

ORIGINAL RESEARCH ARTICLE

Geometrical study of DNA sequences of New World hantavirus and Nipah virus using the ATG genomic walk method

Guennadi Kouzaev*

Department of Electronic Systems, Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology, Trondheim, Norway

Abstract

Viral genome analysis is essential for characterizing genetic variation and evolutionary patterns, while complementary computational approaches may provide additional insights beyond conventional sequence-based methods. In this paper, the complete RNAs of New World hantavirus and Nipah virus are analyzed using a DNA walk method that traces the distribution of ATG triplets along sequences. It provides visual one-dimensional models (ATG trajectories) of genomes on a plane and quantitative estimates based on: (i) point-defined relative values of the divergence of ATG trajectories, (ii) relative global-defined standard deviation of ATG triplet numbers in a studied set of sequences, and (iii) relative global-defined standard deviation of fractal dimension values of inter-ATG distances in the same sequence set. These calculated parameters for the two species mentioned are compared with those of previously studied viruses, allowing them to be arranged in the following order based on their increasing sensitivity to variations in ATG distribution: Severe Acute Respiratory Syndrome Coronavirus 2, Middle East Respiratory Syndrome Coronavirus, Nipah, Ebola, and dengue viruses, which is in good accordance with their observed and recognized mutability. The method verified in this paper can be used as a supplementary pre-screening qualitative and quantitative tool for conventional sequencing analysis.

*Corresponding author:

Guennadi Kouzaev
(guennadi.kouzaev@ntnu.no)

Citation: Kouzaev G. Geometrical study of DNA sequences of New World hantavirus and Nipah virus using the ATG genomic walk method. *Gene Protein Dis.* 2026;5(3):026270023.
doi: 10.36922/GPD026270023

Received: June 29, 2026**Revised:** September 18, 2026**Accepted:** September 24, 2026**Published online:** September 30, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: ATG genomic walks; Hantavirus; Nipah; Severe Acute Respiratory Syndrome Coronavirus 2; Middle East Respiratory Syndrome Coronavirus; Ebola; Dengue viruses

1. Introduction

According to data from the International Committee on Taxonomy of Viruses (ICTV)¹, about 16,000 virus species are registered out of the millions that exist worldwide. Only 200–270 cause illness in humans, and researchers may discover additional pathogenic species in the future. Many trigger sudden outbreaks or even pandemics, depending on their contagiousness, the affected region's connectivity to other areas, the level of medical care, and vaccine availability.

Clinicians identify viral diseases through symptom review, physical exams, antibody tests, imaging, blood counts, and other tests. More advanced detection involves analyzing RNA or DNA using relatively expensive instrumentation and software.^{2,3}

Genomic sequences are strings of nucleotides of the four bases adenine (A), guanine (G), cytosine (C), and uracil (U). In place of uracil (U), thymine (T) is used; this does not affect the mathematical processing.

The typical genome length of single-stranded RNA viruses ranges from 2,000 to 30,000 nucleotides. The genome length of a double-stranded DNA virus can be substantially longer: Up to 250,000 bases. The giant viruses can have 2,500,000 nucleotides. Millions of viral samples and mutated variants, represented as ordered strings of letters, are recorded in freely accessible databases, such as GenBank⁴, Global Initiative on Sharing All Influenza Data (GISAID)⁵, National Center for Biotechnology Information Virus⁶, Virus Host DB⁷, and the Bacterial and Viral Bioinformatics Resource Center.⁸

Given the large number of nucleotides in sequences, only computational methods can extract useful information from genomes. Some operations on genomic data are performed directly on these letter strings. Others require recoding them into numbers and then applying numerical methods, such as big-data and signal-processing algorithms.^{9,10}

One of the most important techniques is visualizing genomic data, which enhances understanding of complex molecular biology.¹¹ Among these methods is the DNA walk representation.^{12–17} The basic idea is to traverse the nucleotide sequence and mark points in four-dimensional space for each of the four nucleotides. Joining these points with lines creates a trajectory (walk) in this space. Given the high dimensionality of this space, other methods have been considered to map these trajectories onto a two-dimensional plane or three-dimensional space. Many of them are discussed in a previous review.¹⁷ Unfortunately, mapping can yield rather complicated trajectories, making it difficult to understand viral peculiarities without numerical algorithms.

One method for representing genomic walks analytically was proposed in 2021.¹⁶ In this approach, the counts of consecutive occurrences of repeating patterns (for instance, ATG triplets), as well as separate A, C, G, and T nucleotides in complete sequences, are defined, and their trajectories are plotted on a single image.

To date, 200 Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), 20 Middle East Respiratory Syndrome Coronavirus (MERS-CoV), 27 dengue, and 12 Ebola viral samples using the ATG walk method have been modeled ([Appendices 2–5](#)). The first two show more compact trajectories, and the relative divergence of each studied set is at most 2%. The ATG curves of the second two viruses demonstrate divergence parameters of around

14% and 9% for the studied viral sets.

In addition to visual analysis of ATG trajectories and simple calculations, fractal dimension analyses were conducted on inter-ATG distance distributions. Dengue and Ebola viruses showed higher standard deviations in fractal dimension parameters. These findings are consistent with the fact that dengue and Ebola viruses can form not only multiple lineages but also separate subfamilies, which have not been seen in SARS-CoV-2 and MERS-CoV to date. This conclusion was made in 2021¹⁶, but the world of viruses is highly variable, and further studies using this method are needed to understand genome evolution and prepare for possible outbreaks and pandemics of viral diseases.

This paper aims to study the genomes of the New World hantavirus and Nipah virus using the ATG walk method for:

- Further validation of the ATG walk method applied to these viruses.
- Development of comparable sequence parameters allowing quantitative estimation of virus sensitivity regarding ATG variations.
- Estimation of the variability of these viruses in comparison to SARS-CoV-2, MERS-CoV, dengue, and Ebola viruses.

2. Materials and methods

In this research, we obtained complete genomic sequences of New World hantaviruses and Nipah viruses from GenBank.⁴ The processed data are compared with previously studied GenBank and GISAID⁵ available data for SARS-CoV-2, MERS CoV, Ebola, and dengue viruses.^{16,17}

2.1. Metric-based algorithm

We generated the ATG walk data using our metric-based algorithm. As stated above, for both DNA and RNA descriptions by characters, the alphabet consists of four nucleotides. We used this designation for statistical studies of RNA and DNA when their physicochemical properties were outside the scope of the research. Protein descriptions require 20 amino acids with their symbolic designations. Non-hieroglyphic human language texts have several dozen characters. Because modern programming languages can process multi-character strings, artificial intelligence applications can achieve interesting results in DNA research.

Computers use characters encoded in binary. For instance, the UTF-8 format allows encoding all 1,112,064 valid character code points and is widely used on the World Wide Web.¹⁸ In this format, the first 128 characters (United States American Standard Code for Information

Interchange) require only one byte (eight binary digits). If DNA sequences are initially represented by binary units, calculating their numerical properties can reduce time.

Because the DNA/RNA chains are now written as binary sequences, they can be characterized quantitatively using a suitable technique. One of them is to calculate a metric distance between the binary-represented symbols and a base “moving” along the characters’ chains. In particular, this distance can be the Hamming distance.^{19,20} Other metric formulas used in complex systems and coding theory also apply.^{21–24}

This procedure provides quantitative values for each move, and you can then apply any pertinent signal-processing algorithm to these results. If needed, these DNA/RNA-computed data can be quantitatively compared with human-language texts processed similarly. This approach could help us better understand DNA/RNA sequences.

Consider an algorithm that calculates the Hamming distance between a moving base and a genomic sequence of characters. This base can be ATG triplets or any given series of symbols, including a single one. Hamming distance is a metric for comparing two binary data strings; it is the number of bit positions in which the two bits differ. To calculate the Hamming distance $d_H(A, B)$ between two strings A and B , the operation $\text{XOR}(A \oplus B)$ is calculated, and the total number of ‘1s’ in the resultant string is counted.

In MATLAB, a function `pdist2(A,B,'hamming')` calculates this distance between binary strings. MATLAB

also offers other metrics. It gives a percentage of “1s” in the resultant string of Hamming’s distance calculation. If the compared strings are different in length, then Levenshtein distance²² is used; several MATLAB functions are available online.²⁴ Each symbol in an RNA sequence A of length N (number of nucleotides in a sequence) was compared to the ones in the triplet base $B: \{\text{ATG}\}$. This number should be divisible by 3 because of using triplets. Otherwise, the studied sequence is extended by additional “A”s to make the number of nucleotides divisible by 3, which does not change our ATG-walk results; otherwise, Levenshtein code²² can be used. Then, a $3N -$ string C is created using a MATLAB function `pdist2(A,B,'hamming')`. A numerical algorithm can process this new string to obtain helpful information. Particularly, we calculated the positions x_i of ATG triplets in an RNA sequence A . After that, we calculated the inter-ATG length $l_{i,i+1}^{(\text{ATG})}$ set, measured in the number of base acids.

The number of these triplets is calculated using our code and verified with a MATLAB (R2025b, The MathWorks, Inc., USA) function `count(C, 'ATG')` applied directly to the character string C . To further validate the algorithm, we compared triplet positions with those in arbitrarily chosen control RNA sequences. Figure 1 provides a clearer view of the results: nucleotide trace trajectories are shown as curves, and the positions of Adenine in ATG triplets are shown as straight vertical lines.

The advantage of this method is that base acid points are placed along simple curves, often close to the distorted sigmoid function²⁵ traces typically observed in biological

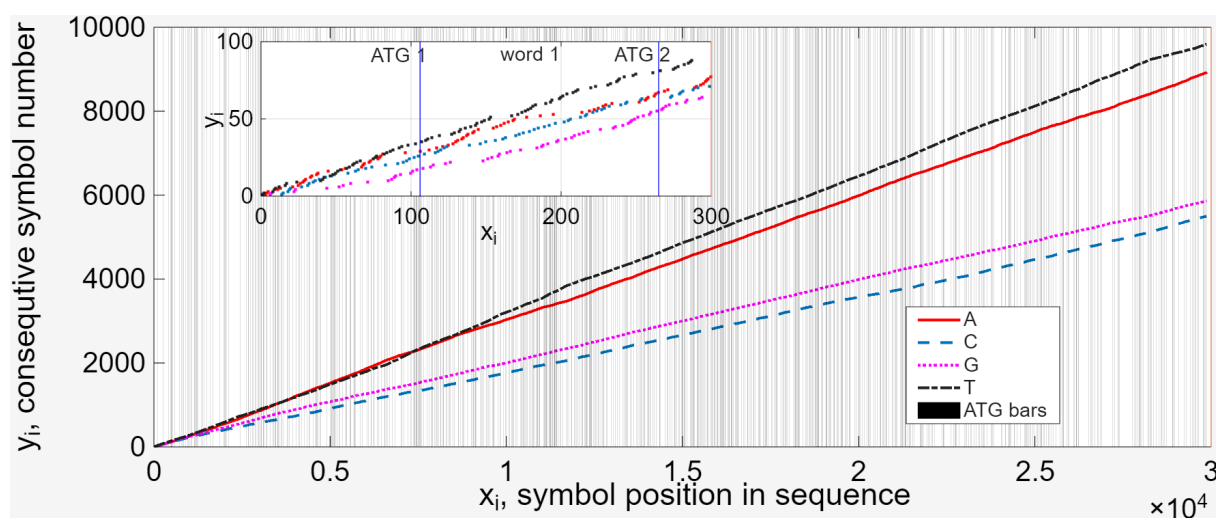


Figure 1. Combined picture¹⁶ for nucleic acid trajectories and positions of ATG triplets of a Severe Acute Respiratory Syndrome Coronavirus 2 complete MN988668 sequence (Genbank⁴). The body inset shows a magnified picture of four nucleotide positions at the beginning of this sequence, and the position of the first letter in the first two ATG triplets.

growth models.²⁶ These curves, plotted for different viral samples, enable visual discovery of genomic mutations or the application of correlation algorithms to the plots or their underlying digital data.²⁵

This paper focused mainly on ATG walks.^{17,27} In most cases, these triplets start codons in genomic sequences. Building and analyzing trajectories of these triplets allows detection of mutations, coupled with changes in codon number, variations in their length, and changes in the density of ATG distributions along the studied sequences. In the future, the algorithm is expected to be modified to exclude non-starting ATG triplets.

2.2. Study of fractal properties of ATG walks using FracLab code

Many previous studies have examined the fractality of nucleotide distributions and DNA/RNA sequences.^{28–39} Fractality means a figure has a geometric shape with detailed structure at arbitrarily small scales, and small-scale genomic motifs repeat at large scales. Thus, nucleotide distribution along a genome is not entirely random because of this long-range fractal correlation. The measure of self-similarity is the fractal dimension. Hausdorff also defines fractal dimension as a measure of the roughness of a figure or its surface. Moreover, the large genomic datasets are often patterned, and each pattern can have its own fractal dimension; i.e., the sequences can be multifractals.³⁴ It is typical in genomics as well, but it is also common in the theory of nonlinear dynamical systems, signal processing, and brain tissue morphology.^{35–43}

Several definitions and formulas exist for calculating fractal dimension. For instance, the Hausdorff dimension D_H is used for geometric figures (Equation 1):

$$D_H = \frac{\log(N)}{\log\left(\frac{1}{r}\right)} \quad (1)$$

where N is the number of smaller self-similar copies the shape is divided into, and r is the scaling factor by which each copy is reduced.

The regularization fractal dimension^{40,41} is another practical measure used to analyze real-world signals, time series, and images, and it is used in this paper. It detects roughness and abnormalities in data sets, as well as the self-similarity of patterns.

To apply any of these methods, a genomic sequence must be converted into a digital sequence representing an imaginable continuous function $f(t)$. Then these digits are treated as signal samples⁴⁴, and a fractal dimension calculation algorithm can be used.⁴⁵ Formal definition

of regularization dimension $\dim_R(\Gamma)$ can be found in Equation 2⁴⁰:

$$\dim_R(\Gamma) = 1 + \left(\lim_{a \rightarrow 0} \frac{\log(L_a)}{-\log(a)} \right) \quad (2)$$

where a is a scale, $L_a = \int_K \sqrt{1 + f'_a(t)^2} dt$, K is the finite interval for $f(t)$, and Γ is the set of samples.

In this contribution, we computed the fractal properties of genomic sequence distributions using the FracLab (2.2, French National Institute for Research in Computer Science and Automation, France) software package⁴⁶, where these parameters can be calculated. Although much research has used and validated this software, we tested it again by calculating the Weierstrass function, synthesized according to a given fractal dimension.⁴⁷ It was found that the recommended default parameters⁴⁶ (Figure 2) provide good accuracy in the computation of the mentioned function, and they were used further in our work.

Following ideas for applying signal-processing methods in genomics, we used sets of inter-ATG lengths $l_{i,i+1}^{(ATG)}$ in sequences as “sampling signals,” processed with the FracLab 2.2 software tool. An example of such a “signal” is given in Figure 3.

3. Modeling results for the New World hantavirus sequences

Hantavirus belongs to the family Hantaviridae, subfamily Mammantavirinae, and genus *Orthohantavirus*, which includes around 60 species that can cause disease in humans.^{1,48–50} The first outbreak was recorded in Korea in the 1950s, although it had long been known.⁵⁰ People can be infected through contact with rodents, bats, shrews, and moles, via their urine, droppings, saliva, or by inhaling contaminated airborne particles. Humans with Old World hantaviruses suffer from potentially fatal hemorrhagic fever with renal syndrome, and the mortality rate is between 0.3% and 10%.⁵¹

For New World hantavirus, the rodent-borne infection is most often associated with hantavirus cardiopulmonary syndrome.⁵⁰ The mortality rate reaches 40%–50%.⁵¹ One of the strains, called Andes hantavirus (*Orthohantavirus andesense*) or ANDV, can be transmitted from humans to humans⁵², and the last outbreak⁵³ was caused by this strain.

Before considering the modeling results, some peculiarities of hantavirus genomes should be noted. They consist of a short (S), medium (M), and long (L) segments

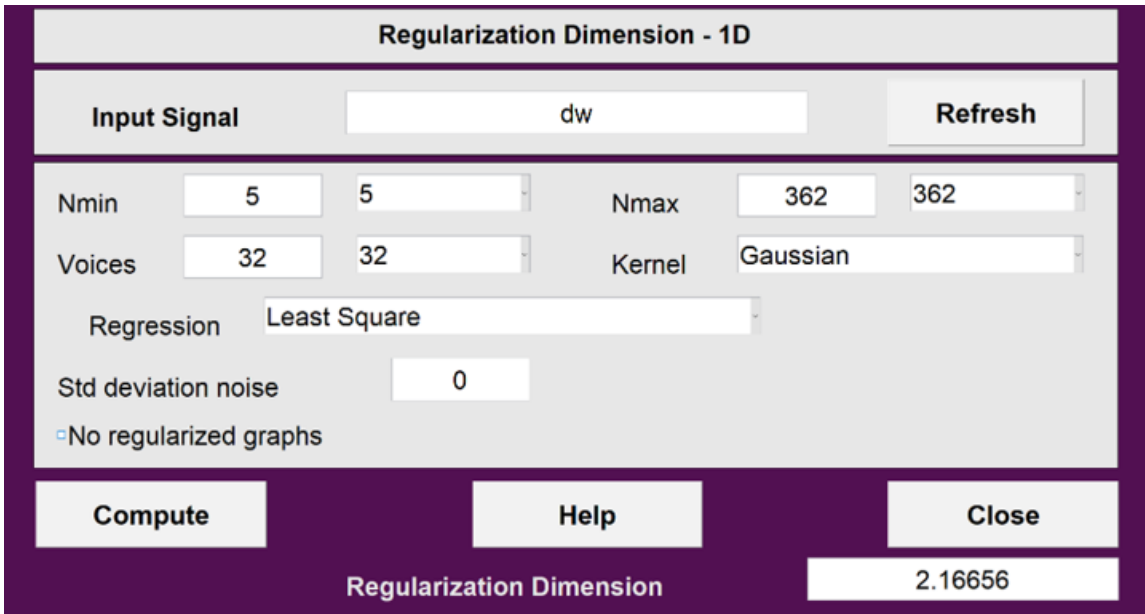


Figure 2. Control panel of FracLab 2.2 used to compute the fractal regularization dimension. The calculated dimension value is given for the MN988668 SARS-CoV-2 (GenBank) virus RNA sequence.

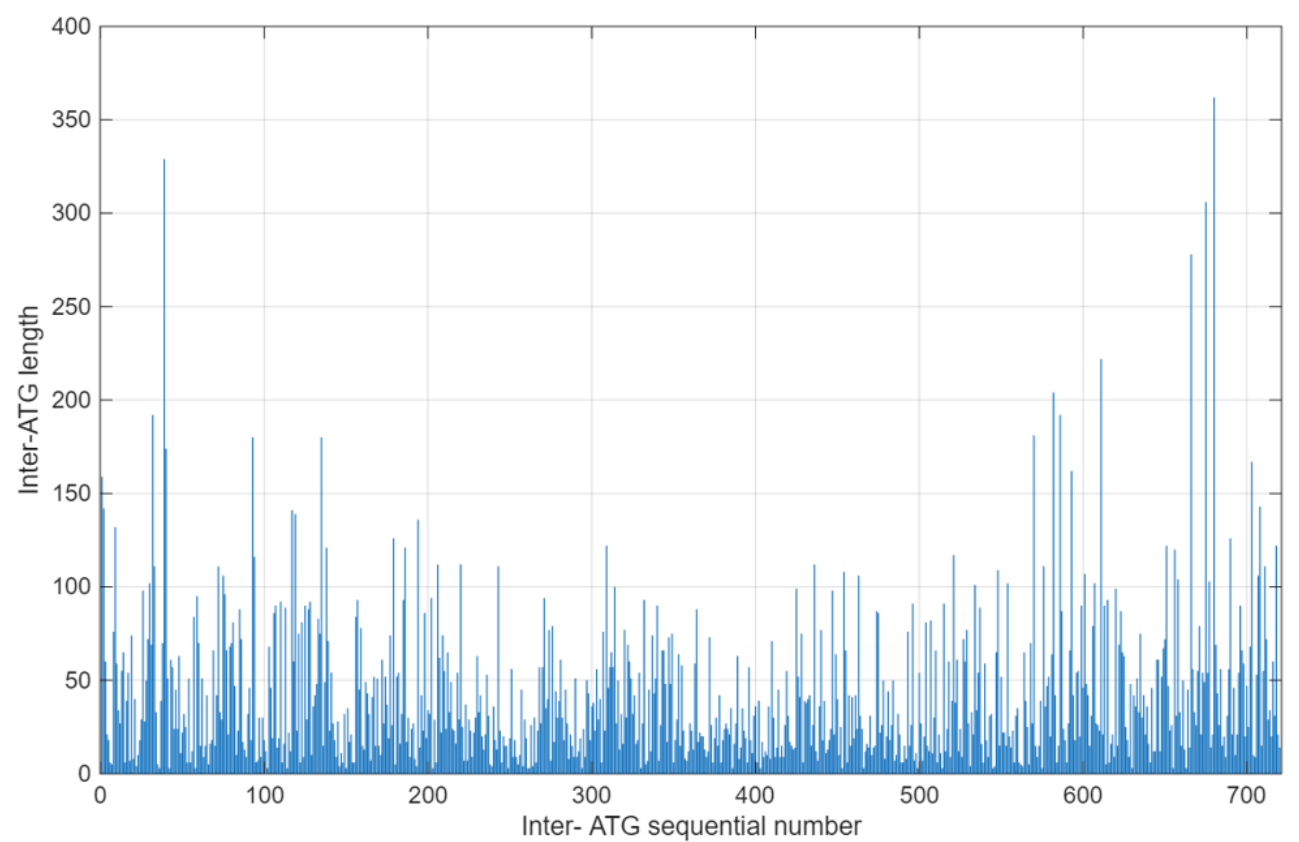


Figure 3. Distribution of inter-ATG distances in the genomic sequence hCoV-19/Norway/7448/2020 (Global Initiative on Sharing All Influenza Data)

as separate closed-loop molecules. The last is the longest and encodes the viral RNA-dependent RNA polymerase.⁵¹ This massive protein copies the genome and transcribes viral messenger RNA inside the infected cell. The segmented genome “design” may lead to a combinatorial mutation (genetic reassortment) if different viral species infect a cell. Other environment-induced mutations occur during translation and are frequent because cells have limited proofreading ability against the virus’s single-stranded RNAs.

In this paper, we modeled the complete genomes of L-sections only using the ATG walk techniques described in the Introduction. Specifically, we analyzed the most samples available from GenBank, including 20 ANDV sequences, one human-host Choclo virus, Monongahela virus, Sin Nombre virus, New York hantavirus, Bayou virus, and two mammal-host Choclo viruses (Table A1). Information on the last seven viruses is included in previous studies.^{54–59}

Figure 4A shows ATG walks (L-segment) for 28 New World hantavirus sequences (Table A1), and Figure 4B illustrates the walks only for 20 ANDV for comparison.

Each y_i point in Figure 4 corresponds to a consecutive number of an ATG triplet, with the symbol A placed at a certain x_i in the sequence.

To estimate the divergence between the curves, the relative divergence parameter $\delta_{i,N} = 2 \frac{y_N - y_1}{y_N - y_1} \cdot 100\%$ was calculated for both figures at the point $x_i = 6,218$ where y_1 is the lowest value among the traces, and the highest value is marked as y_N .

For Figure 4A, $\delta y_{1,28} = 9.7\%$, and for Figure 4B, $\delta y_{\text{Andes}} = 8.65\%$. It was observed that the 20 ANDV ATG traces are more tightly clustered than all 28 hantavirus ATG traces. This clustering is less pronounced than in SARS-CoV-2 (1.5%) and MERS-CoV (2%).²⁷

In addition to trajectory calculations, we computed fractal dimension parameters for inter-ATG distances across all 28 sequences. As shown in Figure 5, the density distribution of ATG triplets is nonuniform, and codon lengths vary in a pseudo-random manner. This distribution exhibits fractal properties¹⁶, which can be quantified. The fractal dimension of the inter-ATG length distribution

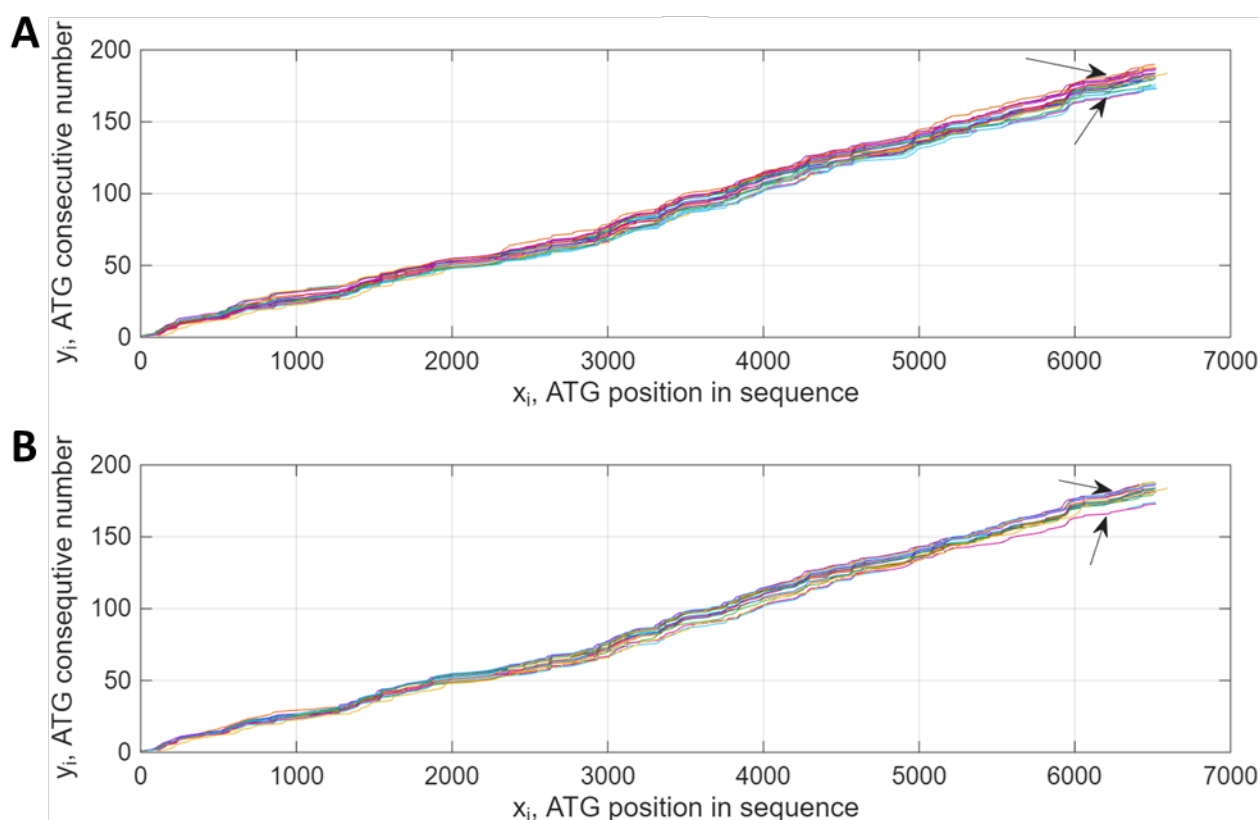


Figure 4. Distributions of ATG triplets in complete genomic sequences for (A) 28 New World hantavirus samples; (B) 20 Andes virus samples

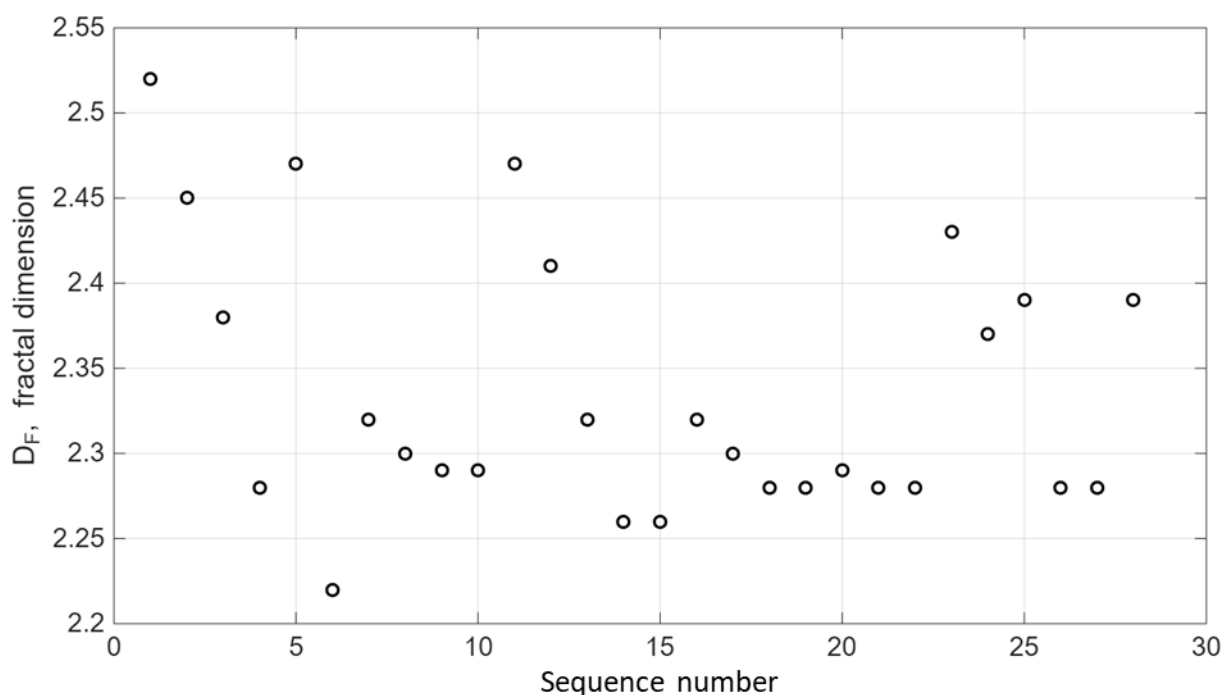


Figure 5. Fractal dimension parameter calculation results for 28 New World hantavirus complete genomic sequences (**Table A1**)

was computed using FracLab 2.2.⁴⁶ **Figure 5** presents the fractal dimension results for each hantavirus sequence, numbered as in **Table A1**.

The standard deviation D_F of this set of points is 0.076, which is higher than that of SARS-CoV-2 (0.0105) and MERS-CoV (0.0216). **Section 5** provides a more detailed discussion of the results.

4. Modeling results for the Nipah virus sequences

The Nipah virus, or *Henipavirus nipahense*, was first identified in Malaysia in 1998 and has since been found in several other Southeast Asian countries, including Bangladesh, India, Indonesia, the Philippines, and Singapore, in humans, pigs, and dogs.^{60–63} Most recently, at the beginning of 2026, this infection was registered in India once more.⁶⁰

This zoonotic virus originated in bats and spread among humans and animals through contaminated fruits. The disease can also spread between people through biological fluids and respiratory droplets in crowded hospital settings. The virus mainly affects the brain (encephalitis) and the lungs. The mortality rate ranges from 40% to 80%^{61–69}, reaching 91% during the 2018–2019 outbreak in India. Currently, no effective treatments or vaccines are available,

although many approaches are under development and research.

Because mutations occur randomly, the virus varies; for example, the genome of the Malaysian strain (NiV-M) evolves at a rate of $4.5\text{--}6.5 \times 10^{-4}$ substitutions per site per year. Additionally, a potential new cluster was suggested by studying bat and human virus genomes during the Kerala (India) 2018–2019 outbreak.⁵¹ However, only a few genomes are available in GenBank for them. Therefore, monitoring genomic variation is essential because of this virus's high mutation rate and the need to develop adaptive treatments and vaccines for this severe infection. However, only a limited number of human-origin Nipah virus sequences are available, and this study analyzes most of these samples. **Table A2** lists these genomes and their calculated parameters.

To date, all Nipah sequences are classified into several main groups: NiV-M (Malaysia), NiV-BD1, and NiV-BD2 (Bangladesh). Another group mentioned in the literature was identified after the 2018–2019 outbreak in Kerala, India.⁶⁹ This classification was based on analyses of phylogenetic trees generated using various mathematical methods and software tools^{61–71}, as well as observations of disease symptoms, death rates, and other factors.

We derived the ATG walks from 31 complete viral

genome samples from NiV-M, NiV-BD, and Kerala. Figure 6 and Table A2 show the simulation results.

Most traces are highly similar, although the cumulative differences increase toward the end of the complete sequences. RNAs carry spatial information, and differences between sequences also reflect how nucleotides are distributed within them. In addition to visual analysis, the method enables quantitative estimates. For this purpose, we used the relative width of a set of the 31 trajectories mentioned above. For comparison, the normalized cumulative difference of ATG curves was calculated at two points $x_a=12,620$ and $x_b=17,930$, respectively. It gives, correspondingly, $\delta y_{1,31} = 2(y_{31} - y_1)/(y_1 + y_{31}) \approx 5.45\%$ and 2.88% after more than 27 years of observation (1999–2025, Table A2).

In addition to the estimates of divergence for ATG walks, a comprehensive statistical parameter—the fractal regularization dimension of inter-ATG distances—was computed for all 31 genomes (Figure 7). These distances were measured in nucleotide counts, and their distributions along the genomes were analyzed using FracLab 2.2.^{16,17,27}

The calculated standard deviation for the studied Nipah virus samples was 0.0538, which was higher than that of the reference SARS-CoV-2 (0.0105) and MERS-CoV (0.0215) viruses.

5. Discussion

When conducting a comparative study of viruses, it should be noted that only a tiny fraction of them are available to be sequenced, and the world of these species is highly variable due to mutations caused by replication errors, environmental pressures, and recombination, in which RNA fragments from different viruses are combined within an infected cell.

Our algorithm can track single-nucleotide mutations, but the paper focuses on imaging ATG (AUG) triplet trajectories in complete genomic sequences; it tracks both the starting and coding triplets. Comparing ATG walks allows us to identify severe mutations that alter the number of ATGs and, therefore, the number of codons, as well as variations in their distribution along the studied sequences. For instance, the replication mechanism can misread any nucleotide in an ATG triplet, causing the codon or coding ATGs to disappear from the translated sequence.⁷² Because proofreading in mammalian cells often fails for viral RNAs, these mutated virions are released from infected cells. Genomic sequencing is not error-free, but the number of errors is decreasing as sample preparation improves, driven by advances in equipment and software. Today, these errors are below 1%.^{73,74} Alternatively, accuracy can be improved with multiple reads.

In this paper, we analyzed GenBank datasets of selected

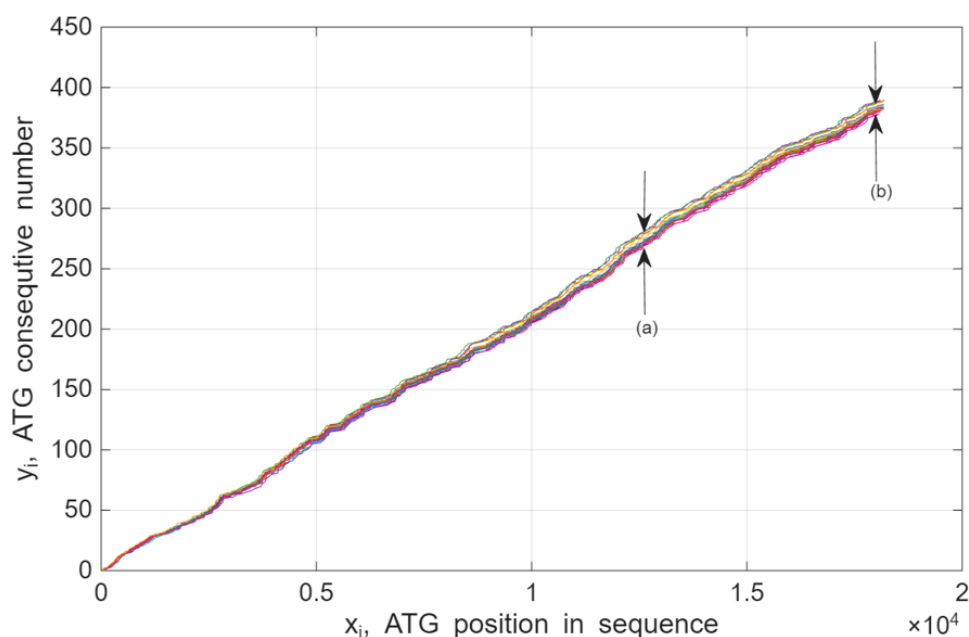


Figure 6. ATG walks for 31 complete genomic sequences. For comparison, the normalized cumulative difference between the ATG curves was calculated at the two points marked (a) and (b), respectively.

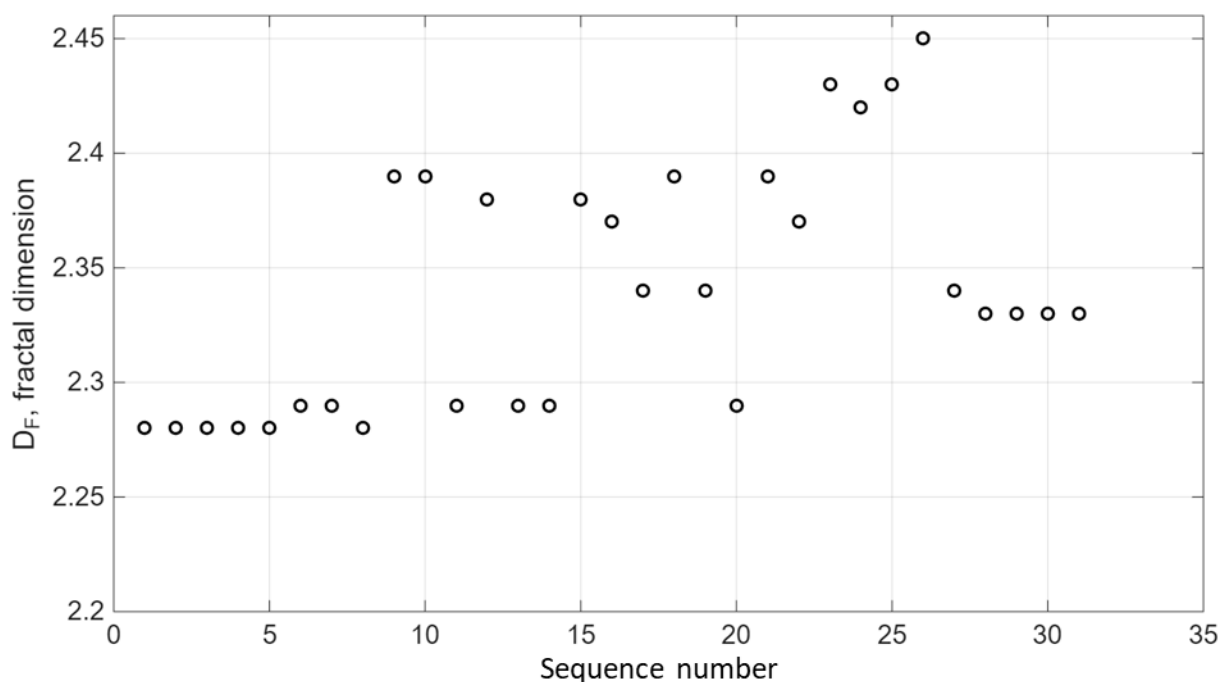


Figure 7. Fractal dimension parameter calculation results for thirty-one Nipah complete genomic sequences (Table A2)

complete genomes of New World hantaviruses and Nipah viruses (Table 1), along with previously obtained data for SARS-CoV-2 (Appendix 2), MERS-CoV (Appendix 3), dengue (Appendix 4), and Ebola (Appendix 5) viruses.^{1,27}

The viruses are compared based on their length and the relative standard deviation of ATG triplet counts $\delta_{\text{ATG}} = 100\% \cdot \sigma(N_{\text{ATG}}) / \bar{N}_{\text{ATG}}$ in the studied set. This is calculated as the ratio of the standard deviation of the number of ATG triplets in a set to the average number N of these triplets in the same set. The relative divergence values for the evaluated ATG trajectory sets are listed in row 6 and are defined at the ends of the ATG sets for each virus sequence. Additionally, the relative fractal dimension parameters $\delta_F = 100\% \cdot \sigma(D_F) / \bar{D}_F$ for the sets of ATG trajectories appear in row 7, which is the ratio of the standard deviation $\sigma(D_F)$ to the average fractal dimension $\bar{D}_F$ of the set under study.

Although MERS-CoV showed an increased mutation rate, it remained relatively stable.⁷⁵ The Ebola and dengue viruses were split into multiple strains many years ago. Our study of their ATG traces is consistent with this, showing substantial divergence among the walk curves.¹⁶

The New World hantavirus is divided into 30 lineages and strains⁵⁰, and only six of which are modeled here, given

the availability of their complete sequences of L-segments. The Nipah virus, divided into only three species (NiV-M, NiV-BD, and presumably Kerala's), is more stable than the New World hantavirus. The Ebola virus is represented here by only four of the six known strains to date: Ebola (Zaire), Bombali, Sudan, and Bundibugyo.

Notably, the dengue virus, represented in a previous study by all four strains, exhibits the highest counts across all modeled viruses.¹⁶ The computed values are consistent with the mutation peculiarities of these six viruses; they increase with viral mutability (Figure 8).

A question may arise about the representativeness of the results, given the limited number of available genomes studied. Consider, for instance, the hantavirus, dengue, Ebola, and Nipah viruses. Studying more samples of these viruses increases the relative divergence parameter δ_{1N} , but it does not change the conclusion about their general instability. The influence of adding new species on δ_{ATG} and δ_F is currently under study.

Given the randomness of mutations, future genome variations may alter the above classification, so it should be corrected. This strongly supports ongoing genomic surveillance to detect viruses that may have acquired new pathogenic properties.

Table 1. ATG walk observation results for SARS-CoV-2, MERS-CoV, dengue, Ebola, New World hantavirus, and Nipah viruses

Characteristics	SARS-CoV-2 and beta-CoV	MERS-CoV	Nipah virus	New World hantavirus L-segment	Ebola virus	Dengue virus
Observation years	2010 (bats); 2020–August 2025 (humans)	2012–2020	1999–2025	1990–2026	1976–2019	1974–2019
Typical length of sequences	~ 30,000	~ 30,000	~ 18,200	~ 6,500	~ 19,000	~ 10,000
Number of studied genomes	200	20	31	28	16	26
$\delta_{ATG}, \%$	0.6	0.76	0.98	2.44	2.58	3.65
$\delta_{y_{1,N}}, \%$	1.5	2	2.88	7.76	9	14
$\delta_F, \%$	0.5	0.9	2.3	3.3	7.4	8.4

Abbreviations: MERS-CoV: Middle East Respiratory Syndrome Coronavirus; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2.

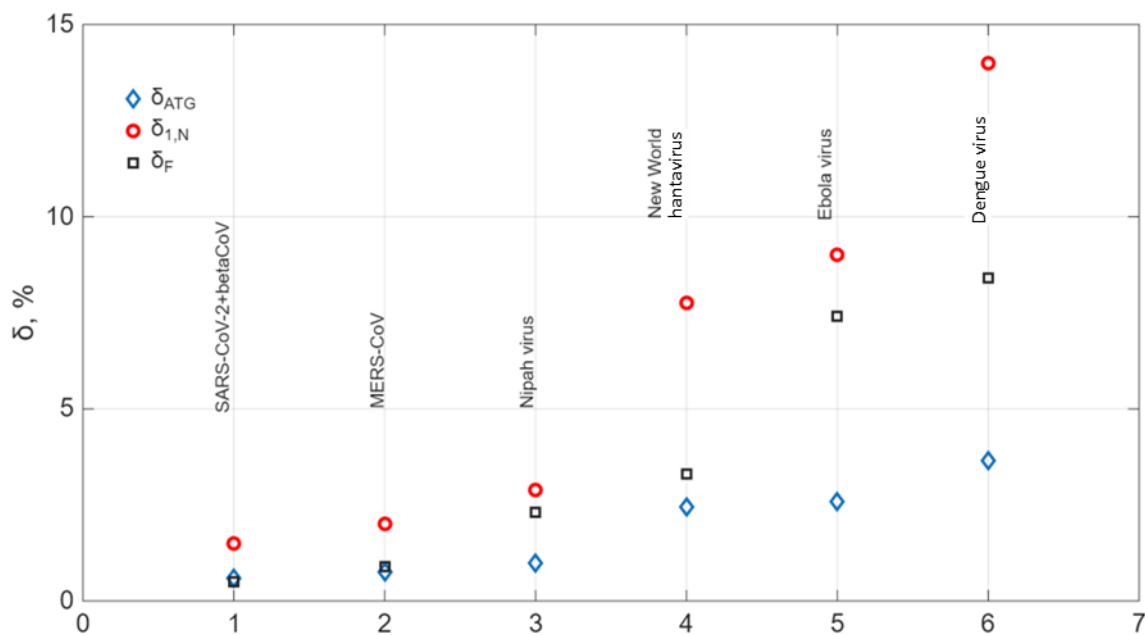


Figure 8. Relative parameter graph for the six modeled viruses

Abbreviations: MERS-CoV: Middle East Respiratory Syndrome Coronavirus; SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2.

6. Conclusion

This study applied the recently proposed ATG-walk method¹⁶, which enables quantitative comparison of different viruses, to the hantavirus and Nipah virus genomes to identify and analyze mutations, including variations in ATG numbers and their distribution along the RNA sequences.

For these viruses, we plotted ATG traces (walks) for 28 and 31 sequences, respectively, available in GenBank. Then, for their quantitative comparison, three quantitative

parameters were introduced:

- Point-defined relative divergence of trace sets $\delta_{1,N}$
- Globally defined relative standard deviation of ATG triplets δ_{ATG} in sets
- Globally defined relative standard deviation of fractal dimension values δ_F for inter-ATG distances in sets.

Comparison of these quantitative measures allowed us to arrange the studied viruses, along with the earlier modeled ones, in the following order: SARS-CoV-2, MERS-CoV, Nipah virus, New World hantavirus, Ebola virus, and dengue virus, with increasing mutation counts,

which is in good agreement with their observations. By comparing the genomes of all six viruses, we can generalize certain properties across them. Among them:

- Continuous mutations, random in their nature, along with variations in the number of ATG triplets, codons and their contents, and non-coding elements, can lead to an increase in the divergence of ATG trajectories at a certain rate.
- In some cases, specific ATG trajectories form distinct clusters (dengue, Ebola, and, presumably, Nipah), seen visually and associated with the virus mutations.
- To date, a link between the virus genomes capable of causing strong mutations and increased quantitative parameters δ_{ATG} , δ_{IN} , and δ_F considered here is confirmed once more.

The viral landscape is highly unpredictable, and ongoing surveillance of viral genomes is strongly advised to prevent pandemics, as demonstrated with SARS-CoV-2. In conclusion, the ATG walk method can serve as a supplementary pre-screening tool for conventional sequencing analysis, although further research is needed to improve this technique.

Acknowledgments

The author thanks the Faculty of Information Technology and Electrical Engineering at the Norwegian University of Science and Technology (NTNU) for supporting his 2026 spring sabbatical research stay in Canada. The author is grateful to the researchers who deposited genomic sequences in GenBank®.

Funding

None.

Conflict of interest

The author declares no competing interests.

Author contributions

This is a single-authored article.

Ethics approval and consent

Not applicable.

Consent for publication

Not applicable.

Availability of data

The genomic virus data are from GenBank (www.ncbi.nlm.nih.gov/genbank) and GISAID (<https://gisaid.org/>). Other data are available from the corresponding author

upon reasonable request.

References

1. *The International Committee on Taxonomy of Viruses (ICTV)*. Accessed on June 12, 2026. <https://ictv.global/>
2. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107(1):1-8. doi: 10.1016/j.ygeno.2015.11.003
3. *The Sequencing Buyer's Guide 8th Ed*. Front Line Genomics. Accessed on April 2, 2026. <https://frontlinegenomics.com/sequencing-buyers-guide-8th>
4. *GenBank*®. Accessed on June 2, 2026. <https://www.ncbi.nlm.nih.gov/genbank/>
5. *Global initiative on sharing all influenza data (GISAID)*. Accessed on June 22, 2026. <https://gisaid.org/>
6. *NCBI Virus*. National Library of Medicine. Accessed on June 22, 2026. <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>
7. *Virus Host DB*. <https://www.genome.jp/virushostdb/>
8. *Bacterial and Viral Bioinformatics Resource Center*. Accessed on June 21, 2026. <https://www.bv-brc.org/>
9. Lorenzo-Ginori J, Rodriguez-Fuentes A, Abalo R, Rodriguez R. Digital Signal Processing in the Analysis of Genomic Sequences. *C BIO*. 2009;4(1):28-40. doi: 10.2174/157489309787158134
10. Ishtyaq Mahmud M, Banerjee T. Artificial Intelligence in genomics: a comprehensive survey of methods, resources, challenges, and prospects. *Briefings in Bioinformatics*. 2026;27(3). doi: 10.1093/bib/bbag229
11. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T. Visualizing genomes: techniques and challenges. *Nat Methods*. 2010;7(S3):S5-S15. doi: 10.1038/nmeth.1422
12. Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *Journal of Biological Chemistry*. 1983;258(2):1318-1327. doi: 10.1016/s0021-9258(18)33196-x
13. Gates MA. Simpler DNA representation. *Nature*. 1985;316:219. doi: 10.1038/316219a0
14. Nandy A, Dey S, Basak SC, Bielińska-Wąż D, Wąż P. Characterization of the Zika virus genome - a bioinformatics study. *Curr Comp Aided Drug Design*. 2016;12(2):87-97. doi: 10.2174/1573409912666160401115812
15. Bielińska-Wąż D, Wąż P, Panas D. Applications of 2D and 3D-dynamic representations of DNA/RNA sequences for

- description of genome sequences of viruses. *Comb Chem High Throughput Screen*. 2022; 25(3):429-438.
doi: 10.2174/1386207324666210804120454
16. Belinsky A, Kouzaev G. Visual and quantitative analyses of virus genomic sequences using a metric-based algorithm. *WSEAS Trans Circ Syst*. 2022;21:321-348.
doi: 10.37394/23201.2022.21.35
 17. Belinsky A, Kouzaev GA. DNA walks in virus genomics. *JP J Biostatistics*. 2024;24(2):251-286.
doi: 10.17654/0973514324017
 18. The Unicode Consortium. Chapter 2: General Structure. In: *The Unicode Standard*. 6th ed. The Unicode Consortium; 2010.
 19. Hamming RW. Error-detecting and error-correcting codes. *Bell Syst Techn J*. 1950;29(2):147-60.
doi: 10.1002/j.1538-7305.1950.tb00463.x
 20. Waggner B. *Pulse Code Modulation Techniques*. Berlin-Heidelberg: Springer-Verlag; 1995.
 21. Navarro G, Raffinot M. *Flexible Pattern Matching in Strings: Practical Online Search Algorithms for Texts and Biological Sequences*. Cambridge: Cambridge University Press; 2002.
 22. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. 1966;10(8):707-710.
 23. Gabidullin E. Theory of codes with maximum rank distance. *Probl Inf Trans*. 1985;21(1):1-76.
 24. Polityko E. *Calculation of distance between strings*. MATLAB Central File Exchange. Retrieved March 3, 2021. <https://www.mathworks.com/matlabcentral/fileexchange/17585-calculation-of-distance-between-strings>
 25. Kyurchiev N, Markov S. *Sigmoid functions: Some approximations and modelling aspects*. Lambert Acad. Publ; 2015.
 26. Ling Y, He B. Entropic analysis of biological growth models. *IEEE Trans. Biomed Eng*. 1993;40(12):1193-1200.
doi: 10.1109/10.250574
 27. Kouzaev GA. ATG walks in virus genomics. In: *Proceeding of the 2nd International Conference Infectious Diseases and Applied Microbiology and Beneficial Microbes*, Vienna, Austria, September 25-26, 2025;46. https://www.researchgate.net/publication/396045850_ATG_Walks_in_Virus_Genomics
 28. Cattani C. Fractals and hidden symmetries in DNA. *Math Problems Eng*. 2010;2010(1).
doi: 10.1155/2010/507056
 29. Voss RF. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys Rev Lett*. 1992;68(25):3805-3808.
doi: 10.1103/physrevlett.68.3805
 30. Abramson G, Cerdeira HA, Bruschi C. Fractal properties of DNA walks. *Biosystems*. 1999;49(1):63-70.
doi: 10.1016/s0303-2647(98)00032-x
 31. Hao B, Lee HT, Zhang S. Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals*. 2000;11(6):825-836.
doi: 10.1016/S0960-0779(98)00182-9
 32. Su Z-Y, Wu T, Wang S-Y. Local scaling and multifractality spectrum analysis of DNA sequences-GenBank data analysis. *Chaos, Solitons and Fractals*. 2009;40(4):1750-1765.
doi: 10.1016/j.chaos.2007.09.078
 33. Durán-Meza G, López-García J, del Río-Correa JL. The self-similarity properties and multifractal analysis of DNA sequences. *Appl Math Nonlin Sci*. 2019;4(1):267-278.
doi: 10.2478/AMNS.2019.1.00023
 34. Birdi KS. *Fractals in Chemistry, Geochemistry, and Biophysics*. N.Y.: Plenum Press; 1993.
 35. Dewey TG. *Fractals in Molecular Biophysics*. Cambridge: Oxford University Press; 1997.
 36. Feder J. *Fractals*. N.Y.: Plenum Press; 1988.
 37. Grassberger P, Procaccia I. Measuring the strangeness of strange attractors. *Physica D*. 1983;9(1-2):189-208.
doi: 10.1016/0167-2789(83)90298-1
 38. Rasband SN. *Chaotic Dynamics of Nonlinear Systems*. Weinheim: J. Wiley & Sons; 1989.
 39. Henry B, Lovell N, Camacho F. Nonlinear Dynamics Time Series Analysis. In: *Nonlinear Biomedical Signal Processing, Dynamic Analysis and Modeling*. IEEE; 2000.
doi: 10.1109/9780470545379.ch1
 40. Roueff F, Véhel JL. A regularization approach to fractional dimension estimation. In: *Proceedings of Fractals 98*. World Scientific; 1998:1-14.
 41. Véhel JL, Legrand P. Signal and image processing with Fraclab, In: *Thinking in Patterns*. World Scientific; 2004: 321-322.
doi: 10.1142/9789812702746_0032
 42. Kouzaev GA. *Application of Advanced Electromagnetics, Components and Systems*. In: *Lecture Notes in Electrical Engineering*. Berlin-Heidelberg: Springer; 2013.
 43. Guidolin D, Tortorella C, De Caro R, Agnati LF. Does a self-similarity logic shape the organization of the nervous system? In: Di Leva A, ed. *The Fractal Geometry of the Brain*. Springer; 2016:137-156. https://link.springer.com/chapter/10.1007/978-1-4939-3995-4_9
 44. Vaidyanathan P, Yoon BJ. The role of signal-processing concepts in genomics and proteomics. *J Franklin Inst*.

- 2004;341(1-2):111-135.
doi: 10.1016/j.jfranklin.2003.12.001
45. Karmi A, Najafi A, Gifani P, Khakabimamaghani S. Fractal analysis of DNA by nonlinear genome signal processing for exon and intron separation. *Annual Research & Review in Biology*. 2014;4(4): 699-708.
doi: 10.9734/ARRB/2014/6453
 46. FracLab 2.1. *A fractal analysis toolbox for signal and image processing*. <https://project.inria.fr/fracclab/>
 47. Monge-Álvarez J. *Weierstrass Cosine Function (WCF)*. MATLAB Central File Exchange. Retrieved March 21, 2021. <https://www.mathworks.com/matlabcentral/fileexchange/50292-weierstrass-cosine-function-wcf>
 48. Bradfute SB, Calisher CH, Klempa B, Klingstrom J, Kuhn JH, Laenen L, *et al*. ICTV Virus Taxonomy Profile: Hantaviridae 2024. *J Gen Virology*. 2024;105(4):001975.
doi: 10.1099/jgv.0.001975
 49. Kuhn JH, Schmaljohn CS. A brief history of Bunyaviral family *Hantaviridae*. *Diseases*. 2023;11(1):38.
doi: 10.3390/diseases11010038
 50. Zupanc TA, Saksida A, Korva M. Hantavirus infections. *Clinical Microbiology and Infection*. 2019;21:e6-e16.
doi: 10.1111/1469-0691.12291
 51. Muyangwa M, Martynova EV, Khaiboullina SF, Morzunov SP, Rizvanov AA. Hantaviral proteins: structure, functions, and role in Hantavirus infection. *Front Microbiol*. 2015;6:1326.
doi: 10.3389/fmicb.2015.01326
 52. Martínez VP, Di Paola N, Alonso DO, *et al*. “Super-spreaders” and person-to-person transmission of Andes virus in Argentina. *N Engl J Med*. 2020;383(23):2230-2241.
doi: 10.1056/NEJMoa2009040
 53. Velavan TP, Schmidt-Chanasit J. When rare zoonoses travel: Andes virus, Hantavirus cardiopulmonary syndrome, and the preparedness gap. *Int J Infect Disease*. 2026;169:108778.
doi: 10.1016/j.ijid.2026.108778
 54. Nelson R, Cañate R, Pascale JM, *et al*. Confirmation of Choclo virus as the cause of hantavirus cardiopulmonary syndrome and high serum antibody prevalence in Panama. *J Med Virol*. 2010;82(9):1586-1593.
doi: 10.1002/jmv.21864
 55. Gonzalez P, Salazar JR, Salinas TP, *et al*. Two decades of wildlife pathogen surveillance: Case study of Choclo orthohantavirus and its wild reservoir *oligoryzomys costaricensis*. *Viruses*. 2023;5(6):1390.
doi: 10.3390/v15061390
 56. Albariño CG, Guerrero LW, Chakrabarti AK, Rollin PE, Nichol ST, Newton ILG. Complete genome sequences of Monongahela Hantavirus from Pennsylvania, USA. *Microbiol Resour Announc*. 2018;7(11):e00928-18.
doi: 10.1128/MRA.00928-18
 57. Meier K, Thorkelsson SR, Durieux Trouillette Q, *et al*. Structural and functional characterization of the Sin Nombre virus L protein. *PLoS Pathog*. 2023;19(8):e1011533.
doi: 10.1371/journal.ppat.1011533
 58. McMullan LK, Albariño CG, Ksiazek TG, Nichol ST, Spiropoulou CF. Complete genome sequences of a Hantavirus isolate from New York. *Genome Announc*. 2018;6(12):e00188-18.
doi: 10.1128/genomeA.00188-18
 59. Ortega E, Simonson S, Shedroff E, *et al*. Bayou Hantavirus Cardiopulmonary Syndrome, Louisiana, USA, 2022–2023. *Emerg Infect Dis*. 2025;31(2):401-403.
doi: 10.3201/eid3102.241069
 60. World Health Organization (WHO). *Nipah virus fact sheet*. Accessed on April 2, 2026. <https://www.who.int/news-room/fact-sheets/detail/nipah-virus>
 61. Madhukalya R, Yadav U, Parray HA, *et al*. Nipah virus: pathogenesis, genome, diagnosis, and treatment. *Appl Microbiol Biotechnol*. 2025;109(1):158.
doi: 10.1007/s00253-025-13474-6
 62. Asokan S, Luke MS, Atiyah HM, *et al*. Nipah virus as a pandemic threat: Current knowledge, diagnostic gaps, and future research priorities. *Diagnostic Microbiology and Infectious Disease*. 2026;114(2):117141.
doi: 10.1016/j.diagmicrobio.2025.117141
 63. Branda F, Ceccarelli G, Giovanetti M, *et al*. Nipah Virus: A Zoonotic Threat Re-Emerging in the Wake of Global Public Health Challenges. *Microorganisms*. 2025;13(1):124.
doi: 10.3390/microorganisms13010124
 64. Tan FH, Sukri A, Idris N, *et al*. A systematic review on Nipah virus: global molecular epidemiology and medical countermeasures development. *Virus Evolution*. 2024;10(1).
doi: 10.1093/ve/veae048
 65. Whitmer SLM, Lo MK, Sazzad HMS, *et al*. Inference of Nipah virus evolution, 1999–2015. *Virus Evolution*. 2020;7(1).
doi: 10.1093/ve/veaa062
 66. Rahman MdM, Miah M, Hossain ME, *et al*. Development of a culture-independent whole-genome sequencing of Nipah virus using the MinION Oxford Nanopore platform. *Microbiol Spectr*. 2025;13(6).
doi: 10.1128/spectrum.02492-24
 67. Cortes-Azuero O, Lefrancq N, Nikolay B, *et al*. The genetic diversity of Nipah virus across spatial scales. *J Infect Dis*.

- 2024;230(6):e1235-e1244.
doi: 10.1093/infdis/jiae221
68. Hauser N, Gushiken AC, Narayanan S, Kottlil S, Chua JV. Evolution of Nipah Virus Infection: Past, Present, and Future Considerations. *TropicalMed*. 2021;6(1):24.
doi: 10.3390/tropicalmed6010024
69. Sudeep AB, Yadav PD, Gokhale MD, *et al*. Detection of Nipah virus in Pteropus medius in 2019 outbreak from Ernakulam district, Kerala, India. *BMC Infect Dis*. 2021;21(1).
doi: 10.1186/s12879-021-05865-7
70. Lv C, He J, Zhang Q, Wang T. Vaccines and Animal Models of Nipah Virus: Current Situation and Future Prospects. *Vaccines*. 2025;13(6):608.
doi: 10.3390/vaccines13060608
71. Faus-Cotino J, Reina G, Pueyo J. Nipah Virus: A Multidimensional Update. *Viruses*. 2024;16(2):179.
doi: 10.3390/v16020179
72. Turki TG, Ghayyib AA. Mechanisms leading to the loss of the AUG codon function as a translation initiator: A review. *Gene Protein Dis*. 2026.
doi: 10.36922/GPD025350065
73. Ma X, Shao Y, Tian L, *et al*. Analysis of error profiles in deep next-generation sequencing data. *Genome Biol*. 2019;20(1):50.
doi: 10.1186/s13059-019-1659-6
74. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genome Bioinform*. 2021;3(1):1-9.
doi: 10.1093/nargab/lqab019
75. Dempsey R, Goldswain H, Newman J, *et al*. Characterisation of Naturally Occurring MERS-CoV Spike Mutations and Their Impact on Fusion and Neutralisation. *Viruses*. 2026;18(3):377.
doi: 10.3390/v18030377

Appendix

Appendix 1

Table A1. New World hantavirus complete genomic sequence data for L-segments

No	GenBank number and L-segment name	Collection year	Number of nucleotides	Number of ATG triplets	D_F
1	PP058739.1: <i>Orthohantavirus andesense</i> isolate 13796_ L Argentina 1990	1990	6,537	184	2.52
2	MG717393.1: <i>Orthohantavirus sinnombreense</i> strain New York 1 segment L, complete sequence, Host: White-footed mouse	1994	6,561	190	2.45
3	NC_005217.1: Sin Nombre virus map viral genome L segment, USA	1995	6,562	180	2.38
4	MN258156.1: <i>Orthohantavirus andesense</i> isolate Epilink/96_2/6/19, Argentina	1996	6,546	184	2.28
5	MH539865.1: <i>Monongahela hantavirus</i> isolate USA_ PA_1997 segment L, complete sequence	1997	6,562	184	2.47
6	PP058738.1: <i>Orthohantavirus andesense</i> isolate 807266_ L Paraguay 1997	1997	6,537	184	2.22
7	PP058737.1: <i>Orthohantavirus andesense</i> isolate 15631_ L Argentina 1998	1998	6,540	181	2.32
8	KT983773.1: Choclo virus strain 588 segment L, complete sequence, Panama	2000	6,562	173	2.30
9	AF291704.5: Andes virus strain Chile-9717869	2002	6,562	182	2.29
10	NC_003468.2: Andes virus strain Chile-971869	2002	6,562	182	2.29
11	PP210993.1: MAG: <i>Orthohantavirus chocloense</i> isolate MSB: Mamm:264444, Panama	2006	6,505	176	2.47
12	PP210996.1: MAG: <i>Orthohantavirus chocloense</i> isolate MSB: Mamm:264422, Panama	2006	6,551	176	2.41
13	MN850096.1: <i>Orthohantavirus andesense</i> strain ARCH14/ NRC1, Argentina	2014	6,531	183	2.32
14	MN850095.1: <i>Orthohantavirus andesense</i> strain AREB14/ P3, Argentina	2014	6,528	188	2.26
15	MN850094.1: <i>Orthohantavirus andesense</i> strain AREB14/ P2, Argentina	2014	6,562	188	2.26
16	MN850093.1: <i>Orthohantavirus andesense</i> strain AREB14/ P1, Argentina	2014	6,464	187	2.32
17	MN258158.1: <i>Orthohantavirus andesense</i> isolate NRC- 3/18_06/01/1997, Argentina	2018	6,540	186	2.3
18	MN258165.1: <i>Orthohantavirus andesense</i> isolate Epuyen/18-19_Patient_13_12/24/18 Argentina	2018	6,546	184	2.28
19	MN258166.1: <i>Orthohantavirus andesense</i> isolate Epuyen/18-19_Patient_14_12/26/18, Argentina	2018	6,543	184	2.28
20	MN850097.1: <i>Orthohantavirus andesense</i> isolate Epuyen/18-19_Patient_14_12/26/18, Argentina	2018	6,543	184	2.29
21	PP003836.1: <i>Orthohantavirus andesense</i> isolate BA18-RR1620, Argentina	2018	6,558	174	2.28
22	PP003837.1: <i>Orthohantavirus andesense</i> isolate BA18- LD1633, Argentina	2018	6,587	173	2.28

(Cont'd...)

Table A1. (Continued)

No	GenBank number and L-segment name	Collection year	Number of nucleotides	Number of ATG triplets	D_F
23	OQ092243.1: <i>Andes orthohantavirus</i> strain VARS/22-01, Brazil	2022	6,538	180	2.43
24	PP639088.1: <i>Orthohantavirus bayoui</i> isolate 202204786/10039/Bayou virus/Louisiana_USA/2022 segment L, complete sequence	2022	6,539	180	2.37
25	PP639090.1: <i>Orthohantavirus bayoui</i> isolate 202300391/10142/Bayou virus/Louisiana_USA/2023 segment L, complete sequence	2023	6,537	180	2.39
26	PV808478.1: <i>Orthohantavirus andesense</i> isolate CHI-Hu13724_P2, Chile	2024	6,555	184	2.28
27	PV808475.1: <i>Orthohantavirus andesense</i> isolate CHI-Hu13724, Chile	2024	6,562	184	2.28
28	ANDV/Switzerland/Hu-3337/2026_L	2026	6,537	180	2.390

Abbreviations: ANDV: Andes Hantavirus; MAG: Metagenome-assembled genome;

Table A2. Nipah virus (*Henipavirus nipah*) complete genomic sequence data

No	GenBank number and virus complete name	Collection year	Number of nucleotides	Number of ATG triplets	D_F
1	AJ564623.1: Nipah virus complete genome, isolate NV/MY/99/UM-0128, Malaysia	1999	18,246	384	2.28
2	MK673562.1: <i>Henipavirus nipahense</i> isolate 808651, Malaysia	1999	18,231	383	2.28
3	AF212302.2: Nipah virus, Malaysia	2001	18,246	384	2.28
4	AY029767.1: Nipah virus isolate UMMC1, Malaysia	2001	18,246	384	2.28
5	AY029768.1: Nipah virus isolate UMMC2, Malaysia	2001	18,246	384	2.28
6	MK673565.1: <i>Henipavirus nipahense</i> isolate 810405, Bangladesh	2004	18,235	384	2.29
7	MK673567.1: <i>Henipavirus nipahense</i> isolate 810482, Bangladesh	2004	18,245	383	2.29
8	AY988601.1: Nipah virus from Bangladesh	2005	18,252	384	2.28
9	FJ513078.1: Nipah virus isolate Ind-Nipah-07-FG from India	2007	18,252	378	2.39
10	FJ513078.1: Nipah virus isolate Ind-Nipah-07-FG from India	2007	18,252	378	2.39
11	MK673568.1: Nipah virus isolate Ind-Nipah-07-FG from India	2008	18,237	386	2.29
12	JN808863.1: Nipah virus isolate NIVBGD2008RAJBARI, Bangladesh	2008	18,242	372	2.38
13	MK673571.1: <i>Henipavirus nipahense</i> isolate 201200903, Bangladesh	2011	18,240	382	2.29
14	MK673573.1: <i>Henipavirus nipahense</i> isolate 201200919, Bangladesh	2011	18,236	382	2.29
15	MK673578.1: <i>Henipavirus nipahense</i> isolate 201206119, Bangladesh	2011	18,235	381	2.38
16	MK673581.1: <i>Henipavirus nipahense</i> isolate 201206316, Bangladesh	2012	18,235	381	2.37
17	MK673591.1: <i>Henipavirus nipahense</i> isolate 201206316, Bangladesh	2014	18,157	380	2.34
18	MK673592.1: <i>Henipavirus nipahense</i> isolate 201601241, Bangladesh	2014	18,234	380	2.39
19	MK673590.1: <i>Henipavirus nipahense</i> isolate 201601191, Bangladesh	2014	18,208	383	2.34
20	MK673584.1: <i>Henipavirus nipahense</i> isolate 201601126, Bangladesh	2015	18,229	385	2.29

(Cont'd...)

Table A2. (Continued)

No	GenBank number and virus complete name	Collection year	Number of nucleotides	Number of ATG triplets	D_F
21	MK673585.1: <i>Henipavirus nipahense</i> isolate 201601136, Bangladesh	2015	18,245	381	2.39
22	MH523642.1: <i>Henipavirus nipahense</i> strain MCL-18-H-1088, India	2018	18,242	390	2.37
23	MH396625.1: <i>Henipavirus nipahense</i> strain MCL-18-H-1088, India	2018	18,210	389	2.43
24	MH523640.1: <i>Henipavirus nipahense</i> strain MCL-18-H-1197, India	2018	18,132	390	2.42
25	MH523641.1: <i>Henipavirus nipahense</i> strain MCL-18-H-1209, India	2018	18,027	388	2.43
26	PP981674.1: <i>Henipavirus nipahense</i> isolate NiV/human/BGD/FP0260922/2022, Bangladesh	2022	18,077	381	2.45
27	PP981675.1: <i>Henipavirus nipahense</i> isolate NiV/human/BGD/RS0560822/2022, Bangladesh	2022	18,096	381	2.34
28	PP981676.1: <i>Henipavirus nipahense</i> isolate NiV/human/BGD/RS0644723/2023, Bangladesh	2023	18,092	378	2.33
29	PQ368168.1: <i>Henipavirus nipahense</i> isolate NiV/human/BGD/NINS00123U/2023, Bangladesh	2023	18,090	381	2.33
30	PV132707.1: <i>Henipavirus nipahense</i> isolate NiV/human/BGD/PMSKH0000424/2024, Bangladesh	2024	18,128	381	2.33
31	PX130166.1: <i>Henipavirus nipahense</i> isolate NiV/human/BGD/FP0372425/2025, Bangladesh	2025	18,128	381	2.33

Appendix 2. List of 200 Severe Acute Respiratory Syndrome Coronavirus 2 complete genomic sequences.

BetaCoV/bat/Cambodia/RShSTT182/2010, BetaCoV/bat/Cambodia/RShSTT200/2010, MN988668, hCoV-19/Norway/7448/2020, hCoV-19/England/205341113/2020, hCoV-19/Brazil/SP-918/2020, hCoV-19/Mexico/CMX-INER-0222/2020, hCoV-19/Zambia/ZMB-24503/2020, hCoV-19/Brazil/RS-00674-HM_LMM52649/2020, hCoV-19/South Africa/KRISP-K004282/2020, hCoV-19/South Africa/KRISP-K004540/2020, hCoV-19/Taiwan/TSGH-34/2020, hCoV-19/Australia/ACT0090/2020, hCoV-19/Norway/0179/2021, hCoV-19/USA/TX-HHD-2102044112/2021, hCoV-19/Austria/CeMM3224/2021, hCoV-19/Japan/NGY-NNH-075/2021, hCoV-19/Croatia/HRV000_49/2021, hCoV-19/Netherlands/UT-RIVM-13868/2021, hCoV-19/Belgium/rega-3528/2021, hCoV-19/Spain/RI-IBV-99016064/2021, hCoV-19/Ireland/D-NVRL-e84IRL94434/2021, hCoV-19/USA/NY-PRL-2021_02_25_00C23/2021, hCoV-19/USA/CA-LACPHL-AF00513/2021, hCoV-19/India/ILSGS00925/2021, hCoV-19/India/MH-ICMR-NIV-INSACOG-296092/2021, hCoV-19/India/MH-ICMR-NIV-INSACOG-GSEQ-1320/2021, hCoV-19/Brazil/SP-2441/2021, hCoV-19/South Korea/KDCA3504/2021, hCoV-19/Argentina/PAIS-G0233/2021, hCoV-19/Russia/IVA-CRIE-L188N0202/2021, hCoV-19/Canada/ON-S2383/2021, hCoV-19/Peru/LAL-INS-1603/2021, hCoV-19/Peru/LIM-INS-443/2021, hCoV-19/England/WSFT-25C69B5/2021, MZ622698.1, MZ571142.1, MZ477758.1, hCoV-19/USA/CA-CDC-LC0934366/2022, hCoV-19/Canada/ON-HSC-22-24664-v1/2022, hCoV-19/Canada/ON-HSC-22-26431-v1/2022, hCoV-19/Netherlands/OV-LabMicTA-1347/2022, hCoV-19/USA/un-CRIP-0933-000JHH00/2022, hCoV-19/Netherlands/OV-LabMicTA-4265/2023, hCoV-19/USA/un-CRIP-206283495/2023, hCoV-19/USA/un-CRIP-206394699/2023, hCoV-19/USA/un-CRIP-206480476/2023, hCoV-19/USA/un-CRIP-206486213/2023, hCoV-19/American Samoa/AS-H2531481/2025, hCoV-19/Anhui/AQCDC-114/2024, hCoV-19/Anhui/AQCDC-115/2024, hCoV-19/Anhui/BZCDC-411/2024, hCoV-19/Anhui/HFCDC-328/2024, hCoV-19/Anhui/HFCDC-332/2024, hCoV-19/Anhui/HFCDC-338/2024, hCoV-19/Anhui/HFCDC-339/2024, hCoV-19/Anhui/HFCDC-343/2024, hCoV-19/Anhui/HFCDC-346/2024, hCoV-19/Anhui/HFCDC-348/2024, hCoV-19/Anhui/HFCDC-350/2024, hCoV-19/Anhui/HFCDC-351/2024, hCoV-19/Anhui/HFCDC-357/2024, hCoV-19/Anhui/HFCDC-358/2024, hCoV-19/Anhui/HFCDC-360/2024, hCoV-19/Anhui/HFCDC-362/2024, hCoV-19/Anhui/BZCDC-409/2024, hCoV-19/Argentina/INEI127460/2025, hCoV-

19/Argentina/INEI127499/2025, hCoV-19/Barbados/BPH-3001801/2025, hCoV-19/Brazil/SP-IB_CEVC_2500336/2025, hCoV-19/Brazil/SP-IB_CEVC_2500337/2025, hCoV-19/Canada/AB-ABPHL-130567/2025, hCoV-19/Canada/AB-ABPHL-130568/2025, hCoV-19/Canada/AB-ABPHL-130579/2025, hCoV-19/USA/MI-MDHHS-SC45833/2025, hCoV-19/USA/MI-MDHHS-SC45834/2025, hCoV-19/USA/MI-MDHHS-SC45836/2025, hCoV-19/USA/NE-NPHL-RP250121/2025, hCoV-19/USA/NY-Wadsworth-25055746-01/2025, hCoV-19/Canada/ON-PHL-202505441/2025, hCoV-19/France/GES-RELAB-IPP05396/2025, hCoV-19/Japan/FR049-0235/2025, hCoV-19/USA/CA-SFDPH-EP250001128/2025, hCoV-19/USA/OR-OSPHL10461/2025, hCoV-19/Canada/QC-L00863786001/2025, hCoV-19/Heilongjiang/HLJCDC-0238/2025, hCoV-19/USA/NJ-PHEL-V25011078/2025, hCoV-19/USA/MI-UM-11870944/2025, hCoV-19/Yunnan/YNCDC-0708/2025, hCoV-19/USA/CA-CDPH-500147095/2025, hCoV-19/Germany/BW-RKI-I-1148598/2025, hCoV-19/Japan/BRL25-0187/2025, hCoV-19/USA/FL-IDPH-S25WGS0661/2025, hCoV-19/France/OCC-CHU-TLS-6005958233/2025, hCoV-19/France/OCC-CHU-TLS-6005858773/2025, hCoV-19/Japan/FR050-0204/2025, hCoV-19/Japan/FR050-0207/2025, hCoV-19/France/OCC-CHU-TLS-6005967779/2025, hCoV-19/USA/OR-OSPHL-10525/2025, hCoV-19/USA/CA-LACPHL-AY13792/2025, hCoV-19/USA/CA-LACPHL-AY13703/2025, hCoV-19/USA/CA-LACPHL-AY13697/2025, hCoV-19/Italy/APU-UNIBA_Ba1389/2025, hCoV-19/Canada/ON-PHL-202506816/2025, hCoV-19/Denmark/DCGC-692592/2025, hCoV-19/Spain/PV-HUD-98032698/2025, hCoV-19/Brazil/RJ-FIOCRUZ-3747/2025, hCoV-19/France/ARA-RELAB-IPP06587/2025|EPI_ISL_20238894|2025-09-30, hCoV-19/Japan/EIS13-858/2025|EPI_ISL_20237990|2025-09-29, hCoV-19/Scotland/CLIMB-CM7YJCHI/2025|EPI_ISL_20240195|2025-09-28, hCoV-19/Scotland/CLIMB-CM7YJHKP/2025|EPI_ISL_20240200|2025-09-29, hCoV-19/Scotland/CLIMB-CM7YJQ87/2025|EPI_ISL_20240184|2025-09-30, hCoV-19/Spain/CT-LabRefCat-079390345743/2025|EPI_ISL_20240422|2025-09-14, hCoV-19/USA/IL-IDPH-S25WGS0732/2025|EPI_ISL_20240444|2025-09-01, hCoV-19/USA/IL-IDPH-S25WGS0738/2025|EPI_ISL_20240451|2025-09-04, hCoV-19/USA/IL-IDPH-S25WGS0741/2025|EPI_ISL_20240454|2025-09-12, hCoV-19/USA/IL-IDPH-S25WGS0760/2025|EPI_ISL_20240458|2025-09-29, hCoV-19/Canada/AB-ABPHL-134534/2025|EPI_ISL_20246194|2025-10-25, hCoV-19/Canada/AB-ABPHL-134535/2025|EPI_ISL_20246195|2025-10-25, hCoV-19/Canada/BC-BCCDC-784136/2025|EPI_ISL_20244584|2025-10-05, hCoV-19/France/PAC-RELAB-HCL725000246801/2025|EPI_ISL_20243927|2025-10-15, hCoV-19/Italy/UMB-LAOP-176-4-21 /2025|EPI_ISL_20245329|2025-10-20, hCoV-19/Netherlands/ZH-EMC-9396/2025|EPI_ISL_20244139|2025-10-19, hCoV-19/SouthKorea/KDCA300477/2025|EPI_ISL_20246376|2025-10-20, hCoV-19/South Korea/KDCA300478/2025|EPI_ISL_20246377|2025-10-18, hCoV-19/Spain/MD-HGUGM-66853663/2025|EPI_ISL_20244311|2025-10-23, hCoV-19/USA/WI-WSLH-2501941/2025|EPI_ISL_20245936|2025-10-1, hCoV-19/Australia/SA30712266/2025|EPI_ISL_20249691|2025-11-03, hCoV-19/Australia/SA31011468/2025|EPI_ISL_20249697|2025-11-06, hCoV-19/Australia/SA31112961/2025|EPI_ISL_20249702|2025-11-07, hCoV-19/Australia/SA31200024/2025|EPI_ISL_20249703|2025-11-08, hCoV-19/Canada/NL-NML-784770/2025|EPI_ISL_20276507|2025-11-05, hCoV-19/Canada/NL-NML-786877/2025|EPI_ISL_20276535|2025-11-07, hCoV-19/Italy/LOM_Sacco_Var-R62139/2025|EPI_ISL_20276558|2025-10-11, hCoV-19/Spain/CT-HUB09736/2025|EPI_ISL_20277352|2025-11-5, hCoV-19/USA/CA-LACPHL-AY15591/2025|EPI_ISL_20249556|2025-11-02, hCoV-19/USA/NY-NYULH-24213/2025|EPI_ISL_20276582|2025-11-22, hCoV-19/Germany/BE-RKI-I-1150076/2025|EPI_ISL_20316427|2025-12-08, hCoV-19/Germany/BW-RKI-I-1150082/2025|EPI_ISL_20316433|2025-12-15, hCoV-19/Germany/HB-RKI-I-1150083/2025|EPI_ISL_20316434|2025-12-11, hCoV-19/Germany/NW-RKI-I-1149912/2025|EPI_ISL_20316200|2025-12-12, hCoV-19/Germany/NW-RKI-I-1150035/2025|EPI_ISL_20316321|2025-12-02, hCoV-19/Germany/NW-RKI-I-1150071/2025|EPI_ISL_20316356|2025-12-02, hCoV-19/Germany/NW-RKI-I-1150081/2025|EPI_ISL_20316432|2025-12-12, hCoV-19/Italy/APU-UNIBA-Ba-1452/2025|EPI_ISL_20316041|2025-12-11, hCoV-19/Italy/APU-UNIBA-Ba-1461/2025|EPI_ISL_20316151|2025-12-10, hCoV-19/Italy/APU-UNIBA-Ba-1463/2025|EPI_ISL_20316560|2025-12-15, hCoV-19/Japan/TKYF252072/2025|EPI_ISL_20316143|2025-12-09, hCoV-19/Japan/TKYF252145/2025|EPI_ISL_20316150|2025-12-13, hCoV-19/Belgium/Sciensano-LS-S1432/2026|EPI_ISL_20421395|2026-01-17, hCoV-19/Brazil/MS-FIOCRUZ-FMS1347/2026|EPI_ISL_20429671|2026-01-29, hCoV-19/Canada/BC-BCCDC-795604/2026|EPI_ISL_20421601|2026-01-22, hCoV-19/Canada/BC-BCCDC-795617/2026|EPI_ISL_20421611|2026-01-26, hCoV-19/Canada/BC-BCCDC-795629/2026|EPI_ISL_20421621|2026-01-29, hCoV-19/env/Italy/PIE-ARPA-BL0059TO/2026|EPI_ISL_20347757|2026-01-26, hCoV-19/Japan/FR055-0001/2026|EPI_ISL_20432149|2026-01-27, hCoV-19/Liaoning/LNCDC-20-6/2026|EPI_ISL_20433615|2026-01-29, hCoV-19/Liaoning/LNCDC-20-7/2026|EPI_ISL_20433614|2026-01-28, hCoV-19/USA/IL-RIPHL_33801_G/2026|EPI_ISL_20421910|2026-01-22, hCoV-19/USA/IL-RIPHL_33803_G/2026|EPI_ISL_20421912|2026-01-23, hCoV-19/Brazil/RS-FIOCRUZ-

1369/2026|EPI_ISL_20435085|2026-02-06, hCoV-19/Brazil/RS-FIOCRUZ-1371/2026|EPI_ISL_20435087|2026-02-22, hCoV-19/env/Italy/PIE-ARPA-BL00123TO/2026|EPI_ISL_20379808|2026-02-09, hCoV-19/Liaoning/LNCDC-23-1/2026|EPI_ISL_20433619|2026-02-06, hCoV-19/Liaoning/LNCDC-25-1/2026|EPI_ISL_20433617|2026-02-05, hCoV-19/Liaoning/LNCDC-25-2/2026|EPI_ISL_20433622|2026-02-05, hCoV-19/Liaoning/LNCDC-26-2/2026|EPI_ISL_20433623|2026-02-04, hCoV-19/Liaoning/LNCDC-26-3/2026|EPI_ISL_20433621|2026-02-04, hCoV-19/USA/NY-GBW-GKISBBBI18851/2026|EPI_ISL_20368581|2026-02-12, hCoV-19/USA/NY-NYULH-24584/2026|EPI_ISL_20378445|2026-02-05, hCoV-19/USA/NY-NYULH-24595/2026|EPI_ISL_20378456|2026-02-07, hCoV-19/USA/NY-NYULH-24609/2026|EPI_ISL_20389679|2026-02-09, hCoV-19/USA/NY-NYULH-24619/2026|EPI_ISL_20389689|2026-02-13, hCoV-19/USA/MN-MDH-50347/2026|EPI_ISL_20433430|2026-03-20, hCoV-19/USA/MN-MDH-50330/2026|EPI_ISL_20433478|2026-03-18, hCoV-19/USA/MN-MDH-50332/2026|EPI_ISL_20433464|2026-03-24, hCoV-19/USA/MN-MDH-50342/2026|EPI_ISL_20433448|2026-03-23, hCoV-19/USA/MN-MDH-50346/2026|EPI_ISL_20433432|2026-03-25, hCoV-19/USA/MN-MDH-50353/2026|EPI_ISL_20433375|2026-03-16, hCoV-19/USA/MN-MDH-50360/2026|EPI_ISL_20433474|2026-03-26, hCoV-19/USA/MN-MDH-50364/2026|EPI_ISL_20433426|2026-03-26, hCoV-19/USA/MN-MDH-50381/2026|EPI_ISL_20433428|2026-03-27, hCoV-19/USA/MN-MDH-50387/2026|EPI_ISL_20433438|2026-03-24, hCoV-19/USA/MN-MDH-50398/2026|EPI_ISL_20433460|2026-03-30, hCoV-19/Netherlands/UT-RIVM-148280/2026|EPI_ISL_20432182|2026-04-07, hCoV-19/South Korea/KDCA301115/2026|EPI_ISL_20440580|2026-04-06, hCoV-19/South Korea/KDCA301118/2026|EPI_ISL_20440582|2026-04-07, hCoV-19/Canada/NB-CHUDGLD-01D-MM00079R/2026|EPI_ISL_20421776|2026-04-01, hCoV-19/Germany/NW-RKI-I-1150691/2026|EPI_ISL_20431475|2026-04-05, hCoV-19/Hong Kong/VB26112288/2026|EPI_ISL_20422257|2026-04-02, hCoV-19/Netherlands/LI-RIVM-148285/2026|EPI_ISL_20432157|2026-04-09, hCoV-19/Netherlands/NB-RIVM-148281/2026|EPI_ISL_20432166|2026-04-01, hCoV-19/Netherlands/NB-RIVM-148288/2026|EPI_ISL_20432168|2026-04-09, hCoV-19/Netherlands/UT-RIVM-148279/2026|EPI_ISL_20432181|2026-04-07, hCoV-19/Netherlands/UT-RIVM-148286/2026|EPI_ISL_20432183|2026-04-05, hCoV-19/Netherlands/UT-RIVM-148287/2026|EPI_ISL_20432184|2026-04-08, hCoV-19/USA/IL-RIPHL_16275_G/2026|EPI_ISL_20432037|2026-04-08, hCoV-19/USA/IL-RIPHL_16278_G/2026|EPI_ISL_20432038|2026-04-05, hCoV-19/USA/MN-MDH-50377/2026|EPI_ISL_20433458|2026-04-01, hCoV-19/USA/MN-MDH-50397/2026|EPI_ISL_20433362|2026-04-02

Appendix 3. List of 20 Middle East Respiratory Syndrome Coronavirus complete genomic sequences.

MF598617.1, MF598595.1, NC_019843.3, KY673148.1, KT225476.2, MG923479.1, MK967708.1, MT361640.1, KT326819.1, MK129253.1, OL622035.1, OP712625.1, MH734114.1, MG923468.1, KJ361503.1, KM210277.1, KF958702.1, OP654179.1, MW086535.1, MZ268405.1

Appendix 4. List of 26 dengue virus complete genomic sequences.

KY672944.1, KY672937.1, MW386865.1, MG560269.1, MG560267.1, MN566112.1, MK268692.1, MH069499.1, KY672955.1, KY672954.1, LC379197.1, MN018389.1, MK005258.1, NC_001475.2, KY863456.1, MH544649.1, MH823209.1, KY921907, KF041255.1, LC379196, LC379195, KJ579245.1, MG272274.1, KY672960.1, KX224312.2, MG272272.1

Appendix 5. List of 16 Ebola virus complete genomic sequences.

MG572235.1, KU174137.1, KY786025.1, KY786026.1, KY785936.1, MH121162.1, MK952150.1, MH121169.1, NC_039345.1, MW056492.1, MF319186.1, MK340750.1, MW056493.1, MK028856.1, MK028834.1, MK028835.1