Covid-19 Mutation Rate between Nucleotides and Future Mutation Rate Prediction using Pair-Wise Sequence Alignment

Sara A. Shehab
Faculty of Computers and Artificial Intelligence
Sadat City

ABSTRACT
Covid-19 a novel coronavirus has created a global pandemic. This is an RNA virus that all the world is immobilized with its infectious. 10 million people have been infected and 600K dead. Due to the mutation in the human body this RNA virus does, the mutation rate was estimated for infected people. NCBI Gen-Bank contains the gene sequences for infected people. These data set were tested to evaluate the nucleotide percent mutation rate. Furthermore, based on the length of the data set, the data set selected for different regions to evaluate mutation rate. The results conclude that for all regions, Thymine (T) and Adenine (A) are mutated for a huge amount of data compared to other nucleotides (C) Cytosine (G) Guanine. To predict the improvement of this virus and future mutation rate, Fact Dynamic Sequence Alignment pairwise algorithm has been applied. the predicted increment percentage in mutation rate is 0.1% for 600th patients in the future from G to T and T to C and G, C to G, while T is mutated to A with decrement of 0.1% and A mutated to C with decrement of 0.1%. these results indicate that this method can be applied efficiently to predict future gene Mutation.

Keywords
Covid-19, mutation rate, gene mutation, pair-wise sequence alignment INTRODUCTION

1. INTRODUCTION
In 2019 The World Health Organization (WHO) has declared a pandemic coronavirus disease (COVID-19) [1]. Many efforts done to stop the spread of this virus. Pandemic is defined as occurs over wide areas and affects a huge amount of population .H1N1 is the last reported pandemic in the world in 2009 [2].in January 2020, there is a new virus with unknown source was identified [3][4], it was named corona virus, from the analysis of this virus for infected cases the samples obtained. This was the cause of the outbreak, according to genetics.. In February 2020, WHO give a name of 'COVID-19' to the novel virus disease [5]. The virus is known as SARS-CoV-2, and the sickness is known as COVID-19. [6]. Coronaviruses are a type of virus that causes ailments like respiratory and gastrointestinal problems. Middle East Respiratory Syndrome (MERS-CoV), severe acute respiratory syndrome, and other respiratory infections ranging from the common cold to more serious illnesses (SARS-CoV) [7]. A novel coronavirus (nCoV) is a new strain of coronavirus that has never been seen in humans before. Scientists name coronaviruses after they figure out exactly what they are (as in the case of COVID-19, the virus causing it is SARS-CoV-2). Coronaviruses are named from the way they appear under a microscope. The virus is made up of a core of genetic material surrounded by a protein spiked envelope. This gives it a crown-like appearance. Corona is a Latin word that means "crown."

Coronaviruses are zoonotic, which means they can be passed from animals to people. MERS-CoV was found to be transmitted from dromedary camels to people, while SARS-CoV was found to be transmitted from civet cats to humans. [7]. The origins of SARS-CoV-2 (COVID-19) is still unknown, however investigations into the zoonotic genesis of the outbreak are underway [8]. Although more evidence is currently being gathered, existing knowledge suggests that human-to-human transmission is taking place. COVID-19's transmission pathways are unknown at this time, although evidence from other coronaviruses and respiratory disorders suggests that the virus could spread via large respiratory droplets and direct or indirect contact with contaminated secretions [9]. In busy settings and indoor rooms with poor ventilation, especially affected people spending a long time with others, such as a shopping mall, restaurant, etc., airborne transmission can occur. In addition, while performing medical operations, airborne transmission happens in medical care settings (aerosol-generating procedures) [9][10].

2. DATA SET ANALYSIS
The full sequence of SARS-CoV-2 is currently available in the NCBI Gen-Bank, which has a large amount of sequencing data. All of the sequences come from a human body that has been impacted by Covid-19. This information was gathered from 33 different countries. In this dataset, the four fundamental nucleotides A=adenine, G=guanine, C=cytosine, and T=thymine are represented by a 3068 data set and a reference data set of length 29903. To identify the mutation between nucleotides, the input data set will be aligned with the reference files using pairwise sequence alignment techniques.

3. GENE MUTATION
A gene mutation is a change in the DNA sequence that makes up a gene that causes it to differ from the sequence seen in most people. Mutations can range in size from a single DNA building block (base pair) to a vast stretch of a chromosome that contains many genes. In the general population, most disease-causing gene mutations are infrequent. Other genetic alterations, on the other hand, are more common. Polymorphisms are genetic variations that occur in greater than 1% of the population. They're prevalent enough to be classified as a typical DNA variant. Many of the typical variances between people, such as eye color, hair color, and blood type, are caused by polymorphisms. Although many polymorphisms have no effect on a person's health, some of

them may influence the chance of acquiring specific illnesses. A gene mutation is a change in the DNA sequence or base pairs on a chromosome that affects a gene's function. If a mutation occurs in one of these two ways, the likelihood of it being passed down to future generations increases. If a mutation develops only in isolated cells, such as in cancer, the chances of it being passed on are little to none. Mutations can be divided into five categories:

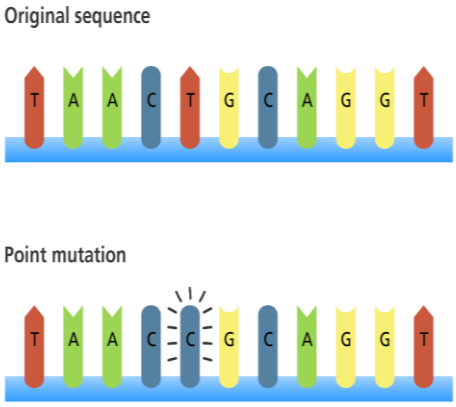
A point mutation (Fig.1(a)) is a change in one nucleotide in the DNA sequence.

Substitution (Fig.1(b)) occurs when one or more bases in a sequence are replaced by the same number of bases (for example, a cytosine for an adenine).

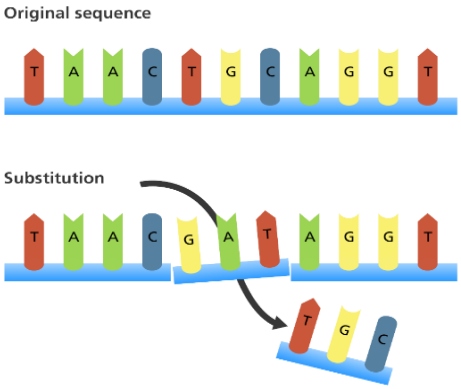
Inversion (Fig. 1(c)) occurs when a chromosomal segment is reversed end to end.

Insertion (Fig.1(d)) is when a base is introduced to the sequence.

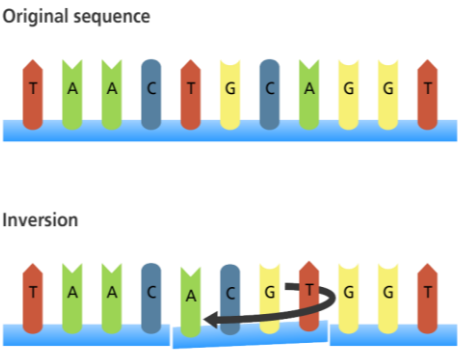
When a base is removed from a sequence, it is called deletion (Fig.1(e)).



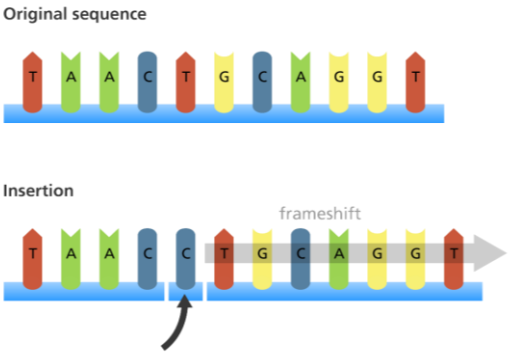
(a) Point Mutation



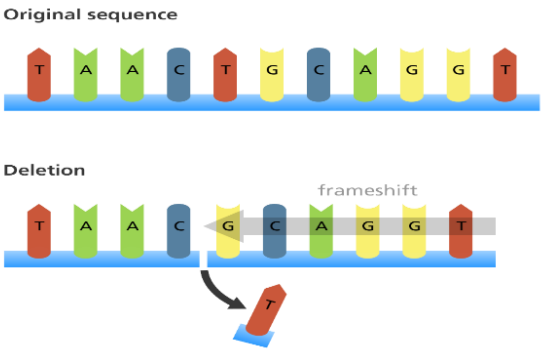
(b) Substation Mutation



(c) Inversion Mutation



(d) Insertion Mutation



(e) Deletion Mutation

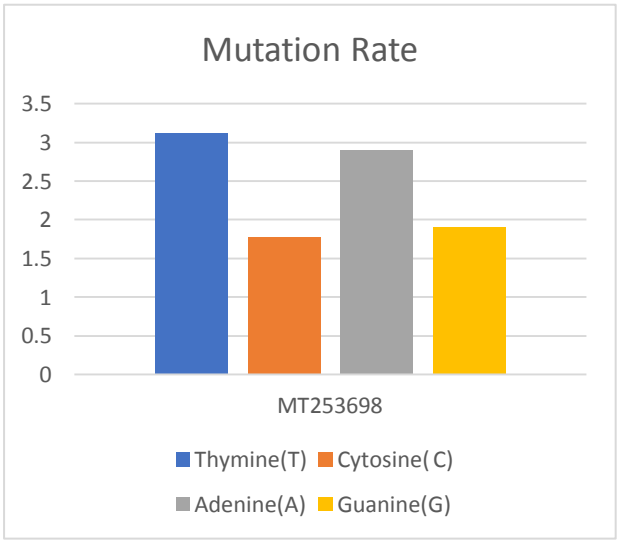
Fig 1 Mutation Types

4. MUTATION RATE EVALUATION

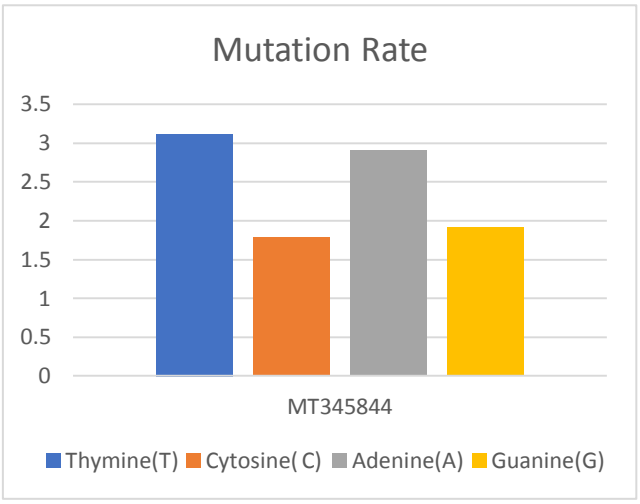
There are three types of substantiation which are missense, nonsense, and silent. A missense mutation means changes in the resulting amino acids. The nonsense mutation means that the protein becomes nonfunctional in the process of gene translation. A silent mutation is the case that the amino acids remain unchanged whatever the changing in codon. to evaluate the mutation rate between nucleotides, the mutation rate equation is used :-

$$Mutation\ Rate = \frac{Mutation}{Slen} \times 100 \quad (eq.1)$$

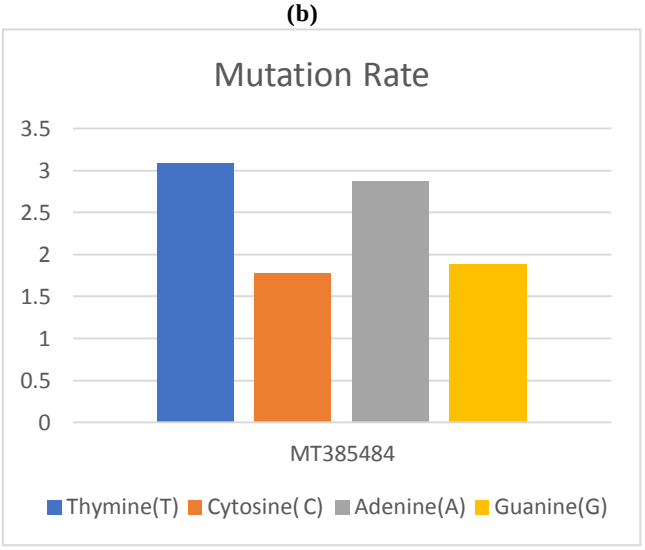
Here, the Mutation rate is the output of the applying method, Slen is the length of the data set which is 3068 for the full data set, the data selected from China is 40, the data selected from Australia is 918 , and the data selected from USA is 1903 . The output list the evaluation mutation rate for infected people. In this process, nucleotide mutation rate for the selected data set is calculated. six infected people selected with IDs 'MT345844', 'MT385484',' MT451042', 'MT253698', 'MT451810', and 'MT470120'. The results indicate that Thymine (T) and Adenine (A) record a huge percent mutation rate when compared to other nucleotides. Whereas (C) Cytosine and (G) Guanine were not changed much. The mutation rate was calculated for the different data sets. The results show that (T) Thymine and (A) Adenine for all different data have a high percent mutation rate . This indicates that this virus occurs when changes in its gene sequences. Finally, the mutation rate for different infected people has been shown in Fig.2. The mutation rate for ID 'MT253698' has been shown in Fig.2 (a). It shows that a huge percent of Thymine (T) is being mutated to other nucleotides. Also, a huge amount of Adenine (A) is mutated to other nucleotides whereas the Cytosine (C) and Guanine (G) were not changed much. After that, the mutation rate has been calculated for ID' MT345844' and ID 'MT385484'as shown in Fig.2 (b) and (c). the results conclude that all rates have a common factor of having the high mutation rate of T and A. Finally, the nucleotide mutation rate for other infected people evaluated as shown in Fig.2 (d), (e),3 -and (f). It shows the same results huge amount of (T) Thymine and (A) Adenine when compared to (C) Cytosine and (G) Guanine. All results indicate that for different data sets Thymine (T), Adenine (A), Cytosine(C) and Guanine (G) are almost equal.



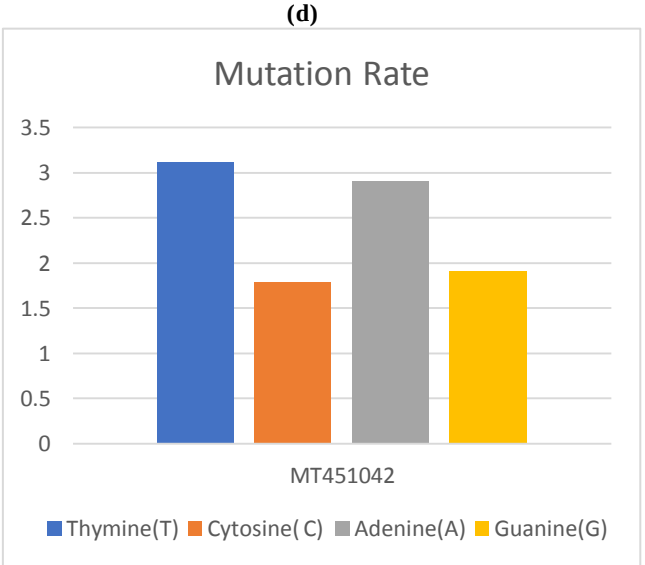
(a)



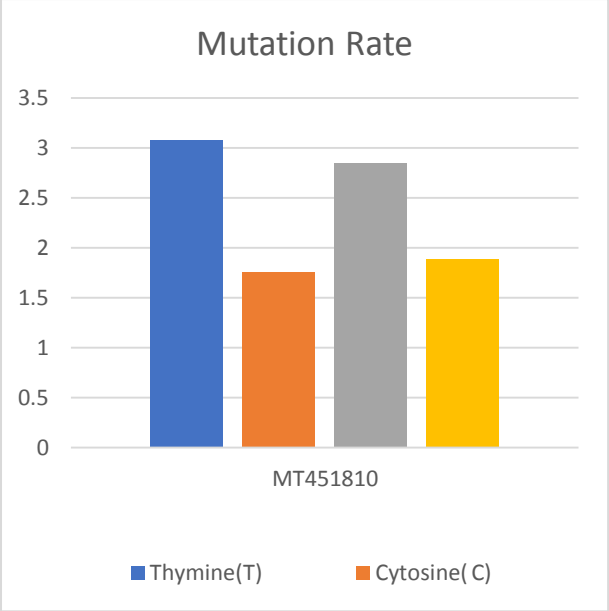
(b)



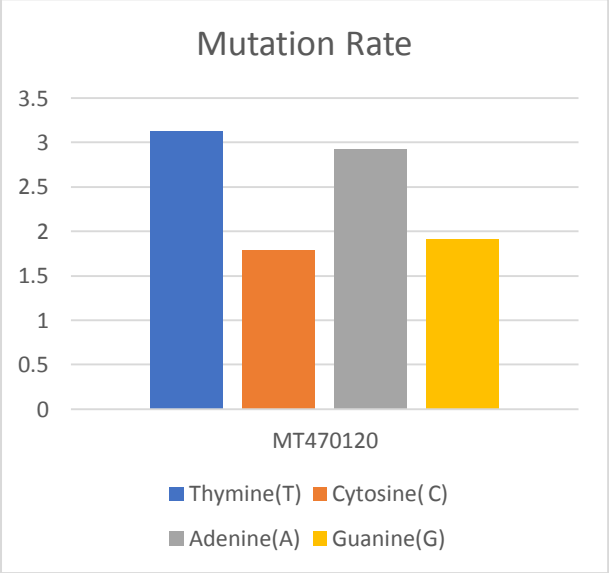
(d)



(e)



(F)



(g)

Fig 2 Mutation Rate for different PatientID

5. MUTATION RATE FUTURE PREDICTION BETWEEN NUCLEOTIDES USING SEQUENCE ALIGNMENT

In the processed data set, the affected people data set aligned with the reference data using the Fast Dynamic Sequence Alignment Algorithm. When aligning the infected people sequence with the reference data set the sequence alignment method is used as shown:-

Using the processed nucleotide data set, data from 21st January to 12th January 2021 is selected randomly. These data are arranged from older to earlier that makes it easy to evaluate the time series for data. Nearly, 3068 patients are in this data set for 100 days. Fast Dynamic Algorithm for Sequence Alignment (FDASA) is a type of pairwise alignment algorithm used to align two DNA sequences. the data set and reference data set used for aligning the process. The aligned sequences indicate the mutation rate for 100 days as seen in Fig.3. Fig.4 detect The mutation rate for 600th patients in future time. This method will be perfectly used to detect the mutation rate for patients in the future.

This type of mutation is called substation missense that means the behavior of the virus is changed when the nucleotide is change as it affected the generation of protein. This method was implemented with java ide to train the data set. Using the last six's patient mutation rate , this method is effectively predict the future mutation rate for detected patient as shown in Fig.5

Input: two DNA sequences with length M,N.

Output: Alignment between two sequences with the region of similarity and difference.

Processing:

1. Initialization

Gap= -1, Match= +1, Mismatch= -1

2. Detect the three main diagonals in the matrix.

3. Main Iteration

For each cell in the three main diagonals:

if (nucle in col =nucle in row)

$C(i, j) = (C(i-1, j-1) + match)$

if (nucle in col ≠ nucle in row)

C (i, j) =max of three rounded cells except diagonal

4. Termination

C (M, N) is the optimal score

5. Performance

Time: O (3M + 1)

Space: O (3M + 1) .

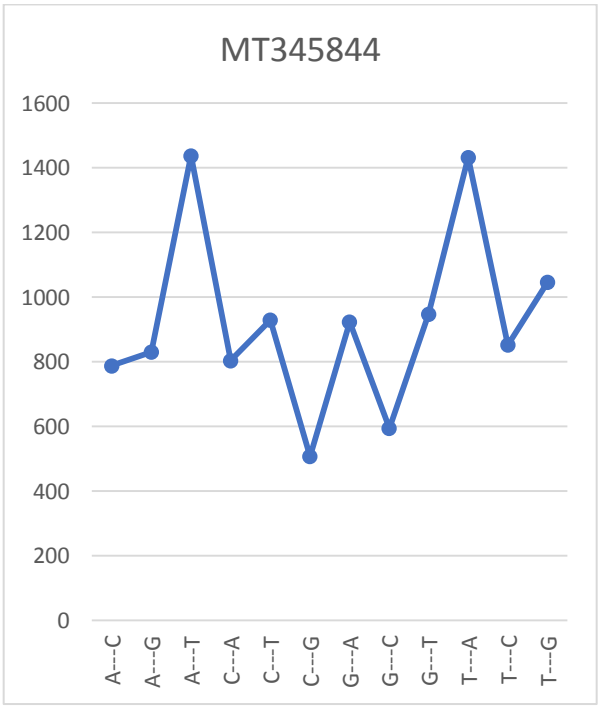


Fig 3 Mutation rate for next 100 days

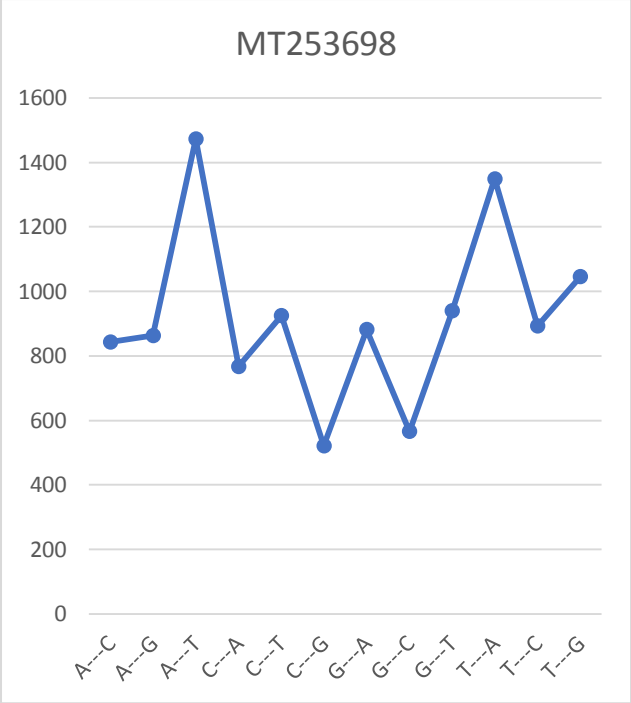


Fig 4 Mutation rate for 600th patient in future

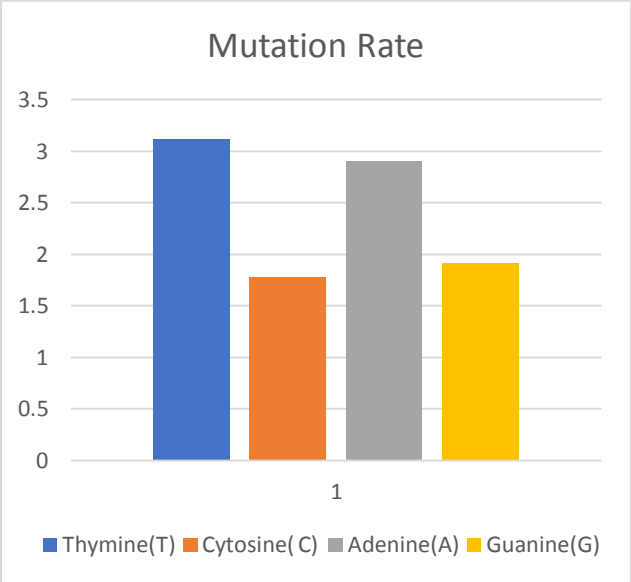


Fig 5 Mutation rate for 600th patient in future

6. CONCLUSION

In this twenty-first century, Covid-19 pandemic is the most restricted problem in the world. This virus becomes deadly and difficult due to the great mutation in nucleotides. The spreading power of this virus is temporarily limited with lock-down as the result of not inventing the vaccine , the mutation rates cannot be controlled till now. This paper discussed the mutation rate in nucleotides. The results detect that all infected people have a huge amount of (T) Thymine and (A) Adenine when compared to (G) Guanine and (C) Cytosine mutation rates. Fast pairwise alignment algorithm is used to predict the mutation rate in the person's body and detect if he affected with virus or not in the future. The Mutation rate for the 600th patient will be predicted in the future. The processed Data included from Gen-Bank with infected people. This method is a good way to predict day- bias mutation rates. In this paper, the substation, insertion, and deletion mutation rates are used.

7. REFERENCES

[1] World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (Accessed 14 March 2020).

[2] Merriam Webster Dictionary. Pandemic. Available from:https://www.merriam-webster.com/dictionary/pandemic (Accessed 14 March 2020).

[3] World Health Organisation. Novel Coronavirus – China.Disease outbreak news: Update 12 January 2020.

[4] Wikipedia. Timeline of the 2019–20 coronavirus pandemic in November 2019 – January 2020. Available from https://en.wikipedia.org/wiki/Timeline_of_the_2019%E2%80%9320_coronavirus_pandemic_in_November_2019_%E2%80%93_January_2020. [last accessed 17 March 2020]

[5] World Health Organization. Director-General's remarks at the media briefing on 2019-nCoV on 11 February 2020. 2020/2/18)[2020-02-21]. https://www. who. int/dg/speeches/detail/who-director-generals-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020. 2020.

[6] Public Health England. COVID-19: epidemiology, virology and clinical features. Available from:https://www.gov.uk/government/publications/wuhan-novel-coronavirus-background-information/wuhan-novel-coronavirus-epidemiology-virology-and-clinical-features (Accessed 14 March 2020)

[7] World Health Organization. Coronavirus. Available from: https://www.who.int/health-topics/coronavirus (Accessed 14 March 2020)

[8] Chan JF, Lau SK, To KK, Cheng VC, Woo PC, Yuen KY. Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease. Clinical microbiology reviews. 2015 Apr 1;28(2):465-522.

[9] Public Health England. COVID-19: epidemiology, virology and clinical features. Available from:https://www.gov.uk/government/publications/wuhan-novel-coronavirus-background-information/wuhan-novel-coronavirus-epidemiology-virology-and-clinical-features. (Accessed 14 March 2020)

[10] Public Health England. COVID-19: epidemiology, virology and clinical features. https://www.gov.uk/government/publications/wuhan-novel-coronavirus-background-information/wuhan-novel-coronavirus-epidemiology-virology-and-clinical-features. Accessed 14 March 2020.