Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

The seasonal behaviour of COVID-19 and its galectin-like culprit of the viral spike

Kelsey Caetano-Anollés[a], Nicolas Hernandez[b], Fizza Mughal[b], Tre Tomaszewski[b], and Gustavo Caetano-Anollés[b,*]

[a]*Callout Biotech, Albuquerque, NM, United States*

[b]*Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois, Urbana, IL, United States*

**Corresponding author: e-mail address: gca@illinois.edu*

1 Introduction

Many diseases have a seasonal cycle occurring once or multiple times a year, or once every several years. This regular temporal behaviour is observed because the emergence of infectious diseases depends on a multiplicity of factors, including the seasons, weather and geography, the number of susceptible individuals in a population, the severity of the pathogen, host physiology and phenology, and the genomic makeup of the virus, which can convert a non-infectious pathogen to one that is infectious for a specific or general population (Dowell, 2001; Martinez, 2018). The transmission of viral infections associated with seasonal variations has been referred to as 'seasonal forcing' (Rohani, Earn, & Grenfell, 1999) and remains an active field of exploration. Here we review evidence supporting the seasonal behaviour of infectious diseases. We focus on viral diseases and the ravaging coronavirus disease 2019 (COVID-19).

COVID-19, and the fatal spread of its novel culprit, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pathogen, has emerged as the largest pandemic of the 21st century thus far (Wang, Horby, Hayden, & Gao, 2020). The first case was reported in the Chinese city of Wuhan in December 2019 (Wu, Zhao, et al., 2020). Since then the virus has spread over the world with more than 217 million reported cases and 4.5 million deaths worldwide despite administration of over 5 billion vaccine doses as of September 1, 2021. Coronaviruses are a highly diverse and globally distributed group of enveloped viruses with single-stranded RNA genomes. They can infect humans and other mammals and avian species causing

Methods in Microbiology, ISSN 0580-9517, https://doi.org/10.1016/bs.mim.2021.10.002
Copyright © 2021 Elsevier Ltd. All rights reserved.

respiratory, gastrointestinal, hepatic, and neurological diseases. They are part of the *Coronaviridae* family, subfamily *Orthocoronaviridae*, which consists of four coronavirus genera (α, β, γ, and δ) (Coronaviridae Study Group, 2020). To date, seven human coronaviruses (HCoVs) have been identified as highly circulating viruses: the α-coronaviruses HCoVs-NL63 and HCoVs-229E and the β-coronaviruses HCoVs-OC43, HCoVs-HKU1, which cause seasonal and usually mild respiratory tract infections, and the β-coronaviruses MERS-CoV, SARS-CoV-1 and SARS-CoV-2, which cause severe and life-threatening respiratory disease (Ashour, Elkhatib, Rahman, & Elshabrawy, 2020). SARS-CoV-2 infects humans and animals across different regions of the world. Many studies have been devoted to the causes of the COVID-19 outbreak, its geographic distribution, factors that modify virus infectivity, the effects of seasonal changes on viral transmissibility, and the genomic makeup of the virus. Like many other respiratory diseases, COVID-19 shows an anticipated connection between climate and disease dynamics (Smit et al., 2020). It is therefore important to study the seasonal behaviour of the SARS-CoV-2 virus and host responses that are associated with infection since this knowledge can help mitigation efforts.

2 The seasonal behaviour of viruses

2.1 Seasonality of viral diseases

Hirsch (1883) was one of the first to lay out foundations for a geographical and historical pathology, which he presented in handbook format almost a century and a half ago while describing temporal and spatial patterns of spread of major viral and bacterial diseases. Inspired by the work of Finke (1792-1795) and the traditions of Hippocrates, Galen and Avicenna, he analysed patterns in onset and spread of acute infectious diseases, including influenza, dengue, hantavirus sweating sickness, smallpox, measles, and scarlet and yellow fever. Recurrent patterns were carefully recorded and often elaborated. For example: '*The recurrence of the epidemic of measles at one particular place …depends solely on two factors, the time of importation of the morbid poison, and the number of persons susceptible of it*' (Hirsch, 1883). It took however an additional half a century to analyse periodicities with statistical and modelling methods (Brownlee, 1918; Soper, 1929), and even longer to start dissecting possible culprits. For example, the idea that recurrent patterns do not necessarily arise from chains of host-to-host pathogen transmission but rather from changes in susceptibility to pathogens present year-round in host populations was only recently elaborated (Dowell, 2001). Despite advances, we are still far from fully understanding why and how diseases wax and wane with the seasons.

There appears to be an epidemic calendar for many viral infectious diseases (see Glossary). This calendar results in infection peaks occurring at different times during one or multiple years in patterns that are remarkably consistent. In the Northern Hemisphere, for example, influenza often occurs during winter (the 'flu season'), chickenpox during spring, and poliomyelitis during the summer. Fig. 1 shows early and more recent examples of periodic incidence recorded for measles and influenza,

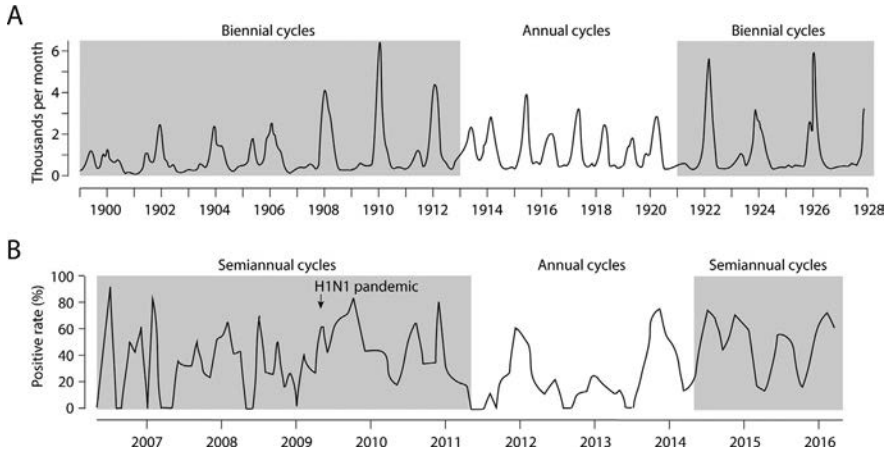

FIG. 1

The seasonal behaviour of measles and influenza. (A) The plot shows number of cases of measles in Glasgow from 1900 to 1927. A noisy succession of biennial and annual cycles can be observed spanning three periods. (B) Composite influenza virus activity in Chengdu, a populous subtropical city in Southwestern China, from 2006 to 2015. Monthly positive rates (in relative scale) were dissected into influenza A subtypes and influenza B lineages (not shown). The start of H1N1 types of the 2019 pandemic and noisy semiannual and annual cycles are indicated.

Panel (A) data from Soper, H.E. (1929). The interpretation of periodicity in disease prevalence. Journal of the Royal Statistical Society, 92*(1), 34–73. Panel (B) data from Zhou, L., Yang, H., Kuang, Y., Li, T., Xu, J., Li, S., et al. (2019). Temporal patterns of influenza A subtypes and B lineages across age in a subtropical city, during pre-pandemic, pandemic, and postpandemic seasons.* BMC Infectious Diseases, 19, *89.*

two classical acute respiratory diseases caused by RNA viruses with significantly different initial R reproduction metrics (12–18 and 1–2, respectively). While seasonality appears not significantly constrained by infection spread potential (number of infections caused by an infected person), the recurrence and stochastic behaviours showcased in Fig. 1 are suggestive of the many research challenges that face the study of seasonal behaviour.

A number of seasonal cycles of viral and bacterial diseases have been recently documented in concert with their likely drivers (Martinez, 2018). Their ubiquity suggests seasonality may be a unifying feature of epidemics in general. In some cases there are clear cycles that impose temporal restrictions to infection despite vaccinations or other mitigation strategies. For example, measles has been long associated to school attendance, which fosters aggregation of school children and virus transmission (Fine & Clarkson, 1982; Soper, 1929). However, in some countries, measles has been also associated to agricultural cycles (Duncan, Duncan, & Scott, 1997) and the likely culprits of measles seasonality appear much more complex than anticipated (Conlan & Grenfell, 2007). In fact, human physiology may be also an important player.

Seasonal variations play a vital role in determining the time at which an infection might occur, its transmission, and its potential to become an epidemic (Rohani et al., 1999). Important elements affecting seasonal cyclicity include biotic and abiotic mechanisms. Seasonal cycles are sometimes associated with pathogen life cycles that involve viral spread via insect vectors or maintenance in animal or environmental reservoirs. To illustrate, MERS-CoV is recurrently introduced into the Middle Eastern population during the camel calving season (Dudas, Carvalho, Rambaut, & Bedford, 2018). Seasonal peaks of the Marburg virus disease coincide with the birthing season of bats, which occurs twice a year (Amman et al., 2012). Similarly, cases of rabies coincide with seasonal infection cycles in bats (George et al., 2011). In zoonotic diseases such as these, seasonality may result from contact with wildlife or livestock. In the case of Ebola, for example, the pathogen presence in wildlife peaks in the dry season showing also an environmental effect (Groseth, Feldmann, & Strong, 2007). Complex environmental and life cycle interactions also occur in African sleeping sickness. The distribution of the tsetse fly vector, which expands during the rainy season, affects the seasonal behaviour of the disease (Knight, 1971). Seasonal changes in the incidence of vector-borne diseases such as yellow fever, Zika and Lyme disease are also expected to be driven by seasonal fluctuations in the population of the tick and mosquito (*Aedes aegyptii*) vectors (e.g. Ferguson et al., 2016; Soper, 1967). This abiotic modulation of vector population dynamics adds complexity to biotic mechanisms of seasonality. Abiotic mechanisms are also central for some seasonal cycles. Factors include geographic location and environmental conditions. For example, latitude influences the time of emergence and the magnitude of disease outbreaks, such as in poliomyelitis (Paccaud, 1979) and influenza (Cox & Fukuda, 1998). At worldwide level, early suggestions that influenza outbreaks move across the Earth every year along a sinuous curve parallel with the 'midsummer' curve of vertical solar radiation (Hope-Simpson, 1981) have been recently supported by an exhaustive modelling study (Deyle, Maher, Hernandez, Basu, & Sugihara, 2016). Primarily, the seasons arise from Earth's tilted axis relative to the plane of its orbit, which changes the amount of sunlight reaching its surface at various latitudes over the course of a year. Worldwide analysis and modelling of the effects of meteorological factors of humidity and temperature suggest periodicity of influenza follows seasonal periodicities of the planet but also show a U-shaped nonlinear effect of absolute humidity on influenza, which is mediated by temperature (Deyle et al., 2016). This global effect produces a trade-off of wetter air and low temperature that likely promotes viral spread by protecting the proteins and lipids of the viral envelope. This trade-off appears optimal at ~24 °C. Results align with previous experiments that used guinea pig models of transmission to demonstrate the coordinated effects of temperature and humidity on the incidence of influenza disease (Lowen, Mubareka, Steel, & Palese, 2007). A number of environmental factors imposed by climate conditions are also expected to play critical roles in epidemics as they influence pathogen survival during transition periods spanning hosts. These factors directly impact the survival of the virus pathogen. These effects are particularly significant for respiratory viral diseases

(Moriyama, Hugentobler, & Iwasaki, 2020). For example, short exposure of viruses to temperature or UV light exposure in droplets or other forms of fomite airborne transmission can have significant effects on viral load and emergence of epidemics (e.g. for influenza; Weber & Stilianakis, 2008). Alternatively, the environment can also affect host susceptibility to infection via physiology or behaviour. Seasonal variations are expected to be impacted by phenological effects of life history, including cycles of migration or hibernation, and endogenous rhythm responsible for seasonal changes in immunity, reproduction and metabolism, all of which are impacted by the environment (Dowell, 2001; Martinez, 2018). Environmental factors also affect antiviral defenses including intrinsic barriers such as mucus production and epithelial integrity, inducible innate immune defense mechanisms, and adaptive immunity (Moriyama et al., 2020).

2.2 Drivers of seasonality in trade-off performance spaces

When exploring epidemic calendars, 'seasonal drivers' (see Glossary) have been categorized into those that involve the environment, host behaviour, host phenology, and exogenous biotic factors (Martinez, 2018). A more fine-tuned categorization of drivers was used in a survey of infectious diseases that included: (a) vector seasonality; (b) seasonality in livestock, domestic animals and wildlife; (c) seasonal climate (e.g. temperature, humidity); (d) seasonal non-climatic abiotic environment (e.g. water salinity); (e) seasonal co-infection; (f) seasonal exposure-behaviour-contact; (g) seasonal biotic environment (e.g., algal density in waterbodies); and (h) seasonal flareups-symptoms-remission-latency (see tables 1–4 in Martinez, 2018). Other categorizations stressed within-host and anthropogenic factors (Kronfeld-Schor et al., 2021).

Here we propose that many of these multiple driving forces are in 'trade-off' relationships and can be better described within a framework of a 'triangle of viral persistence' modulated by environment, physiology and behaviour (Fig. 2), with environment impinging on both the life cycle of the virus (including 'virion' and 'virocell'; defined in Glossary) and efficient viral transmission (e.g. enteric, respiratory), and physiology and behaviour embodying functional beneficial characteristics of the hosts, vectors, and reservoirs involved. A 'trade-off' exists when one trait cannot increase without a decrease in another (Garland, 2014). These trade-offs are usually caused by limitations in matter-energy and information unfolding in time and space. The triangle is based on a 'persistence' framework describing the impact of the environment on biological systems (Yafremava et al., 2013). Its main premise is that the environment constrains evolution of physiologies over the system's initial and boundary conditions. We here extend this framework, which is based on trade-offs between *'economy'*, *'flexibility'* and *'robustness'* in performance spaces (defined in Glossary), to the calendar of viral life cycles. Similar triangles have been proposed for the persistence of networks that model biological systems, the persistence of molecular communication and molecular systems, and the persistence of viruses (Table 1). In particular, the triangle of viral persistence

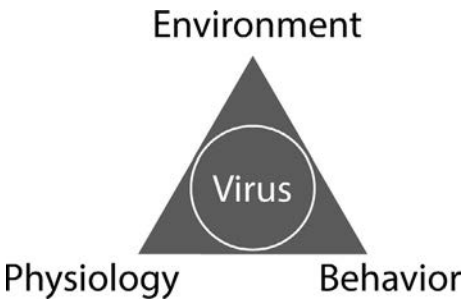

FIG. 2

A triangle diagram of viral persistence explains main drivers of seasonality as a landscape of trade-offs between factors related to environment, physiology and behaviour. The sphere in the middle represents viral quasispecies clouds in Pareto fronts within a three-dimensional fitness landscape. Viral interaction with environment, physiology and behaviour result in approaches that achieve solution goals to increase persistence of viral life cycles.

Table 1 Triangles describing persistent systems.

Persistent system	Trade-off strategies[a]			Reference
General	*Economy*	*Flexibility*	*Robustness*	Yafremava et al., 2013
Network	Heterogeneity	Modularity	Randomness	Caetano-Anollés et al., 2021
Communication	Persuasion (Quantity)	Attraction (Quality)	Relevance (Relation)	Caetano-Anollés, 2021
Molecular	Catalyst	Machine	Gatekeeper	Caetano-Anollés & Caetano-Anollés, 2015
Virus life cycle	Dependency	Propagation	Dormancy	Nasir, Kim, & Caetano-Anollés, 2017
Epidemic calendar	Behaviour	Physiology	Environment	Present proposal

[a]*Definitions of Economy, Flexibility and Robustness strategies of the general framework can be found in the Glossary.*

(Nasir et al., 2017) explains how the three viral strategies of dependency (via symbiotic-like associations of dissimilar entities tending towards mutualism, commensalism, amensalism and parasitism), propagation (through lysis, budding and transport), and dormancy (covert existence through latency via episomes or endogenized genetic material) take advantage of the engineering strategies of economy, flexibility and robustness, respectively. Symbiotic-like co-dependencies impose altruistic, cooperative or antagonistic relationships that involve the budgeting of resources (economy). Efficient propagation through lytic interactions requires establishing life

cycles and evolutionary arm races that foster co-evolutionary flexibility. Dormancy inside cells shields against environmental variations and is a form of robustness.

The sphere in Fig. 2 represents a cloud of points in the trade-off space (morphospace) of a 'viral quasispecies', a dynamical virus collective defining a multiplicity of virions and virocells (see Glossary). These points locate in Pareto fronts, which are boundaries in multidimensional performance spaces that provide optimal fitness solutions. Performance spaces are here defined as worlds of traits, which in our case associate with strategies of economy, flexibility and robustness. Traits are the results of processes such as number of cases in an epidemic, a geographical distribution of vectors, or a direct measurement of seasonal behaviour (e.g. amplitudes of physiological rhythms). These processes are sometimes drivers (e.g. schooling in measles epidemics). The geometries of fitness solutions, which have been mathematically elaborated (Sheftel, Shoval, Mayo, & Alon, 2013), correspond to planes in three-dimensional performance spaces as described in Table 1. However, geometries can be simpler, such as line segments arising from trade-offs between only two drivers of strategy, or more complex, such as polygons describing trade-offs between a multiplicity of drivers. Mathematical solutions to the analysis of performance spaces is well developed and can provide new tools for the exploration of seasonality.

To explain drivers of seasonality, the triangle of viral persistence links *environment* to dormancy and robustness, *physiology* to propagation and flexibility, and *behaviour* to dependency and economy (Table 1). These epidemic calendar associations become clear when seasonal drivers are assigned to strategies:

(i) *Environment:* Robustness embodies mechanisms that use information to maintain structure and function in the face of environment-induced damage and change. A clear example is the survival of respiratory viruses exposed to temperature, absolute humidity or UV radiation. During a pandemic, the evolving viral quasispecies will develop protein 'variants' with sets of co-occurring amino acid changes that will strive to survive the environmental challenges imposed by weather or geographic location. One proposed mechanism is protection of integrity of the viral envelope (Minhaz Ud-Dean, 2010), which could be accomplished by favouring mutations in protein regions of intrinsic disorder (Tomaszewski et al., 2020). Following a selective sweep, mutations take over the quasispecies population in search for new cycles of better variant proteins. While these processes of diversification occur during replication, the actual driver results from a 'passive' interaction between the environmental factors (e.g. temperature and absolute humidity) and the viral envelope (e.g. influenza proteins). This passive interaction is dormant, in that the virus is in a state where its ability to change is inactive or inoperative (akin to a seed entering a dormant state to accomplish efficient plant propagation). This example serves the purpose of illustration. However, the environment affects other forms of viral transmission, such as the spread of enteroviruses through sewage water or the spread of a number of RNA viruses through vectors. Here, environmental effects are indirect. Diseases transmitted by

mosquitoes and tick require that vectors avoid being dormant in 'overwintering' states (e.g. for Culex mosquitoes and West Nile encephalitis disease; Nasci et al., 2001). In tropical and subtropical regions, vector number and generation time depend on warm weather and rainy seasons, but also on the more stochastic effects of snow melts and floods (see examples in Fisman, 2012). Viral zoonoses are also dependent on environmental factors such as rainfall (e.g. hantaviruses and rodents in the American Southwest; Jonsson, Moraes Figuereido, & Vapalahti, 2010). In these cases, the passive element is the vector or reservoir but the effects are ultimately on the virocells and the opportunities to fulfill life cycles through different hosts.

(ii) *Physiology:* Flexibility reflects structural and functional mechanisms requiring processing of information needed to respond and adapt to internal and external challenges. Indeed, it is within the physiology of the host where the viral life cycle fully achieves the goal of viral propagation. However, physiological drivers of seasonal behaviour can be multiple and involve not only the host but vectors and reservoirs. Early focus on these drivers (Dowell, 2001) and later elaborations (Kronfeld-Schor et al., 2021; Martinez, 2018) highlight the effects of host susceptibility to pathogen infection. Seasonal changes of immunity and inflammation, among a multitude of 'omic' analytes, are now being studied at multiomic profiling levels and with unprecedented depth (Sailani et al., 2020; Wyse, O'Malley, Coogan, McConkey, & Smith, 2021). New data confirm previous results, which suggested the possible role of cellular time-keeping processes, including circadian clocks, circannual rhythms, metabolic cycles, photoperiod effects, and other chronobiological processes. For example, there is evidence of widespread seasonal transcriptome regulation in white human blood cells and adipose tissue with endogenous patterns of immune function that reverse in Northern and Southern Hemispheres (Dopico et al., 2015). Similarly, the physiologies of a central circadian clock that supports daily oscillations of cellular processes and behaviours synchronizes with clocks in peripheral tissues that regulate immunity and other important cellular responses (Borrmann, McKeating, & Zhuang, 2021). Circadian molecular pathways shape viral infection, and vice versa, infections perturb circadian rhythms. This tight integration of cell physiology and viral life cycles impacts the environment and behaviour strategies. For example, integration will fine tune the viral replication, assembly, maturation and release phases of the cellular infection process that ultimately produces the infective virion particles. This outcome then impinges on viral protein interaction with the environment and the behaviour of host, vectors and reservoirs. It is here where fine tuning at physiology level gets tested at physiological, ecological and evolutionary time scales.

(iii) *Behaviour:* Economy reflects the budget of matter–energy costs of a system. In human endeavours, economy plays a central role in decision-making. For example, the opening-closing of schools is intimately linked to economic

drivers, as has been made evident in the current socioeconomical and political landscape of COVID-19. The fact that school cycles are linked to the seasonal incidence of measles (Fine & Clarkson, 1982; Soper, 1929) comes as no surprise. The same goes for other respiratory illnesses (e.g. varicella, influenza). Many cultural, socioeconomic, and lifestyle factors affect seasonal patterns of infection. Some of these factors change the frequency of interactions between/within hosts, vectors and reservoirs and finally the host-virus interface. Because human activity has been a dominant influence on climate and the environment during the past centuries, anthropogenic influences are expected to impact epidemic calendars. These include changing climate, land and freshwater use, food availability and consumption, demographics, travel, settlement and urbanization, technology developments, war and famine, antimicrobial drug use, and even breakdown of public health measures (e.g. Jones et al., 2008; Kronfeld-Schor et al., 2021). They also include effects on geography, such as river fragmentations and dewatering (Farah-Pérez, Umaña-Villalobos, Picado-Barboza, & Anderson, 2020). All these effects can be seen as dependency relationships that change behaviour at the level of the organism and in doing so can impact epidemic cycles.

While some seasonal drivers may be significant contributors to the epidemic calendar, there may be multiple trade-off relationships that may be relevant. These can be dissected with Pareto fronts in multidimensional performance spaces. For example, cold weather leads to indoor crowding and higher levels of person-to-person contact (Moriyama et al., 2020). This pushes the persistence of the seasonal patterns of respiratory virus transmission towards the economy-driven 'behaviour' vertex of the triangle. However, indoor heating systems favour low relative humidity (Moriyama et al., 2020), which pushes the seasonal response towards the robustness-driven 'environment' vertex. In addition, vitamin D serum levels, which are influenced by sunlight exposure and decrease during winter (Kasahara, Singh, & Noymer, 2013; Klingberg, Olerod, Konar, Petzold, & Hammarsten, 2015), help the immune system fight the viral infections (Cannell et al., 2006). This pushes the seasonal response towards the flexibility-driven 'physiology' vertex. In all cases, a number of physiological mechanisms are expected to be relevant, some defining human 'crowding' responses, others defining the effects of absolute humidity in viral makeup, and yet others impacting immunological rhythms.

We end by noting that despite the increasing acceptance of seasonal patterns of disease, only a few diseases have been systematically characterized. Mechanisms of seasonal forcing have only been investigated for diseases of significant public interest, including measles and influenza (e.g. Mantilla-Beniers, Bjørnstad, Grenfell, & Rohani, 2010; Shaman, Pitzer, Viboud, Grenfell, & Lipsitch, 2010). There is still much to learn about the causal factors of seasonal periodicities. This necessitates both the application and development of statistical and data mining methodologies that can dissect causative from aleatory effects.

2.3 Finding significant correlations

The effects of history and geography on viral diseases requires identification of culprits. Generally, this starts by first identifying variables and then correlations between variables that are statistically significant. Prior to the introduction of the mathematical concepts of 'regression' and 'correlation' by Galton and Pearson at the end of the 1800s, the only way to establish a relationship between variables was to infer a causative connection (Rodgers & Nicewander, 1988). This occurred without measurements of associations between variables that would test cause-effect relationships. Today, Pearson's product-moment correlation coefficient r and the regression equation are the most widely used statistical measures of relationship for studying observations. Alternative correlation indices such as Spearman's rho, the point-biserial correlation (r_{pb}), and the phi coefficient are simply computations of r for special types of data. Indices measure the strength and direction of an association between two variables (bivariate association) with no assumption of causality. They all assume dimensionless values ranging from –1 to +1, with values of 1 describing a perfect positive linear correlation, values of –1 describing a perfect negative correlation, and values of 0 describing the absence of a correlation. For Pearson's r, variables should be measured on a continuous scale (e.g. interval or ratio), variables should be both paired and independent from other observations, the association should be linear, variables should be approximately normally distributed, variances and co-variances should be finite, variances along the line of best fit should remain homogeneous (fulfilling homoscedasticity), and there should be no outliers in the data. Some of these requirements can be lifted, such as normality either in the marginal distribution or in the bivariate surface when the number of data points N exceeds 20. Generally, r values above [0.1], [0.3], and [0.5] indicate small (weak), medium (moderate), and large (strong) associations, respectively. However, cut-offs are arbitrary and any association above [0.1] with significant P-values should be interpreted within the context of the scientific questions posed.

Of the 13 different ways to conceptualize correlation (Rodgers & Nicewander, 1988), we highlight correlation as *'bivariate ellipses of isoconcentration'* because they have become new tools of statistical inquiry (Friendly, Monette, & Fox, 2013). When Galton made one of the first uses of a scatter plot in his famous description of the relationship between the height of children and the average height of their parents (Galton, 1886) he realized that an ellipse was associated with regression. Since then, 'data', 'concentration' or 'density' ellipses that capture for example 95% of observations in a scatterplot have been used to effectively describe bivariate marginal relationships in multivariate data. The geometrical approach detects unusual patterns such as curvilinear relationships or extreme outliers, which disrupt correlation measurements. Modern analysis methods now use Kernel density estimations of smoothed bivariate surfaces to explore skewness and multimodality in the data. The goal is to make inferences about the underlying probability density function everywhere in the scatter plot by smoothing out contributions from each data point to spaces surrounding them. Aggregation of the smoothed contributions

describe the structure of the data and its density function. For example, the scatter plots of Fig. 3 show the relationship between the number of COVID-19 cases and temperature, latitude or altitude of countries with reported cases. It is part of a larger ongoing study (Hernandez and Caetano-Anollés, ms. in preparation) jumpstarted by an initial exploration during the first wave of the pandemic (Burra et al., 2021). The strong association of COVID-19 cases with temperature and latitude and its absence with elevation suggests the geographical coordinate is associated with the onset of the disease. Remarkably, the bimodal density patterns reveal countries located at latitudes spanning 30–60°N or S and with average temperatures of 10–20 °C have been most impacted. Bimodality was also present in Kernel density plots describing the strong negative correlations between latitude and temperature ($r = -0.749$; rho $= -0.709$), suggesting bimodality is at least partially driven by planetary environmental factors. As we will describe below, these observations are relevant.

We note that correlations must first be confirmed as real, then every possible causative relationship must be systematically explored. Five criteria must be fulfilled to validate a causal relationship in scientific inquiry (Chambliss & Schutt, 2012): (i) *Empirical association:* Typically this involves finding a correlation between dependent and independent variables to determine if they vary together (covariation); without an empirical association of the variables there cannot be a causal effect; (ii) *Temporal priority:* This criterion requires that there should be variation in the independent variable before variation in the dependent variable, i.e. cause must come before its effect; (iii) *Non-spuriousness:* Here the goal is to eliminate falsehood in explanation of empirical associations by identifying other variables that could explain empirical associations and subjecting them to a process of falsification; (iv) *Mechanism:* Understanding a mechanism that connects independent and dependent variables adds further support to the cause-and effect phenomenon; and finally (v) *Context:* Establishing the external environment in which the cause-and-effect occurs, for example, with networks of cause-effect relationships, the causal-effect relationships are strengthened.

A number of tools help explore causal relationships, such as using hypothesis testing methods that compare null and auxiliary hypotheses to a primary hypothesis, or using simple controlled experiments (e.g. split-run testing in randomized experiments) that compare *ad minimum* two versions of a single variable. Here, Fisher's statistical principles of experimental design—comparison, replication, randomization, blocking, orthogonality, and factorials—are helpful. They seeded advances in algebra and combinatorics and development of impactful methods such as the Taguchi methods for robust engineering design that for example transformed the automotive industry (Montgomery, 2013).

2.4 Distinguishing error from chaos in time series

The epidemiological periodicities that are evident in the examples of Fig. 1 suggest the calendar of epidemics can be quite erratic. Measles for example peaks somewhere between November and December for the Northern Hemisphere and the flu

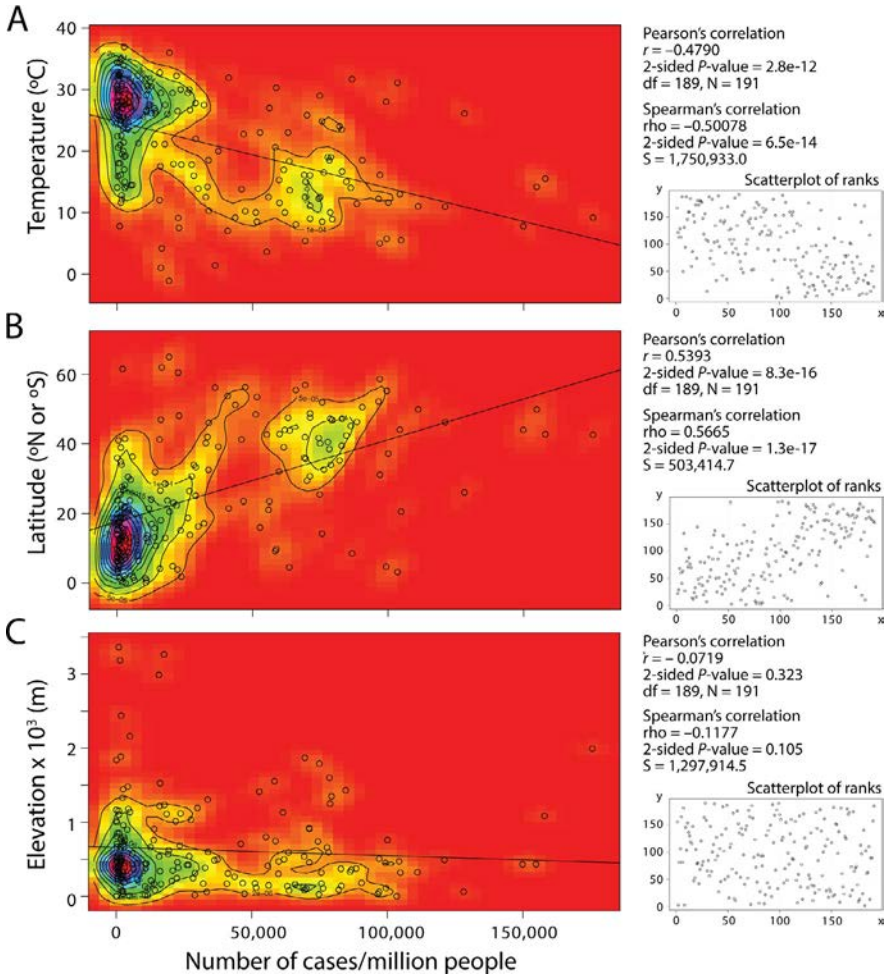

FIG. 3

Bivariate kernel density plots describing the relationship between the number of confirmed COVID-19 cases (per million people) and temperature (A), the geographical coordinate of latitude (B), or the average elevation (in metres) (C) of 191 countries as of May 23, 2021. Kernel bandwidths were 5497.2, 2.1, 4.5 and 105.6 for cases, temperature, latitude and elevation, respectively. Pearson product-moment correlation *r* values and Spearman's rank correlation rho values are given with their corresponding *P*-values, together with regression lines. Correlations show strong associations in (A) and (B), and negligible association in (C). Scatterplots of ranks from the Spearman's rank correlation show how rank transformations diminish homoscedasticity effects. Calculations used the University of Basel online statistical server (Wessa, 2021). Note that countries located at latitudes spanning 30–60°N or S and with average temperatures of 10–20 °C have been relatively most impacted. They include Montenegro, Czechia, Slovenia, Luxembourg, Serbia, US, Israel, Netherlands, Belgium, France, Croatia, Georgia, Hungary, Portugal, Cyprus, Lebanon, Lichtenstein, Argentina, Spain, Uruguay, Armenia, North Macedonia, Slovakia, Jordan, Chile, Italy, Malta, United Kingdom, Monaco, Moldova, Bosnia and Herzegovina, Turkey, Bulgaria, Palestine, and Ukraine (in decreasing order of cases per million).

ARTICLE IN PRESS

season peaks somewhere between November and March. Besides stochastic behaviour, there appears to be higher-level recurrent events imposed on the typical cycles in measles and the flu that must be explained, such as semiannual or biennial behaviour. The existence of pandemic or major epidemic events complicate the recurrent patterns, as showcased by the incidence of influenza A and B in a major metropolis of Northwestern China following the 2009 H1N1 pandemic (Zhou et al., 2019; Fig. 1B). Similar complications have been also highlighted for the city of Mexico (Bjørnstad & Viboud, 2016).

Statistical predictive and modelling methodologies must be developed when principles driving a dynamical system are poorly understood. Such is the case of the seasonal behaviour of viruses. Time series analysis is a methodology readily used in public health and biomedicine applications (Zeger, Irizarry, & Peng, 2006). It addresses experimental situations where a sequence of observations is made over time to monitor equally spaced temporal changes of a phenomenon, such as the number of cases of viral disease in a population. The goal of the methodology is to generate simple descriptions, explanations, predictions or control of the underlying process. The plots of Fig. 1 for example are simple descriptions of a temporal recurrence. Data displays and summary statistics are created to better understand response variations over time. In some cases, the time series can be decomposed into components to reveal autocorrelations that can help us better describe the process (e.g. dissecting trends, seasonality, and residual effects). Sometimes a time series can be decomposed into a numerous series of components, using the Fourier transform or the discrete wavelet transform, for example. The time series is often re-expressed as cosine waves of arbitrary amplitude and phase. Amplitudes can be plotted against frequency to produce 'periodograms', and a 'spectrum' of the stochastic process can be estimated as the expected squared amplitude. When the goal of statistical analysis is explanation, the dependence of the process is modelled against one or more predictor time series using regression analysis. These predictor series are assumed non-random. Because neighbouring values of the process tend to be correlated, we can make predictions from past process responses or from predictor covariate time series with autoregressive models or distributed lag models. Time series regression models conditioned on past observations can take the form of marginal or conditional models, which often dissect a regression model for the mean and a model for the autocorrelation function. In the presence of Gaussian processes, autoregressive moving-average models display autocorrelation functions that decay or oscillate to zero with increasing lag. In some cases, the autoregressive process exhibits error and the models used to describe it are stationary and do not exhibit long-term trends and are therefore only appropriate in the absence of trends. Similarly, in cases where past responses influence the present through non-linear functions, we must apply nonlinear time series modelling approaches that for example can take into consideration response interactions. Finally, one popular approach to dissect the noisy time-series data that embody seasonal disease cyclicity, is the use of nonparametric autoregression methods (Hardle, 1990), including nearest-neighbours, splines, local polynomials and neural nets. These methods recursively

dissect error from chaotic behaviour that may govern cyclic disease patterns in biology (Sugihara, Grenfell, & May, 1990). For processes underlying periodic behaviour, a general form of spectral density may be assumed with increasing sample size that can dissect the autocorrelation structure of the stationary process (Hardle, Lütkenpohl, & Cheng, 1997). A recent analysis of global environmental drivers of influenza showcases the power of using nonparametric autoregression in the form of a smoothing spline to determine seasonal cycles (Deyle et al., 2016). The spline works on time series variables with and without sinusoidal waveforms.

3 The seasonal behaviour of COVID-19

Human coronaviruses are part of a group colloquially known as 'winter viruses', viruses that show peak incidences in the winter months. These viruses include influenza virus (Tamerius et al., 2011), human respiratory syncytial virus (Midgley et al., 2017), and rhinovirus pathogens (Morikawa et al., 2015). The incidence of endemic human coronaviruses increases during winter (Killerby et al., 2018; Li, Wang, & Nair, 2020; Monto et al., 2020; Nickbakhsh et al., 2020). Consequently, coronavirus seasonality as a group already suggests SARS-CoV-2 will likely exhibit a similarly marked seasonal behaviour once it becomes endemic.

The 2003 outbreak of SARS-CoV-1 and the 2012 outbreak of MERS-CoV did not advance enough to showcase seasonal behaviour. In contrast, the worldwide spread of the COVID-19 pathogen was already rampant during the first wave of the pandemic, which peaked in April 2020 some few months after its emergence in Wuhan (Fig. 4A). However, this same fact throws a monkey wrench onto detecting seasonal periodicities, as has been evident with the mismatched periodicities of influenza (e.g. Fig. 1B). As previously discussed, seasonal epidemics may arise because of the combination of environmental effects, the high transmissibility of the virus, the initial susceptibility of the human population, and the nature and degree of the immune response to infection. Seasonal patterns embody stable oscillations that can be dissected with epidemiological models. Under these circumstances, some degree of immunity occurs at the initial stage of an epidemic when the transmission rate is high and the immune response is weak (Grenfell & Bjørnstad, 2005; Hethcote, Stech, & Van Den Driessche, 1981). In such a case, even robust and substantial environmental drivers cannot stop transmission, as has been observed with previous influenza outbreaks. Researchers assume the same patterns of a seasonal epidemic are valid for COVID-19 and that the virus could produce seasonal oscillations if it becomes endemic (Kissler, Tedijanto, Goldstein, Grad, & Lipsitch, 2020). However, the pandemic continues to be rampant worldwide and new amino acid variants are arising that complicate vaccine mitigation strategies introduced massively at the beginning of year 2021. We anticipate robust oscillating patterns will be difficult to detect under these conditions.

Seasonality can be uncovered by focusing on correlations arising in a wide diversity of geographical areas or locations, preferably at worldwide distribution level, or

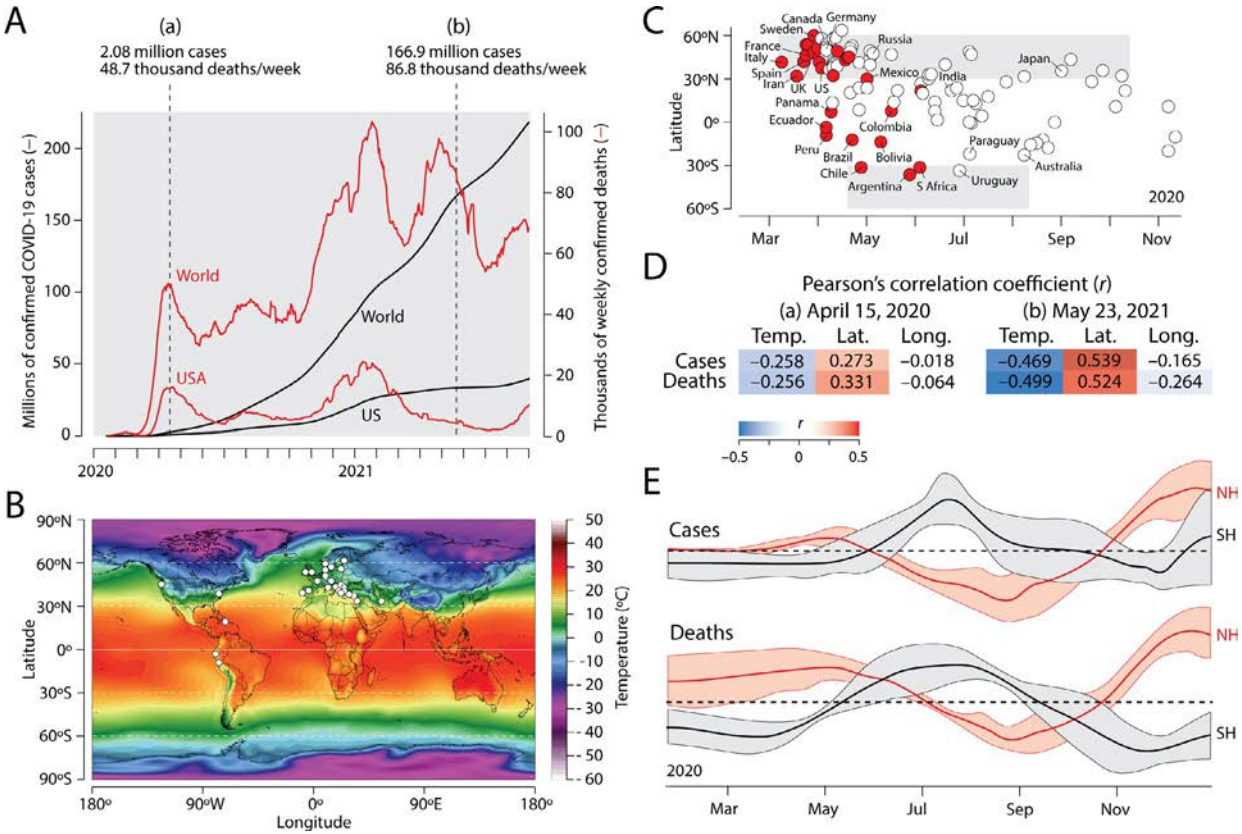

FIG. 4

See figure legend on next page.

ARTICLE IN PRESS

by studying time series describing seasonal recurrences on one or more particular geographical area or location. We will refer to these two distribution approaches 'spatial' and 'temporal', respectively, though spacetime is involved in both, discussing evidence supporting the seasonal behaviour of COVID-19.

3.1 Spatial distribution approaches

The amount of solar radiation (electromagnetic energy) received by any region of the world varies with geographic coordinates (especially latitude), time of day, the seasons, local weather and landscape. About half of incoming radiation in the form of photons (89 out of 174 petawatts) is absorbed by clouds, oceans and landmasses and converted into heat, but also electricity and chemicals that link to energy chains of

FIG. 4 The seasonal behaviour of COVID-19. (A) The plot shows the numbers of confirmed cases (black) and weekly deaths (red) since the start of the recorded COVID-19 pandemic. Three waves are clearly evident for worldwide weekly deaths recorded at the time of data collection from CSSE at John Hopkins University, September 1, 2021. Data gathered in April 15, 2020 (Burra et al., 2021) (a) and May 23, 2021 (Hernandez and Caetano-Anollés, ms. in preparation) (b) are highlighted in the timeline and its analysis summarized in panel (D). (B) Contour plot of temperatures taken 2 metres above Earth's surface for the entire world during the first wave of the pandemic (November 2019 to March 2020). Data was extracted from the NCEP/NCAR reanalysis (v. 1) project and rendered online (Climate Change Institute, University of Maine; climatereanalyzer.org). White circles represent countries with substantial outbreaks (>10 cumulative confirmed deaths per million people) as of April 15, 2020. With exception of Peru, Ecuador and the Dominican Republic, all country outbreaks occurred in yellow or green-shaded contours with temperatures ranging 5–11 °C. (C) A timeline of countries with substantial outbreaks (>10 cumulative confirmed deaths per million people) showing COVID-19 spread massively in a 30°N to 60°N corridor of latitudes during the first wave of the pandemic and then spread through a 30°S to 50°S corridor of latitudes during the onset of the winter season of the Southern Hemisphere. Red circles are countries exhibiting >300 confirmed deaths per million people as of November 15, 2020. (D) Effect of temperature (Temp.) and geographical coordinates of Latitude (Lat.) and Longitude (Long.) on number of cases and deaths worldwide normalized to account for population differences of the 211 and 191 countries analysed in 2020 and 2021, respectively. Correlations were considered significant when *P*-values were less than 0.05 and association strengths had Pearson's correlation coefficients *r* were higher than 0.1. (E) Seasonal oscillations detected in 5 countries of the Northern Hemisphere (NH), USA, India, Russia, France and the United Kingdom, and 5 countries of the Southern Hemisphere (SH), Brazil, Argentina, South Africa, Peru and Chile. Average seasonal signals and 95% confidence intervals are given in red and black shades for the two hemispheres, respectively. Seasonal signals were detected in the time series of cases and deaths using an adaptive one-dimensional time series analysis, the Ensemble Empirical Mode Decomposition (EEMD) method.

Panel (E) data from Liu, X., Huang, J., Li, C., Zhao, Y., Wang, D., Huang, Z., et al. (2021). The role of seasonality in the spread of COVID-19 pandemic. Environmental Research, 195, *110874.*

convection, circulation and biomass. The Earth's 23.5° tilted axis and its elliptical orbit around the sun make radiation strike the surface at different angles following the ecliptic of the celestial sphere. This causes differences in solar energy that create cyclic variations in temperature at daily and yearly levels. We note that temperature also varies with differences in altitude and topographical surface. For example, continents are generally warmer than oceanic regions in the Northern Hemisphere, while this reverses in the Southern Hemisphere tempered by the scarcity of land masses. If temperature variation impacts the triangle of viral persistence, a viral pandemic of COVID-19 proportions is expected to leave increasingly stronger worldwide seasonal signatures as the disease progresses. In addition, since distance from the Equator measured as latitude is a strong predictor of temperature, any temperature effects on seasonal behaviour detected by studying correlations of geographic coordinates with epidemiological data should be strengthened. The same can be said with other variables, such as elevation, which is correlated with UV radiation.

Indeed, initial studies during the first wave of the pandemic (Fig. 4A) suggested temperature and humidity levels of geographical areas (countries, regions, cities) were inversely related to epidemiological variables such as confirmed cases or deaths, while latitude showed a positive relationship. In many instances however results were inconclusive. A GRADE tool analysis of 17 initial explorations (mostly non-peer reviewed at that time) of data downloaded prior to March 1, 2020 (generally much earlier) revealed that only 4 were low risk (Mecenas, Bastos, Vallinoto, & Normando, 2020). Of these four, the only formally published study concluded that weather factors alone could not explain variability of the viral reproductive number in Chinese provinces or cities (Poirier et al., 2020). The meta-analysis suggested warm and wet climates reduced COVID-19 spread but also highlighted the importance of controlling confounding variables. Other initial studies of epidemiological variables were of significance. Demongeot, Flet-Berliac, and Seligmann (2020) showed that higher temperatures decreased infection rates in both French administrative regions and in 21 countries spread throughout continents. An association analysis of temperatures of countries or regions with epidemiological variables collected at the beginning of March 2020 and a time series predictive analysis with the Autoregressive Integrated Moving Average (ARIMA) model showed COVID-19 speed of contagion decreased with increasing temperatures. In addition, a focused study of how weather parameters of temperature, humidity and rainfall affected disease incidence ending March 2020 in Jakarta, Indonesia, revealed only average temperature exhibited statistically significant associations (Tosepu et al., 2020). The use of a log-linear generalized additive model to study cases and deaths in 166 countries with March 27, 2020 data showed that a temperature increase of only 1 °C reduced the number of daily reported cases by 3.08%, while a 1% increase in relative humidity reduced the number of daily reported cases by 0.85% (Wu, Jing, et al., 2020). The analysis concluded high temperature and humidity significantly affected COVID-19 epidemiological variables. Similarly, data from 429 cities also showed a negative correlation between viral transmission and temperature, reporting that a 1 °C temperature increase reduced the reported cases by 0.86% (Wang, Jiang, et al., 2020).

Temperature, population age, and tourism were the most important factors affecting COVID-19 risk measured with cases and deaths in 154 countries at the end of March 2020 (Yang & Lee, 2021). The study used a Pearson's correlation matrix but also modelled nonlinear associations of temperature and risk with random forest statistical methods. In contrast, Kassem (2020) reported the impact of temperature on cases per million in 43 countries divided into three groups according to first time disease introduction. There was only a slightly significant ($P < 0.1$) inverse relationship in only one of the groups.

While temperature and humidity were the initial focus, an analysis of 50 cities worldwide with and without community spread as of March 10, 2020 showed community transmission was restricted to a 30°N to 50°N latitude corridor with consistent weather patterns of 5–11 °C average temperatures combined with low absolute humidity (Sajadi et al., 2020). Poole (2020) also observed a similar corridor spanning 25°N to 55°N latitudes. Indeed, a world temperature map showed countries with significant COVID-19 outbreaks during the first wave of the pandemic (exhibiting >10 cumulative deaths per million people as of April 15, 2020) were located within a 30°N to 60°N latitude corridor in regions with 5–11 °C average surface temperatures (Fig. 4B). Moreover, a timeline of countries with significant outbreaks (>10 cumulative deaths per million people) describing how COVID-19 was advancing from China to the rest of the world for the most part of year 2020 shows that the disease began spreading through a latitude corridor in the Northern Hemisphere followed by one in the Southern Hemisphere (Fig. 4C). Almost half of outbreak countries were in the Northern corridor, 40% of which showed >300 cumulative deaths per million people as of November 15, 2020. A transition from the Northern corridor to Southern latitudes began to occur at the beginning of April with outbreaks in Peru and Ecuador (when temperatures started to lower in the Andean mountain range), then Brazil (beginning with the elevated Sao Paulo) but reached the Southern corridor in May and fully in June (when winter formally started in the Southern hemisphere). We also explored the worldwide role of temperature and geographical coordinates of latitude and longitude on epidemiological parameters for 211 countries at the peak of the first wave, April 15, 2020 (Burra et al., 2021). Both temperature and latitude were significantly correlated to COVID-19 cases and deaths with weak to moderate association values (Fig. 4D). Repeating the analysis during the third wave, a year later, showed these associations were becoming stronger (Figs 3 and 4D). This corroborates the impact of the geographical temperature-latitude correlate on COVID-19 seasonal behaviour, which is likely associated with the planetary effect of solar radiation. Thus, latitude, temperature and humidity appeared important factors during the first wave of the pandemic, consistent with the behaviour of a seasonal respiratory virus.

Given the strong temperature-latitude association we detected, COVID-19 seasonality is likely related to planetary-level environmental effects of solar radiation on viral transmission, directly through germicidal effects or indirectly through 'physiology' effects (vitamin D-related or through other immune regulatory activities). The low-energy infrared (IR) wavelengths transmit heat and could be responsible

for germicidal effects of temperature. Laboratory studies have shown that high temperatures and humidity levels shorten SARS-CoV-2 half-life in aerosols and other media, with virus persistence dramatically decreasing with small temperature increases (Chin et al., 2020; Matson et al., 2020; van Doremalen et al., 2020). The high-energy ultraviolet (UV) wavelengths (especially the UV-C spectrum) could have direct germicidal action on viral survival, as shown in laboratory experiments (Heilingloh et al., 2020; Seyer & Sanlidag, 2020). The expectation that UV radiation levels could be correlated with COVID-19 cases and deaths was recently fulfilled (Karapiperis et al., 2021). A machine learning-based analysis of data from 12 countries distributed across latitudes with comparable economic and demographic indices and epidemic surveillance statistics showed year-round UV radiation levels strongly associated with incidence rates, more so than with other variables. Remarkably, time series showed high UV levels depressed the number of cases in a hemisphere-related manner. These types of studies link geographical to temporal effects of seasonality and may lift some misconceptions about weather and seasonality (Carlson, Gomez, Bansal, & Ryan, 2020) by providing new methods of hypothesis testing.

3.2 Temporal distribution approaches

While studies have shown that COVID-19 affects some geographical regions more than others, a seasonal behaviour must also unfold with time in one location, with cases waxing and waning with the seasons (e.g. measles and influenza in individual cities; Fig. 1). This is particularly important because correlations often fail to portray cause-and-effect relationships, especially in nonlinear dynamic systems where persistent trends can often reverse with time or when drivers enslave specific responses (Sugihara et al., 2012). Time series analysis provides the tools to dissect seasonal cycles such as those of influenza (Deyle et al., 2016). Novel implementations of these methodologies have been applied to modelling and forecasting of COVID-19 cases, deaths and recoveries (e.g., Abotaleb & Makarovskikh, 2021; Barría-Sandoval, Ferreira, Benz-Parra, & López-Flores, 2021; Gecili, Ziady, & Szczesniak, 2021). However, seasonal oscillation signatures derived from time series are expected to be weak at the start of a pandemic but stronger as infections reach a seasonal endemic equilibrium. While such transition has been recently modelled across a range of immunity durations and demographic and social-mixing structures (Li, Metcalf, Stenseth, & Bjørnstad, 2021), there is no clear understanding of important seasonality factors (Smit et al., 2020). Liu et al. (2021) however recently extracted weak seasonal signals embedded in confirmed COVID-19 cases and deaths (as of December 31, 2020) in sets of five countries of the Northern and Southern Hemispheres. Using the Ensemble Empirical Mode Decomposition (EEMD), a method widely used in image processing and geophysical, financial and biomedical applications (Colominas, Schlotthauer, & Torres, 2014), the spectral analysis-based Hilbert-Huang transform decomposed (partitioned) multidimensional signal in the time series into different intrinsic 'mode' functions along with a trend. Systematic introduction of 'white' noise series to the original data followed by decomposition

in iterative steps allowed to obtain ensemble means of the mode functions for signal extraction. Remarkably, the signal curves showed COVID-19 cases and deaths were higher in colder climates and seasonality more pronounced at higher latitudes in an hemisphere-related sinusoidal pattern (Fig. 4E). This provides further confirmation of the seasonal behaviour of COVID-19.

4 Viral genomic make up and seasonality

RNA viruses are considered the most common aetiological agents of human disease. They represent 25–44% of all human pathogens, rivalling noxious bacteria (Jones et al., 2008). They trigger important pandemics such as those caused by the interspecies-transmitted Chikungunya and Zika viruses, while also embodying highly communicative zoonotic pathogens such as Ebola and Influenza A. RNA viruses exhibit exceptionally short generation times, high infection rates, and high levels of mutation and recombination, all leading to high genetic variability, rapid genome evolution, wide host range, enhanced virulence, evasion of host immunity, and increased resistance to antivirals.

RNA viruses have genome sizes ranging 2–32 kb, with sizes inversely proportional to mutation rates (Sanjuán & Domingo-Calap, 2016). RNA viruses make up 4 out of the 7 groups of the 'Baltimore classification' (see Glossary), which organizes viral genomes according to pathways used for mRNA synthesis (Baltimore, 1971). These groups include double-stranded RNA (dsRNA), positive sense single-stranded RNA (plus-ssRNA), and negative sense single-stranded RNA (minus-ssRNA) viruses and one of the two groups of RNA transcribing viruses, the single-stranded RNA viruses with DNA intermediate (ssRNA-RT). RNA viruses can be enveloped or nonenveloped with icosahedral, icosahedral nucleocapsid, helical nucleocapsid, or multilayered capsid virion morphologies. Genomes can be unsegmented, segmented, bi-segmented or tri-segmented. Such structural and functional diversity also translates into a wide diversity of life cycle strategies.

Coronaviruses belong to the order *Nidovirales*, a group or enveloped viruses with plus-ssRNA genomes that are capped and polyadenylated (Coronaviridae Study Group, 2020). Their genomes serve as mRNAs, which are directly translated into a functional proteome in the cytoplasm upon viral entry. Genomes also serve as templates for replication, folding into complex assemblies with modified cell membranes to construct scaffolds for replication, transcription and translation. The *Nidovirales* group expresses structural proteins coded in the 5′ region of the genome separately from the rest through sub-genomic mRNA.

Coronaviruses harbour the largest known non-segmented RNA genomes reported to date (~26–32 kb in length), a property that arises from evolutionary expansive trends that are unfolding in the *Nidovirales* (Lauber et al., 2013). Large genomes result in large proteome repertoires of ~30 proteins of large size, many with published atomic structures. The genome and proteome of SARS-CoV-2 is a clear example (Fig. 5A). Note however that the SARS-CoV-2 genome coding capacity has been established with

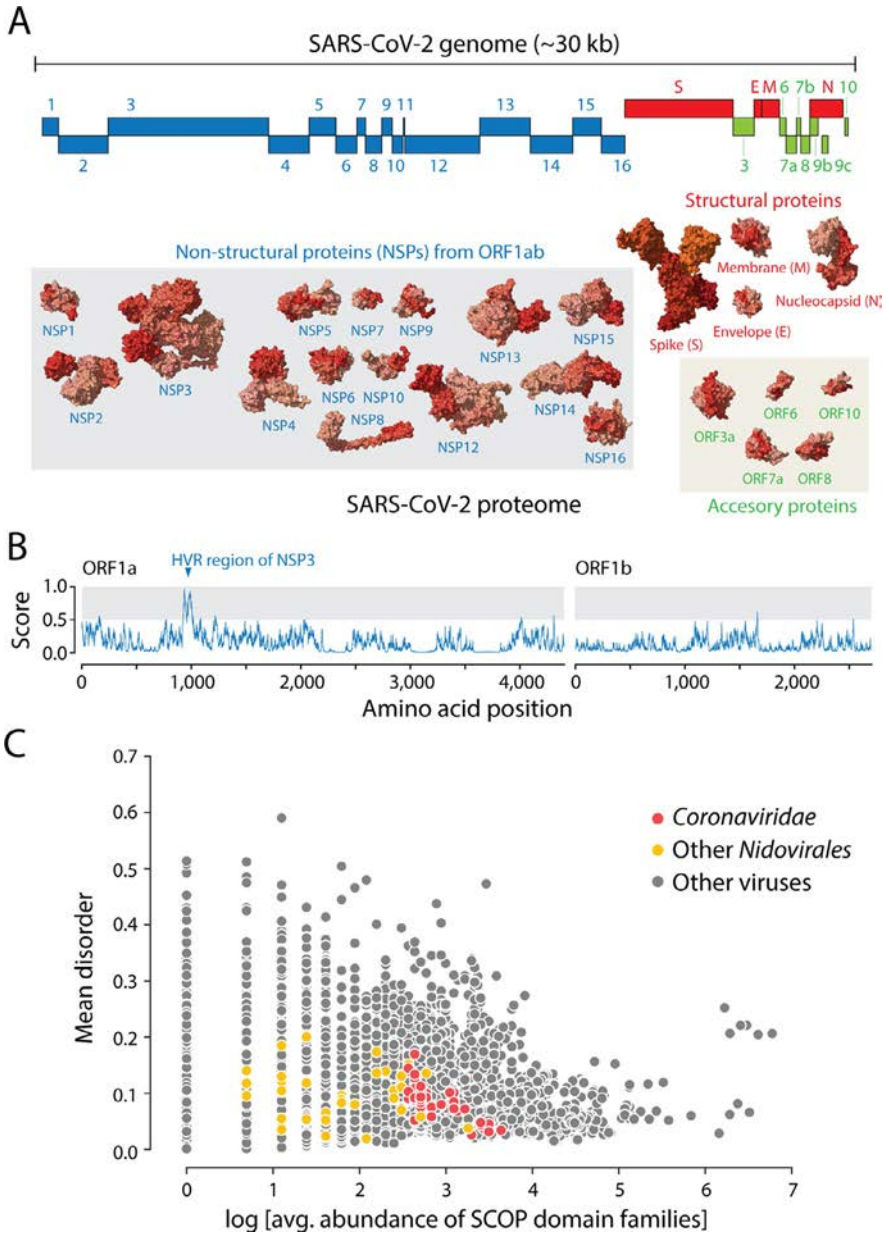

FIG. 5

The coronavirus proteome and protein intrinsic disorder (PID). (A). A diagram describing the SARS-CoV-2 genome and the corresponding proteome illustrated with models of molecular structure for its typical 29 non-structural, structural and accessory proteins. (B) Mapping PID with the IUPred2 algorithmic tool along the ORF1 and ORF2 predicted amino acid sequences, which encode overlapping polyproteins (pp1a and pp1ab) that cleave into 16 non-structural proteins (NSPs) of the SARS-CoV-2 reference (Wuhan) genome. Disorder scores above 0.5 indicate significant PID levels. Only the hypervariable region (HVR) of the large NSP3 papain-like protease scaffold shows significant PID. (C) Scatterplots showing the relationship between mean disorder and average abundance of structural domains defined at SCOP family level identified using HMMER in 6044 viral proteomes. *Nidovirales*, especially the *Coronaviridae* family, showed proteomes with significantly low levels of PID.

ribosome-profiling techniques and that a complement of 23 unannotated viral ORFs were identified that add to proteome complexity (Finkel et al., 2020). Coronaviruses adapt to host environments by relying on the low fidelity of a replication complex, which assembles around the NSP12 RNA dependent RNA polymerase with its extra N-terminal β-hairpin domain in interaction with an hexadecameric complex of disordered NSP7 and NSP8 primase cofactors (Gao et al., 2020; Kirchdoerfer & Ward, 2019). Coronaviruses also adapt to external environments by modulating flexibility of regular protein structure, such as the case of the structural transmembrane spike (S) protein, and by minimizing protein intrinsic disorder (PID) of the unstructured regions of their proteins (see definition in Glossary), especially modulating PID of the structural nucleocapsid (N), membrane (M) and envelope (E) proteins (Goh, Dunker, Foster, & Uversky, 2020a, 2020b, 2021; Tomaszewski et al., 2020). The flexibility of proteins is an intrinsic molecular property of regular structure present in helices and strands or in irregular motifs such as loops. This flexibility is necessary for the proper establishment of molecular functions. Protein flexibility of the spike (S) protein has been shown to have been enhanced by the widely distributed D614G mutation that unfolded early during the first wave of the pandemic. The coronavirus spike is a trimer of highly glycosylated S-protein protomers, each harbouring an N-terminal S1 subunit sequence with an N-terminal domain (NTD) and a receptor-binding domain (RBD) and a C-terminal S2 subunit holding a 'fusion' region with fusion peptide (FP) and internal fusion peptide (IFP) sequences, 2 heptad repeat (HR) sequences, and a transmembrane (TM) domain (see Glossary). The subunits are processed by host proteases upon viral entry. *In silico* modelling of the S-protein suggests the D614G mutant variant breaks a D614-T859 side chain hydrogen bond between the neighbouring S1 and S2 subunits of pairs of the three protomers of the spike, thereby enhancing flexibility and modifying their interactions (Korber et al., 2020). A more recent conformational dynamics study of the spike glycoprotein using cryo-EM showed D614G is heavily involved in the interaction with residues K854, K354 and Y837 of the fusion peptide (FP) region through side-chain atoms, contributing to linkage and/or allostery between the S1 and S2 subunits of neighbouring protomers (Xu et al., 2021). Flexibility and its associated molecular functionalities are also endowed by PID in regions lacking significant constraints on internal degrees of freedom of the polypeptide chain typical of fixed three-dimensional (3D) structure under native cellular conditions (Dunker, Brown, Lawson, Iakoucheva, & Obradović, 2002). These regions are highly dynamic and exhibit conformations that resemble either random-coils, molten globules or flexible linkers. Low PID is responsible for the rigidity of the coronavirus shell and the tropism-delimited transmission (respiratory or orofecal) mode of the virus (Goh et al., 2020b). PID has been proposed as a player in a pangolin-mediated spread of an attenuated human SARS-CoV-2 precursor that preceded the COVID-19 pandemic (Goh et al., 2021). Remarkably, SARS-CoV-2 retains the flexibility of the respiratory mode while enhancing the environmental resilience of the virion for efficient dispersion (Goh et al., 2020b). Fig. 5B illustrates the very low PID levels that we recently detected in the proteome of SARS-CoV-2 (Tomaszewski et al., 2020) with an analysis of ORFs encoding non-structural proteins (NSPs).

Remarkably, these low PID levels are not only a rigidity property of coronaviruses alone but of the entire *Nidovirales* group. A global analysis of 6044 viral proteomes showed levels of mean intrinsic disorder were less than 0.2 for the proteomes of the *Nidovirales* (Fig. 5C).

In Section 4, we focus on possible molecular culprits of seasonal behaviour. We study the variants of the SARS-CoV-2 proteome encoded in thousands of genomes sampled during the first wave of the pandemic. Analysis reveals that a number of molecular regions associated with clades and the highly infectious 'variants of concern' (VOCs; see Glossary) that arose in later waves are located in PID regions we initially identified that are necessary to sustain the structural integrity of virion structure. We also show that the N-terminal domain of the S-protein, which decorates the outer shell of the virion particle (crowning its 'corona' appearance), contains a galectin-like structure. Tracking the prevalence of mutations in this structure follows a seasonal pattern. We propose the galectin-like structure is a frequent target of mutations because it helps the virus evade or modulate physiological responses of the host to further its spread and survival. We end by suggesting that these regions are good molecular candidates for sensors of environment and physiology.

4.1 Mutational change in rapidly expanding viral populations

Viruses in general and RNA viruses in particular are considered an ideal system to study evolution because of their high error rates of replication, large populations, and short multiplication times (Manrubia & Lázaro, 2006).

RNA viruses have mutation rates of about 10^{-3} to 10^{-6} misincorporations per nucleotide in each replication round, orders of magnitude higher than the typical 10^{-6} to 10^{-8} observed for DNA viruses (Batschelet, Domingo, & Weissmann, 1976; Sanjuán & Domingo-Calap, 2016). This means that each newly synthesized viral genome will carry an average of one to two mutations when aligned with the parental sequence, matching known error rates of 1.1 nucleotide deletion/substitution per genome per round of replication (Drake, Charlesworth, Charlesworth, & Crow, 1998). We note that substitution rates are not only raw measurements of genetic changes (point mutations, insertions, deletions, rearrangements) occurring in a genome but also depend on the generation time and viral population sizes, not to mention consideration of fitness. There are three main reasons for high mutation levels: (1) copying errors during the viral replication process, (2) recombination in the viral genome, and (3) host-induced RNA editing systems. In RNA viruses, high mutation rates largely arise from their RNA-dependent RNA polymerases lacking proofreading activity, which allows them to adapt to environmental changes (Domingo, 2002). In addition, single-strand RNA viruses such as coronaviruses are more prone to mutations than double-strand genome viruses as single nucleic acid strands are more susceptible to chemical damage and oxidative deamination (Seronello et al., 2011). Higher mutation rates enable viral adaptation to be faster over time (Berngruber, Froissart, Choisy, & Gandon, 2013). However, not all mutations are adaptive and helpful; most of the mutations reported in RNA viruses are

deleterious and hinder their viability (Johnson & Barton, 2002). This puts RNA viruses close to an 'error threshold' of too many deleterious mutations for successful persistence (Domingo, Sheldon, & Perales, 2012). Remarkably, coronavirus proofreading abilities are crucially enhanced by the 3′-to-5′ exonuclease domain of the NSP14 protein (Minskaia et al., 2006), which increases 12- to 15-fold template copying fidelity (Eckerle et al., 2010) and enables both genome expansion and mutation robustness (Smith & Denison, 2013). Genetic recombination and genomic reassortment also contribute to viral diversity and emergence of new phenotypes. In recombination, two or more segments from two different parental genomic molecules combine and produce a new mutant or recombinant phenotype that may be more pathogenic and virulent than the parental molecules (Posada, Crandall, & Holmes, 2002). Recombination may accelerate viral evolution, but more research is needed to evaluate the exact role it plays (Rhodes, Wargo, & Hu, 2003). Genomic reassortment occurs in tripartite genomic RNA viruses like influenza in which new variants arise after the simultaneous entry of two or more different viruses into a host cell (Vijaykrishna, Mujerki, & Smith, 2015). Recombinants produced by genetic recombination exhibit new properties that can significantly enhance viral diversity (Earn, Dushoff, & Levin, 2002). In fact, the high recombination rates of RNA viruses are directly associated with their virulence (Khatchikian, Orlich, & Rott, 1989), the host range (Gibbs & Weiller, 1999), host immunity (Malim & Emerman, 2001), and resistance to antivirals (Nora et al., 2007).

While mutation frequency determines genetic variations in a viral population and speeds up viral evolution (Sanjuán & Domingo-Calap, 2016), mutational robustness is dependent on population size and how viruses manage the detrimental effects of mutations. While most random mutations (approximately 40%) are lethal, deleterious mutations that are non-lethal (approximately 30%) lead to a reduced effectiveness of the virus (Sanjuán, Moya, & Elena, 2004) or can have positive effects, as has been reported in bacteriophage φ6 (Burch & Chao, 2004), vesicular stomatitis virus (Sanjuán et al., 2004), and HIV-1 (Bonhoeffer, Chappey, Parkin, Whitcomb, & Petropoulos, 2004). However, beneficial mutations may occur if the environment changes rapidly, if a multiplicity of ecological niches are present, or if other factors adding to the viral mutational landscape enhances novelty generation, including rearrangement/recombination, changes in genome size, host-encoded modifiers, and replication mode (Sanjuán & Domingo-Calap, 2016). Regardless of how viruses manage mutations, changes in population size can influence the frequency of a given mutation and the identity of the viral population as a whole. For example, small populations let genetic drift cause stochastic fixation of mutations in sequences that are conserved and are present in the parental genome. Because most of the progeny will acquire these mutations, the identity of the population is maintained, very much as it occurs in higher organisms. In other words, a small population size guarantees 'survival of the fittest' by competition and natural selection of the best-adapted variants among a limited number of genomes (Earn et al., 2002). When populations are large, a swarm of genetic variants revolving around a consensus sequence is formed in a phenomenon known as 'the survival of the flattest' (Lauring & Andino, 2010).

Here, viruses with larger numbers of variants are retained, limited only by the 'error threshold' and the catastrophic possibility of viral extinction. Viral identity is lost and replaced by a 'quasispecies', a group of different variants that show genetic linkage through mutation, have shared functions, and collectively contribute to the characteristics of the population.

4.2 A structuring quasispecies as paradigm of viral mutational change

For the last 30 years, the quasispecies concept provided a population-based framework to elaborate viral evolution, especially that of RNA viruses (Lauring & Andino, 2010). Although this framework is based on classical population genetics, its main objective is to explain the results of error-prone replication, describe viral evolutionary dynamics, and test predictions in model organisms (Domingo et al., 2006). As it is clearly evident from tracking the expanding COVID-19 pandemic with millions of viral sequences (all publicly available in the GISAID viral data repository), the single reference sequence (accession NC_045512.2, version March 30, 2020; previously 'Wuhan seafood market pneumonia virus') rapidly transformed into a cloud of diverse amino acid variants in each and every host, which can be described using accepted nomenclature from the Human Genome Variation Society—for example, the first major and now universal variant D614G of the spike makes explicit that a glycine (G) has substituted an aspartic acid (D) in site 614 of the reference sequence. These variants are organized around that 'master sequence', which in our case represented the original viral infection that jumpstarted the pandemic. Indeed, initial SARS-CoV-2 mutations recurrently resulted in genetically related variants that assembled into phylogenetic groups with common origin, i.e., clades, all unified by a single origin (root) in a phylogenetic tree (see Glossary) meticulously tracked by for example the GISAID repository (Elbe & Buckland-Merrett, 2017). Fig. 6 shows an example maximum likelihood tree reconstructed from 3615 sampled genomes using the Nextstrain online system (Hadfield et al., 2018) with groups labelled as either GISAID clades or emerging VOCs. So far, variants have been clustered into ten clades: L, S, O, V, G, GH, GR, GV, GRY (Mercatelli & Giorgi, 2020) and GK. Clade L arose in Wuhan and dominated SARS-CoV-2 spread prior to January 2020, a time when clades S and O appeared. Clades V and G appeared at the middle of January 2020, followed by clades GH, GR appearing at the end of February, clades GV and GRY appearing in June and September 2020, respectively, and clade GK appearing in October 2020. While clades L, V and S are becoming extinct, clades G, GH, GR, GV, GRY and GK embody the current genomic makeup of the virus (Singh, Pandit, McArthur, Banerjee, & Mossman, 2021). This makeup is now producing VOCs that are structuring in more complex ways the viral quasispecies. To illustrate, note how the subtree defining the GK clade that embodies VOC delta embeds viruses from other clades (mostly GRY and GV), suggesting an important role of horizontal processes of genetic exchange such as recombination (Fig. 6). This questions the suitability of using phylogenetic trees without reticulations to describe this new

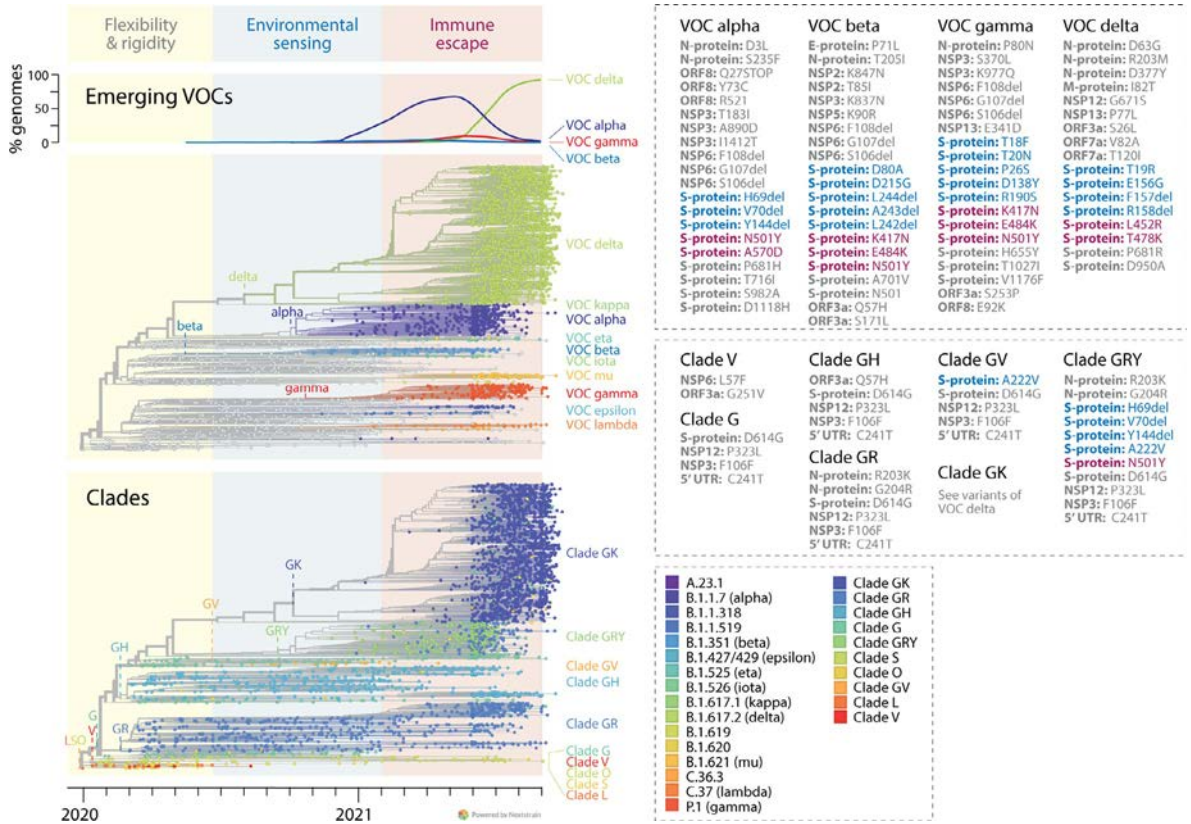

FIG. 6

The mutational landscape of the SARS-CoV-2 virus. A maximum likelihood phylogenetic tree of 3615 genomes sampled from more than a million GISAID worldwide sequences between December 2019 and September 2021 obtained from Nextstrain (https://nextstrain.org) shows groups of taxa (leaves of the tree indicated with circles) labelled as either GISAID clades or emerging variants of concern (VOCs) and their origins along branches of the trees. Proportions (given as smoothed percentages) of genomes holding major emerging VOCs along the evolutionary timeline is shown above the trees. The timeline of clades and VOCs and their signature mutations (listed on the right) show three successive phases driven by proteome flexibility and rigidity, environmental sensing, and vaccine-driven immune escape. Variants coloured in grey, blue and purple are linked to the three successive drivers, respectively. Variant lists of VOCs do not include the 4-variant D614G haplotype nor the N-protein R203K and G204R mutations, which are widespread in all VOCs. See Bernasconi, Mari, Casagrandi, and Ceri (2021) for a 'lineage directory' of markers with changes in more than 75% of sequences.

ARTICLE IN PRESS

diversification process. In fact, a recent analysis of mutations in SARS-CoV-2 genome sequences in the United States appear to suggest that mutations are being accumulated and sequences convert to VOCs through serial 'super spreader' founder events in a sea of mutational bursts (Tasakis et al., 2021). Similarly, the rise of VOCs appears to coincide with a major global shift in the selective landscape that involves convergence between lineages (Martin et al., 2021).

An increasing body of molecular evidence suggests we are witnessing major episodes of structuring of the evolving SARS-CoV-2 quasispecies. These episodes can be identified by studying the accumulation of sets of highly prevalent variants, which we here call 'markers'. These markers recurrently combine with a highly diverse set of other variants of the master sequence. This marker-linked 'cloud' of variant combinations define a mutational landscape that the virocell explores to optimize. Fig. 6 dissects possible diversification mechanisms by showing a timeline of clades and VOCs together with their most characteristic marker mutations. Based on these markers, we elaborate a sequence of structuring episodes triggered during different phases of the pandemic by: (i) enhancing viral survival through a careful balance of protein flexibility and rigidity, (ii) facilitating viral sensing of the environment as winter approached the Northern Hemisphere, and (iii) evading vaccines and therapeutics that were massively introduced at the beginning of year 2021. This last phase involves strategies of 'immune escape' (see Glossary), i.e., the search of mutants that override the innate and adaptive immunity mechanisms of the host. The next subsections will provide genomic evidence to support our contentions.

4.3 The mutational landscape of evolving SARS-CoV-2 viruses

A vast body of literature has been amassed describing the millions of SARS-CoV-2 genomes that have been sequenced so far. These efforts have been abundantly reviewed (e.g. Harvey et al., 2021; Huang, Yang, Xu, Xu, & Liu, 2020; Majumdar & Niyogi, 2021; Singh et al., 2021). Many mutations have been studied in different viral proteins that might affect viral replication, transmissibility, and virulence. One clear example has been the mutations of the S-protein of the spike (Harvey et al., 2021; Huang et al., 2020), but also of the nucleocapsid protein, the N-protein.

Prompted by the initial arrival of COVID-19 to the United States, we developed an entropy-based algorithm to explore the mutational landscape of the SARS-CoV-2 virus (Tomaszewski et al., 2020), which detected 27 high-entropy mutations at the amino acid level in an initial set of ~25,000 genomic sequences retrieved on May 7, 2020 from the GISAID worldwide repository (Table 2). Our method uses Shannon informational entropy to describe the amount of variation existing in discrete per-location nucleotide or amino acid composition data. We used it to quantify population diversity and selection with two independent state variables, as suggested and tested with H3N2 influenza by Pan and Deem (2011). The first variable uses mutational entropy as a measure of molecular biodiversity, being maximal for example when all amino acids in an amino acid site of the sampled viral proteomes are equally represented. Entropy is zero when only one amino acid overtakes that site

Table 2 The SARS-CoV-2 mutational landscape at the peak of the first wave of the pandemic.[a]

Protein	Site	Entropy Δ[b]	Str	Location	Region	IUPred2	Anchor2
M	175	–0.136	loop	surface	ordered	0.2167	0.2647
N	13	0.050	loop	surface	disordered	0.8781	0.9749
	193	–0.073	loop	surface	disordered	0.7912	0.4214
	197	–0.064	loop	surface	disordered	0.8158	0.4259
	203	**0.157**	**loop**	**buried**	**disordered**	**0.7573**	**0.5969**
	204	**0.143**	**loop**	**buried**	**disordered**	**0.7605**	**0.621**
S	**614**	**–0.220**	**loop**	**surface**	**ordered**	**0.1266**	**0.3609**
NSP1	75	–0.016	loop	surface	ordered	0.3184	0.4256
NSP2	85	0.055	loop	surface	ordered	0.0771	0.261
	212	–0.120	loop	surface	ordered	0.2748	0.2408
	559	–0.151	loop	surface	ordered	0.1921	0.345
	585	–0.163	loop	surface	ordered	0.4245	0.4187
NSP3	58	–0.101	helix	surface	ordered	0.0414	0.2217
	153	–0.020	loop	surface	disordered	0.8074	0.9536
NSP4	308	–0.071	loop	surface	TM	0.0004	0.0003
NSP5	15	0.114	helix	surface	ordered	0.0813	0.0843
NSP6	**37**	**–0.121**	**loop**	**surface**	**ordered**	**0.0003**	**0**
NSP12	**323**	**–0.222**	**helix**	**surface**	**ordered**	**0.0137**	**0.0299**
NSP13	504	–0.226	loop	surface	ordered	0.115	0.4329
	541	–0.231	loop	surface	ordered	0.3053	0.3068
3a	13	0.082	loop	surface	ordered	0.0587	0.029
	57	**0.066**	**helix**	**pore**	**TM**	**0.001**	**0.0014**
	196	–0.072	loop	surface	ordered	0.115	0.2522
	251	**–0.217**	**loop**	**surface**	**disordered**	**0.4703**	**0.4235**
8	24	0.105	helix	surface	ordered	0.0395	0.004
	62	–0.015	loop	surface	ordered	0.1635	0.0205
	84	**–0.259**	**loop**	**surface**	**ordered**	**0.0405**	**0.0098**

[a]*An examination of relative entropy delta values and the structure (Str: loop, helix, strand), location (surface, buried, pore) and region [ordered, disordered, transmembrane (TM)] involving mutation sites are provided together with IUPred2 scores of intrinsic disorder and Anchor2 estimates of binding propensity. IUPred2 and Anchor2 scores above 0.5 support significant intrinsic disorder and binding energy of that amino acid site, respectively. Data from Tomaszewski et al. (2020).*
[b]*Negative values of relative entropic delta imply entropy reversals that result in increased fitness and trends towards fixation. Positive values imply entropy expansion fostering trends towards site diversification. Rows in bold are sites that have been confirmed as positively selected and exhibiting population level expansions as of March 2021 (from table 2, Singh et al., 2021).*

ARTICLE IN PRESS

(suggesting fixation) and its value increases when mutations are diluted in the viral cloud. The second variable is a relative measure of entropy, 'relative entropy delta', which measures selection pressure by comparing viral diversity at two different time points of a time series. Entropy reversal trends suggest mutations are positively selected to enhance fitness while entropy expansions suggest other evolutionary forces are at play. We note entropy expansions are expected to precede reversals since they generate novelties which are then culled in a reversal phase.

Out of the 27 high-entropy mutations identified, 19 were entropy reversals (Table 2). Five of these have been confirmed as positively selected and exhibiting population level expansions (as of March 2021; Singh et al., 2021). Another 4 positively selected mutations were associated with the PID-rich serine/arginine-rich linker of the N-protein, R203K and G204R, and the viroporin protein encoded by ORF3a, Q57H and G251V. They were in either entropic expansion or reversal modes, respectively. Most of 27 high-entropy mutations were located on the surface of the molecules once mutations were traced onto three-dimensional atomic models (Tomaszewski et al., 2020). The only exceptions were the R203K and G204R mutations of the N-protein, which were buried, and the G57H mutation of the ORF3a viroporin, which was located on an internal transmembrane domain. Remarkably, seven of the high-entropy mutations were located on PID regions of significant disorder of the N-protein, the NSP3 papain-like protease, and the ORF3a viroporin (Table 2). Three of these were positively selected.

Our analysis identified the well-known D614G mutation of the S-protein had coordinated entropic trends with the P323L mutation of the NSP12 polymerase that mediates viral replication. The D614G mutation, which is already fixed worldwide in the current viral population, was part of a haplotype of four mutations altering the S and NSP12 proteins, 5′ UTR, and silently the NSP3 papain-like protein (F106F) during the first wave. This haplotype defined the G-clade that originated in China and was established in Europe in the middle of January 2020 following its first report in Germany (Korber et al., 2020; Phan, 2020). A time series of variant appearance revealed that the D614G and P323L haplotype marker of the G-clade had spread through continents during the January-April period of the first wave visualized through genomic analysis of the thousands of sequences we had available at the time. The haplotype is believed to have increased infectivity (Becerra-Flores & Cardozo, 2020; Korber et al., 2020), and as previously discussed, to have enhanced protein flexibility (Korber et al., 2020).

We also revealed accumulation of high-entropy variants enhancing flexibility in PID regions and other regions that were of interest. The N-protein plays critical roles in maintaining viral structure and viability once the virus enters the cell, helping viral replication, transcription and packaging (e.g. Woo, Lee, Lee, Kim, & Cho, 2019). The N-proteins holds two major RNA-binding domains, an N-terminal domain (NTD) and a C-terminal domain (CTD) connected by a central serine/arginine-rich linker and flanked by terminal sequences. We found these linker and terminal sequences were PID regions and their contribution to the entire molecule made

the N-protein a 'highly disordered' molecule. All identified high-entropy mutations (Table 2) occurred in loop regions of the N-terminal and linker PID regions. Two particularly important mutations were in the linker, R203K and G204R, which were already present in January but later defined the GR-clade that appeared late February and the GRY-clade of September. Their high Shannon entropy values of −0.8 suggested a coordinated active exploration of these two mutation sites for novel genotypes, which was later reflected by its worldwide spread through clades GR and GRY later in year 2020 and in VOC delta in 2021 (Fig. 6). In fact, the terminal and linker PID regions flanking the high-entropy mutations we detected were particularly targeted by major VOCs alpha and delta. This strongly suggests PID flexibility is providing an important functionality to the SARS-CoV-2 molecules. Three more focused studies revealed a coordinated mutational push associated with the N-protein that followed that of the S-protein. Ye, West, Silletti, and Corbett (2020) identified 10,983 amino acid substitutions in 38,318 genome sequences of SARS-CoV-2 N-protein retrieved June 3, 2020, prior to the appearance of the GRY clade. Remarkably, 94% of mutations localized to the linker and terminal regions, with 9567 of them located in the central serine/arginine-rich linker. Similarly, Rahman et al. (2021) analysed 61,485 sequences of the N-protein retrieved on July 17, 2020 (prior to the appearance of clade GRY) to study the functional roles and structural importance of amino acid sites in diagnosis and vaccine development. They detected 1,034 unique mutations, a third of which occupied primer binding sites used in PCR diagnostic reactions. A group of 684 amino acid substitutions in 317 positions revealed that only 165 targeted the NTD and CTD RNA binding domains while the rest were located in the disordered linkers and terminal sequences. Troyano-Hernáez, Reinosa, and Holguín (2021) retrieved 105,276 genomic sequences available in September 2020 and studied mutations in all structural proteins of SARS-CoV-2. Despite extremely high conservation of structural proteins (>99%), they found 291, 142, 2,671 and 890 mutational changes in the M, E, S and N proteins, all of which were targeting 74–89% of available amino acid sites. Rates of 1.76, 2.18, 2.35 and 2.5 mutations/site for the M, E, S and N proteins, respectively, showed significant mutational targeting of the nucleocapsid components of the virion. The study also displayed the centrality of the D614G mutation of the S-protein, with a prevalence of 81.5%, and of the linked R203K and G204R mutations of the N-protein, with a prevalence of 37%. This showed these central flexibility-driven mutations were widely circulating before the emergence of VOCs in October 2020 (Fig. 6). These protein flexibility-driven heterogeneities that are becoming established with time are expected. They embody diverse 'quasispecies' dynamic behaviours unfolding in the presence of extreme genetic variation (Domingo et al., 2002) that followed the first wave of the pandemic. However, further studies are required to explore the precise mechanism and functional role of nucleocapsid PID regions that are selectively targeted by mutation/deletions in replication and viral pathogenesis. These deletions may influence the tertiary structure and function of the protein and subsequently affect virus-host interaction, immune modulation, and the pathogenic nature of the virus (Phan, 2020).

The N-protein and other high-entropy mutations established likely pathways of significant mutational change associated with protein flexibility and disorder (Tomaszewski et al., 2020), some of which are currently being used over a year later by widely distributed VOCs (Fig. 6). VOC alpha markers D3L and S235F and VOC delta markers D63G, R203M and D377Y are all located in PID regions highlighted by our initial high-entropy mutants. The large NSP3 papain-like protease that initiates cleavage of the non-structural ORFs exhibited a high-entropy mutation (P153L, listed in Table 2) that targeted the N-terminal PID-rich hypervariable region (HVR) of the molecule (Fig. 5B). This PID region now holds marker T183I for VOC alpha. Similarly, the flexible ORF3a and ORF8 accessory viroporin proteins showed high-entropy mutations in regions that embody markers for VOC alpha, gamma and delta, including the S26L mutation in a 5′-terminal disordered tail of the ORF3a viroporin and mutations Q27STOP, Y73C, and R52I of the 5′ terminal region of the ORF8 viroporin we previously identified as significant.

While the search for variants that would modulate protein flexibility and rigidity continued over time, the remarkable appearance in the middle of 2020 of a number of markers in clades GV and GRY (labelled in blue) that were located in the N-terminal NTD region of the S-protein suggested a different mutational landscape was also unfolding (Fig. 6). These markers included A222V and three deletions, H69del, V70del, and Y144del. Later on, VOC alpha expressed two additional related deletions (H69-V70del and Y145del) and VOC delta added the terminal T19R marker. These and many other NTD-associated markers appeared as the pandemic was again approaching winter in the Northern Hemisphere. The S-protein has a solvent exposed region that mainly holds the NTD and RBD domains of the S1 subunit, and a buried region comprising the transmembrane region of the S2 subunit, which associates predominantly with the lipid bilayer of the capsid (Fig. 7D). In proteins there is a continuum between rigidity, flexibility and disorder, which has been only recently characterized (Akhila et al., 2020). A simple exploration of PID levels however already shows the S-protein is a significantly rigid structure (Fig. 7B). The N-terminal S1 subunit was less ordered than the transmembrane S2 region, with scores ranging from 0 to 0.3. The Anchor2 algorithm of IUPred2 predicted binding regions based on sequence, identifying segments that had the capacity to gain energy by interacting with a globular partner protein. Two segments of the NTD had such binding properties, flanking the central domain structure. Some significant mutations occurred in the more flexible subtending region, such as L18F, T19R and H69_V70del. We also traced the location of some significant variants onto a crystallographic atomic model of one of the three S-protein monomers of the spike (Fig. 7D). Most mutations were located on the surface of the molecule, including those in the NTD region, which localized to loops of its 13-stranded 3-layered β-sandwich galectin fold structure with two antiparallel β-sheets stacked against each other through hydrophobic interactions (Peng et al., 2012). While RDB carries the central role of binding to Angiotensin Converting Enzyme 2 (ACE2) on the host cells, NTD appears to have a binding role related to immune responses that is uncertain but likely associated to temperature and environment. While RBD is

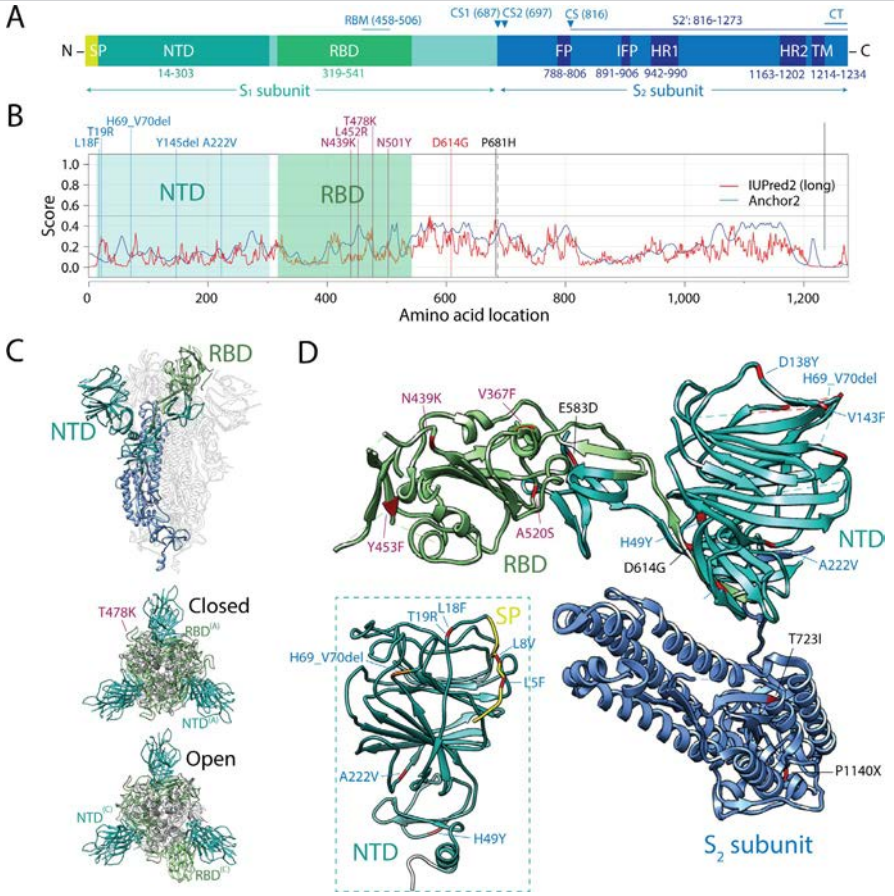

FIG. 7

The rigid yet flexible structure of the viral spike. (A) A diagram describing the different regions of the S-protein molecule, including their lengths in amino acids. SP, signal peptide; NTD, N-terminal domain; RDB, receptor-binding domain; RBM, receptor-binding motif; CS, cleavage site; FP, fusion peptide; IFP, internal fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; TM, transmembrane domain; CT, cytoplasmic tail. (B) The S-protein is a significantly rigid structure. Mapping PID (red line) and gain-loss of binding energy (blue line) with IUPred2 along the S-protein sequence shows the protein is highly ordered with scores <0.3–0.4 in the N-terminal S subunit that holds the NTD and RBD domains. Disorder scores above 0.5 indicate significant PID levels and proteins with 0–10% of PID residues are considered 'highly ordered' (the S-protein has none). Significant variant markers are indicated, which usually coincide with more flexible regions. (C) The S-protein is given in the context of the spike homotrimer described with an atomic model (PDB entry *6vxx*). Chain A is highlighted with its differently coloured NTD and RBD domains. Top views of the spike complex are shown in the bottom in closed (resting state; PDB entry; *7df3*) and open (ACE2 triggered; PDB entry *7dk3*) states. Note that only one of the three S-proteins has the RBD upward conformation exposing the S2 subunit of a different protomer. The T478K mutation of the RBM is shown to highlight its external loop location. (D) Representative mutations are mapped onto a crystallographic atomic model of one of the three protomers of the spike (PDB entry *6vxx*). The model is displayed in cartoon (ribbon) format with the polypeptide visualized from the top of the coronavirus crown (slightly tilted when compared to the top views in panel C) and showing the tripartite structure of the spike monomer. The NTP, RDB and membrane bound domains coloured in light sea green, light green and blue are clearly evident. Note that the signal peptide (SP) and some small segments are missing in the model. The inset shows a complete NTD model obtained with I-Tasser by the Zhang lab (GenBank: QHD43418.1) displaying the SP and N-terminal region of the domain, with locations of most important mutations highlighted.

immunodominant (i.e. is the preferred epitope), changes in the NTD region alter antigenicity (reviewed by Harvey et al., 2021). In fact, an 'NTD supersite' consisting of NTD residues 14–20, 140–158 and 245–264 is involved in antibody neutralization by inhibiting conformational changes or interactions with auxiliary receptors of its structurally plastic structure. NTD is also central from an evolutionary perspective. An analysis of 'evolutionarily important' residues of the S-protein detected by sequence conservation showed they were concentrated in two domains, the NTD and the RBD, both of which had a role of host cell binding in a number of related viruses (Saputri et al., 2020). While many evolutionarily important residues were detected in NTD, suggesting an involvement in host receptor binding, the study showed 'importance' correlated with structural flexibility. Remarkably, all atom molecular dynamics simulations of the SARS-CoV-2 S-protein revealed protein flexibility was dependent on external temperature (Rath & Kumar, 2020). In the simulations, the RBD was less mobile than the NTD, with flexibility limited mostly to the receptor binding motif. Increasing temperature made a number of charged polar amino acid residues on the top layer of the more flexible NTD less solvent exposed. These residues involved loops delimited by the N-terminal β-strand, β8-β9, β9-β10 and β14-β15, which harbour most mutations highlighted in Fig. 7D (including the inset). Similarly, increases in temperature closed the flexible binding motifs of the RBD in the trimer, which are usually in an open conformation. This conformation sealed the visibility of the trimeric pore when visualized from the top of the spike, burying the receptor binding residues necessary for contacting the ACE2 receptor. Thus, temperatures above 40 °C resulted in the inactivation of the spike, possibly in a reversible manner and without loss of secondary structure. The combination of mobile rigid elements encompassed by deformable regions that act as hinges may be the mechanism behind this environmental modulation. Given these results, the putative binding properties of NTD are likely to have a regulatory role related to environmental responses, which we here sought to confirm.

Finally, the significant appearance in clade GRY of the N501Y mutation (Fig. 6) revealed the early accumulation of a marker in one of the six key contact residues of RBD shown to increase both ACE2 receptor affinity and infectivity and virulence (Starr et al., 2020). A cryo-EM analysis of receptor-triggered dynamic transformation from the closed prefusion state of the spike to the fusion-prone open state (Fig. 7C) exhibited enhanced sensitivity of ACE2 through the T470-478 loop and Y505 site of RDB viral determinants (Xu et al., 2021). Remarkably, T478K has become a significant marker of VOC delta. The appearance of VOC alpha at the end of 2020 brought also the P681H marker of the S-protein, one of four residues comprising the insertion that creates a S1/S2 furin cleavage site between S1 and S2 in spike, which is known to promote viral entry and transmission in animal models. Harvey et al. (2021) carefully reviewed how mutations associated with the main binding function of the S-protein and other functional regions that appeared in this later phase of the pandemic contributed to 'immune escape' by changing antigenicity and transmissibility but also reducing the efficiency of vaccines, convalescent plasma, and post vaccination serum. None of these markers were significant in

our initial exploration of high-entropy mutations during the first wave of the pandemic, showing they represented novel targets of dynamic change in the viral population.

Thus, the three structuring phases proposed in Fig. 6, *'flexibility/rigidity'*, *'environmental sensing'*, and *'immune escape'*, are compatible with the gradual appearance of markers over time associated with multiple PID regions, a putative molecular sensor located in evolutionarily important sites of the N-terminal NTD region of the S-protein, and the central receptor-binding function of the RBD region, respectively.

4.4 Viral genomic change and seasonal behaviour

Because COVID-19 transmission was initially restricted to a 30°N to 50°N latitude corridor of the Northern Hemisphere, seasonal behaviour had to unfold in countries of those regions during the first wave of the pandemic. We reasoned NTD mutations in the S-protein that were emerging in the middle of 2020 were likely associated with a molecular region that was already acting as environmental sensor at a much earlier time. To test this hypothesis, we used recently published genomic data (Showers, Leach, Kechris, & Strong, 2021) to track spike mutations as they emerged during almost the entire span of year 2020. In that study, variants were identified in 139,835 filtered SARS-CoV-2 genomic sequences retrieved worldwide on November 14, 2020. Their weekly prevalence was then used to construct heat maps with cells ordered in time from the first week of January to the last of October of 2020. To dissect a possible environmental sensing role of the NTD region of the spike, we dissected the time series of S-protein variants in search of seasonal patterns.

A focus on year 2020 provided the best opportunity to find a seasonal pattern in the data, especially because vaccine introduction in 2021 was expected to foster viral immune evasion strategies and offset any possible signal. If SARS-CoV-2 acted as a 'winter virus', it would push prevalence of mutations in molecular 'sensor' regions in a seasonal manner in search for better modulation of sensing functions. That should translate into mutational exploration through bursts of weekly prevalence of variants occurring first during the winter. Most of these would not be retained or would be revisited at low levels in the timeline. Our expectation was that this prevalence effect should disappear during the summer but should be recovered later in the year with new bursts and variants. To study bursts, we reordered the time series by time of significant appearance of variants in the time series and by grouping variants according to their location in regions of the S-protein. Using different filtering criteria, the most relevant variants could be followed in time, making evident any detectable seasonal pattern present in the data.

Our expectations were fulfilled. The rearranged heat map of widely abundant variants worldwide made explicit a chronology of major mutational spreads and bursts within a 'sea' of amino acid changes expressed at low level in the sampled viral population. For example, Fig. 8A shows a time series describing the weekly abundance of individual variants present in more than 2% of sequenced genomes worldwide ordered by time of significant appearance in the 'N-terminal' region

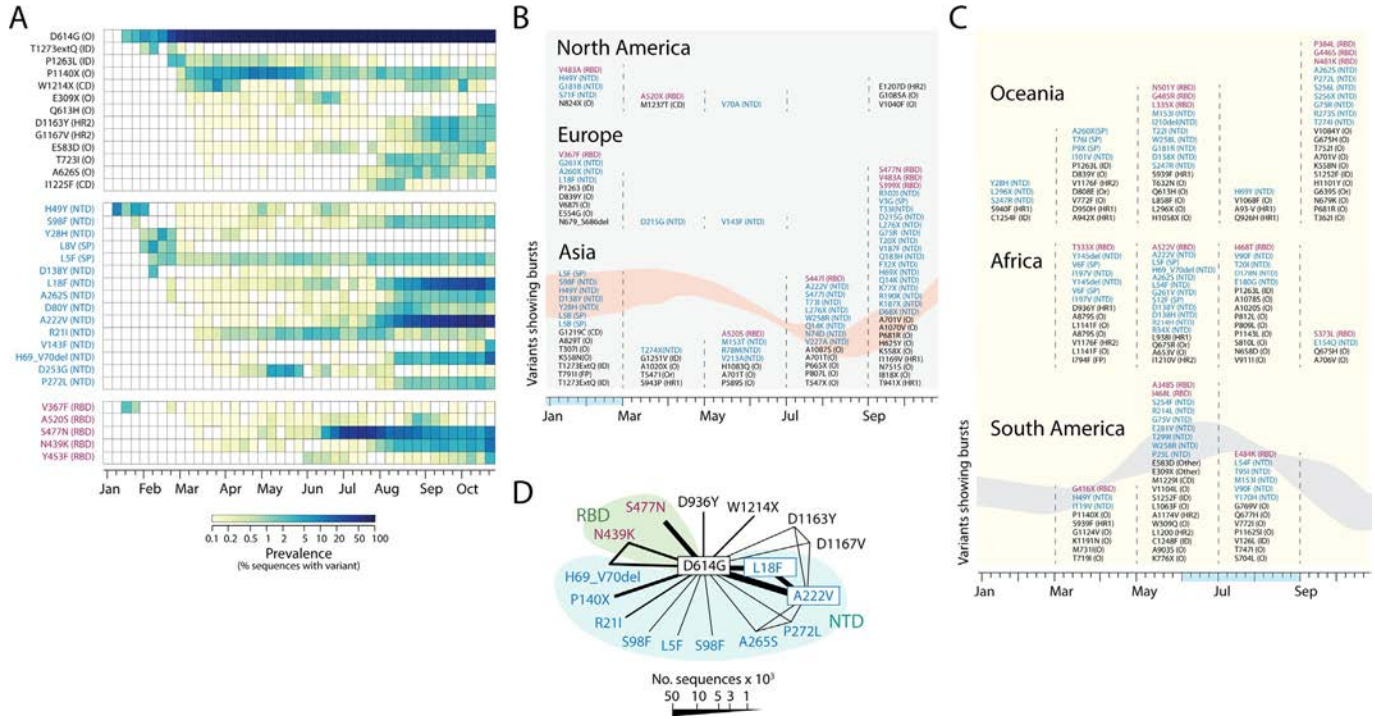

FIG. 8

A timeline of S-protein variants of the spike during the emergence of the COVID-19 pandemic. (A) Heatmap describing the gradual weekly accumulation of variants grouped by their appearance in either the N-terminal domain (NTD), the receptor-binding domain (RBD), or other locations of the protein molecule (see diagram of domain locations in the S1 and S2 subunits above). Only variants that were present in at least 2% in sequenced genomes for at least a week were included and ordered by time of significant appearance. Major domains holding the mutations are given in parentheses following the name of the variant; SP, signal peptide; NTD and RDB, N-terminal and receptor-binding domains; FP, fusion peptide; HR1, heptad repeat 1; HR2, heptad repeat 2; CD, cytoplasmic domain; ID, intracellular domain; and O, other. Prevalence levels in the heatmap at given in a log scale. (B) and (C) Stacked bar plots constructed with names of variants associated with bursts of weakly variant prevalence initiated during 2-month periods. Only variants appearing in at least 2% of sequences in any given continent are considered and coloured according to regions. Bursts were defined as variants appearing with prevalence higher than 0.3% in one or multiple weeks, which then withered throughout the timeline. (D) A network of co-occurrence of variants in sequences describes the most abundant variant combinations appearing in the timeline. Nodes are variants and links represent their co-occurrence in at least 500 sequences out of the 137,605 analysed. All data was retrieved from supplementary tables in Showers et al. (2021).

ARTICLE IN PRESS

[NTD plus signal peptide (SP)], the RBD region, and in 'Other' regions, including the S2 subdomain of the S-protein. The time series showed the early and growing prevalence of the D614G variant (90% during the week of June 19) but also of the NTD L18F and A222V substitutions, which spread in July and early August, respectively. Most RBD substitutions peaked later. For example, S477N peaked in July (with 44% prevalence worldwide and massively present in Oceania) but quickly decreased in incidence. RBD substitution N439K and NTD deletion H69_V70del were significant but were clearly late introductions.

More importantly, the timeline revealed a cyclic pattern of prevalence for the N-terminal region, but not so much for RDB or 'Other' regions. With the exception of D614G and SP variant L5F, which progressed towards fixation, all mutants introduced during the January and February winter months of the Northern Hemisphere showed significant bursts of genomic abundance lasting one or a few weeks, which then declined significantly in later months with some resurfacing by the end of 2020. These included the first 6 mutations in the N-terminal SP and NTD regions (H49Y, S98F, Y28H, L8V and L5F), two mutations in the intracellular domain (ID) (T1273extQ and P1263L), one in the cytoplasmic domain (CD) (W1214X) and one in the RDB region (V367F). ID and RDB mutations later vanished as the year progressed. With only one exception (D253G of the NTD), mutations appearing between the months of March and July had low prevalence and accumulated without significant bursts in the timeline. No significant late mutation entries appeared after July as the Northern Hemisphere was entering summer, again suggesting a seasonal component.

Remarkably, dissecting the timeline along geography uncovered bursts occurring preferably in the N-terminal SP and NTD regions that increased during the winter seasons of the Northern Hemisphere for Europe, North America and Asia (Fig. 8B), and the winter season of the Southern Hemisphere for Oceania, South America and in part Africa (Fig. 8C). This expected seasonal pattern, which matches patterns shown in Fig. 4E, provides strong support to assumptions and hypotheses. The hemisphere-related sinusoidal pattern was particularly evident in Asia and South America. In Asia, a significant number of bursts in January-March driven by the start of the pandemic in China and involving either the NTD or 'Other' regions of the S-protein was followed by marked decreases and then significant increases in activity towards the end of the year with massive numbers of bursts involving the NTD region. In South America, the pattern reverses with most bursts occurring during the months of May and June, fewer bursts observed in July and August, and no burst observed in September and October, with patterns mostly driven by variants in the 'N-terminal' and 'Other' categories. A similar behaviour was observed for Africa and Oceania but with one difference for the number of bursts during September and October for Oceania, which was maximal for that set. We note that the absence of entries during the months of January and February for South America and Africa may be related to the low numbers of sequenced genomes available for those regions during that time. We also note that all of North America and Europe and most of Asia (with exception of East Timor, Indonesia, and Maldives) is located on the Northern

Hemisphere and that most of South America (90%), all of Oceania and only 30% of Africa is located in the Southern Hemisphere, which holds only 15% of the world population. For Africa, only the Mediterranean countries and South Africa are within the seasonal corridor of the disease. Despite possible geographical uncertainties, the hemisphere-related patterns appear consistent.

How would bursts exert their influence on seasonality? Quantifying variants present in a sequence provides insight into the kind of changes occurring in an individual burst, which depend on the functional needs of proteins at that time. We therefore explored the combinatorial landscape of variants arising during the 2020 timeline. Out of 137,605 variant sequences analysed (Showers et al., 2021), 9.8% contained only one variant (D614G), while 44.7% contained 2, 30.8% contained 3, and 14.4% contained 4–9. The average of 2.53 ± 0.94 (SE) variants per sequence shows the poor mutational saturation of the sampled viral population as a whole. A network of variants with links describing their co-occurrence in sequences revealed that the vast majority of sequences had variants that co-occurred with D614G. Thus, D614G represents the central hub of the network. In fact, only 43 variant combinations representing 128 sequences lacked D614G variants (3 of which however co-occurred with other variants of the 614 site). This is surprising and suggests D614G plays a very central functional role, possibly of flexibility, that is complemented by each new variant. A network of co-occurrence of variants in sequences with variant combinations present in at least 500 sequences is shown in Fig. 8D. Remarkably, the network reveals that variants located in the NTD region, especially A222V and L18F, were driving network co-occurrence in 2020. This again supports the central role of NTD as environmental sensor. Note that 7,088 sequences contained all three D614G, A222V and L18F variants, the top most common 3-variant combination. As expected, many of the central network players we mapped earlier onto the atomic model of the S-protein (Fig. 7) were also central hubs in this highly represented network.

4.5 Genomic change levels appear unlinked to temperature or latitude

When analysing the effects of temperature and latitude on epidemiological parameters we also explored if genomic change levels for individual countries correlated with temperature and latitude (Burra et al., 2021). We filtered genomic sequences and collected 55,455 sequences from 211 countries on August 5, 2020. Variant accumulation and mutation rates were calculated for the entire genome and for specific regions known for significant pathways of change, including the S1 subunit, the S2 subunit and the RBD region of the S-protein, ORF pp1a, and the NSP2 protease regulator. Pearson correlation analysis showed no significant associations. The only weak borderline correlations were observed when comparing total genomic change with temperature ($r = -0.09$) and latitude ($r = 0.09$), both of which had highly significant P values of $\sim 4e-100$. There were several possible reasons for lack of significant association. Perhaps cyclic patterns observed in Fig. 4 were not fully manifested at the time of data collection or the confounding hemisphere-related temporal

effects were not appropriately dissected in the genomic set. Alternatively, and given the results of Fig. 8, variant bursts rather than variant levels could harbour better signatures of seasonal behaviour. It is unclear whether the evolution of the SARS-CoV-2 quasispecies is driven by mutational load and not specific molecular sensors, as we here propose. Further exploration is therefore warranted.

4.6 Galectin homologues appear likely molecular culprits

While adaptive immunity is the central line of defense of vertebrates, the evolutionary older innate immunity response is the first step taken against an invading pathogen (Janeway, Travers, Walport, & Shlomchik, 2001). This strategy recruits immune cells to sites of infection through production of chemical factors (e.g. cytokine cascades), activates proinflammatory reactions and clearance of antibody complexes, removes foreign structures through specialized white cells, triggers the adaptive immune system through the process of antigen presentation, and acts as a physical and/or chemical barrier to the infectious agent. Innate immunity does not respond to antigens. Instead, it recognizes pathogen-associated and damage-associated molecular patterns (PAMPs and DAMPs). PAMPs are small evolutionarily-conserved molecular motifs of microbes (e.g. LPSs, endotoxins) that activate innate immune responses when recognized by toll-like or other pattern recognition receptors. DAMPs include heat shock proteins, interleukin-1α, defensins and anexins. Lectins are one important class of pattern recognition receptors that bind to carbohydrate moieties of glycoproteins or glycolipids. They often have dual roles as PAMPs and DAMPs (Sato, St-Pierre, Bhaumik, & Nieminen, 2009). In particular, the *galectin* family of lectins are highly evolutionarily-conserved effectors that mediate a multiplicity of biological processes, making them molecular targets for therapeutic intervention (Dings, Miller, Griffin, & Mayo, 2018). Central roles include control of cell-cell and cell-matrix interactions, adhesion, proliferation, apoptosis, pre-mRNA splicing, immunity and inflammation. They are also involved in multiple processes of cancer initiation and development, including apoptosis, adhesion and migration, cell transformation, invasion and metastasis, immune escape and angiogenesis (Dings et al., 2018). Galectin functions are multifaceted but generally involve binding to carbohydrate moieties of glycoconjugates on the surfaces of cells. Galectins are defined by conserved peptide sequence elements in one or more well-defined carbohydrate recognition domain (CRD) and are grouped into prototype, chimera, and tandem repeat groups depending on having a single CRD core, a CRD core with collagen-like N-terminal tail (galectin-3), or a tandem repeat of two CRDs, respectively. Note that galectins can be monomeric or dimeric. Structurally, CRDs are about 130 amino acids long and have β-sandwich structures composed of 11 β-strands in antiparallel arrangement. Six of these β-strands (β1, β3, β4, β5, β6, and β10), define the sugar-binding face of the fold structure.

Remarkably, a DALI analysis of the structural neighbourhood of the SARS-CoV-2 S-protein of the viral spike using a representative crystallographic structural entry as query revealed a significant structural match of the N-terminal NTD domain to the relatively small galectin CRD structures (Fig. 9). The DALI server is an

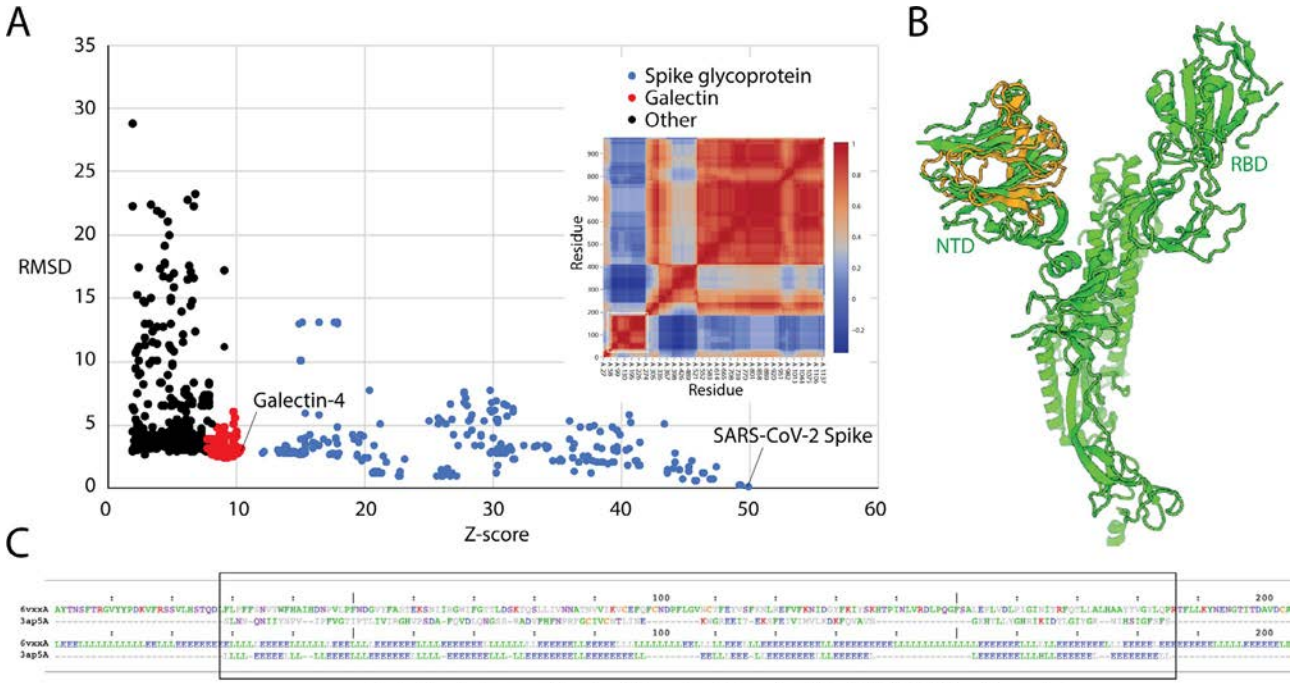

FIG. 9

See figure legend on next page.

ARTICLE IN PRESS

established fully automated nonhierarchical alignment tool for structure comparison that uses a distance matrix alignment to construct lists of structural neighbours (Holm, 2020). DALI provides structural alignments as either three-dimensional or two-dimensional comparisons by explicitly rotating and translating one domain structure over another or by mapping molecular structure into a matrix of intramolecular distances, respectively. Since structure is far more conserved than sequence, structural similarities [as root-mean square deviation (RMSD) scores] can better dissect deeper homology relationships than sequences, especially when these are established between domain regions of different sizes (using Z-scores) (see Glossary for technical definitions). Note that better structural matches get lower RMSD values and larger structural alignments get larger Z-scores. As expected, the RMSD vs

FIG. 9 Galectins, likely culprits of COVID-19 seasonal behaviour. (A) The DALI structural neighbourhood of the spike protein of SARS-CoV-2 (PDB entry *6vxxA*) contains 2310 structures (downloaded June 3, 2020) spread through a scatter plot describing their RMSD and Z-score values (see Glossary for definitions). RMSD describes rigid-body structural superposition while the Z-score describes a length-rescaled distance matrix alignment that maximizes one-to-one atomic correspondences between two structures with a weighted sum of similarities of intramolecular distances. The plot reveals galectins (red circles) exhibit a good structural match (RMSD ~ 2.5–5 Å) to a small segment (Z-score ~ 8–10) of the spike protein query (which self-aligns with RMSD = 0 and Z-score = 50). The inset shows the match correlation matrix, which describes the correlation of matched structures along amino acid residue positions of the query structure (proceeding from N-terminus to the C-terminus in the x axis). The matrix can be used to evaluate modular segments in the alignment of the entire neighbourhood. Structural modular domains are expected to show strong positive correlation (red hues) within domains and negative correlations (blue hues) between domains. The highlighted white square at the N-terminus of the molecule in the left-bottom part of the correlation matrix reveals that the region of the N-terminal domain (NTD) of domain S1 of the spike, which is homologous to the galectin fold (see panel B) is indeed modular. The matrix also shows significant structural modularity of the receptor binding domain (RDB) and the S2 subunit of the spike molecule. (B) Structural alignment of the monomer of galectin-4 (brown; PDB entry *3ap5A*) to the N-terminal domain of the coronavirus protein (green; PDB entry *6vxxA*) shows there is a good structural match in their three-dimensional atomic models (rendered in ribbon format), confirming structural similarities (putative homologies) between the NTD and the galectin fold. Alignments with other galectins show similar structural matches (not shown). (C) A protein sequence and secondary structure alignment details the structural match at the N-terminal region of the spike molecule. Symbols of amino acids in the pairwise sequence alignment between the query SARS-CoV-2 spike protein and galectin-4 at the top are given in the one-letter amino acid notation of the IUPAC-IUB Commission of Biochemical Nomenclature. Symbols in the pairwise alignment of secondary structures at the bottom follow DSSP nomenclature: H/h, helix; E/e, strand; L/l, coil. Uppercase means structurally equivalent positions with the query sequence.

Z-score scatter plot detected a multiplicity of structural clusters (blue circles in the plot), most of them related to spike proteins of other RNA viruses (Fig. 9A). While these spike structures showed good structural matches (RMSD ~ 2–5 Å), alignments of protein segments of decreasing length (Z ~ 10–50) showcased the remarkable structural diversity that exist in the world of spike proteins. The novelty is the existence of a cluster of small structures with RMSD ~ 2.5–5 Å and Z ~ 8–10 (red circles) showing a small segment located on the N-terminal region of the SARS-CoV-2 spike protein (see correlation matrix in Fig. 9A) matched known crystallographic structures of galectin proteins. Fig. 9B and C show structural and secondary structure alignments between galectin-4 and the spike protein we used as query. Note alignment of the extended strands (E) that make the sequential loops of β-strands.

But why are galectin-like structures relevant to the seasonal behaviour of COVID-19? The answer can be found in the coral reefs of the Pacific and Indian Oceans and the devastating phenomenon of coral bleaching that threatens marine resources and the well-being of the millions that live in coastal communities. Scleractinian corals establish an endosymbiosis with dinoflagellates of the genus *Symbiodinium* that allows them to grow and persist in warm, well-illuminated and oligotrophic waters. Coral bleaching has been attributed to symbiosis disruption under environmental stress resulting from rising ocean water temperatures and other factors of global change (e.g. Kleypas, Castruccio, Curchitser, & McLeod, 2015). A number of lectins have been identified in corals as part of the innate immune response system of coral communities (Kvennefors et al., 2010), some implicated in thermal and disease stress responses (Ricci et al., 2019; Vidal-Dupiol et al., 2014). In particular, galectin proteins with antimicrobial immunity functions were found in greater abundance in healthy coral symbioses (Ricci et al., 2019). Remarkably, a recent study showed galectin functions of the scleractinian coral *Pocillopora damicornis* were dependent on temperature effects, establishing galectins as possible environmental sensors (Wu et al., 2019). A cloned galectin that encoded a 293 amino acid-long protein harbouring a CRD and collagen-like N-terminal tail with signal peptide was able to recognize and agglutinate or stabilize coral pathogens and symbionts. Binding activities were only optimal between 25–30 °C. Results suggested galectins may be involved in heat bleaching of scleractinian corals through temperature-regulated recognition of pathogens and symbionts.

We end by noting that lectins are a diverse family of proteins with the capacity to selectively recognize carbohydrate moieties associated for example with receptor proteins. It appears that the SARS-CoV-2 S-protein, besides binding ACE2, also interacts with the CRD of C-type lectin receptors, which facilitate viral entry by dysregulating host immune responses and acting as entry receptors (Rahimi, 2021). A number of studies have recently surfaced that point into the centrality of lectin-mediated recognition (Gao et al., 2021; Soh et al., 2021; Thépaut et al., 2021). Curiously, these lectins interact with both ACE2 and with the NTD domain of the S-protein (Soh et al., 2021), perhaps facilitating the process of receptor presentation by trans-infection (Thépaut et al., 2021).

5 Conclusions and prospects

The relationship between weather and diseases has been a major concern for centuries. Understanding the reasons and conditions under which pathogens evolve to spread and achieve higher infection rates is considered a major challenge in epidemiology. This understanding is necessary for prediction, mitigation and informed public health decision making. Here we propose a triangle of viral persistence will help identify seasonal and other drivers of the wax and waning of viral diseases. Timing of winter-peaking diseases, but not their amplitude, scales with latitude when analysing weather and health insurance claim databases across U.S. counties (Schober, Rzhetsky, & Rust, 2021). This latitude scaling suggests seasonality across many respiratory pathogens in temperate zones responds to annual changes in irradiance and possibly a human circannual clock. Our discussions have shown this prediction applies to the SARS-CoV-2 disease and is corroborated by a growing number of studies. The survival of the COVID-19 virus depends on the environment. The capsid is surrounded by a lipid bilayer, which supports structural proteins that are expected to be sensitive to environmental factors such as temperature, humidity and solar radiation (Moriyama et al., 2020). These environmental factors affect the ability of the virus to survive in droplets and surfaces and may thereby reduce its transmission rate.

We also propose that molecular components of the virus can act as sensors of environment and physiology, and could represent molecular culprits of seasonality. Through analysis of variants of the SARS-CoV-2 proteome encoded in thousands of genomes, we searched for possible structures capable of being modulated by the environment. The search resulted in the identification of a galectin-like structure within the N-terminal domain of the S-protein. This galectin-like structure may be a frequent target of mutations because it helps the virus evade or modulate physiological responses of the host to further its spread and survival. Regions such as these appear to be good molecular candidates for sensors of environment and physiology. They are rigid enough to sustain environmental damage and flexible enough to enable recognition and signaling. Tracking mutations in these molecular regions may provide key insights into predictable changes that occur in seasonal viral transmission patterns as well as viral mutations throughout the course of an epidemic.

In short, while it has been observed that seasonal changes have a significant impact on the transmissibility and infectivity of SARS-CoV-2, the exact mechanisms through which this phenomenon occurs have not yet been elucidated until now. The study of SARS-CoV-2 virus seasonal behaviour and host responses that are associated with infection is crucial, since this knowledge can help mitigation efforts. Results of our analyses show that the SARS-CoV-2 virus pushes prevalence of mutations in molecular 'sensor' regions in a seasonal manner in search for better modulation of sensing functions. Although the data analysed and presented here show clear seasonality of SARS-CoV-2, there is still much to learn about the causal factors of seasonal periodicities. This necessitates both the application and development of statistical and data mining methodologies that can dissect causative from aleatory effects. Systematic

characterization of the seasonal patterns of infectious disease and transmission patterns in general can help inform us on cultural, socioeconomic, and life style factors that affect our susceptibility to disease, as well as public policy decisions on things such as school openings and closings during periods of peak virus transmission.

Acknowledgements

We dedicate this work to the frontline medical professionals who have been saving the life of others with limited protective equipment, selflessly, and at their own peril during an age of misinformation. COVID-19 research in the laboratory of G.C.-A. is supported by the Office of Research and Office of International Programs in the College of Agricultural, Consumer and Environmental Sciences at the University of Illinois at Urbana-Champaign.

Financial disclosures and conflicts of interest

The authors have no financial disclosures or conflicts of interests to disclose.

References

Abotaleb, M., & Makarovskikh, T. (2021). System for forecasting COVID-19 cases using time-series and neural networks models. *Engineering Proceedings*, *5*, 46.

Akhila, M. V., Narwani, T. J., Floch, A., Maljković, M., Bisoo, S., Shinada, N. K., et al. (2020). A structural entropy index to analyze conformations in intrinsically disordered proteins. *Journal of Structural Biology*, *210*, 107464.

Amman, B. R., Carroll, S. A., Reed, Z. D., Sealy, T. K., Balinandi, S., Swanepoel, R., et al. (2012). Seasonal pulses of Marburg virus circulation in juvenile Rousettus aegyptiacus bats coincide with periods of increased risk of human infection. *PLoS Pathogens*, *8*(10), e1002877.

Ashour, H. M., Elkhatib, W. F., Rahman, M., & Elshabrawy, H. A. (2020). Insights into the recent 2019 novel coronavirus (SARS-CoV-2) in light of past human coronavirus outbreaks. *Pathogens*, *35*(3), 235–241.

Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological Reviews*, *35*(3), 235–241.

Barría-Sandoval, C., Ferreira, G., Benz-Parra, K., & López-Flores, P. (2021). Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques: A comparative study. *PLoS One*, *16*(4), e0245414.

Batschelet, E., Domingo, E., & Weissmann, C. (1976). The proportion of revertant and mutant phage in a growing population, as a function of mutation and growth rate. *Gene*, *1*(1), 27–32.

Becerra-Flores, M., & Cardozo, T. (2020). SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *International Journal of Clinical Practice*, *74*, e13525.

Bernasconi, A., Mari, L., Casagrandi, R., & Ceri, S. (2021). Data-driven analysis of amino acid change dynamics reveals SARS-CoV-2 variant emergence. *Scientific Reports*, *11*, 21068.

Berngruber, T. W., Froissart, R., Choisy, M., & Gandon, S. (2013). Evolution of virulence in emerging epidemics. *PLoS Pathogens*, *9*(3), e1003209.

Bjørnstad, O. N., & Viboud, C. (2016). Timing and periodicity of influenza epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(46), 12899–12901.

Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M., & Petropoulos, C. J. (2004). Evidence for positive epistasis in HIV-1. *Science*, *306*(5701), 1547–1550.

Borrmann, H., McKeating, J. A., & Zhuang, X. (2021). The circadian clock and viral infections. *Journal of Biological Rhythms*, *36*, 9–22.

Brownlee, J. (1918). An investigation into the periodicity of measles epidemics in London from 1703 to the present day by the method of the periodogram. *Philosophical Transactions of the Royal Society London B*, *208*, 225–250.

Burch, C. L., & Chao, L. (2004). Epistasis and its relationship to canalization in the RNA virus φ6. *Genetics*, *167*(2), 559–567.

Burra, P., Soto-Díaz, K., Chalen, I., Gonzalez-Ricon, R. J., Istanto, D., & Caetano-Anollés, G. (2021). Temperature and latitude correlate with SARS-CoV-2 epidemiological variables but not with genomic change worldwide. *Evolutionary Bioinformatics*, *19*, 1176934321989695.

Caetano-Anollés, G. (2021). The language of biomolecular communication. In G. Caetano-Anollés (Ed.), *Untangling molecular biodiversity* (pp. 283–345). Singapore: World Scientific Publishing.

Caetano-Anollés, D., & Caetano-Anollés, G. (2015). Computing the origin and evolution of the ribosome from its structure—Uncovering processes of macromolecular accretion benefitting synthetic biology. *Computational and Structural Biotechnology Journal*, *13*, 427–447.

Caetano-Anollés, G., Mughal, F., Aziz, M. F., Koç, I., Caetano-Anollés, K., Caetano-Anollés, D., et al. (2021). A 'double tale' of module creation in evolving networks. In G. Caetano-Anollés (Ed.), *Untangling molecular biodiversity* (pp. 91–168). Singapore: World Scientific Publishing.

Cannell, J. J., Vieth, R., Umhau, J. C., Holick, M. F., Gran, W. B., Madronich, S., et al. (2006). Epidemic influenza and vitamin D. *Epidemiology and Infection*, *134*, 1129–1140.

Carlson, C. J., Gomez, A. C., Bansal, S., & Ryan, S. J. (2020). Misconceptions about weather and seasonality must not misguide COVID-19 response. *Nature Communications*, *11*(1), 1–4.

Chambliss, D. F., & Schutt, R. K. (2012). *Making sense of the social world* (4th ed.). London: SAGE Publications.

Chin, A. W. H., Chu, J. T. S., Perera, M. R. A., Hui, K. P. Y., Yen, H.-L., Chan, M. C. W., et al. (2020). Stability of SARS-CoV-2 in different environmental conditions. *The Lancet Microbe*, *1*, e10.

Colominas, M. A., Schlotthauer, G., & Torres, M. E. (2014). Improved complete ensemble EMD: A suitable tool for biomedical signal processing. *Biomedical Signal Process and Control*, *14*, 19–29.

Conlan, A. J. K., & Grenfell, B. T. (2007). Seasonality and the persistence and invasion of measles. *Proceedings of the Royal Society B: Biological Science*, *274*, 1133–1141.

Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. (2020). The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology*, *5*, 536–544.

Cox, N. J., & Fukuda, K. (1998). Influenza. *Infectious Disease Clinics of North America*, *12*, 27–38.

Demongeot, J., Flet-Berliac, Y., & Seligmann, H. (2020). Temperature decreases spread parameters of the new Covid-19 case dynamics. *Biology (Basel)*, *9*, 94.

Deyle, E. R., Maher, M. C., Hernandez, R. D., Basu, S., & Sugihara, G. (2016). Global environmental drivers of influenza. *Proceedings of the National Academy of Science of the United States of America*, *113*, 13081–13086.

Dings, R. P. M., Miller, M. C., Griffin, R. J., & Mayo, K. H. (2018). Galectins as molecular targets for therapeutic intervention. *International Journal of Molecular Sciences*, *19*, 905.

Domingo, E. (2002). Quasispecies theory in virology. *Journal of Virology*, *76*(1), 463–465.

Domingo, E., Martin, V., Perales, C., Grande-Perez, A., Garcia-Arriaza, J., & Arias, A. (2006). Viruses as quasispecies: Biological implications. *Current Topics in Microbiology and Immunology*, *299*, 51–82.

Domingo, E., Sheldon, J., & Perales, C. (2012). Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews*, *76*, 159–216.

Domingo, E., Ruiz-Jarabo, C. M., Sierra, S., Arias, A., Pariente, N., Baranowski, E., et al. (2002). Emergence and selection of RNA virus variants: Memory and extinction. *Virus Research*, *82*(1–2), 39–44.

Dopico, X. C., Evangelou, M., Ferreira, R. C., Guo, H., Pekalski, M. L., Smyth, D. J., et al. (2015). Widespread seasonal gene expression reveals annual differences in human immunity and physiology. *Nature Communications*, *6*, 7000.

Dowell, J. (2001). Seasonal variation in host susceptibility and cycles of certain infectious diseases. *Emerging Infectious Diseases*, *7*(3), 369–374.

Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, *148*(4), 1667–1686.

Dudas, G., Carvalho, L. M., Rambaut, A., & Bedford, T. (2018). MERS-CoV spillover at the camel-human interface. *eLife*, *7*(e31257), 1–37.

Duncan, C. J., Duncan, S. R., & Scott, S. (1997). The dynamics of measles epidemics. *Theoretical Population Biology*, *2*, 155–163.

Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradović, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, *41*, 6573–6582.

Earn, D. J., Dushoff, J., & Levin, S. A. (2002). Ecology and evolution of the flu. *Trends in Ecology and Evolution*, *17*(7), 334–340.

Eckerle, L. D., Becker, M. M., Halpin, R. A., Li, K., Venter, E., Lu, X., et al. (2010). Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathogens*, *6*(5), e1000896.

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

Farah-Pérez, A., Umaña-Villalobos, G., Picado-Barboza, J., & Anderson, E. P. (2020). An analysis of river fragmentation by dams and river dewatering in Costa Rica. *River Research and Applications*, *36*, 1442–1448.

Ferguson, N. M., Cucunubá, Z. M., Dorigatti, I., Nedjati-Gilani, G. L., Donnelly, C. A., Basáñez, M. G., et al. (2016). Countering the Zika epidemic in Latin America. *Science*, *353*(6297), 353–354.

Fine, P. E. M., & Clarkson, J. A. (1982). Measles in Engand and Wales—I: An analysis of factors underlying seasonal patterns. *International Journal of Epidemiology*, *11*, 5–14.

Finke, L. L. (1792-1795). *Versuch einer allgemeinen medicinisch -praktischen Geographie*. *3 Vols*. Leipzig: Weidmannschen Buchhandlung.

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., et al. (2020). The coding capacity of SARS-CoV-2. *Nature*, *589*, 125–130.

Fisman, D. (2012). Seasonality of viral infections: Mechanisms and unknowns. *Clinical Microbiology and Infection*, *18*, 946–954.

Friendly, M., Monette, G., & Fox, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. *Statistical Science*, *28*(1), 1–39.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, *15*, 246–263.

Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., et al. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*, *368*, 779–782.

Gao, C., Zeng, J., Jia, N., Stavenhagen, K., Matsumoto, Y., Zhang, H., et al. (2021). SARS-CoV-2 spike protein interacts with multiple innate immune receptors. *bioRxiv*. https://doi.org/10.1101/2020.07.29.227462.

Garland, T., Jr. (2014). Trade-offs. *Current Biology*, *24*(2), PR60–PR61.

Gecili, E., Ziady, A., & Szczesniak, R. D. (2021). Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. *PLoS One*, *16*(1), e0244173.

George, D. B., Webb, C. T., Farnsworth, M. L., O'Shea, T. J., Bowen, R. A., Smith, D. L., et al. (2011). Host and viral ecology determine bat rabies seasonality and maintenance. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(25), 10208–10213.

Gibbs, M. J., & Weiller, G. F. (1999). Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(14), 8022–8027.

Goh, G. K.-M., Dunker, A. K., Foster, J. A., & Uversky, V. N. (2020a). Rigidity of the outer shell predicted by a protein intrinsic disorder model sheds light on the COVID-19 (Wuhan-2019-nCoV) infectivity. *Biomolecules*, *10*, 331.

Goh, G. K.-M., Dunker, A. K., Foster, J. A., & Uversky, V. N. (2020b). Shell disorder analysis predicts greater resilience of the SARS-CoV-2 (COVID-19) outside the body and in body fluids. *Microbial Pathogenesis*, *144*, 104177.

Goh, G. K.-M., Dunker, A. K., Foster, J. A., & Uversky, V. N. (2021). Shell disorder analysis suggests that pangolins offered a window for a silent spread of an attenuated SARS-CoV-2 precursor among humans. *Journal of Proteome Research*, *19*, 4543–4552.

Grenfell, B., & Bjørnstad, O. (2005). Epidemic cycling and immunity. *Nature*, *433*(7024), 366–367.

Groseth, A., Feldmann, H., & Strong, J. E. (2007). The ecology of Ebola virus. *Trends in Microbiology*, *15*(9), 408–416.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextrain: Real-time tracking of pathogen evolution. *Bioinformatics*, *34*(23), 4121–4123.

Hardle, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press.

Hardle, W., Lütkenpohl, H., & Cheng, R. (1997). A review of nonparametric time series analysis. *International Statistical Review*, *65*(1), 49–72.

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews. Microbiology*, *19*, 409–424.

Heilingloh, S. S., Aufderhorst, U. W., Schipper, L., Dittmer, U., Witzke, O., Yang, D., et al. (2020). Susceptibility of SARS-CoV-2 to UV radiation. *American Journal of Infection Control*, *48*(10), 1273–1275.

Hethcote, H. W., Stech, H. W., & Van Den Driessche, P. (1981). Nonlinear oscillations in epidemic models. *SIAM Journal on Applied Mathematics*, *40*(1), 1–9.

Hirsch, A. (1883). Handbook of geographical and historical pathology. *Acute infective diseases: Vol. I*. London: The New Sydenham Society.

Holm, L. (2020). DALI and the persistence of protein shape. *Protein Science*, *29*, 128–140.

Hope-Simpson, R. E. (1981). The role of season in the epidemiology of influenza. *Journal of Hygiene (London)*, *86*(1), 35–47.

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

Janeway, C., Travers, P., Walport, M., & Shlomchik, M. (2001). *Immunobiology* (5th ed.). New York: Garland Science.

Johnson, T., & Barton, N. H. (2002). The effect of deleterious alleles on adaptation in asexual populations. *Genetics*, *162*(1), 395–411.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., et al. (2008). Global trends in emerging infectious diseases. *Nature*, *451*, 990–993.

Jonsson, C. B., Moraes Figuereido, L. T., & Vapalahti, O. (2010). A global perspective on Hantavirus ecology, epidemiology, and disease. *Clinical Microbiology Reviews*, *23*(2), 412–441.

Karapiperis, C., Kouklis, P., Papastratos, S., Chasapi, A., Danchin, A., Angelis, L., et al. (2021). A strong seasonality pattern for COVID-19 incidence rates modulated by UV radiation levels. *Viruses*, *13*, 574.

Kasahara, A. K., Singh, R. J., & Noymer, A. (2013). Vitamin D (25OHD) serum seasonality in the United States. *PloS One*, *8*, e65785.

Kassem, A. Z. E. (2020). Does temperature affect COVID-19 transmission? *Frontiers in Public Health*, *8*, 554964.

Khatchikian, D., Orlich, M., & Rott, R. (1989). Increased viral pathogenicity after insertion of a 28S ribosomal RNA sequence into the haemagglutinin gene of an influenza virus. *Nature*, *340*(6229), 156–157.

Killerby, M. E., Biggs, H. M., Haynes, A., Dahl, R. M., Mustaquim, D., Gerber, S. I., et al. (2018). Human coronavirus circulation in the United States 2014-2017. *Journal of Clinical Virology*, *101*, 52–56.

Kirchdoerfer, R. N., & Ward, A. B. (2019). Structure of the SARS-CoV NSP12 polymerase bound to NSP7 and NSP8 co-factors. *Nature Communications*, *10*, 2342.

Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., & Lipsitch, M. (2020). Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*, *368*(6493), 860–868.

Kleypas, J. A., Castruccio, F. S., Curchitser, E. N., & McLeod, E. (2015). The impact of ENSO on coral heat stress in the western equatorial Pacific. *Global Change Biology*, *21*, 2525–2539.

Klingberg, E., Olerod, G., Konar, J., Petzold, M., & Hammarsten, O. (2015). Seasonal variations in serum 25-hydroxy vitamin D levels in a Swedish cohort. *Endocrine*, *49*, 800–808.

Knight, G. (1971). The ecology of African sleeping sickness. *Annals of the Association of American Geographers*, *61*(1), 23–44.

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, *182*, 812–827.

Kronfeld-Schor, N., Stevenson, T. J., Nickbakhsh, S., Schernhammer, E. S., Dopico, X. C., Dayan, T., et al. (2021). Driversof infectious disease seasonality: Potential implications for COVID-19. *Journal of Biological Rythms*, *36*(1), 35–54.

Kvennefors, E. C., Leggat, W., Kerr, C. C., Ainsworth, T. D., Hoegh-Guldberg, O., & Barnes, A. C. (2010). Analysis of evolutionarily conserved innate immune components in coral links immunity and symbiosis. *Developmental and Comparative Immunology*, *34*, 1219–1229.

Lauber, C., Goeman, J. J., Parquet, M. C., Thi Nga, P., Snijder, E. J., Morita, K., et al. (2013). The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathogens*, *9*(7), e1003500.

Lauring, A. S., & Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. *PLoS Pathogens*, *6*(7), e1001005.

Li, R., Metcalf, J. E., Stenseth, N. C., & Bjørnstad, O. N. (2021). A general model for the demographic signatures of the transition from pandemic emergence to endemicity. *Science Advances*, *7*, eabf9040.

Li, Y., Wang, X., & Nair, H. (2020). Global seasonality of human seasonal coronaviruses: A clue for postpandemic circulating season of severe acute respiratory syndrome coronavirus 2? *Journal of Infectious Diseases*, *222*, 1090–1097.

Liu, X., Huang, J., Li, C., Zhao, Y., Wang, D., Huang, Z., et al. (2021). The role of seasonality in the spread of COVID-19 pandemic. *Environmental Research*, *195*, 110874.

Lowen, A. C., Mubareka, S., Steel, J., & Palese, P. (2007). Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathogens*, *3*(10), 1470–1476.

Majumdar, P., & Niyogi, S. (2021). SARS-CoV-2 mutations: The biological trackway towards viral fitness. *Epidemiology and Infection*, *149*, e110.

Malim, M. H., & Emerman, M. (2001). HIV-1 sequence variation: Drift, shift, and attenuation. *Cell*, *104*(4), 469–472.

Manrubia, S. C., & Lázaro, E. (2006). Viral evolution. *Physics of Life Reviews*, *3*(2), 65–92.

Mantilla-Beniers, N. B., Bjørnstad, O. N., Grenfell, B. T., & Rohani, P. (2010). Decreasing stochasticity through enhanced seasonality in measles epidemics. *Journal of the Royal Society Interface*, *7*(46), 727–739.

Martin, D. P., Weaver, S., Tegally, H., San, E. J., Shank, S. D., Wilkinson, E., et al. (2021). The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. *medRxiv*. https://doi.org/10.1101/2021.02.23.21252268.

Martinez, M. E. (2018). The calendar of epidemics: Seasonal cycles of infectious diseases. *PLoS Pathogens*, *14*(11), e1007327.

Matson, M. J., Yinda, C. K., Seifert, S. N., Bushmaker, T., Fischer, R. J., van Doremalen, N., et al. (2020). Effect of environmental conditions on SARS- CoV-2 stability in human nasal mucus and sputum. *Emergent Infectious Diseases*, *26*, 2276–2278.

Mecenas, P., Bastos, R. T. R. M., Vallinoto, A. C. R., & Normando, D. (2020). Effects of temperature and humidity on the spread of COVID-19: A systematic review. *PLoS One*, *15*(9), e0238339.

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and genomic distribution of SARS-CoV-2 mutations. *Frontiers in Microbiology*, *11*, 1800.

Midgley, C. M., Haynes, A. K., Baumgardner, J. L., Chommanard, C., Demas, S. W., Prill, M. M., et al. (2017). Determining the seasonality of respiratory syncytial virus in the United States: The impact of increased molecular testing. *Journal of Infectious Diseases*, *216*, 345–355.

ARTICLE IN PRESS

Minhaz Ud-Dean, S. M. (2010). Structural explanation for the effect of humidity on persistence of airborne virus: Seasonality of influenza. *Journal of Theoretical Biology*, *263*(3), 822–829.

Minskaia, E., Hertzig, T., Gorbalenya, A. E., Campanacci, V., Cambillau, C., Canard, B., et al. (2006). Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 5108–5113.

Montgomery, D. (2013). *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons, Inc.

Monto, A. S., DeJonge, P., Callear, A. P., Bazzi, L. A., Capriola, S., Malosh, R. E., et al. (2020). Coronavirus occurrence and transmission over 8 years in the HIVE cohort of households in Michigan. *Journal of Infectious Diseases*, *222*, 9–16.

Morikawa, S., Kohdera, U., Hosaka, T., Ishii, K., Akagawa, S., Hiroi, S., et al. (2015). Seasonal variations of respiratory viruses and etiology of human rhinovirus infection in children. *Journal of Clinical Virology*, *73*, 14–19.

Moriyama, M., Hugentobler, W. J., & Iwasaki, A. (2020). Seasonality of respiratory viral infections. *Annual Review of Virology*, *7*, 83–101.

Nasci, R. S., Savage, H. M., White, D. J., Niller, J. R., Cropp, B. C., Godsey, M. S., et al. (2001). West Nile virus in overwintering Culex mosquitoes, New York City, 2000. *Emerging Infectious Diseases*, *7*, 742–744.

Nasir, A., Kim, K. M., & Caetano-Anollés, G. (2017). Long-term evolution of viruses: A Janus-faced balance. *BioEssays*, *39*, 1700026.

Nickbakhsh, S., Ho, A., Marques, D. F., McMenamin, J., Gunson, R. N., & Murcia, P. R. (2020). Epidemiology of seasonal coronaviruses: Establishing the context for the emergence of coronavirus disease 2019. *Journal of Infectious Diseases*, *222*, 17–25.

Nora, T., Charpentier, C., Tenaillon, O., Hoede, C., Clavel, F., & Hance, A. J. (2007). Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *Journal of Virology*, *81*(14), 7620–7628.

Paccaud, M. (1979). World trends in poliomyelitis morbidity and mortality, 1951-1975. *World Health Statistics Quarterly*, *32*, 198–224.

Pan, K., & Deem, M. W. (2011). Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. *Journal of the Royal Society Interface*, *8*, 1644–1653.

Peng, G., Xu, L., Lin, Y. L., Chen, L., Pasquarella, J. R., Holmes, K. V., et al. (2012). Crystal structure of bovine coronavirus spike protein lectin domain. *Journal of Biological Chemistry*, *287*, 41931–41938.

Phan, T. (2020). Genetic diversity and evolution of SARS-CoV-2. *Infection, Genetics and Evolution*, *81*, 104260.

Poirier, C., Luo, W., Majumder, M. S., Liu, D., Mandl, K. D., Mooring, T. A., et al. (2020). The role of environmental factors on transmission rates of the COVID-19 outbreak: An initial assessment in two spatial scales. *Scientific Reports*, *10*(1), 1–11.

Poole, L. (2020). Seasonal influences on the spread of SARS-CoV-2 (COVID19), causality, and forecastabililty (3-15-2020). *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3554746.

Posada, D., Crandall, K. A., & Holmes, E. C. (2002). Recombination in evolutionary genomics. *Annual Review of Genetics*, *36*(1), 75–97.

Rahimi, N. (2021). C-type Lectin CD209L/L-SIGN and CD209/DC-SIGN: Cell adhesion molecules turned to pathogen recognition receptors. *Biology*, *10*, 1.

Rahman, M. S., Islam, M. R., Rubayet Ul Alam, A. S. M., Islam, I., Hoque, M. N., Akter, S., et al. (2021). Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences. *Journal of Medical Virology*, *93*(4), 2177–2195.

Rath, S. L., & Kumar, K. (2020). Investigation of the effect of temperature on the structure of SARS-CoV-2 Spike protein by molecular dynamics simulations. *Frontiers in Molecular Biosciences*, *7*, 583523.

Rhodes, T., Wargo, H., & Hu, W.-S. (2003). High rates of human immunodeficiency virus type 1 recombination: Near-random segregation of markers one kilobase apart in one round of viral replication. *Journal of Virology*, *77*(20), 11193–11200.

Ricci, C. A., Kamal, A. H. M., Chakrabarty, J. K., Fuess, L. E., Mann, W. T., Jinks, L. R., et al. (2019). Proteomic investigation of a diseased Gorgonian coral indicates disruption of essential cell function and investment in inflammatory and other immune processes. *Integrative and Comparative Biology*, *59*(4), 830–844.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *43*(1), 59–66.

Rohani, P., Earn, D. J., & Grenfell, B. T. (1999). Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, *286*(5441), 968–971.

Sailani, M. R., Metwally, A. A., Zhou, W., Rose, S. M. S.-F., Ahadi, S., Contrepois, K., et al. (2020). Deep longitudinal multiomics profiling reveals two biological seasonal patterns in California. *Nature Communications*, *11*, 4933.

Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. (2020). Temperature, humidity, and latitude analysis to estimate potential spread and seasonality of coronavirus disease 2019 (COVID-19). *JAMA Network Open*, *3*, e2011834.

Sanjuán, R., & Domingo-Calap, P. (2016). Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, *73*(23), 4433–4448.

Sanjuán, R., Moya, A., & Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(22), 8396–8401.

Saputri, D. S., Li, S., van Eerden, F. J., Rozewicki, J., Xu, Z., Ismanto, H. S., et al. (2020). Flexible, functional, and familiar: Characteristics of SARS-CoV-2 spike protein evolution. *Frontiers in Microbiology*, *11*, 2112.

Sato, S., St-Pierre, C., Bhaumik, P., & Nieminen, J. (2009). Galectins in innate immunity: Dual functions of host soluble β-galactosidase-binding lectins as damage-associated molecular patterns (DAMPs) and as receptors for pathogen-associated molecular patterns (PAMPs). *Immunological Reviews*, *230*, 172–187.

Schober, A. F., Rzhetsky, A., & Rust, M. J. (2021). Seasonal disease in the United States has the hallmarks of an entrained circannual clock. *medRxiv*. https://doi.org/10.1101/2021.05.26.21257655.

Seronello, S., Montanez, J., Presleigh, K., Barlow, M., Park, S. B., & Choi, J. (2011). Ethanol and reactive species increase basal sequence heterogeneity of hepatitis C virus and produce variants with reduced susceptibility to antivirals. *PLoS One*, *6*(11), e27436.

Seyer, A., & Sanlidag, T. (2020). Solar ultraviolet radiation sensitivity of SARS-CoV-2. *The Lancet Microbe*, *1*(1), e8–e9.

Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T., & Lipsitch, M. (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biology*, *8*(2), e1000316.

Sheftel, H., Shoval, O., Mayo, A., & Alon, U. (2013). The geometry of the Pareto front in biological phenotype space. *Ecology and Evolution*, *3*, 1471–1483.

ARTICLE IN PRESS

Showers, W. M., Leach, S. M., Kechris, K., & Strong, M. (2021). Analysis of SARS-CoV-2 mutations over time reveals increasing prevalence of variants in the spike protein and RNA-dependent RNA polymerase. *bioRxiv*. https://doi.org/10.1101/2021.03.05.433666.

Singh, J., Pandit, P., McArthur, A. G., Banerjee, A., & Mossman, K. (2021). Evolutionary trajectory of SARS -CoV-2 and emerging variants. *Virology Journal*, *18*, 166.

Smit, A. J., Fitchett, J. M., Engelbrecht, F. A., Scholes, R. J., Dzhivhuho, G., & Sweijd, N. A. (2020). Winter is coming: A southern hemisphere perspective of the environmental drivers of SARS-CoV-2 and the potential seasonality of COVID-19. *International Journal of Environmental Research and Public Health*, *17*(16), 5634.

Smith, E. C., & Denison, M. R. (2013). Coronaviruses as DNA wannabes: A new model for the regulation of RNA virus replication fidelity. *PLoS Pathogens*, *9*, e1003760.

Soh, W. T., Liu, Y., Nakayama, E. E., Ono, C., Torii, S., Nakagami, H., et al. (2021). The N-terminal domain of spike glycoprotein mediates SARS-CoV-2 infection by associating with L-SIGN and DC-SIGN. *bioRxiv*. https://doi.org/10.1101/2020.11.05.369264.

Soper, H. E. (1929). The interpretation of periodicity in disease prevalence. *Journal of the Royal Statistical Society*, *92*(1), 34–73.

Soper, F. L. (1967). Dynamics of Aedes aegyptii distribution and density. *Bulletin of the World Health Organization*, *36*, 536–538.

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., et al. (2020). Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, *5*, 1295–1310.

Sugihara, G., Grenfell, B., & May, R. M. (1990). Distinguishing error from chaos in ecological time series. *Philosophical Transactions of the Royal Society, London B*, *330*(1257), 235–251.

Sugihara, G., Nay, R., Ye, H., Hsieh, C.-H., Deyle, E., Fogarty, M., et al. (2012). Detecting causality in complex ecosystems. *Science*, *338*(6106), 496–500.

Tamerius, J., Nelson, M. I., Zhou, S. Z., Viboud, C., Miller, M. A., & Alonso, W. J. (2011). Global influenza seasonality: Reconciling patterns across temperate and tropical regions. *Environmental Health Perspectives*, *119*, 439–445.

Tasakis, R. N., Samaras, G., Jamison, A., Lee, M., Paulus, A., Whitehouse, G., et al. (2021). SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial founder events and mutation bursts. *PLoS One*, *16*(7), e0255169.

Thépaut, M., Luczkowiak, J., Vivès, C., Labiod, N., Bally, I., Lasala, F., et al. (2021). DC/L-SIGN recognition of spike glycoprotein promotes SARS- CoV-2 trans-infection and can be inhibited by a glycomimetic antagonist. *PLoS Pathogens*, *17*(5), e1009576.

Tomaszewski, T., DeVriers, R. S., Dong, M., Bhatia, G., Norsworthy, M. D., Zheng, X., et al. (2020). New pathways of mutational change in SARS-CoV-2 proteomes involve regions of intrinsic disorder important for virus replication and release. *Evolutionary Bioinformatics*, *16*, 1176934320965149.

Tosepu, R., Gunawan, J., Effendy, D. S., Ahmad, L. O. I. E., Lestari, H., Bahar, H., et al. (2020). Correlation between weather and Covid-19 pandemic in Jakarta, Indonesia. *Science of the Total Environment*, *725*, 138436.

Troyano-Hernáez, P., Reinosa, R., & Holguín, Á. (2021). Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to September 2020: A global and regional approach by epidemiological week. *Viruses*, *13*, 243.

van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, B. N., et al. (2020). Aerosol and surface stability of SARS- CoV-2 as compared with SARS-CoV-1. *New England Journal of Medicine*, *382*, 1564–1567.

Vidal-Dupiol, J., Dheilly, N. M., Rondon, R., Grunau, C., Cosseau, C., Smith, K. M., et al. (2014). Thermal stress triggers broad *Pocillopora damicornis* transcriptomic remodeling, while *Vibrio coralliilyticus* infection induces a more targeted immuno-suppression response. *PLoS One*, *9*, e107672.

Vijaykrishna, D., Mujerki, R., & Smith, G. J. D. (2015). RNA virus reassortment: An evolutionary mechanism for host jumps and immune evasion. *PLoS Pathogens*, *11*(7), e1004902.

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

Wang, M., Jiang, A., Gong, L., Lu, L., Guo, W., Li, C., et al. (2020). Temperature significant change COVID-19 transmission in 429 cities. *medRxiv*. https://doi.org/10.1101/2020.02.22.20025791.

Weber, T. P., & Stilianakis, N. I. (2008). Inactivation of influenza A viruses in the environment and modes of transmission: A critical review. *Journal of Infection*, *57*, 361–373.

Wessa, P. (2021). *Free statistics software (v1.2.1), Office for Research Development and Education*. http://www.wessa.net.

Woo, J., Lee, E. Y., Lee, M., Kim, T., & Cho, Y.-E. (2019). An in vivo cell-based assay for investigating the specific interaction between the SARS-CoV N-protein and its viral RNA packaging sequence. *Biochemistry and Biophysics Research Communications*, *520*, 499–506.

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., et al. (2020). Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. *Science of the Total Environment*, *729*, 139051.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*, 265–269.

Wu, Y., Zhou, Z., Wang, J., Luo, J., Wang, L., & Zhang, Y. (2019). Temperature regulates the recognition activities of galectin to pathogen and symbiont in the sclerotinian coral *Pocillopora damicornis*. *Developmental and Comparative Immunology*, *96*, 103–110.

Wyse, C., O'Malley, G., Coogan, A., McConkey, S., & Smith, D. (2021). Seasonal and daytime variation in multiple immune parameters in humans: Evidence from 329,261 participants of the UK Biobank cohort. *iScience*, *24*, 102255.

Xu, C., Wang, Y., Liu, C., Zhang, C., Han, W., Hong, X., et al. (2021). Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. *Science Advances*, *7*(1), eabe5575.

Yafremava, L. S., Wielgos, M., Thomas, S., Nasir, A., Wang, M., Mittenthal, J. E., et al. (2013). A general framework of persistence strategies for biological systems helps explain domains of life. *Frontiers in Genetics*, *4*, 16.

Yang, H.-Y., & Lee, J. K. W. (2021). The impact of temperature on the risk of COVID-19: A multinational study. *International Journal of Environmental Research and Public Health*, *18*, 4052.

Ye, Q., West, A. M., Silletti, S., & Corbett, K. D. (2020). Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Science*, *29*(9), 1890–1901.

Zeger, S. L., Irizarry, R., & Peng, R. D. (2006). On time series analysis of public health and biomedical data. *Annual Reviews of Public Health*, *27*, 57–79.

ARTICLE IN PRESS

Zhou, L., Yang, H., Kuang, Y., Li, T., Xu, J., Li, S., et al. (2019). Temporal patterns of influenza A subtypes and B lineages across age in a subtropical city, during pre-pandemic, pandemic, and postpandemic seasons. *BMC Infectious Diseases*, *19*, 89.

Further reading

Callaway, E. (2020). The coronavirus is mutating-does it matter? *Nature*, *585*(7824), 174–177.

Johannes, L., Jacob, R., & Leffler, H. (2019). Galectins at a glance. *Journal of Cell Science*, *131*(9), jcs208884.

Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. *Annual Review of Virology*, *3*, 237–261.

Modenutti, C. P., Capurro, J. I. B., Di Lella, S., & Martí, M. A. (2019). The structural biology of galectin-ligand recognition: Current advances in modeling tools, protein engineering, and inhibitor design. *Frontiers in Chemistry*, *7*, 823.

Glossary

Amino acid variant A mutant protein with an amino acid replaced by another in a particular site of its polypeptide sequence.

Baltimore classification of viruses A classification system of viruses that groups viruses according to seven types of pathways used for mRNA synthesis: Group I: double-stranded DNA viruses (dsDNA); Group II: single-stranded DNA viruses (ssDNA);Group III: double-stranded RNA viruses (dsRNA); Group IV: positive sense single-stranded RNA viruses (plus-ssRNA); Group V: negative sense single-stranded RNA viruses (minus-ssRNA); Group VI: single-stranded reverse transcribing RNA viruses with a DNA intermediate (ssRNA-RT); and Group VII: double-stranded reverse transcribing DNA viruses with an RNA intermediate (dsRNA-RT).

Clade A group of taxa with a common evolutionary origin.

Coronavirus spike protein The spike (S) protein is the largest structural protein of coronaviruses. The S-protein assemble into trimers to form spike structures that project outward from the surface of the virion and give the typical 'solar corona' appearance of the virion under negative stained transmission electron microscopy. The highly glycosylated 1200–1400 amino acid long protein is typically composed of two regions known as the S1 and S2 subunits. The N-terminal S1 subunit interacts with receptor molecules on the surface of cells and contains an N-terminal domain (NTD) and a C-terminal domain (CTD). While both domains can interact with receptors, the CTD of the SARS-CoV-2 virus is known as the receptor-binding domain (RBD). The C-terminal S2 subunit embeds the spike in the viral envelope and mediates viral entry by fusion of viral and host cell membranes. It holds a 'fusion' region with hydrophobic fusion peptide (FP) and internal fusion peptide (IFP) sequences, a 'fusion core' region composed of two heptad repeat (HR) sequences that undergo major fusion-triggered conformational changes, and a C-terminal transmembrane (TM) domain that acts as anchor.

Economy The budget of a system in terms of matter-energy costs to maintain its functioning mechanisms.

Epidemic calendar A calendar of seasonality of viral and bacterial diseases.

Flexibility The structural and functional mechanisms responding to 'internal' and 'external' changes imposed on the system; these mechanisms require processing of information and delimit the potential of the system to adapt to environmental change.

Galectins A class of proteins known to bind specifically to β-galactoside sugars with broad specificity and harbour a disulphide bond-dependency for carbohydrate binding and stability, making them S-type lectins.

***Human coronaviruses* (HCoVs)** A group of coronaviruses that can cause infections of the respiratory system and mild-to-severe illnesses, including common cold, pneumonia, severe acute respiratory syndrome, kidney failure and even death.

Immune escape The ability of a virus to evade innate or adaptive immunity mechanisms.

Morphospace Spaces describing worlds of phenotypes, including observable molecular, cellular, and organismal characteristics resulting from the interaction of genotypes with the environment.

Persistence Maintenance of identity through time, including the continuance of a feature or lineage of an organism or virus.

Phylogenetic tree A specific hypothesis of history (a phylogeny) in the form of a network without reticulations.

Phylogeny A hypothesis of history and genealogical relationship among a group of entities (taxa) in the form of a tree or network with specific connotations of ancestry and an implied time axis.

***Protein Intrinsic Disorder* (PID)** The lack or fixed or ordered three-dimensional molecular structure that endows a molecule with flexibility, typically in the absence of macromolecular interacting partners. PID can be a property of the entire protein in intrinsically disordered proteins (IDPs) or of particular regions within protein domains or in linker or terminal sequences. PID can be predicted with high confidence at per-residue level directly from amino acid sequence with a number of algorithmic implementations, such as the IUPred energy-based prediction method.

Robustness The mechanisms that use information to passively maintain structure and function despite external influence; these mechanisms protect the system from malfunction by resisting damage and change.

***Root-mean-square deviation* (RMSD)** When referring to atomic positions, RMSD measures the average distance (in Å) between the atoms of superimposed proteins. Typically, RMSDs of the C_α atomic coordinates of the protein backbone are used to measure similarities of three-dimensional structures after optimal rigid body superposition. RMSD is used as quantitatively measure of structural similarity between one or more proteins, with lower values implying better structural match.

SARS-CoVs Strains of coronaviruses that cause severe acute respiratory syndrome (SARS), a respiratory illness that materialized in two major outbreaks, the 2002–2004 SARS-CoV-1 outbreak and the ongoing 2019 SARS-CoV-2 pandemic outbreak.

Seasonal driver Causative agent of the seasonal behaviour of infectious diseases.

Seasonal forcing Seasonal variations in the onset and spread of infectious diseases.

Spike protein A round, flattened or button-shaped glycoprotein (previously known as a peplomer protein; from Greek, peplos 'robe', meros 'part') that projects outward from the lipid bilayer of the surface of an enveloped virus. Spike proteins play important roles, including attachment of the virion to receptor sites on cell surfaces, mediating release of the nucleocapsid with its genetic material, and harbouring hemagglutinating or enzymatic activities.

ARTICLE IN PRESS

***Variant of concern* (VOC)** A variant of a virus exhibiting a set of mutations (amino acid variants) that increase virus transmissibility, morbidity, mortality, risk, immune escape, diagnostic test evasion, or other criteria of significance.

Viral quasispecies A dynamical virus collective that is structured at a population level by exhibiting a large number of variants that arise continually via mutation and are subjected to natural selection.

Virion The infective submicroscopic particle form of a virus, complete with a capsid protein coat structure encasing an infectious nucleic acid.

Virocell A cell with physiologies transformed by viral infection geared towards production of virions rather than cellular multiplication.

Z-score In bioinformatics, the Z-score is simply the comparison of an actual alignment score with the scores obtained on a set of random sequences by a Monte-Carlo process. In the case of DALI, the Z-score is an optimized similarity score defined as the sum of equivalent C_α-C_α atomic distances among two proteins. The maximum of the total score defines a common structural core where every atom in the alignment makes a net positive contribution. Scores are rescaled according to length because large structural alignments get higher scores than smaller alignments. Significant similarities have Z-scores > 2, which usually correspond to similar folds.