Chapter 3

Analysis Of Non-Synonymous Variations in Genome Sequences of *Escherichia Coli*

CHAPTER 3
ANALYSIS OF NON-SYNONYMOUS VARIATIONS IN GENOME SEQUENCES OF *Escherichia coli*

3.1. Abstract

In this work, we have normalized the estimation of non-synonymous transition (*Nti*) and non-synonymous transversion (*Ntv*) and also calculated the *Nti'* to *Ntv'* ratio through an improved estimator across all the codons in 2481 genes in *E. coli*. Our analysis in 40,586 non-synonymous variations unveils many surprising findings including the role of codon degeneracy. Between the *Nti* values of *two-fold degenerate* codons and *four-fold degenerate* codons, the latter was observed with 2.4-fold higher values than the former (p<0.01), suggesting a novel understanding of the role of codon degeneracy in non-synonymous variations. Even though the reason behind the *ti* bias in FFD codons is unknown. Our study in 64*64 matrix and 20*20 matrix revealed many insightful significances of codon as well as amino acid exchangeabilities in *E. coli*. Among the most frequently occurring amino acid changes, we observed some surprisingly high preference of *Ntv* in amino acids like Ser→Ala, Phe→Leu/Tyr, Cys→Ser and Lys→Gln. We observed Gly→Asp to be three times more frequent than Gly→Glu despite the involvement of similar G→A *ti* in both. It is noteworthy that Asp and Glu are not much different while it comes to their physiochemical properties as amino acids. Similarly, despite the involvement of G→A *ti,* we observed Gly→Ser (SB) to be 4.71 times more frequently occurring than Gly→Arg (SB). Many of our observations pointed towards the role of intrinsic factors such as economy, stereochemistry, and hydrophobicity of amino acids for a firm understanding of amino acid exchangeabilities. Our observations like frequent AUG→AUA (Met→Ile) and frequent Cys→Ser (SB) over Cys→Ser (FB)

subsequently assisted us to understand the structure and evolution of genetic code table logically. The confounding impact of mutation and selection or either of them is yet to be cited adequately while it comes to the amino acid exchangeability in proteins. The detrimental impact of non-synonymous transition and non-synonymous transversion requires ample amount of evolutionary data to conclude any outcome. However, the extrapolation of our work in other organisms will help us to understand the non-synonymous variations in a distinctive pattern.

3.2. Introduction

The quest for whether codons are assigned to amino acids or amino acids are assigned to codons remains unresolved, despite advancements in our understanding of molecular evolution. The intra-class substitution between nitrogenous bases (R←→R and Y←→Y) known as transition (*ti*) is principally observed in the genomes over the inter-class substitutions among nitrogenous bases (R←→Y) also known as transversion (*tv*) (Sen et al., 2022; Beura et al., 2023). This mutational bias between *ti* and *tv* seems to have influenced the codon assignments in the genetic code table. The rate difference between *ti* and *tv* is believed to play a key role in demonstrating the synonymous variations. An amino acid is assigned to the synonymous codons in such a manner that *ti* is observed more frequently in the synonymous polymorphisms in the genetic code table (Duchêne et al., 2015; Lewis Jr et al., 2016). The mutational bias between *ti* and *tv* (Gerber and Keller, 1999; Lewis et al., 2016; Bhagwat et al., 2016), seems to have influenced the assignments of codons in the genetic code table considering non-synonymous variations. For an example, Ala→Thr require G→A *ti* in the 1st position of Ala codons (GCN→ACN) whereas Ala→Gly requires C→G *tv* in the 2nd position of Ala codons (GCN→GGN), as the rate difference between a *ti* and a *tv* is known (Sen et al., 2022), Ala→Thr changes are more frequently anticipated than Ala→Gly. Table 3.1 represents the possible amino acid exchangeabilities through *ti* and *tv* concerning each amino acids through single nucleotide variations (SNVs).

Table 3.1. Theoretical accountability of amino acids exchangeability shows the respective amino acid's exchangeability through *ti* and *tv*. The estimated numbers show the theoretical possibilities of *Nti* and *Ntv* for each amino acid.

Amino acids	Through *ti*	Through *tv*	Estimated numbers	
			Nti	*Ntv*
Phe	Ser, Leu	Tyr, Cys, Leu, Ile, Val	4	12
Leu	Ser, Pro	Trp, His, Gln, Arg, Ile, Met, Val	8	25
Ile	Thr, Val, Met	Leu, Phe, Met, Asn, Lys, Arg, Ser	7	14
Met	Thr, Val, Ile	Leu, Phe, Lys, Arg	3	6
Val	Ile, Met, Ala	Leu, Phe, Asp, Glu, Gly	8	16
Ser	Pro, Phe, Leu, Gly, Asn	Tyr, Cys, Trp, Thr, Ala, Arg, Ile	12	25
Pro	Leu, Ser	Thr, Ala, His, Gln, Arg	8	16
Thr	Ala, Ile, Met	Pro, Ser,Asn, Lys, Ser, Arg	8	16
Ala	Thr, Val	Pro, Ser,Asp, Glu, Gly	8	16
Tyr	Cys, His	Asn, Asp,Pro, Leu	4	8
His	Arg, Tyr	Asn, Asp, Gln, Pro, Leu	4	12
Gln	Arg	Pro, Leu, Lys, His, Glu	2	12
Asn	Asp, Ser	Lys, Thr, Ile, His, Tyr	4	12
Lys	Glu, Arg	Asn, Thr Ile, Met, Gln	4	10
Asp	Gly, Asn	Glu, Ala, Val, His, Tyr	4	12
Glu	Lys, Gly	Asp, Ala, Val, Gln	4	10
Cys	Tyr, Arg	Trp, Ser, Phe, Gly	4	10
Trp	Arg	Cys, Ser, Leu, Gly	1	6
Arg	Cys, Trp, His, Gln, Gly, Lys	Pro, Leu,Ser, Thr, Ile	11	23
Gly	Ser, Arg, Asp. Glu	Arg, Cys, Trp, Ala, Val	8	15

The potential substitutions in a codon are always higher than the actual substitutions (Beura et al., 2024 [under review]). The codon substitution model (CSM) allows for the substitution of each nucleotide by the remaining three nucleotides (1 *ti* and 2 *tv*) at each position of a codon resulting in nine different combinations of codons through SNVs. The 61 sense codons in the genetic codon table do not share a common substitution pattern of non-synonymous sites. The role of codon degeneracy comes into limelight while calculating the possible non-synonymous sites for each codon. A typical *two-fold degenerate* (TFD) codon has an estimated *Nti* as 2 and *Ntv* as 6. Whereas a typical *four-fold degenerate* (FFD) has an estimated *Nti* as 2 and *Ntv* as 4 (Table 3.2). Such discrepancies can also be observed in the family box (FB) and split box (SB) of *six-fold degenerate* (SFD) codons of Leu, Arg and Ser. Therefore, the estimated numbers of *Nti* and *Ntv* varies among codons of different degeneracy. The stronger purifying selection on the non-synonymous substitution suggests for the role of pre-termination codons (PTC) while analysing different non-synonymous substitutions (Hurst and Pál 2001). PTC are such codons in the genetic codon table which are prone to converting into stop codon by a SNVs (Modiano et al. 1981; Aziz et al. 2022). For example, there are only two *zero-fold degenerate* codons (AUG and UGG), hence the *Nti* and *Ntv* possibilities for both AUG and UGG should have been three and six respectively, but in the case of UGG, two *Nti* result into UAG and UGA through one SNVs each at 2nd and 3rd position respectively, hence the *Nti* possibilities for UGG is estimated as one whereas for AUG it is considered as three, which correctly attributes to the pretermination nature of UGG. Hence the position of the stop codons in the genetic codon table eventually affects the premature termination nature of the PTC. Table 3.2 represents the estimated numbers of non-synonymous *ti* (*Nti*) and non-synonymous *tv* (*Ntv*) for each codon after excluding the pre-termination nature of codons.

Table 3.2. The estimated numbers of *Nti* and *Ntv* calculated for each codon is represented. The FFD codons and TFD codons (represented in blue and orange shades) can be seen with variable numbers of *Ntv* numbers between them.

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

Our procedure accounts for the PTC and normalizes the actual substitutions out of potential substitutions. Interestingly, the theoretical estimates of the genetic code table identify 74.5% variations (116 *Nti* + 276 *Ntv*) as non-synonymous by nature. However, such a high proportion of non-synonymous variations is not feasible for the sustainability of the majority of life forms. The anticipation of frequent *Ntv* is common if the numbers are to be considered without the impact of selection. Therefore, it has been hypothesized that amino acid assignment in the genetic codon table has been done considering the mutation rate. Interestingly, researchers in the past decades have worked extensively on the transitional and transversional aspect of non-synonymous polymorphism in different organisms and the resulting physiochemical changes through amino acid substitutions (Zhang 2000). The among-species variations in amino acid exchangeabilities are probably a result of proteome-wide changes in the physicochemical environments of amino acid residues during evolution (Zou and Zhang, 2019). Still many researchers believe that the non-synonymous variations resulting in specific amino acid changes are species-specific (Dang et al., 2010; Chen et al., 2019; Weber and Whelan, 2019 Zou and Zhang, 2021). It has been already reported that *ti* variation is more frequent than *tv* variation due to the structural similarity in the intra-nucleotide class (Zhang, 2000; Freudenberg-Hua et al., 2003; Schrider et al., 2013). However, the selection involving both the synonymous and non-synonymous variations regarding *ti* and *tv* bias are likely to be non-identical as the highest possibilities are estimated in case of non-synonymous *tv* (Beura et al., 2024 [under review]). It is already known that non-synonymous changes are usually under stronger purifying selection than the synonymous ones (Schmidt et al., 2008). Therefore, the selection on protein structure and function considering more frequently happening mutations are thought to be less deleterious than the mutations occurring less frequently to enhance the fitness of an organism (Charlesworth, B and Charlesworth, 1998; Agrawal and Whitlock, 2012). The CSM proposed by Goldman and Yang allows for variable rate of selection acting on different codon positions and assumes that the rate of non-synonymous variations depends on the properties of amino acids involved (Goldman and Yang, 1994). Answering questions about the cumulative impact of selection, mutation, or each individually on coding sequences requires extensive study across different clades. Subsequently, the evolution of the genetic code table as per the assignment of codons and amino acids can be better understood by following the amino acid exchangeability study which also addresses the challenge of deciphering the detrimental effects of non-synonymous *ti* or *tv* variations.

In this study, we note that, codon degeneracy is a key contributor to the overall $\frac{Nti}{Ntv}$ across coding sequences in *E. coli*. Our study also reveals the most/least frequent amino acid exchangeabilities and most/least frequently mutating codons in *E. coli*. We also observe that Nti is favoured for many amino acids, while *Ntv* is preferred only for frequent changes in a few specific amino acids. The majority of the cumulative ratio of $\frac{Nti}{Ntv}$ involving individual amino acids revealed the higher prevalence of *Nti*, sometimes reaching up to tenfold differences. Overall, our study highlights the role of codon degeneracy in shaping the genetic code table as it has evolved to its present state. The extrapolation of our work to other organisms might untangle the detrimental impact of *Nti* and *Ntv* in species-specific prospects of evolution.

3.3. Materials and Methods

3.3.1. Coding sequence information

We performed a SNVs analysis of 2481 coding sequences across 157 strains in *E. coli* (Thrope et al., 2017). The coding sequences were chosen based on their proper annotation, alignments, and high prevalence of non-synonymous variations. The codons having variations in more than one position and strains having ambiguous nucleotides in individual coding sequences were not considered in the study. We have followed the procedure of calculation of mutations through a consensus sequences-based approach already explained in Chapter 2 of this thesis (Sen et al., 2022; Aziz et al., 2022; Beura et al., 2024 [under review]).

3.3.2. Finding *Nti′ and Ntv′ in* coding sequences and the estimation of $\frac{Nti'}{Ntv'}$ for all the codons

We have developed a normalization procedure to overcome the dilemma of codon degeneracy and PTC in coding sequences. As discussed, all the codons have different *Nti* and *Ntv* possibilities based on their theoretical calculations. The possibilities of acquisition of non-synonymous variation varies among codons to some extent even in a similar degeneracy class (Table 2). In general, PTC codons have few *Nti* or *Ntv* possibilities that lead to stop codons. Now to overcome the phenomenon of stronger purifying selections in such scenarios, it is essential to normalize the *Nti* and *Ntv* variations with regards to their observed and total estimated changes in each codon now termed as Nti' and Ntv'. Further, we have also applied the approach of normalization for the estimation of improved $\frac{\boldsymbol{Nti}}{\boldsymbol{Ntv}}$ (Beura et al., 2024 [under review]).

$$\textbf{a)}\ \boldsymbol{Nti'} = \frac{\boldsymbol{Nti_o}}{\boldsymbol{Nti_e}}$$

$$\textbf{b)}\ \boldsymbol{Ntv'} = \frac{\boldsymbol{Ntv_o}}{\boldsymbol{Ntv_e}}$$

$$\textbf{c)}\ \frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}} = \frac{\left({}^{\boldsymbol{Nti_o}}/_{\boldsymbol{Nti_e}}\right)}{\left({}^{\boldsymbol{Ntv_o}}/_{\boldsymbol{Ntv_e}}\right)}$$

$\boldsymbol{Nti_o}$ = Non-synonymous transitions observed
$\boldsymbol{Ntv_o}$= Non-synonymous transversions observed

$\boldsymbol{Nti_e}$= Non-synonymous transitions estimated
$\boldsymbol{Ntv_e}$= Non-synonymous transversions estimated
Where $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ = Improved $\frac{\boldsymbol{Nti}}{\boldsymbol{Ntv}}$

3.3.3. Visualizing the frequent amino acid exchangeability in *E. coli* coding sequences

We studied the frequent amino acid replacements by considering the unidirectional changes (FROM→TO) concerning each amino acid only to understand the pattern of amino acid exchangeability in *E. coli*. We analysed the amino acid exchangeability through individual codon-wise and amino acid-wise changes by performing the 64*64 matrix and 20*20 matrix respectively. For instance, if UUU has 100 non-synonymous variations, and UUU→CUU was observed ten times, then the frequency of UUU→CUU was calculated as 10/100= 0.1. Suppose UUC has fifty non-synonymous variations, and UUC→CUC was observed twenty times, then the frequency was calculated as 20/50=0.4. Therefore, the UUU→CUU and UUC→CUC 64*64 matrix normalization value can be mentioned as 0.1 and 0.4 respectively. While analysing the above example for 20*20 matrix, for Phe amino acid, now we have a total 100+50= 150 non-synonymous variations (UUU+UUC), a total of 10+20=30 non-synonymous changes led to Leu family box (FB). Hence, the frequency of Phe→Leu FB can be calculated as 30/150= 0.2. It is noteworthy that, we have calculated the family box (FB) and split box (SB) codons and their encoded amino acids separately *in six-fold degenerate* codons.

OriginPro 2022, OriginLab Corporation, Northampton, MA, USA was used to draw the Box-plot/Scatter plots as well as to perform the Mann-Whitney test (Mann and Whitney., 1947) to find out the *p*-value. We used https://github.com/CBBILAB/CBBI.git for the estimation of $\frac{Nti'}{Ntv'}$ values (Beura et al., 2024 [under review]).

3.4. Results

3.4.1. Frequency of non-synonymous variations in different degenerate codons of *E. coli*

In this study, we considered the computational analysis of 881,244 codons and observed 40,634 non-synonymous SNVs that consists of 23,600 *Nti* and 16,986 *Ntv* out of 157,4349 estimated *Nti* and 395,4112 estimated *Ntv*. The summary table encapsulates the comprehensive results across various parameters (Table 3.3). We found out *Nti'* and *Ntv'* for the 61 codons that enabled us to compare across these codons (Table 3.4). We noted that, AGG exhibited the maximum *Nti'* value at 0.03358, whereas CUG displayed the minimum *Nti'* value at 0.00119, a 28.12 times more frequent value was observed between both. While analysing between *zero-fold degenerate* codons AUG and UGG, we observed *Nti*' value as 0.00923 and 0.00276 respectively whereas the *Ntv*' was noted as 0.00248 and 0.00229 respectively. AUG exhibited *Nti'* values four times more frequent than UGG, while the *Ntv'* values were similar. It is pertinent to note that Met has higher possibilities of amino acids exchangeability through *ti* than Trp, whereas Trp is more restricted for amino acid exchangeability through *ti.* Interestingly, among the TFD codons, we observed that GAC exhibited the maximum *Nti'* value at 0.01712 whereas the minimum *Nti'* value was observed in UUC at 0.00254, indicating a 6.72 times more frequent value between both. Similarly, GAG was noted with the maximum *Ntv*' values at 0.0082 and UUC was noted with the minimum *Ntv*' values at 0.00216. Notably among the TFD codons, we observed a two-fold magnitude between the maximum *Nti'* and *Ntv'* values. Similarly, among the FFD codons, GUC was noted with the maximum *Nti' values* as 0.0311 whereas the minimum *Nti'* values was noted in GGG at 0.01256, notably a 2.47-fold difference was observed. Interestingly, the maximum *Ntv'* value was observed in ACU at 0.00748, 4.1 times more frequent value was observed between the maximum *Nti'* values and *Ntv'* values among the FFD codons. Among the SFD codons, we observed the maximum *Nti'* values in AGG at 0.03357 and the minimum *Nti'* values was observed in CUG at 0.00119, interestingly 28.12 times more frequent value was observed between both. Similarly, among the SFD codons, the maximum *Ntv'* value was observed in UCA as 0.01031 and the minimum *Ntv'* value was observed in CGU at 0.002633. It is pertinent to note that among the *Ntv'* values, we observed the highest *Ntv'* value in UCA at 0.01031, whereas the lowest *Ntv'* value was observed in AUU at 0.00198. Nevertheless, we noted more frequent occurrences of *Nti*' values than *Ntv*' values in all codons, except UUR, CUR, and UCA. The preliminary observations insisted us to unravel in deep regarding the amino acid exchangeabilities.

Table 3.3. The overall result in summary is represented

Results in summary		
	Total codons considered	881244
	Total NS variations	40586
Estimated	*Nti*	1574349
	Ntv	3954112
Observed	*Nti*	23600
	Ntv	16986
Min	*Nti'*	0.001194
Max	*Nti'*	0.033575
Min	*Ntv'*	0.001977
Max	*Ntv'*	0.010308
Min	*Nti'/Ntv'*	0.297
Max	*Nti'/Ntv'*	13.402
Mean	*Nti'/Ntv'* (overall)	3.827024

	Nti'/*Ntv*' (FFD)	5.9812
	Nti'/*Ntv*' (TFD)	2.1478
	Nti'	0.014876
	Ntv'	0.004304
	Nti' (FFD)	0.0214
	Nti' (TFD)	0.0089
	Ntv' (FFD)	0.00403
	Ntv' (TFD)	0.00423
Codon count	TFD	294423
	FFD	302884
Nti	TFD (obs)	5011
Ntv		7297
Nti	FFD (obs)	13048
Ntv		4792

Table 3.4. The codons, codon count, estimated and observed *Nti* and *Ntv*, normalized *Nti'* and *Ntv'* along with the ratio of *Nti'* to *Ntv*' is represented.

		Estimated		Observed			Non-normalized	Normalized		
Codons	Codon count	*Nti*	*Ntv*	*Nti*	*Ntv*	Total NS	frequency	*Nti'*	*Ntv'*	*Nti'*/*Ntv*'
UUU	18823	37646	112938	107	256	363	0.0193	0.00284	0.00227	1.254
UUC	15308	30616	91848	78	199	277	0.0181	0.00255	0.00217	1.176
UUA	11055	11055	44220	45	257	302	0.0273	0.00407	0.00581	0.700
UUG	11799	11799	58995	26	325	351	0.0297	0.00220	0.00551	0.400
CUU	8856	17712	35424	235	144	379	0.0428	0.01327	0.00407	3.264
CUC	9745	19490	38980	224	164	388	0.0398	0.01149	0.00421	2.732
CUA	2852	2852	11408	6	43	49	0.0172	0.00210	0.00377	0.558
CUG	51099	51099	204396	61	822	883	0.0173	0.00119	0.00402	0.297
AUU	27008	54016	135040	471	267	738	0.0273	0.00872	0.00198	4.410
AUC	23532	47064	117660	377	355	732	0.0311	0.00801	0.00302	2.655
AUA	2337	7011	9348	147	50	197	0.0843	0.02097	0.00535	3.920
AUG	25087	75261	150522	695	374	1069	0.0426	0.00923	0.00248	3.717
GUU	15988	31976	63952	858	135	993	0.0621	0.02683	0.00211	12.711
GUC	13673	27346	54692	851	127	978	0.0715	0.03112	0.00232	13.402
GUA	9400	18800	37600	468	111	579	0.0616	0.02489	0.00295	8.432
GUG	25013	50026	100052	758	279	1037	0.0415	0.01515	0.00279	5.434
UCU	7141	14282	28564	82	137	219	0.0307	0.00574	0.00480	1.197
UCA	5093	10186	10186	93	105	198	0.0389	0.00913	0.01031	0.886
UCC	7857	15714	31428	110	173	283	0.0360	0.00700	0.00550	1.272
UCG	8082	16164	24246	159	167	326	0.0403	0.00984	0.00689	1.428
CCU	5588	11176	22352	213	90	303	0.0542	0.01906	0.00385	4.953
CCC	4333	8666	17332	209	86	295	0.0681	0.02412	0.00496	5.291
CCA	7017	14034	28068	263	113	376	0.0536	0.01874	0.00403	4.655

CCG	22655	45310	90620	735	316	1051	0.0464	0.01622	0.00349	4.652
ACU	7320	14640	29280	235	219	454	0.0620	0.01605	0.00748	2.146
ACC	21826	43652	87304	728	530	1258	0.0576	0.01668	0.00607	2.747
ACA	4738	9476	18952	266	99	365	0.0770	0.02807	0.00522	5.374
ACG	12681	25362	50724	645	270	915	0.0722	0.02543	0.00532	4.778
GCU	13052	26104	52208	654	250	904	0.0693	0.02505	0.00479	5.232
GCC	23256	46512	93024	1403	530	1933	0.0831	0.03016	0.00570	5.294
GCA	17361	34722	69444	1008	369	1377	0.0793	0.02903	0.00531	5.463
GCG	32699	65398	130796	1910	713	2623	0.0802	0.02921	0.00545	5.358
UAU	13130	26260	52520	93	122	215	0.0164	0.00354	0.00232	1.525
UAC	11010	22020	44040	62	109	171	0.0155	0.00282	0.00248	1.138
CAU	11046	22092	66276	335	231	566	0.0512	0.01516	0.00349	4.351
CAC	8915	17830	53490	253	210	463	0.0519	0.01419	0.00393	3.614
CAA	12797	12797	76782	127	538	665	0.0520	0.00992	0.00701	1.416
CAG	26303	26303	157818	338	1076	1414	0.0538	0.01285	0.00682	1.885
AAU	13324	26648	79944	312	303	615	0.0462	0.01171	0.00379	3.089
AAC	19580	39160	117480	396	343	739	0.0377	0.01011	0.00292	3.464
AAA	29253	58506	146265	282	542	824	0.0282	0.00482	0.00371	1.301
AAG	8226	16452	41130	137	217	354	0.0430	0.00833	0.00528	1.578
GAU	27772	55544	166632	777	894	1671	0.0602	0.01399	0.00537	2.607
GAC	17488	34976	104928	599	483	1082	0.0619	0.01713	0.00460	3.720
GAA	36002	72004	180010	634	973	1607	0.0446	0.00881	0.00541	1.629
GAG	15626	31252	78130	377	641	1018	0.0651	0.01206	0.00820	1.470
UGU	4174	8348	20870	49	59	108	0.0259	0.00587	0.00283	2.076
UGC	5646	11292	28230	55	101	156	0.0276	0.00487	0.00358	1.361
UGG	13056	13056	78336	36	179	215	0.0165	0.00276	0.00229	1.207
CGU	19750	39500	79000	674	208	882	0.0447	0.01706	0.00263	6.481
CGC	20676	41352	82704	876	308	1184	0.0573	0.02118	0.00372	5.688
CGA	2562	2562	7686	80	23	103	0.0402	0.03123	0.00299	10.435
CGG	4093	8186	12279	215	48	263	0.0643	0.02626	0.00391	6.719
AGU	6681	13362	40086	265	203	468	0.0700	0.01983	0.00506	3.916
AGC	14081	28162	84486	568	490	1058	0.0751	0.02017	0.00580	3.478
AGA	944	1888	3776	59	27	86	0.0911	0.03125	0.00715	4.370
AGG	551	1102	2755	37	17	54	0.0980	0.03358	0.00617	5.441
GGU	22637	45274	90548	589	194	783	0.0346	0.01301	0.00214	6.072
GGC	28184	56368	112736	863	238	1101	0.0391	0.01531	0.00211	7.252
GGA	5953	11906	17859	153	47	200	0.0336	0.01285	0.00263	4.883
GGG	9510	19020	38040	239	87	326	0.0343	0.01257	0.00229	5.494

In order to investigate the possible role of codon degeneracy, we observed frequent FFD codons among the top Nti' frequency values (Table 3.4). GUC was noted with the maximum Nti' frequency at 0.03112, whereas the maximum Nti' frequency among TFD codons was observed in GAC at 0.01713, notably 1.8 times more frequent Nti' value was observed in the former. It suggested towards the prevalence of higher Nti' values in FFD codons. We then separately analysed the cumulative mean Nti' values among FFD codons and also among TFD codons. The mean Nti' in FFD codons was observed as 0.0214 whereas in TFD codons it was observed as 0.0089, notably a 2.4-fold higher frequency was observed in the former (Table 3.3). Interestingly, the comparative study involving Ntv' values revealed no such significant mean magnitude fold difference between the two degenerate codons. The box-plot in Figure 3.1 represents the significant Nti' difference observed in case of TFD and FFD codons (p<0.01), whereas Ntv' between both the degenerate codon was not observed to be significantly different (p>0.01). It apparently provides the probable role of codon degeneracy on Nti' frequency values in *E. coli*.

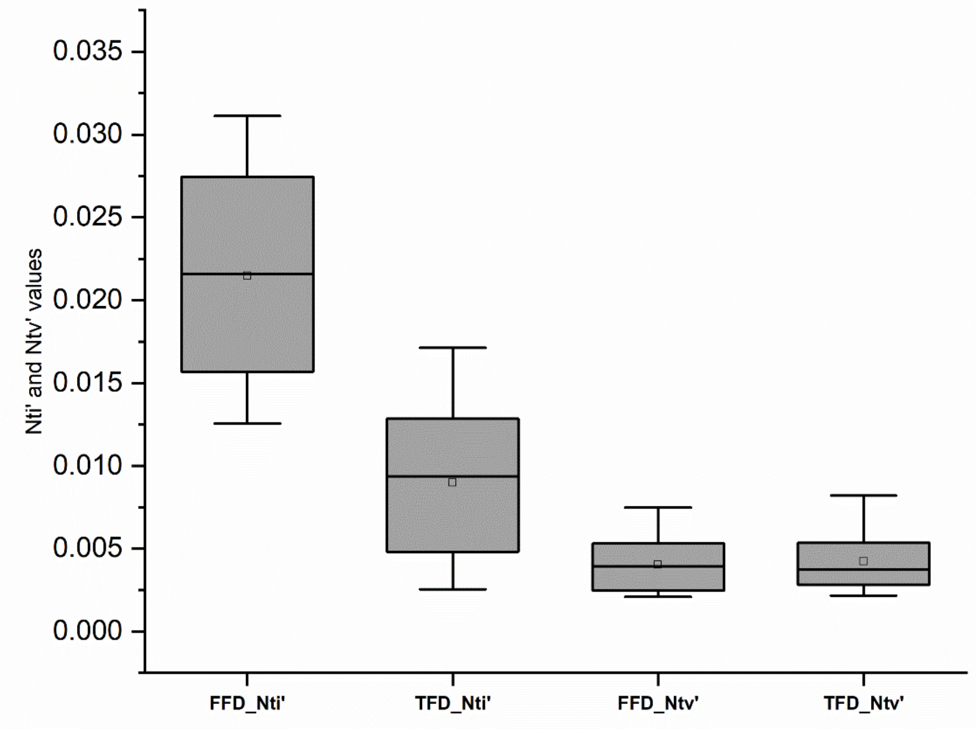

Fig. 3.1. The box-plot illustrates the Nti' and Ntv' values of TFD codons and FFD codons. Here, y-axis shows the ti' and Ntv' values. The comparative study between Nti' values

between the TFD codons and FFD codons shows a significant difference between both the mean values (*p*<0.01). The Ntv' values between the TFD codons and FFD codons are not significantly different (*p*>0.01) as the mean values are much closer between both.

Table 3.4 provides the $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ across all the 61 sense codons. The maximum $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ was observed in GUC as 13.402 whereas the minimum $\frac{Nti'}{Ntv'}$ was observed in CUG as 0.297. The mean $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ across the codons was observed as 3.892, which again coincides with our previous observations (Beura et al., 2024 [under review]). The mean $\frac{Nti'}{Ntv'}$ in TFD codons was observed as 5.98 whereas the mean $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ in FFD codons was observed as 2.14 (Table 3.3). To further substantiate the impact of codon degeneracy on the non-synonymous variations, a box-plot analysis of $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ between the TFD codons and FFD codons was performed (Fig. 3.2). It revealed a significant difference (*p*<0.01) between both the degenerate codons. This observation inspired us to delve into one of the most contentious topics in molecular evolution research: the detrimental effects of *Nti* and *Ntv* on coding sequences. However, the noteworthy aspect of our study highlighted the influence of codon degeneracy on non-synonymous variations which was not known before.

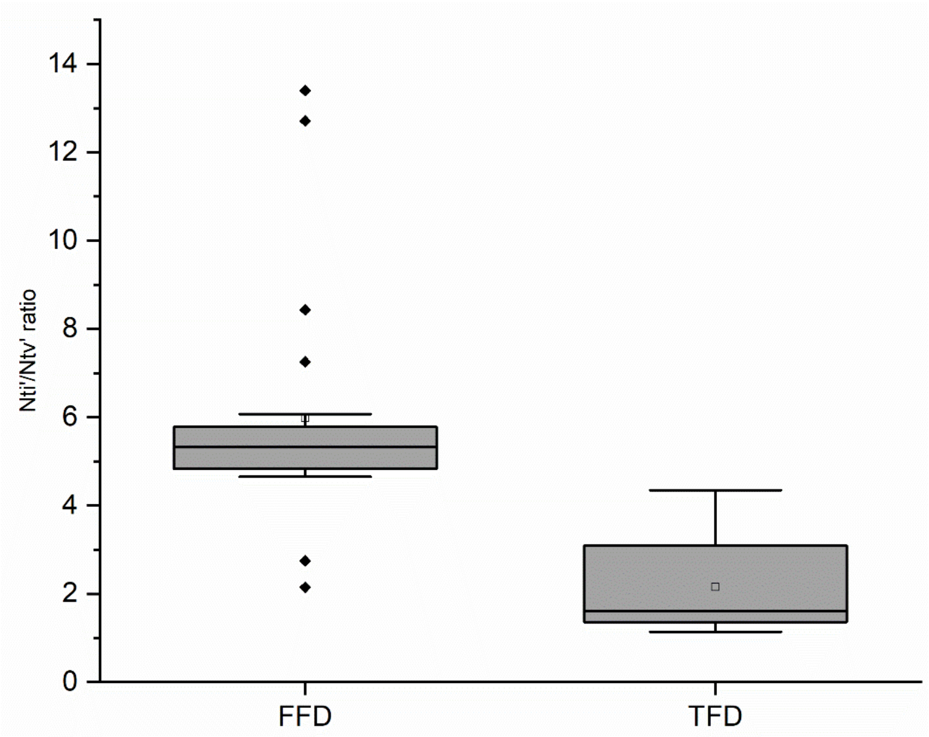

Fig. 3.2. The box-plot illustrates the $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ values comparison between TFD codons and FFD codons. The *y*-axis shows the $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ ratio. The result shows a significant difference between the $\frac{\boldsymbol{Nti'}}{\boldsymbol{Ntv'}}$ values of both the degenerate classes (p<0.01).

3.4.2. The study of amino acid exchangeability in *E. coli*

We have examined the exchangeability of amino acids in *E. coli* down to the individual amino acid level concerning FB and SB separately for *six-fold degenerate* codons. The 64*64 codon matrix revealed the codon-to-codon changes for all the 61 codons involving SNVs only for each codon (Appendix VI). The maximum codon exchangeability frequency value was observed in case of CGA→CAA at 0.776. Similarly, the minimum codon exchangeability frequency value was observed in case of AUG→AAG at 0.009. Notably, we observed an eighty-six-fold difference between this minimum frequency and the maximum frequency. Our observation in the 20*20 amino acid matrix had uncovered several noteworthy insights into amino acid exchangeability in *E. coli*. The normalized frequency of amino acid exchangeability is provided in Table 3.5. The maximum frequency value in the 20*20 amino acid matrix was noted in case of Arg (SB) → Lys at 0.607 which involves the mechanism of *ti* and the maximum frequency value for any amino acid change involving a *tv* was observed in case of Ser→Ala at 0.445. The former involves the amino acid exchangeability between two positively charged amino acids (Lys and Arg) whereas the latter raises the possibility of selection as Ala and Ser are not similar by nature. We also observed the least frequent amino acid changes in Ile→Arg, which can be correctly attributed by the different nature of both the amino acids. We have delved into understanding alterations occurring at the level of individual amino acids, as outlined below, and endeavoured to unravel the pattern of substitutions implicated in these alterations. A detailed result is represented in Table 3.6 highlighting the type of substitutions, positions in those concerned codons of certain amino acids, frequencies of exchangeability and resulting amino acid changes.

- **Phe:**
  - The most significant change was noted for Leu SB and FB, both equally registering a frequency value of 0.24, followed closely by Tyr at 0.215 (U→A).
  - The frequency of change to Ile was notes as 0.0953 whereas frequency to Tyr was notes as 0.2125, despite being U→A *tv in* both cases, Phe→Tyr was observed to be more frequently changing. Aromaticity of Phe and Tyr could be explained behind the exchangeability between both.
- **Leu (SB):**
  - The most significant changes were noted for Phe with a frequency value of 0.182 resulting from a *tv*. It is pertinent to note that, UUR→UUY changes were frequently observed.

  - The frequency of change to Val was notes as 0.1516 whereas frequency to Trp was notes as 0.092, despite being U→G *tv in* both cases, Leu (SB)→Val was observed to be 16 times more frequently changing than Leu (SB)→Trp.

- **Leu (FB):**
  - The exchangeability frequency between Leu (FB) and Phe was recorded to be similar and hydrophobic nature could be the possible reason behind this exchangeability.
  - The most frequent amino acid change was observed in Phe and Met with a frequency value of 0.25 and 0.21 respectively, however both involves C→U *ti* and C→A *tv*.
- **Ile:**
  - The most significant changes were noted for Val with a frequency value of 0.443 resulting from a *ti*. It is noteworthy that, Ile is more closely related to Val than Leu.
  - The frequency values in Leu (FB) were noted as 0.180 and in Leu (SB) it was noted as 0.017. Remarkably, the change to Leu (FB) was observed to be ten times more frequent compared to changes to Leu (SB).

- **Met:**
  - One of the highest SB exchanges was observed in AUG→AUA resulting in more frequent Met→Ile changes.
  - The frequency value of Ile exchange was noted as 0.4808. However, in an interesting scenario, we observed that AUG→AUA was noted with four times more frequency value than AUG→AUU. The frequent Met→Ile could be

explained by the similar hydrophobicity nature of both amino acids.

- **Val:**
    - The highest frequency changes were observed in Ile, with a value of 0.488, followed by Met and Ala, both registering frequency values of 0.133. Notably, all these changes are resulted from *Nti*.
    - However, both Val and Ile are known to be strongly similar in structure and hydrophobicity.
- **Ser (FB):**
    - The most frequent change was observed in Ala, with a value of 0.307 followed by Thr with a value of 0.19. Surprisingly, Pro recorded a value of 0.16. The unusual U→G/A *tv* over U→C *ti* raised the possibility of selection in Ser (FB).
    - Interestingly, the frequency of Cys was noted as 0.0322 and the frequency of Trp was noted as 0.003, ten times more frequent Cys changes was observed although C→G *tv* was the mechanism involved in these cases.
- **Pro:**
    - The most frequent changes were observed in Ser (FB) and Leu (FB) with values 0.44 and 0.26 respectively, both are resulted by C→U *ti.*
    - The frequency of Gln changes was noted as 0.166 whereas the frequency in case of His was noted as 0.026, C→A *tv* was involved in both cases yet, Pro→Gln were observed to be favored 6.32 times more than Pro→His. The preference of His over Gln could be explained as a part of selection at peptide level.
- **Thr:**
    - The most frequent changes were observed in Ala as the values were noted as 0.28, followed by Ile and Met with 0.22 and 0.12 frequency values respectively.

All these changes were due to *ti*.

- Among *tv* changes, Ser (FB) was noted with highest frequency values 0.12. Despite being a polar amino acid, Thr exchanges with Ile and Ala which could be understood through proteomic investigations.

- **Ala:**
    - The most frequent changes were observed in Thr and Val with frequency values of 0.38 and 0.34 respectively despite the latter being closer to Ala by hydrophobicity. However, changes to Ser (FB) were observed to be the most frequent among all the *tv* changes with a value of 0.13.
    - Between Ser (FB) → Ala and Ala→Ser (FB), the former was observed to be two times more frequent than the latter.
    - Interestingly, changes to Ser (FB) (0.13) noted seven times more frequency than changes to Pro (0.019).
- **Tyr:**
    - The most frequent change was observed in Phe with a frequency value of 0.415 which is the result of a A→U *tv*. Surprisingly, we have observed twice more frequent changes involving Tyr→Phe than Phe→Tyr. However, the aromaticity of amino acids could be explained behind thus exchangeability.
    - Among *ti* changes, the most frequent changes were observed in His and Cys with values 0.2461 and 0.1554 respectively.
- **His:**
    - The most frequent changes were observed in Tyr with a value of 0.418.
    - In a noteworthy comparison between changes leading to Asn and Asp, we observed a frequency of eight times higher in Asn compared to changes leading

to Asp, despite both being resulted by *Ntv*. The preference of not exchanging His with a negatively charged amino acid Asp could be a possibility behind this discrepancy.

- **Gln:**
  - The most frequent changes were observed in Leu (FB) with a value of 0.240 followed by Arg (FB) with a value of 0.220. The maximum exchangeability value of a polar amino acid with a non-polar amino acid like Leu needs further understanding in proteomic study.
  - Remarkably, changes leading to Lys and Glu, we observed a frequency two times higher in Lys compared to changes leading to Glu, despite both being resulted by *Ntv*.
- **Asn:**
  - The most frequent changes were observed in Ser (SB) and Asp frequency values 0.31 and 0.20, both are resulted by *Nti*.
- **Lys:**
  - The most frequent amino acid changes were observed in Arg (SB) with a frequency value of 0.24. This could be explained by the fact that both Lys and Arg are positively charged amino acids.
  - Interestingly, among *Ntv* changes, the most frequent change was observed in Gln with a frequency value of 0.19.
  - Despite having a *Nti* possibility, changes to Glu are less preferred than changes to Gln.
- **Asp:**
  - The most frequent amino acid changes were observed in Asn with a frequency

value of 0.3727 followed by Gly with 0.1271. Both have a similar mechanism of *Nti*. Both, Asp and Asn are structurally similar.

  - Among *Ntv* changes, Glu changes were noted with the highest frequency value of 0.33.

- **<u>Glu:</u>**
  - The most frequent amino acid changes were observed in Lys with a frequency value of 0.2758 followed by Gly with 0.1093. Both have a similar mechanism of *Nti*. Interestingly, these two amino acids are known as negatively charged hence the intra SB exchange was preferentially observed between the Asp and Glu exchangeability in a reversible manner.
  - Among *Ntv* changes, Glu changes were noted with the highest frequency value of 0.36.
- **<u>Cys:</u>**
  - The most frequent amino acid changes were observed in Ser (SB) with a frequency value of 0.3409 which is resulted by U→A *tv*.
  - Interestingly, Ser (FB) recorded 0.05 frequency value despite being a G→C *tv*.
  - Among *Nti* changes, Tyr recorded the highest change with a frequency value of 0.2841.

- **<u>Trp:</u>**
  - The most frequent amino acid changes were observed in Arg (SB) with a frequency value of 0.2698 resulting through a *ti*.
  - Among the *Ntv* changes, the most frequent amino acid changes were observed in Cys with a frequency value of 0.2512. Interestingly, Trp was noted with the

lowest amino acid acceptability among the 20 amino acids in this study.

- **Arg (SB):**
  - The most frequent amino acid changes were observed in His with a frequency value of 0.3795.
  - Interestingly, we observed four times more frequent amino acid exchangeability in His (0.37) while comparing with Gln (0.09).
  - Similarly, we observed ten times more frequent amino acid exchangeability in Cys (0.25) while comparing with Trp (0.027).
- **Ser (SB):**
  - The most frequent amino acid changes were observed in Asn with a frequency value of 0.3991 resulting through a *ti*.
  - It is interesting to note that the frequency values for changes leading to Gly and Cys were similar, despite Gly being changed by *Nti* and Cys being changed by *Ntv*.
- **Arg (SB):**
  - The most frequent amino acid changes were observed in Lys with the maximum *Nti* frequency value of 0.607.
  - The highest *Nti* frequency was observed in the changes involving Arg (SB)→Lys.

- **Gly:**
  - The most frequent amino acid changes were observed in Ser (SB) and Asp with frequency values of 0.363 and 0.238 respectively.
  - Interestingly, changes to Asp (0.238) were observed to be three times higher

than changes to Glu (0.085), even though the mechanism of change is similar for both the cases.

- Similarly, changes to SB of Ser (0.363) were observed to be 4.71 times more frequent than changes to SB of Arg (0.077).
- Among the *Ntv* comparison between Ala and Arg (FB), we observed a three times more frequent values in Ala (0.085) than FB of Arg (0.023). It suggests that Gly prefers Nti for its amino acid frequent exchangeability with Ser (SB) and Asp.

The summary of the amino acid exchangeability is represented in table 3.7.

Table 3.7. Summary of the amino acid exchangeability in *E. coli*

Amino acids	Most frequent changes to	Least frequent changes to	Most irreversible changes with	Remark
Phe	Leu (SB)	Ser (FB)	Tyr	Aromatic amino acids exchangeability
Leu (SB)	Phe	Trp	N/A	UUR-UUY *tv* changes
Leu (FB)	Phe	His	Pro	Higher *tv* in CUR
Ile	Val	Arg (SB)	Met	Most changes to Ile, Ile are closer to Val
Met	Ile	Lys	Ile	AUG to AUA frequent *ti*
Val	Ile	Asp	Ile	Structural similarity
Ser (FB)	Ala	Trp	Pro	UCA with maximum *Ntv* values
Pro	Ser (FB)	Arg (FB)	Ser and Leu (FB)	Second least favoured changes to Pro
Thr	Ala	Arg (SB)	Ala	Changes from Ala is favoured despite different hydrophobicity
Ala	Thr	Pro	Val	Changes from Val less favoured despite similar hydrophobicity
Tyr	Phe	Ser (FB)	His	Presence of aromatic ring similarity with His
His	Tyr	Asp	Tyr	Presence of aromatic ring similarity with Tyr
Gln	Leu (FB)	Pro	Leu (FB) and Arg (FB)	Changes to Glu is less favoured despite similarities in structure
Asn	Ser (SB)	Ile	N/A	Similar exchangeability with Ser (SB)
Lys	Arg (SB)	Met	N/A	Changes to Glu are less favoured than Gln
Asp	Asn	Val	N/A	Similar exchangeability with Glu/Negatively charges
Glu	Asp	Val	N/A	Similar exchangeability with Asp/Negatively charges
Cys	Ser (SB)	Trp	Ser (SB) and Tyr	Favouring Ser changes through U to A tv over G to C
Trp	Arg (SB)	Ser (FB)	Cys and Arg (FB)	Least changes to Trp
Arg (FB)	His	Pro	His	Favouring His over Gln due to slight positive charge
Ser (SB)	Asn	Ile	N/A	Similary hydrophilicity with Asn
Arg (SB)	Lys	Trp	Arg (SB)	Positively charged amino acids
Gly	Ser (SB)	Trp	Asp and Ser (SB)	Probable role of selection due to the preferrence of Asp over Glu and Ser over Arg

Table 3.5. The 20*20 amino acid matrix represents the frequency of amino acid changes from one amino acid to the other amino acid

Amino Acids	Total Changes	Phe	Leu (SB)	Leu (FB)	Ile	Met	Val	Ser (FB)	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Trp	Arg (FB)	Ser (SB)	Arg (SB)	Gly
Phe	640		0.242	0.233	0.095		0.094	0.056				0.213							0.067					
Leu (SB)	653	0.182			0.090	0.101	0.152	0.109												0.009				
Leu (FB)	1699	0.254			0.144	0.213	0.085		0.056				0.020	0.039							0.048			
Ile	1667	0.048	0.017	0.181		0.114	0.443			0.122					0.026	0.004						0.043	0.001	
Met	1069		0.081	0.090	0.481		0.132			0.075						0.009							0.011	
Val	3587	0.023	0.040	0.060	0.488	0.140					0.133						0.003	0.020						0.033
Ser (FB)	1026	0.114	0.157						0.162	0.192	0.307	0.032							0.032	0.004				
Pro	2014			0.261				0.444		0.081	0.078		0.026	0.166							0.013			
Thr	2992				0.224	0.127		0.121	0.047		0.282				0.084	0.041						0.073	0.023	
Ala	6837						0.341	0.139	0.020	0.387							0.024	0.055						0.060
Tyr	386	0.415						0.054					0.246		0.070		0.060		0.155					
His	1029			0.060					0.051			0.420		0.066	0.162		0.027				0.152			
Gln	2079			0.241					0.042				0.141			0.182		0.083			0.224			
Asn	1354				0.026					0.097		0.038	0.103			0.132	0.203					0.320		
Lys	1178				0.047	0.0263				0.119				0.199	0.149			0.115					0.240	
Asp	2753						0.025				0.069	0.048	0.019		0.373			0.338						0.127
Glu	2625						0.061				0.096			0.090		0.276	0.368							0.109
Cys	264	0.133						0.057				0.284								0.027	0.110	0.341		0.049
Trp	215		0.195					0.047											0.251		0.167		0.270	0.070
Arg (FB)	2432			0.097					0.003				0.380	0.094					0.258	0.028		0.102		0.028
Ser (SB)	1526				0.057					0.091					0.399				0.130		0.077		0.098	0.147
Arg (SB)	140				0.043	0.036				0.043						0.607				0.007		0.207		0.079
Gly	2410						0.060				0.085						0.239	0.085	0.056	0.006	0.027	0.364	0.077	

Legend. In the left column, the amino acids represent the FROM amino acids, while the top row lists the TO amino acid changes. For example, the normalization value of the Phe → Leu (SB) exchangeability would be 0.242. Similarly, values for exchanges such as Phe → Leu (SB)/Ile/Val/Ser (FB)/Tyr/Cys can be derived from the first row. Hence, values for other amino acids can be interpreted in a similar pattern.

Table 3.6. Individual amino acid level analysis shows the frequencies and exchangeability (From→To) of amino acids in different positions of the codons of the amino acids. The colour scales shows the higher frequency values for each amino acids.

Amino acid	Individual amino acid wise study								
Phe	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Leu (FB)	0.2328	U→C	Ser (FB)	0.0563	U→C	Leu (SB)	0.2422	U/C→A/G
	Ile	0.0953	U→A	Tyr	0.2125	U→A			
	Val	0.0938	U→G	Cys	0.0672	U→G			
Leu (SB)	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Ile	0.0904	U→A	Ser (FB)	0.1087	U→C	Phe	0.1822	A/G→U/C
	Met	0.1011	U→A	Trp	0.0092	U→G			
	Val	0.1516	U→G						
Leu (FB)	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Phe	0.2537	C→U	Pro	0.0559	U→C	N/A		
	Ile	0.1442	C→A	His	0.0200	U→A			
	Met	0.2131	C→A	Gln	0.0394	U→A			
	Val	0.0848	C→G	Arg (FB)	0.0483	U→G			
Ile	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Phe	0.0480	A→U	Thr	0.1224	U→C	Met	0.1140	U/C/A→G
	Leu (SB)	0.0174	A→U	Asn	0.0264	U→A			
	Leu (FB)	0.1806	A→C	Lys	0.0036	U→A			
	Val	0.4433	A→G	Ser (SB)	0.0432	U→G			
				Arg (SB)	0.0012	U→G			
Met	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Leu (SB)	0.0814	A→U	Thr	0.0748	U→C	Ile	0.4808	G→U/C/A
	Leu (FB)	0.0898	A→C	Lys	0.0094	U→A			
	Val	0.1319	A→G	Arg (SB)	0.0112	U→G			
Val	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Phe	0.0231	G→U	Ala	0.1330	U→C	N/A		
	Leu (SB)	0.0401	G→U	Asp	0.0025	U→A			
	Leu (FB)	0.0599	G→C	Glu	0.0195	U→A			
	Ile	0.4882	G→A	Gly	0.0335	U→G			
	Met	0.1399	G→A						
Ser (FB)	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Pro	0.1618	U→C	Phe	0.1140	C→U	N/A		
	Thr	0.1920	U→A	Leu (SB)	0.1569	C→U			
	Ala	0.3070	U→G	Tyr	0.0322	C→A			
				Cys	0.0322	C→G			
				Trp	0.0039	C→G			

Pro	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Ser (FB)	0.4444	C→U	Leu (FB)	0.2607	C→U	N/A		
	Thr	0.0814	C→A	His	0.0263	C→A			
	Ala	0.0780	C→G	Gln	0.1663	C→A			
				Arg (FB)	0.0129	C→G			
Thr	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Ser (FB)	0.1207	A→U	Ile	0.2236	C→U	N/A		
	Pro	0.0468	A→C	Met	0.1270	C→U			
	Ala	0.2824	A→G	Asn	0.0839	C→A			
				Lys	0.0411	C→A			
				Ser (SB)	0.0729	C→G			
				Arg (SB)	0.0234	C→G			
Ala	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Ser (FB)	0.1392	G→U	Val	0.3408	C→U	N/A		
	Pro	0.0199	G→C	Asp	0.0243	C→A			
	Thr	0.3869	G→A	Glu	0.0546	C→A			
				Gly	0.0603	C→G			
Tyr	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	His	0.2461	U→C	Phe	0.4145	A→U	N/A		
	Asn	0.0699	U→A	Ser (FB)	0.0544	A→C			
	Asp	0.0596	U→G	Cys	0.1554	A→G			
His	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Tyr	0.4198	C→U	Leu (FB)	0.0603	A→U	Gln	0.0661	U/C→A/G
	Asn	0.1623	C→A	Pro	0.0505	A→C			
	Asp	0.0272	C→G	Arg (FB)	0.1516	A→G			
Gln	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Lys	0.1818	C→A	Leu (FB)	0.2405	A→U	His	0.1414	A/G→U/C
	Glu	0.0827	C→G	Pro	0.0423	A→C			
				Arg (FB)	0.2237	A→G			
Asn	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Tyr	0.0377	A→U	Ile	0.0258	A→U	Lys	0.1322	U/C→A/G
	His	0.1027	A→C	Thr	0.0968	A→C			
	Asp	0.2031	A→G	Ser (SB)	0.3198	A→G			
Lys	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Gln	0.1986	A→C	Ile	0.0467	A→U	Asn	0.1494	A/G→U/C
	Glu	0.1154	A→G	Met	0.0263	A→U			
				Thr	0.1188	A→C			
				Arg (SB)	0.2402	A→G			
Asp	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Tyr	0.0483	G→U	Val	0.0254	A→U	Glu	0.3378	U/C→A/G
	His	0.0193	G→C	Ala	0.0694	A→C			
	Asn	0.3727	G→A	Gly	0.1271	A→G			
Glu	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv

	Lys	0.2758	G→A	Val	0.0613	A→U	Asp	0.3676	A/G→U/C
	Gln	0.0895	G→C	Ala	0.0964	A→C			
				Gly	0.1093	A→G			
Cys	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Arg (FB)	0.1098	U→C	Phe	0.1326	G→U	Trp	0.0265	U/C→G
	Ser (SB)	0.3409	U→A	Ser (FB)	0.0568	G→C			
	Gly	0.0492	U→G	Tyr	0.2841	G→A			
Trp	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Arg (FB)	0.1674	U→C	Leu (SB)	0.1953	G→U	Cys	0.2512	G→U/C
	Arg (SB)	0.2698	U→A	Ser (FB)	0.0465	G→C			
	Gly	0.0698	U→G						
Arg (FB)	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Cys	0.2578	C→U	Leu (FB)	0.0970	G→U	N/A		
	Trp	0.0275	C→U	Pro	0.0029	G→C			
	Ser (SB)	0.1024	C→A	His	0.3795	G→A			
	Gly	0.0275	C→G	Gln	0.0938	G→A			
Ser (SB)	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Cys	0.1304	A→U	Ile	0.0570	G→U	Arg (SB)	0.0983	A/G→U/C
	Arg (FB)	0.0773	A→C	Thr	0.0911	G→C			
	Gly	0.1468	A→G	Asn	0.3991	G→A			
Arg (SB)	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Trp	0.0071	A→U	Ile	0.0429	G→U	Ser (SB)	0.2071	A/G→U/C
	Gly	0.0786	A→G	Met	0.0357	G→U			
				Thr	0.0429	G→C			
				Lys	0.6071	G→A			
Gly	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv	Exchanged by	Frequency	ti/tv
	Cys	0.0560	G→U	Val	0.0598	G→U	N/A		
	Trp	0.0062	G→U	Ala	0.0855	G→C			
	Arg (FB)	0.0274	G→C	Asp	0.2386	G→A			
	Ser (SB)	0.3639	G→A	Glu	0.0855	G→A			
	Arg (SB)	0.0772	G→A						

3.5. Discussion

Theoretically, codons offer more possibilities for *Ntv*, but it has been observed that *Nti* are more common in various intra-species and inter-species studies, including those involving humans (Zhang, 2000; Freudenberg-Hua et al., 2003). *Tvs* are more frequently purged out of the population during the process of selection (Vartanian et al., 1996; Hurst and Pal, 2001). In this study, we tried to understand the much debatable topic in the blooming field of molecular evolution. Our endeavour in the normalization of *Nti* and *Ntv*, enabled us to do a comparative study among different codons of *E. coli* in terms of their frequency of non-synonymous changes. It has also been suggested that the structure of the genetic code table allows frequent *Nti* over *Ntv* which are believed to be less deleterious (Zou and Zhang, 2021). Accordingly, we observed more frequent *Nti* changes than *Ntv* in the FFD amino acids. Interestingly, FFD codons have an overall 40% of cytosine proportion excluding the 3rd position and TFD codons have 18.51% of cytosine content including the 3rd position. This could be one of the possible explanations for the frequent *Nti* observed in FFD codons making the FFD codons susceptible to cytosine deamination (C→U). However frequent amino acid exchangeabilities such as Phe→Tyr, Tyr→Phe, Ser (FB)→Ala, Asp→Glu, Glu→Asp, Cys→Ser (SB) were observed that are attributed by frequent *Ntv* changes. Regarding such higher *Ntv* changes, it is noteworthy that the functional groups of FFD amino acids are usually less bulky than the functional groups of the amino acids mentioned above. TFD codons code for amino acids which have high economic importance (Akashi and Gojobori, 2002), positively/negatively charged and other cellular functions such as cellular signalling and enzymatic activities. Considering the frequent Gly changes to other amino acids, we have observed three times more frequent changes to Asp than Glu. Despite being sharing similar property, both Asp and Glu were expected to be exchanged at a similar rate by Glu. Hence the role of selection even if Gly→Asp/Glu resulted by the similar mechanism (G→A *ti*) could not be denied. In another comparative analysis regarding the exchangeability of Gly→Ser (SB) and Gly→ Arg (SB), we observed a 4.71 times more frequent changes in Gly→Ser (SB). The smaller functional group of Ser might be a reason behind the frequent Gly→Ser frequent changes conferring to stereochemistry. In a comparison between Ala→Thr and Ala→Val, surprisingly we observed 1.13 times more frequent amino acid exchangeability in the former, whereas depending upon the hydrophobicity the adverse result was anticipated. Hence, selection along with intrinsic factors such as hydrophobicity, stereochemistry and economy of the amino acid might be playing a contributing role in shaping up the evolutionary prospects in microbes.

The evolutionary aspect of the distant placement of Ser FB and Ser SB in the genetic code table raises questions over their disparity in usage in *E. coli* (Inouye et al., 2020). We observed that, the *Ntv* changes in Cys→Ser prefers SB changes six times more frequently than FB changes. Astonishingly, Thr→Ser (FB) and Thr→Ser (SB) could not show any disparity like Cys→Ser change. More frequent changes were observed between AUG→AUA, which is the prime contributing factor behind the higher *Nti*' values in Met. Such observations raise primary concern over the structure and evolution of the genetic code table. Our most irreversible changes study revealed many such scenarios of amino acid exchangeabilities such as higher Arg→Lys over Lys→Arg, Trp→Cys over Cys→Trp etc which indicate the role of selection at amino acid levels for the survival of the organism.

Interestingly, Leu (FB and SB) recorded lower $\frac{Nti'}{Ntv'}$ value. But the higher $\frac{Nti'}{Ntv'}$ values in the SB of Arg and Ser over their FB raises concerns regarding their evolutionary history and conservation in *E. coli*. Coincidently, the top row in the classical genetic code table i.e. UNN encoded amino acids recoded the least $\frac{Nti'}{Ntv'}$ value. The amino acid acceptance value of Ile was observed to be the highest among all the amino acids which lucidly makes Ile the most neutral amino acid in our work. Whereas Trp and Pro were among the least amino acids acceptance values. Hence further investigations are required to understand the exchangeability of amino acids in organisms. The non-synonymous *ti* bias between TFD codons and FFD codons is due to selection or mutation, or the confounding impact of both will be an interesting area of research in the future.

3.6. Bibliography

- Agrawal, A. F., & Whitlock, M. C. (2012). Mutation load: the fitness of individuals in populations where deleterious alleles are abundant. *Annual Review of Ecology, Evolution, and Systematics*, *43*, 115-135.
- Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences*, *99*(6), 3695-3700.
- Aziz, R., Sen, P., Beura, P. K., Das, S., Tula, D., Dash, M. & Ray, S. K. (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*(4), dsac023.
- Beura, P. K., Aziz, R., Sen, P., Das, S., Namsa, N. D., Feil, E. J., ... & Ray, S. K. (2022). Synonymous and non-synonymous transitions/transversions vividly disclose purifying selection in Escherichia coli coding sequences. *bioRxiv*, 2022-11.
- Beura, P. K., Sen, P., Aziz, R., Satapathy, S. S., & 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. *Journal of Genetics*, *102*(1), 22.
- Bhagwat, A. S., Hao, W., Townes, J. P., Lee, H., Tang, H., & Foster, P. L. (2016). Strand-biased cytosine deamination at the replication fork causes cytosine to thymine

mutations in Escherichia coli. *Proceedings of the National Academy of Sciences*, *113*(8), 2176-2181.

- Charlesworth, B., & Charlesworth, D. (1998). Some evolutionary consequences of deleterious mutations. *Genetica*, *102*, 3-19.
- Chen, Q., He, Z., Lan, A., Shen, X., Wen, H., & Wu, C. I. (2019). Molecular evolution in large steps—codon substitutions under positive selection. *Molecular biology and evolution*, *36*(9), 1862-1873.
- Dang, C. C., Le, Q. S., Gascuel, O., & Le, V. S. (2010). FLU, an amino acid substitution model for influenza proteins. *BMC evolutionary biology*, *10,* 1-11.
- Duchêne, S., Ho, S. Y., & Holmes, E. C. (2015). Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. BMC evolutionary biology, 15, 1-10.
- Freudenberg-Hua, Y., Freudenberg, J., Kluck, N., Cichon, S., Propping, P., & Nöthen, M. M. (2003). Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. *Genome research*, *13*(10), 2271-2276.
- Gerber, A. P., & Keller, W. (1999). An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science*, *286*(5442), 1146-1149.
- Goldman, N., & Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, *11*(5), 725-736.
- Hurst, L. D., & Pál, C. (2001). Evidence for purifying selection acting on silent sites in BRCA1. *TRENDS in Genetics*, *17*(2), 62-65.

- Inouye, M., Takino, R., Ishida, Y., & Inouye, K. (2020). Evolution of the genetic code; Evidence from serine codon use disparity in Escherichia coli. *Proceedings of the National Academy of Sciences*, *117*(46), 28572-28575.
- Lewis Jr, C. A., Crayle, J., Zhou, S., Swanstrom, R., & Wolfenden, R. (2016). Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proceedings of the National Academy of Sciences*, *113*(29), 8194-8199.
- Modiano, G., Battistuzzi, G., & 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?. *Proceedings of the National Academy of Sciences*, *78*(2), 1110-1114.
- Schrider, D. R., Houle, D., Lynch, M., & Hahn, M. W. (2013). Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. *Genetics*, *194*(4), 937-954.
- Schmidt, S., Gerasimova, A., Kondrashov, F. A., Adzuhbei, I. A., Kondrashov, A. S., & Sunyaev, S. (2008). Hypermutable non-synonymous sites are under stronger negative selection. *PLoS genetics*, *4*(11), e1000281.
- Sen, P., Aziz, R., Deka, R. C., Feil, E. J., Ray, S. K., & Satapathy, S. S. (2022). Stem region of tRNA genes favors transition substitution towards keto bases in bacteria. *Journal of Molecular Evolution*, *90*(1), 114-123.
- Thorpe, H. A., Bayliss, S. C., Hurst, L. D., & Feil, E. J. (2017). Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, *206*(1), 363-376.

- Weber, C. C., & Whelan, S. (2019). Physicochemical amino acid properties better describe substitution rates in large populations. *Molecular biology and evolution*, *36*(4), 679-690.
- Vartanian, J. P., Henry, M., & Wain-Hobson, S. (1996). Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions. *Nucleic acids research*, *24*(14), 2627-2631.
- Zou, Z., & Zhang, J. (2019). Amino acid exchangeabilities vary across the tree of life. *Science Advances*, *5*(12), eaax3124.
- Zou, Z., & Zhang, J. (2021). Are nonsynonymous transversions generally more deleterious than nonsynonymous transitions?. *Molecular biology and evolution*, *38*(1), 181-191.
- Zhang, J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of molecular evolution*, *50*, 56-68.