SAS Journal of Medicine
Abbreviated Key Title: SAS J Med
ISSN 2454-5112
Journal homepage: https://saspublishers.com

OPEN ACCESS

Sars-Cov-2 Mutations: Platforms Used for Their Analysis and Evolution

Angel San Miguel Hernández*, Julia San Miguel Rodríguez

Clinical Analysis Service, Rio Hortega Valladolid University Hospital (*) International University of La Rioja (UNIR)

DOI: 10.36347/sasjm.2021.v07i06.010 | **Received:** 23.04.2021 | **Accepted:** 05.06.2021 | **Published:** 11.06.2021

***Corresponding author:** Angel San Miguel Hernández

Abstract

Review Article

The biology of the SARS-CoV-2 virus leads to constant changes in its genome through mutations, so the appearance of variants is an expected fact. Since evolutionary adaptation and diversification has been observed globally throughout the pandemic. Most of the mutations that arise do not provide a selective advantage to the virus, or phenotypic changes that imply alterations in the behavior of the infection. There is concern with the appearance of one or more variants that escape the effect of the neutralizing antibodies generated after a previous infection or vaccination, which could lead to cases of reinfection or loss of vaccine efficacy. These variants could be associated with a decrease in the sensitivity of certain diagnostic techniques, although most of these techniques use several different genes to avoid this effect. The genomic sequencing of SARS-CoV-2 and the platforms used to monitor the different varieties and observe how they are circulating around the world are reviewed.

Keywords: SARS-Cov-2, mutations, platforms, pandemic.

INTRODUCTION

When viruses penetrate the cells of an infected organism, it makes use of the machinery of those cells to replicate its genetic material, DNA or RNA, and produce thousands of copies, following the instructions contained in it. In all this process errors or mutations can occur, so that some of the new replicas of the hereditary molecule are slightly different from the original. A new genetic variant of the virus would thus emerge. On very few occasions this mutation confers any advantage on it; such as, for example, favoring their ability to spread. In that case, the new variant would expand more rapidly, progressively replacing the existing ones until it became the majority. And the normal thing is that during an epidemic, different variants of the same virus coexist in different proportions. Thus, the natural selection process is one of the mechanisms that drives the evolution of beings capable of reproducing themselves and transmitting their inheritance to the next generation; another is genetic drift [1]. When it acts on a population of viruses, it is when variants with different transmission capacity arise. But that's not their only way of acting from viruses; in many cases it can occur, for example, under the influence of an environmental factor that, in comparative terms, favors the survival and proliferation of certain genetic variants over others. In this way, the dominant ones leave more offspring, so their genetic traits will end up becoming the majority in the population. We call this factor selective pressure. Antivirals and vaccines can act, as far as viruses are concerned, as selective pressures and act in this way when they prevent or hinder the reproduction of certain variants but not others. In that case, they would suppress or make the former a minority, leaving free way for the latter to proliferate. This is what happens when a variant of a virus is resistant to the action of an antiviral or a vaccine. And it is not difficult for such resistance to emerge. In addition, antivirals are often administered when an infection has already occurred and there are already millions of viral particles in the host organism. In such circumstances there are millions of viruses that can potentially mutate and become resistant to the drug. And on the other hand, the effect of an antiviral, such as that arising with an antibiotic in bacteria, is usually based on the action on a single cellular process, and the probability that a genetic variant resistant to such action will emerge is not low.

As vaccines are administered before an infection occurs, so that the defenses they generate can act before the pathogen proliferates in the body, thus preventing the emergence of millions of potential resistant variants when multiplying. And because the vaccine induces the production of antibodies that act against different targets, called epitopes, in pathogens. The probability that, due to mutation, genetic variants will arise that modify all the epitopes and, in this way, avoid the action of the antibodies is very low, although it is not null. Therefore, it is very important to avoid the transmission of SARS-Cov-2, and thus it would prevent individuals from becoming infected; furthermore, by limiting their proliferation, the probability that variants will emerge that may be more easily transmissible or that generate resistance to vaccines. Therefore, the genome of every organism is the set of all its genetic information and that defines its main biological characteristics. The first genome of the SARS-CoV-2 coronavirus was obtained in January 2020 and was a step to better understand how the virus behaves and acts. Since then, more than 40,000 SARS-CoV-2 genomes have been sequenced around the world, information that is making it possible to monitor and track how the virus spreads thanks to genetic and molecular epidemiology studies. STRUCTURE AND GENES OF SARS-CoV-2 The causative agent of Covid-19 is a beta coronavirus and belongs to a family of viruses that can cause respiratory symptoms ranging from the common cold to severe pneumonia [1]. SARS-CoV-2 is a virus belonging to the genus β-coronavirus, whose genetic material is a chain of ribonucleic acid (RNA). The genetic material of the virus of approximately 30 kb, is protected from the environment by four main structural proteins (Figure 1), which are:

- **Protein E (envelope)** is a transmembrane component of the virus, which is involved in the exchange of ions between the interior of the viral particle and the environment. It is also involved in the process of entering the virus into the host cell through interactions with cell membrane proteins.
- **The M protein (membrane),** which is a glycoprotein that acts as the membrane of the viral particle and supports the E protein. The M protein is the most abundant in viral particles and consists of three transmembrane domains. This protein is crucial in the assembly of viral particles, as it provides the scaffolding that gives the virus its shape and structure.
- **The S protein (spike),** which is a transmembrane glycoprotein that protrudes from the viral particle, having an ectodomain greater than 10 nm. This protein is made up of up to 300 monomers, whose S1 and S2 subunits are those that lead to the anchoring of the virus in the receptors of host cells.
- **The protein N (nucleocapside),** which is linked to the genetic material of the virus, forming the «nucloecápsid», which protects the RNA from the environment, and is essential for the release of RNA in the cytoplasm of the infected cell.

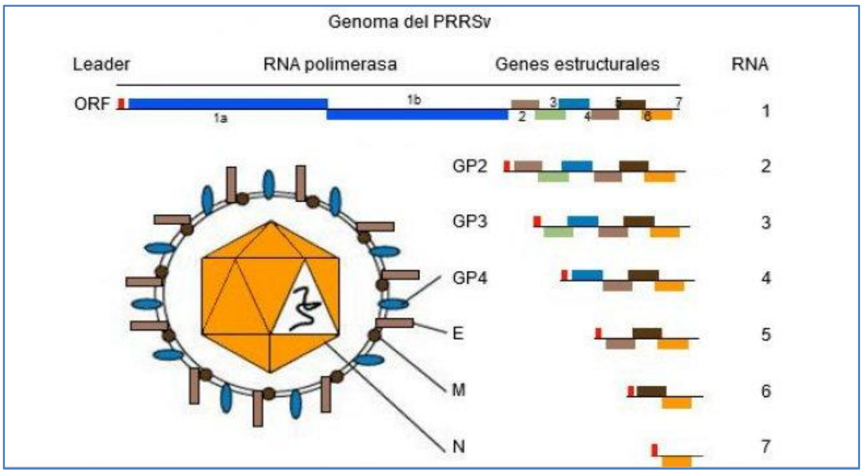

Fig-1: Schematic diagram of the structure and structural genes of SARS-CoV-2.

The life cycle of SARS-Cov-2 is shown in Figure 2. When the peak protein of SARS-CoV-2 binds to the host cell receptor, the virus enters and the envelope is detached, allowing the Genomic RNA is present in the cytoplasm. The ORF1a and ORF1b RNAs are made up of genomic RNA, and are then translated into pp1a and pp1ab proteins, respectively. These proteins are cleaved by protease to form a total of 16 non-structural proteins. Some of them form a replication / transcription complex (RNA-dependent RNA polymerase, RdRp), which uses the (+) strand genomic RNA as a template. The genomic strand RNA (+) produced through the replication process becomes the genome of the new viral particle. The subgenomic RNAs produced through transcription are translated into structural proteins (S: peak protein, E: envelope protein, M: membrane protein, and N: nucleocapsid protein) that form a viral particle. Spike, envelope, and membrane proteins enter the endoplasmic reticulum, and the nucleocapsid protein combines with genomic positive strand RNA to become a nucleoprotein complex. They fuse into the entire viral particle in the compartment of the Golgi apparatus and endoplasmic reticulum, and are excreted in the extracellular region through the Golgi apparatus and the vesicle.

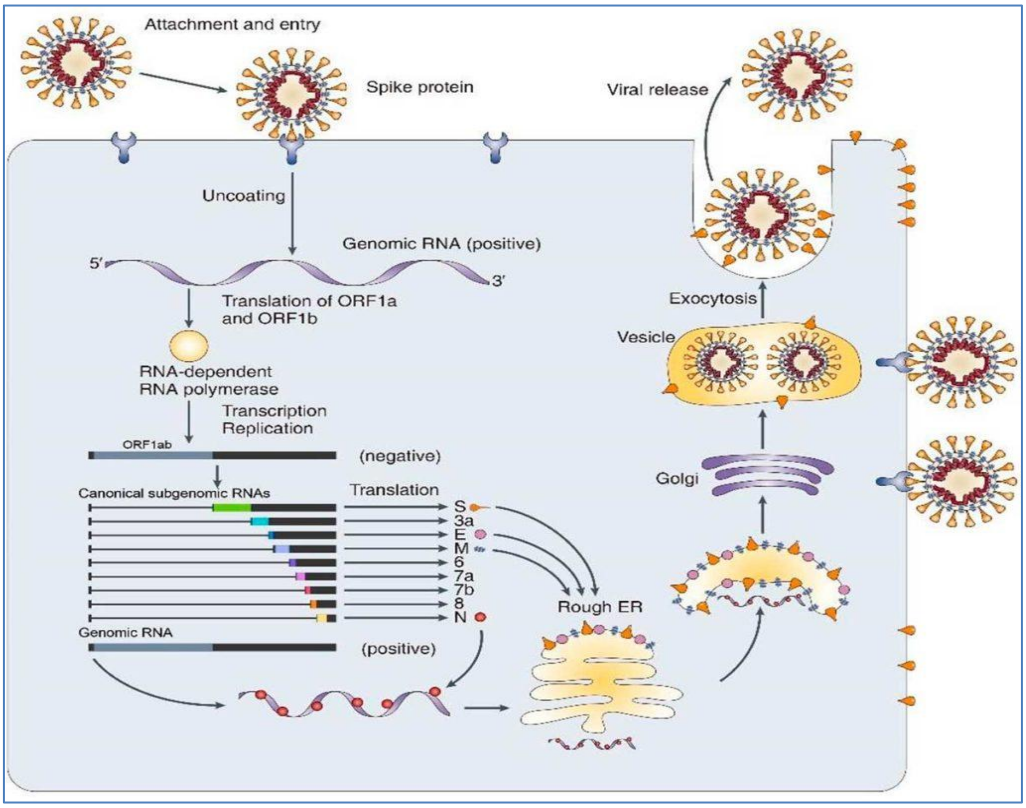

Fig-2: Life cycle of SARS-CoV-2.

Genomic sequencing of sars-cov-2 to see its evolution

All viruses are generating copies of their genome while infecting other organisms and as in this process small changes or genetic mutations are produced in the genome, their analysis allows us to trace how the virus is transmitted between people. By investigating these mutations in SARS-CoV-2, researchers have been able to establish what are known as phylogenetic clusters of the coronavirus, which are different branches or types of the virus that explain its origin, evolution and spread. Since there is enough information about how the virus has spread on the planet, and about what mutations and characteristics they have in different geographical locations.

Thus, several families of the new coronavirus have been differentiated, called "Phylogenetic Clades", characterized by different mutations. All the major clades of the virus, which help explain its origin and distribution; and all the clades that are in almost all the countries have been found, with different variations in the frequency of each one of them. One of these variants of SARS-CoV-2 has become the majority genetic form in many countries of the world, and in our environment in particular.

The analysis of the mutations suffered by the virus is also making it possible to investigate whether its transmission capacity over time and infection is attenuated or becomes stronger. Therefore, the genetic characteristics and evolution of the virus have continued to be studied. The analysis of how it is transmitted, a discipline known as genomic epidemiology is carried out, which is considered essential to know the diversity of the virus in a specific territory, evaluate its spread and facilitate decision-making and containment measures to prevent its expansion.

The SARS-CoV-2 coronavirus, discovered in january 2020 after being isolated from samples of patients affected by a new disease now known as Covid-19, evolves and undergoes genetic changes, like all viruses. Knowing these changes, which explain their behavior, is essential to improve the management of the virus and the management of the disease. Since the discovery of SARS-CoV-2 until now, the sequencing of its genome and the knowledge of the different variants that circulate around the world, is allowing to know more about its origin, influence and distribution [2].

The sequencing of the 29,899 nucleotides of its genome was a reference on which the researchers could work and track the SARS-Cov-2. In this way, thousands and thousands of complete genomes have been sequenced around the world [3].

Having the genomic information of these viruses, together with basic epidemiological information, such as where and when the sample was obtained, makes it possible to monitor how the virus spreads throughout the world thanks to molecular and genomic epidemiology. This is possible due to a characteristic of viruses related to their need to use the cellular machinery of the organism they infect in order to multiply and survive.

However, those copies that are generated are not always equal copies and mistakes or mutations are made and some of the 29.899 nucleotides that make up its genome are modified. These small errors or mutations that are produced in the genome are what make it possible to trace how the virus is transmitted between people.

In the first cases in China, the genomes of the viruses obtained from other market workers who had also fallen ill were sequenced. These genomes turned out to be exact copies of the genome found in the first patient, except for a small number of nucleotides, between 1 and 5. This similarity between the sequences allows defining that these viruses come from the same ancestor and form what is known as a phylogenetic cluster.

Once a sufficient number of genomes have been sequenced, it is possible to estimate how much the virus is capable of mutating, that is, how many mutations the virus accumulates in a given time, which is known as the rate of evolution. And it is estimated that the more than 100 first sequences available at the beginning of february 2020 were already enough to reliably calculate this rate of evolution, since it was not altered when adding new sequences to the analysis [3].

It was estimated that the SARS-CoV-2 coronavirus accumulated mutations at a rate of between 1.19 and 1.31 × 10−**3** substitutions/site/year, similar to that of the other epidemic coronaviruses, since in the case of SARS-CoV a rate between 0.80 × 10−**3** and 2.38 × 10−**3** and for MERS-CoV between 0.88 × 10−**3** and 1.37 × 10−**3** (5-7). Knowing this rate of evolution, which would approximately mean the accumulation of one mutation each 10 days, it allows estimating the geographical and temporal location for the ancestor of a specific phylogenetic cluster. Thus, by analyzing the first sequenced genomes, it has been possible to determine that the origin of the epidemic took place at the end of November 2019 [7-10].

As different mutations have appeared in the virus genome over time, the spread of SARS-CoV-2 around the world can be monitored and traced. One of the first cases of SARS-CoV-2 infection detected in Europe took place at Munich airport in a woman who had been in contact with her parents living in Wuhan and who was a worker for a German company. The genome of the virus detected in one of these employees was a nearly identical copy of other viruses sequenced in Shanghai and very similar to the first sequenced genome on the Wuhan market [11].

Specifically, the genome of the virus sequenced in Germany (BavPat1) presented mutations C3037T, in the ORF1ab and A23403G gene, which caused the change of amino acids D614G in the spike gene of the virus, with respect to the reference genome (Wuhan- Hu-1). This amino acid change had already been detected in other viruses found in Shanghai and would end up being one of the mutations found in most of the genomes of the sequenced viruses in Europe during the following weeks [11].

The epidemiological investigation of this chain of transmission made it possible to detect that one of the 12 cases related to this first positive in Europe would end up becoming the first case detected in Spain, in an individual who flew from Munich to the Canary Islands on january 28 2020 [11].

Another of the cases of infection detected in Europe had its origin in a conference of a gas company in Singapore on January 20-22, 2020, attended by more than 100 people from around the world. As of February 9, at least 7 attendees had tested positive for SARS-CoV-2 infection in Singapore, Malaysia, South Korea and the United Kingdom. One of them is related to a European outbreak of at least 13 other people with a positive diagnosis for Covid-19 and that had its origin in a French ski resort, affecting people from the United Kingdom, France and Spain [12]. Specifically, this case is related to the second positive detected in our country in Mallorca [12].

The United Kingdom is the country that has made the greatest effort in whole genome sequencing worldwide, with more than 20,000 sequences [13]. This information, together with the genomes of viruses from infections around the world, has allowed researchers to make a first approximation of the importation and establishment of virus lineages throughout the country. They have been able to detect at least 1,356 independent introductions of the virus into the UK from other regions of the world, although it is claimed that there is a possible underestimation of the number. Around 25% of these introductions appear to have become extinct as no new sequences belonging to these lineages have been found during a period of 4 weeks [13]. 80% of imports from other countries occurred mainly during the month of March 2020, reaching their peak in the middle of the same month.

By calculating the time of the most recent common ancestors of all these lineages, they determined as a median March 25, 2020 [13]. In addition, they have determined that one in three imports into the United Kingdom come from travelers from Spain since February 16, but mainly during the first half of March, date from which imports from Spain began to decline until reaching disappear the first week of April 2020 [13]. The 50 oldest transmission chains detected in this study and 49 of the 50 chains with the highest number of sequences are related to G614 variants, that is, this variant arrived earlier and with greater intensity in the United Kingdom, which could suggest a founder effect. in a country where this variant is clearly the majority.

In Scotland, it was possible to sequence the genome of 20% (n = 452) of all individuals with a diagnosis of COVID-19 available on 1 April in the whole country (n = 2310) [14]. This information has made it possible to obtain an estimate of the epidemic and the imports and spread of the virus in Scotland. Thus, 113 independent introductions have been detected. 51% of the clusters identified were associated with undetected introductions related to viruses that were already circulating in other European countries such as Spain, Italy and Austria. The authors state that the undocumented introductions occurred before the first cases detected [14].

Different nomenclatures existing in the genomic epidemiology of sars-cov-2

The analysis of the complete genome of pathogens has been revealed as an important tool in the study of the molecular epidemiology of infectious diseases. The existence of publication and sequence analysis platforms, such as those already mentioned and others such as NextStrain, have allowed the visualization in real time of the sequences available from the different countries, facilitating the study of the distribution of the virus and the identification of mutations. that could lead to possible adaptations to the human host or changes in the characteristics of the virus (van Dorp et al., 2020). This has facilitated the definition of a nomenclature for the different clades or branches with the variants that have appeared in its development and expansion.

The main proposals for phylogenetically classifying SARS-CoV-2 sequences are those of the NextStrain and GISAID platforms.

The NextStrain platform proposes five large clades, and two of them emerged already in 2019: clade 19A, which is considered the root group, and 19B, defined by changes C8782T and T28144C. The three remaining clades group circulating virus sequences in 2020: 20A, which is distinguished from 19A by the substitutions C3037T, C14408T and A23403G, 20B, characterized by three consecutive changes G28881A, G28882A and G28883C and 20C, which presents the substitutions C1059T and G25563T. In the first two clades, 19A and 19B, the sequences that circulated during the first months in Asia are grouped, while clade 20A comprises the European sequences of early 2020. The other two clades of 2020 comprise the majority sequences in Europe (20B) and North America (20C).

While the classification proposed by the GISAID platform is based on the combination of nine genetic markers that allows 95% of the SARS-CoV-2 sequences to be classified into six well-defined phylogenetic groups ranging from the two initial groups S and L, until the subsequent evolution of clade L in groups V and G, the latter being finally divided into clades GH and GR. These names refer to mutations that serve to describe the group. For example, the D614G change in the spicule characterized the group described as clade G. The unification of criteria when naming the different clades is a pending task, all these proposals are still under evaluation, since, for example, currently none of these nomenclatures reflects some phenotypic property of the virus, such as antigenic variants, although the virus is antigenically similar so far.

The distribution of the different clades in the countries that make up the WHO European region is very varied, highlighting in Spain the high proportion of clades 19B/S and 20A/G. This fact may be due to a founder effect or sampling biases; but also for the duration of travel restrictions and the different measures implemented in each country. In the case of early travel restrictions, the early incidence would probably be reduced and therefore the most prevalent clades would be 19A/L/V/O compared to 20A/G clades that later became the dominant clades.

Currently, they have managed to sequence thousands and thousands of complete SARS-CoV-2 genomes in the world. This has allowed platforms such as Nextstrain (https://nextstrain.org/) to be able to show the genomic epidemiology of the virus in almost real time and how it has been transmitted over time by the different affected countries (Figure 3) [15].In the Nextstrain platform, 5 large phylogenetic clades are defined to classify the genomes that are being sequenced and that are named based on the estimated year in which they emerged (19 or 20) followed by a letter (Table 1).

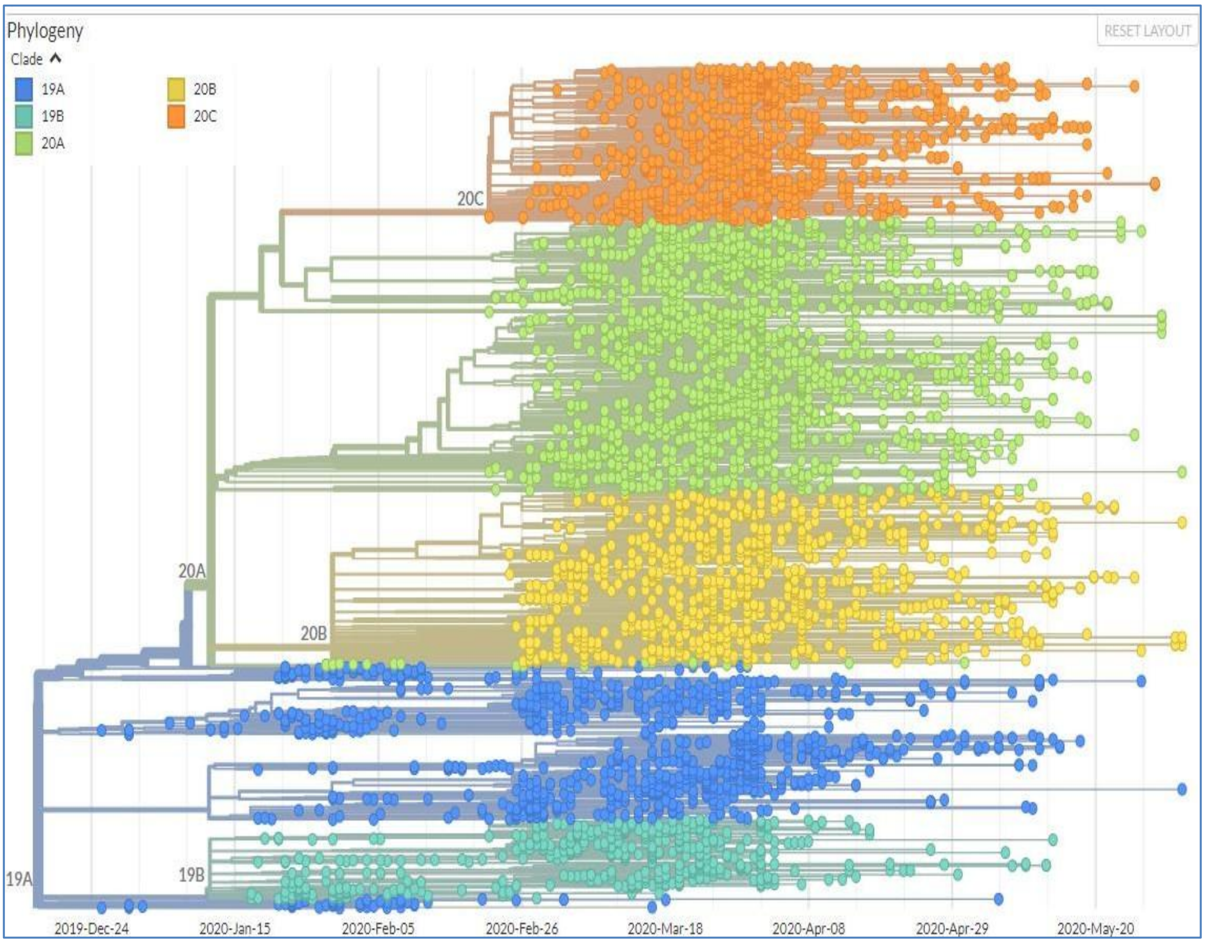

Fig-3: Main phylogenetic clades that are defined in the Nextstrain platform (https://nextstrain.org/)

Table-1: The Nextstrain platform shows the five major phylogenetic clades to classify the genomes that are being sequenced and that are named based on the estimated year in which they emerged 2019 or 2020 followed by a letter.

-19 A: Considered the root clade from which all the others arise and which reached a global frequency between 47% -65% in january 2020.

-19 B: Characterized by the C8782T and T28144C mutations and also reached a high prevalence in Asia in January 2020 (28-33%).

-20 A: Characterized by the C14408T and A23403G mutations, reaching a global frequency of 41-46% in April-May 2020, mainly by countries in North America, Asia and Europe.

-20 B: Characterized by the consecutive mutations G28881A, G28882A and G28883C, which reached a prevalence of around 20% between March-April 2020, mainly due to sequences from Europe.

-20 C: Characterized by C1059T and G25563T, reaching a global frequency of around 20% in april 2020 mainly by sequences from the US.

While in the GISAID database (https://www.gisaid.org/), in which a large part of the genomes that are being sequenced globally are deposited, it classifies said genomes in a different way (16). The large clades defined by GISAID are:

Table-2: Large clades that have been defined by GISAID (https://www.gisaid.org/)

- **Clado S:** Characterized by the presence of the L84S mutation in the virus NS8 protein.
- **Clade V:** Characterized by the presence of the G251V mutation in the virus NS3 protein.
- **Clado G:** Characterized by the presence of the D614G mutation in the spike of the virus.
- **Clado GH:** Which also has the D614 mutation in the virus spike but also has the Q57H mutation in the NS3 protein.
- **Clado GR:** Which also has the D614 mutation in the virus spike but also has the G204R mutation in the virus nucleocapsid protein.

And there is also a third nomenclature proposed by an important group of evolutionary virologists based on the definition of hierarchical lineages A and B that proposes a dynamic system to define the appearance of new local outbreaks with epidemiological importance [17]. Although they are different systems, there are certain equivalences, not perfect, between the different clades defined by the three nomenclature systems. Thus, lineage B would include viruses from clades G, GH and GR of the GISAID system and clades 20A, 20B and 20C of the Nextstrain system. For its part, lineage A would include clades S and V of the GISAID system and clades 19A and 19B of Nextstrain.

There is also a Spanish public initiative called NextSpain (https://evosalut1.uv.es/) that facilitates the analysis and interpretation of results on the epidemiology and evolution of SARS-CoV-2 in Spain following the Nextstrain procedures. In the Figure it can be seen how the Spanish genomes are distributed in the clades defined by Nextstrain.

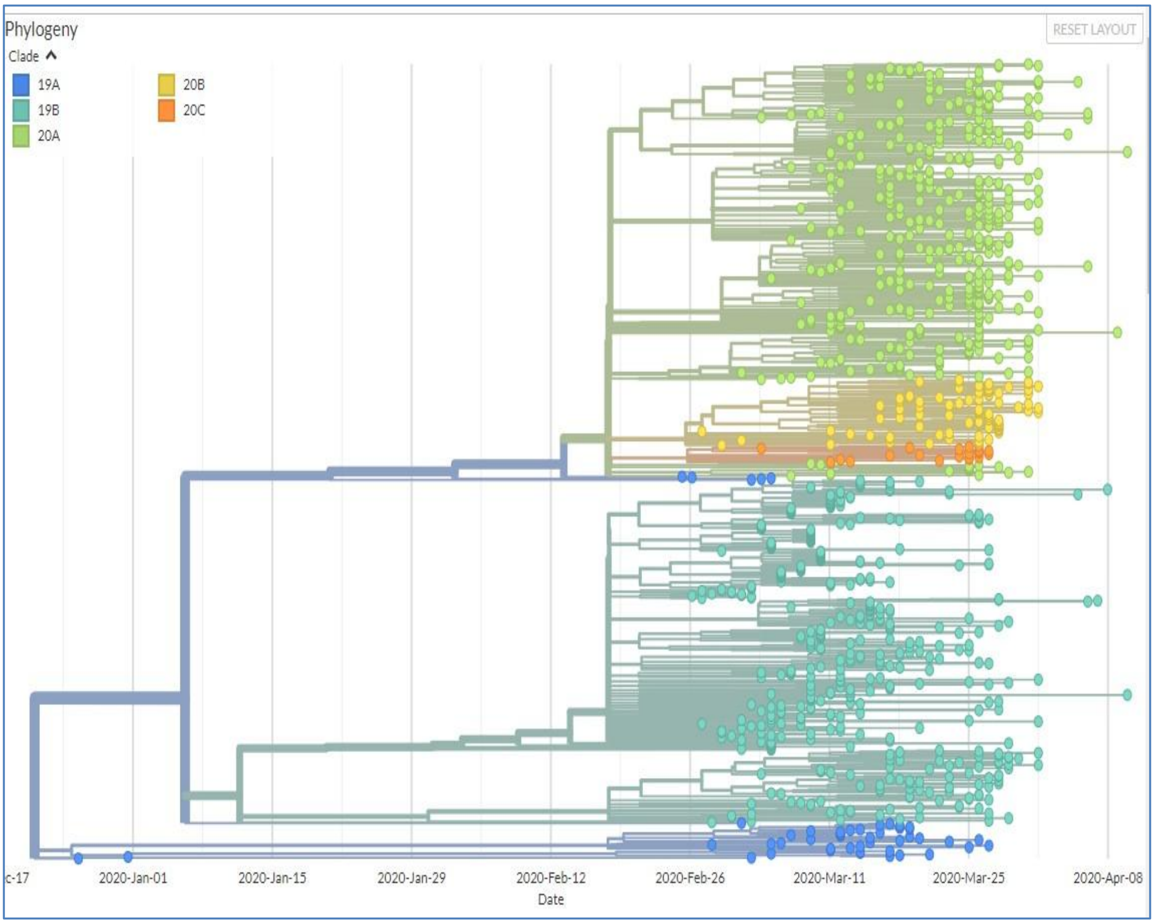

Fig-4: Phylogenetic clades of sequences from Spain defined by Nextstrain and included in NextSpain (https://evosalut1.uv.es/).

In Europe the situation seems to be dominated by clades characterized by the D614G mutation in the virus spike (clades G), in such a way that they reach a frequency higher than 70% in countries such as Austria, Belgium, Denmark, Finland, France, Germany, Greece, Iceland, Italy, Luxembourg, Portugal, Scotland, Sweden and Switzerland. Countries such as England (64%), Holland (63%), Wales (61%) and Spain (52.9%) have the lowest rates on the continent.

But the frequency of the L84S mutation in the NS8 protein in most European countries is less than 1% (Belgium, Denmark, England, Finland, France, Italy, Norway, Sweden, Switzerland and Wales). In other countries such as Luxembourg (2%), Portugal (3%), Iceland (3%), Germany (3%), Scotland (4%) and Greece (8%) this frequency is somewhat higher.

Spain is the European country with the highest frequency of viruses of this phylogenetic clade, reaching 40% and adding 3 out of 5 viruses of this group sequenced in Europe and it seems that these variants arrived earlier and in greater intensity in Spain and have managed to stay in time and this is what is known as the founder effect.

In our country, the oldest sequenced genomes come from samples from Madrid, Valencia, Segovia and Granada from the end of February 2020 and from Guadalajara, Burgos, Álava, Vizcaya, La Rioja, Tenerife and Orense from the first week of March [16]. Of these 15 sequences, 9 belong to clade 19B, 3 to clade 20A and one to clusters 19A, 20B and 20C [16]. The fact that clade 19B has been found in at least 9 different provinces suggests that a founder effect could have been produced that would explain the high prevalence of this clade in Spain in contrast to what was observed in other countries.

SARS-Cov-2 MUTATIONS

One of the most important genes in the virus genome is the S gene, which encodes the spike of the virus. And it is this protein that facilitates the entry of the virus into the target cell through the binding of the S1 subunit to the cellular ACE2 receptor and the subsequent fusion of the viral and cellular membranes through the S2 subunit. The D614G mutation reached a global prevalence of more than 40% in april 2020, raising alarms about the possible biological significance of this mutation in which a polar amino acid with a charged side chain, aspartic acid, is replaced by a small amino acid. no side chain like glycine.

In recent studies, viruses with the D614G mutation in the spicule gene began to spread in Europe in late January-early February 2020 and the moment it is introduced into new territory it is capable of becoming the form dominant [20]. The researchers defend that this fact has taken place globally and also in countries such as England, France, Germany, Italy, the Netherlands, Japan, the USA or Australia, and even at the level of cities such as New York or Washington.

However, analyzing the data of the first 300 complete genomes of Spain, this trend does not seem to be fulfilled, since the G614 viruses have not succeeded in displacing the D614 viruses and both genetic forms seem to coexist with similar frequencies. A possible explanation for this fact in Spain could be a founding effect; that is, it is possible that the D614 viruses (and especially the clade 19B viruses) will arrive earlier and in greater numbers at different points of the Spanish geography. In fact, among the 15 oldest genomes in Spain (detected during the last week of February and the first of March 2020), 10 of them were D614 (67%) and 5 G614 (33%) [18].

And different hypotheses are studied about the possible advantages of the G614 on viruses with respect to the D614 variants that originally circulated in the city of Wuhan [19]. These advantages would be based on the structural and / or immunological characteristics. From a structural point of view, the G614 mutation could act by decreasing the interaction between the S1 and S2 subunits, facilitating the release of S1 [19].

In another study, it was analyzed how both integrated variants behave in a pseudotyped retrovirus in cell cultures of HEK293T cells [20]. In which a greater infectivity has been observed in pseudoviruses with the G614 variant and a greater incorporation of protein S in the pseudovirion compared to the D614 variant. Although binding to the receptor and its neutralization by plasma from convalescent individuals was equivalent, the authors conclude that the G614 mutation appears to provide stability to the virus protein S and could be the cause of more efficient transmission [20].

From an immunological point of view, the SARS-CoV equivalent of the G614 mutation appears to integrate into the immunodominant epitope LYQDVNC, which is recognized by antibodies isolated from patients recovered after SARS-CoV infection. (twenty-one). Therefore, this mutation could confer resistance against protective responses mediated by antibodies directed against D614, causing greater susceptibility to reinfection by G614 virus [19].

A third mechanism that could be involved is a phenomenon called antibody-dependent infection amplification (ADE), because the epitope where it is inserted is an immunodominant ADE epitope [21]. Although this type of ADE phenomena have previously been observed in SARS-CoV infection, there is currently no evidence that it plays any role in the case of SARS-CoV-2 [22].

Regarding the possible greater virulence of the G614 variants, there does not seem to be much scientific evidence so far. There is a study that has linked the lethality of SARS-CoV-2 infection with the prevalence of the G614 mutation in different countries of the world, including Spain [23]. These authors found that the estimated lethality correlates with the proportion of G614 virus ($p < 0.02$), although in the case of Spain the data seem to be an outlier in this correlation [23]. However, the biases that have been committed both in the calculation of the fatality rate and the prevalence of the G614 mutation make the conclusions of this study not very robust.

On the other hand, in a cohort of 453 individuals infected by SARS-CoV-2 from Sheffield, England, no differences were observed in the severity of the disease caused by G614 and D614 variants, since the proportions of patients infected by both variants they were similar both in outpatients, as in hospitalizations or in ICU units [19].

Therefore, the circulation of the G614 variant is the one that has been imposed in most countries of the world and in Europe in particular. The explanations for this fact focus on two hypotheses: One of them may be due to the selective advantage of the G614 variants over the D614 or a founder effect, where the variant that arrives earlier and with greater force is the one that ends up prevailing.

Most scientists point to a selective advantage for viruses with the G614 mutation; even greater infectivity has been demonstrated in laboratory tests. However, these results should be viewed with caution, as it is necessary to demonstrate that these results actually represent an advantage in the transmission of the virus.

On the other hand, the analysis of the more than 1600 imports of the virus in the United Kingdom has shown that the majority variant G614 arrived earlier and in greater intensity than the D614 variants. In the case of Spain, the analysis of the oldest variants detected so far (last week of february and first of march), would indicate a high percentage of sequences of clade 19B formed by D614 variants, which could generate simultaneous outbreaks in at least 9 Spanish provinces.

This fact could be in favor of a possible founder effect, and would also explain the high prevalence of clade 19B observed in Spain compared to other European countries. In any way, whether due to an evolutionary advantage or a founder effect or even a combination of both, the monitoring of transmission clusters through genomic epidemiology is essential to know the genetic diversity of the virus present in a specific territory; evaluate their dispersion; study whether the protection measures used have been successful; help in the study of possible new outbreaks, and facilitate decision-making when imposing containment measures, among other issues. With the confirmation in several countries that they have registered the variant of SARS-CoV-2 that has caused many cases of contagion in the United Kingdom, confusion has also arisen regarding the terms used to refer to it. All viruses constantly mutate, but at different rates and with different repercussions. The coronavirus uses special proteins to enter our body and as different infections occur, different errors occur in copying and then mutations or changes occur in the genetic code of viruses, such as the one caused by COVID -19.

In order to understand all this, it should be noted that, when carrying out genetic sequencing or analysis of the virus, from samples taken in different regions of the world, it is when scientists identify certain characteristics by which these mutations can be grouped into variants or lineages. In the case of the United Kingdom, a different mutation of the coronavirus was detected. And for the analysis, phylogenetic trees are created, which are like family trees where all the relatives that shed viruses such as SARS-CoV-2 are expressed.

Each of the branches that arise directly from SARS-CoV-2 are called lineages, which are designated with a series of numbers and a letter to identify them, considering their order of appearance and their genetic makeup.

Lineage B.1.1.7 has different mutations in its genome, approximately 23, but the main one is the one that develops in position 501 of its genetic code, where the amino acid asparagine (N) has been replaced by tyrosine (Y). The abbreviation for this mutation is N501Y, which is also sometimes referred to as S: N501Y, to specify that it is in the peak protein of the virus.

In addition to the UK variant, the 501Y.V2 variant has been identified in South Africa. But both the UK and South African variants share a mutation, that of position 501.

In the United States, so far, a hundred cases of variant B.1.1.7. Have been confirmed, according to the CDC (Centers for Disease Control and Prevention).

We can say, from the outset, that SARS-CoV-2 is one of the different strains of coronaviruses. The two types or strains of coronavirus best known so far are: SARS-CoV-2 found in Wuhan, China, since the end of 2019 and previously it was SARS-CoV, which causes severe acute respiratory syndrome (SARS). So, each of the new types or species of coronavirus is called a strain.

It is imprecise to say that the variant registered in the United Kingdom is either a new one, or to mention that a new strain arrived, since for that to happen, the virus would have to undergo a drastic change or mutation in its gene chain, which has not happened so far. Hence, for now, it is still considered that the vaccines developed so far are still useful to face the virus, and prevent more cases of deaths from Covid-19.

Regarding variant B.1.1.7, reported by the United Kingdom in December, the communication of the number of cases has been worrying, which can be up to 75% more contagious. However, until now there is no scientific evidence that it causes more harm in COVID-19 patients or makes the virus more lethal and that it cannot affect the effectiveness of the vaccines developed so far. Although all this is a loop, the greater the number of cases, the more patients are admitted to hospitals, more go to the ICU and later more patients die.

Genetic sequencing is a technology that allows knowing and deciphering the genetic code that all living beings have and in which it is a matter of reading that code, which contains essential information for its development and operation, as if it were from a genetic instruction book treated. These hallmarks, which define the characteristics and genetic signature of biological organisms, are inscribed in molecules called nucleic acids, made up of nucleotides.

Throughout 2020, thousands and thousands of complete coronavirus genomes have been sequenced, thanks to the analysis of samples from patients affected by the Covid-19 disease. Achieving this sequencing is essential to better understand the virus and define its characteristics and behavior. From the outset, the sequencing allowed it to be classified, defined and included as a new member of the already known virus families.

The genomic sequencing of SARS-CoV-2 has made it possible to find out its origin, know how it is transmitted, investigate its capacity for diffusion and contagion, and obtain the necessary information for the future development of drugs and vaccines.

Therefore, complete genome sequencing is on the way to establishing itself as a reference technique in the field of molecular identification and characterization, driven by the enormous sequencing capacity of the different current massive sequencing platforms and the reduction of costs per nucleotide.

At present, all research centers are capable of genetic sequencing and there are different technologies to carry it out. Sanger sequencing, one of the first to be developed and key to automating the sequencing process, remains a reference. Over time, new technologies have emerged that allow more information to be obtained from the sequenced organism more quickly. These include technologies such as Illumina and IonTorrent, considered part of the second generation of genomic sequencing, and Pacific Bioscience and Oxford Nanopore, which are already part of a third generation of this technology.

Genetic sequencing is a technique that is developed in practically all research laboratories today and there are different technologies to achieve it. The reference method currently is Sanger sequencing by capillary electrophoresis, but in the last 20 years there has been a great development of new methods that are known as high-throughput sequencing and that include second and third generation methods. All these methods coexist with each other, since they have different applications.

1.- Sanger sequencing by capillary electrophoresis. It is a modification of the strategy designed in 1975 by the scientist Sanger in which chemically modified nucleotides were used so that when added to a new chain that is being formed, it could not continue, functioning as stop nucleotides. The development of fluorescent techniques, the improvement in the enzymes necessary to carry out the process and the introduction of capillary electrophoresis made it possible to automate the process and reach current equipment [24, 25].

2.- Second generation high performance methods. The characteristics of these new methods is their ability to carry out millions of sequencing reactions simultaneously. The development of these methods allowed great advances in the Human Genome Project. And there are currently two major technologies, Illumina and IonTorrent, which are based on different principles.

3.- Third generation high performance methods. These methods go one step further and are capable of sequencing DNA molecules without prior amplification, without the need to follow the sequencing-by-synthesis strategy that second-generation methods follow. These types of methods include technologies such as Pacific BioSciences or Oxford Nanopore.

Spain, like many other countries in the world, has sequenced numerous complete genomes, so that researchers have more tools to advance in different applications against SARS-Cov-2; among them we have:

- Classification of the virus. The analysis of the SARS-CoV-2 genome has made it possible to locate this new virus in the tree of diversity of viruses known to date through what is known as phylogenetic analysis. Thus, the sequencing of the genome of the virus isolated from one of the first patients detected in the Chinese city of Wuhan has made it possible to classify the sequenced virus as a new member of the Coronaviridae family, Orthocoronavirinae subfamily, Betacoronavirus genus, Sarbecovirus subgenus, coronavirus species related to acute severe respiratory syndrome (SARS) (2)

- Origin of the virus. The sequencing of complete genomes of the virus has made different hypotheses about the origin of the virus. This is possible because it has been possible to compare its sequence with that of other viruses isolated in animals and whose genome has been previously sequenced and published so that any researcher can study it. In this way, the SARS-CoV-2 genome that currently most closely resembles is a virus isolated in bats with which it shares 96% of its genome. It has also been determined that it is very similar to another coronavirus isolated in a pangolin and with which it shares 90% of the genome [27, 28].

- Transmission of the virus. Comparing the genomes that are being sequenced around the world as well as information on when and where these samples were obtained allows researchers to identify the possible start of the epidemic, trace possible transmission routes between cities and countries, and monitor its geographical spread. And they allow to know how much the virus is mutating.

- Pathogenicity of the virus. Thanks to the sequencing of the virus genome and its subsequent analysis, it has been possible to identify that its genome is composed of a single RNA strand of positive polarity formed by approximately 30,000 nucleotides. At least 6 open reading frames (ORFs) are known, including ORF1a and ORF1b, which encode two polyproteins that are processed by at least three viral proteases to produce 16 non-structural proteins. The other ORFs code for structural proteins that include spike, membrane, envelope, and nucleocapsid proteins [29]). Thus, we can know the specific sequence of the proteins that make up the virus envelope, crucial for the process of assembly and release of the virus. Or the spike glycoprotein, which we know is composed of two subunits (S1 and S2) and which occurs in the viral particle as a homotrimer and thus binds to the cellular receptor ACE2. In addition, by analyzing its sequence we can know that the S2 subunit contains a fusion peptide, a transmembrane domain and another cytoplasmic domain and that it is highly conserved, which could be an important therapeutic target. We also know that the S1 subunit contains the receptor-binding domain and that it is much less conserved with respect to other coronaviruses, with an amino acid identity of 40%. And also the existence of other genes that do not present homology with any other coronavirus gene found so far, such as ORF3b or the presence of other genes such as ORF8, which code for a protein structurally different from others found in SARS-CoV. Further study of these genes may be key to understanding the pathogenicity of the virus [29].

- Design of antiviral drugs. The study of the sequence of the virus allows us to know which positions of the virus are key to infect human cells. In this sense, thanks to the information generated in other related viruses, such as SARS or MERS, we know that residues L455, F486, Q493, S494, N501 and Y505 of the spike protein are key for binding to the human receptor. ACE2 [30]. Also that certain rare variants already observed in SARS-CoV-2 viruses such as V483A, G476S, L455I, F456V and S494P, have previously been associated with a slightly lower affinity for the receptor, as well as an altered antigenicity in equivalent positions in the viruses. MERS and SARS-CoV [31-33]. Just as the spike protein of SARS-CoV-2 is quite different from that of SARS-CoV, having an identity of 76%, there are other proteins that are much more conserved and that constitute other possible targets of the virus on which to design antiviral drugs. These proteins are its protease and its polymerase with which it shares 96-97%, respectively. Therefore, any drug designed against these SARS-CoV proteins could have a high probability of also working against SARS-CoV-2.

-Design of vaccines. There are different strategies for the development of vaccines against SARS-CoV-2. One of the vaccines such as China consists of integrating the SARS-CoV-2 spike gene into the genome of another virus (adenovirus). And with this it is achieved that the immune system of the vaccinated person reacts against said SARS-CoV-2 protein and that it acts at the moment that this vaccinated person becomes infected by the virus. For its part, another of the US vaccines is an RNA molecule also based on the SARS-CoV-2 spike gene.

There are many other strategies that include inactivated or attenuated viruses where the vaccine product would consist of inoculating the complete virus so that it is not capable of triggering the disease but a powerful immune response against the virus. All these strategies require knowing the genome or parts of it in order to start designing the vaccine itself [34].

How the SARAS-Cov-2 genomic sequencing study is carried out is summarized schematically in figure 5. In which, the genome or the complete genomic sequence of SARS-Cov-2 (RNA) has a code of 30,000 letters that are needed to be able to form copies of itself and to replicate or multiply in the target cell.Once the RNA of the sample is collected, the RNA of the virus must be purified and separated from the RNA of the patient, following different strategies and then Determines the molecule's sequencing through chemical processes and bioinformatics analysis. Once the genomic sequence of the virus has been obtained, this information can be used for different applications, such as the design of vaccines, the transmission of the virus worldwide, studying the origin of the virus, to see the pathogenicity of the virus or for the design of drugs.

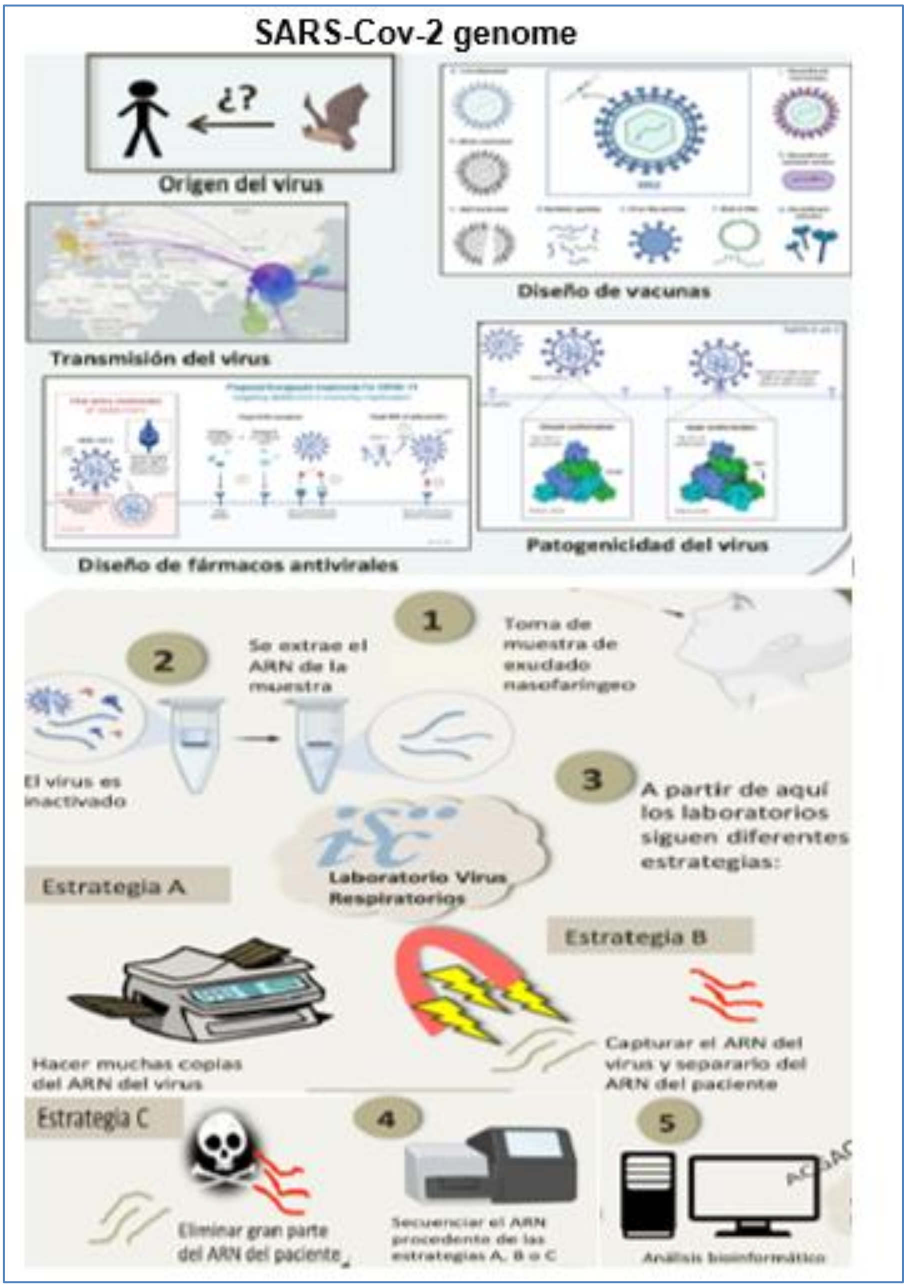
SARS-Cov-2 genome
¿?
Origen del virus
Diseño de vacunas
Transmisión del virus
Patogenicidad del virus
Diseño de fármacos antivirales
1
Toma de muestra de exudado nasofaringeo
2
Se extrae el ARN de la muestra
El virus es inactivado
3
A partir de aquí los laboratorios siguen diferentes estrategias:
Laboratorio Virus Respiratorios
Estrategia A
Estrategia B
Hacer muchas copias del ARN del virus
Capturar el ARN del virus y separarlo del ARN del paciente
Estrategia C
Eliminar gran parte del ARN del paciente
4
Secuenciar el ARN procedente de las estrategias A, B o C
5
Análisis bioinformático

Fig-5: Scheme of the genomic sequencing of SARS-Cov-2. Taken from Francisco Díez-Fuertes. https://www.isciii.es/InformacionCiudadanos/DivulgacionCulturaCientifica/DivulgacionISCIII/Paginas/Divulgacion/InformeCoronavirusSequenciacion.aspx

Bioinformatic analysis of the genomic sequences of sars cov-2

Bioinformatics is a fundamental discipline in the analysis of big data derived from high-throughput technologies such as genome sequences generated by massive sequencing. And in this pandemic, it has had great relevance, generating the methodology and providing the necessary tools for the analysis, processing and interpretation of the data obtained in the sequencing of the viral genome. And it was very quickly possible to obtain the first complete sequence of the virus genome, and to be deposited in public databases such as GeneBank, GISAID, etc [2].

At the time when the WHO determined in January 2020 that SARS-CoV-2 was a public health emergency of global importance, the bioinformatics community assumed the responsibility of creating standardized and efficient analysis protocols, adapted to the characteristics of the genome of the SARS-CoV-2 virus. The first analysis protocols for the assembly of the SARS-CoV-2 genome were those developed by the ARTIC network and by the GalaxyProject community. And the number of samples that could be sequenced with nanopore-based devices is much lower than that obtained with Illumina, which is why the analysis methods were adapted to the data generated by these sequencers, also expanding it to others. Viral genome enrichment methods, such as the use of capture probes.

With the efforts of many bioinformaticians (https://github.com/virtual-biohackathons/covid-19-bh20), a great variety of analysis protocols and collaborations have emerged around the world, resulting in tools such as Viralrecon (https://github.com/nf-core/viralrecon) to reconstruct the viral genome from the massive sequencing data.

There have been two approaches on which the different protocols for obtaining the sequence of viral genomes are based, those based on reference genomes and those for de novo assembly.

Those based on reference genomes consist of mapping the readings of the samples on the genome of the Wuhan SARS-CoV-2 virus, with the subsequent determination and filtering of variants between both sequences and generation of a consensus genome that contains the own variants of the analyzed sample [38]. But this approach would have a clear disadvantage, which is that it would not allow the identification of structural variants in the viral genome that were not in the reference genome used. However, the SARS-CoV-2 virus does not seem to have varied enough since its appearance that this strategy is not effective enough, the main reason why it is the most widely used worldwide.

Those based on de novo assembly, although less widely used, would allow obtaining these structural variants and consist of de novo assembly of the readings obtained from the sequencer, without using a reference genome. Although different programs were already available for this type of assembly, some have been optimized for the reconstruction of the SARS-CoV-2 genome, such as CoronaSpades [39].

This would be the analysis option to choose if an enrichment approach is made using probes and it is not recommended in the case of amplicons because the differences in sequencing depth between them can generate artifacts in the assembly. Once we have determined the consensus genomes, they can be uploaded to the existing public repositories, such as GISAID or ENA (https://www.ebi.ac.uk/ena/browser/about), so that they are available to the scientific community.

These and other platforms and repositories are making many efforts to unify quality criteria and analysis, so that the comparison of virus genomic sequences reveals robust phylogenetic relationships that also facilitate understanding the temporal evolution of the virus and thus be able to determine the transmission chains in the world [37]. In order to help know the circulating virus variants, which is necessary information to maintain effective PCR-based viral diagnostic tools, monitor a possible vaccine and know its efficacy, or identify viral quasispecies with possible impact on the future development of the pandemic.

CONCLUSIONS

As we know, we are witnessing the fourth wave of Covid-19 and how the different restriction and containment measures proposed by different governments are not enough and how this cannot be stopped until the Covid-19 vaccine arrives and is applied in all countries, until the famous herd immunity is achieved. Which is why people are experiencing the famous pandemic fatigue. As long as this does not occur and at the moment in which the restriction measures are relaxed, we will be able to attend a fourth and another fifth wave of SARS Cov-2.

With the new variants of SARS-Cov-2 found (British, South African, Brazilian, Indian, etc.), since as they progress, we would be talking about sub-pandemics, within the global SARS-Cov-2 pandemic, so there are to see how they evolve.

It is thought that some of the new variants of SARS-Cov-2, mainly the South American and the Brazilian, Indian and others that may be emerging, will affect the protection produced by the different vaccines and these vaccines will have to change over time or as in the case of influenza, the different variants of the virus remain as seasonal viruses.

REFERENCE

1. Kennedy, D.A., Read, A.F. (2017). Why does drug resistance readily evolve but vaccine resistance does not?. Proc. R. Soc. B 284: 20162562.
2. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... & Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265-269.
3. Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva, A. A., ... & Ziebuhr, J. (2020). Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol, 5(4), 536-544.
4. Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P., & Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus evolution, 6(2), veaa061.
5. Li, X., Zai, J., Zhao, Q., Nie, Q., Li, Y., Foley, B. T., & Chaillon, A. (2020). Evolutionary history, potential intermediate animal host, and cross- species analyses of SARS- CoV- 2. Journal of medical virology, 92(6), 602-611.
6. Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y. P., ... & Fu, Y. X. (2004). Moderate mutation rate in the SARS coronavirus genome and

its implications. BMC evolutionary biology, 4(1), 1-9.

7. Cotten, M., Watson, S. J., Kellam, P., Al-Rabeeah, A. A., Makhdoom, H. Q., Assiri, A., ... & Memish, Z. A. (2013). Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. The Lancet, 382(9909), 1993-2002.
8. Xu, Y. (2020). Unveiling the Origin and Transmission of 2019-nCoV. Trends in Microbiology, 28(4):239-240.
9. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. Nature medicine, 26(4), 450-452.
10. Benvenuto, D., Giovanetti, M., Salemi, M., Prosperi, M., De Flora, C., Junior Alcantara, L. C., ... & Ciccozzi, M. (2020). The global spread of 2019-nCoV: a molecular evolutionary analysis. Pathogens and global health, 114(2), 64-67.
11. Böhmer, M. M., Buchholz, U., Corman, V. M., Hoch, M., Katz, K., Marosevic, D. V., ... & Zapf, A. (2020). Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. The Lancet Infectious Diseases, 20(8), 920-928.
12. Hodcroft, E. B. (2020). Preliminary case report on the SARS-CoV-2 cluster in the UK, France, and Spain.
13. Oliver Pybus & Andrew Rambaut with Louis du Plessis, Alexander E Zarebski, Moritz U G Kraemer, Jayna Raghwani, Bernardo Gutiérrez, Verity Hill, John McCrone, Rachel Colquhoun, Ben Jackson, Áine O'Toole, Jordan Ashworth, on behalf of the COG-UK consortium. Preliminary analysis of SARS-CoV-2 importation & establishment of UK transmission lineages. 2020 https://virological.org/t/preliminary-analysis-of-sars-cov- 2-importation-establishment-of-uk-transmission-lineages/507
14. Robson, S., Scarlett, G., Bourgeois, Y. X. C., Beckett, A. H., & Loveson, K. (2020). Genomic epidemiology of SARS-CoV-2 spread in Scotland highlights the role of European travel in COVID-19 emergence.
15. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., ... & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), 4121-4123.
16. Elbe, S., & Buckland‐ Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges, 1(1), 33-46.
17. Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., ... & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology, 5(11), 1403-1407.
18. Yeh, T. Y., & Contreras, G. P. (2020). Faster de novo mutation of SARS-CoV-2 in shipboard quarantine. Bull World Health Organ.
19. Korber, B., Fischer, W., Gnanakaran, S. G., Yoon, H., Theiler, J., Abfalterer, W., ... & COVID, S. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. BioRxiv.
20. Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Rangarajan, E. S., Izard, T., ... & Choe, H. (2020). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. BioRxiv.
21. Wang, Q., Zhang, L., Kuwahara, K., Li, L., Liu, Z., Li, T., ... & Liu, G. (2016). Immunodominant SARS coronavirus epitopes in humans elicited both enhancing and neutralizing effects on infection in non-human primates. ACS infectious diseases, 2(5), 361-376.
22. Jaume, M., Yip, M. S., Cheung, C. Y., Leung, H. L., Li, P. H., Kien, F., ... & Peiris, J. M. (2011). Anti-severe acute respiratory syndrome coronavirus spike antibodies trigger infection of human immune cells via a pH-and cysteine protease-independent FcγR pathway. Journal of virology, 85(20), 10582.
23. Becerra‐ Flores, M., & Cardozo, T. (2020). SARS‐ CoV‐ 2 viral spike G614 mutation exhibits higher case fatality rate. International journal of clinical practice, 74(8), e13525.
24. Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of molecular biology, 94(3), 441-448.
25. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J., & Hood, L. E. (1985). The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic acids research, 13(7), 2399-2412.
26. Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., ... & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The lancet, 395(10223), 507-513.
27. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., & Garry, R. F. (2020). The proximal origin of SARS-CoV-2. Nature medicine, 26(4), 450-452.
28. Lam, T. T. Y., Jia, N., Zhang, Y. W., Shum, M. H. H., Jiang, J. F., Zhu, H. C., ... & Cao, W. C. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature, 583(7815), 282-285.
29. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2020). Features, evaluation and treatment coronavirus (COVID-19) StatPearls Publishing. Treasure Island, FL, USA.
30. Wan, Y., Shang, J., Graham, R., Baric, R. S., & Li, F. (2020). Receptor recognition by the novel

coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. Journal of virology, 94(7).

31. Kleine-Weber, H., Elzayat, M. T., Wang, L., Graham, B. S., Müller, M. A., Drosten, C., ... & Hoffmann, M. (2019). Mutations in the spike protein of Middle East respiratory syndrome coronavirus transmitted in Korea increase resistance to antibody-mediated neutralization. Journal of virology, 93(2).
32. Wu, K., Peng, G., Wilken, M., Geraghty, R. J., & Li, F. (2012). Mechanisms of host receptor adaptation by severe acute respiratory syndrome coronavirus. Journal of Biological Chemistry, 287(12), 8904-8911.
33. Rockx, B., Donaldson, E., Frieman, M., Sheahan, T., Corti, D., Lanzavecchia, A., & Baric, R. S. (2010). Escape from human monoclonal antibody neutralization affects in vitro and in vivo fitness of severe acute respiratory syndrome coronavirus. The Journal of infectious diseases, 201(6), 946-955.
34. Amanat, F., & Krammer, F. (2020). SARS-CoV-2 vaccines: status report. Immunity, 52(4), 583-589.
35. van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., ... & Balloux, F. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution, 83, 104351.
36. Díez-Fuertes, F., Iglesias-Caballero, M., García-Pérez, J., Monzón, S., Jiménez, P., Varona, S., ... & Casas, I. (2021). A founder effect led early SARS-COV-2 transmission in Spain. Journal of virology, 95(3).
37. Alm, E., Broberg, E. K., Connor, T., Hodcroft, E. B., Komissarov, A. B., Maurer-Stroh, S., ... & Pereyaslov, D. (2020). Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance, 25(32), 2001410.
38. Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., ... & Andersen, K. G. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome biology, 20(1), 1-19.
39. Meleshko, D., Korobeynikov, A. (2020). CoronaSPAdes: from biosynthetic gene clusters to coronaviral assemblies," bioRxiv, 2020:p. 2020.07.28.224584.