Ph.D. degree in Systems Medicine

Curriculum in Human Genetics

European School of Molecular Medicine (SEMM)

University of Milan and University of Naples “Federico II”

Disciplinary sector: BIO/11

An Improved Genomic Surveillance Approach To Dissect The SARS-CoV-2 Pandemic

Antonio Grimaldi
Telethon Institute of Genetics and Medicine (TIGEM)
University ID No.: R12419

Supervisor: Prof. Davide Cacchiarelli

External advisor: Prof. Nicola Elvassore

External advisor: Prof. Diego Di Bernardo

Internal examiner: Prof. Sandro Banfi

External examiner: Prof. Carlo Federico Perno

Academic Year 2021-2022

Index

Abstract 1

1. Introduction 2

1.1. Coronaviruses 2

1.1.1. History and classification of Coronaviruses. 2

1.1.2. COVID-19: the first coronavirus pandemic 6

1.1.3. The origin of COVID-19: SARS-CoV-2. 12

1.1.4. SARS-CoV-2 virion structure 14

1.1.4.1. Spike protein 15

1.1.4.2. Envelope (E) protein 20

1.1.4.3. Membrane protein 21

1.1.4.4. Nucleocapsid (N) protein 22

1.1.5. SARS-CoV-2 genome 24

1.1.6. SARS-CoV-2 pathogenesis 28

1.1.7. SARS-CoV-2 lineages, mutations, and variants of concern. 34

1.2. Next generation sequencing for pathogens genomic surveillance 39

1.2.1. First-generation DNA sequencing 39

1.2.2. Second generation sequencing 40

1.2.3. Third generation sequencing 44

1.2.4. Genomic surveillance of pathogens' genome 45

1.3. Aim of the Thesis 47

2. Materials and Methods 49

2.1. Samples collection, RNA extraction, and SARS-CoV-2 testing 49

2.2. SARS-CoV-2 sequencing and computational analysis 50

2.2.1. Library generation and sequencing for SARS-CoV-2 genotyping 50

2.2.2. Data analysis for SARS-CoV-2 genome reconstruction 52

2.2.3. Discovering new SARS-CoV-2 variants 53

2.3. Host mRNA-seq and computational analysis 54

3. Results 56

3.1. A systematic approach allows the generation of large and robust genomic data in a cost-effective manner....56

3.2. Characterization of SARS-CoV-2 genome evolution in Campania. 63

3.3. Identification of new variants based on the analysis of single mutations. 70

3.4. Transcriptional profiling of SARS-CoV-2 infected patients reveals a gene signature. 81

4. Discussion....87

4.1. Development of an NGS approach for easily monitoring SARS-CoV-2 genome variability 87

4.2. Genomic surveillance allow the identification of new genetic variants 88

4.3. Integrating clinical metadata and genomic data to identify intra-host viral evolution....89

4.4. Main advantages and drawbacks of the proposed NGS workflow 90

4.5. Identification of variant-independent gene signature....91

4.6. Conclusions....92

5. Appendix 94

6. Bibliography....98

7. Figures and Figure legends....108

Abstract

Genomic surveillance of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the only approach to rapidly monitor and tackle emerging variants of concern (VOC) of the COVID-19 pandemic. Such scrutiny is crucial to limit the spread of VOC that might escape the immune protection conferred by vaccination strategies. It is also becoming clear now that efficient genomic surveillance would require monitoring the host gene expression to identify prognostic biomarkers of efficacy and disease progression. Here we applied an integrated workflow for RNA extracted from nasal swabs to obtain in parallel the entire genome of SARS-CoV-2 and host respiratory epithelium transcriptome, representing the majority of Italian processed genomic samples. In addition, we have matured and applied novel proof-of-principle approaches to prioritize possible gain-of-function mutations by leveraging patients' metadata and isolated patient-specific signatures of SARS-CoV-2 infection. The goals mentioned above have all been achieved in a cost-effective manner that does not require automation, in an effort to allow any lab with a benchtop sequencer and a limited budget to perform integrated genomic surveillance on premises.

1. Introduction

1.1. Coronaviruses

1.1.1. History and classification of Coronaviruses.

While describing the epidemiology of the common cold in a population of students, Tyrrell, Kendall, and Boyne in 1960 identified a new pathological agent. The authors tried isolating the viruses responsible for the illness starting from patients' nasal washings. Some of the washes did not yield any isolable virus; however, inoculation of these samples to volunteers caused the development of the common cold, even when filtered through bacteria-tight membranes[1]. The infectious agent was called B814 and during the following four years, Tyrrell and Boyne demonstrated that the pathogen was able to propagate, infect volunteers treated with tetracycline and that ether inactivated it[2]. These findings proved that B814 was a virus. In the same period, McIntosh, Hamre, and others identified several viruses similar to B814, including 229E and OC43[3,4]. All of them were ether sensitive and able to propagate in the presence of inhibitors of DNA-replication, suggesting that the viruses had a lipid envelope and an RNA genome[2–5]. Serological tests, however, excluded that any of these viruses were related to known orthomyxoviruses (the other main class of human enveloped viruses, comprising *Influenza* and related viruses), thus indicating the potential discovery of a new class of viral agents[2–4]. Finally, in 1968, Almeida and Tyrell demonstrated through electron microscopy that the newly characterized pathogens were morphologically and functionally distinct from *Influenza* viruses[6]. Their structures, instead, resembled those of viruses recently discovered in mice (*Mouse Hepatitis Virus*, MHV)[7] and chickens (*Infectious Bronchitis Virus*, IBV)[3,6]. Indeed, all these agents were described as pleomorphic particles, ranging from 80-120 nm in diameter and surrounded by 20 nm long spikes projecting from the main body[3]. These structures were petal-shaped, longer, and less abundant than the ones found in *Orthomyxoviruses* and gave the virion the general appearance of a solar corona[3,6] (Fig.1).

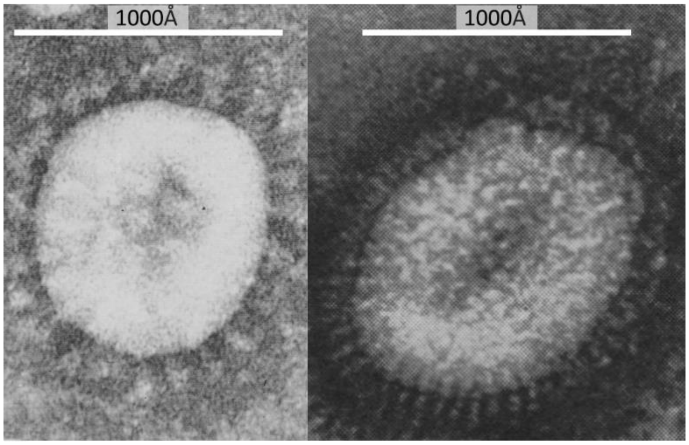

Figure 1. Electron micrograph of negative stained virions showing the global differences between Coronaviruses (B814, left) and Orthomyxoviruses (Influenza A2, right). While the size of the two virions is comparable, the spikes are longer and fewer in the former. Modified from Almeida et al. (1966)[6] and McIntosh et al. (1967)[3].

Because of this, in 1968, Almeida coined the term *Coronaviruses* and suggested the designation of a new virus family (*Coronaviridae*) to group them[5,8]. Currently, all human coronaviruses are grouped under the *Orthocoronavirinae* subfamily (*Coronaviridae* family in the *Nidovirales* order). Serology and genomic studies further divide this taxon into 4 main genera: *Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus*[9]. Human-infecting coronaviruses, however, have been identified only in *Alphacoronavirus* (HCoV-229E and HCoV-NL63) and *Betacoronavirus* (HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV and SARS-CoV-2)[9,10] (Fig.2).

Coronaviruses have been believed to cause only mild to moderate respiratory or gastro-intestinal symptoms for over 30 years until the isolation of the acute respiratory syndrome coronavirus (SARS-CoV)[11,12]. This virus was the aetiological agent of a severe pneumonia outbreak in China in 2002-2003 and boosted the research towards identifying new coronaviruses[13]. This yielded to the isolation of the *Human coronavirus NL-63* (HCoV-NL63)[14] and *Human coronavirus HKU1* (HCoV-HKU1)[15] in 2004 and 2005, respectively. Both the viruses were mainly associated with mild respiratory manifestations, with only a few cases of more severe lower respiratory tract symptoms associated with an HCoV-NL63 strain[13]. Ten years after the discovery of SARS-CoV, another virus raised concern for its high virulence. It was associated with the Middle East Respiratory Syndrome (MERS) outbreak of 2012 and thus named MERS-CoV[16]. As SARS-CoV, also MERS-CoV was highly infective, able to localize in lower respiratory airways, and had originated in bats[10]. In 2017 Hu, Zeng and Yang reported the existence of a genetic pool in bat coronaviruses related to SARS-CoV (globally known as SARS-related Coronaviruses, SARSr-CoV) containing all the genetics "building blocks" of SARS-CoV genome in a cave in Yunnan district, China[17]. Serological proofs of actual bat coronaviruses spilling over to humans were also identified[18]. These and other findings[19] highlighted the risk of the emergence of new human coronavirus outbreaks[20,21]. This was actually the case in 2019, when *SARS coronavirus 2* (SARS-CoV-2) was identified as the aetiological agent of Coronavirus Disease 2019 (COVID-19) pandemic[22,23].

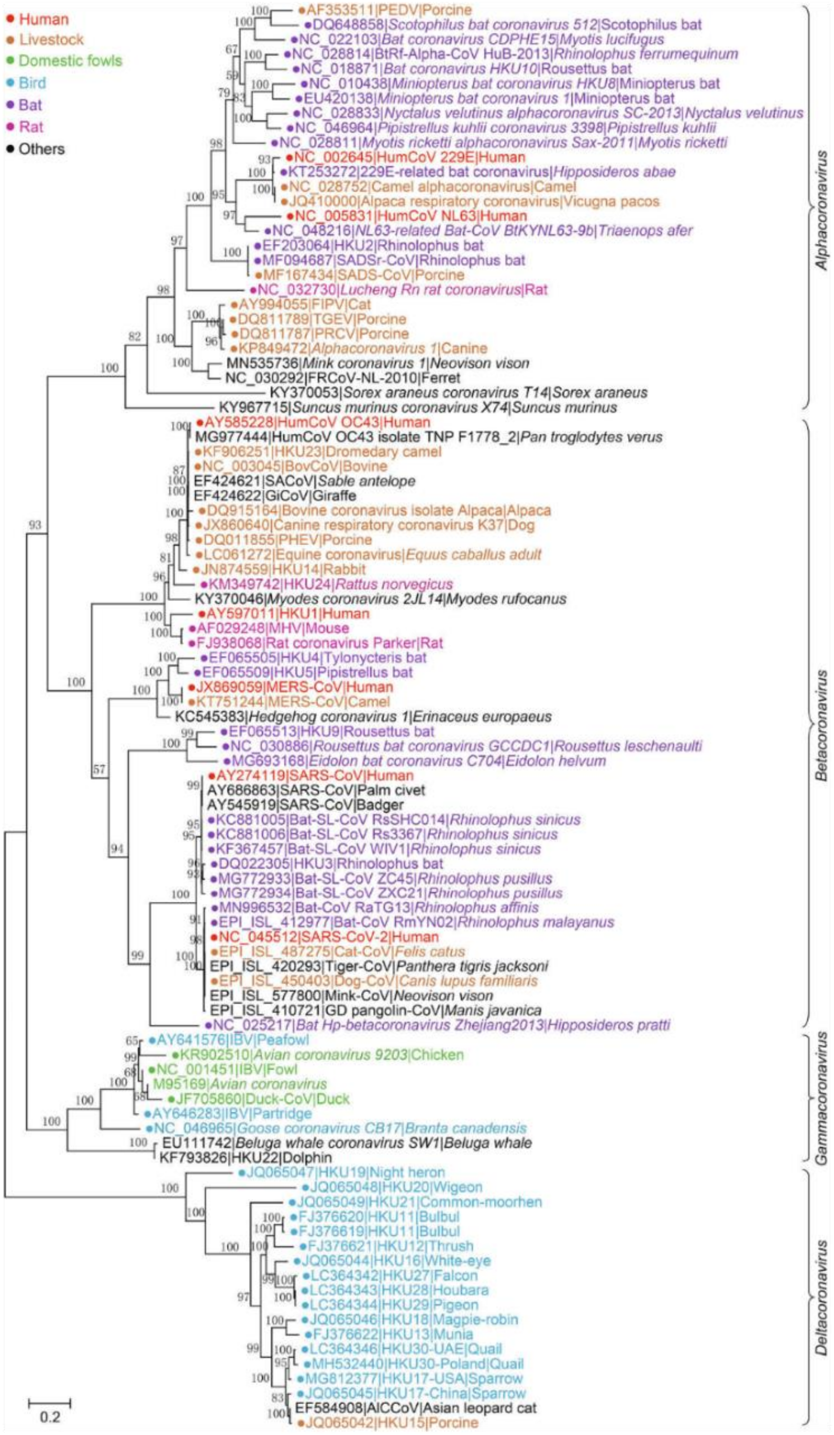
● Human
● Livestock
● Domestic fowls
● Bird
● Bat
● Rat
● Others
●AF353511|PEDV|Porcine
●DQ648858|Scotophilus bat coronavirus 512|Scotophilus bat
●NC_022103|Bat coronavirus CDPHE15|Myotis lucifugus
●NC_028814|BtRf-Alpha-CoV HuB-2013|Rhinolophus ferrumequinum
●NC_018871|Bat coronavirus HKU10|Rousettus bat
●NC_010438|Miniopterus bat coronavirus HKU8|Miniopterus bat
●EU420138|Miniopterus bat coronavirus 1|Miniopterus bat
●NC_028833|Nyctalus velutinus alphacoronavirus SC-2013|Nyctalus velutinus
●NC_046964|Pipistrellus kuhlii coronavirus 3398|Pipistrellus kuhlii
●NC_028811|Myotis ricketti alphacoronavirus Sax-2011|Myotis ricketti
●NC_002645|HumCoV 229E|Human
●KT253272|229E-related bat coronavirus|Hipposideros abae
●NC_028752|Camel alphacoronavirus|Camel
●JQ410000|Alpaca respiratory coronavirus|Vicugna pacos
●NC_005831|HumCoV NL63|Human
●NC_048216|NL63-related Bat-CoV BtKYNL63-9b|Triaenops afer
●EF203064|HKU2|Rhinolophus bat
●MF094687|SADSr-CoV|Rhinolophus bat
●MF167434|SADS-CoV|Porcine
●NC_032730|Lucheng Rn rat coronavirus|Rat
●AY994055|FIPV|Cat
●DQ811789|TGEV|Porcine
●DQ811787|PRCV|Porcine
●KP849472|Alphacoronavirus 1|Canine
MN535736|Mink coronavirus 1|Neovison vison
NC_030292|FRCoV-NL-2010|Ferret
KY370053|Sorex araneus coronavirus T14|Sorex araneus
KY967715|Suncus murinus coronavirus X74|Suncus murinus
Alphacoronavirus
●AY585228|HumCoV OC43|Human
MG977444|HumCoV OC43 isolate TNP F1778_2|Pan troglodytes verus
●KF906251|HKU23|Dromedary camel
●NC_003045|BovCoV|Bovine
EF424621|SACoV|Sable antelope
EF424622|GiCoV|Giraffe
●DQ915164|Bovine coronavirus isolate Alpaca|Alpaca
●JX860640|Canine respiratory coronavirus K37|Dog
●DQ011855|PHEV|Porcine
●LC061272|Equine coronavirus|Equus caballus adult
●JN874559|HKU14|Rabbit
●KM349742|HKU24|Rattus norvegicus
KY370046|Myodes coronavirus 2JL14|Myodes rufocanus
●AY597011|HKU1|Human
●AF029248|MHV|Mouse
●FJ938068|Rat coronavirus Parker|Rat
●EF065505|HKU4|Tylonycteris bat
●EF065509|HKU5|Pipistrellus bat
●JX869059|MERS-CoV|Human
●KT751244|MERS-CoV|Camel
KC545383|Hedgehog coronavirus 1|Erinaceus europaeus
●EF065513|HKU9|Rousettus bat
●NC_030886|Rousettus bat coronavirus GCCDC1|Rousettus leschenaulti
●MG693168|Eidolon bat coronavirus C704|Eidolon helvum
●AY274119|SARS-CoV|Human
AY686863|SARS-CoV|Palm civet
AY545919|SARS-CoV|Badger
●KC881005|Bat-SL-CoV RsSHC014|Rhinolophus sinicus
●KC881006|Bat-SL-CoV Rs3367|Rhinolophus sinicus
●KF367457|Bat-SL-CoV WIV1|Rhinolophus sinicus
●DQ022305|HKU3|Rhinolophus bat
●MG772933|Bat-SL-CoV ZC45|Rhinolophus pusillus
●MG772934|Bat-SL-CoV ZXC21|Rhinolophus pusillus
●MN996532|Bat-CoV RaTG13|Rhinolophus affinis
●EPI_ISL_412977|Bat-CoV RmYN02|Rhinolophus malayanus
●NC_045512|SARS-CoV-2|Human
●EPI_ISL_487275|Cat-CoV|Felis catus
EPI_ISL_420293|Tiger-CoV|Panthera tigris jacksoni
●EPI_ISL_450403|Dog-CoV|Canis lupus familiaris
EPI_ISL_577800|Mink-CoV|Neovison vison
EPI_ISL_410721|GD pangolin-CoV|Manis javanica
●NC_025217|Bat Hp-betacoronavirus Zhejiang2013|Hipposideros pratti
Betacoronavirus
●AY641576|IBV|Peafowl
●KR902510|Avian coronavirus 9203|Chicken
●NC_001451|IBV|Fowl
M95169|Avian coronavirus
●JF705860|Duck-CoV|Duck
●AY646283|IBV|Partridge
●NC_046965|Goose coronavirus CB17|Branta canadensis
EU111742|Beluga whale coronavirus SW1|Beluga whale
KF793826|HKU22|Dolphin
Gammacoronavirus
●JQ065047|HKU19|Night heron
●JQ065048|HKU20|Wigeon
●JQ065049|HKU21|Common-moorhen
●FJ376620|HKU11|Bulbul
●FJ376619|HKU11|Bulbul
●FJ376621|HKU12|Thrush
●JQ065044|HKU16|White-eye
●LC364342|HKU27|Falcon
●LC364343|HKU28|Houbara
●LC364344|HKU29|Pigeon
●JQ065046|HKU18|Magpie-robin
●FJ376622|HKU13|Munia
●LC364346|HKU30-UAE|Quail
●MH532440|HKU30-Poland|Quail
●MG812377|HKU17-USA|Sparrow
●JQ065045|HKU17-China|Sparrow
EF584908|AICCoV|Asian leopard cat
●JQ065042|HKU15|Porcine
Deltacoronavirus
0.2

Figure 2. Maximum likelihood tree based on RdRP proteins in Coronaviruses. The tree shows the classification of Coronaviruses in 4 main genera (*Alpha-*, *Beta-*, *Gamma-* and *Deltacoronavirus*) and the corresponding distribution of human-infecting viruses (red). For each virus, the accession ID of a representative genomic sequenced is shown. Colours represents the host and the number near each node are the bootstrap values. The bar indicates the genetic distance (number of substitutions per RdRPp residue). Modified from Zhou et al. (2021)[10].

1.1.2. COVID-19: the first coronavirus pandemic

COVID-19 is the second pandemic recognized in the XXI century, the first documented to be caused by a coronavirus in history[24]. As of August 2022, it affected over 601 million people and caused at least 6.49 million casualities[25] (Fig. 3).

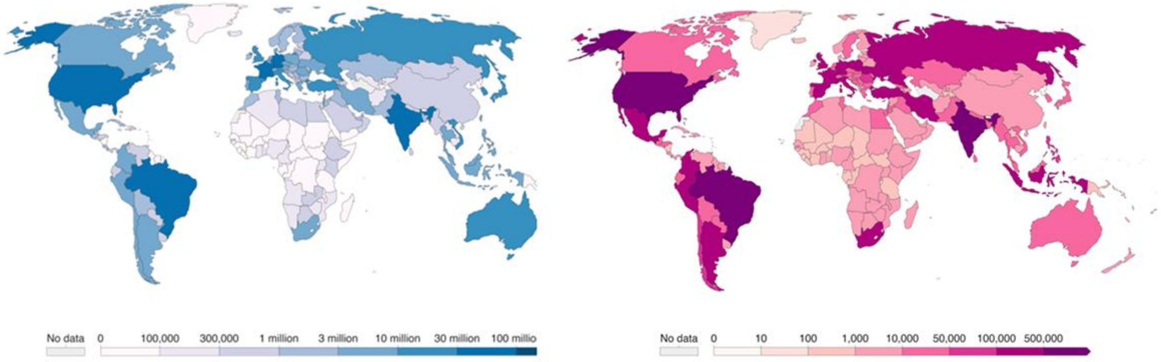

Figure 3. Worldwide cumulative confirmed COVID-19 cases (left) and deaths (right) updated to August 30, 2022. Due to limited testing, variability in diagnostic protocols, or difficulties in attributing the cause of death, these numbers are probably an underestimation. Modified from https://ourworldindata.org/coronavirus[25]

The first COVID-19 cases were reported in Wuhan, Hubei province, China, in late December 2019 as a viral pneumonia of unknown aetiology. However, the onset of the symptoms in the first patient was traced back to the beginning of the month[26]. One of the first reports describing the illness showed that the pathology caused only unspecific symptoms at its onset, including cough, myalgia, and fever. Nevertheless, half of the patients in the study developed dyspnoea, and 1/3 were admitted to the intensive care unit because of the severe degree of hypoxaemia[26]. Later, sequencing studies revealed that the virus causing the disease was a previously unknown *betacoronavirus* named SARS-CoV-2[22,23]. Animal-to-human transmission was believed to be the only root of the infection. However, In February 2020, Chan and co-workers reported the first cases of person-to-person viral transmission in family and nosocomial settings[27]. Interhuman contagion confirmed the fear of a wide spread of the disease, which in January 2020 was not restricted to Wuhan anymore. Indeed, despite several interventions acting to limit viral dissemination (including FFP1/2 masks, handwashing, travel restrictions, and even city lockdown), SARS-CoV-2 rapidly spread all around the world[28]. These data and the overall fatality rate, initially close to 3%, fuelled global concern about the new disease[29], and in March 2020, WHO declared COVID-19 a world pandemic[30] (Fig 4).

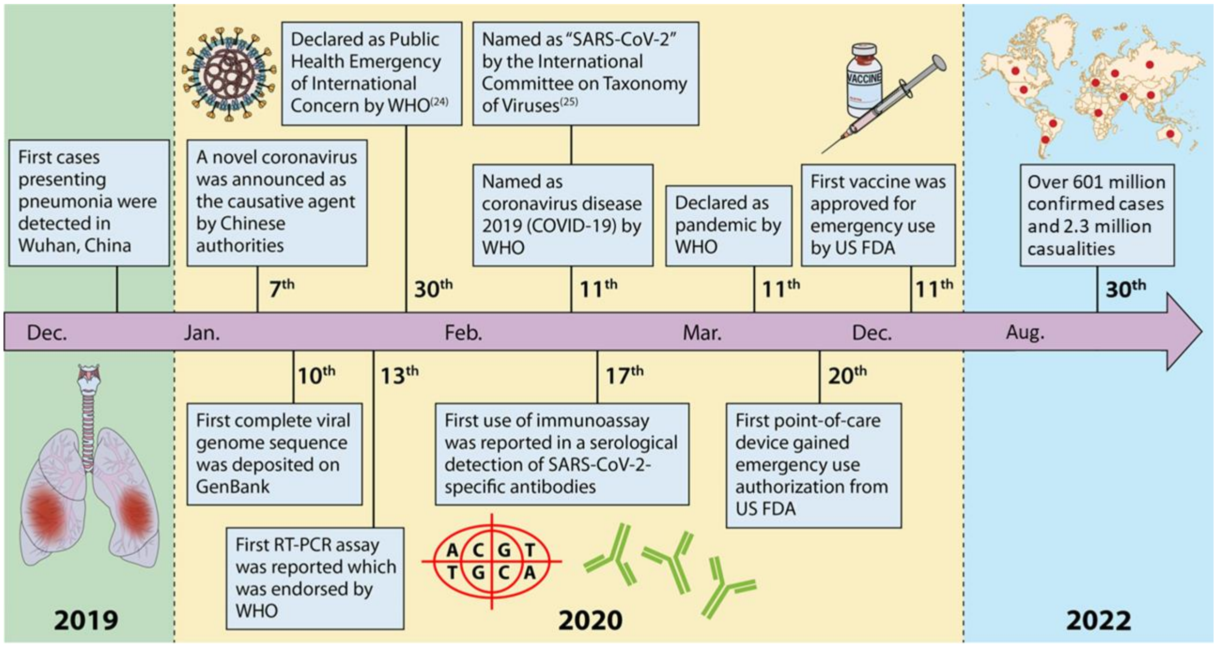
First cases presenting pneumonia were detected in Wuhan, China
A novel coronavirus was announced as the causative agent by Chinese authorities
Declared as Public Health Emergency of International Concern by WHO(24)
Named as "SARS-CoV-2" by the International Committee on Taxonomy of Viruses(25)
Named as coronavirus disease 2019 (COVID-19) by WHO
Declared as pandemic by WHO
First vaccine was approved for emergency use by US FDA
VACCINE
Over 601 million confirmed cases and 2.3 million casualities
7th
30th
11th
11th
11th
30th
Dec.
Jan.
Feb.
Mar.
Dec.
Aug.
10th
13th
17th
20th
First complete viral genome sequence was deposited on GenBank
First RT-PCR assay was reported which was endorsed by WHO
First use of immunoassay was reported in a serological detection of SARS-CoV-2-specific antibodies
First point-of-care device gained emergency use authorization from US FDA
A C G T
T G C A
2019
2020
2022

Figure 4. COVID-19 spread from its first detection in December 2019 till the end of August 2022. The figure shows some of the most important events during the pandemic. Modified from Safiabadi et al. (2021)[28].

From a clinical perspective, COVID-19 is a respiratory disease potentially affecting both upper and lower airways. The first symptoms appear in most patients by 11.5 days post-infection, with a median incubation period of 5.1 days. Symptomatology varies depending on disease severity, which is divided into 4 main classes according to 2022 WHO guidelines[31] (Fig. 5):

- **Asymptomatic or pre-symptomatic infection.** Patients tested positive for SARS-CoV-2 but not displaying any symptomatology. The absence of symptoms could be due to early detection of the virus when the disease is still incubating. However, a percentage of individuals between 17.9% and 33.3% usually remain asymptomatic and never develop any symptom[31,32].

- **Non-severe illness.** While symptomatic, patients do not display symptoms of respiratory distress or sign of pneumonia (such as abnormal chest imaging). Most symptomatic patients (up to 70%) experience fever, shortness of breath, and cough. Other common clinical manifestations include myalgia and headache (35%), sole throat, anosmia, and dysgeusia (loss of smell and taste, respectively)[31,32].

- **Severe illness.** Patients display pneumonia, signs of severe respiratory distress (including the inability to complete full sentences, respiratory frequency >30 breaths per minute), and oxygen saturation is lower than 90% in room air[31,32]. This saturation threshold is arbitrary, and WHO guidelines recommend cautiously evaluating all cases with oxygen saturation ≤94%[31]. It has been shown that up to 14% of symptomatic patients can develop severe COVID-19[33].

- **Critical illness.** Patients require life-sustaining therapies such as mechanical ventilation or vasopressor therapy. Symptoms include sepsis, septic shock, multiple organ failure, and Acute Respiratory Distress Syndrome (ARDS)[31]. ARDS is a life-threatening severe respiratory failure (mortality rate as high as 50%) characterized by arterial hypoxemia that is refractory to oxygen administration[34]. Up to 5% of all symptomatic patients display critical illness[33].

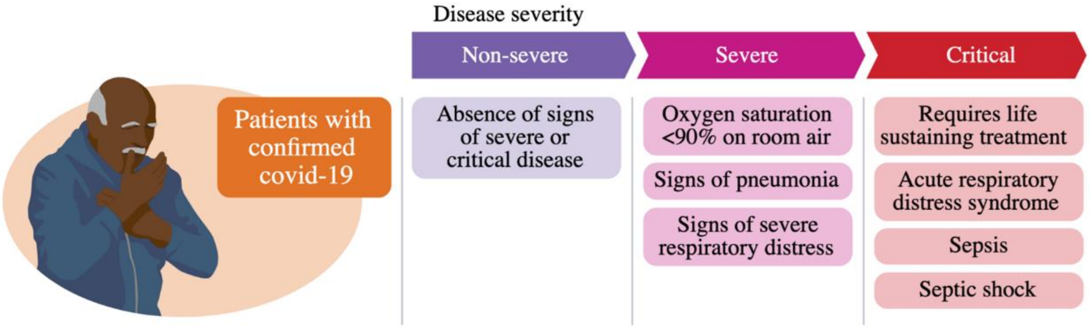
Disease severity
Non-severe
Severe
Critical
Patients with confirmed covid-19
Absence of signs of severe or critical disease
Oxygen saturation <90% on room air
Signs of pneumonia
Signs of severe respiratory distress
Requires life sustaining treatment
Acute respiratory distress syndrome
Sepsis
Septic shock

Figure 5. Scheme showing COVID-19 classification based on disease severity. Each category has a defining set of diagnostic features and requires specific treatments. Modified from "*Therapeutics and COVID-19: living guideline*"[31]

Several factors reshaped the frequency of symptomatic or severe infections during the pandemic. Among these, the beginning of the vaccination campaign played a pivotal role [35]. Vaccinations proved to be highly effective in protecting against both hospitalisation and death, decreasing the odds of developing a severe illness by 6 times [35]. Nevertheless, the protection against infection (but not hospitalisation) usually decreases six months after the inoculation, at least for the first two doses[36,37]. As a result, the gap between the frequencies of vaccinated and unvaccinated COVID-19 patients varies with a six-month periodicity[35]. SARS-CoV-2 genetic variants (further discussed in paragraph 1.1.7) also have a role in vaccine effectiveness. The ones identified during 2020-2021, in particular (i.e., alpha, beta, gamma, and delta), were associated with increased risk of hospitalisation[38]. None of them, however, significantly affected the vaccine-elicited protection against the severe disease[38]. Interestingly, infections due to the main 2022 variant (named omicron) showed an opposite trend. This genetic variant is associated with a significative decrease in vaccine effectiveness[39] that is balanced by its intrinsic lower capability to induce hospitalisation[40].

Several other risk factors have been associated with severe and critical COVID-19. Older age is probably the most known. It has been shown that hospitalized patients over 59 years of age had 5.1 times more chances of developing severe or critical illness than those aged 30-59[41]. Similarly, the adults to children ratio in patients displaying severe or critical symptoms is as high as 22:1[41]. Intriguingly, also male gender appears to increase the odds of developing severe COVID-19[41,42]. Indeed, the probability of severe COVID-19 in males can be up to twice that of women[43]. The reasons underlying these statistics have not been clearly defined yet. Since most immunity-related genes lie on chromosome X, their bi-allelic expression in females might contribute to a more robust immune response to SARS-CoV-2 infection[43,44]. Another potential factor might be the differential expression of oestradiol which has been negatively correlated with pro-inflammatory cytokine levels[41,43,44]. Finally, several comorbidities, including hypertension, diabetes, and obesity, have also been associated with worse outcomes in COVID-19 patients[41] (Fig. 6).

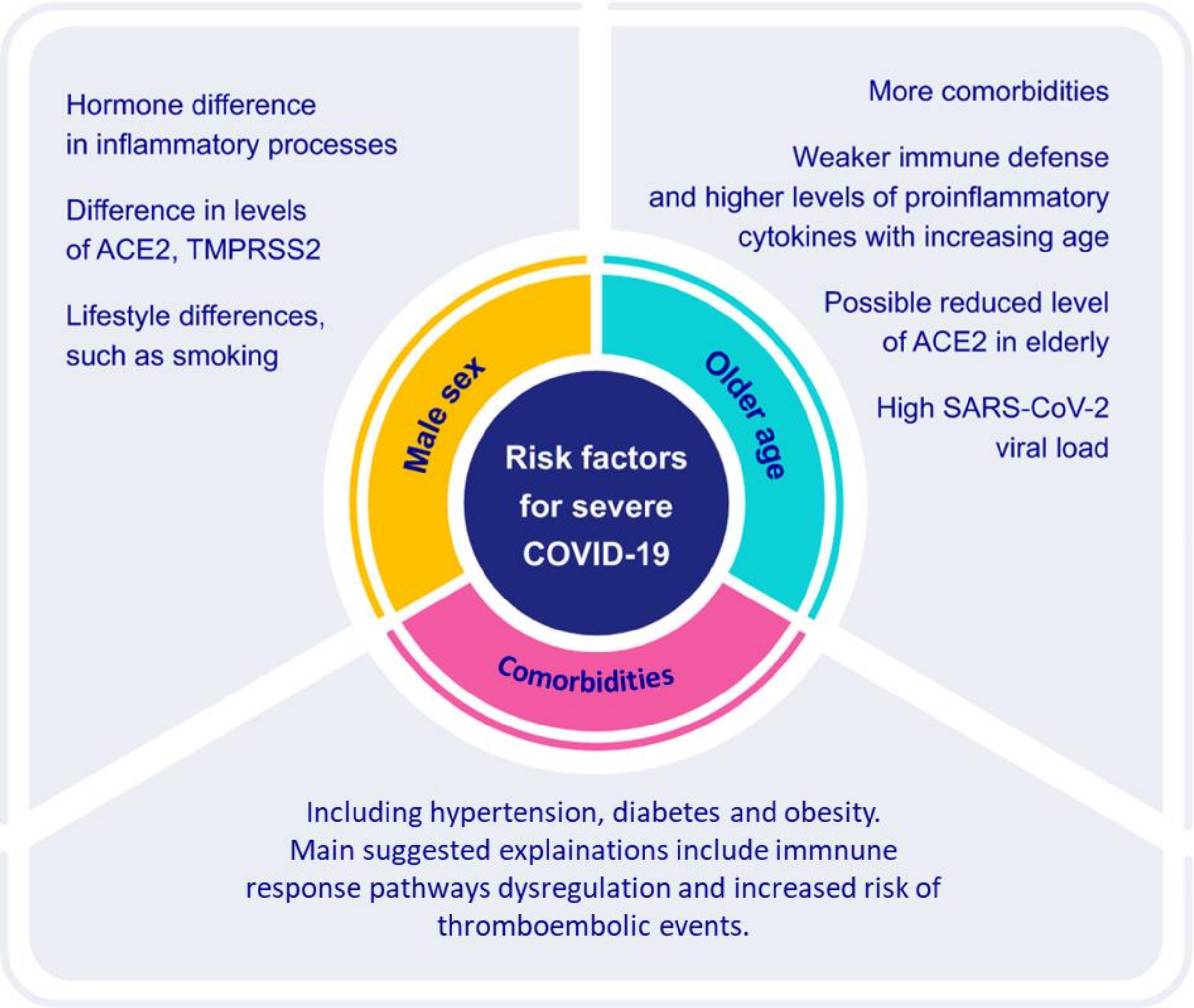

Figure 6. Main severe COVID-19 risk factors. Older age is generally associated with an increase in comorbidities, weak immune defense, and higher levels of proinflammatory molecules. In addition, ACE2 levels are decreased in the elderly and might be part of the mechanism causing higher risks of severe illness. Differences in sex hormones involved in inflammatory processes are among the leading causes of males' higher risk of developing severe symptomatology. Also, the expression levels of ACE2 and TMPRSS2 vary in males and females and might play a role. Other risk factors are hypertension, diabetes, and obesity. Modified from Gao et al. (2021)[41].

1.1.3. The origin of COVID-19: SARS-CoV-2.

SARS-CoV-2 is the causative agent of COVID-19. It is a *betacoronaviru*s belonging to the *sarbecovirus* subgenera and is a SARS-related Coronavirus. Like all the other coronaviruses infecting humans, also SARS-CoV-2 is believed to have an animal origin[13,45]. Nevertheless, its original natural host and the mechanisms allowing the initial cross-species transmission of the virus are not fully understood. Its genome has been initially found to be similar to one of several *coronaviruses* isolated in pangolins[46]. Subsequent analyses, however, have ruled out the possibility that pangolins were the original host of the virus[47]. Indeed, the genome of a bat-infecting SARSr-CoV (named RaTG13) identified in 2013 appeared to have a high similarity to that of SARS-CoV-2 (96,3%)[48]. Interestingly, this virus was sampled in a cave in Yunnan province (China) following the identification of three patients that worked there and displayed a severe respiratory syndrome[49].

A more recent work by Temmam and co-workers[50] identified at least three bat SARSr-CoVs genomes (named BANAL-20-52, BANAL-20-103, BANAL-20-236) sharing the highest degree of genetic similarity with SARS-CoV-2 so far. These viruses have the same ability to infect human cells as the first isolates of SARS-CoV-2 and are sensible to SARS-CoV-2 specific immunoglobulins. Indeed, bats of the genera *Rhinolophus* are considered the main animal reservoir of coronaviruses and the species where almost all human-infecting *betacoronaviruses* originated[51]. Currently, it is hypothesized that SARS-CoV-2 is derived from multiple recombination events in *Rhinolophus* SARS-related coronaviruses[50] (Fig. 7).

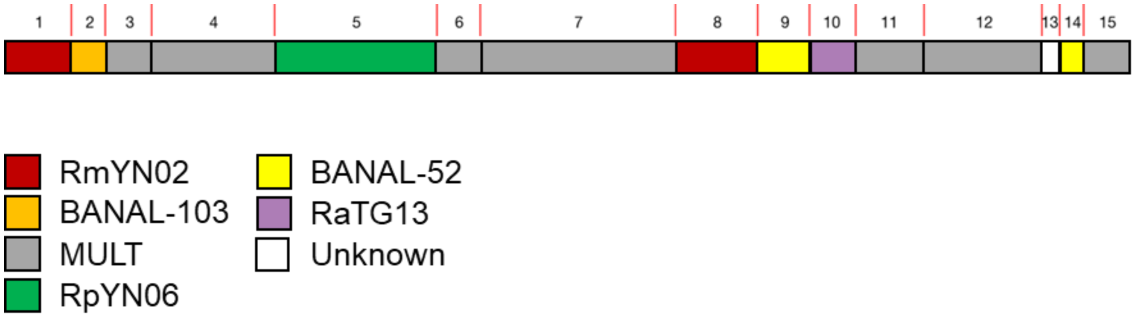

Figure 7. Representation of the possible recombinant origin of the SARS-CoV-2 genome. RNA alignment with other *Sarbecoviruses* reveals 15 possible fragments (numbers on top) deriving from genome recombination. Red bars represent putative break points. Colours are the most similar SARSr-CoV genome(s). "MULT" is used when multiple sequences are equally similar to the fragment. "Unknown" indicates a region of unresolved phylogeny. Modified from Temman et al. (2022)[50].

It is unclear how the virus spread from *Rinolophus* bats to humans, mainly because of the ecological separation of bats from humans[52]. However, current data suggest that at least two independent zoonotic events give rise to SARS-CoV-2 pandemic, the first occurring in mid-November 2019 (and responsible for the SARS-CoV-2 "B lineage"), the subsequent at the end of the same month (generating the "A lineage", see paragraph 1.1.7)[45]. Both events have been proposed to occur in the Huanan seafood wholesale market (hereafter Huanan), in Wuhan[45,53], amongst the most important wholesale wet markets in central China. Indeed, about 33% of the earliest cases were directly exposed to the Huanan market, and the remaining patients lived close to it (Fig. 8)[53].

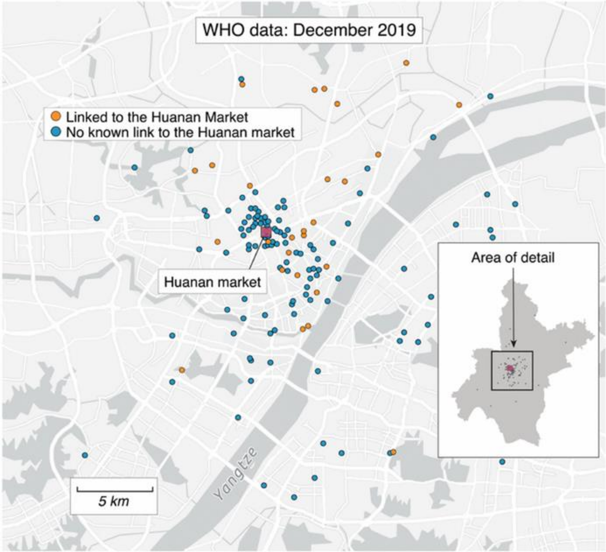
WHO data: December 2019
● Linked to the Huanan Market
● No known link to the Huanan market
Huanan market
Area of detail
5 km
Yangtze

Figure 8. Map of the first 155 cases reported in Wuhan. The inset shows the map of all December 2019 cases in the city, grey dots are the ones non displayed in the main panel. In both the maps the red square indicates the location of Huanan market. Modified from: Worobey et al. (2022)[53].

Later analysis of environmental samples in the Hunan market, including instruments, gloves, and animal cages, tested positive for the presence of SARS-CoV-2[45,53]. In addition, the Huanan market was known for live wildlife trade, including raccoon dogs, pangolins, and other mammals susceptible to SARSr-CoV infection[45,53,54]. These animals were caught in the same areas where also *Rinolophus* SARSr-CoVs extensively circulated[19,53], enforcing the hypothesis that they acted as intermediary hosts for the initial viral spill over that started the pandemic.

1.1.4. SARS-CoV-2 virion structure.

SARS-CoV-2 is an enveloped virus with a positive-sense, single-stranded RNA genome. The virion is pleiomorphic or roughly spherical and 60-140 nm in diameter[22]. One of the most known characteristics of all coronaviruses is the solar corona-like shape of the virions when observed by negative-staining electron microscopy. This appearance is due to a fringe of petal-shaped or globular proteins projecting 9-14 nm apart from the viral envelope and known as peplomer or Spike (S) proteins[22].

The envelope alone is about 80 nm in diameter and 7.8 nm thick (while regular biomembranes thickness is about 4 nm)[55,56]. Its lipid composition derives from the remodeling of host intracellular membranes, mostly Endoplasmic Reticulum (ER) and Golgi complex[55]. Besides the Spike proteins, the viral envelope contains two other transmembrane polypeptides, namely the M (Membrane, 15-30 kDa) and E (Envelope, 8-12 kDa) proteins, with M being the most abundant structural protein in the virion [12]. Encapsulated in the envelope lies the viral nucleocapsid composed of N (Nucleocapsid, 43-50 kDa) proteins dimers wrapped around the RNA genome with a helicoidal symmetry (an uncommon property for positive-sense RNA viruses) (Fig. 3)[12,57]. Each of the proteins listed above (collectively known as structural proteins) has a precise structure and functional role in the viral life cycle.

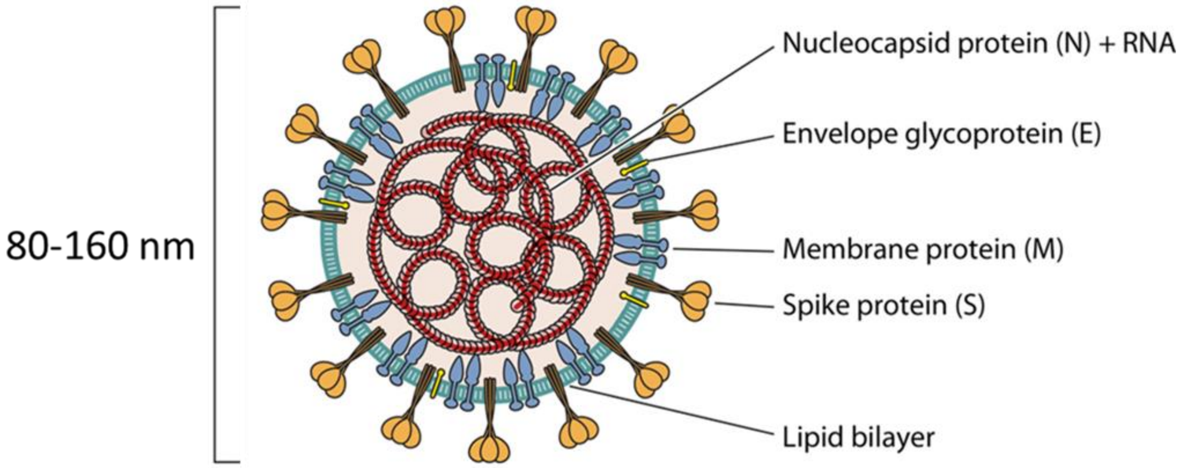

Figure 9. Diagram of SARS-CoV-2 virion showing main structural proteins and features. Adapted from Safiabadi et al. (2021)[28].

1.1.4.1. Spike protein

The Spike (S) protein is a 660 kDa transmembrane glycoprotein[58], the main determinant of coronaviruses tropism, mediating membranes fusion and viral entry upon binding to a specific host receptor, Angiotensin I-converting enzyme 2 (ACE2)[12]. S protein is a pear- or petal-shaped homotrimer[5,58], with each protomer composed of 2 subunits: S1 and S2, the former being important for host receptor recognition, the latter for mediating membrane fusion and anchoring the protein to viral envelope[52,58]. These subunits are derived from the cleavage of the S protein during its maturation. The cleavage site is often referred to as S1/S2 or furin-like site and is located between residues 685-686[58,59]. The S1 subunit comprises two independent domains, the N- and C- terminal domains (S1-NTD and S1-CTD, respectively)[60]. S1-CTD (residues 319-541)[61] is the domain directly interacting with ACE2 and is thus also referred to as Receptor Binding Domain (RBD)[60]**.** RBD can assume two different conformations, "up" and "down". In the "up" conformation, the RBD exposes the receptor binding sites (composing the Receptor Binding Motif, RBM) and can bind ACE2[58]. S1-NTD (residues 14-305) is the most variable region among S proteins of SARSr-CoVs[59]. Indeed, while SARS-CoV-2 spike protein shares 90.6% of its residues with the one of BANAL-20-306 (see paragraph **1.1.3**), this value collapses to just 65.8% when comparing the two NTDs [50,62]. However, the role of this domain in SARS-CoV-2 is not well understood. One hypothesis is that NTD might bind some glycans in the early phases of viral entry in host cells[63,64].

The S2 subunit comes right after the S1 domain and is closer to the viral membrane. It can be divided into several regions:

- **S2′ cleavage site**, corresponding to residues 815-816[59]. It is a dibasic peptide (Lysine-Arginine) recognized and cut mainly by human transmembrane protease, serine 2 (TMPRSS2)[59]. This cleavage (occurring right downstream of the arginine residue) activates the S protein, allowing a complex set of conformational changes in both S1 and S2 and culminating in the exposure of a downstream region known as fusion peptide[59].
- The **Fusion peptide** (FP) is a small peptide rich in non-polar residues[59]. This characteristic mediates the strong affinity of this region for biological

membranes. Indeed, upon S activation, the fusion peptide is inserted in the host membranes, destabilizing them, and mediating their fusion with viral envelope[58,59].

- **Heptad-repeat regions (HR1/HRN and HR2/HRC)** are two heptad repeats located downstream the FP and separated by 180 residues[59]. Each HR is a heptamer repat having the general structure (*abcdefg*)$_n$ with *a* and *d* usually being non-polar residues[65]. In S homotrimer, HR1 and HR2 fold into a 6-helix bundle that is proposed to facilitate FP insertion in host membranes upon S activation[65].
- The S2 C-terminal region contains the **Transmembrane (TM) domain and the Cytoplasmic (CT) tail.** The TM anchors the Spike protein to the viral envelope[61]. It consists of a single transmembrane alpha helix rich in hydrophobic residues that are believed to play critical roles in S trimerization and membrane fusion[61]. Downstream to the TM and inserted into the virion lumen, there is the CT tail. It is a stretch of hydrophilic residues rich in palmitoylated cysteines required for correct viral trafficking and syncytium formation in cell-to-cell viral transmission[59,61] (Fig. 10).

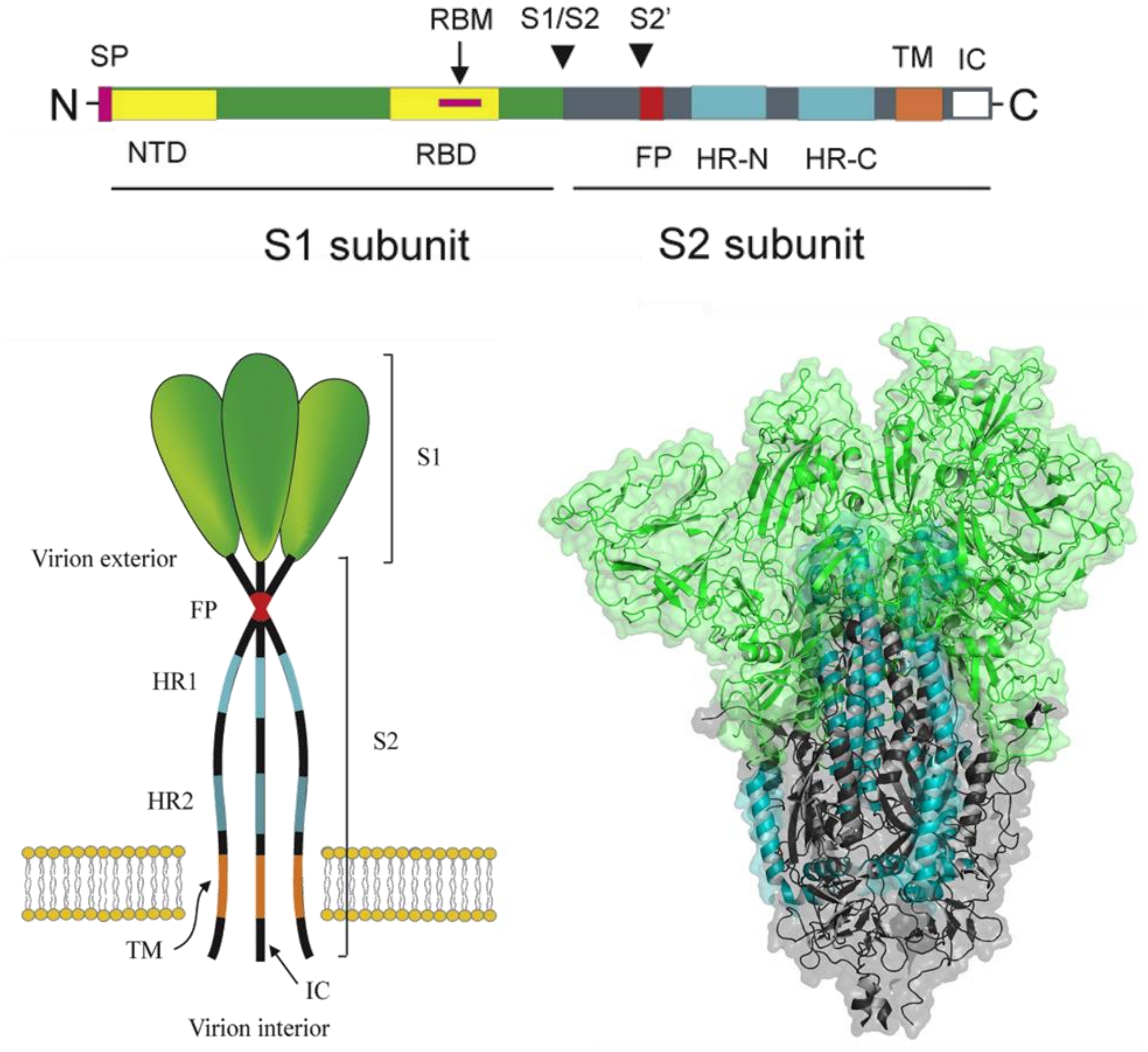

Figure 10. Model of Spike protein structure and organization within the virion membrane. Top: the spike protein is split into two subunits (S1 and S2) by the cleavage in the S1/S2 site. The S1 domain contains a Signal Peptide (SP), the N-terminal Domain (NTD) and the Receptor Binding Domain (RBD). The Receptor Binding Motif (RBM) is part of the RBD and directly interacts with spike protein receptor. S2 Domain contains several regions crucial for spike activation, including a cleavage site (S2′), the Fusion Peptide (FP) and two Heptad Repeats (HR-N and HR-C). The transmembrane region anchors the protein to virion membrane while the intracellular tail (IC) is in virion interior. Bottom: Tertiary structure scheme (left) of spike protein homotrimer and corresponding structure determined by crystallography (right, PDB: 7BNN). Colours are the same as in Top panel. Adapted from Artica et al. (2020)[52].

Studies on intact virions revealed that 26 spike molecules are present on the surface of each viral particle and project from its surface at a 40° angle[58]. Each of these proteins exists as a metastable molecule in the so-called "prefusion" conformation[58]. This is the state of S protein before its activation. It can be further classified depending on the conformation of the three RBDs[66]. The closed state (referred to as "RBD down") corresponds to all RBDs in "down" conformation and represents 31% of all prefusion S trimers[66]. RBDs are exposed in the "open" conformation, acquired when one ("one RBD up", 55% of prefusion S proteins) or two RBDs ("two RBDs up", 14%) have an "up" conformation[66].

Spike protein rapidly undergoes a radical structural change upon receptor binding[56,67] and acquires the "postfusion" conformation[56]. In this state, S protein shifts from a petal-shaped to a needle-like structure (Fig. 11), with TM and FP bridged toghether[56]. This structure, however, can also be identified in free virions, representing about 3% of all S proteins expressed[58].

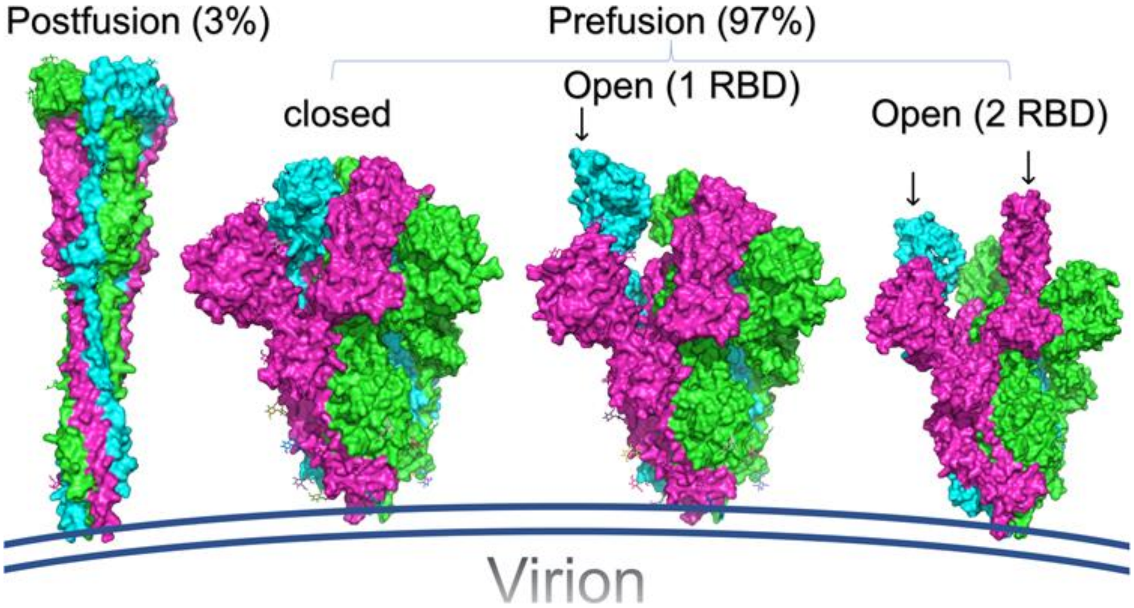

Figure 11. The structures of the prefusion and postfusion structures of SARS-CoV-2 spikes. Arrows indicates the position of open Receptor Binding Domains (RBDs). Adapted from Ismail and Elfiky (2020)[165].

1.1.4.2. Envelope (E) protein

Envelope protein is a small (75 residues) transmembrane protein playing several functions in coronaviruses[52]. It is mainly composed of three alpha helices, the most N-terminal (H1) being the protein's transmembrane domain (TMD). The second (H2) and third (H3) helices are respectively buried in the envelope and exposed to the virion interior (while still being in contact with the membrane) (Fig. 12)[68].

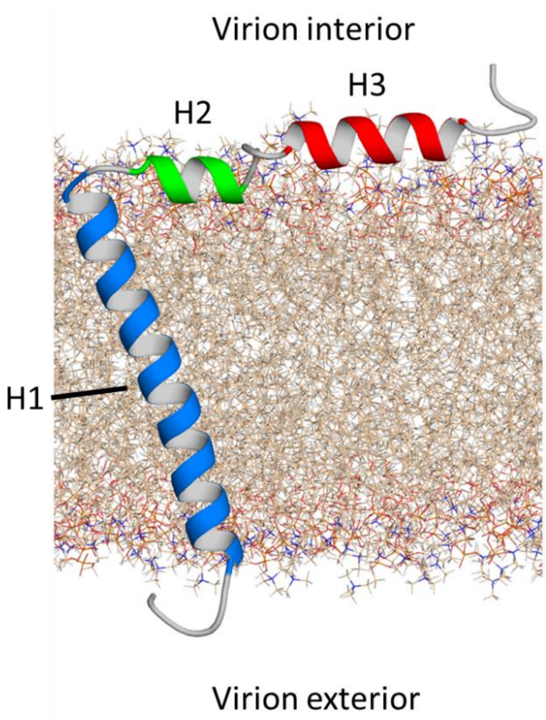

Figure 12. Three-dimensional structure of Envelope protein and organization within virion envelope. H1, H2 and H3 are three predicted alpha-helix regions of the protein. Adapted from Kuzmin et al. (2021)[68].

E proteins cause the curvature of the membranes and thus play a role in virion budding and shaping. Consequently, ty are highly expressed in infected cells, especially on the ER-Golgi intermediate compartment (ERGIC) membranes, the cellular compartment where SARS-CoV-2 assembles and buds[68]. Nonetheless, most of these proteins are not incorporated in the virion, where the expression of the E protein is relatively low[52].

They indeed play a role on host membranes: here, the E protein oligomerizes in a homopentamer, forming an ion channel known as virioporin[52].

1.1.4.3. Membrane protein

The Membrane (M) protein is an integral membrane glycoprotein, the most abundant structural protein in SARS-CoV-2 virion[52]. It is a multi-pass polypeptide with a molecular mass of about 30 kDa and spanning three times virion envelope[52]. The protein's N- and C-terminal regions are located in the virion exterior and interior, respectively[52]. The C-terminal region of the protein folds in a beta-sheet sandwich domain (BD) that, together with the three transmembrane domains (TMs), are essential for M dimerization[69] (Fig. 13).

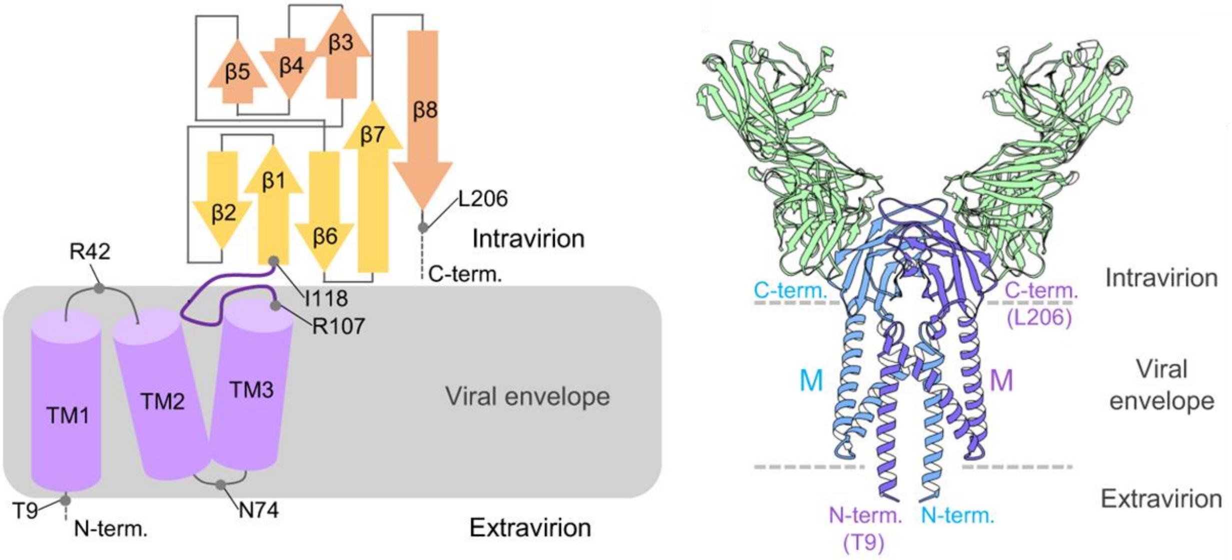

Figure 13. Structural features of SARS-CoV-2 M protein. Left: each M protein is divided in two main regions: the membrane core, composed of 3 transmembrane helices (TM1-3) and the intravirion tail, composed of several beta-sheets structures. Disordered structures are indicated as dashed lines. Left: Cristal structure of M protein dimer and its organization within virion membrane. Modified from Zhang et al. (2022)[69].

M dimers represent the scaffold of coronaviruses envelope and constitute the lattice where the other structural proteins (S and E) are interspersed[52]. Furthermore, M dimers have the capability to bend bio-membranes, inducing the curvature required for viral budding[69]. Its role is so crucial that the sole expression of M and E in human cells determines the production of viral-like particles similar in shape and size to SARS-CoV-2 virions[52].

M protein also interacts with S and N proteins and plays a vital role in the correct assembly of the virion. Indeed, during viral assembly, the M protein recruits the other structural proteins at the level of the membrane[69]. Finally, with its C-terminal region rich in basic residues, the M protein interacts with the N protein and the nucleocapsid in the mature virion.

1.1.4.4. Nucleocapsid (N) protein

N protein is a 419 residues-long protein mainly responsible for SARS-CoV-2 genome packaging[70]. It is divided into 5 regions: the disordered N-terminal domain (NTD), the

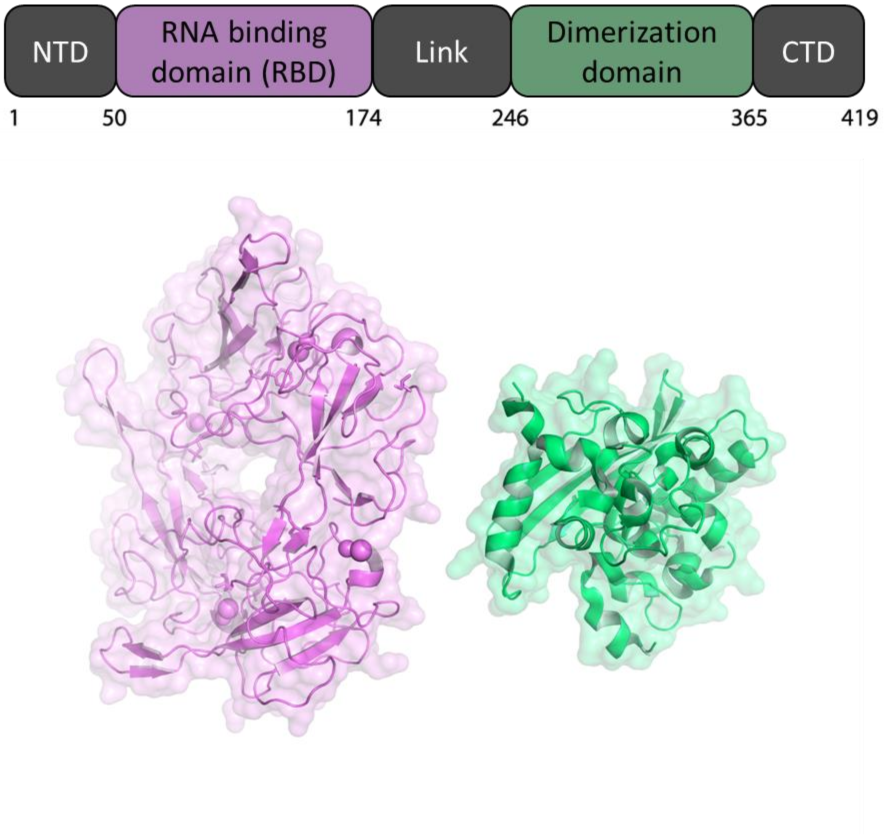

Figure 14. Schematic representation of N protein domains (top) and crystal structures of RNA binding domain (RBD, purple) and dimerization domain (green) (bottom).

RNA binding domain (RBD), a dimerization domain, and a disordered C-terminal domain (CTD) (Fig. 14)[70]. The RBD and the dimerization domain are linked by an intrinsically disordered region (link)[70]. All these regions, including the disordered ones, can bind RNA and are responsible for nucleoprotein-RNA phase separation *in vitro*. Such process, in *vivo*, allows the generation of condensates having a high local concentration of RNAs and proteins[70,71]. Here also, components of viral RNA transcription might be recruited, increasing viral genome transcription and replication efficiency[71]. It has been suggested that this process is the same allowing the packaging of viral genome in the nucleocapsid[70].

1.1.5. SARS-CoV-2 genome

Coronaviruses have the most extended genome amongst RNA viruses, ranging from 26 to 32 kb[72]. For comparison, Influenza A viruses have a genome of about 13.5 kb (considering the sum of all genomic fragments)[73]. SARS-CoV-2 genome is a 30 kb long, non-segmented, single strand RNA molecule. It also has a 5′-cap structure and 3′-poly(A) tail, which allows it to be bound by host ribosomes and translated as soon as it penetrates cytoplasm (it is a positive sense genome)[57]. Viral RNA codes for 29 proteins[74] and can be roughly divided into 4 regions (Fig. 15):

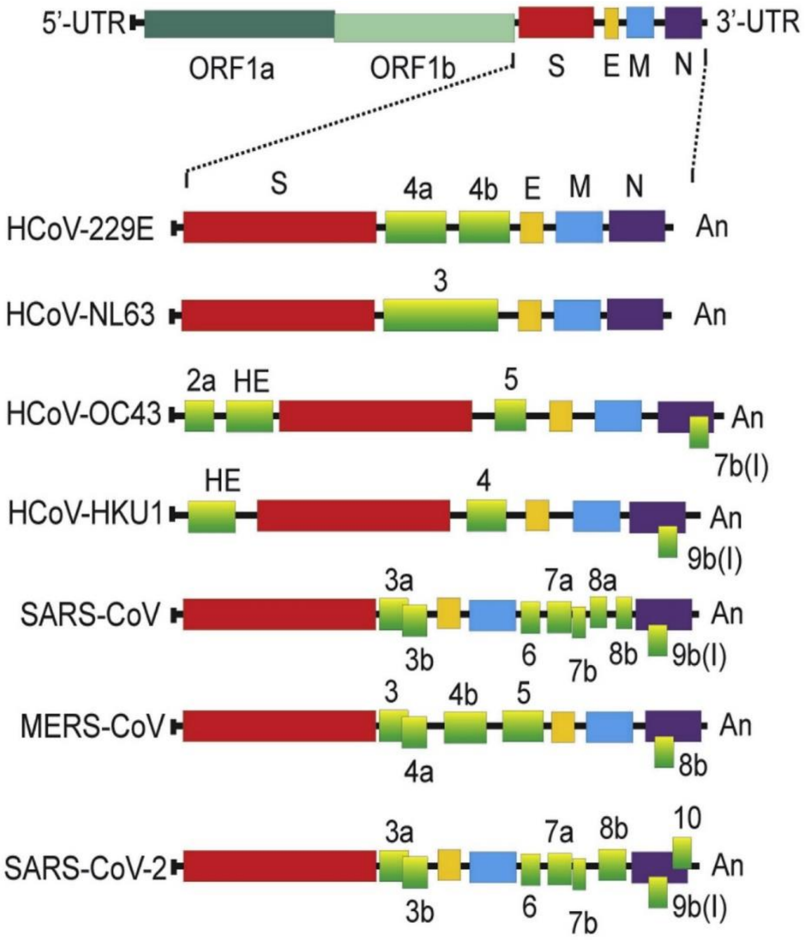

Figure 15. Genetic organization of SARS-CoV-2. A comparison with other human-infecting coronaviruses' genome is shown. Accessory genes are shown in green. Modified from Safiabadi et al. (2021)[28]

5′-end, represents the first 266 nucleotides and is not translated[75]. Instead, it folds in a set of conserved secondary structures (stem and loops, SLs in Fig. 16) with a pivotal role in genome replication, gene expression, and sub-genomic RNAs synthesis[76].

Replicase gene, occupies two-thirds of the viral genome. It contains two long, partially overlapping ORFs named ORF1a and ORF1b. Their translation produces two possible outcomes: the orf1a or orf1ab polyproteins (also referred to as pp1a and pp1ab, respectively), the latter being produced thanks to a -1 ribosomal frameshifting[74]. This process is allowed by a pseudoknot structure and a slippery sequence (5′-UUUAAAC-3′) at the end of ORF1a[75]. Most of the time (up to 75%), the ribosomes, while translating ORF1a, rapidly unwind the pseudoknot structure and release the RNA at the level of ORF1a STOP codon[57,77]. In some cases, however, the pseudoknot structure causes the ribosomes to pause at the level of the slippery sequence. Here the ribosomes slide 1 nucleotide back, moving to ORF1b reading frame, escaping ORF1a STOP codon and continuing to add aminoacidic residues until they reach ORF1b STOP codon[57]. orf1a and orf1ab polyproteins are then fragmented into 11 or 16 polypeptides, respectively. These proteins correspond to viral non-structural proteins 1-16 (nsp1 to nsp16). They do not have a role in virion structure but are functional components required for the viral infection cycle[52,74]. The functions and characteristics of each non-structural protein are summarized in Table 1 and will be further analysed in paragraph **1.1.6**.

Protein	Function	Ref.
nsp1	Inhibits host innate immune response by promoting cellular mRNA degradation and preventing its translation.	78,79
nsp2	Is dispensable in SARS-CoV and MHV. Might suppress micro RNAs processing and bind prohibitins.	80–82
nsp3	Papain-like protease (PL^{Pro}) is required for pp1a and pp1ab cleavage. It is a multidomain protein with several other roles, including ADP-ribose polymerase cytokine expression, induction, and inhibition of host immune response	83,84
nsp4	Transmembrane protein required for replication organelle generation by assembling Double Membrane Vesicles (DMVs).	85
nsp5	Main protease, cleaves pp1ab and pp1a	86,87
nsp6	Zippers the endoplasmic reticulum, playing a pivotal role in SARS-CoV-2 replication organelle generation.	85
nsp7	Interacts with nsp8, acting as processivity clamp for RNA polymerase.	88
nsp8	Part of the processivity clamp of RNA polymerase.	88
nsp9	Part of the replicase complex. Binds the RNA as a dimer and is requires for viral viability.	89
nsp10	Enhances nsp16 and nsp14 activities.	90
nsp12	RNA dependent RNA polymerase.	91
nsp13	RNA helicase and 5′ triphosphatase	92
nsp14	Viral 3′-5′ exonuclease and main factor for RNA proof-reading. Works also as N7 Methyl transferase and adds 5′ cap to viral RNAs	90,93
nsp15	Viral endoribonuclease, targets viral poly(U) tracts enhancing host response escape.	94
nsp16	Metilates viral RNA enhancing host response escape.	95

Table 1. Modified from [57]

Structural and accessory genes region, lies downstream of the replicase region and occupies the remaining 1/3 of the viral genome. Structural genes are a set of nested ORFs known as S, M, E, and N that code for the structural proteins spike, membrane, envelope, and nucleoprotein, respectively[52]. Accessory genes are interspersed among the structural ones and code for 8 different polypeptides (see Fig. 15)[52]. These proteins are not required for the viral infection cycle; nevertheless, they may play a pivotal role in enhancing infection efficiency or immune escaping[57]. Transcriptional regulatory sequences (TRSs) are upstream of each structural and accessory gene. They are believed to assume secondary structures able to regulate the expression of these genes[57].

3′-UTR, Coronaviruses 3′-UTR is an important *cis*-regulatory element for viral replication. It consists of 324 nucleotides and a poly(A) tail[75,76]. Some 3′-UTR secondary structures, but not the primary ones, are believed to be conserved among coronaviruses[76]. They mostly consist of a bulged stem and loop and a pseudoknot structure (PK)[76]. These structures are necessary for viral replication[96]. The region downstream of the PK (known as Hypervariable Region, HVR) varies among coronaviruses for both primary and secondary structures, except for an octa nucleotide 5′-GGA AGA GG-3′ which is instead universally conserved[96] (Fig. 16).

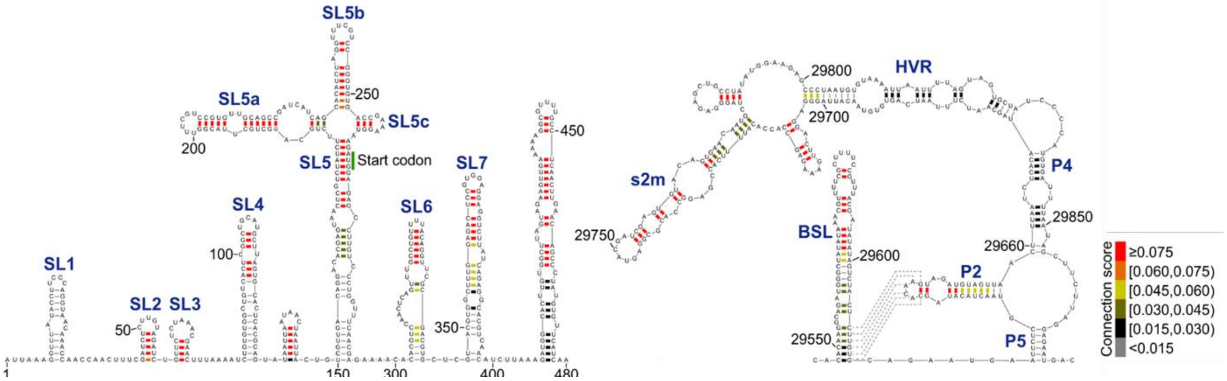

Figure 16. Secondary structures of SARS-CoV-2's first 500 nucleotides and 3′-UTR. Hydrogen bound colours represents base pairing probabilities as connection scores. SL1-7: stem and loops; s2m: stem and loop II-like motif; P1-5: pseudoknot stems; dashed lines: pseudoknot, BSL: bulged stem and loop. Green bar represents the first codon of ORF1ab. Modified from Cao et al. (2020)[75]

It is worth noting that the general gene order of the SARS-CoV-2 genome, 5′UTR-replicase-S-M-E-N-3′UTR, is conserved among all coronaviruses (see Fig. 15).

1.1.6. SARS-CoV-2 pathogenesis

SARS-CoV-2 transmission is mainly mediated by respiratory droplets/aerosol produced by asymptomatic and symptomatic infected individuals[97]. While the airborne and direct transmission route is the best characterized so far, other mechanisms of infection have been proposed, including contact with contaminated surfaces and the oro-fecal transmission[97]. Indeed, once in contact with the host, SARS-CoV-2 infects mainly the upper and lower respiratory tract and the gastrointestinal system[97]. The tropism of the virus is mediated by the expression of its host receptor, ACE2. This integral-membrane enzyme is mainly expressed in the respiratory airways, small and large intestine, and in several other tissues, including the kidney, gallbladder, pancreas, testis, and placenta[98].

Viral entry in host cells requires the binding of viral S protein with ACE2 and event mediated by the RBD on the S1-CTD of the spike protein[56,99]. The spike protein requires the proteolytic cleavage at the S2′ site for its activation[56]. This event is mainly mediated by TMPRSS2 on the host cell cytoplasmic membrane, a serine protease broadly expressed in the respiratory airways, including lungs[98,99]. The cleavage mediated by TMPRSS2 induces the S protein to shift to its post-fusion conformation[59]. The fusion peptide is thus exposed and inserted into the host membrane, destabilizing its architecture and mediating its fusion with the viral envelope[59]. The membrane fusion eventually allows the nucleocapsid to access the host cytoplasm[99].

While being one of the main mechanisms for accessing host cells′ translation machinery, the TMPRSS2-dependent mechanism is not the only one exploited by the virus[99]. If cells do not express high levels of TMPRSS2, or the ACE2-virion complex simply does not encounter this serine-protease, ACE2 binding can induce a clathrin-mediated endocytosis[99]. The virus is thus led into the cells via the endosome. However, the maturation of the endosome to late-endolysosome and its consequent acidification activates the host's membrane protease cathepsin-L[99]. As TMPRSS2, cathepsin-L cleaves the S2′ site, including S activation and fusion of the viral envelope with the endolysosome membrane. This fusion results in SARS-CoV-2 genome release in the cytoplasm[99] (Fig.17).

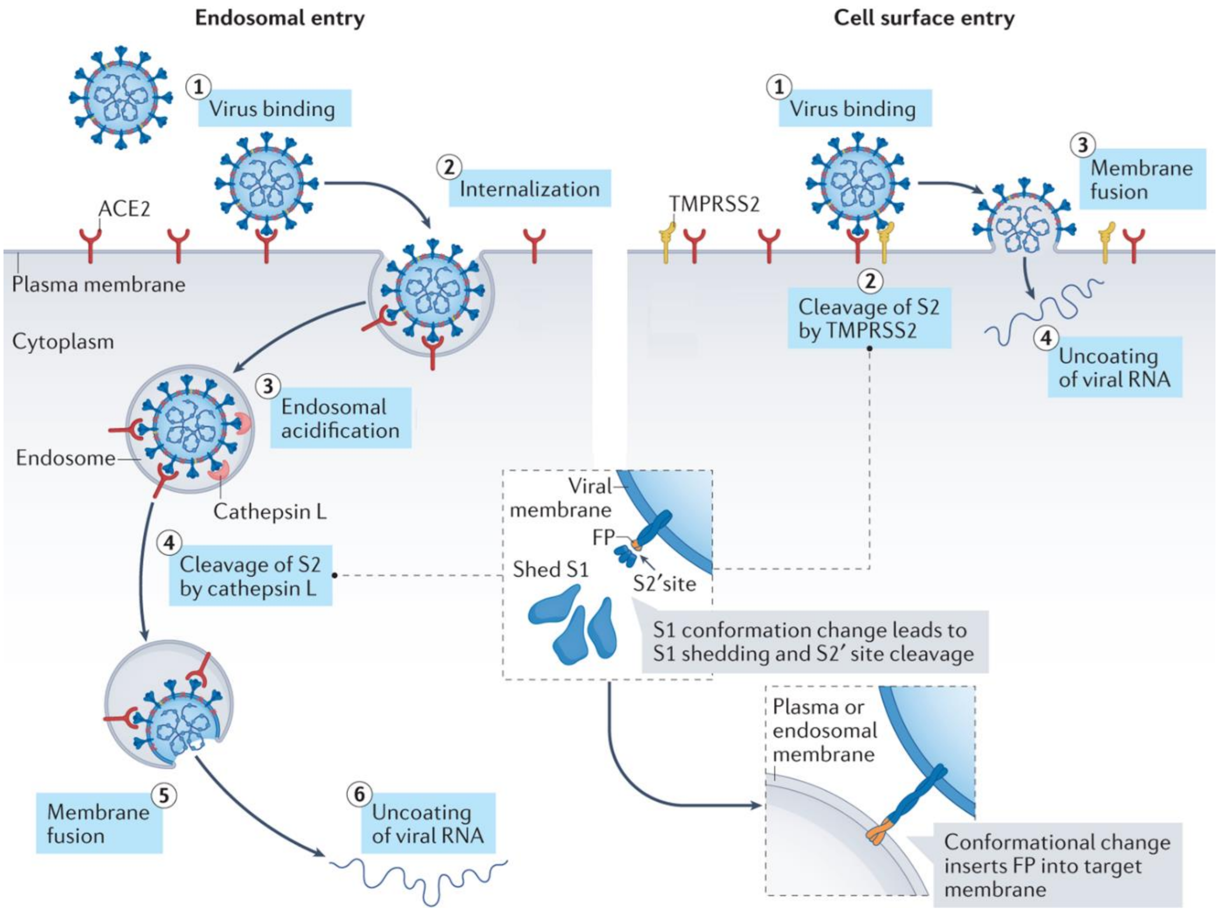

Figure 17. SARS-CoV-2 entry in host cells. SARS-CoV-2 exploits two possible mechanisms for host cell infection. In both the cases, the virus binds its receptor ACE2 (1). If target cell expresses low levels of TMPRSS2 (left) the virion is internalized (2). Subsequent endosomal acidification (3) leads to Cathepsin L activation which cleave the S2′ site on the spike protein. Such cleavage, in turn, activates the spike protein, which leads to the fusion of virion envelope and endosomal membrane (5). Viral RNA is thus released in the cytoplasm and uncoated (6). This process occurs at the level of cell membrane if Spike protein is cleaved by TMPRSS2 (right). Such cleavage is functionally equivalent to the one of Cathepsin L and induce virion envelope fusion with host cell membrane (3) and subsequent viral RNA release in the cytoplasm (4). Adapted from Jackson et al. (2021)[99].

Once entered the cytoplasm, the viral genome is immediately translated. The polyproteins pp1a and pp1ab are thus produced in the ratio 1.4-2.2:1 and co-translationally or post-translationally cleaved in 15 (nsp1-10 and nsp12-16) or 11 (nsp1-11) proteins[77]. The initial cleavage is an autoproteolytic event that releases nsp1 and is mediated by the Papain-Like protease nsp3 (PL^{pro}). This enzyme also mediates its release and the one of nsp2. The main protease nsp5 (M^{pro}), instead, mediates the release of nsp5-16. Finally, nsp4 release requires the cleavage operated by both PL^{pro} and M^{pro}, respectively, acting on nsp4 N- and C-terminus (Fig. 18)[77].

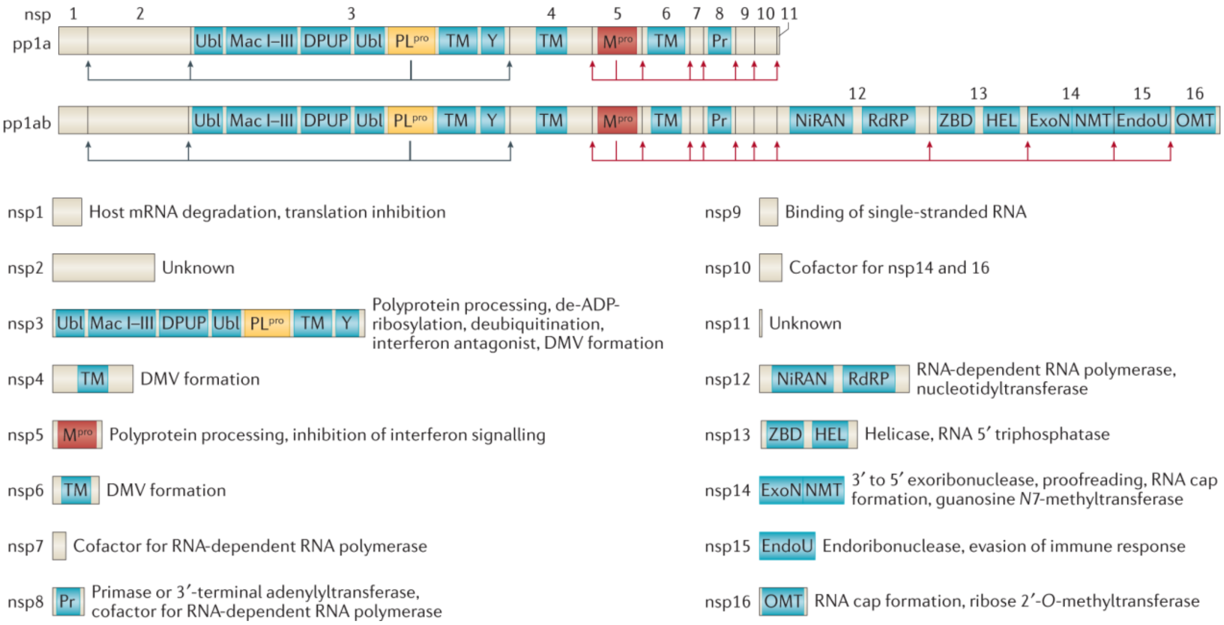
nsp
1
2
3
4
5
6
7
8
9
10
11
pp1a
Ubl
Mac I–III
DPUP
Ubl
PLpro
TM
Y
TM
Mpro
TM
Pr
pp1ab
12
13
14
15
16
NiRAN
RdRP
ZBD
HEL
ExoN
NMT
EndoU
OMT
nsp1 Host mRNA degradation, translation inhibition
nsp2 Unknown
nsp3 Polyprotein processing, de-ADP-ribosylation, deubiquitination, interferon antagonist, DMV formation
nsp4 DMV formation
nsp5 Polyprotein processing, inhibition of interferon signalling
nsp6 DMV formation
nsp7 Cofactor for RNA-dependent RNA polymerase
nsp8 Primase or 3′-terminal adenylyltransferase, cofactor for RNA-dependent RNA polymerase
nsp9 Binding of single-stranded RNA
nsp10 Cofactor for nsp14 and 16
nsp11 Unknown
nsp12 RNA-dependent RNA polymerase, nucleotidyltransferase
nsp13 Helicase, RNA 5′ triphosphatase
nsp14 3′ to 5′ exoribonuclease, proofreading, RNA cap formation, guanosine N7-methyltransferase
nsp15 Endoribonuclease, evasion of immune response
nsp16 RNA cap formation, ribose 2′-O-methyltransferase

Figure 18. Schematic structure of pp1a and pp1ab displaying the main domains and activity of each non-structural protein after its release from the polyprotein. Main proatease (M^{pro}) and papain-like protease (PL^{pro}) cleavage sites are shown as red and black arrows, respectively. DMV, double-membrane vesicle; DPUP, Domain Preceding Ubl2 and PLpro; EndoU, endoribonuclease; ExoN, exoribonuclease; HEL, helicase; Mac I–III, macrodomains 1–3; NiRAN, nidovirus RdRP-associated nucleotidyltransferase; NMT, guanosine N7-methyltransferase; OMT, ribose 2′-O-methyltransferase; Pr, primase or 3′-terminal adenylyl-transferase; RdRP, RNA-dependent RNA polymerase; TM, transmembrane domains; Ubl, ubiquitin-like domain; Y, Y and CoV-Y domain; ZBD, zinc-binding domain. Adapted from V'kovski et al. (2021)[77]

Once released, nsp1 starts to interfere with host mRNAs translation, while nsp3, 4, and 6 operate a complex manipulation of the host's endomembranes to generate the replication organelle (RE)[77]. This structure, typical of coronaviruses, is produced by leveraging ER perinuclear-membranes and consists of a net of interconnected double-membrane vesicles (DMVs)[77,85]. Here, nsp6 is responsible for ER zippering and generates the membranous filaments interconnecting the DMVs[85]. The lumen of these vesicles is proposed to shield viral RNA from host defense factors and to create a confined environment where replicase factors are accumulated, and viral RNA is synthetized[77]. nsp12 works as RNA-dependent RNA polymerase and acts as the main functional subunit of the transcriptase-replicase complex (RTC) together with nsp7-10 and nsp13-16[77].

SARS-CoV-2 replication requires the initial synthesis of a negative-sense copy of the genome. This RNA molecule works as a template for synthesizing new positive-sense genomes. This process is relatively accurate for an RNA virus with a predicted error rate of one every 10^6 bases. This exceptionally high fidelity (up to 100 times higher than in other RNA viruses) is mediated by the proof-reading activity of nsp10 and-nsp14[74].

The positive-sense genome also works as a template to produce sub-genomic RNAs (sgRNAs), smaller RNAs mainly used to synthesize structural and accessory proteins[77]. Their synthesis is mediated by a discontinuous transcription process unique to the *Nidovirales* family. During the negative-sense RNA synthesis, indeed, the RTC stops at the level of transcription regulatory sequences (TRSs), located upstream of most structural genes. The RTC then switches its template and continues the RNA synthesis from a leader region (TRS-L) located in the 5′-UTR of the genome[77]. The mechanism requires the annealing of the TRSs with TRS-L (Fig. 19) and produces a set of negative sense sgRNAs sharing the same leader sequence at their 3′. These are then transcribed in the positive sense, polycistronic sgRNAs which are then translated into one protein (the 5′-most ORF)[77].

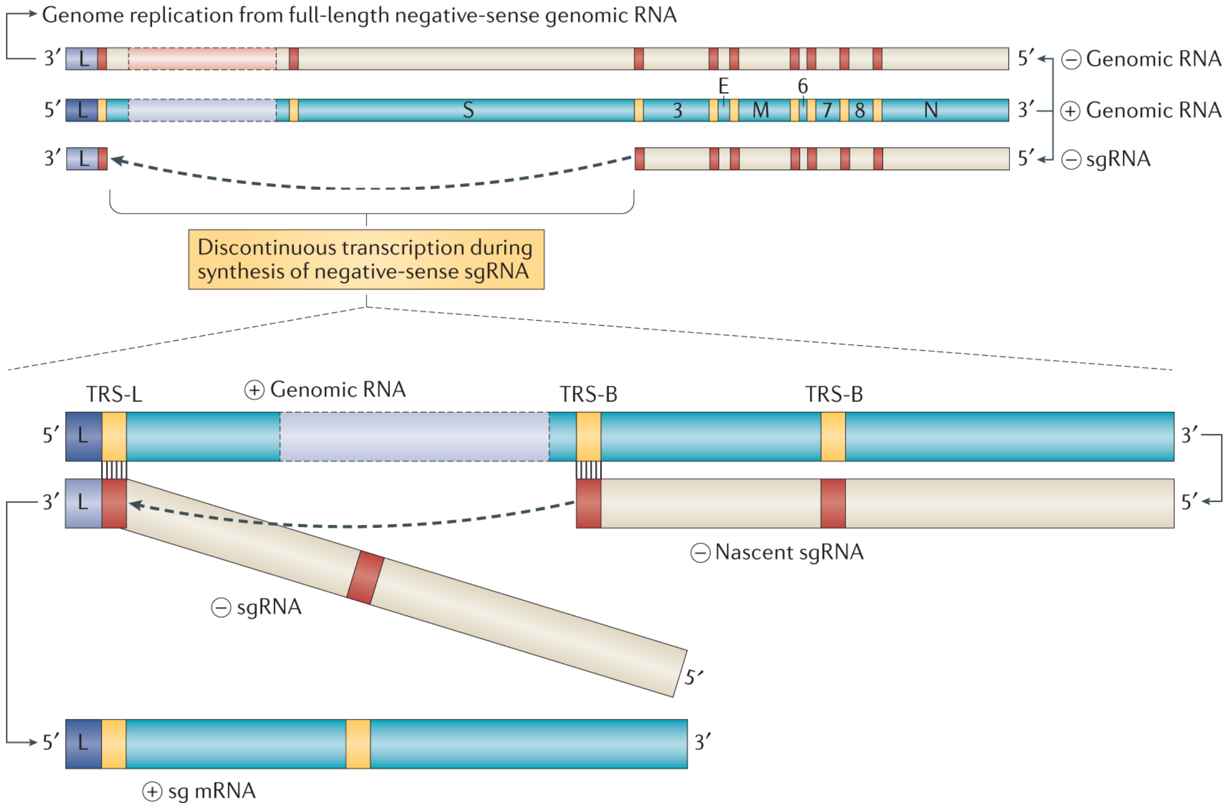

Figure 19. Schematic mechanism of sub-genomic RNA (sgRNA) production in SARS-CoV-2. While the positive-sense genomic RNA can be fully replicated to a negative-sense RNA, discontinuous transcription can also occur. The process is regulated by transcription regulatory sequences located in gene bodies (TRS-B) and at the 3′-end of negative-sense RNA (TRS-L). Modified from V'kovski et al. (2021)[77].

Structural proteins are translated on the surface of rough ER and then translocated towards the ERGIC. Here M proteins recruit the N-wrapped genome, E, and the Spike proteins. M and E induce the invagination of the membrane, and the consequent budding of the newly produced virion into the lumen of the compartment[99,100]. The virion, incapsulated in a host vesicle, is then translocated to the Golgi, where structural proteins are glycosylated, and the S protein cleaved in its S1/S2 site by furin (or a furin-like) protease[99]. The mature virion is finally released outside the cell through the lysosomal exocytotic pathway[100] (Fig. 20).

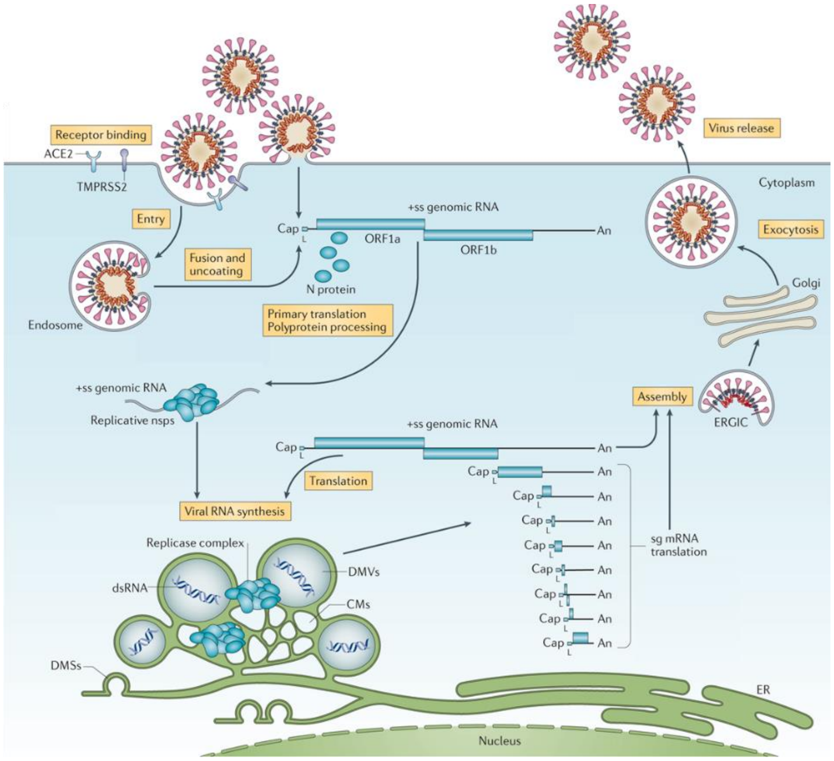

Figure 20. Schematic overview of SARS-CoV-2 infection cycle. Once entered in host cells upon receptor binding, viral RNA is released in the cytosol. Here the genome is immediately translated in a polyprotein that is then processed to produce replicative the non-structural proteins (nsps). These proteins are required for the generation of the replicative organelle, composed of several Double-Membrane Vesicles (DMVs) and deriving from the Endoplasmic Reticulum (ER). Here viral genome is replicated and sub-genomic RNAs produced. The corresponding sub genomic mRNAs (sg mRNAs) are finally translated in the structural proteins which assembly the virion at the level of the ERGIC. Virions are finally released through an exocytosis pathway. Adapted from V'kovski et al. (2021)[77].

1.1.7. SARS-CoV-2 lineages, mutations, and variants of concern.

Over the course of 2 years since its first detection, SARS-CoV-2 acquired several mutations. Current data[101] suggest a mutational rate of about 30 substitutions per year. This value, while being half that of seasonal influenza[101], caused the rise of several SARS-CoV-2 genetic variants[102].

Several systems have been proposed for the definition, classification, and nomenclature of such variants, including the Pango (Phylogenetic Assignment of Named Global Outbreak), GISAID (Global Initiative on Sharing Avian Influenza Data), and Nextstrain systems[103]. The most popular among these is probably the Pango dynamic nomenclature system[104]. It classifies SARS-CoV-2 genetic variants into a hierarchical system composed of *"lineages"*. For a new lineage to be designated, it must satisfy genetic and phylogenetic criteria[102]:

1. The genetic variant should exhibit at least one nucleotide substitution relative to a reference genome (e.g., an ancestral lineage or the first SARS-CoV-2 genome identified).

2. At least 5 viruses belonging to the putative genetic variant should have been fully sequenced (i.e., more than 95% of the genome is known).

3. The putative lineage should show evidence of ongoing transmission. For this reason, the genomes included in the lineage should display at least one shared substitution.

4. Finally, all the genomes of the lineage should cluster in the same phylogenetic group and the lineage-defining node should have a bootstrap value >70%.

The base name for each lineage is given by one or more English alphabet letters (e.g., lineage A, B, P, Q)[102]. Since lineages are grouped in a hierarchical system, each can potentially represent the ancestor for new lineages emerging subsequently. In these cases, numbers and points are added to the lineage base name (e.g., lineages

P.1.1 and P.1.2 are both derived from P.1, which in turn is a sub-lineage of lineage P)[102]. A maximum of three sub-levels are used for each lineage, after which a new base name is used (e.g., the lineage B.1.1.7.2 is called Q.2)[102].

More than 1200 lineages have been identified so far[104]. As described in paragraph 1.1.3, the SARS-CoV-2 pandemic started thanks to two independent zoonotic infections. These events started the spread of the first two SARS-CoV2 lineages, named lineage B (the first to infect humans) and A (the first identified)[102]. The two viruses differ for only two nucleotides, in position 8782 and 28144. Lineage A's nucleotides in those positions are T and C, respectively, while lineage B displays C and T[13]. Since the beginning of the pandemic, these SARS-CoV-2 variants have spread around the globe, and competition for the new host allowed the continuous natural selection of those mutations, causing increased infectivity, virulence, or both[102]. The cornerstone example is the D614G (Aspartate 614 to Glycine) mutation on the Spike protein (Spike D614G) in lineage B viruses (and later in lineage A for convergent evolution)[105]. Although this mutation appeared only in February 2020, its frequency rapidly increased until the substitution fixed in the global population[102,105] (Fig. 21).

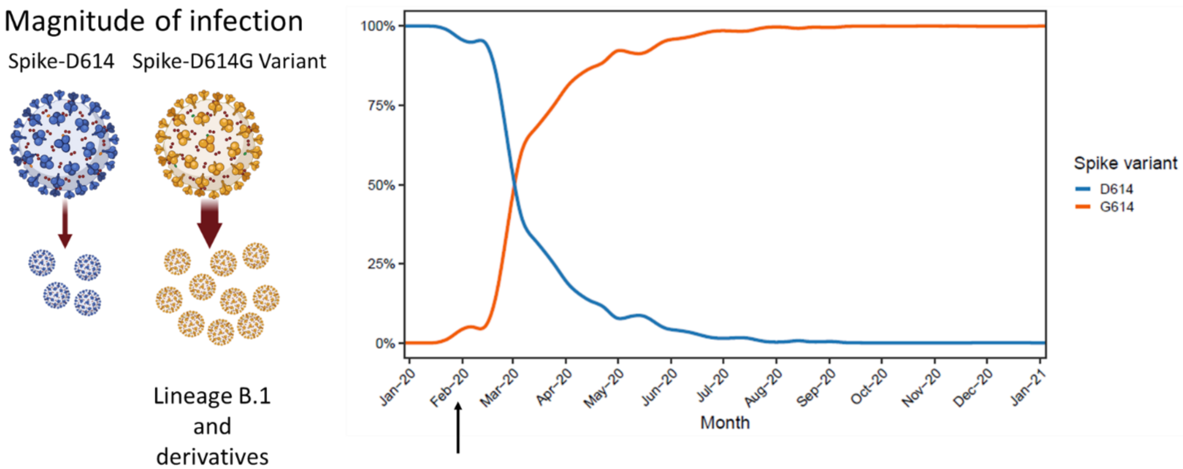

Figure 21. Spike D614G substitution increased SARS-CoV-2 infectivity. D614G substitution caused the arise of the lineage B.1, characterized by higher replicative efficiency and infectivity than the ancestral virus (left). Such features explain the rapid increase in the frequency of the mutation worldwide (right). The black arrow indicates the spike D614G mutation onset. Data from GISAID.

Mechanistically, the D614G substitution abolishes a hydrogen bond between the S1 and S2 subunits of the Spike protein[106]. This, in turn, allosterically alters RBD conformation, which is more prone to shift towards an "up" conformation in the mutant than in the WT genotype. Thus, in the D614G variant, the spike protein is more prone to bind its receptor, which guarantees a competitive advantage over the wild-type (WT) genotype in terms of infectivity[105]. Indeed, it has been shown that when the WT virus and the D614G variant co-infected the same cells, most viral particles produced harboured the Spike D614G mutation[105,107]. This mutation defined the B.1 lineage and guaranteed its initial advantage over A lineages allowing it to spread and diverge into several other lineages.

Starting from the end of 2020, some B.1 sub-lineages caused international concern because of their increased infectivity and virulence[102]. A systematic approach to universally name these variants, especially outside the scientific world, was then needed[103]. At the beginning of 2021, the World Health Organisation and American Centre for Disease Control introduced the concepts of Variant of Interest (VOI), Variant of Concern (VOC) and Variant of High Consequences (VOHC)[108].

- A VOI is defined as a genetic variant identified in multiple countries (or, however, causing multiple COVID-19 cases) and exhibiting phenotypic changes relative to a reference variant[109].
- A VOC is a VOI exhibiting increased transmissibility, virulence, or even changes in COVID-19 symptoms. It can also cause decreased effectiveness of the current measures for containing, diagnosing or treating COVID-19[109].
- A VOHC is a VOC causing more severe disease and increasing hospitalizations, even in vaccinated patients. A VOHC, indeed, can significantly reduce vaccine effectiveness in protecting against the severe disease and the efficacy of current treatments. As of August 2022, no SARS-CoV-2 genetic variant has never been designated as VOHC[110].

The nomenclature for all these variants simply uses Greek letters for both VOCs and VOIs (e.g., the Alpha VOC and the Iota VOI). Table 2 summarizes some of the features of the 6 VOCs identified so far.

Who label	Earliest identification	Become dominant in Europe	Pango lineages
Alpha	Sep-2020	Yes	B.1.1.7/Q
Beta	May-2020	No	B.1.351
Gamma	Nov-2020	No	P.1
Delta	Oct-2020	Yes	B.1.617.2/AY
Omicron	Nov-2021	Yes	B.1.1.529/BA

Table 2. Modified from [57]

The first VOC identified was the Alpha variant (lineage B.1.1.7), harbouring, in addition to D614G, several other mutations on the spike protein[111]. Among these, the N501Y substitution fell in the RBD and was shown to increase the affinity of the spike protein to its receptor ACE2[112]. This feature significantly increased the infectivity and fitness of the variant. Indeed, in winter 2021, the B.1.1.7 lineage became the variant accounting for about 90% of all the infections[113].

The Beta (lineage B.1.351) and Gamma (P.1) variants were identified shortly after the Alpha and were responsible for outbreaks in South Africa (Beta) and Brazil (Gamma)[114]. These variants shared a set of mutations in the spike RBD, including E484K, N501Y, and K417N/T (in the Beta and Gamma variants, respectively)[114]. Interestingly, this set of mutations, as well as the Spike E484K substitution alone, conferred to these lineages a lower sensibility to monoclonal antibodies and to natural and vaccine-induced human sera[114–116].

All these variants, however, were outcompeted by the Delta variant (lineage B.1.617.2 and sub-lineages, identified as AY). First identified in India in October 2020, this variant rapidly spread worldwide, being responsible for almost 100% of infections starting from September 2021[113,114].

Finally, the Omicron variant (B.1.1.529 and its sub-lineages indicated as BA and corresponding to the so-called Omicron 2 and Omicron 5) raised at the end of 2021, probably in South Africa, and rapidly become the dominant variant. At the time of writing, August 2022, the Omicron variant is the unique VOC circulating, responsible for the totality of SARS-CoV-2 cases in the world[114] (Fig. 22).

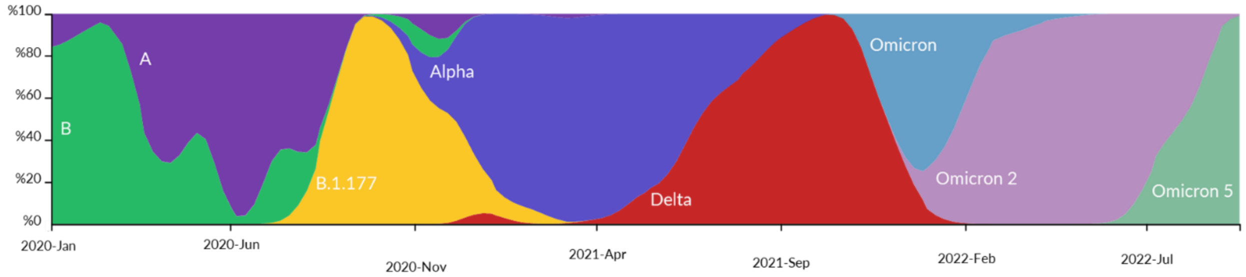

Figure 22. Worldwide relative frequencies of the main SARS-CoV-2 lineages and Variant of Concern identified so far (last update: end of August 2022). The data for this graph have been obtained from Nextstrain[101].

SARS-CoV-2 genome is continuously shaped under the force of selective pressure that, several times during the pandemic, promoted variants with advantageous features over all others. Unfortunately, such advantages translated into higher of the virus[114] highlighting the need to not only correctly diagnose and treat SARS-CoV-2 infections but also monitor the genetic variability of the virus, especially now that new selective stimuli, such as vaccines and antibodies treatments, have been introduced in the environment. An effective strategy to reach this goal is the so-called "genomic surveillance", an approach leveraging the power of Next Generation Sequencing technologies.

1.2. Next generation sequencing for pathogens genomic surveillance

As discussed in the previous paragraph, identifying new SARS-CoV-2 lineages requires disclosing the global information contained in viral genetic material. Such a goal can be accomplished by leveraging the new DNA sequencing technologies. These approaches are generally divided into three main categories: first, second and third generation sequencing[117].

1.2.1. First-generation DNA sequencing

First-generation sequencing approaches represented the first technologies to read the genetic information held in nucleic acids. Currently, the term refers to the methodology developed by Frederik Sanger in 1977 ("Sanger sequencing") and subsequently optimized and automatized. It is based on the so-called "chain termination" by fluorescent-labeled 3'-dideoxynucleotides (DDNs)[117]: the DNA, indeed, is replicated *in vitro* by using a mix of common nucleotides and DDNs. These modified nucleotides miss the 3'-OH group required by DNA polymerase to catalyse the synthesis of the phosphodiester bound. Thus, when the DNA polymerase adds a DDN to the nascent chain, the reaction is stopped. The earlier the DDN is added to the chain, the shorter it will be. In addition, since each of the four DDNs is labelled with a different fluorophore, the fluorescence emission of the terminated chain informs on the nature of the last nucleotide (i.e., the DDN) added to it. The products are run on a capillary electrophoresis system at the end of the reaction. Such technology has the capability of resolving DNA strand lengths differing for a single base. Finally, an automatized system records both the fluorescence and the length of each fragment, reconstructing the nucleotide sequence of the original DNA molecule[117].

While highly reliable, this approach requires that the input DNA is homogeneous in sequence. The higher the complexity (i.e., the number of molecules with different sequence) the trickier it is to reconstruct the template sequence. In addition, first-generation sequencing approaches have limited throughput, allowing to sequence, on average, a maximum of one hundred samples per run[117].

1.2.2. Second generation sequencing

At the beginning of the 21st century, the advancements in computational power, optic resolution, and microfabrication led to the development of a sequencing approach able to address the drawbacks of first-generation sequencing[118]. This new technology is now known as Next Generation Sequencing (NGS) and represents the second-generation of sequencing approaches[118]. It allows to massively parallel each sequencing reaction, with hundreds of samples processed at the same time[117,118]. In addition, NGS allows to digitalize the information obtained from the sequencer to the extent that each input DNA molecule generates an individual digital sequence (read) that can be independently analysed[118]. This way, a heterogenous mixture of DNA sequences can be analysed in the same run, since the signal originating from each template can be digitally separated from all the others. Such results are usually obtained thanks to two NGS-specific features:

1. All samples, prior to sequencing, are processed to produce a collection of ready-to-sequence DNA fragments known as a "sequencing library" (or simply "library"). Libraries generation often requires several steps, including the addition of DNA sequences (known as adapters) to the ends of each DNA input molecule. Among other things, the adapters contain a variable sequence that differs in different samples and acts as a molecular barcode (known as barcode or index)[118]. After sequencing, each read will be associated with a specific barcode and, thus, with the sample it originated from. It is thus safe to mix (to "multiplex") several hundreds of samples in the same sequencing run[118].

2. During sequencing, DNA molecules are segregated into wells or specific positions of functionalized surfaces able to accommodate individual input

molecules. The molecules are then *in situ* amplified and/or sequenced. The signal coming from different wells/positions is thus registered in a digital file and assigned to different reads[118].

NGS approaches have been declined in several ways since their first introduction[117,118]. Nevertheless, the most adopted technology nowadays is the one developed initially by Solexa and then acquired by Illumina[117]. In Illumina sequencing, DNA libraries are denatured to single strands and loaded on a functionalized surface known as a flow cell. Small oligonucleotides are bound to the flow cell surface and are complementary to libraries' adapters. The input DNA thus binds to the oligos on the surface of the flow cell. Such oligos are extended using the input DNA as template, producing complementary copies of the libraries covalently bound to the flow cell. The flow cell-bound molecules, in turn, bend and act as a template for neighbour oligos. The process, iterated several times, is called bridge amplification and generates clusters of *in situ*, clonally amplified DNA molecules. Each cluster generates the signal required for a single read. Current Illumina flow cells are able to generate up to several billion clusters (i.e., reads)[118].

The DNA in each cluster is sequenced during the synthesis of a new strand thanks to modified nucleotides, a process known as Sequencing by Synthesis (SBS)[117]. The nucleotides are modified with a blocking group masking their 3'-OH group and a base-specific fluorophore. The block is reversible and can be removed during sequencing[118]. The overall process can be divided into the following steps:

1. A primer is annealed to the flow cell-bound DNA molecules.

2. The DNA polymerase extends the primer by adding a modified nucleotide. Since the nucleotides have their 3'-OH group blocked, only one nucleotide will be added.

3. Unbound nucleotides are washed away, and a CCD camera records the fluorescence deriving from the incorporated nucleotides in each DNA cluster. Since each base is associated with a different fluorophore, the signal read by the CCD camera is associated with a specific nucleotide.
4. Finally, the fluorophore and the blocking are cleaved from the last incorporated nucleotide, unmasking its 3'-OH group[118] and allowing the process to be repeated.

With this process, it is possible to sequence up to 500-700 bases and to read both ends of each DNA sequenced, a process known as Paired-end sequencing (PE)[119].

A technology similar to Illumina has been recently developed by MGI and is called DNA Nanoball Sequencing, DNB Seq[120]. While still a second-generation approach, it does not rely on any PCR during sequencing. DNB Seq, instead, uses the Rolling Circle Replication (RCR) to produce giant DNA concatemers that are functionally equivalent to Illumina's DNA clusters[120]. Briefly, DNA libraries are first converted to single-strand, circular DNA molecules (sscDNA). Then, these molecules are linearly amplified with RCR, a process similar to the one used by mitochondria to replicate their genome. It exploits the activities of a φ29 phage's DNA polymerase (φ29 polymerase) that binds to sscDNA and catalyses the synthesis of a complementary strand[120]. However, at the end of the first replication cycle, the enzyme does dissociate from the template and starts a new synthesis displacing the newly generated DNA strand[121]. The result is a concatemer containing 300-500 tandem copies of the original DNA that folds in a 300 nm globular structure, the DNB (Fig. 23)[120].

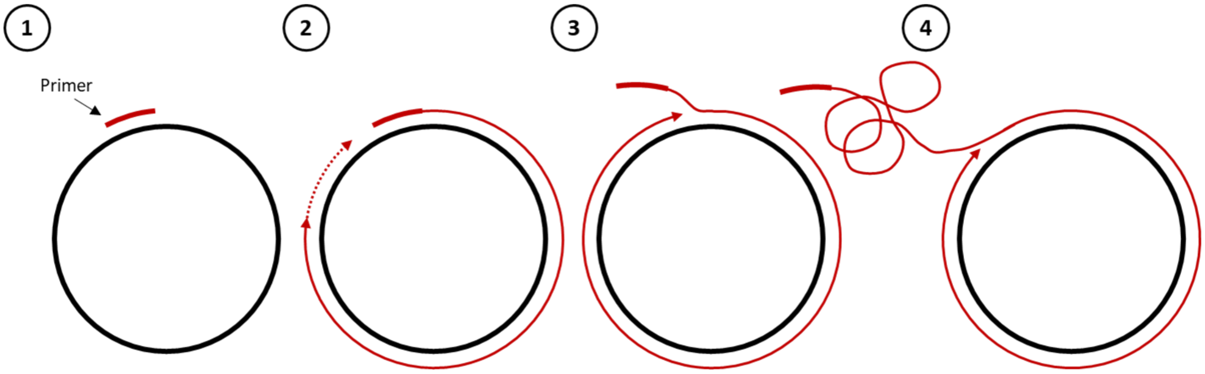

Figure 23. Schematic overview of DNA nanoballs generation by rolling circle replication. A single strand DNA (black circle) is annealed to a primer, indicated by the black arrow (1). A DNA polymerase catalyses the extension of such a primer (2), and the displacement of the previously synthesized DNA strand while performing a new round of synthesis (3). This process is repeated several times until a DNA concatemer is obtained (4).

The main advantage of RCR over PCR is that φ29 polymerase always uses the very same DNA template for its amplification. This means that any potential mutation introduced by the enzyme will not be amplified since no new DNA strand is used template[121]. Furthermore, φ29 polymerase is extremely processive and robust, capable of synthesizing up to 70 kb[121] with an error rate as low as 1 mismatch in 10^7 nucleotides condensated[122].

DNBs are rich in negative charges on their surfaces, a feature that prevents DNBs from aggregating and allows their binding on the surface of the sequencing flow cell. DNB Seq flow cells, indeed, are composed of individual micro-spots, 300 nm in diameter, and covered with polyamines. The positive charge and the size of each spot allow the binding of only one DNB, which, in turn, produces the signal for one read. Once bound to the flow cell, all tandem copies in the DNB are sequenced simultaneously using an SBS approach with fluorescent-labeled nucleotides[120].

DNB sequencing is generally cheaper than Illumina and produces data that are qualitatively comparable[123–126]

NGS revolutionized the concept of DNA sequencing, allowing the beginning of the "genomics era". The possibility to multiplex several hundreds of samples in the same sequencing run significantly reduces the sequencing costs per sample. Nevertheless, these technologies still require expensive sequencers and are bound to DNA pre-amplification before their sequencing. Finally, NGS approaches produce small reads, impacting the capability to reconstruct long and complex (e.g., rich in repeats) genomes[127].

1.2.3. Third generation sequencing

Third-generation sequencing (TGS) approaches are usually defined as technologies able to sequence single molecules without requiring any kind of DNA amplification[118]. In addition, these approaches generate 20 kb-long reads, far longer than the one obtained with NGS[117]. The leading technologies adopted are: Single Molecule Real Time sequencing (SMRT sequencing) developed by Pacific Bioscience (PacBio) and Nanopore sequencing by Oxford Nanopore Technologies (ONT)[117].

- SMRT sequencing adopts modified flow cells (SMRT flow cells) composed of nano well arrays. Each nano well can accommodate only one DNA molecule. A DNA polymerase is immobilized at the bottom of each nano well and replicates a single DNA strand *in situ* by using fluorescent-labelled nucleotides. The hole of each nano well has a diameter smaller than the wavelength of the light used for fluorophores excitation. Such feature causes the detector to only record the fluorescence coming from the very bottom of the nano well where the DNA polymerase is located and where the last nucleotide have just been incorporated. The light pulses coming out from each nano well are continuously recorded by an ultrahigh-resolution camera, and their interpretation is performed in real time[117].

- Nanopore technology does not rely on DNA synthesis for sequencing. Single-strand DNA travels instead through engineered bacterial nanopores inserted in artificial membranes. Sensors monitor changes in electrical current on the membrane while the DNA translocates through the protein nanopore. Each DNA base, indeed, disrupts the electric current in a slightly different manner, thus allowing to reconstruct sample sequence in real time while translocating. In addition, each Nanopore flow cell comprises thousands of nanopores, each producing a sequence of single DNA molecules. Finally, this technology is capable to also detect modified DNA bases or to even directly sequence RNA molecules[117].

The rapid evolution of sequencing approaches in last two decades boosted the development of new approaches to old issues, including studying pathogens' whole genomes to counteract their spread[128].

1.2.4. Genomic surveillance of pathogens' genome

Genomic surveillance leverages the power of genomics to unveil the dynamics behind pathogens origins, spread, and virulence. Such new approach to study pathogens has been boosted by the advent of NGS and the subsequent decrease in costs, workaround time, and complexity associated with genomics analysis[129]. Monitoring pathogens genetic variability holds the power to tackle their rapidly evolving nature, allowing the identification of new lineages, phenotypes, and resistance mechanisms to medical intervention as soon as they appear[129]. In addition, the direct comparison of pathogen genomes isolated from different patients is a powerful tool for performing contact tracing of infections. Using such a strategy, S. Mate et al., demonstrated that Ebola virus RNA persists in human semen for several months after healing, but also that such semen can mediate human-to-human sexual transmission of disease[130]. A systematic implementation of a genomic surveillance program also allowed the complete reconstruction of the 2013-2016 Western Africa Ebola outbreak dynamics, unveiling the origin of the epidemic, how it spread in Western Africa countries and what were the factors sustaining viral transmission[131,132]. Similarly, by applying a TGS approach, Giovanetti, Faria et al., demonstrated that the 2015-2017 Zika virus outbreak in the Americas probably originated in northeast Brazil in 2014 and that from there, it started to spread to nearby regions and countries[133]. Using a similar approach, it was also possible to recapitulate the routh of infection that allowed the virus to spread to the Amazon region, Brazil, and USA[134,135].

A key advantage of genomic surveillance approaches relies on the intrinsically digital nature of gnome sequences. The possibility to efficiently collect and rapidly share such information has proven critical for world-threatening diseases, such as Influenza and COVID-19[113]. Among the several databases collecting genomic data, the Global Initiative on Sharing Avian Influenza Data (GISAID) is particularly important. It was developed in 2015 to collect and timely share Influenza H5N1 data, but rapidly evolved and adapted for the analysis of both seasonal Influenza and SARS-CoV-2 sequences[136]. At the time of writing (end of August 2022), GISAID collects 356,990 influenza virus sequences and has become a reference point for studying this virus[113].

With over 13 million SARS-CoV-2 sequences uploaded on GISAID, the study of viral genome variability has proven its power during COVID-19 pandemic, emerging as a critical tool for the containment and mitigation of virus spread[137]. Indeed, the implementation of NGS and TGS methodologies had a pivotal role in the discovery of SARS-CoV-2[22], the tracking of genetic variants[105,111,138–141], and the development of new vaccines against Omicron subvariants[142,143].

1.3. Aim of the Thesis

While being one of the first countries to be hit by the pandemic, an alarmingly low number of SARS-CoV-2 sequences were produced in Italy in 2020, especially from the south of the country[113]. These data contrasted with the results obtained by the United Kingdom in the same period, which allowed the discovery of the Alpha VOC (Fig. 24)[111].

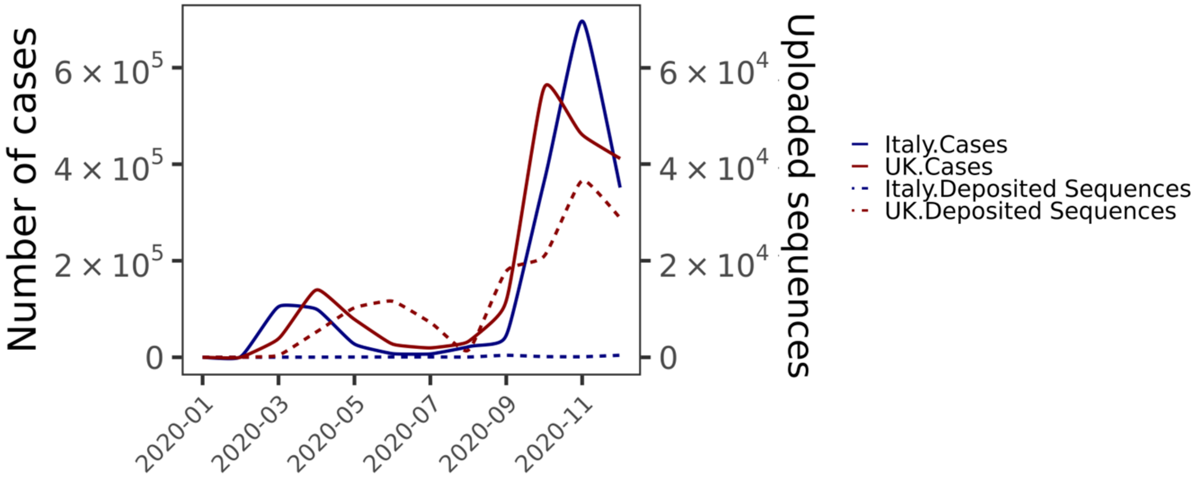

Figure 24. Time distribution of infection (continuous lines) in Italy (blue) and United Kingdom (UK, red). The number of sequencies produced and deposited on GISAID are shown as a dashed blue (sequences originated in Italy) and red line (sequences originated in UK).

Along with genotyping of SARS-CoV-2, it is becoming relevant to acquire insights into the cellular response of the host cells upon infection to monitor disease progression and potentially identify new biomarkers. Currently, most studies concerning the elucidation of the molecular bases of viral infection have been carried out *in vitro* virus-infected models[144–146]. While these approaches promise an easy-to-handle and rapid solution, they lack generalization, as they do not account for the physiological interaction between infected host cells and their microenvironment. However, due to the exiguous quality of nasal swab RNA, combined with the high costs of sampling a large cohort of COVID-19 positive patients, it has been challenging to obtain consistent patient gene expression data so far.

Now more than ever, providing simple and cost-effective tools to comprehensively profile both virus genome and host transcriptome is imperative to effectively track and isolate focal areas of variants eluding vaccination. To fill this gap, we propose an integrated genomic framework allowing the reconstruction of the SARS-CoV-2 complete genome and obtaining global host gene expression signature upon infection from the same diagnostic extract. We developed a customized and affordable amplicon-based approach that can be potentially implemented in laboratories equipped with benchtop sequencers. As proof of principle, we applied such workflow to perform an efficient surveillance of SARS-CoV-2 genome in Campania region, the most densely inhabited region in the South of Italy. Our effort allowed us to establish novel methods to identify emerging variants and to recognize molecular signatures associated with coronavirus infection, thus leading to a powerful tool for disease prevention, diagnosis, and, potentially, personalized treatment.

2. Materials and Methods

2.1. Samples collection, RNA extraction, and SARS-CoV-2 testing

Sample handling, diagnostics, and logistics were carried out by Ospedale Cotugno as the regional reference centre for infectious diseases and Istituto Zooprofilattico Sperimentale del Mezzogiorno (IZSM), as coordinator of the Coronet network of Regione Campania. All samples were randomly collected in Campania, Italy, as part of the institute's diagnostic activity during 2020 and 2021. In most cases, after a first diagnosis, a second RNA extraction and qPCR was performed by IZSM to generate uniform qPCR results. RNA extraction was performed by using either the Maelstrom 9600 (TANBead), GeneQuality X120 (AbAnalitica) or Abbott m2000sp automatic platforms according to manufacturer's specifications. SARS-CoV-2 abundance in each sample was tested by using either the Allplex 2019-nCoV Assay (Seegene), Real Quality RQ-2019-nCoV kit (AbAnalitica) or Abbott RealTime SARS-CoV-2 Amplification Kit by detecting at least two of the N, E or RdRP, SARS-CoV-2 genes. In all analyses where Ct value was employed, the average Ct of the three genes was calculated and used.

A total of 22228 were used for SARS-CoV-2 whole genome sequencing. Out of these, 387 samples were used to investigate host gene expression and were divided into two cohorts depending on the viral variant identified: the first cohort included 162 samples assigned to the B.1.x variant, and the second included 225 samples assigned to the Delta variant. In addition, 300 RNA extracts from SARS-CoV-2 negative swabs were also sequenced.

2.2. SARS-CoV-2 sequencing and computational analysis

2.2.1. Library generation and sequencing for SARS-CoV-2 genotyping

The procedure we developed and optimized for SARS-CoV-2 genome analysis is based on the concept of "amplicon sequencing". Briefly, PCRs are conducted to specifically enrich a set of genomic loci, represented in this case by the SARS-CoV-2 genome. Such continuous amplification of the same loci is known to increase the odds of environment and sample cross-contamination[147]. We thus adopted a strict protocol to minimize the chances of contamination. All procedures for library generation were performed with standard filtered low-retention tips, and each step was performed in separate PCR hoods located in different rooms with dedicated pipettes and thermocyclers. Dividing the pre- and post-amplification steps into distinct areas is a known strategy to reduce the risk that libraries prepared in a previous round may contaminate raw samples[148]. In addition, before and after each step, all surfaces were decontamination by using a combination of UV irradiation, 0.5% bleach, and DNAzap (Thermofisher). Finally, our protocol also implemented two steps of Uracil DNA Glycosylase (UDG)-mediated decontamination[147]. Such enzyme, added to the samples before each PCR step, specifically degrades U-containing DNA molecules. However, uridylate residues are incorporated only during the last amplification by adding UTP to the nucleotide mixture. Thus, prior to each amplification, UDG specifically degrades any amplicon carry-over from previous rounds of library preparation.

Libraries were always prepared in multiples of 96 samples arrayed in a 96-well plate, and at least 5 blank samples (water) were added to each plate to monitor cross-contaminations. Library generation for SARS-CoV-2 genome sequencing was performed using a modified and optimized version of the amplicon-based ATOPlex RNA Library Prep kit (MGI Tech) starting from 5, 2,5, and 1,25 uL of unquantified extracted RNA. In addition, the volume of reagents was reduced to ½, ¼, and ⅛, of the recommended volumes, respectively. Such strategy is based on the idea that reagents final concentration is not affected by proportionally scaling reaction volumes and does not influence enzymatic activities.

The sequencing strategy was also optimized to increase the sequencing throughput from 96 libraries per run to 384 by manually loading the 4 flow-cell lanes of a PE 100 cycles 320G flow-cell (MGI Tech). Increasing the number of samples loaded in each flow cell is expected to decrease the number of reads assigned to each sample. Nevertheless, it is possible to predict the number of reads assigned to each sample, *R*, by using the formula:

$$R = \frac{O}{N}$$

With *O* being the maximum throughput of the flow cell ($1.8 \cdot 10^9$ paired-end reads) and *N* being the number of samples loaded. Thus loading 384 samples yields $R{=}4.7{\cdot}10^6$ paired-end reads. However, this *R* value is enough to reach a 10^4X coverage, as can be computed with the following formula:

$$X = \frac{2 \cdot R \cdot L}{S} = \frac{2 \cdot 4.7 \cdot 10^6\ reads\ \cdot 100\ bases/read}{3 \cdot 10^5\ bases} = \frac{9.4\ \cdot 10^8}{3 \cdot 10^5} = 3.13 \cdot 10^4$$

With *X* being the coverage, *L* the length of each read and *S* the length of SARS-CoV-2 genome. Factor 2 is added to convert paired-end reads to individual reads. These numbers suggest that further multiplexing can be achieved by increasing the indexing up to 768 libraries per run, as shown by randomly subsampling 1,25 million reads per sample. Similarly, two 96 library pools can be sequenced on two lanes of a PE100 cycles SP/S1 Novaseq flow-cell (Illumina). The two sequencing technologies show comparable performances[124–126].

One-step tests were performed by merging the 1st and 2nd PCR step of the ATOPlex RNA Library Prep kit. We prepared a PCR reaction mixture containing all the components of the 1st PCR step plus the “PCR Primer block” and the “PCR additive” of the 2nd PCR. The PCR was then conducted using the program suggested in the original protocol[149]. The number of cycles was increased from 13 to 25 to decrease the number of unincorporated primers at the end of the amplification. For the same reason, the concentration of the "PCR Primer Pool'' component was decreased to 1/75 of the original one. As soon as the reaction cooled down to 4 °C, the indexing primers were added, and the reaction was allowed to continue for 15 cycles. All the reagents, except the PCR Primer Pool'' were used at the same concentration as suggested by the original user manual[149].

2.2.2. Data analysis for SARS-CoV-2 genome reconstruction

FASTQ files generated by the MGI sequencer (DNBSEQ-G400) were used as input for the pre-processing pipeline. The pipeline was adapted from MGI-tech-bioinformatics[150] , and a threshold coverage of at least 30X was used to call each base in the consensus sequence. It was further parallelized and automated to process 100 samples/h using Nextflow[151]. SARS-CoV-2 viral load was implied as the percentage of reads aligning to the viral genome with respect to the co-amplified Lambda phage genome added as spike-in at the beginning of the library preparation. A co-amplified host GAPDH locus was used in the pipeline for internal positive control. Only samples with a minimum SARS-CoV-2/Lambda reads ratio of 10%, 50000 SARS-CoV-2 reads, and at least 50% of genotyped bases were considered for GISAID submission. Blank samples generally displayed around 1% SARS-CoV-2/Lambda reads ratio and almost never exceeded 10%. Upon GISAID submission, only samples labelled as complete by GISAID and with >95% of genotyped bases were used for further analysis. Furthermore, to normalize sequencing statistics when comparing the three solutions developed, only samples with Ct values lower than or equal to 33 were selected.

The phylogenetic analysis was generated using the Nextstrain[101] standard pipeline on a random subsample of sequences generated until 2022-03-30. Tree visualization and manipulation was performed using R (v. 4.1.0) with the packages ape (v. 5.5), ggTree (v. 3.0.2), phangorn (v. 2.7.1), castor (v. 1.6.8). Since our original dataset contained few

Omicron VOC samples, BA.1.21.1 tree was generated by using the omicron sequences produced and a random sample of sequences from GISAID assigned to other lineages (GISAID epi set: EPI_SET_20220509ow).

2.2.3. Discovering new SARS-CoV-2 variants

To identify new lineages, we first manually explored the mutation distribution of concern in our dataset. A mutation of concern was labelled as "expected" if its frequency in a specific lineage was higher than 30% over the total number of world-wide samples assigned to that specific lineage. Otherwise, it was identified as unexpected.

Another approach we adopted was the automatic analysis of mutation trends. The overall idea underlying this approach is that a unique combination of mutations defines each SARS-CoV-2 lineage. Thus, if a lineage is growing in frequency in a given timeframe, so would also the frequency of its lineage-defining mutations. Such substitutions should therefore show remarkably similar temporal trends. To automatize this kind of analysis, we chose to simply group (i.e., cluster) mutations based on their trends and using PAM clustering method as follows. The input for the algorithm was a mutation x month matrix indicating the frequency of each mutation in each month. Out of all the mutations detected, only those reaching 5% of incidence at least once during the analysis period were used for clustering. The number of clusters was chosen using the silhouette method (factoextra v. 1.0.7). This yielded to 3 optimal clusters. A further round of clustering on the first two clusters (k=28 and k=3, chosen with the silhouette method) resulted in a total of 32 groups. The same number of clusters was chosen for both the analysis performed in May 2021 and January 2022. Finally, clusters too similar were manually merged.

All SARS-CoV-2 sequencing data are available through the GISAID database. All analyses were performed by exporting sample metadata and PANGO lineage from GISAID (at the date of 2022-03-22, GISAID accession number: EPI_SET_20220718pm)), by including only full genomes and excluding those at low coverage as described above. RNA-seq gene expression data are available at GEO Datasets (GSE184610).

The data analysis pipeline is freely available for non-commercial use upon the signature of an institutional MTA at: https://gitlab.com/nextgd/ngdx-atoplex-panel-covid-19-pipeline.

2.3. Host mRNA-seq and computational analysis

RNA-seq was performed by using the 3'DGE mRNA-seq clinical grade sequencing service (Next Generation Diagnostic srl)[152] which included library preparation, quality assessment, and sequencing on a NovaSeq 6000 sequencing system using a single-end, 100 cycle strategy (Illumina Inc.). Before library preparation, a 40-60uL unquantified swab RNA extract (1-5 ng/ul estimate) was treated with DNAse I (Life Technologies), purified and concentrated to a final volume of 5uL, all volume was then used in the library preparation reaction. The libraries were generated according to manufacturer's specifications or by halving the original recommended volumes without compromising library quality. One or two sets of 96 library pools were sequenced on a SE100 cycles SP Novaseq flow-cell (Illumina).

Illumina NovaSeq raw data were initially analyzed by Next Generation Diagnostic srl proprietary 3'DGE mRNA-seq pipeline (v2.0) which involves a cleaning step by quality filtering and trimming, alignment to the reference genome and counting by gene[153–155].

Samples were considered qualitative and retained based on the number of detected genes (≥ 5000) and the percentage of reads assigned to genes (≥ 20%). Data was normalized via the cpm function from the edgeR[156] package (v. 3.34.1). Principal component analysis was conducted by prcomp function from R (v. 4.2) on normalized, log-transformed counts.

Correlation analysis between Ct values and gene expression was performed on genes that were expressed (i.e., CPM > 1) in at least 70% of the entire dataset (8100 genes for B.1 and 5525 for Delta). The test was performed using the function cor.test from R (v. 4.2). Anti-correlation was defined for results with p-value < 10-4. Pathway and gene sets enrichment analysis was conducted using the enrichR[157] package (v. 3.34.1).

3. Results

3.1. A systematic approach allows the generation of large and robust genomic data in a cost-effective manner.

Besides screening and diagnosis, one of the major needs related to the SARS-CoV-2 pandemic is to collect and analyse a considerable number of viral genomes, to guarantee a rapid geographical and continuous surveillance of VOC. To achieve this goal, we developed a systematic workflow that allows the collection, whole genome sequencing (WGS), cloud data processing, and sharing of up to 4500 SARS-CoV-2 genomes per week. Our approach is based on optimizing an amplicon-based workflow (see Methods) (Fig. 25).

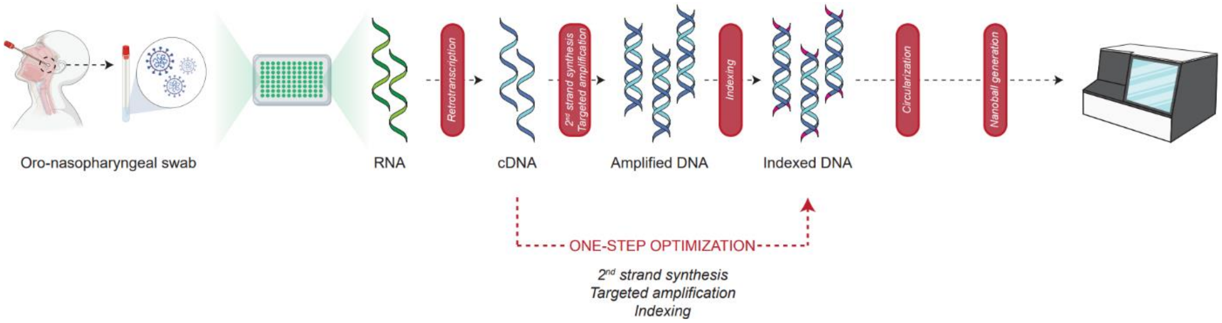

Figure 25. Schematic representation of the workflow set up to collect, process and analyse a considerable number of viral genomes. Top: Oro-nasopharyngeal swabs are performed to diagnose the presence of SARS-CoV-2 genome in patients and extract its RNA. Subsequently, viral RNA is retrotranscribed and subjected to two PCR steps to amplify and index it. After circularization and nanoball generation, the library is then sequenced and analysed. Bottom: As an alternative and faster approach, an optimized strategy enables the amplification and indexing to occur in one PCR step.

To both increase processivity and efficiently reduce costs, the protocol was tested and validated with a decreasing amount of input RNA for the generation of the libraries. In particular, we tested 5 µL, 2.5 µL, 1.25 µL of unquantified RNA and proportionally scaled down the reaction volumes to ½, ¼, and ⅛ (solution A, B, and C, respectively).

Being able to rapidly process the RNA sample to the final viral genome consensus is critical for retrieving meaningful data on the SARS-CoV-2 genome surveillance in a territory. We addressed this point by both optimizing the steps required for library generation and adjusting the number of samples sequenced in each run. Notably, we merged the targeted and the indexing amplification steps in a single PCR (Fig. 1A, see Methods). In parallel, we tested the performance and efficacy of our sequencing flow against an increasing number of samples per run, as such may have an adverse effect on the quality of the resulting viral sequences. We compared Quality Check (QC) statistics obtained by sequencing at a depth of ~ 9, 4,5, and 2,75 million 100bp paired-end reads per sample. This translates into sequencing two (192 samples), four (384 samples) or eight (768 samples) 96-well plates per run. The features of all the solutions tested in this work are summarized in the following Table 3.

	Standard	Solution A		Solution B			Solution C		One-step
RNA volume (μL)	10	5		2.5			1.25		2.5
Reads*/sample ($\cdot 10^6$)	20	10	5	10	5	2.5	5	2.5	5
Samples/flowcell	96	192	384	192	384	768	384	768	384
Processing time (h)**	51	26	14	26	14	7	14	7	13
Relative cost/sample	100%	51%	41%	35%	26%	21%	18%	13%	26%

Table 3.

To compare all the solutions mentioned above, we looked at the sequencing statistics of randomly selected swabs sharing an average Ct value <33. As shown in the following Fig. 26, neither the genome coverage nor the number of sequences passing our quality filters (see methods) and submitted to GISAID was highly affected by volume or sequencing depth reductions. Indeed, the percentage of the genome having a coverage of at least 100X remained extremely high (Fig. 26, left panel) in all solutions tested, and the percentage of samples passing filters was always higher than 75% (Fig. 26, right panel). Only the One-step solution experienced a decrease in the percentage of samples passing filters. While the efficiency of this solution was about ~ 60%, it allowed removing a magnetic beads purification step and thus reduced the hands-on time for library generation by ~40% (see Table 3).

Therefore, by finely tuning the starting amount of RNA, the library generation steps and the number of samples loaded in each sequencing run, we were able to decrease both the processing time and the costs of required for SARS-CoV-2 genotyping.

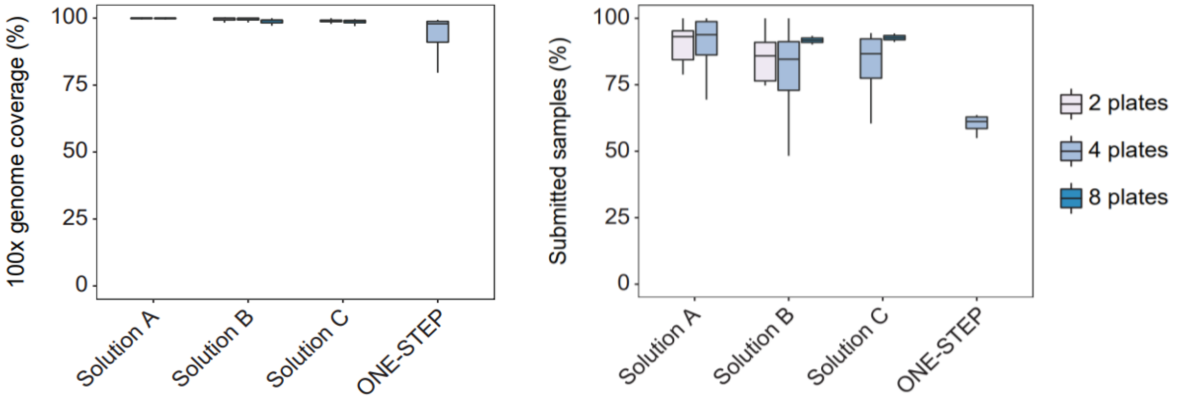

Figure 26. Boxplot showing the % of viral genome reaching at least 100X coverage in each sample (left). Such high coverage is reflected in the number of samples passing our quality filters and submitted on the GISAID platform. All samples are divided by each tested solution. For comparison purposes, only samples with average Ct value < 33 were considered.

Since solution B represented the best trade-off between sequencing costs and the percentage of viral genomes retrieved, we decided to apply it as proof of principle, to apply it for the genomic surveillance of the Campania region, Italy. Using such optimized SARS-CoV-2 WGS workflow, during 2021, we were able to process and sequence 22228 samples, 17193 of which generated high-quality genomes (defined as those complete genomes with a percentage of Ns lower than 5%). Furthermore, a strong correlation between the number of reads detected in each sample and the Ct values obtained from a diagnostic qPCR was observed, as shown in Fig. 27. SARS-CoV-2 reads showed a proportional rate relative to Ct in the intervals between 40 and 25 Ct while reaching saturation <25 Ct.

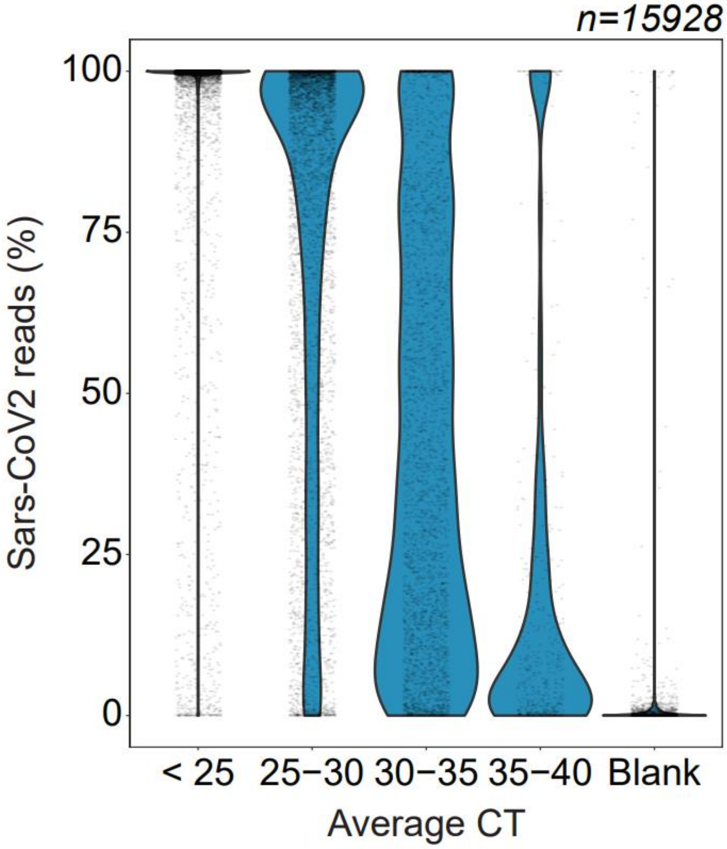

Figure 27. Violin plot showing the distribution of the percentage of SARS-CoV-2 reads detected for different ranges of CTs. n:sample size.

To further evaluate the effectiveness of our approach relative to viral titre, we tested how such parameter affected sequencing efficiency (Fig. 28). As expected, the probability of retrieving a genomic consensus decreased with the viral titre (Fig.28, left panel). However, the overall efficiency of our methodology was higher than 75% for Ct values lower than 33, demonstrating the excellent reliability of the approach. For this reason, whenever possible we focused on sequencing samples with Ct values <33 during our genome surveillance activity.

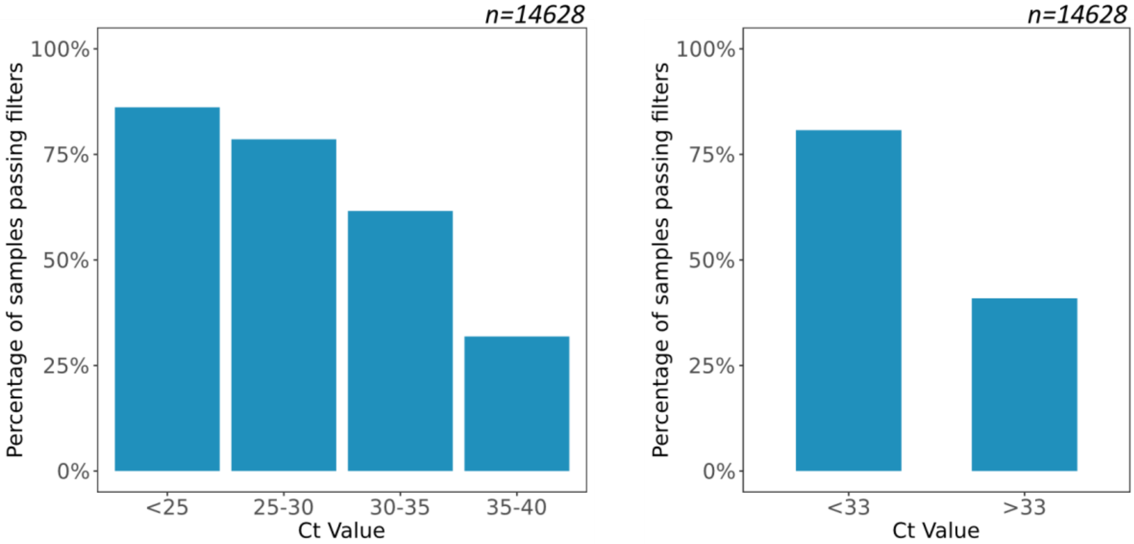

Figure 28. Efficiency of SARS-CoV-2 genome retrieval depending on samples CT value. Samples were divided in 5 (left) or two (right) Ct classes. n: sample size.

Altogether, these observations suggest that our WGS approach reliably quantifies the viral load and provides crucial metadata to correlate higher virus titre to specific virus lineages and a transcriptional response from host cells (see next paragraph). The robustness of our approach was further established by analysing the mean coverage level in 2000 randomly selected samples divided per Ct value. Fig. 29 shows that the coverage appeared to be homogeneous across most of the SARS-CoV-2 sequence for all Ct values tested. However, a drop in genome coverage was detected at the end of the genome (after position 29854) and at genomic window 14779-14840. The former is probably associated with poorly priming of viral poly(A) tail and has been observed in several other sequencing approaches[158]. On the other hand, the latter is probably due to low amplicon generation from a specific couple of primers. Nevertheless, the coverage is always higher than 30X (dashed red line in Fig.29) for all samples associated to a Ct value ≤35.

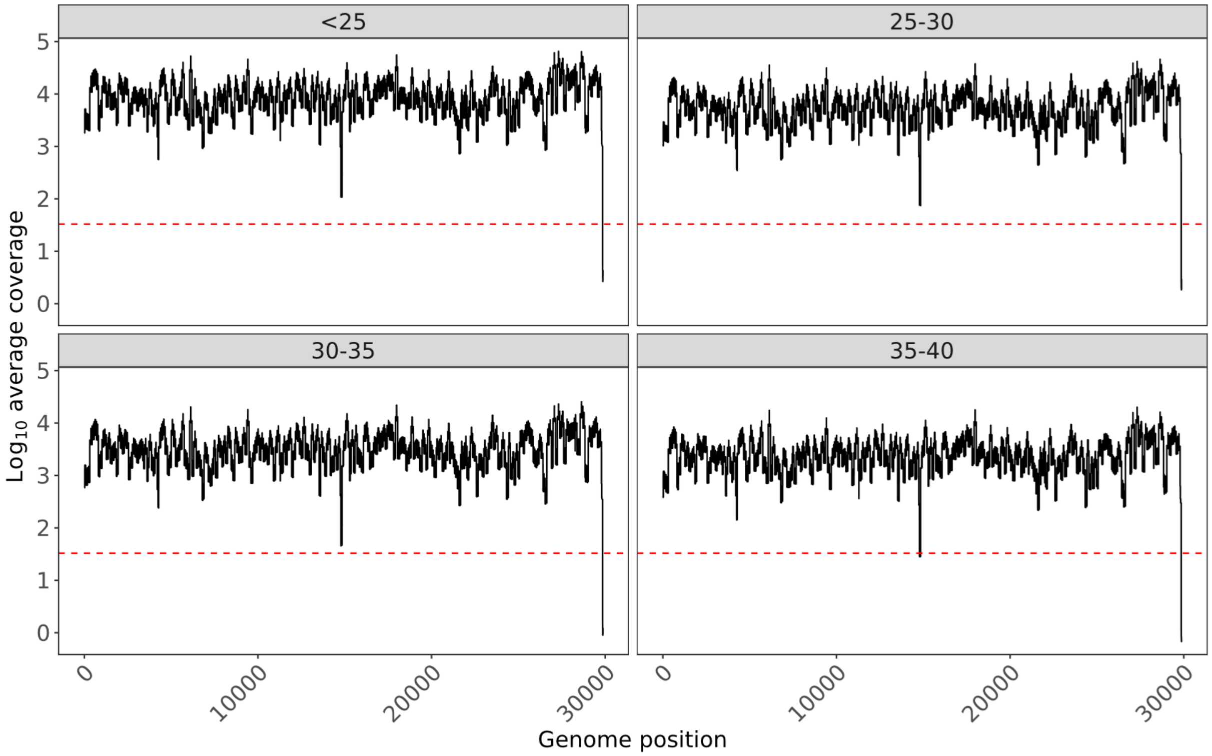

Figure 29. Average coverage across SARS-CoV-2 genome obtained in different Ct classes (plot titles). The 30X coverage threshold chosen for confident base calling is shown as a red dashed line.

Looking at the whole dataset, sequencing coverage was extremely high, with each base covered, on average, by over 5000 reads. This piece of evidence confirmed the absence of significant biases in the single nucleotide evaluation. Hence, we investigated the SNPs information derived from our genomic screening and determined missense and synonymous mutations to be the most frequent across the entire genome, although few positions appeared to be more prone to mutate (Fig.30).

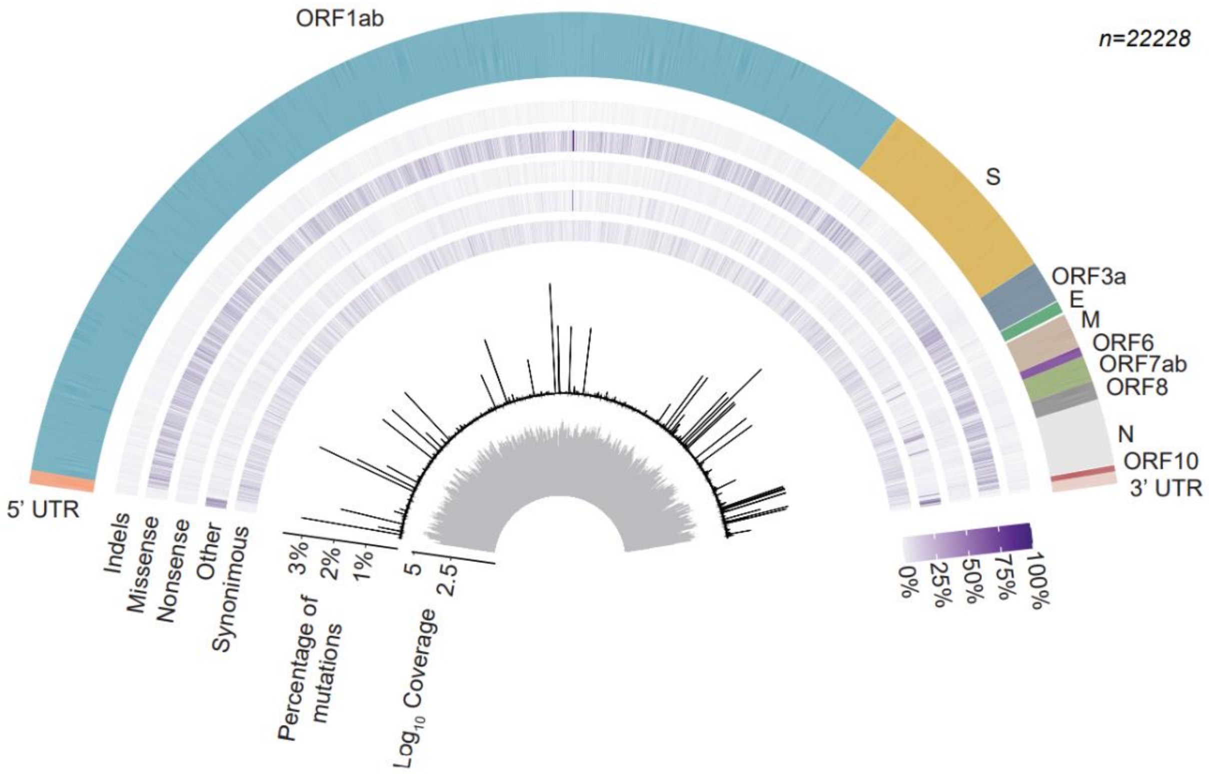

Figure 30. Variant annotation, cumulative frequency, and sequencing coverage of each position of SARS-CoV-2 genome. n: sample size.

Indeed, 6970 mutations were efficiently detected in our dataset, 194 of which were previously unknown and 40 only identified in Campania (Fig. 31, left panel). Interestingly, out of the 194 mutations first collected in the region, 20 fall within the Spike gene[159]. Taken altogether the mutations detected allowed us to identify 156 different SARS-CoV-2 pangolin lineages (Fig. 31, right panel), some of which were retrieved for the first time thanks to our activity (following paragraphs). Looking at Figure 31, it is worth noting that Delta VOC and its subvariant (red bars) accounted for most of the SARS-CoV-2 lineages identified in our territory during the period under interest.

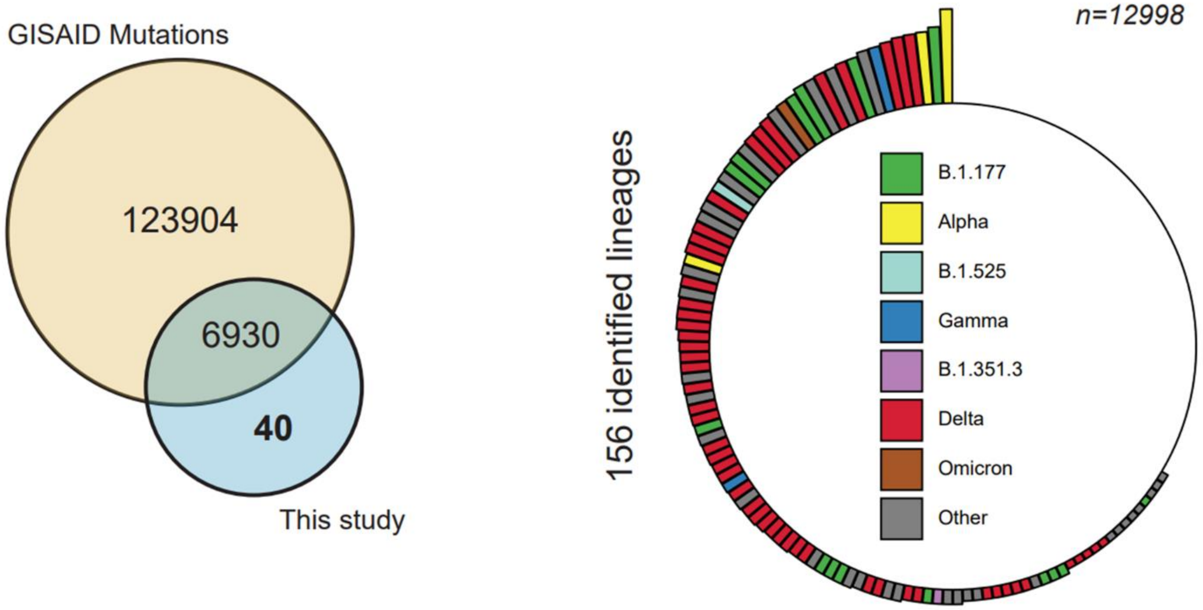

Figure 31. Mutations and lineages identified during 2021 genome surveillance. Left: Venn diagram showing the intersection between mutations detected in all the sequenced genomes worldwide (yellow) and the mutations found in this study (light blue). Right: representation of all the 156 lineages identified in this study. The length of the bars is indicative of the number of samples for each lineage in the logarithmic scale. Coloured bars indicate VOC

3.2. Characterization of SARS-CoV-2 genome evolution in Campania.

As aforementioned, from the end of December 2020 to the first week of 2022, we sequenced, uploaded to GISAID, and analysed 17193 SARS-CoV-2 genomes. Our workflow was tested throughout the Campania region, which includes the major southern Italian metropolitan areas and some of the most densely inhabited cities in Europe. Samples collection started in March 2020 and, in order to depict an accurate picture of the SARS-CoV-2 pandemic, positive swabs were selected reflecting population demographics of sex, age, and the geographical distribution across the area of interest (Fig. 32)

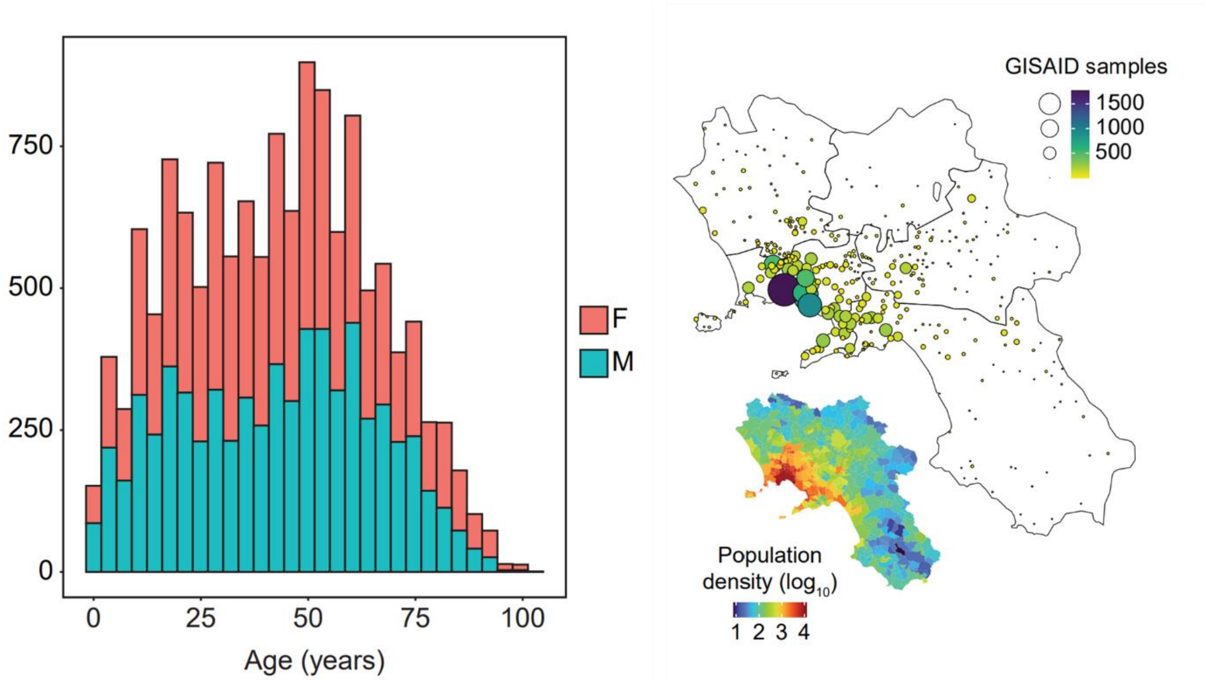

Figure 32. Demographic (left) and geographic (right) distribution in Campania of the patients swabbed for SARS-CoV-2 genome analysis (left). The population density (bottom right) of the region is also shown as reference.

The results of our work of collection, sequencing, and sharing of SARS-CoV-2 genetic data are reflected by the numbers of sequences deposited on GISAID, the reference database for sharing pathogens data (see paragraph 1.2.4). Our dataset, indeed, represents almost half the sequences uploaded from the south of Italy and 28% of all sequences produced in the country (Fig. 33)

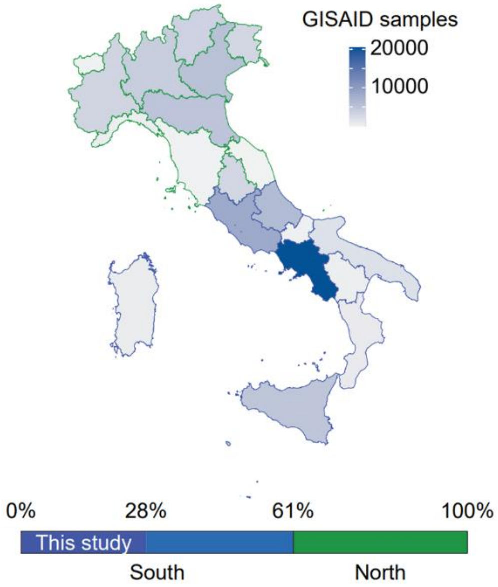

Figure 33. Geographic map representing Italian regions, coloured by the number of genomes deposited on the GISAID platform. Bottom: percentage of genomes deposited on GISAID over the total Italian sequences, divided in Northern (green) and Southern (blue) regions. 20% of the overall Italian sequences has been produced by this study (dark blue).

Sequencing numbers were evenly distributed during the pandemic; thus, we were able to sequence, in most months during 2021, at least 5% of all COVID-19 positive samples (Fig. 34). Such value made Campania one of the few Italian regions to be compliant with ECDC recommendations, reaching a sequencing coverage comparable to that of North-European countries.

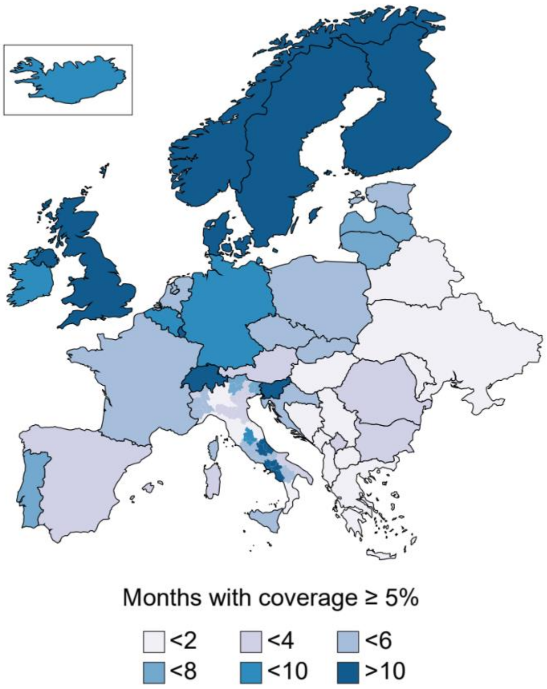

Figure 34. Geographic map representing European States, colored by the number of 2021 months with at least 5% of viral genomes compared to new cases. 5% is the recommended

The analysis of samples collected during the pandemic allowed us to unveil the full dynamics of the SARS-CoV-2 outbreak in Campania. First, we reconstructed the distribution of all the VOCs that arrived in Campania; notably, the delta (represented by B.1.617.2 and AY.* lineages) and alpha (B.1.1.7 and Q.*) VOCs, represented the vast majority of variants detected (71.6%, see Figure 31). In accordance with worldwide data, the first VOCs arriving in the region, starting from December 2020, were the B.1.1.7 and P.1 (gamma variant). Next, other VOCs were detected in the region, including B.1.351, P.1 and B.1.1.529 lineages (i.e., Beta, Gamma, and Omicron VOCs, respectively). We also identified three main Variants of Interest (VOI); the B.1.427, B.1.525, B.1.621 lineages (i.e., Epsilon, Eta, Mu, respectively) (Fig. 35).

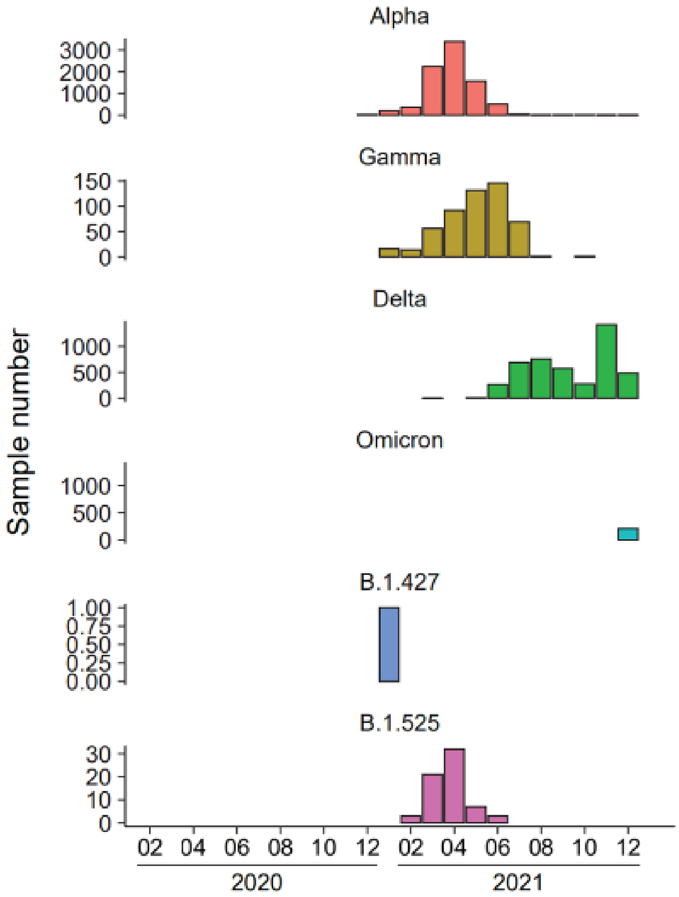

Figure 35. Bar charts showing the distribution in time of the samples assigned to variants of concern and of interest identified in the during our surveillance program.

Out of the 156 viral lineages identified in the region, 5 were first recorded in Campania territory, namely B.1.1.187, B.1.177.33, B.1.177.75, C.18 and P.1.1. In particular, C.18 viral variant was first collected in July 2020. In contrast, its first record outside Campania was registered 3 months later, suggesting a possible epidemiological origin from our territory of investigation. Similarly, over 82% of B.1.1.187 samples collected during the pandemic derived from Italy, all from Campania. Our analysis also showed that the first gamma VOC sub-variant identified (pangolin lineage P.1.1) was first sampled in Campania by our activity (Tables 4 and Appendix Table 1) and that it was explicitly enriched in Italy, with sequences from Campania representing about 20% of all P.1.1 samples identified.

Pangolin designation	First collection date in Campania	First collection date outside Campania
B.1.1.187	2020-06-24	2020-07-08
B.1.177.75	2020-09-02	2020-09-21
B.1.177.88	2020-12-26	2021-01-21
C.18	2020-07-01	2020-09-09
P.1.1	2021-01-04	2021-01-07

Table 4.

Looking at the whole picture, we determined that the main infection peaks in the region were associated with the spread of specific viral lineages (Fig. 36 top and middle overlays). Indeed, while the first wave of infections was primarily due to the ancestral B.1 lineage, the second one (autumn 2020) was led by the B.1.177 lineage (also referred to as European or Spanish variant) and its sub-lineages. Interestingly, the time window between the first two infection peaks was characterized by two of the lineages mentioned above to be firstly detected in Campania; B.1.1.187 and C.18, associated with most of the COVID-19 cases during late spring and summer 2020. These two variants distinguish the pandemic in Campania relative to the rest of Italy (Fig. 36 lower level, red arrow). In the same period in the rest of the country, infections were predominantly associated with other B.1 sub-lineages, including B.1.1, B.1.1.305 and B.1.1.229. Finally, the two infection peaks of 2021 were due to the spread of alpha and delta variants. These two VOCs succeeded one after the other and accounted for almost all COVID-19 cases in the first (alpha) and second (delta) half of 2021. Interestingly, from December 2021 B.1.1.529 (omicron) variants started to emerge.

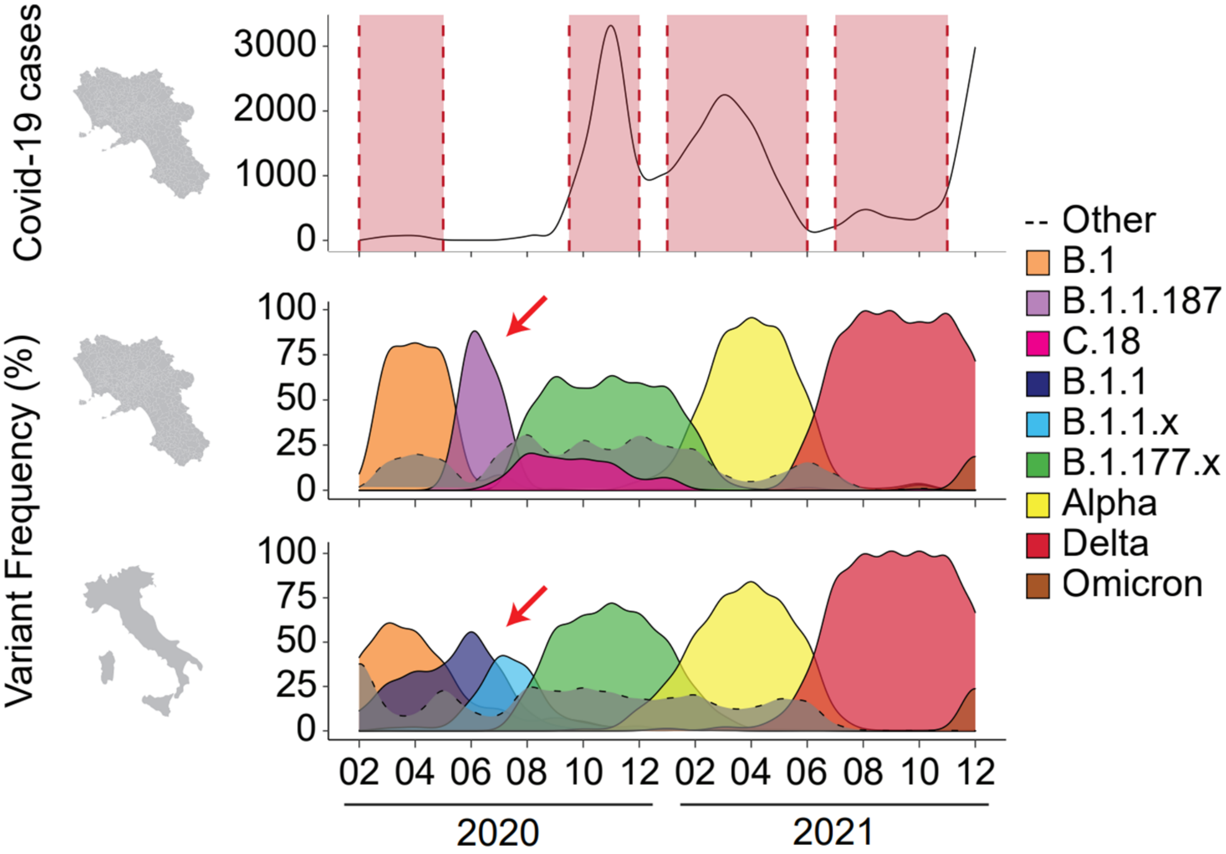

Figure 36. Density plots showing the distribution, in time, of the most frequent variants described in this thesis (middle) or in Italy (bottom) relative to the Campania infection curve (top) and infection waves (red-colored areas). The red arrow highlights different variants dynamics between regional and national level, in a certain period.

Since during this succession of variants in the regional territory, none of them ever reappears after being undermined by the subsequent one, it is fair to suppose that each variant has been substituted by one with higher fitness and capability to spread. To test this hypothesis, we looked at the viral loads in the upper respiratory ways of patients infected by the predominant variants in Campania. We observed a clear trend towards an increase of viral titre in patients during the pandemic, with a Ct value difference between omicron and the ancestral B.1 variant of -7,8 (q value < 2·10-16 pairwise Mann-Whitney test, Fig. 37 left panel). A similar trend towards decreasing Ct values was also observed when considering all the variants identified in the region (Mann-Kendall test, p value=8,99·10-10, Fig. 37 right panel).

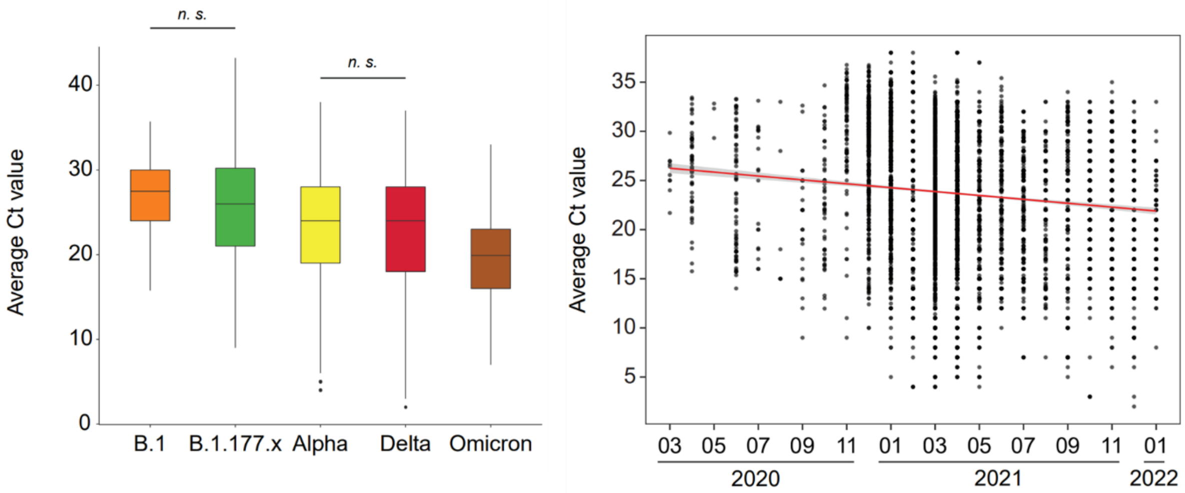

Figure 37. Time course of Ct values during pandemic in Campania. Left: Distribution of the average CT value across different Variants of Concern (VOC). Only not significant (n.s.) pairwise comparisons are reported (Bonferroni adjusted p-value > 0,05). Right: Distribution of Ct values, in time, for all the samples collected during the pandemic in Campania. The trend line (red) and 95% confidence interval (light gray) are shown.

3.3. Identification of new variants based on the analysis of single mutations.

Since the comparative analysis of our dataset with GISAID world data allowed us to retrospectively identify viral variants firstly sampled in Campania (B.1.1.187 and C.18), we were interested in exploring whether it was possible to unveil new viral lineages circulating in the territory. To achieve this goal, we explored several approaches. We mainly focused on the concept that a new SARS-CoV-2 variant is characterised by a specific set of mutations; therefore, we generated approaches based on:

1) mutations associated with higher infectivity found in unexpected variants.

2) An increasing incidence of a set of mutations in a short time window.

3) The appearance of new mutations in samples collected by patients with persistent infections.

First, we explored SARS-CoV-2 “mutations of concern” genotyped in unexpected lineages. Interestingly, we found that the Spike E484K substitution had an unexpected distribution in the lineage identified at the beginning of 2021. (Fig. 38).

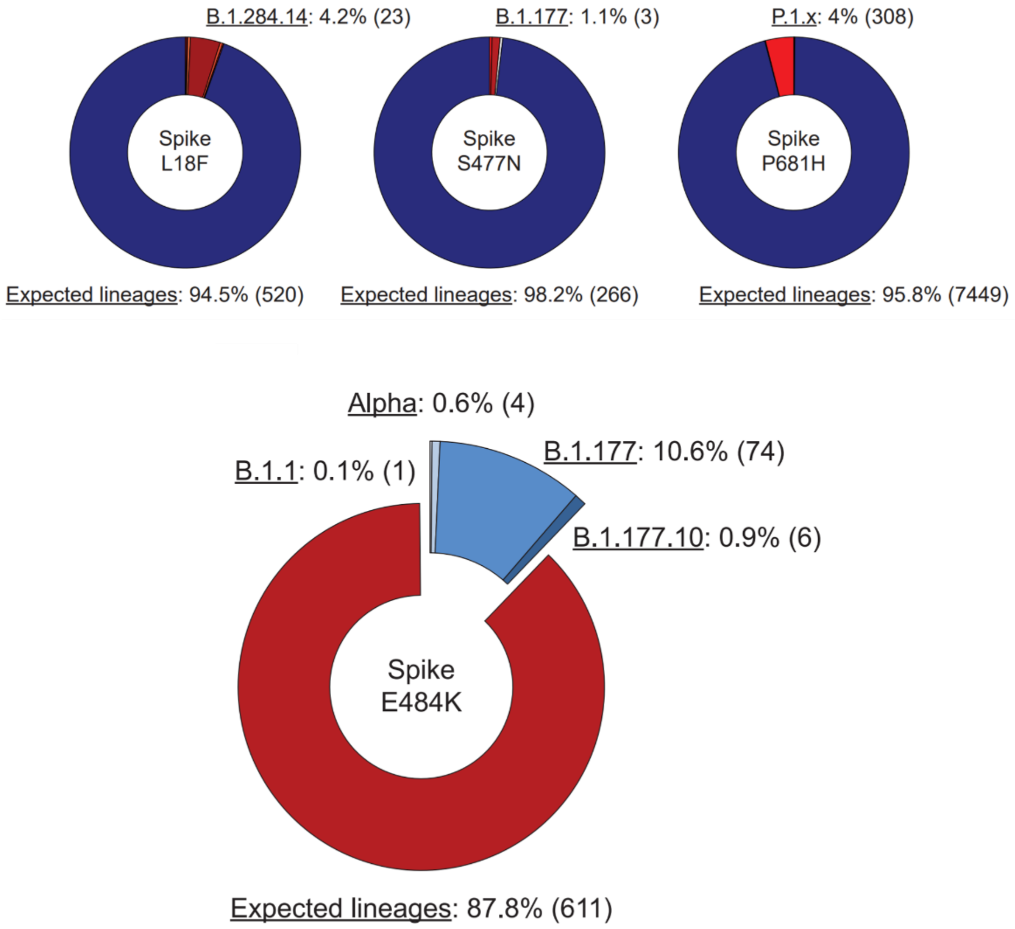

Figure 38. Donut charts representing the number of analysed genomes presenting some mutation of concern, namely Spike L18F, S477N, P681H (top) and E484K (bottom) divided by lineage. The definition of Expected lineage is described in the Methods chapter.

This mutation is typically found in P.1.x and B.1.351.x viral lineages and has been associated with a decreased sensibility to both monoclonal and BNT162b2 vaccine-induced antibodies[114–116]. However, as of May 2021, ~21% carrying this mutation were associated with the B.1.177.x lineage (Fig. 38, lower panel). To further investigate this finding, we performed a phylogenetic analysis over our entire dataset using Nextstrain[101] and found that all B.1.177.x samples carrying the Spike E484K substitution (B.1.177^{E484K} samples) clustered in a specific and monophyletic clade branching within the B.1.177.x lineage (Fig. 39).

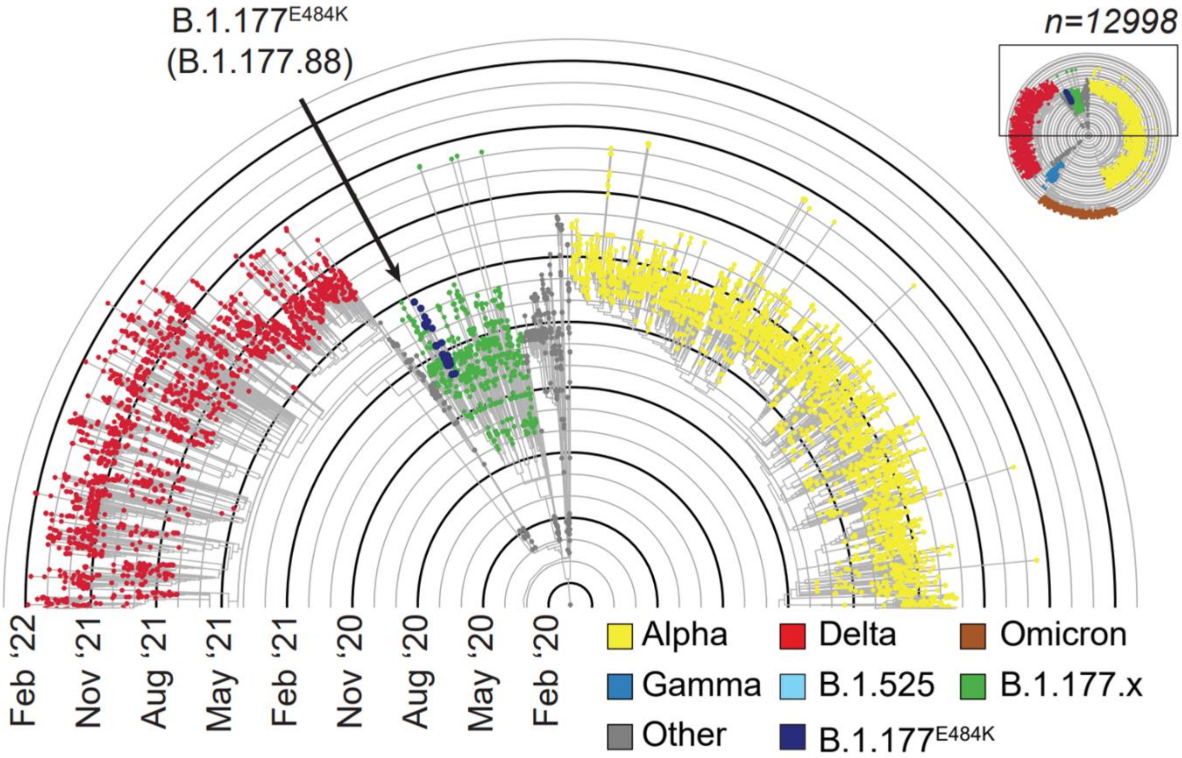

Figure 39. Section of the maximum likelihood phylogenetic tree representation of a random subset of our data (n=12998), coloured by lineages. The identified lineage is reported (blue dots, right) and zoomed in (left). n:sample size.

We further confirmed this finding by looking at the distribution of B.1.177^{E484K} samples in the phylogenetic tree containing all high-quality SARS-CoV-2 genomes from GISAID. This data points to the fact that B.1.177^{E484K} samples cluster in a monophyletic clade with an extremely high (0,99) support value, thus confirming regional level incidence (Fig.40 left panel). Additionally, since the GISAID database revealed that B.1.177^{E484K} samples had been identified for the first time in Campania through our program, we investigated their geographic distribution in the regional territory to trace the epidemiological link (Fig. 40 right panel). Surprisingly, these samples originated from a specific area between Naples and Salerno called "Agro Nocerino-Sarnese".

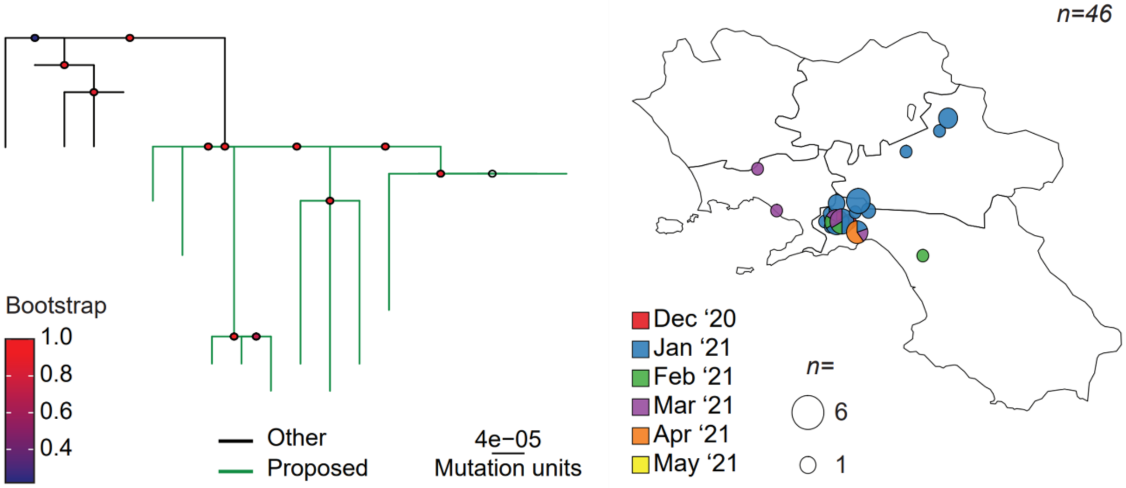

Figure 40. Right: phylogeny of the proposed B.1.177^{E484K} lineage. Samples belonging to the proposed lineage are in green. Bootstrap values for each node are shown as node points. Left: Geographic distribution of genomic variants belonging to the identified lineage, coloured by the collection date. The size of each pie chart is proportional to the number of samples in each geographic position. n:sample size

Combining these results, we hypothesized that B.1.177^{E484K} variant had probably arisen in this area in December (Treetime divergence inferred interval: 2020-11-22~2020-12-21) and then spread nearby in Campania and in other confining Italian regions (mainly Lazio and Basilicata). Altogether, these observations allowed us to define a new SARS-CoV-2 lineage, which is now recognized by the Pangolin nomenclature B.1.177.88.

To identify any new variant rapidly growing in the territory as soon as it appears, we exploited another approach based on the incidence over time of each of the 6441 amino acid mutations we identified. Since, by definition, a viral variant is defined by a specific combination of mutations, we looked at mutations that displayed similar trends in the same period and grouped them in clusters. In order to identify any potential new alpha subvariant growing in Campania, we applied this methodology to the data collected until May 2021, when the variant reached its maximum. Confirming the robustness of our approach, several clusters clearly reflected the trends of known lineages; for instance, cluster 6 consisted of those substitutions that characterize the B.1.177.x lineage (namely N A220V and Spike A222V) and presented the same trend over time. Similarly, cluster 29 reflected the trend of B.1.1.7 lineage (Fig. 41).

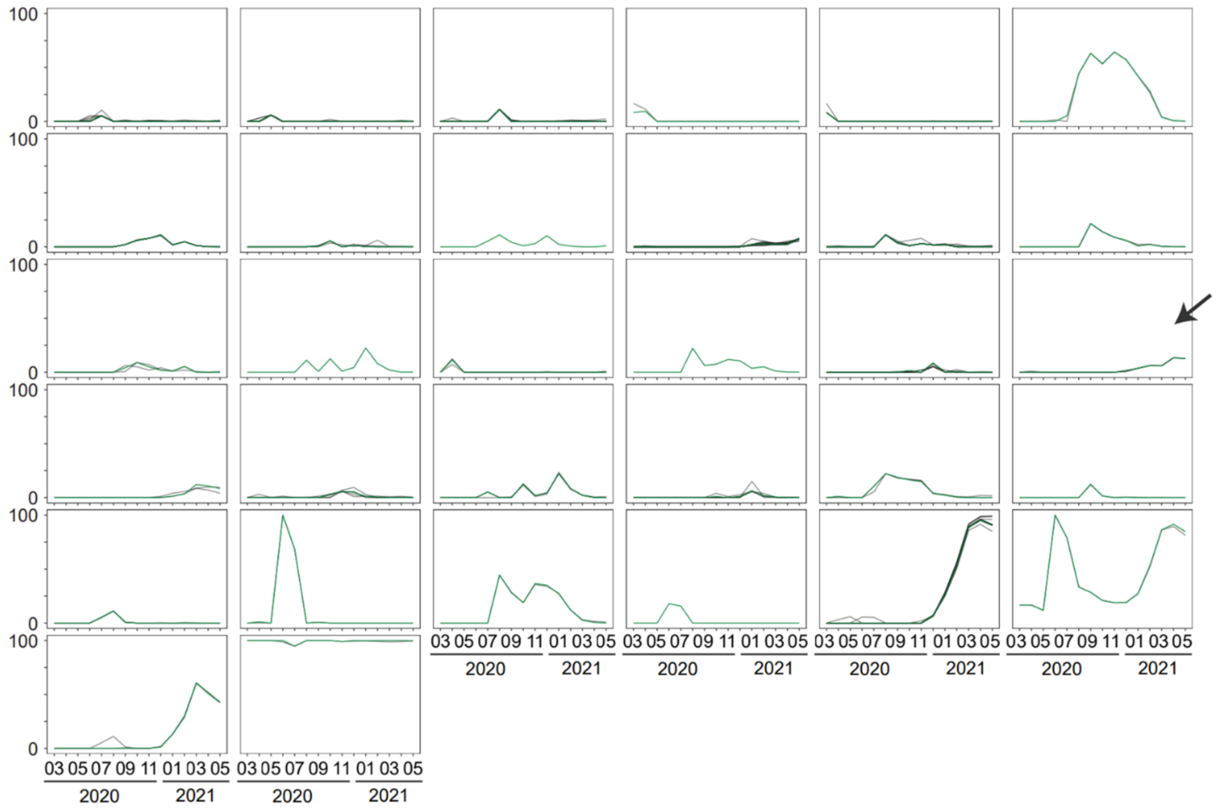

Figure 41. Results from the clustering analysis for samples collected until May 2022, displayed as line plots of frequency over time (trends). The arrow indicates the cluster investigated in next figure.

Among these, cluster 18 was particularly interesting. As shown in the following Fig. 42, it consisted of 4 mutations (NSP2 Y316C, NSP3_T1306I, NS7a T120I, NS8 H112Y) with the exact same frequency behaviour over time, thus suggesting a possible SARS-CoV-2 haplotype.

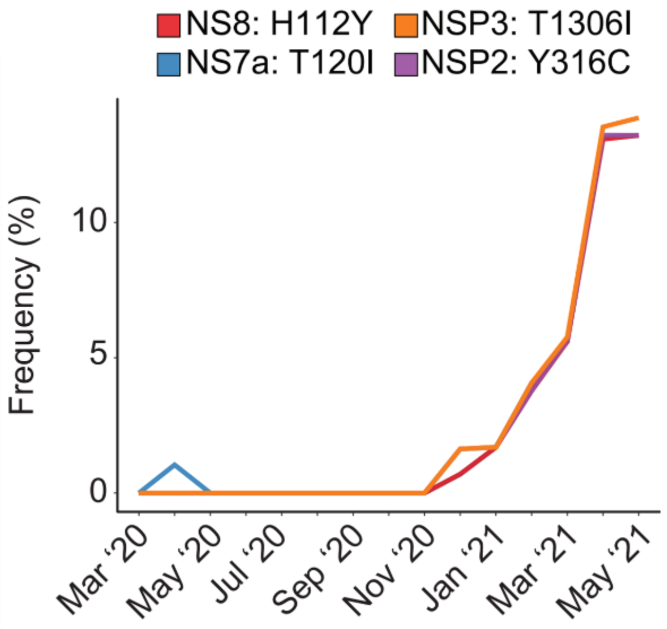

Figure 42. Line plot showing the frequency trend of the selected mutations in Campania.

A further investigation revealed that these SNPs define a set of samples assigned to the Alpha VOC and specifically localized in Campania in May 2021 (Alpha^{YTTH} samples). We carried out a phylogenetic analysis that confirmed Alpha^{YTTH} as monophyletic (Fig. 43, left panel), and this finding was, again, corroborated by the high bootstrap (0.8) value associated with the base node defining the clade (Fig. 43, right panel).

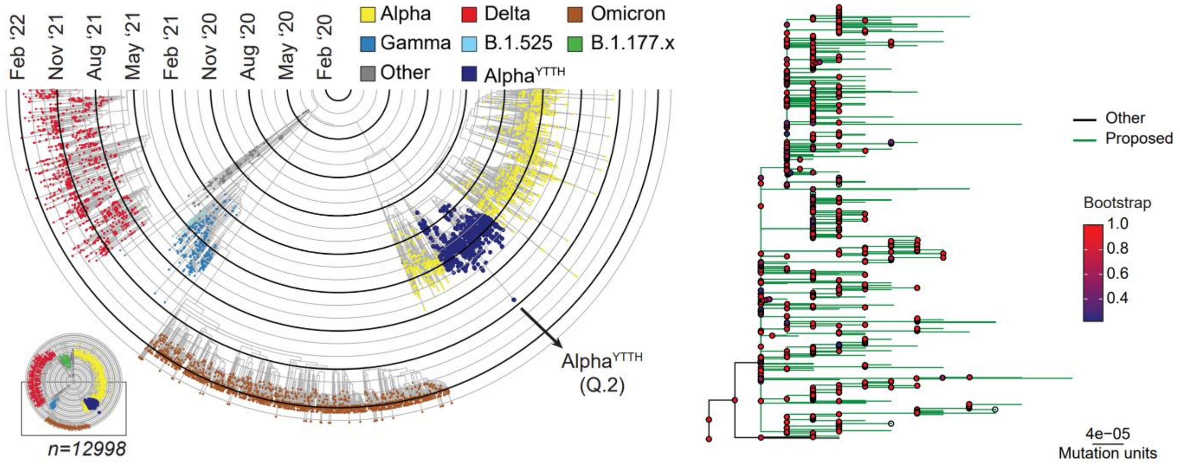

Figure 43. Right: Section of the maximum likelihood phylogenetic tree representation of a random subset of our data (n=12998), coloured by lineages. The identified lineage Alpha^{YTTH} is reported (blue) and zoomed in (top). n:sample size. Phylogeny of the proposed Alpha^{YTTH} lineage. Samples belonging to the proposed lineage are in green. Bootstrap values for each node are shown as node points.

Whilst Alpha^{YTTH} genomes did not show any geographic enrichment, its temporal distribution was indicative of an inland origin (Treetime divergence inferred interval: 2020-12-01~2020-12-06), followed by its spread first to the Neapolitan coast and then towards the Southern Neapolitan province (Fig. 44).

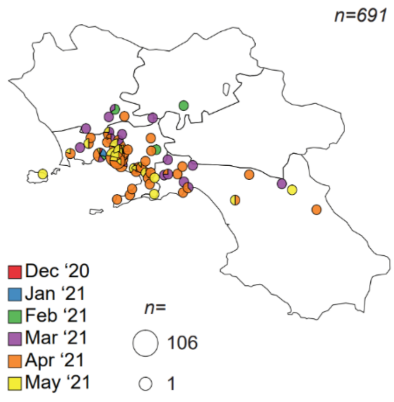

Figure 44. Geographic distribution of genomic variants belonging to the identified lineage, colored by the collection date. The size of each pie chart is proportional to the number of samples in each geographic position. n:sample size.

The Alpha^{YTTH} variant has been recognized, upon our alert, as one of the first B.1.1.7 sub lineages by the Pangolin system and is now referred to as the Q.2 lineage.

We also applied the same approach with the final goal to identify any possible Omicron subvariant spreading in Campania ad the beginning of 2022. For this reason, we applied again the same approach to the data produced till January 2022. This allowed us to discover a set of samples, assigned to the Omicron VOC, that was characterized by 2 interesting mutations: NS7b E3Stop and Nsp12 L749M (Fig. 45, left panel). These samples were first collected at the end of 2021, and rapidly spread in Campania at the beginning of 2022, accounting for over 10% of all the infections in the region between January and March 2022. Also in this case, the phylogenetic analysis revealed that the samples clustered in a statistical significative monophyletic group (Fig. 45, right panel) that has been successfully assigned to a new SARS-CoV-2 subvariant, named BA.1.21.1.

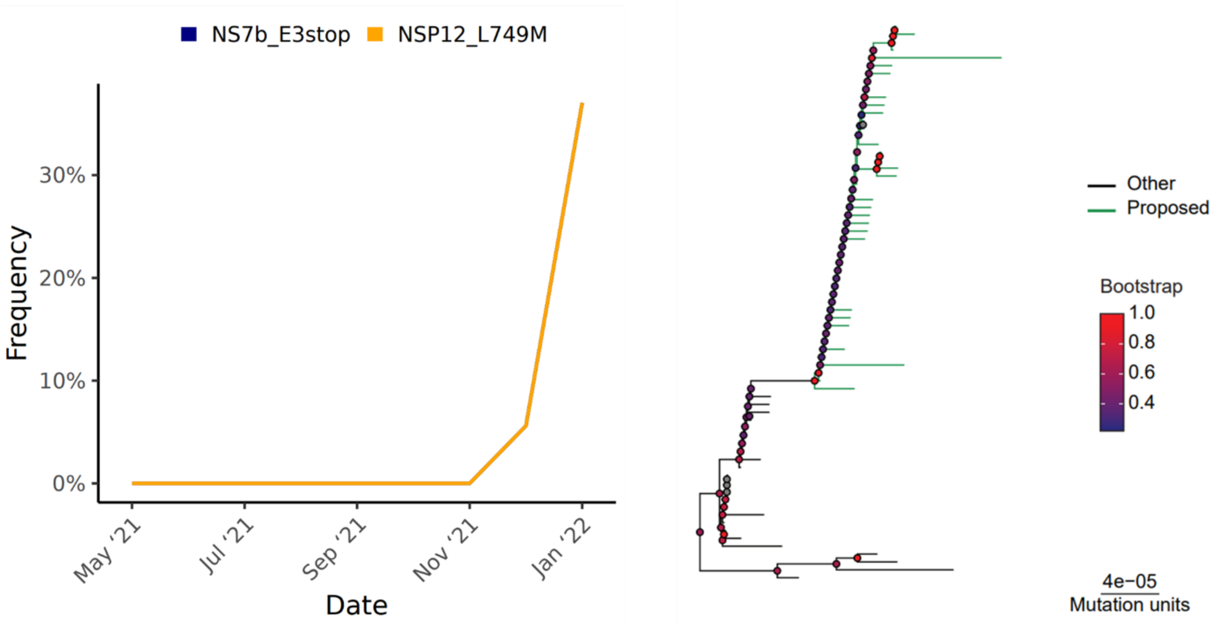

Figure 45. Line plot showing the frequency trend of the selected omicron subvariant mutations in Campania (left) and corresponding phylogenetic reconstruction (right). Samples belonging to the proposed lineage are in green. Bootstrap values for each node are shown as node points.

Several reports showed the accumulation of mutations in the SARS-CoV-2 genome during persistent infections[160]. However, the frequency of such events is still overlooked. In order to possibly address this question and identify potential new variants, we analysed swabs collected from 20 patients multiple times for over 40 days during prolonged infections. Age and immunological status highly varied across the patients (Table 1): patients' age ranged from 13 to 88 (average 62) years, and while most of them were affected by simple or bilateral pneumonia, six suffered a more severe respiratory failure, and only one showed no COVID-related symptomatology. It is worth noting that, although most samples were collected during 2021, none of the patients had completed a three-dose SARS-CoV-2 vaccination cycle, 4 had only one vaccine dose, and most had no vaccination at all (13/20) (Table 5).

Patient	Immune compromised	Main clinical symptoms	Comorbidities	Age	Vaccine	Outcome
1	Yes	Pneumonia	LNH	64	None	Healed
2	No	Pneumonia	Hemoperitoneum, anaemia	30	None	Deceased
3	Yes	Respiratory failure	Pulmonary hypertension, NHL	64	(x2)	N/A
4	No	ARDS	Diabetes, hypertension, ischemic heart disease	76	None	Healed
5	No	Mild respiratory failure	Necrotizing-hemorrhagic pancreatitis	60	None	Deceased
6	No	ARDS	Hypertension, dyslipidaemia	61	None	Deceased
7	No	Bilateral pneumonia	T2D, obesity, hypertension	59	None	Healed
8	No	N/A	Atrial fibrillation , T2D	78	None	Deceased
9	Yes	ARDS	Anaemia, ALS, COPD	73	(x1)	Healed
10	No	ARDS	Sepsis, anaemia, pulmonary hypertension	64	None	Healed
11	No	Bilateral pneumonia	None	88	None	Deceased
12	No	Bilateral pneumonia asthenia,	Hypertension, T2D, HCV, dyslipidemia, obesity	68	None	Healed
13	Yes	Pneumonia	NHL	73	(x1)	Healed
14		N/A		87	N/A	Healed
15	No	Respiratory failure	Psoriasis	44	None	Heled
16	No	Pneumonia, dyspnoea	Hypothyroidism, severe obesity	71	(x1)	Healed
17	N/A	N/A	N/A	26	N/A	Healed
18	Yes	Asymptomatic	Ewing sarcoma	13	None	Healed
19	No	Cough, dyspnea, pneumonia	Mixed dyslipidaemia, obesity, hyperthyroidism, hypovitaminosis D	64	None	Healed
20	Yes	Pneumonia	Thymoma, Good's syndrome	60	(x1)	Healed

Table 5.

Sequencing of the viral genetic material confirmed no shift from one viral variant to another over time, but each had a set of patient-specific mutations. However, looking at the individual mutations, in one patient (#8), there was an actual increase in the number of amino acid substitutions, as confirmed by 2 independent sequencing runs on 2 subsequent timepoints (Fig. 46).

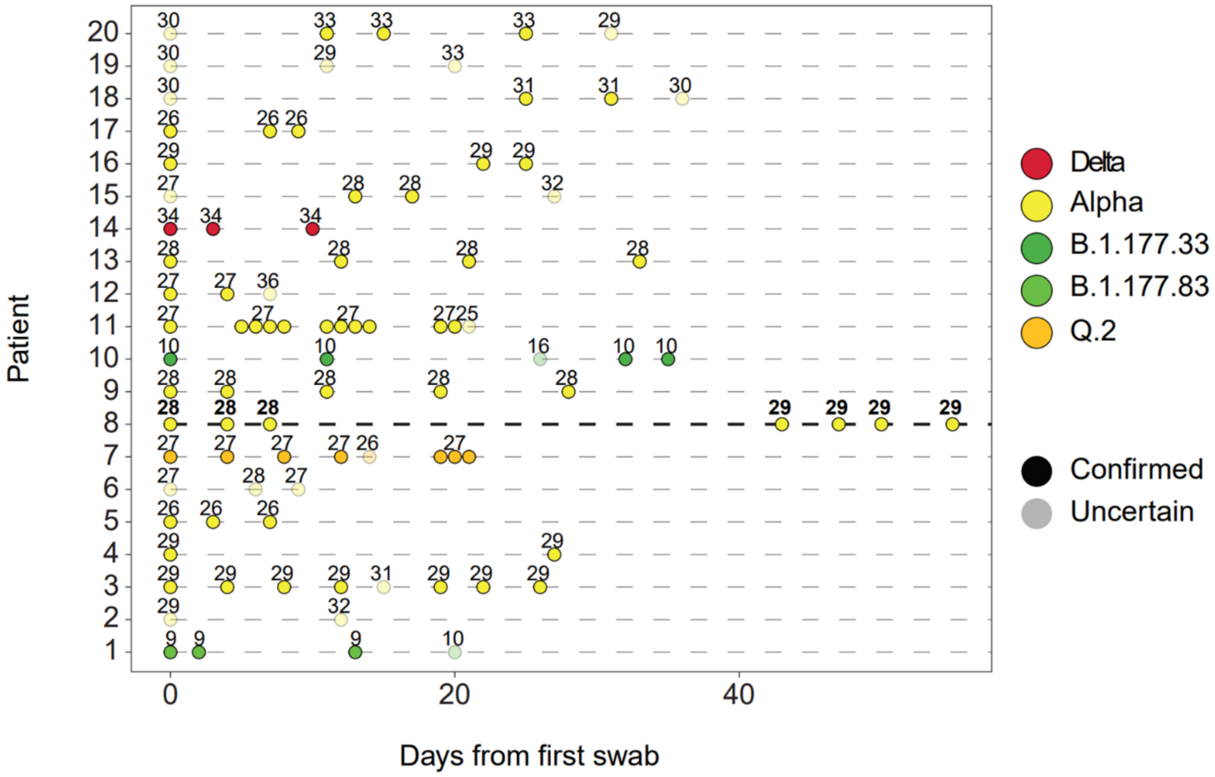

Figure 46. Genomic characterization of twenty patients with long COVID-19 infection. The number of detected mutations is reported as a function of the number of days from the first swab. The assigned lineage (colours) and consistency (transparency) are also displayed.

The acquisition of the mutation (NSP13 R339C) was recorded only after 40 days from the first swab. It did not correlate with an increase in the viral load or a worsening of the symptoms (Fig. 46). These results suggest that in specific conditions, such as over 40 days of persistent infection, the SARS-CoV-2 genetic consensus sequence can actually change, although the rate of such an event, as well as its biological significance, are not known yet (Fig.47).

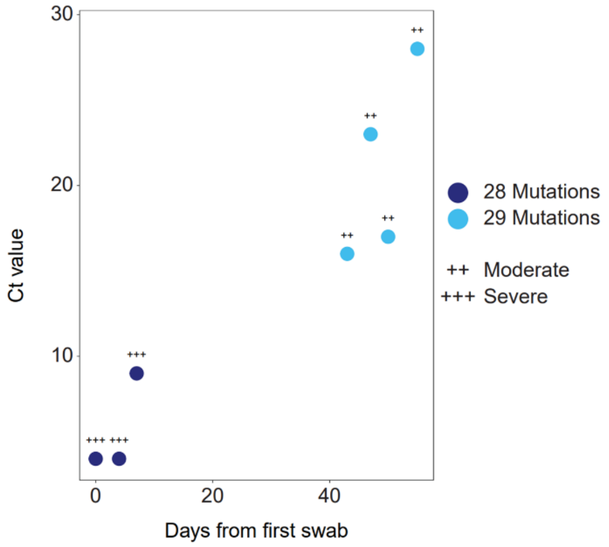

Figure 47. Patient 8 genomic characterization relative to the number of detected mutations (colours), the infection load (y axis) and symptoms severity (+++: severe; ++: moderate).

Tracking new variants based on mutations arising in specific conditions is a novel approach for SARS-COV-2 surveillance. Here we showed that by combining this approach with deep profiling of viral variability, new SARS-CoV-2 variants could be unveiled, even at the regional level.

3.4. Transcriptional profiling of SARS-CoV-2 infected patients reveals a gene signature.

The comprehensive gene expression profiling of the respiratory epithelium of patients positive for SARS-CoV-2 infection holds great promise in terms of preventive, diagnostic, and therapeutic advancements. For this reason, we implemented an RNA-seq workflow adapted to work with diagnostic swabs, known to have low quantity and RNA quality. We processed around 700 samples in two batches to analyze the differential molecular host response to B.1 and Delta variants infection. After filtering, the B.1 final dataset comprehended 116 SARS-COV-2 positive samples to be compared with 88 negative ones. On the other hand, the Delta dataset was composed of 43 and 95 SARS-COV-2 positive and negative samples, respectively (Fig. 48).

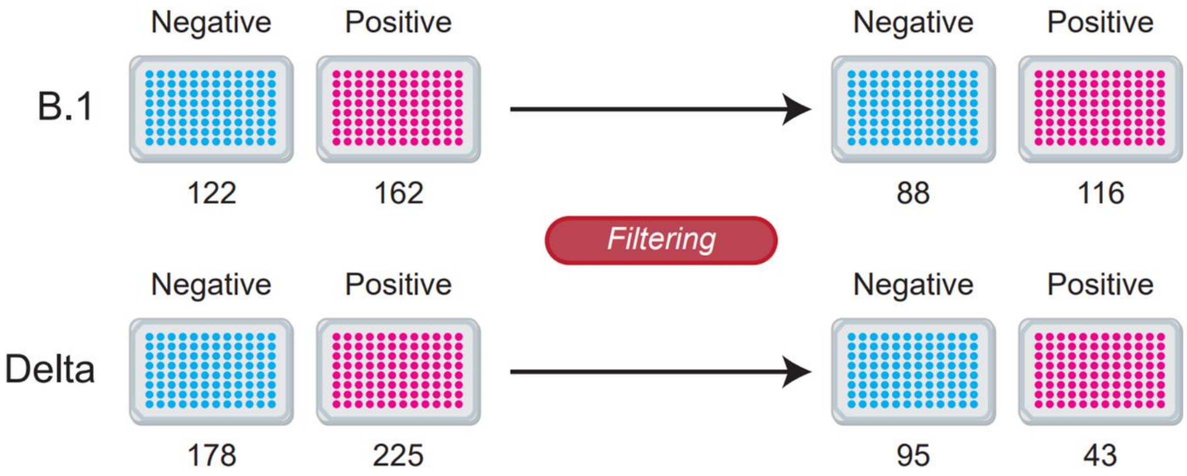

Figure 48. Schematic representation of RNA-seq data structure, pre- and post-filtering. The number of samples per condition pre- and post-filtering are shown.

Although the cohort of patients was numerous, in both cases, many confounding variables influenced the possibility of comparing positive and adverse conditions. Inter-patient heterogeneity, different viral loads, and swab-related variability are some factors that prevented us from finding a solid variance solely related to the presence or absence of the infection (Fig. 49).

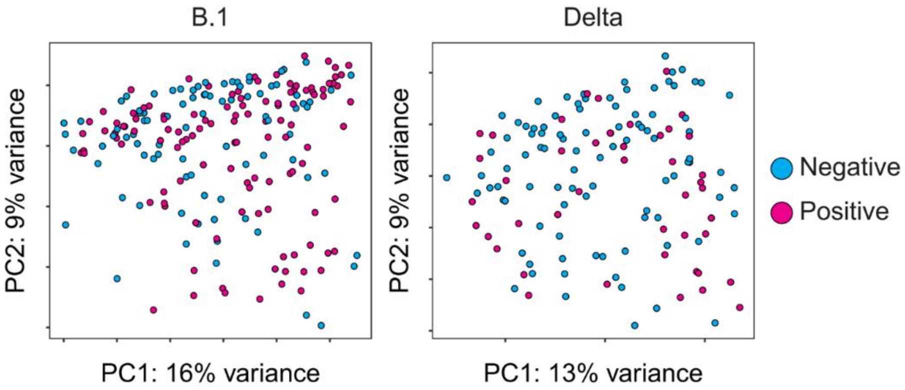

Figure 49. Principal Component Analysis plots of B.1 and Delta datasets coloured by SARS-COV-2 infection positivity.

Therefore, we decided to take advantage of the Ct values associated with positive samples and perform a correlation analysis between gene expression and viral load, starting with the B.1 dataset (Fig. 50). After filtering non-expressed genes (see Methods), a Pearson correlation test revealed a signature of 161 genes (Appendix Table 2) significantly anti-correlated with Ct values(p-value < 0,0001, Fig. 50 bottom). Furthermore, among the 10 most anti-correlated genes, many downstream targets of interferon antiviral response (e.g., IFI44L, OAS2, PARP9, IFITM3, IFIT1) were found, as already reported from in vitro experiments and single-cell studies[161].

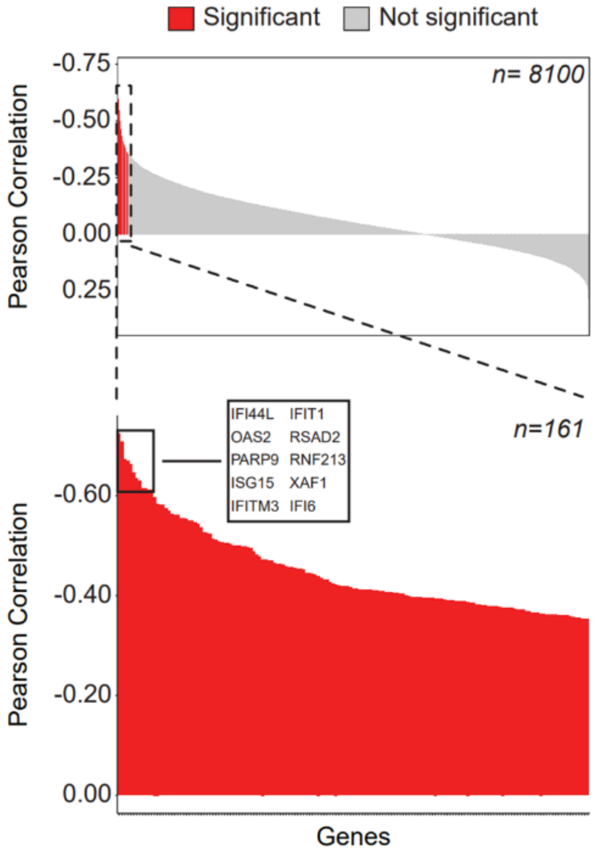

Figure 50. Correlation analysis between CTs and gene expression of B.1 patients, performed on 8100 genes, is shown as a barplot. For each gene (x axis), its correlation value (y axis) and significance (p-value < 0,0001, red) is reported. Bottom: highlight of the significant results. (161 genes). The top 10 most anti-correlated genes are reported (black box).

We confirmed an enhanced antiviral immune response by performing pathway and gene signatures enrichment analyses. Indeed, together with COVID-19- and Bronchitis-related signatures, the most significant results comprehended the Interferon Alpha pathway and its inducers, IRFs. Additionally, STAT3-regulated genes were enriched, which were recently found to be aberrantly activated upon SARS-CoV-2 infection22 (Fig. 51).

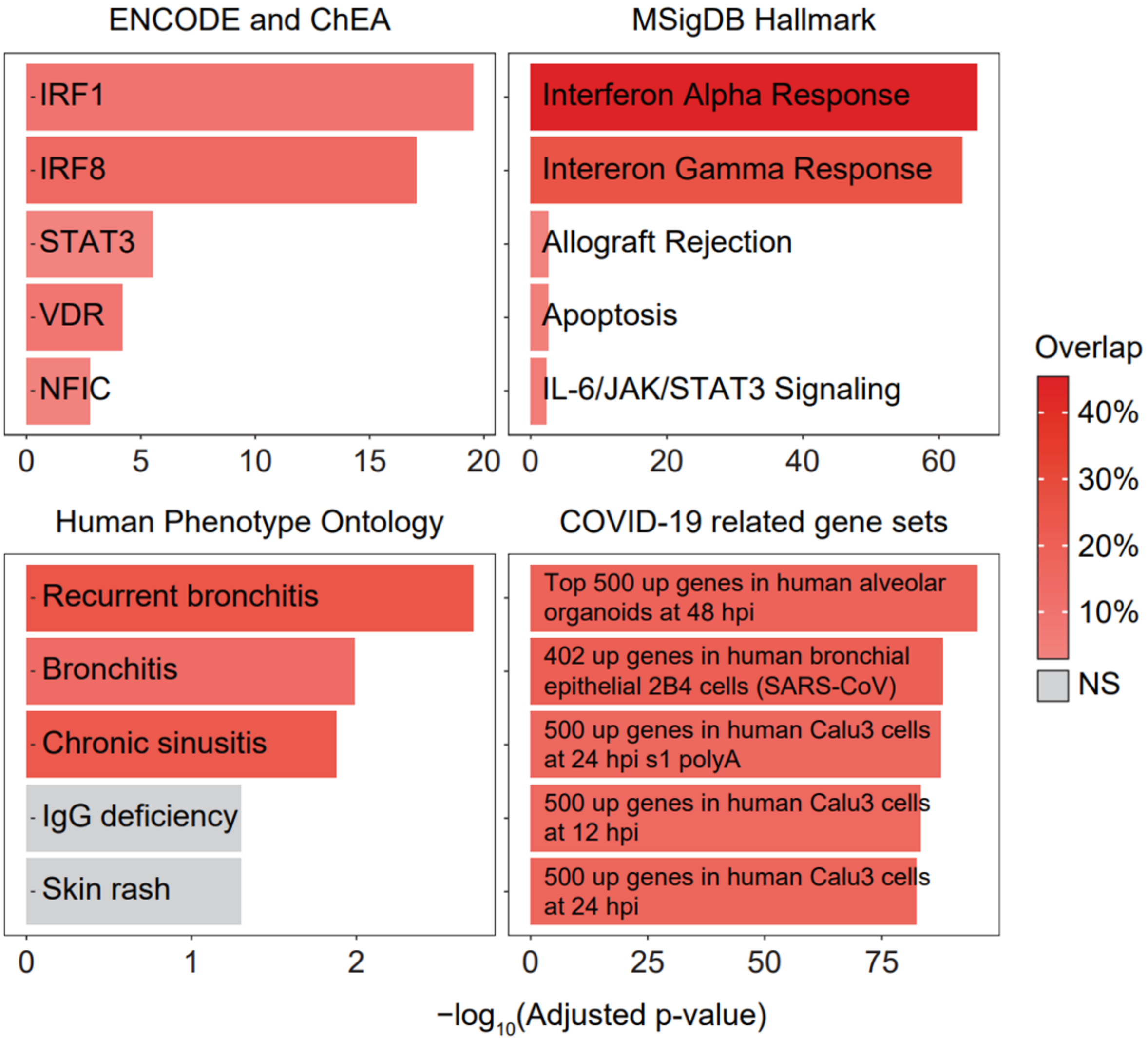

Figure 51. Pathway and gene set enrichment analysis performed for different databases using the gene signature previously identified. Each barplot shows the significance (x axis) and the percentage of overlap (fill colour) between the input signature and the tested public gene sets.

Interestingly, when looking at the expression levels of these genes in our cohort, negative patients displayed a transcriptional behaviour comparable to samples with the lowest viral load. As shown in the following Fig. 52, these data suggest that not only the gene signature we identified is characteristic of SARS-CoV-2 infected patients, but also that its expression levels correlate with the viral titre in patients' upper respiratory ways.

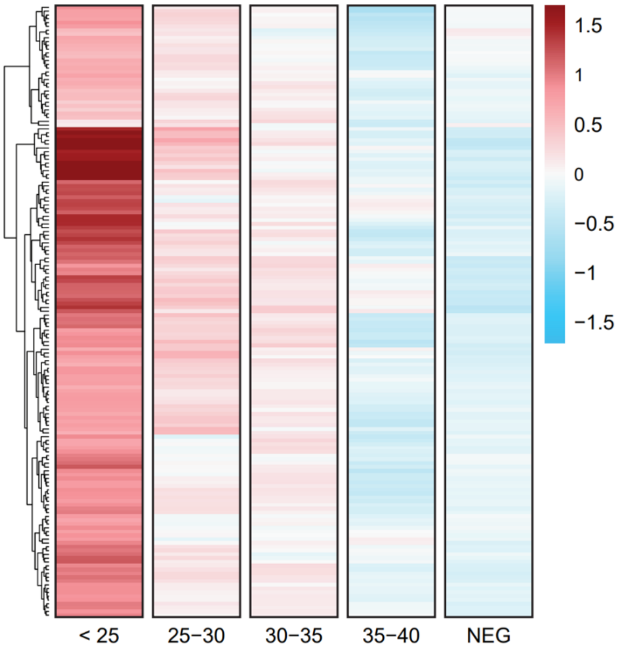

Figure 52. Heatmap of z-scored, log2-transformed and normalized gene counts for the 161 significantly correlated genes from Fig. 50. Values have been averaged in 4 groups of samples depending on the CT (x axis) or whether they were negative.

In order to confirm our finding and exclude that the gene signature identified was only typical of the ancestral SARS-CoV-2 variants, we applied the same approach to the Delta dataset. We thus retrieved a molecular signature of 16 genes (Fig. 52, left panel), way smaller than the previous one, most probably due to the restricted number of patients. Nevertheless, 81,25% of genes was common to the B.1 signature, belonged to the same pathways (IFIT3, OAS3, IFI6 - Fig. 4D) (Fig.53, central panel), and, as expected, their expression levels in negative patient were similar to the ones in patients with the lowest viral titre (Fig. 53, right panel).

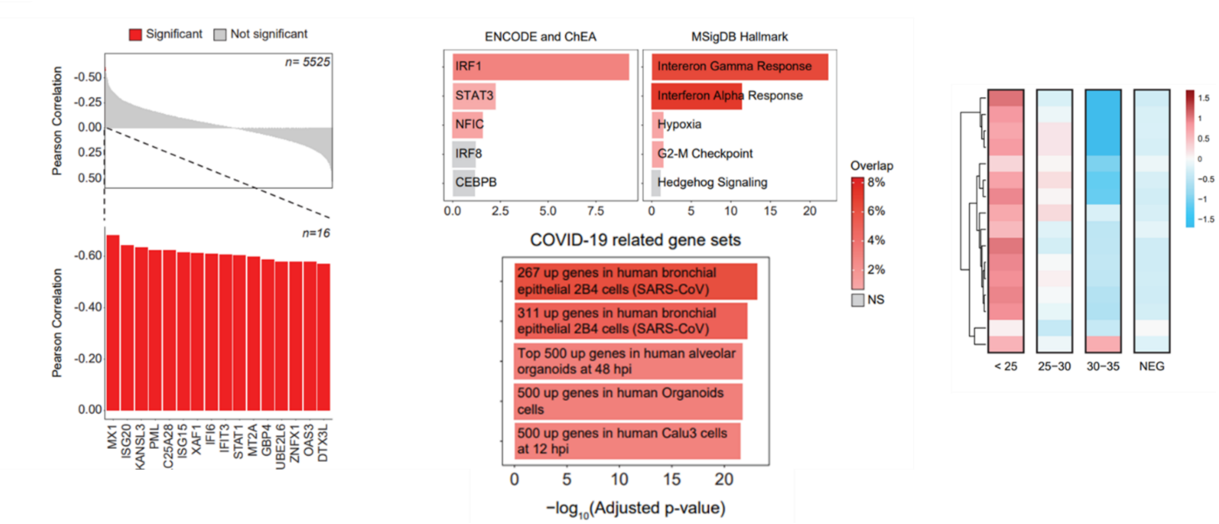

Figure 53. Gene signature analysis in patients infected by Delta variant. Left: correlation analysis between CTs and gene expression of Delta patients, performed on 5525 genes, is shown as a barplot. For each gene (x axis), its correlation value (y axis) and significance (p-value < 0.0001, red) are reported. The 16 significant genes are highlighted in the inset. Centre: pathway and gene set enrichment analysis performed for different databases using the gene signature previously identified. Each barplot shows the significance (x axis) and the percentage of overlap (fill colour) between the input signature and the tested public genesets. Left) Heatmap of z-scored, log2-transformed and normalized gene counts for the 16 significantly correlated genes from the analysis of Delta dataset. Values have been averaged in 3 groups of samples depending on the CT (x axis) or whether they were negative.

In conclusion, our CT-based approach overcomes all the technical and biological variability related to the direct use of regular swabs extracts. It establishes a robust gene signature preserved across different viral lineages and could be used as biomarkers for disease monitoring, prevention, and non-conventional treatments.

4. Discussion

4.1. Development of an NGS approach for easily monitoring SARS-CoV-2 genome variability

Genomic surveillance using Next Generation Sequencing approaches has proven its extreme efficacy in some of the most notorious outbreaks of the 21st century[132,134,162]. Indeed, the technology allows specifically identifying the pathogen genome variability directly from clinical specimens, not relying on the traditional and time-consuming isolation and in vitro cultivation steps. As further proof of its potential, genome sequencing is now considered the standard typing procedure for the influenza virus and has already been used for decision-making in terms of vaccine development by CDC26. Similarly, with the emergence of SARS-CoV-2 and its rapid evolution towards more and more infective variants[104,139] genomic surveillance had a critical role in monitoring virus evolution and detecting new mutations. Nevertheless, several countries still lack an efficient or homogenous integrated program for SARS-CoV-2 genome sequencing[163]. Such observation depends on several factors in a country-specific manner. The amount of resources NGS technologies rely on is probably an essential factor for low-income countries. However, at the end of 2020, some high-income countries also lacked a consistent sequencing program[163]. This was the case of Italy. While being the site of the first European COVID-19 outbreak, during 2020 genomic surveillance remained at low negligible levels in the country, and thus far below the 5% of all COVID-19 patients as recommended by ECDC[48]. Furthermore, as of August 2022, the number of sequences uploaded on GISAID is still highly inhomogeneous, with several regions that uploaded on GISAID less than 1000 genomes since the beginning of the pandemic. The approach we optimized aims at generating an affordable and easy program that can be translated to monitoring viral variability at regional level; a strategy that could allow emerging economies to perform efficient surveillance. This approach does not rely on any specific automation and can be implemented by a three-person team on any short-read sequencer. We proved its applicability to as few as 1.25 μL of unquantified RNA, enabling us to scale down the library reaction volumes and thus the costs while not affecting the sequencing metrics. The percentage of consensus sequences retrieved in each sequencing is only slightly affected when pooling up to 384 samples per flow cell, corresponding to 5 million reads per sample. Similar results were obtained when simulating 2,5 million reads per sample, thus demonstrating that, in principle, it is possible to sequence up to 1536 samples per run when using two flow cells in parallel.

We also tested a fast protocol which, by merging two PCR steps in one step, allows to speed up the library generation times by 40%. While in this case the general sequencing quality is lower, the solution still enables retrieving a consensus sequence for about 60% of the analyzed samples. Such a number is still enough for screening purposes to identify the main circulating variants and might be applied when time is a critical factor, as during major infection waves. It indeed allows to easily sequence over 4000 samples in one month, allowing the detection of variants with a frequency of about 1% and comply with ECDC requirements to sequence at least 5% of all positive cases.

4.2. Genomic surveillance allow the identification of new genetic variants

As proof of principle, we applied the framework to the 2021 genomic surveillance of the most densely populated region in Italy, Campania. As a result, the region is now the one with the highest number of sequences deposited on GISAID and one of the few to align to ECDC recommendation in 2021.

One of the biggest limitations of lacking microbiological surveillance relies on the inability to detect and isolate new emerging variants, increasing the chances of new waves of infections. The most adopted evolutionary model to study SARS-CoV-2 relies on the assumption that each lineage spread from an original ancestor originated in a given space and time. Therefore, profoundly profiling the pandemic dynamics at the regional level is critical for detecting such ancestors as soon as they start to emerge and spread. We demonstrated that new mutations can be identified even looking at a confined territory as Campania is. In addition, we observed several lineages potentially originating in Campania, including VOCs sub variants (as the P.1.1, BA.1.21.1) that spread world-wide. Some of these variants were designated by the pangolin system only after our alert. We indeed propose three principles to investigate and, potentially, define novel lineages. First, evaluate the presence of known pathogenic SNPs in unexpected lineages. Second, observe the co-occurrence of mutations with increasing frequency over time. Finally, look at the emergence of new mutations in prolonged hospitalized infections. The latter approach emphasizes the synergy between the healthcare centres, which provide clinical metadata, and the sequencing facilities that generate the viral consensus sequences. Thanks to these approaches, we were able to discover and describe three new lineages (B.1.177.88, Q.2, BA.1.21.1) and helped to guide local policymakers in the establishment of localized containment areas in the Region. For instance, the "Agro Nocerino-Sarnese" area was quarantined after pointing out the emergence of the B.1.177.88 variant; this decision prevented the spreading of the variant, and it disappeared a few weeks later.

4.3. Integrating clinical metadata and genomic data to identify intra-host viral evolution

Among the others, we also collected several samples from patients with persistent SARS-CoV-2 infections over time. Interestingly none of the patients completed an entire SARS-CoV-2 vaccination cycle, and the vast majority had no vaccination at all. In only one patient out of 20, we were able to actually detect the rising of a new mutation (in viral NTP/helicase NSP13 R339C) in the viral genome. The identification of this mutation can be associated with two possible events: i) the virus actually acquired a new mutation in the host or, less likely but possible, ii) the mutation pre-existed at low frequencies as part of the quasispecies infecting the host and was then fixed in the viral population. It is worth noting that the mutation identified is extremely rare world-wide and it was identified on GISAID only 23 times in the same variant of the patient under investigation (alpha). However, while this observation alone does not necessarily imply the identification of a new lineage, it strongly suggests that viral populations in patients with persistent infections can potentially evolve.

4.4. Main advantages and drawbacks of the proposed NGS workflow

In conclusion, in this study, we propose a cost-effective and rapid workflow for SARS-CoV-2 genome sequencing whose cost per sample is 5 times lower than the standard application for SARS-CoV-2 WGS (using solution B). Moreover, our approach is based on PCR enrichment and amplification of viral genomes, thus not requiring any specialized skill and suitable to be performed after a minimum training. Finally, the possibility to pool 384 samples or more in each sequencing flow cell, allows a 3-person team (two wet scientists and one bioinformatician) to deliver the sequences of over 760 samples in as few as 6 days (with ~ 5h of hands-on time). Taken all together, these properties make our approach not only highly valuable in monitoring the COVID-19 pandemic, as we showed at the regional level, but also easily transferable to other genomic centres.

The main limitation of the approach is its amplicon-based nature, which requires the monitoring of the primers used as the viral genome mutates over time. As shown in Fig. 29, indeed, a couple of our primers responsible for amplifying the genomic window between positions x-y display lower efficiency relatively to all the others. Nevertheless, our approach still allows a coverage higher than 30X for samples having a Ct value<35 and a 28X coverage for all the others. Such values are way higher than the one generally for confidently calling a base (10X)[158,164]. Furthermore, other strategies used for SARS-CoV-2 genome sequencing, e.g., probe-based enrichment and metagenome WGS, are either more time-consuming and expensive (probe-based enrichment) or deeply affected by host and microbial genetic material (metagenome).

The use of short-reads, as for all second-generation sequencing strategies, has a potential impact on the capability to discern viral recombination from patient co-infection, the former being a central feature in coronaviruses[48]. While such issue cannot be solved without a long-read approach, we argue that any possible spread of recombinant strains would be recognized by the co-occurrence of the same mutations associated with different variants in several samples, as proven by the detection of two XA recombinant variants in our dataset.

The general lack of bioinformatics skills required for raw data analysis is a critical factor for NGS technologies implementation in clinical diagnostic laboratories. While offering a simple and cheap approach for SARS-CoV-2 genome sequencing, our workflow also relies on the use of bioinformatics tools for data interpretation. We addressed this problem by developing a comprehensive pipeline that requires minimum informatics skills. Once started, the pipeline performs all the analysis required for the production of the consensus sequence and automatically performs the upload of high-quality sequences to GISAID.

4.5. Identification of variant-independent gene signature

Eventually, we identified molecular signatures from COVID-19 patients' gene expressions that agree with identified biomarkers reported in previous studies. Our approach extends the scope of SARS-CoV-2 genomic surveillance, as it allows for examining in-vivo samples characterized by the predominance of degraded RNA molecules. This competence enables overcoming the limitation of in-vitro and single-cell studies, such as model-specific variations and a small number of samples limit, respectively. Gene expression data from COVID-19 patients might have a pivotal role as a bridge between genomic data and translational medicine. On the one hand, finding a gene signature that describes and defines the patient status after SARS-CoV-2 infection may be helpful in understanding the pathogenesis of the virus in different patients and patients' status. On the other hand, it might be used to evaluate new treatments. In this study, we propose a cost-effective and rapid workflow to produce these data and retrieve biologically relevant biomarkers. Furthermore, the RNA-seq analysis implemented in our workflow offers a comparison between molecular signatures from RNAs of different SARS-COV-2 variants for the first time, proving that the transcriptional host response of the upper airways changes in the same direction, regardless of the viral variant they have been infected by. We also envision integrating this approach with other types of metadata (e.g., patient symptomatology) to achieve the goals mentioned above.

4.6. Conclusions

Here we developed a fast and cost-effective approach for SARS-CoV-2 genomic surveillance. The proposed strategy allows to scale viral genome sequencing down to 10 times less per sample. In addition, this protocol minimizes the hands-on time and does not require intensive training or any particular automation. Taken altogether, these features allowed us to profile the SARS-CoV-2 pandemic in Campania (Italy) during 2020-2021. We thus identified the main variants leading each infection wave in the regional territory and discovered 3 new SARS-CoV-2 lineages specifically originated in Campania, demonstrating the potential of genomic surveillance. We also added a further layer of information by integrating viral genotype with host upper respiratory airways transcriptome upon infection. This integrative point of view revealed a gene-expression signature correlated with viral loads and characterizing real-world infected patients. Finally, we showed that the host airway epithelium response to SARS-CoV-2 infection is not significantly different in B.1 and delta variant infected patients. In conclusion, we believe that the proposed approach can significantly help to fight against the pandemic by democratizing viral genome profiling through next-generation sequencing.

5. Appendix

	B.1.1.187	B.1.177.33	B.1.177.75	C.18	P.1.1
Italy (Campania)	87 (87)	506 (400)	658 (346)	348 (334)	2384 (637)
United Kingdom	14	64	18	2	9
Bulgaria	4	0	9	0	0
Finland	1	0	0	0	2
Argentina	0	0	0	0	4
Australia	0	1	0	0	0
Austria	0	1	4	0	0
Belgium	0	0	1	0	8
Bosnia and Herzegovina	0	0	0	0	2
Brazil	0	0	0	0	200
Chile	0	0	0	0	9
Colombia	0	0	0	0	2
Croatia	0	0	0	0	2
Denmark	0	1	6	1	6
Ecuador	0	0	0	0	3
France	0	3	1	5	13
French Guiana	0	0	0	0	1
Germany	0	16	243	0	289
Greece	0	0	0	0	1
Hong Kong	0	0	0	2	0
Iceland	0	7	2	0	0
Kosovo	0	0	1	0	0
Latvia	0	0	0	0	1
Lithuania	0	1	0	0	2
Luxembourg	0	1	1	0	46
Malta	0	0	0	0	7
Mexico	0	0	0	0	9
Norway	0	1	0	0	1
Peru	0	0	0	0	2
Philippines	0	0	0	0	1
Poland	0	0	0	0	12
Portugal	0	0	0	0	1
Romania	0	1	0	0	7
Serbia	0	0	2	0	0
Slovenia	0	2	0	3	3
South Korea	0	0	1	0	0
Spain	0	6	6	1	6
Sweden	0	1	0	1	0
Switzerland	0	12	19	11	12

	B.1.1.187	B.1.177.33	B.1.177.75	C.18	P.1.1
Thailand	0	0	0	0	1
Turkey	0	0	0	0	4
USA	0	0	0	0	94
Total	106	624	972	374	3144

Appendix Table 1

Gene	Correlation	p-value
IFI44L	-0,72377	2,72E-18
OAS2	-0,70729	1,66E-17
PARP9	-0,67231	2,60E-16
ISG15	-0,66887	5,66E-16
IFITM3	-0,66217	5,77E-16
IFIT1	-0,64625	4,16E-11
RSAD2	-0,6341	1,96E-11
RNF213	-0,62904	3,98E-14
XAF1	-0,61478	3,43E-13
IFI6	-0,61412	2,90E-13
ZNFX1	-0,61178	3,78E-13
DDX60L	-0,61158	4,20E-12
IFI44	-0,59707	2,13E-10
PARP14	-0,58377	6,08E-12
IFIT3	-0,5816	6,25E-11
HERC6	-0,58129	1,00E-09
MX1	-0,57415	1,97E-11
OAS3	-0,56969	6,18E-10
BST2	-0,56323	1,23E-10
OAS1	-0,5616	1,74E-10
SP100	-0,56117	5,67E-11
MX2	-0,55456	5,74E-10
GBP1	-0,55423	2,31E-10
HELZ2	-0,55367	4,05E-09
TRIM22	-0,54955	1,27E-09
IFIT2	-0,54894	7,01E-09
EPSTI1	-0,54688	5,30E-10
CMPK2	-0,54357	2,58E-08
STAT1	-0,53709	5,11E-10
IRF7	-0,52662	6,97E-08
UBE2L6	-0,52415	1,83E-09
NMI	-0,52251	9,19E-09
PARP12	-0,51299	3,77E-07
SAMD9	-0,51105	1,17E-08
NUB1	-0,50752	7,09E-09
EIF2AK2	-0,50716	7,29E-09
IFI27	-0,50542	8,36E-09
ZC3HAV1	-0,50519	8,51E-09
DDX60	-0,50337	5,14E-08
DTX3L	-0,5004	1,44E-08
PLSCR1	-0,5	7,56E-08
SLFN5	-0,49954	1,32E-08
C19orf66	-0,49884	9,49E-08
IFIH1	-0,49745	3,38E-07
LY6E	-0,49681	1,41E-08
IFI35	-0,49412	1,75E-07
PSMA3-AS1	-0,48864	5,30E-08
SMCHD1	-0,48183	7,50E-08
SP110	-0,47876	2,09E-07
XRN1	-0,47233	1,46E-07
BAZ1A	-0,47132	9,29E-08
PARP10	-0,47063	7,13E-06
HERC5	-0,46993	5,00E-06
ITGAL	-0,46538	8,18E-06
SP140L	-0,46334	2,37E-07
STAT2	-0,46247	6,01E-07
LARP7	-0,46092	1,92E-07
ZNF337	-0,45854	1,15E-05
ESF1	-0,45667	2,90E-07
LAP3	-0,45628	5,45E-07
ISG20	-0,45484	1,09E-06
HPS5	-0,45285	1,56E-06
GBP4	-0,45278	6,82E-07
NLRC5	-0,45145	9,40E-07
IRF1	-0,44564	5,36E-07
IFI16	-0,44482	6,34E-07
ZZZ3	-0,44251	2,56E-06
ARID4B	-0,44126	7,13E-07
IRF9	-0,43905	2,94E-05
SPATS2L	-0,43433	1,24E-06
TYMP	-0,43139	1,49E-06
BLNK	-0,43118	2,74E-05
LINC00685	-0,42721	3,31E-05
TAP1	-0,42338	4,99E-06
UTRN	-0,42245	2,32E-06
TAP2	-0,42033	5,93E-06
BCL2	-0,41823	6,15E-05
MLLT6	-0,41814	3,32E-06
NRDE2	-0,41754	7,66E-06
FUBP1	-0,41654	3,65E-06
GPATCH8	-0,41372	5,24E-06
SECTM1	-0,41324	2,35E-05
ZC3H13	-0,4118	4,38E-06
PPM1K	-0,41164	9,60E-06
USP15	-0,4116	8,73E-06
GALM	-0,41147	6,57E-06

Gene	Correlation	p-value
LGALS9	-0,41118	4,54E-06
APOL6	-0,41017	4,81E-06
PHF11	-0,40992	8,70E-06
CD2	-0,40851	3,60E-05
C9orf114	-0,40774	1,19E-05
AMMECR1	-0,40688	5,18E-05
SPN	-0,40645	3,28E-05
IARS	-0,4059	2,79E-05
SAMD9L	-0,4057	8,28E-06
IL2RG	-0,40569	3,41E-05
PML	-0,40302	1,53E-05
TNFSF10	-0,40282	1,07E-05
FYB	-0,39961	1,16E-05
LRIF1	-0,39802	5,41E-05
PILRB	-0,39746	2,05E-05
ZC3H15	-0,39702	1,02E-05
TRIM25	-0,3963	1,99E-05
CDK11A	-0,39593	2,03E-05
NKTR	-0,39556	1,11E-05
PPP4R2	-0,39491	1,50E-05
UBA7	-0,3947	9,04E-05
ABCF1	-0,39432	1,19E-05
NF1	-0,39417	3,18E-05
HLA-E	-0,39277	1,29E-05
HRASLS2	-0,39263	7,60E-05
DDX27	-0,39102	1,55E-05
TARS	-0,39063	2,06E-05
SART3	-0,3901	1,94E-05
ARAP2	-0,38942	2,85E-05
AKAP13	-0,38872	1,62E-05
TMC6	-0,38854	7,70E-05
ATM	-0,38829	6,55E-05
CARD16	-0,38775	6,72E-05
CHORDC1	-0,38755	6,23E-05
KMT2B	-0,38455	3,09E-05
RUBCN	-0,38422	7,93E-05
BTN3A1	-0,38204	9,54E-05
MAU2	-0,38121	4,69E-05
RBCK1	-0,38058	3,78E-05
UBR2	-0,38058	5,26E-05
BDP1	-0,37853	2,79E-05
ZCCHC2	-0,37845	4,21E-05
DFFA	-0,37767	4,38E-05
USP8	-0,37764	2,93E-05
BAZ2A	-0,37647	3,11E-05
RAB10	-0,37588	3,77E-05
WARS	-0,37583	4,09E-05
GIGYF2	-0,37509	3,35E-05
HSP90AB1	-0,37464	3,42E-05
SUPT16H	-0,37359	3,91E-05
SRRM1	-0,37194	3,94E-05
ADAR	-0,37151	4,71E-05
LUC7L3	-0,37126	4,08E-05
TRIM14	-0,369	9,93E-05
DDX18	-0,36841	4,72E-05
TOP1	-0,36542	5,50E-05
STK4	-0,36503	5,60E-05
CWF19L2	-0,36437	5,79E-05
PRRC2C	-0,36427	5,82E-05
ELF1	-0,36294	6,23E-05
RANBP2	-0,36234	9,31E-05
SPEN	-0,36204	7,56E-05
HNRNPH3	-0,36115	7,33E-05
BOD1L1	-0,36104	6,85E-05
RSBN1L	-0,36011	7,17E-05
SETX	-0,3601	8,31E-05
ANKRD12	-0,35998	7,22E-05
KMT2E	-0,35945	7,41E-05
TAOK1	-0,35888	8,20E-05
SUZ12	-0,35721	9,56E-05
SYNRG	-0,35508	9,19E-05
PSMB8	-0,35471	9,35E-05
PPP2R2A	-0,3537	9,83E-05
SMARCA2	-0,35366	9,84E-05
PRPF38B	-0,35358	9,88E-05

Appendix Table 2

6. Bibliography

1. Kendall, E. J. C. *et al.* Virus Isolations from Common Colds Occurring in a Residential School. *Br Med J* **2**, 82 (1962).

2. Tyrrell, D. A. J., Bynoe, M. L., &h, D. T. M., Obst, D. R. C. O. G. & Brit, M. Cultivation of a Novel Type of Common-cold Virus in Organ Cultures. *Br Med J* **1**, 1467 (1965).

3. Mcintosh, K., Dees, J. H., Becker, W. B., Kapikian, A. Z. & Chanock, R. M. RECOVERY IN TRACHEAL ORGAN CULTURES OF NOVEL VIRUSES FROM PATIENTS WITH RESPIRATORY DISEASE.

4. Hamre, D. & Procknow, J. J. A new virus isolated from the human respiratory tract. *Proc Soc Exp Biol Med* **121**, 190–193 (1966).

5. J. D. Almeida; D . M. Berry; C. H. Cunningham; D. Hamre; M. S. Hofstad; L. Mallucci; K. McIntosh; D. A. J. Tyrrell. Virology: Coronaviruses. *Nature 1968 220:5168* **220**, 650–650 (1968).

6. Almeida, J. D. & Tyrrell, D. A. The morphology of three previously uncharacterized human respiratory viruses that grow in organ culture. *J Gen Virol* **1**, 175–178 (1967).

7. GLEDHILL, A. W. & ANDREWES, C. H. A Hepatitis Virus of Mice. *Br J Exp Pathol* **32**, 559 (1951).

8. Tyrrell, D. A. J. *et al.* Coronaviridae. *Intervirology* **5**, 76 (1975).

9. Walker, P. J. *et al.* Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch Virol* **166**, 2633–2648 (2021).

10. Zhou, Z., Qiu, Y. & Ge, X. The taxonomy, host range and pathogenicity of coronaviruses and other viruses in the Nidovirales order. *Animal Diseases 2021 1:1* **1**, 1–28 (2021).

11. Peiris, J. S. M. *et al.* Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361**, 1319 (2003).

12. Liu, D. X., Liang, J. Q. & Fung, T. S. Human Coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae). *Encyclopedia of Virology* 428 (2021) doi:10.1016/B978-0-12-809633-8.21501-X.

13. Ye, Z. W. *et al.* Zoonotic origins of human coronaviruses. *Int J Biol Sci* **16**, 1686 (2020).

14. van der Hoek, L. *et al.* Identification of a new human coronavirus. *Nat Med* **10**, 368–373 (2004).

15. Woo, P. C. Y. *et al.* Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. *J Virol* **79**, 884–895 (2005).

16. Zaki, A. M., van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *New England Journal of Medicine* **367**, 1814–1820 (2012).

17. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog* **13**, e1006698 (2017).

18. Li, H. *et al.* Human-animal interactions and bat coronavirus spillover potential among rural residents in Southern China. *Biosaf Health* **1**, 84–90 (2019).

19. Lin, X. D. *et al.* Extensive diversity of coronaviruses in bats from China. *Virology* **507**, 1 (2017).

20. Fan, Y., Zhao, K., Shi, Z. L. & Zhou, P. Bat Coronaviruses in China. *Viruses 2019, Vol. 11, Page 210* **11**, 210 (2019).

21. Fung, T. S. & Liu, D. X. Human Coronavirus: Host-Pathogen Interaction. *https://doi.org/10.1146/annurev-micro-020518-115759* **73**, 529–557 (2019).

22. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine* **382**, 727–733 (2020).

23. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

24. Liu, Y. C., Kuo, R. L. & Shih, S. R. COVID-19: The first documented coronavirus pandemic in history. *Biomed J* **43**, 328–333 (2020).

25. Ritchie, H. *et al.* Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).

26. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).

27. Chan, J. F. W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* **395**, 514–523 (2020).

28. 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).

29. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *The Lancet* **395**, 470–473 (2020).

30. WHO announces simple, easy-to-say labels for SARS-CoV-2 Variants of Interest and Concern. https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern.

31. Therapeutics and COVID-19: living guideline. *Therapeutics and COVID-19: living guideline* (2022).

32. Guimarães, P. O. *et al.* Features, Evaluation, and Treatment of Coronavirus (COVID-19). *New England Journal of Medicine* **385**, 406–415 (2022).

33. Berlin, D. A., Gulick, R. M. & Martinez, F. J. Severe Covid-19. *New England Journal of Medicine* **383**, 2451–2460 (2020).

34. Fan, E., Brodie, D. & Slutsky, A. S. Acute Respiratory Distress Syndrome: Advances in Diagnosis and Treatment. *JAMA* **319**, 698–710 (2018).

35. Saban, M., Myers, V. & Wilf-Miron, R. Changes in infectivity, severity and vaccine effectiveness against delta COVID-19 variant ten months into the vaccination program: The Israeli case. *Prev Med (Baltim)* **154**, 106890 (2022).

36. Hall, V. *et al.* Protection against SARS-CoV-2 after Covid-19 Vaccination and Previous Infection. *New England Journal of Medicine* **386**, 1207–1220 (2022).

37. Tartof, S. Y. *et al.* Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study. *Lancet* **398**, 1407 (2021).

38. Chi, W. Y. *et al.* COVID-19 vaccine update: vaccine effectiveness, SARS-CoV-2 variants, boosters, adverse effects, and immune correlates of protection. *Journal of Biomedical Science 2022 29:1* **29**, 1–27 (2022).

39. Andrews, N. *et al.* Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. *New England Journal of Medicine* **386**, 1532–1546 (2022).

40. Nyberg, T. *et al.* Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: a cohort study. *The Lancet* **399**, 1303–1312 (2022).

41. Gao, Y. dong *et al.* Risk factors for severe and critically ill COVID-19 patients: A review. *Allergy* **76**, 428–455 (2021).

42. Berenguer, J. *et al.* Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain. *Clin Microbiol Infect* **26**, 1525–1536 (2020).

43. Gemmati, D. *et al.* COVID-19 and Individual Genetic Susceptibility/Receptivity: Role of ACE1/ACE2 Genes, Immunity, Inflammation and Coagulation. Might the Double X-chromosome in Females Be Protective against SARS-CoV-2 Compared to the Single X-Chromosome in Males? *Int J Mol Sci* **21**, (2020).

44. Delshad, M., Sanaei, M.-J., Pourbagheri-Sigaroodi, A. & Bashash, D. Host genetic diversity and genetic variations of SARS-CoV-2 in COVID-19 pathogenesis and the effectiveness of vaccination. *Int Immunopharmacol* **111**, 109128 (2022).

45. Pekar, J. E. *et al.* The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. *Science* eabp8337 (2022) doi:10.1126/SCIENCE.ABP8337/SUPPL_FILE/SCIENCE.ABP8337_DATA_S1_TO_S3.ZIP.

46. Lam, T. T. Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature 2020 583:7815* **583**, 282–285 (2020).

47. Frutos, R., Serra-Cobo, J., Chen, T. & Devaux, C. A. COVID-19: Time to exonerate the pangolin from the transmission of SARS-CoV-2 to humans. *Infection, Genetics and Evolution* **84**, 104493 (2020).

48. Li, X. *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv* **6**, (2020).

49. Zhou, P. *et al.* Addendum: A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature 2020 588:7836* **588**, E6–E6 (2020).

50. Temmam, S. *et al.* Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature 2022 604:7905* **604**, 330–336 (2022).

51. Forni, D., Cagliani, R., Clerici, M. & Sironi, M. Molecular Evolution of Human Coronavirus Genomes. *Trends Microbiol* **25**, 35 (2017).

52. Artika, I. M., Dewantari, A. K. & Wiyatno, A. Molecular biology of coronaviruses: current knowledge. *Heliyon* **6**, e04743 (2020).

53. Worobey, M. *et al.* The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic. *Science* abp8715 (2022) doi:10.1126/SCIENCE.ABP8715/SUPPL_FILE/SCIENCE.ABP8715_MDAR_REPRODUCIBILITY_CHECKLIST.PDF.

54. Freuling, C. M. *et al.* Susceptibility of Raccoon Dogs for Experimental SARS-CoV-2 Infection. *Emerg Infect Dis* **26**, 2982 (2020).

55. Lefkowitz, E. J. *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* **46**, D708–D717 (2018).

56. Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730-738.e13 (2020).

57. Fehr, A. R. & Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Coronaviruses* **1282**, 1 (2015).

58. Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730-738.e13 (2020).

59. Zhu, C. *et al.* Molecular biology of the SARs-CoV-2 spike protein: A review of current knowledge. *J Med Virol* **93**, 5729 (2021).

60. Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894-904.e9 (2020).

61. Huang, Y., Yang, C., Xu, X. feng, Xu, W. & Liu, S. wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica 2020 41:9* **41**, 1141–1149 (2020).

62. Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **50**, W276–W279 (2022).

63. Schuurs, Z. P. *et al.* Evidence of a putative glycosaminoglycan binding site on the glycosylated SARS-CoV-2 spike protein N-terminal domain. *Comput Struct Biotechnol J* **19**, 2806–2818 (2021).

64. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369**, 650–655 (2020).

65. Xia, X. Domains and Functions of Spike Protein in SARS-Cov-2 in the Context of Vaccine Design. *Viruses* **13**, (2021).

66. Ke, Z. *et al.* Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature 2020 588:7838* **588**, 498–502 (2020).

67. Hwang, S. S. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1255–1260 (2020).

68. Kuzmin, A., Orekhov, P., Astashkin, R., Gordeliy, V. & Gushchin, I. Structure and dynamics of the SARS-CoV-2 envelope protein monomer. *Proteins: Structure, Function, and Bioinformatics* **90**, 1102–1114 (2022).

69. Zhang, Z. *et al.* Structure of SARS-CoV-2 membrane protein essential for virus assembly. *Nature Communications 2022 13:1* **13**, 1–12 (2022).

70. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nature Communications 2021 12:1* **12**, 1–17 (2021).

71. Savastano, A., Ibáñez de Opakua, A., Rankovic, M. & Zweckstetter, M. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nat Commun* **11**, (2020).

72. Woo, P. C. Y., Lau, S. K. P., Huang, Y. & Yuen, K. Y. Coronavirus Diversity, Phylogeny and Interspecies Jumping: *https://doi.org/10.3181/0903-MR-94* **234**, 1117–1127 (2009).

73. Chaitanya, K. v. Structure and Organization of Virus Genomes. *Genome and Genomics* 1 (2019) doi:10.1007/978-981-15-0702-1_1.

74. Hartenian, E. *et al.* The molecular virology of coronaviruses. *Journal of Biological Chemistry* vol. 295 12910–12934 Preprint at https://doi.org/10.1074/jbc.REV120.013930 (2020).

75. Cao, C. *et al.* The architecture of the SARS-CoV-2 RNA genome inside virion. *Nature Communications 2021 12:1* **12**, 1–14 (2021).

76. Yang, D. & Leibowitz, J. L. The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus Res* **206**, 120 (2015).

77. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H. & Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology 2020 19:3* **19**, 155–170 (2020).

78. Tanaka, T., Kamitani, W., DeDiego, M. L., Enjuanes, L. & Matsuura, Y. Severe acute respiratory syndrome coronavirus nsp1 facilitates efficient propagation in cells through a specific translational shutoff of host mRNA. *J Virol* **86**, 11128–11137 (2012).

79. Huang, C. *et al.* Alphacoronavirus transmissible gastroenteritis virus nsp1 protein suppresses protein translation in mammalian cells and in cell-free HeLa cell extracts but not in rabbit reticulocyte lysate. *J Virol* **85**, 638–643 (2011).

80. Zou, L., Moch, C., Graille, M. & Ment Chapat, C. The SARS-CoV-2 protein NSP2 impairs the silencing capacity of the human 4EHP-GIGYF2 complex. *iScience* **25**, 104646 (2022).

81. Graham, R. L., Sims, A. C., Baric, R. S. & Denison, M. R. The nsp2 proteins of mouse hepatitis virus and SARS coronavirus are dispensable for viral replication. *Adv Exp Med Biol* **581**, 67–72 (2006).

82. Cornillez-Ty, C. T., Liao, L., Yates, J. R., Kuhn, P. & Buchmeier, M. J. Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 2 Interacts with a Host Protein Complex Involved in Mitochondrial Biogenesis and Intracellular Signaling. *J Virol* **83**, 10314–10318 (2009).

83. Neuman, B. W. *et al.* Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. *J Virol* **82**, 5279–5294 (2008).

84. Egloff, M.-P. *et al.* Structural and Functional Basis for ADP-Ribose and Poly(ADP-Ribose) Binding by Viral Macro Domains. *J Virol* **80**, 8493–8502 (2006).

85. Ricciardi, S. *et al.* The role of NSP6 in the biogenesis of the SARS-CoV-2 replication organelle. *Nature 2022 606:7915* **606**, 761–768 (2022).

86. Lu, Y., Lu, X. & Denison, M. R. Identification and characterization of a serine-like proteinase of the murine coronavirus MHV-A59. *J Virol* **69**, 3554–3559 (1995).

87. Roe, M. K., Junod, N. A., Young, A. R., Beachboard, D. C. & Stobart, C. C. Targeting novel structural and functional features of coronavirus protease nsp5 (3CLpro, Mpro) in the age of COVID-19. *J Gen Virol* **102**, 1558 (2021).

88. Zhai, Y. *et al.* Insights into SARS-CoV transcription and replication from the structure of the nsp7-nsp8 hexadecamer. *Nat Struct Mol Biol* **12**, 980–986 (2005).

89. Zeng, Z. *et al.* Dimerization of Coronavirus nsp9 with Diverse Modes Enhances Its Nucleic Acid Binding Affinity. *J Virol* **92**, 692–710 (2018).

90. Bouvet, M. *et al.* In vitro reconstitution of SARS-coronavirus mRNA cap methylation. *PLoS Pathog* **6**, 1–13 (2010).

91. Jiang, Y., Yin, W. & Xu, H. E. RNA-dependent RNA polymerase: Structure, mechanism, and drug discovery for COVID-19. *Biochem Biophys Res Commun* **538**, 47–53 (2021).

92. Romano, M., Ruggiero, A., Squeglia, F., Maga, G. & Berisio, R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. *Cells* **9**, (2020).

93. Eckerle, L. D. *et al.* Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog* **6**, 1–15 (2010).

94. Hackbart, M., Deng, X. & Baker, S. C. Coronavirus endoribonuclease targets viral polyuridine sequences to evade activating host sensors. *Proc Natl Acad Sci U S A* **117**, 8094–8103 (2020).

95. Decroly, E. *et al.* Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2'O)-methyltransferase activity. *J Virol* **82**, 8071–8084 (2008).

96. Sola, I., Mateos-Gomez, P. A., Almazan, F., Zuñiga, S. & Enjuanes, L. RNA Biology RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. (2011) doi:10.4161/rna.8.2.14991.

97. Harrison, A. G., Lin, T. & Wang, P. Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. *Trends Immunol* **41**, 1100 (2020).

98. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, (2015).

99. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology 2021 23:1* **23**, 3–20 (2021).

100. Ghosh, S. *et al.* β-Coronaviruses Use Lysosomes for Egress Instead of the Biosynthetic Secretory Pathway. *Cell* **183**, 1520 (2020).

101. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

102. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403–1407 (2020).

103. Li, X. & Hospital, H. M. Concerns on the multiple nomenclature systems for SARS-CoV-2. *J Med Virol* **94**, 1224–1226 (2022).

104. O'Toole, Á., Pybus, O. G., Abram, M. E., Kelly, E. J. & Rambaut, A. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* **23**, 1–13 (2022).

105. Baric, R. S. Emergence of a Highly Fit SARS-CoV-2 Variant. *New England Journal of Medicine* **383**, 2684–2686 (2020).

106. Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *Cell* **183**, 739 (2020).

107. WT, H. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **19**, 409–424 (2021).

108. Konings, F. *et al.* SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. *Nature Microbiology 2021 6:7* **6**, 821–823 (2021).

109. COVID-19 Weekly Epidemiological Update. https://www.who.int/publications/m/item/covid-19-weekly-epidemiological-update.

110. Thakur, V. *et al.* Waves and variants of SARS-CoV-2: understanding the causes and effect of the COVID-19 catastrophe. *Infection* **50**, 309–325 (2022).

111. Rambaut, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological* (2020).

112. Liu, Y. *et al.* The N501Y spike substitution enhances SARS-CoV-2 infection and transmission. *Nature* **602**, 294 (2022).

113. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).

114. Rani Rajpal, V. *et al.* A comprehensive account of SARS-CoV-2 genome structure, incurred mutations, lineages and COVID-19 vaccination program. *https://doi.org/10.2217/fvl-2021-0277* **17**, 687–706 (2022).

115. Planas, D. *et al.* Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. *Nat Med* 1–8 (2021) doi:10.1038/s41591-021-01318-5.

116. Chen, R. E. *et al.* Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nat Med* (2021) doi:10.1038/s41591-021-01294-w.

117. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J* **18**, 9–19 (2020).

118. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1 (2016).

119. Modi, A., Vai, S., Caramelli, D. & Lari, M. The Illumina Sequencing Protocol and the NovaSeq 6000 System. *Methods in Molecular Biology* **2242**, 15–42 (2021).

120. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science (1979)* **327**, 78–81 (2010).

121. Blanco, L. *et al.* Highly Efficient DNA Synthesis by the Phage φ 29 DNA Polymerase: Symmetrical Mode of DNA Replication. *Journal of Biological Chemistry* **264**, 8935–8940 (1989).

122. Pugh, T. J. *et al.* Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Res* **36**, e80 (2008).

123. Senabouth, A. *et al.* Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genom Bioinform* **2**, (2020).

124. Naval-Sanchez, M. *et al.* Benchmarking of ATAC Sequencing Data From BGI's Low-Cost DNBSEQ-G400 Instrument for Identification of Open and Occupied Chromatin Regions. *Front Mol Biosci* **9**, 900323 (2022).

125. Rao, J. *et al.* Performance of copy number variants detection based on whole-genome sequencing by DNBSEQ platforms. *BMC Bioinformatics* **21**, (2020).

126. Jeon, S. A. *et al.* Comparison between MGI and Illumina sequencing platforms for whole genome sequencing. **43**, 713–724 (2021).

127. Ari, Ş. & Arikan, M. Next-generation sequencing: Advantages, disadvantages, and future. *Plant Omics: Trends and Applications* 109–135 (2016) doi:10.1007/978-3-319-31703-8_5/TABLES/1.

128. González-Candelas, F., Francés-Cuesta, C. & García-González, N. The power and limitations of genomic surveillance of bacteria. *https://doi.org/10.2217/fmb-2019-0259* **14**, 1345–1348 (2019).

129. Hill, V., Ruis, C., Bajaj, S., Pybus, O. G. & Kraemer, M. U. G. Progress and challenges in virus genomic epidemiology. *Trends Parasitol* **37**, 1038–1049 (2021).

130. Mate, S. E. *et al.* Molecular Evidence of Sexual Transmission of Ebola Virus. *New England Journal of Medicine* **373**, 2448–2454 (2015).

131. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309 (2017).

132. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369 (2014).

133. Faria, N. R. *et al.* Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature 2017 546:7658* **546**, 406–410 (2017).

134. Giovanetti, M. *et al.* Genomic and Epidemiological Surveillance of Zika Virus in the Amazon Region. *Cell Rep* **30**, 2275-2283.e7 (2020).

135. Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature 2017 546:7658* **546**, 401–405 (2017).

136. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 1 (2017).

137. Sharif, N. *et al.* Genomic surveillance, evolution and global transmission of SARS-CoV-2 during 2019–2022. *PLoS One* **17**, e0271074 (2022).

138. Faria, N. R. *et al.* Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological.Org* 1–9 (2021).

139. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).

140. Tegally, H. *et al.* Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. *Nature* (2021) doi:10.1038/s41586-021-03402-9.

141. Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J Transl Med* **18**, 179 (2020).

142. Ema-ecdc. ECDC-EMA statement on booster vaccination with Omicron adapted bivalent COVID-19 vaccines.

143. Adapted vaccine targeting BA.4 and BA.5 Omicron variants and original SARS-CoV-2 recommended for approval | European Medicines Agency. https://www.ema.europa.eu/en/news/adapted-vaccine-targeting-ba4-ba5-omicron-variants-original-sars-cov-2-recommended-approval.

144. Hoagland, D. A. *et al.* Modulating the transcriptional landscape of SARS-CoV-2 as an effective method for developing antiviral compounds. *bioRxiv* 2020.07.12.199687 (2020) doi:10.1101/2020.07.12.199687.

145. Katsura, H. *et al.* Human Lung Stem Cell-Based Alveolospheres Provide Insights into SARS-CoV-2-Mediated Interferon Responses and Pneumocyte Dysfunction. *Cell Stem Cell* **27**, (2020).

146. E, W. *et al.* Transcriptomic profiling of SARS-CoV-2 infected human cell lines identifies HSP90 as target for COVID-19 therapy. *iScience* **24**, (2021).

147. Davidi, D. *et al.* Amplicon residues in research laboratories masquerade as COVID-19 in surveillance tests. *Cell Reports Methods* **1**, 100005 (2021).

148. Mifflin, T. E. Setting up a PCR laboratory. *CSH Protoc* **2007**, pdb.top14 (2007).

149. Xiao, M. *et al.* Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome Med* **12**, 1–15 (2020).

150. GitHub - MGI-tech-bioinformatics/SARS-CoV-2_Multi-PCR_v1.0: SARS-CoV-2 analysis pipeline for multiplex-PCR MPS(Massive Parrallel Sequencing) data. https://github.com/MGI-tech-bioinformatics/SARS-CoV-2_Multi-PCR_v1.0.

151. di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319 (2017).

152. Ma, F. *et al.* A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library preparation methods 06 Biological Sciences 0604 Genetics. *BMC Genomics* **20**, (2019).

153. BBMap download | SourceForge.net. https://sourceforge.net/projects/bbmap/.

154. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

155. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

156. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

157. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, (2021).

158. Gerber, Z. *et al.* A comparison of high-throughput SARS-CoV-2 sequencing methods from nasopharyngeal samples. *Scientific Reports 2022 12:1* **12**, 1–8 (2022).

159. Guruprasad, L. Human SARS CoV-2 spike protein mutations. *Proteins: Structure, Function and Bioinformatics* (2021) doi:10.1002/prot.26042.

160. Corey, L. *et al.* SARS-CoV-2 Variants in Patients with Immunosuppression. *New England Journal of Medicine* **385**, 562–566 (2021).

161. Mulay, A. *et al.* SARS-CoV-2 infection of primary human lung epithelium for COVID-19 modeling and drug discovery Graphical abstract. *CellReports* **35**, 109055 (2021).

162. Andersen, K. G. *et al.* Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. *Cell* **162**, 738–750 (2015).

163. Chen, Z. *et al.* Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics 2022 54:4* **54**, 499–507 (2022).

164. Liu, T. *et al.* A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples. *iScience* **24**, 102892 (2021).

165. Ismail, A. M. & Elfiky, A. A. SARS-CoV-2 spike behavior in situ: a Cryo-EM images for a better understanding of the COVID-19 pandemic. *Signal Transduction and Targeted Therapy 2020 5:1* **5**, 1–2 (2020).

7. Figures and Figure legends

Figure 1. Electron micrograph of negative stained virions showing the global differences between Coronaviruses (B814, left) and Orthomyxoviruses (Influenza A2, right). While the size of the two virions is comparable, the spikes are longer and fewer in the former. Modified from Almeida et al. (1966)[6] and McIntosh et al. (1967)[3]. 3

Figure 2. Maximum likelihood tree based on RdRP proteins in Coronaviruses. The tree shows the classification of Coronaviruses in 4 main genera (*Alpha-*, *Beta-*, *Gamma*-and *Deltacoronavirus*) and the corresponding distribution of human-infecting viruses (red). For each virus, the accession ID of a representative genomic sequenced is shown. Colours represents the host and the number near each node are the bootstrap values. The bar indicates the genetic distance (number of substitutions per RdRPp residue). Modified from Zhou et al. (2021)[10]. 5

Figure 3. Worldwide cumulative confirmed COVID-19 cases (left) and deaths (right) updated to August 30, 2022. Due to limited testing, variability in diagnostic protocols, or difficulties in attributing the cause of death, these numbers are probably an underestimation. Modified from https://ourworldindata.org/coronavirus[25] 6

Figure 4. COVID-19 spread from its first detection in December 2019 till the end of August 2022. The figure shows some of the most important events during the pandemic. Modified from Safiabadi et al. (2021)[28]. 7

Figure 5. Scheme showing COVID-19 classification based on disease severity. Each category has a defining set of diagnostic features and requires specific treatments. Modified from "*Therapeutics and COVID-19: living guideline*"[31]9

Figure 6. Main severe COVID-19 risk factors. Older age is generally associated with an increase in comorbidities, weak immune defense, and higher levels of proinflammatory molecules. In addition, ACE2 levels are decreased in the elderly and might be part of the mechanism causing higher risks of severe illness. Differences in sex hormones involved in inflammatory processes are among the leading causes of males' higher risk of developing severe symptomatology. Also, the expression levels of ACE2 and TMPRSS2 vary in males and females and might play a role. Other risk factors are hypertension, diabetes, and obesity. Modified from Gao et al. (2021)[35]. 11

Figure 7. Representation of the possible recombinant origin of the SARS-CoV-2 genome. RNA alignment with other *Sarbecoviruses* reveals 15 possible fragments (numbers on top) deriving from genome recombination. Red bars represent putative break points. Colours are the most similar SARSr-CoV genome(s). "MULT" is used when multiple sequences are equally similar to the fragment. "Unknown" indicates a region of unresolved phylogeny. Modified from Temman et al. (2022)[44]. 13

Figure 8. Map of the first 155 cases reported in Wuhan. The inset shows the map of all December 2019 cases in the city, grey dots are the ones non displayed in the main panel. In both the maps the red square indicates the location of Huanan market. Modified from: Worobey et al. (2022)[47]. 14

Figure 9. Diagram of SARS-CoV-2 virion showing main structural proteins and features. Adapted from Safiabadi et al. (2021)[28]. 15

Figure 10. Model of Spike protein structure and organization within the virion membrane. Top: the spike protein is split into two subunits (S1 and S2) by the cleavage in the S1/S2 site. The S1 domain contains a Signal Peptide (SP), the N-terminal Domain (NTD) and the Receptor Binding Domain (RBD). The Receptor Binding Motif (RBM) is part of the RBD and directly interacts with spike protein receptor. S2 Domain contains several regions crucial for spike activation, including a cleavage site (S2'), the Fusion Peptide (FP) and two Heptad Repeats (HR-N and HR-C). The transmembrane region anchors the protein to virion membrane while the intracellular tail (IC) is in virion interior. Bottom: Tertiary structure scheme (left) of spike protein homotrimer and corresponding structure determined by crystallography (right, PDB: 7BNN). Colours are the same as in Top panel. Adapted from Artica et al. (2020)[46]. 18

Figure 11. The structures of the prefusion and postfusion structures of SARS-CoV-2 spikes. Arrows indicates the position of open Receptor Binding Domains (RBDs). Adapted from Ismail and Elfiky (2020)[159]. 19

Figure 12. Three-dimensional structure of Envelope protein and organization within virion envelope. H1, H2 and H3 are three predicted alpha-helix regions of the protein. Adapted from Kuzmin et al. (2021)[62]. 20

Figure 13. Structural features of SARS-CoV-2 M protein. Left: each M protein is divided in two main regions: the membrane core, composed of 3 transmembrane helices (TM1-3) and the intravirion tail, composed of several beta-sheets structures. Disordered structures are indicated as dashed lines. Left: Cristal structure of M protein dimer and its organization within virion membrane. Modified from Zhang et al. (2022)[63]. 21

Figure 14. Schematic representation of N protein domains (top) and crystal structures of RNA binding domain (RBD, purple) and dimerization domain (green) (bottom). 22

Figure 15. Genetic organization of SARS-CoV-2. A comparison with other human-infecting coronaviruses' genome is shown. Accessory genes are shown in green. Modified from Safiabadi et al. (2021)[28] 24

Figure 16. Secondary structures of SARS-CoV-2's first 500 nucleotides and 3'-UTR. Hydrogen bound colours represents base pairing probabilities as connection scores. SL1-7: stem and loops; s2m: stem and loop II-like motif; P1-5: pseudoknot stems;

dashed lines: pseudoknot, BSL: bulged stem and loop. Green bar represents the first codon of ORF1ab. Modified from Cao et al. (2020)[69] 27

Figure 17. SARS-CoV-2 entry in host cells. SARS-CoV-2 exploits two possible mechanisms for host cell infection. In both the cases, the virus binds its receptor ACE2 (1). If target cell expresses low levels of TMPRSS2 (left) the virion is internalized (2). Subsequent endosomal acidification (3) leads to Cathepsin L activation which cleave the S2′ site on the spike protein. Such cleavage, in turn, activates the spike protein, which leads to the fusion of virion envelope and endosomal membrane (5). Viral RNA is thus released in the cytoplasm and uncoated (6). This process occurs at the level of cell membrane if Spike protein is cleaved by TMPRSS2 (right). Such cleavage is functionally equivalent to the one of Cathepsin L and induce virion envelope fusion with host cell membrane (3) and subsequent viral RNA release in the cytoplasm (4). Adapted from Jackson et al. (2021)[93]. 29

Figure 18. Schematic structure of pp1a and pp1ab displaying the main domains and activity of each non-structural protein after its release from the polyprotein. Main proatease (M^{pro}) and papain-like protease (PL^{pro}) cleavage sites are shown as red and black arrows, respectively. DMV, double-membrane vesicle; DPUP, Domain Preceding Ubl2 and PLpro; EndoU, endoribonuclease; ExoN, exoribonuclease; HEL, helicase; Mac I–III, macrodomains 1–3; NiRAN, nidovirus RdRP-associated nucleotidyltransferase; NMT, guanosine N7-methyltransferase; OMT, ribose 2′-O-methyltransferase; Pr, primase or 3′-terminal adenylyl-transferase; RdRP, RNA-dependent RNA polymerase; TM, transmembrane domains; Ubl, ubiquitin-like domain; Y, Y and CoV-Y domain; ZBD, zinc-binding domain. Adapted from V’kovski et al. (2021)[71] 30

Figure 19. Schematic mechanism of sub-genomic RNA (sgRNA) production in SARS-CoV-2. While the positive-sense genomic RNA can be fully replicated to a negative-sense RNA, discontinuous transcription can also occur. The process is regulated by transcription regulatory sequences located in gene bodies (TRS-B) and at the 3′-end of negative-sense RNA (TRS-L). Modified from V’kovski et al. (2021)[71].32

Figure 20. Schematic overview of SARS-CoV-2 infection cycle. Once entered in host cells upon receptor binding, viral RNA is released in the cytosol. Here the genome is immediately translated in a polyprotein that is then processed to produce replicative the non-structural proteins (nsps). These proteins are required for the generation of the replicative organelle, composed of several Double-Membrane Vesicles (DMVs) and deriving from the Endoplasmic Reticulum (ER). Here viral genome is replicated and sub-genomic RNAs produced. The corresponding sub genomic mRNAs (sg mRNAs) are finally translated in the structural proteins which assembly the virion at the level of the ERGIC. Virions are finally released through an exocytosis pathway. Adapted from V’kovski et al. (2021)[71]. 33

Figure 21. Spike D614G substitution increased SARS-CoV-2 infectivity. D614G substitution caused the arise of the lineage B.1, characterized by higher replicative efficiency and infectivity than the ancestral virus (left). Such features explain the rapid increase in the frequency of the mutation worldwide (right). The black arrow indicates the spike D614G mutation onset. Data from GISAID. 35

Figure 22. Worldwide relative frequencies of the main SARS-CoV-2 lineages and Variant of Concern identified so far (last update: end of August 2022). The data for this graph have been obtained from Nextstrain[95]. 38

Figure 23. Schematic overview of DNA nanoballs generation by rolling circle replication. A single strand DNA (black circle) is annealed to a primer, indicated by the black arrow (1). A DNA polymerase catalyses the extension of such a primer (2), and the displacement of the previously synthesized DNA strand while performing a new round of synthesis (3). This process is repeated several times until a DNA concatemer is obtained (4). 43

Figure 24. Time distribution of infection (continuous lines) in Italy (blue) and United Kingdom (UK, red). The number of sequencies produced and deposited on GISAID are shown as a dashed blue (sequences originated in Italy) and red line (sequences originated in UK). 47

Figure 25. Schematic representation of the workflow set up to collect, process and analyse a considerable number of viral genomes. Top: Oro-nasopharyngeal swabs are performed to diagnose the presence of SARS-CoV-2 genome in patients and extract its RNA. Subsequently, viral RNA is retrotranscribed and subjected to two PCR steps to amplify and index it. After circularization and nanoball generation, the library is then sequenced and analysed. Bottom: As an alternative and faster approach, an optimized strategy enables the amplification and indexing to occur in one PCR step. 56

Figure 26. Boxplot showing the % of viral genome reaching at least 10X coverage in each sample (left). Such high coverage is reflected in the number of samples passing our quality filters and submitted on the GISAID platform. All samples are divided by each tested solution. For comparison purposes, only samples with average Ct value < 33 were considered. 58

Figure 27. Violin plot showing the distribution of the percentage of SARS-CoV-2 reads detected for different ranges of CTs. n:sample size. 59

Figure 28. Efficiency of SARS-CoV-2 genome retrieval depending on samples CT value. Samples were divided in 5 (left) or two (right) Ct classes. n: sample size. 60

Figure 29. Average coverage across SARS-CoV-2 genome obtained in different Ct classes (plot titles). The 30X coverage threshold chosen for confident base calling is shown as a red dashed line. 61

Figure 30. Variant annotation, cumulative frequency, and sequencing coverage of each position of SARS-CoV-2 genome. n: sample size. 62

Figure 31. Mutations and lineages identified during 2021 genome surveillance. Left: Venn diagram showing the intersection between mutations detected in all the sequenced genomes worldwide (yellow) and the mutations found in this study (light blue). Right: representation of all the 156 lineages identified in this study. The length of the bars is indicative of the number of samples for each lineage in the logarithmic scale. Coloured bars indicate VOC 63

Figure 32. Demographic (left) and geographic (right) distribution in Campania of the patients swabbed for SARS-CoV-2 genome analysis (left). The population density (bottom right) of the region is also shown as reference. 64

Figure 33. Geographic map representing Italian regions, coloured by the number of genomes deposited on the GISAID platform. Bottom: percentage of genomes deposited on GISAID over the total Italian sequences, divided in Northern (green) and Southern (blue) regions. 20% of the overall Italian sequences has been produced by this study (dark blue). 65

Figure 34. Geographic map representing European States, colored by the number of 2021 months with at least 5% of viral genomes compared to new cases. 5% is the recommended .sequencing **capacity limit by ECDC. Only for Italy, individual regions are displayed** 66

Figure 35. Bar charts showing the distribution in time of the samples assigned to variants of concern and of interest identified in the during our surveillance program. 67

Figure 36. Density plots showing the distribution, in time, of the most frequent variants described in this thesis (middle) or in Italy (bottom) relative to the Campania infection curve (top) and infection waves (red-colored areas). The red arrow highlights different variants dynamics between regional and national level, in a certain period.69

Figure 37. Time course of Ct values during pandemic in Campania. Left: Distribution of the average CT value across different Variants of Concern (VOC). Only not significant (n.s.) pairwise comparisons are reported (Bonferroni adjusted p-value > 0,05). Right: Distribution of Ct values, in time, for all the samples collected during the pandemic in Campania. The trend line (red) and 95% confidence interval (light gray) are shown. 70

Figure 38. Donut charts representing the number of analysed genomes presenting some mutation of concern, namely Spike L18F, S477N, P681H (top) and E484K (bottom) divided by lineage. The definition of Expected lineage is described in the Methods chapter. 71

Figure 39. Section of the maximum likelihood phylogenetic tree representation of a random subset of our data (n=12998), coloured by lineages. The identified lineage is reported (blue dots, right) and zoomed in (left). n:sample size. 72

Figure 40. Right: phylogeny of the proposed B.1.177^{E484K} lineage. Samples belonging to the proposed lineage are in green. Bootstrap values for each node are shown as node points. Left: Geographic distribution of genomic variants belonging to the identified lineage, coloured by the collection date. The size of each pie chart is proportional to the number of samples in each geographic position. n:sample size 73

Figure 41. Results from the clustering analysis for samples collected until May 2022, displayed as line plots of frequency over time (trends). The arrow indicates the cluster investigated in next figure. 74

Figure 42. Line plot showing the frequency trend of the selected mutations in Campania. 75

Figure 43. Right: Section of the maximum likelihood phylogenetic tree representation of a random subset of our data (n=12998), coloured by lineages. The identified lineage Alpha^{YTTH} is reported (blue) and zoomed in (top). n:sample size. Phylogeny of the proposed Alpha^{YTTH} lineage. Samples belonging to the proposed lineage are in green. Bootstrap values for each node are shown as node points. 76

Figure 44. Geographic distribution of genomic variants belonging to the identified lineage, colored by the collection date. The size of each pie chart is proportional to the number of samples in each geographic position. n:sample size. 76

Figure 45. Line plot showing the frequency trend of the selected omicron subvariant mutations in Campania (left) and corresponding phylogenetic reconstruction (right). Samples belonging to the proposed lineage are in green. Bootstrap values for each node are shown as node points. 77

Figure 46. Genomic characterization of twenty patients with long COVID-19 infection. The number of detected mutations is reported as a function of the number of days from the first swab. The assigned lineage (colours) and consistency (transparency) are also displayed. 80

Figure 47. Patient 8 genomic characterization relative to the number of detected mutations (colours), the infection load (y axis) and symptoms severity (+++: severe; ++: moderate). 81

Figure 48. Schematic representation of RNA-seq data structure, pre- and post-filtering. The number of samples per condition pre- and post-filtering are shown. 82

Figure 49. Principal Component Analysis plots of B.1 and Delta datasets coloured by SARS-COV-2 infection positivity. 82

Figure 50. Correlation analysis between CTs and gene expression of B.1 patients, performed on 8100 genes, is shown as a barplot. For each gene (x axis), its correlation value (y axis) and significance (p-value < 0,0001, red) is reported. Bottom: highlight of the significant results. (161 genes). The top 10 most anti-correlated genes are reported (black box). 83

Figure 51. Pathway and gene set enrichment analysis performed for different databases using the gene signature previously identified. Each barplot shows the significance (x axis) and the percentage of overlap (fill colour) between the input signature and the tested public gene sets. 84

Figure 52. Heatmap of z-scored, log2-transformed and normalized gene counts for the 161 significantly correlated genes from Fig. 50. Values have been averaged in 4 groups of samples depending on the CT (x axis) or whether they were negative. 85

Figure 53. Gene signature analysis in patients infected by Delta variant. Left: correlation analysis between CTs and gene expression of Delta patients, performed on 5525 genes, is shown as a barplot. For each gene (x axis), its correlation value (y axis) and significance (p-value < 0.0001, red) are reported. The 16 significant genes are highlighted in the inset. Centre: pathway and gene set enrichment analysis performed for different databases using the gene signature previously identified. Each barplot shows the significance (x axis) and the percentage of overlap (fill colour) between the input signature and the tested public genesets. Left) Heatmap of z-scored, log2-transformed and normalized gene counts for the 16 significantly correlated genes from the analysis of Delta dataset. Values have been averaged in 3 groups of samples depending on the CT (x axis) or whether they were negative. 86