Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Theoretical Biology 530 (2021) 110885

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Inverted repeats in coronavirus SARS-CoV-2 genome manifest the evolution events

Changchuan Yin [a,*], Stephen S.-T. Yau [b *]

[a] Department of Mathematics, Statistics, and Computer Science, The University of Illinois at Chicago, Chicago, IL 60607-7045, USA
[b] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 7 June 2021
Revised 13 August 2021
Accepted 25 August 2021
Available online 31 August 2021

Keywords:
COVID-19
SARS-CoV-2
2019-nCoV
Coronavirus
Genome
Inverted repeat
Palindrome
Evolution

ABSTRACT

The world faces a great unforeseen challenge through the COVID-19 pandemic caused by coronavirus SARS-CoV-2. The virus genome structure and evolution are positioned front and center for further understanding insights on vaccine development, monitoring of transmission trajectories, and prevention of zoonotic infections of new coronaviruses. Of particular interest are genomic elements Inverse Repeats (IRs), which maintain genome stability, regulate gene expressions, and are the targets of mutations. However, little research attention is given to the IR content analysis in the SARS-CoV-2 genome. In this study, we propose a geometric analysis method and using the method to investigate the distributions of IRs in SARS-CoV-2 and its related coronavirus genomes. The method represents each genomic IR sequence pair as a single point and constructs the geometric shape of the genome using the IRs. Thus, the IR shape can be considered as the signature of the genome. The genomes of different coronaviruses are then compared using the constructed IR shapes. The results demonstrate that SARS-CoV-2 genome, specifically, has an abundance of IRs, and the IRs in coronavirus genomes show an increase during evolution events.

1. Introduction

Coronavirus (CoV) SARS-CoV-2 is the causative agent of the global COVID-19 pandemic, which initially emerged in Wuhan, China, in December 2019. The spread of SARS-CoV-2 is a severe threat to global health, and there are over 157 million confirmed cases and more than 3.27 million deaths worldwide as of May 10, 2021, according to WHO (Max Roser et al., 2020). From the current global COVID-19 status, it is predictable that SARS-CoV-2 may exist with humans for a long-time. Understanding the virus genomic structures and evolution may offer deep insights into infection mechanism, vaccine design, therapeutic drug development, tracking the transmission, and prevention of future infections. We also aim to understand the zoonotic origin and evolution of the coronavirus SARS-CoV-2 from genome structure perspective.

Coronaviruses (CoVs) belong to the family Coronaviridae, consisting of a group of positive-sensed, single-stranded RNA viruses. To date, seven human CoVs (HCoVs) have been identified and researched (Cui et al., 2019; Fan et al., 2019). The first coronaviruses that infected humans, designated HCoV-229E and HCoV-OC43, were alpha-CoVs, reported in the 1960s. The individual symptoms evoked by HCoV-229E, HCoV-OC43 are mild. The five beta-CoVs are human-CoV/OC43, human-CoV/HKU1, severe acute respiratory syndrome coronavirus (SARS-CoV, 2003), Middle East respiratory syndrome coronavirus (MERS-CoV, 2012), and SARS-CoV-2 (2019). SARS-CoV, MERS-CoV, and SARS-CoV-2 are highly virulent, causing severe lower respiratory tract infection and extrapulmonary manifestations. In addition, SARS-CoV-2 is more transmissible compared to SARS-CoV and MERS-CoV, and is characterized by asymptomatic carriers, long latency period, high infectivity, and relatively high mortality. Two human coronaviruses, HCoV-HKU1 and HCoV-NL63 (Woo et al., 2005; Van Der Hoek et al., 2004), were identified after the SARS-CoV epidemic as common causes of human respiratory tract infections. In addition to these seven human CoVs, SARS-related bat-CoVs and animal CoVs were discovered. bat-CoV/HKU2 (2007) was identified Chinese horseshoe bat *Rhinolophus* (Lau et al., 2007. Swine acute diarrhea syndrome coronavirus (swine-CoV/SADS, 2017) is highly pathogenic and has broad hosts with inherent infection potential to humans(Edwards et al., 2020). The most closely related virus to SARS-CoV-2 is RaTG13 (2013), identified from a *Rhinolophus affinis* bat (Zhou et al., 2020b). RaTG13 shares 96.1% nucleotide identity with SARS-CoV-2. Recently, bat-CoV/RmYN02 (2020) was discovered in *Rhinolophus malayanushe* (Zhou et al., 2020a), exhibiting 93.3% nucleotide sequence identity with SARS-CoV-2.

* Corresponding authors.
E-mail addresses: cyin1@uic.edu (C. Yin), yau@uic.edu (S.S.-T. Yau).

https://doi.org/10.1016/j.jtbi.2021.110885
0022-5193/

A new bat-CoV/RpYN06 (2020) is also a relative of SARS-CoV-2 (Zhou et al., 2021).

A genome holds both explicit and implicit structural elements that confer optimal and precise functions. The explicit structures include protein-coding regions, regulatory regions, 5'- and 3'-UTRs, and the poly-A tail (Kim et al., 2020). These explicit structures of SARS-CoV-2 genome have been extensively studied. However, the implicit structures are often unexplored. The implicit genomic structures are critical to maintaining the global and local structures so that genome replication, transcription, and translation can occur in an optimal environment. Examples of the implicit structures in a genome include hidden genomic signals such as periodicity-3 in protein-coding regions, and periodicity-10.5 in dinucleotide-rich regions, tandem repeats. Particularly, RNA virus genomes contain special implicit structures, inverted repeats (IR), which can stabilize the genomes and regulate gene expressions.

An inverted repeat (IR) is a sequence that matches its downstream reverse complement sequence. The initial sequence and the reverse complement may have a spacer, which can vary from zero to thousands of bases. An IR of zero spacer is specially named a palindrome. For example, the inverted repeat, 5'-TTTACGTAAA-3' is a palindrome, the palindrome-first is 5'-TTTAC-3', and the palindrome-second is 5'-GTAAA-3'. When the spacer in an inverted repeat is non-zero, the repeat is generally inverted. For convenience, we still denote the initial sequence in a general IR as a palindrome-first and the downstream reverse complement as a palindrome-second. For example, in the general IR, 5'-AAAGGCT...AGCCTTT-3', we still name the palindrome-first as 5'-AAAGGCT-3', and the palindrome-second as 5'-AGCCTTT-3'. Through self-complementary base pairing, an IR can form a stem-loop structure in an RNA molecule, where the palindrome-first and palindrome-second make a stem, and the spacer makes a loop. Note that an IR may not have perfect complementary base pairing in palindrome-first and palindrome-second sequences so the stem formed by an imperfect IR can have mismatches, insert, or deletions.

Biologically, IRs have important functions and are prevalent in numerous virus, bacterial and eukaryotic genomes. IRs in a genome greatly enhance the stability of genome structure and help regulate gene expression. IR sequences are the principal elements in archaeal and bacterial CRISPR-CAS systems (Mojica et al., 2005), which function as adaptive antiviral defense systems. In DNA genomes, IRs from transcription can form double-strand RNA (dsRNA) to repress gene expression (Muskens et al., 2000). IRs delimit the boundaries of transposons in genome evolution and form stem-loop structures in retaining genome stability and flexibility. IRs are described as hot-spots of eukaryotic and prokaryotic genomic stability (Voineagu et al., 2008), replication (Pearson et al., 1996), and gene silencing (Adelman et al., 2002; Buchon and Vaury, 2006; Selker, 1999). Previous studies showed that IRs play key roles in cellular and virus evolution (Lavi et al., 2018), genetic diversity (Čechová et al., 2018), and cancer (Zou et al., 2017).

Despite the importance of IRs in genomes, few investigations on IRs in SARS-CoV-2 genome have been conducted. Our previous study revealed that SARS-CoV-2 genome has an abundance of inverted repeats and the IRs are mainly located in the gene of the Spike protein and N protein (Yin and Yau, 2021). Limanskaya showed that coronaviruses including SARS-CoV genomes have predominated IRs in the genes encoding replicase and spike glycoproteins of coronaviruses (Limanskaya, 2009). Bartas et al. revealed that Nidovirales including SARS-CoV-2 genomes have enriched IRs (Bartas et al., 2020). Goswami et al.'s study showed that SARS-CoV-2 hot-spot mutations are predominantly distributed within IRs and CpG island loci (Goswami et al., 2020). Yet the dynamic changes of IRs and the impacts on virus evolution have not been fully investigated.

To further reveal the evolutionary history of SARS-CoV-2 related coronaviruses, in this study, we present a novel geometric method to represent genome shape by Delaunay triangulations of IR points formed by palindrome-firsts and palindrome-seconds. The IR shape of a genome can be used as the signature of a genome. The IRs shapes of different genomes are compared for similarity by the Hausdorff distance. Therefore, we may track the IRs in virus evolution. Moreover, this study demonstrates the prevalence of IRs in SARS-CoV-2 and human-CoV genomes and highlights the significance of IRs in the evolution of SARS-CoV-2.

2. Materials and methods

2.1. Barcoding genomes using inverted repeats

To ensure consistency in comparing different coronavirus genomes, we only extract the IRs from genomes with the perfect complementary base-pairing of the palindrome-first and palindrome-second sequences. Note that a short IR of length P can occur inside a long IR of length Q ($Q > P$), for example, IRs of length 10 contain IRs of lengths 9, 8, etc. We only keep the IR of length Q and exclude the IR of length P. In some cases, two different IRs can have overlapped sequences, we still consider these two IRs are unique.

The retrieved genome IRs are mapped on the protein genes of a genome according to the positions of the palindrome-first and palindrome-second sequences of the IRs.

In the IRs of a genome, the palindrome-first and palindrome-second sequences have a strong tendency to form a stem structure when the length of the palindrome-first or palindrome-second sequence is long or when the spacer between the palindrome-first or palindrome-second sequences is short. The positions of the formed stems can be determined by the positions of the palindrome-first or palindrome-second sequences. We define an IR stem point in a genome as follows.

Definition 2.1. Let RNA sequence of length m be at the position $(x, x+m)$ in a genome, and the sequence has a downstream reverse complement sequence at position $(y, y+m)$ in the same genome, the first segment is called palindrome-first and the downstream sequence is called palindrome-second. An IR stem point is formed by the palindrome-first and palindrome-second in the genome. The coordinate of the IR stem point is (x, y).

Therefore, we can represent an RNA genome structure as the abstract IR points in the genome. Each point represents a possible stem formed by the palindrome-first or palindrome-second sequences of an IR in the genome. Because the distribution of the palindrome-first or palindrome-second sequences along a genome is similar to a bar-code, we propose and name the IR distribution as the barcode of the genome. The IR barcode can be used to characterize the genome since each genome contains a unique distribution of IRs.

2.2. Delaunay triangulation of IRs in a genome

The IR points constructed by the palindrome-first and palindrome-second sequences represent the stem structures that stabilize the whole genome and regulate gene expressions. If we connect the IR points as a graph, we can visualize a virus genome by the abstract shape of the IR structures. Here, we propose to transform the IR point sets into the characteristic Delaunay triangulation by tessellations. Because of the stem forming of an RNA virus genome by IRs, the interconnected Delaunay triangles from IR points can be considered the virtual structure of an RNA virus genome.

The Delaunay triangulation provides a unique way of triangulating the convex hull spanned by the set of IR points. The Delaunay tessellation is defined such that inside the interior of the circumcircle of each Delaunay triangle, no other IR points are present. Tessellation adapts to both the local density and geometry of the point distribution. Where the density is high, the triangles are small, and vice versa. The sizes of the triangles are, therefore, a measure of the local density of the point distribution. Therefore, we may use the Delaunay triangulation of the IR joining points in a genome to represent the distribution structure of IRs in an RNA virus genome.

The colors of a triangulation graph are to illustrate the shape of the graph. The darker blue colors are for smaller coordinates for lower and left positions, while the darker red colors are for larger coordinates to upper and right positions.

2.3. Metric of genome IR barcodes: Hausdorff distance

In an RNA genome, the stem structure formed by an IR palindrome-first and palindrome-second can be characterized as an IR point. Because of the distinctive distribution of IR points in a genome, the numbers of IR points in two different genomes are usually different. To compare two given genomes using IR barcodes, we propose measuring the Hausdorff distance of the IR point sets from the genomes. The resulting Hausdorff distance represents the dissimilarity of two genomes which are represented by IR points, or equally, IR barcodes.

The Hausdorff distance, named after Felix Hausdorff (1868–1942), is the maximum distance of a set to the nearest point in the other set. The Hausdorff distance can determine the degree of resemblance between two objects or point sets that are superimposed on one another (Huttenlocher et al., 1993). The Hausdorff distance is defined as

$$d_H(X,Y) = \max\left\{ \sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y) \right\},$$

where sup denotes the supremum and inf denotes the infimum.

Because two IR graphs may not have the same number of nodes, the direct one-to-one comparison of two graph nodes is impossible. The similarity of two IR graphs can be measured by the Hausdorff distance. If two IR graphs are similar, then more parts can be superimposed, the Hausdorff distance is small, otherwise, the Hausdorff distance is large.

The following is the procedure for the IR graph analytics in a genome. (1) By string matching, scan the whole genome for inverted repeats using sliding-window approach. The coordinates of an IR is two positions of the palindrome-first and the palindrome-second of the IR. The space of an IR can be any length. (2) Delaunay triangulation of IR in a genome as a graph using the coordinates of the IRs. (3) To compare two genomes in the context of IRs, the Hausdorff distance of the IR graphs from the two genomes is computed.

2.4. Computer programs and genome data

To identify and analyze IRs in genomes, the complete genomes of coronaviruses were scanned for IRs by in–house computer programs in this study. The computer programs for IR analysis including IR retrieval, similarity, visualization, and shape construction were written in Python. The core algorithm in IR scanning was string fuzzy search by sliding-windows along a sequence. The IR retrieving programs were compared and validated with Palindrome analyzer, a web-based server for IRs (Brázda et al., 2016). The retrieved IRs in SARS-CoV-2 genome is provided as supplementary materials.

This study depends on the complete genomes of human and bat coronaviruses that were downloaded from NCBI GenBank or the GISAID repository (http://www.GISAID.org) (Shu and McCauley, 2017). The NCBI GenBank/GISAID access numbers and references of the human and bat coronaviruses are as follows. SARS-CoV-2 (GenBank: NC_045512.2) (Wu et al., 2020), SARS-CoV/BJ01 (GenBank: AY278488), SARSr-CoV/RaTG13 (GenBank: MN996532) (Zhou et al., 2020b), bat-CoV/RmYN02 (GISAID: EPI_ISL_412977) (Zhou et al., 2020a), bat-CoV/RpYN06 (GISAID: EPI_ISL_1699446) (Zhou et al., 2021), MERS-CoV (GenBank: NC_019843) (Zaki et al., 2012), SARSr-CoV/ZC21 (GenBank: MG772934) (Hu et al., 2018), SARSr-CoV/ZC45 (GenBank: MG772933) (Hu et al., 2018), swine-CoV/SADS (GenBank: MG557844) (Zhou et al., 2018), pangolin-CoV/MP789 (GenBank: MT121216) (Lam et al., 2020; Liu et al., 2020), civets-SCoV/SZ3 (GenBank: AY304486) (St-Jean et al., 2004), and human-CoV/OC43 (GenBank: AY585229) (St-Jean et al., 2004).

3. Results

3.1. SARS-CoV-2 genome has abundant IRs

Since long IRs are deemed to have great influence on the stability of genomes of various organisms, we only examine the relatively long perfect IRs in SARS-CoV-2 genome. The longest IR identified in SARS-CoV-2 genome is 15 bp sequence, the palindrome-first sequence 5'-ACTTACCTTTTAAGT-3' is at 8474–8489 (nsp3 gene), and the palindrome-second sequence 5'-ACTTAAAAGGTAAGT-3' is at 13295–13310 (nsp10 gene). Using the proposed IR identification and analysis methods, we produce the distribution of IRs in SARS-CoV-2 genome, the IR points formed by the positions of the palindrome-first and the palindrome-second sequences, and the Delaunay triangulation of IR points of SARS-CoV-2 genome (Fig. 1). Fig. 1(a) shows the barcode signature of the genome by palindrome-first and the palindrome-second sequences of 12–15 bp, and Fig. 1(b) shows the coordinates of the IR points formed by the palindrome-first and the palindrome-second sequences. Fig. 1(b) and (d) are the construction and shape rendering of the Delaunay triangulation of IR points, respectively.

3.2. Relevance analysis of SARS-CoV-2 and its related CoVs by IR distributions

As IRs can stabilize an RNA genome and control virus replication and transcription (Pearson et al., 1996), the genomic IRs analysis may provide deep insights on the evolution of CoVs, especially on the zoonotic origin of SARS-CoV-2. To infer the zoonotic origin of SARS-CoV-2, we here compare the IR distributions from typical bat-CoV. We evaluate and compare the distributions of IRs of 12–15 bp in four SARS CoV genomes: SARS-CoV-2 (Fig. 1), SARSr-CoV/ZC21, SARSr-CoV/RaTG13, SARSr-CoV/RmYN02, and SARS-CoV/BJ01 (Fig. 2), and four bat-CoV genomes: bat-CoV/Pangolin, MERS-CoV, swine-CoV/SADS, and bat-CoV/ZC45 (Fig. 3).

From the qualitative and quantitative analysis, we observe that SARS-CoV-2 strain (Fig. 1(d)) is more closely related to SARSr-CoV/ZC21 and SARSr-CoV/RaTG13 (Fig. 2(a,b)) than SARSr-CoV/RmYN02 and SARS-CoV (Fig. 2(c,d)).

The IR distribution results (Table 1) show that human CoVs such as human-CoV/229E and human-CoV/NL63, and human-CoV/HKU1 have a higher number IRs than bat-CoVs. The reason for the increased IRs is possibly due to the mutations and recombinations in CoV genomes during the infection periods in human hosts. Therefore, we may use the IR changes to infer the evolution trend of CoVs from bat to human host. SARS-CoV-2 may probably have

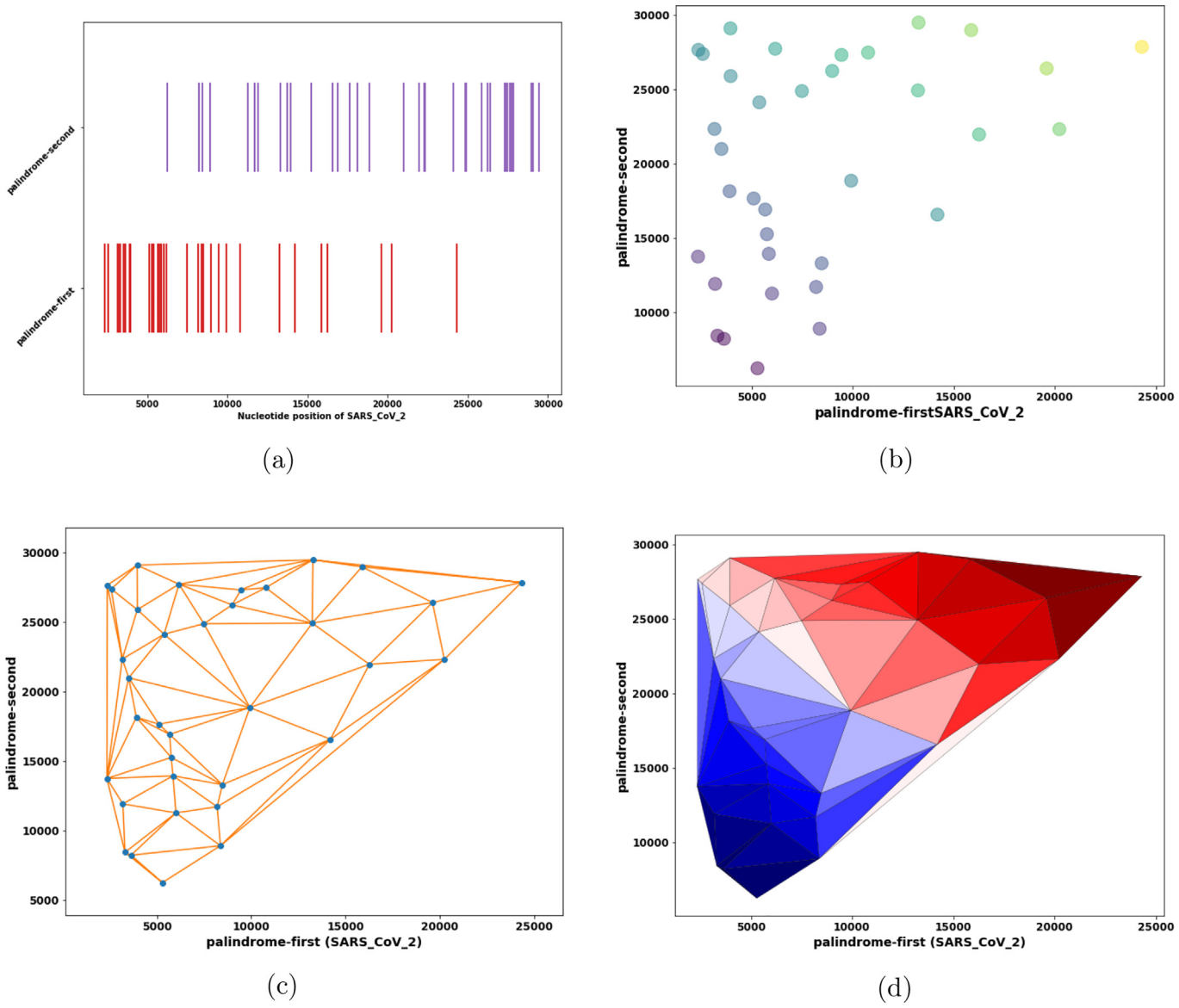


Fig. 1. Distribution of IRs consists of the set of palindrome-first sequences and the palindrome-second sequences in SARS-CoV-2 genome. The lengths of IRs are from 12 to 15 bp. (a) The positional distribution of the palindrome-first sequences and the palindrome-second sequences. (b) Scatter IR points formed by the positions of the palindrome-first sequences and the palindrome-second sequences in the genome. (c) Delaunay triangulation of the IR points. (d) Delaunay triangulation of the IR points with colored in adjacent triangles. The darker blue colors indicate smaller coordinates for lower and left positions, while the darker red colors indicate larger coordinates to upper and right positions.

evolved from SARSr-CoV/RaTG13 from the distribution of IRs. We also observed that SARS-CoV-2 has fewer long IRs than SARS-CoV/BJ01 and SARSr-CoV/RmYN01.

To investigate the evolutionary relationship of the CoV genomes from the IRs distribution, we survey the IRs in these human-CoVs and relate these CoVs to SARS-CoV-2. The IR structures of the human-CoVs are illustrated in Fig. 4.

3.3. Evolution distances of SARS-CoV-2 and its related CoVs by IR points

IR distribution in the CoV genomes may give clues on the zoonotic evolution of SARS-CoV-2. The evolution proximities of the CoVs from the Hausdorff distances of the genomics IR points. To this end, we compare the differences of IRs in bat and human CoVs so that the patterns of zoonotic evolution can be revealed. Table 1 lists the Hausdorff distances between SARS-CoV-2 and bat and human-CoVs, and also the frequencies of IRs of at least 12 bp in each CoV genome. The IR numbers are counted by both the palindrome-first and palindrome-second sequences of the IRs. From the IR distances, we observe that SARS-CoV-2 strain is more closely related to SARSr-CoV/ZC21 and SARSr-CoV/RaTG13, although SARS-CoV-2 shares a higher sequence identify with SARSr-CoV/RaTG13 than SARSr-CoV/ZC21 (Zhou et al., 2020b). One possible reason for this discordance of IR distribution is due to the high level of synonymous mutations in SARr-CoV/RaTG13 (Li et al., 2020). In addition, a recent study shows that the region ORF1a of SARSr-CoV/ZC21 is closer to SARS-CoV-2 (Li et al., 2020), suggesting recombinations occurred in the CoV genomes. Therefore, we may consider SARS-CoV-2, SARSr-CoV/RaTG13 and SARSr-CoV/ZC21 to be three very close strains. From Table 1, we show that Pangolin-CoV genome contains the highest frequency of IRs and is far from SARS-CoV-2. The result also shows that the human CoVs (human-CoV/OC43, human-CoV/229E, human-CoV/

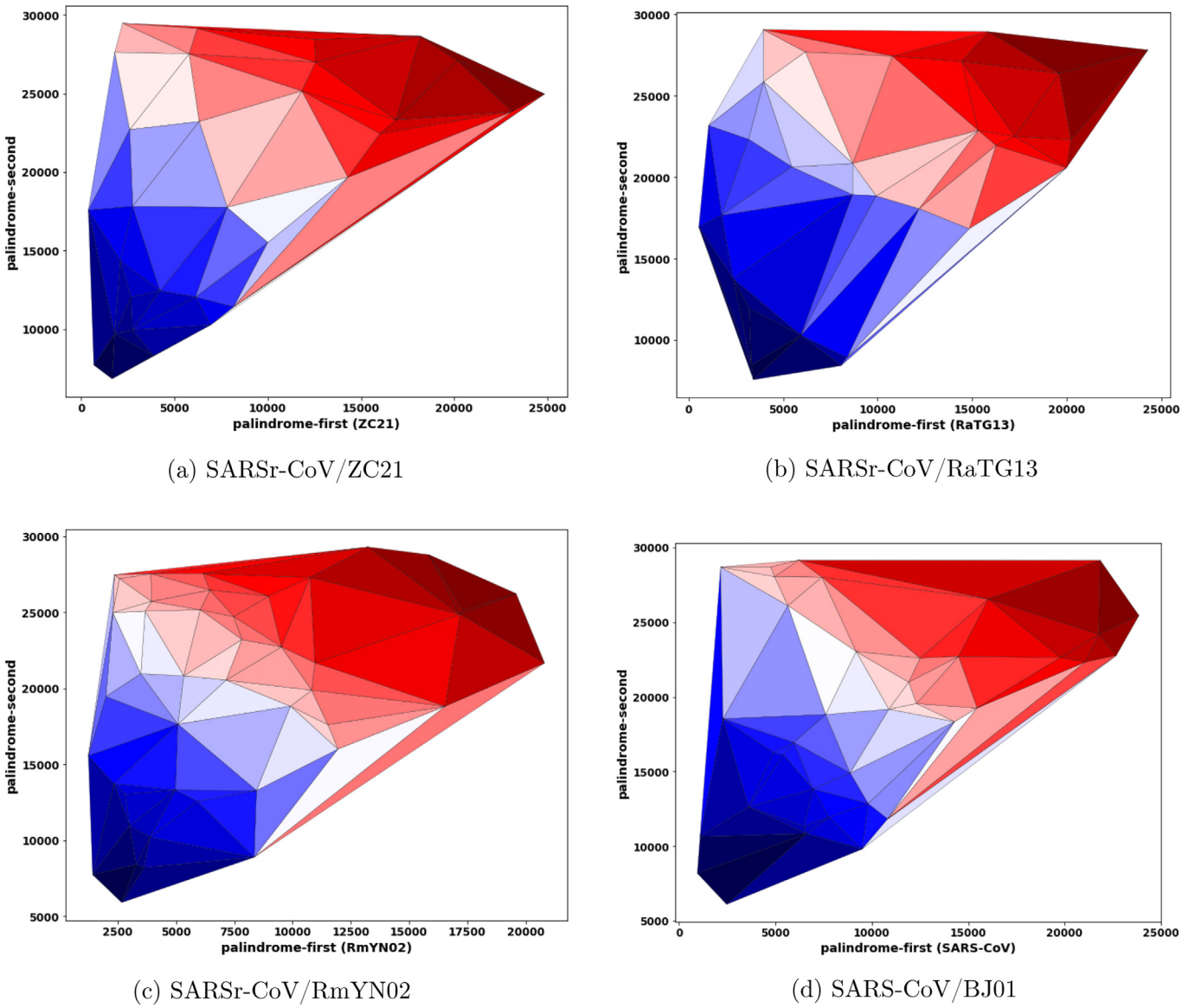


Fig. 2. Delaunay triangulation of the IR points in SARS-CoV-2 related bat-CoV genomes. (a) SARSr-CoV/RaTG13. (b) SARSr-CoV/ZC21. (c) SARSr-CoV/RmYN02. (d) SARS-CoV/BJ01. The darker blue colors indicate smaller coordinates for lower and left positions, while the darker red colors indicate larger coordinates to upper and right positions.

NL63, human-CoV/HKU1) have high frequencies of long IRs in the genome. The reason is probably due to the increasing long IRs during the human infection periods. The additional control example is that swine-SADS-CoV has the lowest IRs because swine-SADS-CoV was in its original native state.

The zoonotic evolution of a CoV means the period of its infecting humans or close related hosts from wild state. The wild state is when the virus is with the native bat. When a CoV is in its wild and native state, the IR number is small. When the CoV infects human or mammalian hosts, under the host immune response and natural selection, the virus may undergo mutations, therefore resulting in increased IRs for survival in the host. We speculate that the IRs in CoV genomes are increasing during human host-interaction evolution. This speculation is also supported by Goswami et al.'s study that SARS-CoV-2 hot-spot mutations are predominantly distributed within IR locis (Goswami et al., 2020).

The relative evolution proximities among the bat and human-CoVs are further inferred by the Hausdorff distances of the IR points in the CoV genomes (Fig. 5.). The results also exhibit that the closest relative of SARS-CoV-2 is SARSr-CoV/RaTG13, followed by bat-CoV/RpYN06. In addition, the results also indicate that SARS-CoV (2003) is closely related to civets-SCoV, supporting the theorem that Civets could be the intermediate and transmission host of SARS-CoV (Hu et al., 2015. We note that swine-SADS-CoV is an outlier and in the original native state of bat host. swine-SADS-CoV arrived directly from the native bat host before jumping to swine.

Note that Fig. 5 is the multidimensional scaling (MDS) projection of the tested CoVs, and the planer MDS layout is based on pair-wise distances of the CoVs, and takes account all the relationships of the CoVs. The MDS coordinates are for proximity illustration and can have some differences in some points. The actual distance between SARS-CoV-2 and MERS-CoV is 5174.9437, and the distance between SARS-CoV-2 and SARS-CoV/BJ01 is 4024.2620 (Table 1).

We observed that long IRs of at least 12 bp in CoV genomes increase over evolution time. For example, swine-CoV/SADs is very virulent and is considered in the raw native state, it contains the

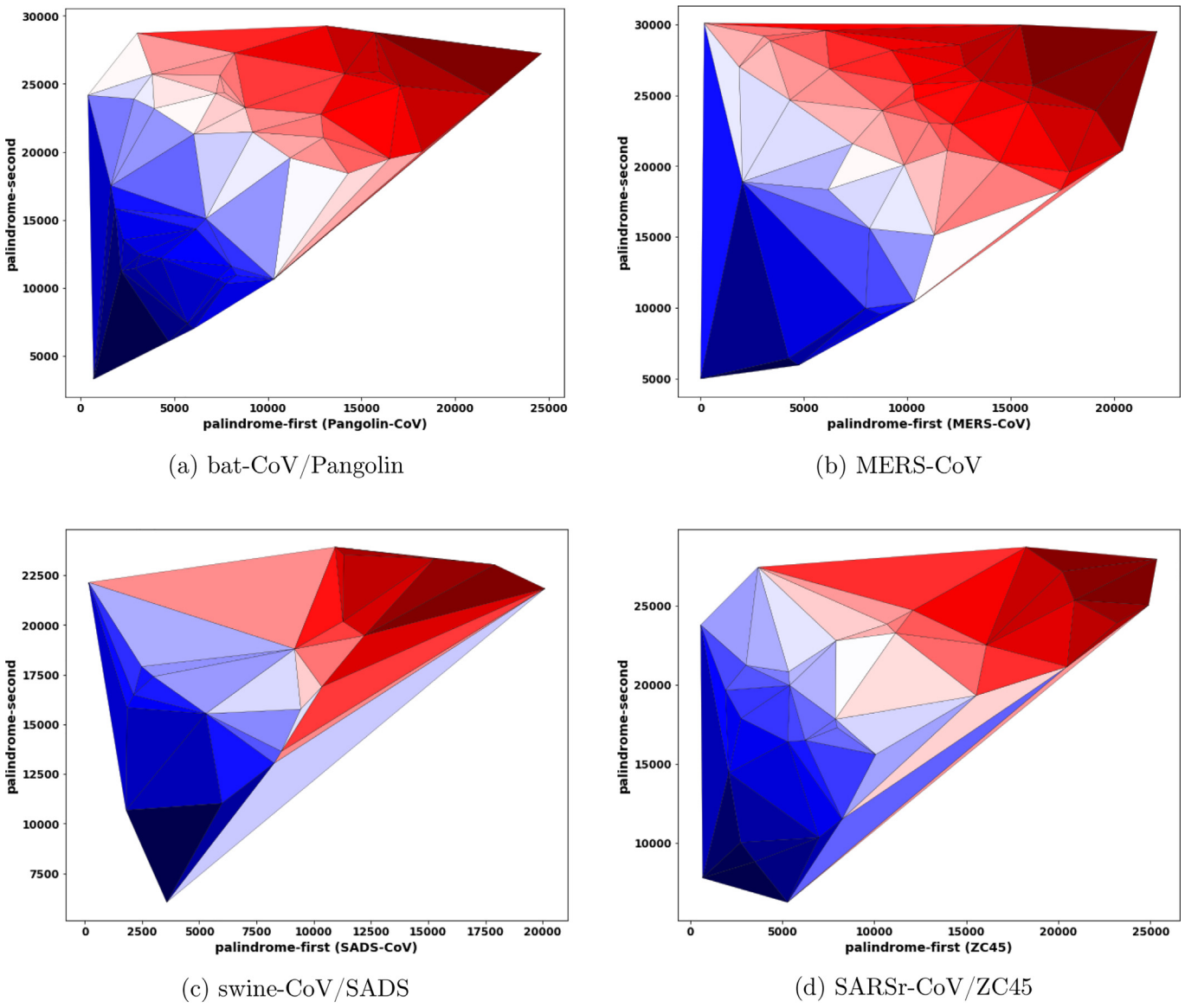


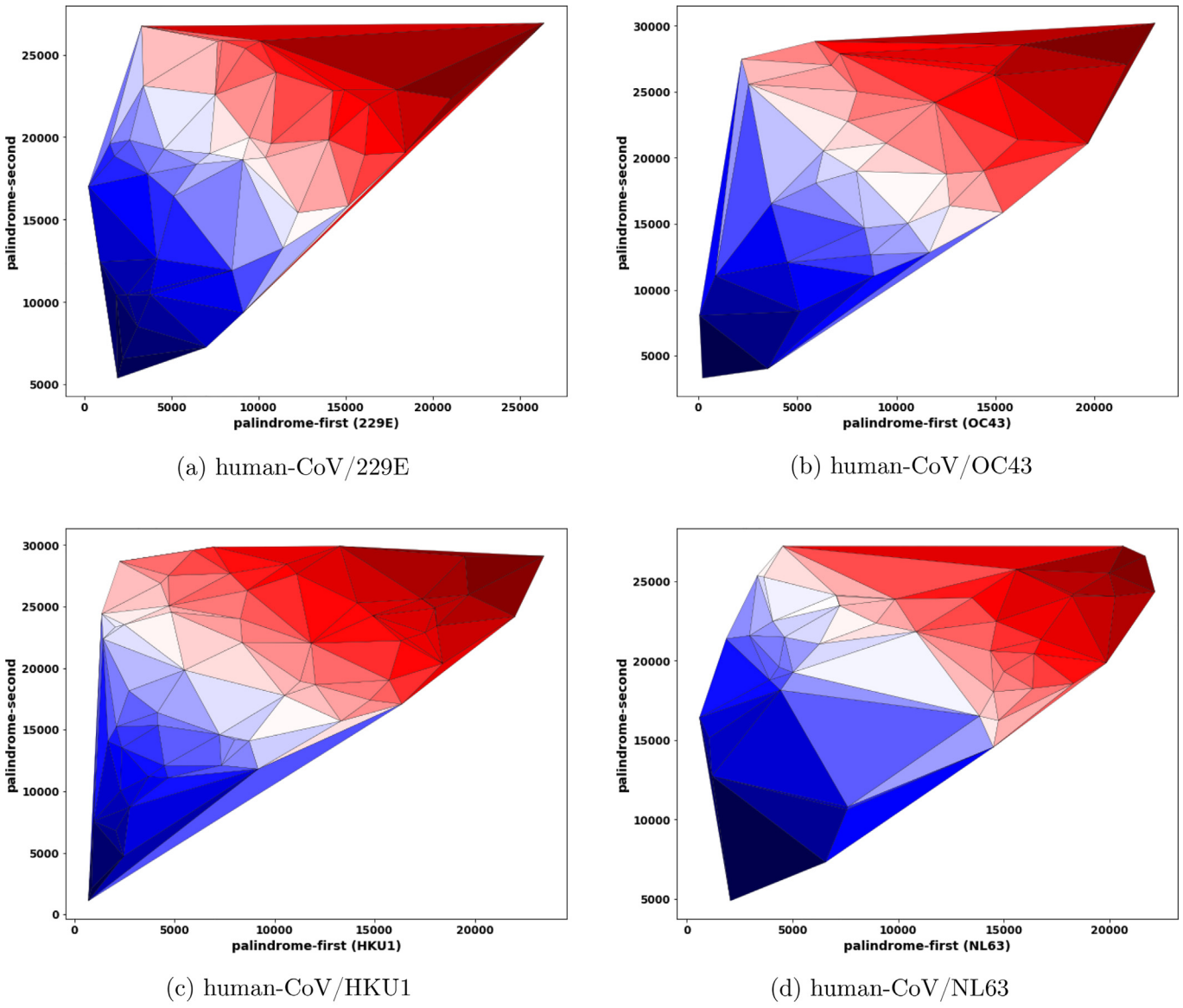
Fig. 3. Delaunay triangulation of the IR points in bat-CoV genomes. (a) bat-CoV/Pangolin. (b) MERS-CoV. (c) swine-CoV/SADS. (d) SARSr-CoV/ZC45. The darker blue colors indicate smaller coordinates for lower and left positions, while the darker red colors indicate larger coordinates to upper and right positions.

Table 1
The Hausdorff distances between SARS-CoV-2 and SARS-related CoVs.

Virus	GenBank/GISAID number	Distance	IR12+Frequency
SARS-CoV-2	NC_045512	0	35
SARSr-CoV/RaTG13	MN996532	3947.8937	30
SARSr-CoV/ZC21	MG772934	3569.9604	31
SARSr-CoV/ZC45	MG772933	4894.4083	32
pangolin-CoV/MP789	MT121216	5464.1787	43
bat-CoV/RmYN02	EPI_ISL_412976	4990.1383	44
bat-CoV/RpYN06	EPI_ISL_1699446	4048.291	32
bat-CoV/HKU2	NC_009988	4930.6329	26
SARS-CoV/BJ01	AY278488	4024.2620	38
civets-SCoV/SZ3	AY304486	4024.11	38
MERS-CoV	NC_019843	5174.9437	38
swine-SADS-CoV	MG557844	7914.0691	21
human-CoV/OC43	AY585229	5869.5071	36
human-CoV/229E	NC_002645	6163.2648	45
human-CoV/NL63	NC_005831	4403.8698	46
human-CoV/HKU1	AY597011	6889.9608	62

least number of IRs (Fig. 3(c)), whereas the human-CoVs, which have evolved in humans for long period, the IRs in human-CoVs are enriched (Fig. 4).

To investigate the dynamic characteristics of IRs in CoV genomes during evolution, we count the IRs of different lengths from bat CoV genomes identified in the last decades (Fig. 6). The result shows that SARS-CoV (2003) and swine-SAD-Cov (2017) genomes have the least amount of IRs, while SARS-CoV-2 (2019) and its related SARSr-CoV/RaTG13 (2013) and SARSr-CoV/RmYN02 (2019) have high amounts of IRs. From the repeat analysis, we postulate that during evolution and human infection, mutations and recombinations may occur and accumulate IRs in genomes under natural selection. The mutations and recombinations by the IR segments may be one of the driving forces for virus fast evolution. A global view of the genome profile using long IRs demonstrates that the genome structures are constant with variations. The IRs steadily increase during evolution, reflecting the cumulative abundance of these IR sequences due to possible recombination events. Therefore, the close relevance of human and bat CoVs can



(a) human-CoV/229E (b) human-CoV/OC43

(c) human-CoV/HKU1 (d) human-CoV/NL63

Fig. 4. Delaunay triangulation of the IR points in four human CoV genomes. (a) human-CoV/229E. (b) human-CoV/OC43. (c) human-CoV/HKU1. (d) human-CoV/NL63. The darker blue colors indicate smaller coordinates for lower and left positions, while the darker red colors indicate larger coordinates to upper and right positions.

be examined through the IRs bar-coding profiles. In addition, from these repeat analyses, we may infer that during evolution, the recombinations may occur and produce accumulative inverted repeats under natural selection.

4. Discussions

In this study, we present a novel geometric method to represent the genome architecture using the IR contents in the genome. The method maps the IR points into a geometric shape so that the IR distribution can be visualized as a graph, and genomes can be compared using the Hausdorff distance of these graph shapes. Therefore, a linear genome sequence can be represented as a graph according to the IRs, which are k-mer sequences. The graph representation uses both the critical genomic IR structures and k-mer sequences. The graph representation contains the localization, frequency, and distribution of IRs in a genome. The genomic graphs formed by the IRs in this method can be further characterized using graph and topological analytics and algorithms.

Previous numerous researches have suggested that SARS-CoV mutations during human infection are generated in most cases by immune response through RNA editing. The implicit relationship between RNA editing and inverted repeat (IR) is that the RNA editing events often happen on the dsRNA regions, which are formed by IRs.

This study investigates the contents and distribution of IRs in SARS-CoV-2 and other CoV genomes. The correlation between IRs and CoV evolution is positive, indicating IRs may be essential for zoonotic evolution, genome structures, and functions. Therefore, we may consider IRs for the therapeutic potential of SARS-CoV-2. This study shows that SARS-CoV-2 genome has increasing numbers of IRs during evolution. However, the origin and functions of the accumulated IRs in SARS-CoV-2 genome remain unknown. As ADAR RNA editing targets dsRNA, this suggests that dsRNA encompasses the entire SARS-CoV-2 genome (Giorgio et al., 2020). While dsRNA in human transcripts is often driven by IRs, the most likely source of dsRNA in the viral transcripts is replication, where both positive and negative strands are present and can result in wide regions of dsRNA. IRs delimit the boundaries in transposons in gen-

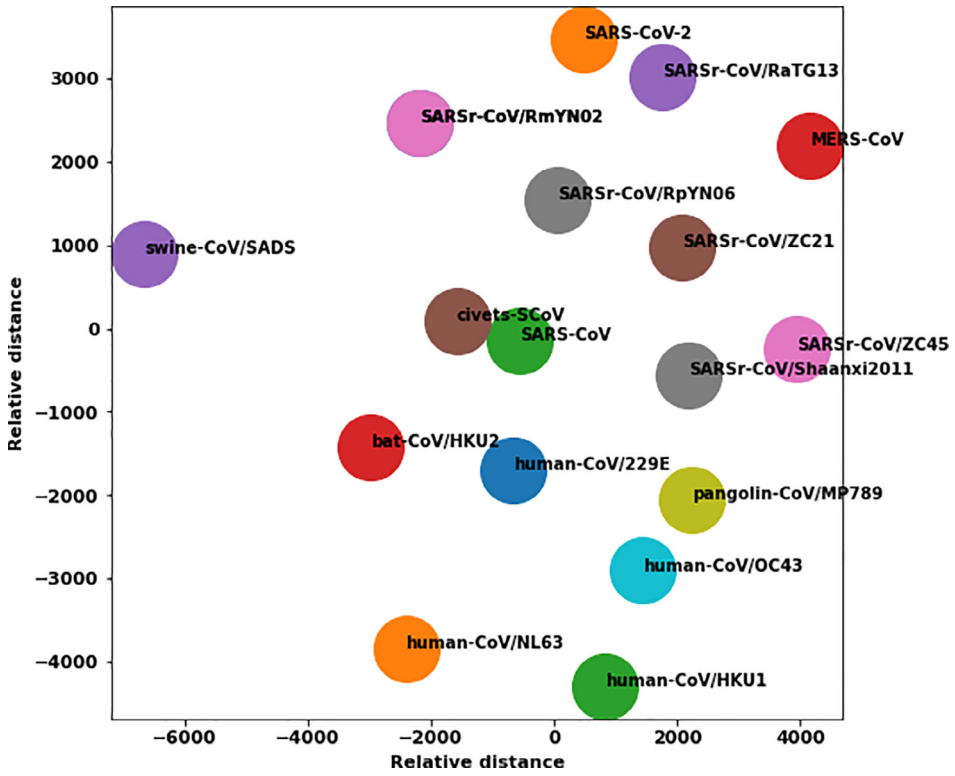


Fig. 5. Proximities of bat and human-CoVs measured by the Hausdorff distances of IR points in the CoV genomes. The proximities are inferred by MDS projection of the pairwise Hausdorff distances of the CoV genomes.

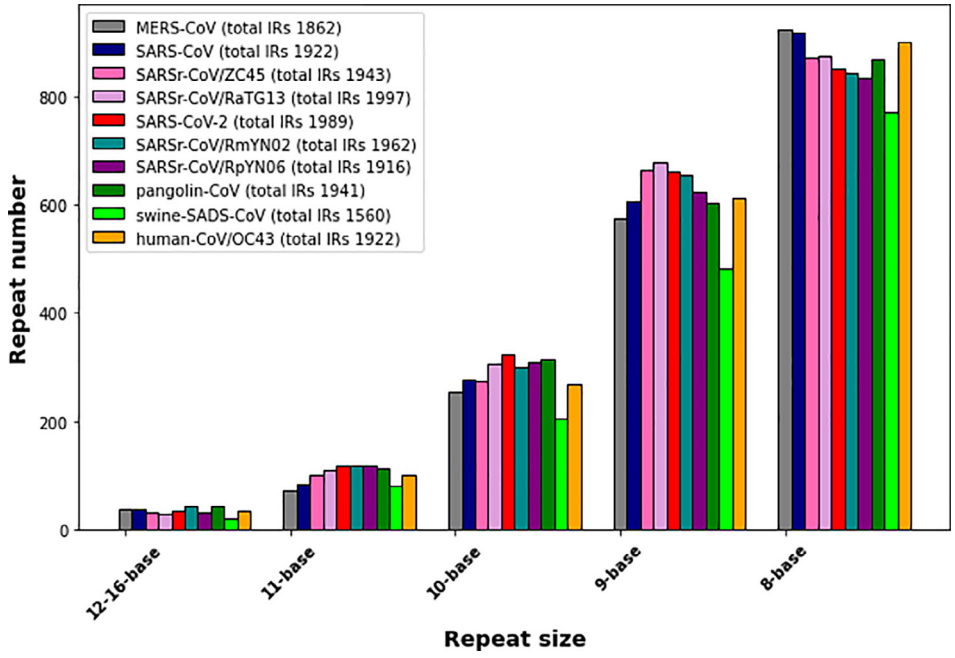


Fig. 6. Frequencies of IRs of different lengths in the bat and human-CoV genomes. The repeat numbers are counted by palindrome-first sequences only. The total numbers of all IRs of each genomes are in the legends.

ome evolution and form stem-loop structures that retain genome stability and flexibility. Additionally, as we see increasing perfect IRs during evolution, our results support a model in which imperfect IRs are corrected to perfect IRs in a preferential manner(Lavi et al., 2018).

Importantly, previous studies have suggested that SARS-CoV-2 mutations are generated in most cases by human immune response through RNA editing (Giorgio et al., 2020). Two human defense mechanisms are through the apolipoprotein B mRNA editing catalytic polypeptide-like proteins (APOBEC) and adenosine

deaminase acting on RNA (ADAR). ADAR acts as Adenosine deamination (A-tol) in double-stranded RNA (dsRNA). A recent study showed that the IRs are the hot spots of mutations in SARS-CoV-2 genomes. Here, we propose the connection between high mutations in IRs and the ADAR editing system, especially, the IR regions are dsRNA and the targets of ADAR editing. As a result of ADAR RNA editing during infection, if mutations are in the middle of a short IR, then the number of short IRs is decreased; if mutations are at the ends of short IRs, the number of short IR is also decreased, but the numbers of long IRs are increased. This analysis is in the agreement of the distributions of bat and human-CoVs (Fig. 6). Therefore, the ADAR RNA editing in virus evolution may account for high-frequency mutations in IRs of SARS-CoV-2 genome.

This study only considers the perfect IRs in genomes. If considering matching pairs in the IRs, we can expect that much longer inverted repeats can be identified, and the number of IRs in the virus genome will increase significantly with evolution. The imperfect IRs are the natural forms of the repeats to maintain the genome structures. Because the perfect IR distribution and types in a genome are unique, and extracting the perfect IRs is parameter-free, the perfect IRs can be considered as a unique genomic signature. The signatures from perfect IRs are consistent, and therefore can be used in distinguishing the closely related viruses and differing virus mutation variants. The quantitative comparison of the signature can also provide phylogenetic taxonomy when appropriate numerical metrics for the signatures are realized. Therefore, the perfect IRs can be an effective barcode to distinguish genotypes.

IRs rendered stems and cruciforms play fundamental roles in genomic stability, replication, gene expression, nucleosome structure, and recombination (Brázda et al., 2011; Gallaher, 2020). IRs also contribute the pathogenicity of MERS-CoV (Xie et al., 2017), EBV virus (Bridges et al., 2019), and Marek's disease virus (Vychodil et al., 2021). This study demonstrates the increasing IRs in CoVs during zoonotic evolution. Therefore, the IR changes in a coronavirus during epidemiological time may probably indicate the pathogenicity of the virus. The IRs in coronavirus genomes can be potential therapeutic and pharmaceutical targets. These topics need future investigations.

5. Abbreviations

- COVID-19: coronavirus disease 2019
- CRISPR: clusters of regularly interspaced short palindromic repeats
- dsRNA: double strand RNA
- MERS-CoV: Middle East Respiratory Syndrome coronavirus
- NCBI: National Center for Biotechnology Information (USA)
- IR(s): inverted repeat(s)
- SARS: severe acute respiratory syndrome
- SARS-CoV-2: severe acute respiratory syndrome coronavirus 2

CRediT authorship contribution statement

Changchuan Yin: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Stephen S.-T. Yau:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We sincerely appreciate the researchers who sequenced and shared the complete genome sequences of SARS-CoV-2 and bat-CoV/RmYN02 coronaviruses from GISAID (https://www.gisaid.org/). This research is partially supported by the National Natural Science Foundation of China (NSFC) grant (91746119, to S.S.-T. Yau) and Tsinghua University Spring Breeze Fund (2020Z99CFY044, to S.S.-T. Yau).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.jtbi.2021.110885.

References

Adelman, Z.N., Sanchez-Vargas, I., Travanty, E.A., Carlson, J.O., Beaty, B.J., Blair, C.D., Olson, K.E., 2002. RNA silencing of dengue virus type 2 replication in transformed C6/36 mosquito cells transcribing an inverted-repeat RNA derived from the virus genome. Journal of Virology 76 (24), 12925–12933.

Bartas, M., Brázda, V., Bohálová, N., Cantara, A., Volná, A., Stachurová, T., Malachová, K., Jagelská, E.B., Porubiaková, O., Červeň, J., et al., 2020. In-depth bioinformatic analyses of nidovirales including human SARS-CoV-2, SARS-CoV, MERS-CoV viruses suggest important roles of non-canonical nucleic acid structures in their lifecycles. Frontiers in Microbiology 11, 1583.

Brázda, V., Laister, R.C., Jagelská, E.B., Arrowsmith, C., 2011. Cruciform structures are a common DNA feature important for regulating biological processes. BMC Molecular Biology 12 (1), 1–16.

Brázda, V., Kolomazník, J., Lỳsek, J., Hároníková, L., Coufal, J., Šťastnỳ, J., 2016. Palindrome analyser–A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. Biochemical and Biophysical Research Communications 478 (4), 1739–1745.

Bridges, R., Correia, S., Wegner, F., Venturini, C., Palser, A., White, R.E., Kellam, P., Breuer, J., Farrell, P.J., 2019. Essential role of inverted repeat in Epstein-Barr virus IR-1 in B cell transformation; geographical variation of the viral genome. Philosophical Transactions of the Royal Society B 374 (1773), 20180299.

Buchon, N., Vaury, C., 2006. RNAi: a defensive RNA-silencing against viruses and transposable elements. Heredity 96 (2), 195–202.

Čechová, J., Lỳsek, J., Bartas, M., Brázda, V., 2018. Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability. Bioinformatics 34 (7), 1081–1085.

Cui, J., Li, F., Shi, Z.-L., 2019. Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology 17 (3), 181–192.

Di Giorgio, S., Martignano, F., Torcia, M.G., Mattiuz, G., Conticello, S.G., 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Science Advances 6 (25), eabb5813.

Edwards, C.E., Yount, B.L., Graham, R.L., Leist, S.R., Hou, Y.J., Dinnon, K.H., Sims, A.C., Swanstrom, J., Gully, K., Scobey, T.D., et al., 2020. Swine acute diarrhea syndrome coronavirus replication in primary human cells reveals potential susceptibility to infection. Proceedings of the National Academy of Sciences 117 (43), 26915–26925.

Fan, Y., Zhao, K., Shi, Z.-L., Zhou, P., 2019. Bat coronaviruses in china. Viruses 11 (3), 210.

Gallaher, W.R., 2020. A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. Archives of Virology 165 (10), 2341–2348.

Goswami, P., Bartas, M., Lexa, M., Bohálová, N., Volná, A., Červeň, J., Červeňová, V., Pečinka, P., Špunda, V., Fojta, M., et al., 2020. SARS-CoV-2 hot-spot mutations are significantly enriched within inverted repeats and CpG island loci. Briefings in Bioinformatics.

Hu, B., Ge, X., Wang, L.-F., Shi, Z., 2015. Bat origin of human coronaviruses. Virology Journal 12 (1), 1–10.

Hu, D., Zhu, C., Ai, L., He, T., Wang, Y., Ye, F., Yang, L., Ding, C., Zhu, X., Lv, R., et al., 2018. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerging Microbes & Infections 7 (1), 1–10.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (9), 850–863.

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N., Chang, H., 2020. The architecture of SARS-CoV-2 transcriptome. Cell 181 (4), 914–921.

Lam, T.T.-Y., Jia, N., Zhang, Y.-W., Shum, M.H.-H., Jiang, J.-F., Zhu, H.-C., Tong, Y.-G., Shi, Y.-X., Ni, X.-B., Liao, Y.-S., et al., 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 583 (7815), 282–285.

Lau, S.K., Woo, P.C., Li, K.S., Huang, Y., Wang, M., Lam, C.S., Xu, H., Guo, R., Chan, K.-H., Zheng, B.-J., et al., 2007. Complete genome sequence of bat coronavirus HKU2 from chinese horseshoe bats revealed a much smaller spike gene with a different evolutionary lineage from the rest of the genome. Virology 367 (2), 428–439.

Lavi, B., Levy Karin, E., Pupko, T., Hazkani-Covo, E., 2018. The prevalence and evolutionary conservation of inverted repeats in proteobacteria. Genome Biology and Evolution 10 (3), 918–927.
Li, X., Giorgi, E.E., Marichannegowda, M.H., Foley, B., Xiao, C., Kong, X.-P., Chen, Y., Gnanakaran, S., Korber, B., Gao, F., 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Science Advances 6 (27), eabb9153.
Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X., Jiang, W., 2020. The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive rna modification. Future Virology 15 (6), 341–347.
Limanskaya, O.Y., 2009. Bioinformatic analysis of inverted repeats of coronaviruses genome. Biopolymers & Cell 25 (4), 307.
Liu, P., Jiang, J.-Z., Wan, X.-F., Hua, Y., Li, L., Zhou, J., Wang, X., Hou, F., Chen, J., Zou, J., et al., 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathogens 16, (5) e1008421.
Max Roser, Hannah Ritchie, E.O.-O., Hasell, J., 2020. Coronavirus Pandemic (COVID-19). Our World in Data. Https://ourworldindata.org/coronavirus..
Mojica, F.J., García-Martínez, J., Soria, E., et al., 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. Journal of Molecular Evolution 60 (2), 174–182.
Muskens, M.W., Vissers, A.P., Mol, J.N., Kooter, J.M., 2000. Role of inverted DNA repeats in transcriptional and post-transcriptional gene silencing. Plant Gene Silencing, 123–140.
Pearson, C.E., Zorbas, H., Price, G.B., Zannis-Hadjopoulos, M., 1996. Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. Journal of Cellular Biochemistry 63 (1), 1–22.
Selker, E.U., 1999. Gene silencing: repeats that count. Cell 97 (2), 157–160.
Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22 (13).
St-Jean, J.R., Jacomy, H., Desforges, M., Vabret, A., Freymuth, F., Talbot, P.J., 2004. Human respiratory coronavirus OC43: genetic stability and neuroinvasion. Journal of Virology 78 (16), 8824–8834.
Van Der Hoek, L., Pyrc, K., Jebbink, M.F., Vermeulen-Oost, W., Berkhout, R.J., Wolthers, K.C., Wertheim-van Dillen, P.M., Kaandorp, J., Spaargaren, J., Berkhout, B., 2004. Identification of a new human coronavirus. Nature Medicine 10 (4), 368–373.
Voineagu, I., Narayanan, V., Lobachev, K.S., Mirkin, S.M., 2008. Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. Proceedings of the National Academy of Sciences 105 (29), 9936–9941.
Vychodil, T., Conradie, A.M., Trimpert, J., Aswad, A., Bertzbach, L.D., Kaufer, B.B., 2021. MarekâĂŹs disease virus requires both copies of the inverted repeat regions for efficient in vivo replication and pathogenesis. Journal of Virology 95 (3), e01256–20.
Woo, P.C., Lau, S.K., Chu, C.-M., Chan, K.-H., Tsoi, H.-W., Huang, Y., Wong, B.H., Poon, R.W., Cai, J.J., Luk, W.-K., et al., 2005. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. Journal of Virology 79 (2), 884–895.
Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579 (7798), 265–269.
Xie, Q., Cao, Y., Su, J., Wu, J., Wu, X., Wan, C., He, M., Ke, C., Zhang, B., Zhao, W., 2017. Two deletion variants of Middle east respiratory syndrome coronavirus found in a patient with characteristic symptoms. Archives of Virology 162 (8), 2445–2449.
Yin, C., Yau, S.S.-T., 2021. Inverted repeats in coronavirus SARS-CoV-2 genome and implications in evolution. Communications in Information and Systems.
Zaki, A.M., Van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D., Fouchier, R.A., 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. New England Journal of Medicine 367 (19), 1814–1820.
Zhou, P., Fan, H., Lan, T., Yang, X.-L., Shi, W.-F., Zhang, W., Zhu, Y., Zhang, Y.-W., Xie, Q.-M., Mani, S., et al., 2018. Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. Nature 556 (7700), 255–258.
Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E.C., et al., 2020a. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. Current Biology.
Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al., 2020b. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798), 270–273.
Zhou, H., Ji, J., Chen, X., Bi, Y., Li, J., Hu, T., Song, H., Chen, Y., Cui, M., Zhang, Y., et al., 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. bioRxiv..
Zou, X., Morganella, S., Glodzik, D., Davies, H., Li, Y., Stratton, M.R., Nik-Zainal, S., 2017. Short inverted repeats contribute to localized mutability in human somatic cells. Nucleic Acids Research 45 (19), 11213–11221.