Comparison between synonymous transition/transversion with non-synonymous transition/transversion reveals different purifying selection on coding sequences

Pratyush Kumar Beura[1], Ruksana Aziz[1], Piyali Sen[2], Saurav Das[1], Nima Dondu Namsa[1,3], Edward J Feil[4], Siddharatha Sankar Satapathy*[2,3], Suvendra Kumar Ray*[1,3]

Department of [1]Molecular Biology and Biotechnology, [2]Computer Science and Engineering, [3]Centre for Multidisciplinary Research, Tezpur University, Tezpur, Assam, India, 784028, [4]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, United Kingdom

For correspondence*

Siddhartha Shankar Satapathy: ssankar@tezu.ernet.in

Suvendra Kumar Ray: suven@tezu.ernet.in

ABSTRACT

Base substitution is a major event in genome variation. Therefore, a correct estimation of the base substitution event is important to understand the process of selection. Here, we developed a method (link) to estimate $\frac{ti}{tv}$ ratio by accounting codon degeneracy as well as pretermination codons in a coding sequence. A distinct observation of the influence of codon degeneracy and pretermination codon on $\frac{ti}{tv}$ was observed in 2516 genes of *Escherichia coli* analyzed in this study. The mean Sti, Stv, Nti and Ntv values revealed the decreasing order of their variation dynamics, respectively in *E. coli* genome. While a strong positive correlation was observed

between Sti, and Stv values, the correlation between Sti, and Nti was poor in coding sequences. In several instances coding sequences with same Sti values were found to have different Nti values. This indicated that the varying strength of purifying selection on Nti in genes is responsible for the poor correlation above. Our approach is convenient to visualize the base substitution variation dynamics as well as selection in coding sequences. Also, the proposed method is useful to estimate different $\frac{ti}{tv}$ ratios accurately in coding sequences and is insightful from evolutionary perspectives.

INTRODUCTION

The genetic code table is a beautiful gift of nature to study molecular evolution (1,2). The different degenerate codons influence the transition (*ti*) to transversion (*tv*) ratio ($\frac{ti}{tv}$) as follows: synonymous variation in a *two-fold degenerate* (TFD) codon occurs *via* only one transition whereas the same in a *four-fold degenerate* (FFD) codon occurs via one transition and two transversions (Supplementary Figure S1A). In addition to the above, compositional variation of pretermination codons (PTC) (Supplementary Figure S1B and S1C) (3) among the coding sequences is also likely to influence the $\frac{ti}{tv}$ ratio because non-synonymous variations leading to termination codons purged out rapidly in a population due to strong purifying selection (4-6). Considering in general, a *ti* being more frequent than a *tv* (7-13) which is known as Kimura's two parameter model (14), and a stronger purifying selection on non-synonymous variation than a synonymous variation (15), the $\frac{ti}{tv}$ ratios of TFD codons and FFD codons are expected to be different. Accordingly, it has been recently reported that the synonymous variation in TFD codons is relatively higher than the FFD codons in *Escherichia coli* (16). Therefore, in any

organism, the $\frac{ti}{tv}$ ratio is likely to vary across the coding sequences having differences in degenerate codon composition. Hence, a comparative study of $\frac{ti}{tv}$ among coding sequences in a genome will be always insightful.

Considering the number of possibilities of the two types of base variations in genomic DNA (17) (Figure 1). The ratio of *ti* to *tv* is theoretically expected to be 0.5. But the ratio is generally observed more than 0.5 in the genomes (7,8). Because of the following factors such as geometry of DNA double helix (18), modification of DNA bases such as deamination of cytosine and adenine (19-21), secondary structure in the RNA (9) and codon degeneracy are known to elevate *ti* phenomenon in a genome over *tv* (16,22). In a standard genetic code table out of 549 total single nucleotide variations (SNVs) involving 61 sense codons (9 SNVs per codon), only 134 SNVs are synonymous (24.4%). Among synonymous (*Sti* and *Stv*) and non-synonymous (*Nti* and *Ntv*) variations, the theoretically estimated ($_e$) numbers of *Sti*, *Stv*, *Nti* and *Ntv* in a genetic codon table are 62, 72, 121 and 294, respectively (see Supplementary Table S1). Yet the observed synonymous variations are always higher than the expected proportion across genes (23-26) due to stronger purifying selection on non-synonymous variations than the synonymous variations (15). Hence, by accounting the codon degeneracy as well as PTC, the *Sti*, *Stv*, *Nti* and *Ntv* in the genetic codon table can be estimated theoretically for each codon (see Supplementary Table S2).

The codon substitution model suggests that sites within codons evolve at different rates and consequently they should not be equally treated (22,27). In the codon substitution model, using likelihood approach the $\frac{ti}{tv}$ ratio was estimated in accordance with codon degeneracy class which was implemented in estimating *dN*/*dS* in coding sequences (22,28,29). Recently, by

accounting $\frac{ti}{tv}$ substitution rate difference in codons as well as nonsense variations in PTC, the *dN*/*dS* values have been improved (16). This finding implies that factors such as codon degeneracy and PTC composition should be considered while analyzing the evolution of protein-coding genes. The ratio of *ti* to *tv* has been studied for reliable estimation of sequence distance and phylogeny reconstruction (30). Using maximum likelihood approach, the $\frac{ti}{tv}$ ratio (commonly known as kappa κ) has been used to mainly compare the $\frac{ti}{tv}$ values among different species (30). The parameter κ has not been implemented to compare coding sequences within a genome in the context of their compositional differences of different degenerate classes before. As the selective pressure on synonymous and non-synonymous variations are heterogeneous (31), it is important to analyze the $\frac{ti}{tv}$ values among different genes within an organism by accounting synonymous sites and non-synonymous sites regarding their compositional differences of different degenerate codons.

Given that there is a significant variation of substitution rates among protein coding sequences within a genome (32), an appropriate methodology for the estimation of *ti* to *tv* ratio is required. We are presenting an improved equation that builds upon the research conducted by Yasuo Ina in 1998 (33). Considering the differential selective constraint in synonymous and non-synonymous variations it is crucial to separately analyze the synonymous $\frac{ti}{tv}$ and non-synonymous $\frac{ti}{tv}$ to gain a comprehensive understanding of the overall $\frac{ti}{tv}$. The improved equations used in this study are helpful to estimate the *ti* and *tv* rate differences in coding sequences by accounting the compositional difference in terms of codon degeneracy and PTC. Though it is known that TFD codons only undergo synonymous *ti*, the $\frac{Sti}{Stv}$ ratio among the coding sequences

in relation to their compositional differences (TFD, FFD and PTC) has not been reported previously. Further, we observed a correlation between TFD: FFD and $\frac{Sti}{Stv}$ values and the PTC% and $\frac{Nti}{Ntv}$ values. We tried to establish the biological significance of codon composition and $\frac{ti}{tv}$ ratio across 2516 coding sequences of *E. coli*. Interestingly, coding sequences with similar synonymous variation dynamics are found to have a profound variation regarding their non-synonymous variation dynamics. The improved method developed in the present study may be applied to study different $\frac{ti}{tv}$ ($\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$) across coding sequences.

MATERIALS AND METHODS

Coding sequences of *E. coli* genome

In this study, we carried out a SNVs based computational analysis of 2516 coding sequences across 157 strains of *E. coli* (34). The information regarding the dataset selected for the present study is provided in Supplementary Table S3. Although the public database has ~4500 coding sequences, we excluded coding sequences showing length variation and improper alignment/annotation. Codons having variations in more than one position or variations leading to ambiguous nucleotides were not considered in the analyses of the variations in this study. A detailed procedure for estimating variation is represented in (see Supplementary Table S4) using example of a hypothetical sequence. This procedure is based on intra-species genome sequence comparison to find out single nucleotide variations in bacterial genomes. A reference sequence has been generated considering the most frequent nucleotide occurrence at a site in the gene sequence. The logic is that the most abundant nucleotide in a species is the ancestral nucleotide in the species (9,16,35,36).The observed SNVs (SNV_o) at different positions of each codon were categorized into Sti_o, Stv_o, Nti_o and Ntv_o, where S for synonymous, N for non-synonymous, ti_o for

observed transition and tv_o for observed transversion, and subsequently ti_o ($Sti_o + Nti_o$) and tv_o ($Stv_o + Ntv_o$) for each gene were calculated (see Supplementary Table S5). After finding observed variation values for a gene, we calculated the theoretically estimated variation values Sti_e, Stv_e Nti_e, Ntv_e, ti_e and tv_e. We derived the reference sequence for each gene based on the methodology explained (see Supplementary Table S3). The reference sequences were further used to find the codon counts of each gene. Then we multiplied the count of each codon by the theoretically estimated variation values for each codon. For example, if the codon frequency of GGG is observed to be 10 in a gene, then the total theoretically estimated value for *Sti*, *Stv*, *Nti* and *Ntv* are 10, 20, 20 and 40 (as estimated *Sti*, *Stv*, *Nti* and *Ntv* for each GGG is 1, 2, 2 and 4, respectively). We then summed up the individual variation values for 61 sense codons to get the total theoretically estimated individual substitution values (Sti_e, Stv_e, Nti_e and Ntv_e) for the gene (see Supplementary Table S6). The values of Sti_o, Stv_o, Nti_o, Ntv_o, ti_o, and tv_o and the values of Sti_e, Stv_e Nti_e, Ntv_e, ti_e, and tv_e of a coding sequence are then applied in the modified equation to find out different $\frac{ti}{tv}$ ratio (see Supplementary Table S7). A detailed workflow is provided in Figure 2.

Improved equations

We proposed an improved mathematical equation for the calculation of different $\frac{ti}{tv}$ ratio as described below.

$$\frac{ti'}{tv'} = \frac{\left({ti_o}/{ti_e}\right)}{\left({tv_o}/{tv_e}\right)} \qquad \textbf{Equation 1}$$

$$\frac{Sti'}{Stv'} = \frac{\left({Sti_o}/{Sti_e}\right)}{\left({Stv_o}/{Stv_e}\right)} \qquad \textbf{Equation 2}$$

$$\frac{Nti'}{Ntv'} = \frac{\left({Nti_o}/{Nti_e}\right)}{\left({Ntv_o}/{Ntv_e}\right)} \qquad \textbf{Equation 3}$$

Where, $\frac{ti'}{tv'}$ is improved $\frac{ti}{tv}$ ratio, $\frac{Sti'}{Stv'}$ is improved $\frac{Sti}{Stv}$ ratio, $\frac{Nti'}{Ntv'}$ is improved $\frac{Nti}{Ntv}$ ratio, ti_o is transition observed, ti_e is the theoretically estimated transition, tv_o is transversion observed, tv_e is transversion estimated, Sti_o is synonymous transition observed, Sti_e is the theoretically estimated synonymous transition, Stv_o is synonymous transversion observed, Stv_e is the theoretically estimated synonymous transversion, Nti_o is non-synonymous transition observed, Nti_e is the theoretically estimated non-synonymous transition, Ntv_o is non-synonymous transversion observed, Ntv_e is the theoretically estimated non-synonymous transversion. We calculated $\frac{ti}{tv}$ ratios for all the genes using a program written in the C language. (to be modified later)

It is noteworthy that in a genome the number of potential replacements is greater for non-synonymous alterations, whereas the frequency of actual variations is higher for synonymous alterations among coding sequences. The improved estimators $\frac{ti'}{tv'}$, $\frac{Sti'}{Stv'}$, and $\frac{Nti'}{Ntv'}$ account for the observed variations out of the total possible variations, on the contrary, the conventional method only takes an account of the ti_o and tv_o. The possible sites for *Sti* and *Stv* are also different among genes, which is dependent upon the codon composition belonging to different degeneracy classes. The conventional $\frac{ti}{tv}$ ratio does not adequately account for the limitations of variable selection against synonymous and non-synonymous variations. Hence, all these parameters make the requirement of an improved estimator instead of the conventional estimator.

Comparative representation of different $\frac{ti}{tv}$ through conventional and improved method

We have done a comparative analysis of $\frac{ti}{tv}$ calculated using the existing and the improved method to support the validation of method proposed in this study. Also, we have calculated the TFD: FFD and the percentage of PTC (PTC%) of each gene and performed statistical analysis between different parameters. A summary of the different parameters observed throughout the dataset is represented in Table 1. We also calculated the percentage (%) change in $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ values obtained from conventional to improved method as follows.

(a) $$\frac{Sti}{Stv}\ \%\ \text{change} = \frac{\frac{Sti}{Stv} - \frac{Sti'}{Stv'}}{\frac{Sti}{Stv}} \times 100$$

(b) $$\frac{Nti}{Ntv}\ \%\ \text{change} = \frac{\frac{Nti}{Ntv} - \frac{Nti'}{Ntv'}}{\frac{Nti}{Ntv}} \times 100$$

We found out $\frac{ti}{tv}$ in all these 2516 genes using maximum likelihood (ML) approach available in MEGA-X (37). Using our improved approach, we computed the normalized *Sti*, *Stv*, *Nti*, and *Ntv* values for the dataset. These normalized values were determined by dividing the observed values by the corresponding estimated values in the dataset and further analyzed in the variation dynamics study.

Statistical analyses

Comparison among the conventional, ML approach and the improved $\frac{ti}{tv}$ values were done using different statistical methods to demonstrate the performance and accuracy of the improved method. OriginPro, Version 2022, OriginLab Corporation, Northampton, MA, USA was used to draw the Box-plot as well as to perform the Mann-Whitney test (38) to find out the *p*-value. The

correlation plot was drawn for R^2 value and Pearson's correlation coefficient (39) was also calculated between the $\frac{ti}{tv}$ values obtained using the conventional and improved method.

Software and programme

Programme written in Python-language is available at GitHub in the following link. (). Further queries regarding computer programme may contact SSS (ssankar@tezu.ernet.in).

RESULTS

Impact of codon degeneracy and PTC on $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ values in *E. coli*

We found out the total number of TFD, FFD, and PTC codons across 2516 coding sequences of *E. coli*. The percentage of TFD, FFD, and PTC was calculated (see Supplementary Table S3). The percentage (%) compositional differences in the 2516 coding sequences represented as TFD%, FFD%, and PTC% is shown in a box-plot (Figure 3). A variable range of TFD: FFD ratio in the coding sequences of *E. coli* was observed. The minimum to maximum TFD:FFD ratio was observed as 0.229 and 3.333 in *sugE* and *mntS*, respectively. It suggests that FFD codons are present approximately four times more frequent than TFD codons in certain coding sequences. Similarly, TFD codons are present approximately three times more frequent than FFD codons in the selected coding sequences. Among 2516 coding sequences, only 35 coding sequences possessed the TFD:FFD ratio of 1.000. Hence, this observation suggests that there is a compositional asymmetry of TFD and FFD codons in the coding sequences. The minimum to maximum PTC% was observed in the gene *atpE* as 15.00% and in the gene *chaB* as 57.14% (Figure 3). The compositional variation

(TFD, FFD and PTC) observed in coding sequences were considered to validate the accuracy of different $\frac{ti}{tv}$ ratio using the conventional and improved method.

We estimated $\frac{ti}{tv}$ by accounting codon degeneracy as well as PTC in coding sequences and compared with the conventional method of $\frac{ti}{tv}$ in coding sequences. Theoretically, a coding region entirely composed of FFD codons will result into a twice $\frac{Sti'}{Stv'}$ value than $\frac{Sti}{Stv}$ value, similarly a coding region composed of same proportion of FFD and TFD codons will result into equal values of $\frac{Sti'}{Stv'}$ and $\frac{Sti}{Stv}$**.** The increase in TFD codon proportion will directly impact the difference between $\frac{Sti}{Stv}$ values and $\frac{Sti'}{Stv'}$ values. We have observed a similar pattern as well (Supplementary Figure S2). The complete set of observed and estimated values of different variations is enlisted (see Supplementary Table S3). These values were used to calculate the $\frac{ti}{tv}, \frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$. Further the improved equation was used to calculate the $\frac{ti'}{tv'}, \frac{Sti'}{Stv'}$ and $\frac{Nti'}{Ntv'}$. We compared the values derived using the conventional method ($\frac{ti}{tv}$, $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$) and improved method ($\frac{ti'}{tv'}$, $\frac{Sti'}{Stv'}$ and $\frac{Nti'}{Ntv'}$) (see Supplementary Table S3). A box-plot showing the comparison between the values of different $\frac{ti}{tv}$ ratio obtained through the conventional and improved method is represented in Figure 4. The minimum % change in $\frac{Sti}{Stv}$ between conventional and improved method was observed in the gene *garP* as -30.99% and maximum in the gene *yfhL* as 30.76%. Similarly, the minimum % change in $\frac{Nti}{Ntv}$ between conventional and improved method was observed in the gene *thiS* as -194.675% and maximum in the gene *bhsA* as -129.600%. The detailed % change in $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ is represented **(**see Supplementary Table S3). While it is evident that, TFD codons and FFD codons have different

probabilities of obtaining synonymous variations, the composition of TFD: FFD was cross-checked against the % change in $\frac{Sti}{Stv}$ for each gene. Likewise, the impact of PTC% in estimating the $\frac{Nti}{Ntv}$ values have been considered in this study.

The synonymous variation in the genes is likely to be different considering the differential composition of TFD codons and FFD codons, because when FFD codon proportion is higher in a gene, $\frac{ti}{tv}$ value is going to be lower. Similarly, when the FFD codon proportion is lower, $\frac{ti}{tv}$ value is going to be higher. So, this observation is distinct, and the improved estimator is taking care of the compositional differences (estimated numbers) of different degenerate codons. The Pearson correlation coefficient value between TFD: FFD and % change in $\frac{Sti}{Stv}$ (conventional to improved) was obtained to be 0.891, signifying a strong correlation. Similarly, the Pearson correlation coefficient value between PTC% and % change in $\frac{Nti}{Ntv}$ was obtained to be -0.488, a moderate negative correlation between the two variables. The results were found to be statistically significant at $p<0.01$, for $\frac{ti}{tv}$ & $\frac{ti'}{tv'}$, $\frac{Sti}{Stv}$ & $\frac{Sti'}{Stv'}$ and $\frac{Nti}{Ntv}$ & $\frac{Nti'}{Ntv'}$. Further the correlation graph between % change in $\frac{Sti}{Stv}$ and TFD: FFD and % change in $\frac{Nti}{Ntv}$ and PTC% also substantiate the role of codon composition and $\frac{ti}{tv}$ ratio (Figure 5A & 5B). In the trendline analyses the % change in $\frac{Sti}{Stv}$ was perceived to be directly influenced by the TFD: FFD with a R^2 value of 0.795. A weak negative association was obtained between % change in $\frac{Nti}{Ntv}$ and PTC% with a R^2 value of 0.216. Hence, it is pertinent to note that the overall % change in $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ were directly influenced by the TFD and PTC codon composition across the coding sequences. We performed a simulation study to substantiate the impact of the compositional difference of degeneracy class in estimating $\frac{Sti}{Stv}$ by using our improved method in coding sequences. For detailed methodology and results (see Supplementary Table S8,

Supplementary Figure S3). The results of comparative analysis between ML based method and improved method are presented in supplementary files for 2516 genes (see Supplementary Table S9, Supplementary Figure S4).

***Sti, Stv, Nti* and *Ntv* variation dynamics in coding sequences through improved approach**

A total of 2516 coding sequences were subjected to analysis for finding out the different variation dynamics such as *Sti*, *Stv*, *Nti*, and *Ntv*. The mean *Sti*, *Stv*, *Nti*, and *Ntv* values were 0.113, 0.034, 0.014 and 0.004 respectively (see Supplementary Table S11, Figure 6). The mean *Sti*, *Stv*, *Nti*, and *Ntv* values revealed the decreasing order of their variation dynamics, respectively. This comparison of the values suggests that *Sti* is 3.3 times more frequent than *Stv*, while *Nti* is 3.7 times more frequent than *Ntv*. Moreover, the ratio of *Sti* to *Nti* was 8.2, and the ratio of *Stv* to *Ntv* was 9.3. This indicates a similar purifying selection on *Nti* and *Ntv* and this difference was found to be statistically significant ($p<0.01$) between *ti* and *tv* of both synonymous and non-synonymous variations. Table 2 presents the Pearson correlation coefficient values for the pairs *Sti*, *Stv*, *Nti*, and *Ntv* in the 2516 coding sequences, obtained using both conventional and improved methods. It shows the correlations between these variables under both approaches. By using the conventional approach, the Pearson correlation between *Sti* and *Nti* was calculated to be 0.52, while the correlation between *Stv* and *Ntv* was found to be 0.54. However, with the improved method, the correlation values decreased substantially to 0.19 for *Sti* and *Nti*, and 0.25 for *Stv* and *Ntv*, respectively. Our method is more appropriate in estimating the distinctive selection patterns between *Sti* and *Nti* even though the pattern of substitution is similar in both. As part of our comparative study, we investigated the coding sequences of similar *Sti* and different *Nti* values. In a set of 20 coding sequences with similar *Sti* values such as 0.131, while the *Nti* values varied significantly, ranging from 0.000 to 0.030. This wide range of *Nti* values, with a maximum 30-fold variation, highlights the presence of a variable

Nti pattern within the respective coding sequences. Additionally, these results suggest that different substitution types undergo distinct selection processes.

DISCUSSION

Codon degeneracy influences the transition to transversion ratio in coding sequences because a TFD codon undergoes synonymous variation only by transition while a FFD codon undergoes synonymous variation by transition as well as transversion. In addition, there are only 18 pretermination codons in the genetic code table where non-sense substitution is rare to observe. Coding sequences in a genome are with varying with composition of codons belonging to different degeneracy as well as PTC. Therefore, it is inappropriate to compare coding sequences regarding the $\frac{ti}{tv}$ ratio without accounting degeneracy and pretermination nature of codons. In this study we have developed a methodology to estimate $\frac{ti}{tv}$ ratio by accounting degeneracy and pretermination nature of codons, which has been implemented to analyze 2516 coding sequences in *E. coli*. Our result suggested that coding sequences are different regarding their $\frac{ti}{tv}$ ratio, which is independent of the composition of degenerate codons and PTC. In fact, the $\frac{Sti}{Stv}$ values are observed to be variable among the coding sequences, which is intriguing. We did not observe any significant correlation of $\frac{Sti}{Stv}$ values with the gene expression represented as CAI values. The role of context dependent mutation (40-43), which has been described recently, might be a contributing factor that needs to be investigated in future.

Our study on *ti* and *tv* has revealed that the base variation dynamics among the coding sequences is in the decreasing order such as *Sti* > *Stv* > *Nti* > *Ntv*. The lower value of *Nti* and *Ntv* indicates the role purifying selection (5) on non-synonymous variations. The strength of

purifying selection on *Nti* and *Ntv* seems comparable, which is unlike the report by Zou and Zhang (44). Our study is on *E. coli* genes, whereas their research is on 90 clades representing all domains of life. In our approach we have considered degeneracy and PTC, which has not been implemented in their study. As transitions are more frequent events than transversion, transition can be considered alone to understand the selection on coding sequences. Considering *Sti* representing the variation dynamics in a coding sequence, the *Nti* representing the strength of purifying selection acting on the coding sequence, a comparison between *Sti* and *Nti* in a coding sequence might represent the strength of purifying selection acting on it. In several instances, we have observed that coding sequences with equal *Sti* dynamics are highly different regarding their *Nti* dynamics, which indicates disproportionate purifying selection on coding sequences.

Various methods have been implemented in the estimation of the *ti* to *tv* ratio by molecular evolutionary biologists to understand the mutation as well as selection bias. The phylogeny-based method and Bayesian methods have been implemented by researchers in the inter-species study of $\frac{ti}{tv}$ (45,46). Usually the *dN*/*dS* approach has been used to visualize the selection in coding sequences (16) that considers both *ti* and *tv*. However, our method implemented in this study is simple, succinct, easy to follow for researchers in the field with limitations in mathematics and provides an insight to understand the evolution of protein coding sequences. We also did a comparison with the ML method. There are several genes where we could observe the $\frac{ti}{tv}$ values estimated by the improved method is more in concordance with the observed *ti* and *tv* values than the ML method. It has been said that the likelihood methods are computer intensive and difficult to apply to large datasets and can fall into local traps (47). Furthermore, the extrapolation of our method in various inter-species studies, as well as

phylogenetic tools can enhance the horizon of our understanding regarding mutational bias in coding sequences.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article and Supplementary Data.

ACKNOWLEDGMENTS

PKB is grateful to Tezpur University for the institutional fellowship. RA is thankful for the JRF fellowship from the DBT grant (BT/511/NE/TBP/2013) and (BT/403/NE/U-Excel/2013). PS is grateful to UGC, GoI New Delhi, for the JRF. SD is thankful to DBT for MSc. Scholarship. SSS and SKR are thankful to DBT, GoI for the twinning grant (BT/PR16361/NER/95/192/2015 date 18-10-2016) to them. NDN, SSS & SKR thankfully acknowledge the DBT funded Bioinformatics and Computational Biology Centre at Tezpur University.

Authors contributions

P.K.B, S.S.S. and S.K.R: Idea, design, executed, data analysis, writing, and discussion; P.S and R.A.: executed, data analysis, and writing; S.D: writing and discussion. N.D.N, E.F: Writing, Discussion, and critical analysis.

Author Disclosure Statement

The authors declare no potential conflict of interest. The authors also declare no competing financial interests.

FUNDING

The authors received no specific funding for this work.

REFERENCE

1. Crick, F.H.C., Barnett, L., Brenner, S. and Watts-Tobin, R.J. (1961) General nature of the genetic code for proteins. *Nature*, **30**;192:1227-32. https://doi.org/ 10.1038/1921227a0.
2. Khorana, H.G., Büchi, H., Ghosh, H.,Gupta, N., Jacob, T. M., Kössel, H., Morgan, R., Narang, S.A., Ohtsuka, E. and Wells, R.D. (1966) Polynucleotide synthesis and the genetic code. Cold Spring Harbor Symp Quant Biol. **31**:39-49. https://doi.org/10.1101/sqb.1966.031.01.010.
3. Modiano, G., Battistuzzi, G. and Motulsky, A.G. (1981) Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha-and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? *Proc Natl Acad Sci*, **78**:1110–1114. https://doi.org/ 10.1073/pnas.78.2.1110
4. van den Eynden, J., Basu, S. and Larsson, E. (2016) Somatic mutation patterns in hemizygous genomic regions unveil purifying selection during tumor evolution. *PLoS Genet*, **12**: e1006506. https://doi.org/10.1371/journal.pgen.1006506
5. Morales, A.C., Rice, A.M., Ho, A.T.l., Mordstein, C., Mühlhausen, S., Watson, S., Cano, L., Young, B., Kudla, G. and Hurst, L.D. (2021) Causes and consequences of purifying selection on SARS-CoV-2. *Genome Biol Evol*, **13**: evab196. https://doi.org/10.1093/gbe/evab196
6. Li, W-H., Gojobori, T. and Nei, M. (1981) Pseudogenes as a paradigm of neutral evolution. *Nature*, **292**:237–239. https://doi.org/10.1038/292237a0

7. Duchêne, S., Ho, S.Y.W. and Holmes, E.C. (2015) Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evol Biol*, **15**:1–10. https://doi.org/10.1186/s12862-015-0312-6
8. Stoltzfus, A. and Norris, R.W. (2016) On the causes of evolutionary transition: transversion bias. *Mol Biol Evol*, **33**:595–602. https://doi.org/10.1093/molbev/msv274
9. Sen, P., Aziz, R., Deka, R.C., Feil. E. J., Ray, S.K. and Satapathy S.S. (2022) Stem region of tRNA genes favors transition substitution towards keto bases in bacteria. *J Mol Evol*, **90**:114–123. https://doi.org/10.1007/s00239-021-10045-x
10. Seplyarskiy, V.B., Kharchenko, P., Kondrashov, A.S. and Bazykin, G.A. (2012) Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. *Mol Biol Evol*, **29**:1943–1955. https://doi.org/10.1093/molbev/mss071
11. Lewis Jr, C.A., Crayle, J., Zhou, S. Swanstrom, R. and Wolfenden, R. (2016) Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proc Natl Acad Sci*, **113**:8194–8199. https://doi.org/ 10.1073/pnas.1607580113.
12. Lyons, D.M. and Lauring, A.S. (2017) Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. *Mol Biol Evol*, **34**:3205–3215. https://doi.org/ 10.1093/molbev/msx251
13. Schroeder, J.W., Randall, J.R., Hirst, W.G., O'Donnel, M.E. and Simmons, L.A. (2017) Mutagenic cost of ribonucleotides in bacterial DNA. *Proc Natl Acad Sci*, **114**:11733–11738. https://doi.org/10.1073/pnas.1710995114
14. Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, **16**:111–120. https://doi.org/ 10.1007/BF01731581

15. Hurst, L.D. and Pál, C. (2001) Evidence for purifying selection acting on silent sites in BRCA1. *TRENDS in Genetics*, **17**:62–65. https://doi.org/ 10.1016/s0168-9525(00)02173-9
16. Aziz, R., Sen, P., Beura, P.K., Das, S., Tula, D., Dash, M., Namsa, N.D.N., Deka, R.C., Feil, E.J. Satapathy, S.S. *et al* (2022) Incorporation of transition to transversion ratio and nonsense mutations, improves the estimation of the number of synonymous and non-synonymous sites in codons. *DNA Research*, **29**: dsac023. https://doi.org/10.1093/dnares/dsac023
17. Fersht, A.R. and Knill-Jones, J.W. (1981) DNA polymerase accuracy and spontaneous mutation rates: frequencies of purine. purine, purine. pyrimidine, and pyrimidine. pyrimidine mismatches during DNA replication. *Proc Natl Acad Sci*, **78**(7):4251–4255.https://doi.org/10.1073/pnas.78.7.4251
18. Topal, M.D. and Fresco, J.R. (1976) Complementary base pairing and the origin of substitution mutations. *Nature*, **263**(5575):285-289. https://doi.org/10.1038/263285a0.
19. Kino K, Sugiyama H (2001) Possible cause of G. C> C. G transversion mutation by guanine oxidation product, imidazolone. *Chem. Biol*, 8, 369–378. https://doi.org/10.1016/S1074-5521(01)00019-9
20. Rocha, E.P.C., Touchon, M. and Feil. E. J. (2006) Similar compositional biases are caused by very different mutational effects. *Genome Res*, **16**, 1537–1547. https://doi.org/ 10.1101/gr.5525106
21. Bhagwat, A. S., Hao, W., Townes, J.P., Lee, H., Tang, H. and Foster, P.L. (2016) Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli. *Proc Natl Acad Sci*, **113**, 2176–2181. https://doi.org/10.1073/pnas.1522325113

22. Muse, S.V. and Gaut, B.S. (1994) A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome. *Mol Bio Evol*, **1 1(5)**:7 15-724. https://doi.org/ 10.1093/oxfordjournals.molbev.a040152
23. Tamura, K. (1992) The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. *Mol Biol Evol*, **9**:814–825. https://doi.org/10.1093/oxfordjournals.molbev.a040763
24. Wolfe, K.H., Sharp, P.M. and Li, W-H. (1989) Rates of synonymous substitution in plant nuclear genes. *J Mol Evol*, **29**:208–211. https://doi.org/10.1007/BF02100204
25. Moriyama, E.N. and Powell, J.R. (1997) Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. *J Mol Evol*, **45**:378–391. https://doi.org/ 10.1007/pl00006243
26. Ngandu, N.K., Scheffler, K., Moore, P., Woodman, Z., Martin, D. and Seoighe, C. (2008) Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. *Virol J*, **5**:1–11. https://doi.org/ 10.1186/1743-422X-5-160
27. Shapiro, B., Rambaut, A. and Drummond, A.J. (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol*, **23**, 7–9. https://doi.org/10.1093/molbev/msj021
28. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, **11**(5), 725-736. https://doi.org/ 10.1093/oxfordjournals.molbev.a040153.
29. Arenas, M. (2015) Trends in substitution models of molecular evolution. *Front Genet*, **6**:319. https://doi.org/10.3389/fgene.2015.00319

30. Yang, Z. and Yoder, A.D. (1999) Estimation of the Transition/Transversion Rate Bias and Species Sampling. *J Mol Evol* **48**:274–283. https://doi.org/10.1007/PL00006470
31. Bartolomé. C., Maside. X., Yi. S., Grant. A.L. and Charlesworth, B. (2005) Patterns of selection on synonymous and nonsynonymous variants in Drosophila miranda. *Genetics,* **169**(3):1495-507. https://doi.org/10.1534/genetics.104.033068
32. Wu, C-I. and Li, W.H. (1985) Evidence for higher rate of nucleotide substitution in rodents than in man. *Proc Natl Acad Sci,* **82**(6):1741-1745. https://doi.org/ 10.1073/pnas.82.6.1741
33. Ina, Y. (1998) Estimation of the transition/transversion ratio. *J Mol Evol,* **46**:521–533. https://doi.org/10.1007/PL00006333
34. Thorpe, H.A., Bayliss, S.C. and Hurst, L.D. and Feil, E.J. (2017) Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics,* **206**:363–376. https://doi.org/10.1534/genetics.116.195784
35. Sen, P. (2022) Computational Analysis of Codon Usage Bias Single Nucleotide Polymorphism and RNA Secondary Structures in Microbial Genome Sequences. Ph.D. Thesis, Tezpur University http://hdl.handle.net/10603/473607
36. Beura, P.K., Sen, P., Aziz, R., Satapathy, S.S. and Ray, S.K. (2023) Transcribed intergenic regions exhibit a lower frequency of nucleotide polymorphism than the untranscribed intergenic regions in the genomes of Escherichia coli and Salmonella enterica. *J Genet,* **102**:22. https://doi.org/10.1007/s12041-023-01418-w
37. Kumar, S., Stecher, G., and Li, M., Knyaz, C. and Tamura, K. (2018) MEGAX: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol,* **35**(6):1547–1549 https://doi.org/ 10.1093/molbev/msy096

38. Mann, H.B. and Whitney, D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Statist*, **18**(1), 50-60. https://doi.org/ 10.1214/aoms/1177730491

39. Pearson, K. (1896) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc R Soc Lond*, **60**(359-367), 489-498. https://doi.org/10.1098/rspl.1896.0076

40. Sung, W., Ackerman, M.S., Gout, J.F., Miller, S.F., Williams, E., Foster, P.L. and Lynch, M. (2015) Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. *Mol Biol Evol.* 32:1672–1683. doi: 10.1093/molbev/msv055.

41. Zhu, Y., Neeman, T., Yap, V.B. and Huttley. G.A. (2017) Statistical methods for identifying sequence motifs affecting point mutations. *Genetics.* doi: 10.1534/genetics.116.195677.

42. Aikens. R.C., Johnson. K.E. and Voight. B.F. (2019) Signals of variation in human mutation rate at multiple levels of sequence context. *Mol Biol Evol.* 36:955–965. doi: 10.1093/molbev/msz023.

43. Ling. G., Miller, D., Nielsen, R. and Stern. A. A. (2020) Bayesian framework for inferring the influence of sequence context on point mutations. *Mol Biol Evol.* 37:893–903. doi: 10.1093/molbev/msz248.

44. Zou, Z. and Zhang, J. (2021) Are Nonsynonymous Transversions Generally More Deleterious than Nonsynonymous Transitions? *Mol Biol Evol*, **38**(1):181–191. https://doi.org/ 10.1093/molbev/msaa200

45. Purvis, A. and Bromham, L. (1997) Estimating the transition/transversion ratio from independent comparisons with an assumed phylogeny. *J Mol Evol*, **44**, 112–119. https://doi.org/ 10.1007/pl00006117

46. Huelsenbeck, J. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinform* **17**, 754–755. https://doi.org/ 10.1093/bioinformatics/17.8.754.

47. Golding, B. and Felsenstein, J. (1990) A maximum likelihood approach to the detection of selection from a phylogeny. *J Mol Evol*, **31**:511-523. https://doi.org/10.1007/bf02102078

Figure_legend_Table

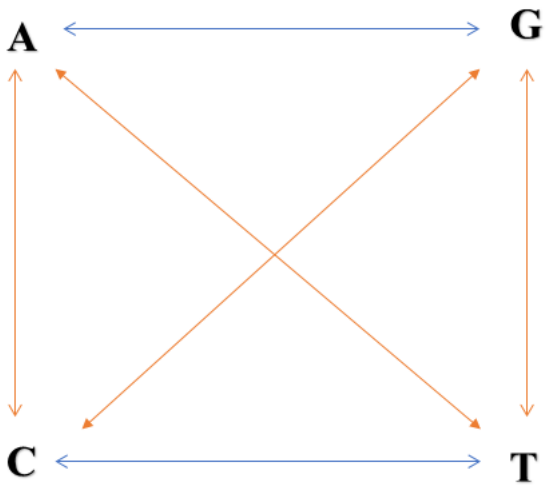


Figure 1. Schematic representation of the different substitutions occurring in the DNA. The transition *(ti)* is marked with blue arrows and (*tv)* is marked with orange arrows. A and G are categorized under purine (R) and C and T are categorized under pyrimidine (Y). Out of the twelve possible base substitutions, four are *ti* in which a purine (or pyrimidine) nucleotide is replaced by

another purine (or pyrimidine) nucleotide, and eight are *tv* in which a purine (or pyrimidine) nucleotide is replaced by another pyrimidine (or purine) nucleotide.

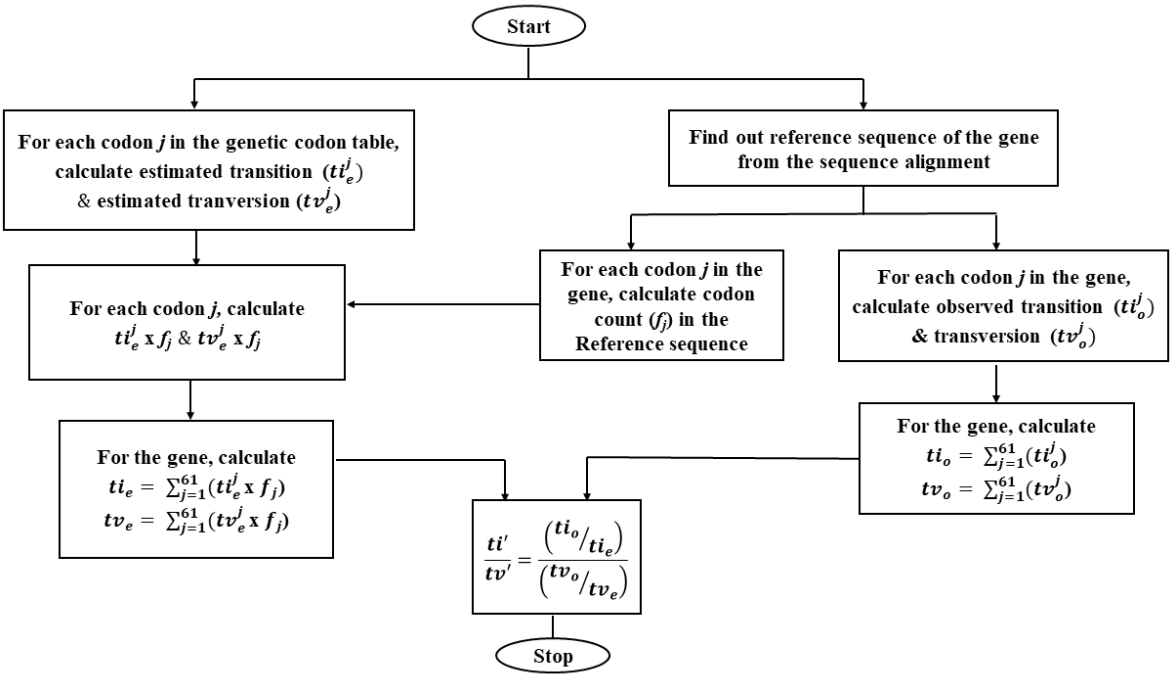


Figure 2. Schematic representation of the workflow showing the steps to execute our improved method. By considering each codon as *j*, ti_e^j and tv_e^j can be calculated for each codon as presented in Supplementary Table S2. The reference sequence needs to be calculated from the coding region (gene/query). From the reference sequence the observed ($_o$) substitutions are calculated for the gene. The other usage of the reference sequence is to find out the codon count of the gene. Then, the respective codon count is multiplied by estimated transition and transversion for each codon. The next step is to find out the sum of *ti* and sum of *tv* in all codons to represent the estimated ($_e$) *ti* and *tv* for the gene. After getting the observed and estimated values for *ti* and *tv*, now the improved equation is applied to find out the $\frac{ti'}{tv'}$ of the gene. This scenario is only showing the $\frac{ti'}{tv'}$ ratio, whereas in similar way the $\frac{Sti'}{Stv'}$ & $\frac{Nti'}{Ntv'}$ can be calculated.

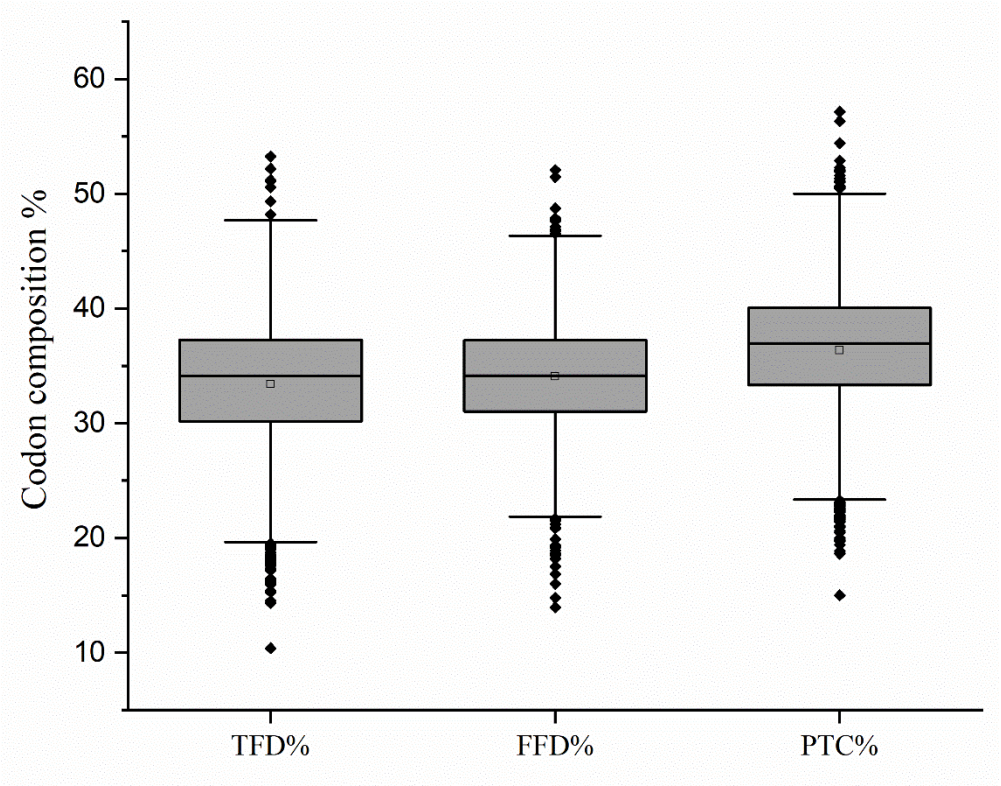


Figure 3. The box-plot presents the percentage (%) of TFD, FFD and PTC codons in the coding sequences considered to validate the improved method proposed in this study. The minimum and maximum TFD% in the dataset was observed as 10.38% and 53.25%, similarly the minimum and maximum FFD% in the dataset was observed as 13.95% and 52.05%. The minimum and maximum PTC% in the dataset was observed as 15.00% and 57.14%.

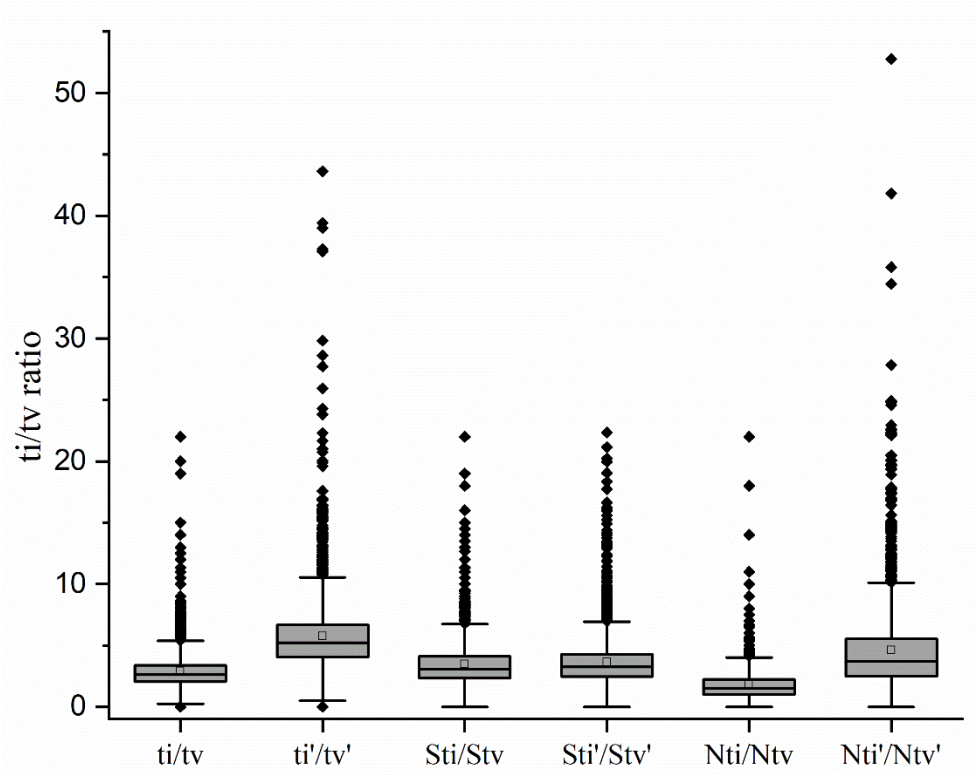
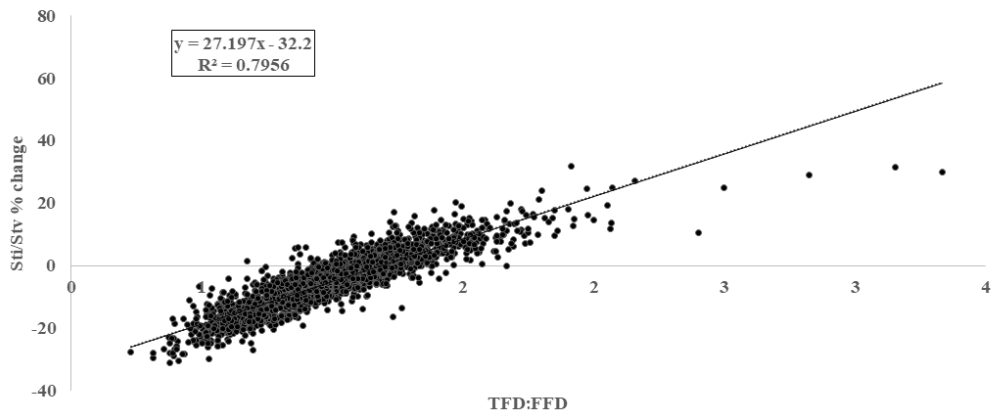


Figure 4. The box-plot presents the different $\frac{ti}{tv}$ ratio values. Here we have compared the different $\frac{ti}{tv}$ ratios of the entire set of genes through conventional and improved method. $\frac{ti}{tv}$ shows the conventional method and $\frac{ti'}{tv'}$ shows the result of improved method. Similarly, $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ represent the conventional values and $\frac{Sti'}{Stv'}$ and $\frac{Nti'}{Ntv'}$ represent the improved values. This illustration shows that the median values between $\frac{ti}{tv}$ & $\frac{ti'}{tv'}$, $\frac{Sti}{Stv}$ & $\frac{Sti'}{Stv'}$ and $\frac{Nti}{Ntv}$ & $\frac{Nti'}{Ntv'}$ are statistically different (p-value < 0.01).

A



B

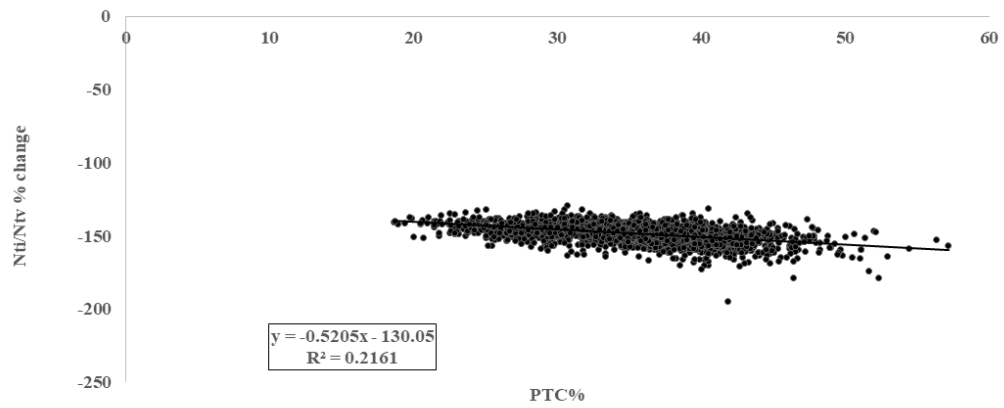


Figure 5. Regression plot of coding sequence considered to validate the improved method. The x-axis shows the codon composition (TFD: FFD and PTC%) and y-axis shows the % change in $\frac{Sti}{Stv}$ and $\frac{\mathbf{Nti}}{\mathbf{Ntv}}$. (A) % change in $\frac{Sti}{Stv}$ is showing a strong correlation with the TFD: FFD with a R^2 value of 0.795. (B) % change in $\frac{\mathbf{Nti}}{\mathbf{Ntv}}$ is showing a weak negative correlation with the PTC% with a a R^2 value of 0.216.

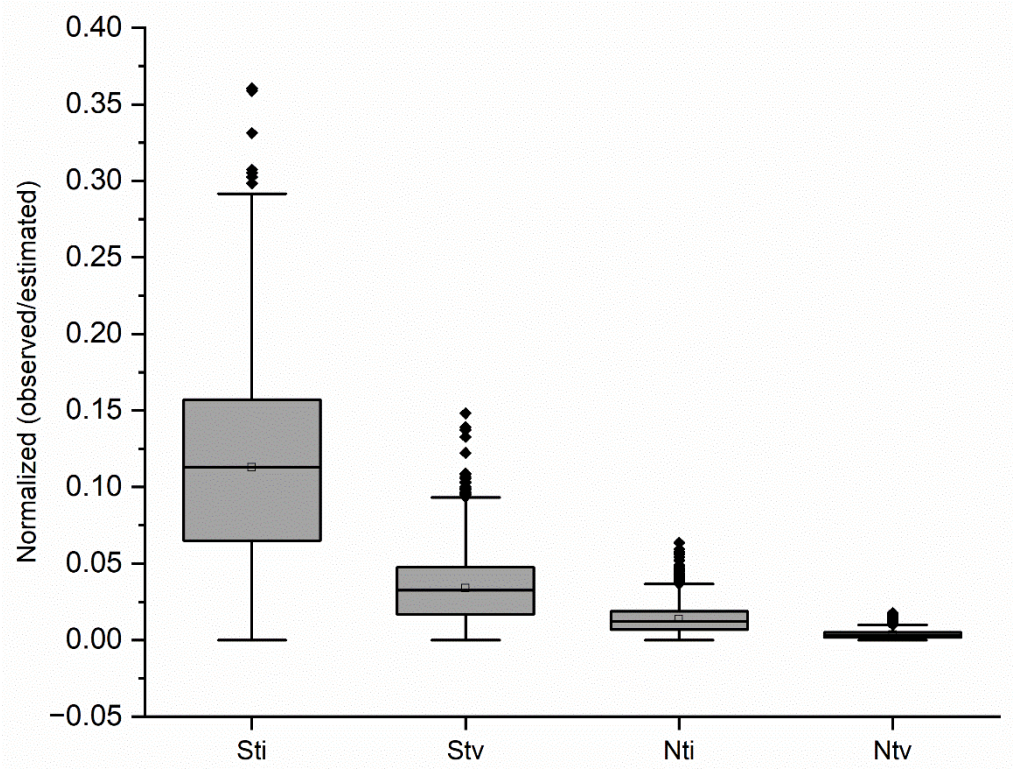


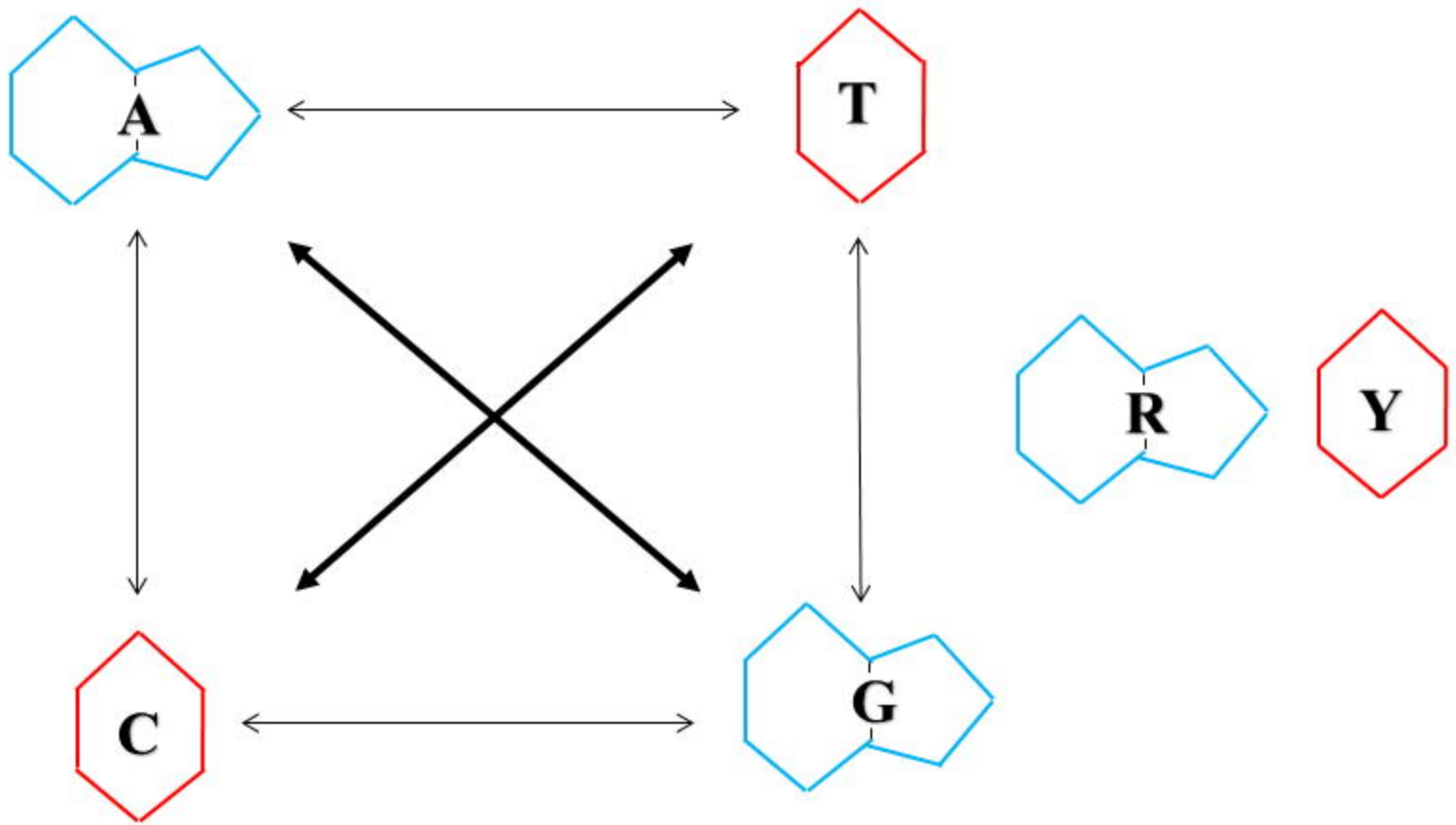
Figure 6. The figure illustrates the box-plot of the normalized *Sti*, *Stv*, *Nti*, and *Ntv* values. The mean *Sti* and *Stv* were observed as 0.113 and 0.031 respectively, similarly the mean *Nti* and *Ntv* values were observed as 0.014 and 0.004. *Sti* was observed to be highly frequent than *Stv* and *Nti* was observed to be highly frequent than *Ntv*. The pairs such as (*Sti* and *Stv)* and (*Nti* and *Ntv*) values show a statistically significant difference ($p<0.01$).

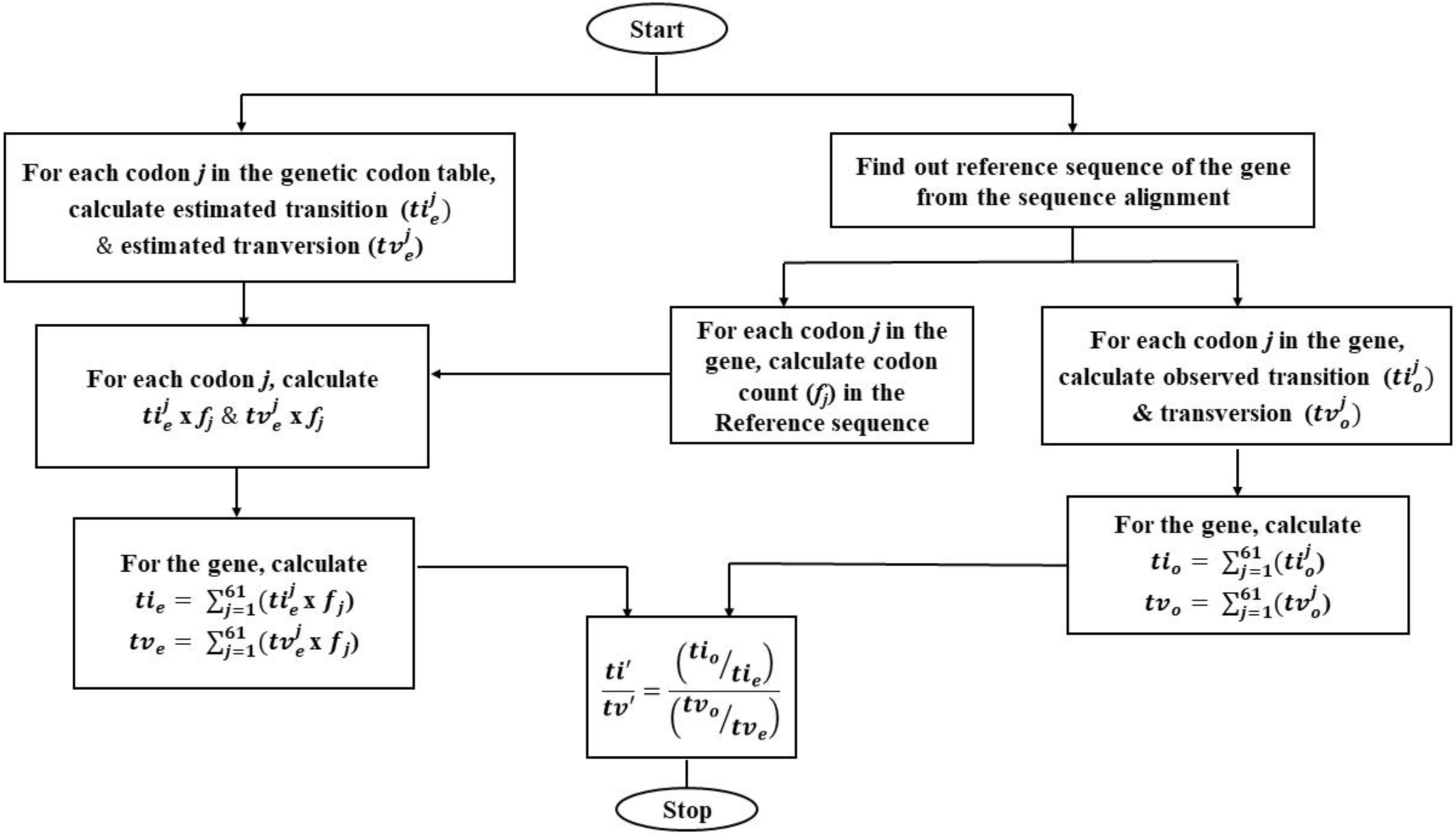
Table 1. Summary of the 2516 coding sequences selected for the present study. It represents the minimum to maximum range of different parameters like GC%, TFD/FFD, PTC% and the mean and standard deviation (S.D.) of the TFD%, FFD% and PTC%

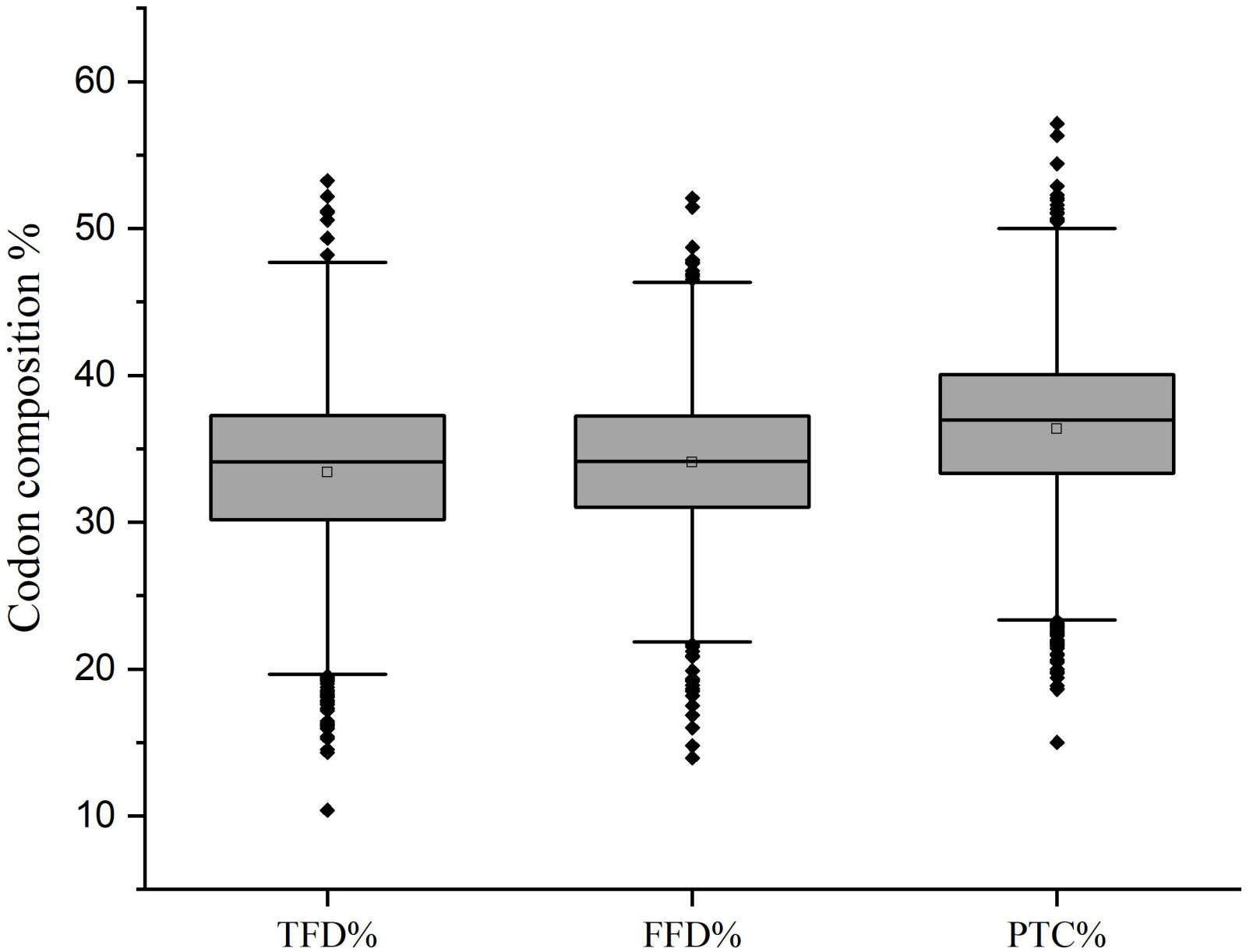
Genes considered	Size bp (Min-Max)	GC% (Min-Max)	TFD/FFD (Min-Max)	PTC% (Min-Max)	TFD%	FFD%	PTC%	
					33.384	34.084	36.350	Mean
2516	93-4554	32.95-63.80	0.22-3.33	15.00-57.14	5.726	4.761	5.625	S. D

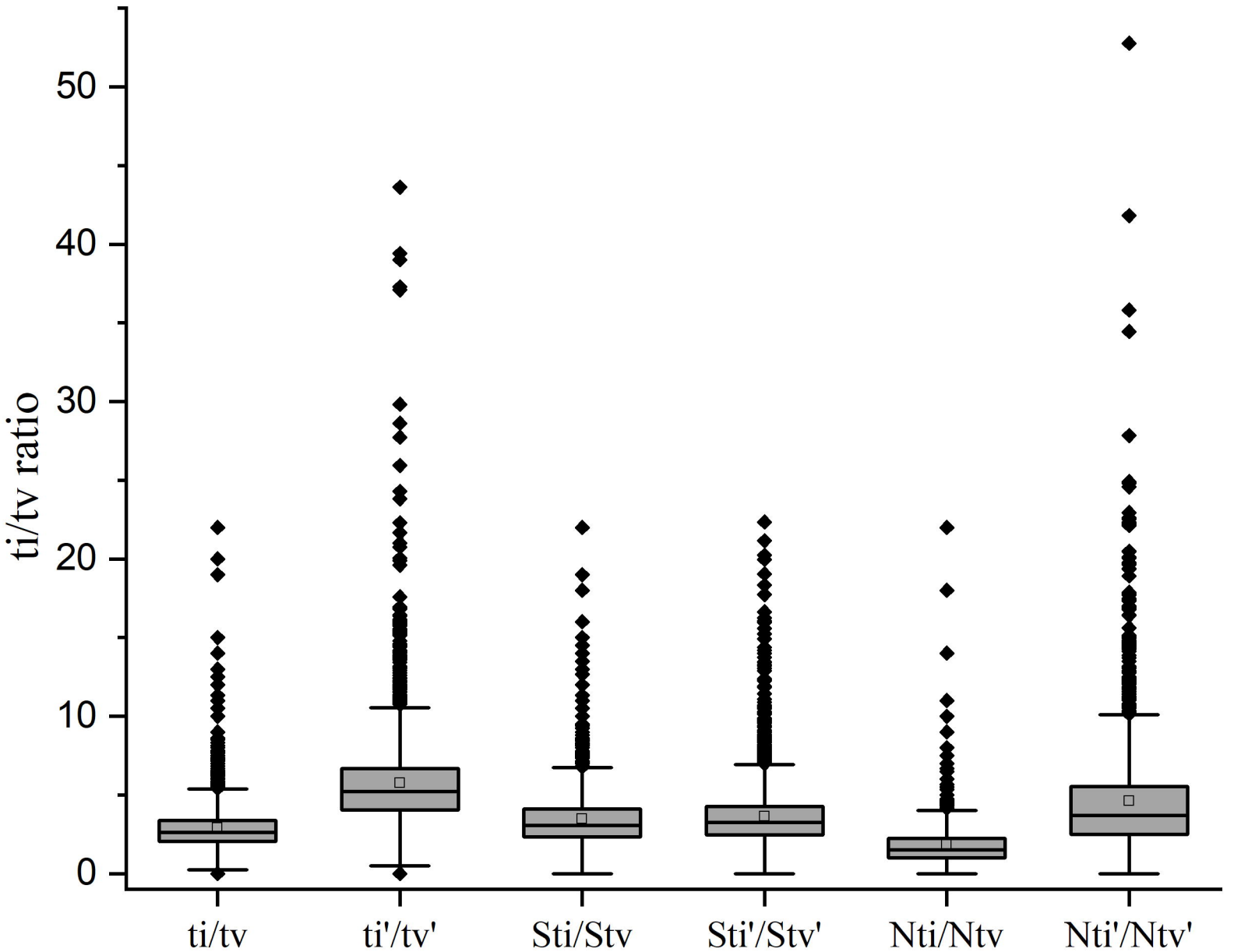
Table 2. Comparative Pearson correlation coefficient study among Sti, Sti, Nti and Ntv through conventional and improved estimator

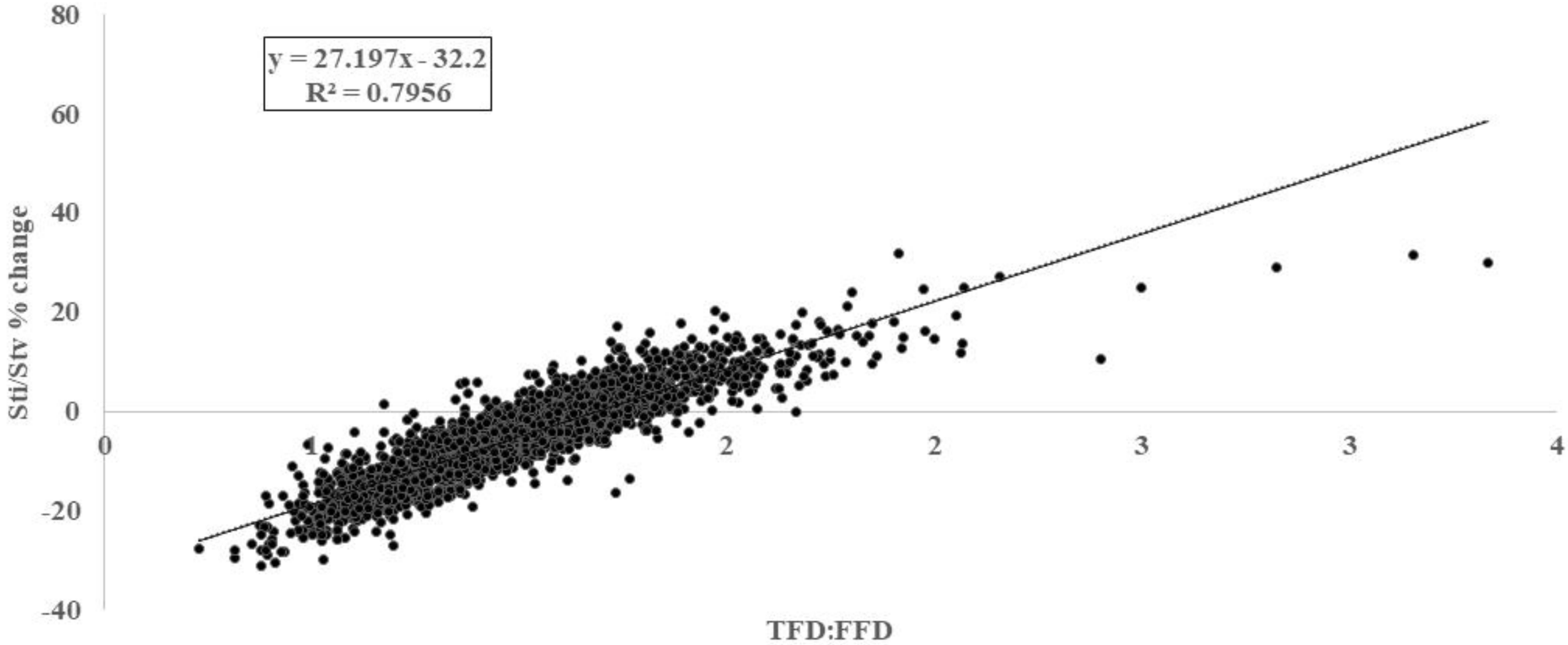
Pearson correlation coefficient	Conventional method	Improved method
	Observed values	Normalized values (Obs/Est)
Sti & Stv	0.920	0.795
Nti & Ntv	0.789	0.612
Sti & Nti	0.527	0.192
Stv & Ntv	0.544	0.257

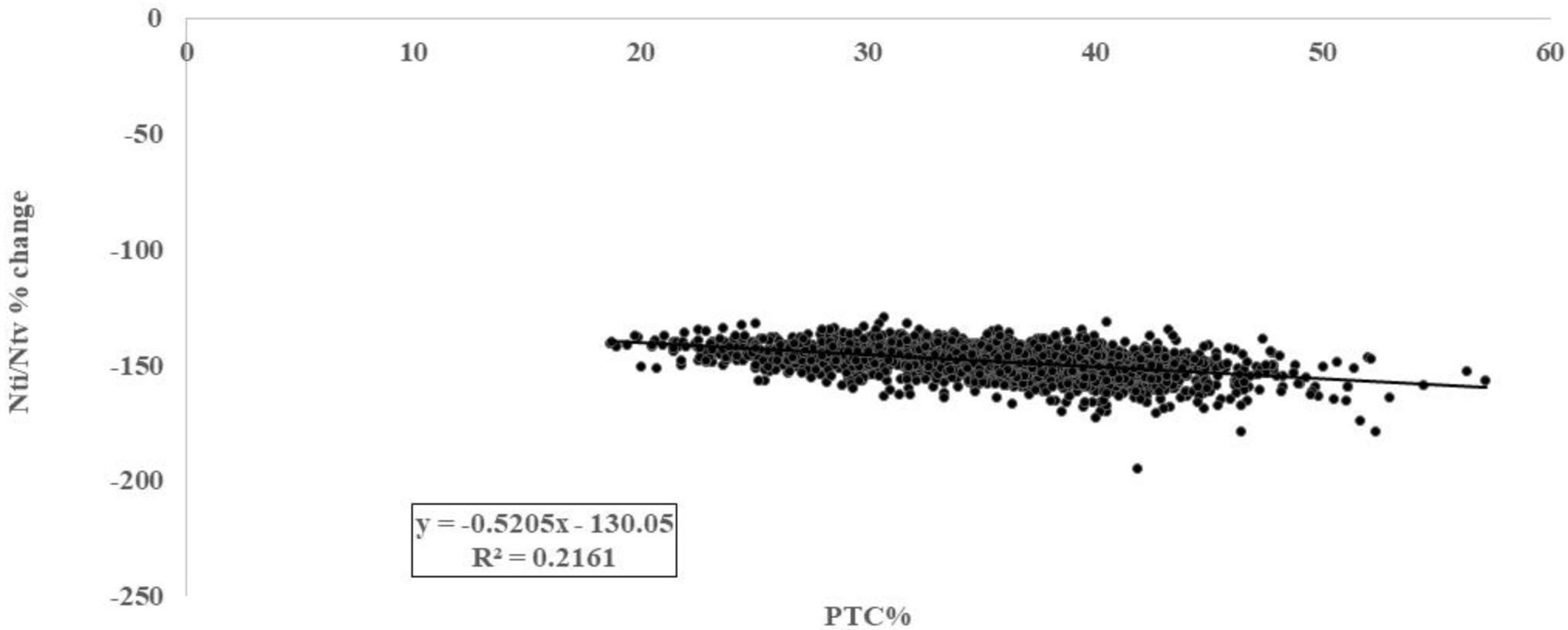

A
T
R
Y
C
G

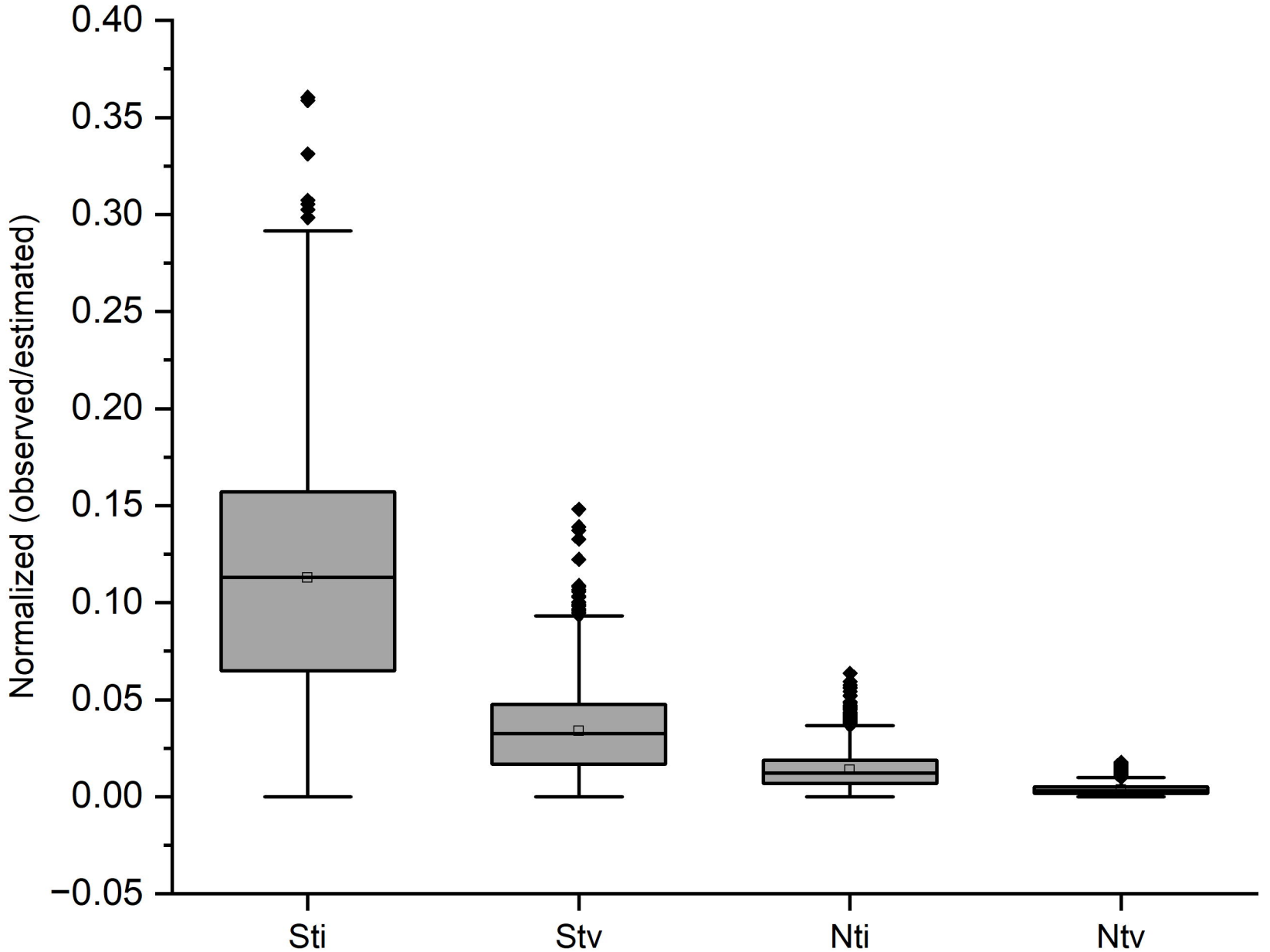

Start
For each codon j in the genetic codon table, calculate estimated transition (ti_e^j) & estimated tranversion (tv_e^j)
Find out reference sequence of the gene from the sequence alignment
For each codon j in the gene, calculate codon count (f_j) in the Reference sequence
For each codon j in the gene, calculate observed transition (ti_o^j) & transversion (tv_o^j)
For each codon j, calculate ti_e^j x f_j & tv_e^j x f_j
For the gene, calculate $ti_o = \sum_{j=1}^{61}(ti_o^j)$ $tv_o = \sum_{j=1}^{61}(tv_o^j)$
For the gene, calculate $ti_e = \sum_{j=1}^{61}(ti_e^j \text{ x } f_j)$ $tv_e = \sum_{j=1}^{61}(tv_e^j \text{ x } f_j)$
$\frac{ti'}{tv'} = \frac{(ti_o/ti_e)}{(tv_o/tv_e)}$
Stop


Codon composition %
60
50
40
30
20
10
TFD%
FFD%
PTC%


ti/tv ratio
0
10
20
30
40
50
ti/tv
ti'/tv'
Sti/Stv
Sti'/Stv'
Nti/Ntv
Nti'/Ntv'


y = 27.197x - 32.2
R² = 0.7956
80
60
40
20
0
-20
-40
Sti/Stv % change
0
1
2
2
3
3
4
TFD:FFD


0
-50
-100
-150
-200
-250
0
10
20
30
40
50
60
Nti/Ntv % change
PTC%
y = -0.5205x - 130.05
R² = 0.2161


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
Normalized (observed/estimated)
Sti
Stv
Nti
Ntv