Communications in Information and Systems
Volume 21, Number 1, 125–145, 2021

Inverted repeats in coronavirus SARS-CoV-2 genome and implications in evolution

Changchuan Yin* and Stephen S.-T. Yau†

The coronavirus disease (COVID-19) pandemic, caused by the coronavirus SARS-CoV-2, has caused 60 million infections and 1.38 million fatalities. Genomic analysis of SARS-CoV-2 can provide insights on drug design and vaccine development for controlling the pandemic. Inverted repeats in a genome greatly impact the stability of the genome structure and regulate gene expression. Inverted repeats involve cellular evolution and genetic diversity, genome arrangements, and diseases. Here, we investigate the inverted repeats in the coronavirus SARS-CoV-2 genome. We find that SARS-CoV-2 genome has an abundance of inverted repeats. The inverted repeats are mainly located in the gene of the Spike protein. This result suggests the Spike protein gene undergoes recombination events, therefore, is essential for fast evolution. Comparison of the inverted repeat signatures in human and bat coronaviruses suggest that SARS-CoV-2 is mostly related SARS-related coronavirus, SARSr-CoV/RaTG13. The study also reveals that the recent SARS-related coronavirus, SARSr-CoV/RmYN02, has a high amount of inverted repeats in the spike protein gene. Besides, this study demonstrates that the inverted repeat distribution in a genome can be considered as the genomic signature. This study highlights the significance of inverted repeats in the evolution of SARS-CoV-2 and presents the inverted repeats as the genomic signature in genome analysis.

1. Introduction

1.1. Human coronaviruses

The novel human coronavirus SARS-CoV-2 (formerly, 2019-nCoV) first emerged in Wuhan, China, in December 2019, the causative agent for Coronavirus Disease-2019 (COVID-19) pandemic, has claimed 1.38 million lives

*ORCID: 0000-0002-4147-4195.
†ORCID: 0000-0001-7634-7981.

across the globe as of Nov. 24, 2020 [23]. Understanding the molecular structure and evolution of SARS-CoV-2 genome is of urgency for tracing the origin of the virus and provides insights on vaccine development and drug design for controlling the current COVID-19 pandemic.

Human coronaviruses (CoVs) are common viral respiratory pathogens that cause mild to moderate upper-respiratory tract illnesses. Three common CoVs, 229E, and OC43 identified in 1965, NL63 in 2004 can cause the common cold. Four typical human CoVs found in recent years are Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) in 2002, HKU1 in 2005, and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in 2012. Among these human CoVs, SARS-CoV and MERS-CoV are highly pathogenic and caused severe and fatal infections. MERS symptoms are very severe, usually including fever, cough, and shortness of breath which often progress to pneumonia. SARS symptoms often include fever, chills, and body aches which usually progressed to pneumonia. The current coronavirus SARS-CoV-2, which causes a worldwide COVID-19 pandemic, is milder than SARS-CoV, but can cause severe syndromes and fatality in people with cardiopulmonary disease, people with weakened immune systems, infants, and older adults.

SARS-CoV-2 is a beta coronavirus, like MERS-CoV and SARS-CoV. All three of these coronaviruses have their origins in bats. Yet the zoonotic origin of SARS-CoV-2 is still unconfirmed. [53, 52]'s study showed that the bat SARS-related coronavirus strain SARSr-CoV/RaTG13, identified from a bat *Rhinolophus affinis* in Yunnan province, China, in July 2012, shares 96.2% nucleotide identity. A recent study identified a new SARSr-CoV/RmYN02 (2019) from *Rhinolophus malayanus*, which is closely related to SARS-CoV-2 [51]. SARSr-CoV/RmYN02 shares 93.3% nucleotide identity with SARS-CoV-2 and comprises natural insertions at the S1/S2 cleavage site of the Spike protein. The unique S1/S2 cleavage in the Spike protein in SARS-CoV-2 may confer the zoonotic spread of SARS-CoV-2. However, the originating relationship among these CoVs is not entirely clear.

1.2. Coding structures of SARS-CoV-2 genome

SARS-CoV-2 coronavirus contains a linear single-stranded positive RNA genome (Fig. 1). The SARS-CoV-2 RNA genome of 29.9kb has a total of 11 genes with 11 open reading frames (ORFs) [49], consisting of the leader sequence (5'UTR), the coding regions, and 3'UTR pseudoknot stem-loop [44]. The coding regions include ORF1ab and genes encoding 16 non-structural

proteins [10] and structural proteins (spike (S), envelope (E), membrane (M), and nucleocapsid (N)) [12], and several accessory proteins.

ORF1ab encodes replicase polyproteins required for viral RNA replication and transcription [7, 5]. Nonstructural protein 1 (nsp1) likely inhibits host translation by interacting with 40S ribosomal subunit, leading to host mRNA degradation through cleavage near their 5'UTRs. Nsp1 promotes viral gene expression and immunoevasion in part by interfering with interferon-mediated signaling. Nonstructural protein 2 (nsp2) interacts with host factors prohibitin 1 and prohibitin 2, which are involved in many cellular processes including mitochondrial biogenesis. The third non-structural protein (nsp3) is Papain-like proteinase. Nsp3 is an essential and the largest component of the replication and transcription complex. The Papain-like proteinase cleaves non-structural proteins 1-3 and blocks the host's innate immune response, promoting cytokine expression [33, 18]. Nsp4 encoded in ORF1ab is responsible for forming double-membrane vesicle (DMV). The other non-structural proteins are 3CLPro protease (3-chymotrypsin-like proteinase, 3CLpro) and nsp6. 3CLPro protease is essential for RNA replication. The 3CLPro proteinase accounts for processing the C-terminus of nsp4 through nsp16 in coronaviruses [1]. Together, nsp3, nsp4, and nsp6 can induce DMV [2].

SARS-coronavirus has a unique RNA replication facility, including two RNA-dependent RNA polymerases (RNA pol). The first RNA polymerase is a primer-dependent non-structural protein 12 (nsp12), and the second RNA polymerase is nsp8, nsp8 has the primase capacity for *de novo* replication initiation without primers [37]. Nsp7 and nsp8 are essential proteins in the replication and transcription of SARS-CoV-2. Nsp7 is responsible for nuclear transport. The SARS-coronavirus nsp7-nsp8 complex is a multimeric RNA polymerase for both *de novo* initiation and primer extension [30, 37]. Nsp8 also interacts with ORF6 accessory protein. The nsp9 replicase protein of SARS-coronavirus binds RNA and interacts with nsp8 for its functions [36]. Helicase (nsp13) possesses helicase activity, thus catalyzing the unwinding dsRNA or structured RNA into single strands. Importantly, nsp14 may function as a proofreading exoribonuclease for virus replication, hence, SARS-CoV-2 mutation rate remains low.

Furthermore, the SARS-CoV-2 genome encodes several structural proteins. The structural proteins possess much higher immunogenicity for T cell responses than the non-structural proteins [19]. The structural proteins include spike (S), envelope (E), membrane protein (M), and nucleoprotein (N) [22, 31]. The Spike glycoprotein has two domains S1 and S2. Spike protein S1 attaches the virion to the host cell membrane through the receptor ACE2,

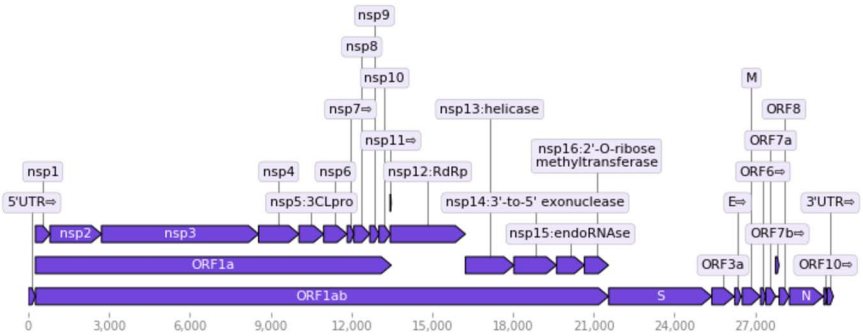


Figure 1: The structural diagram of SARS-CoV-2 genome (GenBank: NC_045512). The diagram of SARS-CoV-2 genome was made using DNA Feature Viewer [54].

initiating the infection [42, 43]. After being internalized into the endosomes of the cells, the S glycoprotein is then cleaved by cathepsin CTSL. The spike protein domain S2 mediates fusion of the virion and cellular membranes by acting as a class I viral fusion protein. Especially, the spike glycoprotein of coronavirus SARS-CoV-2 contains a furin-like cleavage site [9]. A recent study indicates that SARS-CoV-2 is more infectious than SARS-CoV according to the changes of S protein-ACE2 binding affinity [6]. The envelope (E) protein interacts with membrane protein M in the budding compartment of the host cell. The M protein holds dominant cellular immunogenicity [21]. Nucleoprotein (ORF9a) packages the positive-strand viral RNA genome into a helical ribonucleocapsid (RNP) during virion assembly through its interactions with the viral genome and a membrane protein M [13]. Nucleoprotein plays an important role in enhancing the efficiency of subgenomic viral RNA transcription and viral replication.

1.3. Non-coding structures of the SARS-CoV-2 genome

In addition to the coding regions, SARS-CoV-2 genome contains hidden structures that can retain genome stability, regulate gene replication and expression, and control virus life cycles. The non-coding genome structures include leader sequences, transcriptional regulatory sequences (TRS), G-quadruplex structures, frame-shifting regions, and repeats. The first non-coding structure is the 5' leader sequence of about 265 bp is the unique characteristic in coronavirus replication and plays critical roles in the gene

expression of coronavirus during its discontinuous sub-genomic replication [20].

SARS-CoV-2 contains G-quadruplex structures [16]. It is well established that sequences with G-blocks (adjacent runs of Guanines) can potentially form non-canonical G-quadruplex (G4) structures [8, 24]. The G4 structures are formed by stacking two or more G-tetrads by Hoogsteen hydrogen bonds and often are the sites of genomic instability, serving one or more biological functions [3].

An inverted repeat is a single-stranded sequence of nucleotides followed by downstream its reverse complement downstream. The intervening sequence between the initial sequence and the reverse complement is called a spacer. When the spacer sequence is zero, the inverted repeat is called a palindrome. For example, the inverted repeat, 5'-ATTCGCGAAT-3' is a palindrome, the palindrome-first sequence is 5'-ATTCG-3', and the palindrome-second sequence is 5'-CGAAT-3'. When the spacer in an inverted repeat is non-zero, the repeat is generally inverted. In a generally inverted repeat, we still denote the initial sequence as a palindrome-first sequence and the downstream reverse complement as a palindrome-second sequence. For example, in the general inverted repeat, 5'-TTTAGGT...ACCTAAA-3', the palindrome-first sequence is 5'-TTTAGGT-3', and the palindrome-second sequence is 5'-ACCTAAA-3'. Through self-complementary base pairing, an inverted repeat can form a stem-loop (hairpin) structure in an RNA molecule, where the palindrome-first and palindrome-second sequences make a stem, and the spacer sequence makes a loop. It should be noted that an inverted repeat may not have perfect complementary base pairing in palindrome-first and palindrome-second sequences, so the stem formed by an imperfect inverted repeat can have mismatches, insert, or deletions. Inverted repetitive sequences are principal components of the archaeal and bacterial CRISPR-CAS systems [25], which function as adaptive antiviral defense systems.

Inverted repeats have important biological functions in viruses. Inverted repeats delimit the boundaries in transposons in genome evolution and form stem-loop structures in retaining genome instability and flexibility. Inverted repeats are described as hotspots of eukaryotic and prokaryotic genomic instability [40], replication [29], and gene silencing [32]. Therefore, inverted repeats involve cellular evolution and genetic diversity, mutations, and diseases.

Despite the paramount roles of the non-coding structures, the non-coding structures are not immediately visible as the coding regions. This

study is to identify one of the crucial non-coding structures, inverted repeats in SARS-CoV-2 genome, and investigate the cohort of the inverted repeats and the virus evolution.

2. Materials and methods

2.1. Identification of inverted repeats

The complete genomes of coronaviruses were scanned for inverted repeats using Palindrome analyzer [4]. Palindrome analyzer (http://bioinformatics.ibp.cz/) is a web-based server for retrieving palindromic and inverted repeats in DNA or RNA sequences. Palindrome server describes the features of inverted repeats including similarity analysis, localization, and visualization.

2.2. Inverted repeat analysis

To ensure consistency in comparing coronavirus genomes, we only extracted the inverted repeats with the perfect complementary base pairing of the palindrome-first and palindrome-second sequences. Noted that a short inverted repeat of length P can be inside a long inverted repeat of length Q ($Q > P$), in this case, we only extracted the inverted repeats of length Q and excluded the inverted repeat of length P.

The retrieved inverted repeats were mapped on the protein genes in a genome according to the positions of the palindrome-first and palindrome-second sequences of the inverted repeats.

The distributions of inverted repeats on protein genes in the different genomes are assessed by the Wasserstein distance, known as the earth mover's distance. The Wasserstein distance corresponds to the minimum amount of work required to transform one distribution into the other. The $p-th$ Wasserstein distance between two probability distributions μ and ν is defined as follows [39],

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x-y|\, d\pi(x,y) \right)^{1/p},$$

where $\Gamma(\mu,\nu)$ denotes the set of probability distributions on $\mathbb{R}\times\mathbb{R}$ with marginals μ and ν.

2.3. Genome data

The following complete genomes of SARS-CoVs and SARS-related coronaviruses (SARSr-CoVs) were downloaded from NCBI GenBank: SARS-CoV-2 (GenBank: NC_045512.2) [44], SARS-COV/Tor2 (GenBank: NC_004718), SARSr-CoV/RaTG13 (GenBank: MN996532) [53], SARSr-CoV/RmYN02 (GISAID: EPI_ISL_412977) [51, 34], and MERS-CoV (GenBank: NC_019843) [50].

3. Results

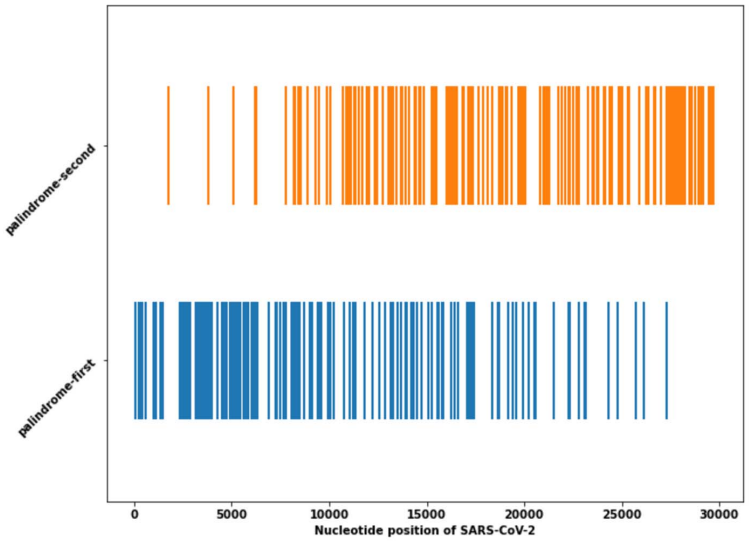
3.1. Inverted repeats in SARS-CoV-2 genome

Long inverted repeats are deemed to greatly influence the stability of the genomes of various organisms. The longest inverted repeats 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). The repeats of 11-15 bp are predominantly located in the gene of the Spike (S) protein (Fig. 2(a) and (b)). The other three protein genes (nsp3, RdRp, and N protein) are also enriched with long inverted repeats.

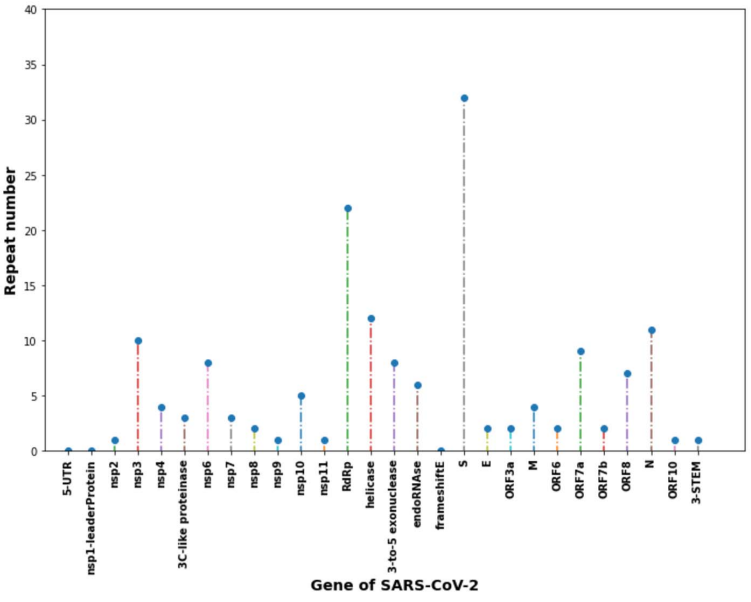
Long inverted repeats often contribute to the stability of a genome because of stable stems formed by the long inverted repeats. The results also suggest the recombinations took place at the gene of the Spike protein during evolution. Together, four protein genes (S, nsp3, RdRp, and N protein) of abundant inverted repeats are evolving dramatically and are critical for virus survival, therefore, can be the pharmaceutical targets [11].

The relation of virus genomes may provide insights on the zoonotic origin and evolution of the viruses. To examine the close relevance of human and bat CoVs, we evaluate and compare the distributions of inverted repeats of 11-15 bp in four CoV genomes: SARS-CoV-2 (Fig. 2(a)), SARS-CoV (Fig. 3(a)), MERS-CoV (Fig. 4(a)) SARSr-CoV/RaTG13 (Fig. 5(a)), and SARSr-CoV/RmYN02 (Fig. 6(a)). The repeat numbers of the inverted repeats of 11-15 bp on each protein gene in the genomes are shown in Fig. 2(b), Fig. 3(b), Fig. 4(b), Fig. 5(b), and Fig. 6(b). The repeat numbers are counted by both the palindrome-first and palindrome-second sequences of the inverted repeats.

It is noted that the long inverted repeats of 11-15 bp in nsp3 protein gene are unbalanced in different CoVs. SARS-CoV-2 nsp3 (Fig. 2(b)) gene contains much lower number of inverted repeats than highly pathogenic CoVs,

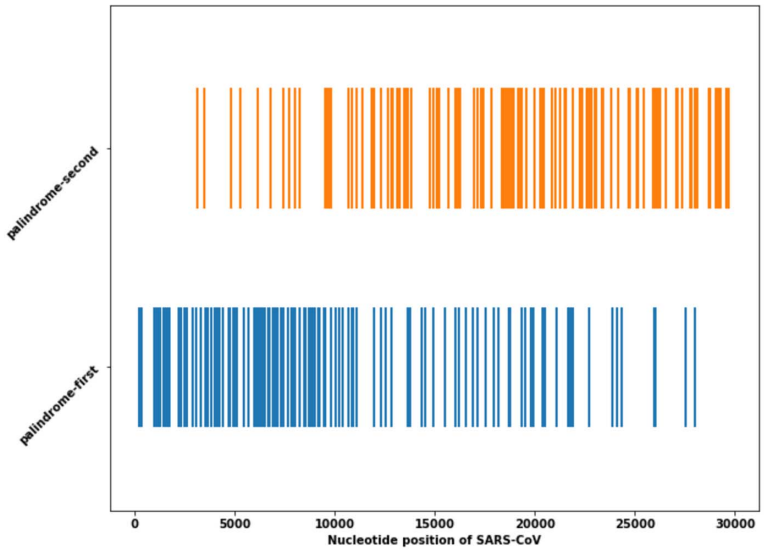


(a)

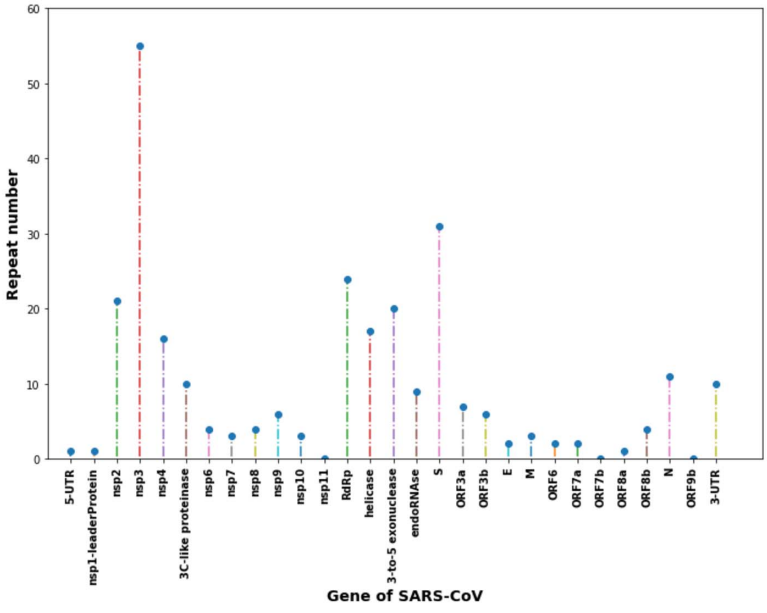


(b)

Figure 2: Distributions of inverted repeats consisting of first half sequences and second sequences on SARS-CoV-2 genome (NC_045512). (a) Inverted repeats of 11-15 bp. (b) Repeat numbers of inverted repeats of 12-15 bp in the protein genes of the genome. In (b), the repeat numbers are counted by both palindrome-first and palindrome-second sequences.

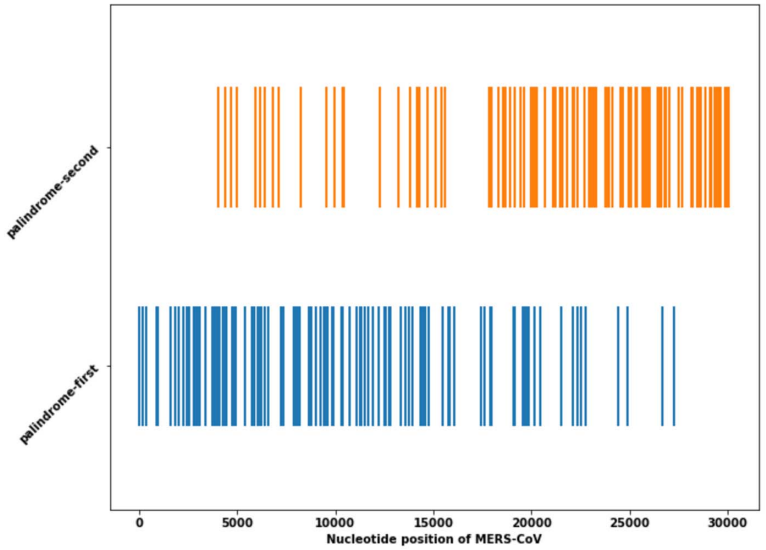


(a)

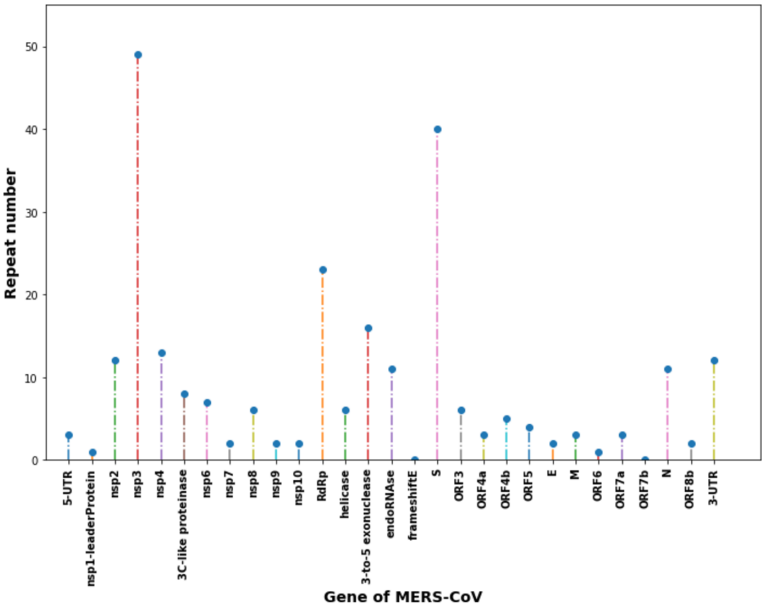


(b)

Figure 3: Distributions of inverted repeats consisting of palindrome-first and palindrome-second sequences on SARS-CoV genome (NC_004718). (a) Inverted repeats of 11-15 bp. (b) Repeat numbers of inverted repeats of 12-15 bp in the protein genes of the genome.

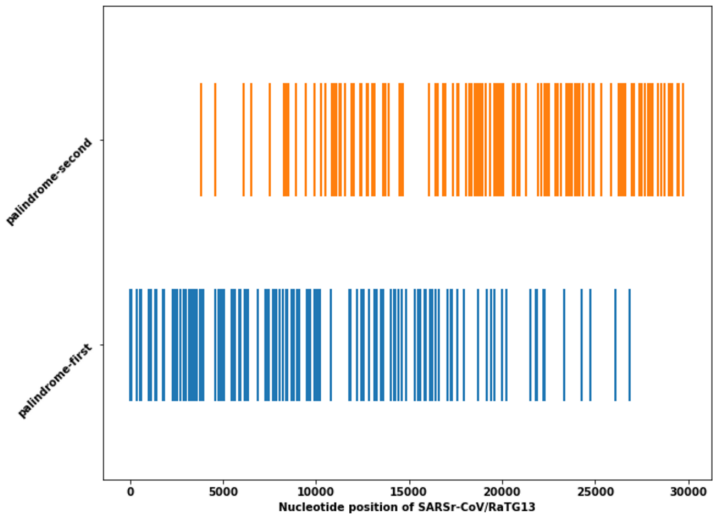


(a)

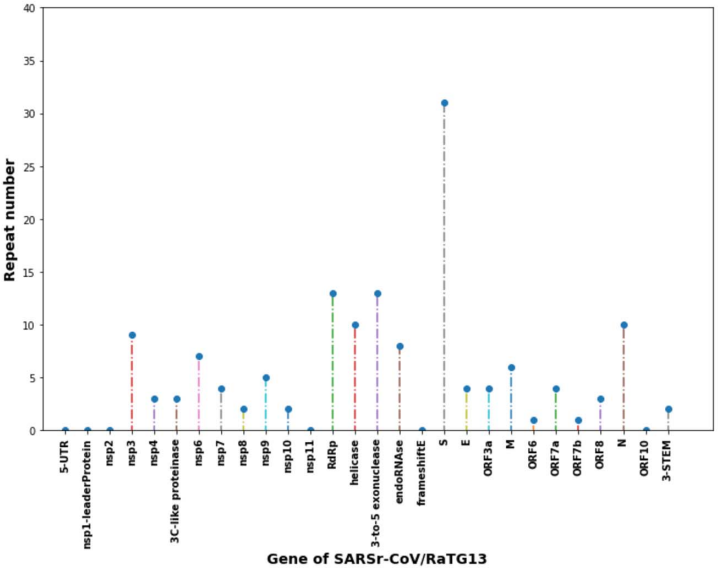


(b)

Figure 4: Distributions of inverted repeats consisting of palindrome-first and palindrome-second sequences on MERS-CoV genome (NC_019843). (a) Inverted repeats of 11-15 bp. (b) Repeat numbers of inverted repeats of 12-15 bp in the protein genes of the genome.

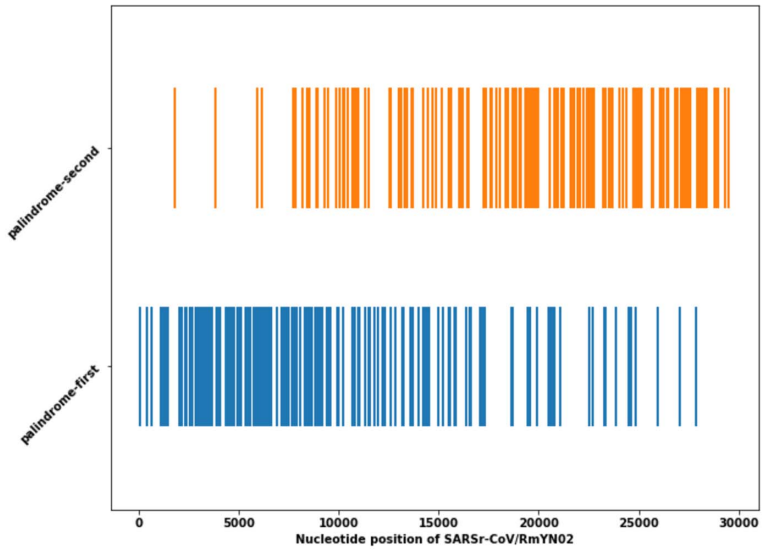


(a)

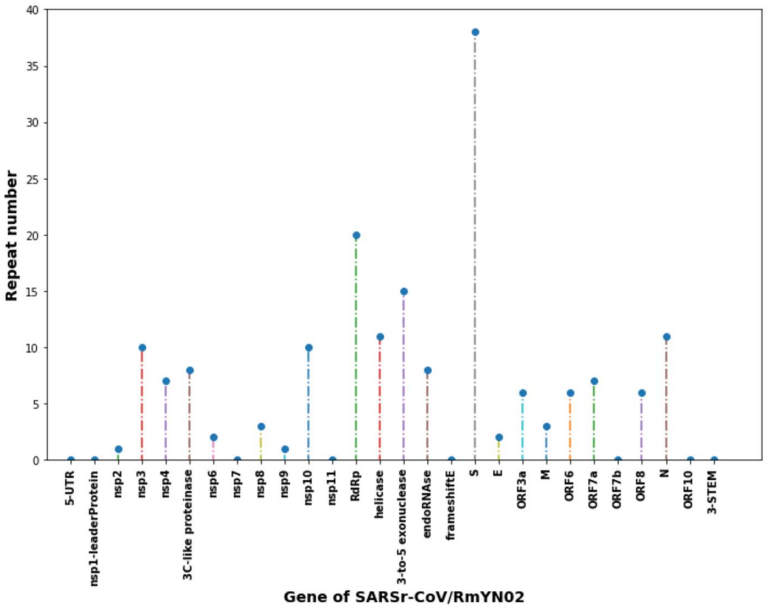


(b)

Figure 5: Distributions of inverted repeats consisting of palindrome-first and palindrome-second sequences on SARSr/RaTG13 genome (MN996532). (a) Inverted repeats of 11-15 bp. (b) Repeat numbers of inverted repeats of 12-15 bp in the protein genes of the genome.



(a)



(b)

Figure 6: Distributions of inverted repeats consisting of palindrome-first and palindrome-second sequences on SARSr-CoV/RmYN02 genome (EPI_ISL_412977). (a) Inverted repeats of 11-15 bp. (b) Repeat numbers of inverted repeats of 12-15 bp in the protein genes of the genome.

SARS-CoV (Fig. 3(b)) and MERS-CoV (Fig. 4(b)), while SARS-CoV-2-related CoVs, SARSr-CoV/RaTG13 and SARSr-CoV/RmYN02, have similar levels of long repeats in the nsp3 gene (Fig. 5(b) and Fig. 6(b)). The biased inverted repeats of nsp3 genes of CoVs may correlate with the pathogenicity level of the CoVs. Nsp3 is the largest multiple-function protein is essential for anchoring the replication/transcription complex (RTC) [18]. Nsp3 is the scaffold protein to bind viral RNA, N protein, Nsp4, Nsp6, and host proteins and interacts with host ER membranes, forming convoluted membranes (CMs) and double-membrane vesicles (DMV). The exact roles of the ample long inverted repeats in SARS-CoV and MERS-CoV are not clear. Probably the long inverted repeats can stabilize local RNA structures, protecting the RNAs being degraded by host RNAases during initial infection.

We observed that 3-to-5' exonuclease has a relatively low repeat number in SARS-CoV-2. However, the repeat number of 3-to-5' exonuclease is relatively high in SARS-CoV, MERS-CoV, and RaTG13. The 3-to-5' exonuclease (nsp14) is responsible for proof-read in RNA replication so that the mutation rate is relatively low in CoVs. The abundance change in nsp14 probably have impacts on the replication accuracy, therefore, affecting mutation rate in SARS-CoV-2.

Taking account of the inverted repeats of wide ranges 8-15 bp, we computed the pairwise Wasserstein distances of the repeat numbers of protein genes in three closely related SARSr-CoVs: the distance between SARS-CoV-2 and SARSr-CoV/RaTG13 is 6.8571, the distance between SARS-CoV-2 and SARSr-CoV/RmYN02 is 5.7143, the distance between SARSr-CoV/RaTG13 and SARSr-CoV/RmYN02 is 6.3571, and the distance between SARS-CoV-2 and MERS-CoV is 27.5. Therefore, we conclude that SARS-CoV-2 strain is more closely related to SARSr-CoV/RaTG13 (2013) than SARSr-CoV/RmYN02 (2019). Both SARS-CoV-2 and SARSr-CoV/RmYN02 may evolve from SARSr-CoV/RaTG13. In addition, the Wasserstein distances suggest that SARS-CoV-2 (2019) is the farthest from MERS-CoV (2012), followed by SARS-CoV (2002). We also observe that the Spike protein gene in SARSr-CoV/RmYN02 (Fig. 6(b)) have more long inverted repeats than the counterparts of SARS-CoV-2 (Fig. 2(b)) and SARSr-CoV/RaTG13 (Fig. 5(b)). Unsurprisingly, the Spike protein in SARSr-CoV/RmYN02 contains natural insertions at the S1/S2 cleavage site. This cleavage site may originate from some recombination events of the Spike genes as the result of inverted repeats.

The total frequencies of inverted repeats of different lengths in the human and bat CoVs also suggest that SARS-CoV-2 is closely related SARSr-

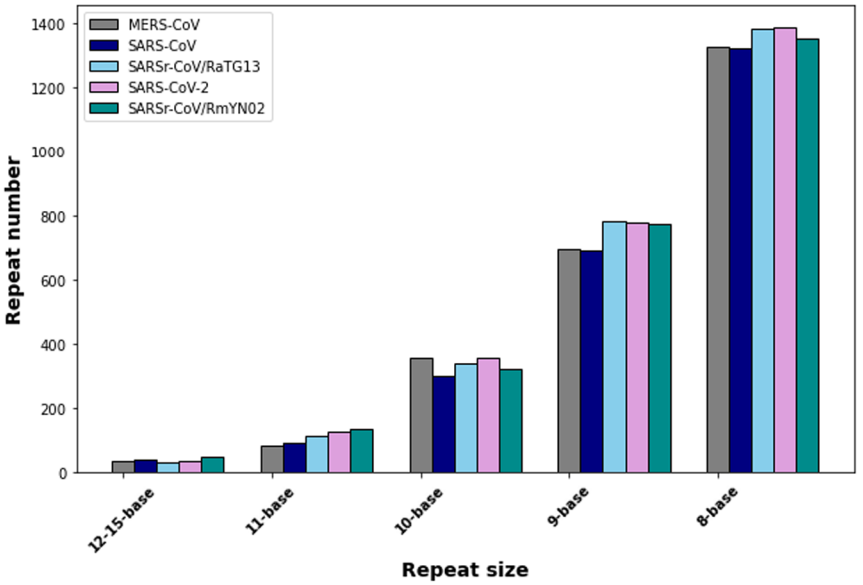


Figure 7: Frequencies of inverted repeats of different lengths in the coronavirus genomes: SARS-CoV-2, SARS-CoV, MERS-CoV, SARSr-CoV/RaTG13, and SARSr-CoV/RmYN02. The repeat numbers are counted by palindrome-first sequences only.

CoV/RaTG13 (Fig. 7). Notedly, Fig. 7 shows that the inverted repeats of all lengths are increasing from SARS-CoV (in 2003) to SARS-CoV-2 (in 2019).

From these repeat analyses, we may infer that during evolution, the recombinations may occur and produce accumulative inverted repeats under natural selection. We see that recombinations can be one of the driven forces for fast evolution.

4. Discussions

The COVID-19 pandemic has caused substantial health emergencies and economic stress in the world. Vaccine development is critical to mitigating the pandemic. The facts revealed in this study that three proteins nsp3, RdRp, and the Spike protein are rich with inverted repeats suggest that these three proteins are functional significance for virus survivals, and shall be the targets of drug design and vaccine development.

If we relax the matching pairs in the inverted repeats, we expect that much longer inverted repeats can be identified, and the number of inverted

repeats in the virus genome will be increased significantly. The imperfect inverted repeats are the natural forms of the repeats to maintain the genome structures. Because the perfect inverted repeat distribution and types in a genome are unique and extracting the perfect inverted repeats are parameter-free, the perfect inverted repeats can be considered as the genomic signature. The signatures from perfect inverted repeats are consistent, therefore, can be used for 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 inverted repeats can be an effective barcode to delimit species and genotypes.

This research has limitations. All the results are based on theoretical analysis of genomes and lack experimental supports. The repeats are abundant in nsp3, RdRp, S, N, and helicase genes, however, the impacts of the inverted repeats on the structures and functions of the proteins are not determined. The functional impact of the inverted repeats through collaboration with experiments is our future research direction.

Acknowledgments

We sincerely appreciate the researchers worldwide who sequenced and shared the complete genome data of SARS-CoV-2 and other 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), Tsinghua University Spring Breeze Fund (2020Z99CFY044, to S.S.-T. Yau), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (042202008, to S.S.-T. Yau).

Competing interests

We declare we have no competing interests.

Abbreviations

- COVID-19: coronavirus disease 2019
- SARS: severe acute respiratory syndrome
- SARS-CoV-2: severe acute respiratory syndrome coronavirus 2
- MERS-CoV: Middle East Respiratory Syndrome coronavirus
- CRISPR: clusters of regularly interspaced short palindromic repeats
- ACE2: angiotensin-converting enzyme 2
- NCBI: National Center for Biotechnology Information (USA)

References

[1] Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R., Hilgenfeld, R., 2003. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science 300 (5626), 1763–1767.

[2] Angelini, M. M., Akhlaghpour, M., Neuman, B. W., Buchmeier, M. J., 2013. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. MBio 4 (4), e00524–13.

[3] Bochman, M. L., Paeschke, K., Zakian, V. A., 2012. DNA secondary structures: stability and function of G-quadruplex structures. Nature Reviews Genetics 13 (11), 770–780.

[4] 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.

[5] Cavasotto, C. N., Lamas, M. S., Maggini, J., 2020. Functional and druggability analysis of the SARS-CoV-2 proteome. European Journal of Pharmacology, 173705.

[6] Chen, J., Wang, R., Wang, M., Wei, G.-W., 2020. Mutations strengthened SARS-CoV-2 infectivity. Journal of Molecular Biology 432 (1), 5212–5226.

[7] Chen, Y., Liu, Q., Guo, D., 2020. Emerging coronaviruses: genome structure, replication, and pathogenesis. Journal of Medical Virology 92, 418–423.

[8] Choi, J., Majima, T., 2011. Conformational changes of non-B DNA. Chemical Society Reviews 40 (12), 5893–5909.

[9] Coutard, B., Valle, C., de Lamballerie, X., Canard, B., Seidah, N., Decroly, E., 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Research 176, 104742.

[10] Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Melamed, S., Weiss, S., Isrealy, T., et al., 2020. The coding capacity of SARS-CoV-2. Nature.

[11] Gao, K., Nguyen, D. D., Wang, R., Wei, G.-W., 2020. Machine intelligence design of 2019-nCoV drugs. bioRxiv.

[12] Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., et al., 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature, 1–13.

[13] He, R., Leeson, A., Ballantine, M., Andonov, A., Baker, L., Dobie, F., Li, Y., Bastien, N., Feldmann, H., Strocher, U., et al., 2004. Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. Virus Research 105 (2), 121–125.

[14] Hoang, T., Yin, C., Zheng, H., Yu, C., He, R. L., Yau, S. S.-T., 2015. A new method to cluster DNA sequences using fourier power spectrum. Journal of Theoretical Biology 372, 135–145. MR3331829

[15] Jehan, Z., Vallinayagam, S., Tiwari, S., Pradhan, S., Singh, L., Suresh, A., Reddy, H. M., Ahuja, Y., Jesudasan, R. A., 2007. Novel noncoding RNA from human Y distal heterochromatic block (Yq12) generates testis-specific chimeric CDC2L2. Genome Research 17 (4), 433–440.

[16] Ji, D., Juhas, M., Tsang, C. M., Kwok, C. K., Li, Y., Zhang, Y., 2020. Discovery of G-quadruplex-forming sequences in SARS-CoV-2. Briefings in Bioinformatics.

[17] Le, F., Groshan, K., Zeng, X. P., Richelson, E., 1997. Characterization of the genomic structure, promoter region, and a tetranucleotide repeat polymorphism of the human neurotensin receptor gene. Journal of Biological Chemistry 272 (2), 1315–1322.

[18] Lei, J., Kusov, Y., Hilgenfeld, R., 2018. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral Research 149, 58–74.

[19] Li, C. K.-f., Wu, H., Yan, H., Ma, S., Wang, L., Zhang, M., Tang, X., Temperton, N. J., Weiss, R. A., Brenchley, J. M., et al., 2008. T cell responses to whole SARS coronavirus in humans. The Journal of Immunology 181 (8), 5490–5500.

[20] Li, T., Zhang, Y., Fu, L., Yu, C., Li, X., Li, Y., Zhang, X., Rong, Z., Wang, Y., Ning, H., et al., 2005. siRNA targeting the leader sequence of SARS-CoV inhibits virus replication. Gene Therapy 12 (9), 751–761.

[21] Liu, J., Sun, Y., Qi, J., Chu, F., Wu, H., Gao, F., Li, T., Yan, J., Gao, G. F., 2010. The membrane protein of severe acute respiratory syndrome coronavirus acts as a dominant immunogen revealed by a clustering region of novel functionally and structurally defined cytotoxic T-lymphocyte epitopes. Journal of Infectious Diseases 202 (8), 1171–1180.

[22] Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., et al., 2003. The genome sequence of the SARS-associated coronavirus. Science 300 (5624), 1399–1404.

[23] Max Roser, Hannah Ritchie, E. O.-O., Hasell, J., 2020. Coronavirus Pandemic (COVID-19). Our World in Data https://ourworldindata.org/coronavirus.

[24] Métifiot, M., Amrane, S., Litvak, S., Andreola, M.-L., 2014. G-quadruplexes in viruses: function and potential therapeutic applications. Nucleic Acids Research 42 (20), 12352–12366.

[25] 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.

[26] Nakahori, Y., Mitani, K., Yamada, M., Nakagome, Y., 1986. A human Y-chromosome specific repeated DNA family (DYZ1) consists of a tandem array of pentanucleotides. Nucleic Acids Research 14 (19), 7569–7580.

[27] Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., et al., 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nature Genetics 36 (1), 40–45.

[28] Pathak, D., Premi, S., Srivastava, J., Chandy, S. P., Ali, S., 2006. Genomic instability of the DYZ1 repeat in patients with Y chromosome anomalies and males exposed to natural background radiation. DNA Research 13 (3), 103–109.

[29] 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.

[30] Prentice, E., McAuliffe, J., Lu, X., Subbarao, K., Denison, M. R., 2004. Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. Journal of Virology 78 (18), 9977–9986.

[31] Ruan, Y., Wei, C. L., Ling, A. E., Vega, V. B., Thoreau, H., Thoe, S. Y. S., Chia, J.-M., Ng, P., Chiu, K. P., Lim, L., et al., 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. The Lancet 361 (9371), 1779–1785.

[32] Selker, E. U., 1999. Gene silencing: repeats that count. Cell 97 (2), 157–160.

[33] Serrano, P., Johnson, M. A., Chatterjee, A., Neuman, B. W., Joseph, J. S., Buchmeier, M. J., Kuhn, P., Wüthrich, K., 2009. Nuclear magnetic resonance structure of the nucleic acid-binding domain of severe acute respiratory syndrome coronavirus nonstructural protein 3. Journal of Virology 83 (24), 12998–13008.

[34] Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22 (13).

[35] Smith, J. O., 2007. Mathematics of the discrete Fourier transform (DFT). W3K: Charleston, SC, USA, 7–9.

[36] Sutton, G., Fry, E., Carter, L., Sainsbury, S., Walter, T., Nettleship, J., Berrow, N., Owens, R., Gilbert, R., Davidson, A., et al., 2004. The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. Structure 12 (2), 341–353.

[37] Te Velthuis, A. J., van den Worm, S. H., Snijder, E. J., 2012. The SARS-coronavirus nsp7+ nsp8 complex is a unique multimeric RNA polymerase capable of both de novo initiation and primer extension. Nucleic Acids Research 40 (4), 1737–1747.

[38] Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Bioinformatics 13 (3), 263–270.

[39] Vallender, S., 1974. Calculation of the Wasserstein distance between probability distributions on the line. Theory of Probability & Its Applications 18 (4), 784–786. MR0328982

[40] 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.

[41] Voss, R., 1992. Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. Physical Review Letters 68, 3805–3808.

[42] Wan, Y., Shang, J., Graham, R., Baric, R. S., Li, F., 2020. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. Journal of Virology 94 (7).

[43] Wong, S. K., Li, W., Moore, M. J., Choe, H., Farzan, M., 2004. A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2. Journal of Biological Chemistry 279 (5), 3197–3201.

[44] 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.

[45] Yin, C., Chen, Y., Yau, S. S.-T., 2014. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. Journal of Theoretical Biology 359, 18–28. MR3248415

[46] Yin, C., Yau, S. S.-T., 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. Journal of Theoretical Biology 247 (4), 687–694. MR2479617

[47] Yin, C., Yau, S. S.-T., 2015. An improved model for whole genome phylogenetic analysis by Fourier transform. Journal of Theoretical Biology 359 (21), 18–28. MR3385919

[48] Yin, C., Yau, S. S.-T., 2017. A coevolution analysis for identifying protein-protein interactions by Fourier transform. PLoS One 12 (4), e0174862.

[49] Yoshimoto, F. K., 2020. The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. The Protein Journal, 1.

[50] 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.

[51] Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E. C., et al., 2020. 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.

[52] 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., 2020. Addendum: A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798).

[53] 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., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798), 270–273.

[54] Zulkower, V., Rosser, S., 2020. DNA features viewer, a sequence annotations formatting and plotting library for Python. bioRxiv.

Changchuan Yin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
USA
E-mail address: cyin1@uic.edu

Stephen S.-T. Yau
Department of Mathematical Sciences
Tsinghua University
Beijing 100084
China
E-mail address: yau@uic.edu

Received November 25, 2020