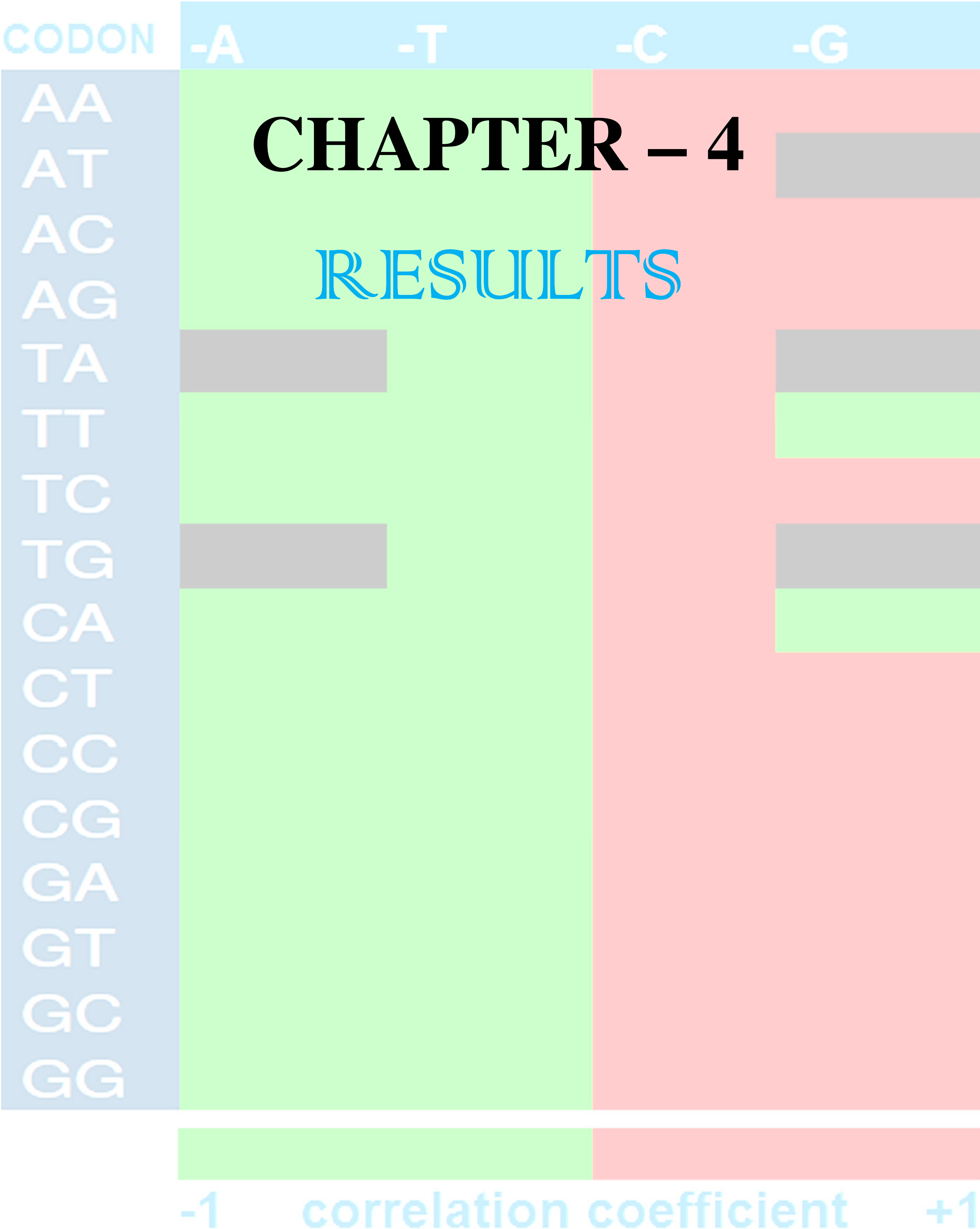
CHAPTER – 4

RESULTS



4. Results

4.1. Nucleotide composition in proto-oncogenes/oncogenes

The overall nucleotide compositions in the complete coding sequences of eighty two human proto-oncogene/oncogenes were analyzed (Table 3). The mean value of A base (520.8) was observed at the highest, followed by C (517.6), G (502.2) and T (416.6) among all the coding sequences. The GC content for the selected coding sequences ranged from 38.1% to 71.0%, with a mean of 54.1% and a standard deviation of 0.086. The AT content for the selected coding sequences varied from 29.0% to 61.9%, with a mean of 45.9% and a standard deviation of 0.086. We calculated the GC content at different codon positions (Figure 4) and it was found that the GC content at each codon position variation among the genes.

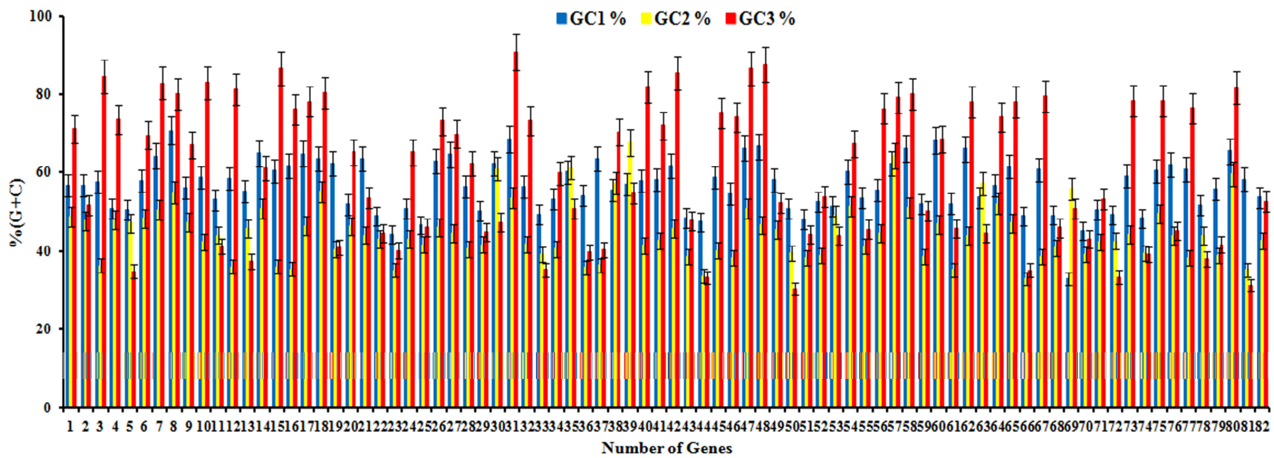


Figure 4: Percentage of GC content at three codon positions in the coding sequences of proto-oncogenes/oncogenes

It is a well known fact that the nucleotide at the third codon position varies considerably due to the wobble hypothesis which allows the cell to identify the codons that encode 61 different amino acid using less than 61 *tRNA* molecules.

In our analysis, the comparison of the nucleotide composition at third codon position (A_3, T_3, G_3, C_3) confirmed that the mean value of C_3 (187.9) was the highest followed by G_3 (178.7), T_3 (150.9) and A_3 (134.9). The mean value of GC_3% (ranged from 18.2% to 90.8%) and AT_3% (ranged from 9.2% to 69.6%) were 54.9% and 45.1% respectively with an exact standard deviation of 0.20. The average value of (GC_1+GC_2) % ranged from 39.8% to 62.9% with a mean of 50.8% and a standard deviation of 0.057 which could be presumed that G/C ending codons might be preferred in coding sequences of proto-oncogenes/oncogenes.

Table 3: Nucleotide composition analysis in the coding sequences of eighty-two proto-oncogenes/oncogenes

CDS N0.	A	T	G	C	A_3	T_3	G_3	C_3	AT %	GC %	GC_1 %	GC_2 %	GC_3 %	AT_3 %	GC_{12} %
1	817	580	982	1014	161	165	377	428	41.2	58.8	56.6	48.7	71.2	28.8	52.7
2	955	745	935	914	280	292	302	309	47.9	52.1	56.8	47.8	51.6	48.4	52.3
3	324	260	437	422	24	50	197	210	40.5	59.5	57.6	36.4	84.6	15.4	47
4	104	85	124	131	18	21	55	54	42.6	57.4	50.7	48	73.6	26.4	49.3
5	272	184	171	189	96	81	50	45	55.9	44.1	50.4	47.1	34.9	65.1	48.7
6	609	429	711	759	128	126	265	317	41.4	58.6	57.9	48.3	69.6	30.4	53.1
7	114	132	238	236	13	28	98	101	34.2	65.8	64.2	50.4	82.9	17.1	57.3
8	206	215	372	548	37	52	136	222	31.4	68.6	70.7	55	80.1	19.9	62.9
9	458	385	479	631	77	137	172	265	43.2	56.8	56.1	47.3	67.1	32.9	51.7
10	849	623	1178	1166	96	119	521	536	38.6	61.4	58.8	42.4	83.1	16.9	50.6
11	660	579	526	536	230	220	162	155	53.8	46.2	53.1	44.1	41.3	58.7	48.6
12	871	552	1024	997	115	99	463	471	41.3	58.7	58.5	36.1	81.4	18.6	47.3
13	845	743	627	734	280	335	173	195	53.8	46.2	55.3	45.7	37.4	62.6	50.5
14	146	132	176	224	45	43	59	79	41	59	65	50.9	61.1	38.9	58
15	196	149	267	276	25	14	129	128	38.9	61.1	60.5	36.1	86.8	13.2	48.3
16	192	175	252	251	25	44	118	103	42.2	57.8	61.7	35.5	76.2	23.8	48.6
17	162	162	267	288	27	37	114	115	36.9	63.1	64.8	46.4	78.2	21.8	55.6
18	170	104	239	303	29	24	86	133	33.6	66.4	63.6	55.1	80.5	19.5	59.4
19	605	615	583	543	174	287	180	141	52	48	62.4	40.5	41	59	51.5
20	316	269	334	365	65	84	125	154	45.6	54.4	51.9	46.3	65.2	34.8	49.1
21	149	87	150	124	44	35	57	34	46.3	53.7	63.5	44.1	53.5	46.5	53.8
22	186	121	135	122	51	53	46	38	54.4	45.6	48.9	43.1	44.7	55.3	46
23	455	223	271	179	128	97	96	55	60.1	39.9	44.4	35.1	40.2	59.8	39.8
24	313	259	326	320	59	82	110	155	47	53	50.7	43.1	65.3	34.7	46.9
25	349	324	235	310	102	117	94	93	55.3	44.7	46.6	41.6	46.1	53.9	44.1
26	759	719	1120	1170	140	196	429	491	39.2	60.8	63	46.1	73.2	26.8	54.5
27	309	276	402	468	63	83	152	187	40.2	59.8	64.7	44.7	69.9	30.1	54.7
28	361	275	316	407	80	90	139	144	46.8	53.2	56.5	40.6	62.5	37.5	48.6
29	972	748	707	729	263	318	240	231	54.5	45.5	50.1	41.6	44.8	55.2	45.9
30	503	345	588	532	155	191	134	176	43.1	56.9	62.5	61	47.3	52.7	61.7

Table 3 continued

CDS N0.	A	T	G	C	A_3	T_3	G_3	C_3	AT %	GC %	GC_1 %	GC_2 %	GC_3 %	AT_3 %	GC_{12} %
31	113	95	227	282	8	14	86	131	29	71	68.6	53.6	90.8	9.2	61.1
32	575	483	699	712	104	116	290	313	42.9	57.1	56.4	41.8	73.3	26.7	49.1
33	146	118	106	80	42	55	27	26	58.7	41.3	49.3	39.3	35.3	64.7	44.3
34	678	526	656	609	166	164	247	246	48.8	51.2	53.2	40.6	59.9	40.1	46.9
35	379	294	545	363	88	171	102	166	42.6	57.4	60.2	61.3	50.9	49.1	60.7
36	759	486	503	448	226	215	171	120	56.7	43.3	54.2	35.9	39.8	60.2	45.1
37	379	272	331	242	126	117	97	68	53.2	46.8	63.5	36.5	40.4	59.6	50
38	99	27	113	85	22	10	50	26	38.9	61.1	55.6	57.4	70.4	29.6	56.5
39	86	26	86	81	30	12	28	23	40.1	59.9	57	67.7	54.8	45.2	62.4
40	111	92	165	145	9	22	68	72	39.6	60.4	57.9	41.5	81.9	18.1	49.7
41	339	235	373	409	71	55	145	181	42.3	57.7	58.2	42.7	72.1	27.9	50.4
42	226	129	299	342	32	16	144	140	35.6	64.4	61.7	45.8	85.5	14.5	53.8
43	846	768	678	639	220	290	219	248	55.1	44.9	48.4	38.6	47.8	52.2	43.5
44	207	146	134	83	67	60	37	26	61.9	38.1	47.4	33.7	33.2	66.8	40.5
45	352	289	451	438	53	73	190	194	41.9	58.1	58.8	40.2	75.3	24.7	49.5
46	111	100	143	123	19	22	54	64	44.2	55.8	54.7	38.4	74.2	25.8	46.5
47	212	144	381	376	25	24	143	179	32	68	66.3	50.9	86.8	13.2	58.6
48	195	126	291	360	17	23	137	147	33	67	66.7	46.6	87.7	12.3	56.6
49	1025	639	767	1031	291	260	318	285	48.1	51.9	58.1	45.4	52.3	47.7	51.8
50	463	374	317	247	150	175	79	63	59.7	40.3	50.7	39.6	30.4	69.6	45.2
51	1232	1155	913	927	353	433	284	339	56.5	43.5	48.1	38.3	44.2	55.8	43.2
52	396	253	273	338	103	91	98	128	51.5	48.5	52.6	39	53.8	46.2	45.8
53	406	252	497	552	44	45	221	259	38.5	61.5	57.5	42.5	84.4	15.6	50
54	355	346	467	572	95	94	175	216	40.3	59.7	60.3	51.4	67.4	32.6	55.9
55	613	410	428	472	182	167	147	145	53.2	46.8	53.5	41.3	45.6	54.4	47.4
56	319	225	345	431	54	50	136	200	41.2	58.8	55.5	44.5	76.4	23.6	50
57	103	92	199	227	21	22	69	95	31.4	68.6	62.3	64.3	79.2	20.8	63.3
58	278	198	433	486	40	52	163	210	34.1	65.9	66.2	51.2	80.2	19.8	58.7
59	544	436	443	422	127	180	163	145	53.1	46.9	52	38.5	50.1	49.9	45.3
60	580	468	847	808	134	150	324	293	38.8	61.2	68.4	46.8	68.5	31.5	57.6
61	185	132	150	103	55	48	46	41	55.6	44.4	52.1	35.3	45.8	54.2	43.7
62	420	465	751	737	55	118	304	314	37.3	62.7	66.1	43.9	78.1	21.9	55
63	1499	1505	1460	1779	516	635	424	506	48.1	51.9	53.7	57.2	44.7	55.3	55.5
64	284	244	346	479	48	69	120	214	39	61	56.8	52.1	74.1	25.9	54.4
65	153	121	228	224	24	29	96	93	37.7	62.3	61.6	47.1	78.1	21.9	54.3
66	1043	911	669	584	342	351	209	167	60.9	39.1	48.9	33.1	35.2	64.8	41
67	195	185	281	281	30	34	115	135	40.3	59.7	60.8	38.5	79.6	20.4	49.7
68	441	381	302	379	123	148	109	121	54.7	45.3	49.1	40.9	45.9	54.1	45
69	439	371	345	363	102	121	139	144	53.4	46.6	50.8	33.2	55.9	44.1	42
70	50	38	49	16	18	11	15	7	57.5	42.5	45.1	39.2	43.1	56.9	42.2
71	530	469	461	487	134	169	169	177	51.3	48.7	50.4	42.4	53.3	46.7	46.4
72	591	493	387	389	197	216	94	113	58.3	41.7	49.2	42.6	33.4	66.6	45.9
73	638	631	962	988	100	132	382	459	39.4	60.6	59.2	44.2	78.4	21.6	51.7
74	2073	1988	1547	1436	639	789	464	456	57.7	42.3	48.3	39.6	39.2	60.8	43.9
75	412	467	721	764	54	116	287	331	37.2	62.8	60.5	49.5	78.4	21.6	55
76	383	241	297	336	108	122	119	70	49.6	50.4	62.1	43.9	45.1	54.9	53
77	73	70	103	99	16	11	45	43	41.4	58.6	60.9	38.3	76.5	23.5	49.6
78	633	428	388	465	194	201	126	117	55.4	44.6	51.6	44	38.1	61.9	47.8
79	331	269	230	271	102	112	69	84	54.5	45.5	55.9	39	41.7	58.3	47.4
80	139	101	258	276	23	24	80	131	31	69	65.5	59.7	81.8	18.2	62.6
81	2589	1550	1642	1311	876	748	509	231	58.4	41.6	58.3	35.3	31.3	68.7	46.8
82	1183	940	1083	1015	311	355	383	358	50.3	49.7	53.7	42.7	52.7	47.3	48.2
M	520.8	416.6	502.2	517.6	134.9	150.9	178.7	187.9	45.9	54.1	56.7	44.8	45.1	54.9	50.8
SD	537.6	431.8	422.4	440.3	172.4	189.2	143.0	146.0	.086	.086	.067	.073	0.20	0.20	0.057

M: mean; **SD**: standard deviation; GC_{12}: average of GC contents of first and second codon positions

4.2. Nucleotide composition in tumour suppressor genes

We analyzed the nucleotide composition of the coding sequences of sixty-three tumour suppressor genes (Table 4) which revealed that mean value of A (767.2) was the highest followed by G (655.2), C (654.4) and T (613.2) among all the coding sequences. The overall mean percentage of GC (ranged from 35.8% to 74.7%) and AT (ranged from 25.3% to 64.2%) contents was 50.7% and 49.3% respectively with an equal standard deviation of 0.09. The GC contents at different codon positions were calculated and it was observed that the GC content at each codon position variation among the genes (Figure 5).

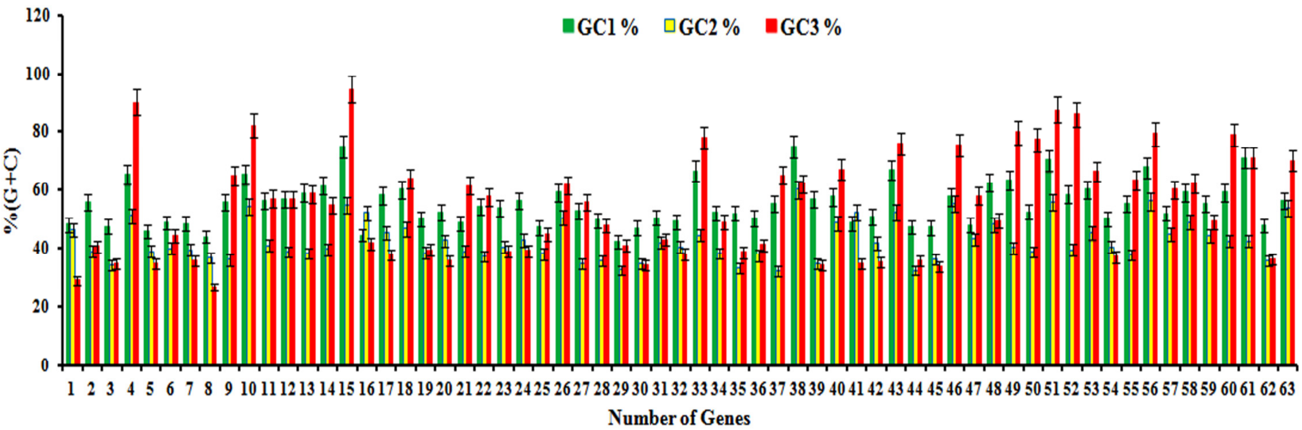


Figure 5: Percentage of GC contents at three codon positions in the coding sequences of tumour suppressor genes

The nucleotide composition at the third position of codon (A_3,T_3,G_3,C_3) showed that the mean value of T_3 (233.5) was the highest followed by C_3 (227.7), G_3 (220.8) and A_3 (214.6). The GC_3 values (ranged from 26.7%-94.7%, mean=54.4%, SD=0.17) was compared with that of AT_3 values (ranged from 5.3%-73.3%, mean=45.6%, SD=0.17) in the coding sequences of tumour suppressor genes.

The average percentage of GC contents of the first and second codon position (GC_{12}) was found to range from 37.4% to 67.6% with a mean value of 48.9% and a standard deviation (SD) of 0.06. Therefore, from the overall nucleotide composition analysis, it could be concluded that GC–ending and AT–ending codons are almost equally distributed in the coding sequences of tumour suppressor genes.

Table 4: Nucleotide composition analysis in the coding sequences of sixty-three tumour suppressor genes

CDS. N0.	A	T	G	C	A_3	T_3	G_3	C_3	AT %	GC %	GC_1 %	GC_2 %	GC_3 %	AT_3 %	GC_{12} %
1	2940	2075	1744	1773	2940	2075	1744	1773	58.8	41.2	48.1	46.4	29.1	70.9	47.3
2	1442	1100	1110	983	1442	1100	1110	983	54.8	45.2	55.9	39	40.6	59.4	47.5
3	2981	2629	1923	1638	2981	2629	1923	1638	61.2	38.8	47.5	34.2	34.8	65.2	40.9
4	484	354	883	964	484	354	883	964	31.2	68.8	65.3	51.1	90.1	9.9	58.2
5	1414	1144	869	827	1414	1144	869	827	60.1	39.9	45.9	39	34.7	65.3	42.5
6	468	423	366	342	468	423	366	342	55.7	44.3	48.8	39.8	44.3	55.7	44.3
7	1931	1356	1241	1064	1931	1356	1241	1064	58.8	41.2	48.4	39.3	35.9	64.1	43.9
8	3767	2823	1882	1785	2988	2205	1426	1331	64.2	35.8	44	36.6	26.7	73.3	40.3
9	612	462	623	550	612	462	623	550	47.8	52.2	55.8	36	64.8	35.2	45.9
10	366	238	561	683	366	238	561	683	32.7	67.3	65.3	54.5	82.1	17.9	59.9
11	709	577	651	712	709	577	651	712	48.5	51.5	56.4	41	57	43	48.7
12	669	508	615	599	669	508	615	599	49.2	50.8	56.8	38.8	56.7	43.3	47.8
13	251	220	258	252	251	220	258	252	48	52	59	38.2	58.7	41.3	48.6
14	130	113	155	109	130	113	155	109	47.9	52.1	61.5	39.6	55	45	50.6
15	168	104	372	433	168	104	372	433	25.3	74.7	74.7	54.9	94.7	5.3	64.8
16	498	452	422	389	499	452	421	389	53.9	46.1	44.5	52.1	41.6	58.4	48.3
17	320	200	225	239	320	200	225	239	52.8	47.2	58.2	45.4	37.8	62.2	51.8
18	1894	1266	1864	2305	1894	1266	1864	2305	43.1	56.9	60.2	46.6	63.9	36.1	53.4
19	877	765	661	559	877	765	661	559	57.4	42.6	50.2	38.4	39.3	60.7	44.3
20	574	467	465	339	574	467	465	339	56.4	43.6	52.4	42.6	35.8	64.2	47.5
21	573	554	537	577	573	554	537	577	50.3	49.7	48.7	39	61.4	38.6	43.9
22	539	544	532	542	539	544	532	542	50.2	49.8	54.4	37.3	57.7	42.3	45.9
23	681	501	548	394	681	501	548	394	55.6	44.4	53.8	40.3	39	61	47.1
24	445	383	382	323	445	383	382	323	54	46	56.2	42.7	39.1	60.9	49.5
25	897	792	682	611	897	792	682	611	56.6	43.4	47.1	38.1	44.9	55.1	42.6
26	83	65	105	92	83	65	105	92	42.9	57.1	59.1	50.4	61.7	38.3	54.8
27	480	429	425	409	480	429	425	409	52.2	47.8	52.8	34.9	55.8	44.2	43.9
28	382	310	305	248	382	310	305	248	55.6	44.4	49.6	35.9	47.7	52.3	42.8
29	163	121	81	97	163	121	81	97	61.5	38.5	42.2	32.5	40.9	59.1	37.4
30	1143	940	723	593	1143	940	723	593	61.3	38.7	47	34.7	34.4	65.6	40.9
31	363	295	283	259	363	295	283	259	54.8	45.2	50.5	42	43	57	46.3
32	464	383	347	279	464	383	347	279	57.5	42.5	49.1	40.5	37.9	62.1	44.8
33	361	317	569	586	364	320	575	589	37	63	66.6	44.5	77.9	22.1	55.6
34	660	556	554	501	660	556	554	501	53.5	46.5	52.2	38.3	48.9	51.1	45.3
35	890	757	649	509	890	757	649	509	58.7	41.3	52	33.3	38.6	61.4	42.7
36	2544	2317	1806	1853	2544	2317	1806	1853	57.1	42.9	50.3	37.5	41.1	58.9	43.9
37	532	346	510	400	532	346	510	400	49.1	50.9	55.4	32.2	65.1	34.9	43.8
38	1426	1198	2353	2691	79	83	136	128	34.2	65.8	75	60.1	62.2	37.8	67.6
39	316	197	224	148	316	197	224	148	58	42	56.9	34.9	34.2	65.8	45.9
40	440	368	475	631	440	368	475	631	42.2	57.8	57.7	48.6	67.1	32.9	53.2

Table 4 continued

CDS N0.	A	T	G	C	A_3	T_3	G_3	C_3	AT %	GC %	GC_1 %	GC_2 %	GC_3 %	AT_3 %	GC_{12} %
41	814	699	616	634	247	354	143	177	54.8	45.2	48.6	52.3	34.7	65.3	50.5
42	1148	895	746	772	372	396	211	208	57.4	42.6	50.8	41.8	35.3	64.7	46.3
43	446	376	682	842	85	105	264	328	35.0	65.0	66.8	52.4	75.7	24.3	59.6
44	428	318	244	222	129	131	67	77	61.6	38.4	47.3	32.4	35.6	64.4	39.9
45	943	757	529	558	315	301	171	142	61.0	39.0	47.1	36.2	33.7	66.3	41.7
46	292	244	373	534	59	59	142	221	37.1	62.9	57.6	55.5	75.5	24.5	56.6
47	236	188	206	213	53	65	73	90	50.3	49.7	48.0	43.1	58.0	42.0	45.6
48	84	140	129	127	25	56	39	40	46.7	53.3	62.5	48.1	49.4	50.6	55.3
49	1230	732	1534	1550	163	176	693	650	38.9	61.1	63.3	40.2	79.8	20.2	51.8
50	292	204	321	314	43	43	149	142	43.9	56.1	52.5	38.7	77.2	22.8	45.6
51	85	98	194	259	14	12	73	113	28.2	71.2	70.3	55.7	87.7	12.3	63.0
52	296	208	414	384	30	31	186	187	38.7	61.3	58.5	39.4	85.9	14.1	49.0
53	335	287	404	429	71	92	147	175	42.7	57.3	60.2	45.2	66.4	33.6	52.7
54	745	532	505	438	229	237	151	123	57.5	42.5	50.1	40.3	37.0	63.0	45.2
55	532	380	518	478	120	112	196	208	47.5	52.2	55.3	37.7	63.5	36.5	46.5
56	372	267	610	728	72	65	216	306	32.3	67.7	67.7	56.1	79.2	20.8	61.9
57	637	494	592	650	144	172	208	267	47.7	52.3	52.0	45.0	60.1	39.9	48.5
58	276	234	307	365	62	86	115	131	43.1	56.9	59.1	49.0	62.4	37.6	54.1
59	332	356	287	390	96	135	104	120	50.4	49.6	55.2	44.4	49.2	50.8	49.8
60	1093	1040	1551	1671	162	214	670	739	39.8	60.2	59.2	42.4	78.9	21.1	50.8
61	141	106	212	183	33	29	79	73	38.5	61.5	71.0	42.5	71.0	29.0	56.8
62	1400	1188	952	759	412	503	293	225	60.2	39.8	47.7	35.6	36.1	63.9	41.7
63	301	238	371	440	61	74	143	172	39.9	60.1	56.4	53.8	70.0	30.0	55.1
M	767.2	613.2	655.2	654.4	214.6	233.5	220.8	227.7	49.3	50.7	55.3	42.5	54.4	45.6	48.9
SD	741.9	585.9	513.3	544.5	245.5	258.9	171.9	184.1	0.09	0.09	0.07	0.06	.178	.178	.065

M: mean; **SD**: standard deviation; GC_{12}: average of GC contents of first and second codon positions

4.3. Nucleotide skewness among proto-oncogenes/oncogenes

The variation in base composition within each coding sequence was calculated from their differences in usage for the GC, AT, keto, amino, purine, and pyrimidine bases of the eighty-two proto-oncogenes/oncogenes (Table 5). The magnitude of skew values revealed that base composition bias is linked to transcription process (Fujimori *et al.* 2005, Touchon and Rocha 2008). In our analysis, forty-three proto-oncogenes/oncogenes out of eighty-two showed negative GC skew value and the remaining thirty-nine showed positive GC skew value (Figure 6). Positive GC skew indicates richness of G over C and negative GC skew reveals richness of C over G (Tillier and Collins 2000).

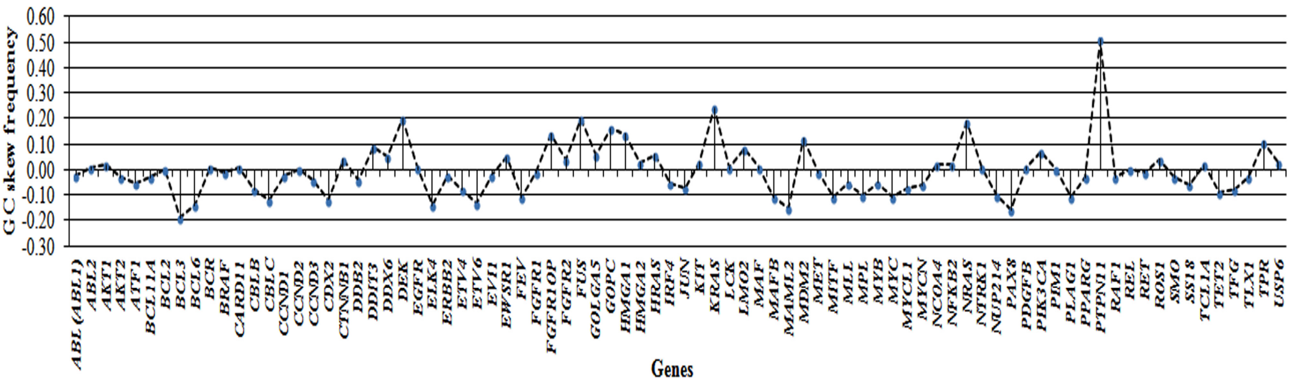


Figure 6: GC skew values of eighty-two proto-oncogenes/oncogenes

In case of AT3 skew, nearly all the selected proto-oncogenes/oncogenes showed positive value which indicates richness of A over T (Tillier and Collins 2000). Negative value of keto skew was observed as reported earlier by Zhang & Gerstein in human genome (Zhang and Gerstein 2003). The keto skew values were negative for forty-three proto-oncogenes/oncogenes. A wide variation of purine, pyrimidine and amino skews was observed which might affect the expression patterns of these genes in the cell.

Table 5: Skew values of eighty-two proto-oncogenes/oncogenes

CDS NO.	GC SKEW (G-C/G+C)	AT SKEW (A-T/A+T)	KETO SKEW (A-C/A+C)	PURINE SKEW (A-G/A+G)	PYRIMIDINE SKEW (T-C/T+C)	AMINO SKEW (T-G/T+G)
1	-0.02	0.17	-0.11	-0.09	-0.27	-0.26
2	0.01	0.12	0.02	0.01	-0.10	-0.11
3	0.02	0.11	-0.13	-0.15	-0.24	-0.25
4	-0.03	0.10	-0.11	-0.09	-0.21	-0.19
5	-0.05	0.19	0.18	0.23	-0.01	0.04
6	-0.03	0.17	-0.11	-0.08	-0.28	-0.25
7	0.00	-0.07	-0.35	-0.35	-0.28	-0.29
8	-0.19	-0.02	-0.45	-0.29	-0.44	-0.27
9	-0.14	0.09	-0.16	-0.02	-0.24	-0.11
10	0.01	0.15	-0.16	-0.16	-0.30	-0.31
11	-0.01	0.07	0.10	0.11	0.04	0.05
12	0.01	0.22	-0.07	-0.08	-0.29	-0.30
13	-0.08	0.06	0.07	0.15	0.01	0.08
14	-0.12	0.05	-0.21	-0.09	-0.26	-0.14
15	-0.02	0.14	-0.17	-0.15	-0.30	-0.28
16	0.00	0.05	-0.13	-0.14	-0.18	-0.18
17	-0.04	0.00	-0.28	-0.24	-0.28	-0.24
18	-0.12	0.24	-0.28	-0.17	-0.49	-0.39
19	0.04	-0.01	0.06	0.02	0.06	0.03
20	-0.04	0.08	-0.07	-0.03	-0.15	-0.11
21	0.09	0.26	0.09	0.00	-0.18	-0.27
22	0.05	0.21	0.21	0.16	0.00	-0.05
23	0.20	0.34	0.44	0.25	0.11	-0.10
24	0.01	0.09	-0.01	-0.02	-0.11	-0.11
25	-0.14	0.04	0.06	0.20	0.02	0.16
26	-0.02	0.03	-0.21	-0.19	-0.24	-0.22
27	-0.08	0.06	-0.20	-0.13	-0.26	-0.19
28	-0.13	0.14	-0.06	0.07	-0.19	-0.07
29	-0.02	0.13	0.14	0.16	0.01	0.03
30	0.05	0.19	-0.03	-0.08	-0.21	-0.26

Table 5 continued

CDS NO.	GC SKEW (G-C/G+C)	AT SKEW (A-T/A+T)	KETO SKEW (A-C/A+C)	PURINE SKEW (A-G/A+G)	PYRIMIDINE SKEW (T-C/T+C)	AMINO SKEW (T-G/T+G)
31	-0.11	0.09	-0.43	-0.34	-0.50	-0.41
32	-0.01	0.09	-0.11	-0.10	-0.19	-0.18
33	0.14	0.11	0.29	0.16	0.19	0.05
34	0.04	0.13	0.05	0.02	-0.07	-0.11
35	0.20	0.13	0.02	-0.18	-0.11	-0.30
36	0.06	0.22	0.26	0.20	0.04	-0.02
37	0.16	0.16	0.22	0.07	0.06	-0.10
38	0.14	0.57	0.08	-0.07	-0.52	-0.61
39	0.03	0.54	0.03	0.00	-0.51	-0.54
40	0.06	0.09	-0.13	-0.20	-0.22	-0.28
41	-0.05	0.18	-0.09	-0.05	-0.27	-0.23
42	-0.07	0.27	-0.20	-0.14	-0.45	-0.40
43	0.03	0.05	0.14	0.11	0.09	0.06
44	0.24	0.17	0.43	0.21	0.28	0.04
45	0.01	0.10	-0.11	-0.12	-0.20	-0.22
46	0.08	0.05	-0.05	-0.13	-0.10	-0.18
47	0.01	0.19	-0.28	-0.28	-0.45	-0.45
48	-0.11	0.21	-0.30	-0.20	-0.48	-0.40
49	-0.15	0.23	0.00	0.14	-0.23	-0.09
50	0.12	0.11	0.30	0.19	0.20	0.08
51	-0.01	0.03	0.14	0.15	0.11	0.12
52	-0.11	0.22	0.08	0.18	-0.14	-0.04
53	-0.05	0.12	0.07	0.12	-0.05	0.00
54	-0.10	0.01	-0.23	-0.14	-0.25	-0.15
55	-0.05	0.20	0.13	0.18	-0.07	-0.02
56	-0.11	0.17	-0.15	-0.04	-0.31	-0.21
57	-0.07	0.06	-0.38	-0.32	-0.42	-0.37
58	-0.06	0.17	-0.27	-0.22	-0.42	-0.37
59	0.02	0.11	0.13	0.10	0.02	-0.01
60	0.02	0.11	-0.16	-0.19	-0.27	-0.29
61	0.19	0.17	0.28	0.10	0.12	-0.06
62	0.01	-0.05	-0.27	-0.28	-0.23	-0.24
63	-0.10	0.00	-0.09	0.01	-0.08	0.02
64	-0.16	0.08	-0.26	-0.10	-0.33	-0.17
65	0.01	0.12	-0.19	-0.20	-0.30	-0.31
66	0.07	0.07	0.28	0.22	0.22	0.15
67	0.00	0.03	-0.18	-0.18	-0.21	-0.21
68	-0.11	0.07	0.08	0.19	0.00	0.12
69	-0.03	0.08	0.09	0.12	0.01	0.04
70	0.51	0.14	0.52	0.01	0.41	-0.13
71	-0.03	0.06	0.04	0.07	-0.02	0.01
72	0.00	0.09	0.21	0.21	0.12	0.12
73	-0.01	0.01	-0.22	-0.20	-0.22	-0.21
74	0.04	0.02	0.18	0.15	0.16	0.12
75	-0.03	-0.06	-0.30	-0.27	-0.24	-0.21
76	-0.06	0.23	0.07	0.13	-0.16	-0.10
77	0.02	0.02	-0.15	-0.17	-0.17	-0.19
78	-0.09	0.19	0.15	0.24	-0.04	0.05
79	-0.08	0.10	0.10	0.18	0.00	0.08
80	-0.03	0.16	-0.33	-0.30	-0.46	-0.44
81	0.11	0.25	0.33	0.22	0.08	-0.03
82	0.03	0.11	0.08	0.04	-0.04	-0.07

4.4. Nucleotide skewness among tumour suppressor genes

The variation in base composition in each coding sequence of tumour suppressor gene was calculated from their differences in usage for the GC, AT, keto, amino, purine, and pyrimidine bases (Table 6). Skew value reveals that the base composition bias is linked to transcription process (Fujimori *et al*. 2005, Touchon and Rocha 2008). In our analysis, twenty-nine tumour suppressor genes out of sixty-three showed the negative GC skew value and the remaining twenty-four showed positive GC skew value (Figure 7).

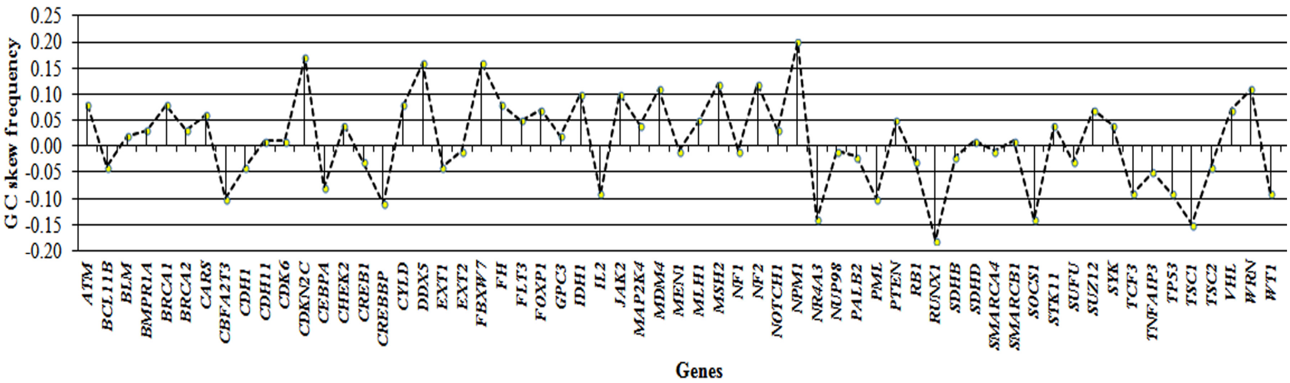


Figure 7: GC skew values of sixty-three tumour suppressor genes

Positive GC skew indicates the richness of G over C whereas negative GC skew reveals the richness of C over G for tumour suppressor genes (Tillier and Collins 2000). In case of AT3 skew, nearly all the selected tumour suppressor genes showed positive value which depicts the richness of A over T (Tillier and Collins 2000). Negative value of keto skew was observed as reported earlier by Zhang & Gerstein in human genome (Zhang and Gerstein 2003). The keto skew values were negative for twenty-three tumour suppressor genes out of sixty-three. A wide variation of purine, pyrimidine and amino skew was observed which might affect the expression patterns of these genes.

Table 6: Skew values of sixty-three tumour suppressor genes

CDS NO.	GC SKEW (G-C/G+C)	AT SKEW (A-T/A+T)	KETO SKEW (A-C/A+C)	PURINE SKEW (A-G/A+G)	PYRIMIDINE SKEW (T-C/T+C)	AMINO SKEW (T-G/T+G)
1	-0.01	0.17	0.25	0.26	0.08	0.09
2	0.06	0.13	0.19	0.13	0.06	0.00
3	0.08	0.06	0.29	0.22	0.23	0.16
4	-0.04	0.16	-0.33	-0.29	-0.46	-0.43
5	0.02	0.11	0.26	0.24	0.16	0.14
6	0.03	0.05	0.16	0.12	0.11	0.07
7	0.08	0.17	0.29	0.22	0.12	0.04
8	0.03	0.14	0.36	0.33	0.23	0.20
9	0.06	0.14	0.05	-0.01	-0.09	-0.15
10	-0.10	0.21	-0.30	-0.21	-0.48	-0.40
11	-0.04	0.10	0.00	0.04	-0.10	-0.06
12	0.01	0.14	0.06	0.04	-0.08	-0.10
13	0.01	0.07	0.00	-0.01	-0.07	-0.08
14	0.17	0.07	0.09	-0.09	0.02	-0.16
15	-0.08	0.24	-0.44	-0.38	-0.61	-0.56
16	0.04	0.05	0.12	0.08	0.07	0.04
17	-0.03	0.23	0.14	0.17	-0.09	-0.06
18	-0.11	0.20	-0.10	0.01	-0.29	-0.19
19	0.08	0.07	0.22	0.14	0.16	0.07
20	0.16	0.10	0.26	0.10	0.16	0.00
21	-0.04	0.02	0.00	0.03	-0.02	0.02
22	-0.01	0.00	0.00	0.01	0.00	0.01
23	0.16	0.15	0.27	0.11	0.12	-0.04
24	0.08	0.07	0.16	0.08	0.08	0.00
25	0.05	0.06	0.19	0.14	0.13	0.07
26	0.07	0.12	-0.05	-0.12	-0.17	-0.24
27	0.02	0.06	0.08	0.06	0.02	0.00
28	0.10	0.10	0.21	0.11	0.11	0.01
29	-0.09	0.15	0.25	0.34	0.11	0.20
30	0.10	0.10	0.32	0.23	0.23	0.13
31	0.04	0.10	0.17	0.12	0.06	0.02
32	0.11	0.10	0.25	0.14	0.16	0.05
33	-0.01	0.06	-0.24	-0.22	-0.30	-0.28
34	0.05	0.09	0.14	0.09	0.05	0.00
35	0.12	0.08	0.27	0.16	0.20	0.08
36	-0.01	0.05	0.16	0.17	0.11	0.12
37	0.12	0.21	0.14	0.02	-0.07	-0.19
38	-0.07	0.09	-0.31	-0.25	-0.38	-0.33
39	0.20	0.23	0.36	0.17	0.14	-0.06
40	-0.14	0.09	-0.18	-0.04	-0.26	-0.13

Continued

CDS NO.	GC SKEW (G-C/G+C)	AT SKEW (A-T/A+T)	KETO SKEW (A-C/A+C)	PURINE SKEW (A-G/A+G)	PYRIMIDINE SKEW (T-C/T+C)	AMINO SKEW (T-G/T+G)
41	-0.01	0.08	0.12	0.14	0.05	0.06
42	-0.02	0.12	0.20	0.21	0.07	0.09
43	-0.10	0.09	-0.31	-0.21	-0.38	-0.29
44	0.05	0.15	0.32	0.27	0.18	0.13
45	-0.03	0.11	0.26	0.28	0.15	0.18
46	-0.18	0.09	-0.29	-0.12	-0.37	-0.21
47	-0.02	0.11	0.05	0.07	-0.06	-0.05
48	0.01	-0.25	-0.20	-0.21	0.05	0.04
49	-0.01	0.25	-0.12	-0.11	-0.36	-0.35
50	0.01	0.18	-0.04	-0.05	-0.21	-0.22
51	-0.14	-0.07	-0.51	-0.39	-0.45	-0.33
52	0.04	0.17	-0.13	-0.17	-0.30	-0.33
53	-0.03	0.08	-0.12	-0.09	-0.20	-0.17
54	0.07	0.17	0.26	0.19	0.10	0.03
55	0.04	0.17	0.05	0.01	-0.11	-0.15
56	-0.09	0.16	-0.32	-0.24	-0.46	-0.39
57	-0.05	0.13	-0.01	0.04	-0.14	-0.09
58	-0.09	0.08	-0.14	-0.05	-0.22	-0.13
59	-0.15	-0.03	-0.08	0.07	-0.05	0.11
60	-0.04	0.02	-0.21	-0.17	-0.23	-0.20
61	0.07	0.14	-0.13	-0.20	-0.27	-0.33
62	0.11	0.08	0.30	0.19	0.22	0.11
63	-0.09	0.12	-0.19	-0.10	-0.30	-0.22

4.5. Codon usage patterns in the coding sequences of proto-oncogenes/oncogenes

The correlation coefficient between codon usage and GC bias was analyzed using a heat map (Figure 8) in order to find out the relationship between the codon usage variation and GC constraints among the selected coding sequences of human proto-oncogenes/oncogenes.

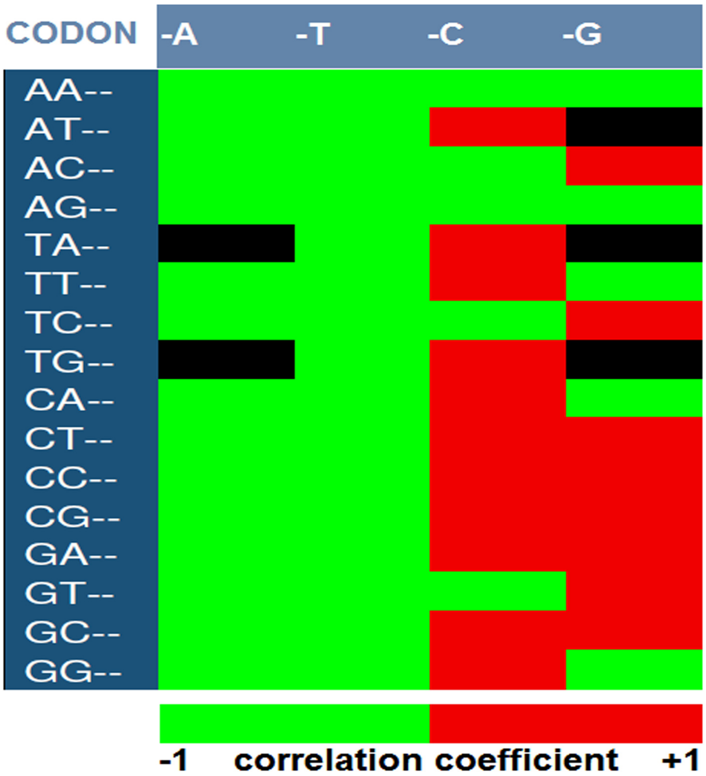


Figure 8: Heat maps of correlation coefficient of codons with GC_{3s}. The color coding red represents the positive correlation, green as negative correlation. The black fields are stop codons (TAA, TAG, TGA) and non-degenerate codons (ATG, TGG) in the coding sequences of the 82 proto-oncogenes/oncogenes under study.

In our analysis, nearly all the codons ending with G/C base showed a strong positive correlation ($p<0.01$) with GC bias and all the A/T–ending codons showed negative correlation with GC bias. But, five C–ending codons (AAC, ACC, AGC, TCC, GTC) as well as five G –ending codons (AAG, AGG, TTG, CAG, GGG) showed a negative correlation between codon usage and GC bias. These codons display non-linear upward usage profiles due to the effect of GC bias (Figure 9). The codon TTG encoding leucine amino acid exhibited significant negative correlation ($p<0.01$) with GC_{3s}.

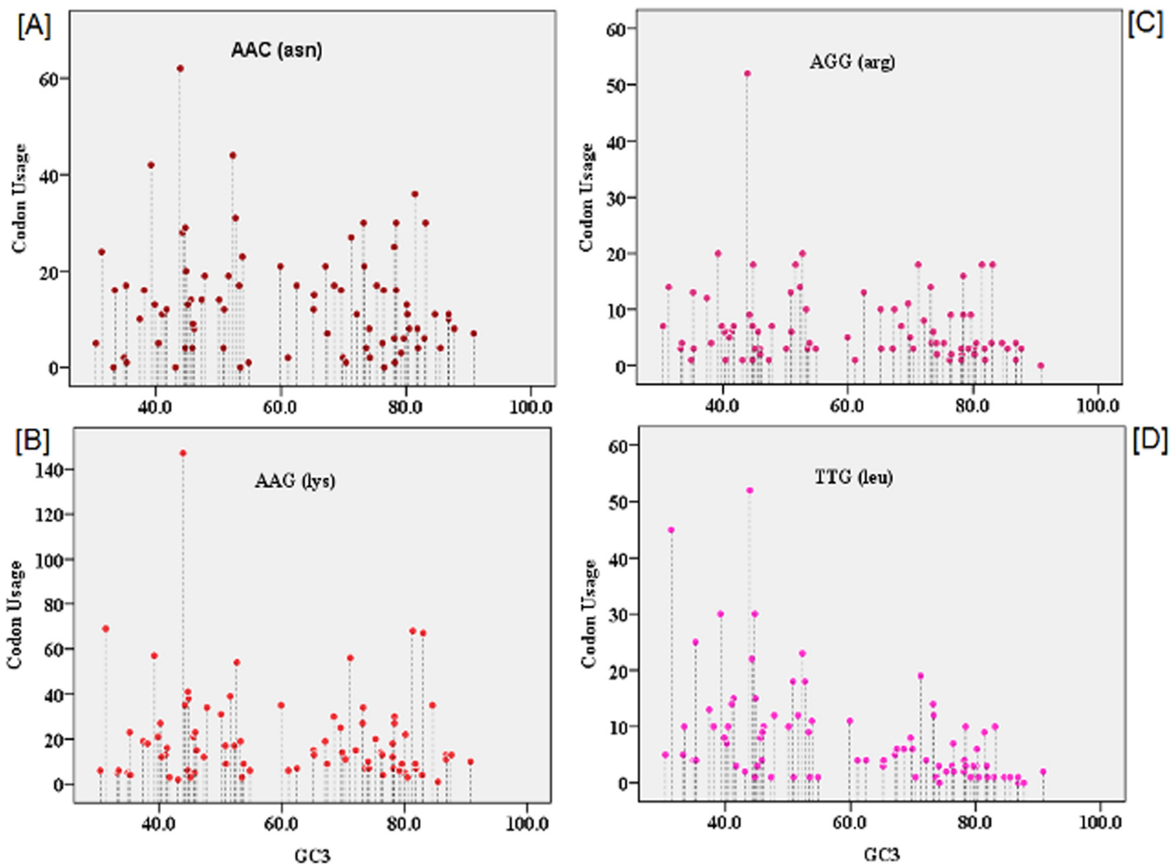


Figure 9: Scatter plots of negatively correlated codons with GC3s. [A] Scatter plots of codon usage frequency for the asparagine codon AAC *vs* GC_{3s}. [B] Scatter plots of codon usage frequency for the lysine codon AAG *vs* GC_{3s}. [C] Scatter plots of codon usage frequency for the arginine codon AGG *vs* GC_{3s}. [D] Scatter plots of codon usage frequency for the leucine codon TTG *vs* GC_{3s}.

4.6. Codon usage patterns in the coding sequences of tumour suppressor genes

To investigate the relationship between the codon usage variation and GC constraints among the selected coding sequences of human tumour suppressor genes, we analyzed the Pearson's correlation coefficient between codon usage and GC bias using heat map (Figure 10).

In our analysis, nearly all G- and C- ending codons showed strong positive correlation with GC bias, and conversely for A- and T- ending codons. However, one C–ending codon AAC (asparagine) and three G-ending codons namely AAG (lysine), AGG (arginine) and TTG (leucine) showed negative correlation between codon usage and GC bias. These codons exhibit non-linear upward usage profiles as a function of GC bias (Figure 11). The codons AGG and TTG encoding amino acid arginine and leucine respectively displayed significant negative correlation ($p<0.01$) with GC_{3S}.

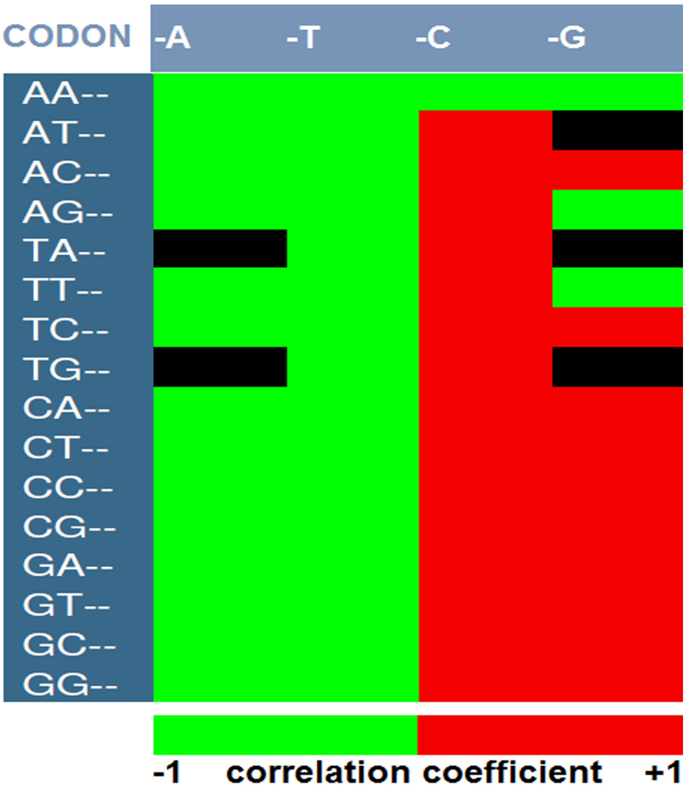


Figure 10: Heat maps of correlation coefficient of codons with GC_{3s}. The color coding red represents the positive correlation, green as negative correlation. The black fields are stop codons (TAA, TAG, TGA) and non-degenerate codons (ATG, TGG) in the coding sequences of the 63 tumour suppressor gene under study.

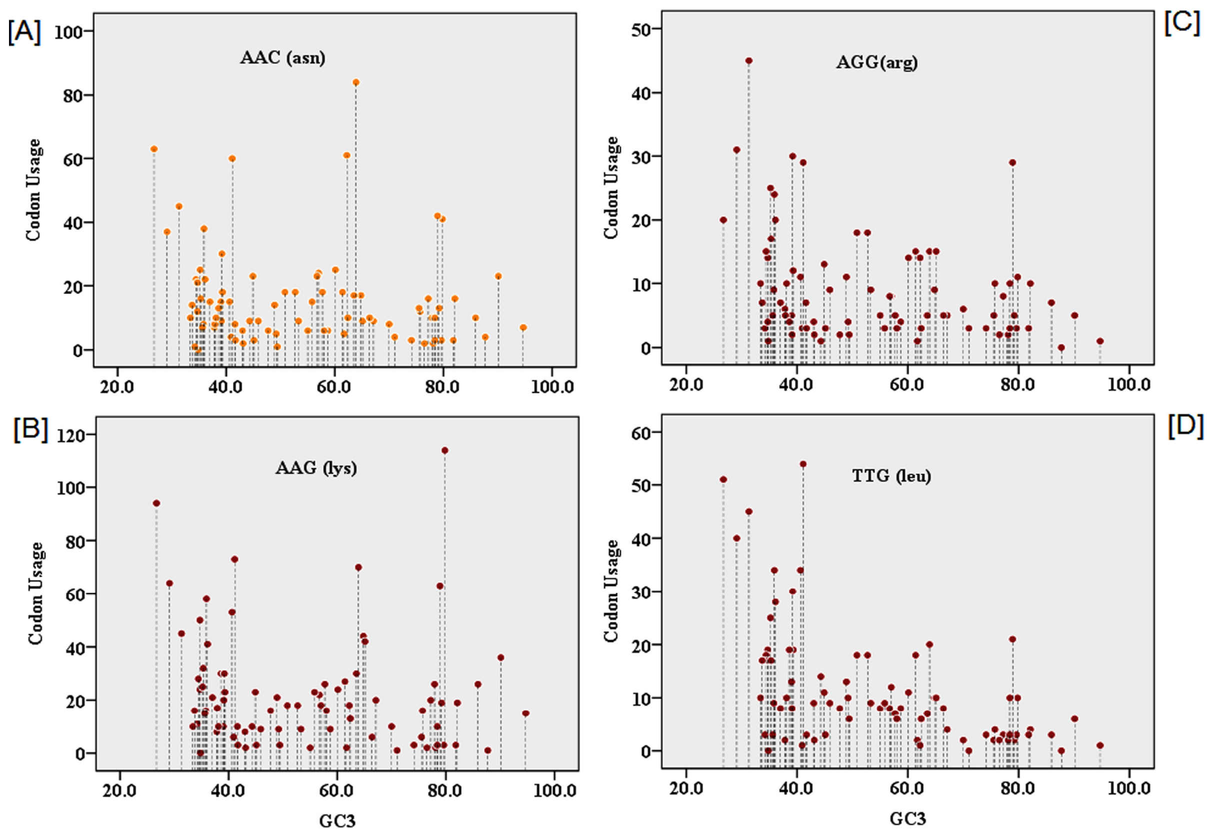


Figure 11: Scatter plots of negatively correlated codons with GC3s. [A] Scatter plots of codon usage frequency for the asparagine codon AAC *vs* GC_{3s}. [B] Scatter plots of codon usage frequency for the lysine codon AAG *vs* GC_{3s}. [C] Scatter plots of codon usage frequency for the arginine codon AGG *vs* GC_{3s}. [D] Scatter plots of codon usage frequency for the leucine codon TTG *vs* GC_{3s}.

4.7. Relationship between codon usage bias and compositional constraints in the coding sequences of proto-oncogenes/oncogenes

In order to investigate the relationship between codon usage variation and compositional constraints, we accomplished Pearson's correlation analysis between the values of A, T, G, C, GC and A_3, T_3, G_3, C_3, GC_3 values, respectively (Table 7).

Significant positive correlation was observed in nucleotide composition, suggesting that nucleotide constraint may influence the codon usage pattern in proto-oncogenes/oncogenes. In addition, a neutrality plot was drawn (Figure 12) to estimate the extent of directional mutation pressure against selection in the codon usage bias of human proto-oncogenes/oncogenes (Sueoka 1988). In neutrality plot analysis, when a gene is placed on the slope of unity there exists a significant correlation between its GC_{12} (the average value of GC_1 and GC_2) and $GC_{3,}$ indicating that the gene is under neutral mutation through random selective pressure. But if the gene is under directional mutation pressure, the gene would fall below the slope of unity, i.e. closer to X-axis and further from the Y-axis. Therefore, a regression line with a slope less than 1, indicates that a non neutral mutation pressure affects the codon usage among different genes of the same genome (Necsulea and Lobry 2006, Sueoka and Kawanishi 2000). In our analysis, we observed a significant positive correlation (Pearson $r=0.543$, $p<0.01$) between GC_{12} and GC_{3s} which suggested that intragenomic GC mutation bias plays an important role in shaping the codon usage pattern of these genes. The neutrality plot (as shown in figure 12) reveals the results of equilibrium coefficient of mutation and selection. In the plot (Figure 12) some of the points are located in the diagonal distribution and the values of GC_3 are in a wide range of distribution, indicating GC_{12} and GC_3 are definitely the affects of mutation bias model (Sueoka 1988). Accordingly, mutation bias might just play a minor role in shaping the codon bias, whereas the natural selection seems to probably dominate the codon bias. Further, we quantify the natural selection and the mutation pressure using regression coefficient. The regression coefficient of GC_{12} to GC_3 of human proto-oncogene/oncogene is 0.176, indicating the relative neutrality is 17.6 %, while the relative constraint is 82.4 % for GC_3. These results suggest that natural selection played a major role while mutation pressure played a minor role in the codon usage patterns of these genes.

Table 7: Correlation analysis between A, T, G, C, GC contents and A_3, T_3, G_3, C_3, GC_3 contents in all the selected coding sequences of proto-oncogenes/oncogenes

	A_3	T_3	G_3	C_3	GC_3
A	r=0.902**	r=0.876**	r=0.799**	r=0.727**	r=-0.302*
T	r=0.883**	r=0.915**	r=0.766**	r=0.717**	r=-0.301*
G	r=0.638**	r=0.660**	r=0.958**	r=0.942**	r=0.049
C	r=0.624**	r=0.632**	r=0.952**	r=0.953**	r=0.113
GC	r=-0.551**	r=-0.508**	r=0.140	r=0.258	r=0.912**

**p<0.01 *p<0.05

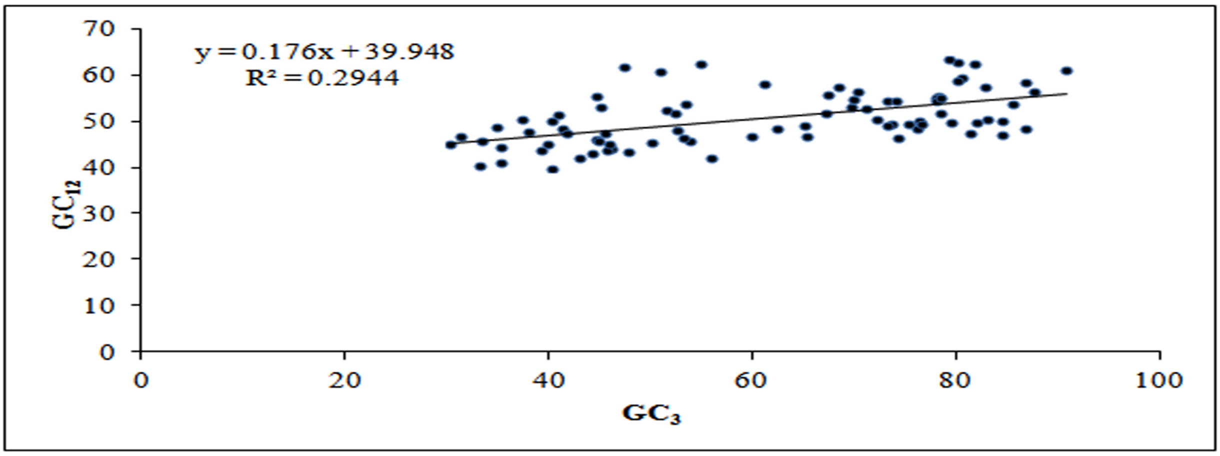


Figure 12: Neutrality plot of GC_{12} versus GC_3. The regression line is y = 0.176x + 39.48; R^2 = 0.294; GC_{12}: average GC content at first and second codon positions.

4.8. Relationship between codon usage bias and compositional constraints in the coding sequences of tumour suppressor genes

Correlation analysis between the values of A, T, G, C, GC and corresponding A_3, T_3, G_3, C_3, GC_3 values, respectively, were performed in order to investigate the relationship between codon usage variation and compositional constraints (Table 8).

We observed a significant positive correlation between nucleotide compositions, suggesting that nucleotide constraint might influence the codon usage pattern in tumour suppressor genes. Moreover, we performed a neutrality plot analysis of GC_{12} versus GC_3 (Figure 13) and in our analysis, we observed a significant positive correlation (Pearson r=0.724, $p<0.01$) between GC_{12} and GC_3 values which suggested that the intragenomic GC mutation bias might influence the codon usage of these genes. The linear regression coefficient of GC_{12} on GC_{3s} was 0.259 indicating the relative neutrality is 25.9 %, while the relative constraint is 74.1 % for GC_3. These results suggest that natural selection played a major role while mutation pressure played a minor role in the codon usage patterns of these genes.

Table 8: Correlation analysis between A, T, G, C, GC contents and A_3, T_3, G_3, C_3, GC_3 contents in all the selected coding sequences of tumor suppressor genes

	A_3	T_3	G_3	C_3	GC_3
A	r=0.983**	r=0.982**	r=0.755**	r=0.621**	r=-0.410**
T	r=0.981**	r=0.986**	r=0.729**	r=0.606**	r=-0.431**
G	r=0.826**	r=0.803**	r=0.950**	r=0.910**	r=-0.097
C	r=0.743**	r=0.713**	r=0.953**	r=0.957**	r=0.013
GC	r=-0.445**	r=-0.470**	r=0.144	r=0.308*	r=0.940**

**p<0.01 *p<0.05

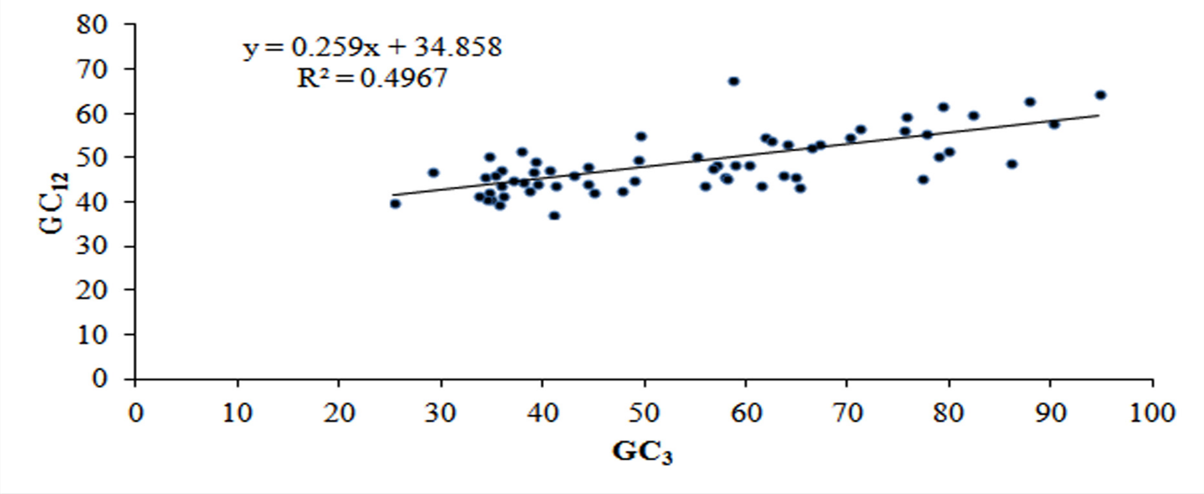


Figure 13: Neutrality plot of GC_{12} versus GC_3. The regression line is y = 0.259x + 34.86; R^2 = 0.496; GC_{12}: average GC content at first and second codon positions.

4.9. Effective number of codons (*ENC*) and its relationship with GC_{3s} values in the coding sequences of proto-oncogenes/oncogenes

The *ENC* values of coding sequences ranged from 38 to 60 indicating relatively weak codon usage bias among these genes. The GC_3 values ranged from 30.4% to 90.8%. The correlation coefficient between *ENC* and GC_{3s} values showed a significant negative correlation (Pearson r=-0.835, p<0.01) for each gene, suggesting that genes with higher GC_{3s} values and lower *ENC* values had strong bias. Further, to investigate the relationship between codon usage variations among genes, we plotted the values of *ENC* versus GC_{3s} (Figure 14) as per Wright (1990). The comparison of actual versus expected distribution in the absence of selection reveals that apart from compositional bias other forces exist which may influence the codon bias. Conversely, if GC_{3s} had been solely responsible for codon usage variation among the genes, the *ENC* values would have fallen on the *ENC*-GC_{3s} curve (Wright 1990). The curve line (shown in figure 7) indicates the expected position of genes in which GC_3 value is the determining factor in shaping the codon usage pattern. Most of the genes in our analysis were lying on and close to the reference line, which suggests that GC_{3s} value was the major determining factor of codon usage pattern of these genes. Thus, it is evident that compositional bias might also play a role in defining the codon usage variation in human proto-oncogenes/oncogenes.

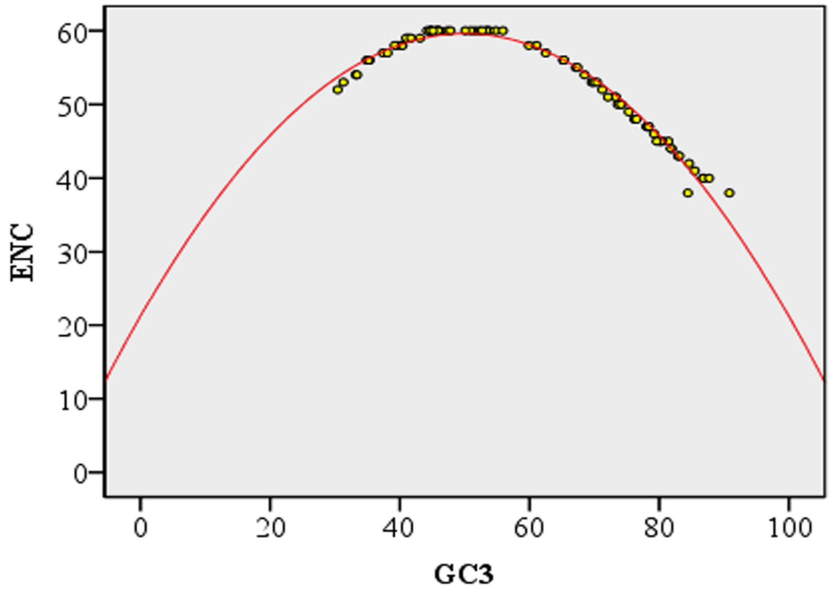


Figure 14: Distribution of effective number of codons (*ENC*) and GC3s in the coding sequences of proto-oncogenes/oncogenes. The curve (shown in red) indicates the expected curve between *ENC* and GC contents under random codon usage

4.10. Effective number of codons (*ENC*) and its relationship with GC_{3s} values in the coding sequences of tumour suppressor genes

The *ENC* values in the coding sequences of tumour suppressor genes ranged from 35 to 60 indicating relatively weak codon usage bias among these genes. The GC_3 values ranged from 25.4% to 94.7%. We also calculated the correlation coefficient between *ENC* and GC_{3s} values and the result showed a significant negative correlation (Pearson $r=-0.695$, $p<0.01$) for each gene. Further, in order to investigate the relationship between codon usage variations among genes, we plotted the values of *ENC* versus GC_{3s} (Figure 15) as per Wright (1990).

The comparison of actual versus expected distribution in the absence of selection reveals that apart from compositional bias other forces exist which might influence codon bias. Conversely, if GC_{3s} had been solely responsible for codon usage variation among the genes, the *ENC* values would have fallen on the *ENC*-GC_{3s} curve (Wright 1990). The solid curve line (shown in figure 8) indicates the expected position of genes in which GC_3 value is the determining factor in shaping the codon usage pattern. Most of the genes in our analysis were lying on and close to the reference line, representing that GC_{3s} value was the sole determining factor of codon usage pattern of these genes. Thus, it is evident that compositional bias might play a significant role in defining the codon usage variation in human tumour suppressor genes.

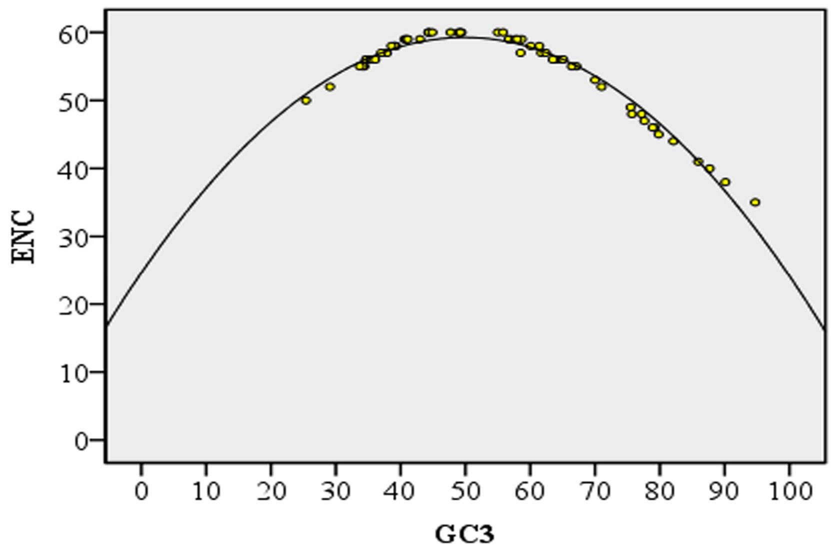


Figure 15: Distribution of effective number of codons (*ENC*) and GC3s in the coding sequences of tumour suppressor genes. The curve indicates the expected curve between *ENC* and GC contents under random codon usage.

4.11. Frequency of optimal codons for the corresponding amino acids among proto-oncogenes/oncogenes

The frequently used optimal codon (*Fop*) values for the proto-oncogenes/oncogenes were analyzed in order to determine the occurrence of the highest and the lowest frequently used codons for the corresponding amino acid. The average percentage of the *Fop* value of codons showed that the highest frequently used codons were GCC, AGA, AAC, GAC, TGC, CAG, GAG, GGC, CAC, ATC, CTG, AAG, TTC, CCC, AGC, ACC, TAC and GTG for the amino acids alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, phenylalanine, proline, serine, threonine, tyrosine and valine respectively (Table 9).

4.12. Frequency of optimal codons for the corresponding amino acids among tumour suppressor genes

To determine the occurrence of the highest and the lowest frequently used codons for each amino acid, the frequently used optimal codon (*Fop*) values for the tumour suppressor genes were analyzed. The average *Fop* values of codons showed that the most frequently used codons were GCC, AGA, AAC, GAC, TGC, CAG, GAG, GGC, CAC, ATC, CTG, AAG, TTT, CCT, AGC, ACA, TAC and GTG for the amino acids alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, phenylalanine, proline, serine, threonine, tyrosine and valine respectively (Table 10).

Table 9: The average of *Fop* values showing the highest percentage of codon usage for the corresponding amino acid among proto-oncogenes/oncogenes

Amino Acid	Syn. Codon	Avg. Fop	Highest % of Usage Codon
Ala	GCA	0.23	
	GCC	0.40	40%
	GCG	0.11	
	GCT	0.26	
Arg	CGT	0.08	
	CGC	0.20	
	CGA	0.12	
	CGG	0.19	
	AGA	0.21	21%
	AGG	0.20	
Asn	AAC	0.60	60%
	AAT	0.40	
Asp	GAT	0.44	
	GAC	0.56	56%
Cys	TGC	0.58	58%
	TGT	0.42	
Gln	CAA	0.26	
	CAG	0.74	74%
Glu	GAA	0.4	
	GAG	0.60	60%
Gly	GGA	0.25	
	GGC	0.34	34%
	GGG	0.24	
	GGT	0.17	
His	CAT	0.40	
	CAC	0.60	60%
Ile	ATA	0.16	
	ATC	0.50	50%
	ATT	0.34	
Leu	TTA	0.08	
	TTG	0.15	
	CTA	0.07	
	CTC	0.20	
	CTG	0.39	39%
	CTT	0.11	
Lys	AAA	0.40	
	AAG	0.60	60%
Phe	TTT	0.43	
	TTC	0.57	57%
Pro	CCA	0.26	
	CCC	0.33	33%
	CCG	0.14	
	CCT	0.27	
Ser	TCA	0.14	
	TCC	0.23	
	TCG	0.06	
	TCT	0.17	
	AGC	0.26	26%
	AGT	0.14	
Thr	ACA	0.28	
	ACC	0.35	35%
	ACG	0.13	
	ACT	0.24	
Tyr	TAC	0.55	55%
	TAT	0.45	
Val	GTA	0.13	
	GTC	0.23	
	GTG	0.48	48%
	GTT	0.16	

Table 10: The average of *Fop* values showing the highest percentage of codon usage for the corresponding amino acid among tumour suppressor genes

Amino Acid	Syn. Codon	Avg. Fop	Highest % of Usage Codon
Ala	GCA	0.26	
	GCC	0.38	38%
	GCG	0.10	
	GCT	0.26	
Arg	CGT	0.09	
	CGC	0.15	
	CGA	0.12	
	CGG	0.17	
	AGA	0.26	26%
	AGG	0.21	
Asn	AAC	0.54	54%
	AAT	0.46	
Asp	GAT	0.49	
	GAC	0.51	51%
Cys	TGC	0.53	53%
	TGT	0.47	
Gln	CAA	0.31	
	CAG	0.69	69%
Glu	GAA	0.48	
	GAG	0.52	52%
Gly	GGA	0.30	
	GGC	0.31	31%
	GGG	0.22	
	GGT	0.17	
His	CAT	0.45	
	CAC	0.55	55%
Ile	ATA	0.19	
	ATC	0.45	45%
	ATT	0.36	
Leu	TTA	0.10	
	TTG	0.24	
	CTA	0.08	
	CTC	0.15	
	CTG	0.32	32%
	CTT	0.11	
Lys	AAA	0.45	
	AAG	0.55	55%
Phe	TTT	0.53	53%
	TTC	0.47	
Pro	CCA	0.28	
	CCC	0.28	
	CCG	0.14	
	CCT	0.30	30%
Ser	TCA	0.16	
	TCC	0.20	
	TCG	0.06	
	TCT	0.18	
	AGC	0.24	24%
	AGT	0.16	
Thr	ACA	0.32	32%
	ACC	0.30	
	ACG	0.14	
	ACT	0.24	
Tyr	TAC	0.54	54%
	TAT	0.46	
Val	GTA	0.15	
	GTC	0.22	
	GTG	0.41	41%
	GTT	0.22	

4.13. Relative synonymous codon usage in the coding sequences of proto-oncogenes/oncogenes

The relative synonymous codon usage values of 59 codons in the coding sequences for proto-oncogenes/oncogenes were analyzed excluding the codons ATG (methionine) and TGG (tryptophan). In our calculation, the *RSCU* value greater than unity indicates the usage of most abundant codon whereas *RSCU* value less than one reveals the usage of least abundant codon. In addition, *RSCU* value greater than 1.6 and less than 0.6 indicates the over-represented and the under-represented codon respectively (Wong *et al*. 2010). The overall *RSCU* values in the selected coding sequences of proto-oncogenes/oncogenes revealed that 27 codons were most frequently used among the 59 codons and the most predominantly used codons were G/C –ending compared to A/T –ending (Table 11). Besides, it was observed that C –ending codon (15) was mostly favored compared to G –ending codon (7) in the coding sequence of proto-oncogenes/oncogenes. In addition, the most over-represented codon in the coding sequence of proto-oncogenes/oncogenes was CTG encoding leucine amino acid.

4.14. Relative abundance of dinucleotide in the coding sequences of proto-oncogenes/oncogenes

Literature suggested that dinucleotide bias can influence the overall codon usage patterns in a variety of organisms (Chiusano *et al*. 2000, Karlin and Burge 1995). The relative abundance of 16 dinucleotides from the coding sequences of eighty-two proto-oncogenes/oncogenes was calculated in order to assess the effect of dinucleotides on the codon usage patterns in these genes under study (Figure 16). The outcome of the analysis showed that CpG dinucleotides were under-represented (mean ± standard deviation = 0.51±0.206) whereas GpC dinucleotides were over-represented ones (mean ± standard deviation = 1.04±0.113). Moreover, *RSCU* values (Table 11) of codons containing the CpG dinucleotide (TCG, CCG, ACG, GCG, CGA, and CGT) except CGC and CGG were the least used codons (*RSCU*<1.0) for their corresponding amino acid. Similarly, the GpC dinucleotide containing codons (TGC, CGC, GGC, GCC, AGC, GCT) except GCA and GCG were over-represented (*RSCU*>1.0) in proto-oncogenes/oncogenes under study. Besides, the dinucleotide TpG containing four codons (TGC, CTG, TGC, GTG) and CpA containing four codons (CCA, CAC, CAG, and ACA) were over-represented and most of them were also used as preferred codons for their corresponding amino acid based on *RSCU* analysis except TTG and TGT for TpG dinucleotide and TCA, CAA, GCA, CAT for CpA dinucleotide in human proto-oncogenes/oncogenes. Previous study reported that the CpA and TpG dinucleotides were over-represented in different organisms. This could be due to the effect of CpG dinucleotide (Bird 1980). Our results also revealed that most of the codons containing CpA and TpG dinucleotides were over-represented in the selected proto-oncogenes/oncogenes under study.

Table 11: Overall relative synonymous codon usage patterns (*RSCU*) in the coding sequences of proto-oncogenes/oncogenes

Amino Acid	Codon	N	RSCU[a]	Amino Acid	Codon	N	RSCU[a]
Ala	GCA	776	0.89	Leu	TTA	366	0.44
	GCC*	1319	1.64		TTG	614	0.87
	GCG	344	0.44		CTA	331	0.39
	GCT*	877	1.03		CTC*	878	1.21
Arg	CGT	221	0.43		CTG*	1799	2.42
	CGC*	505	1.22		CTT	620	0.67
	CGA	370	0.75	Lys	AAA	1206	0.78
	CGG*	531	1.16		AAG*	1557	1.22
	AGA*	564	1.22	Phe	TTT	782	0.85
	AGG*	526	1.22		TTC*	830	1.08
Asn	AAC*	1077	1.21	Pro	CCA*	1012	1.04
	AAT	931	0.74		CCC*	1172	1.36
Asp	GAT	1152	0.84		CCG	412	0.52
	GAC*	1289	1.14		CCT*	969	1.04
Cys	TGC*	567	1.16	Ser	TCA	660	0.78
	TGT	456	0.80		TCC*	979	1.40
Gln	CAA	770	0.50		TCG	261	0.42
	CAG*	1908	1.50		TCT	861	0.98
Glu	GAA	1498	0.78		AGC*	1087	1.58
	GAG*	1970	1.22		AGT	730	0.84
Gly	GGA	879	0.96	Thr	ACA*	834	1.11
	GGC*	1137	1.38		ACC*	883	1.43
	GGG	792	0.98		ACG	313	0.51
	GGT	627	0.68		ACT	680	0.95
His	CAT	513	0.69	Tyr	TAC*	766	1.14
	CAC*	812	1.23		TAT	619	0.79
Ile	ATA	339	0.43	Val	GTA	350	0.49
	ATC*	909	1.55		GTC	615	0.89
	ATT	726	0.95		GTG*	1268	1.95
					GTT	493	0.64

[a] mean values of RSCU based on the synonymous codon usage frequencies of proto-oncogenes/oncogenes; N: Total number of preferred codon; **RSCU*>1

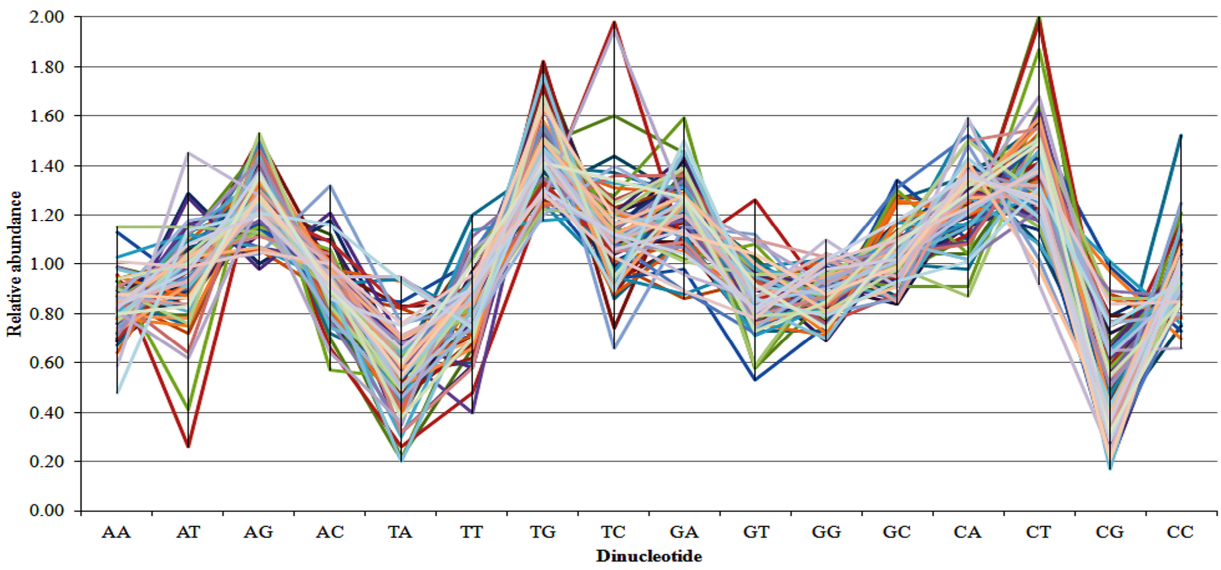


Figure 16: Line diagram showing the relative abundance of sixteen dinucleotides in the coding sequence of proto-oncogenes/oncogenes in human

4.15. Trends of codon usage variation among proto-oncogenes/oncogenes

To determine the trends in codon usage variation among human proto-oncogene/oncogenes, we performed correspondence analysis (COA) based on the *RSCU* values. In our analysis, we observed that the first principal axis (f_1) accounted for 46.68% of all variations within the gene set, whereas the second axis (f_2) accounted for only 7.86% (Figure 17). The major trends of variation in the entire data can be illustrated in terms of correlation analysis by using primary axis. Therefore, a correlation analysis of the axis 1 (f_1) with the major indices, namely *ENC*, GC, GC3s and *CAI* values (r= 0.781, -0.933, -0.985 and -0.798, respectively, $p<0.01$) was done using Spearman's rank correlation method. The above results indicate significant correlation between the primary axes against the four major indices. Hence, the results revealed that nucleotide composition and mutation bias might play major roles in the codon usage of human proto-oncogenes/oncogenes.

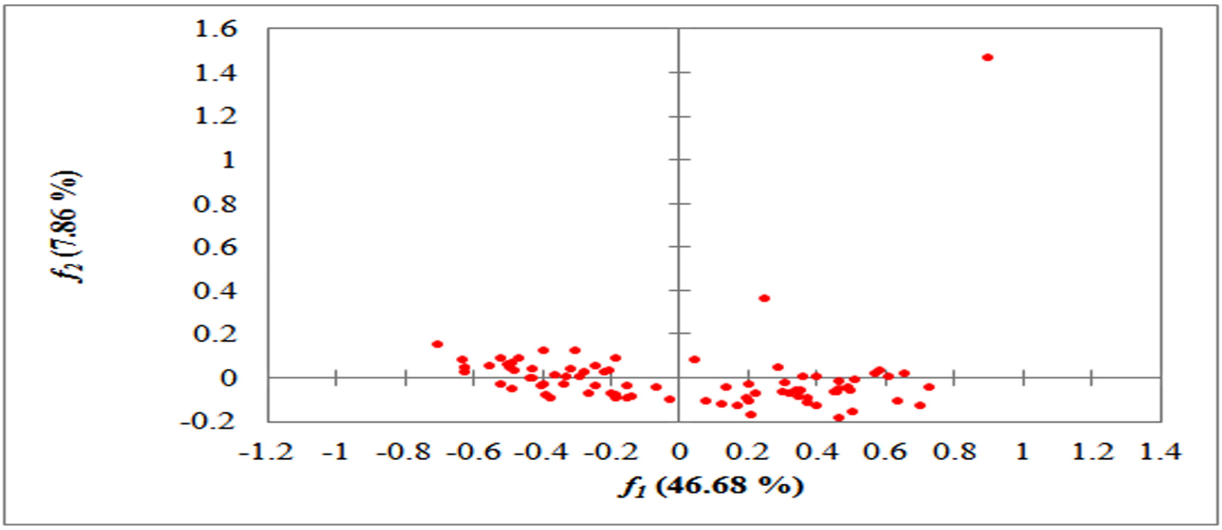


Figure 17: Correspondence analysis of *RSCU* values in human proto-oncogenes/oncogenes. Each point in the plot represents the distribution of a gene corresponding to the coordinates of the primary and secondary axes of variation.

4.16. Relative synonymous codon usage in the coding sequences of tumour suppressor genes

The relative synonymous codon usage values of 59 codons in the coding sequences for tumour suppressor genes were analyzed excluding the codons ATG (methionine) and TGG (tryptophan). The overall *RSCU* values in the selected coding sequences of tumour suppressor genes revealed that 32 codons were most frequently used among the 59 codons and the most predominantly used codons were C/T–ending compared to A/G–ending ones (Table 12).

Besides, it was observed that C–ending codon (12) was mostly favored compared to T –ending codon (8) followed by G–ending (7) and A–ending codon (5) in the coding sequences of tumour suppressor genes. The most over-represented codon (highest *RSCU* value) was CTG encoding the amino acid leucine.

4.17. Relative abundance of dinucleotide in the coding sequences of tumour suppressor genes

The relative abundance of 16 dinucleotides from the coding sequences of sixty three tumour suppressor genes were calculated in order to assess the effect of dinucleotides on the codon usage patterns in these genes under study (Figure 18). The results of the analysis showed that CpG dinucleotides were under-represented (mean ± standard deviation = 0.49±0.210) whereas GpC dinucleotides were over-represented (mean ± standard deviation = 1.20±0.111). Moreover, *RSCU* values (Table 12) of codons containing the CpG dinucleotide (TCG, CCG, ACG, GCG, CGA, CGC, CGG and CGT) were the least used codons further (*RSCU*<1.0) for their corresponding amino acid. Similarly, the GpC dinucleotide containing codons (TGC, GCA, GGC, GCC, AGC, GCT) except CGC and GCG were over-represented (*RSCU*>1.0) in tumour suppressor genes under study. In addition, the dinucleotide TpG containing four codons (TGC, CTG, TGC, GTG) and CpA containing four codons (CCA, CAC, CAG, GCA and ACA) were over-represented and most of them were also used as preferred codons for their corresponding amino acid based on *RSCU* analysis except TTG and TGT for TpG dinucleotide and TCA,CAA, CAT for CpA dinucleotide in tumour suppressor genes. The over-representation of CpA and TpG dinucleotides was also reported earlier in different organisms (Bird 1980).

Table 12: Overall relative synonymous codon usage patterns (*RSCU*) in the coding sequences of tumour suppressor genes

Amino Acid	Codon	N	RSCU[a]	Amino Acid	Codon	N	RSCU[a]
Ala	GCA*	935	1.07	Leu	TTA	704	0.69
	GCC*	1192	1.47		TTG	778	0.91
	GCG	316	0.41		CTA	421	0.48
	GCT*	949	1.05		CTC	790	0.98
Arg	CGT	230	0.50		CTG*	1461	2.00
	CGC	418	0.91		CTT	844	0.95
	CGA	327	0.73	Lys	AAA	1770	0.91
	CGG*	440	1.00		AAG*	1642	1.09
	AGA*	767	1.59	Phe	TTT*	1114	1.07
	AGG*	539	1.28		TTC	751	0.90
Asn	AAC*	1055	1.06	Pro	CCA*	924	1.11
	AAT	1329	0.94		CCC*	921	1.13
Asp	GAT*	1524	1.00		CCG	408	0.51
	GAC*	1216	1.00		CCT*	1041	1.25
Cys	TGC*	516	1.04	Ser	TCA	892	0.98
	TGT	585	0.96		TCC*	829	1.18
Gln	CAA	840	0.58		TCG	221	0.35
	CAG*	1810	1.42		TCT*	968	1.08
Glu	GAA	2130	0.98		AGC*	934	1.41
	GAG*	1740	1.02		AGT*	894	1.00
Gly	GGA*	943	1.21	Thr	ACA*	1004	1.25
	GGC*	960	1.25		ACC*	828	1.20
	GGG	649	0.85		ACG	298	0.53
	GGT	579	0.70		ACT*	866	1.03
His	CAT	705	0.92	Tyr	TAC*	709	1.06
	CAC*	712	1.09		TAT	680	0.91
Ile	ATA	516	0.51	Val	GTA	479	0.56
	ATC*	900	1.35		GTC	599	0.86
	ATT*	1013	1.14		GTG*	1142	1.66
					GTT	755	0.91

[a]mean values of RSCU based on the synonymous codon usage frequencies of tumour suppressor genes; N: Total number of preferred codon; *$RSCU > 1$

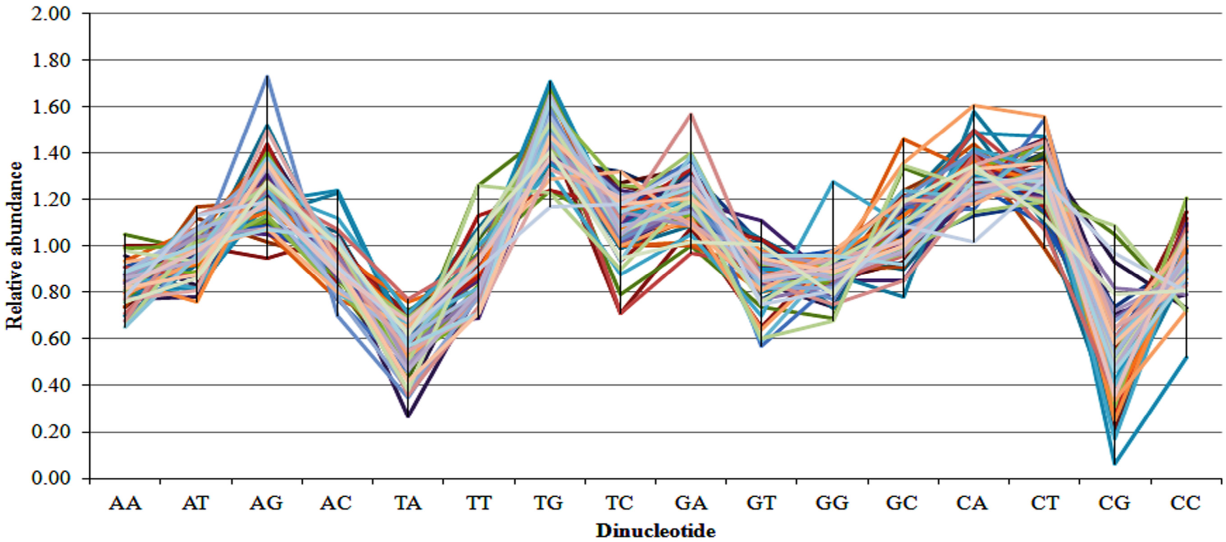


Figure 18: Line diagram showing the relative abundance of sixteen dinucleotides in the coding sequence of tumour suppressor genes in human

4.18. Trends of codon usage variation among tumour suppressor genes

Correspondence analysis (COA) based on *RSCU* values was performed in order to determine the trends in codon usage variation among human tumour suppressor genes. We observed that the first principal axis (f_1) accounted for 56.10% of all variations within the gene set, whereas the second axis (f_2) accounted for only 5.17% variation (Figure 19). The major trends of variation in the entire data can be illustrated in terms of correlation analysis by using primary axis. Therefore, a correlation analysis of the axis 1 (f_1) with the major codon bias indices, namely *ENC*, GC, GC3s and *CAI* values (r= -0.385, 0.947, 0.982 and 0.837, respectively, $p<0.01$) was done using Spearman's rank correlation method. The above results indicate significant correlation between the primary axis and the four major indices. Hence, the results revealed that both nucleotide composition and mutation bias might play major roles in the codon usage of human tumour suppressor genes.

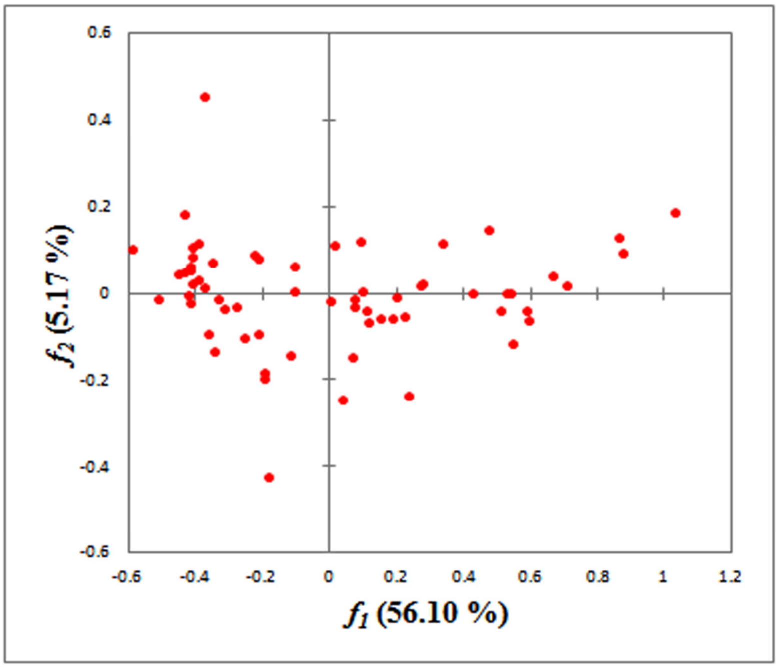


Figure 19: Correspondence analysis of RSCU values in human tumour suppressor genes. Each point in the plot represents the distribution of a gene corresponding to the coordinates of the primary and secondary axes of variation.

4.19. Prediction of proto-oncogenes/oncogenes expression level and codon usage bias

Codon adaptation index (*CAI*) is a quantitative measure used for the prediction of gene expression on the basis of extent of bias towards codon sequence (Behura and Severson 2012, Gupta *et al*. 2004). In our analysis, *CAI* values ranged from 0.715 to 0.887 (Figure 20) with a mean value of 0.801 and a standard deviation of 0.039. The magnitude of *CAI* values indicates that most of the genes selected in the present study are highly expressive in cell.

Moreover, a significant positive correlation was observed between *CAI* and GC_{3s} (r=0.847, P<0.01) and between *CAI* and GC (r=0.715, P<0.01) content values. Furthermore, significant negative correlation (r=-0.627, P<0.01) also observed between *ENC* and *CAI* values (Figure 21).

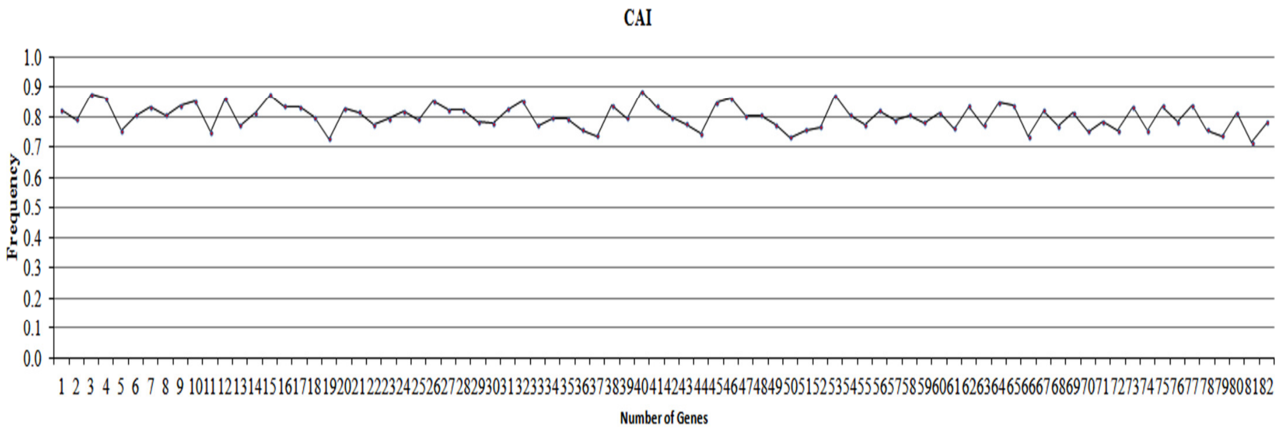


Figure 20: Trends of the *CAI* values in the coding sequences of proto-oncogenes/oncogenes

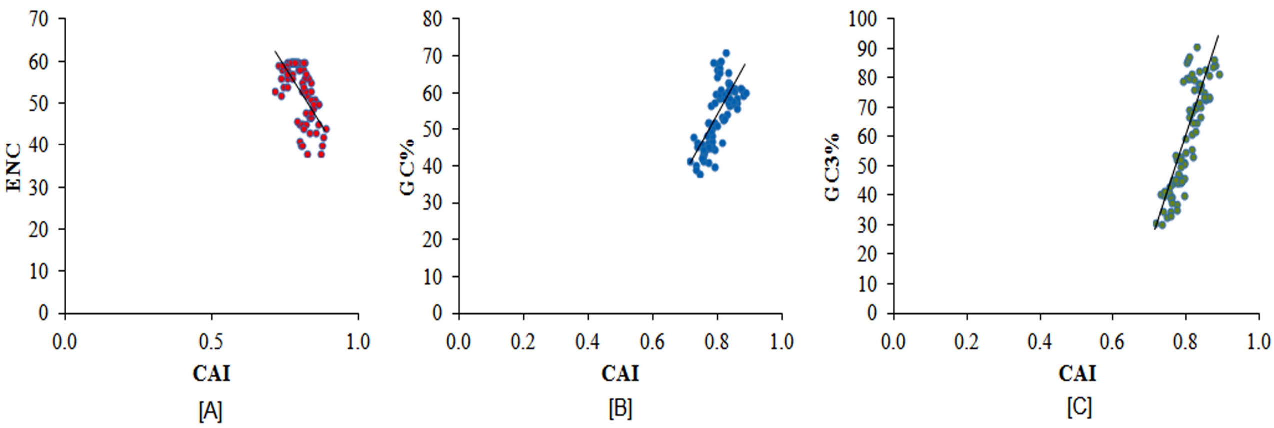


Figure 21: Correlation analysis between *CAI* values and *ENC*, *GC*, GC_3 values. [A] Correlation between *CAI* and *ENC*. [B] Correlation between *CAI* and GC contents. [C] Correlation between *CAI* and GC_3 values.

4.20. Relationship between codon usage and *CAI* values of proto-oncogenes/oncogenes

To investigate the relationship between the codon usage variation and the level of gene expression among the selected coding sequences of human proto-oncogenes/oncogenes, the correlation coefficient between codon usage and *CAI* was estimated using heat map (Figure 22). The results showed that almost all G/C –ending codons are positively correlated with *CAI* and vice versa for A/T –ending codons. This indicates that gene expression increases with the increase in usage of G/C –ending codons. However, two G–ending codons namely TTG (leucine) and CAG (glutamine) showed negative correlation between codon usage and gene expression as measured by *CAI*.

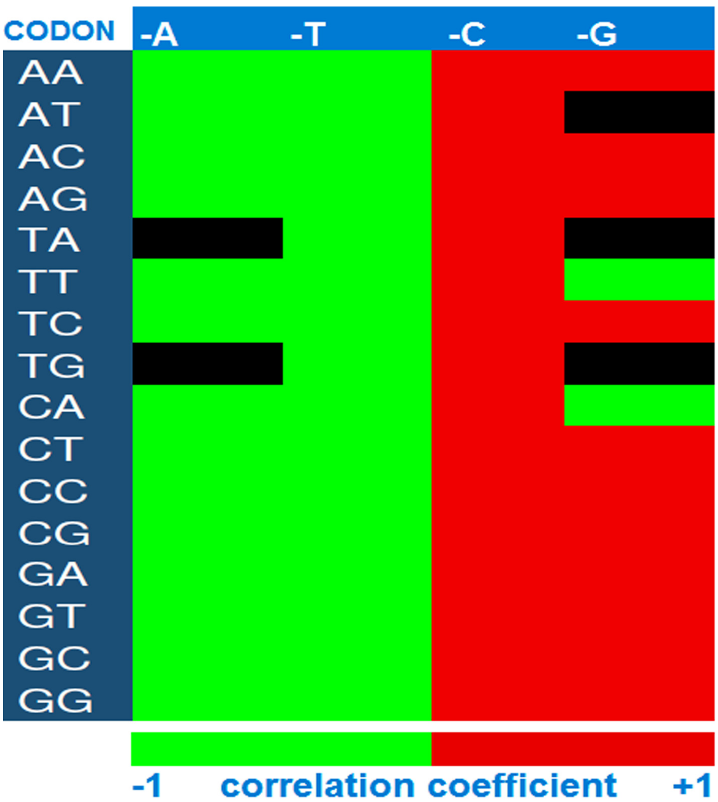


Figure 22: Heat maps of correlation coefficient of codons with *CAI*. The color coding red represents the positive correlation, green as negative correlation. The black fields are stop codons (TAA, TAG, TGA) and non-degenerate codons (ATG, TGG) in the coding sequences of the eighty-two proto-oncogenes/oncogenes under study.

4.21. Prediction of translational efficiency as measured by *tAI* and its relationship with codon usage in proto-oncogenes/oncogenes pool

The analysis of *tAI* for the coding sequences in a genome acquires importance only when the translational selection process plays a major role in the shaping of the codon usage of the particular genome (Man *et al*. 2006). We estimated *tAI* values for human proto-oncogenes/oncogenes pool (Jung and McDonald 2011) which ranged from 0.326 to 0.414 with a mean of 0.360. We accomplished a correlation analysis between codon usage and *tAI* and represented the same in a heat map (Figure 23A). The results showed that nearly all codons ending with G/C base were positively correlated with *tAI* and vice versa for all A/T ending codons. Besides, we observed a significant negative correlation ($r=-0.545$, $p<0.01$) between *tAI* and effective number of codons (*ENC*) and significant positive correlation between *tAI* and GC3s, an indicator of the extent of base composition bias ($r=0.382$, $p<0.01$) (Figure 23 -B & C). In addition, a significant positive correlation (Pearson, $r=0.676$, $p<0.01$) was observed between the two parameters *tAI* and *CAI* of the coding sequences which indicated that the expression of proto-oncogenes/oncogenes, was significantly influenced by the genomic *tRNA* pool.

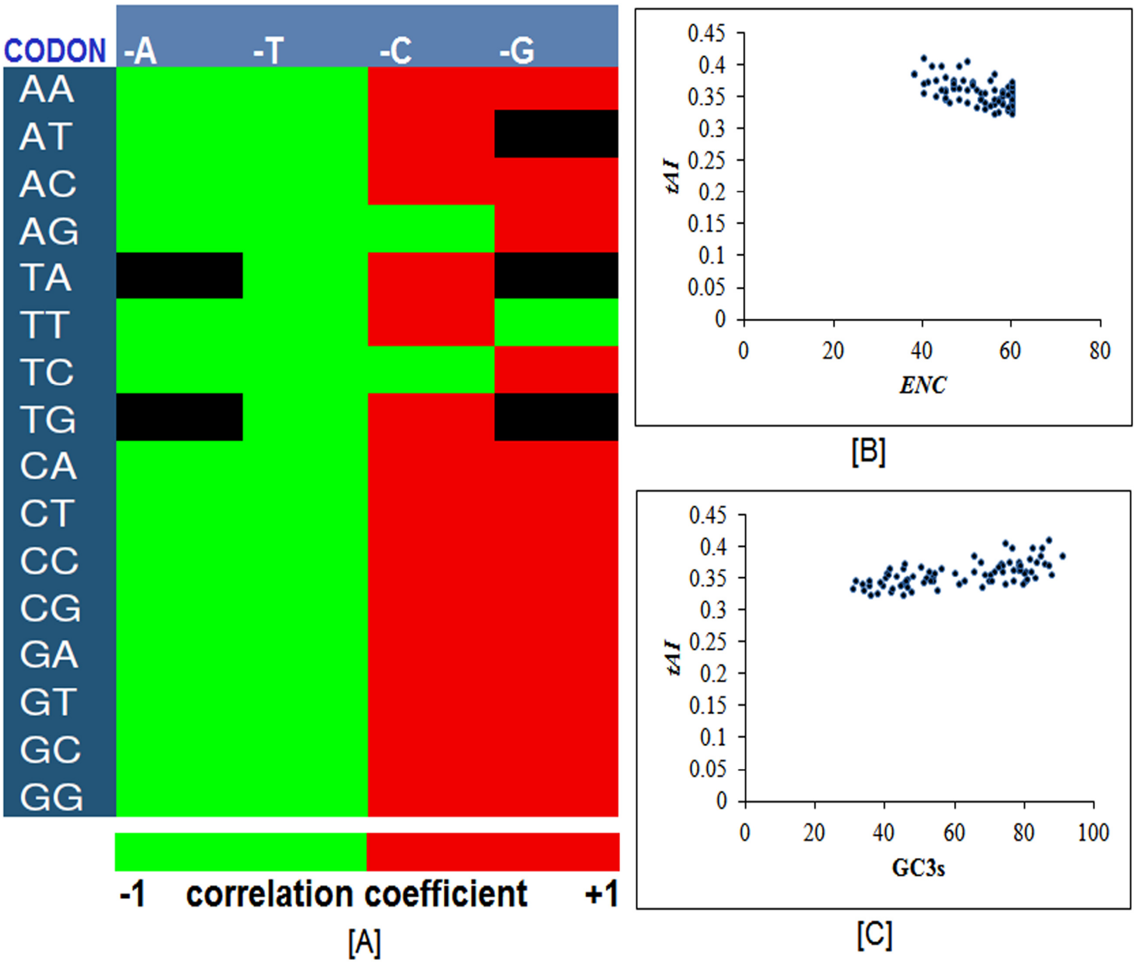


Figure 23. Relationship between *tAI* and codon usage bias in proto-oncogene/oncogene pool. [A] Correlation coefficient between codon usage and *tAI*. The rectangular red color box indicates the positive correlation, green as negative correlation; black boxes are non-degenerate codons (ATG, TGG) and three termination codons (TAA, TAG, TGA). [B] Correlation coefficient between *tAI* and *ENC*. [C] Correlation coefficient between *tAI* and *GC3s*.

4.22. Prediction of tumour suppressor gene expression level and codon bias

The expression of tumour suppressor gene was measured using *CAI* values which ranged from 0.713 to 0.858 (Figure 24) with a mean of 0.775 and a standard deviation of 0.037. The *CAI* values indicate that most of the tumour suppressor genes selected for the present study are highly expressed in cells.

In addition, we observed significant negative correlation between *CAI* and *ENC* (r=-0.515, P<0.01) values, significant positive correlation between *CAI* and GC_3 (r=0.839, P<0.01) as well as between *CAI* and GC contents (r=0.730, P<0.01) in tumour suppressor genes (Figure 25).

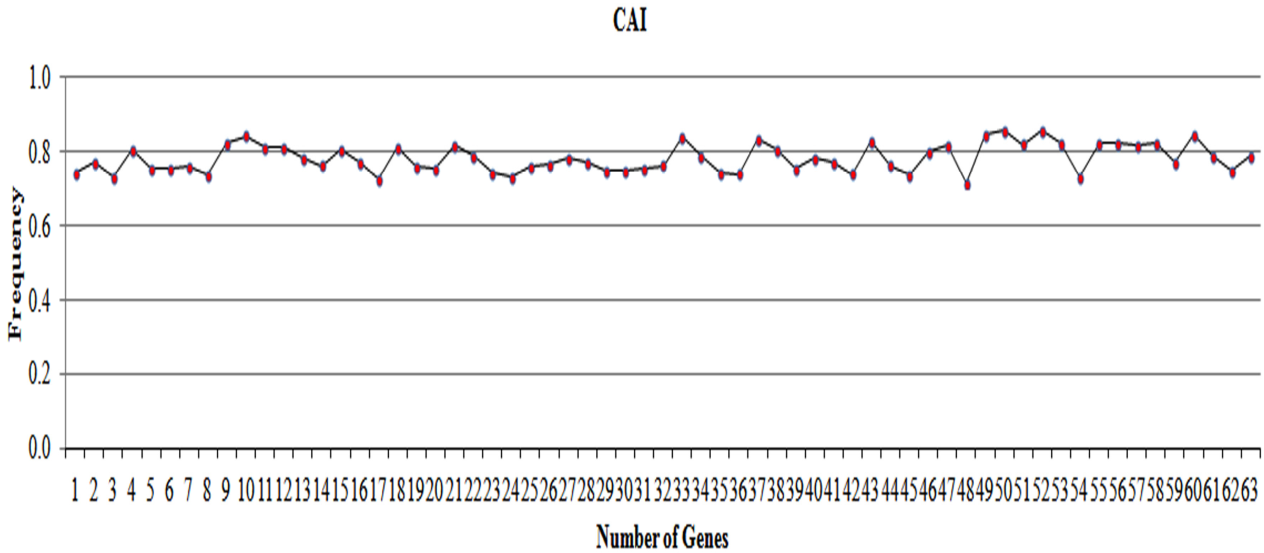


Figure 24: Trends of the *CAI* values in the coding sequences of tumour suppressor genes

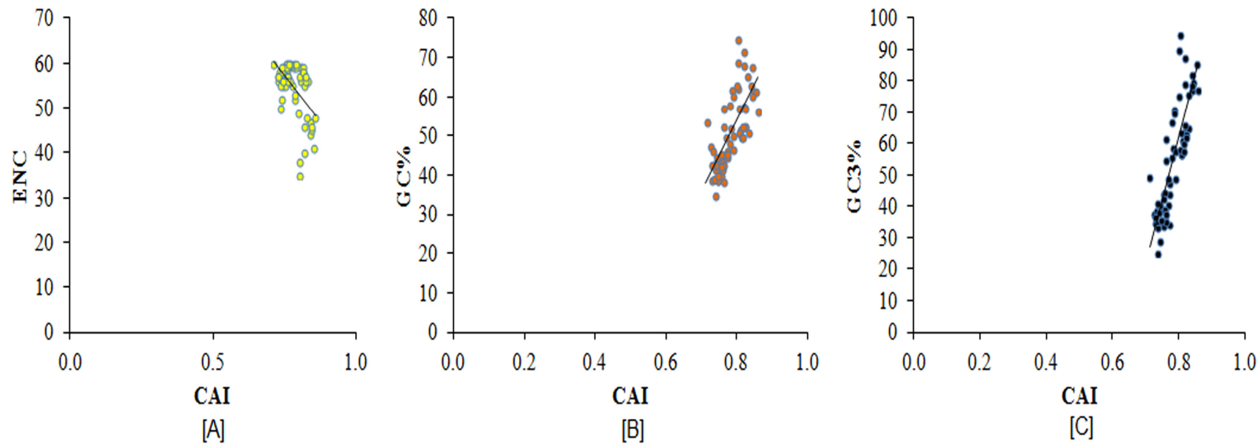


Figure 25: Correlation analysis between *CAI* values and *ENC*, GC, GC_3 values. [A] Correlation between *CAI* and *ENC*. [B] Correlation between *CAI* and GC contents. [C] Correlation between *CAI* and GC_3 values.

4.23. Relationship between codon usage and *CAI* values of tumour suppressor genes

To investigate the relationship between the codon usage variation and gene expression level among the selected coding sequences of human proto-oncogenes/oncogenes, the correlation coefficient between codon usage and *CAI* was analyzed using heat map (Figure 26). The result showed that almost all the G/C –ending codons are positively correlated with *CAI* and vice versa for A/T–ending codons. This suggests that gene expression increases with the increase in usage of G/C–ending codons. However, two G–ending codons namely AGG (arginine) and TTG (leucine) showed negative correlation between codon usage and gene expression as measured by *CAI*.

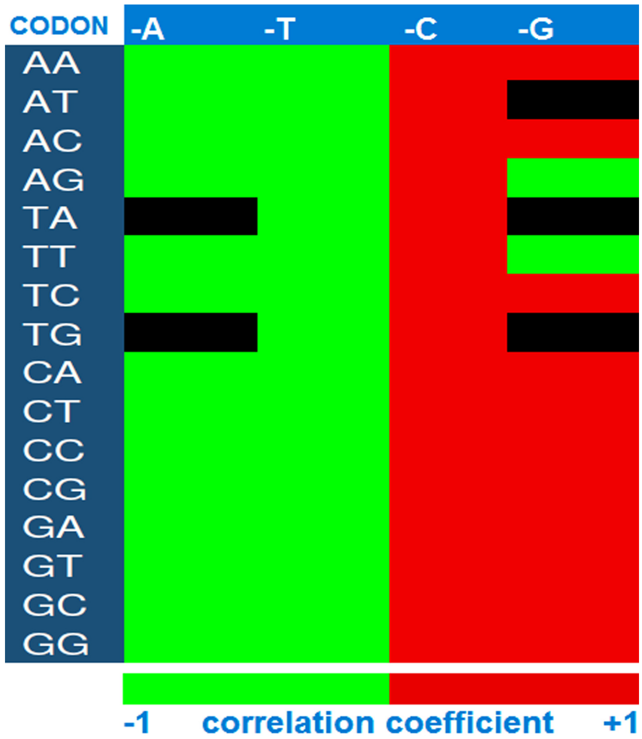


Figure 26: Heat maps of correlation coefficient of codons with *CAI*. The color coding red represents the positive correlation, green as negative correlation. The black fields are stop codons (TAA, TAG, TGA) and non-degenerate codons (ATG, TGG) in the coding sequences of the sixty-three tumour suppressor genes under study.

4.24. Prediction of translational efficiency as measured by *tAI* and its relationship with codon usage in tumour suppressor gene pool

We calculated *tAI* values for human tumour suppressor gene pool (Jung and McDonald 2011) which ranged from 0.382 to 0.402 with a mean value of 0.356. We performed a correlation analysis between codon usage and *tAI* and represented it in a heat map (Figure 27 A). The results showed that nearly all codons ending with G/C base were positively correlated with *tAI* and vice versa for all A/T ending codons.

Besides, we observed a significant positive correlation (r=-0.382, p<0.01) between *tAI* and effective number of codons (*ENC*) as well as between *tAI* and GC3s, an indicator of the extent of base composition bias (r=0.511, p<0.01) (Figure 27-B & C). Moreover, we observed a significant positive correlation (r=0.625, p<0.01) between the two parameters *tAI* and *CAI* for coding sequences. This revealed that the expression of tumour suppressor genes (*CAI*) was significantly influenced by the genomic *tRNA* pool.

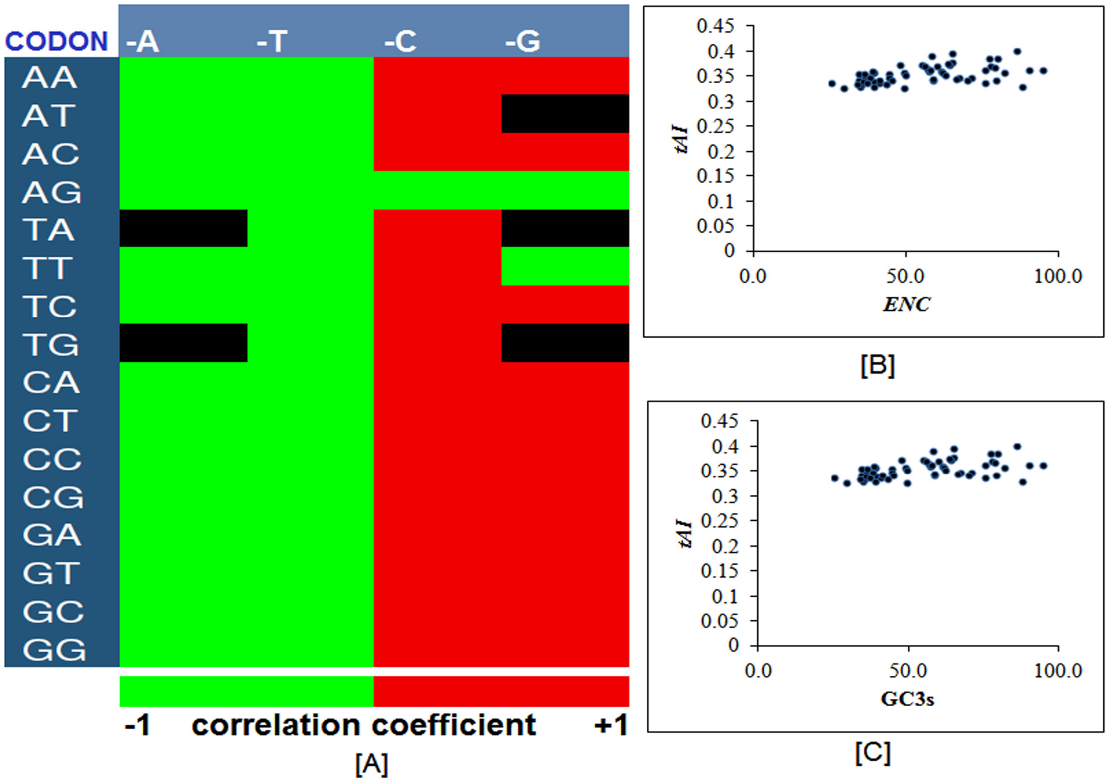


Figure 27. Relationship between *tAI* and codon usage bias in tumour suppressor gene pool. [A] Correlation coefficient between codon usage and *tAI*. The rectangular red color box indicates the positive correlation, green as negative correlation; black boxes are non-degenerate codons (ATG, TGG) and three termination codons (TAA, TAG, TGA). [B] Correlation coefficient between *tAI* and *ENC*. [C] Correlation coefficient between *tAI* and *GC3s*.