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.

Gene Reports 23 (2021) 101020

Contents lists available at ScienceDirect

Gene Reports

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

Genome-wide *in silico* identification and characterization of Simple Sequence Repeats in diverse completed SARS-CoV-2 genomes

Rasel Siddiqe, Ajit Ghosh *

Department of Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh

ARTICLE INFO

Keywords:
Microsatellite
SARS-CoV-2 virus
Simple sequence repeat
Genome sequence
Comparative genomics

ABSTRACT

Simple sequence repeats (SSRs) or, Microsatellites are short repeat sequences that have been extensively studied in eukaryotic (plants) and prokaryotic (bacteria) organisms. Compared to other organisms, the presence and incidence of SSR on viral genomes are less studied. With the emergence of novel infectious viruses over the past few decades, it is imperative to study the genetic diversity in such viruses to predict their evolutionary and functional changes over time. Following the emergence of SARS-CoV-2, we have assembled 121 complete genomes reported from 31 countries across the six continents for the identification and characterization of SSR repeats. Using two independent SSR identification tools, we have found remarkable consistency in the diversity of microsatellites pattern (38–42 per genome) found in the 121 analyzed SARS-CoV-2 genomes indication their important role for genome stability. Among the identified motifs, trinucleotide and hexanucleotide repeats were found to be the most abundant form followed by mono- and di-nucleotide. There were no tetra- or penta-nucleotide repeats in the analyzed SARS-CoV-2 genomes. The discovery of microsatellites in SARS-CoV-2 genomes may become useful for the population genetics, evolutionary analysis, strain identification and genetic variation.

1. Introduction

Coronavirus disease 2019 (COVID-19) is an acute respiratory infectious disease caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It belongs to the subfamily Coronavirinae of the family Coronaviridae of the order Nidovirales and genus Betacoronavirus (Saha et al., 2020; Weiss and Leibowitz, 2011). According to the serotype and genomic characteristics, coronaviruses could be divided into four major genera that include alpha and beta causing infection primarily to mammals, and gamma and delta forms predominantly infect birds (Tang et al., 2015). Coronaviruses are enveloped, unsegmented single positive-stranded RNA virus with a genomic length varying from 26 to 32 kilobases (Wang et al., 2020). Genome of SARS-CoV-2 possesses 14 ORFs which codes for 27 proteins (Wu et al., 2020). In the recent years, there are three large scale epidemic outbreaks of coronaviruses include SARS-CoV of 2003, MERS-CoV of 2012 and SARS-CoV-2 of 2019 (Khan et al., 2020; Zhou et al., 2020). COVID-19 was initially reported from China but spread all over the world rapidly (Guo et al., 2020). The total number of COVID-19 cases diagnosed so far exceeds 63 million worldwide as on 30th November 2020 with a total death of more than 1.4 million (https://www.worldometers.info/coronavirus/).

SARS-CoV-2 has caused a state of alarm across the world due to its high infection rate and mortality among the elderly and immune-deficient individuals. Due to very limited knowledge of this novel virus, high rate of transmission has occured to all the age groups and diverse demographics population. Thus, the study of genome sequence and comparative genomics has attracted much attention. Moreover, the advancements in sequencing technologies and analysis tools boost-up the process at an unprecedented speed. The first three novel coronaviruses (GISAID accession ID: EPI_ISL_402119, EPI_ISL_402120 and EPI_ISL_402121) were sequenced from Wuhan (Wu et al., 2020). Currently, over 94,000 SARS-CoV2 viral genomes have already been sequenced and deposited for in the public domain like GenBank database (Benson et al., 2000) and GISAID database (Shu and McCauley, 2017). To understand the molecular genetics, evolutionary genomics and other important features of these viruses, development of a reliable biomarker like SSR could be an excellent tool.

Abbreviations: COVID-19, coronavirus disease 2019; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SSR, simple sequence repeats; RD, relative density; RA, relative abundance; SpliMNPV, Spodoptera littoralis multiple nucleopolyhedrovirus; HCV, hepatitis C virus.

* Corresponding author.
E-mail address: aghosh-bmb@sust.edu (A. Ghosh).

https://doi.org/10.1016/j.genrep.2021.101020
Received 3 October 2020; Received in revised form 6 December 2020; Accepted 29 December 2020
Available online 26 January 2021

Simple sequence repeats (SSRs) are short tandem repeat sequences found across the genomes of all organisms. SSRs are essentially sequences of varying lengths containing repeats of 1–6 nucleotides. There are several characteristics associated with SSR sequences such as they are present ubiquitously in any genome (Li et al., 2004); their accumulation has been associated with the variation in genome size (Gao and Qi, 2007); they could exist in both coding and non-coding sequences (Riley and Krieger, 2009); they are highly variable and polymorphic in nature (Kim et al., 2008). SSRs are found to be associated with the recombination hotspots and random integration. This could be considered as an explanation of the fact that pathogenic organisms use this variability to combat host immune responses (Zhao et al., 2012).

One of the extensive applications of SSR has been considered to use as a genetic marker (Heesacker et al., 2008; Temnykh et al., 2001). A few notable results have also been found using SSR in genome mapping, along with ecological and evolutionary biology. Although several independent studies have focused on SSR in viral genomes, a distinct distribution pattern is yet to be established (Chen et al., 2011). Viral SSRs are capable of generating genomic diversity that in turn manifest phenotypic changes (Li et al., 2004). Genome features including length and GC content largely influence their occurrence (Dieringer and Schlötterer, 2003; Kelkar et al., 2008). Here, we have investigated the distribution, size and GC content variability among 121 SARS-CoV-2 genome sequence isolated from different countries and identified the prevalence of SSR markers.

2. Methods and materials

2.1. Genome sequence collection

Complete genome sequences of SARS-CoV-2 (121) were acquired from the NCBI Virus database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/). Sequences were collected from 31 countries (Table S1) and selected according to the date of data deposition ranging from early January 2020 to late June 2020. The sequence data were processed in FASTA format.

2.2. Simple sequence repeat identification

Two SSR identification tools were used in the study. First, Simple Sequence Repeat Identification Tool (SSRIT, https://archive.gramene.org/db/markers/ssrtool) was used to detect perfect SSR motifs in the given sequences at FASTA format. The minimum number of repeats was set to 5 for dimers, 3 for Trimeric, Tetrameric and Pentameric repeats: and 2 for hexameric repeats. Thus, the resulting configuration is 5-3-3-3-2 for the minimum number of repeats.

As SSRIT cannot detect monomeric repeats, we have employed a second tool IMEx-web: Imperfect Microsatellite Extraction Webserver (http://43.227.129.132:8008/IMEX/). IMEx advanced mode was used to identify the perfect microsatellites in the complete genomic sequences. Minimum repeat numbers for monomers was set to 10; 5 for dimers; 3 for Trimer, Tetramer and Pentamer repeats; and 2 for Hexameric repeats. The resulting configuration was 10-5-3-3-3-2 for the minimum number of repeats.

2.3. Calculation of relative density (RD) and relative abundance (RA)

To accurately assess the significance of SSR in a genome, the Relative Density (RD) and Relative Abundance (RA) of the matrix has been calculated using the following equation.

$$\text{Relative Density (RD)} = \text{Total length in repeats (bp)}/\text{Genome size (bp)}*1000$$

$$\text{Relative Abundance (RA)} = \text{Number of repeats}/\text{Genome size(bp)}*1000$$

2.4. Programming

Python - Programming Language (IDLE 5.8.2) was used to manage and keep track of data collected in CSV format. Subsequently, all of the data analysis of various repeats in individual sequences was carried out using Python. Lastly, Python Module, Matplotlib, PyPlot was used to generate the bar charts.

2.5. Statistical analysis

Correlation between total relative abundance, relative density against genome size was established using Microsoft Excel 2016.

3. Results

3.1. Collection and distribution of SARS-Cov-2 genome sequences

We analyzed the presence of perfect SSRs over 6 bp long, from a pool of 121 completely sequenced SARS-Cov-2 genomes, with an average size of 29,855 bases ranging from 29,574 to 29,945 bases. All these sequences were sampled from 31 different countries over 6 continents (Table 1). A maximum number of 12 sequences were taken from china, while minimum one genome sequence was taken from Nepal, Turkey, Sweden, Peru, Ukraine, and South Africa to make sure the presence of diversity. The list of genomic sequences including their accession number, size, attributed region and GC content are summarized in Table 1.

3.2. Incident frequency of SSRs

Incident frequency of SSRs in the 121 genomes varied at a negligible level (Fig. 1) regardless of regional variation or SSR search tool specialization. No tetrameric or pentameric repeats were observed in any of the sequences. Both IMEx and SSRIT provided almost identical data for di, tri, and hexameric repeats with few exceptions. The total number of SSR found in each sequence ranged between 38−42 and 38–41 repeats with monomeric repeats detected by only IMEx or without monomeric repeats as detected by SSRIT, respectively. Thus, the total number of repeats mainly varied in the sequences having monomeric repeats detected by IMEx. Sequences such as S2 (MT635672) show an equal number of repeats from both IMEx and SSRIT which doesn't contain any monomeric repeats (Fig. 1). The average number of trimeric repeats is ~20 (19.95041322) with the highest value being 20 and the lowest is 18. The average number of hexameric repeats was ~18 (17.97520661), with the highest value of 18 and the lowest value of 16. Almost all of the sequences contained 2 dimeric, 20 trimeric and 18 hexameric repeats except 4 sequences MT635672 (S2), MT502774 (S6), MT372482 (S30), MT451783 (S74) which had lesser number of trimeric repeats and three including MT372482 (S30), MT039890 (S49), MT447176 (S66) had lesser number of hexameric repeats (Figs. 2 and 3).

3.3. Calculation of RA and RD

Relative abundance (RA) and Relative density (RD) of SSR was calculated as the number of repeats per kilobase pair (kb) and total length in repeats per kb, respectively (Figs. 2 and 3). Relative abundance was calculated for each type of repeats (i.e: monomeric, dimeric, trimeric, hexameric denoted by RA1, RA2, RA3 and RA6) as well as for the total number of repeats in a sequence (Tables 2 and 3). All the identified SSR repeats from IMEx and SSRIT tools were analyzed with little variation among all the 121 genome sequences. Similarly, relative density (RD) was calculated as the total length of repeats divided by the genome size in kb for all the repeats detected by both IMEx and SSRIT tools. There is more variation in RD values using IMEx analyzed SSRs due to the inconsistency of monomeric repeats (Fig. 2B). The highest value of total RA and RD from the IMEx tool is 1.42 and 14.89; while the

Table 1
List of analyzed completed SARS-CoV-2 genomes along with their attributed regions, genome size and G/C content.

No	Accession	Size (bp)	Country	G/C content	No	Accession	Size (bp)	Country	G/C content	No	Accession	Size (bp)	Country	G/C content
S1	MT476385	29,902	BGD	37.96	S41	MT499208	29,873	POL	37.99	S81	MT121215	29,945	CHN	37.91
S2	MT635672	29,832	BGD	37.99	S42	MT499209	29,903	POL	37.95	S82	MN938384	29,838	CHN	38.02
S3	MT607246	29,903	BGD	37.95	S43	MT499210	29,899	POL	37.94	S83	MT259229	29,864	CHN	38.01
S4	MT577359	29,816	BGD	38.01	S44	MT450872	29,782	SRB	38.01	S84	MT259230	29,866	CHN	38.01
S5	MT539160	29,758	BGD	38.01	S45	MT459979	29,782	SRB	38.01	S85	MT446312	29,879	CHN	37.99
S6	MT502774	29,859	BGD	38.01	S46	MT324062	29,903	ZAF	37.96	S86	MT123290	29,891	CHN	38.00
S7	MT126808	29,876	BRA	38.00	S47	MT304475	29,882	KOR	37.98	S87	MT281577	29,903	CHN	37.97
S8	MT350282	29,903	BRA	37.96	S48	MT304474	29,882	KOR	37.98	S88	MT470176	29,903	FRA	37.96
S9	MT256924	29,782	COL	38.01	S49	MT039890	29,903	KOR	37.96	S89	MT470177	29,903	FRA	37.97
S10	MT470219	29,903	COL	37.96	S50	MT292571	29,782	ESP	38.01	S90	MT470178	29,903	FRA	37.96
S11	MT371568	29,740	CZE	37.87	S51	MT292574	29,782	ESP	38.00	S91	MT470179	29,903	FRA	37.96
S12	MT371572	29,756	CZE	38.00	S52	MT292569	29,782	ESP	38.02	S92	MT320538	29,882	FRA	37.99
S13	MT371573	29,756	CZE	38.00	S53	MT359865	29,890	ESP	37.98	S93	MT459847	29,812	GRC	38.01
S14	MT358641	29,903	DEU	37.97	S54	MT371047	29,903	LKA	37.96	S94	MT459924	29,818	GRC	38.01
S15	MT318827	29,870	DEU	38.00	S55	MT371048	29,903	LKA	37.96	S95	MT459899	29,818	GRC	38.00
S16	MT358642	29,903	DEU	37.96	S56	MT371050	29,903	LKA	37.97	S96	MT459897	29,818	GRC	38.01
S17	MT358638	29,903	DEU	37.97	S57	MT093571	29,886	SWE	38.00	S97	MT459867	29,818	GRC	38.01
S18	MT459985	29,903	GUM	37.95	S58	MT374114	29,901	TWN	37.96	S98	MT459862	29,812	GRC	38.01
S19	MT459986	29,903	GUM	37.96	S59	MT374102	29,901	TWN	37.97	S99	MT270814	29,764	HKG	38.02
S20	MT459987	29,890	GUM	37.96	S60	MT370516	29,900	TWN	37.97	S100	MT215195	29,764	HKG	38.03
S21	MT320891	29,822	IRN	38.00	S61	MT066176	29,870	TWN	38.01	S101	MT365031	29,891	HKG	37.99
S22	MT447177	29,793	IRN	38.01	S62	MT066175	29,870	TWN	38.01	S102	MT365030	29,891	HKG	37.99
S23	MT276597	29,851	ISR	38.02	S63	MT447155	29,805	THA	38.02	S103	MT114412	29,889	HKG	37.99
S24	MT276598	29,870	ISR	38.00	S64	MT447159	29,834	THA	38.01	S104	MT230904	29,891	HKG	37.98
S25	MT077125	29,785	ITA	38.02	S65	MT447165	29,671	THA	37.97	S105	MT415321	29,903	IND	37.97
S26	MT066156	29,867	ITA	38.01	S66	MT447176	29,840	THA	37.99	S106	MT415320	29,901	IND	37.97
S27	MT428551	29,900	KAZ	37.96	S67	MT327745	29,832	TUR	38.01	S107	MT477885	29,899	IND	37.96
S28	MT428552	29,903	KAZ	37.97	S68	MT466071	29,903	URY	37.97	S108	MT012098	29,854	IND	38.02
S29	MT428553	29,903	KAZ	37.96	S69	MT192772	29,891	VNM	37.98	S109	MT050493	29,851	IND	38.01
S30	MT372482	29,865	MYS	37.64	S70	MT192773	29,890	VNM	37.98	S110	MT467260	29,800	IND	38.01
S31	MT372481	29,898	MYS	37.94	S71	MT007544	29,893	AUS	37.97	S111	MT467253	29,800	IND	37.99
S32	MT372480	29,868	MYS	37.94	S72	MT450935	29,805	AUS	38.02	S112	LC542976	29,903	JPN	37.97
S33	MT072688	29,811	NPL	38.02	S73	MT450932	29,802	AUS	38.02	S113	LC529905	29,903	JPN	37.97
S34	MT396266	29,880	NLD	37.98	S74	MT451783	29,802	AUS	37.73	S114	LC542809	29,903	JPN	37.96
S35	MT457399	29,876	NLD	37.99	S75	MT451755	29,812	AUS	37.94	S115	MT444626	29,840	USA	37.94
S36	MT457396	29,877	NLD	38.00	S76	LR757998	29,866	CHN	37.99	S116	MT380730	29,882	USA	37.98
S37	MT240479	29,836	PAK	37.99	S77	LR757996	29,868	CHN	38.00	S117	MT380731	29,882	USA	37.99
S38	MT262993	29,836	PAK	38.02	S78	MT253710	29,781	CHN	38.02	S118	MT159712	29,882	USA	37.99
S39	MT500122	29,819	PAK	38.02	S79	MT253700	29,781	CHN	38.02	S119	MT159717	29,882	USA	37.99
S40	MT263074	29,856	PER	38.01	S80	MT049951	29,903	CHN	37.97	S120	MN985325	29,882	USA	38.00
										S121	MT326173	29,574	USA	37.95

Country tri-letter code legend in Supplementary Table 3.

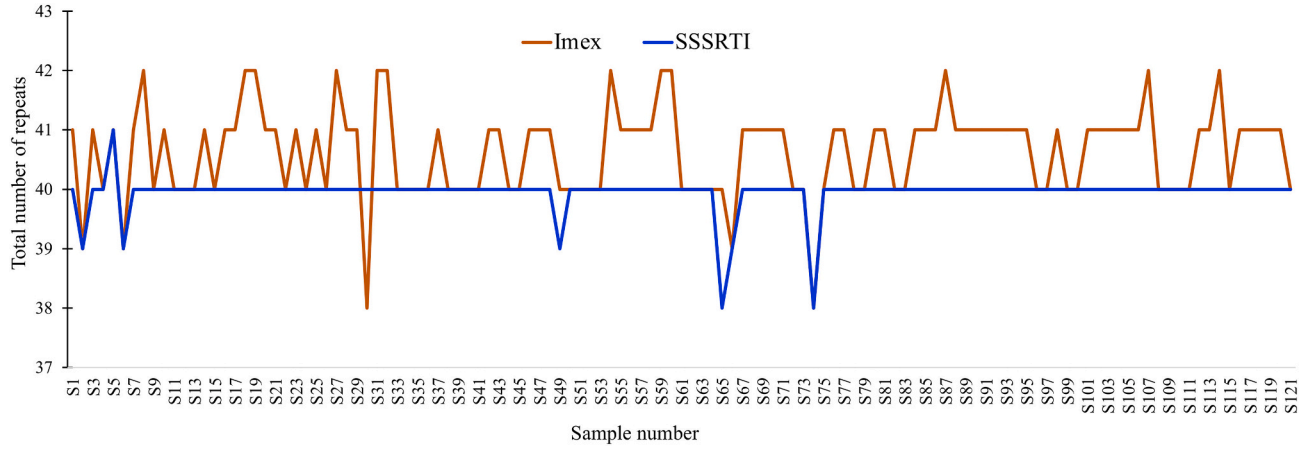

Fig. 1. Comparison of the total number of SSR repeats using IMEx and SSRIT tools. SSRIT tool cannot detect the presence of monomeric repeats in the identified genome, while IMEx can. That creates a variation in the total number of identified SSR motifs and presented in the figure.

lowest value is 13.29 and 1.27, respectively (Table 2). Likewise, the highest value of total RA and RD for SSRIT tool is 1.37 and 14.36; while the lowest is 1.27 and 13.45, respectively (Table 3).

3.4. Motifs types in analyzed genomes

Monomeric repeats from the IMEx tool analysis showed that 50 sequences do not contain any monomeric repeat while remaining 59 have only one and the rest 12 sequences have 2 monomeric repeats. Out of

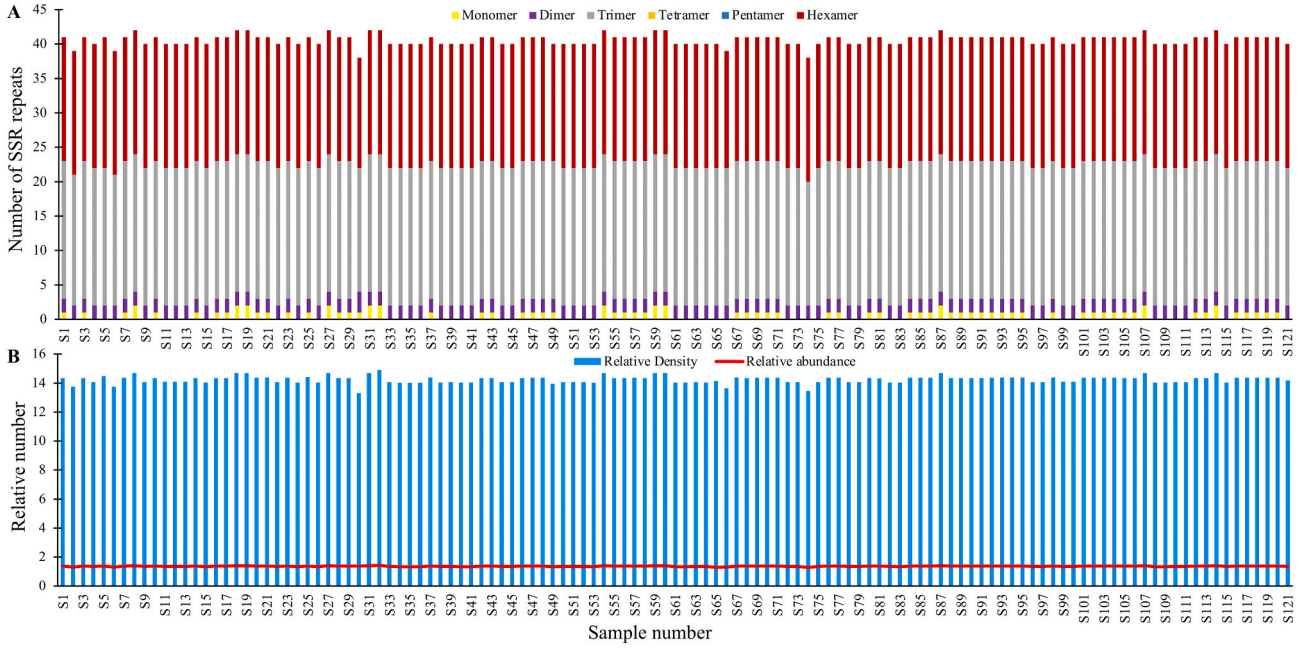

Fig. 2. Analysis of SSRs found in IMEx tool. (A) Analysis of total SSR per genome and (B) relative density and abundance of the identified SSR repeats present.

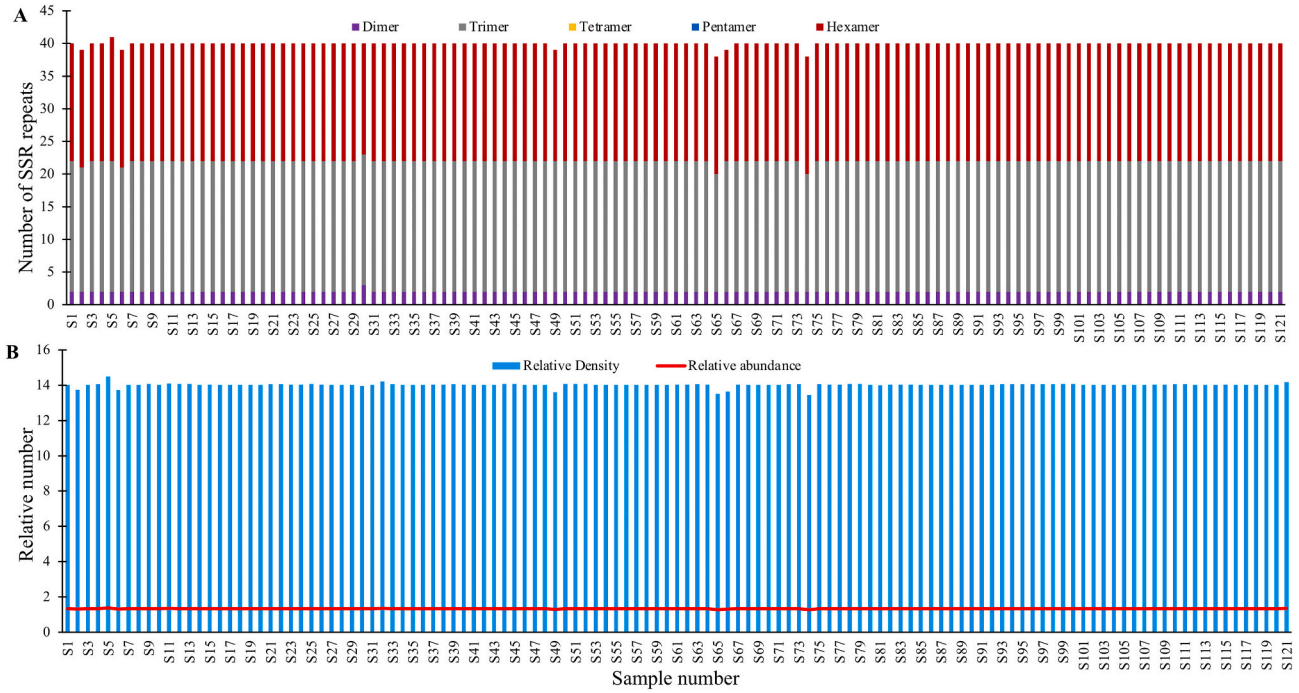

Fig. 3. Analysis of SSRs found in SSRIT tool. (A) Analysis of total SSR per genome and (B) relative density and abundance of the identified SSR repeats present.

these 59 sequences with only one monomeric repeat, 45 contained $(A)_n$ while the rest 14 contained $(T)_n$ (Table S2). However, two monomeric repeats containing 12 sequences have both $(A)_n$ and $(T)_n$ repeats (Fig. 4). All except one sequence S30 (MT372482) contained two predominant dimeric repeats of $(TC)_n$ and $(GT)_n$ motif. A third dimeric repeat was found only in the sequence S30 (MT372482) which possessed $(AT)_n$. Among the trimeric repeats, motifs $(TTC)_n$ and $(CTT)_n$ occurred twice in all the analyzed sequences (Fig. 4). The occurrence of motifs is counted across all sequences. For instance, if a motif is repeated twice in each sequence, the total occurrence of the motif is 242 (total number of sequences X2). Motifs $(AAG)_n$ and $(GAA)_n$ were also repeated twice in all of the sequences except S6 (MT502774) and S2 (MT635672) which had $(AAG)_n$ once and S30 (MT372482) which had both $(AAG)_n$ and $(GAA)_n$ once. Motifs $(AGT)_n$ and $(CTG)_n$ were present once in each sequence except sequence S74 (MT451783). Motif $(CAA)_n$ was the only trimer that was repeated four times in a cluster in every sequence, while other trimeric repeats repeated three times. Nineteen different hexametric motifs were identified in the analyzed sequences. Among them, $(TAGTCA)_n$ and $(TACTTG)_n$ was absent in S30 (MT372482); $(GTTTTCT)_n$ and $(GGCTTT)_n$ was missing in S49 (MT039890) and S66 (MT447176). Exceptionally, $(AATAGG)_n$ motif was only found to be present in one sequence S74 (MT539160). All other hexameric repeats

Table 2
Relative abundance (RA) and relative density (RD) in SARS-CoV-2 genome sequence calculated for IMEx derived SSR repeats.

No	RA1	RA2	RA3	RA6	RA	Length	RD	No	RA1	RA2	RA3	RA6	RA	Length	RD
S1	0.033	0.067	0.669	0.602	1.371	429	14.34	S61	0.000	0.067	0.670	0.603	1.339	419	14.02
S2	0.000	0.067	0.637	0.603	1.307	410	13.74	S62	0.000	0.067	0.670	0.603	1.339	419	14.02
S3	0.033	0.067	0.669	0.602	1.371	429	14.34	S63	0.000	0.067	0.671	0.604	1.342	419	14.05
S4	0.000	0.067	0.671	0.604	1.342	419	14.05	S64	0.000	0.067	0.670	0.603	1.341	419	14.04
S5	0.000	0.067	0.672	0.638	1.378	431	14.48	S65	0.000	0.067	0.674	0.607	1.281	419	14.12
S6	0.000	0.067	0.636	0.603	1.306	410	13.73	S66	0.000	0.067	0.670	0.570	1.307	407	13.63
S7	0.033	0.067	0.669	0.602	1.372	429	14.35	S67	0.034	0.067	0.670	0.603	1.374	429	14.38
S8	0.067	0.067	0.669	0.602	1.405	439	14.68	S68	0.033	0.067	0.669	0.602	1.371	429	14.34
S9	0.000	0.067	0.672	0.604	1.343	419	14.06	S69	0.033	0.067	0.669	0.602	1.372	429	14.35
S10	0.033	0.067	0.669	0.602	1.371	429	14.34	S70	0.033	0.067	0.669	0.602	1.372	429	14.35
S11	0.000	0.067	0.672	0.605	1.345	419	14.08	S71	0.033	0.067	0.669	0.602	1.372	429	14.35
S12	0.000	0.067	0.672	0.605	1.344	419	14.08	S72	0.000	0.067	0.671	0.604	1.342	419	14.05
S13	0.000	0.067	0.672	0.605	1.344	419	14.08	S73	0.000	0.067	0.671	0.604	1.342	419	14.05
S14	0.033	0.067	0.669	0.602	1.371	429	14.34	S74	0.000	0.067	0.604	0.604	1.275	401	13.45
S15	0.000	0.067	0.670	0.603	1.339	419	14.02	S75	0.000	0.067	0.671	0.604	1.342	419	14.05
S16	0.033	0.067	0.669	0.602	1.371	429	14.34	S76	0.033	0.067	0.670	0.603	1.373	429	14.36
S17	0.033	0.067	0.669	0.602	1.371	429	14.34	S77	0.033	0.067	0.670	0.603	1.373	429	14.36
S18	0.067	0.067	0.669	0.602	1.405	439	14.68	S78	0.000	0.067	0.672	0.604	1.343	419	14.06
S19	0.067	0.067	0.669	0.602	1.405	439	14.68	S79	0.000	0.067	0.672	0.604	1.343	419	14.06
S20	0.033	0.067	0.669	0.602	1.372	429	14.35	S80	0.033	0.067	0.669	0.602	1.371	429	14.34
S21	0.034	0.067	0.671	0.604	1.375	429	14.38	S81	0.033	0.067	0.668	0.601	1.369	429	14.32
S22	0.000	0.067	0.671	0.604	1.343	419	14.06	S82	0.000	0.067	0.670	0.603	1.341	419	14.04
S23	0.033	0.067	0.670	0.603	1.373	429	14.37	S83	0.000	0.067	0.670	0.603	1.339	419	14.03
S24	0.000	0.067	0.670	0.603	1.339	419	14.02	S84	0.033	0.067	0.670	0.603	1.373	429	14.36
S25	0.034	0.067	0.671	0.604	1.377	429	14.40	S85	0.033	0.067	0.669	0.602	1.372	429	14.35
S26	0.000	0.067	0.670	0.603	1.339	419	14.02	S86	0.033	0.067	0.669	0.602	1.372	429	14.35
S27	0.067	0.067	0.669	0.602	1.405	439	14.68	S87	0.067	0.067	0.669	0.602	1.405	439	14.68
S28	0.033	0.067	0.669	0.602	1.371	429	14.34	S88	0.033	0.067	0.669	0.602	1.371	429	14.34
S29	0.033	0.067	0.669	0.602	1.371	429	14.34	8S9	0.033	0.067	0.669	0.602	1.371	429	14.34
S30	0.033	0.100	0.603	0.536	1.373	397	13.29	S90	0.033	0.067	0.669	0.602	1.371	429	14.34
S31	0.067	0.067	0.669	0.602	1.405	439	14.68	S91	0.033	0.067	0.669	0.602	1.371	429	14.34
S32	0.068	0.068	0.678	0.611	1.425	439	14.89	S92	0.033	0.067	0.669	0.602	1.372	429	14.35
S33	0.000	0.067	0.671	0.604	1.342	419	14.05	S93	0.034	0.067	0.671	0.604	1.375	429	14.39
S34	0.000	0.067	0.669	0.602	1.339	419	14.02	S94	0.034	0.067	0.671	0.604	1.375	429	14.38
S35	0.000	0.067	0.669	0.602	1.339	419	14.02	S95	0.034	0.067	0.671	0.604	1.375	429	14.38
S36	0.000	0.067	0.669	0.602	1.339	419	14.02	S96	0.000	0.067	0.671	0.604	1.341	419	14.05
S37	0.034	0.067	0.670	0.603	1.374	429	14.37	S97	0.000	0.067	0.671	0.604	1.341	419	14.05
S38	0.000	0.067	0.670	0.603	1.341	419	14.04	S98	0.034	0.067	0.671	0.604	1.375	429	14.39
S39	0.000	0.067	0.671	0.604	1.341	419	14.05	S99	0.000	0.067	0.672	0.605	1.344	419	14.07
S40	0.000	0.067	0.670	0.603	1.340	419	14.03	S100	0.000	0.067	0.672	0.605	1.344	419	14.07
S41	0.000	0.067	0.670	0.603	1.339	419	14.02	S101	0.033	0.067	0.669	0.602	1.372	429	14.35
S42	0.033	0.067	0.669	0.602	1.371	429	14.34	S102	0.033	0.067	0.669	0.602	1.372	429	14.35
S43	0.033	0.067	0.669	0.602	1.371	429	14.34	S103	0.033	0.067	0.669	0.602	1.372	429	14.35
S44	0.000	0.067	0.672	0.604	1.343	419	14.06	S104	0.033	0.067	0.669	0.602	1.372	429	14.35
S45	0.000	0.067	0.672	0.604	1.343	419	14.06	S105	0.033	0.067	0.669	0.602	1.371	429	14.34
S46	0.033	0.067	0.669	0.602	1.371	429	14.34	S106	0.033	0.067	0.669	0.602	1.371	429	14.34
S47	0.033	0.067	0.669	0.602	1.372	429	14.35	S107	0.067	0.067	0.669	0.602	1.405	439	14.68
S48	0.033	0.067	0.669	0.602	1.372	429	14.35	S108	0.000	0.067	0.670	0.603	1.340	419	14.03
S49	0.033	0.067	0.669	0.569	1.338	417	13.94	S109	0.000	0.067	0.670	0.603	1.340	419	14.03
S50	0.000	0.067	0.672	0.604	1.343	419	14.06	S110	0.000	0.067	0.671	0.604	1.342	419	14.06
S51	0.000	0.067	0.672	0.604	1.343	419	14.06	S111	0.000	0.067	0.671	0.604	1.342	419	14.06
S52	0.000	0.067	0.672	0.604	1.343	419	14.06	S112	0.033	0.067	0.669	0.602	1.371	429	14.34
S53	0.000	0.067	0.669	0.602	1.338	419	14.01	S113	0.033	0.067	0.669	0.602	1.371	429	14.34
S54	0.067	0.067	0.669	0.602	1.405	439	14.68	S114	0.067	0.067	0.669	0.602	1.405	439	14.68
S55	0.033	0.067	0.669	0.602	1.371	429	14.34	S115	0.000	0.067	0.670	0.603	1.340	419	14.04
S56	0.033	0.067	0.669	0.602	1.371	429	14.34	S116	0.033	0.067	0.669	0.602	1.372	429	14.35
S57	0.033	0.067	0.669	0.602	1.372	429	14.35	S117	0.033	0.067	0.669	0.602	1.372	429	14.35
S58	0.033	0.067	0.669	0.602	1.371	429	14.34	S118	0.033	0.067	0.669	0.602	1.372	429	14.35
S59	0.067	0.067	0.669	0.602	1.405	439	14.68	S119	0.033	0.067	0.669	0.602	1.372	429	14.35
S60	0.067	0.067	0.669	0.602	1.405	439	14.68	S120	0.033	0.067	0.669	0.602	1.372	429	14.35
								S121	0.000	0.068	0.676	0.609	1.353	419	14.16

Relative abundance (RA), RA1 = Monomeric repeats, RA2 = Dimeric repeats, RA3 = Trimeric repeats, RA6 = Hexameric repeats, RD = Relative Density.

were found precisely once in every sequence. These SSR markers were found to be distributed in the ORF1ab, S, ORF3ab, ORF7a, and N regions of the SARS-CoV-2 genome (Fig. 5). Maximum 24 motifs were present in the ORF1ab region, followed by 5 motifs each in S, ORF3ab, and N regions, and only one motif present in ORF7a region.

3.5. Correlation studies

The correlation between genome size/GC content with the value of relative abundance (RA) and relative density (RD) of SSRs was determined. Correlation coefficient of IMEx tool detected SSRs repeats showed a positive correlation with the total RA 0.52 ($R^2 = 0.271$, $P < 0.05$) and RD 0.419 ($R^2 = 0.176$, $P < 0.05$). While that with G/C content is −0.102 ($R^2 = 0.010$, $P > 0.1$) and 0.147 ($R^2 = 0.022$, $P > 0.1$) for RA and RD, respectively. Surprisingly, total RA and RD correlation coefficients obtained from the SSRIT tool correlate negatively with the genome size as −0.0595 ($R^2 = 0.003$, $P > 0.1$) and − 0.107, ($R^2 = 0.011$, $P > 0.1$), respectively. Further analysis suggested that the RA and RD are

Table 3
Relative abundance (RA) and relative density (RD) in SARS-CoV-2 genome sequence calculated for SSRIT derived SSR repeats.

No.	RA2	RA3	RA6	RA	Length	RD	No.	RA2	RA3	RA6	RA	Length	RD
S1	0.067	0.669	0.602	1.338	419	14.01	S61	0.067	0.670	0.603	1.339	419	14.02
S2	0.067	0.637	0.603	1.307	410	13.74	S62	0.067	0.670	0.603	1.339	419	14.02
S3	0.067	0.669	0.602	1.338	419	14.01	S63	0.067	0.671	0.604	1.342	419	14.05
S4	0.067	0.671	0.604	1.342	419	14.05	S64	0.067	0.670	0.603	1.341	419	14.04
S5	0.067	0.672	0.638	1.378	431	14.48	S65	0.067	0.607	0.607	1.281	401	13.51
S6	0.067	0.636	0.603	1.306	410	13.73	S66	0.067	0.670	0.570	1.307	407	13.63
S7	0.067	0.669	0.602	1.339	419	14.02	S67	0.067	0.670	0.603	1.341	419	14.04
S8	0.067	0.669	0.602	1.338	419	14.01	S68	0.067	0.669	0.602	1.338	419	14.01
S9	0.067	0.672	0.604	1.343	419	14.06	S69	0.067	0.669	0.602	1.338	419	14.01
S10	0.067	0.669	0.602	1.338	419	14.01	S70	0.067	0.669	0.602	1.338	419	14.01
S11	0.067	0.672	0.605	1.345	419	14.08	S71	0.067	0.669	0.602	1.338	419	14.01
S12	0.067	0.672	0.605	1.344	419	14.08	S72	0.067	0.671	0.604	1.342	419	14.05
S13	0.067	0.672	0.605	1.344	419	14.08	S73	0.067	0.671	0.604	1.342	419	14.05
S14	0.067	0.669	0.602	1.338	419	14.01	S74	0.067	0.604	0.604	1.275	401	13.45
S15	0.067	0.670	0.603	1.339	419	14.02	S75	0.067	0.671	0.604	1.342	419	14.05
S16	0.067	0.669	0.602	1.338	419	14.01	S76	0.067	0.670	0.603	1.339	419	14.02
S17	0.067	0.669	0.602	1.338	419	14.01	S77	0.067	0.670	0.603	1.339	419	14.02
S18	0.067	0.669	0.602	1.338	419	14.01	S78	0.067	0.672	0.604	1.343	419	14.06
S19	0.067	0.669	0.602	1.338	419	14.01	S79	0.067	0.672	0.604	1.343	419	14.06
S20	0.067	0.669	0.602	1.338	419	14.01	S80	0.067	0.669	0.602	1.338	419	14.01
S21	0.067	0.671	0.604	1.341	419	14.05	S81	0.067	0.668	0.601	1.336	419	13.99
S22	0.067	0.671	0.604	1.343	419	14.06	S82	0.067	0.670	0.603	1.341	419	14.04
S23	0.067	0.670	0.603	1.340	419	14.03	S83	0.067	0.670	0.603	1.339	419	14.03
S24	0.067	0.670	0.603	1.339	419	14.02	S84	0.067	0.670	0.603	1.339	419	14.02
S25	0.067	0.671	0.604	1.343	419	14.06	S85	0.067	0.669	0.602	1.339	419	14.02
S26	0.067	0.670	0.603	1.339	419	14.02	S86	0.067	0.669	0.602	1.338	419	14.01
S27	0.067	0.669	0.602	1.338	419	14.01	S87	0.067	0.669	0.602	1.338	419	14.01
S28	0.067	0.669	0.602	1.338	419	14.01	S88	0.067	0.669	0.602	1.338	419	14.01
S29	0.067	0.669	0.602	1.338	419	14.01	8S9	0.067	0.669	0.602	1.338	419	14.01
S30	0.100	0.670	0.569	1.339	417	13.96	S90	0.067	0.669	0.602	1.338	419	14.01
S31	0.067	0.669	0.602	1.338	419	14.01	S91	0.067	0.669	0.602	1.338	419	14.01
S32	0.068	0.678	0.611	1.357	419	14.21	S92	0.067	0.669	0.602	1.339	419	14.02
S33	0.067	0.671	0.604	1.342	419	14.05	S93	0.067	0.671	0.604	1.342	419	14.05
S34	0.067	0.669	0.602	1.339	419	14.02	S94	0.067	0.671	0.604	1.341	419	14.05
S35	0.067	0.669	0.602	1.339	419	14.02	S95	0.067	0.671	0.604	1.341	419	14.05
S36	0.067	0.669	0.602	1.339	419	14.02	S96	0.067	0.671	0.604	1.341	419	14.05
S37	0.067	0.670	0.603	1.341	419	14.04	S97	0.067	0.671	0.604	1.341	419	14.05
S38	0.067	0.670	0.603	1.341	419	14.04	S98	0.067	0.671	0.604	1.342	419	14.05
S39	0.067	0.671	0.604	1.341	419	14.05	S99	0.067	0.672	0.605	1.344	419	14.07
S40	0.067	0.670	0.603	1.340	419	14.03	S100	0.067	0.672	0.605	1.344	419	14.07
S41	0.067	0.670	0.603	1.339	419	14.02	S101	0.067	0.669	0.602	1.338	419	14.01
S42	0.067	0.669	0.602	1.338	419	14.01	S102	0.067	0.669	0.602	1.338	419	14.01
S43	0.067	0.669	0.602	1.338	419	14.01	S103	0.067	0.669	0.602	1.338	419	14.01
S44	0.067	0.672	0.604	1.343	419	14.06	S104	0.067	0.669	0.602	1.338	419	14.01
S45	0.067	0.672	0.604	1.343	419	14.06	S105	0.067	0.669	0.602	1.338	419	14.01
S46	0.067	0.669	0.602	1.338	419	14.01	S106	0.067	0.669	0.602	1.338	419	14.01
S47	0.067	0.669	0.602	1.339	419	14.02	S107	0.067	0.669	0.602	1.338	419	14.01
S48	0.067	0.669	0.602	1.339	419	14.02	S108	0.067	0.670	0.603	1.340	419	14.03
S49	0.067	0.669	0.569	1.304	407	13.61	S109	0.067	0.670	0.603	1.340	419	14.03
S50	0.067	0.672	0.604	1.343	419	14.06	S110	0.067	0.671	0.604	1.342	419	14.06
S51	0.067	0.672	0.604	1.343	419	14.06	S111	0.067	0.671	0.604	1.342	419	14.06
S52	0.067	0.672	0.604	1.343	419	14.06	S112	0.067	0.669	0.602	1.338	419	14.01
S53	0.067	0.669	0.602	1.338	419	14.01	S113	0.067	0.669	0.602	1.338	419	14.01
S54	0.067	0.669	0.602	1.338	419	14.01	S114	0.067	0.669	0.602	1.338	419	14.01
S55	0.067	0.669	0.602	1.338	419	14.01	S115	0.067	0.670	0.603	1.340	419	14.04
S56	0.067	0.669	0.602	1.338	419	14.01	S116	0.067	0.669	0.602	1.339	419	14.02
S57	0.067	0.669	0.602	1.338	419	14.01	S117	0.067	0.669	0.602	1.339	419	14.02
S58	0.067	0.669	0.602	1.338	419	14.01	S118	0.067	0.669	0.602	1.339	419	14.02
S59	0.067	0.669	0.602	1.338	419	14.01	S119	0.067	0.669	0.602	1.339	419	14.02
S60	0.067	0.669	0.602	1.338	419	14.01	S120	0.067	0.669	0.602	1.339	419	14.02
							S121	0.068	0.676	0.609	1.353	419	14.16

Relative abundance (RA), RA1 = Monomeric repeats, RA2 = Dimeric repeats, RA3 = Trimeric repeats, RA6 = Hexameric repeats, RD = Relative Density.

both positively correlated against G/C content with a coefficient value of 0.310 ($R^2 = 0.096$, $P < 0.05$) and 0.331 ($R^2 = 0.109741269$, $P < 0.05$) respectively. Since the genome sizes of the analyzed viruses are very much similar with little variation to one another, a significant correlation was not expected.

4. Discussion

Due to the advancement of next-generation DNA sequencing technologies, microbial genome could be sequenced in an increasingly efficient, fast, cheap, and multiple copies at a time (Alam et al., 2014a; Atia et al., 2016) and thus, a tremendous surge of over 114,000 SARS-CoV-2 genomic sequences being available in the public database in few months only (https://www.gisaid.org/). This accumulation of a huge dataset has led us to the unravel the genomic complexities and genetic distribution/variation present in SARS-CoV-2 genome isolated from across the globe. These genome sequences represent a potentially valuable resource for mining both clinical and evolutionary significant

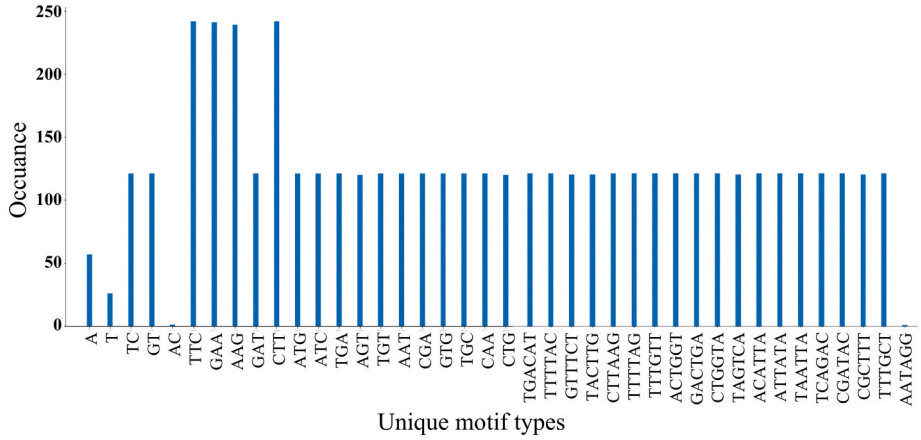

Fig. 4. The differential occurrence of individual SSR motif. The figure showed the occurrence of different unique mono-, di-, tri- and hexanucleotide in all the analyzed 121 SARS-CoV-2 genomes. The figure very clearly illustrates the presence of TTC, GAA, AAG and CTT trinucleotide repeats twice per genome, while the rest of the repeats present only once per genome.

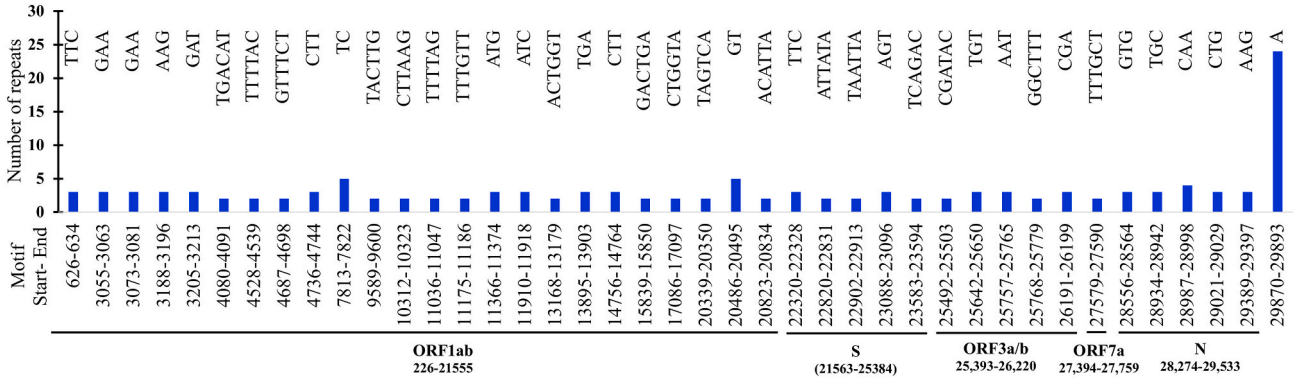

Fig. 5. Distribution of the identified SSR motifs across the genome of SARS-CoV-2. The figure showed the occurrence of different SSR motifs in the ORF1ab, S, ORF3ab, ORF7, and N region of SARS-CoV-2 genomes. The number of repeats of each motif could also be found from this figure.

SSR markers. Their presence and variation across the genome of same species have been studied extensively in different viruses including *Spodoptera littoralis* multiple nucleopolyhedrovirus (SpliMNPV) (Atia et al., 2016), potexvirus (Alam et al., 2014a), Human Immunodeficiency Virus (Chen et al., 2009), Mycobacteriophage (Alam et al., 2019), Hepatitis C (Chen et al., 2011) to identify the correlation between the diversity of repeats, incidence and complexity of repeats, genome size and host range (Zhao et al., 2012). In the present study, we have explored 121 SARS-CoV-2 genomes identified from 31 countries covering 6 continents for the identification, abundance, and composition of SSR repeats and observed a total of 38–42 different types of repeats. The SSRs incidence in SARS-CoV-2 genome is almost similar to potyviruses (23–45 SSRs) (Zhao et al., 2011) and Human immunodeficiency virus isolates (22–48 SSRs) (Chen et al., 2009); but higher than tobamovirus having 11–36 SSRs (Alam et al., 2014a), potexvirus of 11–30 SSRs (Alam et al., 2014a) and geminivirus (4–19 SSRs) (George et al., 2012); and lower than that of *Spodoptera littoralis* multiple nucleopolyhedrovirus with 55 repeats (Atia et al., 2016). Although genome size and hosts play an important factor in determining the occurrence of SSRs (Zhao et al., 2012); SSRs incident frequency varied quite largely across all these studied genomes.

We have calibrated our identification tools so that tandem repeat sequences below 6 bp and above 15 bp are not counted. The minimum number of repeats for each type is 10-5-3-3-3-2 configuration for mono-, di-, tri-, tetra-, penta-, and hexarepeats. We have identified incredible similarity pattern in all of 121 genomes, might be due to the high level of sequence conservancy in SARS-CoV-2. Independent studies on vertebrate and plant genomes have provided a basis for categorizing the most common SSR motifs. The most common SSR motif in animals and invertebrates is $(GT)_n$ (Stallings et al., 1991), whereas in plants it is $(AT)_n$ (Lagercrantz et al., 1993) and in insects, the most common motif is thought to be $(CT)_n$ (Paxton et al., 1996). Dinucleotide repeats AT/TA and AG/GA were found to be the two most prominent form in the largest Closteroviridae RNA virus family (George et al., 2016). Following the similar trend SSR analysis of viral genomes revealed the most common motif to be $(AT)_n$ (Zhao et al., 2012). SARS-CoV-2 deviates from this trend with the most common repeat being trimeric $(TTC)_n$ and $(CTT)_n$ repeats which were present in all of the analyzed genomes for multiple times.

In the case of the SARS-CoV-2 genome, results revealed that the hexameric motif was the most abundant type of repeat (49%) followed by the trinucleotide of 42%, the other two types of mono- and dimeric-repeats present in 4% (Table S3); while tetra- and penta-nucleotide repeats were non-existent. In partial agreement with our results, trinucleotide SSRs were found to be the most frequent types in SpliMNPV and Human Immunodeficiency Virus Type 1 (HIV–1). However, the genome of hepatitis C virus (HCV) possessed predominantly mono-, di- and trinucleotide repeats with the rare presence of other types (Chen et al., 2011). In contrary, the mononucleotide repeats were the most abundant form in 30 alphaviruses (Alam et al., 2014b), Herpes Simplex Virus Type 1 (Deback et al., 2009) and different ssDNA viruses (Jain et al., 2014) genomes followed by di-/tri-nucleotide repeats. Although the presence of tetra– and penta– nucleotides microsatellites is rare in diverse Geminivirus (George et al., 2012) and HCV (Chen et al., 2011), SARS-

CoV-2 genomes showed complete absence of this kind of motifs (Fig. 4).

The level of repetitiveness and incidence of SSR sequences have been readily correlated with genome size and G/C content (Zhao et al., 2012). Several reports established the positive correlation of the SSR content with their respective genome size of fungal (Karaoglu et al., 2005) and plant genomes (Morgante et al., 2002). A weak influence of genome size and GC content had been established on the number, relative abundance and relative density of microsatellites in various analyzed HCV genomes (Chen et al., 2011). Our findings suggest that relative abundance and density is positively correlated with genome size and the correlation is statistically significant. Conversely, the correlation with G/C content is positive but not statistically significant. In establishing distribution patterns of SSRs in SARS-CoV-2, it could be concluded that there is no significant pattern in the distribution of SSRs in viral genomes. It can also be said that the number of SSR present in a genome cannot be considered proportional to the genome size as the sequences used in this study were grossly similar in size (Table 1). Similar kind of study conducted in diverse HIV-1 genomes revealed no direct proportional relationship to the genome size and total SSR contents (Chen et al., 2009). We conducted this study in the hope of documenting and establishing the SSR patterns present in SARS-CoV-2 as well as the particular motifs that are present in the genome. Further studies would perhaps aim at detailing the presence of these repeat motifs in coding and non-coding regions of the genome to predict regions prone to mutations.

5. Conclusion

The relevance of our findings would help to gain knowledge regarding the functional, physiological, and evolutionary significance of various SSR repeats. Repetitive sequences are considered as the hot spots for recombination, as this might play a significant role in the ability of SARS-CoV-2 virus to rapidly adapt to a different kind of environmental and genetic variation of hosts. Genome-wide extraction of microsatellites across 121 SARS-CoV-2 genomes revealed the presence of 38–42 SSRs per genome. Though a complete understanding of the position of these SSRs in the coding region of the genome yet to be completed, the functional variations of this virus in a different region could be assigned.

Role of funding sources

There was no funding received to carry out this work.

CRediT authorship contribution statement

R.S. and A.G. performed all the computational work, wrote the main manuscript, prepared tables, and figures. A.G. conceptualized the idea, supervised the entire study and was involved in the analysis and interpretation of the data. All authors reviewed and approved the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial or personal conflicts.

Acknowledgment

The authors thank NCBI database (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) for providing the curated genomic data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.genrep.2021.101020.

References

Alam, C.M., Singh, A.K., Sharfuddin, C., Ali, S., 2014a. Incidence, complexity and diversity of simple sequence repeats across potexvirus genomes. Gene 537, 189–196.
Alam, C.M., Singh, A.K., Sharfuddin, C., Ali, S., 2014b. In-silico exploration of thirty alphavirus genomes for analysis of the simple sequence repeats. Meta Gene 2, 694–705.
Alam, C.M., Iqbal, A., Sharma, A., Schulman, A.H., Ali, S., 2019. Microsatellite diversity, complexity, and host range of mycobacteriophage genomes of the Siphoviridae family. Front. Genet. 10, 207.
Atia, M.A., Osman, G.H., Elmenofy, W.H., 2016. Genome-wide in silico analysis, characterization and identification of microsatellites in Spodoptera littoralis multiple nucleopolyhedrovirus (SpliMNPV). Sci. Rep. 6, 1–9.
Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L., 2000. GenBank. Nucleic Acids Res. 28, 15–18.
Chen, M., Tan, Z., Jiang, J., Li, M., Chen, H., Shen, G., Yu, R., 2009. Similar distribution of simple sequence repeats in diverse completed human immunodeficiency virus type 1 genomes. FEBS Lett. 583, 2959–2963.
Chen, M., Tan, Z., Zeng, G., 2011. Microsatellite is an important component of complete hepatitis C virus genomes. Infect. Genet. Evol. 11, 1646–1654.
Deback, C., Boutolleau, D., Depienne, C., Luyt, C., Bonnafous, P., Gautheret-Dejean, A., Garrigue, I., Agut, H., 2009. Utilization of microsatellite polymorphism for differentiating herpes simplex virus type 1 strains. J. Clin. Microbiol. 47, 533–540.
Dieringer, D., Schlötterer, C., 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 13, 2242–2251.
Gao, L., Qi, J., 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol. Biol. 7, 41.
George, B., Alam, C.M., Jain, S., Sharfuddin, C., Chakraborty, S., 2012. Differential distribution and occurrence of simple sequence repeats in diverse geminivirus genomes. Virus Genes 45, 556–566.
George, Biju, George, Binu, Awasthi, M., Singh, R.N., 2016. In silico genome-wide identification and analysis of microsatellite repeats in the largest RNA virus family (Closteroviridae). Turk. J. Biol. 40, 589–599.
Guo, Y.-R., Cao, Q.-D., Hong, Z.-S., Tan, Y.-Y., Chen, S.-D., Jin, H.-J., Tan, K.-S., Wang, D.-Y., Yan, Y., 2020. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. Military Med. Res. 7, 1–10.
Heesacker, A., Kishore, V.K., Gao, W., Tang, S., Kolkman, J.M., Gingle, A., Matvienko, M., Kozik, A., Michelmore, R.M., Lai, Z., 2008. SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility. Theor. Appl. Genet. 117, 1021–1029.
Jain, A., Mittal, N., Sharma, P.C., 2014. Genome wide survey of microsatellites in ssDNA viruses infecting vertebrates. Gene 552, 209–218.
Karaoglu, H., Lee, C.M.Y., Meyer, W., 2005. Survey of simple sequence repeats in completed fungal genomes. Mol. Biol. Evol. 22, 639–649.
Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., Makova, K.D., 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18, 30–38.
Khan, S., Siddique, R., Shereen, M.A., Ali, A., Liu, J., Bai, Q., Bashir, N., Xue, M., 2020. Emergence of a novel coronavirus, severe acute respiratory syndrome coronavirus 2: biology and therapeutic options. J. Clin. Microbiol. 58.
Kim, T.-S., Booth, J.G., Gauch, H.G., Sun, Q., Park, J., Lee, Y.-H., Lee, K., 2008. Simple sequence repeats in Neurospora crassa: distribution, polymorphism and evolutionary inference. BMC Genomics 9, 31.
Lagercrantz, U., Ellegren, H., Andersson, L., 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. 21, 1111–1115.
Li, Y.-C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.
Morgante, M., Hanafey, M., Powell, W., 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. 30, 194–200.
Paxton, R., Thorén, P., Tengö, J., Estoup, A., Pamilo, P., 1996. Mating structure and nestmate relatedness in a communal bee, Andrena jacobi (Hymenoptera, Andrenidae), using microsatellites. Mol. Ecol. 5, 511–519.
Riley, D.E., Krieger, J.N., 2009. Embryonic nervous system genes predominate in searches for dinucleotide simple sequence repeats flanked by conserved sequences. Gene 429, 74–79.
Saha, S., Malaker, R., Sajib, M.S.I., Hasanuzzaman, M., Rahman, H., Ahmed, Z.B., Islam, M.S., Islam, M., Hooda, Y., Ahyong, V., 2020. Complete genome sequence of a novel coronavirus (SARS-CoV-2) isolate from Bangladesh. Microbiol. Resour. Announcements 9.
Shu, Y., McCauley, J., 2017. GISAID: global initiative on sharing all influenza data–from vision to reality. Eurosurveillance 22, 30494.
Stallings, R., Ford, A., Nelson, D., Torney, D., Hildebrand, C., Moyzis, R., 1991. Evolution and distribution of (GT) n repetitive sequences in mammalian genomes. Genomics 10, 807–815.
Tang, Q., Song, Y., Shi, M., Cheng, Y., Zhang, W., Xia, X.-Q., 2015. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Sci. Rep. 5, 17155.
Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., McCouch, S., 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11, 1441–1452.
Wang, H., Li, X., Li, T., Zhang, S., Wang, L., Wu, X., Liu, J., 2020. The genetic sequence, origin, and diagnosis of SARS-CoV-2. Eur. J. Clin. Microbiol. Infect. Dis. 1.

Weiss, S.R., Leibowitz, J.L., 2011. Coronavirus pathogenesis. In: Advances in Virus Research. Elsevier, pp. 85–164.
Wu, A., Peng, Y., Huang, B., Ding, X., Wang, X., Niu, P., Meng, J., Zhu, Z., Zhang, Z., Wang, J., 2020. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 27 (3), 325–328.
Zhao, X., Tan, Z., Feng, H., Yang, R., Li, M., Jiang, J., Shen, G., Yu, R., 2011. Microsatellites in different Potyvirus genomes: survey and analysis. Gene 488, 52–56.
Zhao, X., Tian, Yonglei, Yang, R., Feng, H., Ouyang, Q., Tian, You, Tan, Z., Li, M., Niu, Y., Jiang, J., 2012. Coevolution between simple sequence repeats (SSRs) and virus genome size. BMC Genomics 13, 435.
Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.