Report on Distance Mutations in CoVid2 - 2019

Introduction:

The comparisons made for this research are based on the following premises:
1 - From a mathematical-informatic perspective, a bit in electronics has two states (on or off), while the biological "bit" of DNA (to make a broad comparison) has four states (the four nitrogenous bases Adenine Cytosine Guanine Thymine).
2 - From this starting point, a further study of sequence comparison will be carried out, applying a mathematical/statistical model involving frequencies and modes of mutation within them.
3 - The basic software used is MEGA-X for the graphical composition of phylogenetic trees. Some algorithms of the above program have been modified, eliminating calculations related to protein synthesis (i.e., the successive combination of nitrogenous bases not relevant to the primary research); "ad hoc" programmed scripts have also been used to facilitate the formatting of text files with bases in a format readable by the software in question.
4 - Viral sequences were obtained from the worldwide genetic database:
https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/
SARS-CoV-2 (Severe acute respiratory syndrome coronavirus sequences).

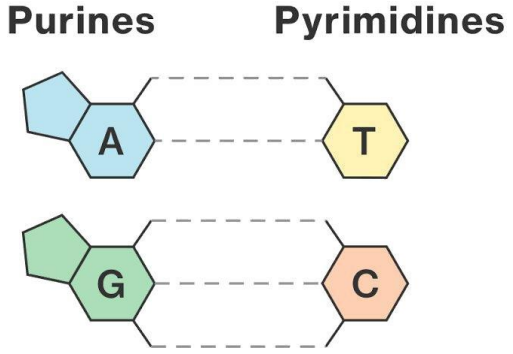
Research objectives:
1 - Comparison of genetic sequences within the same CoVid2 family.
2 - Comparison of identified strains, aimed at identifying types and differences in them, in order to construct a phylogenetic tree.
3 - Localization of any mutations in this tree, compared with the location and dating of identified strains.

Theories and methods of calculating nucleotide distance:

To understand how to calculate the differences between two or more viral sequences, it is necessary to first climb to the step immediately above that of the nitrogenous bases, i.e., the nucleotides:
Nucleotides (derived from Nucleosides through the addition of one or more phosphate groups) are essentially the units that make up nucleic acids (DNA or RNA).
Each of them consists of:
a - a nitrogenous base
b - a molecule of sugar (specifically deoxyribose)
c - a molecule of phosphoric acid (phosphate group)
Nucleotides are as many as the nitrogenous bases from which they derive, in detail:
- Adenosine, derived from Adenine
- Cytidine, derived from Cytosine
- Guanosine, derived from Guanine
- Thymidine, derived from Thymine
- Uridine, derived from Uracil

Please note that in genomic sequences, Uracil is usually replaced by Thymine as software compares sequences of the same family, and generally does not compile a phylogenetic tree with mixed DNA-RNA sequences (*)

The fundamental thing to understand is that nucleotides can bind because it is possible to form a bond between their nitrogenous bases, but not indiscriminately: indeed, adenine binds with thymine, and guanine binds with cytosine.
Nitrogenous bases are then divided, depending on their molecular configuration, into purinic (adenine and guanine) or pyrimidinic (thymine and cytosine).
In the following image, a brief summary of the AT-GC scheme can be found:



Another fundamental concept to understand is the presence/distance of so-called Sites, that is, the regions of DNA where a cytosine is next to a guanine in the linear sequence of bases. They can be classified as CpG or CG, ("CpG" stands for "--C--phosphate--G--") where the deoxycytosine and deoxyguanosine nucleosides are separated only by a phosphate group, which normally bridges the nucleotides in DNA. The notation "CpG" is used to distinguish this linear sequence from the complementary base pairing of CG (cytosine and guanine) on two different strands.
It is precisely on the above-mentioned binding rules of nucleotides that the calculation is based to identify possible mutations in the genetic code, namely:
- Substitutions of nucleotides compatible with others in the same genetic sequence
- Substitutions of purine/pyrimidine bases with other compatible nitrogenous bases
- Substitutions of Sites in CpG Islands repeated in the same sequence

To apply this to a mathematical model, there are obviously multiple methods, depending on how the substitutions of the sequences are set up. However, at a "low level" (i.e. not involving protein synthesis), the "Tajima D" method is often used, which takes its name from the researcher Tajima Fumio: this method is based on the comparisons of the average number of differences of pairs of bases with their respective Sites. The related algorithm is used by multiple genetic comparison and evolution programs.

Here are the algorithms used:
1 - Calculation of distance for adjacent nucleotide substitution:
With p, the mutation distance is indicated, that is, the distance (p) is the proportion of nucleotide Sites in which the two sequences compared differ.
This is obtained by dividing the number of nucleotide differences (nd) by the total number of nucleotides compared (n), i.e.:
$p = nd/n$
where the variation V relative to p is given by:
$V(p) = [p(1 - p)]/n$
In this way, the substitution is consecutive and linear.

This type of function is particularly effective compared to the number of nucleotide substitutions at a Site (d) only when the distance is small, i.e. when the distance p is approximately equal to the number of nucleotide substitutions per Site (d), and is optimal when the latter is generally $p < 0.1$. However, considering the simplicity of comparing this algorithm, appreciable and reliable results can still be obtained even to construct more complex phylogenetic trees where the distance is much longer, provided that all distances of the pairs are still small.
Below is the graph of the Tajima method:

A	-	β	γ	δ
T	α	-	γ	δ
C	α	β	-	δ
G	α	β	γ	-

Where α, β, γ, δ are the substitution ratios.

2 - Calculation of distance for Transition and Transversion:
To what was said above, the so-called calculation of transition and transversion can also be applied: Transition is the substitution of one Purine with another Purine and transversion is the substitution of one Pyrimidine with another Pyrimidine.

Therefore, the proportion of transitional (P) and transversional (Q) differences can be calculated with the following equations:
P = ns/n
Q = nv/n
where ns and nv are respectively the numbers of transitional and transversional differences between the two sequences, with ns + nv = nd. The variants of P and Q are calculated with the same function used for the variation V of the distance previously mentioned. In addition, the ratio of transition to transversion differences (Rd) is calculated from:
Rd = P/Q

And its variation is calculated from
V(Rd) = [c12P + c22Q - (c1P + c2Q)2]/n
Where
c1 = 1/Q
c2 = - P/Q2

This method, integrated with the previous one, manages to provide a more accurate perspective on the phylogenetic tree.

3 - Calculation of Tajima-Nei distance

Since in reality there can still often be a further distancing of nucleotide mutation frequency or in any case when data that imply mutations that are more distant from the simply neighboring base are to be calculated, the following equations, which Dr. Fumio Tajima perfected with Dr. Masatoshi Nei, come to the rescue. In this method, the distance is no longer calculated linearly but logarithmically and is expressed as:

$d = -b \log_e(1 - p/b)$
and
$V(d) = p(1-p)/[(1 - p/b)^2 n]$
where

$$b = \frac{1}{2}(1 - \sum_{i=1}^{4} g_i^2 + p^2 / c)$$

$$c = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{x_{ij}}{2 g_i g_j}$$

In the laest, gi and gj represent the frequency with which the nth nucleotides "i" and "j" are present.
This method proves to be particularly effective when facing objectively mutated sequences that suggest a deep change in the genetic lineage. (***)

Preliminary results:

As the main strain used as a starting point and reference for the subsequent compared sequences, the original Wuhan strain was chosen. This, in addition to being the most widespread based on the analyzed sequences at a global level (as every continent has at least one location where this strain is present), also involves particularly the entire South Asian area.
Below is the graph of the South Asian strain that involves other countries by linear similarity:

Firts results based on linear comparison (circular graph):

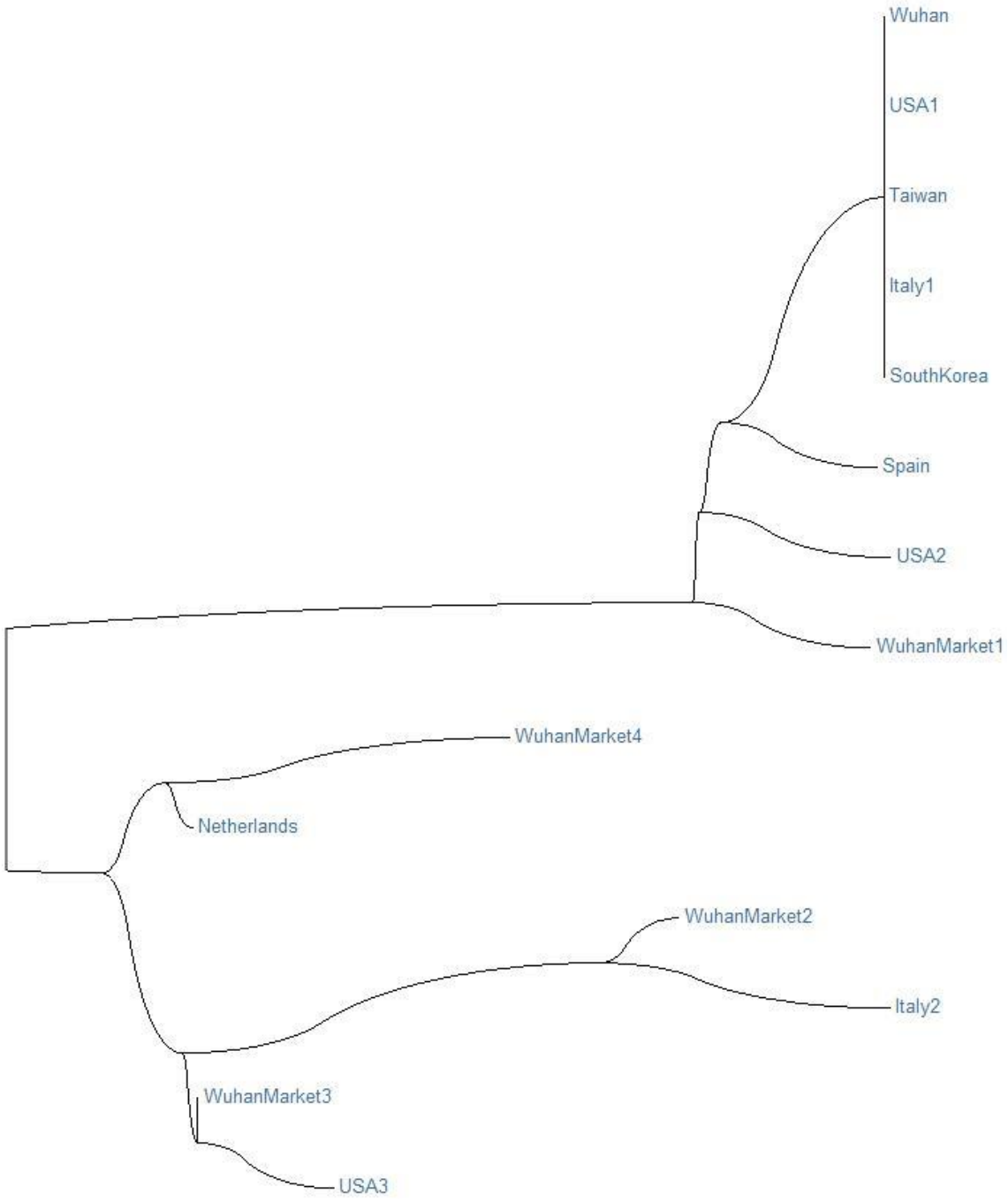
From this first graph, it is yet evident that the Netherlands strain is the furthest, followed by one of the strains from the so-called "Wuhan Seafood Market," identified as "Wuhan Mrkt." The Spanish strain is much closer to the original sequence, while the Wuhan, Taiwan, and South Korea sequences are practically identical (along with one of the many that arrived in the USA, which will be analyzed later).

Anomaly of the "Wuhan Seafood Market" strains

Before delving into the anomaly, it should be mentioned that according to the Chinese government, the sample from the Wuhan seafood market should be the "protovirus," the one from which the pandemic originated. It should also be noted that all the examined eastern sequences are more or less identical (including those from Taiwan and South Korea, isolated by independent researchers), and that, at this point, all the data analyzed have been done only at a comparative-linear level.

The anomaly is as follows: The 2019 sample from the Chinese market where the Beijing government claims this virus spread has a genetic sequence that does not match the South Asian strain.

It should be taken into consideration that there are only four sequences from the "Seafood Market," all included in the genetic database (sequences LR767995-96-97-98), and all published on March 6, 2020:

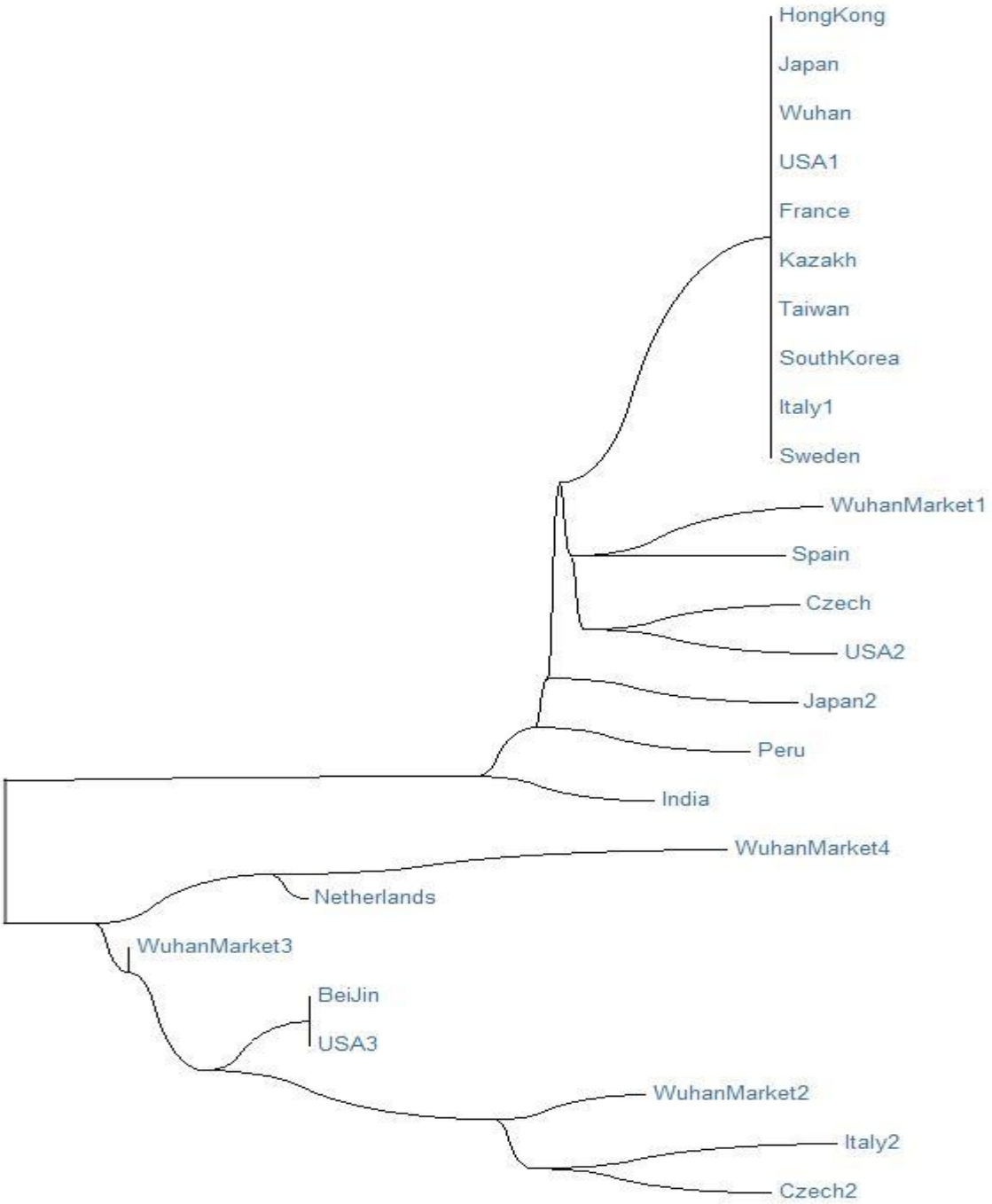


From this graph, it emerges that the four "Seafood Market" strains (Wuhan Market 1-4) are not directly connected to the main South Asian strain.
Since the publication date is the same and covers a one-year period, it's difficult to have a proof from the current data if these strains were mutated or pre-existing.

US Strains:

According to current comparisons (made with a calculation algorithm for nucleotide base similarity), the US appears to be the country most affected by the pandemic, not only at an epidemic level but also in terms of diversity of strains. In fact, every globally present strain analyzed has also been isolated in the US with similar or at least phylogenetically similar characteristics within the same family.

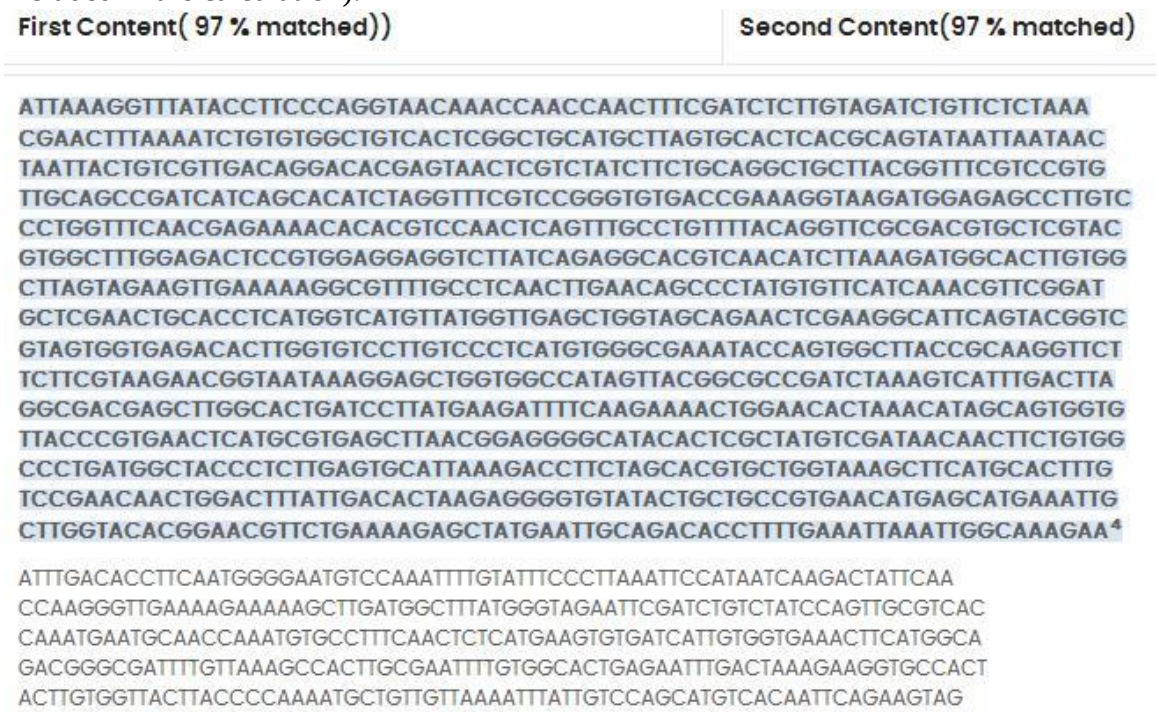
Below is the graph that shows how the US has at least one isolated case in every strain present globally, including the primary South Asian sequence and another strain with similar familial characteristics:



Brazilian strain and stability in the genetic sequence:

As of May 2020, based on the analysis of 32 different samples, the previous data has been confirmed, and it has been found that the most widespread strain in Brazil is stably identical to the primary South Asian common strain. This could support the theory that the virus remains stable within each strain, not presenting significant mutations and having a strong stop-coding sequence clearly visible and marked at the end of the genome, remaining at a fluctuating number of 29890-29903 unchanged nucleotide bases, from which 29860 sites are derived.

It should be noted that the similarity percentage of this strain is based only on the comparative model of the nucleotide bases. Below is the result of the direct comparison (in the image, only the initial part of the code is shown, while the entire sequence is included in the calculation).



The results obtained by applying the Tajima-Nei comparison methods, which involve substitutions of nucleotides and similarity of adjacent sites, to the viral sequences from Wuhan, Italy, and Brazil are shown below:

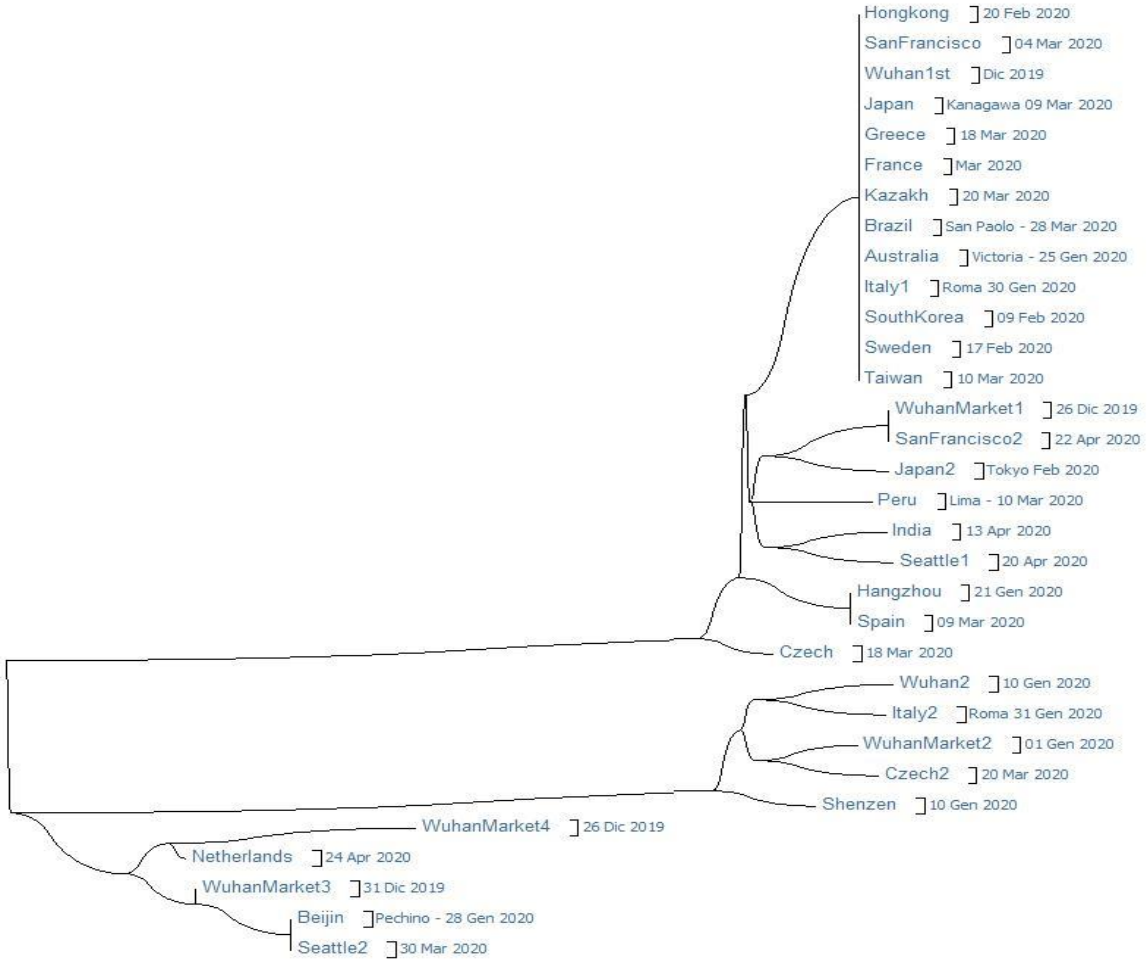
Table. Results from the Tajima's test for 3 Sequences

Configuration	Count
Identical sites in all three sequences	29860
Divergent sites in all three sequences	0
Unique differences in Sequence A	1
Unique differences in Sequence B	1
Unique differences in Sequence C	4

This result shows that on 29866 sites the divergences are equal to 0, the identical sites are 29860 with unique differences between the sequence of Wuhan and that of Italy equal to 1, with differences between the aforementioned sequences and that of Brazil equal to 4.
There is no significant mutation in this strain within its genome.

32-sample phylogenetic tree:

The phylogenetic tree based on the Tajima-Nei model for the 32 viral samples shows a progressive attenuation of the poly(A) polymerase sequence present at the end of the RNA genome (codes for ACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA in the Wuhan-1st sequence in December 2019), which is abbreviated (codes for ACAAAAAAAAAAA in the French sequence of March 2020) or absent in some chronologically later strains (codes for ATTTTAGTAGTGCTATCCCCATGT in the Indian sequence of April 2020).
Below is the graph that reconstructs the phylogenetic tree for the 32 viral samples, showing the dates when each sequence was isolated.
The tree shows the close genetic relationship between the different strains, with some minor variations observed due to random mutations. The presence or absence of the poly(A) polymerase at the end of sequence seems to have no impact on the overall genetic similarity of the strains, or better, from the compared samples it seems undergoing into a progressive attenuation without relevant mutations:



Conclusions:

In conclusion, based on the analyzed samples, a viral model consisting of multiple strains with minimal variations is presented. So far, no significant mutations have been observed in the viral code.

Notes:

- (*) The world database for genetics uses Thymine instead of Uracil for RNA sequences, in order to standardize with programs compatible with the ACGT format, which would not recognize ACGU.
- (**) The Tajima-Nei comparison models used in this study are essentially interchangeable and applicable to both DNA and RNA, as they analyze the sequence, bases, and sites in a progressive mapping at a lower and basic level, regardless of single or double strands.
- (***) This refers to logarithmic distance. It is necessary to specify that in the generated phylogenetic tree, the measurable physical distance between one strain and another must always be referred to the type of algorithm used and obligatorily described (linear, logarithmic, temporal, etc.), in order to have a reference measure.

Acknowledgments:

1 - My Ryzen Octacore, which was essential in graph calculation and rendering timing.
2 - The MEGA X software, a great tool accessible to everyone.
3 - Dr. Fumio Tajima for the excellent evolutionary models developed.
4 - The databases, researchers, and engineers who design machines to extract biological data into mathematical models.
5 - The National Institute for Infectious Diseases Lazzaro Spallanzani for the Italian sequence of the isolated viral code, which was promptly published.

Bibliography:

- Lehninger A.L., Nelson D.L., Cox M.M - Principles of biochemistry
- Multiplication and permutation, Axioms independence – University of Illinois (Urbana-Champaign)
- Evolutionary distance between Nucleotide sequences – Tajima F., Nei M. - University of Texas (Houston)
- MEGA X: Molecular evolutionary genetics analysis across Computing platforms
- Kumar S., Tamura K., Jakobsen I.B., Nei M. - MEGA2: Molecular evolutionary genetics analysis software - Bioinformatics

Virologically yours... Mike Yoshi