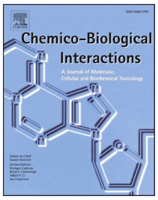
Chemico-Biological Interactions 331 (2020) 109226

Contents lists available at ScienceDirect

Chemico-Biological Interactions

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



Research Article

Deciphering the SSR incidences across viral members of *Coronaviridae* family

Rohit Satyam[a], Niraj Kumar Jha[b,*], Rohan Kar[c], Saurabh Kumar Jha[b], Ankur Sharma[d], Dhruv Kumar[e], Parma Nand[b], Janne Ruokolainen[f], Kavindra Kumar Kesari[f], Mohammad Amjad Kamal[g,h]

[a] *Department of Biotechnology, Noida Institute of Engineering and Technology (NIET), Greater Noida, India*
[b] *Department of Biotechnology, School of Engineering & Technology (SET), Sharda University, Greater Noida, 201310, India*
[c] *Indian Institute of Management Ahmedabad (IIMA), Gujarat, 380015, India*
[d] *Department of Life Science, School of Basic Science & Research, Sharda University, Greater Noida, 201310, India*
[e] *Amity Institute of Molecular Medicine and Stem Cell Research (AIMMSCR), Amity University Uttar Pradesh, Noida, 201313, India*
[f] *Department of Applied Physics, Aalto University, Espoo, Finland*
[g] *King Fahd Medical Research Center, King Abdulaziz University, P. O. Box 80216, Jeddah, 21589, Saudi Arabia*
[h] *Enzymoics, Novel Global Community Educational Foundation, 7 Peterlee Place, Hebersham, NSW, 2770, Australia*

ARTICLE INFO

Keywords:
Simple sequence repeats (SSRs)
Coronaviruses
SARS-CoV-2
Evolution
Mutations

ABSTRACT

Presence of Simple Sequence Repeats (SSRs), both in genic and intergenic regions, have been widely studied in eukaryotes, prokaryotes, and viruses. In the current study, we undertook a survey to analyze the frequency and distribution of microsatellites or SSRs in multiple genomes of *Coronaviridae* members. We successfully identified 919 SSRs with length ≥12 bp across 55 reference genomes majority of which (838 SSRs) were found abundant in genic regions. The *in-silico* analysis further identified the preferential abundance of hexameric SSRs than any other size-based motif class. Our analysis shows that the genome size and GC content of the genome had a weak influence on SSR frequency and density. However, we find a positive correlation of SSRs GC content with genomic GC content. We also report relatively low abundances of all theoretically possible 501 repeat motif classes in all the genomes of *Coronaviridae*. The majority of SSRs were AT-rich. Overall, we see an underrepresentation of SSRs across the genomes of *Coronaviridae*. Besides, our integrative study highlights the presence of SSRs in ORF1ab (nsp3, nsp4, nsp5A_3CLpro and nsp5B_3CLpro, nsp6, nsp10, nsp12, nsp13, & nsp15 domains), S, ORF3a, ORF7a, N & 3′ UTR regions of SARS-CoV-2 and harbours multiple mutations (3′UTR and ORF1ab SSRs serving as major mutational hotspots). This indicates the genic SSRs are under selection pressure against mutations that might alter the reading frame and at the same time responsible for rapid protein evolution. Our preliminary results indicate the significance of the limited repertoire of SSRs in the genomes of *Coronaviridae*.

1. Introduction

Coronaviruses are known to cause mild to severe respiratory, gastrointestinal, and central nervous system infections both in humans and other vertebrates [41]. The viruses weren't considered highly pathogenic until the 2003 outbreak of SARS (Severe Acute Respiratory Syndrome) [42] followed by the emergence of MERS (Middle East Respiratory Syndrome) in Middle Eastern countries [6] and SARS-CoV-2 outbreak of 2019 [13]. The viruses are the members of *Coronaviridae* family (order *Nidovirales*) and have host preferences; for instance, Alpha & Beta-coronaviruses predominantly infect mammals whereas Gamma & Delta majorly infect birds (and sporadically mammals). Bats are believed to be the largest reservoirs of diverse coronaviruses than animal species; domestic and poultry animals being the intermediate hosts, that cause zoonotic transmission of virus finally to the humans [40]. The ecological distribution, evolution, and spillover events of various coronaviruses have been extensively reviewed in some recent reports [5,8,32,36].

[*] Corresponding author. Department of Biotechnology, School of Engineering & Technology (SET) Sharda University, Plot No. 32-34, Knowledge Park III, Greater Noida, Uttar Pradesh, 201310, India.
E-mail addresses: nirajkumarjha2011@gmail.com, niraj.jha@sharda.ac.in (N.K. Jha).

https://doi.org/10.1016/j.cbi.2020.109226
Received 10 May 2020; Received in revised form 5 August 2020; Accepted 11 August 2020
Available online 21 September 2020

Simple sequence repeats (SSRs), refer to tandem repetitions of mono-, di-, tri-, tetra-, penta- and hexanucleotide sequence units of a genome and are widely reported to be the most variable type of short motifs within the viral genome. They are ubiquitously present in a variety of genomic regions including the 3′-UTRs, 5′-UTRs (Untranslated Regions), genic (coding regions), and intergenic regions (non-coding regions) thereby conferring to diverse roles across viral species [44]. SSRs have been widely exploited as neutral markers in multitudes of studies such as ecology and evolutionary genetics, genome mapping, etc. irrespective of their hypermutablility [37,39]. They are characterized by their inherent ability to cause frameshift mutations in genomic regions encoding phenotypic changes and therefore, confer an adaptive advantage in the course of viral mutations [1]; [21,23]. Their highly polymorphic nature results in gain/loss of repeat motifs which makes them altogether important to study the genome evolution.

Despite the deluge of viral genomes in the public databases, the SSR incidences/abundances and their relevance in viral genomes have been given a little attention including coronaviruses in particular. Elucidating the SSR landscape in viral members of *Coronaviridae* and their prospect relevance in evolution and pathogenesis, therefore, became crucial in the current scenario of the COVID-19 outbreak [19].

Thus, the aims of the current study were 1) to analyze various facets of the distribution and dynamics of SSRs in the genomes of *Coronaviridae* members, 2) to identify patterns of SSR incidences across genomes, if any i.e. the underrepresentation/overrepresentation of specific repeat motif classes, 3) the preferential genomic localization of SSRs & 4) to investigate if SSRs serves as mutation hotspots in SARS-CoV-2, a novel SARS strain causing COVID-19 outbreak. The outcome of our study suggests that SSRs are generally underrepresented in *Coronaviridae* members and are characterized by low GC content. Additionally, the attributes of SSRs across genomes under study were quite similar in terms of length (preferentially found to be 12–13 nucleotides long with polyA repeats of varying lengths), GC composition, abundance (SSR frequency didn't exceed 2 irrespective of genome size) and localization. The trends highlighted in the current study are repercussions of the differences in the *Coronaviridae* genome organization and could serve as pitching points to understand the mutation rates in SSRs and how these mutations propagate among the coding and non-coding compartments. Besides, the study attempts to lay the groundwork for the much-needed scientific discussion on SSRs incidences in *Coronaviridae* genomes and endeavours to test their biological significance in pathogenesis, evolution, and immune evasion.

2. Methods

2.1. Identification of microsatellites in Coronaviridae genomes

The 55 complete genomes of *Coronaviridae* families were retrieved on March 23, 2020 (See Supplementary Material 1, Sheet 2 for more details, we used only RefSeq Nucleotides with complete annotations) and were scanned in search of SSRs using a Python package, PERF [2]. A minimum length of SSRs was chosen to be 12 nt [24,34] which represents at least two complete repeating units of a 6-mer motif (hexamer). We used all theoretically possible 501 unique classes of SSRs as described in a study [34,35] to identify their presence/absence in *Coronaviridae* genomes by using the following command: *"PERF -isequence.fasta -a -o sequence_perf_default.tsv"*. The interactive. html pages were used to manually visualize and analyze SSRs prediction data and understand their attributes. The BED files (eg.sequence_perf_default.tsv) so produced by PERF comprise of SSRs genomic coordinates (Column 1–3) followed by repeat class, repeat Length, repeat Strand, motif Number & actual repeat (more details: https://github.com/RKMlab/perf) and were used for the downstream analysis.

2.2. Basic attributes of identified SSRs

For each genome, we computed a few attributes to measure the prevalence of SSRs in the viral genomes of the *Coronaviridae* family. These included SSR frequency (or abundance), SSR density, and SSR GC %. The SSR frequency was defined as the total number of SSRs found in each genome. The SSR density was computed as per the formula

$$SSR_d = \left(\sum SSR_L / G_L\right)*1000$$

Where SSR_L is the length of SSR (in bp) and G_L, the genomic length (in bp), and SSR_d, the SSR density per Kb. This was attempted to normalize and take care of the biases that could crop up due to variable G_L. We use SSR density as a measure of comparison throughout the study unless otherwise mentioned. The SSR GC% for a genome was defined as the GC % of concatenated strings of SSRs retrieved using the coordinates from the BED file. Briefly, we used samtools [14], bedtools [29], and seqkit [31] in combination to compute GC content.

2.3. Class-specific attributes of SSRs

To identify the class-specific trends of SSRs we computed class-specific SSR frequency, SSR base coverage, and SSR density for each of the 501 repeat classes using in-house scripts. The list of 501 SSRs were obtained from Additional file of [35]. The class-specific SSR frequency was computed and collated in the form of a matrix where each row represented repeat class and column represented frequency of that class in each genome. The matrix was visualized and analyzed using Morpheus (https://software.broadinstitute.org/morpheus/) of Broad Institute. The repeat classes were subjected to Hierarchal Clustering using Euclidean distance. The heatmap of the repeat classes present in at least 10 (≥10) members was constructed. The color scale on the heatmap ranged from 0 to 3, 3 being the highest SSR frequency observed. We also checked for variation of repeat class abundance with respect to repeat class lengths using python script from Ref. [34].

2.4. GC% and motif size based SSR composition

The 501 repeat classes were divided into 5 GC cluster groups based on the GC content of the repeat motif. The 60 bp strings formed by repeating the base motif in tandem were constructed and GC content computed using '*seqkit fx2tab -g*' command. The GC cluster group so formed were ≤25%, 26–49%, 50–60%, 61–80% & 81–100%, which encompass 70, 120, 153, 112 & 46 repeat motifs respectively.

Based on the length of repeat motifs, repeat classes were categorized as monomers, dimers, trimers, tetramers, pentamers & hexamers. The motifs with similar length were groped in similar size category.

2.5. Annotation of SSRs as genic or intergenic

The accession list was used to query Batch Entrez Assembly Database to procure GFF files containing genomic feature annotation information of all viruses under the current study. The SSRs annotation was accomplished using an *in-house* developed shell script that computes 4 possible overlap scenarios of SSRs with genic regions. Briefly, the script parses both. tsv files obtained as PERF output and. gff files and performs a coordinate-based comparison [2]. The SSRs overlapping with two or more genes were counted as one while computing the SSR abundances in the genic region. Besides, for all overlapping SSRs with the genic region, the percentage overlap of SSR with the genic region is also reported. The step was critical to negate the skewness stemming in otherwise if the majority of SSRs were found populated within genic-intergenic boundaries. We verified that >95% of exonic SSRs show a complete overlap with exons. We also carried out a variant analysis of SARS-CoV-2 SSRs to decipher if the SSRs serves as the mutational hotspots.

4935 variant sites made available public by NGDC (National

Genomics Data Center) stemming from the analysis of 11641 high-quality human-derived SARS-CoV-2 genome sequences were downloaded last on 28-04-2017. The identified variant sites are graded into three levels (I to III) based on population frequency and mutation density distribution. The class I variants are one with the highest population frequency (>0.05, more credible); class II variant are sites with moderate population frequency and class III being one with <0.05 population frequency, hence low reliability (detailed in table 'Variation Annotation', https://bigd.big.ac.cn/ncov/variation/annotation). A custom shell script was deployed to check if the variants were majorly localized in genic SSRs of SARS-CoV-2.

2.6. Primer designing

The Severe Acute Respiratory Syndrome-related coronavirus, Middle East Respiratory Syndrome-related coronavirus, and Human coronavirus OC43 strains were used for primer designing. The SSRs with 70 bp flanking were retrieved using *samtools* and *seqkit* and were converted to query files using a customized in-house bash script. We used *primer3_core*conda package to retrieve the primer sequence with the custom settings: PRIMER_TASK = generic, PRIMER_PICK_LEFT_PRIMER = 1, PRIMER_PICK_RIGHT_PRIMER = 1, PRIMER_OPT_SIZE = 18, PRIMER_MIN_SIZE = 15, PRIMER_MAX_SIZE = 21, PRIMER_MAX_NS_ACCEPTED = 1, PRIMER_PRODUCT_SIZE_RANGE = 75–100, P3_FILE_FLAG = 1,PRIMER_EXPLAIN_FLAG = 1,PRIMER_MIN_GC = 40, PRIMER_OPT_GC_PERCENT = 50, PRIMER_MAX_GC = 60. The sequence for which the primers couldn't be determined via automated scripts were identified separately by tweaking GC content and other settings (PRIMER_MIN_GC = 30 and SEQUENCE_TARGET = 70,2). The primers so designed were checked for off-targets if any using BLASTn.

3. Results & discussion

3.1. Genome-wide characterization of microsatellites in Coronaviridae family

To screen microsatellites in 55 *Coronaviridae* genomes, we used a PERF package, an exhaustive repeat finding algorithm, to search for all 501 theoretically possible SSR motifs [35] occurrences in the genomes [2].

A total of 919 SSRs with length $\geq$12 bp were identified across 55 reference genomes belonging to two subfamilies: *Coronavirinae* and *Orthocoronavirinae*. The top 4 strains with the largest number of SSRs were human infecting coronaviruses *viz.* Human coronavirus HKU1 (29 S SRs, NC_006577), Human coronavirus OC43 (26 S SRs, NC_006213), Severe acute respiratory syndrome-related coronavirus (25 S SRs, NC_004718), Human coronavirus NL63 (25 S SRs, NC_005831).

The genome size can influence SSRs incidences. Therefore, to account for the variation we calculated the SSR density which is the number of bases covered by SSRs per Kb and plotted the results. As evident from Fig. 1, the SSR density is independent of the genome size. We also computed the correlation coefficient (r) between SSR density and genome length & SSR frequency and genome length. Unlike eukaryotic genomes [34], we found the SSR frequency and SSR density is not correlated with genome size (Pearson, r = 0.149987561 (Density Vs Genome Length) & r = 0.350653216 (Frequency Vs Genome Length).

The highest SSR abundance per Kb was observed in Human coronavirus HKU1 Human coronavirus HKU1, NC_006577 (12.26358351 bp/Kb) *viz*. 1.22% of the genome is covered with SSRs. The HUK1 is followed by Human coronavirus NL63, NC_005831 (11.57768664 bp/Kb), Tor2 SARS-CoV, NC_004718 (11.05845182 bp/Kb) & Human coronavirus OC43, NC_006213 (11.02761784 bp/Kb) (Fig. 2**).** A recent study carried out elsewhere highlights high sequence similarity of CDS of SARS-CoV-2 with 4 coronavirus strains; a Bat Relative (bat-SL-CoVZXC21, MG772934), Tor2 SARS-CoV (NC_004718), and HCoV-EMC MERS-CoV (NC_019843) [12]. We, therefore, decided to compare the SSR repertoire across the aforementioned homologues to SSRs found in SARS-CoV-2 (NC_045512). The Tor2 SARS-CoV SSR density (11.05845182 bp/kb) is higher than SARS-CoV-2 (9.296726081 bp/kb) followed by bat-SL-CoVZXC21 (8.293482259 bp/kb) & least in HCoV-EMC (4.515422159 bp/kb).

The genomic composition of viruses can vary widely and dictates mutational bias toward AT or GC. We, therefore, evaluated genome-wide and SSR localized nucleotide composition across 55 genomes of *Coronaviridae*. Overall, the GC content of *Coronaviridae* genomes was found to range from ~32 to 48%. It has been highlighted in previous studies that *Coronaviridae* genomes have underrepresented CpG ratio which might confer the members of the family, an advantage of immune evasion in vertebrates where immune pathways target CpG rich regions (eg TLRs). Moreover, Coronaviruses exhibits atypical nucleotide composition with high levels of Ts and low levels of Cs, perhaps due to cytokine deamination [7].

The SSRs GC% shows an overall moderate uphill (positive) relationship with genomic GC% (Pearson, r = 0.510175). Interestingly, we found a significant correlation between genomic GC% and SSR's GC% when the organisms were grouped according to their genus (Fig. 3). We observed genus-wise high correlation in *Coronavirinae* (Genomic GC: 40–41; SSRs GC: 29–35, r = 1), & Middle East respiratory syndrome-

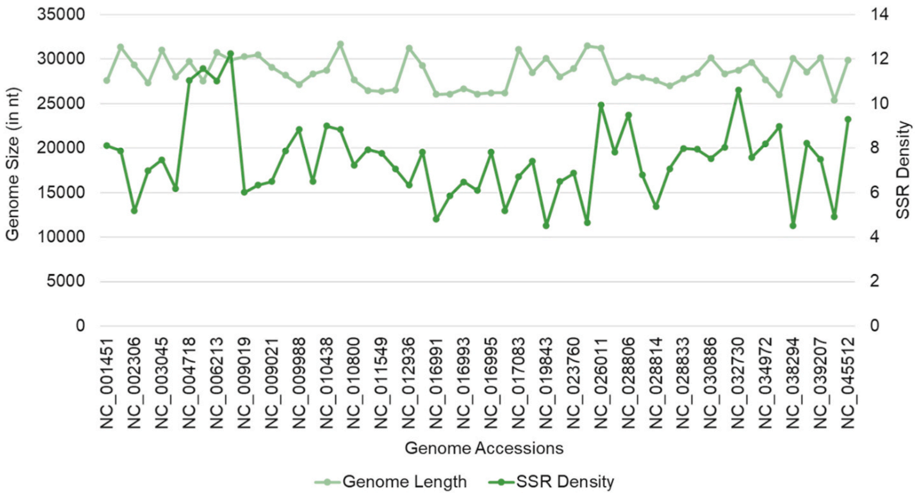


Fig. 1. Overview of SSRs density variation with respect to genome size. No correlation was observed between SSR density with genome size (Pearson, r = 0.149987561).

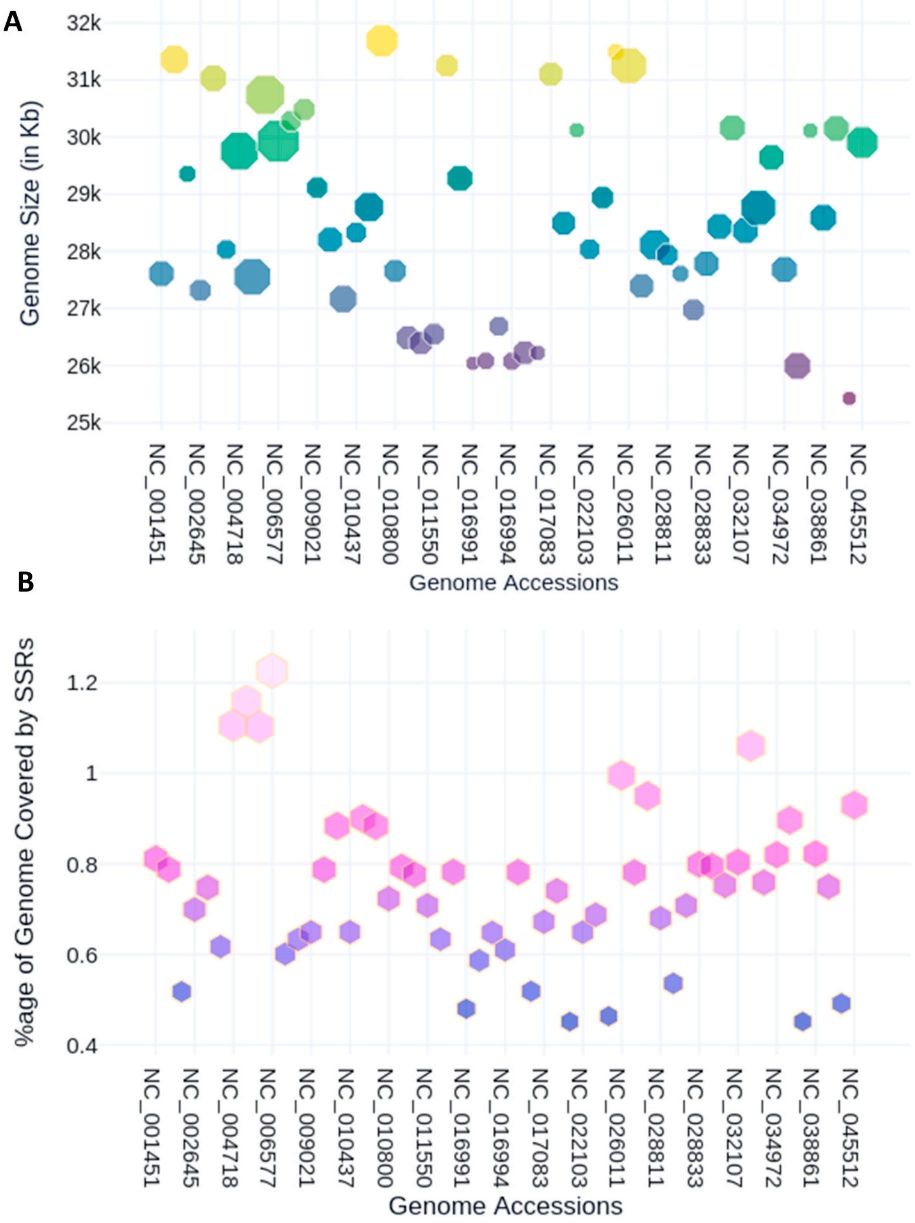


Fig. 2. Overview of A) Percentage of the genome covered by SSRs (SSRs base coverage) and B) their respective genome sizes. The size of octagons (in A) represents the number of bases covered by SSRs in their respective genome. The SSRs form a small fraction of the entire genomes of *Coronaviridae* members.

related coronavirus (Genomic GC: 37–141; SSRs GC: 31–33, r = 0.931311334) followed by Betacoronavirus 1 (Genomic GC: 32–43; SSRs GC: 23–36, r = 0.870653089), Alphacoronavirus 1 (Genomic GC: 36–41; SSRs GC: 30–43, r = 0.835273), Avian coronavirus (Genomic GC: 37–46; SSRs GC: 23–41, r = 0.73215046), Severe acute respiratory syndrome-related coronavirus Genus (Genomic GC: 31–44; SSRs GC: 24–33, r = 0.693393146), Human coronavirus 229E (Genomic GC: 35–40; SSRs GC: 30–38, r = 0.67135439) indicating GC rich SSR abundances. Murine coronavirus and unclassified Betacoronavirus, however, do not such correlation. The bat-SL-CoVZXC21 have similar genomic GC% to that of SARS-CoV-2 but dissimilar SSR GC%.

3.2. SSR abundances across 501 repeat motif classes

We calculated relative abundances of all 501 theoretically possible repeat motifs across 55 genomes and plotted a heatmap based on the observed motif frequencies. Most of the SSRs were 12–13 nt long except a polyA monomeric repeats (polyA) of varying lengths. We observed a distinct pattern that *Coronaviridae* members have intrinsically low abundances of SSRs which do not generally exceed the frequency of 2. The heatmap was plotted for those classes of repeat motifs that were found across a minimum of 10 strains (Fig. 4). We found 16 such classes which were high in A content than in G/C content. 249 Repeat Classes were altogether absent in all the 55 genomes. The polyA monomeric repeats were however found in most of the genomes of *Orthocoronavirinae* subfamily and, majorly in MERS, SARS, and Avian Coronaviruses. The repeats were however found to be localized at the end of the chromosome in the intergenic region (3′UTR). The 3′UTR region, for instance, is reported to be conserved in beta coronaviruses and harbours the *cis*-acting sequences that form potential molecular switch required for viral replication. The sequences fold into secondary or higher-order structures that confer to RNA stability and facilitate both intra- and inter-molecular interactions [43].

Canonically, the SSR frequency is expected to decrease with growing repeat length, as longer repeats have a higher propensity of mutation [34]. Therefore, we looked at the length of each SSR across all 55

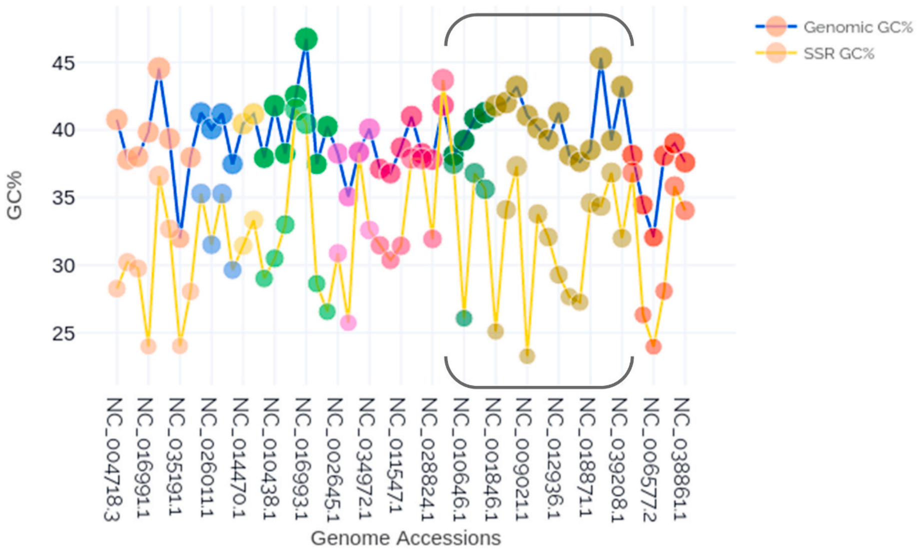


Fig. 3. The SSRs GC% correlation with genomic GC% on genus-based subgrouping. Each genus is represented by a separate color. Genus order from left to right (Severe acute respiratory syndrome-related coronavirus, Middle East respiratory syndrome-related coronavirus, unclassified Coronavirinae, Avian coronavirus, Human coronavirus 229E, Betacoronavirus 1, unclassified Betacorona, Murine coronavirus, Alphacoronavirus 1). Murine coronavirus and unclassified Betacoronavirus (highlighted by braces) however do not such correlation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

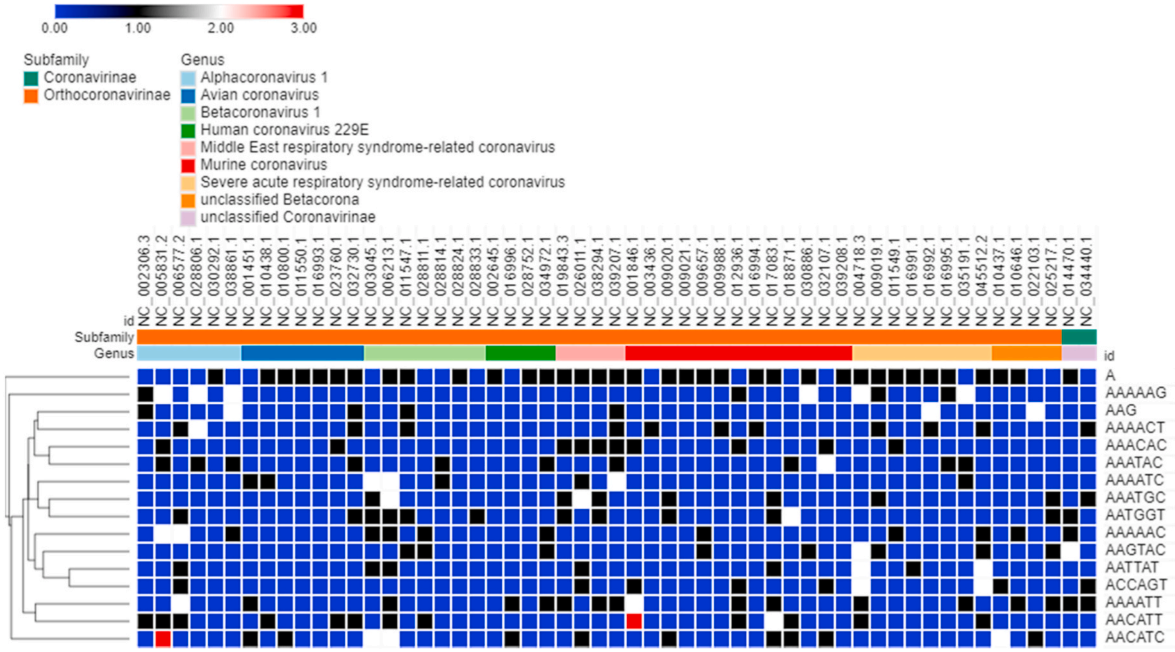


Fig. 4. SSR abundance map of repeat motif classes present in a minimum of≥10 members of *Coronaviridae* members. The scale represents the SSR frequency ranging from 0 (blue) to 3 (red). The genus and subfamily the genomes belong to are represented at the top. PolyA monomeric repeat appears to be the most prominent SSR. The SSRs are underrepresented across all genomes of the *Coronaviridae* family. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

organisms. However, we didn't observe the variation of repeat class abundance with respect to repeat class lengths in the predicted SSRs indicating that SSRs in *Coronaviridae* family exhibit no length preferences.

3.3. *Motif size and GC bases categorization of repeat class reveals hexamers abundances*

To identify if the genomes of *Coronaviridae* atypically favoured repeat classes of certain size categories, we divided the repeat classes into six size categories ranging from monomers to hexamers. We found that *Coronaviridae* genomes were majorly populated with hexamers followed by pentamers in the league. The monomeric repeats were the least abundant (Fig. 5A). The repeat motifs were also clustered into 5 subgroups based on the GC content of the repeat motifs itself as explained in methodology. Maximum SSRs belongs to the repeat classes with ≤25% GC (323 SSRs/919) across all genomes followed by subgroups with 26–49% GC content (314 SSRs/919) and 50–60% (213 SSRs/919). This is in alignment with the fact that *Coronaviridae* genomes are intrinsically AT-rich which highly influences SSRs AT- or GC- richness in different genomic regions in addition to the nucleotide distribution across genomes [7,38].

3.4. *SSR incidences are observed majorly in genic regions in Coronaviridae genomes*

The earlier study highlights the hexanucleotide SSRs abundances in exonic regions of eukaryotic genomes [20] while mono and dinucleotide

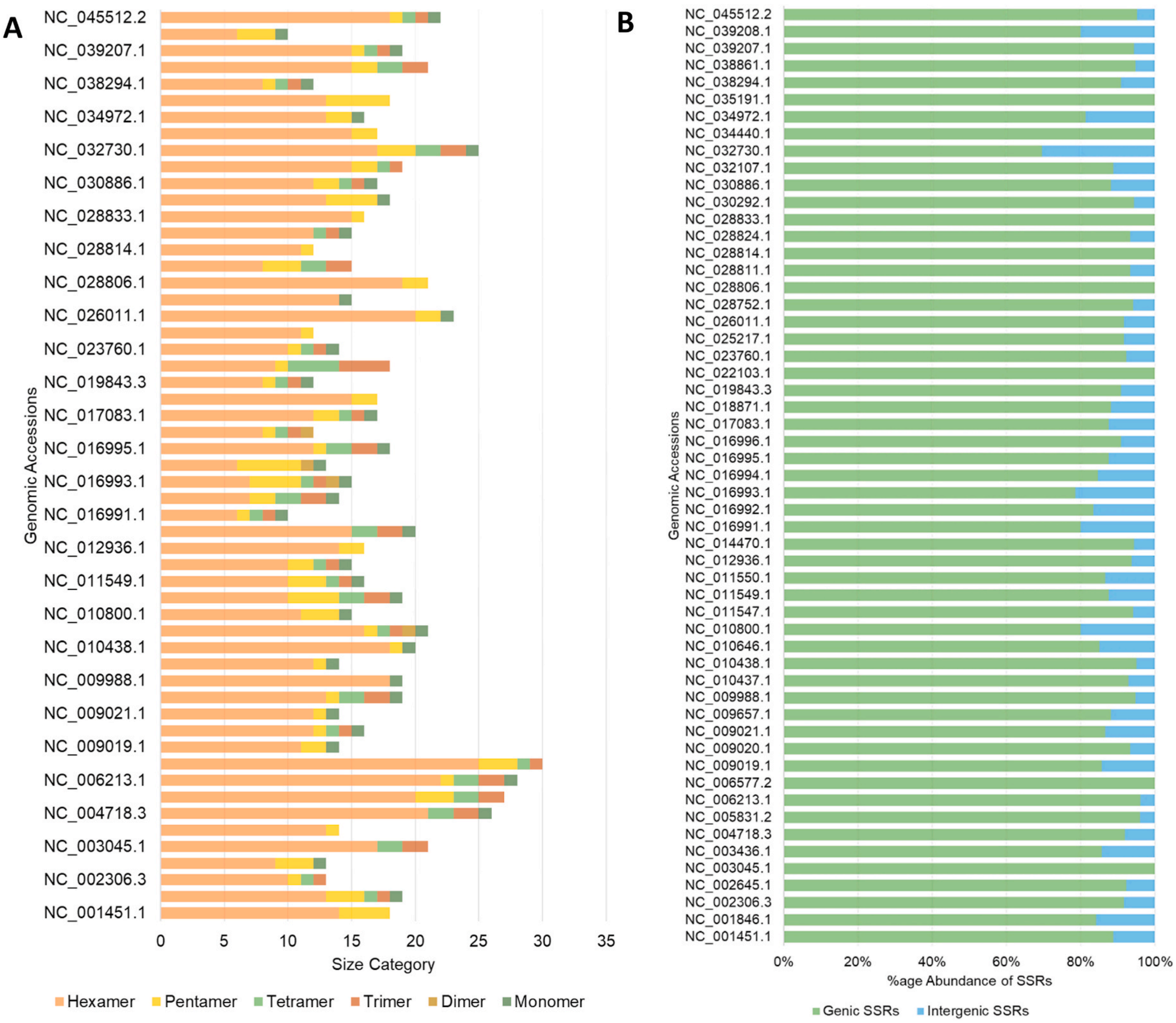


Fig. 5. A) Size base classification of Repeat class motifs shows hexameric motifs enrichment and B) preferential localization of SSRs in genic regions.

repeat in viruses [24]. However, unlike other viruses, the *Coronaviridae* genomes have a different genome architecture. We, therefore, suspected the non-random distribution of SSRs across the genomes as pointed out by several studies [18]. To test our hypothesis, we investigated if there was a significant bias of harbouring SSRs in genic regions. To compute the overlap of SSRs and genic regions we wrote a customized shell script undertaking four overlapping scenarios. Interestingly, we found that 99.4% (833/838 genic SSRs) of the SSRs exhibit a 100% overlap with the genic regions. This led us to infer that the genic regions of *Coronaviridae* are populated with majorly hexanucleotide SSR repeats followed by pentamers (Fig. 5A and B). The comparative under-representation of Dimers, Trimers, and Monomers can be explained based on the destabilization and disruptive effect the repeats impart to the coding region. Moreover, a body of evidence suggests mutations in CDS (Coding Sequence) region can potentially disrupt protein function or could lead to protein truncation [22]. Also, CDS are reported to selectively comprise tri- and hexanucleotide SSR motifs, which can lower the incidences of translational frameshift mutations [10,20,25,35]. Besides, SSRs in the CDS region are under strong evolutionary pressure and prefer not to expand to maintain protein stability encoded by the CDS [28].

The current outbreak of COVID-19 has affected 212 countries and several territories across the globe and various reports underline the ongoing evolution of SARS-CoV-2 [3,27,36]. In our analysis of SARS-CoV-2, 13 SSRs were found to occupy ORF1ab region, 3 SSRs in S gene, 2 in ORF3a, 1 in ORF7a and N gene. The details of SSRs and their genomic coordinates can be found in Supplementary Material 1. The variant analysis of 4938 variants revealed a total of 80 variation sites falling within or on borders of SSRs identified in the reference genome of SARS-CoV-2. Out of 80, 34 falls in intergenic SSR (polyA repeat) irrespective of its smaller size than that of ORF1ab which harbours 13 SSRs with 27 mutations. Most of the variants belonged to Evidence class III (pop. Freq. <0.05/0.01). The detail of mutations with their frequencies can be accessed from Supplementary Material 1. Overall, the variations account for 29.14% of the total bases covered by SSRs in SARS-CoV-2. This indicates the genic SSRs are under selection pressure against non-beneficial mutation. Indeed, it has been reported in earlier studies that tandem repeats are common in protein-coding regions thereby facilitating the rapid evolution of proteins [17,30].

SSRs of SARS-CoV-2 comprised repeats from 17 repeat classes. Therefore, we plotted a class-wise tree-map for the observed frequency of variants (Fig. 6). The intergenic repeat class (polyA repeats) harboured the maximum number of variant loci (34) than the genic classes. The polyA produces a PolyU tail in negative-sense viral RNA. The polyuridine sequence is cleaved by EndoU endonuclease which would otherwise activate the host's immune cells. Mutations in the PolyA region of SRAS-CoV-2 can prevent formation secondary structure PolyU makes with other A/G rich domains in negative-sense RNA, which is otherwise recognized by pattern recognition receptors (PRRs) and might confer a selective advantage in immune evasion [15].

The number of SSRs vs Number of Variations recorded in each gene was plotted (Fig. 7A). To see which type of mononucleotides variations

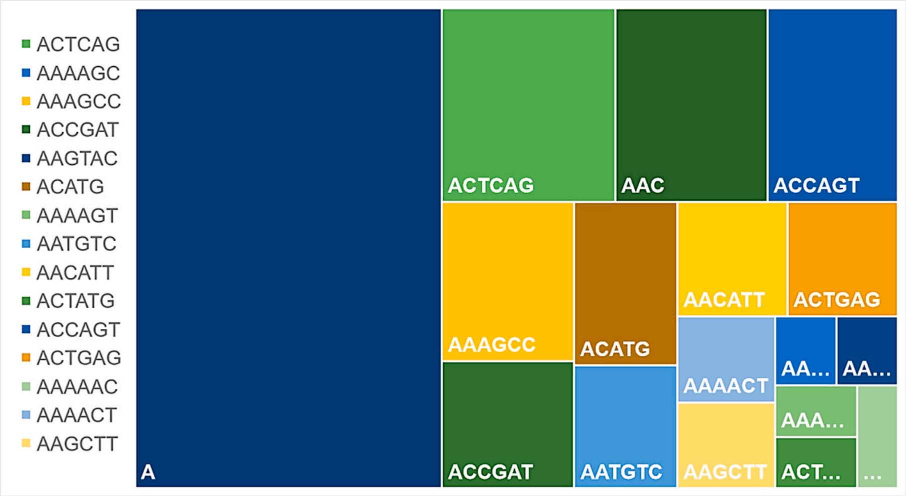


Fig. 6. Class-wise tree-map for the observed frequency of variants. The Box size demarcates the frequency of variants present in each repeat motif class of SSRs in SARS-CoV-2. The polyA monomeric repeat harbours most of the identified genetic variants followed by the 'ACTCAG' class.

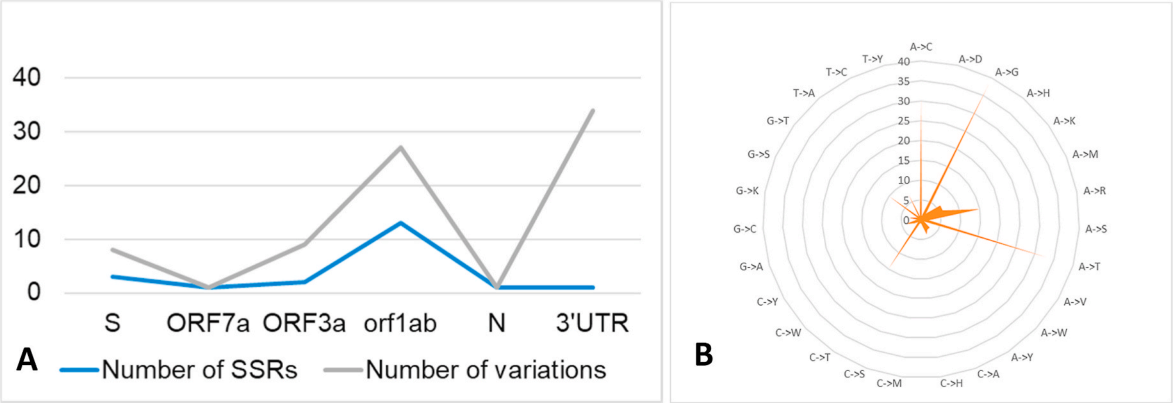


Fig. 7. A) Gene-wise abundances of variations in SAR-CoV-2 genome and B) Radar-plot representing the frequency of mononucleotide variations. We observed A- > G, A- > T & A- > C mutations were more frequent. The eleven "ambiguity" characters were considered separately.

were more prominent; we tried to chart the frequencies of variations in the form of a radar map. We observed A- > G, A- > T & A- > C mutations were more frequent in SSR regions (Fig. 7B). To be transparent, since all these variants belong to Evidence class III as per the limited whole genome set of 11641 high fidelity sequences made available from globally collected samples, and evidence class are subjected to change as more sequences are deposited and analyzed in NGDC.(current analysis identifies intergenic SSR's; polyA repeats (3′UTR region of the genome) to be mutational hotspot comparative to the genic SSRs, see Fig. 7A, followed by ORF1ab). The exoribonuclease (ExoN) coded by coronavirus genomes plays an essential role in high fidelity replication/synthesis of RNA [11]. A study carried out on CoVs lacking ExoN (3′-to-5′ exoribonuclease) showed the accumulation of A- > G and U- > C variations in CoVs viral genomes [4,33]. The mutation in the ExoN coding region (18040.19620 in SARS-CoV-2) might derive the proofreading mechanism haywire leading to progressive accumulation of mutations. This can be offered as a possible explanation that observed A- > G mutation in genic SSRs might be a result of ExoN attenuation due to mutational burden. To check our hypothesis, we revisited the variants dataset to look for mutations in the nsp14 coding region. Surprisingly, we found one high fidelity and three moderate fidelity variations at genomic locations 18060 (C- > T:830; C - > Y:2, synonymous, majorly in the US and Canada), 18877 (C- > T:205, synonymous, majorly in Saudi Arabia and Turkey), 18998 (C- > T:109, missense, majorly in Argentina and US), 18736 (T- > C:88, missense, majorly in New Zealand and Australia). The other possible explanation could be the SARS-CoV-2 RdRp coding domain (also named nsp12) which is a key player of the replication/-transcription machinery. A recent work [26] also highlights the RdRp domain to be a mutational hotspot that can drive replication machinery haywire causing mistakes. The variant dataset, besides, identifies two class I and one class II mutations observed at 14408 (C- > T:4767; C- > Y:5, missense, a hotspot in most of the countries), 14805 (C- > T:636; C - > Y:6, synonymous, majorly in Qatar and Chile), and 15324 (C- > T:234; C - > Y:1, synonymous, majorly in Congo).

To facilitate further research into the SSR repertoire in Coronaviridae members, we identified primers for the SARS-related coronavirus, MERS-related virus, and Human coronavirus OC43 using Primer 3 which are provided in Supplementary Material 2 and can be validated using SSR-PCR [1]. For a few sequences surrounded by AT-rich regions and mononucleotide (A) repeats (polyA tail), the primer designing couldn't be achieved and therefore must be orphaned. Genes harbouring more than one SSR lying juxtaposing to each other and hence having similar primers are demarcated as "Common Primers" [1]. The complete set of SSRs identified and the customized scripts for Batch primer identification are available upon request.

4. Conclusion

The present study screened 55 genomes of *Coronaviridae* family for the incidences, abundances, and composition of microsatellites. The informatic analysis revealed that the SSRs incidences and density were independent of the genome sizes of *Coronaviridae* members. We observe an overall moderate positive correlation between genomic GC% and SSRs GC%. A strong positive correlation in GC percentages was observed when the genomes and SSRs were grouped at the genus level. Our preliminary findings suggest the dearth rather than the complete absence of SSRs in *Coronaviridae* genomes. The underrepresentation of SSRs in *Coronaviridae* genomes can come as an additional explanation for progressively slowing of the nonsynonymous mutation rates in the SARS 2003 outbreak and current SARS-CoV-2 outbreak besides other reasons such as the role of 3′ exonuclease (ExoN) in proofreading activity during replication [16]. The SSRs were found to populate preferentially the genic regions of the genomes analyzed and are predominantly hexameric repeat motifs. Our study highlights SSRs to be present in ORF1ab (nsp3, nsp4, nsp5A_3CLpro and nsp5B_3CLpro, nsp6, nsp10, nsp12, nsp13, & nsp15 domains), S, ORF3a, ORF7a, N &3′ UTR regions of SARS-CoV-2 and harbors multiple mutations (3′UTR and ORF1ab SSRs harboring major number of variants). Though limited in SARS-CoV-2 and other *Coronaviridae* genomes, SSRs have the potential to become mutational hotspots (given to their well-known reputation as hypermutable regions) [23] as the virus explores genotypic space and evolves to find beneficial mutations [9]. However, in-vitro and in-vivo studies are further required for detailed investigation of the role of SSRs in viral genomes of coronaviruses in terms of pathogenesis, evolution and immune evasion.

Author contributions

RS and NKJ designed and wrote the manuscript. RK, SKJ, AS, DK, PN, JR, KKK and MAK analyzed, coordinated and drafted the manuscript. All authors read and approve the final manuscript.

Funding

There are no relevant funding sources to report.

CRediT authorship contribution statement

Rohit Satyam: Conceptualization, Methodology, Software. **Niraj Kumar Jha:** Supervision. **Rohan Kar:** Writing - review & editing. **Saurabh Kumar Jha:** Data curation. **Ankur Sharma:** Software, Validation. **Dhruv Kumar:** Software, Validation. **Parma Nand:** Software, Validation. **Janne Ruokolainen:** Writing - review & editing. **Kavindra Kumar Kesari:** Writing - review & editing. **Mohammad Amjad Kamal:** 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cbi.2020.109226.

References

[1] M.A.M. Atia, G.H. Osman, W.H. Elmenofy, Genome-wide in silico analysis, characterization and identification of microsatellites in spodoptera littoralis multiple nucleopolyhedrovirus (SpliMNPV), Sci. Rep. (2016), https://doi.org/10.1038/srep33741.

[2] A.K. Avvaru, D.T. Sowpati, R.K. Mishra, PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/btx721.

[3] R. Cagliani, D. Forni, M. Clerici, M. Sironi, Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2, J. Virol. (2020), https://doi.org/10.1128/jvi.00411-20.

[4] Y. Chen, H. Cai, J. Pan, N. Xiang, P. Tien, T. Ahola, D. Guo, Functional Screen Reveals SARS Coronavirus Nonstructural Protein Nsp14 as a Novel Cap N7 Methyltransferase, Proceedings of the National Academy of Sciences of the United States of America, 2009, https://doi.org/10.1073/pnas.0808790106.

[5] J. Cui, F. Li, Z.L. Shi, Origin and evolution of pathogenic coronaviruses, in: Nature Reviews Microbiology, 2019, https://doi.org/10.1038/s41579-018-0118-9.

[6] E. De Wit, N. Van Doremalen, D. Falzarano, V.J. Munster, SARS and MERS: recent insights into emerging coronaviruses, in: Nature Reviews Microbiology, 2016, https://doi.org/10.1038/nrmicro.2016.81.

[7] F. Di Giallonardo, T.E. Schlub, M. Shi, E.C. Holmes, Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species, J. Virol. (2017), https://doi.org/10.1128/jvi.02381-16.

[8] J.F. Drexler, V.M. Corman, C. Drosten, Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS, in: Antiviral Research, 2014, https://doi.org/10.1016/j.antiviral.2013.10.013.

[9] S.F. Elena, P. Carrasco, J.A. Daròs, R. Sanjuán, Mechanisms of genetic robustness in RNA viruses, in: EMBO Reports, 2006, https://doi.org/10.1038/sj.embor.7400636.

[10] S. Fujimori, T. Washio, K. Higo, Y. Ohtomo, K. Murakami, K. Matsubara, J. Kawai, P. Carninci, Y. Hayashizaki, S. Kikuchi, M. Tomita, A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription, FEBS (Fed. Eur. Biochem. Soc.) Lett. (2003), https://doi.org/10.1016/S0014-5793(03)01041-X.

[11] S.Y. Fung, K.S. Yuen, Z.W. Ye, C.P. Chan, D.Y. Jin, A tug-of-war between severe acute respiratory syndrome coronavirus 2 and host antiviral defence: lessons from other pathogenic viruses, in: Emerging Microbes and Infections, 2020, https://doi.org/10.1080/22221751.2020.1736644.

[12] A. Grifoni, J. Sidney, Y. Zhang, R.H. Scheuermann, B. Peters, A. Sette, A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2, Cell Host Microbe (2020), https://doi.org/10.1016/j.chom.2020.03.002.

[13] Y.R. Guo, Q.D. Cao, Z.S. Hong, Y.Y. Tan, S.D. Chen, H.J. Jin, K. Sen Tan, D. Y. Wang, Y. Yan, The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak- A n update on the status, in: Military Medical Research, 2020, https://doi.org/10.1186/s40779-020-00240-0.

[14] L. H, H. B, W. A, F. T, R. J, H. N, M. G, A. G, D. R, The sequence alignment/map format and SAMtools, Bioinformatics (2009).

[15] M. Hackbart, X. Deng, S.C. Baker, Coronavirus Endoribonuclease Targets Viral Polyuridine Sequences to Evade Activating Host Sensors, Proceedings of the National Academy of Sciences of the United States of America, 2020, https://doi.org/10.1073/pnas.1921485117.

[16] J.F. He, G.W. Peng, J. Min, D.W. Yu, W.J. Liang, S.Y. Zhang, R.H. Xu, H.Y. Zheng, X.W. Wu, J. Xu, Z.H. Wang, L. Fang, X. Zhang, H. Li, X.G. Yan, J.H. Lu, Z.H. Hu, J. C. Huang, Z.Y. Wan, Y.M.D. Lo, Molecular evolution of the SARS coronavirus, during the course of the SARS epidemic in China, Science (2004), https://doi.org/10.1126/science.1092002.

[17] M. Huntley, G.B. Golding, Evolution of simple sequence in proteins, J. Mol. Evol. (2000), https://doi.org/10.1007/s002390010073.

[18] M.V. Katti, P.K. Ranjekar, V.S. Gupta, Differential distribution of simple sequence repeats in eukaryotic genome sequences, Mol. Biol. Evol. (2001), https://doi.org/10.1093/oxfordjournals.molbev.a003903.

[19] D. Kiselev, A. Matsvay, I. Abramov, V. Dedkov, G. Shipulin, K. Khafizov, Current trends in diagnostics of viral infections of unknown etiology, in: Viruses, 2020, https://doi.org/10.3390/v12020211.

[20] B. Li, Q. Xia, C. Lu, Z. Zhou, Z. Xiang, Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes, Dev. Reprod. Biol. (2004), https://doi.org/10.1016/S1672-0229(04)02004-2. /Beijing Genomics Institute.

[21] Y.C. Li, A.B. Korol, T. Fahima, A. Beiles, E. Nevo, Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review, in: Molecular Ecology, 2002, https://doi.org/10.1046/j.1365-294X.2002.01643.x.

[22] Y.C. Li, A.B. Korol, T. Fahima, E. Nevo, Microsatellites within genes: structure, function, and evolution, in: Molecular Biology and Evolution, 2004, https://doi.org/10.1093/molbev/msh073.

[23] W.H. Lin, E. Kussell, Evolutionary pressures on simple sequence repeats in prokaryotic coding regions, Nucleic Acids Res. (2012), https://doi.org/10.1093/nar/gkr1078.

[24] C. Mashhood Alam, C. Sharfuddin, S. Ali, Analysis of simple and imperfect microsatellites in ebolavirus species and other genomes of filoviridae family, Gene, Cell and Tissue (2015), https://doi.org/10.17795/gct-26404.

[25] D. Metzgar, J. Bytof, C. Wills, Selection against frameshift mutations limits microsatellite expansion in coding DNA, Genome Res. (2000), https://doi.org/10.1101/gr.10.1.72.

[26] M. Pachetti, B. Marini, F. Benedetti, F. Giudici, E. Mauro, P. Storici, C. Masciovecchio, S. Angeletti, M. Ciccozzi, R.C. Gallo, D. Zella, R. Ippodrino, Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant, J. Transl. Med. (2020), https://doi.org/10.1186/s12967-020-02344-6.

[27] T. Phan, Genetic diversity and evolution of SARS-CoV-2, Infect. Genet. Evol. (2020), https://doi.org/10.1016/j.meegid.2020.104260.

[28] W.H. Qi, C. chao Yan, W.J. Li, X.M. Jiang, G.Z. Li, X.Y. Zhang, T.Z. Hu, J. Li, B. S. Yu, Distinct patterns of simple sequence repeats and GC distribution in intragenic and intergenic regions of primate genomes, Aging (2016), https://doi.org/10.18632/aging.101025.

[29] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics (2010), https://doi.org/10.1093/bioinformatics/btq033.

[30] P.A. Romero, F.H. Arnold, Exploring protein fitness landscapes by directed evolution, in: Nature Reviews Molecular Cell Biology, 2009, https://doi.org/10.1038/nrm2805.

[31] W. Shen, S. Le, Y. Li, F. Hu, SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation, PloS One (2016), https://doi.org/10.1371/journal.pone.0163962.

[32] M.A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique, COVID-19 infection: origin, transmission, and characteristics of human coronaviruses, J. Adv. Res. (2020), https://doi.org/10.1016/j.jare.2020.03.005.

[33] E.C. Smith, H. Blanc, M. Vignuzzi, M.R. Denison, Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics, PLoS Pathog. (2013), https://doi.org/10.1371/journal.ppat.1003565.

[34] S. Srivastava, A.K. Avvaru, D.T. Sowpati, R.K. Mishra, Patterns of microsatellite distribution across eukaryotic genomes, BMC Genom. (2019), https://doi.org/10.1186/s12864-019-5516-5.

[35] S. Subramanian, R.K. Mishra, L. Singh, Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions, Genome Biol. (2003), https://doi.org/10.1186/gb-2003-4-2-r13.

[36] X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian, J. Cui, J. Lu, On the origin and continuing evolution of SARS-CoV-2, Nat. Sci. Rev. (2020), https://doi.org/10.1093/nsr/nwaa036.

[37] T. Tsykun, C. Rellstab, C. Dutech, G. Sipos, S. Prospero, Comparative assessment of SSR and SNP markers for inferring the population genetic structure of the common fungus Armillaria cepistipes, Heredity (2017), https://doi.org/10.1038/hdy.2017.48.

[38] F.C. Victoria, L.C. da Maia, A.C. de Oliveira, In silico comparative analysis of SSR markers in plants, BMC Plant Biol. (2011), https://doi.org/10.1186/1471-2229-11-15.

[39] M.L.C. Vieira, L. Santini, A.L. Diniz, C. de F. Munhoz, Microsatellite markers: what they mean and why they are so useful, in: Genetics and Molecular Biology, 2016, https://doi.org/10.1590/1678-4685-GMB-2016-0027.

[40] D. Vijaykrishna, G.J.D. Smith, J.X. Zhang, J.S.M. Peiris, H. Chen, Y. Guan, Evolutionary insights into the ecology of coronaviruses, J. Virol. (2007), https://doi.org/10.1128/jvi.02605-06.

[41] S.R. Weiss, S. Navas-Martin, Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus, Microbiol. Mol. Biol. Rev. (2005), https://doi.org/10.1128/mmbr.69.4.635-664.2005.

[42] W. Xing, G. Hejblum, G.M. Leung, A.J. Valleron, Anatomy of the epidemiological literature on the 2003 SARS outbreaks in Hong Kong and Toronto: a time-stratified review, in: PLoS Medicine, 2010, https://doi.org/10.1371/journal.pmed.1000272.

[43] D. Yang, J.L. Leibowitz, The structure and functions of coronavirus genomic 3' and 5' ends, in: Virus Research, 2015, https://doi.org/10.1016/j.virusres.2015.02.025.

[44] X. Zhao, Y. Tian, R. Yang, H. Feng, Q. Ouyang, Y. Tian, Z. Tan, M. Li, Y. Niu, J. Jiang, G. Shen, R. Yu, Coevolution between simple sequence repeats (SSRs) and virus genome size, BMC Genom. (2012), https://doi.org/10.1186/1471-2164-13-435.