Table of Content:

1.) Project Information:

In this project you will study the difference between human and other two reference genomes based on pairwise alignment result.

In the alignment files, some bases are in lower case. They shall be treated the same as upper case letters.

For a given pairwise alignment between two species, we have following definitions:

Substitution rate = number-of- mismatches / (number-of-mismatches + number-of-matches)

ti/tv = number of transitions / numbers of transversions

Gap rate = number of gaps / (number of matches + number of mismatches + number of gaps)

2.) Introduction:

So far, we have learned what is Genome in class. Genome is encoded into four types of DNA bases and we use the first letter of each to represent them. The four Alphabets are A, C, G and T. So, genomes are long strings containing these four alphabets.

We have learnt that all species have experienced a long history of evolution. And couple of these evolutionary events have occurred because of insertion / deletion (indel; as we cannot establish for a fact whether there was an insertion or deletion and hence, we call it indel) substitution and few more like duplication, translocation and transposition. But in this project, we are just concentrate on substitution.

Substitution is a type where one base is transformed to another (A – G, A – T, A –C, G – A, G – C, G – T, C – A, C – G, C – T, T – A, T – C and T – G). There are two types in substitution:

Transition: When substitution between A – G and C – T occurs.

Transversion: Rest all the types of substitution.

Also, we have learnt that an indel is often referred as a gap and it is represented using '-'. A sequence of contiguous '-' characters is counted as one gap. The gap length refers to the number of '-' characters in the gap.

3.) Programming Language Used:

For this project I have used C++ as my programming language. I took around 2 hours to code and around 3 - 4 hours to debug and modify. And around another 2 hours to write the report.

Used the Command g++ -o <execution file name> <program name.cc> to compile the program and ./<execution file name> to execute

4.) System Information Where the Code Was Executed:

The code was executed in Ubuntu 18.04. The System Specification is as follows:

a.) Processor: Intel core i5 7th Generation, Quad Core Processor
b.) RAM: 8 GB Ram
c.) Hard Disk Size: 1 TB

5.) **Summary / Observation:**

When I executed the
program, I made the following observations:

a.) The cumulative total number of matches present in the input files were : 13793977

b.) The cumulative total number of mis-matches present in the input files were : 189115

c.) The cumulative transition counts were : 129538

d.) The cumulative transversion counts were : 59532

e.) Gap Count is : 17589

f.) Substitution rate is : 0.01352

g.) Ti / Tv ration is : 2.177

h.) Gap rate is 0.001256

Pair Counts	Species 2				
Species 1		A	C	G	T
	A	3433847	7464	31979	4844
	C	7202	3443972	9916	33324
	G	32967	9997	3457570	7292
	T	4616	31664	7512	3431529

Gap Length (Bases)	Gap Count	Gap Frequency
1	8228	0.4678
2	2600	0.1478
3	1489	0.08466
4	1278	0.07266
5	521	0.02962
6	419	0.02382
7	314	0.01785
8	292	0.0166
9	218	0.01239
10	184	0.01046
11	155	0.008812
12	134	0.007618
13	124	0.00705
14	140	0.00796
15	100	0.005685
16	127	0.00722
17	93	0.005287
18	108	0.00614
19	85	0.004833
20	88	0.005003
21	74	0.004207
22	59	0.003354
23	49	0.002786
24	57	0.003241

25	50	0.002843
26	31	0.001762
27	37	0.002104
28	23	0.001308
29	27	0.001535
30	38	0.00216
31	29	0.001649
32	26	0.001478
33	20	0.001137
34	12	0.0006822
35	13	0.0007391
36	17	0.0009665
37	11	0.0006254
38	11	0.0006254
39	15	0.0008528
40	7	0.000398
41	10	0.0005685
42	11	0.0006254
43	12	0.0006822
44	16	0.0009097
45	10	0.0005685
46	16	0.0009097
47	9	0.0005117
48	4	0.0002274
49	6	0.0003411
50	7	0.000398
51	11	0.0006254
52	6	0.0003411
53	4	0.0002274
54	8	0.0004548
55	8	0.0004548
56	7	0.000398

57	5	0.0002843
58	7	0.000398
59	3	0.0001706
60	7	0.000398
61	5	0.0002843
62	5	0.0002843
63	3	0.0001706
64	6	0.0003411
65	4	0.0002274
66	5	0.0002843
67	4	0.0002274
68	5	0.0002843
69	4	0.0002274
70	2	0.0001137
71	5	0.0002843
72	3	0.0001706
73	4	0.0002274
74	7	0.000398
75	1	5.685e-05
76	3	0.0001706
77	2	0.0001137
78	4	0.0002274
79	1	5.685e-05
80	7	0.000398
81	2	0.0001137
82	1	5.685e-05
83	1	5.685e-05
84	1	5.685e-05
85	1	5.685e-05
86	4	0.0002274
87	1	5.685e-05
88	1	5.685e-05

89	5	0.0002843
90	4	0.0002274
91	0	0
92	5	0.0002843
93	2	0.0001137
94	1	5.685e-05
95	2	0.0001137
96	1	5.685e-05
97	2	0.0001137
98	1	5.685e-05
99	1	5.685e-05
100	3	0.0001706
Total	17589	1