

Wuhan-coronavirus homologue map

Kou Amano

Background and Objective

Background: At the beginning of 2020, global risk of infection of coronavirus is spreaded. The pandemic started in 2019 and governments announced a state of emergency. In Japan, the government adopted PCR as a diagnosis method of the infection. But the selection of primers influence the accuracy greatly.

Objective: Therefor, I provide a "map" of homological region of coronavirus to other viruses and animals to help the primer design.

Data

Vertebrates:		
● Bat (NW_017738920.1 .. NW_017802358.1; 63439frgs)	● Dog (NC_006583.3 .. NC_006621.3; 39chrs)	● Rabbit (NC_013669.1 .. NC_013690.1; 22chrs)
● Beluga (NW_022097992.1 .. NW_022103895.1; 5904frgs)	● Ferret (NW_004569142.1 .. NW_004576923.1; 7782frgs)	● Pig (NC_010443.5 .. NC_010462.3; 20chrs)
● Camel (NC_044511.1 .. NC_044547.1; 37chrs)	● Human (NC_000001.4 .. NC_000024.3; 24chrs)	● Rock Pigeon (NW_004973171.1 .. NW_004988092.1; 14922frgs)
● Cat (NC_018723.3 .. NC_018741.3; 19chrs)	● Mouse (NC_000067.6 .. NC_000087.7; 21chrs)	● Turkey (NC_015011.2 .. NC_015042.2; 32chrs)
Viruses:		
● 7554 genomes of viruses or phages	● Wuhan-corona virus genome (MN908947.3)	

Method

BLAST:
DB: makeblastdb -in <<input file>> -out <<DB name>> -dbtype nucl -parse_seqids
Query: megablast -d <<DB name>> -i <<query sequence>> -W 10 > <<output file>>

Self-BLAST:
Fragmentation: fragment bf=<<input file>> S=25 G=25 cs=1 > <<output file>>
DB: same as above.
Query: same as above.

Window-fourrier:
Fragmentation: 30 fragments; 1000 bases / fragment
Conversion: "A" -> 1, "T" -> -1, "G" -> I, "C" -> -I
Fourier transform: Ft (<<each fragment>>)

Selection of frequent homologues:
Vertebrate: base-hit count >= 5; seq-lenght >= 5
Viruses: base-hit count >= 18; seq-lenght >= 5

Result

The homologues

No.	homologue	start	end	count	in CDS
1	AATTTTA	1787	1793	6	Yes
2	GATGAGGATGAAGAAGAAGGT	3047	3067	6	Yes
3	CTACAAAGAAAACAGTTAC	5866	5884	5	Yes
4	ATAAATATTATAATTTG	6944	6960	7	Yes
5	TTGCAT	7425	7430	5	Yes
6	CCATCCATCTTTACTTTGATAAA	7764	7786	7	Yes
7	TTTTTGTTGCTGCTATTTTCTATTTA	8607	8632	9	Yes
8	ATTTCCTGTTTGTTTTGT	11167	11186	7	Yes
9	AAAAGT	12198	12203	5	Yes
10	ATAAAATAGAAGAA	19121	19134	6	Yes
11	TGTTTGTTTTTCTTGTTTTATTGC	21564	21587	14	Yes
12	TATTAATAATATAATGAA	22389	22405	5	Yes
13	TTTTCT	26495	26500	5	No
14	CCTAA	29375	29379	5	Yes

No.	homologue	start	end	count	in CDS
15	CAAGATCTCAATGGTAACTGGTATGATTTTCGGTGATT	14068	14104	21	Yes
16	TGGTAATGCTGC	14775	14786	18	Yes
17	ACAAAACGTAATGTCATCCCTACTATAAECTCAAATGAATCTTAAGTATGCCATTAGTGCAAAGAATAGAGCTCGCACCGTAGCTGGTGTCTCTAT	15031	15125	32	Yes
18	CTTATGGGTTGGGATTATCCTAAATGTGATAGAGCCATGCCTAA	15280	15323	37	Yes
19	GATGCCACAACGCTTATGCTAATAGTGTTTTTAACAT	15490	15527	19	Yes
20	CAAAACAATGTTTTTATGTCTGAAGCAAAATGTTGGACTGAGACTGACCTTACTAAAGGACCTCATGAATTTTGCTCTCAACATACA	15805	15891	26	Yes
21	GGTTGTGATGGTGGCAGTTTGTATGTAATAAACATGCATTCCACACACC	19276	19325	27	Yes
22	TATGAGAGTTATACATTTTGGTGCTGGTTCTGATAAAGGAGTTGCACCAAGGTAC	20850	20903	21	Yes
23	CCGAGGCCACGCGGAGTACGATCGAGTGTACAG	29732	29764	32	No

The plot

