Chapter 4

A Comparative Polymorphism Spectra Analysis in Inter-Operon IGRs And Intra-Operon IGRs

CHAPTER 4

A COMPARATIVE POLYMORPHISM SPECTRA ANALYSIS IN INTER-OPERON IGRS AND INTRA-OPERON IGRS

4.1. Abstract

The temporary exposure of single-stranded regions in the genome during the process of replication and transcription makes the region vulnerable to cytosine deamination resulting in higher rate of C→T transition. Intra-operon intergenic regions undergo transcription along with adjacent co-transcribed genes in an operon, whereas inter-operon intergenic regions are usually devoid of transcription. Hence these two types of intergenic regions (IGRs) can be compared to find out the contribution of replication-associated mutations (RAM) and transcription-associated mutations (TrAM) towards bringing variation in genomes. In our work, we performed a polymorphism spectra comparison between intra-operon IGRs and inter-operon IGRs in genomes of two well-known closely related bacteria such as *Escherichia coli* and *Salmonella enterica*. In general, the size of intra-operon IGRs was smaller than that of inter-operon IGRs in *E. coli* and *S. enterica*. Interestingly, the polymorphism frequency at intra-operon IGRs was 2.5-fold lesser than that in the inter-operon IGRs in *E. coli* genome. Similarly, the polymorphism frequency at intra-operon IGRs was 2.8-fold lesser than that in the inter-operon IGRs in *S. enterica* genome. Therefore, the intra-operon IGRs were often observed to be more conserved. In the case of inter-operon IGRs, the T→C transition frequency was a minimum of two times more frequent than T→A transversion frequency whereas in the case of intra-operon IGRs, T→C transition frequency was similar to that of T→A transversion frequency. The polymorphism was purine-biased and keto-biased more in intra-operon IGRs than the inter-operon IGRs. In *E. coli*, the transition/transversion ratio was observed as 1.639 and 1.338 in inter-operon and in intra-operon IGRs, respectively. In *S. enterica*, the transition/transversion ratio was observed as 2.134 and 2.780 in inter-operon and in intra-operon IGRs, respectively. The observation in this study

indicates that transcribable IGRs might not always have higher polymorphism frequency than non-transcribable IGRs. The lower polymorphism frequency at intra-operon IGRs might be attributed to different events such as the transcription-coupled DNA repair, sequences facilitating translation initiation and avoidance of Rho-dependent transcription termination.

4.2. Introduction

Many of the functionally related bacterial genes are transcribed as a polycistronic unit. As most of the genes are present under operonic units in prokaryotes, two types of untranslated intergenic regions (IGRs) could be found in their genomes; intra-operon IGRs and inter-operon IGRs. Inter-operon IGRs are found in between two separate operonic/cistronic units. Intra-operon IGRs are found between two adjacent open reading frames in an operon (Fig. 4.1), which are popularly known as Intercistronic regions (ICRs). Intra-operon IGRs are the units that undergo replication as well as transcription, whereas inter-operon IGRs are devoid of transcription. However, inter-operon IGRs may not be 100% devoid of transcription because a certain percentage of leaky pervasive transcription in the downstream operon is possible under fluctuating levels of Rho. However, the frequency of such transcription event is very low for which the observed polymorphisms can be considered under replication-associated mutations (RAMs) in such cases. The size of inter-operon IGRs is usually larger than intra-operon IGRs. The small-sized intra-operon IGRs are evolutionarily preferred only to prevent energy wastage in transcribing longer intra-operon IGRs. Sometimes co-transcribed genes can also be found to be overlapped by a few sequences, but they lack intra-operon IGRs. The bacterial chromosome has only 10-15% of regions known as untranslated IGRs and contains many regulatory elements with key functions (Sridhar et al., 2011; Thorpe et al., 2017). In past years, intra-operon IGRs were found to be playing an important role in the formation of hairpin loops as well as Ribosome binding sites (RBS) in bacteriophages (Romantschuk and Müller, 1983). Intra-operon IGRs were also found to be important in assigning RBS and recruitment of EF-G gene (Post and Nomura, 1980).

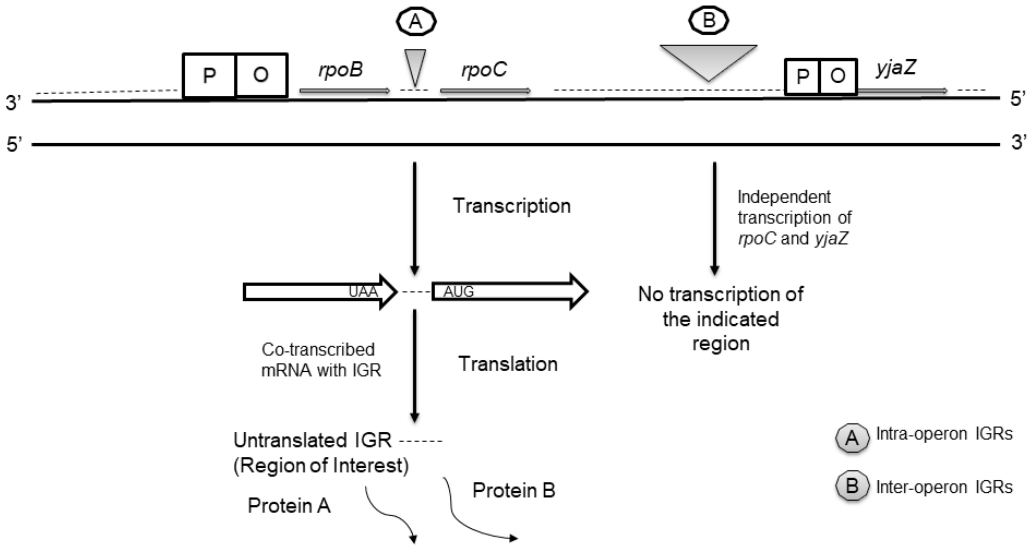

Fig. 4.1. A schematic diagram elucidating the difference between intra-operon IGRs and inter-operon IGRs. The schematic diagram presents two different transcriptions. Scenario (1): transcription of β-operon consisting of P-promoter, O- operator and structural region (*rpoB* and *rpoC*). The region (A) between *rpoB* and *rpoC* is known as intra-operon IGR. This region is co-transcribed along with adjacent genes, but it remains untranslated. Scenario (2): transcription of downstream gene *yjaZ*. The region (B) downstream of (β operon) *rpoC* and upstream of *yjaZ* is known as inter-operon IGR. This region is neither transcribed nor translated.

Among different types of polymorphism, base substitution occurs predominantly in a bacterial genome apart from insertion/deletion (INDELs). Base substitution is also studied by scientists largely by using *Escherichia coli* as a model organism in recent decades. The temporary exposure of single-stranded regions in the genome during events like replication and transcription make the region vulnerable to base substitutions. Scientists have discovered that cytosine deamination is a major reason behind the base substitution in genomic DNA alongside Guanine oxidation leading to G→T substitution. (Kino and Sugiyama, 2001; Rocha et al., 2006; Bhagwat et al., 2016). Similarly, the single-stranded exposure of the non-template strand during transcription makes it prone to C→T base substitutions in the non-template strand has been described recently (Francino and Ochman, 1997; Mugal et al., 2009). Hence C→T/G→A is a major contributor to

the polymorphism spectra in CDS as well as non-CDS regions in chromosomes. The base substitutions are studied as transition and transversion. As nucleotides are divided into two classes; purine (R) and pyrimidine (Y), The intra-class substitution between nucleotides is known as transition (*ti*) (R→R, Y→Y). The inter-class substitution between nucleotides is known as transversion (*tv*) (R→Y, Y→R) (Fig. 4.2). Like coding sequences (CDS), substitutions in IGRs can also affect the expression and regulation of the CDS in pathogenic bacteria (Casali et al., 2014; Laabei et al., 2014).

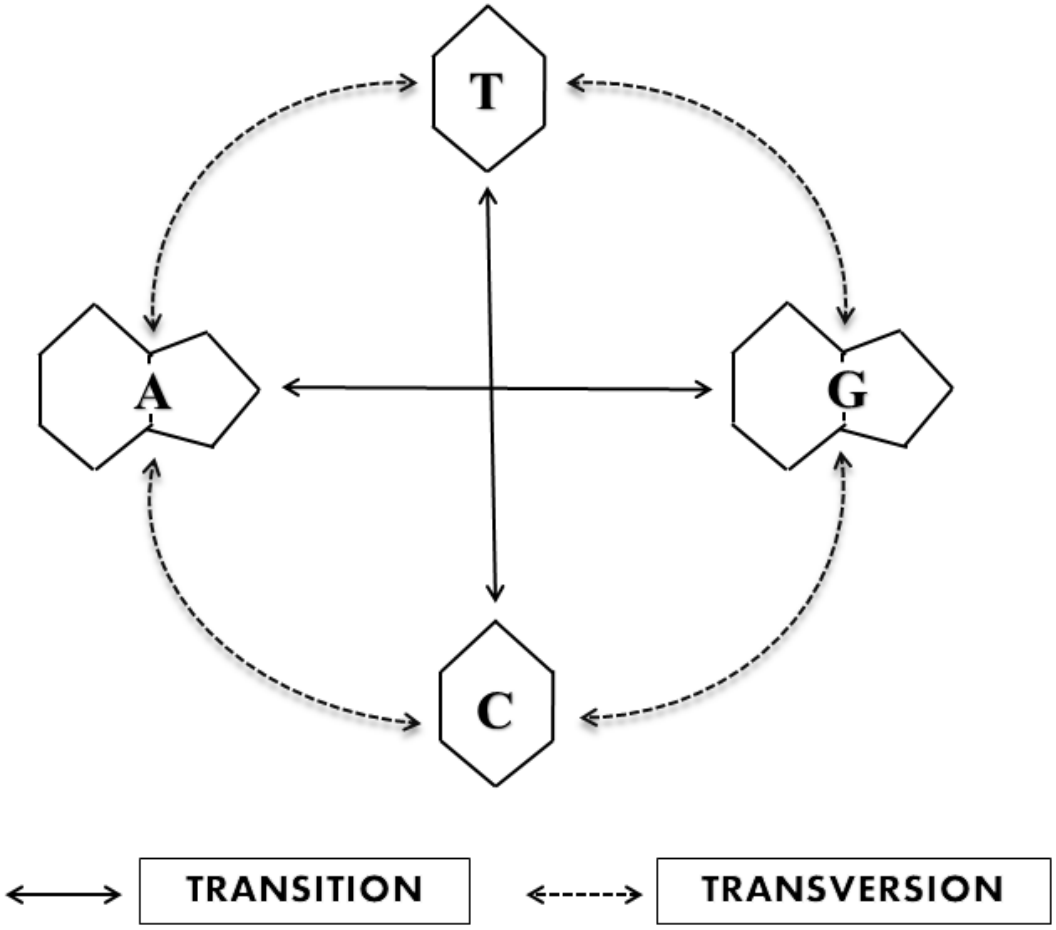

Fig. 4.2. A schematic diagram elucidating transition and transversion in DNA. Purines (A & G) are shown as double-ringed nitrogenous bases and Pyrimidines (T & C) are shown as single-ringed nitrogenous bases. The base substitution between two similar classes of nucleotides is shown as a transition (solid lines with arrow). The base substitution between two different classes of nucleotides is shown as transversion (dotted line with arrow).

As mentioned earlier, intra-operon IGRs undergo transcription making them fundamentally different from inter-operon IGRs. We studied a comparative single nucleotide polymorphisms (SNPs) analysis using computational tools to find out the different effects of transcription and replication on polymorphism in both these regions. As the frequency of transcription is higher compared to a standard replication cycle of any bacteria, the regions containing intra-operon IGRs are likely to be exposed as a single strand more frequently than inter-operon IGRs. Hence, we presumed to get a higher polymorphism frequency value in intra-operon IGRs, and to study the possible effect of transcription-induced polymorphism and/or selection in the IGRs of a bacterial genome. In this study, we considered 157 strains of *E. coli* and 366 strains of *S. enterica* and collected datasets for 134 and 89 intra-operons IGRs respectively for *E. coli* and *S. enterica*, then a comparison with inter-operon IGRs was done for each species.

4.3. Materials and methods

4.3.1. Annotating inter-operon IGRs

In this study, we considered 157 strains of *Escherichia coli* (*E. coli*) and 366 strains of *Salmonella enterica* (*S. enterica*) (Thorpe et al., 2017). We collected datasets for 1120 inter-operon IGRs for *E. coli* and 1150 inter-operon IGRs for *S. enterica* (Sen et al., 2022). While selecting Inter-operon IGRs we have excluded probable promoter/terminator sites (35bp upstream of the start codon and 35bp downstream of the stop codon). The probability of presence of the promoters/terminators towards more upstream in some regions of the chromosome could not be denied. This condition becomes more gene-specific. Hence to find out inter-operon IGRs for the entire chromosome we excluded 35bp as a standard promoter/terminator flanking sequences for all the operonic units.

4.3.2. Annotating intra-operon IGRs

We considered all the non-zero-sized intra-operon IGRs for the study. For example, in the *lac* operon, we found a size of 52 nucleotides of IGRs in between *lacZ* and *lacY* genes. Similarly, we found a size of 64 nucleotides in between *lacY* and *lacA*. We collected 134 intra-operon IGRs for *E. coli* and 89 intra-operon IGRs in *S. enterica* from the available data set (Tables 1 and 2 in supplementary material). A size comparison boxplot was drawn between intra and inter-operon IGRs by using OriginPro 2022. Origin Lab Corporation, Northampton, MA, USA. Mann-Whitney test (Mann and Whitney, 1947) was performed between intra-operon IGRs and inter-operon IGRs by using https://www.socscistatistics.com/tests/mannwhitney/default2.aspx website.

4.3.3. Finding base substitutions in a sequence alignment

A hypothetical example of the detailed procedure is explained (Table 3 in supplementary material). Strains having ambiguous nucleotides (N) were not considered for the study. The reference sequence was derived for each intra-operon IGRs by considering the most frequent nucleotide present at a certain position in the alignment as a reference nucleotide for that position. The procedure followed a methodology that has been already established in our laboratory to derive the reference sequence of coding sequences (Sen et al., 2022). A Python script was written to find out the spectra in both the IGRs for *E. coli* and *S. enterica*. After finding out the individual intra-operon IGRs having polymorphism in each bacterial species, the total spectra were calculated by considering only those intra-operon IGRs possessing polymorphism (Table 4 in supplementary material).

4.3.4. Comparison of polymorphism spectra at the intra-operon IGRs and inter-operon IGRs

We performed the comparative study between the intra-operon and inter-operon IGRs of *E. coli* and *S. enterica*. We also compared the $\frac{ti}{tv}$ ratio in overall spectra and the individual nucleotide-based $\frac{ti}{tv}$ between intra-operon IGRs and inter-operon IGRs. The overall K→M/M→K, R→Y/Y→R, and A/T→G/C or G/C→A/T bias was also observed between both the intra-operon IGRs and inter-operon IGRs. This procedure was also followed in the *S. enterica* dataset.

4.4. Results

4.4.1. SNPs frequency is lower at the intra-operon IGRs than the inter-operon IGRs

It was distinctly observed that the intra-operon IGRs were usually smaller than inter-operon IGRs in both species (Fig. 1 in supplementary material). The total number of polymorphisms in intra-operon IGRs and inter-operon IGRs of *E. coli* and *S. enterica* is reviewed (Table 5 in supplementary material). In *E. coli*, the inter-operon IGRs frequency (total mutations/size) was noted as 0.115, and intra-operon IGRs frequency was noted as 0.045, indicating a more than 2-fold difference between the two IGRs, whereas in *S. enterica* the inter-operon IGRs frequency was noted as 0.193 and intra-operon IGRs frequency was noted as 0.067, indicating more than a 2.5-fold difference between the two IGRs. It suggests that the intra-operon IGRs exhibit lesser nucleotide polymorphism than inter-operon IGRs contrary to the assumptions (Table 4.1). Along the line of the observation, out of 134 intra-operon IGRs, 77 IGRs were found to be conserved across the strains in *E. coli*. Similarly in *S. enterica*, out of 89 intra-operon IGRs, 32 IGRs were found to be conserved across the strains.

Table 4.1. Normalized mutational spectra, $\frac{ti}{tv}$, and mutation frequency of *Escherichia coli (Ec) and Salmonella enterica (Se)* showing numbers and normalized values of base substitutions at inter-operon IGRs and intra-operon IGRs

Polymorphism	Inter operon (*Ec*)	Intra operon (*Ec*)	Inter operon (*Se*)	Intra operon (*Se*)
A→T	0.024	0.014	0.025	0.003
A→C	0.016	0.005	0.024	0.000
A→G	0.049	0.019	0.082	0.028
T→A	0.024	0.020	0.024	0.020
T→C	0.050	0.021	0.082	0.020
T→G	0.016	0.009	0.023	0.006
C→A	0.036	0.017	0.062	0.014
C→T	0.104	0.049	0.202	0.115
C→G	0.014	0.002	0.020	0.002
G→A	0.102	0.022	0.200	0.046
G→T	0.035	0.009	0.061	0.026
G→C	0.013	0.002	0.019	0.003
ti	17425	87	34861	114
tv	10631	65	16335	41
ti/tv	1.639	1.338	2.134	2.780
Total (ti+tv)	28056	152	51196	155
Size	244983	3371	264788	2300
Frequency (Total/size)	0.115	0.045	0.193	0.067

4.4.2. Comparative polymorphism spectra analysis of inter-operon IGRs and intra-operon IGRs

In inter-operon IGRs, transition polymorphisms such as C→T and G→A were two times more frequent than T→C and A →G. In the case of Tv polymorphisms, G→T and C→A were more frequent than the others. The complementary polymorphisms were of similar value. In this study we found out intra-operon IGRs complementary substitutions were of different value, unlike the inter-operon IGRs. The transition substitution of C→T was 0.049 while the other transition polymorphisms such as G→A, A→G, T→C were 0.022. 0.019 and 0.021, respectively. The highest Tv was obtained in T→A with a value of 0.020, almost like three other *ti* spectra A→G,

T→C, and G→A. Among other Tv frequency values, C→A was observed to be the second highest Tv with a value of 0.017 followed by A→T of 0. 014. The complementary substitutions between A→T (0.014) and T→A (0.020); a 1.4-fold value difference was observed, likewise between A→C (0.005) and T→G (0.009) also a fold difference of 1.8 was observed. The fold difference between C→A (0.017) and G→T (0.009) a 2-fold difference was observed. The inter-operon IGRs and intra-operon IGRs polymorphism spectra were found to be significantly different at $p<0.05$ (Fig. 4.3). Similarly, in *S. enterica* in brief, transition polymorphisms such as C→T and G→A were more than two times more frequent than T→C and A →G. In the case of Tv polymorphisms, G→T and C→A were more frequent than the others. The complementary polymorphisms were of similar value. In this study we found out intra-operon IGRs complementary polymorphisms were of different values, unlike the inter-operon IGRs. The transition polymorphism C→T value was 0.115 while the other transition polymorphisms such as G→A, A→G, T→C were 0.046. 0.028 and 0.020, respectively. The highest *tv* was obtained in T→A with a value of 0.020, which was almost like three other *ti* spectra A→G, T→C, and G→A. Among other Tv frequency values, G→T was observed to be the second highest *tv* with a value of 0.026 followed by T→A of 0. 020. The complementary substitutions between A→T (0.003) and T→A (0.020), a six-fold value difference was observed, likewise between G→T (0.026) and C→A (0.014) also a fold difference of 1.8 was observed.

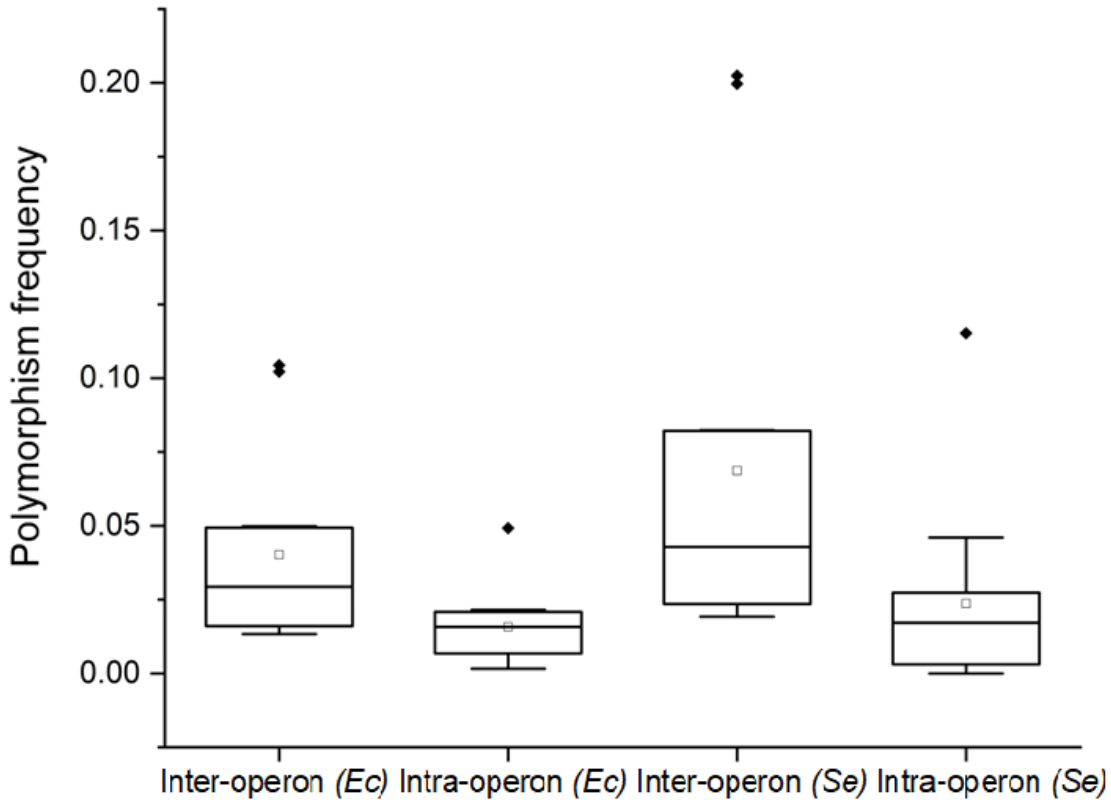

Fig. 4.3. The inter-operon IGRs polymorphism values were significantly higher than intra-operon IGRs in *E. coli* and *S. enterica.* The box-plot shows polymorphism frequency distributions in *y*-axis. Intra-operon and inter-operon IGRs in *E. coli* (*Ec*) and *S. enterica* (*Se*) are shown in *x*-axis. The inter-operon IGRs polymorphism values are significantly higher (p<0.05) than intra-operon IGRs in *E. coli* and *S. enterica*. Mann-Whitney U test is used for statistical significance test.

4.4.3. $\frac{ti}{tv}$ values at inter-operon and intra-operon IGRs in *E. coli* and S. *enterica*

In *E. coli*, the ratio of $\frac{ti}{tv}$ was observed as 1.639 at inter-operon IGRs and 1.338 at intra-operon IGRs (Table 4.1). We then found out $\frac{ti}{tv}$ at individual nucleotides for inter-operon IGRs. For Adenine, A→G *ti* was compared with combined A→T and A→C *tv*. A→G value was 0.049 and the combined A→T and A→C values were 0.040; hence the ratio of $\frac{ti}{tv}$ for A was 1.22 (Table 4.2). The $\frac{ti}{tv}$ ratios for T, C, and G were obtained as 1.25, 2.93, and 2.13 respectively. The $\frac{ti}{tv}$ in the case of G and C were distinctly higher than that for A and T. So, the transition was more

frequent than the transversion in the genome even at individual nucleotide level. In intra-operon IGRs the individual nucleotide-based $\frac{ti}{tv}$ ratio at A, T, C, and G were obtained as 1.00, 0.74, 2.58 and 1.9 respectively (Table 4.2). The ratio at T (0.074) indicated the existence of higher Tv values in T, as explained above T→A Tv value exhibited unexpectedly equal values with T→C. Hence the polymorphism pattern relating to T nucleotide was observed to be different.

Table 4.2. Individual nucleotide-based *ti* to *tv* ratio at inter- and intra-operon IGRs in *E. coli* and *S. enterica* along with their GC%

Species	IGRs	Individual nucleotide-based *ti*/*tv*				GC%
		A	T	C	G	
E. coli	Inter operon	1.22	1.25	2.93	2.13	40.3
	Intra operon	1.00	0.74	2.58	1.90	44.8
S. enterica	Inter operon	1.69	1.74	2.47	2.50	41.6
	Intra operon	10.00	0.77	7.00	1.56	47.6

In *S. enterica*, the ratio of $\frac{ti}{tv}$ was observed as 2.134 at inter-operon IGRs and 2.780 at intra-operon IGRs (Table 4.1). The $\frac{ti}{tv}$ at individual nucleotides were calculated like *E. coli*. A→G value was 0.082 and the combined A→T and A→C values was 0.049; hence the ratio of $\frac{ti}{tv}$ for A was 1.69 (Table 4.2). The $\frac{ti}{tv}$ ratio for T, C, and G were obtained as 1.74, 2.47, and 2.50 respectively. The $\frac{ti}{tv}$ in the case of G and C were distinctly higher than that for A and T similar to *E. coli*. In intra-operon IGRs the individual nucleotide-based $\frac{ti}{tv}$ ratio at A, T, C, and G were obtained as 10.00, 0.77, 7.00 and 1.56 respectively (Table 4.2). The ratio at T (0.077) indicated the existence of higher Tv values in T, as explained above T→A Tv value exhibited unexpectedly equal values with T→C. This similar observation was also seen in the case of *E. coli*. The comparative individual nucleotide-based substitutions between inter and intra-operon IGRs in *S. enterica* resembled the results in *E. coli* (Table 4.2). The $\frac{ti}{tv}$ ratio at T in intra-operon IGRs was

found to be 0.77 which indicated the unusual T→A spectra value has similar values of T→C.

4.4.4. Intra-operon polymorphism is biased towards purine and keto nucleotides

The comparative study of different biases (RY, KM, AT/GC) revealed that there was difference between inter-operon IGRs and intra-operon IGRs in *E. coli* (Table 4.3). In inter-operon IGRs, the R→Y(Pu→Py) substitution was observed as 0.088, whereas Y→R for the same was observed as 0.089. But in the case of intra-operon IGRs, the R→Y was found to be 0.030 and Y→R was found to be 0.048, a 1.6-fold difference was observed in *E. coli*, which was showing a higher proportion of Pu biased polymorphism in intra-operon IGRs. Similarly,while analyzing K→M and M→K (Keto/Amino) biased polymorphism, we had similar observations in the case of inter-operon IGRs for both substitutions. But in the case of intra-operon IGRs, K→M was observed to be 0.065 and M→K was observed to be 0.084 indicating 1.2-fold higher Keto biased polymorphism in intra-operon IGRs. When we calculated the difference between AT and GC biased polymorphism in both IGRs, intra-operon IGRs were found as 0.044 and inter-operon IGRs as 0.146, which recommended a higher AT biased mutation in inter-operon IGRs. Similar observations were also found in *S. enterica* regarding the nucleotide polymorphism spectra at inter-operon IGRs (Table 4.1). Analogous to *E. coli*, the keto and Purine biases were also observed in intra-operon IGRs, as well as higher AT biases were also observed in inter-operon IGRs (Table 4.3). Hence the frequency comparison, individual nucleotide-based $\frac{ti}{tv}$ ratio, skew biased comparison and overall spectra suggested that intra-operon IGRs have a different polymorphism than inter-operon IGRs which possibly explains the role of mutation and/or selection during transcriptional mutagenesis on intra-operon IGRs.

Table 4.3. RY, KM, AT/GC biases present between inter-operon IGRs and intra-operon IGRs in *E. coli* and *S. enterica*

Species and IGRs	R→Y	Y→R	K→M	M→K	A/T→G/C	G/C→A/T	Difference (A/T)-(G/C)
E. coli Inter operon	0.088	0.089	0.189	0.191	0.131	0.277	0.146
E. coli Intra operon	0.030	0.048	0.065	0.084	0.054	0.097	0.044
S. enterica Inter operon	0.128	0.129	0.325	0.330	0.211	0.525	0.313
S*. enterica* Intra operon	0.032	0.042	0.089	0.149	0.054	0.202	0.148

4.5. Discussion

Since intra-operon IGRs undergo both replication and transcription, the probability of getting elevated numbers of polymorphism was expected due to the cumulative effect of both replication and transcription compared to the inter-operon IGRs. But our observation suggests that it might not always be true considering the influence of selection and transcription-coupled DNA repair. Despite replication and transcription, the intra-operon IGRs were observed to be more conserved across the genomes of different strains in *E. coli* and *S. enterica*. However, the correlation between the direction of operon transcription in intra operon-IGRs, with respect to that of replication and obtained polymorphisms will be an interesting study in future. The RNA polymerase-DNA polymerase collision may contribute to the polymorphism spectra in the intra-operon IGRs. Also, it will be very interesting to compare the intra-operon IGRs spectra in high and low-expression genes in the future as highly expressed genes undergo single-stranded separation more frequently than low-expressed genes.

Intra-operon IGRs are under different selection mechanisms than the inter-operon IGRs as follows: transcription coupled repair; Rho-dependent termination; and ribosome binding site for translation initiation. Translation initiation in bacteria involves Shine-Dalgarno (SD) sequence, located near to the initiation codon. Therefore, intra-operon IGRs are likely to possess SD sequence or RBS. So, changes in this region will be under the purifying selection. It is known

that the Rho factor binds to the pyrimidine-rich sequence in mRNA to cause transcription termination. Therefore, polymorphism in intra-operon IGRs resulting in sequences favoring towards the binding of the Rho factor in the region will be under purifying selection. Therefore, polymorphism in intra-operon IGRs favoring both the above processes is usually under strong selection pressure. This might be attributed to the low frequency of polymorphism in these regions. The increase in purine at intra-operon IGRs in favor of ribosome binding site for the downstream gene and decreases the probability of a Rho-dependent termination site for the upstream gene (Bogden et al., 1999). The increase in keto nucleotides might favor any need of secondary structure (Sen et al., 2022).

Usually, transitions are more frequently selected over transversion in a genome. But when a region of a chromosome has an almost equal proportion of transition and transversion relating to one nucleotide, the role of mutation and/or selection cannot be denied. Such transversions are allowable only when it has a benefic effect on the fitness of an organism. For translational efficacy, the prokaryotic RBS near the SD sequences is selected for Purine-rich nucleotides (Shine and Dalgarno, 1974; Omotajo et al., 2015). Hence the frequent T→A unusual base substitutions at intra-operon IGRs might have arisen or been selected to facilitate in translation of the downstream gene.

It is pertinent to note that our investigation in this study is limited to two Gram negative and closely related bacteria *E. coli* and *S. enterica*. In the future, we would like to extend this study to other Gram-negative bacteria and include Gram-positive bacteria such as *B. subtilis* and others It will be interesting to compare the result with B. subtilis where almost terminators are intrinsic unlike *E. coli* and *S. enterica*. The comparative study of inter-operon IGRs and intra-operon IGRs might be investigated in more detail to develop polymorphism signatures to discriminate between these two regions in a genome, which might be helpful for genome annotation and defining operon in the bacterial genome. Therefore, the observation of lower polymorphism frequency,

high T→A transversion and polymorphism bias for purine and keto nucleotides might be useful in predicting inter-operon *vs* intra-operon IGRs.

4.6. Bibliography

- 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.
- Bogden, C. E., Fass, D., Bergman, N., Nichols, M. D., & Berger, J. M. (1999). The structural basis for terminator recognition by the Rho transcription termination factor. *Molecular cell*, *3*(4), 487-493.
- Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., ... & Drobniewski, F. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature genetics*, *46*(3), 279-286.
- Francino, M. P., & Ochman, H. (1997). Strand asymmetries in DNA evolution. *Trends in Genetics*, *13*(6), 240-245.
- 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.
- Laabei, M., Recker, M., Rudkin, J. K., Aldeljawi, M., Gulay, Z., Sloan, T. J., ... & Massey, R. C. (2014). Predicting the virulence of MRSA from its genome sequence. *Genome research*, *24*(5), 839-849.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.
- Mugal, C. F., von Grünberg, H. H., & Peifer, M. (2009). Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Molecular biology and evolution*, *26*(1), 131-142.
- Omotajo, D., Tate, T., Cho, H., & Choudhary, M. (2015). Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC genomics*, *16*, 1-8.

- Post, L. E., & Nomura, M. (1980). DNA sequences from the str operon of Escherichia coli. *Journal of Biological Chemistry*, *255*(10), 4660-4666.
- 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.
- Romantschuk, M. L., & Müller, U. R. (1983). Mutations in the JF intercistronic region of bacteriophages phi X174 and G4 affect the regulation of gene expression. *Journal of virology*, *48*(1), 180-185.
- 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.
- Shine, J., & Dalgarno, L. (1974). The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Sciences*, *71*(4), 1342-1346.
- Sridhar, J., Sabarinathan, R., Balan, S. S., Rafi, Z. A., Gunasekaran, P., & Sekar, K. (2011). Junker: an intergenic explorer for bacterial genomes. Genomics, Proteomics and Bioinformatics, 9(4-5), 179-182.
- 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.