Investigation of Oligonucleotide Usage Variance Between SARS -related Coronaviruses and Common Cold Coronaviruses

Elham Mousavi
Kerman University of Medical Sciences
Majid Nikobin-Boroujeni
Golestan University of Medical Sciences and Health Services
Ali Teimoori
Kerman University of Medical Sciences
Seyed Ali Mohammad Arabzadeh
Kerman University of Medical Sciences
Mohammad Mostakhdem Hashemi
Golestan University of Medical Sciences and Health Services
zahra Arab-Bafrani (z_arab2007@yahoo.com)
Golestan University of Medical Sciences and Health Services School of Health and Paramedicine

Research article

Keywords: SARS-CoV-2, MERS, SARS, Common cold coronaviruses, Oligonucleotide patterns, viral genome

Posted Date: March 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-328801/v1

Abstract

Background: The widespread outbreak of SARS-CoV-2 has become a deal threat for human health. This new emerged virus coupled with severe acute respiratory syndrome (SARS) and middle east respiratory syndrome (MERS) viruses belong to coronoviridae family, which develop SARS in human being. However, prior to the emergence of virulent viruses, the coronaviruses were known as the leading causes of mild common cold. Getting more knowledge about the genome organization of different strains can conduct us how these viruses evolve and become a virulent strain. Here, we reported the difference of oligonucleotide distribution contributing in genome of two groups of coronaviruses, SARS related viruses versa common cold coronaviruses, by employing weighting algorithms approaches.

Results: In this study, we found a few oligonucleotides, which significantly distinguish two viral groups. Among dinucleotide's features, the discrepancy of TC and CC between SARS related viruses and common cold coronaviruses was quite considerable. Furthermore, CC dinucleotide was sequentially repeated in a few multinucleotide patterns including CCA, CCAC, ACCAC, and CACCAC motifs with the highest values, which also discriminated two viral groups.

Conclusions: Theses remarkable oligonucleotides might point towards the existence of some particular RNA elements that might be involved in viral infectivity.

Background

Coronaviridae family is one of the largest groups of RNA viruses, which cause a board range of diseases in animal species and human (1). Although until 2002, it was supposed that human coronaviruses only cause a mild self-limiting respiratory disease, the emergence of severe acute respiratory syndrome (SARS) in 2003, middle east respiratory syndrome (MERS) in 2012 and recently the widespread outbreak of SARS-CoV-2 virus in 2019 has been sparked to be paid more attention to human coronaviruses as those pathogens develop pneumonia, severe acute respiratory syndrome and even death in human being (2). In contrast, some coronavirus strains such as OC43, HKU1, and NL63 and 229E have more probably association with seasonal common cold and are rarely led to severe disease in human (3–6). Although, due to the current global pandemic of SARS-CoV-2, the extensive researches are being done to expand the knowledge about virulence factors in the pathogenicity process (7–9), the mechanism of coronavirus's strains in developing mild to severe illnesses in human is ambiguous yet. Among infectious pathogens, viruses, particularly RNA viruses have regularly evolved their genome to make an adaption by host and avoid host antiviral mechanisms to finally establish a severe disease in host (10). In fact, viral genome consists of particular regions in both coding and non-coding area which interact with both viral and host factors which dictate the progression of disease (10,11). Viruses are able to change their genome during selective pressure to escape from host defense approaches (12,13). To take an example, an antiviral protein in human called, zinc-finger antiviral protein (ZAP), enables the restriction of viral replication by blocking specific sequences enriched by CG dinucleotide in viral genome (14–16), however, some RNA viruses such as HIV are able to decline the abundance of CG in their genome during evolution and eventually hamper ZAP's binding activity (17). Although the genome organization of coronaviruses was structurally determined, extra investigations on genome structure of different coronavirus species can draw up a guide for better understanding of virus evolution and discovering the potential patterns present in viral genome that might have significant role in virus life cycle and pathogenicity. The genome of coronaviruses is a positive single strand RNA with nearby 27-30 kb length, which is the biggest genome among RNA viruses. The genome organization is almost similar in all strains containing gene 1 (ORF1ab) which occupies two-thirds of the genome approximately 20 kb encoding replicase enzyme and a few number of non-structural proteins and ORF2-9, which make up only about 10 kb of genome encoding S-E-M-N (structural proteins) respectively. The 5' and 3' ends of genome contain untranslated regions (UTR) which include leader sequence in 5', several stem loops in both 5' and 3' ends and a ploy A tail in 3' end which plays significant roles

in viral replication and translation (18,19). Generally, the genome of viruses has constructed by distribution of nucleotides, which are able to create particular patterns including dinucleotides, trinucleotides and other multinucleotides that may have a vital role in viral replication and pathogenesis (20). As mentioned above, viruses are able to change their genome during evolutionary process to maximize their power against host defense activity. In this study, we decided to perform a comprehensive analysis on genome of two groups of coronaviruses, SARS related viruses and common cold coronaviruses in order to expand our knowledge about the genome structure of human coronaviruses and oligonucleotide distribution in their genome. To achieve this goal, at first the relative frequency of dinucleotides to multinucleotides were calculated and then, various attribute weighting algorithms were used to determine the discrepancy of oligonucleotide distribution in genome of two groups of human coronaviruses. Findings of current study can conduct us to identify some particular oligonucleotide patterns in genome of coronaviruses, which might have a vital role in evolutionary adaption and viral infectivity.

Results

1. Datasets generation

Generally, 5 datasets were generated based on oligonucleotide feature that each one contained 532 samples (293 viral sequences related to SARS and 239 viral sequence related to common cold) with16 dinucleotides, 64 trinucleotides,256 tetra nucleotides, 1024 penta-nucleotides, and 4096 Hexa- nucleotides respectively as oligonucleotide's attributes in each dataset. Moreover, one dataset containing all 5440 was also created to be analyzed by weighting algorithms (Sup 1).

2. Selection of the most important features

Given the data cleaning was performed to remove useless attributes, all attributes were precious and remained in each dataset. The importance of each contributing attribute in viral genome was evaluated by attribute weighting algorithms in two groups of SARS and common cold coronaviruses. Albeit a significant few number of oligonucleotide attributes were identified between two viral groups that have been presented in table 1 and Sup 2, a considerable oligonucleotide pattern was also observed which discriminated two viral groups. Briefly, CC dinucleotide got a significant value among dinucleotides attributes. Moreover, among trinucleotide features, CC dinucleotide was also repeated with a high value in CCA, GCC, and ACC features. In continue, those features were also sought among tetra oligonucleotides, which three features of CCAC, GCCG, and ACCC were identified with the significant value. Interestingly, our attention was drawn to CCAC and ACCC patterns as being also repeated with a significant weight among a few Panta and hexa-oligonucleotides fig 1. Furthermore, to identify the most important oligonucleotide pattern, all attributes from di to multinucleotides were also run at one dataset by different weighing methods. Remarkably, CACCAC oligonucleotide coupled with a few other features, was highlighted with the highest score (seven value) as shown in table 1. The result of feature selection was provided in Sup 2.

4. The location of the most significant feature on viral genome

As mentioned above, CACCAC pattern was highlighted as one of the significant features table 1. The position of CACCAC pattern was recognized on aligned sequences of each virus species. The mapping and position of this feature were illustrated on reference genomes in fig 2. We found ten conserved motifs of CACCAC in different positions on SARS-CoV-2and SARS genome. However, the only seven-conserved motif of CACCAC was identified on MERS genome. Although the most repetitions located on ORF1, CACCAC Motif was also repeated one time on S ORF of SARS and MERS and tree times on SARS ORF S. Moreover, this motif was also observed in the 3' UTR site of SARS-CoV-2and SARS. However, this motif was also identified on common cold coronaviruses genome; the number of repetition was

quite variable in each species. In addition, the motifs were not quite conserved among some strains especially HKU1 strains.

Discussion

The genome of RNA viruses contains different structures such as cis-acting elements, repeated sequences and RNA motifs, which contribute in the process of viral life cycle (11,26). In fact, these elements are able to interact with viral and cellular factors and regulate viral translation, replication and encapsidation (27,28). For instance, the presence of a particular RNA structure named internal ribosome-entry sites (IRESs) in 5' end of genome in many pathogenic viruses such as hepatitis A virus (HAV) , hepatitis C virus (HCV) and poliovirus allows them to interact with host ribosomal proteins and recruit eukaryotic translation machinery for their own proteins synthesis (29). Moreover, some other features can be involved in virus strategies for induction and regulation of host immune system. A conspicuous example of this sort of features is the existence of (pathogen-associated molecular pattern) PAMP as a small piece of RNA in viral genome. In fact, PAMPs are conserved small sequence of viral genome, or viral replication products, which are recognized by pattern-recognition receptors (PRRs) such as Toll-like receptors (TLRs) or RIGI-like receptors (RLRs) and in the following, host innate immune system, would be activated against the pathogens (30,31). In contrast with this, the presence of some other motifs or RNA elements in genome of some viruses assists them to evade host immune mechanism. As an example, an RNA structure in the 3C protease ORF of poliovirus genome inhibits the function of RNase L, an antiviral endonuclease, that is activated during viral infections as the part of innate immune system (32,33). According to the importance of these elements in viral replication and infectivity, the current study was performed to comprehensively analyze viral sequences of high virulent coronaviruses in comparison to coronaviruses related to common cold to predict a few probable significant RNA motifs. With the development of computational programs, the presence of RNA structures in viral genome has been anticipated by bioinformatics methods. Recently, feature selection techniques such as attribute weighting algorithms have already been used to predict the most important attribute in nucleotide and amino acid level among a large number of protein or genome sequences (23–25). In this study, the relative frequency of contributing oligonucleotides (dinucleotide to hexa nucleotide) in viral genome of different coronavirus strains was calculated as explained in the method section, and then the most important patterns were identified by different attribute weighting algorithms. Given the results, a sequential pattern of CC dinucleotides to CACCAC hexa nucleotides defined by almost 90 percent of all attribute weightings, were identified as the most important features to distinguish SARS and common cold coronaviruses fig 1. A few previous experiments showed that the presence of CCA boxes in viral genome, particularly the genome of positive single strand RNA viruses, would increase significant levels of transcriptional initiation at multiple sites. In fact, viral replicase seems to be able to initiate transcription from CCA boxes without the presence of a unique promoter (34). In the current study, CCA motif was shown as a remarkable feature among trinucleotides and it was repeated sequentially in CCAC, ACCAC, and CACCAC motifs. Furthermore, among all attributes features (di to hex nucleotides), CACCAC was also valued by 70 percent of all weighting models Table 1. There is a possibility that the presence of conserved multiple motifs in genome of SARS related viruses, especially, SARS and SARS-CoV-2 with the most frequency of this motif, might exert a strong influence on viral RNA synthesis. It is noticeable that this motif was also presented as a conserved motif in 3' UTR of SARS-CoV-2 and SARS but it was not observed on other coronaviruses genome in this region. According to the importance of 3' UTR sequences in viral replication and infectivity, the role of this remarkable motif should be evaluated. In this study, some other oligonucleotide features with sizable score was also distinguished between two viral groups as shown in table 1 and sup 2. To understand the biological importance of these features in life cycle of different coronavirus strains, those should be aimed and scrutinized by laboratory techniques in cell culture system and animal models.

Among dinucleotide features, TC and CC dinucleotides, which were confirmed by 80 and 90 percent of all attribute weighting respectively, attracted us too. According to a myriad number of researches, dinucleotide composition

constitutes a genomic signature among a variety of virus species, which might represent a significant impact on viral life cycle and host adaption (20,35). To exemplify, the reduced frequency of UA and UU dinucleotides in HCV genome lead to the interferon (INF) resistance among some HCV genotypes (36). In some other research, it is supposed that frequency of CG and UA enables RNA viruses to escape from host immune system (17). In this study, the relative frequency of TC and CC dinucleotides in SARS related viruses were significantly different from those of common cold coronaviruses. It can be supposed that TC and CC dinucleotides represent an important role in coronaviruses pathogenicity. Interestingly, there is a human enzyme named Apo lipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3 (APOBEC3), which has an effective role in innate antiviral immunity especially about retroviruses and DNA viruses (37,38). The preferred effective sites of two main isoforms of this enzyme, APOBEC3A and APOBEC3 G were reported as TC and CC respectively (39). Both of mentioned dinucleotides were as distinguishing features between two viral groups in the current study. Although in most of studies, the antiviral activity of this enzyme has been identified on retroviruses and DNA viruses. The recent study on NL63 coronavirus showed that replication of RNA viruses can be also restricted by APOBEC3 activity (40). It can be hypothesized that the difference of TC and CC dinucleotides in genome of two groups of coronaviruses is more likely in the result of evolutionary process and thus it can have a substantial role in viral pathogenicity.

Conclusion

To conclude, this mining showed us a few highlighted oligonucleotide features that differed in genome of two groups of common cold and SARS coronaviruses. Those features might contribute to a better understanding of coronaviruses pathogenicity and encountering in innate immunity in the future.

Methods

1. Viral Genome Sequences

For the beginning, the nucleotide database of NCBI was searched for each virus species including SARS-CoV-2, SARS, MERS, HKU1, OC43, NL63, and 229E viruses to obtain full-length genome sequences of each strain. Totally, nearby a hundred full genome sequence of each virus species were retrieved as initial data. However, in the case of NL63, 229E and HKU1, the numbers of deposited full genomes were less than 100. To confirm that, the retrieved sequences belonged to the same species, the multiple sequences alignments were computed using CLUSTAL Omega algorithm in EBI web service. Finally, after checking the aligned sequences and excluding some genomes related to animal species, the final initial data for each human virus strain was created. The more detailed information of viral sequences was summarized in table 2.

2. Oligonucleotide's frequency analyses and attributes extraction

In order to carry out the preliminary analysis, a Hyper Talk program was written in the lab view software, which accepted Fasta text -formatted files. The written program in the software was able to scan the sequences sequentially and build up the overall nucleotide composition, alongside with the frequency of each oligonucleotide in turn. In this study, the frequency of dinucleotides to hexa oligonucleotides for each sequence was computed as an observed oligonucleotide in the lab view software (21). On completion of the scan, the expected numbers of a given oligonucleotide were also calculated using Markov method (22). To avoid the effects of length factor of the sequences and estimate the level of statistical significance of oligonucleotides occurrences, the observed to expected oligonucleotides ratios were obtained (21) and finally, each oligonucleotide odds ratio was considered as an attribute. Totally 5440 attributes (16 dinucleotides, 64 trinucleotides, 256 tetra nucleotides, 1024 penta-nucleotides, and 4096

Hexa-nucleotides) were extracted for each virus sequence by lab view software. List of attributes and calculated values were presented in Sup 3.

In the following, a new dataset was generated for each oligonucleotide feature in two viral groups; the viral sequences related to the Severe Acute Respiratory Syndrome (SARS) including SARS-CoV-2, SARS-CoV, and MERS viruses and the viral sequences related to common cold including OC43, HKU1, NL63, and 229E. Then, the attributes of SARS related viruses were compared with those related to common cold coronaviruses (Sup 1). For this aim, each dataset was imported into Rapid Miner Software [Rapid Miner, Germany] and the following steps were sequentially done. The processes of datasets creation and data mining are outlined in the fig 3.

3. Data Mining

3.1. Data Filtering

In order to get a final cleaned database (FCdb), any duplicated attributes, useless and related attributes with Pearson correlation coefficient greater than 0.9 and also numerical attributes with standard deviations less than or equal to a given deviation threshold (0.1) were excluded from the datasets (23).

3.2. Attribute weighting

Ten different algorithms of attribute weightings named Information Gain, Information Gain Ratio, Rule, Deviation, Chi Squared, Gini Index, Uncertainty, Relief, Support Vector Machine (SVM), and PCA (24,25) were performed on all datasets to achieve the most important nucleotide's attributes that probably discriminate coronaviruses which cause SARS against those which are known as common cold coronaviruses. During execution of attribute weighting program, each attribute gained a value between 0-1 showing its importance. Then, the attribute with a weight higher than 0.7 owning the highest number of weighting algorithms was allocated as the most important attribute. All Attributes and the relevant weighting models have been presented in Sup 1.

Abbreviations

middle east respiratory syndrome (MERS), severe acute respiratory syndrome (SARS), untranslated regions (UTR), internal ribosome-entry sites (IRESs), hepatitis A virus (HAV) , hepatitis C virus (HCV), (pathogen-associated molecular pattern) PAMP, pattern-recognition receptors (PRRs), Toll-like receptors (TLRs), RIGI-like receptors (RLRs), interferon (INF), final cleaned database (FCdb), Support Vector Machine (SVM)

Declarations

Acknowledgments

We are very thankful to all the technicians who provided support during the course of this research.

Funding

The present study was supported by the Kerman University of Medical Sciences (Grant number: 98001248).

Availability of data and materials

The nucleotide database of NCBI was searched to extract each virus sequence.

Authors' contributions

Author contributions ME and ABZ designed the research. ME, NBM and TA performed the experiments. ABZ, ME, and AA contributed to analysis and interpretation of data. ME, HM and ABZ wrote the manuscript.
All authors reviewed the manuscript.

Ethics approval and consent to participate

The study was approved by the Ethical Committee of Kerman University of Medical Sciences (IR.KMU.REC.1398.730)

Competing Interests

The authors have declared that there are no conflicts of interest.

Consent for publication

Not applicable.

References

1. Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. Microbiol Mol Biol Rev. 2005;69(4):635–64.
2. El Zowalaty ME, Järhult JD. From SARS to COVID-19: A previously unknown SARS-CoV-2 virus of pandemic potential infecting humans–Call for a One Health approach. One Heal. 2020;100124.
3. Vabret A, Mourez T. An Outbreak of Coronavirus OC43 Respiratory Infection in Normandy , France. 2003;2002(August 2002):985–9.
4. Abdul-Rasool S, Fielding BC. Understanding human coronavirus HCoV-NL63. Open Virol J. 2010;4:76.
5. Pyrc K, Berkhout B, Van Der Hoek L. The novel human coronaviruses NL63 and HKU1. J Virol. 2007;81(7):3051–7.
6. Dijkman R, Van Der Hoek L. Human coronaviruses 229E and NL63: close yet still so far. J Formos Med Assoc. 2009;108(4):270–9.
7. Liang Q, Li J, Guo M, Tian X, Liu C, Wang X, et al. Virus-host interactome and proteomic survey of PMBCs from COVID-19 patients reveal potential virulence factors influencing SARS-CoV-2 pathogenesis. bioRxiv. 2020;
8. Zhang C, Wu Z, Li J-W, Zhao H, Wang G-Q. The cytokine release syndrome (CRS) of severe COVID-19 and Interleukin-6 receptor (IL-6R) antagonist Tocilizumab may be the key to reduce the mortality. Int J Antimicrob Agents. 2020;105954.
9. Sarkar J, Guha R. Infectivity, virulence, pathogenicity, host-pathogen interactions of SARS and SARS-CoV-2 in experimental animals: a systematic review. Vet Res Commun. 2020;1–10.
10. Kloc A, Rai DK, Rieder E. The roles of picornavirus untranslated regions in infection and innate immunity. Front Microbiol. 2018;9:485.
11. Liu Y, Wimmer E, Paul A V. Cis-acting RNA elements in human and animal plus-strand RNA viruses. Biochim Biophys Acta (BBA)-Gene Regul Mech. 2009;1789(9–10):495–517.
12. Ibrahim A, Fros J, Bertran A, Sechan F, Odon V, Torrance L, et al. A functional investigation of the suppression of CpG and UpA dinucleotide frequencies in plant RNA virus genomes. Sci Rep. 2019;9(1):1–14.
13. Wang Y, Mao J-M, Wang G-D, Qiu Z, Yao Q, Chen K-P. Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. 2020;
14. Ficarelli M, Antzin-Anduetza I, Hugh-White R, Firth AE, Sertkaya H, Wilson H, et al. CpG dinucleotides inhibit HIV-1 replication through zinc finger antiviral protein (ZAP)-dependent and-independent mechanisms. J Virol. 2020;94(6).

15. Miyazato P, Matsuo M, Tan BJY, Tokunaga M, Katsuya H, Islam S, et al. HTLV-1 contains a high CG dinucleotide content and is susceptible to the host antiviral protein ZAP. Retrovirology. 2019;16(1):38.
16. Luo X, Wang X, Gao Y, Zhu J, Liu S, Gao G, et al. Molecular mechanism of rna recognition by zinc-finger antiviral protein. Cell Rep. 2020;30(1):46–52.
17. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature. 2017;550(7674):124–7.
18. Kumar S, Nyodu R, Maurya VK, Saxena SK. Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). In: Coronavirus Disease 2019 (COVID-19). Springer; 2020. p. 23–31.
19. Xu J, Hu J, Wang J, Han Y, Hu Y, Wen J, et al. Genome organization of the SARS-CoV. Genomics Proteomics Bioinformatics. 2003;1(3):226–35.
20. Yin C. Dinucleotide repeats in coronavirus SARS-CoV-2 genome: evolutionary implications. arXiv Prepr arXiv200600280. 2020;
21. Rima BK, McFerran N V. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J Gen Virol. 1997;78(11):2859–70.
22. Pride DT, Wassenaar TM, Ghose C, Blaser MJ. Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics. 2006;7(1):8.
23. KayvanJoo AH, Ebrahimi M, Haqshenas G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Res Notes. 2014 Aug;7(1):565.
24. Ebrahimie E, Ebrahimi M, Sarvestani NR, Ebrahimi M. Protein attributes contribute to halo-stability, bioinformatics approach. Saline Systems. 2011;7(1):1.
25. Ebrahimi M, Lakizadeh A, Agha-Golzadeh P, Ebrahimie E, Ebrahimi M. Prediction of thermostability from amino acid attributes by combination of clustering with attribute weighting: a new vista in engineering enzymes. PLoS One. 2011;6(8):e23146.
26. Newburn LR, White KA. Cis-acting RNA elements in positive-strand RNA plant virus genomes. Virology. 2015;479:434–43.
27. Stewart H, Bingham RJ, White SJ, Dykeman EC, Zothner C, Tuplin AK, et al. Identification of novel RNA secondary structures within the hepatitis C virus genome reveals a cooperative involvement in genome packaging. Sci Rep. 2016;6:22952.
28. Lozano G, Martínez-Salas E. Structural insights into viral IRES-dependent translation mechanisms. Curr Opin Virol. 2015;12:113–20.
29. Kieft JS. Viral IRES RNA structures and ribosome interactions. Trends Biochem Sci. 2008;33(6):274–83.
30. Takeuchi O, Akira S. Innate immunity to virus infection. Immunol Rev. 2009;227(1):75–86.
31. Takeda K, Akira S. Toll-like receptors in innate immunity. Int Immunol. 2005;17(1):1–14.
32. Townsend HL, Jha BK, Han J-Q, Maluf NK, Silverman RH, Barton DJ. A viral RNA competitively inhibits the antiviral endoribonuclease domain of RNase L. Rna. 2008;14(6):1026–36.
33. Han J-Q, Townsend HL, Jha BK, Paranjape JM, Silverman RH, Barton DJ. A phylogenetically conserved RNA structure in the poliovirus open reading frame inhibits the antiviral endoribonuclease RNase L. J Virol. 2007;81(11):5561–72.
34. Yoshinari S, Nagy PD, Simon AE, Dreher TW. CCA initiation boxes without unique promoter elements support in vitro transcription by three viral RNA-dependent RNA polymerases. Rna. 2000;6(5):698–707.

35. Gu H, Fan RLY, Wang D, Poon LLM. Dinucleotide evolutionary dynamics in influenza A virus. Virus Evol. 2019;5(2):vez038.
36. Washenberger CL, Han J-Q, Kechris KJ, Jha BK, Silverman RH, Barton DJ. Hepatitis C virus RNA: dinucleotide frequencies and cleavage by RNase L. Virus Res. 2007;130(1–2):85–95.
37. Stavrou S, Ross SR. APOBEC3 proteins in viral immunity. J Immunol. 2015;195(10):4565–70.
38. Warren CJ, Van Doorslaer K, Pandey A, Espinosa JM, Pyeon D. Role of the host restriction factor APOBEC3 on papillomavirus evolution. Virus Evol. 2015;1(1).
39. Rathore A, Carpenter MA, Demir Ö, Ikeda T, Li M, Shaban NM, et al. The local dinucleotide preference of APOBEC3G can be altered from 5'-CC to 5'-TC by a single amino acid substitution. J Mol Biol. 2013;425(22):4442–54.
40. Milewska A, Kindler E, Vkovski P, Zeglen S, Ochman M, Thiel V, et al. APOBEC3-mediated restriction of RNA virus replication. Sci Rep. 2018;8(1):1–12.

Tables

Table 1. The most important oligonucleotide features that were marked by different weighting algorithms

Di	value	Tri	value	Tetra	value	Panta	value	Hexa	value	all oligonucleotide	value
CC	9	GGG	8	GGGA	8	ACCAC	9	ACGAAA	8	ACGAAA	8
TC	8	CCA	8	AGGG	8	AGTGT	8	TTAAGG	8	AGGGC	8
CT	7	GCC	8	GCCG	8	TCCTT	8	GAGCTA	8	TTAAGG	8
		AGG	7	ACTC	8	AGGGA	8	GACTCA	8	GAGCTA	8
		ACC	7	CTTC	8	TCTAT	8	GTGAAG	8	GACTCA	8
		TCT	7	AAGC	8	TAGGA	8	CACCAC	8	GTGAAG	8
		TTC	7	CCAC	8	GGAGC	8			CACCAC	7
		CTT	7	ATTC	7	GGGAG	8				
				GAGT	7	GAACT	8				
				GGGC	7	AGGGC	8				
				TCTT	7	GGGCT	8				
				ACCC	7	TGTCT	8				
						AAAGC	8				
						CAGCC	8				

Table 2. Human coronavirus strains

Viral sequences of Coronaviridae family				Type of disease	
Genus	Subgenus	Species	The number of extracted sequences	Common cold	Severe respiratory disease
alpha	Duvinacovirus	229E	25	*	
	Setracovirus	NL63	60	*	
Beta	Embecovirus	OC43	121	*	
		HKU1	33	*	
	Merbecovirus	MERS	98		*
	Sarbecovirus	SARS-CoV	99		*
		SARS-CoV2	96		*

Figures

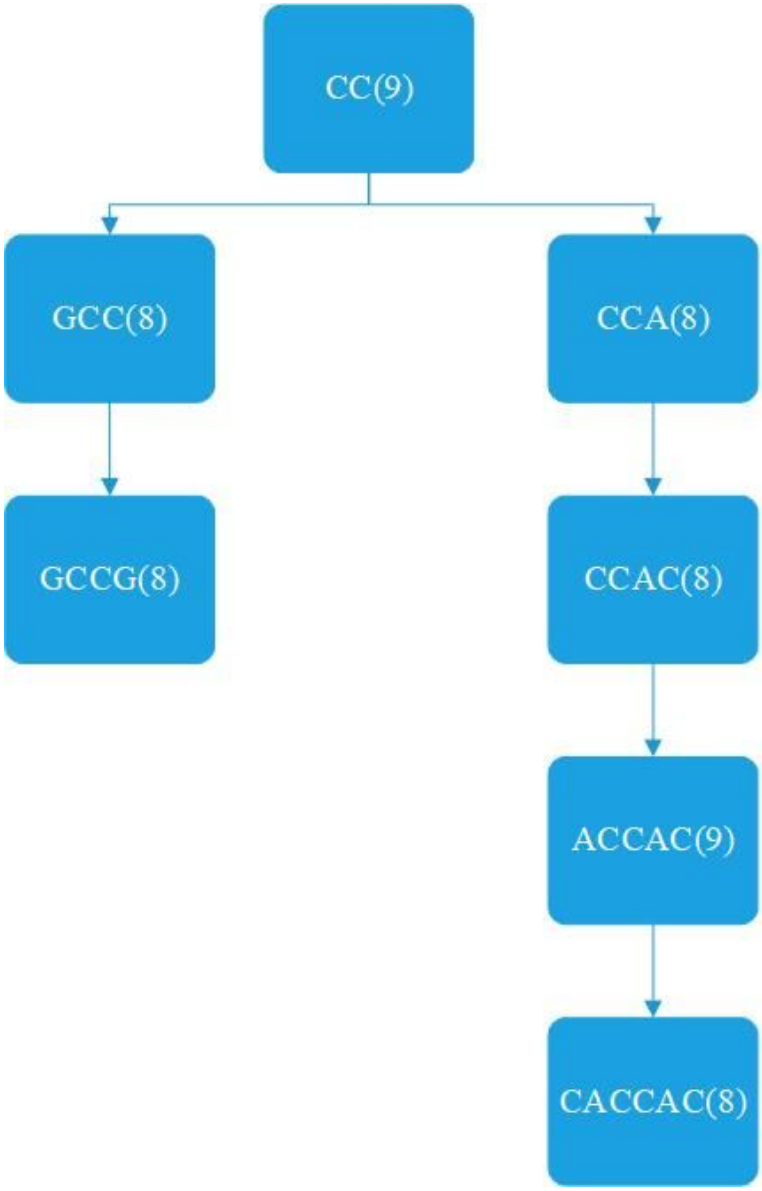

Figure 1

The sequential pattern with the highest value, which discriminates the genome of SARS- related coronaviruses from common cold coronaviruses

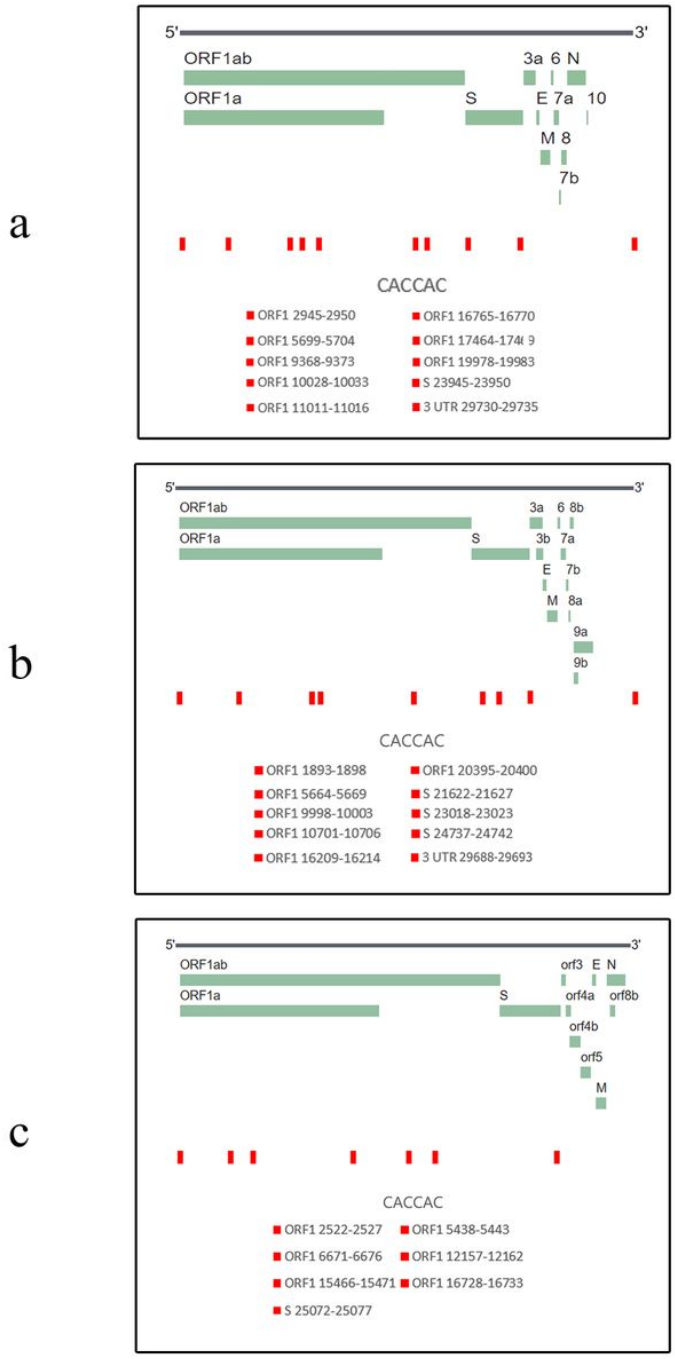

Figure 2

The location of CACCAC feature on viral genome. a. SARS-CoV-2 genome, b. SARS genome, c. MERS genome

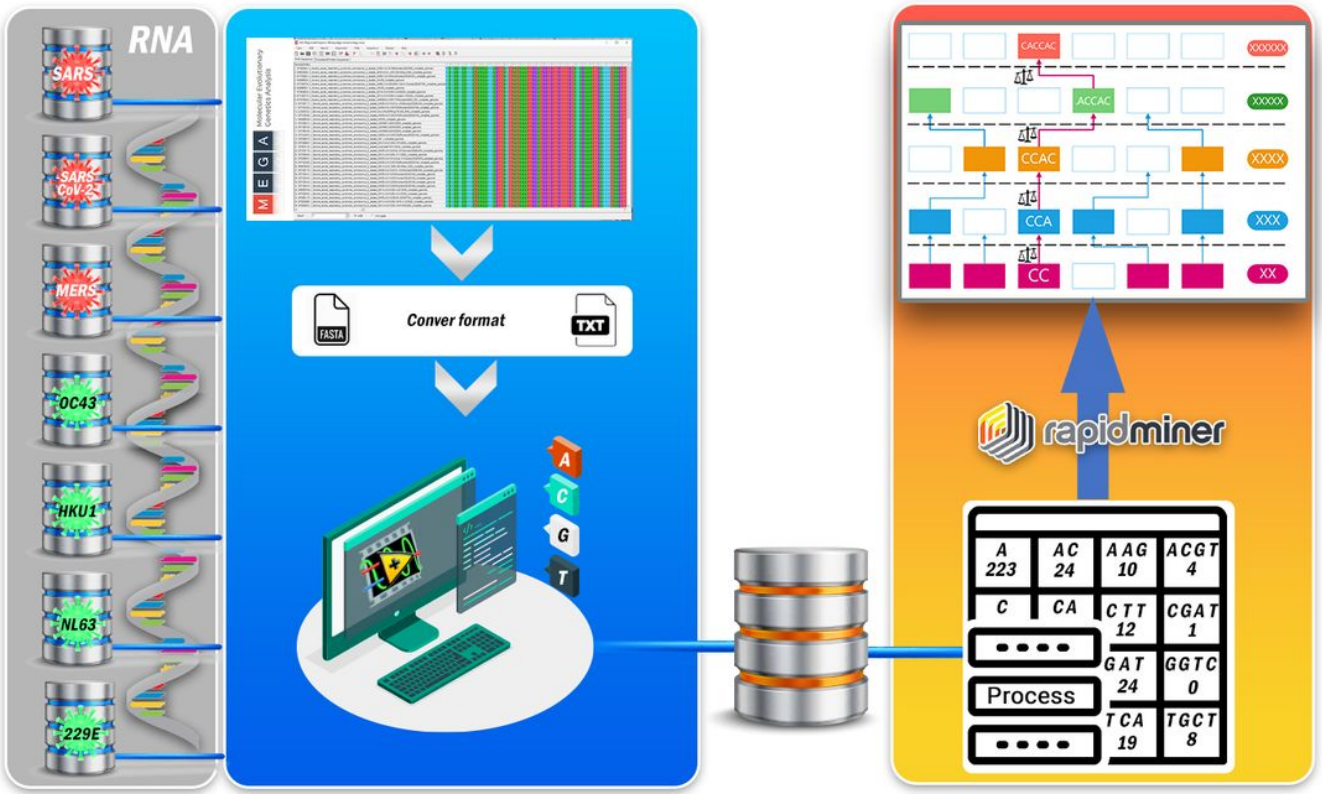

Figure 3

The overview of data mining process on coronaviruses genome

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- sup1.rar
- sup2.rar
- suplementary3.rar