Synonymous and non-synonymous transitions/transversions vividly disclose purifying selection in *Escherichia coli* 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

Full mailing addresses and zip codes:

Pratyush Kumar Beura, Ruksana Aziz, Piyali Sen, Saurav Das, Nima Dondu Namsa, Siddharatha Sankar Satapathy, Suvendra Kumar Ray: Tezpur University, Napaam, Tezpur, Assam, India 784028

Edward J Feil : Milner Centre for Evolution, University of Bath, Claverton Down**,** Bath, UK BA2 7AY

Running head: Substitutions and selection on coding sequences

Keywords: substitution mutation, nucleotide polymorphism, transition to transversion ratio, codon degeneracy, pretermination codon, purifying selection

Abstract

Transition (*ti*) and transversion (*tv*) are major events in genome variation. Therefore, it is crucial to estimate accurately these events for a better understanding of mutational and selection processes in organisms. Here, we developed a method (accessible at https://github.com/CBBILAB/CBBI.git) to estimate $\frac{ti}{tv}$ ratio by accounting degeneracy as well as pretermination nature of codons in coding sequences. Our findings revealed a distinct impact of codon degeneracy and pretermination codons on the $\frac{ti}{tv}$ ratio in the 2516 *Escherichia coli* coding sequences analyzed here. We observed the decreasing order of different base substitution frequencies such as *Sti* (synonymous transition) > *Stv* (synonymous transversion) > *Nti* (non-synonymous transition) > *Ntv* (non-synonymous transversion) in these coding sequences. The correlation between Sti and Stv values was strong (*r* value 0.795) whereas the correlation between *Sti* and *Nti* was weak (*r* value 0.192). Coding sequences with similar *Sti* values exhibited a wide range of *Nti* values. This indicated the varying strength of purifying selection acting on the coding sequences. In concordance with the assumption, the genes with high *Nti* values were observed to have lower CAI values than that of the genes with lower *Nti* values. Our approach is convenient to visualize the base substitution variation dynamics as well as selection in coding sequences. The proposed method is useful to estimate different $\frac{ti}{tv}$ ratios accurately in coding sequences and is insightful from an evolutionary perspective.

Article Summary

Genetic diversity is pivotal in evolution, with base substitution as a key driver. Transition (*ti*) surpasses transversion (*tv*) in genome frequency, making *ti*/*tv* ratios a valuable metric for studying mutation bias. Our improved estimator enhances *ti*/*tv* estimation in coding sequences, accounting in codon degeneracy and pretermination codons. Additionally, we unveil insights into substitution dynamics (*Sti*, *Stv*, *Nti*, *Ntv*) and demonstrate selection's impact on coding sequences by comparing *Sti* and *Nti* values.

Introduction

The genetic code table is a beautiful gift of nature to study molecular evolution (Crick *et al.* 1961; Khorana *et al.* 1966. 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 Fig. 1a). In addition to the above, compositional variation of pretermination codons (PTC) (Supplementary Fig. 1b and 1c) (Modiano *et al.* 1981) 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 (Li *et al.* 1981; Eynden *et al.* 2016; Morales *et al.* 2021). Considering in general, a *ti* being more frequent than a *tv* (Seplyarskiy *et al.* 2012; Duchêne *et al.* 2015; Stoltzfus and Norris 2016; Lewis *et al.* 2016; Lyons and Lauring 2017; Schroeder *et al.* 2017; Sen *et al.* 2022) which is known as Kimura's two parameter model (Kimura 1980), and a stronger purifying selection on non-synonymous variation than a synonymous variation (Hurst and Pál 2001), 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* (Aziz *et al.* 2022). Therefore, in any organism, the $\frac{ti}{tv}$ ratio is likely to vary across the coding sequences having differences in degenerate codon composition (Beura *et al*. 2023). Hence, a comparative study of $\frac{ti}{tv}$ ratios 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 (Fersht and Knill-Jones 1981) (Supplementary Fig. 1d). 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 (Duchêne *et al*. 2015; Stoltzfus and Norris 2016). Because of the following factors such as geometry of DNA double helix (Topal and Fresco 1976), modification of DNA bases such as deamination of cytosine and adenine (Kino and Sugiyama 2001; Rocha *et al.* 2006; Bhagwat *et al*. 2016), secondary structure in the RNA (Sen *et al*. 2022) and codon degeneracy are known to elevate *ti* phenomenon in a genome over *tv* (Muse and Gaut 1994; Aziz *et al.* 2022). 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 (Supplementary Table 1). Yet the observed synonymous variations are always higher than the expected proportion across genes (Wolfe *et al*. 1989; Tamura 1992; Moriyama and Powell 1997; Ngandu *et al*. 2008) due to stronger purifying selection on non-synonymous variations than the synonymous variations (Hurst and Pál 2001). 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 (Supplementary Table 2).

The codon substitution model suggests that sites within codons evolve at different rates and consequently they should not be equally treated (Muse and Gaut 1994; Shapiro *et al.* 2006; Arenas 2015). 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 (Muse and Gaut 1994; Goldman and Yang 1994). 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 (Aziz *et al.* 2022). 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 (Yang and Yoder 1994). 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 (Yang and Yoder 1994). 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 selection pressure on synonymous and non-synonymous variations are heterogeneous (Bartolomé *et al.* 2005), 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 (Wu and Li 1985), 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 (Ina 1998). 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* (Thrope *et al*. 2017). The information regarding the dataset selected for the present study is provided in Supplementary Table 3. 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 (Supplementary Table 4) 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 (Sen et al. 2022; Aziz *et al*. 2022; Sen 2022, Beura *et al.* 2023).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 (Supplementary Table 5). 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 (Supplementary Table 3). 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 (Supplementary Table 6). 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 (Supplementary Table 7). A detailed workflow is provided in Fig. 1.

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, $\boldsymbol{ti_e}$ is the theoretically estimated transition, tv_o is transversion observed, tv_e is transversion estimated, $\boldsymbol{Sti_o}$ is synonymous transition observed, $\boldsymbol{Sti_e}$ is the theoretically estimated synonymous transition, $\boldsymbol{Stv_o}$ is synonymous transversion observed, $\boldsymbol{Stv_e}$ is the theoretically estimated synonymous transversion, $\boldsymbol{Nti_o}$ is non-synonymous transition observed, $\boldsymbol{Nti_e}$ is the theoretically estimated non-synonymous transition, $\boldsymbol{Ntv_o}$ is non-synonymous transversion observed, $\boldsymbol{Ntv_e}$ is the theoretically estimated non-synonymous transversion. We calculated $\frac{ti}{tv}$ ratios for all the genes using a program written in the Python language.

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 (Kumar *et al.* 2018). 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 (Mann and Whitney 1947) to find out the *p*-value. The correlation plot was drawn for R^2 value and Pearson's correlation coefficient (Pearson 1896) was also calculated between the $\frac{ti}{tv}$ values obtained using the conventional and improved method.

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 (Supplementary Table 3). The percentage (%) compositional differences in the 2516 coding sequences represented as TFD%, FFD%, and PTC% is shown in a box-plot (Fig. 2a). 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% (Fig. 2a). 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 Fig. 2). The complete set of observed and estimated values of different variations is enlisted (Supplementary Table 3). 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 $\frac{Nti'}{Ntv'}$) (Supplementary Table 3). 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 Fig. 2b. 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 **(**Supplementary Table 3). 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 (Fig. 3a & 3b). 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 (Supplementary Table 8, Supplementary Fig. 3). The results of comparative analysis between ML based method and improved method are presented in supplementary files for 2516 genes (Supplementary Table 9, Supplementary Fig. 4).

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

The variation dynamics such as *Sti*, *Stv*, *Nti*, and *Ntv* were analyzed in the 2516 coding sequences where the mean values were 0.113, 0.034, 0.014 and 0.004, respectively (Supplementary Table 11, Fig. 4a). The mean values suggested the decreasing order of the variation dynamics: *Sti* > *Stv* >*Nti* > *Ntv*. A comparison among the values suggested that *Sti* was 3.3 times more frequent than *Stv*, while *Nti* was 3.7 times more frequent than *Ntv*. In addition, *Sti* is 8.2 times more frequent than *Nti* while *Stv* was 9.3 times more frequent than *Ntv*. This indicated that the relative purifying selection on *Nti* and *Ntv* in relation to their synonymous counter parts was similar. The Pearson correlation coefficient *r* values between *Sti* and *Stv* was 0.795 while the same between *Sti* and *Nti*, was 0.192 (Table 2). This indicated that selection on non-synonymous variation is independent upon the synonymous variation. In comparison with the conventional method, our improved method is more appropriate in estimating the distinctive selection patterns between *Sti* and *Nti* (Table 2).

As part of our comparative study, we investigated the coding sequences of similar *Sti* and different *Nti* values (Fig. 4b). In a set of 20 coding sequences with similar *Sti* values such as 0.131, the *Nti* values were 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. Similar observation could be found in other Sti values. We compared Sti and Nti values in relation to gene expression represented with CAI values (Sharp and Li 1987; Sen et al. 2019). We studied genes with size more than 500 bp. The top 100 genes with maximum Sti values (mean 0.255) were compared with bottom 100 genes with minimum Sti values (mean 0.006). The mean CAI

values for the top 100 genes were 0.565 whereas the mean CAI values for the bottom 100 genes was 0.509. The two mean CAI values were found to be close. This indicates the Sti values might not be influenced significantly due to gene expression. Similarly, the top 100 genes with maximum *Nti* values (mean 0.036) were compared with bottom 100 genes with minimum *Nti* values (mean 0.001). The mean CAI values for the top 100 genes were 0.491 whereas the mean CAI values for the bottom 100 genes was 0.638. The two mean CAI values of the coding sequences were found to be significantly different ($p<0.01$). The selection due to gene expression is stronger on Nti than on Sti among the strains of *E. coli*.

Discussion

Codon degeneracy influences the *ti* to *tv* 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 (Aziz *et al.* 2022). In addition, there are only 18 pretermination codons in the genetic code table where nonsense substitution is rare to observe. Coding sequences in a genome vary 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 (Sung *et al.* 2015; Zhu *et al.* 2017; Aikens *et al.* 2019; Ling *et al.* 2020), which has been described recently, might be a contributing factor that needs to be investigated in future.

Our investigation into *ti* and *tv* revealed that *Sti* is around three times more frequent than *Stv* while it is eight times more frequent than *Nti*. The difference between *Sti* and *Stv* may be attributed due to the structural changes in DNA happening due to mispairing between either purine: purine or pyrimidine: pyrimidine but the difference between *Sti* and *Nti* may be attributed to the purifying selection acting on the coding sequences because of the non-synonymous changes. The magnitude of the purifying selection acting on non-synonymous substitution in *E. coli* genome is vividly observed. The strength of purifying selection on *Nti* and *Ntv* seems comparable considering the closeness of the $\frac{Sti}{Stv}$ and the $\frac{Nti}{Ntv}$ values. In a recent study by Zou and Zhang (Zou and Zhang 2021) suggests that the purifying selection is more on *Ntv* than on *Nti* which is unlike the observation reported in our study. Our different observation from that of Zou and Zhang might be attributed to the intra-species study limited to *E. coli* whereas Zou and Zhang worked on an inter-species approach covering 90 clades representing all domains of life. Further, in our estimation of *Nti* and *Ntv* we have accounted codon degeneracy and PTC.

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. Additionally, when we compared the mean CAI values of the coding sequences with higher *Sti* values against the mean CAI values of the coding sequences with lower *Sti* values, we did not find any significant difference. On the contrary, when we conducted the same analysis for the *Nti* values, we observed a significant distinction between the CAI values of coding sequences with higher *Nti* values and those with lower *Nti* values. This observation indicates that highly expressed genes tend to have lower *Nti* values. Consequently, it suggests that purifying selection due to gene expression is more pronounced for *Nti* values than for *Sti* values. Since our study on selection relies on intra-species analysis, we anticipate that future research will unveil the underlying mechanisms of selection governing variation dynamics in inter-species studies as well. A probable discrepancy between the process of selection between intra-species and inter-species can be drawn. This is a simple demonstration of observing purifying selection acting on coding sequences. Though the difference between *ti* and *tv* are known since long, we could use it to study selection on coding sequences because of the improved equation.

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 $\frac{ti}{tv}$ (Purvis and Bromham 1997; Huelsenbeck *et al.* 2001). Usually the *dN*/*dS* approach has been used to visualize the selection in coding sequences (Aziz *et al.* 2022) 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 (Golding and Felsenstein 1990). 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 and software availability

A script written in Python-language is available at GitHub in the following link: (https://github.com/CBBILAB/CBBI.git). Further queries regarding the software may contact SSS (ssankar@tezu.ernet.in). The authors confirm that the data supporting the findings of this study are available within the article and Supporting information.

Acknowledgments

PKB is grateful to Tezpur University for its 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. We are also thankful to Prof. Kini R Manjunatha, National University of Singapore for his kind suggestions during his visit to Tezpur University.

Funding

We received no external funding for the work.

Conflicts of interest

The author(s) declare no conflict of interest.

Literature cited

Aikens RC, Johnson KE, Voight BF. Signals of variation in human mutation rate at multiple levels of sequence context. Mol Biol Evol. 2019;36(5): 955–965. doi:10.1093/molbev/msz023.

Arenas M. Trends in substitution models of molecular evolution. Front Genet. 2015; 6:319. doi: 10.3389/fgene.2015.00319

Aziz R. et al. 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. 2022;29: dsac023. doi:10.1093/dnares/dsac023

Bartolomé C, Maside X, Yi S, Grant AL, Charlesworth B. Patterns of selection on synonymous and nonsynonymous variants in Drosophila miranda. Genetics. 2005; 169(3):1495-507. doi: 10.1534/genetics.104.033068

Beura PK, Sen P, Aziz R, Satapathy SS, Ray S, 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. 2023; 102:22. doi:10.1007/s12041-023-01418-w

Beura PK et al. Difference in synonymous polymorphism related to codon degeneracy between co-transcribed genes in the genome of *Escherichia coli*. Curr. Sci. 2023. Doi: 10.18520/cs/v125/i8/

Bhagwat AS et al. Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in Escherichia coli. Proc Natl Acad Sci. 2016;113, 2176–2181. doi:10.1073/pnas.1522325113.

Crick FHC, Barnett L, Brenner S, Watts-Tobin RJ, General nature of the genetic code for proteins. Nature 1961;30; 192:1227-32. doi:10.1038/1921227a0

Duchêne S, Ho SYW, Holmes EC. Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. BMC Evol Biol. 2015; 15:1–10. doi:10.1186/s12862-015-0312-6

Fersht AR, Knill-Jones JW DNA polymerase accuracy and spontaneous mutation rates: frequencies of purine. purine, purine. pyrimidine, and pyrimidine. pyrimidine mismatches during DNA replication. Proc Natl Acad Sci. 1981;78(7):4251–4255. doi:10.1073/pnas.78.7.4251

Golding B, J. Felsenstein. A maximum likelihood approach to the detection of selection from a phylogeny. J Mol Evol. 1990;31:511-523. doi:10.1007/BF02102078

Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994;11(5), 725-736. doi:10.1093/oxfordjournals.molbev.a040153

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

Hurst LD, Pál C. Evidence for purifying selection acting on silent sites in BRCA1. TRENDS in Genetics. 2001;17:62–65. doi:10.1016/s0168-9525(00)02173-9.

Ina Y. Estimation of the transition/transversion ratio. J Mol Evol. 1998; 46:521–533. doi: 10.1007/PL00006333

Khorana HG et al. Polynucleotide synthesis and the genetic code. Cold Spring Harbor Symp Quant Biol. 1966; 31:39-49. doi:10.1101/sqb.1966.031.01.010

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16:111–120. doi:10.1007/BF01731581

Kino K, Sugiyama H. Possible cause of G. C> C. G transversion mutation by guanine oxidation product, imidazolone. Chem. Biol. 2001;8, 369–378. doi:10.1016/S1074-5521(01)00019-9

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGAX: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547–1549. doi:10.1093/molbev/msy096.

Lewis Jr CA, Crayle J, Zhou S, Swanstrom R, Wolfenden R. Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. Proc Natl Acad Sci. 113:8194–8199 (2016). doi:10.1073/pnas.1607580113.

Li W-H, Gojobori T, Nei M. Pseudogenes as a paradigm of neutral evolution. Nature 1981; 292:237–239. doi:10.1038/292237a0

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

Lyons DM, Lauring AS, Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. Mol Biol Evol. 2017;34:3205–3215. doi:10.1093/molbev/msx251

Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist. 1947;18(1), 50-60. doi:10.1214/aoms/1177730491

Modiano G, Battistuzzi G, Motulsky AG. 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. 1981; 78:1110–1114. doi:10.1073/pnas.78.2.1110

Morales AC et al. Causes and consequences of purifying selection on SARS-CoV-2. Genome Biol Evol. 2021;13: evab196. doi:10.1093/gbe/evab196

Moriyama EN, Powell JR. Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J Mol Evol. 1997; 45:378–391. doi:10.1007/pl00006243.

Muse SV, Gaut BS. A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome. Mol Bio Evol. 1994.;1 1(5):7 15-724. doi:10.1093/oxfordjournals.molbev.a040152

Ngandu NK et al. Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol J. 2008; 5:1–11. doi:10.1186/1743-422X-5-160

Pearson K. 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. 1896;60(359-367), 489-498. doi:10.1098/rspl.1896.0076

Purvis A, Bromham L. Estimating the transition/transversion ratio from independent comparisons with an assumed phylogeny. J Mol Evol. 1997;44, 112–119. doi:10.1007/pl00006117

Rocha EPC, M. Touchon, Feil EJ. Similar compositional biases are caused by very different mutational effects. Genome Res. 16, 2006; 1537–1547. doi:10.1101/gr.5525106

Schroeder JW, Randall JR, Hirst WG, O'Donnel ME, Simmons LA. Cost of ribonucleotides in bacterial DNA. Proc Natl Acad Sci. 2017; 114:11733–11738. doi:10.1073/pnas.1710995114

Sen P, Waris A, Ray SK, Satapathy SS. A Web Portal to Calculate Codon Adaptation Index (CAI) with Organism Specific Reference Set of High Expression Genes for Diverse Bacteria Species. ICICSCNT. 2019; doi:10.1007/978-981-15-0633-8_31.

Sen P. "Computational Analysis of Codon Usage Bias Single Nucleotide Polymorphism and RNA Secondary Structures in Microbial Genome Sequences". Ph.D. Thesis, Tezpur University. 2022; http://hdl.handle.net/10603/473607.

Sen P, Aziz R, Deka RC, Feil EJ, Ray SK, Satapathy SS. Stem region of tRNA genes favors transition substitution towards keto bases in bacteria. J Mol Evol. 2022; 90:114–123. doi:10.1007/s00239-021-10045-x.

Seplyarskiy VB, Kharchenko P, Kondrashov AS, Bazykin GA. Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. Mol Biol Evol. 2012; 29:1943–1955. doi:10.1093/molbev/mss071.

Shapiro B, Rambaut A, Drummond AJ, Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol. 2006;23, 7–9. doi:10.1093/molbev/msj021

Sharp PM, Li W-H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. Feb 1987;11;15(3):1281-95. doi:10.1093/nar/15.3.1281

Stoltzfus A, Norris RW, On the causes of evolutionary transition: transversion bias. Mol Biol Evol. 2016; 33:595–602. doi:10.1093/molbev/msv274.

Sung W. et al. Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. Mol Biol Evol. 2015; 32:1672–1683. doi:10.1093/molbev/msv055.

Tamura K. The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. Mol Biol Evol. 1992; 9:814–825. doi:10.1093/oxfordjournals.molbev.a040763

Thorpe HA, Bayliss SC, Hurst LD, Feil EJ, Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. Genetics. 2017;206:363–376. doi:10.1534/genetics.116.195784.

Topal MD, Fresco JR, Complementary base pairing and the origin of substitution mutations. Nature 1976;263(5575):285-289. doi:10.1038/263285a0.

van den Eynden J, Basu S, Larsson E. Somatic mutation patterns in hemizygous genomic regions unveil purifying selection during tumor evolution. PLoS Genet. 12: e1006506 (2016). doi:10.1371/journal.pgen.1006506

Yang Z, Yoder AD. Estimation of the Transition/Transversion Rate Bias and Species Sampling. J Mol Evol. 1999;48:274–283. doi:10.1007/PL00006470

Wolfe KH, Sharp PM, Li W-H. Rates of synonymous substitution in plant nuclear genes. J Mol Evol. 1989;29:208–211. doi:10.1007/BF02100204

Wu C-I, Li W-H. Evidence for higher rate of nucleotide substitution in rodents than in man. Proc Natl Acad Sci. 1985;82(6):1741-1745. doi:10.1073/pnas.82.6.1741

Zhu Y, Neeman T, Yap VB, Huttley GA. Statistical methods for identifying sequence motifs affecting point mutations. Genetics. 2017;205(2):843-856. doi:10.1534/genetics.116.195677

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

Figure_legend_Table

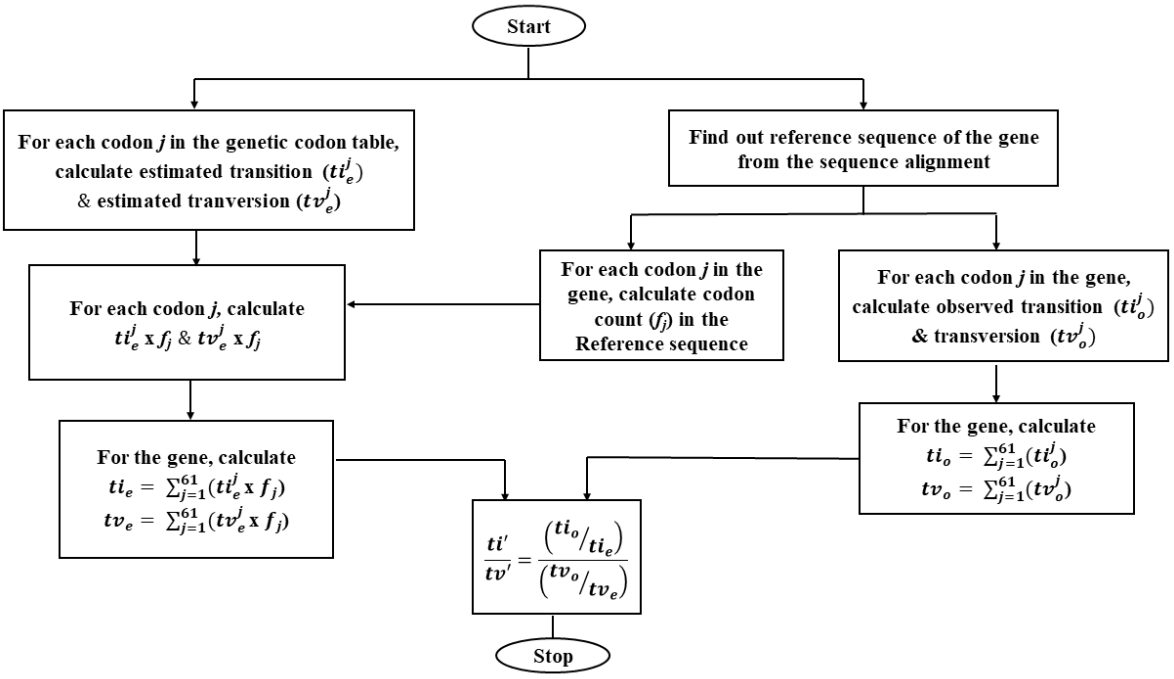
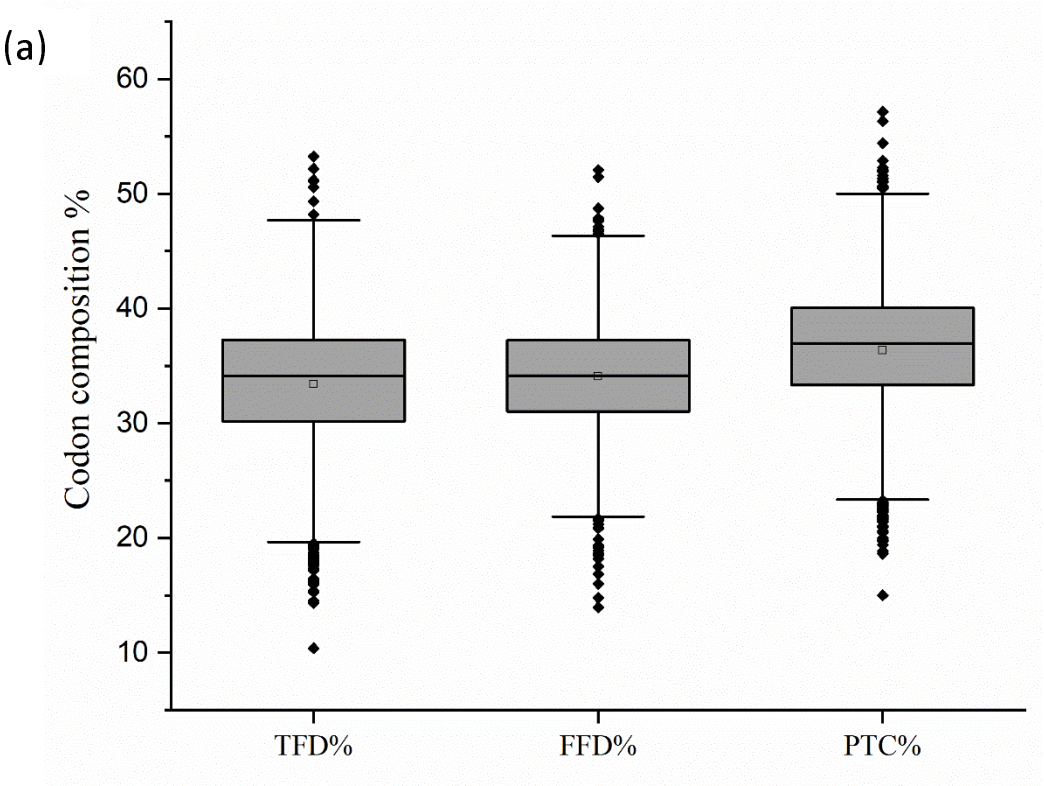


Fig. 1. Schematic representation of the workflow showing the steps to execute our improved method. By considering each codon as *j*, ti_e^j and tv_e^j can be calculated for each codon as presented in Supplementary Table S2. The reference sequence needs to be calculated from the coding region (gene/query). 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.


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

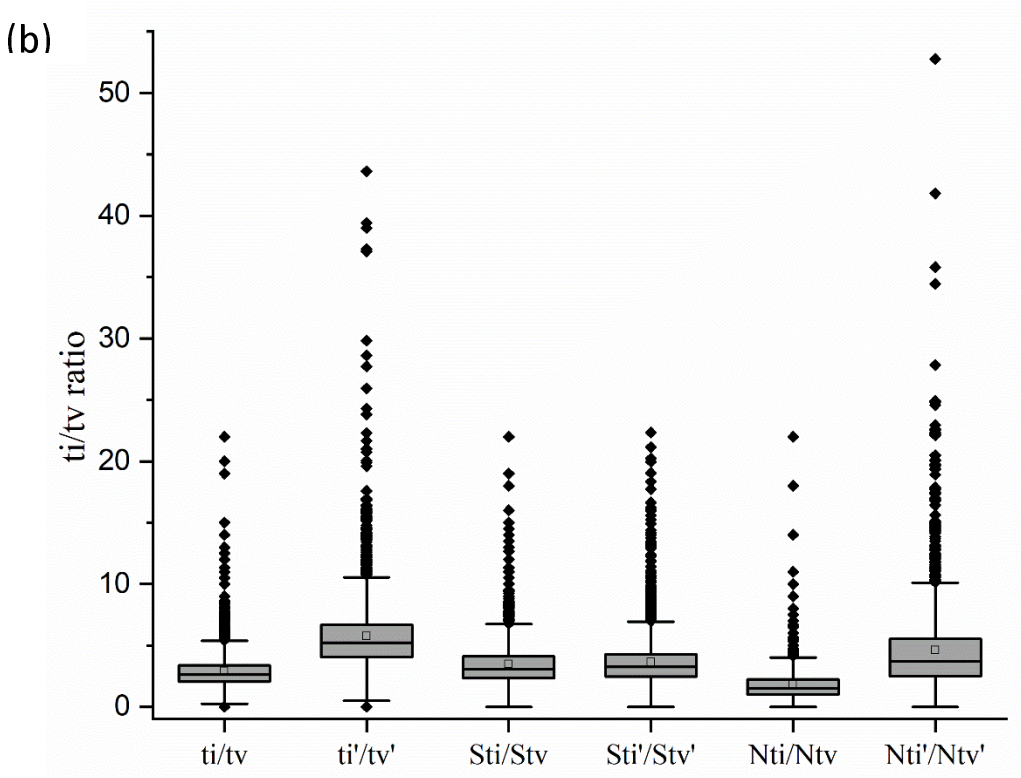
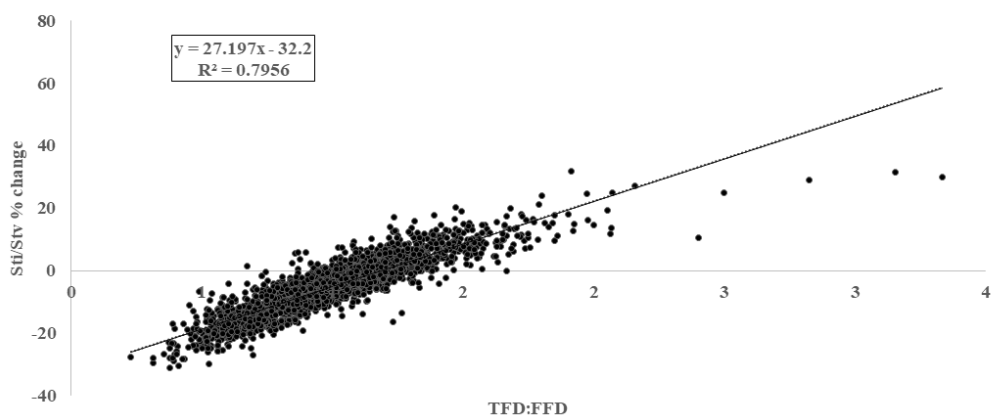

(b)
ti/tv ratio
50
40
30
20
10
0
ti/tv
ti'/tv'
Sti/Stv
Sti'/Stv'
Nti/Ntv
Nti'/Ntv'

Fig. 2. (a) 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%. (b) 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)

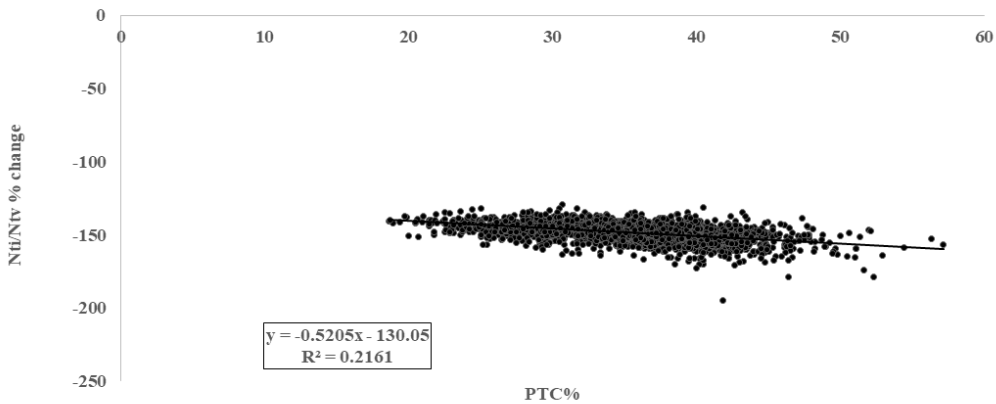
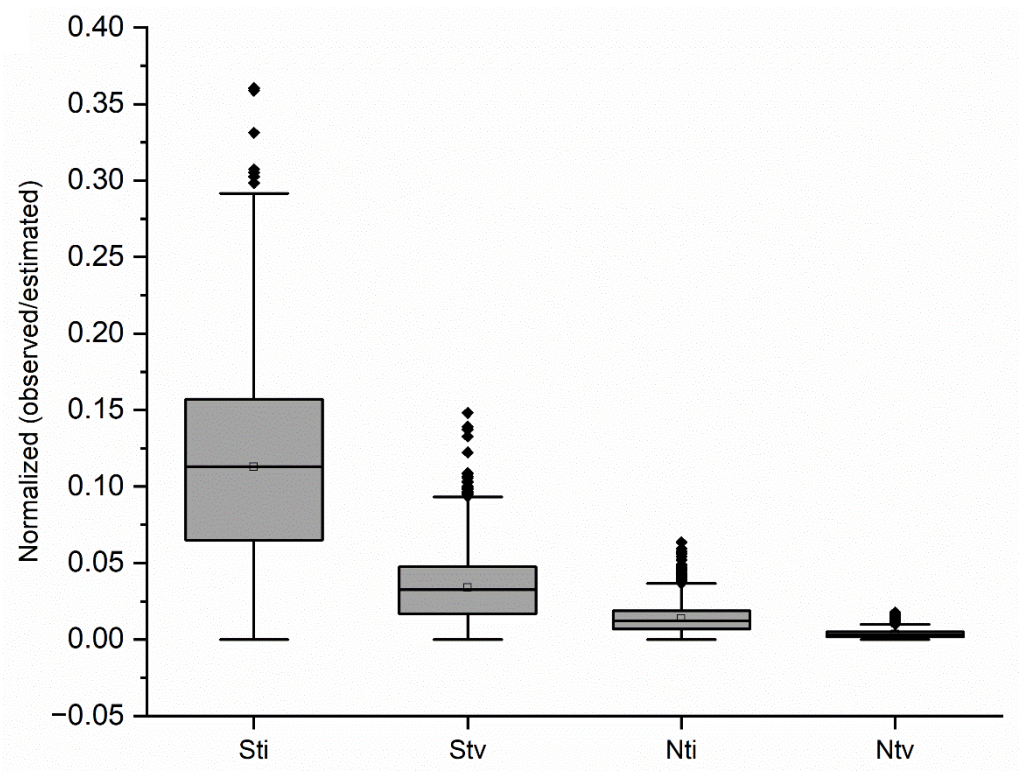


Fig. 3. Regression plot of coding sequence considered to validate the improved method. The x-axis shows the codon composition (TFD: FFD and PTC%) and y-axis shows the % change in $\frac{Sti}{Stv}$ and $\frac{\mathbf{Nti}}{\mathbf{Ntv}}$. (a) % change in $\frac{Sti}{Stv}$ is showing a strong correlation 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.

(a)


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

(b)

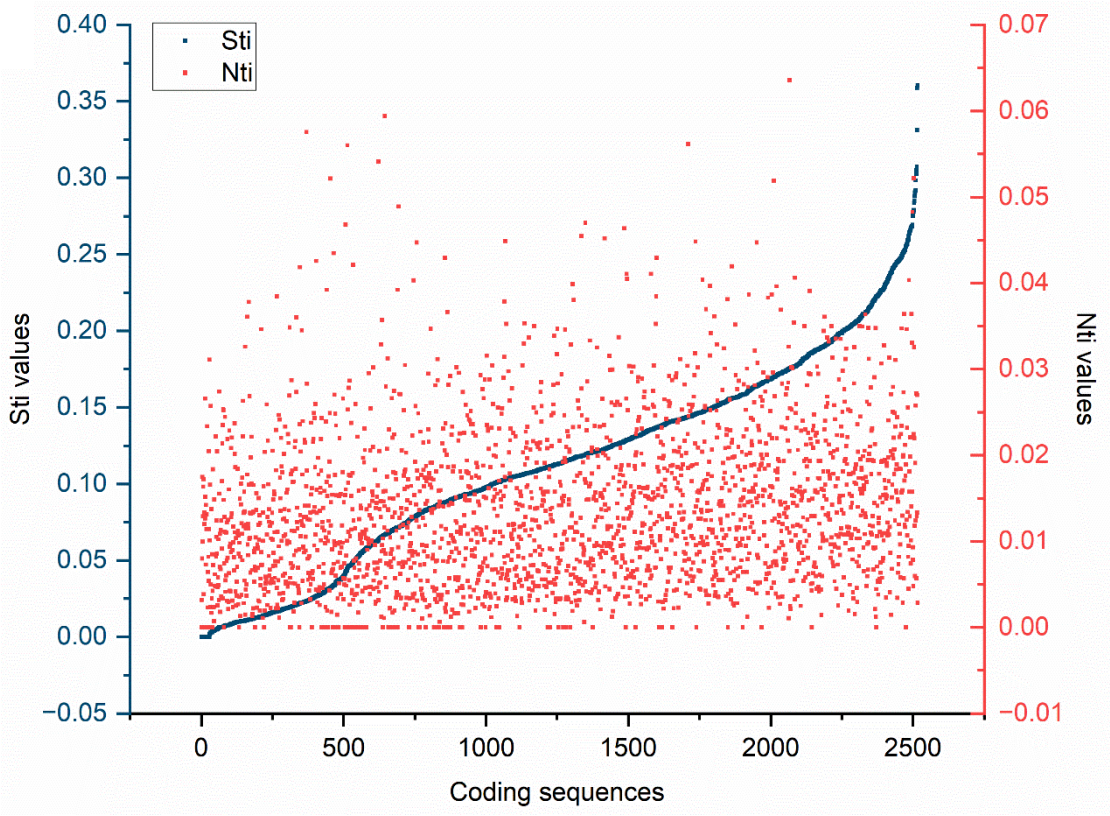

Sti
Nti
Sti values
Nti values
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
0
500
1000
1500
2000
2500
Coding sequences

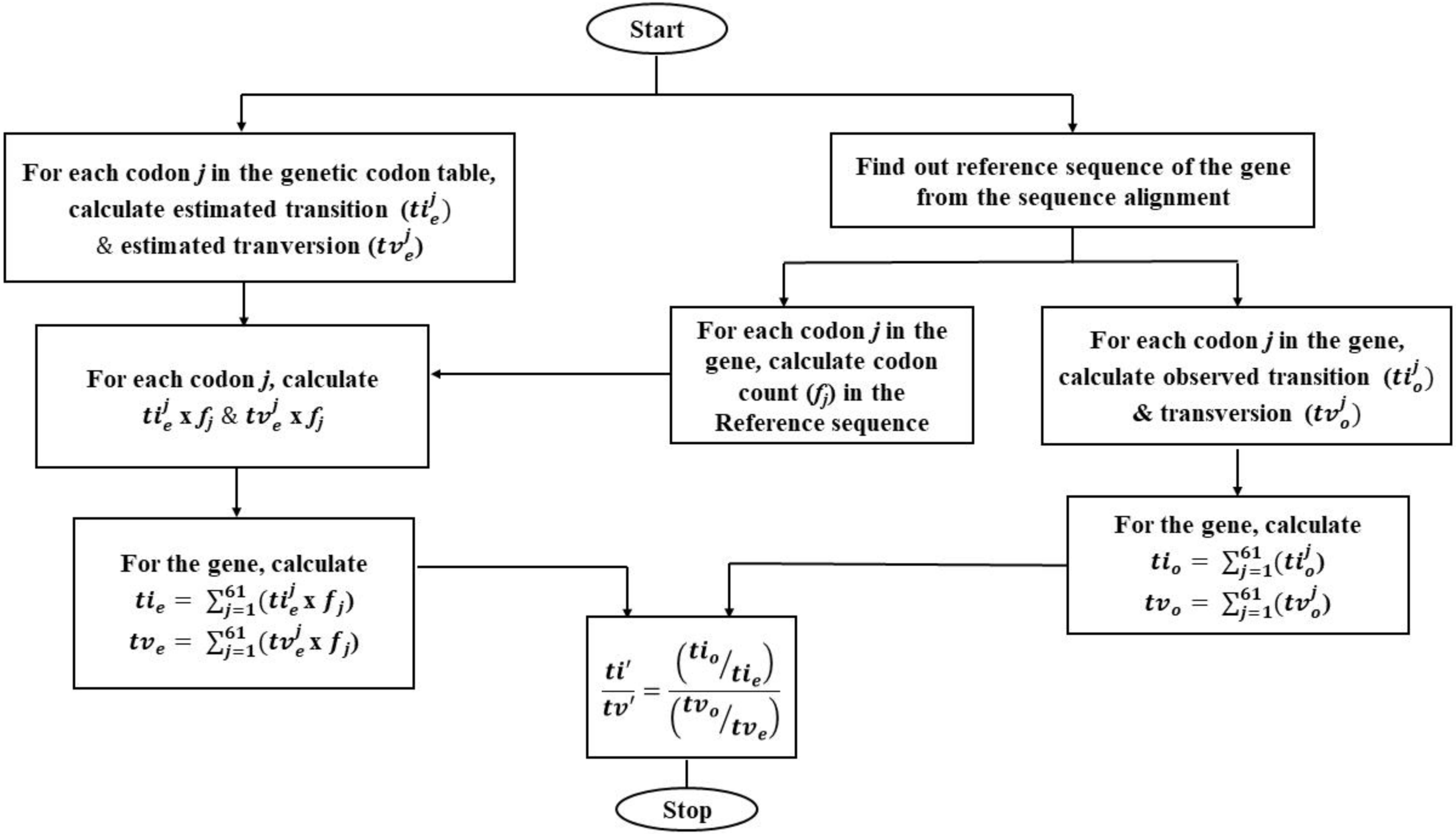
Fig. 4. (a) The figure illustrates the box-plot of the decreasing variation dynamics of *Sti*, *Stv*, *Nti*, and *Ntv* values. The mean *Sti* and *Stv* were observed as 0.113 and 0.031 respectively, similarly the mean *Nti* and *Ntv* values were observed as 0.014 and 0.004. *Sti* was observed to be highly frequent than *Stv* and *Nti* was observed to be highly frequent than *Ntv*. There was a significant difference between *Sti* and *Stv* values as well as between *Nti* and *Ntv* values(p<0.01). (b) The figure illustrates the Sti and Nti values of coding sequences used in the study through a double y-axis plot. The left y-axis shows the Sti values of coding sequences, and the right y-axis shows the Nti values of the coding sequences. The x-axis shows the numbers of coding sequences considered in this study. We have arranged the Sti values in an increasing order. The black line represents the Sti values, and the red dots represent the corresponding Nti values of a coding sequence. It is evident that many coding sequences with similar Sti values have a wide range of Nti values. The Sti and Nti values may represent the mutation and selection relation in a coding sequence.

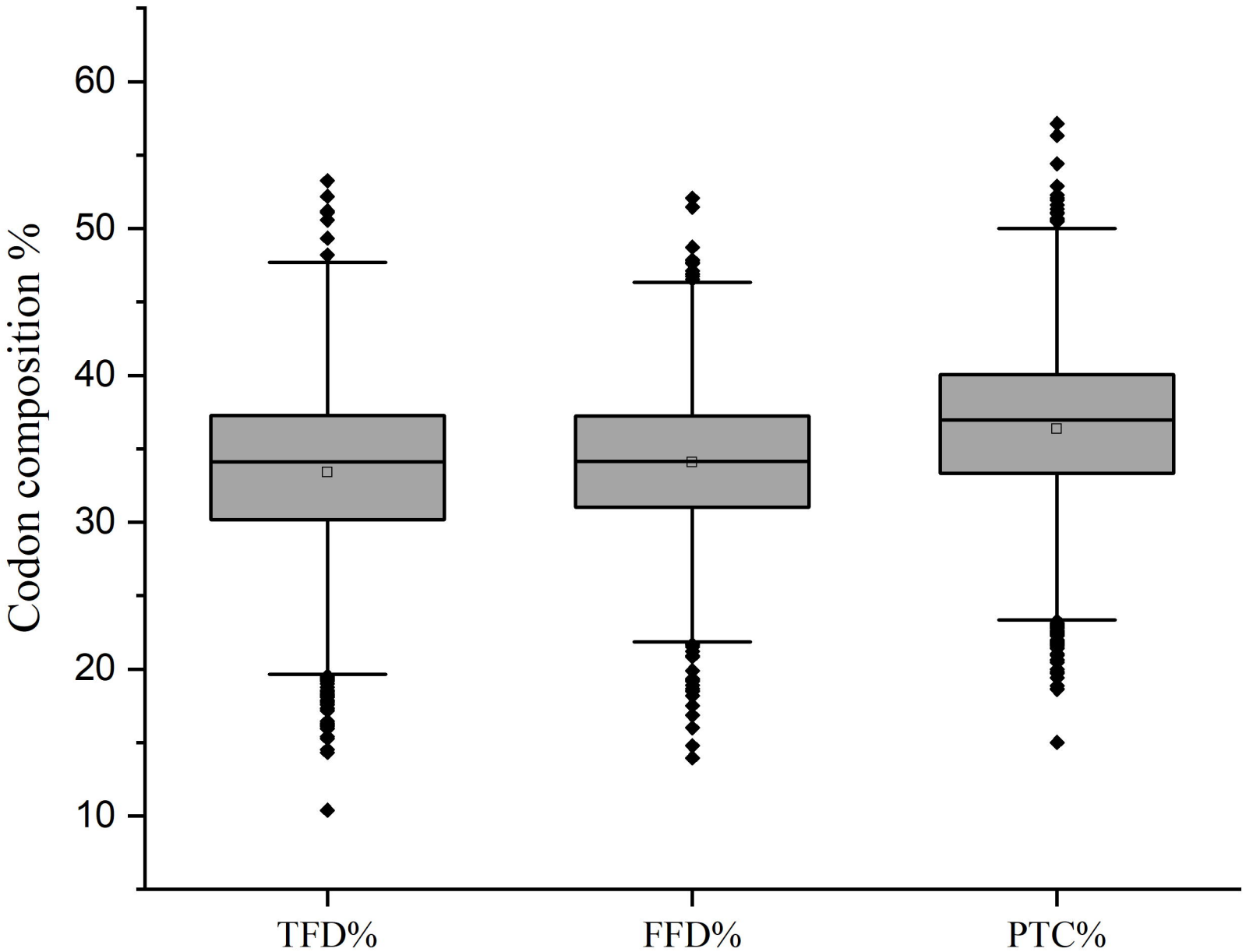
Table 1. Summary of the 2516 coding sequences selected for the present study. It represents the minimum to maximum range of different parameters like GC%, TFD/FFD, PTC% and the mean and standard deviation (S.D.) of the TFD%, FFD% and PTC%

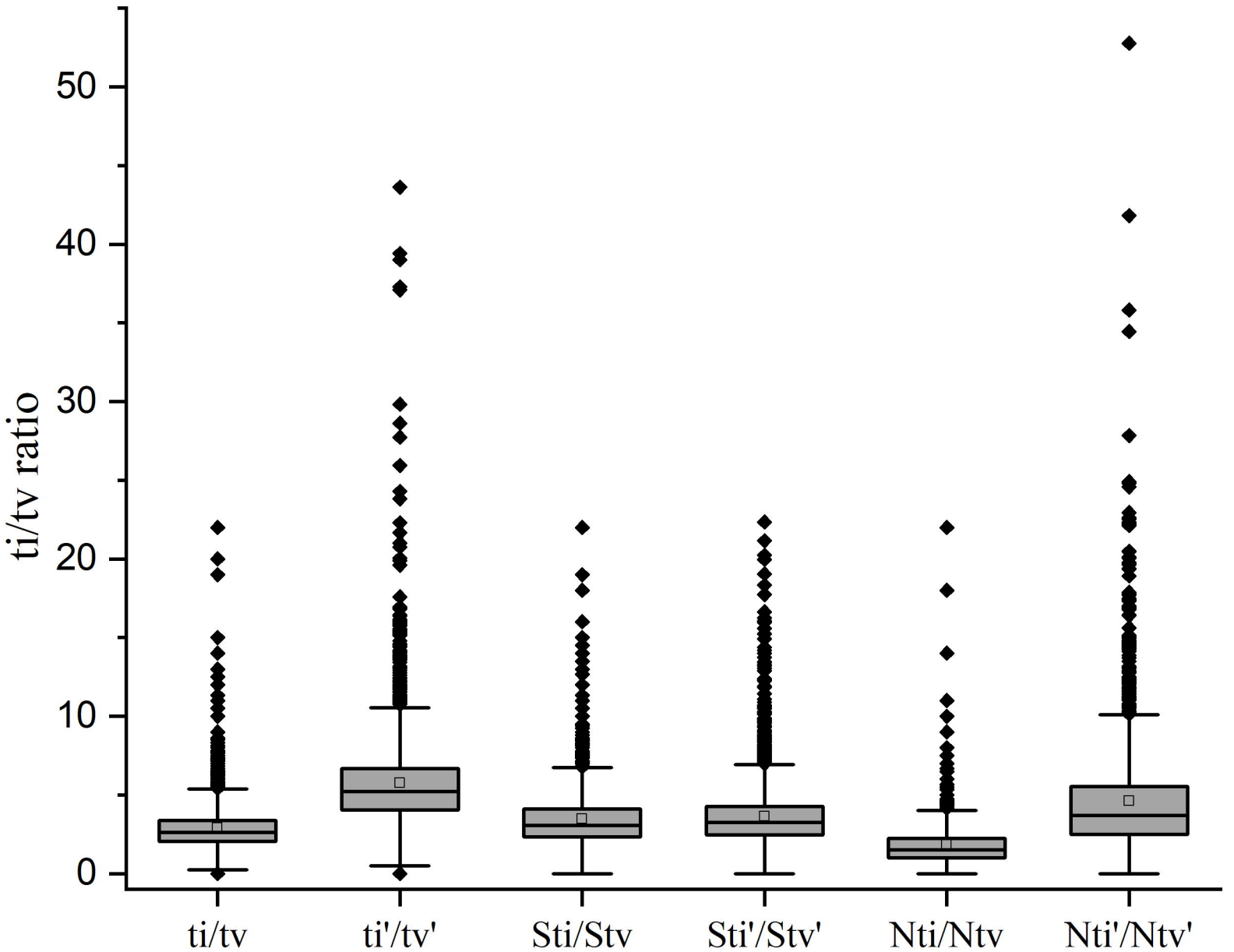
Genes considered	Size bp (Min-Max)	GC% (Min-Max)	TFD/FFD (Min-Max)	PTC% (Min-Max)	TFD%	FFD%	PTC%	
					33.384	34.084	36.350	Mean
2516	93-4554	32.95-63.80	0.22-3.33	15.00-57.14	5.726	4.761	5.625	S. D

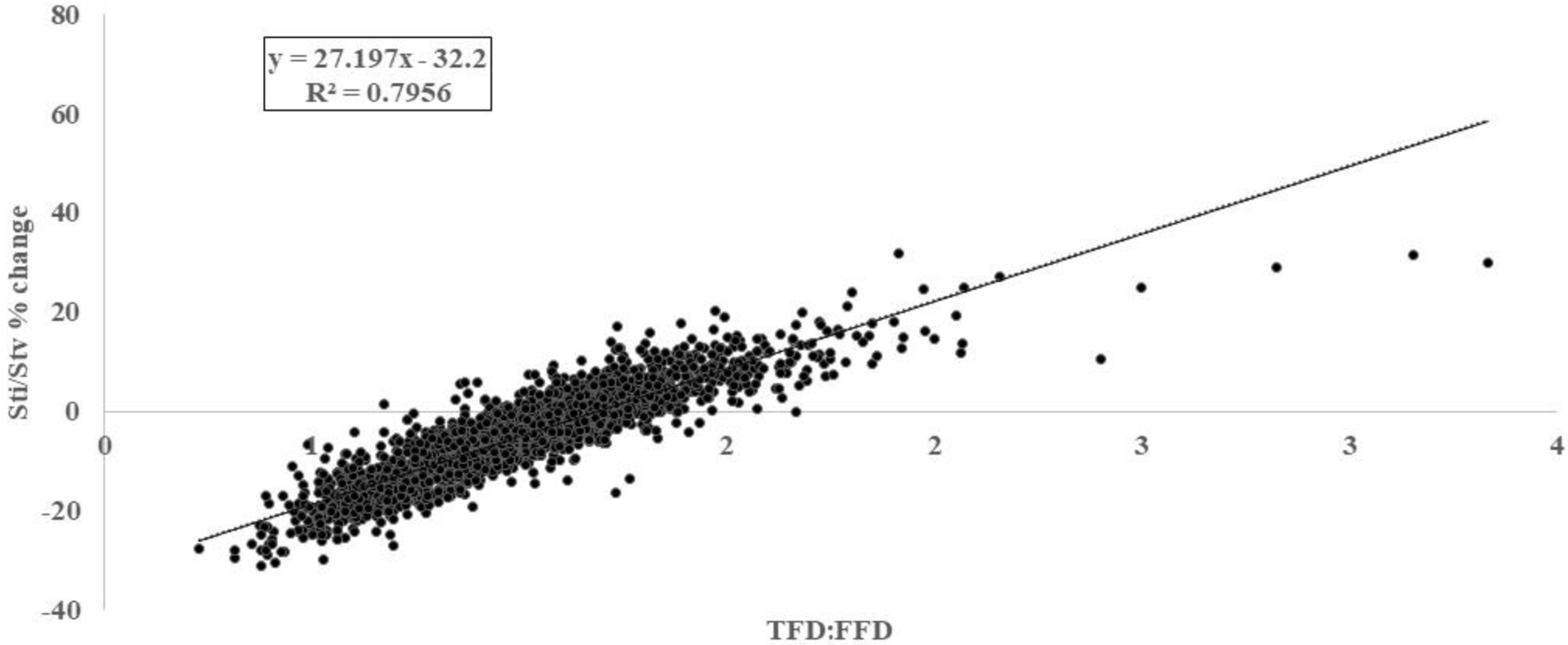
Table 2. Comparative Pearson correlation coefficient study among Sti, Sti, Nti and Ntv through conventional and improved estimator

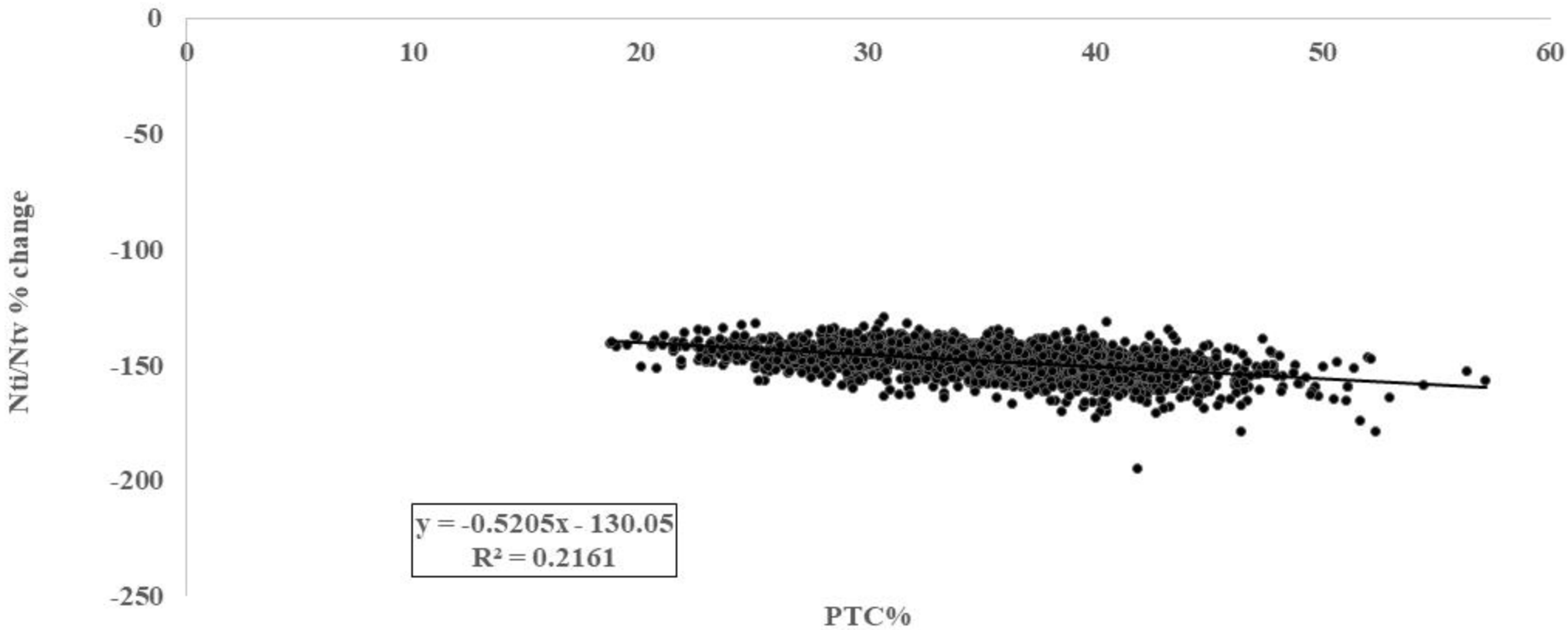
Pearson correlation coefficient	Conventional method	Improved method
	Observed values	Normalized values (Obs/Est)
Sti & Stv	0.920	0.795
Nti & Ntv	0.789	0.612
Sti & Nti	0.527	0.192
Stv & Ntv	0.544	0.257

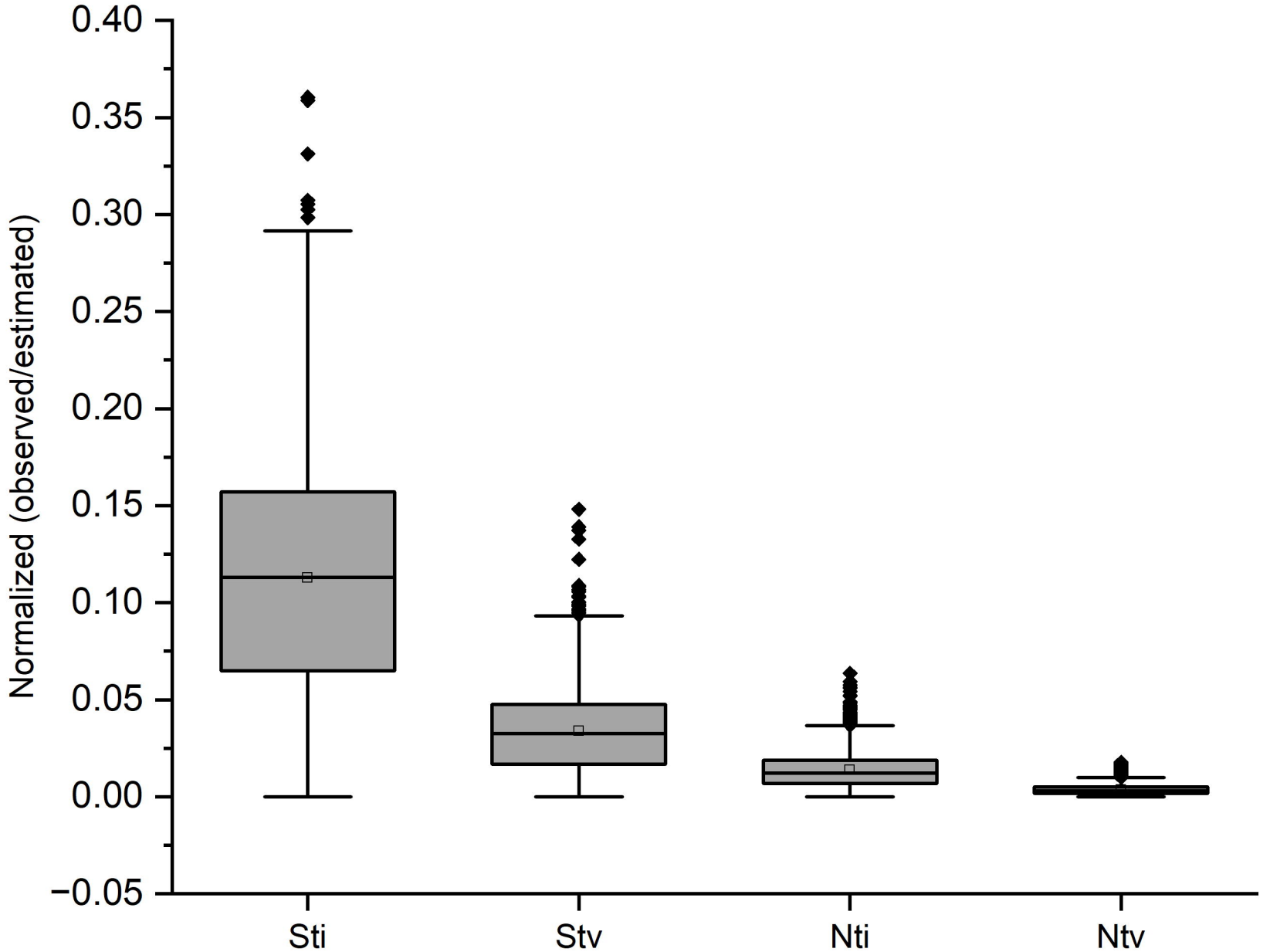

Start
For each codon j in the genetic codon table, calculate estimated transition (ti_e^j) & estimated tranversion (tv_e^j)
Find out reference sequence of the gene from the sequence alignment
For each codon j in the gene, calculate codon count (f_j) in the Reference sequence
For each codon j in the gene, 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

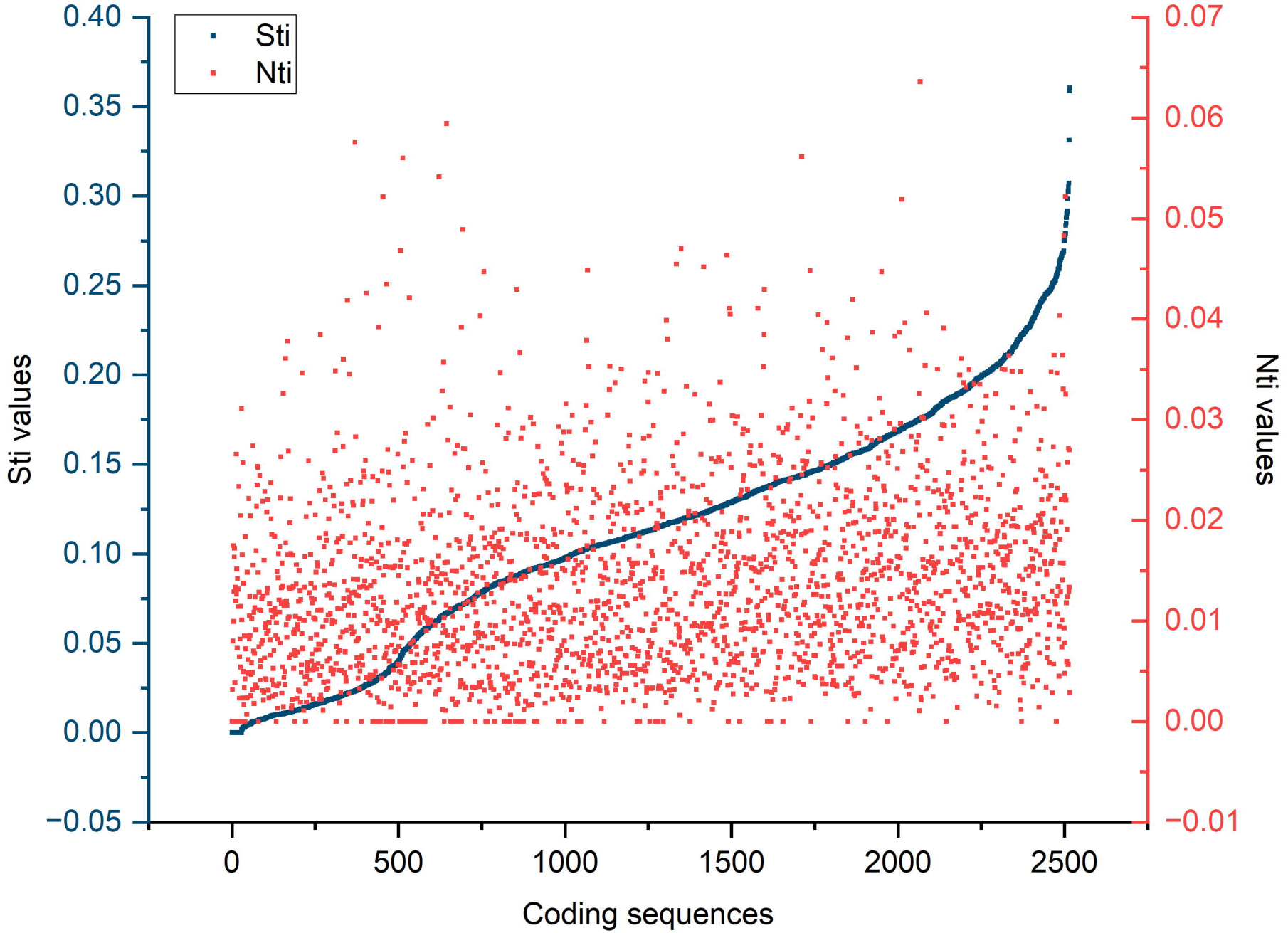

Codon composition %
60
50
40
30
20
10
TFD%
FFD%
PTC%


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


Sti
Nti
Sti values
Nti values
Coding sequences
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
0
500
1000
1500
2000
2500