



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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



## Developing an ultra-efficient microsatellite discoverer to find structural differences between SARS-CoV-1 and Covid-19

Mahmoud Naghibzadeh<sup>a,b,\*</sup>, Hossein Savari<sup>a</sup>, Abdorreza Savadi<sup>b</sup>, Nayyereh Saadati<sup>c</sup>, Elahe Mehrazin<sup>a</sup>

<sup>a</sup> Knowledge Engineering Research Group, Computer Engineering Dept., Ferdowsi University of Mashhad, Mashhad, Iran

<sup>b</sup> High Performance Computing Lab., Computer Engineering Dept., Ferdowsi University of Mashhad, Mashhad, Iran

<sup>c</sup> Ghaem Hospital, Department of Internal Medicine, Mashhad University of Medical Sciences, Mashhad, Iran

### A B S T R A C T

**Motivation:** Recently, the outbreak of Coronavirus-Covid-19 has forced the World Health Organization to declare a pandemic status. A genome sequence is the core of this virus which interferes with the normal activities of its counterparts within humans. Analysis of its genome may provide clues toward the proper treatment of patients and the design of new drugs and vaccines. Microsatellites are composed of short genome subsequences which are successively repeated many times in the same direction. They are highly variable in terms of their building blocks, number of repeats, and their locations in the genome sequences. This mutability property has been the source of many diseases. Usually the host genome is analyzed to diagnose possible diseases in the victim. In this research, the focus is concentrated on the attacker's genome for discovery of its malicious properties.

**Results:** The focus of this research is the microsatellites of both SARS and Covid-19. An accurate and highly efficient computer method for identifying all microsatellites in the genome sequences is discovered and implemented, and it is used to find all microsatellites in the Coronavirus-Covid-19 and SARS2003. The Microsatellite discovery is based on an efficient indexing technique called K-Mer Hash Indexing. The method is called Fast Microsatellite Discovery (FMSD) and it is used for both SARS and Covid-19. A table composed of all microsatellites is reported. There are many differences between SARS and Covid-19, but there is an outstanding difference which requires further investigation.

**Availability:** FMSD is freely available at [https://gitlab.com/FUM\\_HPCLab/fmsd\\_project](https://gitlab.com/FUM_HPCLab/fmsd_project), implemented in C on Linux-Ubuntu system. Software related contact: [hossain\\_savari@mail.um.ac.ir](mailto:hossain_savari@mail.um.ac.ir).

### 1. Introduction

The novel Coronavirus outbreak began in Wuhan, China, in December 2019 (P. [1]) and quickly reached a point such that a pandemic state was declared by the World Health Organization. Although it has been controlled in Wuhan, overall it is in the throughout the world. Knowing the structure of this virus and revealing the hidden properties of its genome could be helpful towards the design of effective treatment procedures of human victims and also the production of a vaccine to provide immunity against this virus. Assessment its severity and better understanding this disease is under way [2].

Finding all microsatellites in Covid-19 is one direction towards analysis of the virus's structure. Lack of a fast, accurate, and memory efficient microsatellite discoverer motivated us to develop such a tool first, and then start analyzing the Covid-19 genome structure. Therefore, this research follows two objectives, development of a general microsatellite discoverer which can be used for different genomes, and

analysis of the structures of both SARS-CoV-1 and that of Coronavirus-Covid-19 using this tool and revealing their differences. The final results will assist future investigations towards drug and vaccine design. Even very minor assistance in these directions can have a great impact on saving lives and elevating the quality of numerous human lives.

Tandem repeats, in genomic sequences, are motifs which are continuously repeated many times in DNA (Deoxyribonucleic acid), gene, genome, or other genomic sequences and, as the name suggests, the orientation of the motif in all repeats are the same. They are classified into three classes, microsatellites, minisatellites and satellites. Different researchers and practitioners disagree where to draw the separating line between microsatellites, minisatellites and satellites; however, the number 7 appears very often as the longest length of the core subsequences of microsatellites [3]. The core subsequence of microsatellites will be called motif, from here on. Another important feature of microsatellites, minisatellites and satellites is the number of repeats of the corresponding motifs. Once again, there is no clear-cut

\* Corresponding author. Knowledge Engineering Research Group, Computer Engineering Dept., Ferdowsi University of Mashhad, Mashhad, Iran.

E-mail address: [naghibzadeh@um.ac.ir](mailto:naghibzadeh@um.ac.ir) (M. Naghibzadeh).

<https://doi.org/10.1016/j.imu.2020.100356>

Received 20 April 2020; Received in revised form 20 May 2020; Accepted 20 May 2020

Available online 21 May 2020

2352-9148/© 2020 The Authors.

Published by Elsevier Ltd.

This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

lower bound for the minimum number of repeats.

A more general case of micro and mini satellites is called Variable Number of Tandem Repeats (VNTR) in which other subsequences may appear between motif repeats [4]. VNTRs are very common in the human genome, and it is estimated that 3% of the human genome is composed of VNTRs [5]. The focus of this research is finding microsatellites with both a fixed number of tandem repeats and those in which the number of repeats vary in different individuals. The variability of the number of repeats in both microsatellites, which is called the number of repeats polymorphism, is a good candidate for an individual's DNA signature [6] [7]. Could it be the case for different varieties of the coronavirus? This has to be investigated after at least the microsatellites are all detected and reported.

The proposed approach to recognition of microsatellites in any genomic sequence as well as that of the novel Coronavirus benefits from a hash indexing method called the K-Mer Hash Index (KMHI) [8]. In this research, a novel improvement is added to the KMHI to make it both time and space efficient, which is briefly discussed in the following paragraph and the details are clarified in the solution approach section. This novelty not only is usable for finding microsatellites, the topic of this research, but also all other kinds of VNTRs in any genomic sequence.

Generally, each row of the KMHI points to a link list of places where the k-mer value corresponding to this row appears in the reference sequence. After the whole input sequence is processed and KMHI and all linked lists are developed, each list is separately processed to find possible microsatellites. The discovery of this research made it possible to remove all linked lists and hence to highly reduce the space requirements and increase the efficiency of the microsatellites' discovery. The processing of each potential microsatellite is done as soon as it is detected during the processing of the input sequence rather than postponing it to after the sequence is completely scanned. The details are described in the solution approach section. The properties and novelties of the presented method, which is named Fast MicroSatellite Discoverer (FMSD), for finding all microsatellites of a given gene, DNA, RNA, or other genome sequences including the Novel Coronavirus [9] and SARS [10] is as follows.

- It finds all microsatellites with core sequences, i.e., motifs, of 1, 2, ..., and 7 characters.
- It uses an efficient indexing method which does not require any space for storing k-mer values or their locations in the sequence.
- It uses extremely low main memory space and no secondary storage except for the input sequence and the output results.
- It is extremely fast.

Potential microsatellites are processed as they are found, not when the input sequence reaches its end. Therefore, although the developed software is not parallel, it can easily become so in the future.

Using FMSD, both SARS and Covid-19 genomes were computationally analyzed and the differences were recognized and highlighted. There seems to be important differences between these two genomes, but their medical significance has to be investigated by geneticists and health specialists.

The remainder of the paper is organized in five sections and one references division. Section 2 reviews related papers, especially the ones used for comparisons. Section 3 is a brief clarification of the problem being solved. Section 4 describes the solution approach. Section 5 details the evaluation, reports the comparison results, and highlights the structural differences with respect to microsatellites between SARS and Coronavirus-Covid-19 as a case study.

## 2. Background

Coronavirus infection growth has been very rapid. In less than three months' time it has reached a pandemic state. The need for research is growing, and its priorities are being set [11]. With respect to urgent drug

discovery, the immediate effort is focused on repurposing drugs [12]. In this respect, the study of virus-host and human-human protein-protein interactome networks is said to be essential [13]; [12]. Another direction is the discovery of known patterns which have been the source of other diseases, in the genome of the Covid-19. It is estimated that there are 4.6 million microsatellites in the human genome [14]. Amplification of many of these microsatellites have been used for many purposes such as disease diagnosis or human observable characteristics segregation. Numerous researches have reported the association between different diseases and certain microsatellites [15,16]. For example, the instability nature of repeats in the microsatellites is associated with many kinds of cancer and a variety of disorders related to the nervous system's degeneration, such as Huntington's disease [17]. Even a repeat polymorphism in one of the genes, i.e., IL-1ra, is shown to be associated with an increased possibility of osteoporotic fractures [18,19]. Having said so, accurately and efficiently locating microsatellites in a given genome or genomic sequence has been and still is an important issue.

Research on finding tandem repeats has a long history and covers diverse classes of problems with different restrictions and assumptions. Generally speaking, the goal is either to find all kinds of tandem repeats [20] or a subset of classes of tandem repeats such as microsatellites [17] with a variable number of repeats in different individuals or without it. From a different point of view, the outcome is either exact or inexact and for the case of inexact, it could be statistical [21], fuzzy [22], based on distances such as Hamming or edit distance [20]. To the best of our knowledge, there is no document with a complete classification of methods for finding tandem repeats and the first objective of this research is presenting a new approach to finding these repeats and not reviewing and classifying existing methods. Therefore, we will limit ourselves to documents which provide a tide background for the concept and introduce those documents which are used in the comparison sections. The software developed here is capable of finding all microsatellites of genomic sequences whatever their length might be. The main objective of this research is to use the developed system to recognize all microsatellites of the novel Coronavirus and that of SARS and highlight the differences. A detailed comparison will be provided.

Kmer-SSR is a package which detects simple sequence repeats of all sizes. The advantage of this method is that it uses a backtracking idea to return to the beginning of the detected sequence and look again for overlapped tandem repeats. This way the precision of the method is increased. The algorithm is an exact one; hence it is not either fuzzy or approximate. This is also the property of the method developed here, which makes it a good candidate to compare our method with [23].

Krait [24] is specifically designed to detect exact microsatellites, similar to what is being done in this paper. However, Krait does not find microsatellites of core sizes equal to 7. To obtain the other six types of microsatellites, the program scans the source sequence six times. It has the capability of using a seed-and-extend concept to detect similar repeats. To check the atomicity of the core sequence, it searches a long list of possible core sequences, which is a time consuming process. Although we noticed it does not detect microsatellites with overlapped cores, it is a good candidate for the comparison with the method which is developed here. However, it is a deficiency for sensitive situations such as Covid-19's genome study.

A software tool called mreps is developed to detect all tandem repeats, including microsatellites, in DNA as well as whole genome. Their claim is that it is capable to identify fuzzy tandem repeats by defining a resolution parameter. It is general in the sense that it identifies tandem repeats with all core sizes. A feature of mreps is that it can tolerate errors between repeated copies [20]. With the generality, of course, comes slowness, usage complexity, and inaccuracy. For the case of Coronavirus which is composed of a short genome, fuzziness and inexactness is a weakness of the method. However, the reason for including fuzziness in the search seems to be to decrease the time requirement. For this reason it is decided to include this software in the comparison part to show the superiority of our method even with respect to time.

The last work we are going to review is PERF [25]. It is based on the preparation of a repeated set in advance, and comparing all possible subsequences of the input sequence against this set, and then perform further processing, if a match happens. In the default option of this tool, all microsatellites of core lengths 1 to 6 are detected. To build the repeat set, all possible combinations of k-mers, for  $k = 1$  to 6, are generated, and each one is extended by contiguous repetition of the k-mer to 12 characters. A moving window with length 12 is initialized with the 12 first characters of the input sequence, and it searches in the repeat set. If it is found, further processing will follow to find all repeats; otherwise, the moving window is pushed one place forward and the process continues. The main difficulty with this approach is the search time in the repeat set, especially if microsatellites of longer core sizes are of users' interest. For example, if the core length is seven nucleotides,  $4^7$  rows have to be added to the repeat set. One other difficulty is that the tool is not capable of detecting microsatellites with a low number of repeats. For example, it would not be able to detect microsatellites with core size 1 in which the number of repeats are less than 12. It is possible to change the default values but, a high search time will persist and other problems may arise.

### 3. Problem definition

A genomic sequence is given as input. It is composed of nucleotides A (adenine), T(thymine), G(guanine), and C(cytosine). For some sequences such as Ribonucleic acid (RNA) T is replaced by U; however, we have assumed the sequence to be composed of A, T, G, and C. By a simple replacement of U by T in the input sequence, the method presented here will hold valid and applicable. The length of the input sequence could be as long as the whole genome and as small as a few nucleotides. The problem is to find all successive tandem repeats of the same core sequence, i.e., same motif. The core sequences of microsatellites are considered to be of length 1, 2, ..., or 7 nucleotides. Further, we assume the core sequence is atomic. For example, ATAT is not atomic because it is composed of two identical ATs, and hence it should be reported as a microsatellite of core sequence AT, assuming it passes other requirements.

Using the developed tool, the existence of microsatellites in Coronavirus-Covid-19 will be carefully investigated, and all such cases will be reported. The same will apply to SARS-CoV-1, and differences will be specified. Possible interpretations of the differences will be discussed, while more complex cases will be left for future investigations.

### 4. Solution approach

The input sequence is a genome or genomic sequence in which all tandem repeats of type microsatellite will be detected and reported. The approach starts with generating an index table from the input sequence. The index table is called K-Mer Hash Index (KMHI).

To generate the KMHI, starting from the beginning of the input sequence, the first six-character subsequence, i.e., characters in positions 0 to 5, is selected and it is converted to an integer using a two-bit code for each of the letters A, C, G, and T. The corresponding codes of the letters are 00, 01, 10, and 11, respectively. For example, the subsequence ACCTGA would be converted to the integer 000101111000 in base 2 which is equal to decimal number 376. Decoding this code will take us to row number 376 of the Table KMHI. Note that the first row is numbered zero. The table has  $212 = 4096$  rows. Although all microsatellites of sizes 1 to 7 are to be found, only one KMHI is produced, and the unit subsequence for the hashing purpose is always 6. Other options such as units of four were examined, and considering the overall performance and flexibility, 6 was selected. In the following, the details of detecting actual microsatellites will be described using an example.

Neither the subsequence itself nor the code is stored in the KMHI table. Depending on the problem being solved, different information may be stored in rows of the KMHI table. This coding system has two

major advantages, a no-space requirement for storing search-key values and negligible decoding time requirement. With the big data production and fast analysis requirements, KMHI is a good candidate to be used in many fields of bioinformatics. Let us say the hash of a search-key value is interpreted as an integer variable  $x$ , then its decoding is equivalent to going to  $KMHI(x)$ , i.e., Row  $x$  of Table KMHI.

The simplest approach is to attach a link list to each row of the KMHI, and as the input is scanned, store the position of each k-mer in the link list which corresponds to this k-mer. After the end of the input sequence is reached, then each linked list is separately processed, and microsatellites are detected. In this approach, for each character of the input sequence, a node of a list is formed, and in each node a location and a pointer is stored. Therefore, the storage requirement forces the software to rely on the secondary storage which makes the time requirement intolerable. Besides, processing the links is another factor which adds to the computation time. In this research, a novel idea was developed which removes all the linked lists and instead, for each k-mer includes only three values in the KMHI table. The software system which is developed around this idea to find all microsatellites of a genomic sequence is called Fast Microsatellite Discoverer (FMSD).

Suppose as the input sequence is being scanned, a 6-mer GTTCGT is seen such that the end character, T, is in Position 410 of the input sequence, i.e., this 6-mer is in Locations 405 to 410 of the input sequence. This 6-mer points to row number 3035 of the KMHI table. Suppose the same 6-mer is repeated 4 positions later, i.e., in Positions 409 to 414. This will point to the same row number 3035 of the KMHI. With respect to this row, we can guess that perhaps a microsatellite, actually GTTC, is forming, because two of them have been already observed. At this point, Row 3035 will store three values, *loc* which is equal to 414, *size* which is equal to 4, and *count* which is equal to 2. Let us assume that the same 6-mer is repeated in positions 413 to 418. Therefore *loc* will become 418, *size* will not change because it shows the core size of this microsatellite, and *count* will change to 3. See Figure 1

A similar case to what was discussed for microsatellites with core subsequences of length 4 is applicable to all microsatellites of sizes one to six. For size seven, a minor extra comparison has to be done. It is clear that size must be 7 but the unit of hashing is six characters. To solve this case, each time an extra comparison for the seventh character has to be performed. There was a compromise between taking the unit of hashing to be seven and reducing the size of the KMHI table by 75%. For the details of the method see Algorithm 1.

To be able to follow the Algorithm 1, it is important to know what every variable is used for. Variable *sequence* is an array of size  $n$ , indexed from zero to  $n-1$ , which holds the input sequence. Table KMHI is the most important variable which is a structure array of size  $212 = 4096$  rows, indexed from 0 to 4095, that is used to hold the location, *loc*, core size, *size*, number of repeats, *count*, of the current potential

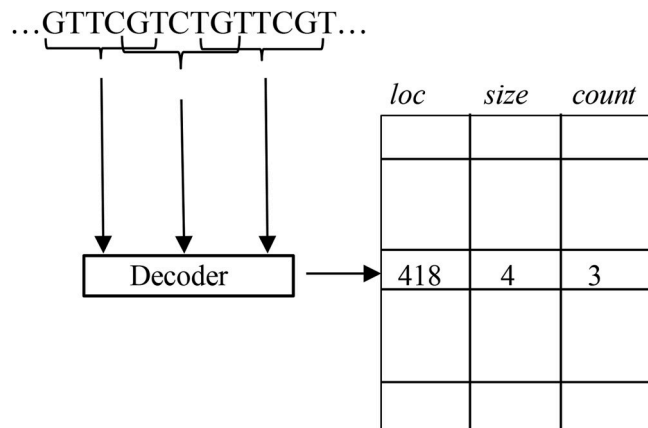


Fig. 1. The structures of hashing and KMHI table.

microsatellite corresponding to this row of the table. This table is considered to be global and all its values are set to zero before entering the algorithm. In each row of the table, after the initial reference, it is always assumed that a potential microsatellite is being detected. If the next reference to this row is not exactly as much as size apart from the previous one then, up to here, is assumed to be a microsatellite, and microsatellite validation routine, *MSValidation*, is called to do the validation such as checking the number of repeats. Otherwise, variables *loc* and *count* are updated.

FMSD recognizes non-atomic tandem repeats and only reports atomic ones. For example, reporting a microsatellite with core sequence TCTC with 10 times repetition is not correct because TCTC is not atomic. The correct microsatellite would be TC with 20 times repetition. Cyclic repetitions are also eliminated. For example, when a microsatellite with core subsequence TCG with 20 times repetition is reported, geneticists can easily interpret that for example, there is a microsatellite with core sequence CGT which is repeated 19 (or 20) times, and there is no need to highly increase the number of reported microsatellites. In addition, if the end of a microsatellite and the start of the next microsatellite have some nucleotides in common, FMSD is able to correctly include the overlapped nucleotides for both microsatellites. It would be the task of the interpreter to decide which one is more important.

The developed method was shown to be very effective. It detects all microsatellites, everything is done in one pass of scanning the input sequence, and there is no need for preprocessing or post processing. It is very fast, and even in the worst case it is at least 2.6 times faster than the fastest state-of-the-art method. It requires negligible main memory, except what it is needed by the input sequence.

## 5. Evaluation

Coronavirus is associated with many properties, some of which are listed below.

- All ages are susceptible with different probabilities [26].
- The fatality rate of Covid-19 is high [27].
- There is no significant gender difference [26].
- Human-to-human transmission is the norm [26].
- Transmissibility is higher than SARS-2003 [27].
- As a consequence, it has forced social distancing throughout the world, and many geographic regions are experiencing a lockdown.
- The commonality to all these direct and indirect properties is in its genome. In the following subsection all microsatellites of both SARS-2003 and Covid-2019 are discovered and reported.

### 5.1. FMSD performance

The software which is developed here is usable for all genomes of all sizes. It is accurate and exact. It discovers the microsatellites with the smallest number of repeats compared to state-of-the-art packages. It is fast while being a sequential software. Since potential microsatellites are

detected as the input is scanned, the method which is developed here can easily become parallel to further improve its run time and make use of all cores of computers. It is also memory efficient. Since correctness and timing are the most important objectives of such algorithms, Table 1 shows the time requirement of different software for different sizes of genomes. The timing of all methods are measured with an Intel Xeon E5-6695 v3 2.3Gigahertz processor with a 64 GB main memory module. It is clearly notable that the time requirement of FMSD is far less than those of other algorithms. In some cases, it is 3600 times faster, and in the worst case it is at least 2.6 times faster than the state-of-the-art method. This offers a great advantage, because most genomic sequences are very long and previous methods may not be capable to perform their tasks in a tolerable period of time. The same discussion is true for main memory utilization; however, since other methods did not reported the memory utilization of their algorithms and the source codes (not the executable code) are not available, comparison is not possible. An important objective of this research is to find all microsatellites of both the SARS and Coronavirus-Covid-19, and provide the required information for the geneticist to analyze their differences towards drug discovery and vaccine production. This is done in the next subsection.

### 5.2. Microsatellites in Covid-19 and SARS-CoV-1

The developed software, FMSD, is applied to both SARS-CoV-1 (accession number AY278741.1) and the novel Coronavirus-Covid-19 (accession number NC\_045512.2) and the results are reported in this section. The results are shown for all microsatellites with atomic core lengths of less than or equal to 7. The minimum number of repeats is set to 3 except for atomic core lengths of 1 and 2 which is set to 7 and 4, respectively. Table 2 lists all microsatellites in both viruses.

A polyAdenine (PolyA) tail is seen in Coronavirus-Covid-19 which is absent in SARS-CoV-1. Yet the longest sequence of Adenine in SARS-CoV-1 is of length 8 which is located somewhere within the genome.

In addition, the maximum length of atomic simple tandem repeats found in these two genomes were 4 nucleotides (TGTT), which is repeated three times in SARS-CoV-1 genome and two times in Coronavirus-Covid-19 genome. This fact as well as the absence of longer microsatellites in either of the genomes suggests a selection against longer simple tandem repeats in viral genomes.

Another microsatellite found in both genomes is CAA (poly glutamine, PolyQ, at protein level). It is shown that PolyQ is involved in transmissibility of cowpox virus, while lacking of such motif in smallpox leads to less survival of this virus [28].

It is worth mentioning that a point mutation in the Covid-19 genome compared to SARS-CoV-1 genome might cause the software to find two different microsatellites in the two genomes. For example, segment TTGTGTGTGTA can be read as 4 repeats of TG or 5 repeats of GT assuming a recent mutation from G to T at the beginning of the segment. Therefore, some differences might not be as significant as they look.

Since Covid-19 recently jumped from animal to human host, it has not yet adapted to its new environment; hence its genome shows a rapid dynamics towards equilibrium. FMSD, as a simple and fast simple

**Table 1**  
Performances of state-of-the-art Microsatellite detectors with respect to processing time.

	Kmer-SSR	Krait <sup>a</sup>	Perf	Mreps	FMSD
<i>Escherichia coli</i> (E.coli) GCF_000005845.2 (4 MB)	6.11s	1s	1.49s	1.40s	0.12s
<i>caenorhabditis elegans</i> GCF_000002985.6 (101.5 MB)	2m:9.64s	5s	17.89s	43.50s	1.85s
<i>darosophila melanogaster</i> (fruit fly) GCF_000001215.4 (145.7M)	3m:3.30s	7s	25.64s	<sup>b</sup>	2.671s
<i>Xenopus tropicalis</i> (tropical clawed frog) GCF_000004195.3 (313.7 MB)	6m:25.30s	15s	51.10s	3m:8.99s	5.49s
GRCh38.p13 GCF_000001405.39 (3.3 GB)	60m:27m:498s	2m:35s	8m:58.89s	<sup>b</sup>	57.58s

<sup>a</sup> This package does not find microsatellites of size 7 core substrings.

<sup>b</sup> For some unknown reason, this package was not able to normally complete its task for some sequences.

**Table 2**  
Microsatellites of SARS-CoV-1 and novel Coronavirus-Covid-19 genomes.

	SARS-CoV-1 (accession number AY278741.1)			Coronavirus-Covid-19 (accession number NC_045512.2)		
	M.S. core	Starts from	No. of Repeats	M.S. core	Starts from	No. of Repeats
1	TTC	625	3	TTC	626	3
2	GAA	3081	3	GAA	3055	3
3	GAA	3186	3	GAA	3073	3
4	TGG	3398	3	AAG	3188	3
5	CA	4932	4	GAT	3205	3
6	TG	5553	5	CTT	4736	3
7	ATG	11296	3	TC	7813	5
8	TG	11474	4	T	11074	8
9	TGC	11562	3	ATG	11366	3
10	AC	13092	4	ATC	11910	3
11	TGA	13825	3	AC	13162	4
12	CTT	14686	3	TGA	13895	3
13	TGTT	19108	3	CTT	14756	3
14	TG	20417	4	GT	20486	5
15	TTA	20964	3	TTC	22320	3
16	CAT	21183	3	GA	22954	4
17	ATT	21833	3	AGT	23088	3
18	A	22516	8	TGT	25642	3
19	T	22568	7	AAT	25757	3
20	GA	22844	4	CGA	26191	3
21	ACA	24251	3	GTG	28556	3
22	CA	25714	4	TGC	28934	3
23	CGA	26063	3	CAA	28987	4
24	TAT	26357	4	CTG	29021	3
25	GTG	28405	3	AAG	29389	3
26	CAA	28836	4	A	29870	34
27	CTG	28870	3			

tandem repeat finder, enables biologists to keep track of the distribution and dynamics of repetitive elements in the genome of Covid-19, which is a great opportunity to watch the biological adaptation dynamics. As an example, repeats of CAG, which codes for glutamine, is shown to be unstable [29].

## 6. Summary

Coronavirus-Covid-19 onset occurred at the end of 2019, and soon after a pandemic state was declared. It acts differently than its predecessor SARS-CoV-1 which appeared in 2003. Being in the same family, it is important to explore the genomic structural differences to look for Covid-19's specific behavior. There is a hope that it might help drug repurposing for symptom relief in the short term, and vaccine design in the long term. To this aim, a highly time and space efficient software called fast microsatellite discoverer, FMSD, was developed first. Its performance was evaluated and reported to be superior than existing systems. A great novelty that makes FMSD suitable is analysis of genomic sequences of any lengths, whether short, long, or very long. Using FMSD, all microsatellites of both SARS-CoV-1 and Covid-19 were discovered and reported. There are many differences. Although the interpretations of some of the differences are provided, further investigations are needed for possible clues to treatment and drug and vaccine design. Note that tandem repeats are reported to be the cause of many diseases.

The links to other software used in this project are also provided. In addition, the links to SARS and Covid19 genomes are provided at the following links:

<https://github.com/rkmlab/perf>  
<https://github.com/ridgelab/Kmer-SSR>  
<https://github.com/lmdu/krait>  
<http://mreps.univ-mlv.fr/howto.html>  
<https://www.ncbi.nlm.nih.gov/nuccore/AY2.78741.1?report=fasta>  
[https://www.ncbi.nlm.nih.gov/nuccore/NC\\_045512.2?report=fasta](https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta).

## Acknowledgement

The authors would like to thank Dr. Hassan Shafiey for his generous genetics comments during the revision of the paper. Dr. Shafiey has received his PhD in Biophysics and his current research interests are computational biology and population genetics.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.imu.2020.100356>.

## References

- [1] Zhou P, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;579:270–3.
- [2] Wilson N, et al. Case-fatality risk estimates for COVID-19 calculated by using a lag time for fatality. *Emerg Infect Dis* 2020;26(6).
- [3] Pickett BD, et al. SA-SSR: a suffix array-based algorithm for exhaustive and efficient SSR discovery in large genetic sequences. *Bioinformatics* 2016;32(17):2707–9.
- [4] Pourcel C, et al. Identification of variable-number tandem-repeat (VNTR) sequences in acinetobacter baumannii and interlaboratory validation of an optimized multiple-locus VNTR analysis typing scheme. *J Clin Microbiol* 2011;49(2):539–48.
- [5] Farnoud F, Schwartz M, Bruck J. Estimation of duplication history under a stochastic model for tandem repeats. *BMC Bioinf* 2019;20(64):1–11.
- [6] Lang T, et al. Genome-Wide distribution of novel Ta-3A1 mini-satellite repeats and its use for chromosome identification in wheat and related species. *Agromomy* 2019;9(60):1–16.
- [7] Parson W. Age estimation with DNA: from forensic DNA fingerprinting to forensic (Epi)Genomics: a mini-review. *Gerontology* 2018;64(4):326–32.
- [8] Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome Res* 2001;11(10):1725–9.
- [9] GenBabk. GenBank; 2019.
- [10] Rota PA, et al. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 2003;300(5624):1394–9.
- [11] Cowling BJ, Leung GM. Epidemiological research priorities for public health control of the ongoing global novel coronavirus (2019-nCoV) outbreak Title'. *Euro Surveill* 2020;25(6).
- [12] Zhou Y, et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 2020;6(14).
- [13] Mir Abolfazl, Naghibzadeh Mahmoud, Saadati Nayyereh. INDEX: incremental depth extension approach for protein–protein interaction networks alignment. *Biosystems* 2017;162:24–34.
- [14] Srivastava S, et al. Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genom* 2019;20(1).
- [15] Kelkar Yogeshwar D, Eckert Kristin A, Chiaromonte Francesca, Makova Kateryna D. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res* 2011;21(12):2038–48.
- [16] Liu Qian, et al. Interrogating the “unsequenceable” genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med* 2017;9(65).
- [17] Kovtun IV, McMurray CT. Features of trinucleotide repeat instability in vivo. *Cell Res* 2008;18:198–213.
- [18] Raje M, Botre C, Ashma R. Genetic epidemiology of osteoporosis across four microsatellite markers near the VDR gene. *Int J Mol Epidemiol Genet* 2013;4(2):101–8.
- [19] Saadati N, Rajabian R. The effect of bisphosphonate on prevention of glucocorticoid-induced osteoporosis. *Iran Red Crescent Med J* 2008;10(1):8–11.
- [20] Kolpakov R, Bana G, Kucherov G. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* 2003;31(13):3672–8.
- [21] Landau GM, Schmidt JP, Sokol D. An algorithm for approximate tandem repeats. *J Comput Biol* 2001;8(1):1–18.
- [22] Genovese LM, et al. A census of tandemly repeated polymorphic loci in genic regions through the comparative integration of human genome assemblies. *Front Genet* 2018;9(155).
- [23] Pickett BD, Miller JB, Ridge PG. Kmer-SSR: a fast and exhaustive SSR search algorithm. *Bioinformatics* 2017;33(24):3922–8.
- [24] Du L, et al. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics* 2018;34(4):681–3.
- [25] Avvaru AK, Sowpati DT, Mishra RK. PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences'. *Bioinformatics* 2018;34(6):943–8.
- [26] Dong Y, et al. Epidemiological characteristics of 2143 pediatric patients with 2019 coronavirus disease in China. *Pediatrics* 2020.
- [27] Wu Zunyou, McGoogan Jennifer M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. *J Am Med Assoc* 2020.
- [28] Schein Catherine H. Polyglutamine repeats in viruses. *Mol Neurobiol* 2019;56(5):3664–75.
- [29] Perutz MF. Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends Biochem Sci* 1999;24(2):58–63.