Codon degeneracy as well as pretermination codons influence transition to transversion ratio in coding sequences

Pratyush Kumar Beura[1], Ruksana Aziz[1], Piyali Sen[2], Saurav Das[1], Nima Dondu Namsa[1,5], Edward J Feil[4], Siddharatha Sankar Satapathy*[2,5], Suvendra Kumar Ray*[1,5]

Departments of [1]Molecular Biology and Biotechnology, [2] Computer Science and Engineering, [5]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

Transition to transversion ratio (*ti/tv*) is important in understanding the mutational bias in the evolution of an organism. Since synonymous polymorphism for *two-fold degenerate* codons depends only on transitions, the *ti/tv* ratios of *two-fold degenerate* codons and *four-fold degenerate* codons are expected to be different for synonymous as well as non-synonymous mutations. Therefore, in any organism, the *ti/tv* ratio is likely to vary among coding sequences having differences in degenerate codon composition. In addition to the above, compositional variation of pretermination codon among the coding sequences is also likely to influence the *ti/tv* ratio because purifying selection is stronger on non-synonymous substitutions leading to termination codons. In this study we have improved the method of finding the ratio between transition to transversion, synonymous transition to synonymous transversion and non-synonymous transition to non-synonymous transversion and developed a computational

algorithm to calculate the *ti/tv* values. Our analyses in 2516 coding sequences in the *E. coli* genome suggested a strong correlation of codon degeneracy with synonymous transition to transversion ratio. Additionally, we have observed a moderate correlation of pretermination codons with non-synonymous transition to transversion ratio. We believe that the improved method developed in this study can be potentially employed in finding the *ti/tv* ratio in coding sequences of other genomes.

Keywords

Substitution mutation, Nucleotide polymorphism, Transition to transversion ratio, Codon degeneracy, Pretermination codon

Introduction

In the universal genetic codon table, of the 64 codons, 61 codons encode for 20 amino acids and 3 codons act as termination signal in the protein synthesis process. The redundancy or the degeneracy nature of the genetic codon suggests that a single amino acid can be encoded by multiple codons. There exists a different magnitude of degeneracy among the 61 sense codons ranging between 0 to 6. For example, AUG and UGG encode Met and Trp, respectively are considered as zero degeneracy. Codons such as UUU and UUC encode Phe and belong to the category of *two-fold degenerate* (TFD) codons. AUU, AUC and AUA encode Ile and belong to the category of three-fold degenerate codons. Codons such as GGU, GGC, GGA and GGG encode *four-fold degenerate* (FFD) codons. Codons such as CUN and UUR encode Leu and are known as *six-fold degenerate* codons. Because of different codon degeneracy the number of synonymous polymorphism and the number of non-synonymous polymorphism are not equal among the codons. Every codon can undergo 9 different single nucleotide polymorphisms

(SNPs), out of which 3 substitutions are transition (*ti*) and 6 substitutions are transversion (*tv*). In a standard genetic codon table out of 549 total SNPs involving 61 sense codons, only 134 SNPs are synonymous (24.4%). Among synonymous (S*ti* and *Stv*) and non-synonymous (*Nti* and *Ntv*) polymorphisms, the theoretically estimated ($_e$) numbers of *Sti*, *Stv*, *Nti* and *Ntv* in a genetic codon table are 62, 72, 121 and 294, respectively (Table S1). Yet the observed synonymous polymorphisms are always higher than the expected proportion across genes (Tamura 1992; Wolfe et al. 1989; Moriyama and Powell 1997; Ngandu et al. 2008) as purifying selection is stronger on non-synonymous polymorphisms than the synonymous polymorphism (Hurst and Pál 2001).

Due to different degeneracy, there is difference among the codons regarding *Sti*, *Stv*, *Nti* and *Ntv* as explained below. A TFD codon has only single synonymous transition (*Sti*), zero synonymous transversion (*Stv*), two non-synonymous transition (*Nti*) and six non-synonymous transversion (*Ntv*). In contrast, a FFD codon has one *Sti*, two *Stv*, two *Nti* and four *Ntv* (Fig. S1A). Hence, it is expected that a gene having higher proportion of FFD codons will obtain more substitutions than TFD codons. The differences in the composition of TFD and FFD among the coding sequences is likely to influence the transition to transversion ratio. It is also known that transition polymorphism is more frequent than a transversion polymorphism (Seplyarskiy et al. 2012; Duchêne et al. 2015; Lewis Jr et al. 2016; Stoltzfus and Norris 2016; Lyons and Lauring 2017; Schroeder et al. 2017; Sen et al 2021), which is known as Kimura's two parameter model (Kimura 1980). Accordingly, it has been reported recently that synonymous polymorphism in TFD codons is relatively higher than that in the FFD in Escherichia coli (Aziz et al 2022).

Among 61 sense codons there are 18 pretermination codons (PTC) which can lead to stop codons by a single substitution at certain positions in the PTC (Modiano et al. 1981) (Fig. S1B).

Out of 18 PTC, UUA, UCA, UAU, UAC and UGG are such 5 PTC in which out of nine possible SNPs, two SNPs can possibly lead to stop codons. For example, of the remaining 13 PTC, out of 9 possible substitutions, 1 SNP in UUG can lead to stop codon. Hence, non-synonymous substitution count is different for UUA and UUG (Fig. S1C). Therefore, PTC compositions are also likely to influence the non-synonymous polymorphism because purifying selection is stronger on non-sense substitutions (van den Eynden et al. 2016; Morales et al. 2021; Li et al. 1981). Because of the variation in the theoretically estimated *Sti*, *Stv*, *Nti* and *Ntv* among codons and codon compositional differences in the coding sequences affects the *ti* to *tv* ratio implicating the importance of codon degeneracy. By acknowledging the disproportion of polymorphism patterns affected by codon degeneracy as well as PTC, now the theoretically estimated polymorphisms can be re-calculated in the form of *Sti*, *Stv*, *Nti* and *Ntv* for each codon (Table 1). Also, we have summarized the total numbers of theoretically estimated synonymous *ti* and *tv* and non-synonymous *ti* and *tv* in the genetic codon table (Table 2).

In this study we observed that compositional difference of degenerate codons (TFD and FFD) and PTC are influencing the *ti/tv* ratio in a gene. Further, we found a correlation between TFD: FFD and *Sti/Stv* values. We found that the percentage of PTC can influence the *Nti/Ntv* ratio. Based on the work by Ina 1998 we tried to establish the biological significance of codon composition and *ti/tv* ratio across 2516 coding sequences of *E. coli*. The new method developed in the present study may be applied to study different *ti/tv* (*Sti/Stv* and *Nti/Ntv*) across coding sequences.

Methods

Coding sequences of *E. coli* genome

In this study, we carried out a SNPs based computational analysis of 2516 coding sequences across 157 strains of *E. coli* (Thorpe et al. 2017). The information regarding the dataset selected for the present study is provided in Table S2. Although the public database has ~4500 coding sequences but we excluded coding sequences showing length variation and improper alignment/annotation. Codons having polymorphisms in more than one position or polymorphisms leading to ambiguous nucleotides were not considered in the analyses of the polymorphisms in this study. A detailed procedure is represented in Table S3 using example of a hypothetical sequence. Further, the observed SNPs (SNP_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 (Table S4). After finding observed polymorphism values for a gene, we calculated the theoretically estimated polymorphism values ti_e, Stv_e Nti_e, Ntv_e, ti_e and tv_e. The reference sequence which was previously derived from the sequence alignments was used to calculate the codon count of each gene. After finding the codon count, each codon count value was multiplied by theoretically estimated polymorphism 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

polymorphism values for 61 sense codons to get the total theoretically estimated ($_e$) individual substitution values (Sti_e, Stv_e, Nti_e and Ntv_e) for the gene (Table S5). 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 (Table S6). A detailed workflow is provided in Fig. 1.

Improved equations

We proposed 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_p 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 ti/*tv* ratios for all the genes using a program written in C language.

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 3.

Statistical analyses

Between the conventional and improved $\frac{ti}{tv}$ values different statistical methods were applied to show the performance and accuracy of improved method. OriginPro 2022, OriginLab Corporation, Northampton, MA, USA was used to draw the Box-plot as well as to perform the Mann-Whitney test (Nachar 2008) to find out the *p*-value. Correlation plot was drawn for R^2 value and Pearson's correlation value was also performed between the $\frac{ti}{tv}$ values obtained using the conventional and improved method (Benesty et al. 2009).

Results

Compositional analysis of TFD, FFD, and PTC codons in 2516 coding sequences of *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 (Table S2). The percent (%) compositional differences in the 2516 coding sequences represented as TFD%, FFD%, and PTC% is shown in a box-plot (Fig. 2). 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 as 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 *atpE* as 15.00% and in *chaB* as 57.14% (Fig. 2). The compositional variation (TFD, FFD and PTC) observed in coding sequences were considered to validate the accuracy of different *ti/tv* ratio using the conventional and improved method.

Impact of codon degeneracy and PTC on $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ values

The complete set of observed and estimated value of different polymorphisms is listed in Table S2. 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 $\frac{Nti'}{Ntv'}$) (Table S2). A box-plot showing the comparison between the values of different *ti/tv* ratio obtained through the conventional and improved method is represented in Figure 3. We also calculated the percentage (%) change in $\frac{Sti}{Stv}$

and $\frac{Nti}{Ntv}$ values obtained from conventional to improved method. The minimum % change in $\frac{Sti}{Stv}$ between conventional and improved method was observed in *garP* as -30.99 and maximum in *yfhL* as 30.76. Similarly, the minimum % change in $\frac{Nti}{Ntv}$ between conventional and improved method was observed in *thiS* as -194.675 and maximum in *bhsA* as -129.600. The detailed % change in $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ is represented in Table S2. While it is evident that, TFD codons and FFD codons have different probabilities of obtaining synonymous polymorphisms, 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 Pearson correlation coefficient was performed to study the correlation of codon degeneracy (TFD: FFD) and PTC% on the % change in $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$. 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 coefficient. 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 strong correlation between TFD: FFD and % change in $\frac{Sti}{Stv}$ suggested that there is an influence of the TFD: FFD codons in overall synonymous $\frac{ti}{tv}$. Similarly, the moderate negative correlation also suggested the contribution of PTC codons in overall non-synonymous $\frac{ti}{tv}$. The results were found to be statistically different 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 4A & 4B). 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. Therefore, the findings of correlation analysis further support the contribution of TFD: FFD and PTC% in $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$, respectively. We showed the role of codon composition in estimating the $\frac{ti}{tv}$ ratio using the improved method and this parameter was not accounted in the conventional method.

Discussion

Various methods have been used in the estimation of the *ti* to *tv* ratio by the molecular evolutionary biologist to understand the mutation as well as selection bias. The phylogeny-based method and Bayesian methods have been implemented by researchers in the study of *ti/tv* (Purvis and Bromham 1997; Huelsenbeck and Ronquist 2001). The work by Ina, 1998 is the most recent one, which has been extended further in this study by accounting codon degeneracy and PTC. We have demonstrated that the coding sequences are highly variable in terms of compositions of different degenerate codons and PTC. The improved equations used in this study is 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 *Sti/Stv* ratio among the coding sequences in relation to their compositional differences (TFD, FFD and PTC) has not been reported previously. Our method also accurately attributes to the importance of PTC in the estimation of *Nti/Ntv* in coding sequences. A species-specific dependence of the deleterious nature of *Nti* and *Ntv* in coding sequences have been

reported recently and the plausible contribution of PTC was not accounted in the study (Zou and Zhang 2021). The observation in the form of codon degeneracy and PTC in estimating *ti/tv* ratio is in concordance with the recent findings on the role of codon degeneracy and PTC in coding sequences to estimate *dN/dS* values (Aziz et al. 2022).

We believe that our improved method is more compatible in the calculation of different *ti* to *tv* in coding sequences. By accounting the possible polymorphism of individual codons, we have done the normalization of the *ti* to *tv* ratio. On normalization using the improved method, if the $\frac{Sti'}{Stv'}$ values are different between two genes, this difference might be attributed to codon degeneracy (TFD:FFD ratio) and gene expression values. Similarly, if $\frac{Nti'}{Ntv'}$ is different between two genes, this difference might be attributed to codon degeneracy (TFD:FFD ratio), PTC% and gene expression values. Though in this work we did not consider the influence of higher codon degeneracy in estimating the observed *Sti* value (unpublished work), it is ascertained that TFD codons only undergo *Sti*, whereas FFD codons can exhibit synonymous polymorphism by *Sti* and *Stv*. However, two genes with almost similar estimated values of polymorphism can have different ratios of *ti* to *tv* because of different intrinsic and extrinsic factors like gene expression, selection, and acceptance of one amino acid instead of the other. Since the magnitude of synonymous polymorphism is limited for a codon in the genetic codon table, the non-synonymous polymorphism for a codon has an ample number of possibilities of its acceptance by the organism.

The findings of the present work provide an insight in the analyses of different degenerate codons and PTC composition in estimating the *ti/tv* ratio across the coding sequences. The

differential usage of amino acid, gene expression, and the evolutionary significance of codon degeneracy and PTC composition is yet to be investigated.

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.

Author Disclosure Statement

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

Funding Information

The authors received no specific funding for this work.

References

Aziz R, Sen P, Beura PK, 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

Benesty J, Chen J, Huang Y, Cohen I (2009) Pearson correlation coefficient. In: Noise reduction in speech processing. Springer, pp 1–4

Duchêne S, Ho SYW, Holmes EC (2015) Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. BMC Evol Biol 15:1–10

Huelsenbeck J, Ronquist F (2001) MrBayes: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Hurst LD, Pál C (2001) Evidence for purifying selection acting on silent sites in BRCA1. TRENDS in Genetics 17:62–65

Ina Y (1998) Estimation of the transition/transversion ratio. J Mol Evol 46:521–533

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

Lewis Jr CA, Crayle J, Zhou S, et al (2016) Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. Proceedings of the National Academy of Sciences 113:8194–8199

Li W-H, Gojobori T, Nei M (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237–239

Lyons DM, Lauring AS (2017) Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. Mol Biol Evol 34:3205–3215

Modiano G, Battistuzzi G, Motulsky AG (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? Proceedings of the National Academy of Sciences 78:1110–1114

Morales AC, Rice AM, Ho AT, et al (2021) Causes and consequences of purifying selection on SARS-CoV-2. Genome Biol Evol 13:evab196

Moriyama EN, Powell JR (1997) Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. J Mol Evol 45:378–391

Nachar N (2008) The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. Tutor Quant Methods Psychol 4:13–20

Ngandu NK, Scheffler K, Moore P, et al (2008) Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol J 5:1–11

Purvis A, Bromham L (1997) Estimating the transition/transversion ratio from independent comparisons with an assumed phylogeny.J. Mol. Evol. 44, 112–119.

Schroeder JW, Randall JR, Hirst WG, et al (2017) Mutagenic cost of ribonucleotides in bacterial DNA. Proceedings of the National Academy of Sciences 114:11733–11738

Sen P, Aziz R, Deka RC, et al (2022) Stem region of tRNA genes favors transition substitution towards keto bases in bacteria. J Mol Evol 90:114–123

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

Stoltzfus A, Norris RW (2016) On the causes of evolutionary transition: transversion bias. Mol Biol Evol 33:595–602

Tamura K (1992) The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. Mol Biol Evol 9:814–825

Thorpe HA, Bayliss SC, Hurst LD, Feil EJ (2017) Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. Genetics 206:363–376

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

Wolfe KH, Sharp PM, Li W-H (1989) Rates of synonymous substitution in plant nuclear genes. J Mol Evol 29:208–211

Zou Z, Zhang j (2021) Are Nonsynonymous Transversions Generally More Deleterious than Nonsynonymous Transitions? Mol. Biol. Evol. 38(1):181–191

Figure_legend_Table

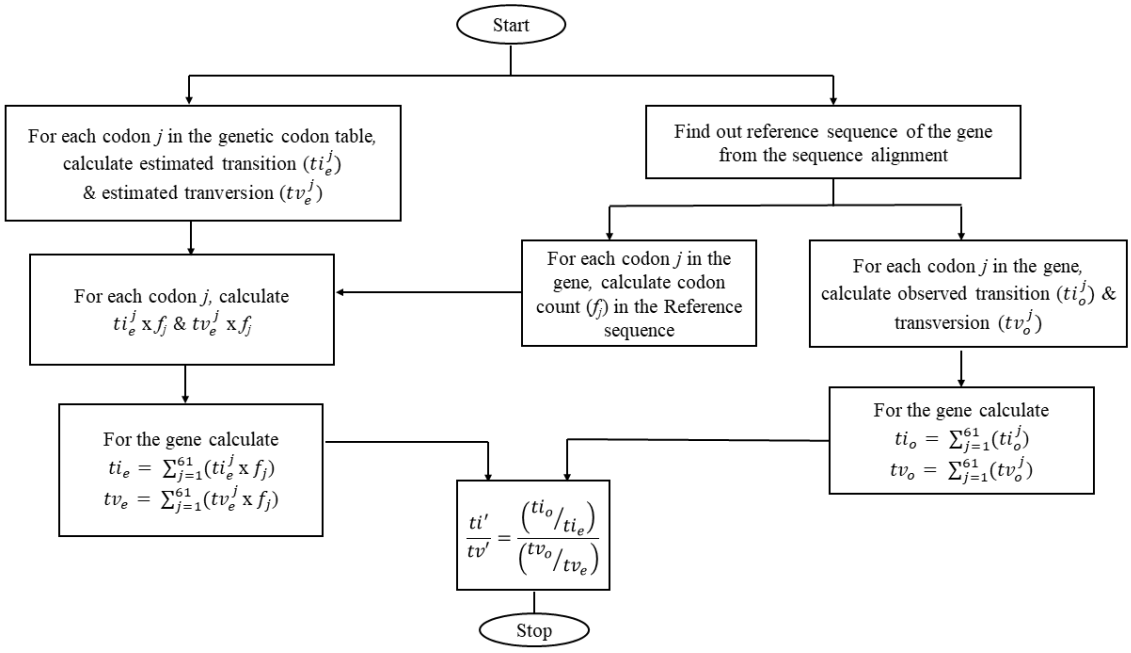


Fig. 1. Schematic representation of the workflow showing the steps to execute our improved method. By considering each codon as *j*, ti_e^j and tv_e^j can be calculated for each codon as presented in Table 1. The reference sequence needs to be calculated from the coding region (gene/query). From the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

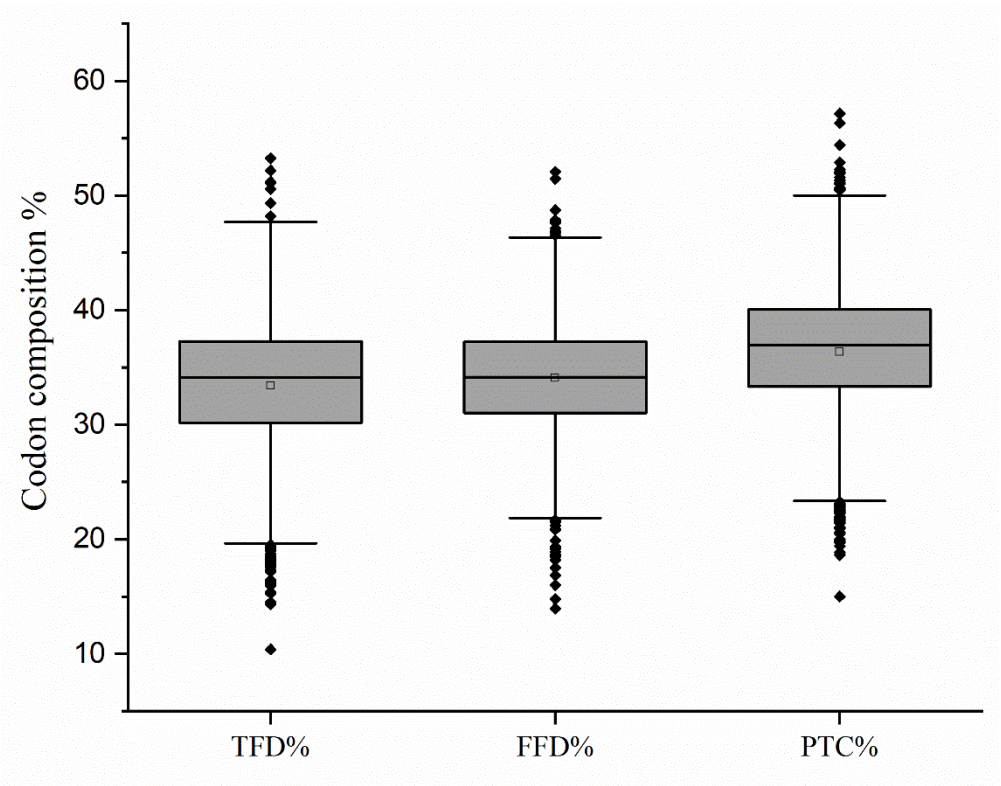


Fig. 2. The box-plot presents the percentage (%) of TFD, FFD and PTC codons in the coding sequences considered to validate the improved method proposed in this study. The minimum and maximum TFD% in the dataset was observed as 10.38 and 53.25, similarly the minimum and maximum FFD% in the dataset was observed as 13.95 and 52.05. The minimum and maximum PTC% in the dataset was observed as 15.00 and 57.14.

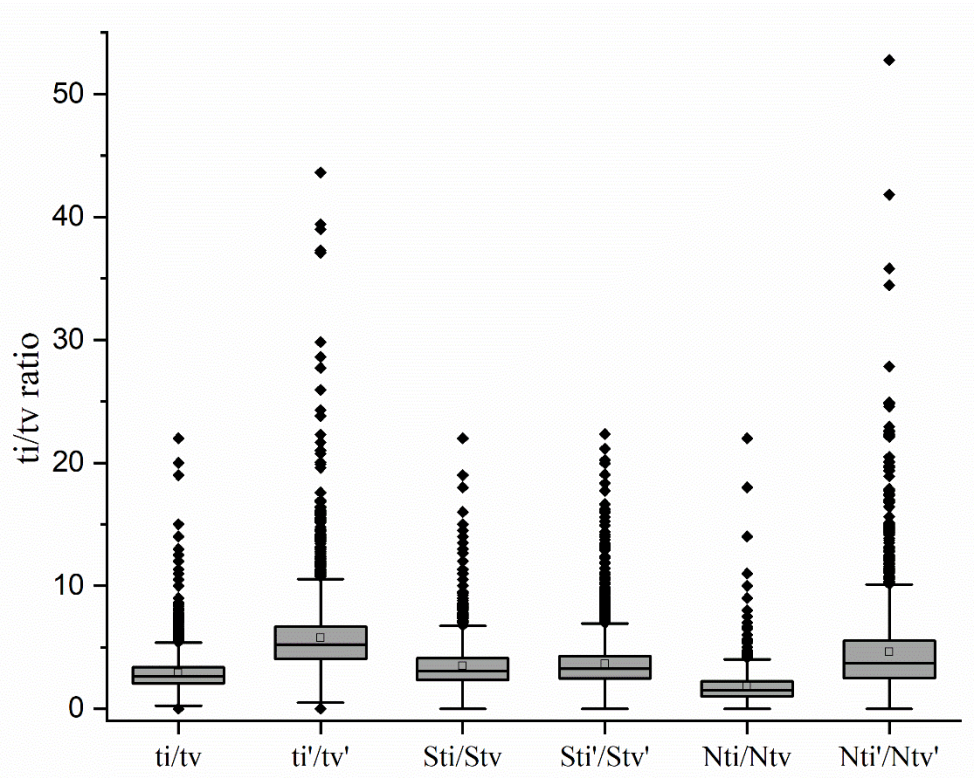
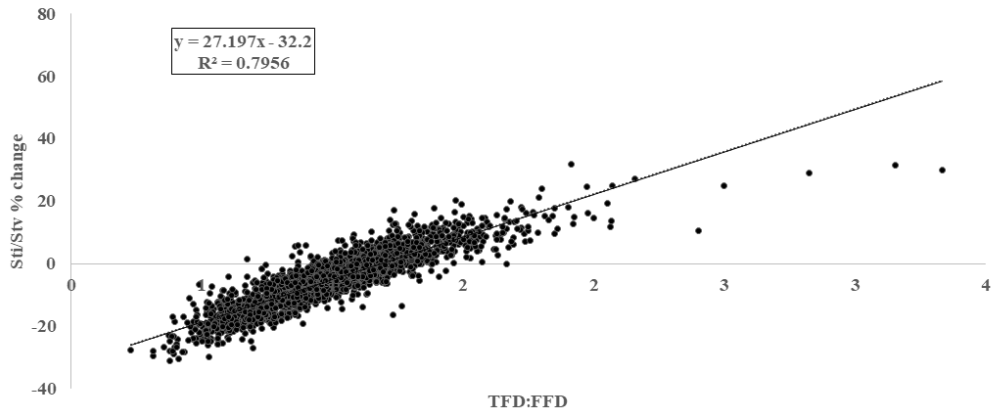


Fig. 3. The box-plot presents the different *ti/tv* ratio values. Here we have compared the different *ti/tv* ratios of the entire set of genes through conventional and improved method. $\frac{ti}{tv}$ shows the conventional method and $\frac{ti'}{tv'}$ shows the result of improved method. Similarly, $\frac{Sti}{Stv}$ 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

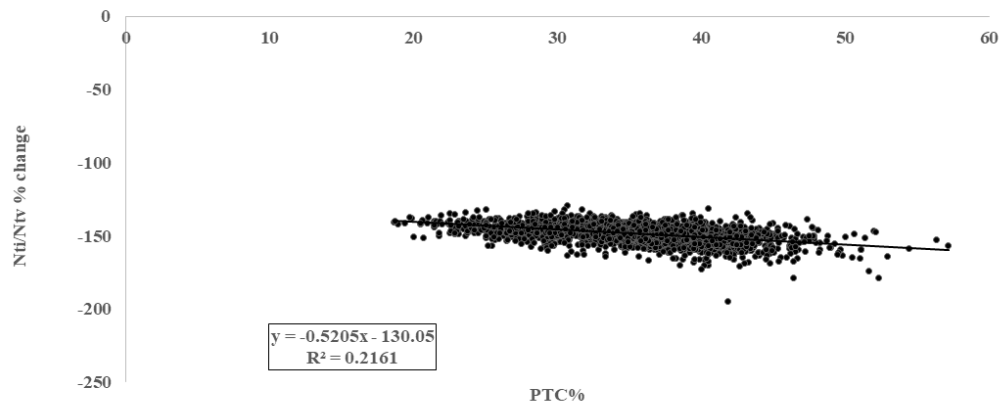


Fig. 4. Regression plot of coding sequence considered to validate the improved method. The x-axis shows the codon composition (TFD: FFD and PTC%) and y-axis shows the % change in $\frac{Sti}{Stv}$ and $\frac{\mathrm{Nti}}{\mathrm{Ntv}}$. (A) % change in $\frac{Sti}{Stv}$ is showing a strong correlation with the TFD: FFD with a R^2 value of 0.795. Among few of the outliers, it can be noted that, *yhfL*, *gadE* and *mntL* obtained the highest % change in $\frac{Sti}{Stv}$ value as 31.72,31.32 and 30.00 respectively, whereas the TFD: FFD codon was noted as 1.91, 3.15 and 3.33, respectively for the genes. Similarly, it can be noted that, *garP*, *proW* and

slyB obtained the lowest % change in $\frac{Sti}{Stv}$ value as -30.99, -30.57 and -29.68, respectively, whereas the TFD: FFD codon was noted as 0.38, 0.41 and 0.53, respectively for the genes. (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. While doing the similar analysis for outliers, the % change in $\frac{Nti}{Ntv}$ value and PTC% it was noted that, *thiS*, *sgrT* and *sbcC* obtained the lowest % change in $\frac{Nti}{Ntv}$ value as -194.67, -179.00 and -178.00, respectively, whereas the PTC% was noted as 41.79, 52.27 and 46.27, respectively for the genes. Similarly, bhsA, *gapA* and *yiaD* obtained the highest % change in $\frac{Nti}{Ntv}$ value as -129.67, -131.55 and -131.94, respectively, whereas the PTC% was noted as 30.68, 40.48 and 25.00, respectively for the genes.

Table 1. Theoretically estimated ($_e$) polymorphism for each codon by excluding the non-synonymous substitutions (*Nti* & *Ntv*) leading to stop codons through PTC

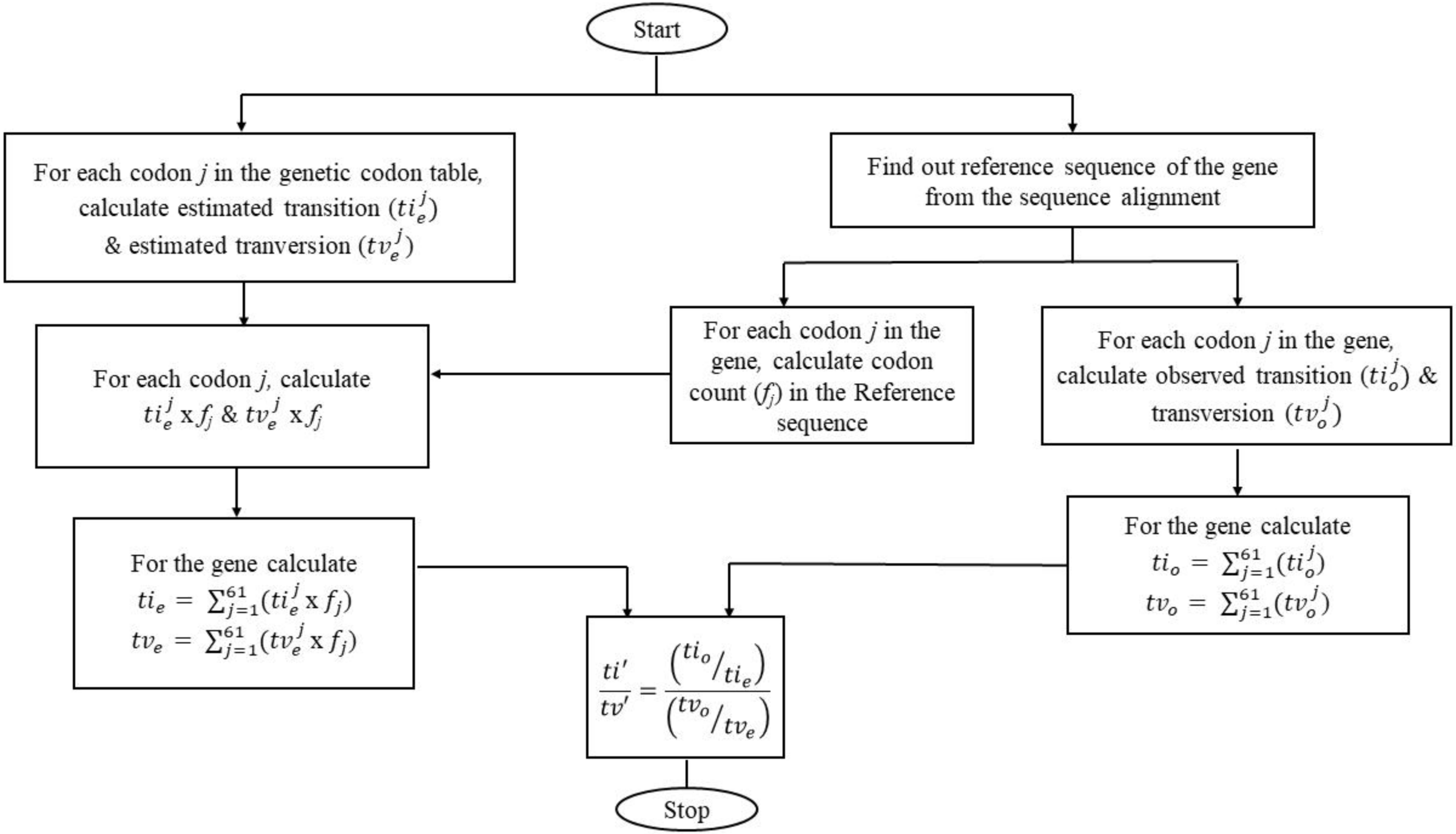
Codon	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Sti_e	Stv_e	Nti_e	Ntv_e
UUU	1	0	2	6	UCU	1	2	2	4	UAU	1	0	2	4	UGU	1	0	2	5
UUC	1	0	2	6	UCC	1	2	2	4	UAC	1	0	2	4	UGC	1	0	2	5
UUA	2	0	1	4	UCA	1	2	2	2	UAA	X	X	X	X	UGA	X	X	X	X
UUG	2	0	1	5	UCG	1	2	2	3	UAG	X	X	X	X	UGG	0	0	1	6
CUU	1	2	2	4	CCU	1	2	2	4	CAU	1	0	2	6	CGU	1	2	2	4
CUC	1	2	2	4	CCC	1	2	2	4	CAC	1	0	2	6	CGC	1	2	2	4
CUA	2	2	1	4	CCA	1	2	2	4	CAA	1	0	1	6	CGA	1	3	1	3
CUG	2	2	1	4	CCG	1	2	2	4	CAG	1	0	1	6	CGG	1	3	2	3
AUU	1	1	2	5	ACU	1	2	2	4	AAU	1	0	2	6	AGU	1	0	2	6
AUC	1	1	2	5	ACC	1	2	2	4	AAC	1	0	2	6	AGC	1	0	2	6
AUA	0	2	3	4	ACA	1	2	2	4	AAA	1	0	2	5	AGA	1	1	2	4
AUG	0	0	3	6	ACG	1	2	2	4	AAG	1	0	2	5	AGG	1	1	2	5
GUU	1	2	2	4	GCU	1	2	2	4	GAU	1	0	2	6	GGU	1	2	2	4
GUC	1	2	2	4	GCC	1	2	2	4	GAC	1	0	2	6	GGC	1	2	2	4
GUA	1	2	2	4	GCA	1	2	2	4	GAA	1	0	2	5	GGA	1	2	2	3
GUG	1	2	2	4	GCG	1	2	2	4	GAG	1	0	2	5	GGG	1	2	2	4

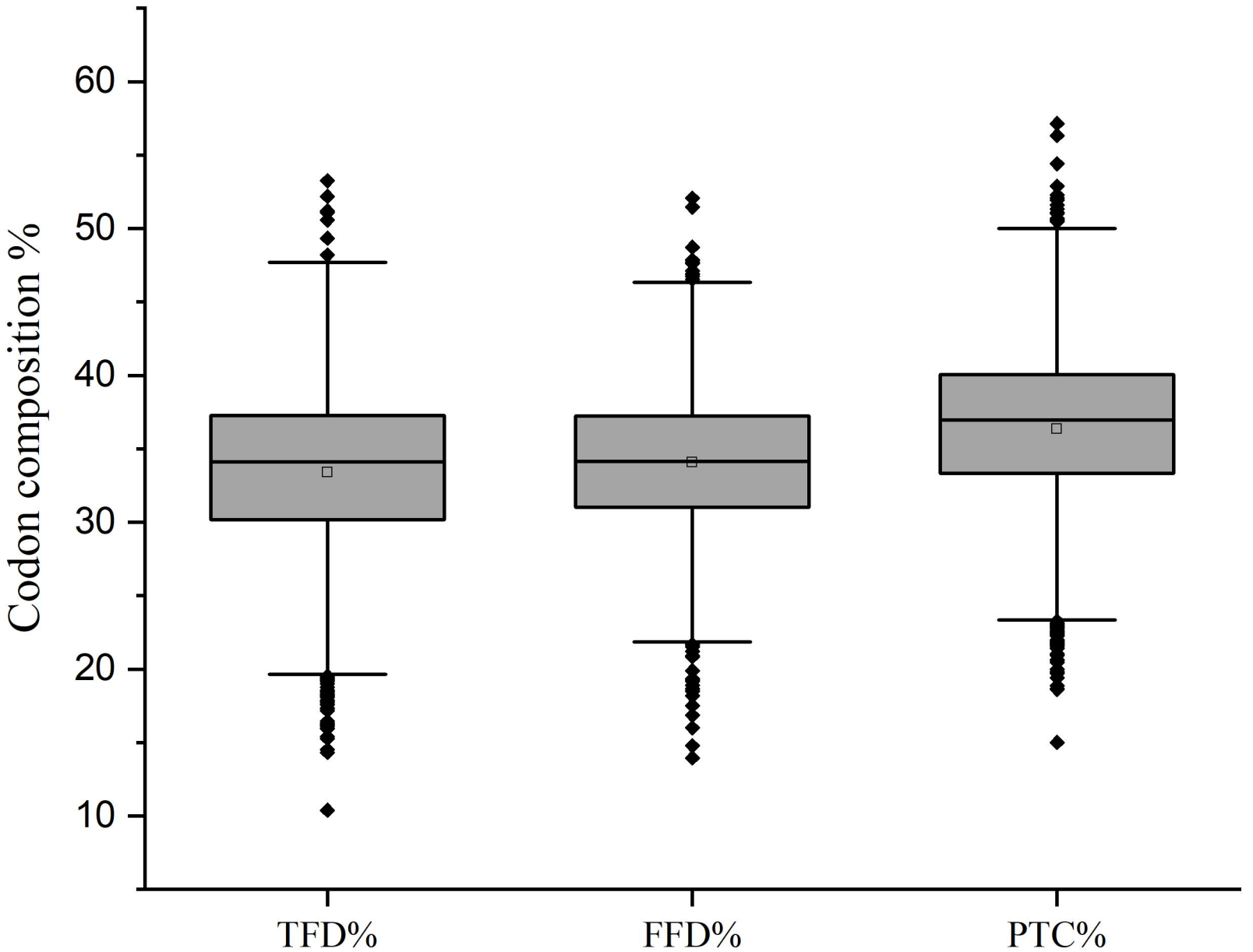
Table 2. Total estimated synonymous *ti* (*tv*) and non-synonymous *ti* (*tv*) substitutions in the genetic codon table. Asterisk(*) showed the total numbers of substitutions and ratio while excluding the substitutions that lead to stop codons in PTC

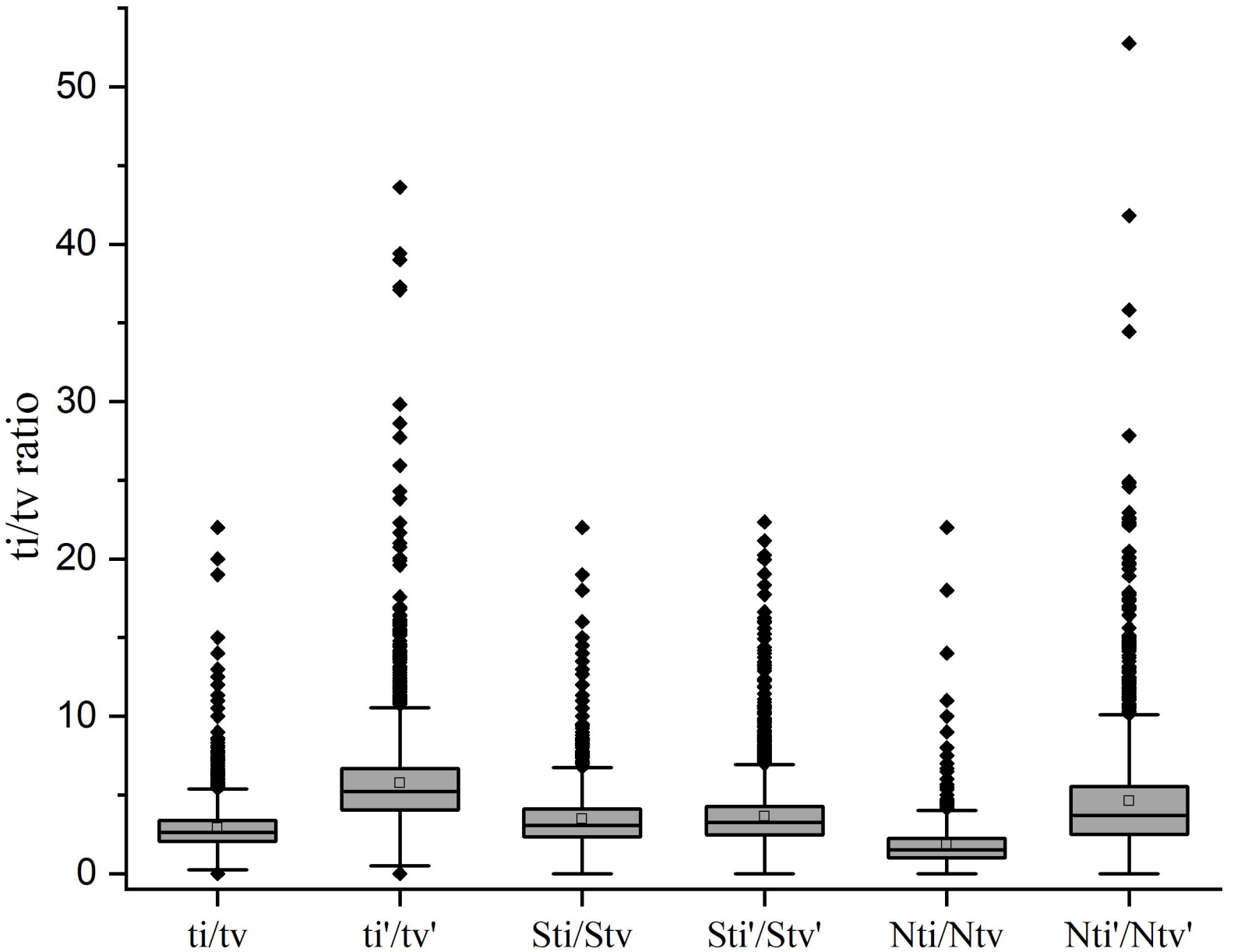
ti & *tv*	Synonymous	Non-synonymous (*)	Total (*)
ti	62	121 (116)	183 (178)
tv	72	294 (276)	366 (348)
ti/tv	0.861	0.412 (0.420)	0.500 (0.511)

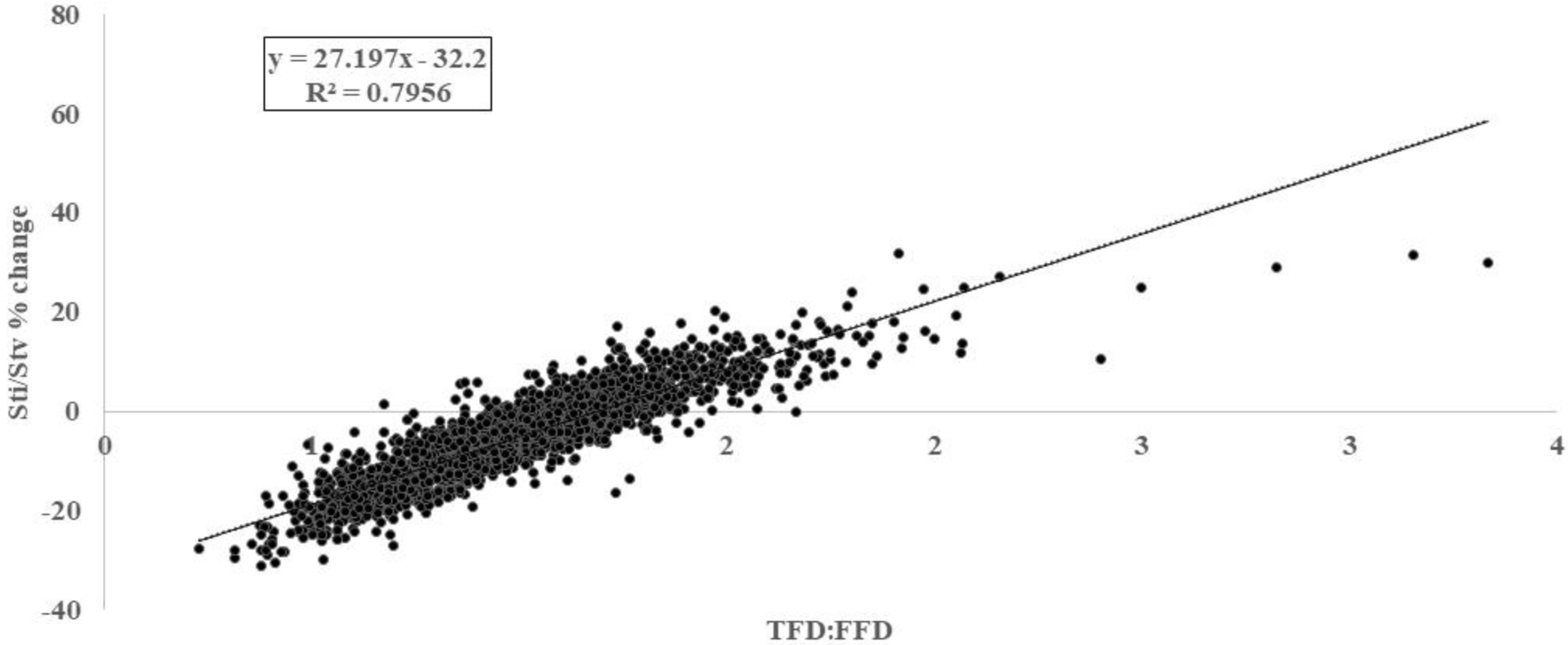
Table 3. Summary of the 2516 coding sequences selected for the present study. It represents the minimum to maximum range of different parameters like GC%, TFD/FFD, PTC% and the mean and standard deviation (S.D.) of the TFD%, FFD% and PTC%

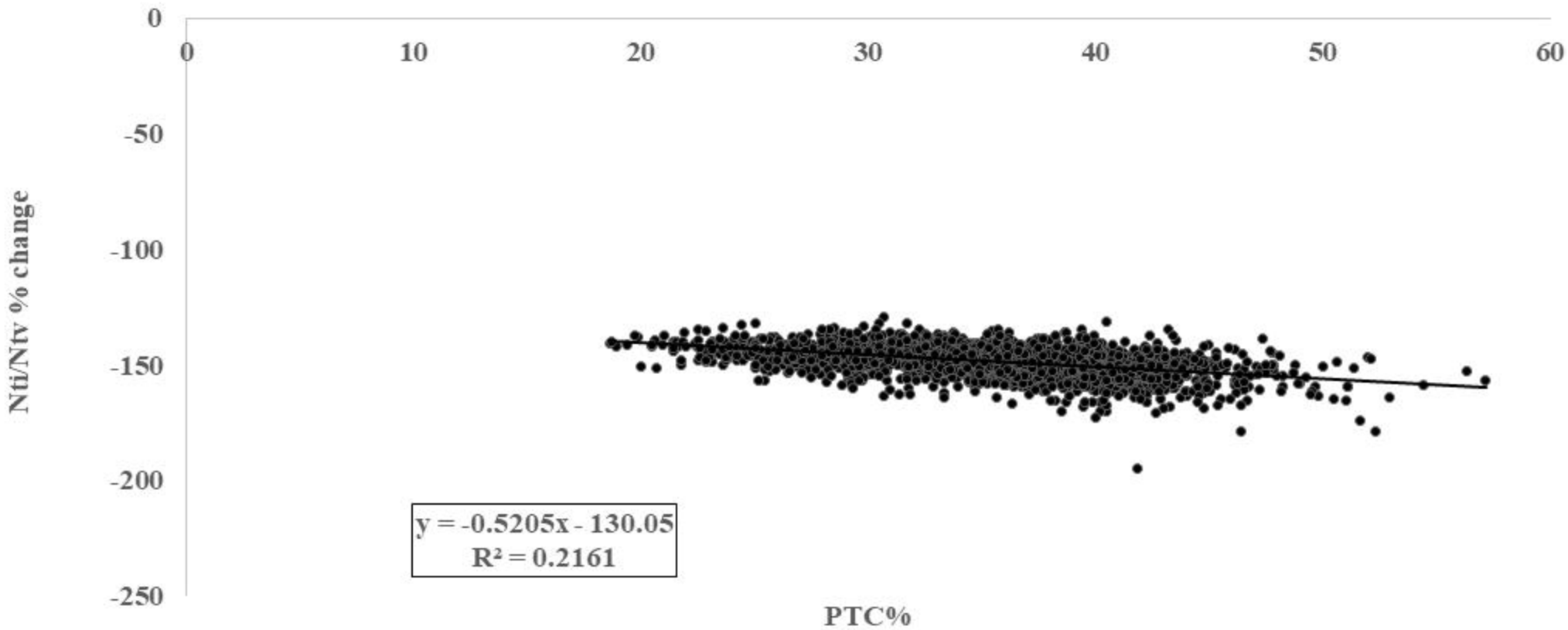
Genes considered	Size bp (Min-Max)	GC% (Min-Max)	TFD/FFD (Min-Max)	PTC% (Min-Max)	TFD%	FFD%	PTC%	
					33.384	34.084	36.350	Mean
2516	93-4554	32.95-63.80	0.22-3.33	15.00-57.14	5.726	4.761	5.625	S. D


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 $ti_e^j \times f_j$ & $tv_e^j \times f_j$
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 $ti_e = \sum_{j=1}^{61}(ti_e^j \times f_j)$ $tv_e = \sum_{j=1}^{61}(tv_e^j \times f_j)$
$\frac{ti'}{tv'} = \frac{(ti_o/ti_e)}{(tv_o/tv_e)}$
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