APPENDICES

Appendix I

Data related to experiments

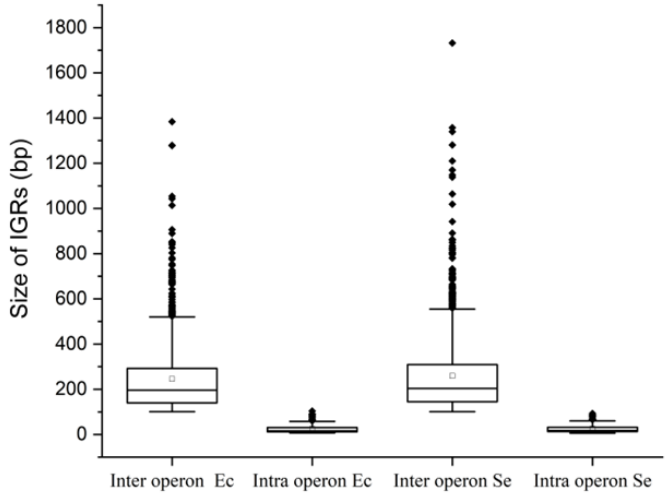
A.1. Estimation Of *ti*/*tv* Ratio by Accounting Degeneracy and Pretermination Nature of Codons

A.1.1. Supplementary materials related to Chapter 2 is available at GitHub link https://github.com/Pratyushbio/Appendices_Thesis_Pratyush_Chapter2/tree/b465a9518bdc71c0f1e87da2fd603296da3cd95a

A.2. **A Comparative Polymorphism Spectra Analysis in Inter-Operon IGRs And Intra-Operon IGRs**

A.2.1. The supplementary files related to Chapter 4 is available.

Supplementary Figure 1. Inter-operon IGRs are larger in size than the intra-operon IGRs in *E. coli* and *S. enterica***.**



Legend. Box plot analysis of the size of intra-operon IGRs and inter-operon IGRs the y-axis shows the size of IGRs in numbers. Ec- *E. coli*, Se- *S. enterica.* In this study total 1120 number of inter-operon IGRs were considered in *E. coli*. The minimum and maximum sizes of inter-operon IGRs were 101 bp and 1384 bp respectively. Total 1150 numbers of inter-operon IGRs were considered in *S. enterica*. The minimum and maximum sizes of inter-operon IGRs were 101 bp and 1732 bp respectively. Total 134 intra-operon IGRs were studied in *E. coli* with sizes ranging from 7 to104 bp. Between *hdeA* and *hdeB* genes, the size of intra-operon IGRs was 104 bp in *E. coli*. Total 89 intra-operon IGRs were studied in *S. enterica* with the sizes ranging from 6 to 93 bp. Between *degQ* and *degS* genes, the size of intra-operon IGRs was 89 bp in *S. enterica*. For inter-operon IGRs we excluded 35 bp upstream of start codon and 35bp downstream of stop codon. Suppose we found 570 bp inter-operon IGRs, we did not consider 35 bp at either end (70

bp total) for further steps, hence only 500 bp was considered as inter-operon IGRs for that region. This 35 bp at either end was present near adjacent genes.

Supplementary table 1. *E. coli* intra operon IGRs list with coordinate details of adjacent genes, size and GC % of IGRs and biological functional information on individual IGRs.

Sl no.	Intra operon	Location		Size	GC%	Information
		Gene-1	Gene-2			
1	*hdeAB*	3655966..3656304	3656408..3656740	104	37.5	Acid stress chaperone unit
2	*tnaAB*	3888730..3890145	3890236..3891483	91	39.33	Unit of trp operon
3	*degQS*	3380743..3382110	3382200..3383267	90	51.14	serine endoprotease
4	*dnaKJ*	12163..14079	14090..14167	87	47.13	Chaperone
5	*cadBA*	4356470..4358617	4358697..4360031	80	42.5	Lysine metabolism
6	*rpoBC*	4181245..4185273	4185350..4189573	77	54.55	RNA polymerase subunits
7	*guaBA*	2630958..2632535	2632604..2634070	69	44.93	nucleotide metabolism
8	*lacYA*	361249..361860	361926..363179	64	34.92	lactose metabolism
9	*manXY*	1902048..1903019	1903082..1903882	63	46.67	mannose metabolism
10	*nagBA*	701603..702751	702811..703611	60	50	Glucosamine metabolism
11	*nhaAR*	17489..18655	18721..19620	60	44.64	Sodium antiporter
12	*xapAB*	2522729..2523985	2524045..2524878	58	44.64	nucleotide metabolism
13	*creCD*	4636696..4638120	4638178..4639530	56	48.08	Histidine kinase regulation
14	*deoBD*	4619603..4620826	4620883..4621602	56	63.27	nucleotide metabolism
15	*lacZY*	361926..363179	363231..366305	52	38	lactose metabolism
16	*atpAG*	3917402..3918265	3918316..3919857	51	48.98	ATP synthase
17	*deoAB*	4618229..4619551	4619603..4620826	50	54.35	nucleotide metabolism
18	*fhuAC*	167484..169727	169778..170575	49	36.96	iron metabolism
19	*livKH*	3595477..3596403	3596451..3597560	46	41.3	Amino acid ABC transporter
20	*hicAB*	1509286..1509462	1509487..1509924	46	57.78	toxin-antitoxin

21	*fliAZ*	2000473..2001024	2001070..2001789	46	38.64	flagella biosynthesis
22	*nrfAB*	4287764..4289200	4289245..4289811	45	59.09	cytochrome metabolism
23	*groSL*	4370688..4370981	4371025..4372671	44	39.53	cochaperonin
24	*eutKR*	2554130..2555183	2555228..2555729	44	46.55	ethanolamine utilization protein
25	*ptsHI*	2533764..2534021	2534066..2535793	43	43.14	phosphocarrier protein
26	*nuoFG*	2397439..2400165	2400218..2401555	41	35.9	NADH dehydrogenase subunit
27	*csgBA*	1103951..1104406	1104447..1104902	39	50	curlin subunit
28	*prpDE*	351215..352666	352706..354592	38	27.03	methylcitrate metabolsim
29	*agaBC*	*3284170..3284646*	*3284685..3285488*	37	39.39	N-acetylgalactosamine transporter
30	*rpsN/rpsH*	3446153..3446545	3446579..3446884	34	41.18	ribosomal units
31	*rpsJ/rplC*	3452297..3452926	3452959..3453270	33	42.42	ribosomal units
32	*prpCD*	*350012..351181*	*351215..352666*	32	40.63	methylcitrate metabolsim
33	*tdcDE*	3260124..3262418	3262452..3263660	32	25	propionic acid metabolism
34	*fucPI*	2934235..2935551	2935584..2937359	31	32.26	L-fucose metabolism
35	*caiTA*	39244..40386	40417..41931	31	34.38	gamma-butyrobetaine antiporter
36	*marAB*	1619574..1619957	1619989..1620207	30	43.33	transcriptional regulators
37	*atoEB*	2324756..2326078	2326109..2327293	29	44.83	fatty acid metabolism
38	*aceBA*	4215478..4217079	4217109..4218413	28	46.43	TCA cycle
39	*fucAO*	2931865..2933013	2933041..2933688	28	39.29	L-fucose fermentation
40	*lsrBF*	1605051..1606073	1606100..1606975	27	37.04	Autoinducer-2 ABC transporter
41	*cmtBA*	3077471..3078859	3078887..3079330	26	34.62	mannitol metabolism
42	*atpGD*	3915993..3917375	3917402..3918265	25	34.62	ATP synthase
43	*ycbKL*	*1112718..1113267*	*1113291-1113940*	24	62.5	periplasmic protein
44	*tdcCD*	3262452..3263660	3263686..3265017	24	45.83	Thr metabolism
45	*gcvTH*	3049160..3049549	3049573..3050667	24	25	Gly metabolism
46	*flgDE*	1131854..1132549	1132574..1133782	23	47.83	flagella biosynthesis
47	*minCD*	1224549..1225361	1225385..1226080	23	30.43	cell division protein
48	*yphED*	1458264..1459052	1459099..1460481	23	43.48	ABC transporter permease
49	*kdpAB*	724988..727036	727059..728732	23	34.78	potassium-transporting ATPase
50	*moaAB*	817044..818033	818055..818567	22	40.91	molybdenum biosynthesis
51	*hemDX*	3987885..3989066	3989088..3989828	22	40.91	uroporphyrinogen biosynthesis

52	*tdcBC*	3263686..3265017	3265039..3266028	22	31.82	Thr metabolism
53	*yphFE*	2676850..2678361	2678384..2679367	21	61.9	ABC transporter
54	*rpmBG*	3811250..3811417	3811438..3811674	21	28.57	50S ribosomal protein
55	*cyoAB*	448650..450641	450663..451610	20	45	cytochrome metabolism
56	*cheBY*	1967048..1967437	1967452..1968501	20	40	chemotaxis
57	*marRA*	1619120..1619554	1619574..1619957	20	35	transcriptional regulators
58	*clpSA*	922913..923233	923264..925540	19	52.63	ATP-dependent protease
59	*eutBC*	2556410..2557297	2557318..2558679	19	47.37	ethanolamine biosynthesis
60	*atpDC*	3915553..3915972	3915993..3917375	19	52.63	ATP synthase
61	*flgEF*	1132574..1133782	1133802..1134557	18	61.11	flagella biosynthesis
62	*carAB*	29651..30799	30817..34038	18	44.44	carbamoyl-phosphate synthetase
63	*hycDE*	2844762..2846471	2846489..2847412	18	33.33	formate hydrogenlyase
64	*hyaBC*	1033254..1035047	1035066..1035773	17	47.06	hydrogenase metabolism
65	*fliIJ*	2016554..2017927	2017946..2018389	17	58.82	flagella biosynthesis
66	*emrAB*	2811427..2812599	2812616..2814154	17	58.82	MDR protein
67	*feoAB*	3540163..3540390	3540407..3542728	17	35.29	iron metabolism
68	*cusFB*	597131..597463	597479..598702	16	43.75	Cu/Ag export
69	*artPI*	902257..902988	903006..903734	16	37.5	Arg metabolism
70	*mglAC*	2236743..2237753	2237769..2239289	16	56.25	carbohydrate metabolism
71	*rplW/rplB*	3450543..3451364	3451382..3451684	16	43.75	ribosomal units
72	*rplV/rpsC*	3449182..3449883	3449901..3450233	16	50	ribosomal units
73	*cydAB*	771458..773026	773042..774181	16	50	cytochrome
74	*speED*	134788..135582	135598..136464	16	31.25	spermidine synthase
75	*rcsDB*	2313488..2316160	2316177..2316827	15	40	phosphorelay system
76	*rplB/rpsS*	3450248..3450526	3450543..3451364	15	40	ribosomal units
77	*fabHD*	1148759..1149712	1149728..1150657	14	21.43	oxoacyl biosynthesis
78	*nuoAB*	2403951..2404613	2404629..2405072	14	57.14	NADH dehydrogenase
79	*lepAB*	2704335..2705309	2705325..2707124	14	57.14	30s ribosomoal biosynthesis
80	*rplM/rpsI*	3377815..3378207	3378223..3378651	14	50	ribosomal units
81	*potAB*	1183617..1184474	1184458..1185594	14	35.71	ABC transporter
82	*fixAB*	42403..43173	43188..44129	13	53.85	electron transfer flavoprotein

83	*citCD*	650487..650783	650798..651856	13	30.77	citrate lyase synthetase
84	*sucAB*	758706..761507	761522..762739	13	30.77	2-oxoglutarate dehydrogenase
85	*oppBC*	1302899..1303819	1303834..1304742	13	61.54	oligopeptide transporter
86	*araGH*	1982554..1983540	1983555..1985069	13	38.46	arabisone operon
87	*nuoHI*	2395908..2396450	2396465..2397442	13	53.85	NADH dehydrogenase
88	*rplR/rpsE*	3444726..3445229	3445244..3445597	13	46.15	ribosomal units
89	*rpsS/rplV*	3449901..3450233	3450248..3450526	13	53.85	ribosomal units
90	*rplx/rplE*	3446899..3447438	3447453..3447767	13	46.15	ribosomal units
91	*rplE/rpsN*	3446579..3446884	3446899..3447438	13	38.46	ribosomal units
92	*atpFH*	3919870..3920403	3920418..3920888	13	76.92	ATP synthase
93	*fdoGH*	4081857..4082759	4082772..4085186	13	46.15	formate dehydrogenase
94	*malFG*	4242626..4243516	4243531..4245075	13	30.77	maltose metabolism
95	*pyrBI*	4470986..4471447	4471460..4472395	13	46.15	aspartate biosynthesis
96	*tauAB*	385232..386194	386207..386974	13	61.54	Taurine metabolism
97	*ubiCA*	4252506..4253003	4253016..4253888	13	23.08	chorismate pyruvate lyase
98	*betBA*	325577..327247	327261..328733	12	41.67	choline, oxygen, and osmotic stress
99	*entEB*	626070..627680	627694..628551	12	50	2,3-dihydroxybenzoate-AMP ligase
100	*gabDT*	2791273..2792721	2792735..2794015	12	58.33	succinate-semialdehyde dehydrogenase
101	*cusBA*	597479..598702	598714..601857	12	41.67	Cu/Ag export
102	*paaAB*	1453927..1454856	1454868..1455155	12	33.33	monooxygenase subunit
103	*fabDG*	1149728..1150657	1150670..1151404	11	27.27	S-malonyltransferase
104	*manYZ*	1903082..1903882	1903895..1904746	11	63.64	mannose metabolism
105	*rpsC/rplP*	3448759..3449169	3449182..3449883	11	72.73	ribosomal units
106	*rpsH/rplF*	3445607..3446140	3446153..3446545	11	45.45	ribosomal units
107	*atpHA*	3918316..3919857	3919870..3920403	11	72.73	ATP synthase
108	*bcsAB*	3690268..3692571	3692618..3695236	11	45.45	cellulose synthase
109	*araBA*	66835..68337	68348..70048	11	45.45	arabisone operon
110	*frdCD*	4379007..4379366	4379377..4379772	11	63.64	fumarate reductase
111	*citEF*	648039..649571	649582..650490	11	36.36	citrate metabolism
112	*proBA*	260388..261491	261503..262756	10	70	gamma-glutamyl kinase
113	*cyoDE*	446815..447705	447717..448046	10	70	cytochrome metabolism

114	*trpCB*	1317222..1318415	1318427..1319785	10	40	trp operon
115	*flgCD*	1131438..1131842	1131854..1132549	10	50	flagella biosynthesis
116	*oppCD*	1303834..1304742	1304754..1305767	10	20	oligopeptide transport system
117	*nuoIJ*	2395342..2395896	2395908..2396450	10	70	NADH dehydrogenase
118	*acrEF*	3413864..3415021	3415033..3418137	10	40	acriflavin resistance protein
119	*glnLG*	4053869..4055278	4055290..4056339	10	50	nitrogen regulation
120	*malPQ*	3547986..3550070	3550080..3552473	10	60	maltodextrin phosphorylase
121	*glyQS*	3722328..3724397	3724407..3725318	10	70	glycyl-tRNA synthetase
122	*dadAX*	1237571..1238869	1238879..1239949	10	50	amino acid metabolism
123	*fadBA*	4027609..4028772	4028782..4030971	10	50	fatty acid metabolism
124	*codBA*	354922..356181	356171..357454	10	40	cytosine metabolism
125	*leuCD*	78848..79453	79464..80864	9	54.55	isopropylmalate dehydratase subunit
126	*flgHI*	1135564..1136262	1136274..1137371	9	66.67	flagella biosynthesis
127	*cheYZ*	1966393..1967037	1967048..1967437	9	55.56	chemotaxis
128	*srlEB*	2826392..2827351	2827362..2827733	9	44.44	glucitol/sorbitol-specific
129	*accBC*	3405436..3405906	3405917..3407266	9	66.67	acetyl-CoA carboxylase carboxyl carrier
130	*rplC/rplD*	3451681..3452286	3452297..3452926	9	44.44	ribosomal units
131	*rplN/rplX*	3447453..3447767	3447778..3448149	9	55.56	ribosomal units
132	*entCE*	624885..626060	626070..627680	8	75	isochorismate synthase
133	*rplF/rplR*	3445244..3445597	3445607..3446140	8	37.5	ribosomal units
134	*nagAC*	700374..701594	701603..702751	7	28.57	N-acetylglucosamine biosynthesis

Supplementary table 2. *S. enterica* intra operon IGRs list with coordinate details of adjacent genes, size and GC % of IGRs and biological functional information on individual IGRs.

Sl no.	Intra operon	Location		Size	GC%	Information
		Gene-1	Gene-2			
1	*degQS*	3538553..3539920	3540013..3541083	93	51.61	serine endoprotease
2	*dnaKJ*	13520..13594	11593..13509	86	51.16	Chaperone
3	*cadBA*	2722916..2724247	2724330..2726474	83	53.01	Lysine metabolism
4	*rpoBC*	4388659..4392687	4392764..4396987	77	53.25	RNA polymerase subunits
5	*guaBA*	2647184..2648761	2648831..2650297	70	45.71	nucleotide metabolism
6	*nhaAR*	46190..47356	47424..48317	68	51.47	Sodium antiporter
7	*tdcCD*	3431184..3432392	3432458..3433789	66	66.67	Thr metabolism
8	*mreBC*	3565538..3566590	3566655..3567698	65	53.85	dynamic cytoskeletal protein
9	*livKH*	3754280..3755206	3755266..3756375	60	53.33	Amino acid ABC transporter
10	*fliAZ*	2027647..2028198	2028257..2028976	59	54.24	flagella biosynthesis
11	*creCD*	4866890..4868314	4868372..4869721	58	48.28	Histidine kinase regulation
12	*xapAB*	2557213..2558469	2558524..2559357	55	45.45	nucleotide metabolism
13	*deoAB*	4844302..4845624	4845676..4846899	52	32.69	nucleotide metabolism
14	*atpAG*	439064..441352	441364..441828	51	60.78	ATP synthase
15	*fhuAC*	223732..225921	225970..226767	49	53.06	iron metabolism
16	*groSL*	4596762..4597055	4597099..4598745	44	40.91	cochaperonin
17	*csgBA*	1225125..1225520	1225525..1226175	42	40.48	curlin subunit
18	*prpCD*	459491..460660	460700..462151	40	35	methylcitrate metabolsim
19	*caiTA*	85510..86652	86687..88204	35	34.29	gamma-butyrobetaine antiporter
20	*rpsN/rpsH*	3612512..3612904	3612938..3613243	34	41.18	ribosomal units
21	*tdcDE*	3428856..3431150	3431184..3432392	34	35.29	propionic acid metabolism
22	*rpsJ/rplC*	3618655..3619284	3619317..3619628	33	45.45	ribosomal units

23	*aceBA*	4424748..4426349	4426381..4427685	32	40.63	TCA cycle
24	*citCD*	68712..69680	69710..70003	30	36.67	citrate lyase synthetase
25	*lsrBF*	4309855..4310877	4310906..4311781	29	34.48	Autoinducer-2 ABC transporter
26	*atpGD*	4096987..4098369	4098396..4099259	27	37.04	ATP synthase
27	*flgDE*	1254883..1255581	1255608..1256819	27	48.15	flagella biosynthesis
28	*gcvTH*	3206432..3206821	3206847..3207941	26	26.92	Gly metabolism
29	*fliIJ*	2043407..2044777	2044799..2045242	22	54.55	flagella biosynthesis
30	*moaAB*	909257..910279	910301..910813	22	31.82	molybdenum biosynthesis
31	*flgEF*	1255608..1256819	1256840..1257595	21	57.14	flagella biosynthesis
32	*kdpAB*	809098..811146	811167..812846	21	42.86	potassium-transporting ATPase
33	*rpmBG*	3944915..3945082	3945103..3945339	21	28.57	50S ribosomal protein
34	*speED*	194195..194989	195010..195870	21	33.33	spermidine synthase
35	*carAB*	75881..77029	77048..80275	19	42.11	carbamoyl-phosphate synthetase
36	*feoAB*	3686962..3687189	3687208..3689526	19	26.32	iron metabolism
37	*artPI*	998942..999673	999691..1000419	18	44.44	Arg metabolism
38	*dsdXA*	4025973..4027310	4027328..4028650	18	38.89	D-serine metabolism
39	*hycDE*	2986879..2988588	2988606..2989529	18	33.33	formate hydrogenlyase
40	*rplV/rpsC*	3615540..3616241	3616259..3616591	18	50	ribosomal units
41	*rplW/rplB*	3616901..3617722	3617740..3618042	18	44.44	ribosomal units
42	*emrAB*	2954763..2955935	2955952..2957490	17	58.82	MDR protein
43	*fucAO*	3051020..3051796	3051801..3052439	17	58.82	L-fucose fermentation
44	*lepAB*	2750161..2751135	2751152..2752951	17	52.94	30s ribosomoal biosynthesis
45	*rcsDB*	2390974..2393643	2393660..2394310	17	52.94	phosphorelay system
46	*rplB/rpsS*	3616606..3616884	3616901..3617722	17	35.29	ribosomal units
47	*cydAB*	848339..849907	849923..851062	16	50	cytochrome
48	*fabHD*	1273112..1274065	1274081..1275010	16	25	oxoacyl biosynthesis
49	*mglAC*	2307594..2308604	2308620..2310140	16	56.25	carbohydrate metabolism
50	*potAB*	1306983..1307846	1307830..1308966	16	31.25	ABC transporter
51	*aceEF*	176242..178905	178920..180809	15	26.67	pyruvate dehydrogenase
52	*atpFH*	4100864..4101397	4101412..4101882	15	73.33	ATP synthase
53	*malFG*	4469458..4470348	4470363..4471907	15	33.33	maltose metabolism

54	*nuoHI*	2454029..2454571	2454586..2455563	15	53.33	NADH dehydrogenase
55	*oppBC*	1830158..1831066	1831081..1832001	15	60	oligopeptide transporter
56	*rplR/rpsE*	3611085..3611588	3611603..3611956	15	46.67	ribosomal units
57	*sucAB*	840555..843356	843371..844579	15	13.33	2-oxoglutarate dehydrogenase
58	*dadAX*	1894120..1895190	1895204..1896502	14	28.57	amino acid metabolism
59	*flgFG*	1256840..1257595	1257609..1258391	14	28.57	flagella biosynthesis
60	*ubiCA*	4477409..4477906	4477920..4478792	14	35.71	chorismate pyruvate lyase
61	*atpHA*	4099310..4100851	4100864..4101397	13	69.23	ATP synthase
62	*eutKR*	2588826..2589320	2589333..2589992	13	69.23	ethanolamine utilization protein
63	*fabDG*	1274081..1275010	1275023..1275757	13	38.46	S-malonyltransferase
64	*fdoGH*	4266598..4267500	4267513..4269927	13	61.54	formate dehydrogenase
65	*manYZ*	673126..673980	673983..674726	13	53.85	mannose metabolism
66	*pyrBI*	4725486..4725947	4725960..4726895	13	46.15	aspartate biosynthesis
67	*rpsC/rplP*	3615117..3615527	3615540..3616241	13	69.23	ribosomal units
68	*acrEF*	3584243..3585400	3585412..3588525	12	50	acriflavin resistance protein
69	*cyoDE*	533857..534747	534759..535088	12	66.67	cytochrome metabolism
70	*flgCD*	1254467..1254871	1254883..1255581	12	66.67	flagella biosynthesis
71	*flgHI*	1258479..1259144	1259156..1260253	12	45.45	flagella biosynthesis
72	*oppCD*	1829139..1830146	1830158..1831066	12	58.33	oligopeptide transport system
73	*proBA*	366286..367389	367401..368651	12	50	gamma-glutamyl kinase
74	*accBC*	3573234..3573704	3573715..3575064	11	63.64	acetyl-CoA carboxyl carrier
75	*araBA*	118078..119580	119591..121300	11	54.55	arabisone operon
76	*citEF*	722639..724168	724178..725086	11	54.55	citrate metabolism
77	*frdCD*	4604467..4604826	4604837..4605232	11	72.73	fumarate reductase
78	*leuCD*	129058..129663	129674..131074	11	45.45	isopropylmalate dehydratase subunit
79	*nuoIJ*	2453464..2454018	2454029..2454571	11	63.64	NADH dehydrogenase
80	*rplC/rplD*	3618039..3618644	3618655..3619284	11	54.55	ribosomal units
81	*entCE*	695065..696240	696250..697860	10	50	isochorismate synthase
82	*fadBA*	4211011..4212174	4212184..4214373	10	50	fatty acid metabolism
83	*glyQS*	3862647..3864716	3864726..3865637	10	80	glycyl-tRNA synthetase
84	*malPQ*	3694897..3696975	3696985..3699378	10	80	maltodextrin phosphorylase

85	*rplF/rplR*	3611603..3611956	3611966..3612499	10	40	ribosomal units
86	*trpCB*	1814226..1815584	1815594..1816787	10	30	trp operon
87	*glnLG*	4235612..4237021	4237030..4238079	9	33.33	nitrogen regulation
88	*ruvAB*	1972565..1973575	1973584..1974195	9	44.44	Holliday junction subunit
89	*sspAB*	3533183..3533683	3533689..3534327	6	50	starvation protein

Supplementary table 3. Methodology for calculation of base substitutions at intra operon IGRs.

Strains	Pos-1	Pos-2	Pos-3	Pos-4	Pos-5	Pos-6	Pos-7	Pos-8	Pos-9	Pos-10
Strain-1	G	T	G	T	C	G	A	A	T	A
Strain-2	G	T	G	T	C	G	A	A	T	A
Strain-3	G	T	G	T	C	G	A	A	T	A
Strain-4	G	T	G	T	C	G	A	A	T	A
Strain-5	G	T	G	T	C	G	A	A	T	A
Strain-6	G	T	G	T	T	G	A	A	T	A
Strain-7	G	T	G	T	C	G	A	A	T	A
Strain-8	G	T	G	T	C	A	A	A	T	A
Strain-9	G	T	T	T	T	A	A	A	T	A
Strain-10	G	T	A	T	C	G	A	A	T	A
Nucleotide count										
A count	0	0	1	0	0	2	10	10	0	10
T count	0	10	1	10	2	0	0	0	10	0
C count	0	0	0	0	8	0	0	0	0	0
G count	10	0	8	0	0	8	0	0	0	0
Reference sequence	G	T	G	T	C	G	A	A	T	A
Mutation			G>T, G->A		C>T	G->A				

Polymorphism spectra	A->T	A->C	A->G	T->A	T->C	T->G	C->A	C->T	C->G	G->A	G->T	G->C	TOTAL
Numbers	0	0	0	0	0	0	0	1	0	2	1	0	1
Normalization	0	0	0	0	0	0	0	1	0	0.67	0.33	0	

A hypothetical intra operon IGRs having 10bp in size is presented here to show the procedure of reference sequence derivation and calculation of mutations in 10 strains, which was followed in our dataset for the work. The nucleotide count sub-section shows the nucleotide counts of individual columns/positions. At position 1 the most frequent nucleotide was found to be G, hence in reference sequence row, G was mentioned under column/position 1. Similarly, the remaining positions were calculated. At position 3, the frequency of G was found to be 8 and frequency of T and A were found to be 1 each, A and C counts were not found. Hence under position-3, reference sequence row G was mentioned. But the mutation row was mentioned as G→T, G→A indicating mutations from the most frequent G to the least frequents (T and A) in that position. At position 5, the most frequent nucleotide was found to be C followed by T, having nucleotide counts of 8 and 2 respectively. Similarly at position 6, the most frequent nucleotide was found to be G followed by A. So, position 5 and 6 were indicating C→T and G→A mutations respectively. Hence it can be said that for this certain scenario we found 4 mutations out of 10 positions. Further this mutation data was considered for the construction of substitution spectra. This methodology was followed for the derivation of reference sequence in our sample intra operon IGRs. Further, in the polymorphism spectra analysis the total numbers of base substitutions can be found along with the normalization value. As described in the Materials and methods section the normalization was done by considering the number of certain base substitution and nucleotide count of the originating nucleotide. Here C→T was found to be 1 in number and C nucleotide was found to be 1. Hence the normalization was done by considering the number of C→T substitution and dividing it with the number of C. Hence 1/1 =1, that was mentioned in normalization row against C→T substitution column. The remaining substitutions were also calculated by this procedure for the entire dataset.

Supplementary table 4. Collection of intra operon IGRs having base substitutions to find out total spectra in *E. coli*.

Operons	A→T	A→C	A→G	T→A	T→C	T→G	C→A	C→T	C→G	G→A	G→T	G→C	Total
tnaAB	3	1	0	3	1	0	1	0	0	0	0	0	9
	0.103	0.034	0	0.12	0.04	0	0.045	0	0	0	0	0	
degQS	0	0	0	0	1	1	0	1	0	0	0	0	3
	0	0	0	0	0.034	0.034	0	0.043	0	0	0	0	
dnaKJ	2	0	2	1	1	0	0	0	0	0	0	0	6
	0.08	0	0.08	0.048	0.048	0	0	0	0	0	0	0	
cadBA	0	0	0	0	0	0	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	0.056	0	0	0	0	

rpoBC	0	0	0	0	0	1	0	0	0	0	0	0	1
	0	0	0	0	0	0.063	0	0	0	0	0	0	
manXY	1	0	0	1	0	0	0	1	1	0	0	0	4
	0.05	0	0	0.048	0	0	0	0.091	0.091	0	0	0	
lacYA	1	1	0	0	3	0	0	0	1	1	1	0	8
	0.083	0.083	0	0	0.158	0	0	0	0.067	0.056	0.056	0	
nagBA	2	0	0	2	0	0	0	2	0	0	0	0	6
	0.1	0	0	0.167	0	0	0	0.143	0	0	0	0	
nhaAR	0	0	0	0	0	1	0	2	0	1	0	0	4
	0	0	0	0	0	0.071	0	0.154	0	0.059	0	0	
creCD	0	0	0	1	0	0	0	2	0	0	0	0	3
	0	0	0	0.067	0	0	0	0.118	0	0	0	0	
deoBD	0	0	0	1	0	1	0	1	0	2	0	0	5
	0	0	0	0.067	0	0.067	0	0.1	0	0.133	0	0	
lacZY	0	0	0	0	1	1	1	0	0	2	0	0	5
	0	0	0	0	0.071	0.071	0.083	0	0	0.154	0	0	
atpAG	0	0	0	1	0	0	0	0	0	0	0	0	1
	0	0	0	0.111	0	0	0	0	0	0	0	0	
deoAB	0	0	0	0	1	0	0	0	0	0	0	0	1
	0	0	0	0	0.077	0	0	0	0	0	0	0	
fhuAC	0	0	1	1	2	0	0	3	0	0	0	0	7
	0	0	0.167	0.053	0.105	0	0	0.25	0	0	0	0	
livKH	1	0	0	0	1	0	0	1	0	4	1	0	8
	0.1	0	0	0	0.091	0	0	0.083	0	0.308	0.077	0	
hicAB	0	0	2	0	1	0	0	0	0	0	0	0	3
	0	0	0.105	0	0.1	0	0	0	0	0	0	0	
fliAZ	0	0	0	0	0	0	0	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0	0.091	0	0	
nrfAB	0	0	0	0	1	0	1	0	0	1	0	0	3
	0	0	0	0	0.167	0	0.077	0	0	0.077	0	0	

groSL	0	0	0	0	0	0	0	0	0	0	1	0	1
	0	0	0	0	0	0	0	0	0	0	0.125	0	
ptsHI	0	0	0	0	1	0	1	0	0	0	0	0	2
	0	0	0	0	0.071	0	0.143	0	0	0	0	0	
xapAB	0	0	1	2	0	0	1	2	0	1	0	0	7
	0	0	0.143	0.083	0	0	0.063	0.125	0	0.091	0	0	
nuoFG	0	0	0	2	1	0	0	0	0	0	0	0	3
	0	0	0	0.154	0.063	0	0	0	0	0	0	0	
csgBA	0	0	0	1	0	0	1	1	0	0	0	0	3
	0	0	0	0.063	0	0	0.143	0.143	0	0	0	0	
prpDE	0	0	0	0	1	0	0	1	0	1	1	0	4
	0	0	0	0	0.167	0	0	0.167	0	0.077	0.077	0	
agaBC	0	0	1	0	0	0	1	0	0	0	0	0	2
	0	0	0.071	0	0	0	0.25	0	0	0	0	0	
prpCD	0	0	0	0	0	1	0	0	0	1	0	0	2
	0	0	0	0	0	0.2	0	0	0	0.25	0	0	
tdcDE	0	0	1	0	1	0	0	0	0	0	0	0	2
	0	0	0.066	0	0.111	0	0	0	0	0	0	0	
fucPI	0	0	0	1	0	0	0	0	0	0	0	0	1
	0	0	0	0.1	0	0	0	0	0	0	0	0	
caiTA	0	0	0	0	0	1	0	0	0	0	2	0	3
	0	0	0	0	0	0.071	0	0	0	0	0.333	0	
marAB	1	0	0	0	0	0	0	0	0	0	1	0	2
	0.091	0	0	0	0	0	0	0	0	0	0.143	0	
atoEB	0	0	1	1	0	0	1	2	0	0	0	0	5
	0	0	0.077	0.333	0	0	0.125	0.25	0	0	0	0	
lsrBF	0	1	0	0	0	0	0	0	0	1	0	0	2
	0	0.125	0	0	0	0	0	0	0	0.167	0	0	
cmtBA	0	1	1	0	0	0	0	0	0	1	0	0	3
	0	0.1	0.1	0	0	0	0	0	0	0.143	0	0	

marRA	0	0	0	0	0	0	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	0.5	0	0	0	0	
eutBC	0	0	0	0	0	0	0	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0	0.2	0	0	
atpDC	0	0	0	0	0	0	1	0	0	0	0	0	1
	0	0	0	0	0	0	0.142	0	0	0	0	0	
fliIJ	0	0	0	0	0	0	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	0.2	0	0	0	0	
cusFB	0	0	1	0	0	0	0	1	0	0	1	0	3
	0	0	0.2	0	0	0	0	0.333	0	0	0.25	0	
mglAC	0	0	0	0	0	0	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	0.2	0	0	0	0	
hdeAB	0	1	0	0	1	0	1	0	0	2	1	0	6
	0	0.027	0	0	0.036	0	0.063	0	0	0.087	0.043	0	
araGH	1	0	0	0	0	0	0	0	0	0	0	0	1
	0.333	0	0	0	0	0	0	0	0	0	0	0	
pyrBI	0	0	0	0	1	0	0	0	0	0	0	0	1
	0	0	0	0	0.166	0	0	0	0	0	0	0	
tauAB	0	0	1	0	0	0	0	0	0	0	0	0	1
	0	0	0.2	0	0	0	0	0	0	0	0	0	
gabDT	0	0	0	0	1	0	0	0	0	0	0	0	1
	0	0	0	0	0.5	0	0	0	0	0	0	0	
cusBA	0	0	1	0	0	0	0	0	0	0	0	0	1
	0	0	0.166	0	0	0	0	0	0	0	0	0	
fabDG	0	1	0	0	0	0	0	0	0	0	0	0	1
	0	0.142	0	0	0	0	0	0	0	0	0	0	
manYZ	0	0	0	0	0	0	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	0.5	0	0	0	0	
araBA	0	0	1	0	0	0	0	0	0	0	0	0	1
	0	0	1	0	0	0	0	0	0	0	0	0	

proBA	1	0	0	0	1	0	0	0	0	0	0	0	2
	0.5	0	0	0	1	0	0	0	0	0	0	0	
flgCD	0	0	0	0	0	0	0	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0	0.25	0	0	
acrEF	0	0	0	0	1	0	0	0	0	0	0	0	1
	0	0	0	0	0.33	0	0	0	0	0	0	0	
malPQ	0	0	0	0	0	0	0	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0	0.5	0	0	
codBA	0	0	0	0	0	0	0	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0	0.33	0	0	
leuCD	0	0	0	0	0	0	0	0	0	0	0	1	1
	0	0	0	0	0	0	0	0	0	0	0	0.25	
flgHI	0	0	0	0	0	0	0	0	0	1	0	0	1
	0	0	0	0	0	0	0	0	0	0.25	0	0	
entCE	0	0	0	0	0	0	0	1	0	0	0	0	1
	0	0	0	0	0	0	0	1	0	0	0	0	
Total spectra	**15**	**5**	**20**	**16**	**17**	**7**	**11**	**31**	**1**	**19**	**8**	**2**	**152**
	0.014	**0.005**	**0.019**	**0.02**	**0.021**	**0.009**	**0.017**	**0.049**	**0.002**	**0.021**	**0.009**	**0.002**	

In case of *E. coli* out of 134 intra operon IGRs, we found substitutions in 57 IGRs. Finally, we considered the 57 IGRs for final spectra constructions. It was also considered for the *S. enterica* species.

Supplementary table 5. The total numbers of individual base substitution present in inter operon IGRs and intra operon IGRs present in *E. coli* and *S. enterica* are listed.

Substitutions	*Numbers of individual base substitutions*			
	Inter operon IGRs		Intra operon IGRs	
	E. coli	*S. enterica*	*E. coli*	*S. enterica*
A->T	1732	1932	15	2
A->C	1185	1827	5	0
A->G	3573	6348	20	20
T->A	1759	1853	16	10
T->C	3642	6359	17	10
T->G	1154	1805	7	3
C->A	1742	3398	11	7
C->T	5107	11117	31	56
C->G	666	1106	1	1
G->A	5103	11037	19	28
G->T	1726	3354	8	16
G->C	667	1060	2	2
Total	28056	51196	152	155

A.4. **A Comparative Synonymous Polymorphism Spectra Analysis in Co-transcribed Gene Pairs**

A.4.1. The supplementary files related to Chapter 5 is available.

Supplementary Table 1. Operons selected for the work, their coordinate and function details. (LeS- Leading strand, LaS- Lagging strand)

Operon	Genes	Start	End	Strand	Function
rpoBC	*rpoB*	4181245	4185273	LeS	DNA-directed RNA polymerase subunit beta
	rpoC	4185350	4189573	LeS	DNA-directed RNA polymerase subunit beta'
lacZY	*lacZ*	363231	366305	LeS	beta-D-galactosidase
	lacY	361926	363179	LeS	galactoside permease
kdpAB	*kdpA*	727059	728732	LeS	potassium-transporting ATPase subunit A
	kdpB	724988	727036	LeS	potassium-transporting ATPase subunit B
araBA	*araB*	68348	70048	LeS	ribulokinase
	araA	66835	68337	LeS	L-arabinose isomerase
bcsAB	*bcsA*	3692618	3695236	LaS	cellulose synthase catalytic subunit
	bcsB	3690268	3692607	LaS	cellulose synthase periplasmic subunit

Supplementary Table 2. Composition, skew, and synonymous sites detailed information of gene pairs.

	Nucleotide composition										Synonymous sites			
Gene	A	T	C	G	SIZE	GC%	AT skew	GC skew	RY skew	KM skew	A	T	C	G
rpoB	1016	891	1034	1088	4029	52.67	0.066	0.025	0.044	-0.018	354	834	834	826
rpoC	1010	938	1090	1186	4224	53.88	0.037	0.042	0.04	0.006	378	978	792	982
lacZ	677	665	845	888	3075	56.36	0.009	0.025	0.018	0.01	268	437	678	707
lacY	239	433	284	298	1254	46.41	-0.289	0.024	-0.144	0.166	127	248	233	284
kdpA	293	451	425	505	1674	55.56	-0.212	0.086	-0.047	0.142	126	271	354	509
kdpB	433	465	533	618	2049	56.17	-0.036	0.074	0.026	0.057	173	334	485	562

araB	349	361	489	502	1701	58.26	-0.017	0.013	0.001	0.015	158	273	412	376
araA	340	328	416	419	1503	55.56	0.018	0.004	0.01	-0.006	79	227	370	358
bcsA	526	638	704	751	2619	55.56	-0.096	0.032	-0.025	0.061	186	368	519	815
bcsB	518	509	618	659	2304	55.43	0.009	0.032	0.022	0.014	191	319	427	716

AT skew was calculated by (A-T)/(A+T) and GC skew was calculated by (G-C)/(G+C). RY and KM skewes were also calculated to know the purine: pyrimidine and keto: amino inequality in genes. RY skew was calculated as [{(A+G) - (T+C)} / (A+T+C+G)] and KM skew was calculated as [{(T+G) -(A+C)}/ (A+T+C+G)]. GC skew was found to be positive for all the genes. but AT skew, RY/KM was found to be of positive and negative values, indicating the inequality in presence of keto/amino, purine/pyrimidine and A/T nucleotides in many co-transcribed pairs. These skew values clearly show the inequality in the presence of complementary bases even in adjacent genes present in the same strand. So, the compositional study indicates that the skew values in an adjacent pair of genes in an operon might not be necessarily identical.

Supplementary Table 3. Methodology for calculation of polymorphism.

Position	1	2	3	4	5	6	7	8	9
ST1	A	A	T	A	G	C	C	T	A
ST2	A	A	T	A	G	T	C	A	A
ST3	A	A	T	A	G	A	C	A	A
ST4	A	A	A	A	G	G	C	A	T
ST5	A	A	A	T	G	G	C	A	T
ST6	A	A	T	T	G	G	C	A	T

ST7	**A**	**A**	**T**	**T**	**G**	**G**	**C**	**A**	**A**
ST8	**A**	**T**	**G**	**T**	**G**	**G**	**C**	**A**	**A**
ST9	**A**	**T**	**G**	**T**	**G**	**G**	**T**	**A**	**T**
ST10	**A**	**T**	**G**	**T**	**G**	**G**	**T**	**A**	**T**
$Count_A$	**10**	**7**	**2**	**4**	**0**	**1**	**0**	**9**	**5**
$Count_T$	**0**	**3**	**5**	**6**	**0**	**1**	**2**	**1**	**5**
$Count_C$	**0**	**0**	**3**	**0**	**0**	**1**	**8**	**0**	**0**
$Count_G$	**0**	**0**	**0**	**0**	**10**	**7**	**0**	**0**	**0**
Reference Sequence	**A**	**A**	**T**	**T**	**G**	**G**	**C**	**A**	**?**
Mutation		**A-->T**	**T>C, T>A**	**T>A**		**G>A, G>T, G>C**	**C>T**	**A>T**	**N/A**

Hypothetical mutation scenario to explain the methodology that has been followed in this work. Strains containing ambiguous nucleotide 'N' were not considered for the study. In position 1 A is found 10 times hence in Reference Sequence its written as A. similarly in 2nd position A was found 7 times and T was found 3 times. Hence it can be said that the mutation was observed in A→T pattern. In the last position we found 5 numbers of A & T each, here we can't assume which one can be taken in reference sequence, so we had to keep in mind the during the work.

Supplementary Table 4. Spectra showing number of synonymous substitutions observed for each gene, before normalization.

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	TOTAL
rpoB	4	2	12	5	20	4	4	76	3	17	6	1	154
rpoC	5	2	12	9	20	8	4	63	1	29	9	3	165
lacZ	2	3	24	4	27	5	9	42	1	48	9	6	180
lacY	1	3	5	1	5	2	1	13	0	14	3	1	49
kdpA	1	1	15	7	30	5	7	42	6	42	6	2	164

kdpB	2	3	12	4	33	3	8	57	9	56	10	2	199
araB	2	1	10	6	19	3	3	46	5	42	3	2	142
araA	1	4	12	5	32	4	8	38	4	30	6	2	146
bcsA	1	2	12	5	26	4	4	57	6	45	3	6	171
bcsB	2	4	13	4	24	0	5	51	3	32	12	7	157

Supplementary Table 5. Correlation study between TFD/FFD & synonymous *ti*/*tv* of whole gene/FFD

Genes	TFD	FFD	TFD/FFD	*ti*/*tv* (synonymous)		whole gene-FFD	Pearson *r* value
				whole gene	FFD		
rpoB	516	410	1.259	4.310	1.192	3.118	**0.677**
rpoC	480	482	0.996	2.929	1.441	1.488	
lacZ	369	331	1.115	3.615	1.939	1.676	
lacY	138	131	1.053	3.083	2.428	0.655	
kdpA	133	223	0.596	3.686	3.091	0.595	
kdpB	180	278	0.647	3.854	2.867	0.987	
araB	182	219	0.831	4.917	3.733	1.184	
araA	190	169	1.124	3.171	1.724	1.447	
bcsA	279	275	1.015	4.344	3.400	0.944	
bcsB	253	272	0.930	3.361	2.040	1.321	

The correlation study between TFD:FFD ratio of all genes and synonymous *ti*/*tv* of whole gene-FFD was taken we got a strong positive Pearson r value as 0.677.

Supplementary Table 6. *ti*/*tv* ratio comparison of individual amino acids of four-fold degenerate codons between co-transcribed gene pairs.

ti/*tv*	*rpoB*	*rpoC*	*lacZ*	*lacY*	*kdpA*	*kdpB*	*araB*	*araA*	*bcsA*	*bcsB*
Val	1.00	0.60	1.57	1.50	6.00	2.29	2.75	1.43	2.50	1.67
Pro	2.00	2.67	2.00	2.00	2.40	N/A	7.00	0.40	3.00	1.86
Thr	1.67	1.38	2.80	2.00	3.00	2.00	4.00	4.25	2.75	1.75
Ala	0.56	1.00	1.27	4.00	2.00	2.17	6.00	1.10	13.00	1.00
Gly	1.75	6.50	3.25	3.00	4.75	4.17	2.40	3.33	2.75	N/A

ti/*tv* ratio comparison of four-fold degenerate amino acids individually in co-transcribed genes shows the difference in ratio even at amino acid level like the A in *araB* & *araA* is showing the difference, hence many such differences in ratio can be found at individual amino acid level. Genes showing NA values had zero transversions for the respective amino acids.

Supplementary Table 7. Codon count and polymorphism count of remaining four pairs of genes.

Amino Acids	Codons	Codon count		polymorphism		Codon count		polymorphism		Codon count		polymorphism		Codon count		polymorphism		Codon count		polymorphism	
		rpoB	*rpoC*	*rpoB*	*rpoC*	*lacZ*	*lacY*	*lacZ*	*lacY*	*kdpA*	*kdpB*	*kdpA*	*kdpB*	*araB*	*araA*	*araB*	*araA*	*bcsA*	*bcsB*	*bcsA*	*bcsB*
V	GUU	41	53	2	5	12	9	2	0	6	11	1	4	9	8	2	2	10	7	2	2
	GUC	13	7	2	1	19	4	3	0	11	19	4	6	5	10	2	6	13	10	3	3
	GUA	31	32	3	7	9	8	4	1	4	6	1	2	5	2	1	2	12	7	7	1
	GUG	24	17	3	3	24	8	9	4	28	24	8	11	20	16	10	7	29	37	2	10
P	CCU	9	4	1	1	11	0	4	0	8	3	6	2	4	3	0	4	2	4	0	1
	CCC	0	0	0	0	9	2	1	0	3	5	3	2	2	1	0	0	2	3	0	1

	CCA	9	8	2	2	6	4	2	3	5	2	4	0	5	3	2	3	7	12	0	4
	CCG	38	45	3	8	36	6	11	0	10	16	4	5	21	12	6	0	38	43	8	14
T	ACU	17	22	2	4	4	2	1	0	3	5	2	4	7	3	1	3	8	1	4	1
	ACC	34	47	8	12	25	7	5	1	13	23	5	6	16	22	7	11	19	25	7	5
	ACA	3	1	2	0	8	2	5	0	2	4	1	1	1	2	1	2	3	3	0	3
	ACG	6	7	1	3	20	8	8	2	12	9	4	4	1	6	1	5	13	12	4	2
A	GCU	19	28	1	7	8	6	4	1	6	10	3	3	9	4	3	3	6	6	2	2
	GCC	9	11	2	3	29	12	6	1	20	28	3	10	29	13	9	9	14	11	1	2
	GCA	22	33	5	5	12	4	4	1	8	13	5	8	19	2	4	2	7	10	0	5
	GCG	28	52	4	6	28	13	11	2	28	43	13	17	23	23	5	7	37	32	11	7
G	GGU	68	85	6	6	24	13	5	1	20	15	10	7	12	16	7	11	17	21	4	5
	GGC	35	29	4	9	34	18	6	4	23	35	7	20	24	21	7	1	23	22	8	6
	GGA	0	0	0	0	4	2	1	1	3	2	0	2	2	1	1	1	2	1	0	0
	GGG	3	1	1	0	9	3	5	2	10	5	6	2	5	1	2	0	13	5	3	2
L	CUU	6	3	0	2	9	5	1	1	8	3	2	0	8	2	1	0	8	4	5	0
	CUC	15	7	3	0	9	4	2	0	8	6	5	2	7	7	1	3	16	5	5	0
	CUA	0	0	0	0	6	3	4	1	2	2	1	1	1	0	0	0	3	6	3	2
	CUG	100	125	6	13	54	32	6	5	48	50	8	15	26	36	8	10	61	50	9	3
S	UCU	23	24	2	1	3	7	0	2	6	5	4	2	5	1	1	0	6	5	1	1
	UCC	31	27	9	2	6	3	0	1	4	5	0	2	7	8	2	0	5	5	4	2
	UCA	0	1	0	0	8	4	1	0	2	1	2	0	2	1	0	0	2	2	1	0
	UCG	3	5	0	0	9	7	0	2	5	7	1	3	3	1	2	0	20	13	3	6
R	CGU	61	75	6	5	19	5	3	0	6	16	3	3	7	11	1	2	24	15	5	1
	CGC	28	24	10	1	37	4	4	0	9	14	5	5	17	15	5	3	22	17	5	7
	CGA	1	0	0	0	3	0	2	0	0	1	0	1	1	8	0	1	2	1	0	1
	CGG	0	0	0	0	7	2	1	0	1	3	0	0	2	1	1	0	12	7	5	2
F	UUU	11	9	1	1	19	32	0	2	25	14	1	1	13	9	3	1	14	13	2	0
	UUC	33	26	2	2	19	24	3	1	10	9	1	0	6	14	0	3	32	21	5	2
Y	UAU	14	7	1	0	13	7	2	0	7	4	1	0	6	4	1	3	14	11	2	0

	UAC	29	27	5	4	18	7	1	0	2	2	1	1	3	7	0	1	16	8	5	1
H	CAU	1	4	0	0	19	4	1	0	4	3	0	1	7	10	0	4	15	6	1	0
	CAC	18	17	5	0	15	0	1	0	2	2	0	1	6	11	2	0	11	2	3	0
Q	CAA	8	2	1	0	15	4	2	1	10	6	3	0	12	8	0	2	10	12	1	0
	CAG	50	48	3	2	43	7	5	0	11	17	4	5	22	18	4	5	26	28	3	1
N	AAU	3	1	0	1	17	8	1	0	9	7	2	0	6	8	3	1	6	13	0	2
	AAC	48	47	4	4	29	8	5	2	15	17	1	4	12	14	0	0	23	32	5	8
K	AAA	56	62	3	3	15	10	1	0	5	27	0	0	16	18	1	3	25	20	2	0
	AAG	24	25	2	3	5	2	1	1	3	4	1	1	0	4	0	2	8	7	3	1
D	GAU	30	34	2	2	41	3	3	0	4	24	1	3	12	19	2	2	19	30	1	2
	GAC	62	47	8	7	23	3	3	0	5	9	1	2	18	14	7	5	12	21	6	7
E	GAA	89	83	2	3	47	9	2	0	11	19	0	2	18	16	3	1	24	16	1	2
	GAG	33	26	1	3	15	2	6	0	4	11	1	4	11	9	5	2	13	11	3	2
C	UGU	5	7	0	0	5	5	2	0	3	3	1	1	2	0	1	2	5	1	2	1
	UGC	2	8	1	3	11	3	1	0	3	2	0	0	12	7	6	2	6	1	0	0

The calculation of codon count, polymorphism and polymorphism frequency was calculated as the following manner. Suppose we observed three synonymous polymorphisms of UUU in *rpoB*, a value of 3 was entered in the polymorphism column against UUU for *rpoB*. Similarly, synonymous polymorphisms in UUU of *rpoC* was also mentioned in the adjacent box for comparison. Then polymorphism frequency was calculated by taking the number of synonymous polymorphisms observed in case of a codon and then dividing the value with the number of that codons present in the gene.

Supplementary Table 8. Polymorphism frequency comparison at codon specific level of all the genes

Amino acids	Codons	Polymorphism frequency in the gene pairs									
		rpoB	*rpoC*	*lacZ*	*lacY*	*kdpA*	*kdpB*	*araB*	*araA*	*bcsA*	*bcsB*

V	GUU	0.049	0.094	0.167	0.000	0.167	0.364	0.222	0.250	0.200	0.286
	GUC	0.154	0.143	0.211	0.000	0.364	0.316	0.400	0.600	0.231	0.300
	GUA	**0.097**	**0.219**	0.444	0.125	0.250	0.333	0.200	1.000	**0.583**	**0.143**
	GUG	0.125	0.176	0.375	0.500	**0.286**	**0.458**	0.500	0.438	**0.069**	**0.270**
P	CCU	0.111	0.250	0.364	0.000	0.750	0.667	0.000	1.333	0.000	0.250
	CCC	0.000	0.000	0.222	0.000	1.000	0.400	0.000	0.000	0.000	0.333
	CCA	0.222	0.250	0.333	0.750	0.800	0.000	0.400	1.000	0.000	0.333
	CCG	**0.079**	**0.178**	**0.306**	**0.000**	0.400	0.313	**0.286**	**0.000**	**0.211**	**0.326**
T	ACU	0.118	0.000	0.250	0.000	0.667	0.800	0.143	1.000	0.500	1.000
	ACC	0.235	0.255	0.200	0.143	0.385	0.261	0.438	0.500	0.368	0.200
	ACA	0.667	0.000	**0.625**	**0.000**	0.500	0.250	1.000	1.000	0.000	1.000
	ACG	0.167	0.429	0.400	0.250	0.333	0.444	**1.000**	**0.833**	0.308	0.167
A	GCU	**0.053**	**0.250**	0.500	0.167	0.500	0.300	**0.333**	**0.750**	0.333	0.333
	GCC	0.222	0.273	0.138	0.083	**0.150**	**0.357**	**0.310**	**0.692**	0.071	0.182
	GCA	0.227	0.152	0.333	0.250	0.625	0.615	**0.211**	**1.000**	0.000	0.500
	GCG	0.143	0.115	**0.393**	**0.154**	0.464	0.395	0.217	0.304	**0.297**	**0.219**
G	GGU	0.088	0.071	**0.208**	**0.077**	0.500	0.467	**0.583**	**0.688**	0.235	0.238
	GGC	**0.114**	**0.310**	0.176	0.222	**0.304**	**0.571**	**0.292**	**0.048**	0.348	0.273
	GGA	0.000	0.000	0.250	0.500	0.000	1.000	0.500	1.000	0.000	0.000
	GGG	0.333	0.000	0.556	0.667	**0.600**	**0.400**	0.400	0.000	0.231	0.400
L	CUU	0.000	0.667	0.111	0.200	0.250	0.000	0.125	0.000	**0.625**	**0.000**
	CUC	0.200	0.000	0.222	0.000	0.625	0.333	0.143	0.429	**0.313**	**0.000**
	CUA	0.000	0.000	0.667	0.000	0.500	0.500	0.000	0.000	1.000	0.333
	CUG	**0.060**	**0.104**	0.111	0.156	**0.167**	**0.300**	0.308	0.278	**0.148**	**0.060**
S	UCU	0.087	0.042	0.000	0.286	0.667	0.400	0.200	0.000	0.167	0.200
	UCC	**0.290**	**0.074**	0.000	0.333	0.000	0.400	0.286	0.000	0.800	0.400
	UCA	0.000	0.000	0.125	0.000	1.000	0.000	0.000	0.000	0.500	0.000
	UCG	0.000	0.000	0.000	0.286	0.200	0.429	0.667	0.000	**0.150**	**0.462**
R	CGU	0.098	0.067	0.158	0.000	0.500	0.188	0.143	0.182	0.208	0.067
	CGC	**0.357**	**0.042**	0.108	0.000	0.556	0.357	0.294	0.200	0.227	0.412
	CGA	0.000	0.000	0.667	0.000	0.000	1.000	0.000	0.125	0.000	1.000
	CGG	0.000	0.000	0.143	0.000	0.000	0.000	0.500	0.000	0.417	0.286

F	UUU	0.091	0.111	0.000	0.063	0.040	0.071	0.231	0.111	0.143	0.000
	UUC	0.061	0.077	0.158	0.042	0.100	0.000	0.000	0.214	**0.156**	**0.095**
Y	UAU	0.071	0.000	0.154	0.000	0.143	0.000	0.167	0.750	0.143	0.000
	UAC	0.172	0.148	0.056	0.000	0.500	0.500	0.000	0.143	**0.313**	**0.125**
H	CAU	0.000	0.000	0.053	0.000	0.000	0.333	**0.000**	**0.400**	0.067	0.000
	CAC	**0.278**	**0.000**	0.067	0.000	0.000	0.500	0.333	0.000	0.273	0.000
Q	CAA	0.125	0.000	0.133	0.250	0.300	0.000	0.000	0.250	0.100	0.000
	CAG	0.060	0.042	**0.116**	**0.000**	0.364	0.294	0.182	0.278	0.115	0.036
N	AAU	0.000	1.000	0.059	0.000	0.222	0.000	0.500	0.125	0.000	0.154
	AAC	0.083	0.085	0.172	0.250	0.067	0.235	0.000	0.000	0.217	0.250
K	AAA	0.054	0.048	0.067	0.000	0.000	0.000	0.063	0.167	0.080	0.000
	AAG	0.083	0.120	0.000	0.500	0.333	0.250	0.000	0.500	0.375	0.143
D	GAU	0.067	0.059	0.073	0.000	0.250	0.125	0.167	0.105	0.053	0.067
	GAC	0.129	0.149	0.130	0.000	0.200	0.222	0.389	0.357	**0.500**	**0.333**
E	GAA	0.022	0.036	0.043	0.000	0.000	0.105	0.167	0.063	0.042	0.125
	GAG	0.030	0.115	**0.400**	**0.000**	0.250	0.364	0.455	0.222	0.231	0.182
C	UGU	0.000	0.000	0.400	0.000	0.333	0.333	0.500	0.000	0.400	1.000
	UGC	0.500	0.375	0.091	0.000	0.000	0.000	0.500	0.286	0.000	0.000

All the 5 pairs of genes (*rpoB/C, lacZ/Y, kdpA/B, araB/A and bcsA/B*) are presented here. The TFD, SFD (Family box) and FFD codons are shown. The frequency values are marked as **bold** show a pair of co-transcribed genes having considerable difference between them in polymorphism frequency in certain codon. Thus, overall TFD, FFD, SFD of all pairs of genes can be compared.

Supplementary Table 9. The total number of codons observed to be different in co-transcribed genes in terms of polymorphism frequencies present in each degeneracy.

Co-transcribed genes	FFD (20)	TFD (18)	SFD (12)
rpoB/C	4	1	3
lacZ/Y	4	2	0
kdpA/B	4	0	1
araB/A	7	1	0
bcsA/B	4	3	4

Appendix II

Intra-operon IGRs in *E. coli* and *S. enterica*, a comparison between the size of IGRs and the number of mutations

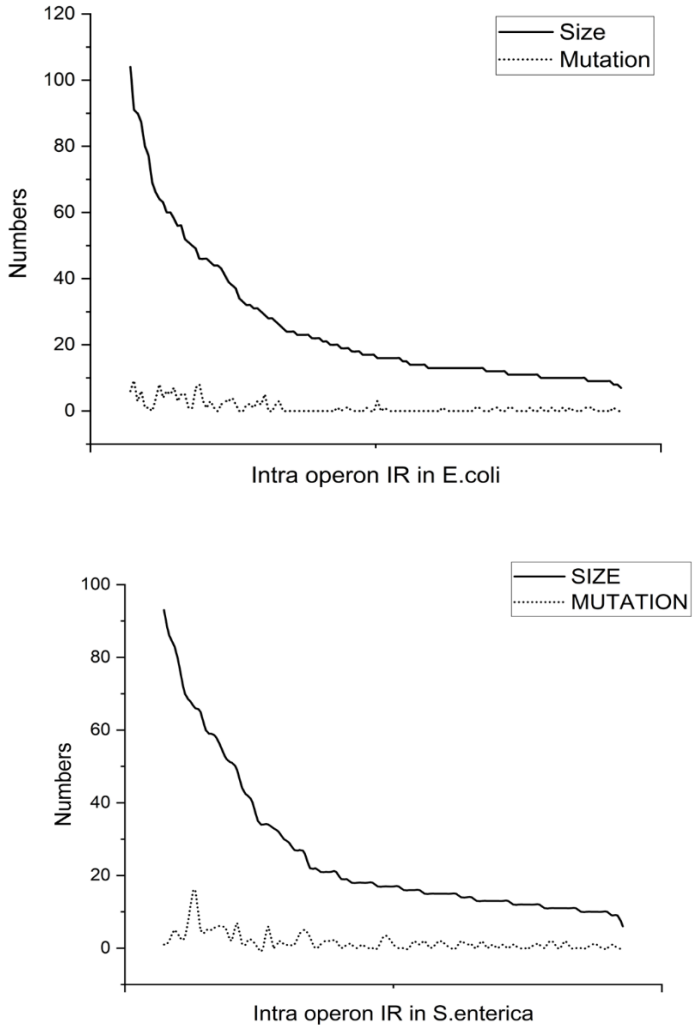


Fig. The comparison between the size and the number of mutations in *E. coli* and *S. enterica.* The figures show a variable number of mutations with an increase in the size of the IGRs in both organisms.

Appendix III

A comparative study of mutation frequency between intra-operon IGRs and inter-operon IGRs in three other bacterial species.

Other bacteria	Intra-operon IGRs taken	Frequency	
		Intra-operon IGRs	Inter-operon IGRs
Staphylococcus aureus	*69*	0.044	0.120
Streptococcus pneumoniae	*53*	0.089	0.122
Klebsiella pneumoniae	*70*	0.053	0.113

Appendix IV

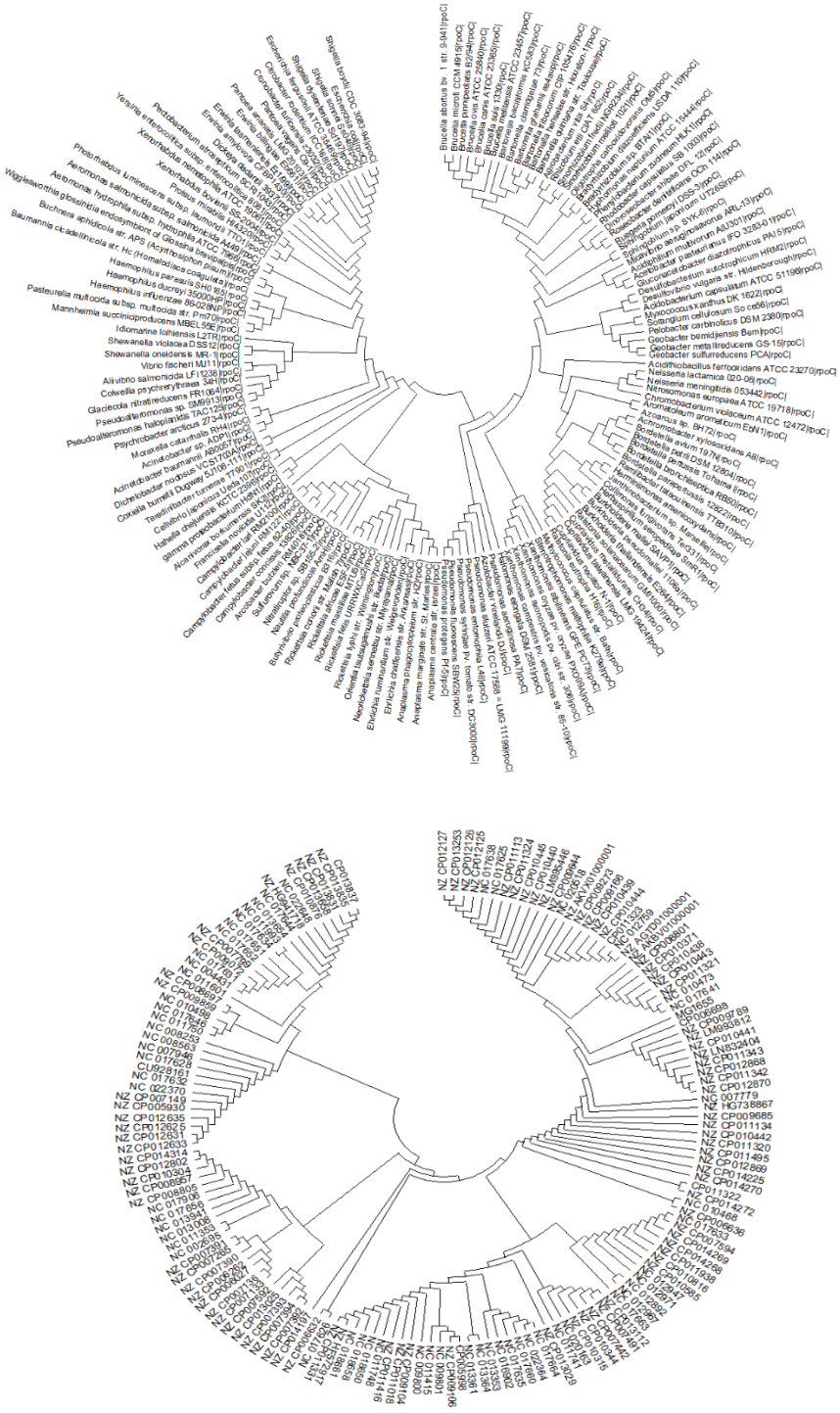


Fig. Inter-species and intra-species comparative study in *rpoC* gene in proteobacteria and *E. coli* strains reveals the difference shown by intra-species study is more prominent than that of the inter-species study.

Appendix V

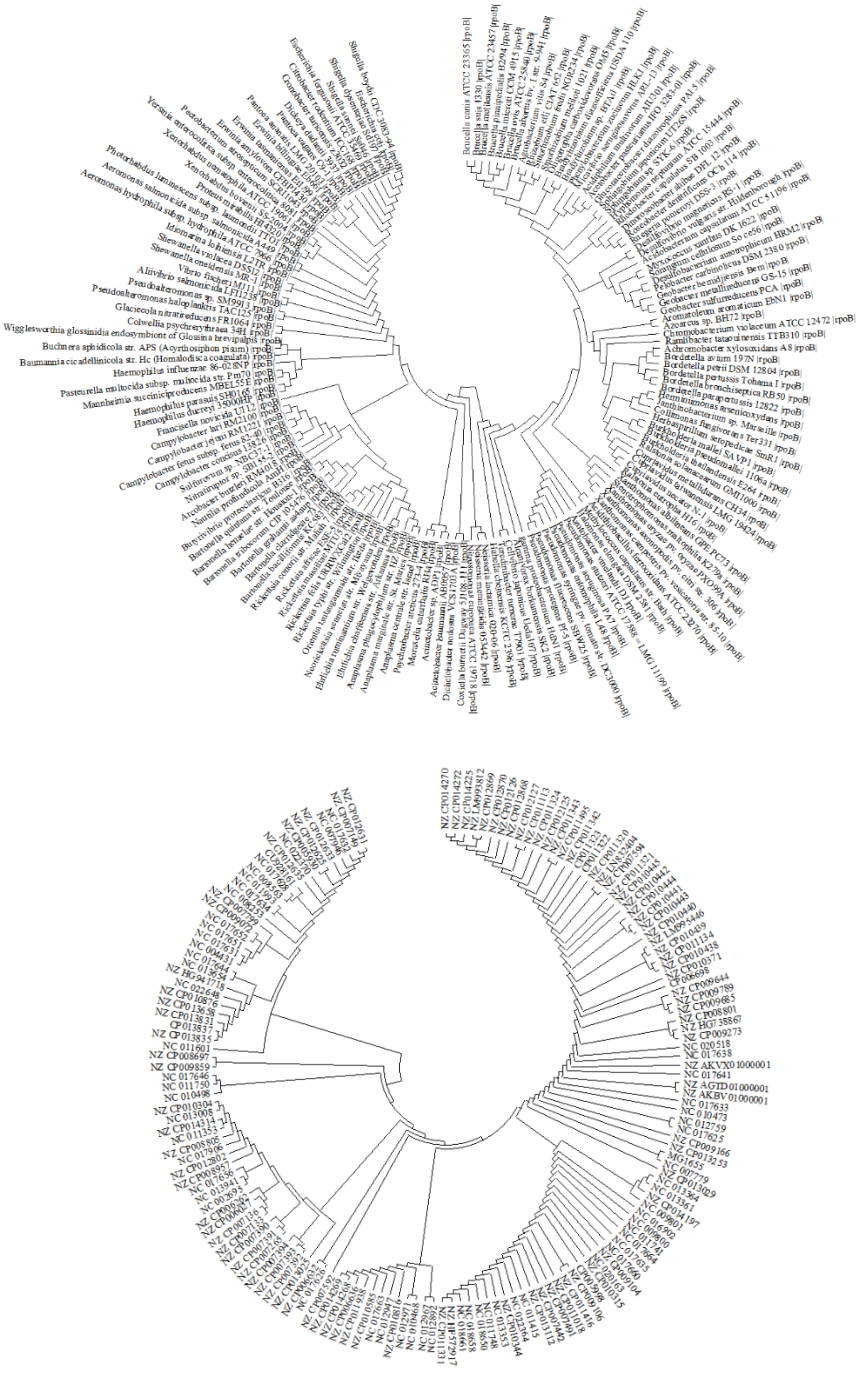


Fig. Inter-species and intra-species comparative study in *rpoB* gene in proteobacteria and *E. coli* strains reveals the difference shown by intra-species study is more prominent than that of the inter-species study.

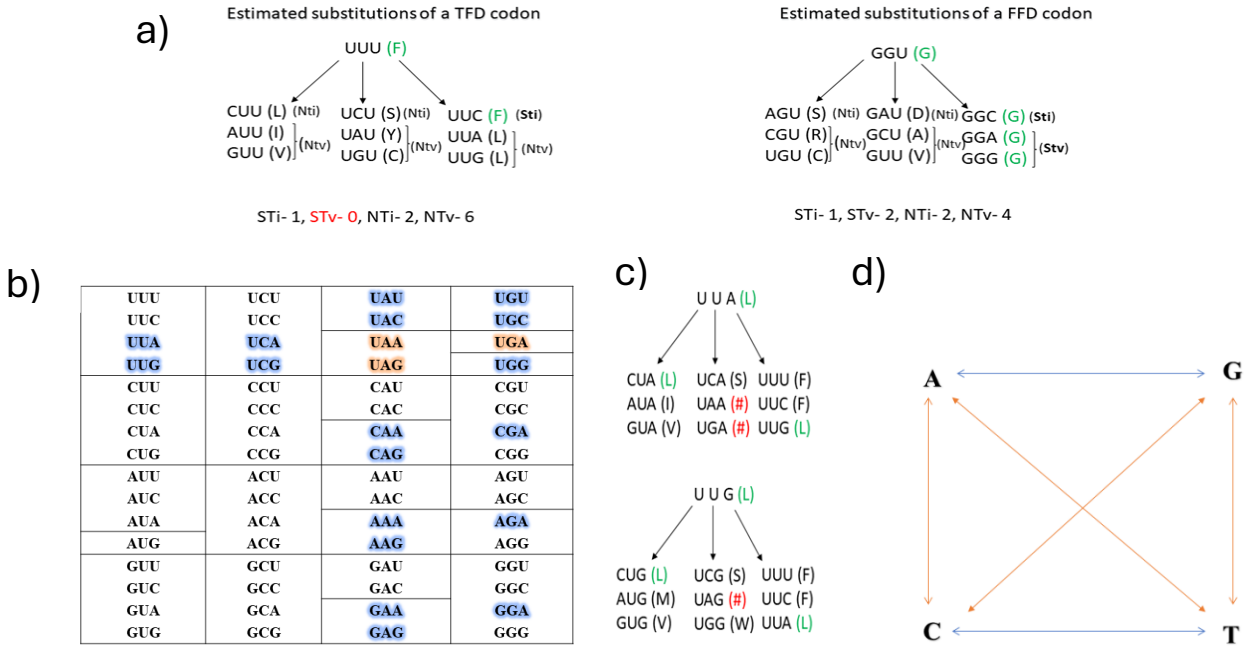
Appendix VI

Place 64*64 Matrix here

Appendix VII

A.3.1. The supplementary files related to Chapter 2 is available.

A.3. **Estimation of ti/tv ratio by accounting degeneracy and pretermination nature of codons**



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

Figure S1. (a) S standard TFD codon and a standard FFD codon provides different substitution scenarios. TFD can result into 1 STi, 0 Stv, 2 Nti and 6 Ntv whereas, FFD codon provides 1 Sti, 2 Stv, 2 Nti and 4 Ntv. (b) Genetic code table highlighting 18 PTC in blue and 3 stop codons in red. It can be noted that, 1 out of the 20 FFD codons is one of the PTCs and 10 out of 18 TFD codons are PTCs. (c) 9 estimated substitutions are presented for a six-fold degenerate codon UUA and UUG coding for Leu. Out of 9 substitutions, 2 in UUA and 1 in UUG are leading to stop codon. Hence non-synonymous substitution count is different for both. Generally, UUA and UUG have 6 possible Ntv but here UUA is giving rise to 2 stop codons and UUG is giving rise to 1 stop codon. So, UUA has 4 Ntv whereas UUG has 5 Ntv possible. This is the way of consideration of non-synonymous substitutions in PTC in our work. (d) Schematic representation of the different substitutions occurring in the DNA. The transition *(ti)* is marked with blue arrows and (*tv)* is marked with orange arrows. A and G are categorized under purine (R) and C and T are categorized under pyrimidine (Y). Out of the twelve possible base substitutions, four are *ti* in which a purine (or pyrimidine) nucleotide is replaced by another purine (or pyrimidine) nucleotide, and eight are *tv* in which a purine (or pyrimidine) nucleotide is replaced by another pyrimidine (or purine) nucleotide.

Codon	Amino acid	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Amino acid	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Amino acid	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Amino acid	Sti_e	Stv_e	Nti_e	Ntv_e
UUU	F	1	0	2	6	UCU	S	1	2	2	4	UAU*	Y	1	0	2	6	UGU*	C	1	0	2	6
UUC		1	0	2	6	UCC		1	2	2	4	UAC*		1	0	2	6	UGC*		1	0	2	6
UUA*	L	2	0	1	6	UCA*		1	2	2	4	UAA	Stop	X	X	X	X	UGA	Stop	X	X	X	X
UUG*		2	0	1	6	UCG*		1	2	2	4	UAG		X	X	X	X	UGG	W	0	0	3	6
CUU		1	2	2	4	CCU	P	1	2	2	4	CAU	H	1	0	2	6	CGU	R	1	2	2	4
CUC		1	2	2	4	CCC		1	2	2	4	CAC		1	0	2	6	CGC		1	2	2	4
CUA		2	2	1	4	CCA		1	2	2	4	CAA*	Q	1	0	2	6	CGA*		1	3	2	3
CUG		2	2	1	4	CCG		1	2	2	4	CAG*		1	0	2	6	CGG		1	3	2	3
AUU	I	1	1	2	5	ACU	T	1	2	2	4	AAU	N	1	0	2	6	AGU	S	1	0	2	6
AUC		1	1	2	5	ACC		1	2	2	4	AAC		1	0	2	6	AGC		1	0	2	6
AUA		0	2	3	4	ACA		1	2	2	4	AAA*	K	1	0	2	6	AGA*	R	1	1	2	5
AUG	M	0	0	3	6	ACG		1	2	2	4	AAG*		1	0	2	6	AGG		1	1	2	5
GUU	V	1	2	2	4	GCU	A	1	2	2	4	GAU	D	1	0	2	6	GGU	G	1	2	2	4
GUC		1	2	2	4	GCC		1	2	2	4	GAC		1	0	2	6	GGC		1	2	2	4
GUA		1	2	2	4	GCA		1	2	2	4	GAA*	E	1	0	2	6	GGA*		1	2	2	4
GUG		1	2	2	4	GCG		1	2	2	4	GAG*		1	0	2	6	GGG		1	2	2	4

Table S1. Estimated (e) polymorphism for each codon are summarized here. Codons superscript by (*) are the pretermination codons where the non-synonymous substitutions giving rise to stop codon can be seen. Further this table is modified in the main text (Table 1), where the possibilities of stop codons from the pretermination codons is excluded.

Codon	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Sti_e	Stv_e	Nti_e	Ntv_e	Codon	Sti_e	Stv_e	Nti_e	Ntv_e
UUU	1	0	2	6	UCU	1	2	2	4	UAU	1	0	2	4	UGU	1	0	2	5
UUC	1	0	2	6	UCC	1	2	2	4	UAC	1	0	2	4	UGC	1	0	2	5
UUA	2	0	1	4	UCA	1	2	2	2	UAA	X	X	X	X	UGA	X	X	X	X
UUG	2	0	1	5	UCG	1	2	2	3	UAG	X	X	X	X	UGG	0	0	1	6
CUU	1	2	2	4	CCU	1	2	2	4	CAU	1	0	2	6	CGU	1	2	2	4
CUC	1	2	2	4	CCC	1	2	2	4	CAC	1	0	2	6	CGC	1	2	2	4
CUA	2	2	1	4	CCA	1	2	2	4	CAA	1	0	1	6	CGA	1	3	1	3
CUG	2	2	1	4	CCG	1	2	2	4	CAG	1	0	1	6	CGG	1	3	2	3
AUU	1	1	2	5	ACU	1	2	2	4	AAU	1	0	2	6	AGU	1	0	2	6
AUC	1	1	2	5	ACC	1	2	2	4	AAC	1	0	2	6	AGC	1	0	2	6
AUA	0	2	3	4	ACA	1	2	2	4	AAA	1	0	2	5	AGA	1	1	2	4
AUG	0	0	3	6	ACG	1	2	2	4	AAG	1	0	2	5	AGG	1	1	2	5
GUU	1	2	2	4	GCU	1	2	2	4	GAU	1	0	2	6	GGU	1	2	2	4
GUC	1	2	2	4	GCC	1	2	2	4	GAC	1	0	2	6	GGC	1	2	2	4
GUA	1	2	2	4	GCA	1	2	2	4	GAA	1	0	2	5	GGA	1	2	2	3
GUG	1	2	2	4	GCG	1	2	2	4	GAG	1	0	2	5	GGG	1	2	2	4

Table S2. Theoretically estimated ($_e$) polymorphism for each codon by excluding the non-synonymous substitutions (*Nti* & *Ntv*) leading to stop codons through PTC. Considering the PTC a total of 62 Sti, 72 Stv, 116 Nti and 276 Ntv are possible in a genetic codon table.

Table S4 A hypothetical example of our methodology showing the derivation of reference sequence from gene alignments and finding the mutations.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Strain 1	A	A	T	A	G	G	C	A	A	C	T	G	G	G	G	A	A	T
Strain 2	A	A	T	A	G	G	C	A	A	C	T	A	G	G	G	A	A	T
Strain 3	A	A	T	A	G	G	C	A	A	C	T	G	G	G	G	A	A	T
Strain 4	A	A	T	A	G	G	C	A	A	C	T	G	A	G	G	A	A	T
Strain 5	A	A	T	T	G	G	C	A	A	C	T	G	G	G	G	A	A	T
Strain 6	A	A	T	T	G	G	C	A	A	C	T	G	G	G	G	A	A	T
Strain 7	A	A	T	T	G	G	C	A	A	C	T	G	G	G	G	A	A	T
Strain 8	A	A	T	T	G	G	C	A	A	C	T	G	G	G	G	A	A	A
Strain 9	A	A	T	T	G	G	C	A	A	C	T	G	G	G	C	A	A	C
Strain 10	A	A	T	T	G	G	C	A	G	C	T	G	G	G	A	A	A	N
A Count	10	10	0	4	0	0	0	10	9	0	0	1	1	0	1	10	10	2
T Count	0	0	10	6	0	0	0	0	0	0	10	0	0	0	0	0	0	6
C Count	0	0	0	0	0	0	10	0	0	10	0	0	0	0	1	0	0	1
G Count	0	0	0	0	10	10	0	0	1	0	0	9	9	10	8	0	0	0
N count	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Ref_seq	A	A	T	T	G	G	C	A	A	C	T	G	G	G	G	A	A	T
Mutation				T>A					A>G			G>A	G>A		G>A, G>C			T>A, T>C

A hypothetical set of strains/set of alignments showing the patterns of substitutions and counts of respective nucleotides. the most frequent nucleotide in a column is regarded as the reference nucleotide in that column/position. the mutation row shows the mutations from most frequent to least frequent nucleotides. The 5th codon (13,14,15) is showing mutations at two positions, hence for the observed substitutions, the concerned mutations are not included, but while counting for the codon count and subsequent estimated substitutions it is included in. In the 18th position strain no. 10 is showing the presence of an ambiguous nucleotide "N" is not considered while calculating the polymorphism.

Table S5. The observed calculation of Stio, Stvo, Ntio and Ntvo as well as tio, tvo from the above hypothetical reference sequence

Codon	Mutation	ti_o	tv_o	Sti_o	Stv_o	Nti_o	Ntv_o
AAT	x	0	0	0	0	0	0
TGG	AGG	0	1	0	0	1	0
CAA	CAG	1	0	1	0	0	0
CTG	TTG/CTA	2	0	2	0	0	0
AAT	AAC/AAA	1	1	1	0	0	1
Total observed		4	2	4	0	1	1

Table S6. The possible calculation of Stie, Stve,Ntie and Ntve from the codon count and referring to Table 1. tie and tve can be calculated by STie+ Ntie and Stve+ Ntve respectively.

For each								Total possi				
Codon	St	Stv_p^j	Nti_p^j	Ntv_p^j		Codon	Codon count (f_j)		Codon	Sti_e	Stv_e	Nti_e	Ntv_e
AAU	1	0	2	6		AAU	2		AAU	2	0	4	12
UGG	0	0	1	6		UGG	1		UGG	0	0	1	6
CAA	1	0	1	6	×	CAA	1	=	CAA	1	0	1	6
CTG	2	2	1	4		CTG	1		CTG	2	2	1	4
GGG	1	2	2	4		GGG	1		GGG	1	2	2	4
								Total Possible		**6**	**4**	**9**	**32**

Table S7. Finding out the value of the above hypothetical sequence by applying the modified equation.

$\frac{ti'}{tv'}$	$\frac{Sti'}{Stv'}$	$\frac{Nti'}{Ntv'}$
What is the value? *ti'= ti_o/ ti_e= 4/15* *tv'= tv_o/ tv_e= 2/36* *= = 4.8*	What is the value? *Sti'= Sti_o/ Sti_e= 4/6* *STv'= Stv_o/ Stv_e= 0/4* *= = N/A*	What is the value? *Nti'= Nti_o/ Nti_e= 1/9* *Ntv'= Ntv_o/ Ntv_e= 1/32* *= 3.55*

Figure S2. The illustration describes about the relationship between $\frac{Sti}{Stv}$ and $\frac{Sti'}{Stv'}$ with the proportion of TFD codons and FFD codons. Suppose in TFD Stio= x_1, Stvo= 0. In FFD Stio = x_2 and Stvo = y_2. Total number of hypothetical codons = 100 (TFD+FFD). In case A, the FFD codon proportion is 100%, hence $\frac{Sti}{Stv} < \frac{Sti'}{Stv'}$. In case B, the TFD codons and FFD codons are at equal proportions hence $\frac{Sti}{Stv} = \frac{Sti'}{Stv'}$. In case C, the FFD proportion is 1% hence $\frac{Sti}{Stv} > \frac{Sti'}{Stv'}$. The hypothetical scenario shows that, the $\frac{Sti'}{Stv'}$ value decreases with increasing proportion of TFD codons (decreasing FFD codons). This additionally validates our improved estimator in terms of the synonymous $\frac{ti}{tv}$ ratio and codon composition.

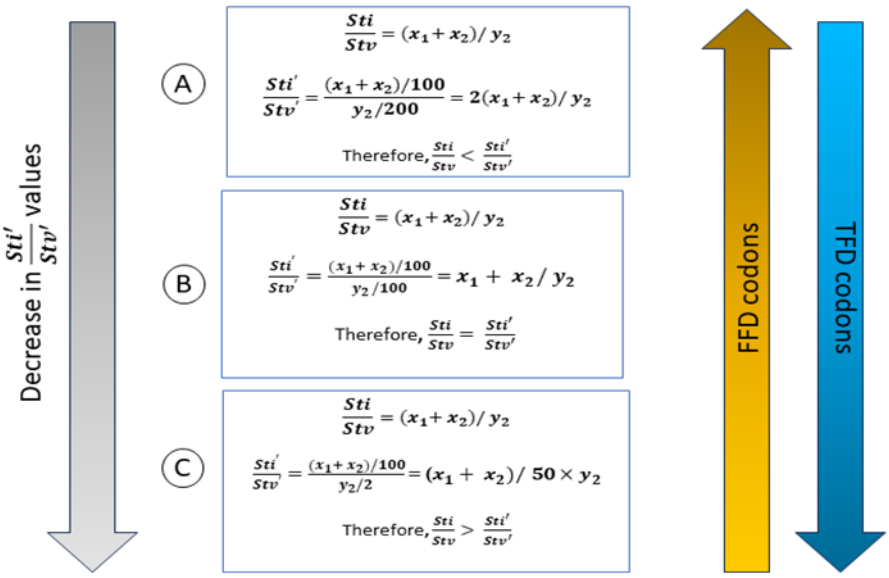


Table S8: Simulation study

We conducted a simulation study to support the results of $\frac{Sti}{Stv}$ using conventional and improved estimators in relation to the codon compositional difference. A hypothetical coding sequence of size 7600 codons was generated for the simulation study. The hypothetical sequence consisted of 200 codons each of the 20 FFD codons and 18 TFD codons. Starting from 100% FFD with gradual 1% increase of TFD and similar proportion decreasing FFD, a total of 101 variable sequences (100% FFD to 100% TFD) were generated of which two extreme sequences were entirely composed of either FFD or TFD codons. We then introduced 760 (10%) substitutions that includes all the 12 base substitutions such as (A→T 20%, A→C 20%, A→G 60%) (T→A 30%, T→C 50%, T→G 20%) (C→T 66%,C→A 21%,C→G 13%) (G→A 55%, G→T 39%, G→C 6%) at the 3rd positions of the randomly generated codons in each coding sequence (32). We calculated the frequency of each nucleotide at the 3rd position of those 10% codons. To determine the number of substitutions to be incorporated we carried out this step separately for FFD codons and TFD codons since we did not consider *tv* in TFD codons to maintain only synonymous substitutions. In total we have generated 44,378 *ti* and 16,195 *tv* in the gene sets while excluding the 16,188 tv in case of TFD codons. The result of the $\frac{Sti}{Stv}$ and $\frac{Sti'}{Stv'}$ were compared using statistical tools.

TFD%	Sti/Stv	Sti'/Stv'		TFD%	Sti/Stv	Sti'/Stv'
0	1.375074221	2.750148442		0.960784	2.702771	2.756826
0.01010101	1.36575985	2.704204503		1	2.776803	2.776803
0.020408163	1.397408551	2.73892076		1.040816	2.798828	2.742852
0.030927835	1.39254274	2.701532917		1.083333	2.883569	2.768226
0.041666667	1.412624026	2.712238131		1.12766	2.901796	2.727689
0.052631579	1.456429892	2.767216794		1.173913	2.988689	2.749594
0.063829787	1.45073174	2.72737567		1.222222	3.046238	2.741614
0.075268817	1.461643653	2.718657194		1.272727	3.114611	2.740858
0.086956522	1.49287556	2.74689103		1.325581	3.174161	2.729778
0.098901099	1.534307783	2.792440165		1.380952	3.254076	2.733424
0.111111111	1.534689329	2.762440791		1.439024	3.320209	2.722571
0.123595506	1.525978549	2.716241816		1.5	3.468866	2.775092
0.136363636	1.559078321	2.743977846		1.564103	3.486519	2.719485
0.149425287	1.5877791	2.762735635		1.631579	3.619426	2.750764

0.162790698	1.585011912	2.726220489		1.702703	3.625621	2.682959
0.176470588	1.603075462	2.725228285		1.777778	3.797594	2.734268
0.19047619	1.632186047	2.742072558		1.857143	3.81815	2.672705
0.204819277	1.660346777	2.75617565		1.941176	3.997627	2.718387
0.219512195	1.657473613	2.718256725		2.030303	4.076444	2.690453
0.234567901	1.679515672	2.720815388		2.125	4.309203	2.75789
0.25	1.721679688	2.7546875		2.225806	4.391439	2.722692
0.265822785	1.747115956	2.760443211		2.333333	4.540348	2.724209
0.282051282	1.743895649	2.720477213		2.448276	4.778031	2.771258
0.298701299	1.762814839	2.714734852		2.571429	4.902287	2.745281
0.315789474	1.794776272	2.728059934		2.703704	5.014443	2.707799
0.333333333	1.822872097	2.734308146		2.846154	5.369541	2.792162
0.351351351	1.834248355	2.714687565		3	5.543588	2.771794
0.369863014	1.881540573	2.747049237		3.166667	5.724494	2.747757
0.388888889	1.895907242	2.730106429		3.347826	6.096127	2.804219
0.408450704	1.920623795	2.727285789		3.545455	6.168705	2.71423
0.428571429	1.97273826	2.761833565		3.761905	6.641057	2.789244
0.449275362	1.987945824	2.743365237		4	6.944156	2.777662
0.470588235	2.018229286	2.744791829		4.263158	7.128451	2.708811
0.492537313	2.045126438	2.740469427		4.555556	7.566684	2.724006
0.515151515	2.089731423	2.758445478		4.882353	8.106635	2.756256
0.538461538	2.125812509	2.763556262		5.25	8.547438	2.73518
0.5625	2.113611125	2.705422239		5.666667	9.1171	2.73513
0.587301587	2.145888594	2.703819629		6.142857	9.791214	2.74154
0.612903226	2.213193117	2.744359465		6.692308	10.67128	2.774533
0.639344262	2.247035877	2.74138377		7.333333	11.40186	2.736447
0.666666667	2.322187812	2.786625374		8.090909	12.44554	2.73802
0.694915254	2.366417112	2.792372193		9	13.37409	2.674818
0.724137931	2.392318542	2.775089509		10.11111	15.2614	2.747052
0.754385965	2.427925693	2.76783529		11.5	17.19773	2.751637
0.785714286	2.448785109	2.742639322		13.28571	20.09748	2.813648
0.818181818	2.516289154	2.767918069		15.66667	23.669	2.840279
0.851851852	2.554453723	2.758810021		19	27.89614	2.789614
0.886792453	2.553124453	2.70631192		24	35.58503	2.846802
0.923076923	2.594511292	2.698291743		32.33333	45.60313	2.736188

Figure S3. The line plot illustrates the outcome of simulation study. The x-axis shows the TFD% and y-axis shows Sti/Stv values. The solid line shows the conventional Sti/Stv value whereas the dash line shows the improved Sti/Stv values. The TFD% can be 0-100% whereas the Sti/Stv can be between 0-?. We have excluded the Sti/Stv value of the gene having 100% TFD as the value was ?. The graph shows a linear curve for the conventional estimator and a strong correlation with TFD%. But the improved estimator performs steadily by accounting for the compositional difference. The mean of the conventional and improved estimator was 7.32 and 2.74 respectively whereas the standard deviation of the conventional and improved estimator was 17.40 and 0.04 respectively.

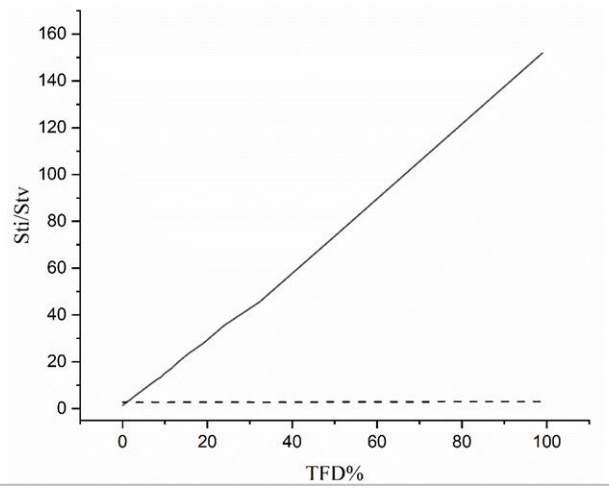


Figure S4. The figure illustrates the box-plot of ti/tv values obtained through ML Based (maximum likelihood), conventional and improved method. The ti/tv ratio between ML based method and improved estimator show a significant difference (p<0.01).

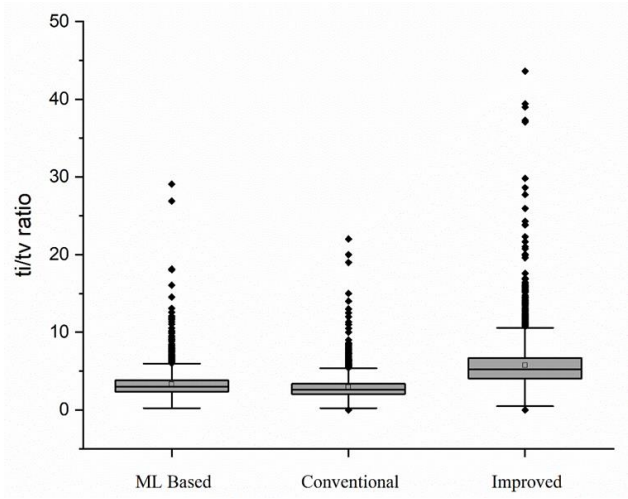


Table 10: A comparative study between ML based and improved method

Gene	ML Based	Improved	ML-Improved	Gene	ML Based	Improved	ML-Improved
	ti/tv	*ti′/tv′*			*ti/tv*	*ti′/tv′*	
yefM	6.95	0.00	6.95	*sadH*	3.64	2.36	1.28
flgC	6.31	0.00	6.31	*sra*	1.25	0.00	1.25
ibpA	7.02	1.93	5.09	*hslJ*	4.18	2.93	1.25
fdnI	7.54	2.99	4.55	*gfcB*	6.54	5.30	1.25
galE	9.15	4.94	4.21	*nhaR*	3.95	2.72	1.23
rbsC	5.92	1.98	3.94	*copA*	3.38	2.16	1.22
bhsA	5.36	1.46	3.90	*rmlA2*	3.66	2.45	1.21
ccmE	5.36	1.94	3.42	*flgB*	1.85	0.67	1.18
cusF	5.26	2.01	3.25	*cadA*	2.47	1.29	1.18
ykgM	4.21	1.29	2.92	*fldC*	3.46	2.31	1.15

xylA	4.74	1.97	2.77	*cheY*	2.11	0.97	1.14
por	5.06	2.49	2.57	*bcsZ*	2.61	1.56	1.05
yhcB	16.05	13.59	2.46	*yfkM*	3.53	2.49	1.04
yjiA	3.85	1.43	2.42	*dinI*	2.34	1.32	1.02
flgH	4.94	2.67	2.27	*estB*	3.38	2.36	1.02
rob	4.16	1.93	2.23	*atpE*	1.00	0.00	1.00
napA	5.53	3.30	2.23	*yjiE*	3.12	2.13	0.99
prpC	4.99	2.83	2.16	*argF*	2.61	1.67	0.94
wfgD	4.69	2.54	2.15	*narY*	3.86	2.93	0.93
gltD	3.49	1.34	2.15	*elbB*	1.68	0.78	0.90
uspE	3.47	1.33	2.15	*sgrT*	1.56	0.69	0.87
srlR	5.34	3.24	2.10	*fruB*	3.08	2.21	0.87
tuf1	4.01	1.96	2.06	*marR*	4.72	3.89	0.83
thiQ	3.43	1.50	1.93	*lutA*	2.76	1.93	0.83
frdB	4.51	2.59	1.92	*yddG*	3.13	2.30	0.83
hisB	4.37	2.46	1.91	*yhjE*	3.43	2.64	0.79
ycgE	4.29	2.39	1.90	*prpR*	3.67	2.91	0.76
ygbM	3.01	1.13	1.88	*rhaS*	4.19	3.44	0.76
pgl	2.85	1.00	1.85	*glxK*	2.44	1.70	0.75
apbE	4.08	2.30	1.79	*galF*	3.18	2.46	0.72
dam	3.67	1.91	1.76	*ycdX*	1.98	1.32	0.66
modE	2.57	0.86	1.71	*rhaB*	2.93	2.28	0.66
ycjO	3.10	1.61	1.49	*bhsA*	4.24	3.60	0.64
gltB	3.17	1.73	1.44	*ybaT*	3.96	3.32	0.64
tfdS	2.56	1.16	1.40	*gor*	3.31	2.68	0.63
fucP	4.34	2.95	1.40	*grxC*	6.38	5.78	0.60
virF	2.80	1.41	1.40	*leuB*	3.94	3.35	0.59
mutS	3.99	2.60	1.39	*gap*	2.50	1.93	0.57
rstA	3.32	1.94	1.38	*cusB*	2.27	1.71	0.56
napC	5.33	3.96	1.38	*proB*	3.85	3.29	0.56
ybhC	3.34	2.01	1.33	*betA*	4.06	3.51	0.56
sucD	3.98	2.65	1.33	*acsA*	4.75	4.20	0.55
cheW	1.78	0.49	1.30	*ytfE*	2.01	1.46	0.55
emrY	2.97	1.68	1.29	*ydeH*	2.01	1.46	0.55
gap	2.52	1.97	0.55	*csgA*	3.37	3.21	0.16
ccmC	2.09	1.55	0.54	*wzzB*	2.85	2.70	0.15
rbsK	1.86	1.32	0.54	*fhuA*	2.32	2.17	0.15
ydfH	1.53	1.00	0.54	*priC*	3.59	3.44	0.15
eco	5.37	4.84	0.53	*ulaE*	3.49	3.35	0.14
ldrD	3.56	3.05	0.52	*yjiG*	3.09	2.97	0.12
sgrR	2.70	2.19	0.51	*hyuA*	3.55	3.44	0.11
flgE	1.77	1.28	0.49	*yedQ*	1.83	1.72	0.11
bioB	2.42	1.96	0.46	*yciC*	3.56	3.46	0.11

lhgO	3.01	2.57	0.44	*gltC*	2.46	2.36	0.10
mqo	3.27	2.84	0.43	*btuB*	2.57	2.48	0.09
ogt	2.80	2.37	0.43	*ilvH*	3.96	3.87	0.09
rhaA	3.11	2.70	0.41	*tusB*	6.65	6.56	0.09
hisH	3.65	3.25	0.40	*uspF*	2.00	1.91	0.09
ycjS	1.97	1.58	0.39	*aroP*	3.30	3.21	0.09
ugpC	2.87	2.48	0.39	*yegS*	5.22	5.14	0.09
malF	4.16	3.77	0.39	*flhE*	2.22	2.14	0.08
arsC	4.64	4.26	0.38	*flgL*	2.72	2.64	0.08
flhC	3.01	2.64	0.38	*mntS*	2.01	1.94	0.07
araB	2.92	2.57	0.36	*pflA*	7.62	7.56	0.06
hcaR	2.63	2.28	0.35	*fldI*	2.14	2.11	0.03
flgA	2.55	2.20	0.35	*torY*	1.80	1.77	0.03
galK	3.16	2.81	0.35	*oppF*	4.22	4.21	0.01
matA	1.76	1.42	0.34				
betB	3.04	2.71	0.33				
glcA	2.31	1.99	0.32				
phoE	2.50	2.19	0.31				
srlR	1.92	1.61	0.31				
cysW	3.20	2.91	0.29				
prfA	2.26	1.97	0.29				
csgC	1.59	1.31	0.28				
nupC	3.82	3.55	0.27				
yidK	3.24	2.98	0.26				
amiD	2.70	2.44	0.26				
cspC	2.27	2.02	0.26				
ybbY	2.50	2.25	0.25				
zntB	2.25	2.00	0.25				
evgA	3.11	2.87	0.24				
slyA	2.99	2.77	0.22				
mepH	4.23	4.02	0.21				
glxR	3.46	3.26	0.20				
prpD	3.05	2.85	0.20				
hisE	3.25	3.05	0.20				
ldhA	3.59	3.40	0.19				
cbl	2.72	2.55	0.17				
yiaB	5.93	5.77	0.16				

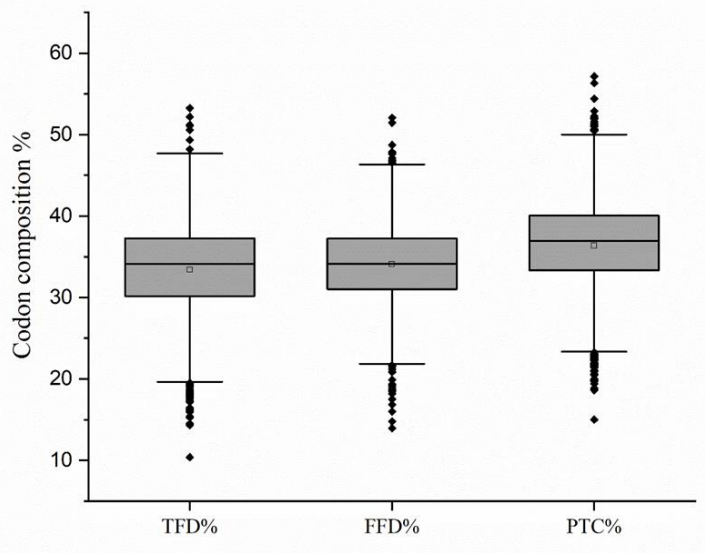


Figure S5. The figure represents the percentage (%) of TFD, FFD and PTC codons considered in this study. The minimum and maximum TFD% in the dataset was observed as 10.38% and 53.25%, similarly the minimum and maximum FFD% in the dataset was observed as 13.95% and 52.05%. The minimum and maximum PTC% in the dataset was observed as 15.00% and 57.14%. The minimum to maximum PTC% was observed in the gene atpE as 15.00% and in the gene chaB as 57.14%

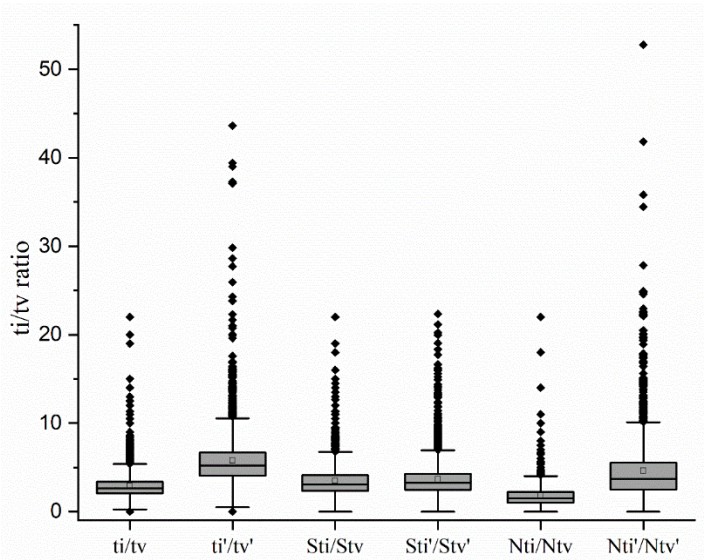
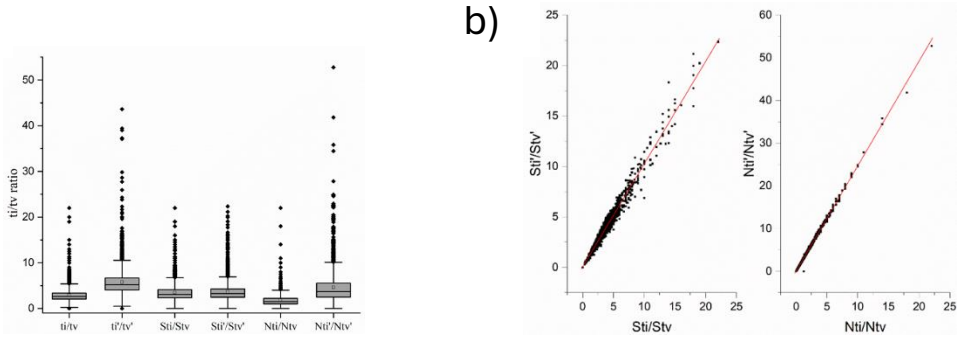


Figure S6. The box-plot presents the different $\frac{ti}{tv}$ ratio values. Here we have compared the different $\frac{ti}{tv}$ ratios of the entire set of genes through conventional and improved method. The x-axis show different $\frac{ti}{tv}$ ratio (synonymous, non-synonymou and overall) whereas the y-axis show different the values. $\frac{ti}{tv}$ shows the conventional method values and $\frac{ti'}{tv'}$ shows the result of improved method values. Similarly, $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ represent the conventional values and $\frac{Sti'}{Stv'}$ and $\frac{Nti'}{Ntv'}$ represent the improved values. This illustration shows that the median values between $\frac{ti}{tv}$ & $\frac{ti'}{tv''}$, $\frac{Sti}{Stv}$ & $\frac{Sti'}{Stv'}$ and $\frac{Nti}{Ntv}$ & $\frac{Nti'}{Ntv'}$ are statistically different (p-value < 0.01).

a) b)



(a) The figure presents distribution of different $\frac{ti}{tv}$ values using box plot. In total 2516 *E. coli* genes were considered while calculating these values. The median values of $\frac{ti}{tv}$, $\frac{Sti}{Stv}$, and $\frac{Nti}{Ntv}$ are statistically different (p-value < 0.01) from $\frac{ti'}{tv'}$, $\frac{Sti'}{Stv'}$ and $\frac{Nti'}{Ntv'}$ respectively. (b) The scatter plots elucidate the conventional and improved ratios. The x-axes represent the conventional $\frac{Sti}{Stv}$ and $\frac{Nti}{Ntv}$ values, whereas the y-axes represent the improved $\frac{Sti'}{Stv'}$ and $\frac{Nti'}{Ntv'}$ values. The y-axis further clarifies the strong correlation between the change in conventional and the improved ratios throughout the coding sequences. Further $\frac{Sti}{Stv}$ & $\frac{Sti'}{Stv'}$ and $\frac{Nti}{Ntv}$ & $\frac{Nti'}{Ntv'}$ are observed to be statistically different (p-value < 0.01).

LIST OF PUBLICATIONS

Published

- Aziz R., Sen P., **Beura P.K.** et al. Incorporation of transition to transversion ratio and nonsense mutations improves the estimation of the number of synonymous and non-synonymous sites in codons. DNA Research. v.29(4); 2022;29: dsac023.

- **Beura P. K.**, Sen P., Aziz R., Satapathy S. S., Ray S. K The transcribed intergenic regions exhibit lower frequency of nucleotide polymorphism than the untranscribed intergenic regions in the genomes of Escherichia coli and Salmonella enterica. Journal of Genetics, 2023:102.22

- **Beura P.K.**, Sen P., Aziz R., Chetia C., Dash M., Satapathy S. S., Ray S. K Difference in synonymous polymorphism related to codon degeneracy between co-transcribed genes in the genome of Escherichia coli. Curr. Sci. 2023: VOL. 125, NO. 8, 871-877.

Manuscripts under review

- **Beura P.K.**, Sen P., Aziz R., Das S., Namsa N.D., Feil E., Satapathy S. S., Ray S. K Synonymous and non-synonymous transitions/transversions vividly disclose purifying selection in Escherichia coli coding sequences.

- Sen P., Aziz R., **Beura P.K.** et al. A consensus sequence based intraspecies single nucleotide variations study reveals difference between intergenic regions and four-fold degenerate sites in bacterial genome.

Manuscripts under preparation

- **Beura P.K.** et al., Non-synonymous ti/tv is variable between four-fold degenerate codons and two-fold degenerate codons in *Escherichia coli*

Conferences/Seminars

- Poster presentation in National level seminar titled "**BIOLOGY IS FASCINATING**" organized by the department of Molecular Biology & Biotechnology, Tezpur University, Tezpur in association with *inSCIgnis* '22 on 1st March 2022.

- Poster presentation in National-level mentoring symposium "**Gurukul in emerging areas in modern biology and medicine**" on 2nd & 3rd March, 2023 organized by Indian National Young Academy of Sciences and department of Molecular Biology & Biotechnology, Tezpur University, Tezpur (INYAS)

- 3rd position in oral presentation in National Seminar on "**EXCITEMENTS IN BIOLOGICAL RESEARCH**" organized by Department of Molecular Biology and Biotechnology, in collaboration with Students' Science Council, Tezpur University.

- Poster presentation in an international conference on "**Molecular Mechanisms in Evolution**" Organized by Ashoka University, Sonipat, Haryana, SMBE Regional meeting in India, 2023, 16-17 Dec 2023, Dehradun

- Oral presentation in National workshop on "**Recent trends in Bioinformatics and Computational Biology**" organized by Center for Bioinformatics and Computational Biology, Tezpur University, 8th Nov-12th Nov 2024.

DNA Research, 2022, 29(4), 1–8
https://doi.org/10.1093/dnares/
dsac023Advance Access Publication Date:
3 August 2022
Research Article

Research Article

Incorporation of transition to transversion ratioand nonsense mutations, improves the estimation of the number of synonymous and non-synonymous sites in codons

Ruksana Aziz[1], Piyali Sen[2], Pratyush Kumar Beura[1], Saurav Das[1], Debapriya Tula[3], Madhusmita Dash[4], Nima Dondu Namsa[1,5], Ramesh Chandra Deka[5,6], Edward J. Feil[7], Siddhartha Sankar Satapathy[2,5,*],and Suvendra Kumar Ray[1,5,*]

[1]Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, 784028 Assam, India, [2]Department of Computer Science and Engineering, Tezpur University, Tezpur, 784028 Assam, India, [3]TCS Innovation, Tata Consultancy Services, Hyderabad, 500081 Telangana, India, [4]Department of Electronics and Communication Engineering, NIT, Papum Pare, 791113 Arunachal Pradesh, India, [5]Center for Multidisciplinary Research, TezpurUniversity, Tezpur, 784028 Assam, India, [6]Department of Chemical Sciences, Tezpur University, Tezpur, 784028 Assam, India, and [7]Department of Biology and Biochemistry, The Milner Centre for Evolution, University of Bath,Bath BA2 7AY, UK

*To whom correspondence should be addressed. Tel. þ91 3712 275117. Email: ssankar@tezu.ernet.in (S.S.S.); Tel. þ913712 275406. Email: suven@tezu.ernet.in (S.K.R)

Received 28 March 2022; Editorial decision 27 June 2022

Abstract

A common approach to estimate the strength and direction of selection acting on protein coding sequences is to calculate the dN/dS ratio. The method to calculate dN/dS has been widely used by many researchers and many critical reviews have been made on its application after the proposition by Nei and Gojobori in 1986. However, the method is still evolving considering the non-uniform substitution rates and pretermination codons. In our study of SNPs in 586 genes across 156 *Escherichia coli* strains, synonymous polymorphism in 2-fold degenerate codonswere higher in comparison to that in 4-fold degenerate codons, which could be attributed to the difference between transition (Ti) and transversion (Tv) substitution rates where the average rate of a transition is four times more than that of a transversion in general. We considered both the Ti/Tv ratio, and nonsense mutation in pretermination codons, to improve estimates of synon-ymous (S) and non-synonymous (NS) sites. The accuracy of estimating dN/dS has been improved by considering the Ti/Tv ratio and nonsense substitutions in pretermination codons.We showed that applying the modified approach based on Ti/Tv ratio and pretermination codonsresults in higher values of dN/dS in 29 common genes of equal reading-frames between *E. coli* and *Salmonella enterica*. This study emphasizes the robustness of amino acid composition with varying codon degeneracy, as well as the pretermination codons when calculating dN/dS values.

Key words: dN/dS, synonymous/non-synonymous sites, pretermination codon, transition, transversion

Journal of Genetics(2023) 102:22

RESEARCH ARTICLE

Transcribed intergenic regions exhibit a lower frequency of nucleotide polymorphism than the untranscribed intergenic regions in the genomes of *Escherichia coli* and *Salmonella enterica*

PRATYUSH KUMAR BEURA[1], PIYALI SEN[2], RUKSANA AZIZ[1], SIDDHARTHA SHANKAR SATAPATHY[2,3]* and SUVENDRA KUMAR RAY[1,3]*

[1]*Department of Molecular Biology and Biotechnology, Tezpur University, Napaam 784 028, India*
[2]*Department of Computer Science and Engineering, Tezpur University, Napaam 784 028, India*
[3]*Center for Multidisciplinary Research, Tezpur University, Napaam 784 028, India*
*For correspondence. E-mail: Siddhartha Shankar Satapathy, ssankar@tezu.ernet.in; Suvendra Kumar Ray, suven@tezu.ernet.in.

Received 8 August 2022; revised 17 October 2022; accepted 1 December 2022

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 a higher rate of C?T transition. Intraoperon intergenic regions undergo transcription along with adjacent co-transcribed genes in an operon, whereas interoperon 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 intraoperon IGRs and interoperon IGRs in genomes of two well-known closely related bacteria such as *Escherichia coli* and *Salmonella enterica*. In general, the size of intraoperon IGRs was smaller than that of interoperon IGRs in *E. coli* and *S. enterica*. Interestingly, the polymorphism frequency at intraoperon IGRs was 2.5-fold lesser than that in the interoperon IGRs in *E. coli* genome. Similarly, the polymorphism frequency at intraoperon IGRs was 2.8-fold lesser than that in the inter-operon IGRs in *S. enterica* genome. Therefore, the intraoperon IGRs were often observed to be more conserved. In the case of interoperon IGRs, the T?C transition frequency was a minimum of two times more frequent than T?A transversion frequency whereas in the case of intraoperon IGRs, T?C transition frequency was similar to that of T?A transversion frequency. The polymorphism was purine-biased and keto-biased more in intraoperon 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 intraoperon IGRs, respectively. In *S. enterica*, the transition/transversion ratio was observed as 2.134 and 2.780 in inter-operon and in intraoperon IGRs, respectively. The observation in this study indicates that transcribable IGRs might not always have higher polymorphism frequency than nontranscribable IGRs. The lower polymorphism frequency at intraoperon 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.

Keywords. intergenic regions; operon; nucleotide polymorphism; replication; transcription.

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, namely intraoperon IGRs and interoperon IGRs. Interoperon IGRs are found in between two separate operonic/cistronic units. Intraoperon IGRs are found between two adjacent open-reading frames in an operon (figure 1), which are popularly known as intercistronic regions (ICRs). Intraoperon IGRs are the units that undergo replication as well as transcription, whereas interoperon IGRs are devoid of transcription. However, inter-operon IGRs may not be 100% devoid of transcription because a

Supplementary Information: The online version contains supplementary material available at https://doi.org/10.1007/s12041-023-01418-w.

Published online: 16 February 2023

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

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

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

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

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

BASE substitution mutation is a major event of molecular evolution in organisms, influenced by different factors such as DNA replication, damage in DNA bases such as deamination of cytosine/adenine, oxidation of guanine[1,2], gene expression, recombination, etc. The asymmetry in DNA replication results in different mutation patterns between the leading strand (LeS) and lagging strand (LaS) in genomes, resulting in the former being enriched with keto nucleo- tides and the latter with amino nucleotides in bacteria[3,4]. In addition, genes near the origin of replication exhibit different mutation patterns than those at the terminus of