Difference in synonymous polymorphism related to codon degeneracy between co-transcribed genes in the genome of *Escherichia coli*

Pratyush Kumar Beura[1], Piyali Sen[2], Ruksana Aziz[1], Chayanika Chetia[1], Madhusmita Dash[3], Siddhartha Shankar Satapathy[2,4] and Suvendra Kumar Ray[1,4,*]

[1]Department of Molecular Biology and Biotechnology, Tezpur University, Napaam 784 028, India
[2]Department of Computer Science and Engineering, Tezpur University, Napaam 784 028, India
[3]Department of Electronics and Communication Engineering, National Institute of Technology, Jote, Papum Pare 791 113, India
[4]Centre for Multidisciplinary Research, Tezpur University, Napaam 784 028, India

In our study, we compared synonymous polymorphism in co-transcribed gene pairs within five well-known *Escherichia coli* operons (*rpoB*/*C*, *lacZ*/*Y*, *kdpA*/*B*, *araB*/*A* and *bcsA*/*B*). Interestingly, the transition to transversion ratio between gene pairs were different due to their compositional differences of two-fold and four-fold degenerate codons. The differences in polymorphism spectra were more pronounced in four-fold and six-fold codons compared to two-fold degenerate codons. Notably, *rpoB* and *rpoC* showed significant distinctions in UCC, GUA, CCG, GCU, GGC and CGC codons. Similar trends were observed in other gene pairs, particularly in higher degenerate codons. Notably, two-fold degenerate codons primarily exhibited synonymous polymorphisms through transitions, while higher degenerate codons encompassed both transition and transversion events. This underscores the intriguing role of degenerate codons in molecular evolution.

Keywords: Base substitution, codon degeneracy, co-transcribed genes, replication and transcription, synonymous polymorphism.

BASE substitution mutation is a major event of molecular evolution in organisms, influenced by different factors such as DNA replication, damage in DNA bases such as deamination of cytosine/adenine, oxidation of guanine[1,2], gene expression, recombination, etc. The asymmetry in DNA replication results in different mutation patterns between the leading strand (LeS) and lagging strand (LaS) in genomes, resulting in the former being enriched with keto nucleotides and the latter with amino nucleotides in bacteria[3,4]. In addition, genes near the origin of replication exhibit different mutation patterns than those at the terminus of replication: in bacteria, the replication terminus region in a chromosome is known to be AT-enriched compared to the origin of replication[5]. The role of transcription in causing mutation asymmetry in genes resulting in higher C → T changes in the non-template strand than those in the template strand has been described recently[6]. Therefore, the expression level of genes has different impacts on mutation rates[7–9]. It is important to accurately compare polymorphism patterns between two genes because it helps us understand further the role of intrinsic factors in polymorphism differences between the genes, if any.

In bacteria, one advantage of studying molecular evolution is that two functionally related genes co-transcribed in an operon are adjacent, localized in the same strand, and similar with respect to their gene expression at the transcription level. Additionally, random drift will be considered minimal in bacterial genes where the population is large, unlike in higher multicellular organisms. Therefore, polymorphism patterns of two adjacent genes in an operon are likely to be similar unless there are some unknown factors causing mutation and/or selection biases in these genes. In addition to replication, localization, and transcription, co-transcribed genes might differ from each other regarding amino acid level selection on their protein structure in a genome.

In this study, we perform a comparative analysis of polymorphism spectra in two adjacently placed co-transcribed genes in the *Escherichia coli* genome by comparing gene sequences across multiple strains. We have considered five pairs of co-transcribed genes (*rpoB*/*C*, *lacZ*/*Y*, *kdpA*/*B*, *araB*/*A* and *bcsA*/*B*) present at different loci in the *E. coli* genome (Figure 1). Surprisingly, the two genes in all five operons were found to be significantly different from each other with respect to their synonymous polymorphism pattern. This indicates that variations arising in the genes are not identical despite the genes being similar regarding replication, transcription and strand localization. This was

[*]For correspondence. (e-mail: suven@tezu.ernet.in)

corroborated by the observation that the phylogeny of the strains using the two co-transcribed genes was not identical. This study also reveals the role of codon degeneracy in synonymous polymorphism. To the best of our knowledge, there are no previous studies comparing the polymorphism spectra between two co-transcribed genes in bacteria.

Materials and methods

Selection of co-transcribed gene pairs from the available dataset

In this study, we have considered 157 strains of *E. coli* for which alignments are available in the public database[10]. The criteria for selection of genes were set (>1200 bp or >400 codons), anticipating a considerable number of synonymous substitutions in each gene. Among five pairs of co-transcribed genes, four pairs were localized in LeS and only one pair (*bcsA*/*B*) was present in LaS. The Supplementary Table 1 provides the list of genes and their detailed information. We wanted to examine only base substitution polymorphism and no other variation, such as insertion or deletion. Hence, the genes selected in this study were identical in size across all strains. For example, the size of *rpoB* was 4029 bp and *lacZ* was 3075 bp across all strains.

Derivation of reference sequence and identifying the overall synonymous spectra

We maintained a common set of strains for each pair of co-transcribed genes in this study. The compositional details along with different skew values and synonymous site values of the reference sequences for all the genes, were determined (Supplementary Table 2). The reference sequence was derived for each gene by considering the most frequent nucleotide present in a certain position in the alignment as the reference nucleotide for that position. This procedure was followed according to the methodology developed in our laboratory[11] (Supplementary Table 3). Only synonymous polymorphisms were considered in this study. The polymorphism values were normalized by dividing them with the synonymous site values of the respective nucleotides in the genes. For each gene, the synonymous site values of each nucleotide were estimated by taking the summation of all possible synonymous substitutions involving the nucleotide in each codon. In the case of UUU, the synonymous site for U at the third position is considered as 1 (one synonymous transition (STi) and zero synonymous transversion (STv)), but for GGU, the synonymous site for U at the third position is considered as 3 (one STi and two STvs). Suppose we observe 30 C → T substitutions, and the synonymous site value of C nucleotides in the gene is 150. Then, the normalized C → T polymorphism frequency is calculated as $30/150 = 0.2$. We estimated the normalized values of all the 12 polymorphisms in the five gene pairs.

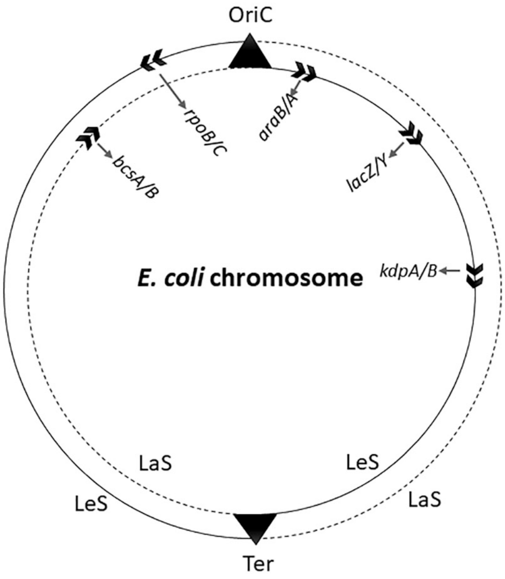

Figure 1. Schematic diagram indicating the relative position of five gene pairs in the leading strand (LeS) and lagging strand (LeS) of the *Escherichia coli* chromosome. LeS and LaS are shown as continuous and dotted lines respectively. The gene locations shown are not to scale.

Calculating the ratio of transition to transversion in gene pairs

Transition (Ti) to transversion (Tv) ratio of all the genes was calculated. Overall, the Ti/Tv ratio of the genes and that at the fourfold degenerate (FFD) site were calculated to show the variation between cotranscribed genes. The overall STi/STv values were obtained from the initial non-normalized synonymous spectra table (Supplementary Table 4). For the FFD site, the Ti and Tv values were estimated by considering synonymous polymorphism for individual genes.

Comparison of polymorphism frequency at the codon level

From the reference sequence of each gene, we determined the codon count using the web portal http://agnigarh.tezu.ernet.in/~ssankar/cbb_tu.html[12]. All the genes were analysed for polymorphism at FFD sites (20 codons), twofold degenerate (TFD) sites (18 codons) and sixfold degenerate (SFD) (12 codons). The SFD amino acid (Leu, Ser, Arg) codons are grouped differently, such as family box (FB; CUN, UCN, CGN) and split box (SB; UUR, AGY, AGR) codons. Due to fewer polymorphisms present in SFD SB, only FB codons of SFD (12 in number) were considered for the comparison of polymorphism. The synonymous polymorphism at the third position of the codons was determined for each box separately. While comparing

Table 1. Synonymous polymorphism spectra of genes presented in a tabular form

Genes	A → T	A → C	A → G	T → A	T → C	T → G	C → A	C → T	C → G	G → A	G → T	G → C
rpoB	0.011	0.006	0.034	0.006	0.024	0.005	0.005	0.091	0.004	0.021	0.007	0.001
rpoC	0.013	0.005	0.032	0.009	0.020	0.008	0.005	0.080	0.001	0.030	0.009	0.003
lacZ	0.007	0.011	0.090	0.009	0.062	0.011	0.013	0.062	0.001	0.068	0.013	0.008
lacY	0.008	0.024	0.039	0.004	0.020	0.008	0.004	0.056	0.000	0.049	0.011	0.004
kdpA	0.008	0.008	0.119	0.026	0.111	0.018	0.020	0.119	0.017	0.083	0.012	0.004
kdpB	0.012	0.017	0.069	0.012	0.099	0.009	0.016	0.118	0.019	0.100	0.018	0.004
araB	0.013	0.006	0.063	0.022	0.070	0.011	0.007	0.112	0.012	0.112	0.008	0.005
araA	0.013	0.051	0.152	0.022	0.141	0.018	0.022	0.103	0.011	0.084	0.017	0.006
bcsA	0.005	0.010	0.063	0.016	0.082	0.013	0.009	0.133	0.014	0.063	0.004	0.008
bcsB	0.010	0.021	0.068	0.013	0.075	0.000	0.012	0.119	0.007	0.045	0.017	0.010

Table 2. Transition/transversion ratio in the whole gene and at four-fold degenerate (FFD) sites

Ti/Tv	*rpoB*	*rpoC*	*lacZ*	*lacY*	*kdpA*	*kdpB*	*araB*	*araA*	*bcsA*	*bcsB*
Whole gene	4.310	2.929	3.615	3.083	3.686	3.854	4.917	3.171	4.344	3.361
FFD	1.192	1.441	1.939	2.428	3.091	2.867	3.733	1.724	3.400	2.040

the two genes for codon-wise polymorphism, we considered only those codons with abundance value ≥30 in a gene, and/or polymorphism observed was ≥5 in a codon. Finally, FFD, TFD and SFD polymorphism frequency information of a pair of genes was compared using Python.

Results

Difference in synonymous polymorphism spectra between co-transcribed genes

We compared the synonymous polymorphism values between five pairs of co-transcribed genes, viz. *rpoB*/*C*, *lacZ*/*Y*, *kdpA*/*B*, *araB*/*A* and *bcsA*/*B* in well-known operons in the *E. coli* genome (Table 1). As transition polymorphisms were more in number than transversion polymorphisms, we compared the genes with respect to transition polymorphism. The differences observed were as follows. In the case of *rpoB* and *rpoC* genes, C → T frequency values were 0.091 and 0.080 respectively, while G → A frequency values were 0.021 and 0.030 respectively. The C → T frequency value was more than fourfold higher than G → A in *rpoB*, whereas in *rpoC,* the C → T frequency was less than threefold higher than G → A. In the case of *kdpA* and *kdpB* genes, the A → G frequency values were 0.119 and 0.069 respectively; in *araB* and *araA*, the T → C frequency values were 0.070 and 0.141 respectively; in *bcsA* and *bcsB*, the G → A frequency values were 0.063 and 0.045 respectively and in *lacZ* and *lacY*, the A → G frequency values were 0.090 and 0.039 respectively. The above differences between the co-transcribed genes with regard to certain transition polymorphisms indicate that the co-transcribed genes are not evolutionarily identical. The occurrence of transversion polymorphisms was not high, and therefore, we avoided using them in the comparative analysis. Some common observations included C → T frequency being the highest among all the polymorphisms and frequency values between the complementary polymorphisms, such as C → T and G → A being significantly different (*P* value <0.05; Mann–Whitney *U* test) in all these genes.

Difference in Ti/Tv ratio between the co-transcribed genes due to difference in codon degeneracy composition

We analysed the STi to STv ratio in all genes (Table 2). It was evident that the two co-transcribed genes were different from each other with regard to the STi/STv ratio as follows: in *rpoB*, the value was 4.3, whereas in *rpoC*, it was 2.9; in *araB*, the value was 4.9, whereas in *araA* it was 3.1; in *bcsA* the value was 4.3, whereas in *bcsB* it was 3.3. We further estimated the STi/STv ratio in the FFD sites in all 10 genes. The value in the FFD sites was significantly lower (*P* value <0.01; Mann–Whitney *U* test) than that in the whole gene. This is because of the difference between Ti and Tv rates in an organism: a Ti is more frequent than a Tv. It is pertinent to note that synonymous polymorphisms in the TFD sites are only possible due to transitions, whereas in the FFD sites, they are possible due to both transition and transversion. We correlated the TFD : FFD ratio and Ti/Tv difference values between the whole gene and the FFD site. The positive correlation (Pearson *r* value of 0.677) suggested that the higher the TFD composition in a gene, the greater the difference in Ti/Tv ratio between the whole gene and the FFD site will be. Therefore, the compositional difference between TFD and FFD among the genes influences their synonymous Ti/Tv ratio (Supplementary Table 5). We then compared the Ti/Tv ratio across amino acid-specific FFD sites (Supplementary Table 6). The two co-transcribed genes were found to be distinctly different from each other in the case of certain amino acid codons. For Gly, the Ti/Tv ratio in *rpoB* and *rpoC* was 1.75 and 6.50 respectively. In

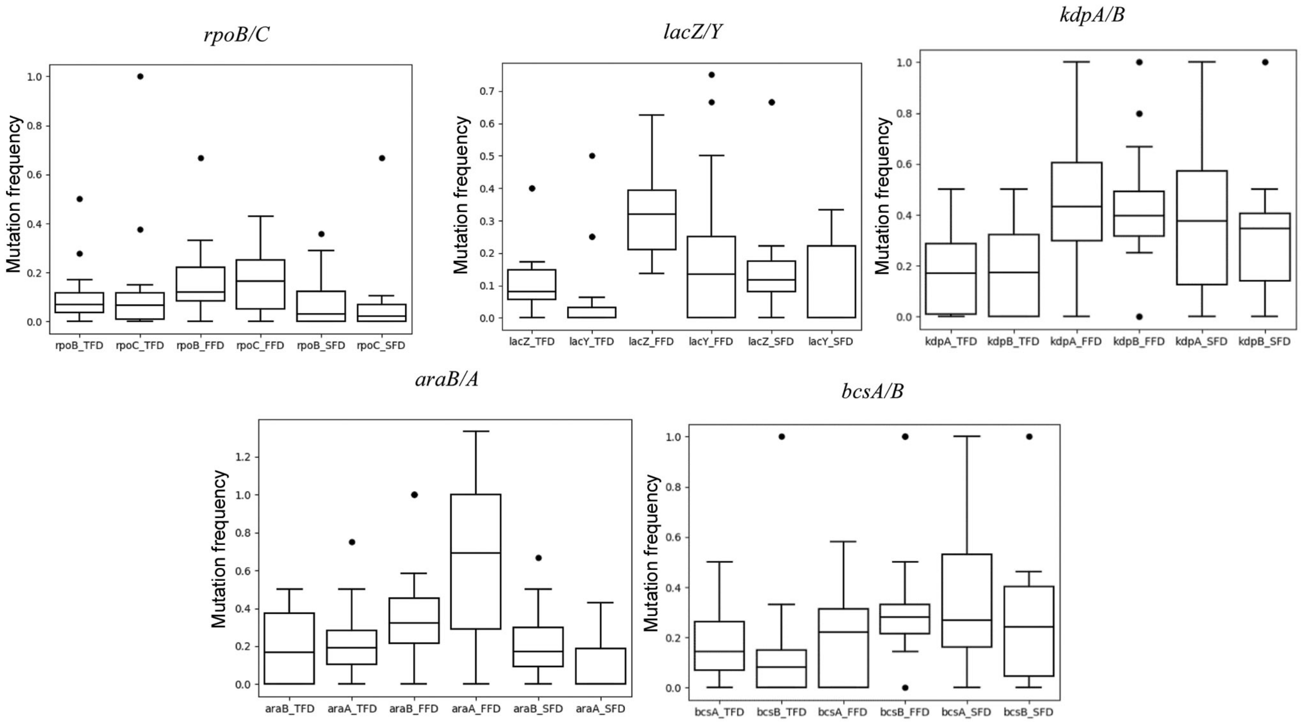

Figure 2. Box plot showing polymorphism frequency of five gene pairs in four-fold degenerate (FFD)/two-fold degenerate (TFD)/six-fold degenerate (SFD). Each plot shows the polymorphism frequency comparison of adjacent genes in the codons of FFD, TFD and SFD (family box) with polymorphism frequency on the *y*-axis and different degenerate codon box names on the *x*-axis.

the case of Ala, the Ti/Tv ratio in *lacZ* and *lacY* was 1.27 and 4.00 respectively.

Difference in synonymous polymorphism between co-transcribed genes in FFD, TFD and SFD (FB) codons

The rate of a transition is usually four times than that of a transversion[11]. The synonymous site value in a FFD codon is 1.000, and that in a TFD codon is 0.667 (ref. 13). Therefore, synonymous polymorphism in the FFD site and SFD (FB) site is expected to be 1.5 times more frequent than that in a TFD site. To compare synonymous polymorphism in the TFD, FFD and SFD (FB) sites between the co-transcribed gene pairs, we performed a codon-wise analysis in each codon (Supplementary Table 7). In *rpoB* and *rpoC*, synonymous polymorphism in the FFD codons such as GUA, CCG, GCU and GGC was found to be different. The codon count of GUA was 31 and 32 in *rpoB* and *rpoC* respectively, but the synonymous polymorphism frequency was 0.097 and 0.219 respectively. Similarly, the codon count of GCU was 19 and 28 in *rpoB* and *rpoC* respectively, but the frequency was 0.053 and 0.250 respectively. Analogously, the polymorphism frequency in CCG was estimated to be 0.079 and 0.178 in *rpoB* and *rpoC* respectively. The polymorphism frequency in GGC was 0.114 and 0.310 in *rpoB* and *rpoC* respectively. This indicates that *rpoB* and *rpoC* genes are different from each other with regard to synonymous polymorphism of certain codons.

Among SFD (FB) codons, a prominent difference was observed in the CUG, UCC and CGC codons. The polymorphism frequency in CUG was 0.060 and 0.104 in *rpoB* and *rpoC* respectively, while in UCC, it was 0.290 and 0.074 in *rpoB* and *rpoC* respectively. Likewise, the polymorphism frequency in CGC was 0.357 and 0.042 in *rpoB* and *rpoC* respectively. Regarding the TFD codons, the only noticeable difference was observed in CAC in the case of *rpoB* and *rpoC*. The polymorphism frequency in CAC was 0.278 and 0.000 in case of *rpoB* and *rpoC* respectively. We correspondingly observed the differences in various codons between the remaining co-transcribed genes according to the criteria. Supplementary Table 8 shows the polymorphism frequency comparison of five pairs of co-transcribed genes, and the FFD, TFD and SFD (FB) amino acids.

Interestingly, in *rpoB* and *rpoC*, out of the 12 codons belonging to SFD (FB), only three exhibited the difference. Similarly, out of the 20 codons belonging to FFD, only four exhibited the difference. Whereas, out of 18 TFD codons, only one exhibited a difference in both the co-transcribed genes. We extrapolated this observation in the remaining gene pairs (Supplementary Table 9). We then generated box plots for each pair of genes with regard to their obtained polymorphism frequencies in a degeneracy-wise manner (Figure 2). The box plots show the degeneracy-wise differences between co-transcribed gene pairs, of which

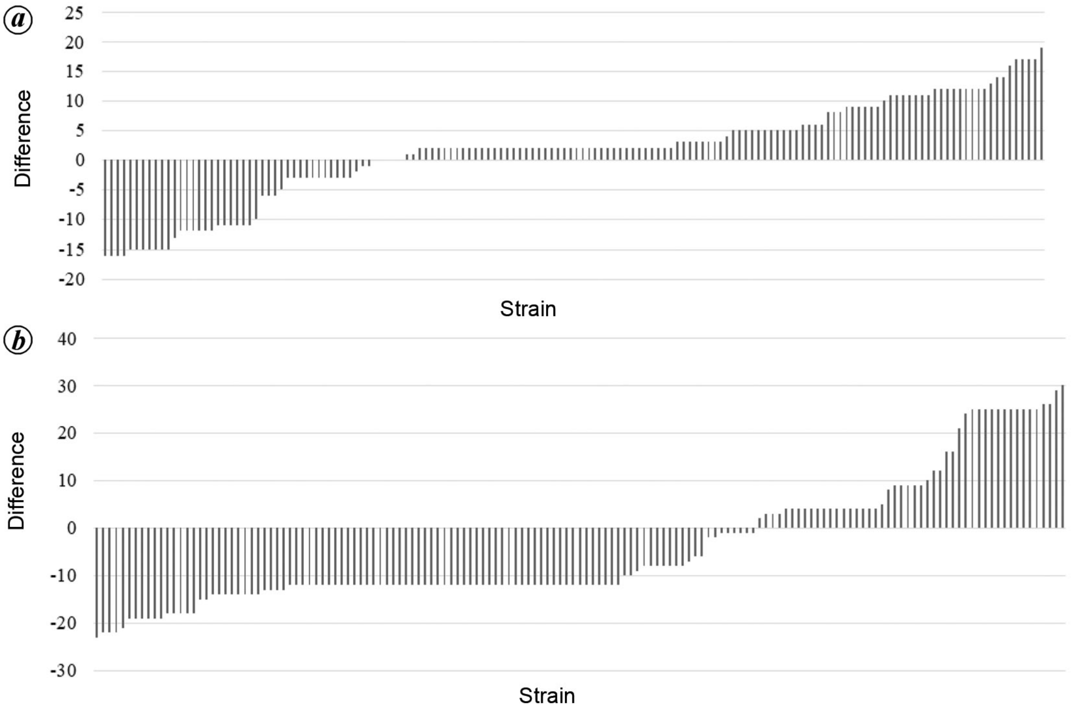

Figure 3. Individual strain-wise polymorphism difference in co-transcribed genes. ***a***, Gene pair *rpoB* and *rpoC*. The figure indicates that more number of individual substitutions are present in *rpoB*. ***b***, Gene pair *kdpA* and *kdpB*. The figure indicates that more number of individual substitutions are present in *kdpB*.

differences in the FFD codons were principally found between a pair of genes. Further, a Mann–Whitney *U* test was performed between FFD and TFD as well as FFD and SFD. The result was found to be significantly different at $P < 0.01$. In an interesting comparison between GAU and GGU codons in *araB* and *araA*, we observed a ~fivefold higher polymorphism frequency value in GGU regardless of the equal codon frequency in GAU and GGU. This is noticeable evidence of higher polymorphism frequency of the FFD codons observed in many cases.

Comparison of phylogeny between co-transcribed gene pairs

Additionally, to get an insight into the polymorphism difference between two co-transcribed gene pairs, we quantified the polymorphism difference between individual strains in the *rpoB/C* and *kdpA/B* gene pairs. We observed total polymorphism in individual strains with regard to co-transcribed pairs. The co-transcribed pairs had different numbers of polymorphisms in individual strains. For example, a strain with zero polymorphism in *rpoB* had 17 polymorphisms in *rpoC*. In Figure 3, the minimum to maximum polymorphism difference between *rpoB* and *rpoC* was –16 to 19. Whereas in the case of *kdpA* and *kdpB* the range was –23 to 30. It is evident that the co-transcribed genes differ, even at the individual strain level. We then generated phylogenetic trees using 10 common strains in the co-transcribed pairs of *rpoB/C* and *kdpA/B*. The comparative study of phylogeny indicated different patterns in the *rpoB*/*C* and *kdpA*/*B* gene pairs (Figure 4). This supports the observation that the co-transcribed genes are not identical with regard to polymorphism patterns in *E. coli.*

Discussion and conclusion

In the present study, co-transcribed gene pairs have been compared with regard to synonymous polymorphism in five operons in *E. coli*. The adjacent co-transcribed genes were found to be different with regard to synonymous polymorphism, though these genes in an operon were similar in case of replication, strand location and transcript-level expression. The compositional difference of degenerate codons between the gene pairs was found to be important for their different Ti/Tv ratios. The TFD codons can undergo synonymous polymorphism only by transition substitution, while the higher degenerate codons can undergo synonymous polymorphism by both transitions and transversion substitutions. The rate of a transition substitution is four times more than that of a transversion substitution, which results in a higher transversion proportion in FFD than TFD codons[14]. Therefore, Ti/Tv is higher in genes with fewer FFD codons. The present study has manifested the role of codon degeneracy due to the difference in polymorphism spectra of two co-transcribed genes. The FFD codons were the common cause of the difference in polymorphism spectra

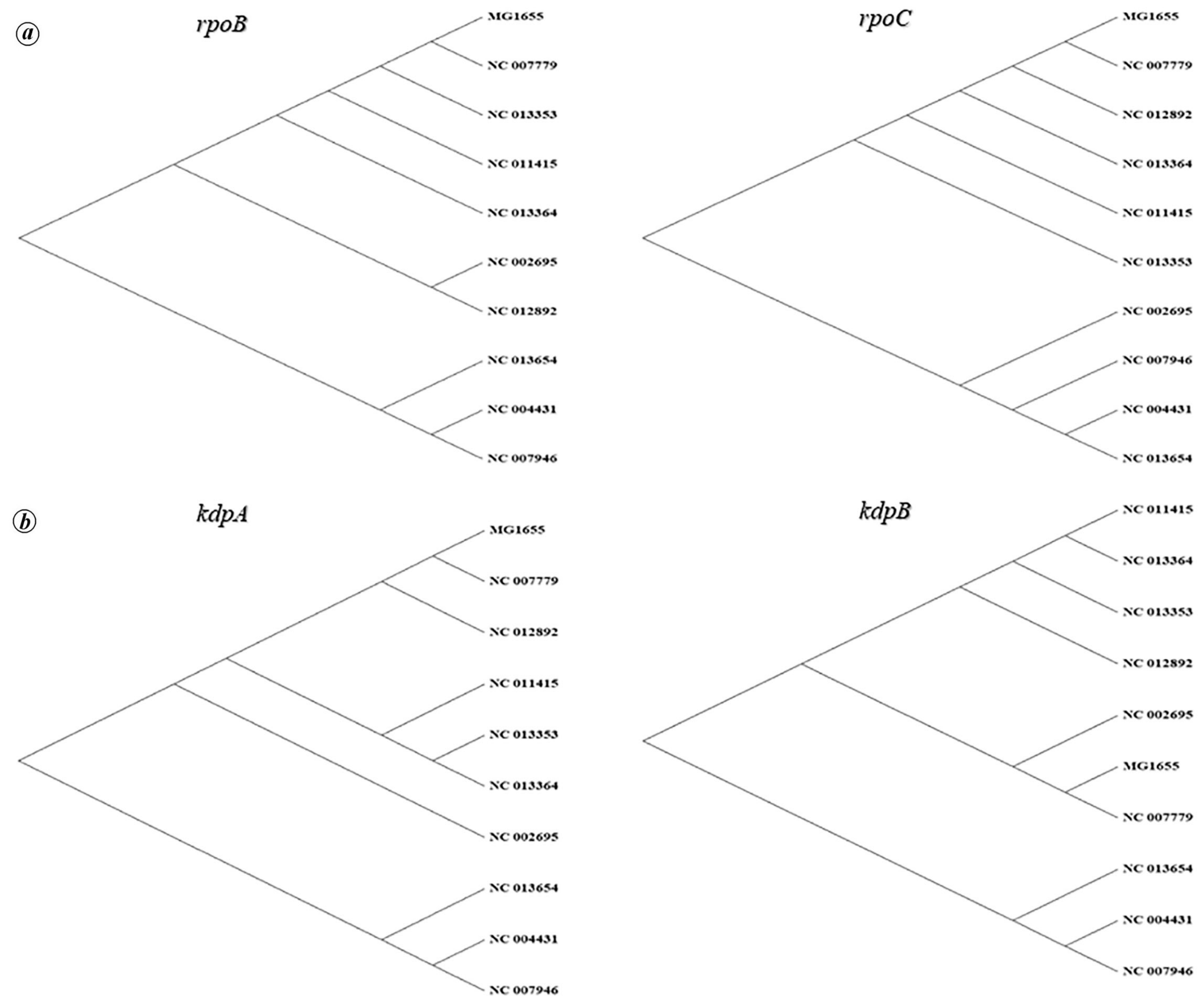

Figure 4. Phylogenetic comparative study between co-transcribed genes in *E. coli* using 10 common strains. Phylogenetic trees of (***a***) *rpoB* and *rpoC* and (***b***) *kdpA* and *kdpB*.

in each gene pair. In the case of the higher degeneracy codons, synonymous transversion is possible, unlike in the TFD codons. It will be examined to investigate whether this is the only reason for the difference or whether there are additional reasons. Moreover, synonymous changes are more diverse than non-synonymous changes, as purifying selection is stronger on non-synonymous changes in a genome. Hence, the selection of TFD codons might be stronger than higher degenerate codons between co-transcribed genes, as we have observed the randomness of synonymous polymorphism in higher degenerate codons between gene pairs. The role of degeneracy on polymorphism difference between the co-transcribed genes indicates a basis towards the neutral theory of evolution in this intra-species genome comparison analysis.

It has already been reported that although synonymous polymorphisms do not affect the amino acid sequence in a protein, they influence its function by protein folding[15]. Whether TFD and FFD of codons contribute differently to protein folding is yet to be determined. The difference between the degenerate codons invokes many fundamental questions in the evolution of the genetic code. Has degeneracy been assigned randomly to amino acids, or does it play any role in protein folding that influences its distribution? Future research will reveal more information about this.

Author disclosure statement: The authors declare no potential conflict of interest and no competing financial interest.

1. Rocha, E. P. C., Touchon, M. and Feil, E. J., Similar compositional biases are caused by very different mutational effects. *Genome Res.*, 2006, **16**(12), 1537–1547.
2. Kino, K. and Sugiyama, H., Possible cause of G.C→C.G transversion mutation by guanine oxidation product imidazolone. *Chem. Biol.*, 2001, **8**, 369–378.

3. Lobry, J. R., Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, 1996, **13**, 660–665.
4. Bulte, J. W., Zhang, S., van Gelderen, P., Herynek, V., Jordan, E. K., Duncan, I. D. and Frank, J. A., Neurotransplantation of magnetically labeled oligodendrocyte progenitors: magnetic resonance tracking of cell migration and myelination. *Proc. Natl. Acad. Sci. USA*, 1999, **96**(26), 15256–15261.
5. Daubin, V. and Perrie, G., G1C3 structuring along the genome: a common feature in prokaryotes. *Mol. Biol. Evol.*, 2003, **20**(4), 471–483.
6. Francino, M. P. and Ochman, H., Strand asymmetries in DNA evolution. *Trends Genet.*, 1997, **13**, 240–245.
7. Park, C., Qian, W. and Zhang, J., Scientific report in highly expressed genes. *EMBO Rep*., 2012, **13**, 1123–1129.
8. Lang, G. I. and Murray, A. W., Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biol. Evol*., 2011, **3**, 799–811.
9. Hodgkinson, A. and Eyre-Walker, A., Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, 2011, **12**, 1–11.
10. Thorpe, H. A., Bayliss, S. C., Hurst, L. D. and Feil, E. J., Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, 2017, **206**, 363–376.
11. Sen, P., Aziz, R., Deka, R. C., Feil, E. J., Satapathy, S. S. and Ray, S. K., Stem region of tRNA genes favors transition substitution towards keto bases in bacteria. *J. Mol. Evol*., 2022, **90**, 114–123.
12. Satapathy, S. S., Sahoo, A. K., Ray, S. K. and Ghosh, T. C., Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc (N̂c) and ENCprime (N̂′ c) measures. *Genes Cells*, 2017, **22**, 277–283.
13. 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 Res.*, 2022, **29**(4), 1–8.
14. Duchêne, S., Ho, S. Y. W. and Holmes, E. C., Declining transition/transversion ratios through time reveal limitations to the accuracy of nucleotide substitution models. *BMC Evol. Biol.*, 2015, **15**, 36.
15. Fung, K. L. and Gottesman, M. M., A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. *Biochim. Biophys. Acta*, 2009, **1794**, 860–871.

ACKNOWLEDGEMENTS. We thank the anonymous reviewers for their useful suggestions and comments that helped us to improve the manuscript. P.K.B. thanks Tezpur University, Napaam for the institute fellowship. P.S. thanks University Grants Commission, New Delhi, for JRF. M.D. thanks NIT, Arunachal Pradesh. S.S.S. and S.K.R. are grateful to DBT, GoI, for the Computational biology and bioinformatics centre at Tezpur University. No funding was received for this work.

Received 23 June 2022; re-revised accepted 16 July 2023

doi: 10.18520/cs/v125/i8/871-877