UNIVERSITAT ROVIRA i VIRGILI

Mutational analysis of SARS-CoV-2 hybridization sites for real-time RT-qPCR primers and *de novo* design of Omicron-specific primers

Nil Novau Ferré

TREBALL FINAL DE GRAU BIOTECNOLOGIA

Tutor acadèmic: Dr. Santi Garcia-Vallvé, Departament de Bioquímica i Biotecnologia, URV (santi.garcia-vallve@urv.cat)

En cooperació amb: Grup de recerca en Quimioinformàtica i Nutrició (QiN), Departament de Bioquímica i Biotecnologia, URV.

Supervisors: Dr. Santi Garcia-Vallvé (santi.garcia-vallve@urv.cat) i Dr. Gerard Pujadas Anguiano (gerard.pujadas@urv.cat), Departament de Bioquímica i Biotecnologia, URV.

Juny 2022

"Research is seeing what everybody else has seen and thinking what nobody else has thought."

Albert Szent-Györgyi

Jo, Nil Novau Ferré , amb DNI 21705347-V, sóc coneixedor de la guia de prevenció del plagi a la URV *Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants* (aprovada el juliol 2017) (http://www.urv.cat/ca/vida- campus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquet TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 7 de Juny de 2022

Novau Ferré, Nil

Firmado digitalmente por Novau Ferré, Nil
Fecha: 2022.06.07 11:25:51 +02'00'

TABLE OF CONTENTS

The current project was carried out in the Cheminformatics and Nutrition (QiN) research group of Rovira i Virgili University's Biochemistry and Biotechnology Department (URV), under the supervision of Dr. Santi Garcia-Vallvé and Dr. Gerard Pujadas Anguiano. The QiN group's research focuses on the use of computational approaches in the development of inhibitors and/or the repurposing of existing compounds or treatments to specific targets. Due to the recent SARS-CoV-2 pandemic, the QiN group has started a new research line to find novel antiviral drugs, as well as different SARS-CoV-2 genome mutational analysis and variant monitoring. This project is part of the previously described research line.

ABSTRACT

Severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2), the cause of a new coronavirus illness (COVID-19), is highly contagious and has spread rapidly worldwide since its discovery. Although many studies have been carried out and different drugs have been developed, there is still no definitive and effective cure for the treatment of the virus. This is why it's critical to keep a close eye on the virus's mutational evolution to spot emerging variants that could develop viral resistance to medium to long-term treatments and vaccines. In this work, we perform a mutation profiling of more than 4,000,000 SARS-CoV-2 genomes. On the other hand, quantitative nucleic acid testing has become the gold standard for determining the cause of an infection. The RT-qPCR tests for SARS-CoV-2 face several difficulties, because as the virus evolves and the target sequences diverge from the selective primer sequences, the approach may lose sensitivity. As we rely on existing RT-PCR primers to detect and manage the spread of the Coronavirus, it is critical to understand how SARS-CoV-2 mutations differ from current primers over time. We measure the number of mismatches between primer sequence and genomic targets to assess the performance of the SARS-CoV-2 primers currently in use. Additionally, we show variant-specific primer design with high sensitivity and specificity to the Omicron variant.

ABBREVIATIONS

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2
COVID-19: COrona VIrus Disease 19
WHO: World Health Organization
CoVs: Coronavirus
VOCs: Variants of Concern
VOIs: Variants of Interest
NTD: N-terminal domain
RBD: Receptor-binding domain
ACE2: Angiotensin-converting enzyme 2
CDC: Centers for Illness Control and Prevention
ART: Antigen rapid test
LFA: Lateral Flow Assay
SGTF: S gene target failure
FNR: False negative rate
NGS: Next Generation Sequencing

Keywords: SARS-CoV-2, COVID-19, mutational analysis, genomic profiling, primer analysis, primer design

1. INTRODUCTION

1.1. SARS-CoV-2 pandemic

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, GenBank NC_045512.2), the causative agent of the COrona VIrus Disease 19 pandemic (COVID-19) infection, was discovered and reported to the World Health Organization (WHO) in Wuhan, China, in December 2019. Since then, the virus has spread rapidly all over the world, and as of May 2022, over 518 million confirmed cases and over six million deaths were reported globally[1]. Even though, new weekly COVID-19 cases have stabilized during these last months (April – May 2022)[1].

The death rate among COVID-19 patients ranged between 2% and 4%, with a high of 13%[2]. As a result, the COVID-19 pandemic has wreaked havoc on health-care systems, closed schools and communities, and dove the world into an economic recession. 2020 was a challenging year, 2021 was difficult with the emergence of multiple variants of SARS-CoV-2 and finally, in 2022 the pandemic appears to have begun its decline. Finally, several countries have started to remove compulsory vaccination, and the measures have been virtually abolished.

1.2. SARS-CoV-2 phylogeny and genomic organisation

It is widely assumed that zoonotic transmission of Coronavirus (CoVs) to humans happens via intermediate host species, where viruses better suited to human receptors can be selected allowing the species barrier to be crossed. The origin of the SARS-CoV-2 outbreak is currently unknown, despite the discovery of similar viruses in bats and pangolins (Figure 1). Severe Acute Respiratory Syndrome (SARS-CoV, GenBank NC_004718.3), Middle East Respiratory Syndrome (MERS-CoV, GenBank NC_019843), and SARS-CoV-2 are all transmitted through zoonotic transmission and spread among humans through intimate contact. Although their similarity, SARS-CoV-2 is less pathogenic than the coronaviruses that cause SARS- and MERS-CoV[3].

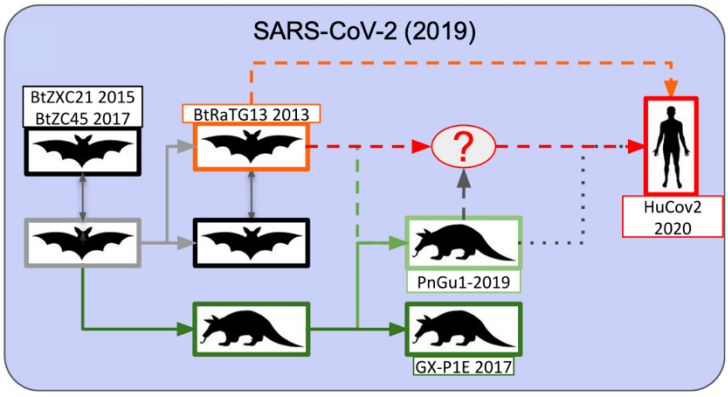

Figure 1 | Unknown intermediate of the virus, which is thought to have originated in a bat or pangolin. Extracted from Sallard *et al.* [24]

1.2.1. Phylogeny organisation

The first coronavirus was found in the 1960s. Coronaviruses are members of the *Coronaviridae* family, which is part of the *Nidovirales* order. They are divided into four genera, which include α-, β-, γ-, and δ- coronaviruses (Figure 2). Between them, α- and β- CoVs infect mammals, γ-coronaviruses infect birds, and δ- coronaviruses infect both mammals and birds[4]. Additionally, β- coronaviruses divide into five different subgenera including *Embecovi*rus (also, known as lineage A), *Sarbecovirus* (lineage B), *Merbecovirus* (lineage C) and *Nobecovirus* (lineage D). Based on this categorization, SARS-CoV-2, SARS-CoV and MERS-CoV are classified in *Sarbecovirus* (the first and second) and *Merbecovirus*, respectively (Figure 2).

It is possible to investigate the phylogenetic relationships of viruses, transmission patterns, evolutionary rates, and the impact of mutations in infection and sickness severity, as well as vaccine development, by comparing multiple genomes. CoVs have the longest RNA viral genomes, ranging from 26 to 32 kb[5]. The SARS-CoV-2 genome, a single-stranded RNA positive sense genome with over 29,903 bp, shares around 79.5% and 50% sequence identity with SARS-CoV and MERS-CoV, with key enzymes and structural components sharing more than 90%

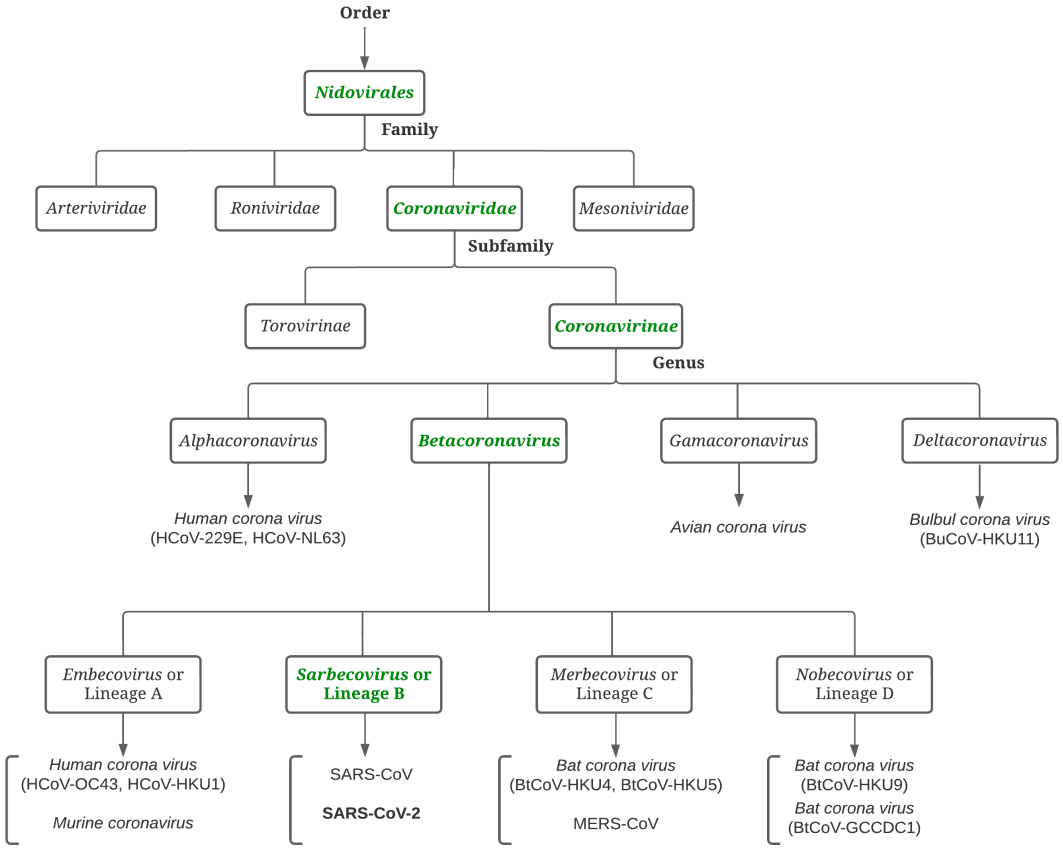

Figure 2 | Tree diagram of the *Nidovirales* order.

sequence identity. Apart from those, the other most related genomes available in public databases are bat-SL-CoVZC45 (GenBank MG772933) with an 87.99% sequence identity and bat-SL-CoVZXC21 (GenBank MG772934) with an 87.23% sequence identity.

SARS-CoV-2 genome is constantly evolving, and several SARS-CoV-2 variants, such as lineage B.1.1.7, and B.1.351 among others, appeared after the outbreak of the COVID-19 pandemic. The identification of specific Variants of Concern (VOCs) and Variants of Interest (VOIs) to prioritize global surveillance and research, and ultimately inform the current response to the COVID-19 pandemic, was prompted by the emergence of variants that posed a greater risk to global public health in late 2020.

1.2.2. Genomic organisation

The genome organization of SARS-CoV-2 was shown to be similar to the related bats and human coronavirus comprising between 6–11 open reading frames (ORFs) encoding ~9680 amino acid polyproteins[6]. It includes four structural proteins: spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins, that have a high sequence resemblance to the SARS-CoV and MERS-CoV homologous proteins (Figure 3A).

As seen in Figure 3B from 5′ to 3′ the genome is as follows: 5′ UTR, the first ORF1ab, the four structural proteins named above, and other accessory proteins encoded by the remaining ORFs. At the overlapping region between ORF1a and ORF1b, the SARS-CoV-2 RNA has a key secondary structural feature: a -1 frameshift-stimulating pseudoknot. This structural feature allows a controlled -1 ribosomal frameshifting at genomic location 13,468, which is required for tightly regulated protein production. This frameshift produces two polypeptides: pp1a and pp1ab. These polypeptides are processed by virally encoded chemo-trypsin-like protease (3CLpro) or main protease (M-pro) and one or two papain-like proteases into 16 nonstructural proteins (nsps, in Figure 3B shown as ns*, where «*» corresponds to the nsp number's). The nsps, which accounts for approximately 70% of the genome, contains two viral cysteine proteases, papain-like protease (nsp3), and chemo-trypsin-like, 3C-like, or main protease (nsp5), an RNA-dependent RNA polymerase (nsp12), a helicase (nsp13), and other proteins that are thought to be involved in SARS-CoV-2 transcription and replication. Complete inhibition of -1 programmed ribosomal frameshifting was observed to significantly diminish SARS-CoV replication by several orders of magnitude[7].

The M protein is regarded as a defensive immunogenic with high capacities to neutralize antibodies, with strong conservation among SARS-CoV-2, SARS-CoV, and MERS-CoV[8]. In addition, the membrane glycoprotein is also required for membrane curvature and packaging of the RNA buds of new particles[9]. N is the only protein that functions primarily to bind to the CoV RNA genome, making up the nucleocapsid. This protein has a critical role in viral assembly[9]. The E protein is required for virus morphogenesis, assembly and budding and it is also involved in the formation of channels in the ER of the host cell[9]. Finally, and probably the best known, the S glycoprotein is a fusion viral protein formed of two subunits, S1 and S2. The S1 subunit, which shares 70% sequence identity with bat SARS-like CoVs and human SARS-CoV, comprises a signal peptide, N-terminal domain (NTD), and receptor-binding domain (RBD). The S2 subunit that shares 99% sequence identity with bat SARS-like CoVs and human SARS-CoV comprises two heptad repeat regions known as HR-N and HR-C, which form the coiled-coil structures surrounded by the protein ectodomain[10]. The surface glycoproteins (S1 and S2) are responsible for binding to the ACE2 (Angiotensin-converting enzyme 2) receptors on the host cell allowing the virus to invade, where S1 bind to the ACE2 receptor and S2 fuses with the host cell membrane[11]. For that reason, spike protein is critical in the commencement of the viral cycle.

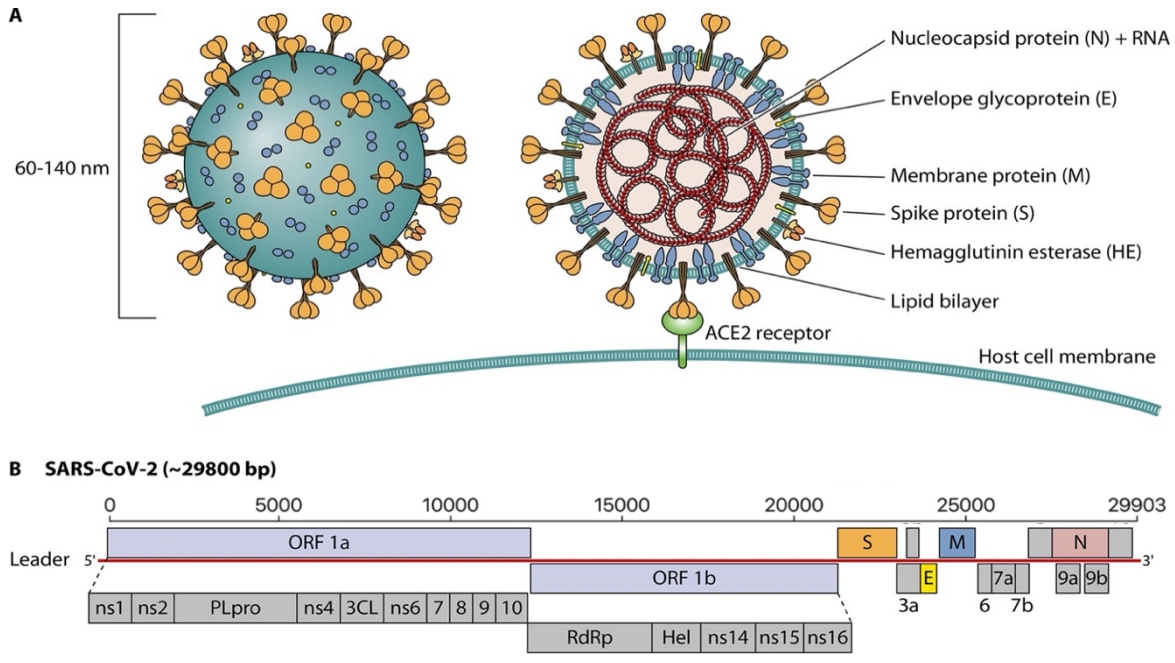

Figure 3 | Genomic organization and proteome of SARS-CoV-2. Figure extracted and adapted from SHS Tali *et al.*[12]

1.3. COVID-19 diagnosis

1.3.1. Viral gene detection by RT-qPCR method

PCR is a very sensitive laboratory technique that has been used in biological and medical sciences to produce both qualitative and quantitative findings. Several modifications to this technique —detailed below— characterize RT-qPCR. RT-qPCR is a diagnostic modification of PCR that is used to detect target RNAs in clinical samples for diagnosing pathogens in molecular diagnostics laboratories. Probe-based RT-qPCR has long been regarded as the gold standard approach for detecting SARS-CoV-2 using upper and lower respiratory tract specimens (nasopharyngeal swab, throat swab, and sputum), and it is now one of the most extensively utilized tests in many countries for population screening, as recommended by the WHO and Centers for Illness Control and Prevention (CDC)[13,14]. It is a dependable and quick technique that produces results in a matter of hours with a high throughput[15].

The RT-qPCR technique consists of two steps: (I) reverse transcription of RNA into complementary DNA (cDNA) and (II) polymerase chain reaction amplification of the cDNA sample using gene-specific primers and fluorescently tagged hydrolysis probes. The first step is used to produce DNA templates, which are then employed in the second step to increase the amount of copies of DNA through repeated heat cycles. Gene-specific primers guide the second reaction, amplifying only the specified fragment of the genome —following the complementarity of bases—, while the probes emit fluorescent signals after each successful amplification of the gene sections. It results in a measurable reaction system[16].

However, the method's sensitivity is its weak point, and any mutation in the area could lead to false negatives. Therefore, a multiplex PCR protocol is used to improve the method, where at least two viral genes are targeted to amplify (Figure 4). This improvement in the technique helps to raise the likelihood of catching the virus, especially in patients with low viral loads. In addition, to minimize false-negative results caused by poor sampling, a human housekeeping gene can be used as an internal control[17] (Figure 4). It is for these reasons that the technique is currently widely used due to its highly reliable results.

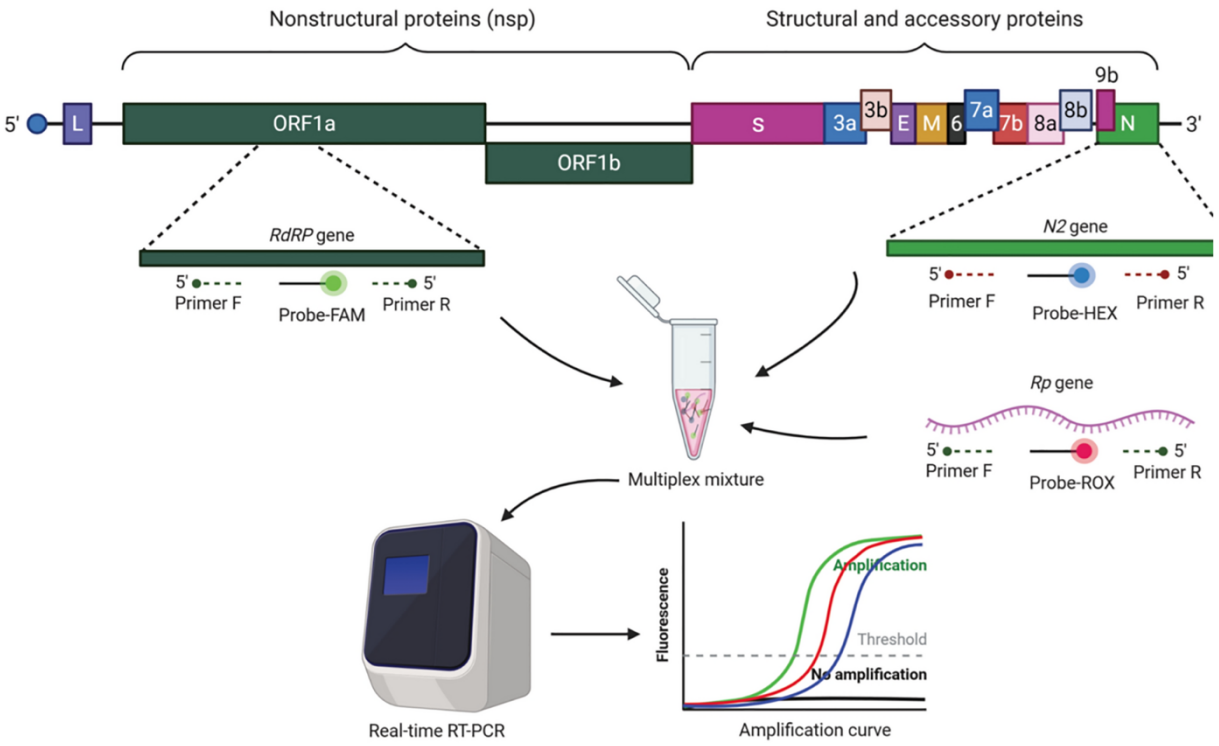

Figure 4 | Multiplex real-time RT-PCR method for the diagnosis of SARS-CoV-2. The image shows an example of amplification of two regions of the SARS-CoV-2 genome: the RdRp gene and the N gene. In addition, the human ribonuclease P (RNase P or RP) gene (responsible for the processing of tRNA molecules) is used as an internal control in multiplex RT-PCR protocols recommended by WHO and CDC. Figure extracted from Huseyin *et al.*[17]

Furthermore, in addition to experimental design, determining the genes to be targeted and the design of multiplex primers and probe sets is a critical goal when developing molecular testing techniques. The formation of an autodimer or heterodimer structure lowers target selectivity and might result in misinterpreted results. Consequently, the experimental design as well as the selection of the best primer and probe sets are crucial and need to be standardised.

Several investigations[18,19] have noted the new coronavirus's genetic diversity and rapid evolution. And, as previously stated, multiple differences in viral RNA sequences, might impact the RT-qPCR results and mutations in the primer and probe target areas of the SARS-CoV-2 genome can cause false-negative results. Although the real-time RT-PCR assay was designed as precisely as possible based on the conserved regions of the viral genomes, the possibility of erring in an outcome due to the several genome mutations has increased the utilization of multiple target gene amplification to avoid these incorrect results[20].

1.3.2. Antigen detection methods

An antigen is a particle, fragment, or molecule that can activate the immune system and cause antibodies to be produced in order to combat diseases and protect the body. Unlike PCR-based approaches, an antigen rapid test (ART), also known more simply as a quick test, is a sort of rapid diagnostic test that determines if an antigen is present or absent for point-of-care and laboratory testing. In the case of diagnosing COVID-19, SARS-CoV-2 antigen assays detect viral components —such as S glycoprotein, M protein, or released N protein— or the virus directly without the need for thermal amplification[21].

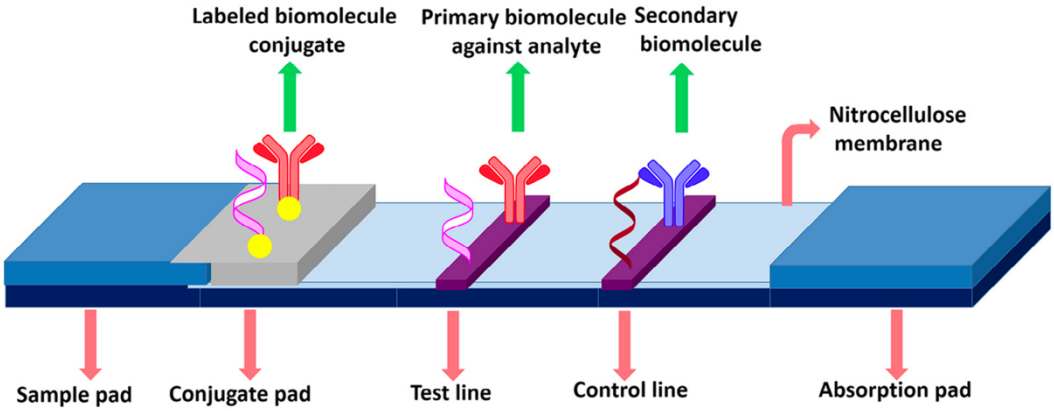

Figure 5 | The basic structure of lateral flow assay. When a sample is dropped on the sample pad, it travels towards the end of the LFA strip due to capillary force. Figure extracted from Bahadir *et al.* [21]

Antigen tests, like PCR-based procedures, only identify the active viral infection, not the state of recovery (in contrast to other diagnostic tests, such as antibody tests). Antigens may be more trustworthy than antibody tests since they come before antibodies and are target-specific. Antigen testing can be run on Lateral Flow Assay (LFA) strips for quick detection (Figure 5) or in an ELISA format for increased sensitivity and high throughput (the simultaneous measurement of 96 samples).

The theory is based on how a liquid sample moves[21] so, the lateral flow antigen assay is inexpensive, may be done by a healthcare provider without specialist training or equipment and have quick turnaround times of less than 5 to 30 minutes. However, the application of an ART kit is limited by its sensibility. A study carried out by Mak *et al.*[22] revealed that the fast antigen detection kit provided by WHO was 100 times less sensitive than RT-qPCR. Moreover, the clinical sensitivity of the ART kit was 68.6% for detecting specimens from COVID-19 patients. Nevertheless, several studies confirm that ART kits are more effective during the acute/recent phase of the COVID-19 disease —*i.e.* in the early and contagious stages of the illness[23].

1.4. SARS-CoV-2 Variants of Concern and Variants of Interest

SARS-CoV-2 was initially thought to be exceptionally well adapted to humans, spreading swiftly with no evidence of natural selection among circulating viruses according to genetic sequencing[24]. With the first reports of emergent SARS-CoV-2 variations associated with greater transmissibility, disease severity, and escape from humoral immunity in the last months of 2020, this changed. Since then, many variants of SARS-CoV-2 have been identified. Until May 2022, the Omicron variant was categorized as VOC because of different characteristics. A VOC, according to the CDC, is the strain that causes increased transmissibility, a more severe disease course, worse treatment efficacy, and a slew of other troubling characteristics[25].

Genetic diversity and global transmission patterns are crucial for understanding pandemic dynamics. And it is important to remark that all the mutations that have prevailed in variants have been due to the advantages they provide to the virus life cycle —*i.e.,* natural selection. Comprehensive genomic diversity analysis of virus sequences from different countries —*i.e.,* VOCs monitoring— will provide insight into the global transmission pattern, virulence and pathogenesis of SARS-CoV-2. Furthermore, this will allow the evaluation of the degree of hybridisation of primers for SARS-CoV-2 detection and the follow-up of cases.

VOCs include several Pango lineages: B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), B.1.617.2 (Delta), and B.1.1.529 (Omicron), while Pango lineages C.37 (Lambda) and B.1.621 (Mu) among others are found in VOIs. Among these, although the disease's pathophysiology is unknown, the greater transmissibility of these variants shows that differences in viral strains are linked to differences in transmission/infectivity and/or severity and underscores the significance of genomic surveillance[26].

In September 2020, the first VOC was detected in the UK and categorized as B.1.1.7 (Table 1). This new variant included nine amino acid modifications in the S gene. One of these, N501Y (Asn501Tyr), improves the affinity of spike for its cellular target ACE2, resulting in increased transmission and likely increased pathogenicity when combined with other less well-characterized alterations[27]. Due to the benefits it provides to the virus' infectious mechanism, this mutation has been sustained in the other VOCs except for Delta.

Table 1 | VOCs and VOIs adapted from the WHO updates on tracking SARS-CoV-2 variants https://www.who.int/en/activities/tracking-SARS-CoV-2-variants

WHO name	Geographic region of first detection	Date first detected	Scientific name (Pango lineage)
Variants of Concern (VOC)			
Alpha variant	United Kingdom	September, 2020	B.1.1.7
Beta variant	South Africa	May, 2020	B.1.351
Gamma variant	Brazil	November, 2020	P.1
Delta variant	India	October, 2020	B.1.617.2
Omicron variant	South Africa	November, 2021	B.1.1.529
Variants of Interest (VOI)			
Epsilon variant	USA	March, 2020	B.1.427/B.1.429
Zeta variant	Brazil	April, 2020	P.2
Eta variant	Multiple countries	December, 2020	B.1.525
Theta variant	Philippines	January, 2021	P.3
Iota variant	USA	November, 2020	B.1.526
Kappa variant	India	October, 2020	B.1.617.1
Lambda variant	Perú	December, 2020	C.37
Mu variant	Colombia	January, 2021	B.1.621

RBD and NTD mutations, as well as deletions, are present in the Beta variant (see Table 5 from VOCs mutation analysis). These alterations increased the transmissibility of the Beta variant by 50% compared to the variants that came before it. Reduced vaccination effectiveness has also been linked to the Beta variant[28]. The Delta form, which was first discovered in India in December 2020, quickly spread across a mostly unvaccinated country, resulting in a large number of infections, hospitalizations, and deaths. This variant is extremely transmissible, with a transmission rate of roughly 60% higher than the Alpha variant[29]. Among the VOCs, the Omicron variant is the highest mutated. This variant has key spike protein mutations that have previously been described in other VOCs (Alpha, Beta, Gamma, and Delta) and VOIs (Kappa, Zeta, Lambda, and Mu) that could speed up S1/S2 cleavage and improve virus-host cell membrane fusion, resulting in increased virus multiplication and infectivity. Furthermore, this variant can avoid convalescent plasma, vaccination sera, and mAbs because of these alterations. In many countries worldwide, the Omicron variant is becoming the most transmissible VOC[30].

There are currently no interest variants circulating, according to WHO. Furthermore, the only variants of concern now circulating are Delta and Omicron, as will be demonstrated in later results. Alpha, Beta and Gamma have been circulating previously, but they have gradually been replaced by the present ones[31].

2. HYPOTHESIS AND OBJECTIVES

In late December 2019, in Wuhan, China, several local health authorities reported clusters of patients with pneumonia of unknown cause; the pathogen was the novel coronavirus (SARS-CoV-2). Since then, an extraordinary amount of research has been published on this topic, owing in part to the scientific attempt to find a cure and effective vaccines. To generate an effective immune response to the virus and to ensure a medium to the long-term effect of these vaccines, their target should not be altered as the virus evolve. Is for that reason that it is critical to keep track of its genetic variations in this regard. Concomitantly, the diagnosis of SARS-CoV-2 infection, which is usually based on quantitative PCR analysis, requires the binding of primers to specific areas of the genome.

More than four million SARS-CoV-2 genomes have been published in the Global Initiative on Sharing Avian Influenza Data (GISAID, www.gisaid.org) as of 6th January 2022, and the number is still growing. A mutational profile of the entire genome, as well as the different primer hybridisation sites, can be derived by analysing this data. This will entail characterizing the various mutations, not just in terms of the number of new mutations discovered, but also in terms of their frequency.

On the other hand, the hybridisation site of the primers is a key point in the test for the detection of the Covid-19 disease. It is for this reason that, with the large number of mutations currently characterised, it is possible to find highly variant-specific areas and areas with lower mutation rates. Furthermore, the best alternative for genotyping SARS-CoV-2 is a next-generation sequencing (NGS); however, this method is not rapid enough and lacks appropriate throughput for the current rise in cases. To this end, it might be interesting to create a speedy and accurate RT-qPCR test that can successfully distinguish Omicron from other SARS-CoV-2 variants.

The main objectives of the present study can be summarized as follows:

1. To characterize a mutational profile of the SARS-CoV-2 genome through:
    a. A global analysis of the genome.
    b. A detailed analysis of the hybridization site of the primers currently in use worldwide.
2. To design specific primers for the most widespread variant in the world today; the Omicron variant

3. MATERIALS AND METHODS

3.1. Data description

GISAID has emerged as a leading source of SARS-CoV-2 genomes, containing the largest number of genomes sequences around the world with metadata about the location and time of collection[32]. SARS-CoV-2 genomes from the GISAID repository were curated collecting high-quality genomes until 6th January 2022. To guarantee the quality of the sequences, we defined high-quality genomes as those that followed the filters: consider only sequences obtained from samples extracted from humans, avoid considering partial sequences by only keeping sequences with a minimum length of 29.000 bp —*i.e.,* complete genomes— and consider only sequences labelled as "high coverage" (*i.e.,* sequences containing: less than 1 % unidentified bases (Ns), less than 0.05% of unique amino acid mutations, to withdraw possible sequencing artefacts, and no insertions and/or deletions, unless verified by the submitter). The complete genome NC_045512.2, isolated from Wuhan-Hu-1 and submitted to the GenBank (https://www.ncbi.nlm.nih.gov/genbank/) database on 17th January 2020 was used as the reference genome. All coding regions of each sequence were then aligned with the respective regions of the reference genome using the blastn algorithm to identify different single nucleotide variations.

Note that all of the RNA sequences in GISAID include thymidine bases instead of uracil nucleotides, which is the right RNA base. This is a conventional sequence analysis procedure, however, it's worth noting that all Ts in the current study will refer to Us in the final RNA sequence.

3.2. Primer information for qPCR

For the comparative analysis of multiple primers for SARS-CoV-2, several primer-probe sets were selected based on sequence information from multiple institutions: the Centers for Disease Control and Prevention (CDC) (USA), Charité – Universitätsmedizin Berlin Institute of Virology (Germany), The University of Hong Kong (Hong Kong), National Institute of Infectious Disease, Department of virology III (Japan), China CDC (China), National Institute of Health (Thailand), among others. The sequences of primer-probe sets and their locations at viral RNA are listed in Table S1.

3.3. openPrimeR

We generated optimised primer sets for amplification of selected segments of the SARS-CoV-2 Spike gene with openPrimeR[33]. To be able to use the library with its full functions, the additional programs MAFFT[34], OligoArrayAux[35], ViennaRNA[36], MELTING[37] and Pandoc[38] were also needed. The reference SARS-CoV-2 genome established for the omicron variant was entered as input and the constraints established and discussed below were adjusted (see General principles of primer/probe design strategy). A FASTA output file was obtained with the different primer options of the restricted area. Once the results were obtained, the primer candidates were checked and re-evaluated using the online server provided by TermoFisher®[39] to re-adjust the T-melting values and check for self-dimerization.

3.4. General principles of primer/probe design strategy

3.4.1. Primer design

To ensure the success of a PCR reaction, multiple primer design criteria have been agreed upon. Primer length should be between 18 and 25 nucleotides. Primer melting temperatures (T-melting) should be between 56°C and 61°C, depending on the GC content. The GC content should be between 40-60%, the closer to 50% the better. The T-melting difference between the two primers should be less than 5°C[40] and the annealing temperature should be 5°C lower than that of the primer with the lower T-melting.

The specificity of the primers will be determined by their ends. Although the 5' end contributes marginally, matching the 3' end will be key to efficient amplification of the region. It has been found that a single mismatch in the last three nucleotides counting from the 3′ end is acceptable, but the second impaired position drops the amplification efficiency[41]. Therefore, for a specific design of the Omicron variant, less attention will be paid to the 5' end of the primer and the focus will be on fixing specificity at the 3' end. These will at most contain two or three G/C at the last 5 bases of the end and the primer will never end with three consecutive G/C residues, as otherwise the complementary dimer or hairpin structure at the 3' end of the primer may cause the PCR reaction to failing, and guanine (G) repeats may prevent complete dissociation of the chain thus reducing the amplification efficiency[42]. However, secondary structures are only predictive, and it may also be necessary to evaluate the dimerisation of the primers upstream and downstream of the amplicon[42] to avoid possible interference in the early stages of the PCR

assay. In addition, to avoid potentially misleading results, the primer must be tested for base complementarity with other sections of the virus genome.

We found multiple areas of polymorphisms in the SARS-CoV-2 genomes and there may be mismatches in primers or probes leading to false negatives and, as the presence of any internal mismatch between primer and plate can greatly decrease PCR efficiency, rather than increase primer degeneracy, the use of multiple primers sets targeting different subgroups is suggested when target sequence polymorphism is high[43].

In order to ensure the correct length of the amplicon, and that qPCR does not need extra time to complete the extension, it has been decided that this should be between 75-150 bp.

3.4.2. Probe design

A similar procedure has been followed for the design of the probe as for the primers. The length of the probes was set between 18 and 25 nucleotides, with a GC content between 40-60%, a higher C than G content, and adjusting the T-melting approximately 10ºC higher than that of the primers used. In addition, it must be in a region close to the forward or reverse primer. G at the 5' end should be avoided to prevent quenching of the 5' fluorophore. Furthermore, the formation of dimers and hairpin structures, as considered in the primer design, should be avoided.

The probe set will be labelled with the 6-carboxyfluorescein [FAM] reporter and the Black Hole Quencher 1 [BHQ1] quencher.

Table 2 | Primers conditions

Requisites	Primers	Probes
GC content	40-60%	40-60%
Calculated primer/probe Tm	58-60%	~10°C higher than the primer Tm
Primer/probe length	20-25 bp	20-25 bp
PCR product length	80-120 bp	80-120 bp
Distance probe to primers	The probe should be in close proximity to the forward or reverse primer	The probe should be in close proximity to the forward or reverse primer
Primer-dimers, hairpins	Avoid	Avoid
3 end rule (3 instability)	Maximum two G or C in the last 5 bp	-
Autoquenching	-	No G on the 5’ end
GC ratio	-	C > G
Degree of degeneracy of bases	Avoid	Avoid

4. RESULTS AND DISCUSSION

4.1. Mutational profile of the SARS-CoV-2 genome

All the results of this work are based on the analysis of 4,616,059 full-length genomic SARS-CoV-2 sequences downloaded from GISAID (www.gisaid.org) from December 2019 to the 6th of January 2022 (see Materials and Methods Data description section). There is a bias in assessing them because sequencing rates in different nations vary greatly. As shown in Table 3, there is a large bias between the different genomes analyzed in the different countries; most of the genomes (55,49%) were deposited only from 2 countries: the USA (30.63%) and the UK (24.86%). On the other hand, surprisingly for us, China, the hypothetical area of origin of the virus, is one of the areas that has contributed the fewest genomes to GISAID, as shown in Figure 6. This bias can be surprising but does not invalidate in any case the results presented here for the design of the primers, because we have been working with the consensus sequences of the variants (see Materials and Methods VOCs description section). In addition, as seen in Figure 7, the temporal distribution of genome-mass sequencing has not been uniform either, with a very considerable increase in March-April-May 2021 and from September to December 2021 —probably due to different waves of infection—.

Table 3 | Top 10 countries with higher contribution to the genomes

Country	Numbers of genomes
USA	1,413,914
United Kingdom	1,147,744
Germany	240,362
Denmark	238,050
Japan	175,032
Canada	152,763
Sweden	109,325
France	101,863
Brazil	74,017
Switzerland	73,923

We found a total of 165,664,116 single nucleotide variations (SNVs) that have been reported in all SARS-CoV-2 genomes analyzed, which supposes an average of 35.89 mutations per genome. Viral mutation rates vary widely, especially due to the differences in the fidelity of the polymerases used in replication[44]. Viruses that encode their genome in RNA, such as SARS-CoV-2, HIV and influenza, tend to pick up mutations quickly as they are copied inside their

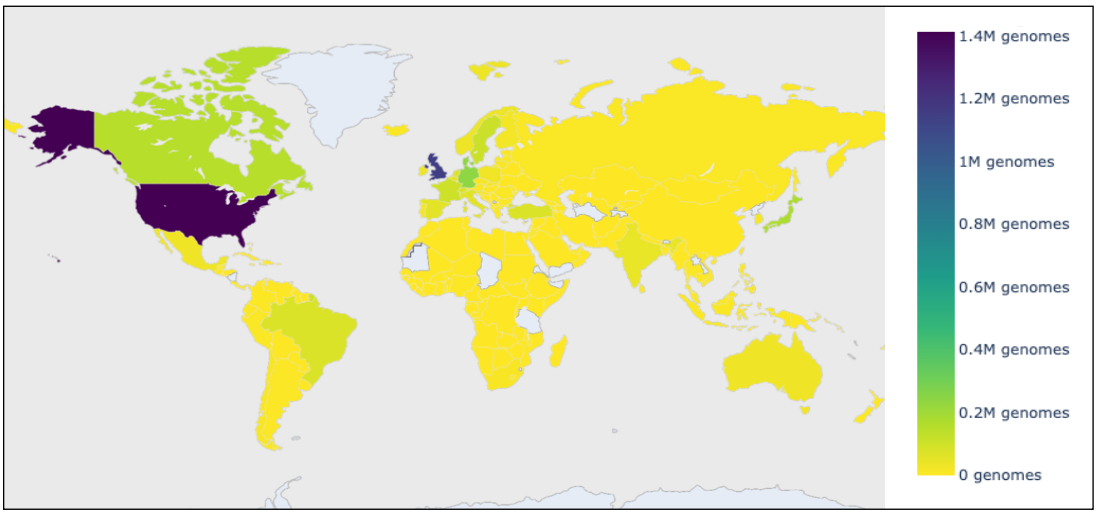

Figure 6 | Number of genomes contributed to GISAID by each country.

hosts because RNA-dependent RNA-polymerases (RdRp) are prone to making errors, but sequencing data suggest that coronaviruses change more slowly than most other RNA viruses. A typical SARS-CoV-2 virus accumulates only a rate of change about half that of influenza and one-quarter that of HIV[45]. Coronaviruses, in common with other members of the *Nidovirales* order, encode a 3′-5′ exonuclease (nsp14, ExoN) that dramatically reduces the effective error rate of the viral RNA-dependent RNA polymerase[46]. Consequently, coronaviruses demonstrate a low-rate substitution rate and minor differences in sequence diversity despite a genome size of nearly 30,000 bases.

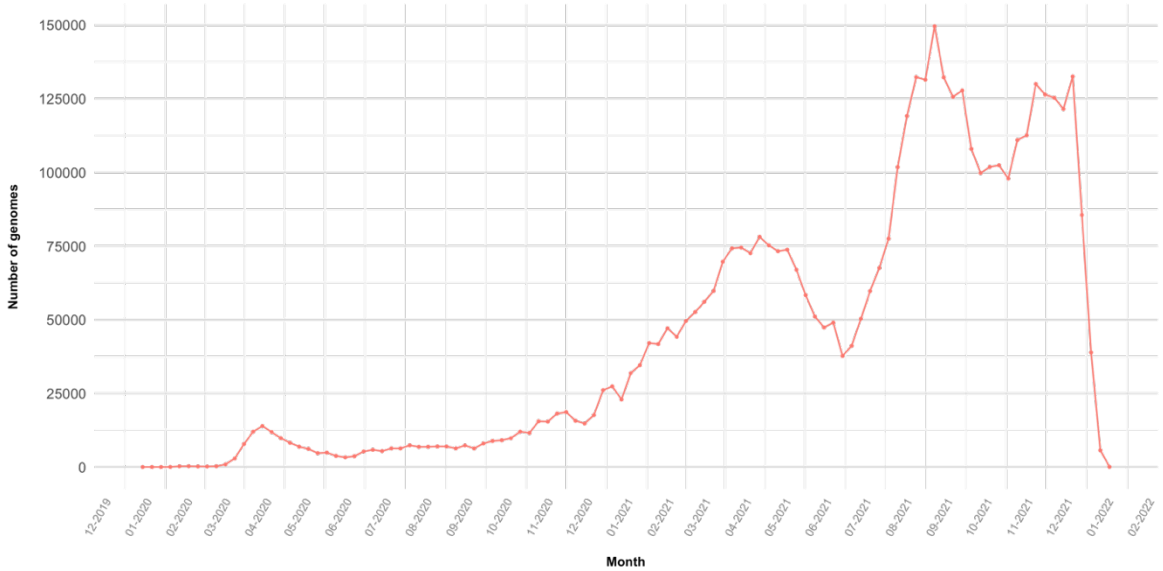

Figure 7 | Number of genomes contributed to GISAID by week. Exponential growth curve of the sequenced genomes published in GISAID database. 2 major peaks can be seen, probably due to waves of infection during early and late 2021.

Table 4 | Unique mutations characterization in percentage in SARS-CoV-2 genome.

Mutated nucleotide (→) Original nucleotide (↓)	A	G	C	T
A	-	11.97%	9.52%	9.3%
G	7.6%	-	5.24%	7.55%
C	6.47%	4.65%	-	7.62%
T	8.75%	8.72%	12.62%	-

It is worth mentioning that the vast majority of these mutations were found to be repeated, and in the large analysis of mutations only 77,371 unique mutations, 14,833 deletions and 1054 insertions were recorded. Bearing in mind that the Wuhan-Hu-1 reference genome is made up of 29,903 bases, about 12,000 unique mutations remain to be detected. It could be, however, that many of these do not code for a functional protein, and therefore will never be detected. A peak of unique mutations was found around January-February 2020 when more than 3 new mutations were detected per published genome (data not shown). From that date onwards, this number has been decreasing until reaching practically 0 new unique mutations per genome currently (data not shown).

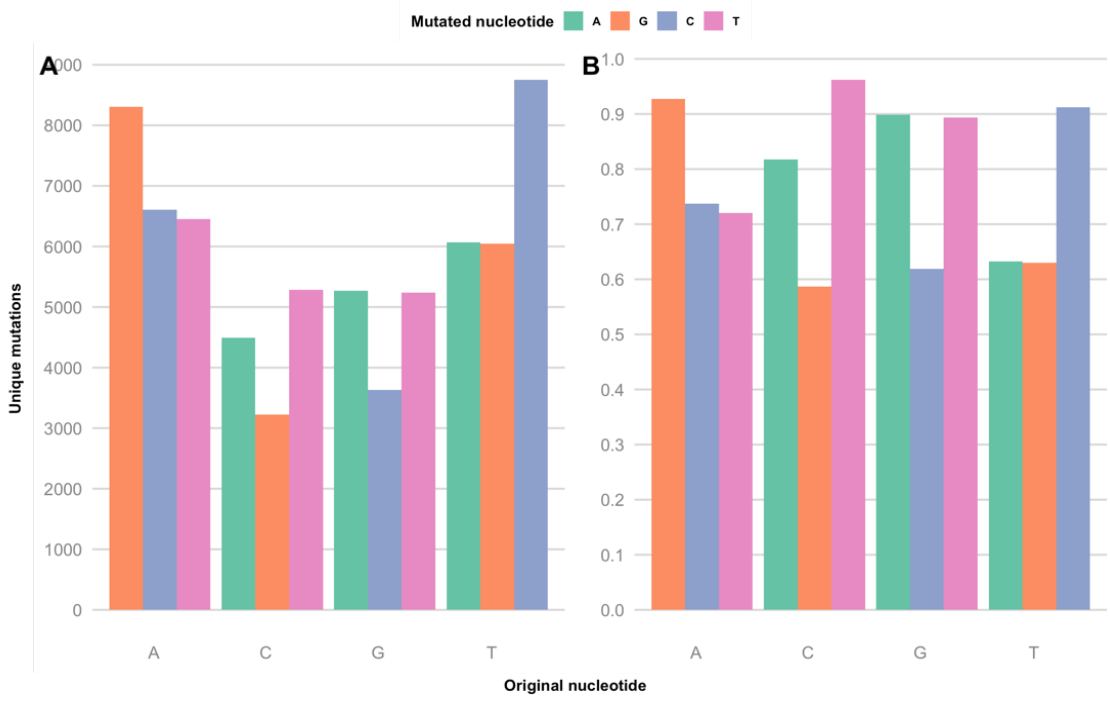

Figure 8 | Unique mutations characterization in SARS-CoV-2 genome plot. The x-axis indicates the original nucleotide and the colour of the bar indicates the base to which it mutates. *E.g.,* the T→C mutation corresponds to the blue bar on the right which has been found a total of ~9000 times (y-axis). Figure A shows the number of unique mutations in the genome. Figure B shows the normalized number of unique mutations by dividing the number of mutations by the total number of bases found in the viral genome.

Among the unique mutations, we counted the number of different transitions and transversions concerning the Wuhan-Hu-1 reference genome. Generally, the most common nucleotide change in the SARS-CoV-2 genome is C→T (as a transition mutation) (Figure 8B) although nowadays this number does not stand out so much from the others because new C→T mutations are hardly detected anymore (Table 4). After that, the G→T, which is the most common transversion, and finally, the third most common event worldwide, G→A (Figure 8B). Surprisingly, for unique mutations that occurred in the SARS-CoV-2 genome, the transition versus transversion ratio was calculated as about 2:3 being transversions more frequent (60.83%) than transitions (39.17%).

The origin of the ~77,000 unique mutations has not followed a regular curve over the two years of the pandemic, as can be seen in Figure 9 and Figure S1. During the first months of the pandemic, all analyses of SARS-CoV-2 genomic sequence data reported a preponderance of C→T transitions in the viral genome and the ratio of C→T to T→C transitions was nearly six times higher than expected (data not shown). These C→T transitions were loosely associated with the base and structure context of RNA deamination favoured by APOBEC3 proteins (host-specific RNA gene editing). To corroborate the hypothesis, multiple studies have been carried out and was seen that the extended context of the mutation, was similar to the A3A-driven editing of cellular mRNA sequences[47]. It was also proposed that the higher number of T→C could be the result of viral protective mechanisms against defective mutations[48]. Nonetheless, this hypothesis still needs to be proved. For the other types of mutations, a more linear trend has been followed as shown in Figure S1.

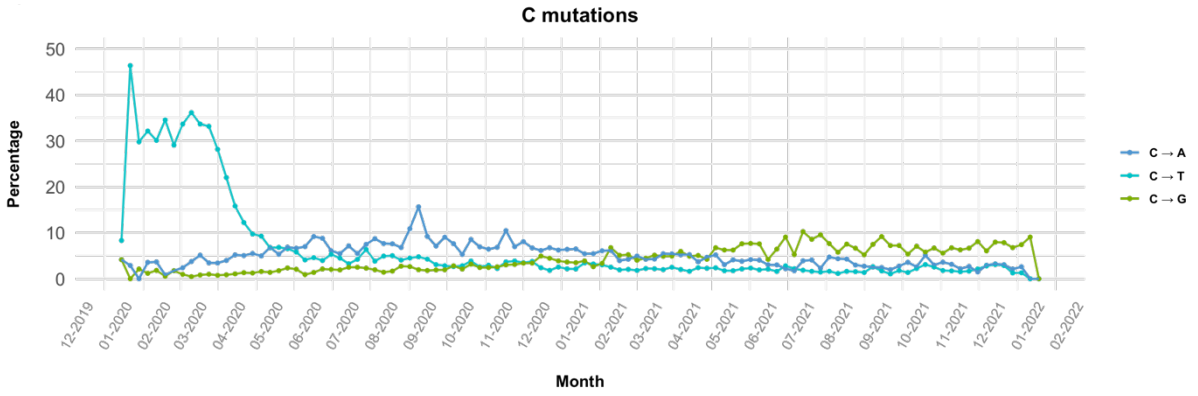

Figure 9 | Evolution of C different mutation percentage of reported unique mutations over the weeks.

The mutation rate in the different genes was also not homogeneous as shown in Figure 10, where the *ORF7a* gene stands out. The protein encoded by *ORF7a* is a non-structural protein which plays a very important role in viral replication in cell culture by modulating the host G0/G1 transition checkpoint. Holland *et al.*[49] found 81-nucleotide deletions in the SARS-CoV-2 AZ-ASU2923 genome that occurred in the ORF7a gene, resulting in a 27 amino acid deletion. Despite recent research on SARS-CoV-2 evolution, there is a lack of research linking the deletions that have occurred in the entire genome of SARS-CoV-2 worldwide. Taking into account other type of SNVs, the number of synonymous mutations per nucleotide is similar in all genes, and several studies suggest that those silent mutations may play a role in viral evolution, by increasing the adaptation of the viral genome to the human codon usage (CU). To this end, the observed continuous adaptation of its CU over time may underlie an increase in the overall efficiency of viral protein production and packaging, with viral genomes with a better adaptation being able to generate more viral particles over time, therefore outperforming other, less adapted viruses[50]. On the other hand, if we focus on the non-synonymous mutations, we find fewer missense and nonsense mutations (*i.e.,* non-synonymous) in genes that encode proteins that have an important role in the replication of the virus (*e.g.,* EndoRNAse, methyltransferase, RdRp, 3C-like protease (M-pro), and helicase). As shown in Figure 10, the significant decrease curve from the *ORF7a* gene to the helicase gene will be key for primer design and analysis.

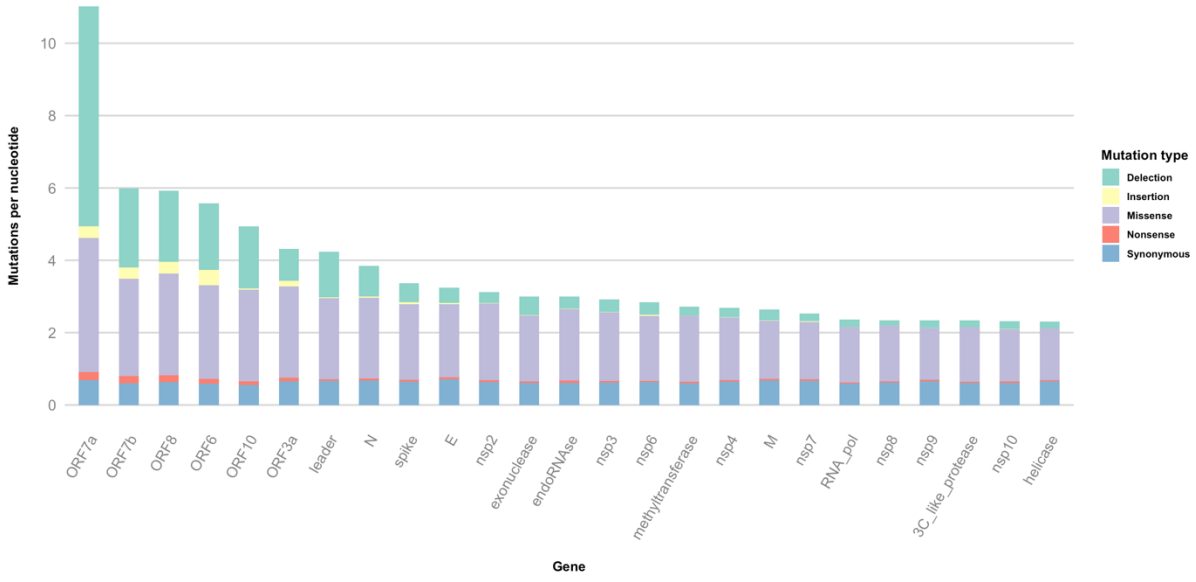

Figure 10 | Number of unique mutations per nucleotide of each gene of the SARS-CoV-2. Mutation rates are calculated by dividing the number of unique mutations of each gene by its length.

As reflected in Figure 11, the most ubiquitous modifications were C3037T (synonymous substitution) and two non-synonymous mutations C14408T (nsp12: P323L), and A23403G (S: D614G) occurring in almost a 100% of the samples. The D614G mutant became predominant across the world within a brief period and all the mutant variants declared as VOC by the WHO and CDC had the D614G mutation common —including the recent B.1.1.529 lineage (Omicron variant)—. Some studies exhibit that patients infected with that spike G614 variant show higher mortality or clinical severity, but it is associated with higher viral load, especially in younger patients[51]. It alters some of the conformations, giving some extra fitness to the virus and increasing the stability of the S trimer. It brings many changes in the S-glycoprotein's characteristic features that make the strain predominant, increasing the transmissibility of the G614 virus leading to a great number of replication events and greater genetic diversity[52]. On the other side, the C14408T mutation replaces a proline (P) with leucine (L) at position 4715 (P4715L) of ORF1ab polyprotein which appears as a replacement of proline with leucine at position 323 (P323L) of RdRp —an exposed residue on the surface of the enzyme. A critical mutation in the RdRp gene has the potential to alter viral replication capability with fidelity, and thereby a mutation in RdRp may contribute to the infectivity of the virus and severity of the disease. This mutation was found to be associated with an overall increase in mutation rate in the viral genome[53].

Furthermore, according to Figure 11, a large number of mutations are found above a 50% frequency, giving rise to different VOI/VOCs that can cause errors in the hybridization of the primers or probes with the viral genome, giving rise to false negatives.

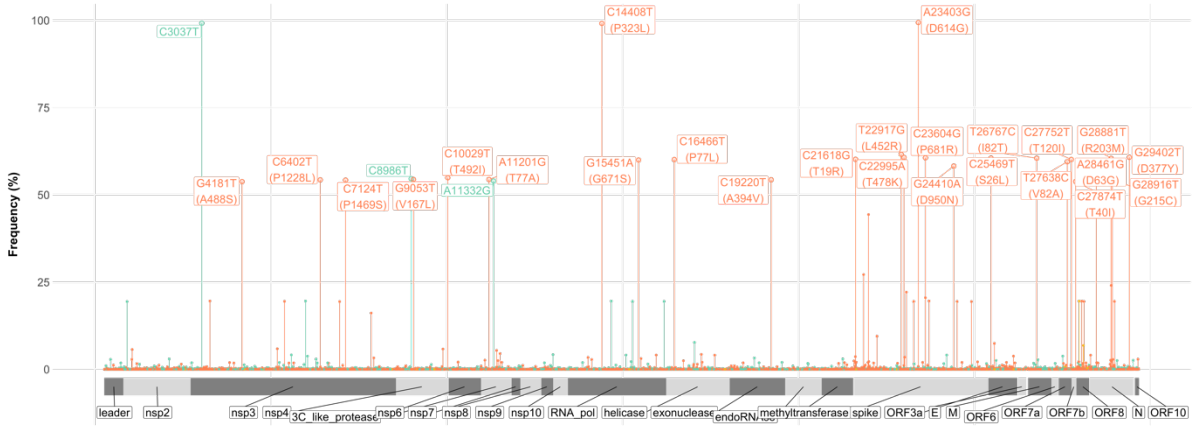

Figure 11 | Mutational profiling of SARS-CoV-2 genome. Unique mutations in the SARS-CoV-2 genome in the specific genomic position (x-axis) and the frequency in which they are found (y-axis). Note that all mutations found with a frequency above 50.0% have a label with information about the mutation and the amino acid change between brackets (if applicable). Frequency is calculated as the number of this specific mutation between the total number of genomes. Synonymous mutations are indicated in aquamarine colour, missense ones in coral, and other types in golden.

4.2. VOCs mutation analysis

Variants of SARS-CoV-2 are further being assessed to determine whether a specific variant's transmissibility, clinical presentation, severity, or impact on countermeasures (i.e., diagnostics, therapeutics, and vaccines) hinders the disease process. It is for that reason that following the evolution of different VOCs and distinct mutations circulating worldwide have been prioritized, including D614G (found in all lineages), N501Y (found in several lineages), E484K (found in several lineages), K417N (found in several lineages), and L452R (found in several lineages) among others (Table 5).

Multiple variants under monitoring, variants of concern (VOC), and variants of interest (VOI) have been identified since the pandemic began. In the present work, however, we will focus on VOCs because these new lineages of SARS-CoV-2 have attracted more attention worldwide due to their increased transmissibility, the danger of significant repercussions, and/or resistance to neutralizing antibodies. The initial SARS-CoV-2 VOC, VOC Alpha (B.1.1.7 Lineage), has been demonstrated to have a 43–90% greater reproduction rate and 75 percent higher infectiousness than earlier strains, becoming the dominant variant in the United Kingdom and endemic worldwide[51]. Mutations in the B.1.1.7 S protein were discovered to have many structural impacts and could boost viral fusion activity and infectivity greatly[51]. More SARS-CoV-2 variants have emerged including VOC Beta (B.1.351 lineage) and VOC Gamma (P.1 lineage) that were first detected in South Africa and Brazil, respectively. According to preliminary modelling, the B.1.351 variant could be 50 percent more transmissible than early SARS-CoV-2 strains[54]. P.1 is 1.7–2.4 times more transmissible than non-P.1 lineage and can circumvent 21–46% of protective immunity produced by previous infection[55]. Following then, the VOC Delta (B.1.617.2 Lineage) mutation spread across India, causing a significant rise in COVID-19 cases in various countries around the world, outcompeting pre-existing lineages such as B.1.1.7 (Alpha)[56]. Finally, on November 24, 2021, the Omicron variant (B.1.1.529

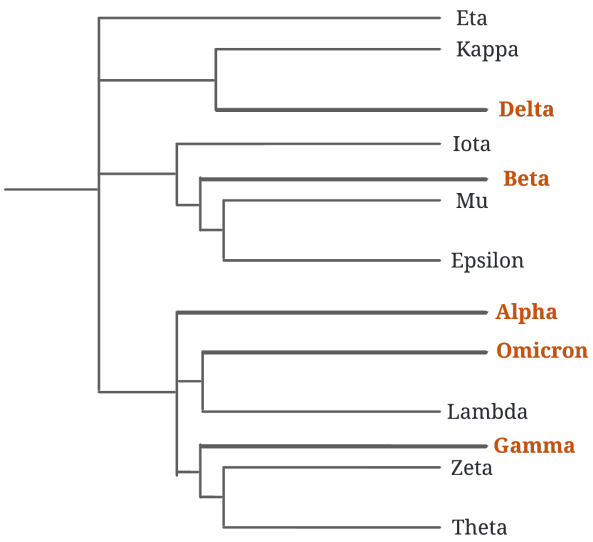

Figure 12 | Omicron early divergence in a phylogenetic mutational tree. Figure adapted from Tiecco *et al.*[57]

Lineage) was discovered for the first time in South Africa. This variant is notable for its strong transmissibility and limited sensitivity to previously tested antibodies (e.g. vaccines, monoclonal antibodies, etc.)[30].

The Omicron variant did not appear to have evolved from one of the previous variants of concern, such as Alpha or Delta. Instead, it appears to have evolved in parallel, and other phylogenetic studies show that the Omicron variant separated from other SARS-CoV-2 strains early[57] (Figure 12).

4.2.1. Molecular profile of Omicron variant

The Omicron variant, which is currently the dominant variant in circulation, has 37 non-synonymous changes in the spike protein, 11 in the NTD, and 15 in the RBD, some of which may be associated with humoral immune escape potential and higher transmissibility[58]. Pango lineages B.1.1.529, BA.1, BA.2, and BA.3 are all part of the Omicron variant. BA.1, which accounts for approximately 99% of the sequences, and BA.3 have the 69–70 deletion in the spike protein, whereas BA.2 does not.

Some mutations in the RBD of the Omicron variant are shared by other SARS-CoV-2 variants as shown in Table 5. These are K417N, E484K, N501Y, D614G, and T478K. Among these, the D614G mutation with aspartic acid substitution to glycine in the S1 subunit of the Omicron variant is the most prominent because it is also found in Alpha, Beta, Gamma, and Delta variants (Table 5).

The Omicron variant SARS-CoV-2 has a deletion of amino acids H69 and V70 in the S-gene of its genome, which can result in S gene target failure (SGTF). For this reason, the PCR (polymerase chain reaction) tests most commonly used to diagnose SARS-CoV-2 infection do not detect the S gene on which these deletions are present. SGTF can be used as a proxy marker for Omicron screening. Although one genetic target has lower sensitivity due to a mutation, assays intended to detect numerous genetic targets should still detect the Omicron variant SARS-CoV-2[59].

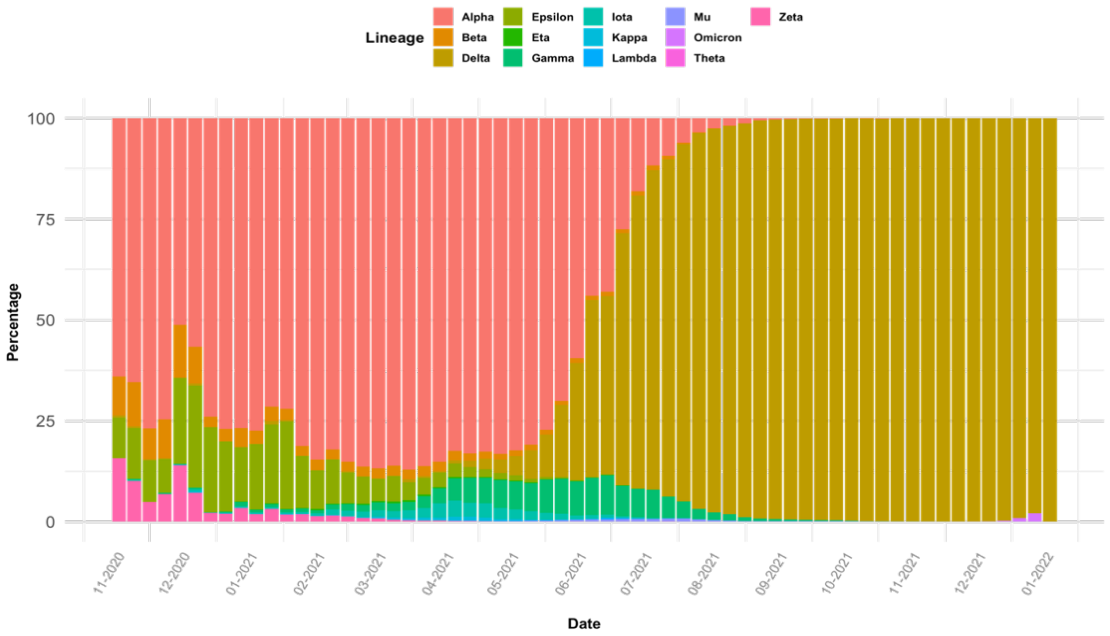

Figure 13 | Evolution of the different variants over time. Percentage calculated by dividing the weekly total of mutations of a variant by the weekly total of mutations. The starting date is November 2020, the approximate month of the characterisation of the first VOC —*i.e.* alpha.

Table 5 | SARS-CoV-2 variants of concern. Information extracted from https://covariants.org/ and https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

Variant	Spike mutations	Origin	Effects
Alpha (B.1.1.7)	Δ69/70, Δ144, N501Y, A570D, D614G, P681H, T716I, S982A, and D1118H	UK, December 2020	- Increased affinity of S protein to ACE-2 - Increased transmissibility - Sensitive to BioNTech/Pfizer, Moderna, Oxford/Astra Zeneca, and Novavax vaccines
Beta (B.1.351)	L18F, D80A, D215G, Δ242–244, R246I, K417N, E484K, N501Y, D614G, and A701V	South Africa, December 2020	- Increased transmissibility - Resistant to Bamlanivimab/Etesivimab - Sensitive to BioNTech/Pfizer and Moderna vaccines after booster shot - Oxford/Astra Zeneca vaccine less effective
Gamma (P.1)	L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, and V1176F	Brazil, January 2021	- Increased transmissibility - Resistant to Bamlanivimab - Reduced neutralization by antibody therapy and convalescent sera - Moderate protection by Coronavac and BioNTech/Pfizer vaccine and good protection by Oxford/Astra Zeneca vaccine
Delta (B.1.617.2)	T19R, FR157–158Δ, L452R, T478K, D614G, P681R, and D950N	India, December 2020	- Increased transmissibility, viral load and infection rate - Increased risk of hospitalization - Slight to moderate reduction in sensitivity to vaccines
Omicron (B.1.1.529)	A67V, HV69–70Δ, T95I, G142D, VYY143–145Δ, N211Δ, L212I, R214EPEins, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, and L981F	South Africa, November 24, 2021	- Increased transmissibility - Decreased sensitivity to all tested antibodies (*e.g.,* vaccines, re-convalescent sera, monoclonal antibodies)

Therefore, a detection pattern that reveals a dropout of the impacted target could be an indicator of the presence of the Omicron variant in a patient who has a positive result. The Alpha variant, as shown in Table 5, has this deletion in its sequence as well; despite this, the Alpha variant was been totally replaced by Delta at the population level, as illustrated in Figure 13.

The mutational profile of this VOC has made it the majority variant in a very short period of time. Even though, it has not been possible to collect this information from the GISAID database because the last date collected was 6th January (Figure 4) (see Materials and Methods section for further details).

4.3. Primers analysis

Clinical laboratories and firms scrambled to develop new methods and reagents for reliable identification of the novel SARS-CoV-2 when the WHO declared the disease a pandemic. As qRT-PCR assay is the gold standard method for diagnosing viruses spread by air, when the first whole-genome sequences of SARS-CoV-2 were published, WHO recommended primers targeting N, E, and ORF1ab. Furthermore, due to the similarity of the SARS-CoV-2 viral genome to that of SARS-CoV, several research laboratories were able to begin designing novel primers for PCR detection at an early stage. Many manufacturers have followed these first recommendations since there was no time to enhance primer design and due to the rise of new SARS-CoV-2 variants emerged. Despite this, our *in silico* analysis of over 4,600,000 genomes submitted to GISAID revealed numerous mismatches in primer binding regions in commonly used diagnostic primers.

A total of 42 primers were collected from different sources and saved in a file with the different data: nucleotides, probe utilized, positions and amplicon length (see Materials and Methods section for further details). Many of the primers used in laboratories for performing PCR tests have not yet been made public. Throughout these new sections, it will be assumed that many of these early-designed primers have been used throughout the pandemic all over the world. Additionally, primers published by the WHO at the beginning of the pandemic will also be used. Thus, a total of 15 primers were selected to be used for mutational analysis (Table S1).

These primers were subjected to a mutational study. First of all, for each batch of primers (*i.e.* forward primer, reverse primer and probe), all the unique mutations were enumerated and sorted

by week (Figure 14). As can be seen in Figure 14, the primers underwent a high mutation rate around March-April 2020. From then on, the mutation rate in the primers has been increasing and has not stopped so far and only some of them have reached a *plateau*, where it seems that the number of mutations is no longer growing or growing more slowly —*e.g.,* Thailand-N or Sigma-Aldrich. The fact that mutations in the other primers continue to grow may mean that, although some of these primers could still be functional after the two years of the pandemic, they will eventually cease to be functional due to low complementarity with the genomes of the new variants.

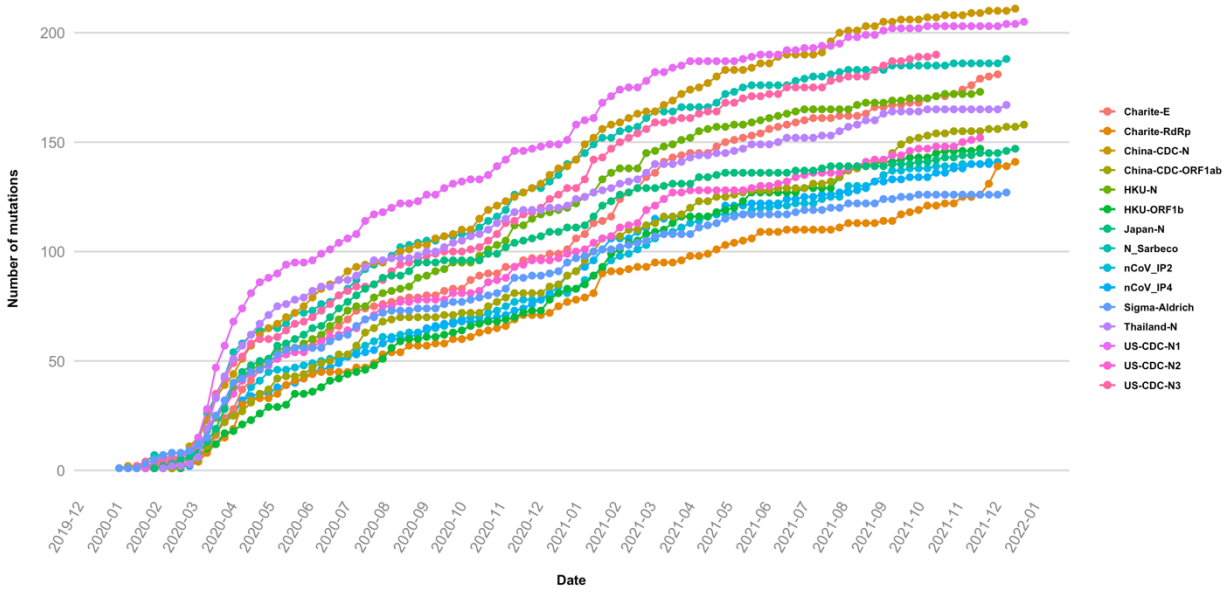

Figure 14 | Evolution of the different primers' mutations over time.

All mutations —including deletions and insertions— were retrieved for the 15 selected primers, and a table was constructed using the synthesized information collected (Table 6). Table 6 demonstrates that several primers have a high incidence of deletions, with some having over 50 —all of them located in the nucleocapsid gene— (*i.e.* HKU-N, US-CDC-N2, Japan-N). These deletions are found more frequently than insertions and were mostly isolated cases because these were not found in a high percentage (data not shown). Even though the number of deletions and insertions is still lower than the number of base substitutions, they may have a greater impact because they involve base deletion or insertion and a probable alteration in the ORF causing a framing error. Furthermore, in most cases, in deletions, more genomic bases are influenced than in a SNV.

Looking at the *total n_found* column (constructed by adding up all the mutation frequencies contained in the different primers), there is a large bias of accumulated frequencies in the different primers. It can be found at frequencies ranging from lower than 1% to frequencies higher than 100%. These primers with a frequency greater than 100% contain mutations that, in addition to being detected on their own, can also be found in combination with other mutations. This means that the frequency of mutations in that location is significantly higher than 100%.

Moreover, as seen in materials and methods, the last 5 positions of the 3' end of the primer are key to the specificity of the primer, and mutations in these positions would decrease its specificity. To ensure that the primer's hybridization is proper and that amplification is produced, the last 5 bases of the extreme 3' must perfectly coincide with the hybridization site[41,42]. It is for this reason that in Table 6, they have been counted in a separate column (columns ending in ...5 last count/n_found). It is shown that many primers accumulate a substantial number of mutations at these sites, with some acquiring more than 50% of the total number of mutations in the primer, as in the case of US-CDC-N2 or HKU-N, with up to 72% of forward primer mutations and reverse primer mutations respectively accumulated in the last 5 positions. These primers are the most likely to fail, and amplification rates will be lower.

As shown in Table 6, most of the primers hybridize to the N gene. Several mutations likely to cause detection failure with the assays used were identified when we aligned the N gene sequence with the different PCR-primer binding sites. The SARS-CoV-2 N gene was found to be one of the most variable regions (Figure 10). As a result, nucleotide variations on the primer binding sites of viral RNA sequences can affect test results. Different studies suggest that the virus's infectivity is increased by a high mutation rate in the nucleocapsid gene[60]. The SARS-CoV-2 N protein, which is abundant in infected cells, performs multiple functions during viral infection, including RNA binding, oligomerization, and genome packaging, as well as transcription, replication, and translation[61]. To increase virus survival, the N protein evades immune response and disrupts other host cellular processes including translation, cell cycle, TGFβ signalling, and apoptosis induction[62]. As a result, searching for primers in this gene is not recommended.

Finally, to see the distribution of the percentage of the mutational profile of the different primers, Figure 15 was constructed. As can be observed, most of the mutations are identified at a frequency of <0.02, although there are some significantly higher outliers, such as Charite-RdRp

and China-CDC-N, which include mutations with a percentage of ~60%. Multiple mutations in the N gene's primer binding sites have an impact on assay performance since it relies heavily on sequence matching between primer-probe and SARS-CoV-2 sequences, which can lead to inconsistency or a high false negative rate (FNR) in test findings[63].

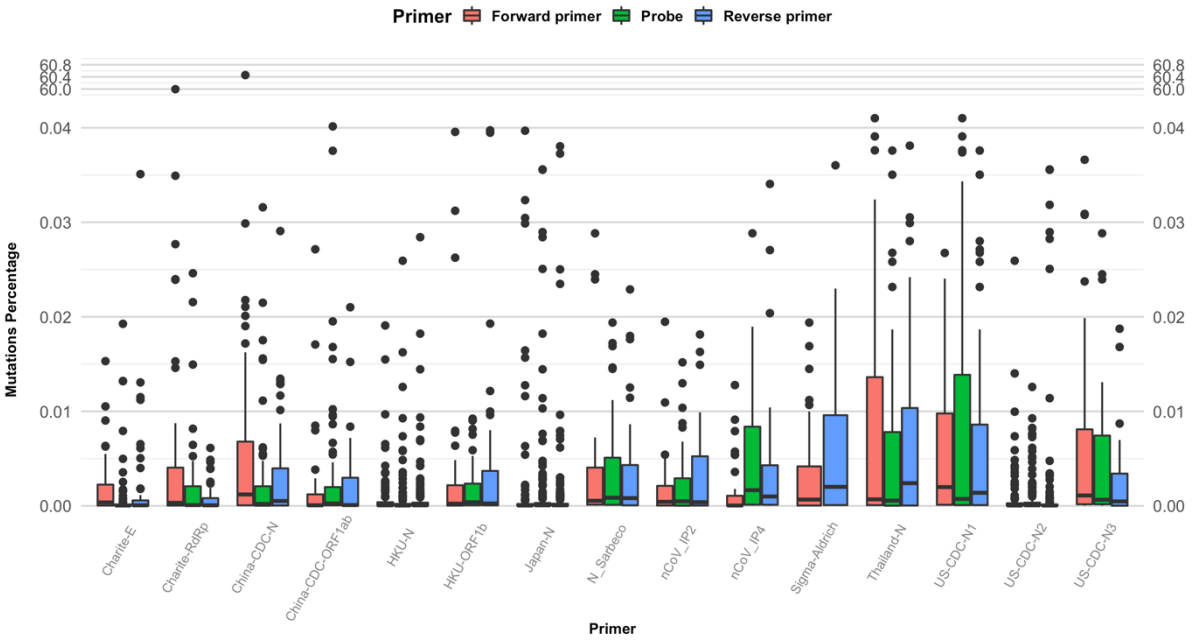

Figure 15 | Mutations of the different primers. For each group of primers a boxplot is shown for the forward primer, one for the reverse primer and one for the probe, with the exception of the primer provided by Sigma-Aldrich, which does not use a probe.

4.4. Omicron primer design

In order to obtain specific primers with a high amplification rate for the Omicron sequences, it is required to find an area of low homology with the reference sequence and the rest of the variants. In addition, it will have to be found in a gene with a lower mutation rate than the N gene. Following these criteria, the S gene region was chosen. The S region is a gene that codes for the S protein that is integrated on top of the surface of the virus. Despite being one of the most mutational sites in comparison to other genes, its mutation rate is still lower than that of the N gene (Figure 10).

Following the procedure outlined above, multiple multialignments of sequences from different VOCs randomly obtained from the GISAID database were made. Based on VOCs mutational analysis, different sequences containing the consensus mutations for each variant were selected, whereas those containing multiple non-characteristic mutations, deletions or undesired

Table 6 | Total number and mutation frequencies of SARS-CoV-2 hybridization sites of real-time RT-PCR primer. The n_found column is the total accumulated frequencies in the primer (or in all, in the case of total n_found). Columns ending with «count» contain two values in brackets separated by an «|». Within these, the first indicates the number of deletions, and the second the number of insertions found. Note that mutations and their frequency have been counted for the last 5 bases of the 3' end in columns ending in ...5 last count/n_found.

Name	Gene	Forward primer position	Forward primer count	Forward primer n_found	Forward primer 5 last count	Forward primer 5 last n_found	Reverse primer position	Reverse primer count	Reverse primer n_found	Reverse primer 5 last count	Reverse primer 5 last n_found	Count probe	N_found probe	Total count	Total n_found
nCoV_IP2	RdRp	12690-12707	41 (3|0)	0,0692	12 (0|0)	0,0374	12780-12797	58 (8|1)	1,2684	13 (2|0)	0,1365	56 (9|0)	0,4619	155	1,7995
nCoV_IP4	RdRp	14080-14098	43 (7|0)	0,1163	13 (3|0)	0,0676	14167-14186	52 (4|0)	0,8441	11 (1|0)	0,0402	53 (2|0)	3,2814	148	4,2419
Charite-E	E	26269-26294	69 (10|0)	0,2165	13 (4|0)	0,0156	26360-26381	59 (14|0)	0,1048	22 (12|0)	0,0026	96 (29|0)	0,1301	224	0,4516
N_Sar-beco	N	28706-28724	52 (5|0)	0,8666	17 (3|0)	0,2327	28814-28833	76 (16|0)	0,8009	27 (12|0)	0,1687	90 (19|0)	0,5115	218	2,1790
Charite-RdRp	RdRp	15431-15452	55 (6|0)	60,388	16 (2|0)	60,103	15505-15528	47 (7|0)	3,1106	14 (5|0)	0,0059	49 (5|0)	0,2408	151	63,739
HKU-ORF1b	ORF1ab	18778-18797	49 (2|0)	0,3101	11 (0|0)	0,0753	18889-18909	57 (7|0)	0,7154	14 (0|0)	0,1465	44 (1|0)	0,1684	105	1,1940
HKU-N	N	29145-29166	131 (72|0)	0,5462	86 (72|0)	0,0811	29236-29254	204 (148|0)	2,0509	147 (133|0)	0,1528	152 (100|0)	0,7374	369	3,3346
China-CDC-ORF1ab	ORF1ab	13342-13362	48 (6|1)	0,2742	10 (2|0)	0,0091	13442-13460	44 (8|0)	0,2344	14 (4|0)	0,0373	82 (18|0)	0,2835	174	0,7922
China-CDC-N	N	28881-28902	103 (31|2)	88,472	37 (20|1)	0,2717	28958-28979	87 (19|2)	22,971	24 (10|0)	0,5491	67 (13|1)	0,3552	257	111,79
US-CDC-N1	N	28287-28306	71(9|0)	5,1656	18 (3|0)	0,0607	28335-28358	79 (11|0)	0,5491	16 (3|0)	0,1101	81 (13|0)	2,1685	231	7,8833
US-CDC-N2	N	29164-29183	137 (87|0)	1,1181	99 (87|0)	0,5144	29213-29230	166 (128|0)	0,6675	119 (114|0)	0,1254	169 (112|0)	0,9886	472	2,7743
US-CDC-N3	N	28681-28702	65 (9|0)	0,9873	14 (5|0)	0,1083	28732-28752	74 (11|1)	0,7726	19 (4|0)	0,1662	68 (6|0)	1,6407	207	3,4007
Japan-N	N	29125-29144	99 (58|0)	0,4866	71 (57|0)	0,3252	29263-29282	214 (169|0)	0,9630	165 (156|0)	0,2572	192 (136|0)	0,5963	505	2,0461
Thailand-N	N	28320-28339	64 (8|0)	1,3329	16 (3|0)	0,0920	28358-28376	73 (13|1)	0,6944	21 (7|0)	0,1728	55 (9|0)	0,3761	192	2,4035
Sigma-Al-drich	N	28750-28771	77 (16|0)	0,3735	24 (10|0)	0,0915	28842-28860	73 (16|1)	2,3848	21 (7|0)	0,1550	0	0	150	2,7584

insertions were excluded. These "consensus" sequences for each VOC were multi aligned with each other in order to test the different sites (Figure 16).

For the primers to be variant-specific, it was decided that they should at least contain several B.1.1.529 lineage characteristic mutations in the last 5 positions of the primer. The spike gene of Omicron is characterized by a significant number of mutations (at least 30 amino acid changes), three small deletions, and one small insertion that is unique to Omicron and allows for focused detection of this variant. Following these criteria, the variant-specific sites identified were: (I) a 3-amino acid deletion zone (VYY143–145Δ), (II) a 3-amino acid insertion (N211Δ–L212I–R214EPEins) and (III) 3 close mutations of serine residues (S371L–S373P–S375F) (Figure 16). Because these are areas that accumulate characteristic mutations of the Omicron variant, these positions were identified as promising hybridization sites for RT-qPCR specific detection.

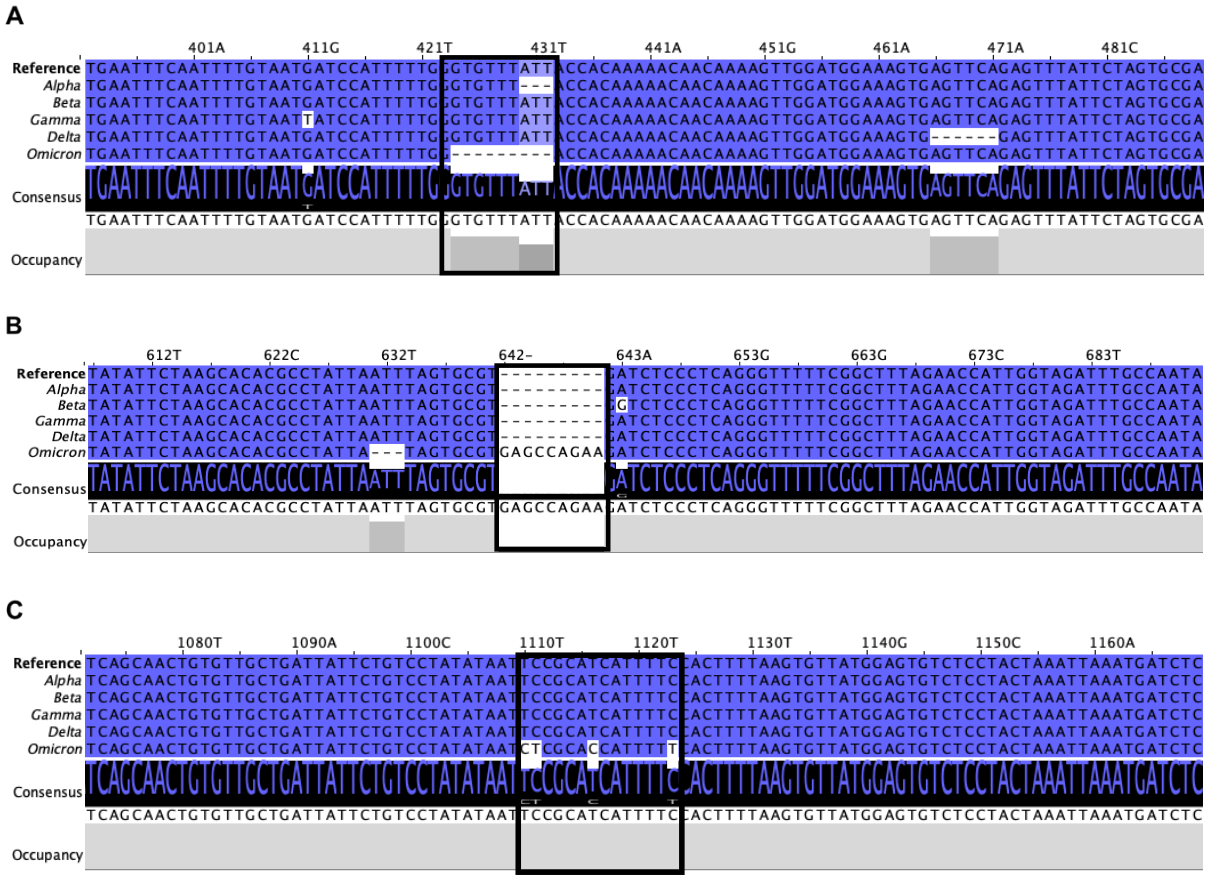

Figure 16 | Alignment of different fragments of the spike gene sequence for different VOCs. The hybridization site alignment of primers for the different consensus mutations of the Omicron variant is shown and the mutated amino acids are marked with a dark rectangle. Figure A shows the VYYY143–145Δ zone. Figure B shows the N211Δ–L212I–R214EPEins zone. Figure C shows the S371L–S373P–S375F zone.

We looked for suitable primers that might hybridize in the chosen location and contained as many of the consensus mutations as possible in their last 5 bases. The methodology described in Materials and methods was followed to accomplish this. For each case, three categories of cases were sought for each mutation: mutations identified in the last 5 nucleotides of the primer forward, reverse, or probe.

4.4.1. VYYY143–145Δ primers

In the case of the first deletion —*i.e.* VYYY143–145Δ, the area where the mutation occurs contains a low percentage of GC, which makes it impossible to design good primers and probes in that area.

4.4.2. N211Δ–L212I–R214EPEins primers

Forward primer containing the insertion

In the case of the area including N211Δ–L212I–R214EPEins, many possible options were found for primer forward design containing the mutation. In the case of the reverse primer and probe, the options were already more limited.

A forward primer containing the R214EPEins in the last 9 bases of that was selected. The reverse primer was one of the only primers that matched the melting temperature, %GC, and length parameters defined above. Finally, the probe could be selected close to the forward primer. Using the primers shown in Table 7, the amplicon length would be 91 bases (Figure 17). All of the parameters listed in Table 2 of materials and procedures are met by this primer set, and it has been checked that they do not hybridize in other non-desired areas of the genome.

Table 7 | Set of designed primers. Note that the positions are shown relative to the reference genome. In the case of an insertion or deletion, the length determined by subtracting the reverse primer's ending position from the forward primer's initial position may differ from the length indicated. The length in the table represents the amplicon's real length.

Target	Type	Position	Length	Sequence (5' → 3')
Spike	F	22,189-22,210		TATTATAGTGCGT**GAGCCAGAA**
	R	22,249-22,273	91	CCTAGTGATGTTAATACCTATTGGC
	P	22,208-22,231		CTCCCTCAGGGTTTTTCGGCTTT

Reverse primer containing the insertion

In this case, the possible reverse primers designed contain the insertion at the 5' end, which does not contribute to optimal primer specificity for the omicron variant. Despite this, they also

contain the deletion (N211Δ) and the mutation (L212I) near to the 3' end, which is why they could be good candidates for identifying the variant sequences. Several forward primers fulfilling the parameters described above were found, but not probes close to either of the two primers.

Unless several primer pairs met the described parameters (data not shown), one of the candidates is shown in Table 8. As can be seen, the length of the amplicon is slightly longer than desired; nevertheless, it would be a fully functional primer pair. On the other side, it has been checked that they do not hybridize in other non-desired areas of the genome.

Table 8 | Set of designed primers. Note that the positions are shown relative to the reference genome. In the case of an insertion or deletion, the length determined by subtracting the reverse primer's ending position from the forward primer's initial position may differ from the length indicated. The length in the table represents the amplicon's real length.

Target	Type	Position	Length	Sequence (5' → 3')
Spike	F	22,092-22,111	131	TGGACCTTGAAGGAAAACAG
	R	22,190-22,211		C**TTCTGGCTC**ACGCACTATAAT

Probe containing the insertion

In this case, we have several possible options. In all cases, they will be probes that start a few bases after the forward primer ends. The insertion site is very good for probe design, just as it has been for forward primer design. Regarding the reverse primer, we have only one option that can be adjusted. The amplicon will be slightly longer than 100 bases (Table 9).

It has not been possible to find a probe close to the reverse primer because the length of the amplicon could not be maintained according to the established criteria (>80 and <120 nucleotides). In all cases, the probes will be close to the forward primer (Figure 17). This primer set meets all the parameters described in Table 2 of materials and methods and it has been checked that they do not hybridize in other non-desired areas of the genome.

Table 9 | Set of designed primers. Note that the positions are shown relative to the reference genome. In the case of an insertion or deletion, the length determined by subtracting the reverse primer's ending position from the forward primer's initial position may differ from the length indicated. The length in the table represents the amplicon's real length.

Target	Type	Position	Length	Sequence (5' → 3')
Spike	F	22,173-22,192		ATTCTAAGCACACGCCTATT
	R	22,249-22,273	118	CCTAGTGATGTTAATACCTATTGGC
	P	22,194-22,214		TAGTGCGT**GAGCCAGAA**GATC

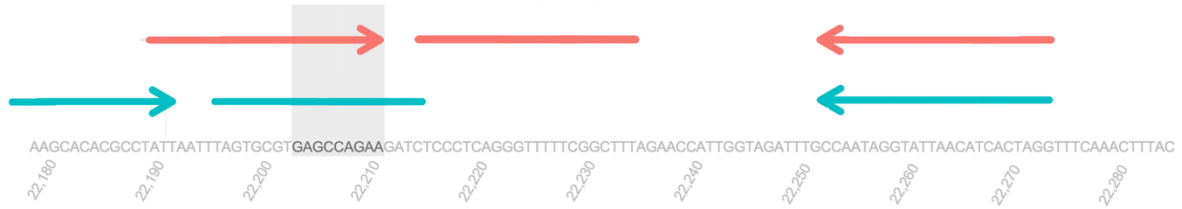

Figure 17 | Hybridisation site of the primers shown in Table 7 and Table 9. The area of the EPE214 insertion is marked in grey. In coral colour is shown the designed primer set where the insertion is contained in the forward primer. The turquoise colour shows the designed primer set where the insertion is contained in the probe.

4.4.3. S371L–S373P–S375F primers

Forward primer containing the mutation

Although the forward primers only contain one specific mutation for the Omicron variant (S373P) at the 5' end, the other two are very close and will therefore also give specificity to the primer (Table 10).

Reverse primers could also be found but, although the amplicon will be slightly longer than 120 bases, they could be used. However, no probes were found that could hybridize due to the low GC content in the area. It has been checked that they do not hybridize in other non-desired areas of the genome.

Table 10 | Set of designed primers. Note that the positions are shown relative to the reference genome. In the case of an insertion or deletion, the length determined by subtracting the reverse primer's ending position from the forward primer's initial position may differ from the length indicated. The length in the table represents the amplicon's real length.

Target	Type	Position	Length	Sequence (5' → 3')
Spike	F	22,657-22,680	142	TTCTGTCCTATATAAT**CT**CGCA**C**C
	R	22,766-22,787		GTCTGACTTCATCACCTCTAAT

Reverse primer containing the mutation

In this case, all reverse primers found that fit the established parameters contain the mutations in the 5' end region. As the mutations are SNVs, they do not have high specificity for the variant. No forward primers or probes are sought.

Probe containing the mutation

A probe matching the parameters and containing the mutations has not been found. No forward or reverse primers have been searched for.

5. CONCLUSION AND FUTURE PERSPECTIVES

The current study gives a complete perspective of the SARS-CoV-2 mutational profile as well as an Omicron-specific primer design.

A mutational profile of the entire genome, by analysing more than 4,000,000 genomes from GISAID, has been carried out. Because detrimental mutations are not evolutionarily conserved, but mutations that improve viral fitness are, mutational profiling allows for the discovery of critical aspects of a genome. Moreover, monitoring changes in SARS-CoV-2 is important because new variants can have a major impact on the severity of Covid-19 disease, the contagion, and the virus's stability.

Furthermore, a mutational analysis of the various primers currently utilized for the detection of SARS-CoV-2 in samples was performed. The different primers examined yielded a total of 3558 single nucleotide mutations, 1380 of which were deletions and 10 of which were insertions. Moreover, it has been discovered that many of them, particularly those identified in the N gene (one of the most mutated), accrue a high rate of mutations —including insertions and deletions— in the last 5 positions of the 3' end (which a correct complementarity in this area is critical for correct amplification). Assessing the mutational profile of the primers used is a critical step in reducing false negatives and ensuring disease monitoring over the world. Because current primers collect a high number of mutations, they are not the ideal option, and the primers must be updated as soon as possible in order to ensure a proper diagnosis of Covid-19. This study provides alternatives to them.

Several substitutions, insertions and deletions characteristic of the Omicron lineage spike-in gene have been analyzed to design several Omicron-specific primers and probes. The best primer pair has been considered to be the one that includes the R214EPEins in the forward primer. This primer pair and probe hybridise to the spike protein region, a region with a lower mutation rate than the N gene —one of the most widely used currently. With an amplicon length of 91 bases, it satisfies all the design criteria and could be a good candidate for use in multiplex PCR and to replace the ones that do not hybridize with the virus's current genomes. On the other side, these new qualitative and variant-specific RT-PCR primers to identify Omicron in respiratory specimens would be excellent candidates to determine which variant the patient is infected with, without resorting to Next Generation Sequencing (NGS).

Furthermore, this analysis differs from previous analysis in that it took into account all deletions and insertions during the analysis. Even though the bulk of them are only found in a few of the genomes retrieved, some of them must be taken into account when evaluating primers and when making the *de novo* design.

In conclusion, the current study lays the door for future SARS-CoV-2 genomic analysis. Using this data, we will be able to continue evaluating and characterizing the virus's mutations in order to find an effective and up-to-date cure and/or vaccine. Furthermore, after two years of the pandemic, and owing to all the data gathered throughout this project, the possibility of rethinking the primers utilized has arisen. Making them variant-specific, it would also allow for increased variant tracking because a simple PCR would identify Omicron variant in respiratory specimens. Finally, a laboratory test could be performed to verify that the variant-specific primers and probes are working properly.

6. ACKNOWLEDGEMENTS:

I would like to first thank my tutor Dr. Santi Garcia-Vallvé. Also, my thanks and appreciations go to my other supervisor, Dr. Gerard Pujadas. Before I joined their group, bioinformatics was a totally unknown world to me. With them, I was able to learn a lot about different bioinformatics approaches, and I was able to continue my education while working on my degree. I would want to express my gratitude for their constant willingness to assist me and the trust they have placed in me. I would also like to thank all of the members of the Cheminformatics and Nutrition research group; it's been a pleasure to be a part of this group.

Finally, I would want to express my gratitude to my family, especially my parents, who have taught me everything you can't learn at school or university and for their unwavering support throughout my career. On the other hand, I would like to express my gratitude to my girlfriend for her unwavering support throughout this period. She has always been there for me, listening and advising me in both happy and bad times. And last, but not least, to my colleagues and friends; those whom I have known for years and those whom I have met at university, they've been there for me the whole time, supporting me and putting up with me.

7. REFERENCE LIST

1. Coronavirus Disease (COVID-19) Situation Reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.
2. Abdelghany, T. M. *et al.* SARS-CoV-2, the other face to SARS-CoV and MERS-CoV: Future predictions. *Biomed. J.* **44**, 86–93 (2021).
3. Verity, R. *et al.* Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect. Dis.* **20**, 669–677 (2020).
4. Paim, F. C. *et al.* Epidemiology of Deltacoronaviruses (δ-CoV) and Gammacoronaviruses (γ-CoV) in Wild Birds in the United States. *Viruses* **11**, (2019).
5. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet (London, England)* **395**, 565–574 (2020).
6. Guo, Y. R. *et al.* The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak- A n update on the status. *Mil. Med. Res.* **7**, 1–10 (2020).
7. Omar, S. I. *et al.* Modeling the structure of the frameshift-stimulatory pseudoknot in SARS-CoV-2 reveals multiple possible conformers. *PLOS Comput. Biol.* **17**, e1008603 (2021).
8. Alharbi, S. N. & Alrefaei, A. F. Comparison of the SARS-CoV-2 (2019-nCoV) M protein with its counterparts of SARS-CoV and MERS-CoV species. *J. King Saud Univ. Sci.* **33**, 101335 (2021).
9. Prates, E. T. *et al.* Potential Pathogenicity Determinants Identified from Structural Proteomics of SARS-CoV and SARS-CoV-2. *Mol. Biol. Evol.* **38**, 702–715 (2021).
10. Kumar, S., Nyodu, R., Maurya, V. K. & Saxena, S. K. Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). *Coronavirus Dis. 2019* 23 (2020) doi:10.1007/978-981-15-4814-7_3.
11. Yang, J. *et al.* Molecular interaction and inhibition of SARS-CoV-2 binding to the ACE2 receptor. *Nat. Commun. 2020 111* **11**, 1–10 (2020).
12. Safiabadi Tali, S. H. *et al.* Tools and Techniques for Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)/COVID-19 Detection. *Clin. Microbiol. Rev.* **34**, (2021).
13. Chu, D. K. W. *et al.* Molecular Diagnosis of a Novel Coronavirus (2019-nCoV) Causing an Outbreak of Pneumonia. *Clin. Chem.* **66**, 549–555 (2020).
14. Corman, V. M. *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Eurosurveillance* **25**, 2000045 (2020).
15. What tests could potentially be used for the screening, diagnosis and monitoring of COVID-19 and what are their advantages and disadvantages? - The Centre for Evidence-Based Medicine. https://www.cebm.net/covid-19/what-tests-could-potentially-be-used-for-the-screening-diagnosis-and-monitoring-of-covid-19-and-what-are-their-advantages-and-disadvantages/.
16. Nolan, T., Hands, R. E. & Bustin, S. A. Quantification of mRNA using real-time RT-PCR. *Nat. Protoc. 2006 13* **1**, 1559–1582 (2006).
17. Tombuloglu, H. *et al.* Multiplex real-time RT-PCR method for the diagnosis of SARS-CoV-2 by targeting viral N, RdRP and human RP genes. *Sci. Reports 2022 121* **12**, 1–10 (2022).
18. Phan, T. Genetic diversity and evolution of SARS-CoV-2. *Infect. Genet. Evol.* **81**, (2020).
19. Shen, Z. *et al.* Genomic diversity of SARS-CoV-2 in Coronavirus Disease 2019 patients. *Clin. Infect. Dis. An Off. Publ. Infect. Dis. Soc. Am.* **71**, 713–720 (2020).
20. Tahamtan, A. & Ardebili, A. Real-time RT-PCR in COVID-19 detection: issues

affecting the results. *Expert Rev. Mol. Diagn.* **20**, 453–454 (2020).
21. Bahadır, E. B. & Sezgintürk, M. K. Lateral flow assays: Principles, designs and labels. *TrAC Trends Anal. Chem.* **82**, 286–306 (2016).
22. Mak, G. C. K. *et al.* Evaluation of rapid antigen detection kit from the WHO Emergency Use List for detecting SARS-CoV-2. *J. Clin. Virol.* **134**, (2021).
23. Porte, L. *et al.* Evaluation of a novel antigen-based rapid detection test for the diagnosis of SARS-CoV-2 in respiratory samples. *Int. J. Infect. Dis.* **99**, 328 (2020).
24. Sallard, E., Halloy, J., Casane, D., Decroly, E. & van Helden, J. Tracing the origins of SARS-COV-2 in coronavirus phylogenies: a review. *Environ. Chem. Lett.* **19**, 769–785 (2021).
25. SARS-CoV-2 Variant Classifications and Definitions. https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html.
26. Villoutreix, B. O., Calvez, V., Marcelin, A. G. & Khatib, A. M. In Silico Investigation of the New UK (B.1.1.7) and South African (501Y.V2) SARS-CoV-2 Variants with a Focus at the ACE2-Spike RBD Interface. *Int. J. Mol. Sci.* **22**, 1–13 (2021).
27. Luan, B., Wang, H. & Huynh, T. Enhanced binding of the N501Y-mutated SARS-CoV-2 spike protein to the human ACE2 receptor: insights from molecular dynamics simulations. *FEBS Lett.* **595**, 1454–1461 (2021).
28. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet. 2021 2212* **22**, 757–773 (2021).
29. Del Rio, C., Malani, P. N. & Omer, S. B. Confronting the Delta Variant of SARS-CoV-2, Summer 2021. *JAMA* **326**, 1001–1002 (2021).
30. Tian, D., Sun, Y., Xu, H. & Ye, Q. The emergence and epidemic characteristics of the highly mutated SARS-CoV-2 Omicron variant. *J. Med. Virol.* **94**, 2376–2383 (2022).
31. Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants.
32. Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly.* **3**, 1049 (2021).
33. Bioconductor - openPrimeR. https://bioconductor.org/packages/release/bioc/html/openPrimeR.html.
34. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
35. OligoArrayAux. http://www.unafold.org/Dinamelt/software/oligoarrayaux.php.
36. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, (2011).
37. Le Novère, N. MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics* **17**, 1226–1227 (2001).
38. Pandoc - About pandoc. https://pandoc.org/#.
39. Multiple Primer Analyzer | Thermo Fisher Scientific - ES. https://www.thermofisher.com/es/es/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html.
40. Chuang, L. Y., Cheng, Y. H. & Yang, C. H. Specific primer design for the polymerase chain reaction. *Biotechnol. Lett.* **35**, 1541–1549 (2013).
41. Cha, R. S., Zarbl, H., Keohavong, P. & Thilly, W. G. Mismatch amplification mutation assay (MAMA): application to the c-H-ras gene. *PCR Methods Appl.* **2**, 14–20 (1992).
42. Bustin, S. & Huggett, J. qPCR primer design revisited. *Biomol. Detect. Quantif.* **14**, 19–28 (2017).
43. Bru, D., Martin-Laurent, F. & Philippot, L. Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. *Appl. Environ. Microbiol.* **74**, 1660–1663 (2008).

44. Duffy, S., Shackelton, L. A. & Holmes, E. C. Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**, 267–276 (2008).
45. Callaway, E. The coronavirus is mutating - does it matter? *Nature* **585**, 174–177 (2020).
46. Eckerle, L. D., Lu, X., Sperry, S. M., Choi, L. & Denison, M. R. High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *J. Virol.* **81**, 12135–12144 (2007).
47. Sharma, S. *et al.* APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. *Nat. Commun. 2015 61* **6**, 1–15 (2015).
48. Wang, R., Hozumi, Y., Zheng, Y. H., Yin, C. & Wei, G. W. Host Immune Response Driving SARS-CoV-2 Evolution. *Viruses* **12**, (2020).
49. Holland, L. A. *et al.* An 81-Nucleotide Deletion in SARS-CoV-2 ORF7a Identified from Sentinel Surveillance in Arizona (January to March 2020). *J. Virol.* **94**, (2020).
50. Ramazzotti, D. *et al.* Large-scale analysis of SARS-CoV-2 synonymous mutations reveals the adaptation to the human codon usage during the virus evolution. *Virus Evol.* **8**, 1–5 (2022).
51. Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75.e11 (2021).
52. Bhattacharya, M., Chatterjee, S., Sharma, A. R., Agoramoorthy, G. & Chakraborty, C. D614G mutation and SARS-CoV-2: impact on S-protein structure, function, infectivity, and immunity. *Appl. Microbiol. Biotechnol.* **105**, 9035–9045 (2021).
53. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **18**, 1–9 (2020).
54. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nat. 2021 5927854* **592**, 438–443 (2021).
55. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, (2021).
56. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nat. 2021 5997883* **599**, 114–119 (2021).
57. Tiecco, G. *et al.* Omicron Genetic and Clinical Peculiarities That May Overturn SARS-CoV-2 Pandemic: A Literature Review. *Int. J. Mol. Sci. 2022, Vol. 23, Page 1987* **23**, 1987 (2022).
58. Ren, S.-Y., Wang, W.-B., Gao, R.-D. & Zhou, A.-M. Omicron variant (B.1.1.529) of SARS-CoV-2: Mutation, infectivity, transmission, and vaccine resistance. *World J. Clin. Cases* **10**, 1 (2022).
59. Enhancing Readiness for Omicron (B.1.1.529): Technical Brief and Priority Actions for Member States - World Health Organization HQ, 23 December 2021 (updated from the last version published on 17 December 2021) - World | ReliefWeb. https://reliefweb.int/report/world/enhancing-readiness-omicron-b11529-technical-brief-and-priority-actions-member-states-0.
60. Wu, H. *et al.* Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host Microbe* **29**, 1788-1801.e6 (2021).
61. de Haan, C. A. M. & Rottier, P. J. M. Molecular Interactions in the Assembly of Coronaviruses. *Adv. Virus Res.* **64**, 165 (2005).
62. Surjit, M. & Lal, S. K. The nucleocapsid protein of the SARS coronavirus: Structure, function and therapeutic potential. in *Molecular Biology of the SARS-Coronavirus* (ed. Sunil K. Lal) 129–151 (Springer Berlin Heidelberg, 2010). doi:10.1007/978-3-642-03683-5_9.
63. Lesbon, J. C. C. *et al.* Nucleocapsid (N) Gene Mutations of SARS-CoV-2 Can Affect Real-Time RT-PCR Diagnostic and Impact False-Negative Results. *Viruses* **13**, (2021).

8. SUPPLEMENTARY DATA

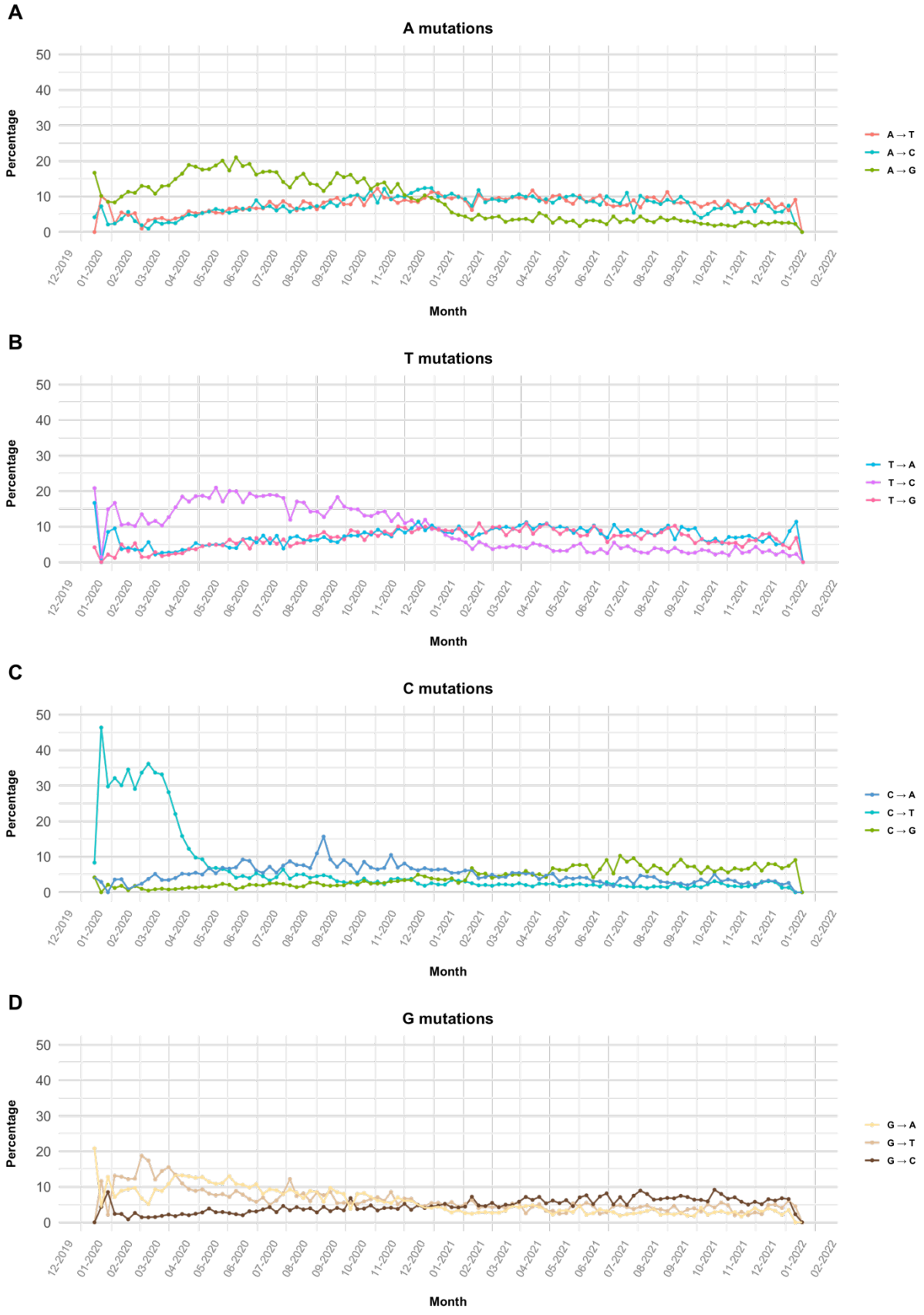

Figure S1 | Percentage of different mutations over the weeks. Calculated by dividing the total number of mutations of each type per week by the total number of mutations detected that week and multiplied by 100.

Table S1 | Primers analyzed

Target	Type	Sequence (5' → 3')	Position	Reference
RdRp	F	ATGAGCTTAGTCCTGTTG	12690-12707	64
	R	CTCCCTTTGTTGTGTTGT	12780-12797	
	P	HEX-AGATGTCTTGTGCTGCCGGTA-BHQ-1	12717-12737	
RdRp	F	GGTAACTGGTATGATTTCG	14080-14098	64
	R	CTGGTCAAGGTTAATATAGG	14167-14186	
	P	FAM-TCATACAAACCACGCCAGG-BHQ-1	14105-14123	
E	F	ACAGGTACGTTAATAGTTAATAGCGT	26269-26294	65
	R	ATATTGCAGCAGTACGCACACA	26360-26381	
	P	FAM-ACACTAGCCATCCTTACTGCGCTTCG-BBQ	26332-26357	
N	F	CACATTGGCACCCGCAATC	28706-28724	65
	R	GAGGAACGAGAAGAGGCTTG	28814-28833	
	P	FAM-ACTTCCTCAAGGAACAACATTGCCA-BBQ	28753-28777	
RdRp	F	GTGARATGGTCATGTGTGGCGG	15431-15452	65
	R	CARATGTTAAASACACTATTAGCATA	15505-15528	
	P	FAM-CAGGTGGAACCTCATCAGGAGATGC-BBQ	15470-15494	
ORF1ab	F	TGGGGYTTTACRGGTAACCT	18778-18797	66
	R	AACRCGCTTAACAAAGCACTC	18889-18909	
	P	FAM-TAGTTGTGATGCWATCATGACTAG-TAMRA	18849-18872	
N	F	TAATCAGACAAGGAACTGATTA	29145-29166	66
	R	CGAAGGTGTGACTTCCATG	29236-29254	
	P	FAM-GCAAATTGTGCAATTTGCGG-TAMRA	29177-29196	
ORF1ab	F	CCCTGTGGGTTTTACACTTAA	13342-13362	67
	R	ACGATTGTGCATCAGCTGA	13442-13460	
	P	FAM-CCGTCTGCGGTATGTGGAAAGGTTATGG-BHQ1	13377-13404	
N	F	GGGGAACTTCTCCTGCTAGAAT	28881-28902	67
	R	CAGACATTTTGCTCTCAAGCTG	28958-28979	
	P	FAM-TTGCTGCTGCTTGACAGATT-TAMRA	28934-28953	
N	F	GACCCCAAAATCAGCGAAAT	28287-28306	68
	R	TCTGGTTACTGCCAGTTGAATCTG	28335-28358	
	P	FAM-ACCCCGCATTACGTTTGGTGGACC-BHQ1	28309-28332	
N	F	TTACAAACATTGGCCGCAAA	29164-29183	68
	R	GCGCGACATTCCGAAGAA	29213-29230	
	P	FAM-ACAATTTGCCCCCAGCGCTTCAG-BHQ1	29188-29210	
N	F	GGGAGCCTTGAATACACCAAAA	28681-28702	68
	R	TGTAGCACGATTGCAGCATTG	28732-28752	
	P	FAM-AYCACATTGGCACCCGCAATCCTG-BHQ1	28704-28727	
N	F	AAATTTTGGGGACCAGGAAC	29125-29144	69
	R	TGGCAGCTGTGTAGGTCAAC	29263-29282	
	P	FAM-ATGTCGCGCATTGGCATGGA-BHQ	29222-29241	
N	F	CGTTTGGTGGACCCTCAGAT	28320-28339	70
	R	CCCCACTGCGTTCTCCATT	28358-28376	
	P	FAM-CAACTGGCAGTAACCA-BQH1	28341-28356	
N	F	GCCTCTTCTCGTTCCTCATCAC	28750-28771	71
	R	AGCAGCATCACCGCCATTG	28842-28860	

9. REFERENCE LIST FOR SUPPLEMENTARY DATA

64. Pasteur, I. Protocol : Real-time RT-PCR assays for the detection of SARS-CoV-2. 1–3 (2020).
65. Corman, V., Bleicker, T., Brünink, S. & Drosten, C. Diagnostic detection of 2019-nCoV by real-time RT-PCR. https://www.who.int/docs/default-source/coronaviruse/protocol-v2-1.pdf.
66. Li Ka Shing Faculty of Medicine, T. U. of H. K. (HKUMed). Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR . https://www.who.int/docs/default-source/coronaviruse/peiris-protocol-16-1-20.pdf.
67. 新型冠状病毒核酸检测引物和探针序列（Specific primers and probes for detection 2019 novel coronavirus）. https://ivdc.chinacdc.cn/kyjz/202001/t20200121_211337.html.
68. Division of Viral Diseases. 2019-Novel Coronavirus (2019-nCoV) Real-time rRT-PCR Panel Primers and Probes. https://www.who.int/docs/default-source/coronaviruse/uscdcrt-pcr-panel-primer-probes.pdf.
69. Shirato, K. *et al.* Development of Genetic Diagnostic Methods for Detection for Novel Coronavirus 2019(nCoV-2019) in Japan. *Jpn. J. Infect. Dis.* **73**, 304–307 (2020).
70. Department of Medical Sciences, M. of P. H. Diagnostic detection of Novel coronavirus 2019 by Real time RT-PCR. https://www.who.int/docs/default-source/coronaviruse/conventional-rt-pcr-followed-by-sequencing-for-detection-of-ncov-rirl-nat-inst-health-t.pdf?sfvrsn=42271c6d_4.
71. Coronavirus causante de la COVID-19 (SARS-CoV-2) Detección, caracterización y producción de una vacuna y un tratamiento. https://www.sigmaaldrich.com/ES/es/life-science/covid.

10. SELF-ASSESSMENT

When Dr. Santi Garcia-Vallvé invited me to join the Cheminformatics and Nutrition group, I knew nothing about bioinformatics. I had only taken one subject in this field and was just ending my second year. Being with the QiN group has allowed me to learn a lot in multiple areas, ranging from big data organization, processing, and presentation. Not to mention working in a group, contributing ideas and suggestions while also being challenged. These features, in my opinion, are just as significant as bioinformatics expertise and can be applied to any other field of research. On the other side, I've learned a lot of protein-analysis software as well as Python and R programming. I am sure that the new strategies I have learned will be extremely useful to me in the future.

In general, and with this current work in particular, I consider my experience in the QiN group to be really helpful on both a professional and personal level. I've learned to work alone while yet working as part of a team and sharing my findings with my coworkers. Moreover, my other intention in writing this assignment was to hone my scientific writing skills by mimicking the style of the scientific publications I read. As a result of this work, I learned a lot about scientific language, critical thinking, and methodology.

After all, and having been able to finish this work, I consider that all the effort put into it has been worthwhile. It has been months of hard work, but in the end, it has been really worth it, especially considering how much I learned along the process. I hope that this is only the start of my research career in the field of bioinformatics.

Thank you so much for taking the time to read my Final Degree Project,

Nil Novau Ferré