Chapter 3

Single nucleotide polymorphism in bacterial genomes

3.1 Introduction

As each of the four DNA bases can mutate to any of the three other bases, there are twelve possible directional substitution mutation types that include four transitions and eight transversions. These directional substitution mutations do not occur at equal frequencies in bacterial genomes for mechanistic reasons such as unequal stability among different base pairs, role of enzymes (polymerases and repair enzymes), differential propensity of bases to damages such as deamination, oxidation, and radiation as well as selective reasons such as differential impact on the structure and function of DNA, RNA, and proteins. There are twice the number of transversions than transitions, but the observed frequency of transitions are double the transversions [187][39][105][204][123][184]. Further, the higher propensity of deamination of C and oxidation of G bases increase the frequency of C→T and G→T base substitutions [211], explaining the universal mutation bias towards A/T in genomes [110][7][69][73][170][223].

Different transition and transversion frequencies also vary between the leading strand (LeS) and the lagging strand (LaS) of replication termed as strand substitution asymmetry [114][47], which is explained with GC (or AT) skew in chromosomes [114][62][91][171][115]. Rapid deamination of cytosine to uracil in single stranded DNA leads to a higher rate of C→T mutation favouring a higher frequency of G and T in the LeS compared to the LaS [46][172]. The resulting GC (or AT) skew violates Chargaff's second parity rule which states that the frequency of G and C (or A and T) should be approximately equal within individual DNA strands in genomes [207][45][151]. Firmicutes are exceptional as they exhibit higher frequencies of A than T in the LeS [172] implicating the selection and strong gene-orientation biases in the genomes of this bacterial group [26]. Assuming four-fold degenerate sites (FFS) evolve under near neutrality, Rocha et al. (2006) determined the polymorphism at four-fold degenerate sites (FFS) in seven different bacterial species to explain patterns of GC skew in genomes [172]. They noted that relative consistency between taxa in terms of base compositional biases does not correspond with the underlying base substitution profiles. Although transcription-associated mutation was known to occur, the emphasis was given towards replication-associated mutation as highly expressed genes were not considered for analysis in their study [35][135][94][50][89]. Considering FFS to be nearly neutral, several researchers have extensively used FFS as a reference to describe strand as well as genome composition in bacteria [136][172][69][73]. However, studies in mammalian as well as avian genomes have indicated significant selective constraint at four-fold degenerate sites [150][100]. The findings of earlier work in the areas of co-translational protein folding indicates selection on synonymous codons and ribosome mediated gene regulation [178][200][87][161]. Research by Charneski et al. (2011) support the role of selection for the atypical AT skew in *S. aureus*, emphasizing the role of strand bias gene distribution over mutation in compositional skew [26]. A similar observation suggesting the selection at the IRs has been reported recently in bacteria [215]. Considering mutation being AT biased in genomes [69][73], it was previously assumed that the entire bacterial genome is under selection [170][159].

Genome composition (G+C%) that varies approximately from 13.0 to 75.0 in bacteria is the other major factor influencing codon usage [143][170]. In organisms with high genome G+C%, synonymous codons ending with G/C are more frequently used in comparison to synonymous codons ending with A/T. The reverse is also true for low genome G+C%. Directional mutation theory was proposed to explain the genome G+C% in bacteria according to which it is the mutation rate from A:T to G:C and the vice versa that defines the genome G+C% in bacteria. There are different mechanisms exists in organisms such as deamination of Cytosine and Adenine, Oxidation of Guanine, Tautomeric forms of the nitrogenous bases that support the directional mutation theory. However, recently, two independent research groups published articles suggesting that mutation is Universally A+T biased in genomes [73][69] and there is constant selection to maintain genome G+C% in bacteria. On the basis of this it has been described that even there is selection on the intergenic regions in bacteria [170], which were earlier assumed to be neutral. More recently Thorpe et al., in 2017 [215] have reported regarding selection at the intergenic regions in bacteria. Therefore, indirect evidences are in favor of selection mechanism existing in bacteria to define its genome G+C%. So it is now well known that both mutation and selection mechanisms are important attributes for the genome composition in bacteria.

There has been no study reported in the literature that directly compares polymorphism spectra between IRs and FFS in bacteria. We have investigated nucleotide polymorphisms by analysing many strains of *Escherichia coli* (*Ec*), *Klebsiella pneumoniae* (*Kp*), *Salmonella enterica* (*Se*) belonging to γ-Proteobacteria and two members of Firmicutes such as *Staphylococcus aureus* (*Sa*) and *Streptococcus pneumoniae* (*Sp*). The Firmicutes are known to exhibit different nucleotide compositional asymmetry between strands as compared to γ-Proteobacteria, which has been ascribed to replication-associated mutations due to the presence of two isoforms of DNA polymerase III alpha subunit, PolC and DnaE in Firmicutes [168][175]. In our polymorphism spectra analysis, the trends between IR and FFS were observed to be different. In IRs, the frequency of C→T (G→A) transitions as well as G→T (C→A) transversions were comparatively higher than the other transition and transversions in these five bacterial species indicating the DNA damages due to cytosine deamination and guanine oxidation in this region. No such trends at FFS was observed. Our findings suggest that polymorphism spectra between IR and FFS should be treated separate.

3.2 Materials and Methods

3.2.1 Segregating the leading and the lagging strands in bacterial chromosomes

We carried out a detailed single nucleotide polymorphism study using computational analysis of genomes of total 157 *Escherichia coli* (*Ec*) strains [215], 208 *Klebsiella pneumoniae* (*Kp*) strains [76], 366 *Salmonella enterica* (*Se*) strains [215], 132 *Staphylococcus aureus* (*Sa*) strains [165] and 264 *Streptococcus pneumoniae* (*Sp*) strains [31]. Considering cumulative GC skew diagram for each bacterium, we segregated respective chromosomes into the leading strand (LeS) and the lagging strand (LaS) as has been described earlier by different researchers [114][62]. The method is mentioned below in brief. We found out abundance values of each nucleotide along a genome sequence using non-overlapping moving window of size 1.0 KB. GC skew was calculated as (G-C)/(G+C). Similarly AT skew was calculated as (A-T)/(A+T), RY skew was calculated as [(A+G)-(C+T)/(A+G+C+T)], and KM skew was calculated as [(G+T)-(A+C)/(A+G+C+T)]. For each of the bacteria, cumulative skew diagrams were generated from these deviations, which was used to identify the leading (LeS) and the lagging (LaS) strands regions in a chromosome (Appendix A.2.2). GC skew is positive in the LeS whereas the same is negative in the LaS. Similarly, KM skew is positive in the LeS whereas the same is negative in the LaS. A schematic view of the LeS and the LaS in a double stranded DNA is presented in Figure 3-1. The LeS and the LaS regions of the Watson strand are aligned with the LaS and the LeS, respectively, of the Crick strand in chromosomes. Using the coordinates of protein coding genes (CDS) from the genome annotation, the CDS were mapped to the LeS and the LaS. Sequences other than those coding for the rRNA, tRNA, protein genes and miscellaneous RNAs were considered as intergenic regions (IRs). IRs in the LeS and the LaS of the Watson strand are aligned opposite to IRs in the LaS and the LeS, respectively, of the Crick strand (Figure 3-1). For polymorphism analysis, IRs in either the Watson or the Crick strand were considered. IRs belonging to either the Watson or the Crick strand were segregated into the LeS and the LaS (Appendix A.2.2 and Figure 3-1). In our analysis, we included only large IRs (size greater than 100 bases) from where 35 bases from both the ends of each IR were ignored to minimize the inclusion of any regulatory regions and considered only the nearly neutral polymorphisms.

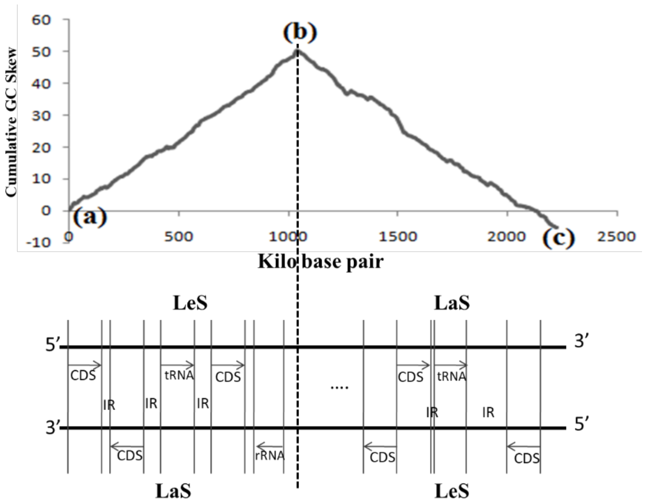

Figure 3-1: A schematic view of the distribution of IRs, CDS, tRNA and rRNA in the leading and lagging strands in double stranded DNA

The figure represents a schematic view of the LeS/LaS and distribution of IRs and CDS, tRNA and rRNA in a double stranded DNA. Considering nucleotide composition of the Watson strand, the skew diagram is generated. In the strand, the region between point (a) and point (b) with positive GC skew is designated as the LeS and the sequence between point (b) and point (c) as the LaS. The LeS and the LaS regions of the Watson strand are aligned with the LaS and the LeS, respectively, of the Crick strand in chromosomes.

3.2.2 Polymorphism analysis at IRs and at FFS in CDS of the bacterial chromosomes

For each bacterium, we extracted alignments of intergenic regions (IRs) and protein-coding regions (CDS) using computer programs written in Python script. Considering the most frequent nucleotide at a position in the sequence alignment, we computed a consensus sequence which was used to identify polymorphisms at different positions. A detailed description of the approach used for analysing polymorphisms in this study is provided in the Appendix A.2.1.

In this study we have done intra-species genome sequence comparison to find out single nucleotide polymorphisms in bacterial genomes. We have done the sequence comparison and the reference sequence have been generated considering the most frequent nucleotide occurrence at a site. The logic is that the most abundant nucleotide in a species is the ancestral nucleotide in the species. Any changes observed in the species is the recent one and not retained for a very long time. This approach has been used by Thorpe et al. 2017 [215]. The following research articles have been published now following the approach [186][6][15].

In the earlier years, when a limited genome sequences were available, homologous sequences used to be compared between two or several closely related species [238]. In case of this inter-species studies, variations between homologous sequences are observed to be more due to species specific selection. Both at synonymous and non-synonymous sites. Therefore, phylogeny is initially constructed and based on the phylogeny nucleotide variation is observed. However, this phylogeny method of nucleotide variation study if applied in case of intra-species it will produce inconsistent result. In support of our argument, we have created phylogeny using *rpoB* and *rpoC* genes both inter-species (Figure 3-2 and Figure 3-3) and intra-species (Figure 3-5 and Figure 3-6). We have noticed that phylogeny prepared using *rpoB* and *rpoC* intra-species is more variable than phylogeny prepared using *rpoB* and *rpoC* inter-species. This is obvious because at inter-species level there is occurrence of strong species-specific purifying selection while it is not present or very low within a species. Using *dnaK* gene phylogeny constructed at intra-species (Figure 3-7) is also producing more variation when compared with that of either *rpoB* or *rpoC*. But, using *dnaK* gene phylogeny constructed at inter-species level (Figure 3-4) is consistent with the phylogeny constructed using *rpoB* and *rpoC*. It indicates that intra-species phylogeny is highly inconsistent across genes while inter-species phylogeny is consistent across genes. Therefore, polymorphism analysis at intra-species level should studied different than that studied at inter-species level.

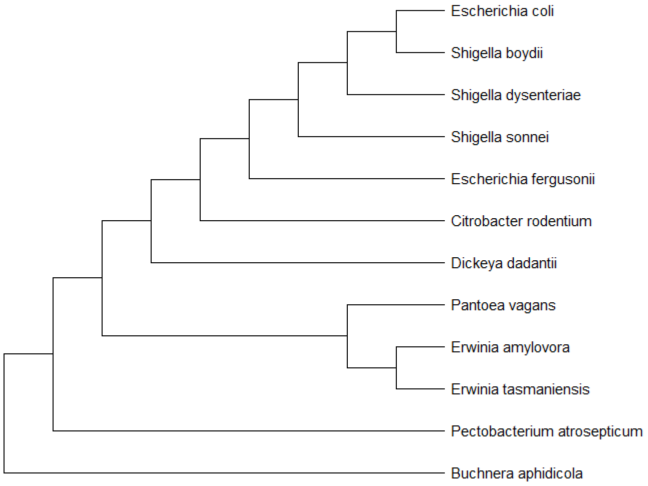

Figure 3-2: Inter-species phylogeny of 12 different bacteria species constructed using *rpoB* gene sequences

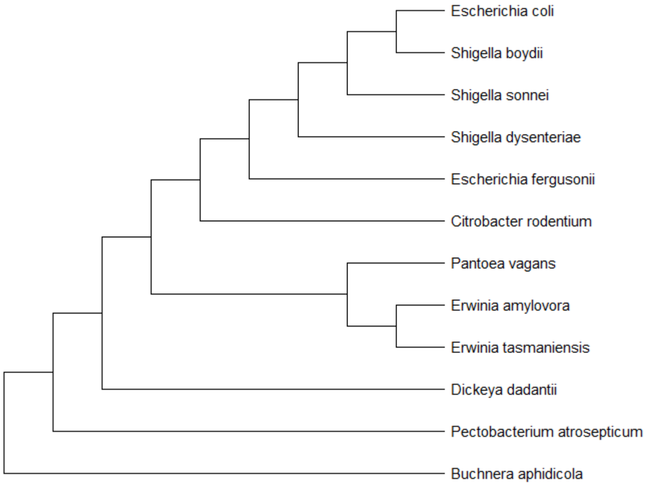

Figure 3-3: Inter-species phylogeny of 12 different bacteria species constructed using *rpoC* gene sequences

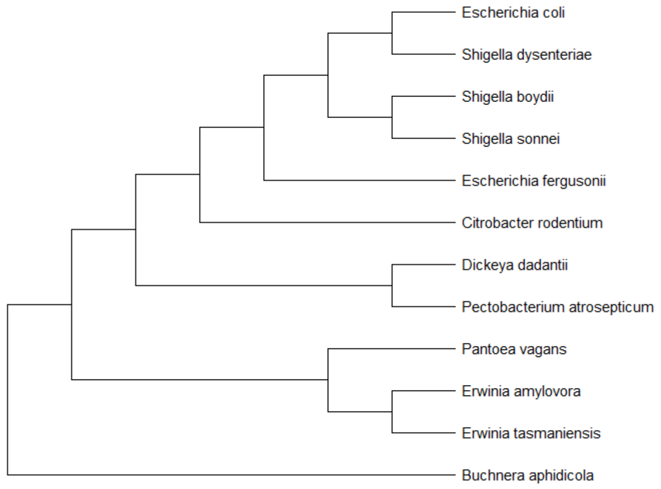

Figure 3-4: Inter-species phylogeny of 12 different bacteria species constructed using rpoB gene sequences

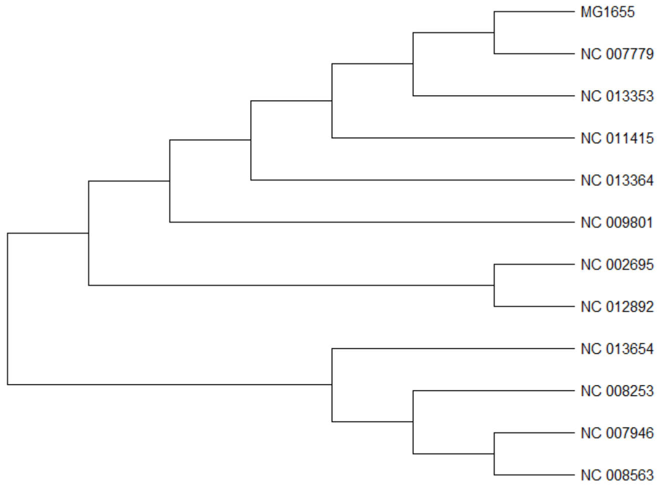

Figure 3-5: Intra-species phylogeny of 12 strains of *E.coli* bacteria using *rpoB* gene sequences

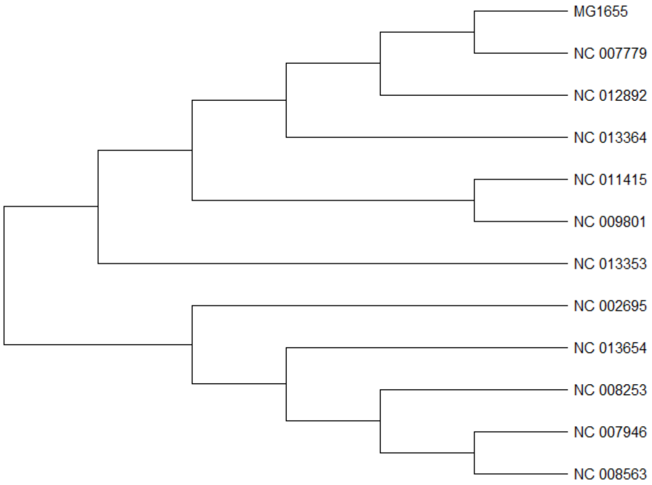

Figure 3-6: Intra-species phylogeny of 12 strains of *E.coli* bacteria using *rpoC* gene sequences

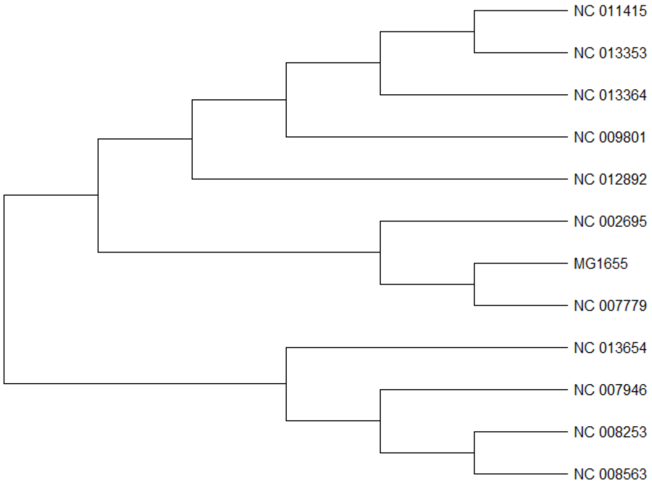

Figure 3-7: Intra-species phylogeny of 12 strains of *E.coli* bacteria using *dnaK* gene sequences

Distribution of pairwise differences in the alignment of *rpoB* gene sequences a sample set of twelve bacteria species and the same distribution in the alignment of *rpoB* gene sequences of twelve strains of *E. coli* are given in the Figure 3-8. It was observed from the distributions that the interspecies pairwise distance is more than 20 times higher than the intraspecies distances. These analysis guided us to follow consensus sequence based approach for estimating polymorphism among the strains of a bacterium in this study. This consensus sequence based methodology is also used in several recent research [186][215][6].

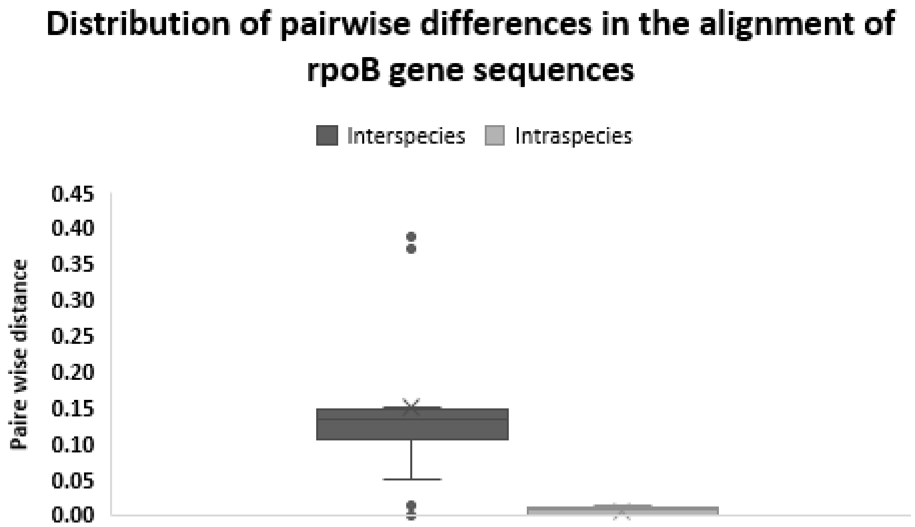

Figure 3-8: Interspecies and intraspecies pairwise difference distribution
Figure presents box and whisker plot of distribution of pairwise differences in the alignment of rpoB gene sequences a sample set of twelve bacteria species and the same distribution in the alignment of rpoB gene sequences of twelve strains of *E. coli*. The y-axis shows the distribution of pairwise difference values. All the sequences considered in interspecies as well as intraspecies comparison were of same size having only base substitution mutations. Software MEGA7 [255][99] is used to estimate pairwise distances.

Phylogenetic relationships among the five bacteria (Appendix A.2.3) were obtained using the reference sequence of *rpoB* and *rpoC* genes by the MEGAX software [99]. The three γ-proteobacteria (*Ec*, *Kp* and *Se*) and two Firmicutes (*Sa* and *Sp*) made different clusters. Further, phylogenetic relationships among the population (Appendix A.2.3) were constructed using the rpoB sequence of all strains of individual organisms [99]. The distribution of polymorphisms among the strains in reference to *rpoB* and *rpoC* genes (Appendix A.2.4) shows that, in general, polymorphism observed in this study is not because of any specific strain but mutations accumulated among all the strains (Appendix A.2.8). A known set of previously published high expression genes [193][185] were considered in this work for the analysis of nucleotide composition and polymorphism at FFS in high expression genes (HEGs) (Appendix A.2.9 and Appendix A.2.10).

In coding sequences (CDS), we considered polymorphism at FFS of the amino acids such as Val, Pro, Thr, Ala and Gly. For example, if a nucleotide change (suppose A→T) observed at the 3rd position of codon, the corresponding codon was found out in the reference sequence (considering the preceding two nucleotides). If the codon codes for Val (i.e., the codon is GTT/GTC/GTA/GTG), then we increase A→T change for Val by 1. Using this approach, we calculated polymorphism at FFS of the five amino acids. Polymorphism frequencies were normalized by dividing the total count of a given change by the total count of the nucleotide in which polymorphism has occurred in the reference sequence. For example, if the total number of C→T change is 20 and the total number of C in the reference sequence (either at IRs or at FFS of that amino acid) is 100, then the normalized frequency is calculated as 20/100 = 0.2. The frequencies of different nucleotide polymorphisms were calculated accordingly. For statistical analysis and determining p-value for significance test, Mann Whitney test is used [125].

3.3 Results

3.3.1 Similar pattern of single nucleotide polymorphism at intergenic regions across these bacteria

Prior to the analysis of nucleotide polymorphism at intergenic regions (IRs), we studied its nucleotide composition in the three γ-Proteobacteria (*Ec*, *Kp*, *Se*) and the two Firmicutes (*Sa*, *Sp*) (Appendix A.2.5). G+C% at IRs was prominently lower than the whole genome in the γ-Proteobacteria than the Firmicutes. The G+C% in the IRs was similar between the LeS and the LaS within a bacterium. Abundance values between complementary nucleotides were more similar than that between two non-complementary nucleotides (Table 3.1). We found out different nucleotide skews (AT/GC/KM/RY) to study compositional biases between the strands in the IRs in reference to the already published results. AT skew was negative in the LeS and positive in the LaS in all bacteria except *Sa* where the pattern was the reverse. The atypical AT skew in IRs was in concordance with the observation reported earlier [26]. It is pertinent to note that the AT skew patterns were not similar in IRs of *Sa* and *Sp*, although both belong to Firmicutes. In contrast to AT skew, GC skew values were found to be positive in the LeS and negative in the LaS of these bacteria. The magnitude of GC skew was observed to be higher than that of AT skew across the five bacteria. Among the five bacteria, the magnitude of both AT and GC skew was observed relatively high in *Sa*. The keto-amino (KM) and purine-pyrimidine (RY) skews were in general positive in the LeS and negative in the LaS of these bacteria (Table 3.1). The strand compositional asymmetry at IRs of these five bacteria was in concordance with the previous findings [115][26].

The disparity of nucleotide composition between the LeS and the LaS was analysed in the context of nucleotide polymorphisms at the IRs. Frequencies of the twelve nucleotide polymorphisms were found out in the LeS and the LaS in these bacteria (Table 3.2). In general, transitions were more frequent than transversions. The *ti*/*tv* values varied from 1.5 to 2.3 among these bacteria (Table 3.2). In a more critical analysis of the transition and transversion frequencies, for example C→T vs. C→G in LeS, it was observed that the former was more than eight times than the latter in *Ec*, while the values were even ten times more in the case of *Se* and *Sp*. This magnitude of difference between the frequencies of C→T and C→G was considerably higher than the theoretical expected four-fold. The frequency values of complementary transition polymorphisms exhibited contrasting

Table 3.1: Compositional features of IRs in LeS and LaS in five bacteria

Nucleotide compositional features	***Ec***		***Kp***		***Se***		***Sa***		***Sp***	
	LeS	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**
A	35099	37981	44856	25471	37686	39691	52274	48613	25008	21804
T	36538	36517	46257	23788	39566	37653	48314	52404	25032	21787
G	26040	23889	40820	19556	29022	26244	21509	17025	13852	9694
C	23398	25521	39052	21513	26341	28585	16212	21998	11371	11450
AT Skew	-0.020	0.020	-0.015	0.034	-0.024	0.026	0.039	-0.038	-0.001	0.000
GC Skew	0.053	-0.033	0.022	-0.048	0.048	-0.043	0.140	-0.127	0.098	-0.083
KM Skew	0.034	-0.025	0.019	-0.040	0.034	-0.033	0.010	-0.008	0.033	-0.027
RY Skew	0.010	-0.001	0.002	-0.003	0.006	-0.002	0.067	-0.063	0.033	-0.027

Table presents count of the nucleotides at IRs and nucleotide composition skews in LeS of IRs as well as LaS of IRs of five bacteria of LeS and LaS of the five bacteria. AT Skew is defined as (A T)/(A+T), GC skew is defined as (G C)/(G+C), KM skew is defined as [(G+T) (A+C)]/[(G+T) + (A+C)], RY skew is defined as [(A+G) (C+T)]/[(A+G) + (C+T)]. Here the A, T, G, C represents respective nucleotide counts.

trends between the strands: C→T frequency was more than that of G→A in the LeS while the reverse was the case in the LaS; A→G frequency was more than that of T→C in LeS while the reverse was the case in the LaS (Figure 3-9). Frequency values of C→T and G→A were about two times more than that of A→G and T→C in both the strands (p-value < 0.01). This indicated that cytosine deamination is a major cause for the high frequency of C→T and G→A in genomes. The higher frequency of C→T than that of G→A in LeS can be attributed to the higher propensity of single stranded DNA towards cytosine deamination over the double stranded DNA. Similarly, higher frequency of A→G than T→C in LeS might be attributed to higher deamination of adenine in the single stranded DNA than the double stranded DNA. The difference between the frequencies of C→T and G→A within a strand was significantly more than that between A→G and T→C (p-value < 0.01). This finding is in concordance with both the following notions that cytosine is more prone to deamination than adenine in DNA and single stranded nature of DNA has a greater impact on cytosine deamination over adenine deamination. Considering the previous explanation that the LeS is remained more exposed as single stranded than the LaS during replication [168], the strand asymmetry regarding transition polymorphisms observed in this study

Table 3.2: Polymorphism spectra at IRs

Substitution spectra and features	***Ec***		***Kp***		***Se***		***Sa***		***Sp***	
	LeS	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**
A→T	0.023	0.024	0.022	0.023	0.024	0.026	0.036	0.035	0.014	0.014
A→G	0.053	0.045	0.048	0.041	0.085	0.079	0.048	0.037	0.049	0.044
A→C	0.015	0.017	0.014	0.016	0.02	0.027	0.012	0.014	0.015	0.016
T→A	0.024	0.024	0.022	0.023	0.022	0.026	0.038	0.033	0.016	0.014
T→G	0.018	0.014	0.015	0.013	0.024	0.023	0.016	0.011	0.015	0.017
T→C	0.045	0.054	0.042	0.048	0.071	0.095	0.042	0.044	0.046	0.05
G→A	0.091	0.114	0.093	0.105	0.175	0.228	0.122	0.134	0.126	0.181
G→T	0.036	0.033	0.04	0.036	0.057	0.065	0.051	0.051	0.043	0.051
G→C	0.012	0.015	0.017	0.017	0.018	0.021	0.016	0.017	0.013	0.015
C→A	0.035	0.037	0.037	0.039	0.057	0.067	0.058	0.045	0.046	0.045
C→T	0.112	0.098	0.095	0.088	0.218	0.188	0.15	0.115	0.187	0.127
C→G	0.014	0.013	0.015	0.015	0.02	0.02	0.02	0.014	0.014	0.015
ti/tv	1.701	1.757	1.527	1.549	2.269	2.145	1.466	1.5	2.318	2.15
AT bias (ti)	2.071	2.141	2.089	2.169	2.519	2.391	3.022	3.074	3.295	3.277
AT bias (tv)	2.152	2.258	2.655	2.586	2.591	2.64	3.893	3.84	2.967	2.909
RY bias	0.005	-0.001	-0.004	-0.002	0.004	-0.003	0.017	-0.014	0.006	-0.005

Table represents 12 different substitution frequencies in LeS and LaS, for example C→T represents total C→T mutations divided by the total count of C in the gene. *ti/tv* is the ratio of transition to transversion. AT bias of transition is calculated as (C→T + G→A) / (A→G + T→C). AT bias of transversion is calculated as (G→T + C→A) / (T→G + A→C). Purine-pyrimidine bias (RY bias) is calculated as (T→A + T→G + C→A + C →G) (A→T + G→T + A→C + G→C).

vindicates replication associated asymmetry in cytosine and adenine deamination between the strands. The pattern of transition polymorphism was similar across these five bacteria indicating the replication associated strand asymmetry is likely to be similar in the IRs of these two groups of bacteria.

Among the transversion polymorphisms, G→T (C→A) was the highest across these bacteria (Table 3.2). The most frequent transversions G→T (C→A) is known to be due to the oxidation of G to form 8oxoG [224], which seems to be occurring equally in both the strands. In *Sa* and *Sp* the frequency of G→T (C→A) was higher than the transitions polymorphisms A→G (T→C). This was contradicting the general notion that frequency of a transition polymorphism is higher than that of a transversion polymorphism. Regarding the other transversion polymorphisms, the order of their frequencies was not consistent across these five bacteria. For example, while A→T (T→A) frequency was the second highest

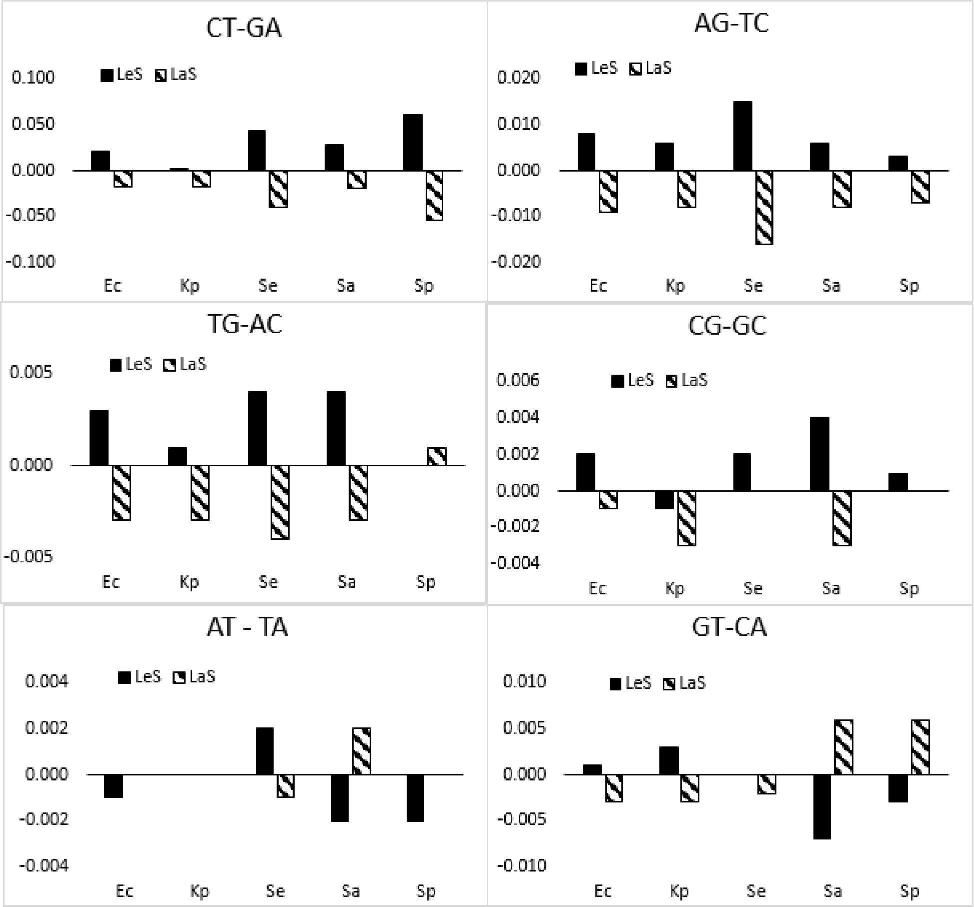

Figure 3-9: Difference between complementary transition polymorphisms in the LeS and the LaS at IRs

The figure presents a two-panel histogram on the six pairs of complementary polymorphism patterns in both the LeS and LaS of IRs. The height of the vertical bar represents the polymorphism frequency difference between a complementary polymorphism pair. Black bars and striped bars present polymorphism frequency differences in LeS and LaS, respectively. The X-axis represents the name of the five organisms: *Escherichia coli (Ec)*, *Klebsiella pneumoniae (Kp), Salmonella enterica (Se), Staphylococcus aureus (Sa)* and *Streptococcus pneumoniae (Sp)*. Y-axis represents frequency values. In general, the pattern of complementary polymorphism in LeS and LaS are found to be in opposite direction.

among the transversions in *Ec*, *Kp*, *Se* and *Sa*; it was the lowest in *Sp*. The reason behind the higher frequency of A→T (T→A) than that of A→C (T→G) or C→G (G→C) is not known. The frequency values of complementary *tv* pairs were similar within a strand. Purine to pyrimidine (R→Y) was observed to be more frequent than pyrimidine to purine (Y→R) in the LeS in all the bacteria except *Kp*. The magnitude of the bias was higher in the Firmicutes than the γ-Proteobacteria. In both *ti* and *tv*, polymorphisms were more than two times biased towards A/T over G/C in three γ-Proteobacteria, and the same were more than three times biased in the Firmicutes. This was in concordance with their genome composition values that polymorphism was more biased to AT in genome with low genome G+C composition.

In conclusion, polymorphism analysis at IRs has revealed that replication associated asymmetry between the LeS and the LaS, can mainly be attributed to higher deamination of cytosine and adenine in the former than the latter. This polymorphism asymmetry can explain the nucleotide composition asymmetry between the strands in IRs except the atypical AT-skew in *Sa*. Further the magnitude of asymmetry regarding G→T polymorphism was not observed to be high between the LeS and the LaS, unlike cytosine deamination. It might be that guanine oxidation process is not different between the strands or the repair of 8-oxoG is more efficient in the LeS than the LaS.

3.3.2 The polymorphism pattern at the four-fold degenerate site (FFS) is different from that at IRs

Researchers have already analysed SNPs at FFS in bacteria to understand the mechanism of molecular evolution [136][132][166][169][172]. However, there has been no report to have a comparative study between the polymorphisms pattern between IRs and FFS. Further, the spectra at FFS have not been analysed in the context of cytosine deamination and guanine oxidation damage. Also, the polymorphisms were not compared across FFS of different amino acid codons within a bacterium. Here we found out polymorphism frequencies at FFS in codons of amino acids such as Val, Pro, Thr, Ala and Gly in the γ-Proteobacteria and Firmicutes (Appendix A.2.6). In general, *ti* was more frequent than *tv* across the five amino acids in these bacteria. However, the *ti*/*tv* values were variable, which indicated differences across these amino acid regarding polymorphism at FFS (Table 3). The transition as well as the transversion polymorphisms were more biased towards A/T in the Firmicutes than the γ-Proteobacteria (p-value < 0.01). Further the magnitudes of bias towards A/T over G/C were not consistent across the five amino acids either in the case of transition or in the case of transversion polymorphisms (Table 3). It is pertinent to note that polymorphisms such as C→T and G→A were not always more frequent than A→G and T→C. This was observed in case of Thr in *Ec* and across the five amino acids in case of *Kp* (Appendix A.2.6). Further, among the transversion polymorphisms, G→T (C→A) were not the most frequent ones in *Ec* and *Kp*. So, the notion that C→T (G→A) is the most frequent transition polymorphism and G→T (C→A) is the most frequent transversion polymorphism is incorrect at FFS, unlike at IRs. Not only the frequency of a polymorphism was variable across the five amino acids, but the order of different polymorphisms regarding their frequency at FFS was also not consistent across the five amino acids (Appendix A.2.6, Figure 3-10). The difference among the amino acids regarding the polymorphisms within a bacterium indicated that the polymorphisms at FFS not necessarily be treated as nearly neutral.

We compared frequency values between complementary polymorphisms within a strand to find out any possible role of strand asymmetry. In general, the difference between C→T and G→A were positive in the LeS and negative in the LaS across the five amino acids in these five bacteria. This indicated replication associated strand asymmetry regarding cytosine deamination. Regarding the other transition polymorphisms, difference between A→G and T→C were high but non-uniform across the five amino acids within a bacterium: the

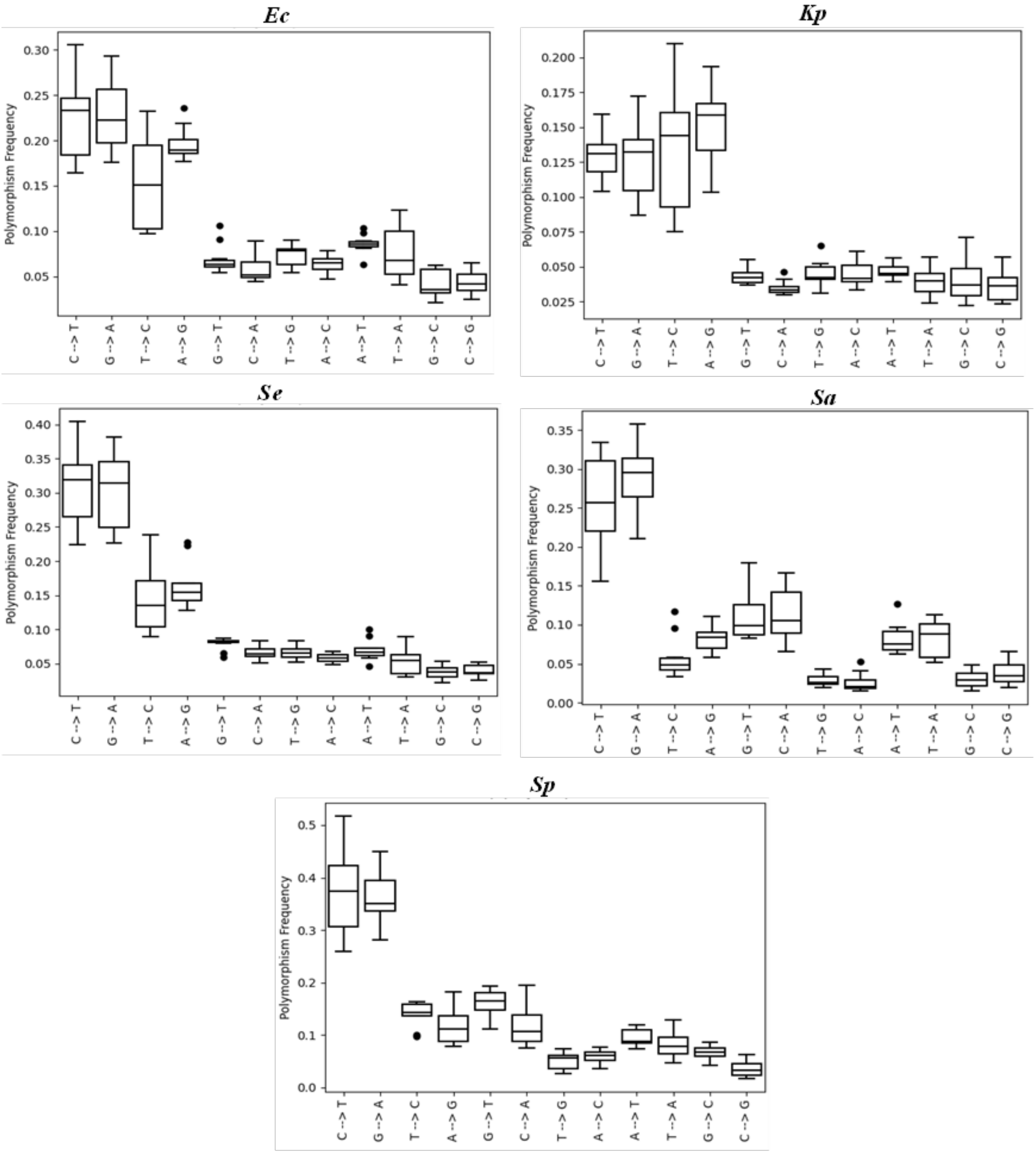

Figure 3-10: Polymorphism frequency at FFS across five amino acids in five bacteria

The figure presents distribution of 12 substitutions of FFS in terms of box-plot of *Ec*, *Kp*, *Se*, *Sa* and *Sp*. The x-axis presents the 12 substitutions and the y-axis presents the polymorphism frequency. The frequency of polymorphism is not uniform across the five amino acids. In case of *Sa* and *Sp*, polymorphism leading to A/T nucleotides at FFS were more frequent than that leading to G/C. In case of *Kp*, transitions polymorphism leading to G/C were more frequent than that leading to C/T. Impact of genome G+C% on polymorphism frequency was evident in this study.

Table 3.3: Comparison between transition-transversion polymorphism at FFS of five bacteria

Bacteria	Polymorphism feature	Val		Pro		Thr		Ala		Gly	
		LeS	LaS	LeS	LaS	LeS	LaS	LeS	LaS	LeS	LaS
Ec	*ti*/*tv*	1.558	1.589	1.288	1.342	1.598	1.739	1.513	1.521	1.732	1.893
	AT bias (*ti*)	1.447	1.320	1.375	1.279	1.020	1.091	1.366	1.333	1.366	1.376
	AT bias (*tv*)	0.915	0.850	1.117	0.945	0.758	0.795	0.904	0.921	1.027	1.096
Kp	*ti*/*tv*	1.701	1.761	1.283	1.464	1.707	1.755	1.631	1.647	1.733	1.787
	AT bias (*ti*)	0.949	0.833	1.076	0.963	0.813	0.745	0.902	0.844	0.932	0.956
	AT bias (*tv*)	0.944	0.866	0.900	0.946	0.973	0.766	0.927	0.846	0.777	0.905
Se	*ti*/*tv*	2.042	2.185	1.604	1.624	1.818	1.982	1.741	1.846	2.462	2.541
	AT bias (*ti*)	2.292	2.165	2.202	2.242	1.843	2.037	2.195	2.098	1.559	1.508
	AT bias (*tv*)	1.168	1.078	1.403	1.278	1.176	1.161	1.091	1.065	1.109	1.220
Sa	*ti*/*tv*	1.383	1.274	1.383	1.296	1.088	1.339	1.359	1.311	1.476	1.406
	AT bias (*ti*)	3.362	2.922	5.130	5.917	4.041	4.786	3.928	4.491	3.077	2.713
	AT bias (*tv*)	3.800	3.280	5.122	7.641	3.770	4.511	3.943	4.347	3.427	3.259
Sp	*ti*/*tv*	1.627	1.669	1.477	1.472	1.435	1.417	1.570	1.398	1.408	1.484
	AT bias (*ti*)	2.097	2.225	3.967	4.051	2.663	2.852	2.619	2.817	2.698	3.168
	AT bias (*tv*)	1.677	1.563	3.059	4.677	2.039	2.604	2.371	3.325	2.520	3.064

ti/*tv* is the ratio of transition to transversion. AT bias of transition is calculated as (C→T + G→A) / (A→G + T→C). AT bias of transversion is calculated as (G→T + C→A) / (T→G + A→C).

difference can be positive in case of an amino acid while negative in case of another amino acid (Figure 3-11). Further, the difference between A→G and T→C values were similar both in the LeS and the LaS. This indicated that the difference was not generated due to replication associated strand asymmetry. Similarly, the inconsistent pattern across the amino acids indicated that the pattern was not due to transcription associated mutation. In case of complementary transversions, the difference values were high but not uniform across the five amino acids within a bacterium. Further, the difference values remain similar both in the LeS and the LaS. Unlike IRs, the difference between complementary polymorphisms were high at FFS and the difference value reflected were more of amino acid specific, not strand specific.

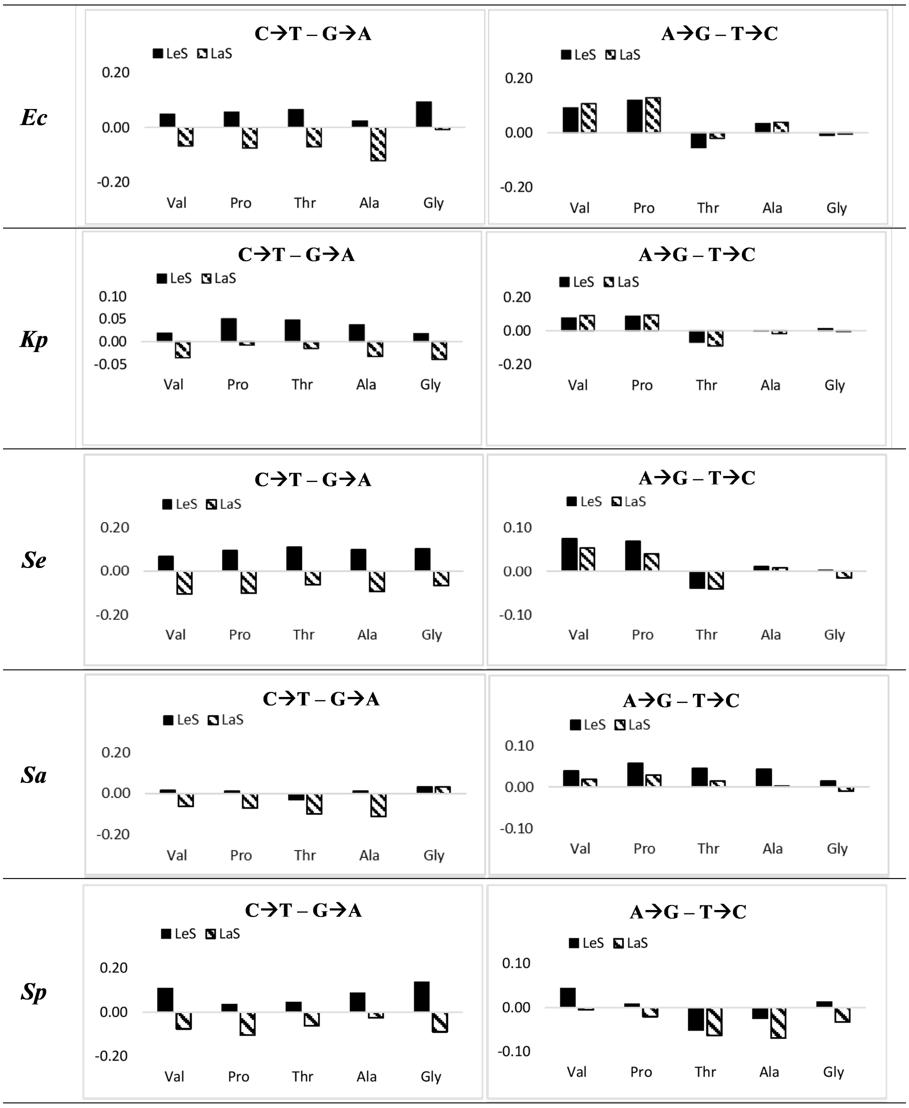

Figure 3-11: Difference between complementary transition polymorphisms in the LeS and the LaS at FFS

The figure presents a two-panel histogram of difference between complementary transition polymorphisms in the LeS and LaS at FFS of five bacteria. The height of the vertical bar represents the polymorphism frequency difference between a complementary polymorphism pair. Black bar and striped bar represent polymorphism frequency differences in LeS and LaS, respectively. The X-axis represents the name of the five amino acids namely Val, Pro, Thr, Ala and Gly. Y-axis represents frequency values. In general, the pattern of complementary transition polymorphism of C→T and G→A in LeS and LaS are found to be in opposite direction, whereas the other complementary transition polymorphism (A→G and T→C) do not show contrasting pattern.

3.3.3 The polymorphism at the four-fold degenerate site coincides with codon usage bias

The amino acid specific polymorphisms at FFS indicated that the polymorphisms were most probably influenced by codon usage bias. So, we compared polymorphisms with nucleotide frequency at FFS of the five amino acids, which represented the codon usage bias of individual amino acids. Wide variation regarding nucleotide frequency at FFS was observed across the five amino acids within a genome (Table 3.4). In general, codon usage bias was observed to be strand independent as the frequency values were similar between the two strands for an amino acid at FFS. This was in concordance with observations by earlier researchers [193][188][227]. However, a moderate impact of strand asymmetry was observed on codon usage bias because in general the frequencies of G and T at the FFS of an amino acid in LeS was more than that in LaS. The reverse was true regarding the frequencies of A and C.

To understand the relation of the polymorphism with codon usage bias, if any, we did a detailed comparative study in individual bacteria. The polymorphism spectra at FFS for individual bacterium is given in Table 3.5. In case of *Ec*, the polymorphism pattern between complementary nucleotide pairs revealed that Pro and Gly often behaved opposite to each other. Considering the previous knowledge that G-ending codon in Pro (CCG) is the most preferred whereas the G-ending codon (GGG) in Gly is the least preferred [180], we analysed both forward and reverse nucleotide polymorphisms. In case of Gly, G→C and G→T transversions were more frequent than C→G and T→G respectively, that supported the higher abundance of GGT/GGC over GGG in the genome. Whereas in case of Pro, G→C and G→T transversions were found to be less frequent than C→G and T→G respectively, that supported the lower abundance of CCT/CCC than CCG in the genome. A→T was more frequent than T→A in Val and Gly that favoured higher abundance of GTT/GGT over GTA/GGA. A→T was less frequent

Table 3.4: Nucleotide frequency at FFS

Bacteria	**Base**	**Val**		**Pro**		**Thr**		**Ala**		**Gly**	
		LeS	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**	**LeS**	**LaS**
Ec	**A**	0.146	0.151	0.172	0.190	0.104	0.116	0.201	0.207	0.093	0.096
	T	0.255	0.250	0.141	0.151	0.166	0.151	0.159	0.152	0.342	0.338
	G	0.397	0.366	0.587	0.525	0.282	0.258	0.400	0.333	0.158	0.129
	C	0.202	0.233	0.100	0.134	0.448	0.475	0.240	0.308	0.406	0.437
Kp	**A**	0.090	0.082	0.068	0.074	0.031	0.034	0.063	0.058	0.062	0.066
	T	0.141	0.123	0.084	0.085	0.085	0.075	0.091	0.080	0.169	0.149
	G	0.497	0.445	0.713	0.661	0.254	0.223	0.463	0.381	0.203	0.166
	C	0.272	0.349	0.134	0.181	0.630	0.668	0.383	0.481	0.566	0.620
Se	**A**	0.155	0.155	0.112	0.122	0.082	0.087	0.116	0.119	0.107	0.105
	T	0.216	0.209	0.152	0.147	0.120	0.106	0.126	0.115	0.236	0.222
	G	0.388	0.334	0.605	0.555	0.371	0.328	0.492	0.416	0.176	0.130
	C	0.241	0.303	0.130	0.177	0.427	0.478	0.266	0.351	0.481	0.544
Sa	**A**	0.348	0.314	0.528	0.482	0.507	0.492	0.475	0.454	0.229	0.195
	T	0.410	0.412	0.318	0.396	0.277	0.311	0.304	0.348	0.562	0.548
	G	0.145	0.112	0.134	0.083	0.182	0.130	0.161	0.099	0.072	0.045
	C	0.098	0.161	0.021	0.039	0.035	0.066	0.060	0.099	0.137	0.213
Sp	**A**	0.202	0.171	0.490	0.453	0.344	0.305	0.269	0.234	0.295	0.330
	T	0.398	0.410	0.340	0.384	0.320	0.377	0.404	0.435	0.436	0.395
	G	0.188	0.122	0.092	0.050	0.113	0.070	0.114	0.061	0.134	0.101
	C	0.212	0.298	0.078	0.113	0.223	0.247	0.212	0.270	0.135	0.173

Table presents amino acid wise nucleotide frequencies in both LeS and LaS at FFS of five bacteria

than T→A in Pro that favoured CCA to be more frequent than CCT. T→C was the most frequent in case of Thr among the five amino acids that made ACC the most preferred codon. Low frequency of T→C in case of Pro and Val, favoured the low frequency of GTC/CCC. While C→T and T→C were of similar frequency in case of Thr, C→T was more than the frequency of T→C in case of Val, Pro and Ala that favoured the higher abundance of T-ending codon (GTT/CCT/GCT) over the C-ending codons. G→A transition was more frequent than A→G in case of Gly, which corresponds to the higher frequency of GGG over GGA. These observations suggested that polymorphism at FFS was influenced by codon usage bias in *Ec*.

A similar comparative study of the polymorphism at FFS and codon usage bias in these amino acids were studied in the other two γ-proteobacteria *Kp* and *Se*. In *Kp*, as the genome is G+C high, G/C-ending codons are more abundant

Table 3.5: Polymorphism spectra at FFS of five amino acids in the leading and the lagging strands of five bacteria

Bacteria	Polymorphism Spectra	Val		Pro		Thr		Ala		Gly	
		LeS	LaS	LeS	LaS	LeS	LaS	LeS	LaS	LeS	LaS
Ec	A→T	0.084	0.081	0.098	0.103	0.085	0.063	0.087	0.089	0.086	0.082
	A→G	0.187	0.203	0.219	0.236	0.177	0.192	0.181	0.193	0.185	0.187
	A→C	0.057	0.066	0.047	0.063	0.078	0.075	0.056	0.061	0.067	0.070
	T→A	0.051	0.055	0.123	0.107	0.070	0.066	0.087	0.104	0.041	0.046
	T→G	0.061	0.054	0.090	0.082	0.079	0.071	0.080	0.078	0.079	0.055
	T→C	0.097	0.097	0.101	0.108	0.232	0.213	0.147	0.155	0.195	0.193
	G→A	0.181	0.232	0.192	0.257	0.176	0.256	0.212	0.293	0.213	0.265
	G→T	0.057	0.054	0.064	0.062	0.069	0.064	0.063	0.060	0.106	0.091
	G→C	0.031	0.034	0.021	0.027	0.052	0.062	0.034	0.036	0.060	0.062
	C→A	0.051	0.048	0.089	0.075	0.050	0.052	0.060	0.068	0.044	0.046
	C→T	0.230	0.164	0.248	0.183	0.241	0.186	0.236	0.171	0.306	0.258
	C→G	0.054	0.046	0.058	0.065	0.034	0.034	0.046	0.038	0.036	0.025
Kp	A→T	0.051	0.044	0.056	0.050	0.039	0.041	0.045	0.045	0.050	0.044
	A→G	0.164	0.193	0.161	0.182	0.103	0.119	0.131	0.139	0.168	0.156
	A→C	0.033	0.041	0.035	0.040	0.042	0.053	0.039	0.045	0.061	0.054
	T→A	0.032	0.033	0.057	0.048	0.042	0.044	0.038	0.045	0.024	0.027
	T→G	0.038	0.041	0.065	0.052	0.031	0.041	0.043	0.046	0.051	0.041
	T→C	0.090	0.101	0.075	0.088	0.170	0.210	0.134	0.155	0.154	0.162
	G→A	0.111	0.141	0.102	0.134	0.087	0.130	0.101	0.141	0.141	0.172
	G→T	0.037	0.038	0.044	0.046	0.041	0.037	0.043	0.041	0.054	0.055
	G→C	0.027	0.037	0.022	0.029	0.042	0.051	0.030	0.037	0.057	0.071
	C→A	0.030	0.033	0.046	0.041	0.030	0.035	0.033	0.036	0.033	0.031
	C→T	0.130	0.104	0.152	0.126	0.135	0.115	0.138	0.107	0.159	0.132
	C→G	0.043	0.039	0.057	0.056	0.023	0.025	0.038	0.034	0.029	0.025
Se	A→T	0.063	0.046	0.100	0.091	0.067	0.065	0.072	0.072	0.062	0.058
	A→G	0.164	0.154	0.168	0.155	0.133	0.128	0.139	0.152	0.228	0.223
	A→C	0.049	0.050	0.053	0.064	0.058	0.060	0.059	0.065	0.055	0.068
	T→A	0.034	0.033	0.073	0.089	0.061	0.049	0.062	0.064	0.031	0.040
	T→G	0.058	0.052	0.066	0.062	0.067	0.064	0.073	0.074	0.083	0.059
	T→C	0.089	0.100	0.099	0.114	0.172	0.167	0.128	0.143	0.226	0.238
	G→A	0.257	0.326	0.246	0.351	0.226	0.331	0.244	0.356	0.303	0.381
	G→T	0.065	0.059	0.084	0.082	0.087	0.080	0.087	0.084	0.084	0.083
	G→C	0.030	0.039	0.022	0.027	0.042	0.044	0.032	0.037	0.053	0.049
	C→A	0.060	0.051	0.083	0.079	0.060	0.064	0.057	0.064	0.069	0.072
	C→T	0.323	0.224	0.342	0.252	0.336	0.270	0.342	0.263	0.405	0.314
	C→G	0.049	0.038	0.052	0.043	0.035	0.026	0.048	0.035	0.035	0.026
Sa	A→T	0.074	0.084	0.066	0.094	0.077	0.074	0.065	0.062	0.097	0.127
	A→G	0.084	0.074	0.090	0.069	0.084	0.063	0.091	0.058	0.111	0.106
	A→C	0.021	0.031	0.015	0.017	0.018	0.021	0.021	0.025	0.041	0.053
	T→A	0.065	0.056	0.113	0.097	0.112	0.087	0.103	0.090	0.056	0.052
	T→G	0.024	0.019	0.034	0.022	0.043	0.024	0.032	0.024	0.034	0.028
	T→C	0.046	0.054	0.033	0.040	0.039	0.049	0.048	0.058	0.096	0.117
	G→A	0.210	0.218	0.310	0.358	0.263	0.318	0.266	0.315	0.303	0.287
	G→T	0.089	0.098	0.087	0.131	0.083	0.099	0.083	0.109	0.151	0.179
	G→C	0.025	0.021	0.015	0.034	0.024	0.036	0.021	0.043	0.048	0.039
	C→A	0.082	0.066	0.164	0.167	0.147	0.104	0.126	0.104	0.106	0.085
	C→T	0.227	0.156	0.321	0.287	0.234	0.218	0.280	0.206	0.334	0.318
	C→G	0.030	0.019	0.051	0.020	0.066	0.039	0.053	0.029	0.039	0.026
Sp	A→T	0.087	0.084	0.085	0.074	0.089	0.102	0.120	0.112	0.120	0.080
	A→G	0.182	0.131	0.109	0.078	0.111	0.080	0.139	0.080	0.173	0.112
	A→C	0.068	0.072	0.041	0.036	0.066	0.063	0.060	0.051	0.077	0.052
	T→A	0.055	0.047	0.128	0.085	0.105	0.072	0.095	0.064	0.096	0.064
	T→G	0.065	0.047	0.060	0.026	0.062	0.033	0.056	0.029	0.073	0.058
	T→C	0.138	0.136	0.101	0.098	0.162	0.143	0.163	0.149	0.161	0.144
	G→A	0.282	0.335	0.399	0.408	0.341	0.349	0.352	0.335	0.383	0.450
	G→T	0.135	0.111	0.150	0.167	0.146	0.162	0.177	0.182	0.183	0.193
	G→C	0.062	0.058	0.043	0.067	0.078	0.067	0.059	0.086	0.070	0.080
	C→A	0.088	0.075	0.159	0.123	0.115	0.088	0.098	0.084	0.195	0.144
	C→T	0.389	0.259	0.434	0.305	0.386	0.287	0.439	0.310	0.518	0.361
	C→G	0.049	0.022	0.040	0.026	0.036	0.019	0.031	0.017	0.063	0.048

over the A/T-ending codons. In general A→G was more frequent than C→T and G→A in all amino acids that favoured higher abundance of G-ending codons in the genome. A→C was more frequent than C→A in all amino acids except Pro that supported CCC being preferred low here. Further, we noticed correlation in preference of C-ending codons and higher frequency of G→C in Thr and Gly, whereas the G-ending codons were preferred in Val and Pro that corresponded to the higher frequency of C→G. A→T was more frequent than T→A in Val and Gly, which supported the T-ending codons being preferred over the A-ending codons in these amino acids. T→G was more frequent than G→T in Pro while T→G was less frequent than G→T in Gly which supported G-ending codon being preferred in Pro but not in Gly. Therefore, the impact of codon usage bias was observed on the polymorphism at FFS of *Kp*. Similarly, in *Se*, G→C transversion was more frequent than C→G in Thr and Gly while the reverse pattern was true in Pro in both the strands. This was in concordance with the preference of C-ending codon in Thr, avoidance of C-ending codon in Pro and G-ending codon in Gly. Polymorphism at A→T was more frequent than T→A in Val and Gly which supported the higher abundance of GTT/GGT in the genome. Therefore, polymorphism pattern at FFS was influenced by codon usage bias in *Se*.

In the two Firmicutes, A- and T-ending codons were the most frequent codons across the five amino acids. In *Sa*, A-ending codons were more frequent than T-ending codons in Pro, Thr and Ala. In concordance, A→T was more frequent than T→A in Val and Gly, while the reverse was the case in Pro, Thr and Ala. In addition, G→A was more frequent than C→T in case of Thr because ACA is the most frequent codon here. In *Sp*, A-ending codon was more frequent than T-ending codons in Pro. It was obvious to observe that A→T was more frequent than T→A in case of Val, Ala and Gly but the reverse was true in case of Pro, which was in concordance with the codon usage bias. Therefore, polymorphism at FFS in *Sa* as well as in *Sp* was influenced by codon usage bias in these bacteria.

3.3.4 The polymorphism spectra at the four-fold degenerate sites in the high expression genes is different from that in the four-fold degenerate sites in the whole genome as well as IRs

It is generally believed that translational selection on codon usage bias is stronger in the high expression genes (HEGs) than the low expression genes (LEGs). Only a small fraction of the genome comprises of high expression gene [193]. Therefore, we analysed polymorphism spectra at FFS of the HEGs, which was found to be different from the whole genome as well as the IRs. It is pertinent to note that codon usage bias in the HEGs is different from the whole genome. Polymorphism spectra at the FFS of the HEGs were significantly lower than that of the whole genome. This can be explained on the basis of stronger selection at HEGs than that at the FFS of the whole genome. Interestingly, some of the polymorphism frequency at HEGs was observed to be higher than that at IRs (Appendix A.2.7). This suggests that over all strong selection at IRs. So, in conclusion it can be said that polymorphism spectra at IRs is different from that at FFS. The selection at IRs is different from that at FFS.

3.4 Discussion

Analysing nucleotide polymorphism in genome sequences of several strains belonging to a single species provides avenue to study mechanisms of molecular evolution. Further the leading and the lagging strands in chromosomes can be easily segregated using computational method facilitating researcher to investigate the replication-associated mutation asymmetry between the strands in bacteria [47]. Despite of low abundance of IRs, these regions can be distinctly identified because of the simplicity of the coding sequences in bacterial chromosomes. In this study, nucleotide polymorphism has been analysed at the intergenic regions (IRs) and four-fold degenerate site (FFS) of three bacterial species namely *E. coli* (*Ec*), *K. pneumoniae* (*Kp*), *S. enterica* (*Se*) belonging to γ-proteobacteria and two other bacterial species namely *S. aureus* (*Sa*) and *S. pneumoniae* (*Sp*) belonging to Firmicutes. Nucleotide compositional asymmetry between the strands is well studied in bacterial chromosomes [114][115]. The higher abundance of the keto nucleotides (G, T) in the LeS than the LaS has been explained based on frequent cytosine deamination in single stranded DNA [166][47][168] which is corroborated by the observation of positive GC skew and negative AT skew in the LeS in bacteria. However, the higher magnitude of GC skew than AT skew in the LeS of most bacterial genomes could not be explained based on the cytosine deamination theory. Rocha et al. (2006) had explained different mutation bias that might lead to similar skew patterns in bacterial genomes. Similarly, the positive AT skew observed in the LeS of *S. aureus* could not be explained by the cytosine deamination theory. Therefore, a detailed investigation has been done in this study to understand the nucleotide compositional asymmetry between the strands in bacteria. In this study, polymorphism analysis at IRs of these five bacteria has revealed that C→T and G→A transition polymorphism are the most frequent ones and display the main difference between the strands. This is in favour of the notion that cytosine deamination is the major cause of polymorphism in genomes and the process is more frequent in the LeS than the LaS. In a small magnitude, we have also observed that A→G and T→C contributes towards strands asymmetry at IRs. In parallel with cytosine deamination theory, it may be hypothesized that more frequent adenine deamination in LeS might result in higher A→G transition in the strand than the complementary strand. It is known that cytosine deamination and adenine deamination have an opposite impact on genome G+C%. Regarding transition polymorphisms at IRs, the five bacteria behave similar in this study. However, in transversions, frequency of a polymorphism is observed to be similar between the strands and the difference value between complementary transversions within strand is very low. Therefore, contribution of transversion polymorphism in strand asymmetry is very low in the IRs of these bacteria. It is pertinent to note that frequency of G→T (C→A) are higher than other transversions. Transversion polymorphisms increases A/T at IRs like transition polymorphisms. But the bias towards A/T of these polymorphisms are more in the two Firmicutes than the γ-Proteobacteria. The G→T (C→A) value is higher as well as A→G (T→C) value is lower in Firmicutes in comparison to the γ-Proteobacteria for which A/T bias is observed to be more in the former than the latter. The polymorphism study at the IRs suggests that the replication associated strand asymmetry is indifferent between the two groups of bacteria. Therefore, the atypical AT skew in the chromosome of *Sa* is not supported by the polymorphism at IRs.

In the two Firmicutes, *Sa* and *Sp* exhibit opposite patterns of nucleotide composition at FFS. The nucleotide A is more frequent than T in Sa, while T is more frequent than A in Sp. The coding sequence is more abundant in the leading strand than the lagging strand of the Firmicutes [168]. Therefore, the abundance of A is more than T in the LeS of *Sa* and the abundance of T is more than A in the LeS of *Sp*. Codon usage bias at FFS of an amino acid is similar between the strands indicating the weak influence of strand specific polymorphism. Therefore, the atypical AT skew in *Sa* can be attributed to codon usage bias, which is due to the selection on codon usage bias. Our findings are in concordance with the earlier observation that selection and gene distribution asymmetry between the strands was associated with the atypical AT skew in *Sa* [26].

The polymorphism pattern at IRs complies with the replication associated mutation (RAM) in these five bacterial genomes. However, the observation of polymorphism frequency at IRs being lower than that at FFS suggests that the IRs are under strong selection. This is in concordance with the conclusion made by Thorpe et al. (2017) [215]. It might be that because of their involvement in gene regulation expression IRs is resistant to change. The polymorphism frequency at FFS is higher than that in IRs might be attributed to additional mutation pressures as well as silent mutations at the former.

Genome G+C% in bacteria varies from 13.0 to 75.0% which accounts a difference of 62% between the minimum and maximum genome composition [159]. Directional mutation bias in support of the neutral theory of evolution has been proposed to explain genome G+C% in organisms [206]. In support of directional mutation theory, Muto and Osawa (1987) demonstrated that synonymous codon usage varies between high and low G+C bacterial genomes [136]. But theoretical analysis of the G+C% of the synonymous codons suggests that the maximum G+C composition difference between two synonymous codons (e.g., GGU and GGC) of an amino acid can be 33.33% with exceptions only in Arg (e.g., CGG, AGU) and Leu (e.g., UUA, CUG) where the difference can be up to 66.67%. Hence, the synonymous codon usage range should be held accountable to a value around 33.33% instead of 62.0% (75.0 - 13.0). It can be argued that the directional mutation theory inadequately explains the genome G+C composition in bacteria as IRs contribute a minor portion of the genome size. It is pertinent to note that the results of earlier mutation analysis in bacterial genomes could not provide substantial evidence in support of the directional mutation theory [69][73][170]. Hence, the possible existence of an unknown selection mechanism responsible for genome G+C% has been hypothesized [170][159]. The role of recombination in determining the genome composition of bacteria have also been implicated [16]. Reported correlation between G+C content and amino acid composition also indicates towards selection pressure at protein level [205][38]. In the genetic code table, synonymous codons are not identical with their G+C content. For example, UUU which codes for phenylalanine is 100% A+T and 0% G+C whereas, the synonymous codon UUC is 66.6% A+T and 33.3% G+C. Similarly, UUA and CUG the codons of Leu are distinctly different regarding their G+C%. It is usually observed that synonymous codons with low G+C% are found in higher proportions in genome with low G+C% while synonymous codons with high G+C% are found in higher proportions in genome with high G+C%. The genome G+C% not only affect the codon usage bias but it also influences amino acid usage. The best example is observed in the case of Lys and Arg where both are basic amino acids. For example, Lys codons are AAA and AAG while Arg codons are CGN and AGR. Because Lys codons are A+T enriched in comparison to Arg codons, organisms with low genome G+C% prefers Lys amino acids than Arg amino acids in homologous protein sequences. G+C% variation at FFS across the amino acids in different genomes reported in this work is an indication of codon usage bias influencing the observed difference in genome composition. The polymorphism difference observed across the amino acids at FFS indicate that amino acid specific translational selection might attribute towards this difference. It is already known that GGG and CCC codons are prone to translational frameshift [140][141]. Interestingly, GGG and CCC codons were not preferred in both the strands of five studied bacteria. Transversions such as G→C is more than C→G in case of Gly while the same is less in case of Pro. GGC codon has been reported to be selected positively in bacterial genomes [178][180].

3.5 Conclusion

It is interesting to note that our analysis using large number of genomes of γ-Proteobacteria and Firmicutes have indicated towards the role of codon usage bias in determining the genome G+C%. We anticipated that future research on translation rate of synonymous codons is expected to uncover the mechanism of genome composition in AT and GC rich bacteria. Large range of genome G+C% is a classic example of the neutral theory of evolution in bacteria which means that there is no specific advantage that could be linked to genome composition [104]. Under this assumption, the low genomic G+C content of endosymbiotic bacteria were considered in favour of the neutral theory. However, recently [36] have discussed that selective advantages favour high genomic AT-contents in intracellular genetic elements. Future understanding of translational decoding by the ribosome might explain the phylogeny specific codon usage bias and genome composition. It is pertinent to note that ribosome mediated gene regulation by co-translational protein folding has been demonstrated to be species specific in *E. coli* and *B. subtilis* [200].