1986).

34. "Analysis of amino acid and codon usage in Paramecium bursaria." *Febs Letters* 589.20(2015):3113–3118.

35. Singh, R. K. , and S. P. Pandey . "Phylogenetic and Evolutionary Analysis of Plant ARGONAUTES. " (2017).

36. Stenico, M. , A. T. Lloyd , and P. M. Sharp . "Codon usage inCaenorhabditis elegans: delineation of translational selection and mutational biases." *Nucleic acids research* 22.13(1994).

37. Lloyd, Andrew T. , and P. M. Sharp . "Evolution of codon usage patterns: The extent and nature of divergence between Candida albicans and Saccharomyces cerevisiae." *Nucleic Acids Research* 20.20(1992):5289-5295.

38. Barth, Dana , and T. U. Berendonk . "The mitochondrial genome sequence of the ciliate Paramecium caudatum reveals a shift in nucleotide composition and codon usage within the genus Paramecium." *BMC Genomics* 12.1(2011):272.

39. Salim, Hannah M. W. , K. L. Ring , and A. R. O. Cavalcanti . "Patterns of Codon Usage in two Ciliates that Reassign the Genetic Code: Tetrahymena thermophila and Paramecium tetraurelia." *Protist* 159.2(2008):283-298.

40. Nasrullah, Izza, Butt, Azeem M., Tahir, Shifa, Idrees, Muhammad, Tong, Yigang, .

Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. *BMC Evol. Biol.* (2015):15, 174.

41. Wright, F. . The 'effective number of codons' used in a gene. *Gene,*(1990):*87*(1), 23-29.

42. Booth, L., Wolfe, B., & Doerder, F. P. . Molecular polymorphism in the mta and mtb mating type genes of tetrahymena thermophila and related asexual species. *Journal of Eukaryotic Microbiology,*(2015): *62*(6), 750-761.

43. Moradian, M. M. , Beglaryan, D. , Skozylas, J. M. , & Kerikorian, V. .. Complete mitochondrial genome sequence of three tetrahymena species reveals mutation hot spots and accelerated nonsynonymous substitutions in ymf genes. *PLOS ONE, 2* ,(2007).

44.Sueoka, N.. Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the dna g+c content of third codon position. *Gene (Amsterdam),* (1999):*238*(1), 0-58.

45. Sueoka, N. . Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America,* (1988):*85*(8), 2653-2657.

46. Sharp, P. M., & Li, W. H. . The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research,* (1987):*15*(3), 1281-95.

47. Stamatakis, A. . Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* (2014):*30*(9), 1312-1313.

48. Ronquist, F. , Teslenko, M. , Van, d. M. P. , Ayres, D. L. , Darling, A. , & Hohna, S. , et al. . Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology,* (2012):*61*(3), 539.

Tables

Table 1 The relative synonymous codon usage frequency (RSCU) of *S. lemnae* and other model organism

Animo acid	Codon	RSCU						Animo acid	Codon	RSCU					
		S.lemnae	*H. sapiens.*	*C. elegans*	*S.cerevisiae*	*P. caudatum*	*T. thermophila*			*S.lemnae*	*H. sapiens*	*C. elegans*	*S.cerevisiae*	*P. caudatum*	*T. thermophila*
Phe	UUU	1.17	0.94	0.72	1.08	1.30	1.50	Ser	UCU	1.48	1.12	1.47	1.83	1.45	1.71
	UUC	0.83	1.06	1.28	0.92	0.70	0.50		UCC	0.42	1.28	0.98	1.09	0.41	0.21
Leu	UUA	1.55	0.47	0.45	1.60	2.18	2.34		UCA	1.76	0.91	1.44	1.16	2.07	0.86
	UUG	1.06	0.78	1.35	2.08	0.91	1.32		UCG	0.22	0.33	0.83	0.51		0.21
	CUU	1.39	0.80	1.86	0.66	1.82	0,92	Pro	CCU	1.58	1.14	0.52	1.18		1.67
	CUC	0.57	1.15	1.38	0.28		0.41		CCC	0.43	1.29	0.23	0.54	2.00	0.33
	CUA	0.99	0.43	0.34	0.79	0.91	0.81		CCA	1.82	1.10	2.75	1.88	2.00	2.00
	CUG	0.44	2.36	0.63	0.59	0.18	0.20		CCG	0.17	0.47	0.51	0.39		
Ile	AUU	1.4	1.10	1.52	1.47	0.93	1.35	Thr	ACU	1.95	1.01	1.34	1.50	1.54	2.00
	AUC	0.63	1.38	1.23	0.89	0.53	0.31		ACC	0.54	1.39	1.02	0.97	0.62	0.38
	AUA	0.96	0.52	0.25	0.63	1.53	1.35		ACA	1.35	1.15	1.15	1.06	1.54	1.63
Met	AUG	1			1.00				ACG	0.16	0.46	0.49	0.47	0.31	
Val	GUU	1.53	0.89	1.67	1.73	1.67	2.13	Ala	GCU	1.86	1.05	1.64	1.73	1.71	2.25
	GUC	0.67	1.06	1.11	0.96	0.67	0.50		GCC	0.55	1.59	1.06	0.97	0.57	
	GUA	1.19	1.60	0.52	0.66	1.00	0.88		GCA	1.43	0.92	0.99	0.95	1.14	1.50
	GUG	0.62	0.44	0.70	0.65	0.67	0.50		GCG	0.16	0.44	0.31	0.35	0.57	0.25
Tyr	UAU	1.31	0.90	0.97	1.02	1.33	1.61	Cys	UGU	0.94	0.94	1.14	1.86	1.67	1.14
	UAC	0.69	1.10	1.03	0.98	0.67	0.39		UGC	1.06	1.06	0.86	0.14	0.33	0.86
Gln	UAA	1.55		1.55				Trp	UGG	1.00			1.00		
	UAG	0.87		0.87				Glu	GAA	1.24	0.86	1.15	1.58	1.89	1.60
Lys	AAA	1.17	0.88	0.84	1.05		1.24		GAG	0.76	1.14	0.85	0.18	0.11	0.40
	AAG	0.83	1.12	1.16	0.95	0.29	0.76	Arg	AGA	4.4	1.29	1.79	4.24	3.00	3.75
Asp	GAU	1.51		1.36	1.46	1.64	1.56		AGG	0.95	1.27	0.26	0.23	2.00	
	GAC	0.49		0.64	0.54	0.36	0.44		CGU	0.19	0.48	1.84	0.97	1.00	0.75
Gly	GGU	1.15	0.64		2.84	1.00	0.47		CGC	0.12	1.10	0.73	0.04		0.75
	GGC	0.69	1.35		0.24		0.47		CGA	0.29	0.65	1.07	0.42		0.75
	GGA	1.76	1.01		0.61	3.00	3.06		CGG	0.05	1.22	0.31	0.10		
	GGG	0.4	1.00		0.31			Asn	AAU	1.42	0.96	1.10	1.11	1.77	1.43
Gln	CAA	1.08	0.54		1.46		1.00		AAC	0.58	1.06	0.90	0.89	0.23	0.57
	CAG	0.5	1.46		0.54		1.00	Ser	AGU	1.32	0.91	0.76	0.95	1.66	1.61
His	CAU	1.44	0.85		1.20	1.60	2.00		AGC	0.8	1.44	0.52	0.30	0.41	1.39
	CAC	0.56	1.15		0.80	0.40									
Ter	UGA	1.00		0.59											
	UAA			1.56											
	UAG			0.60											

[a] The 'RSCU' value represents the pattern of relative synonymous codon usage

Table 2 Preferred codons in genome of *S. lemnae*

Animo acid	Codon	RSCU[a] (high)	RSCU (low)	Animo acid	Codon	RSCU (high)	RSCU (low)	Animo acid	Codon	RSCU (high)	RSCU (low)
Phe	UUU	1.5	0.66	Ser	UCU	1.62	0.74	Arg	AGA	4.64	0.92
	UUC	0.5	1.34		UCC	0.18	1.29		AGG	0.68	1.18
Leu	UUA	2.41	0.51		UCA	2.09	0.7	Gly	GGU	0.98	1.19
	UUG	0.96	1.14		UCG	0.04	0.7		GGC	0.36	1.95
	CUU	1.5	0.51	Pro	CCU	1.28	0.82		GGA	2.52	0.39
	CUC	0.19	0.93		CCC	0.27	0.62		GGG	0.15	0.47
	CUA	0.82	0.34		CCA	2.41	1.13	Lys	AAA	1.37	1.3
	CUG	0.12	2.58		CCG	0.04	1.44		AAG	0.63	0.7
Ile	AUU	1.61	0.66	Thr	ACU	2.55	0.49	Asp	GAU	1.85	0.79
	AUC	0.41	1.99		ACC	0.29	1.6		GAC	0.15	1.21
	AUA	0.99	0.35		ACA	1.13	1.2	Glu	GAA	1.55	1.22
Met	AUG	1	1		ACG	0.03	0.71		GAG	0.45	0.78
Val	GUU	2.24	1.03	Ala	GCU	2.27	0.9	Ser	AGU	1.43	0.64
	GUC	0.31	0.98		GCC	0.25	1.32		AGC	0.64	1.93
	GUA	1.01	0.83		GCA	1.45	1.06	Arg	CGU	0.18	1.13
	GUG	0.44	1.16		GCG	0.03	0.72		CGC	0	1.36
Tyr	UAU	1.68	0.64	Cys	UGU	0.91	0.58		CGA	0.45	0.37
	UAC	0.32	1.36		UGC	1.09	1.42		CGG	0.05	1.05
Gln	UAA	2.16	0.32	TER	UGA	1	1	Asn	AAU	1.71	0.53
	UAG	0.64	0.37	Trp	UGG	1	1		AAC	0.29	1.47
His	CAU	1.76	0.84	Gln	CAA	1.02	1.22				
	CAC	0.24	1.16		CAG	0.18	2.09				

[a] The 'RSCU' value represents the pattern of relative synonymous codon usage

Table 3 Summary of correlation analysis of nucleotide composition and ENC

	A	C	T	G	GC	A3s	C3s	T3s	G3s	GC3s
%A	1.000**	-.583**	-.408**	-.331**	-.619**	.640**	-.568**	-.156**	.166**	-.346**
%C	-.583**	1.000**	-.238**	.115**	.771**	-.544**	.820**	-.313**	-.054**	.616**
%T	-.408**	-.238**	1.000**	-.466**	-.465**	-.069**	-.186**	.688**	-.480**	-.448**
%G	-.331**	.115**	-.466**	1.000**	.722**	-.283**	.218**	-.348**	.451**	.455**
%G+C	-.619**	.771**	-.465**	.722**	1.000**	-.561**	.711**	-.441**	.252**	.721**
ENC	.005	.101**	-.105**	.024**	.086**	-.232**	.106**	-.210**	.374**	.318**

**. Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level

Table 4 Summary of correlation between the first two axes and nucleotide constraints in

S.lemnae genome

Base composition	Axis1	Axis2
A	.075**	-.012
C	-.061**	-.051**
T	-.008	.079**
G	-.036**	-.032**
GC	-.066**	-.056**
A3s	.051**	.028**
C3s	-.060**	-.048**
T3s	.004	.068**
G3s	.014	-.050**
GC3s	-.039**	-.069**
ENC	-.045**	-.065**

**. Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level

Table 5 Correlation analysis among AROMO, GRAVY, the first two axes, GC 3s, ENC and GC in the polyprotein-coding region of *S. lemnae*

		Axis1	Axis2	ENc	GC3s	GC
Gravy	r	-.576**	.547**	-.009	-.001	.010
	p	.000	.000	.218	.850	.155
Aromo	r	-.409**	.484**	.000	-.004	-.005
	p	.000	.000	.968	.602	.452

**. Correlation is significant at the 0.01 level

*. Correlation is significant at the 0.05 level

Figures

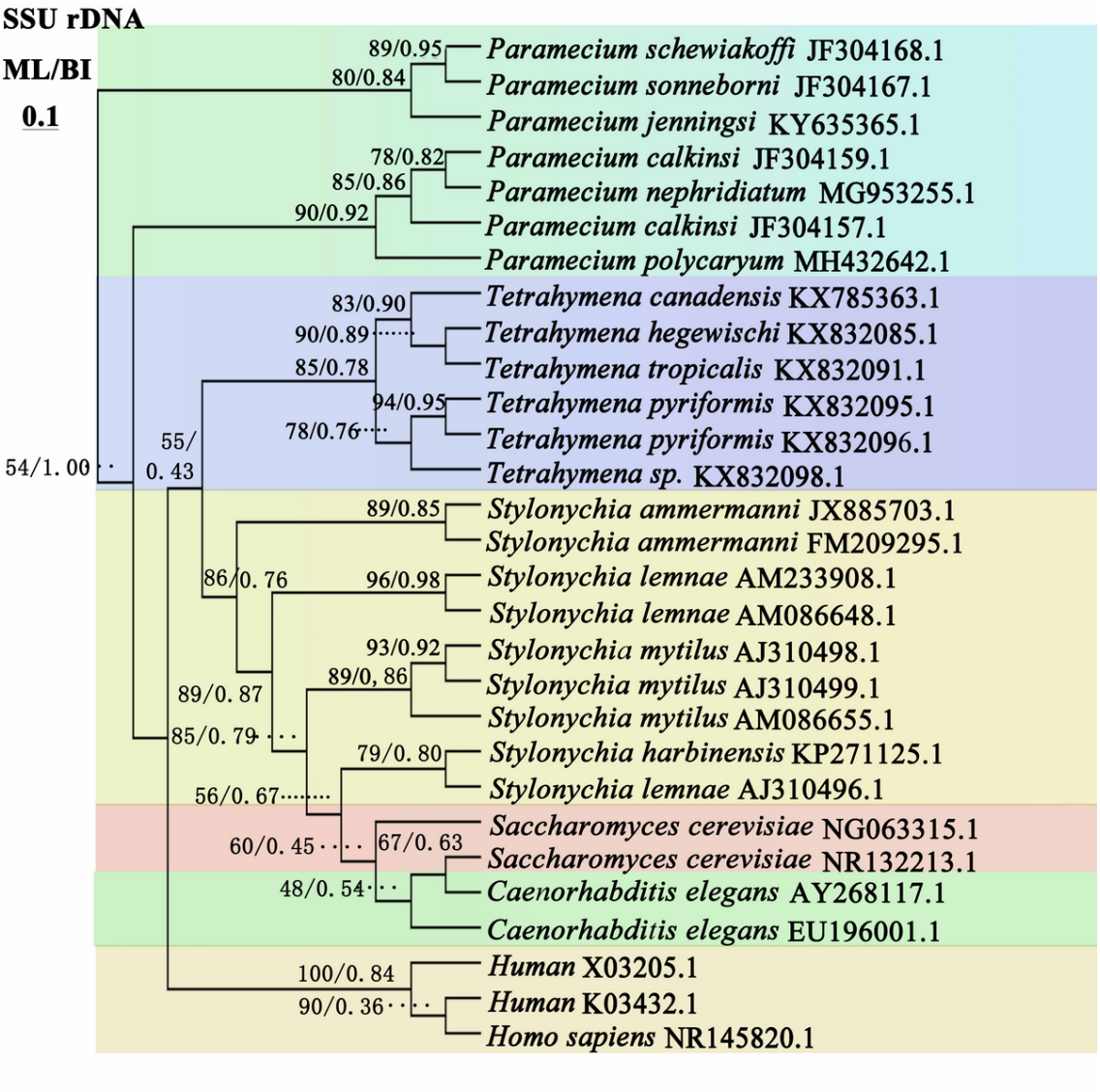
SSU rDNA
ML/BI
0.1
89/0.95
80/0.84
Paramecium schewiakoffi JF304168.1
Paramecium sonneborni JF304167.1
Paramecium jenningsi KY635365.1
78/0.82
85/0.86
90/0.92
Paramecium calkinsi JF304159.1
Paramecium nephridiatum MG953255.1
Paramecium calkinsi JF304157.1
Paramecium polycaryum MH432642.1
83/0.90
90/0.89
85/0.78
94/0.95
78/0.76
Tetrahymena canadensis KX785363.1
Tetrahymena hegewischi KX832085.1
Tetrahymena tropicalis KX832091.1
Tetrahymena pyriformis KX832095.1
Tetrahymena pyriformis KX832096.1
Tetrahymena sp. KX832098.1
54/1. 00
55/
0. 43
89/0.85
Stylonychia ammermanni JX885703.1
Stylonychia ammermanni FM209295.1
86/0. 76
96/0.98
Stylonychia lemnae AM233908.1
Stylonychia lemnae AM086648.1
93/0.92
89/0, 86
89/0. 87
Stylonychia mytilus AJ310498.1
Stylonychia mytilus AJ310499.1
Stylonychia mytilus AM086655.1
85/0. 79
79/0. 80
Stylonychia harbinensis KP271125.1
Stylonychia lemnae AJ310496.1
56/0. 67
Saccharomyces cerevisiae NG063315.1
60/0. 45
67/0. 63
Saccharomyces cerevisiae NR132213.1
48/0. 54
Caenorhabditis elegans AY268117.1
Caenorhabditis elegans EU196001.1
100/0. 84
90/0. 36
Human X03205.1
Human K03432.1
Homo sapiens NR145820.1

Figure 1

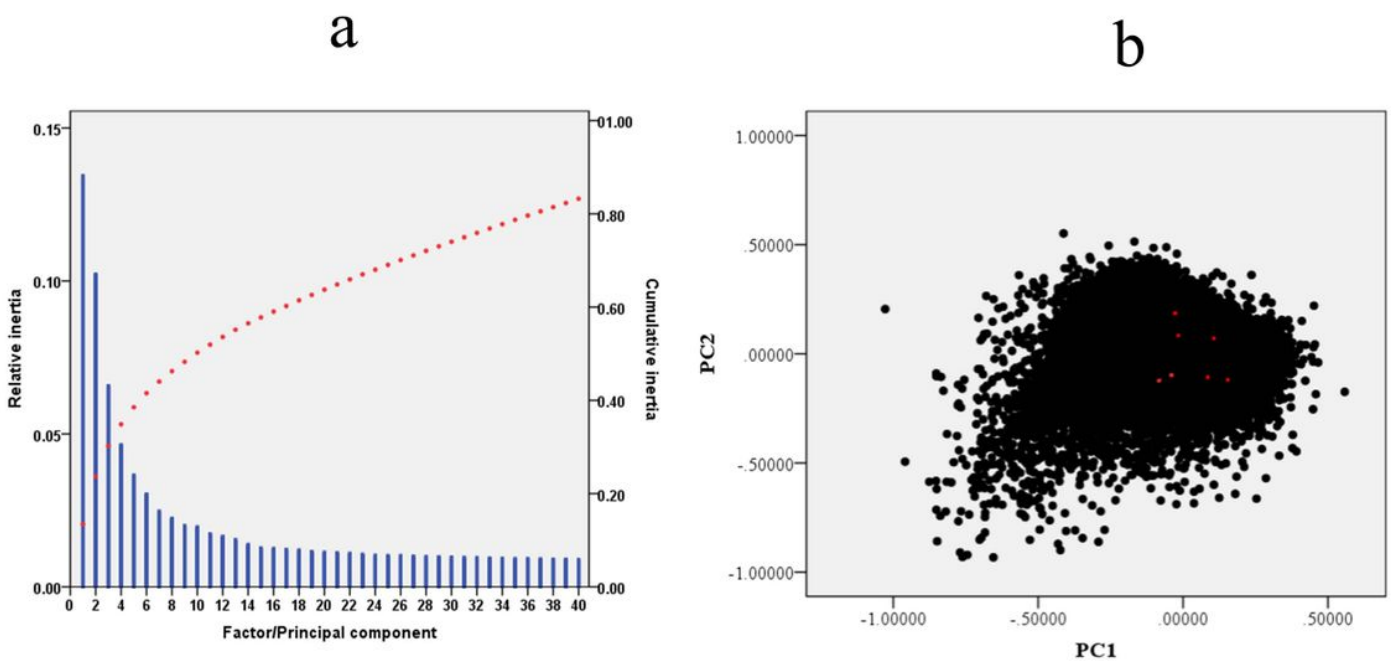

Figure 2

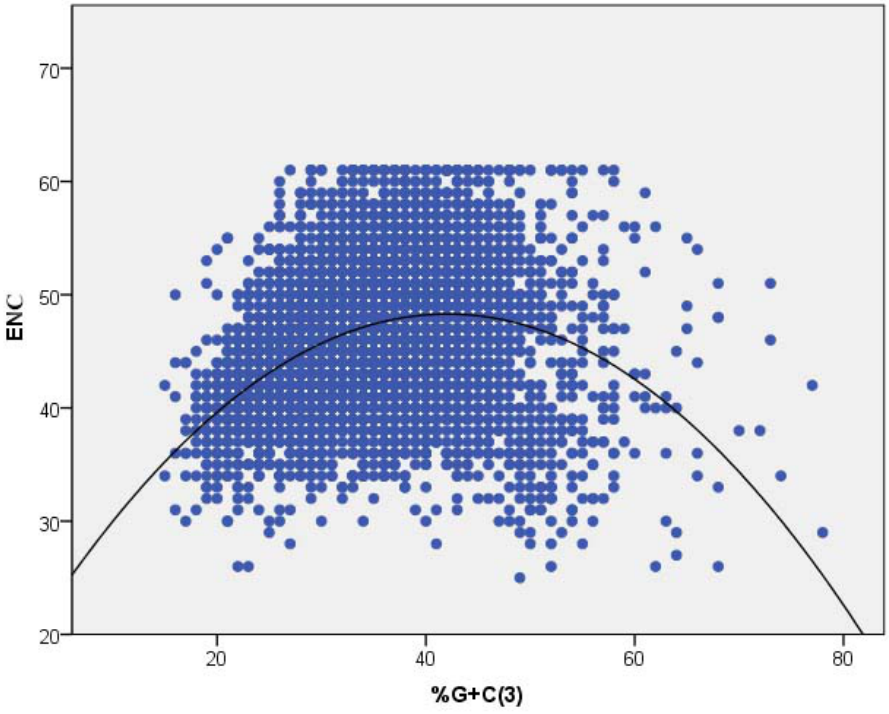

Figure 3

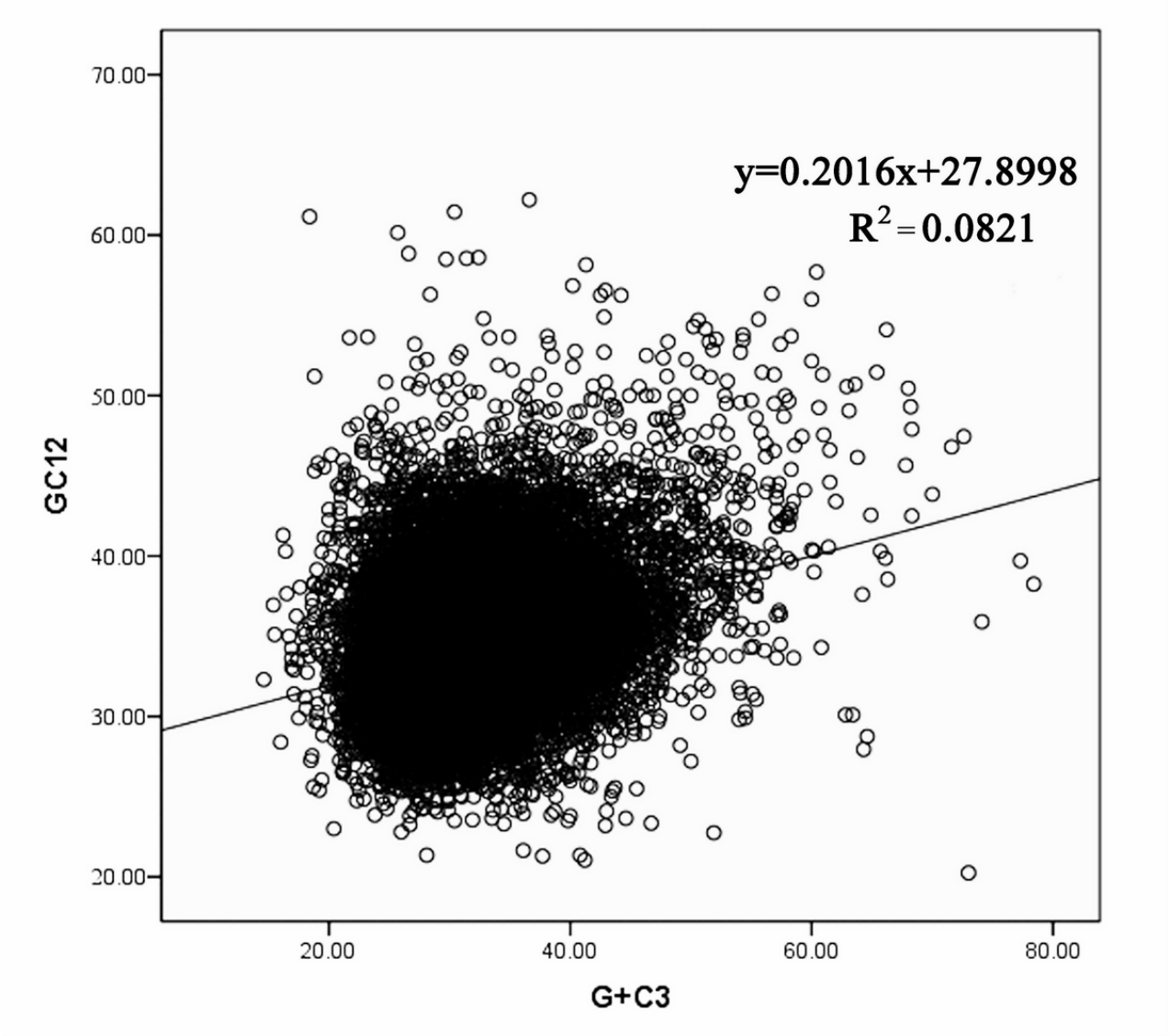

Figure 4

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Fig.S2.png
- supplementarytableS2.doc
- Fig.S1.png