Chapter 5

A Comparative Synonymous Polymorphism Spectra Analysis in Co-Transcribed Gene Pairs

CHAPTER 5

A COMPARATIVE SYNONYMOUS POLYMORPHISM SPECTRA ANALYSIS IN CO-TRANSCRIBED GENE PAIRS

5.1. Abstract

Replication and transcription are two major events behind the different substitution patterns of mutations in genomic DNA. In this work, we have compared the adjacent co-transcribed gene pairs regarding synonymous polymorphism in five well known operons such as *rpoB/C, lacZ/Y, kdpA/B, araB/A* and *bcsA/B* in *Escherichia coli*. Interestingly the co-transcribed genes were found out to be significantly different from each other regarding their polymorphism spectra. The transition to transversion ratio between gene pairs were different due to their compositional differences of *two-fold degenerate* codons and *four-fold degenerate* codons. Further, the polymorphism spectra difference between the gene pairs was more prominent in *four-fold degenerate* and *six-fold degenerate* codons in comparison to the *two-fold degenerate* codons. In case of *rpoB* and *rpoC*, the major difference was observed in case of UCC, GUA, CCG, GCU, GGC and CGC codons. Similarly, in case of the other four pairs of co-transcribed genes, difference was more prominent in the higher degenerate codons than the *two-fold degenerate* codons. It is pertinent to note that, synonymous polymorphism in *two-fold degenerate* codons is restricted only to transitions whereas in the higher degenerate codons synonymous polymorphisms occur through transition as well as transversion. Hence, the role of different degenerate codons invokes many fundamental questions in molecular evolution.

5.2. Introduction

Base substitution mutation is a major event of molecular evolution in organisms, influenced by different factors such as DNA replication, damages in DNA bases such as deamination of cytosine/adenine and oxidation of guanine (Kino and Sugiyama, 2001; Rocha et al., 2006), gene expression, recombination etc. The asymmetry in DNA replication results in different mutation patterns between the leading strand (LeS) and the lagging strand (LaS) in genomes resulting in the former being enriched with keto nucleotides and the latter being enriched with amino nucleotides in bacteria (Lobry, 1996; Bulte et al., 1999). 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 (Daubin, 2003). The role of transcription in causing mutation asymmetry in genes resulting in higher C→T changes in the non-template strand than that in the template strand has been described recently (Francino, 1997). Therefore, expression level of genes has different impact on mutation rates (Lang et al., 2011; Hodgkinson and Eyre-walker, 2011; Park et al., 2012;). It is important to accurately compare polymorphism pattern between two genes because it helps to understand further the role of intrinsic factors in polymorphism difference 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 to each other, localized in the same strand and are similar regarding their gene expression at transcription level. Additionally, random drift will be considered minimal in bacterial genes where the population size is large unlike higher multicellular organisms. Therefore, polymorphism pattern 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 be different from each other regarding the amino acid level selection on their protein structure in a genome.

In this study, our endeavour is to perform a comparative analysis of polymorphism spectra in two adjacently placed co-transcribed genes in *Escherichia coli* (*E. 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 *E. coli* genome (Fig. 5.1). Surprisingly, the two genes in all five operons were found to be significantly different from each other regarding their synonymous polymorphism pattern. It revealed that variations arising in the genes are not identical to each other despite being similar regarding replication, transcription, and strand localization. This was corroborated with the observation that the phylogeny of the strains using the two co-transcribed genes were not identical. This study also revealed the role of codon degeneracy on synonymous polymorphism. So far we understand, this is the first such report of comparison of polymorphism spectra between two co-transcribed genes in bacteria.

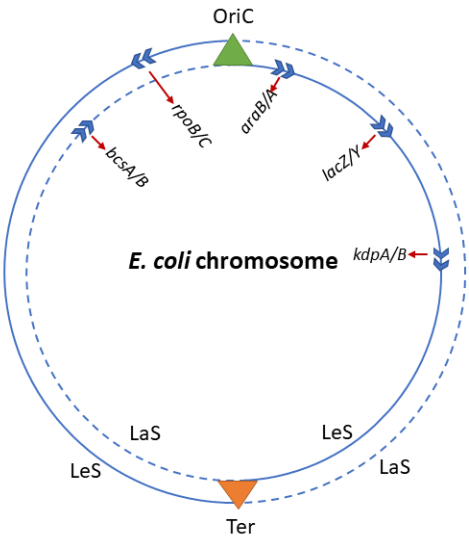


Fig. 5.1. Schematic diagram indicating the relative position of 5 pairs of genes in LeS and LaS of *Escherichia coli* chromosome. The schematic view of the five gene pairs presents in *E. coli*

genome. Leading strand (LeS) and lagging strand (LaS) are shown in continuous and dotted lines respectively. The gene locations are shown not to the scale.

5.3. Materials and Methods

5.3.1. Selection of co-transcribed gene pairs from available dataset

In this work, we considered 157 strains of *E. coli* bacterium for which alignments are available in public database (Thrope et al., 2017). The criteria for the selection of genes were set (>1200bp or >400 codons) anticipating a considerable number of synonymous substitutions in each gene. Out of five pairs of co-transcribed genes, four pairs were localized in the LeS and only one pair (*bcsA/B*) was present in the LaS. The list of genes and their detailed information are enclosed in Supplementary Table 1. We wish to study only the polymorphism due to base substitution polymorphism and have no other variation such as insertion or deletion. Hence the genes selected in this study were identical in terms of size across all the strains. For example, *rpoB* size is 4029 bp across all the strains; *lacZ* size is 3075 bp across all the strains.

5.3.2. Derivation of reference sequence and finding the overall synonymous spectra

We maintained a common set of strains for each pair of co-transcribed genes in the study. The compositional details along with different skew values and synonymous site values of the reference sequences for all the genes were calculated (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 a reference nucleotide for the position. This procedure was followed as per the methodology developed in our laboratory (Sen et al., 2022) (Supplementary Table 3). Only synonymous polymorphisms were considered in this study. The polymorphism values were normalized by dividing the values with the synonymous site values of respective nucleotides present in the gene. For each gene, the synonymous site values of each nucleotide were calculated by taking the summation of all possible synonymous substitutions involving the

nucleotide in each codon. In case of UUU, the synonymous site for U at the 3rd position is considered as 1 [one synonymous transition (*Sti*) and zero synonymous transversion (*Stv*)], but for GGU the synonymous site for U at the 3rd position is considered as 3 (one *Sti* and two *Stv*). Suppose we observed 30 C→T substitutions and the synonymous site values of C nucleotides in the gene is 150. Then the normalized C→T polymorphism frequency is calculated as 30/150=0.2. We estimated normalized values of all the twelve polymorphisms in the five pairs of genes.

5.3.3. Finding the ratio of transition to transversion in gene pairs

Transition (*ti*) to transversion (*tv*) ratio of all the genes were calculated. Overall, $\frac{ti}{tv}$ of the gene and $\frac{ti}{tv}$ at FFD site was calculated for each gene to show the variation between co-transcribed genes. For overall $\frac{Sti}{Stv}$, the values were collected from the initial non-normalized synonymous spectra table (Supplementary Table 4). For FFD site $\frac{ti}{tv}$, the *ti* and *tv* values were calculated by considering synonymous polymorphisms at FFD sites for individual genes.

5.3.4. Polymorphism frequency comparison at the codon level

From the reference sequence of each gene, we computed codon count using the web portal http://agnigarh.tezu.ernet.in/~ssankar/cbb_tu.html (Satapathy et al., 2017). All the genes were analysed for polymorphism at *four-fold degenerate* (FFD) sites (20 codons), *two-fold degenerate* (TFD) sites (18 codons), and *six-fold degenerate* codons (SFD). 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 less numbers of polymorphisms present in SFD SB, only FB codons of SFD (12 codons) were considered for the polymorphism comparison. The synonymous polymorphisms at the 3rd position of codons were calculated for each box separately. While comparing between the two genes regarding codon wise polymorphism, we only considered those codons with abundance values thirty or more in a gene, and/or polymorphism observed are five or more in the codon. Finally, FFD, TFD and SFD polymorphism frequency information of

a pair of gene was compared through the Python script. All the statistical methods used in the study are mentioned in the result section of the text.

5.4. Results

5.4.1. Synonymous polymorphism spectra difference between co-transcribed genes

We compared the synonymous polymorphism values between five pairs of co-transcribed genes such as *rpoB/C*, *lacZ/Y*, *kdpA/B*, *araB/A* and *bcsA/B* in well-known operons in *E. coli* genome (Table 5.1). As transition polymorphisms were more in number than the transversion polymorphisms, we compared the genes regarding the transition polymorphisms. The differences observed were given as follows. In 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 four-fold higher than G→A in *rpoB* whereas in *rpoC*, C→T frequency was less than three-fold higher than G→A. In 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; in *lacZ* and *lacY*, the A→G frequency values were 0.090 and 0.039, respectively. The above difference between the co-transcribed genes regarding certain transition polymorphisms indicated that the co-transcribed genes were not evolutionarily identical. The occurrence of transversion polymorphisms were not high for which we avoided to use those in the comparative analysis. There were some common observations in all the genes such as C→T frequency was the highest among the all the polymorphisms and frequency values between the complementary polymorphisms such as C→T and G→A were significantly different (p value <0.05; Mann-Whitney U test) in these genes.

Table 5.1. Synonymous polymorphism spectra of genes are presented in 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

5.4.2. $\frac{ti}{tv}$ ratio difference between the co-transcribed genes due to codon degeneracy composition difference

We studied synonymous transition (*Sti*) to synonymous transversion (*Stv*) ratio in all genes (Table 5.2). It was evident that the two co-transcribed genes were different from each other with regards to the ratios 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 and in *bcsB* the value was 3.3. We further found out $\frac{Sti}{Stv}$ at FFD sites in all these 10 genes. $\frac{Sti}{Stv}$ values at FFD were observed to be 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 at TFD sites are only possible due to transitions whereas the same at FFD sites are possible by both transition and transversion. We correlated the TFD:FFD ratio and the $\frac{ti}{tv}$ difference values between whole gene and at FFD. The positive correlation (*Pearson r* value 0.677) suggested that higher the composition of TFD in a gene greater will be the $\frac{ti}{tv}$ ratio difference between the whole gene and the FFD. Therefore, the compositional difference between TFD and FFD among the genes influences synonymous $\frac{ti}{tv}$ ratio in genes (Supplementary Table 5). We then compared the $\frac{ti}{tv}$ 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 case of certain amino acid codons: for Gly, $\frac{ti}{tv}$ ratio in *rpoB* and *rpoC* were 1.75 and 6.50 respectively; in case of Ala, $\frac{ti}{tv}$ ratio in *lacZ* and *lacY* were 1.27 and 4.00 respectively.

Table 5.2. $\frac{ti}{tv}$ ratio in whole gene and at FFD sites

$\frac{ti}{tv}$	*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

5.4.3. Synonymous polymorphism difference 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 (Sen et al., 2022). The synonymous site value in a FFD codon is 1.000 and that in a TFD codon is 0.667(Aziz et al., 2022). Therefore, synonymous polymorphism at a FFD site and [SFD (FB)] is expected to be 1.5 times more frequent than that at a TFD site. To compare synonymous polymorphism at TFD, FFD and SFD (FB) between the co-transcribed gene pairs, we performed a codon-wise analyses in each codon (Supplementary Table 7). In *rpoB* and *rpoC*, synonymous polymorphism in FFD codons such as GUA, CCG, GCU, GGC were found to be different. The codon count of GUA was 31 and 32 in *rpoB* and *rpoC* respectively but the synonymous polymorphisms frequencies were 0.097 and 0.219 respectively. Similarly, the codon count of GCU was 19 and 28 in *rpoB* and *rpoC* respectively but the frequencies were 0.053 and 0.250 respectively. Analogously, the polymorphism frequency values in CCG were calculated as 0.079 and 0.178 in *rpoB* and *rpoC* respectively. The polymorphism frequency values in GGC were 0.114 and 0.310 in *rpoB* and *rpoC* respectively. This indicated that *rpoB* and *rpoC* genes are different from each other with regards to synonymous polymorphism of certain codons.

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

Interestingly, in *rpoB* and *rpoC*, out of the twelve codons belonging to the SFD (FB) only three codons were exhibiting the difference. Similarly, out of the twenty codons belonging to the FFD codons, only four codons were exhibiting the difference. Whereas, out of eighteen TFD codons only one codon was exhibiting the difference in both the co-transcribed genes. We extrapolated this observation in the remaining gene pairs (Supplementary Table 9). We then plotted box-plots for each pair of genes with regards to their obtained polymorphism frequencies in a degeneracy-wise manner (Fig. 5.2). The box-plots suggested the degeneracy -wise differences between co-transcribed gene pairs, out of which differences at FFD codons were principally found between a pair of genes. Further we have done a Mann-Whitney U test between FFD and TFD, and between FFD and SFD and 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 ~five-fold higher polymorphism frequency values in GGU regardless of the equal codon frequency in GAU and GGU. This was a noticeable evidence of higher polymorphism frequency of FFD codons which was observed in many cases though.

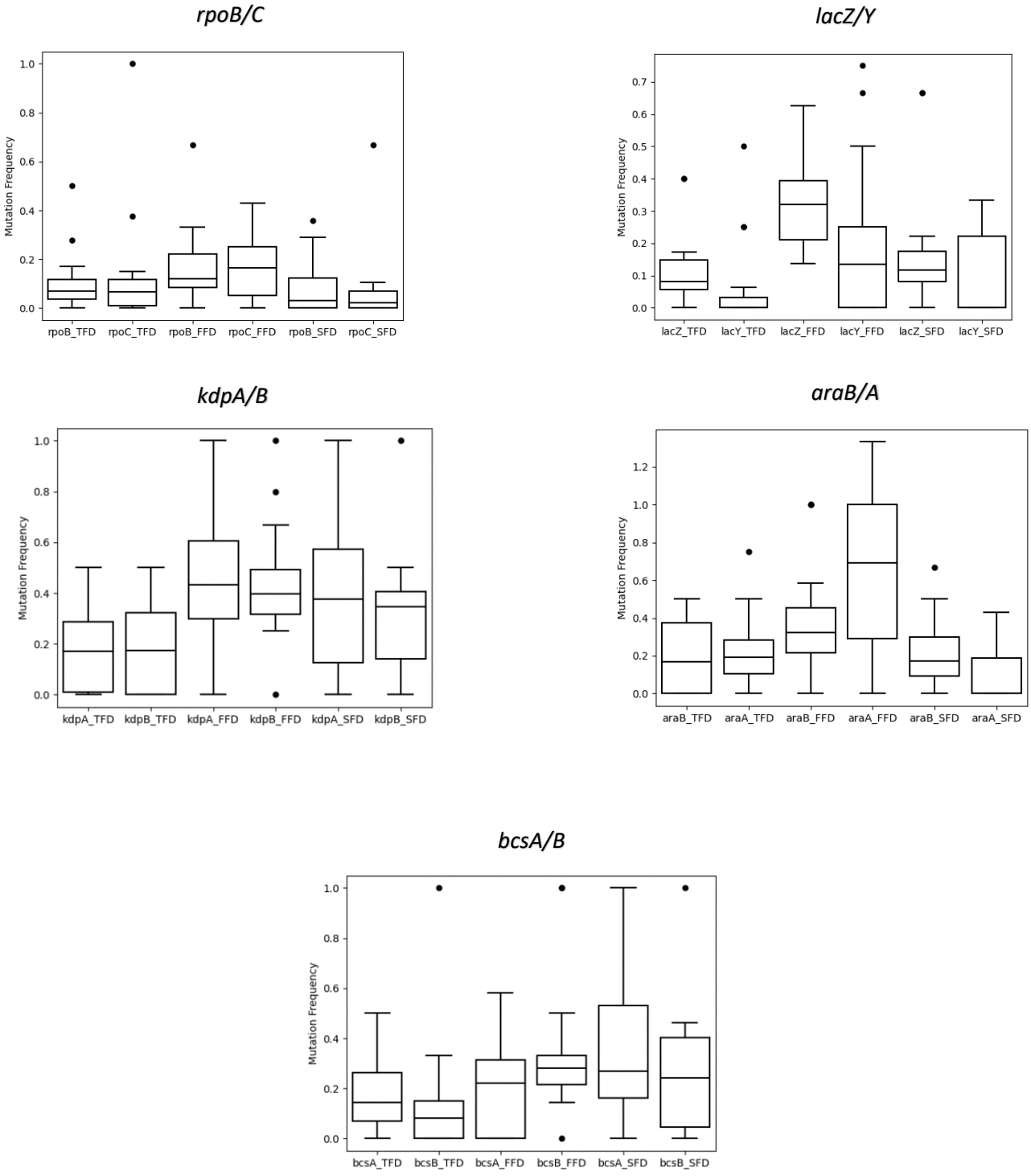
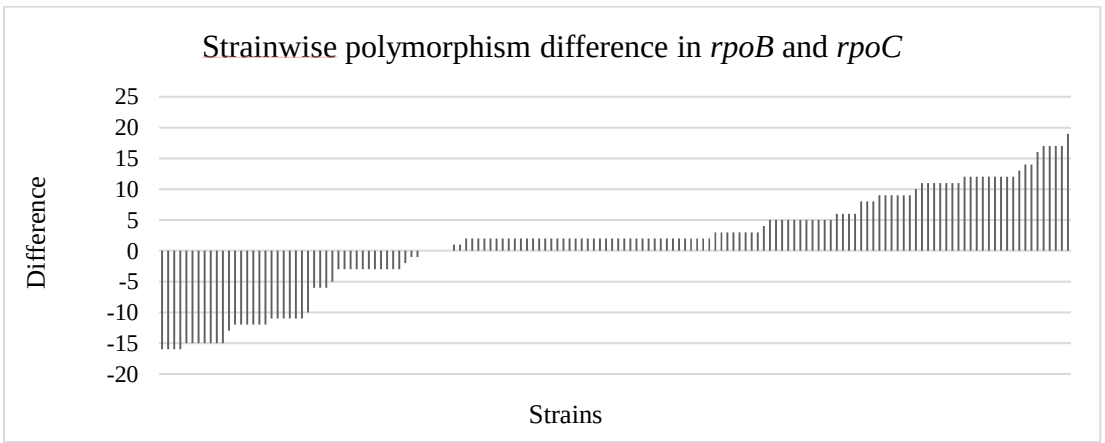


Fig. 5.2. Box plot showing polymorphism frequency of gene pairs in an FFD/TFD/SFD manner. Legend. The box plots present the polymorphism frequency distribution in five pairs of genes. Each graph shows the polymorphism frequency comparison of adjacent genes in FFD, TFD and SFD. Each graph represents polymorphism frequency in *y*-axis and different degenerate codon box names in *x*-axis.

5.4.4. Phylogeny comparison between co-transcribed gene pairs

Additionally, to get an insight regarding the polymorphism difference between two co-transcribed gene pairs, we quantified their polymorphisms difference between individual strains in *rpoB/C* and *kdpA/B* gene pair. We observed total polymorphisms in individual strains with regards to co-transcribed pairs. The co-transcribed pairs had different number of polymorphisms in individual strains. For example, in a strain with zero polymorphisms in *rpoB* was with seventeen number of polymorphisms in *rpoC*. In Figure 5.3, the minimum to maximum polymorphism difference between *rpoB* and *rpoC* were -16 to 19 (*rpoB*-*rpoC*). Whereas, in case of *kdpA* and *kdpB* (*kdpA*-*kdpB*) the range was -23 to 30. It was evident that the co-transcribed genes were different from each other even at the individual strain level. We then generated phylogenetic trees using 10 common strains in co-transcribed pairs of *rpoB/C* and *kdpA/B*. The comparative study of phylogeny indicated different patterns in *rpoB*/*C* and also in *kdpA*/*B* gene pairs (Fig. 5.4). This supported to the above observation that the co-transcribed genes are not identical with regards to polymorphism pattern in *E. coli*.

(a)



(b)

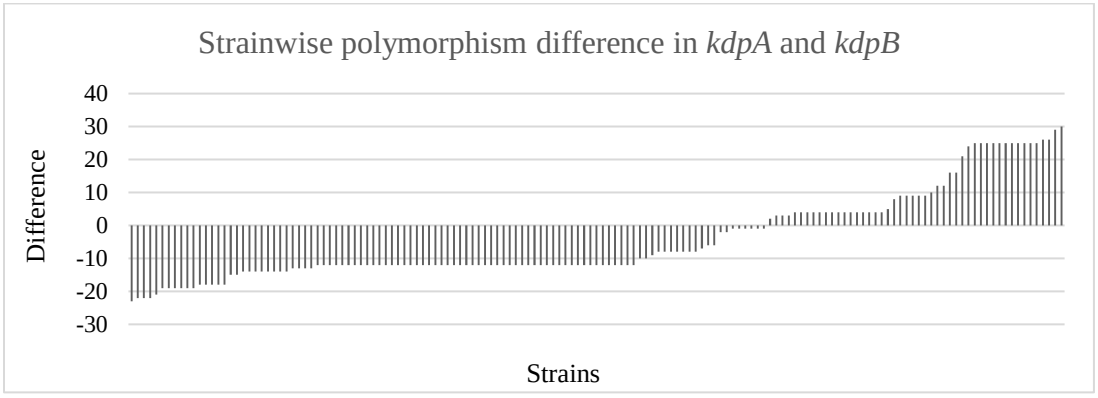
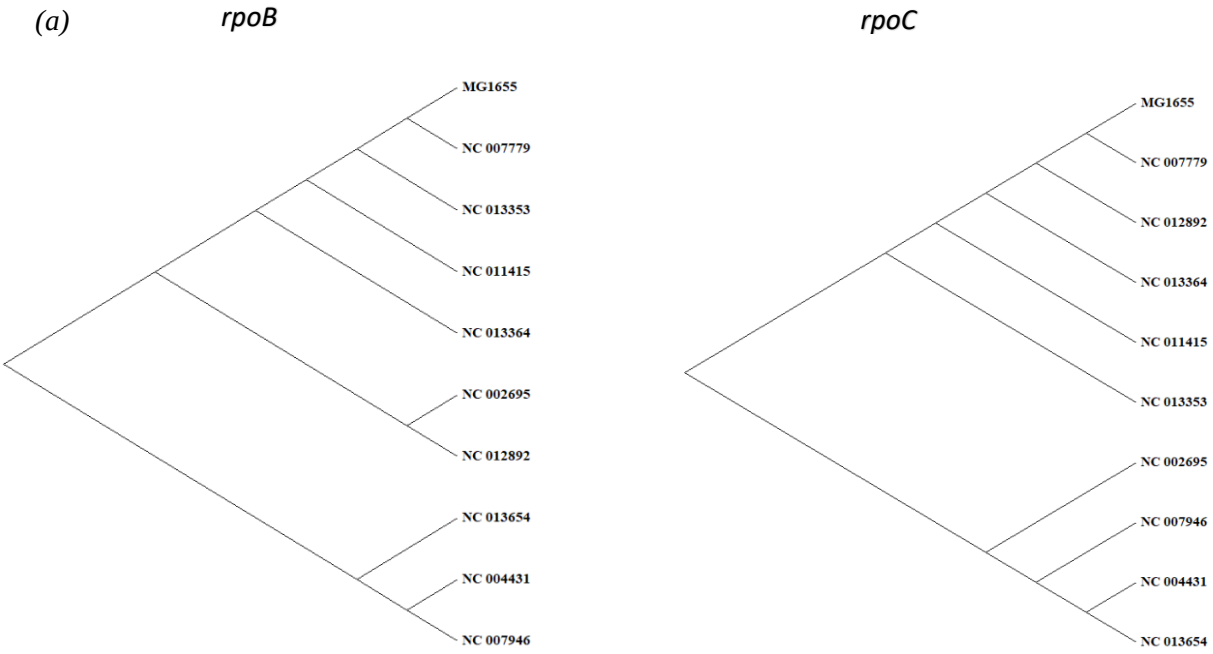


Fig. 5.3. Individual strain wise polymorphism difference in co-transcribed genes. The strainwise polymorphism difference in co-transcribed gene pairs (a) The individual strainwise polymorphism is shown in rpoB and rpoC. It indicates a greater number of individual substitutions are present in rpoB. (b) The individual strainwise polymorphism is shown in kdpA and kdpB. It indicates a greater number of individual substitutions are present in kdpB.

(a)



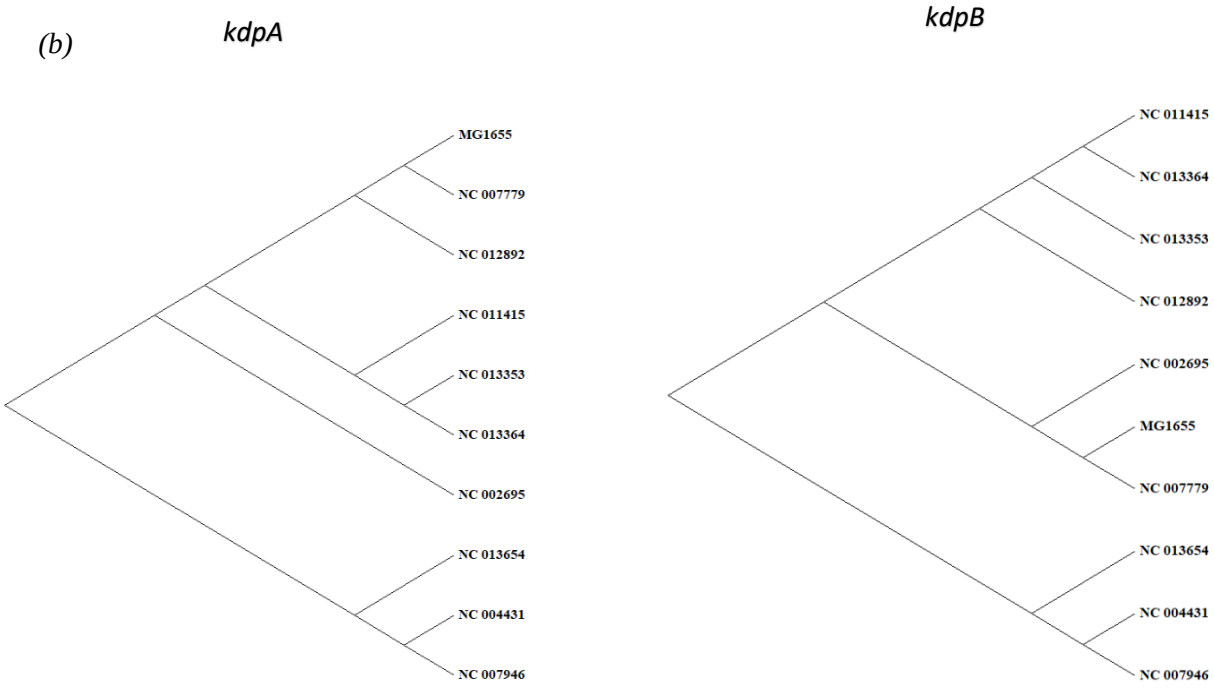


Fig. 5.4. Phylogenetic comparative study between co-transcribed genes in *E. coli*. This illustration shows the comparative phylogenetic study between co-transcribed gene pairs using 10 common strains. (*a*) *rpoB* and *rpoC* phylogenetic trees. (*b) kdpA* and *kdpB* phylogenetic trees.

5.5. Discussion

In the present study co-transcribed gene pairs have been compared regarding synonymous polymorphism in five operons in *E. coli*. The adjacent co-transcribed genes were found out to be different regarding synonymous polymorphism, though these genes in an operon are similar regarding replication, strand location and transcript level expression. The compositional difference of degenerate codons between the gene pairs were found out to be important for their different $\frac{ti}{tv}$ ratios. The TFD codons can undergo synonymous polymorphism only by transition substitution while the higher degenerate codons can undergo synonymous polymorphism by both transition as well as transversion substitutions. The rate of a transition substitution is four times more than that of a transversion substitution (Duchêne et al., 2015) that results in higher

transversion proportion in FFD than TFD codons. Therefore, $\frac{ti}{tv}$ is higher in genes having lower proportion of FFD codons. The current study has manifested the role of codon degeneracy due to the difference in polymorphism spectra of two co-transcribed genes. Yet, among all, FFD codons remained to be the common cause behind the polymorphism spectra difference in each gene pair. In case of the higher degeneracy codons synonymous transversion is possible unlike the TFD codons. It will be interesting to investigate whether this is the only reason for the difference or are there additional reasons involved. 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 on 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 favor towards the neutral theory of evolution in this intra-species genome comparison analysis.

It has already been reported by the researchers that although synonymous polymorphisms do not affect the amino acid sequence in a protein, they influence its function by protein folding (Fung and Gottesman, 2009). Whether TFD codons and FFD codons attribute differently to protein folding is yet to be discovered. The difference between the degenerate codons invokes many fundamental questions in genetic code evolution. 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 secrets about it.

5.6. Bibliography

- 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.
- Bulte, J. W. M., Zhang, S. C., Van Gelderen, P., Herynek, V., Jordan, E. K., Duncan, I. D., & Frank, J. A. (1999). Neurotransplantation of magnetically labeled oligodendrocyte progenitors: magnetic resonance tracking of cell migration and myelination. *Proceedings of the National Academy of Sciences*, *96*(26), 15256-15261.
- Daubin, V., & Perriere, G. (2003). G+ C3 structuring along the genome: a common feature in prokaryotes. *Molecular biology and evolution*, *20*(4), 471-483.
- 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.
- Francino, M. P., & Ochman, H. (1997). Strand asymmetries in DNA evolution. *Trends in Genetics*, *13*(6), 240-245.
- Fung, K. L., & Gottesman, M. M. (2009). A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1794*(5), 860-871.
- Hodgkinson, A., & Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nature reviews genetics*, *12*(11), 756-766.
- Kino, K., & Sugiyama, H. (2001). Possible cause of G· C→ C· G transversion mutation by guanine oxidation product, imidazolone. *Chemistry & biology*, *8*(4), 369-378.

- Lang, G. I., & Murray, A. W. (2011). Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome biology and evolution*, *3*, 799-811.
- Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular biology and evolution*, *13*(5), 660-665.
- Park, C., Qian, W., & Zhang, J. (2012). Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO reports*, *13*(12), 1123-1129.
- Rocha, E. P., Touchon, M., & Feil, E. J. (2006). Similar compositional biases are caused by very different mutational effects. *Genome research*, *16*(12), 1537-1547.
- Satapathy, S. S., Sahoo, A. K., Ray, S. K., & Ghosh, T. C. (2017). Codon degeneracy and amino acid abundance influence the measures of codon usage bias: improved Nc ($\hat{N}c$) and ENCprime ($\hat{N}'c$) measures. *Genes to Cells*, *22*(3), 277-283.
- 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.