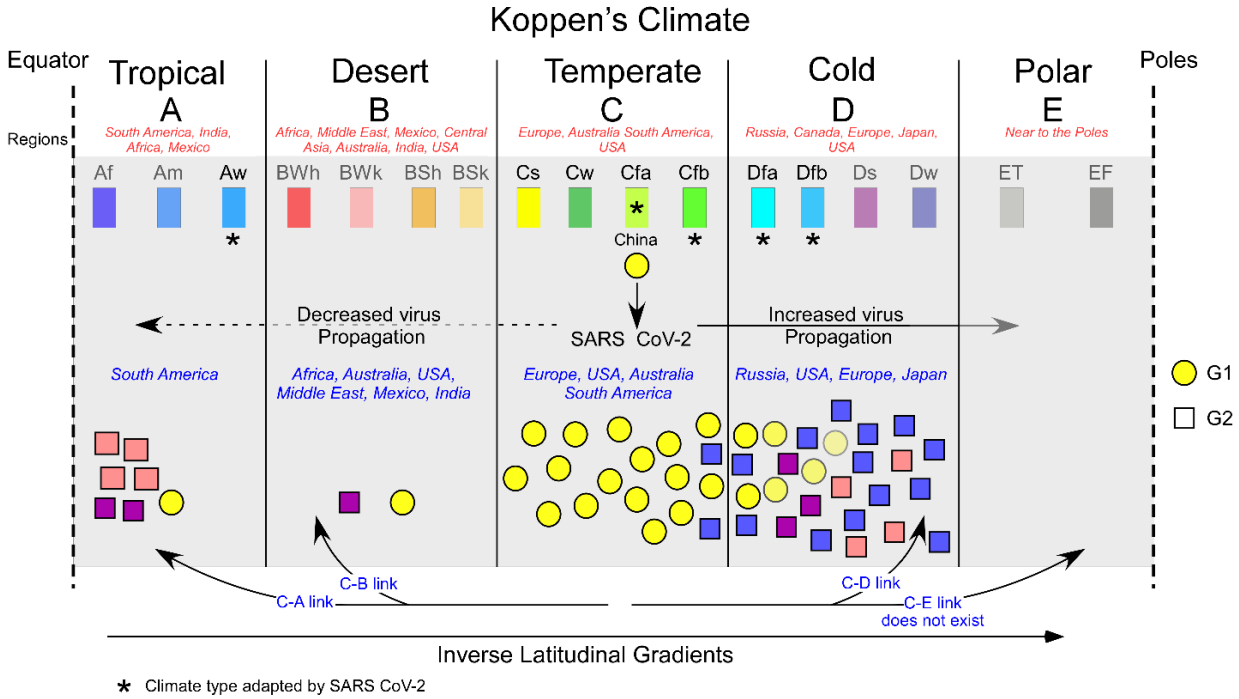
Climatic-niche evolution of SARS CoV-2

Authors: Priyanka Bajaj, Prakash Chandra Arya
Correspondence: prakasha@iisc.ac.in

Graphical Abstract



In Brief:

The authors elucidate adaptation of SARS CoV-2 to different climates by studying phylogenetics and the distribution of strains on Koppen's climate map.

Highlights:

- SARS CoV-2 follows inverse latitudinal gradient.
- Phylogenetic network divides SARS CoV-2 strains into two variant groups, G1 and G2.
- G1 strains is restricted to Koppen's *"temperate"* climate (mainly *Cfa-Cfb*).
- G2 strains has evolved from G1 to sustain in "*humid-continental" (Dfa-Dfb)* and "*tropical-savannah" (Aw)* climate.

Climatic-niche evolution of SARS CoV-2

Priyanka Bajaj[1#] and P. C. Arya[2# *]

1 Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India

2 Centre for Earth Sciences, Indian Institute of Science, Bangalore-560012, India

# Both authors have contributed equally to this work.

*Correspondence: prakasha@iisc.ac.in

Abstract

Adaptation of species to new environments is governed by natural selection that discriminates among genetic variations and favors survival of the fittest. Here, we propose climate plays an important role in the evolution of SARS CoV-2 and the spread of COVID-19 all over the world which was previously not known. To understand the climatic factors responsible for shaping the molecular determinants of the novel coronavirus, genotyping SARS CoV-2 across different latitudes and Koppen's climate is imperative. It seems this virus follows inverse latitudinal biodiversity gradient due to its preference towards Koppen's temperate (C) and cold climate (D). Our molecular phylogenetic analysis revealed division of 176 SARS CoV-2 strains into two variant groups, G1 and G2, well defined by four mutations. Initially, SARS CoV-2 was restricted to a *"humid-subtropical" (Cfa)* climate of southeast China, which soon spread all over the world having C climate. Genomic information superimposed on global Koppen's climate map elucidates that the gradation *"humid-subtropical" (Cfa)* and *"marine-temperate"* (Cfb) to *"humid-continental" (Dfa-Dfb)* climate drives the evolution of G1 into G2 variant group. It seems an early infection in Europe and USA is due to the dominance of C climate. Russia and North America were infected through linkage of C to D climate and South America from C to A

climate. Our study elucidates viruses are sensitive to climate and combined genomic and climatic studies provide crucial information about the pathogenesis and natural spreading pathways during a pandemic which will enable us to take pre-emptive precautionary measures in such outbreaks.

Keywords: SARS CoV-2, molecular phylogeny, virus cluster SNPs, inverse latitudinal gradient, climate zones, Koppen's climate.

Introduction

In December 2019, Corona Virus Disease – 2019 (COVID-19) cases emerged in Wuhan, China. It has affected more than 6 million people worldwide and has claimed over 370,000 lives according to World Health Organization (WHO)[1]. Several studies suggest that SARS coronavirus-2 (SARS CoV-2) has a preference towards higher latitude as most of the cases were reported from temperate regions (Europe, China, USA, etc.)[2,3]. With time, the cases increased in tropical countries as well but still the number of cases in higher latitudes is larger than the tropics. The earliest and the most simple classification of Earth's climate is based on latitudes which divide the Earth's climate into seven climate zones, North Frigid Zone (NFZ), North Temperate Zone (NTZ), North Subtropical Zone (NSTZ), Tropical Zone (TZ), South Subtropical Zone (SSTZ), South Temperate Zone (STZ) and South Frigid Zone (SFZ)[4]. Wladimir Koppen presented a modified classification of Earth's climate based on the precipitation and temperature[5]. Koppen divided Earth's climate into five major climates, A (Tropical), B (Arid), C (Temperate), D (Cold or Continental) and E (Polar)[4,5]. These climates are further subdivided into 30 climate types which are the habitats of a wide variety of plants, animals and microbes[5].

In general, the plant and animal diversity decrease on moving from equator to pole. This pattern of biodiversity is known as the latitudinal biodiversity gradient. The latitudinal gradient is one of the universal feature of nature and has been identified and discussed by several authors[6–9]. Rohde proposed over 30 hypotheses to explain this phenomenon[10]. However, exceptions to this gradient also exist[8,11]. Some groups of organisms, however, show an opposite trend as their population declines towards the equator. Very few studies on the inverse latitudinal gradient exist, and very little is known about their underlying ecological and evolutionary causes[12]. Recently, due to COVID-19 outbreak, a worldwide research has been conducted on SARS CoV-2. Genomic studies of SARS CoV-2 in different climates provides the information about response of virus to the external environment which is useful to take pre-emptive precautionary measure in such outbreaks.

The etiological agent of this year's Severe Acute Respiratory Syndrome (SARS) is novel SARS CoV-2 belonging to the *beta-coronavirus* group, *Nidovirales* order, *coronaviridae* family which is an enveloped virus made up of large single-stranded ~30 kb long positive-sense RNA. These viruses majorly have a conserved genomic organization, consisting of a unique 265 bp long leader sequence, ORF1ab polyprotein, and structural proteins like S (spike glycoprotein), E (Envelope), M (Membrane), and N (Nucleocapsid). ORF1ab encodes replicase, transcriptase and helicase, essential enzymes required for replication, along with non-structural and accessory proteins. Expression of non-structural proteins is facilitated by ribosomal frameshifting and several enzymatic activities are nested within the replicase-transcriptase polyprotein complex[13]. All coronaviruses express structural proteins S, E, M, N; spike glycoprotein being the most immunogenic to T-cell response[14].

Spike glycoprotein of coronaviruses binds to human angiotensin-converting enzyme 2 (hACE2) receptor for viral fusion and entry and is the main target for neutralizing antibodies and development of vaccines[15]. Membrane protein is also antigenic as it stimulates a humoral immune response[16]. E protein is responsible for virus assembly and release of virion particles[17]. Nucleocapsid protein packages RNA genome into a helical ribonucleocapsid protein (RNP) complex during virion assembly and is capable of eliciting an immune response[18]. SARS CoV-2 has a very high mutational rate attributing to the ability of the RNA dependent RNA Polymerase to jump and replicate the viral genome in an error-prone manner[19]. Other factors responsible for generating mutations include nucleic acid damage, diversity-generating retro-elements (DGRs), or editing of the genetic material by host-encoded proteins[20]. Nucleotide substitution has been proposed to be one of the most important mechanisms of viral evolution in nature[21]. However, factors responsible for the generation of these mutations are not well understood. One of the factors is adaptation to new environments dictated by natural selection that discriminates among genetic variations and favors survival of the fittest[22].

Although infection rate will depend on population density, health and hygiene and public awareness of a region but if SARS CoV-2 is adapting to different climate, it can be delimited by analyzing genomic similarity across different latitudes and Koppen's climate. The present study comprises of three parts, (1) latitudinal distribution of COVID-19 is investigated, (2) sequence analysis of 176 SARS CoV-2 strains, (3) distribution of SARS CoV-2 strains in different Koppen's climate to evaluate the effect of climate on SARS CoV-2 evolution. These studies can provide insights on the preferential distribution of

SARS CoV-2 across the globe and the underlying climatic factors that shape the molecular determinants of this virus genome.

Results

Distribution of COVID-19 cases across latitudes

For an overview of the latitudinal preference of SARS CoV-2, we have plotted per million active cases of SARS CoV-2 across different climate zones (Figure 1a). Results show that 81% of the cases belong to NTZ (30°N - 66.5°N), 4% to NSTZ (23.5°N - 30°N), 14% lie in the TZ (23.5°N - 23.5°S), 1% in the STZ (30°S - 66.6°S) and negligible (<0.5%) number of cases have been reported from the rest of the climate zones. Statistical difference exists between number of COVID cases in Temperate Zone versus other climate zones (paired *t*-test two-tail, $P<.001$). The spread of COVID-19 is dominant in the higher latitudes which is usually uncommon as a majority of terrestrial texa prefers to stay near tropical region, suggesting that SARS CoV-2 follows inverse latitudinal gradient in early stage of pandemic. Since a majority of the cases lie in the North Temperate Zone, we have further divided this zone into an interval of 7° latitude i.e. 30°N - 37°N, 37°N - 44°N, 44°N - 51°N, 51°N - 58°N and 58°N - 66.5°N. We found 9% of the cases fall in latitude range 30°N - 37°N, 46% in 37°N - 44°N, 21% in 44°N - 51°N, 14% in 51°N - 58°N and 10% in 58°N - 66.5°N (Figure 1b). The results show a peak of COVID-19 cases in between 37°N to 51°N latitudes, the dominant Koppen's climate between these latitudes is temperate (C) and continental climate (D). The general characteristics of these climates are prevalence of high atmospheric circulation with anticyclones during winters, with an average temperature of ~15°C for C and ~< 10°C for D climate, with relative humidity ranging between ~50-80%. Since the major distribution of SARS CoV-2 is confined within

a latitude range, this trend could be random or there might be a strong underlying cause driven by underlying principles. Since latitude has a very high control on climate, a detailed investigation of the Koppen's climate under the light of genomic sequences is carried out to understand the distribution pattern across the globe.

Molecular phylogeny analysis to infer genomic similarities and their distribution in different climates

We retrieved 176 SARS CoV-2 full-length genome sequences from GISAID database. To probe the genomic similarities between SARS CoV-2, we aligned 176 genomic sequences of viral isolates from countries all over the world to the reference genome[24]. and constructed a phylogenetic tree. Interestingly, our MSA results reveal sixty virus cluster SNPs. SNPs of these virus clusters and their distribution pattern across different climate zones, Koppen's climate and Koppen's climate type is mentioned in Table 1. Climatic parameters for each virus strain used in this study are mentioned in Table S2. Based on phylogenetic clustering, 176 SARS CoV-2 strains are majorly divided into two groups, we named them as G1 (1-58) and G2 (59-176), well defined by four SNPs (Figure 2). First synonymous mutation occurred at 241 position of the unique leader sequence (C241T), second at F924 (C3037T) in nsp3, encoding for papain-like proteinase[32,33], third a non-synonymous mutation took place in ORF1b, P214L (C14408T), that codes for five putative non-structural proteins (nsp13, nsp14, nsp15 and nsp16), functionally involved in replication-transcription complex[34], and fourth a non-synonymous mutation of D614G (A23403G) in S gene, encoding spike glycoprotein (Figure 3a). Mostly, these four mutations are exhibited by 118 strains of G2 group, mainly covering Europe and North America. Since the mortality rate is higher in temperate region (Europe and USA) as

mentioned by WHO (2020), these results suggest these four SNPs could be the key factors in increasing the virulence, transmission and sustainability of the virus in humans in temperate zone.

We explored the extent of genome-wide divergence of G1 and G2 group across different climate zones and Koppen's climate (Figure 3b). 59% of G1 viruses fall in NTZ, 14% in NSTZ, 12% in TZ, 10% in SSTZ and 5% in STZ. 76% of the virus isolates in G2 group are present in the NTZ, 13.5% in TZ, 7.6% in STZ and remaining 2% is equally distributed in NSTZ and SSTZ, showing G2 strain variants evolved to adapt to temperate zones as their population decreased drastically in the subtropical zones. These results show both G1 and G2 strains have a strong preference towards higher latitudes i.e., NTZ, which agrees with the analyzed worldometer data (Figure 3c). These results support that the virus isolates under study follow inverse latitudinal gradient. Mapping viral strains on Koppen's map reveal their prevalence majorly in the C and D climate (Figure 3d). 71% of G1 lie in C climate, 17% in D and the remaining is equally distributed in the A and B climate. 54% of G2 lie in C climate, 36% in D, 9% in A and 1% in B climate pointing towards a preferential shift of the novel coronavirus towards D climate (Figure 3B), alluding G2 is climatically and genomically diverse than G1.

We further analyzed the order in which the phylogenetic clusters evolved from the ancestor 45-57 cluster (containing the reference genome, Strain ID: 50) based on nodes, mutational branches and branch length. The order in which the virus evolved is 44-47 (G1440A, G392D; G2891A, A876T), 1-22 (C8782T, S2839; T28144C, L84S), 33-43 (G26144T, G251V), 23-32, 58-61 (C15324T, N519), 80-115 (G28881A, G28882A, R203K), 116-125 (A20268G, L2167), 126-176 (G25563T, Q57H) and 62-79 (cluster,

acquired mutations). In Figure 3e, looking at the distribution of the viruses in different climate zones, no such preference was observed as the virus evolved but virus clusters in G2 group gradually evolved to sustain in Koppen's D climate which supports our previous observations. Within these major virus clusters, small clusters also exist as shown in Table 1 with their mutational data along with their climatic distribution.

We have examined whether climatic conditions exhibit any selective pressure on each gene (Figure 3f). The pattern suggests inverse latitudinal gradient of SARS CoV-2 and we observed all genes mutate in the NTZ, as expected. Mutations in the M gene are only pertaining to NTZ and NSTZ and Koppen's C and D climate. In particular, there is a surge in the virus strains carrying SNPs in ORF8 in the NSTZ (20%). 77% of the SNPs in ORF8 lie in the C and 20% in the D climate. Overall, the distribution of virus cluster SNPs of ORF1ab, S, ORF3a, and N follow a similar pattern across all the climatic zones and Koppen's climate, implying no difference in selective pressure of the climate in generating mutations in these genes. SNPs are distributed across different climate zones and Koppen's climate, suggesting structural as well as non-structural genes significantly mutate to adapt to the external environment of different climate zones and Koppen's climate (Table 1). S, M and N proteins are immunogenic implicating virus evades immune response by introducing these substitutions.

Apart from non-synonymous mutations, synonymous mutations within the gene can also significantly affect protein function due to codon usage bias[35,36] and through mechanisms such as ribosome stalling[37] and mRNA secondary structure formation[38]. We probed the frequency of derived synonymous versus non-synonymous mutations and observed a very similar distribution pattern of the derived synonymous versus missense mutations

across all climate zones and Koppen's climate (Figure 3g). These analyses suggest novel coronavirus is using varied mechanisms both at the transcriptional as well as translational level to adapt, survive, and, increase infectivity in all types of climates. These findings unequivocally bolster a requirement for further prompt, comprehensive studies that join genomic information, epidemiological information, and climatic distribution with COVID-19 severity.

Distribution of strains across Koppen's climate

We superimposed genomic information on Koppen's climate map to reveal climatic factors involved in shaping the genome of the virus. A full description and criteria of different Koppen's climate type and their symbols are given in Supplementary Table S3. List of Koppen's symbol for each climate type is given in the Supplementary Table S4. The Koppen's climate type is marked within quotations and its standard Koppen's symbol is written within the parentheses (e.g., "*humid-subtropical" (Cfa)*, humid-subtropical is the Koppen's climate type and *'Cfa'* is its standard symbol) throughout the text. We projected the exact locations of 176 strains of SARS CoV-2 on a Gieger-Koppen map (Figure 4). The dominant Koppen climate types for this virus is defined in Supplementary Table S5. We observed, mostly (151/176) virus isolates are distributed in the coastal regions (Chi-square test, $P<.001$). ~85.7% of the total strains are distributed in the coastal regions and 14.3% in the continental region (Figure 5A). Around ~73.86% of the total strains are distributed in *"humid-subtropical" (Cfa)* and *"marine-temperate" (Cfb)* climate type of C climate and *"humid-continental" (Dfa-Dfb)* climate type of D climate. The remaining ~26.14% strains are distributed in other climate types of other Koppen's climate including non 'Cfa-Cfb' of C climate and non 'Dfa-Dfb' of D climate. The *"humid-subtropical" (Cfa)*

climate is found in the eastern coasts of continents and *"marine-temperate" (Cfb)* in the western coasts, these climates are also known as east and west coast climate. It seems that spread of COVID-19 is maximally in areas with *'Cfa'* and *'Cfb'* climate type. We further analyzed the climatic parameters such as temperature and precipitation of each climate. Statistically, significant difference was found in the temperatures (one-way ANOVA, $P<.001$) of G1 (15.82±0.75 ^{o}C (mean±s.e.)) and G2 (11.67±0.68 ^{o}C) strains (Figure 6A). There exists a statistically significant (one-way ANOVA, $P=.03521$) difference in the latitudes of G1 and G2 strains (Figure 6b). However, the difference in precipitation for G1 (1046.95±80 mm) and G2 (896.64±35.48 mm) strains is statistically not significant (one-way ANOVA, $P=.06118$) (Figure 6c). The latitudes and temperature are inversely related to each other (r = -0.6649, Supplementary Figure S1a), which explains the occurrence of G1 strains in lower ~24.14±3.5 (mean±s.e.) and G2 strains in higher latitudes, 34.03±2.7 (Figure 7a). Such relation between latitude and precipitation has not been observed (r = -0.3064, Supplementary Figure S1b) (Figure 7b). We simultaneously probed the effect of all three climate variables on each SARS CoV-2 isolate in G1 and G2 group by generating a mesh plot. (Supplementary Figure S2). Results show G2 strains have a preferential shift towards lower temperature and higher latitude and occur over a wider range of temperature, latitude, and precipitation compared to G1 strains.

We carefully examined the extent of the climate with maximum as well as minimum number of strains along with their neighboring climate, its stretch, termination and gradation to another climate all over the globe. An overview of the map shows, the distribution of 176 strains are mainly concentrated in the western coasts of Europe and North America, and eastern coasts of China, North America, Australia and South America

(Figure 4). Most of the strains are present in the coastal regions. On the map the SARS CoV-2 strains are mainly distributed in the "*humid-subtropical" (Cfa)* and "*marine-temperate" (Cfb)* and "*humid-continental" (Dfa and Dfb)* climate. Two strains from virus clusters (80-115 and 126-176) belonging to South America, are found in "tropical-savanna" *(Aw)* of 'A' climate. The distribution of strains all over the globe is as follows, beginning from the origin of outburst:

China: The entire southeast China has "*humid-subtropical" (Cfa)* climate also referred in the text as Wuhan-type climate. This climate grades into the cold deserts in the north and the land is separated from ocean in the south which isolates SARS CoV-2 strains (n=16) in '*Cfa'* climate. Majority (15/16) of strains from China belong to G1 group (Figure 4). One G2 strain have appeared in towards the C to D climate gradation near Beijing, suggesting G1 population favors (Chi-square test, P <.001) *'Cfa'* climate type of C climate.

Europe: The western coast of Europe consists of *"marine-temperate" (Cfb)* climate, a climate similar to Wuhan's climate i.e., '*Cfa*'. In UK, Spain, France, and Switzerland mainly *'Cfb'* climate persists, whereas Portugal has *"Hot-summer Mediterranean*" *(Csa)* climate. Towards west of Germany the '*Cfb'* climate dominates which grades into *"humid-continental" (Dfa-Dfb)* climate towards east which continues as a belt up to Japan. From the total strains (n=34), around 14.7% of G1 and 85.3% of G2 strains lie in Europe. All G1 strains (n=5) of Europe belong to C climate, of which 60% belongs to Cfb, 40% to Csa. Among G2 strains (n=29), 79.3% of G2 strains belongs to C climate (65.5% Cfa,10.3% Csa, and 3.4% Cfc) and 20.6% to Dfb climate type of D climate.

Russia: Majority (21/23) of strains (n=23) from Russia are present in the "*humid-continental" (Dfa-Dfb)* climate belt which begins from Germany and enters into the continent interior as a long (~8500 km) wide (~600-1700 km) belt, grading to *(Dwb-Dwc)* and tapering towards eastern side but continuing all along the southern boarders of Russia up to Japan. From Russia, ~8.7% of total strains (n=23), belongs to G1 and 91.3% to G2. All G1 (2/23) strains are present in '*Dfb'* climate, one strain is present near the gradation of C to D (Strain ID: 4) climate and another (Strain ID: 31) in the interior of the continent. Of G2 strains, 91.3% of the strains are present in D climate (61.9% Dfb, 9.5% Dfa, 9.5% Dfc, 4.7% Dwb, 4.7% Dwc) and 9.5% in Cfa of C climate, suggesting a strong preference (Chi-square test, $P<.001$) of G2 strains towards D climate.

North America:

USA: Of all the continents North America USA has the most diverse climate, especially towards the western side[1]. The strains are mainly present in the eastern and western coasts of USA. The eastern coast of USA is one of the largest regions (~2.4 x10^6 Km^2) of the world having *"humid-subtropical" (Cfa)* climate (Wuhan's climate). From the eastern coast, the strains enters into the continent interior through a long (~3800 km), wide belt (~600-1000 km) lying in the northern extremities of the USA, extending roughly in NW-SE direction initially following borders between USA and Canada and entering to Canada from southern side. This belt belongs to "*humid–continental" (Dfa-Dfb)* climate; is similar to that of Russia (mentioned above). The strains in the western coast of USA are aligned roughly parallel to the coastline and shows bulging towards the south (Figure 4), the western coast of USA has mainly C climate, the buldged out portion has a bulged *"Hot-summer mediterranean*" *(Csa)* climate, which grades into *"humid-continental" (Dfa-Dfb)*

towards its east. Between the western coast strains and eastern coast strains lies the cold desert, where SARS CoV-2 strains are not present. In USA, ~31% of strains (n=42) belong to G1 while ~69% of the strains belongs to G2. Among G1 strains (n=13), 46.15% of strains belong to C climate (23% Cfa, 23% Cfb), 38.4% to D climate (15.3% Dfa, 23.07% Dfb) and 7.6% in both *"tropical-monsoon" (Am)* and "*cold-desert" (BSk)* climate. The G1 strains of D and A zones mainly lie near boundaries of C and D climate around the eastern and western coasts (Figure 4). Within G2 strains (n =29), ~62% strains lie in C climate (55% Cfa, 6.8% Csb), ~31% in D climate (20.6% Dfa, 10.3% Dfb) and remaining 6.8% are equally distributed in tropical climate (3.4% Am, 3.4% Aw).

Canada: The strains of Canada (n=9) are mainly distributed along the western coasts and towards the southern side. The western coasts of Canada has *"humid-subtropical" (Cfa)* climate and south of Canda has *"humid-continental" (Dfa-Dfb)* climate, which is an extension of '*Dfa-Dfb'* belt initiating from eastern side of USA near New York (mentioned above). 22.2% of the strains belong to G1 and 77.8% to G2. Within G1 (n=2), 50% strains belong to the '*Cfa'* and remaining to '*Dfb'*. Within G2 (n=7) variant group, ~14.2 % strains belong to *"marine-temperate" (Cfb)* and ~85.7% of strains belong to *"humid-continental" (Dfb)* climate.

South America: Majority (18/19) of South America's strains (n=19 of strains are present in the eastern coast of South America, The G1 population is concentrated in the Sao Paulo and one G1 strain (Strain ID: 3) is visible in the western coast of Chile, in both the places C climate is dominant, mainly *"humid-subtropical" (Cfa) and "marine-temperate" (Cfb).* Other than these two places the C climate is not present in the entire South America. The position and extent of South America in the globe is unique in itself as it

connects the C climate with all A (tropical) climate through land. Both G1 and G2 strains are present in the C climate, however G2 strains have shifted towards the *"tropical-savannah" (Aw)* climate towards north, this shift is not visible towards the desert climate in south. Around ~31.5% of G1 strains and 68.4% of G2 strains are present in South America. Within G1 (n=6), ~66% are present in the C climate (33.3% Cfa, 16.6% Cfb, 16.6% Csb), 33.3% in A climate (Aw). Within G2 (n=13), 69.23% of G2 strains are present in C climate (46.15% Cfa, 23% Cfb), while 30.73% of the strains are present in A climate (23% Aw, 7.6% Am).

Africa: Strains from Africa (n=4), are mainly (3/4) from G2 group. One G1 strains belongs to C climate (Cwb). Within G2, 66.66% of strains are present in A climate (Aw), and 33.33% in C climate (Csa).

Australia: All strains (n=11) from Australia are present either in the eastern or in the western coasts consisting of C climate. In the eastern coast the main climate is "*humid-subtropical" (Cfa)* and "*marine-temperate" (Cfb)* climate and western coast consists of *"Hot-summer mediterranean" (Csa)* climate of C climate. The rest of Australia has a desert climate. All strains from Australia are confined to this narrow belt of C climate. ~27% of the strains in Australia belong to G1 and ~73% of the strains belong to G2. Within G1 66.6% of strains are present in Cfa and 33.3% in Cfb. In G2, 50% of strains are present in Cfa, 25% in Cfb and 25 % in Csa climate type.

Japan: Japan has mainly two climates i.e., *"marine-temperate" (Cfb)* towards south and *"humid-continental" (Dfb)* towards north. Strains (n=2), one G1 and another G2 strain from Japan belongs to Cfa climate.

Middle East: The entire Middle East consists of B climate (desert). A very small portion of Middle East consists of C climate in the regions around Turkey. All strains (n=4) of Middle East belong to G1 group, among which 50% belongs to C climate (25% Csa, and 25% Csb) while the remaining are present in the "hot-desert" (BWh) climate.

South Asia and South Asian Islands

The G2 strains are present in India, Thailand and Vietnam, are mainly from "*tropical-savannah" (Aw)* climate, except strains from north-west India with a desert climate (BSh). The G1 strains are present in Philippines and South Korea are having "*tropical-savannah" (Aw)* and "*humid-continental"* (Dfa) climate respectively. The South Korea strain lies in the transition of *'Cfa'* climate (China) to '*Dfa'* climate (South Korea). Most of the strains in the South Asia and South Asian Islands belong to G2. Of total, around 80% of G2 and 20% of G1 strains are present in South Asia and South Asian Islands.

Discussion

In the present work, the genomic sequences, latitudinal gradient and climatic distribution of SARS CoV-2 is studied to understand the reason behind the preferential distribution of the novel coronavirus throughout the world. Unlike majority of terrestrial organisms, the current global distribution pattern of SARS CoV-2 shows inverse latitudinal gradient. Similar results of inverse latitudes were found for pathogenic and infectious diseases[8] and sawflies[11]. We have attempted to provide a plausible cause for the inverse latitudinal gradient through integrating the molecular phylogeny results of SARS CoV-2 and their distribution across different climate zones and climate type of Koppen.

Our genomic analysis of virus strains isolated from several countries all over the world, suggest the novel coronavirus uses multiple mechanisms both at the transcriptional and translational level for evading the immune response, developing drug resistance and increasing pathogenesis. The ORF1a harbors 12 non-structural proteins (nsp) namely nsp1, nsp2, nsp3 (papain-like protease), nsp4, nsp5 (3C-like protease), nsp6, nsp7, nsp8, nsp9, nsp10, nsp11 and nsp12 (RNA-dependent RNA polymerase or RdRp) whereas ORF1b harbors four putative nsp's namely nsp13 (Helicase), nsp14 (3′-5′ exoribonuclease), nsp15 and nsp16 (mRNA cap-1 methyltransferase)[39]. Substitutions in the ORF1ab gene may increase the synthesis of replicase-transcriptase complex, thus, increasing the replication rate of the virus and blocking the host innate-immune response. 614 position in spike glycoprotein lies near the S1/S2 subunit junction where the furin-cleavage site is present (R667) that enhances virion cell-cell fusion[40]. This suggests, aspartate to glycine substitution in the vicinity of the furin-recognition site may result in a conformational change of the spike glycoprotein that favors higher affinity of the Receptor Binding Domain (RBD) to hACE2. A recent article showed retroviruses pseudotyped with Glycine at 614 position infected ACE2-expressing cells markedly more efficiently than those with Aspartic acid due to less S1 shedding and greater incorporation of the S protein into the pseudovirion[41]. It has also been previously reported D614G mutation is increasing at an alarming rate[42,43]. Also, Korber et al. observed that this alteration correlated with increased viral loads in COVID-19 patients[42]. This is consistent with the epidemiological data showing proportion of viruses bearing G614 is correlated to increased case fatality rate on a country by country basis[44] . This substitution coevolved with substitution in the leader sequence, nsp3 and RdRp proteins, suggesting these mutations allow the virus to

transmit more efficiently. This explains these mutations have not emerged merely because of founder's effect but this virus under selection pressure has made itself more stable and infective. We propose climate as one of the selective pressures influencing SARS CoV-2 evolution. Membrane (M) and Nucleocapsid (N) proteins are crucial virus structural proteins. Virus cluster SNPs in ORF1ab, ORF3a, ORF8, S, M and N genes occurring in different climatic zones and Koppen's climate implicates environmental factors of each climatic zones influence SARS CoV-2 genotype which in turn enables them to propagate in their respective climate zone from which they were isolated. However, the actual role of these mutations is not yet determined, and these studies need to be further enlightened by biophysical and biochemical studies. Such mutational insights will aid the design of efficacious vaccines.

To delineate the underlying abiotic factors (temperature, precipitation and latitude) responsible for evolution of SARS CoV-2 (n=176), spreading patterns of G1 and G2 strains were carefully examined on Koppen's map. An elevated spread of COVID-19 in the western and eastern coasts of the continents and a diminished spread in the hot and cold deserts all over the world was observed. The outbreak of COVID-19 was initiated from Wuhan, southeast China. The climate of southeast China is unique in itself, the entire region consists of *"humid-subtropical" (Cfa)* climate type of Koppen's C climate, commonly known as east-coast climate explaining why nearly all strains from China belongs to G1 (Yellow circle in Figure 4) and fall within the boundaries of *"humid-subtropical" (Cfa)* climate. Interestingly, on the Koppen's map this climate is also present in the eastern coasts of Australia, USA, and South America where mainly the G1 strains are distributed (Figure 4). According to the map (Figure 4), the G1 strains are also

distributed in the western coast of North America, Europe and South America where *"marine-temperate" (Cfb)* climate dominates, commonly known as west-coast climate, a climate very similar to *'Cfa'* (Wuhan type climate), the only difference is the hottest month in '*Cfa'* climate is > 22 ^{o}C and, in Cfb is <22 ^{o}C[2], one small cluster of G1 is present in the surrounding regions of Mediterranean sea, Caspian sea, Black sea and Turkey, this region consists of *"Mediterranean" (Csa-Csb)* climate, which is similar to '*Cfa'* and *'Cfb'* climate with more rainfall. Statistically, distribution of G1 strains is in concordance with the distribution of temperate climate and strongly favor C climate (Chi-square test, P<.001) as compared to any other climate. Phylogenetic analysis combined with distribution pattern on Koppen's map reveal influence of C climate in shaping the molecular determinants of G1 variant. This limited climatic range has plausibly restricted G1 population in temperate climate. Across the globe the C climate generally grades into "*cold-desert" (BSk)* or *"humid-continental*" *(Dfa-Dfb)* climate. In the western coast of Europe '*Cfb'* climate grades into *"humid continental*" *(Dfa-Dfb)* climate and this major change in Koppen's climate from C to D seems to drive the evolution of G1 to G2 variant group. Our analysis shows the fall of average annual temperature from C to D climate restricts the ancestor G1 strains in C climate and mutated G2 strains were able to cross the climatic boundaries. In Europe, the gradation of '*Cfb'* to '*Dfa-Dfb'* lie over Germany, towards the west of Germany *'Cfb'* climate prevails which grades eastwards into *Dfa-Dfb* climatic belt, following the southern boundary of Russia and finally grades into *"Dry winter-subarctic" (Dwc)* climate, relatively colder climate than (Dfa-Dfb) and terminates towards the eastern coast of Russia (Primorsky Krai). In Russia most (21/23) of the strains belong to G2, only limited strains belong to G1 (2/23), suggesting G2 strains have adapted to the

D climate (Chi-square test, *P*<.001). All G2 strains of Russia are present in the long belt of '*Dfa-Dfb-Dw'* climate. Similar observations are seen in the western and eastern coast of North America. In the eastern coast of USA (Florida, Georgia, Alabama, North & South Carolina, Virginia etc), '*Cfa*' climate (Wuhan type) prevails which grades into the '*Dfb'* climate, this belt extends from New York to Canada. The G2 strains enter the continent's interior along this *'Dfa-Dfb'* climate belt, running from eastern coast of USA, along the USA-Canada boundary towards the center of Canada (Figure 4). The C climate (*Cfa-Cfb-Csb*) is present all along the western coast of USA which also grades into D climate towards its east, implicating the presence of G2 strains in both C and D climate, contrary to G1 strains that are habituated to C climate. Absence of this virus strains in a cold desert existing between the D climate of western coast and eastern coast, support the preferential spread of COVID-19 in C and D climate (Figure 4). A similar pattern is observed for the occurrence of G1 strains of Canada (2/9) and Alaska (1/4) in C climate and G2 strains in D climate. Our analysis show G2 strains have adapted for both C and D climate type (Chi-square test, *P*<.001). G1 and G2 strain are highly stable in "*humid-subtropical" (Cfa)* and *"marine-temperate" (Cfb)* climate type of C climate which are eastern and western coastal climate that explains the reason for the maximum spread of COVID-19 in coastal regions of the temperate zone. Two C climate belts which were previously unaffected because these belts were surrounded by A and B climate, one of the belt lie in Africa stretching from Zambia-Zimbabwe-Mozambique to Anglo and another stretching all along the foothills of Himalayas, stretching from Kolkata-Bihar-Uttar Pradesh-Delhi-Pakistan to Afghanistan are highly vulnerable to the spread of COVID-19.

In South America, mostly the strains are distributed in the eastern coast as it contains a large region of *'Cfa'* climate (Wuhan type), lying around Sao-Paulo region of Brazil grading into cold desert in the south which restricts further spread towards south. However, towards north of Brazil, the *'Cfa'* climate grades into "*tropical-savannah" (Aw)* climate, the G2 strains (clusters, 80-115 and 126-176) spread towards north from *'Cfa'* to *'Aw'* climate. In the entire world, South America is the only continent connecting C climate to all A climate type (*Af, Am, and Aw*), providing a land bridge between C and A climate. The strains of the South America are a threat to all the countries having '*Aw'* climate type (South Africa, India, Madagascar, Thailand, north-east Australia, Mexico, and South Asian Islands). From the trend observed in the other climate types, it is reasonable to assume that soon SARS CoV-2 will adapt to all A climate types. In the present study, very few strains are located in the B climate (desert), most probably have arrived due to travel. It is possible through the '*Aw'* climate of Africa, a missing link of A-B climate will be established, as a result of which, descendants of G2 might extend their climatic territories into B climate leading to further spread in desert climate. The phylogenetic and climatic data suggest that SARS CoV-2 undergoes genetic modifications to thrive in the external climatic conditions. It is highly possible that the strains mutated to adapt in tropical region will be highly stable in the C and D climate. Studies combining genetic information and distribution pattern with COVID-19 severity are extremely important during a pandemic as it provides information about evolution and the pattern of spread of the viruses in the initial stage. Such information can be used to take pre-emptive measures to significantly reduce the spread of the disease across the globe.

Conclusion

It is reasonable to assume COVID-19 transmission pathway and evolution is influenced by climate. Phylogenetic network classified 176 SARS CoV-2 strains into two variant groups G1 and G2. The G1 strains were habituated to C climate that evolved into G2 by undergoing significant mutations (C241T in leader sequence, F924 in ORF1a, P214L in ORF1b and D614G in S gene), plausibly extended its climatic boundaries from C to D climate, displaying role of natural selection on virus evolution. Mostly SARS CoV-2 has been resistive to exist in B climate. Gradually, strains are adapting to A climate in South America. The strains adapted to *"tropical-savannah" (Aw)* climate are a threat to all the tropical countries, which were initially less affected by COVID-19. The study agrees that viruses are sensitive to their environment and respond towards naturally occurring abiotic factors such as temperature, latitude and humidity to sustain in different climate of the Earth, which also provides insights about seasonal variations possibly being a strong reason for the spread of other viral diseases as well. Here we showed a more refined description of genes based on phylogenetics and their distribution across different climate zones. This finer-grained analysis led to highly relevant insights on evolutionary dynamics of poorly understood SARS CoV-2 genome and provides vital information about the direction of the spread and highlights vulnerable regions of Earth. Such inter-disciplinary studies play an imperative role in designing antiviral strategies and taking pre-emptive precautionary measures to combat COVID-19.

Methodology

Distribution of SARS COV-2 across latitudes

The COVID-19 data is obtained from the 'worldometer' website, a trusted source of COVID-19 database which provides global COVID-19 live statistics[23]. 'Active cases per million population' for different countries were analyzed (assessed on 25 April 2020). To check the latitudinal preference of SARS CoV-2, the countries of the world were segregated based on their latitudes and per million COVID-19 cases were plotted between the latitudes (90°N to 66.5°N), (66.5°N to 23.5°N), (23.5°N to 23.5°S), (23.5°S to 66.5°S), and (66.5°S to 90°S). North Temperate Zone was further divided in an interval of 7° latitude. Distribution of SARS CoV-2 between these latitudes was analyzed and compared.

Molecular phylogenetic analysis

185 full-length SARS CoV-2 genomic sequences from countries across the globe, with genome length more than 29 kb and high coverage were obtained from Global Initiative on Sharing Avian Influenza Data (GISAID) database, accessed till 2 May 2020 and the reference genome was retrieved from GenBank[24] (Table S1). To avoid bias related to the geographical area covered by a country, genomic sequence of strains isolated from different locations from each country was retrieved, depending on the availability of data. The sequences were aligned to the full reference genome[24] by using Biomanager and Seqinr packages of R (version 3.6.3). Among 185 genomes, some partial genomes were discarded. NC_045512 genome sequence was used as reference and the genomic coordinate in this study is based on this reference genome. Based on protein annotations, nucleotide level variants were converted into amino acid codon variants for alignments

when its location within a gene was identified. The amino acid position numbering is according to its position within the specified gene (CDS) as annotated in reference sequence (NC_045512, NCBI)[24]. To ensure comparability, we trimmed the flanks of all sequences. The aligned sequences were used to construct a phylogenetic tree using MEGA X[25]. The evolutionary history was inferred using the Neighbor-Joining method (500 bootstrap tests)[26]. The optimal tree with the sum of branch length = 0.01116462 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method[27] and are in the units of the number of base substitutions per site. All ambiguous positions were removed for each sequence pair (pairwise deletion option). A total of 29408 positions were present in the final dataset. The results are presented in the form of DNA sequencing i.e., U (uracil) is read as T (thymine). We have labeled each virus strain by the GISAID Accession ID and the location from which it was isolated in the format “Location|EPI ISL Accession ID”, in the constructed phylogenetic tree. For ease of visualization, we have marked a new Strain ID (1 to 176) against each SARS CoV-2 isolate in the phylogenetic tree (Figure 2). The same Strain ID is used for the climatic studies in this article. High-frequency SNPs (Single Nucleotide Polymorphisms) distinguishing one virus cluster from the others is referred to as “virus cluster SNPs” throughout this paper.

Mapping virus strain on the Koppen’s climate map

The location of each SARS CoV-2 strain is obtained from the METADATA file provided in GISAID database for each viral isolate (Table S1). The coordinates of the locations were taken from the official website of USGS Earth Explorer[28]. The Gieger-Koppen’s climate

map is used for climatic studies[5]. The Koppen climate type, temperature, precipitation of each strain is assessed from weatherbase[29] and CLIMATE.ORG[30]. The map is georeferenced by using 'Arc-GIS 10.1'[31]. The locations of all strains (n=176) were transferred to the georeferenced map[31]. On the map, the G1 strains were symbolized as 'Yellow-circle', and G2 as 'Square' (Figure 4). Each strain in the map is labelled as per their Strain ID (1 to 176) (Figure 4), the map combines information of the phylogeny, climate, and global distribution of SARS CoV-2. These locations were classified into coastal and continental region, we define the coastal region as land region < 500 km from the ocean/sea and the continental region as land lying >500 km from the coastline measured through google maps.

Statistical analysis

Two-tailed paired t-test and Chi-square test were performed in Microsoft Excel (2016) to test null hypothesis H1, H2, H3 and H4 related to latitudinal preference (H1), climatic preference (H2 and H3) and regional preference (H4) of SARS CoV-2. H1: SARS CoV-2 follows latitudinal biodiversity gradient. H2: Majority of G1 strains do not lie in temperate climate (C). H3: Majority of G2 strains do not fall in temperate (C) and cold (D) climate. H4: The virus isolates are equally distributed in coastal and continental region. Histograms depicting the distribution of coronavirus in coastal region, continental region, Koppen's climate and climate type were plotted using R (version 3.6.3). SigmaPlot10 was used to generate box plot, regression plot, and mesh plot to statistically compare frequency distribution of latitude, temperature and precipitation of G1 and G2 strains. We performed one-way ANOVA to estimate statistical differences in the latitude, temperature and precipitation between G1 and G2 virus populations. Various scatterplots between

latitude, temperature and precipitation of G1 and G2 strains were plotted in R (version 3.6.3). Values were considered statistically significant for *P* values below 0.05. Exact *P* values are provided in appropriate figures.

Potential caveats

We acknowledge several caveats about our analyses. Our data from the tropics is limited because at the time of data collection (SARS CoV-2 strains) from all over the world, the strains from the tropical countries were very limited, from few tropical regions strains were available (e.g., Ghana (Africa); India, Mexico, Nepal, Pakistan) but the data has been discarded due to the travel history of the strains, the remaining strains without travel history have large gaps in genomic sequences which were not suitable for the present study. Also, case history of each patient is not reported in the METADATA file as collecting all information from each patient is time-consuming. Hence, there are chances patients from whom these strains were isolated may have a migratory history. All the data from different individual locations without travel history and large gaps in genomic sequences have been incorporated. To overcome this, the inverse latitude gradients were studied based on the total number of COVID-19 cases all over the globe.

Data availability

Refer Supplementary Table S1-S4.

Code availability

The code is available from the corresponding author on request.

Acknowledgement

We gratefully acknowledge the authors and originating and submitting laboratories of the sequences from GISAID's EpiFlu (TM) Database on which this research is based. A table of the contributors is available in, Supplementary Table S1. We thank Prof. Raghavan Varadarajan, Prof. Raman Sukumar, Dr. Teena Jangid and Chetankumar Jalihal of Indian Institute of Science for proofreading the article.

Conflict of Interest

Authors declare no conflict of interest.

References

1. WHO. *Coronavirus disease (COVID-19)*. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200601-covid-19-sitrep-133.pdf?sfvrsn=9a56f2ac_4 (2020).

2. Sajadi, M. *et al.* Temperature and Latitude Analysis to Predict Potential Spread and Seasonality for COVID-19. *SSRN Electron. J.* (2020) doi:10.2139/ssrn.3550308.

3. Ficetola, G. F. & Rubolini, D. Climate affects global patterns of COVID-19 early outbreak dynamics. *medRxiv* 2020.03.23.20040501 (2020) doi:10.1101/2020.03.23.20040501.

4. Allaby, M. *ATMOSPHERE: A Scientifi c History of Air, Weather, and Climate*. *Journal of Chemical Information and Modeling* vol. 53 (2017).

5. Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* **11**, 1633–1644 (2007).

6. Pianka, E. R. Latitudinal Gradients in Species Diversity: A Review of Concepts. *Am. Nat.* **100**, 33–46 (1966).

7. Rohde, K. Latitudinal gradients in species diversity and Rapoport's rule revisited: A review of recent work and what can parasites teach us about the causes of the gradients? *Ecography (Cop.).* **22**, 593–613 (1999).

8. Guernier, V., Hochberg, M. E. & Guégan, J. F. Ecology drives the worldwide distribution of human diseases. *PLoS Biol.* **2**, 740–746 (2004).

9. Lawrence, E. R. & Fraser, D. J. Latitudinal biodiversity gradients at three levels: Linking species richness, population richness and genetic diversity. *Glob. Ecol. Biogeogr.* 770–788 (2020) doi:10.1111/geb.13075.

10. Rohde, K. Nordic Society Oikos Latitudinal Gradients in Species Diversity : The Search for the Primary Cause Author (s): Klaus Rohde Published by : Wiley on behalf of Nordic Society Oikos Stable URL : http://www.jstor.org/stable/3545569 Accessed : 09-05-2016 20 : **65**, 514–527 (1992).

11. Kouki, J. *et al.* Finnish Zoological and Botanical Publishing Board Reversed latitudinal gradient in species richness of sawflies (Hymenoptera , Symphyta) Reversed latitudinal gradient in species richnes sawflies (Hymenoptera , Symphyta). **31**, 83–88 (1993).

12. Kindlmann, P., Schödelbauerová, I. & Dixon, A. F. G. Inverse latitudinal gradients in species diversity. in *Scaling Biodiversity* (eds. Storch, D., Marquet, P. & Brown, J.) 246–257 (Cambridge University Press, 2012). doi:10.1017/CBO9780511814938.014.

13. Fehr, A. R. & Perlman, S. Coronaviruses: An overview of their replication and pathogenesis. in *Coronaviruses: Methods and Protocols* 1–23 (2015). doi:10.1007/978-1-4939-2438-7_1.

14. Li, C. K. *et al.* T Cell Responses to Whole SARS Coronavirus in Humans. *J. Immunol.* **181**, 5490–5500 (2008).

15. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structural biology: Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science (80-.).* **309**, 1864–1868 (2005).

16. Liu, J. *et al.* The Membrane Protein of Severe Acute Respiratory Syndrome Coronavirus Acts as a Dominant Immunogen Revealed by a Clustering Region of Novel Functionally and Structurally Defined Cytotoxic T-Lymphocyte Epitopes. *J. Infect. Dis.* **202**, 1171–1180 (2010).

17. Ruch, T. R. & Machamer, C. E. The coronavirus E protein: Assembly and beyond. *Viruses* **4**, 363–382 (2012).

18. Chang, C. K., Hou, M. H., Chang, C. F., Hsiao, C. D. & Huang, T. H. The SARS coronavirus nucleocapsid protein - Forms and functions. *Antiviral Research* (2014) doi:10.1016/j.antiviral.2013.12.009.

19. Drexler, J. F. *et al.* Genomic Characterization of Severe Acute Respiratory Syndrome-Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-Dependent RNA Polymerase Gene Sequences. *J. Virol.* **84**, 11336–11349 (2010).

20. Sanjuán, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cellular and Molecular Life Sciences* (2016) doi:10.1007/s00018-016-2299-6.

21. Lauring, A. S. & Andino, R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog.* **6**, 1–8 (2010).

22. Racevska, E. Natural Selection. in *Encyclopedia of Animal Cognition and Behavior* (eds. Vonk, J. & Shackelford, T.) 1–14 (Springer International Publishing, 2018). doi:10.1007/978-3-319-47829-6_542-1.

23. Coronavirus Update (Live): 8,522,724 Cases and 453,714 Deaths from COVID-19 Virus Pandemic - Worldometer. https://www.worldometers.info/coronavirus/.

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

25. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

26. Nei, M. & Saitou, N. The neighbor-joining method: a new method for reco... [Mol Biol Evol. 1987] - PubMed result. *Mol Biol Evol* 406–425 (1987).

27. Tamura, K., Nei, M. & Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11030–11035 (2004).

28. EarthExplorer. https://earthexplorer.usgs.gov/.

29. Travel Weather Averages (Weatherbase). https://www.weatherbase.com/.

30. Climate data for cities worldwide - Climate-Data.org. https://en.climate-data.org/.

31. Herbei, M., Ciolac, V., Smuleac, A. & Ciolac, L. Georeferencing of Topographical Maps Using the Software ArcGIS. *Res. J. Agric. Sci.* **42**, 595–606 (2010).

32. Harcourt, B. H. *et al.* Identification of Severe Acute Respiratory Syndrome Coronavirus Replicase Products and Characterization of Papain-Like Protease Activity. *J. Virol.* (2004) doi:10.1128/jvi.78.24.13600-13612.2004.

33. Serrano, P. *et al.* Nuclear Magnetic Resonance Structure of the Nucleic Acid-Binding Domain of Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 3. *J. Virol.* **83**, 12998–13008 (2009).

34. Snijder, E. J., Decroly, E. & Ziebuhr, J. *The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing*. *Advances in Virus Research* vol. 96 (Elsevier Inc., 2016).

35. Angov, E. Codon usage: Nature's roadmap to expression and folding of proteins. *Biotechnology Journal* vol. 6 650–659 (2011).

36. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529**, 358–363 (2016).

37. Tsai, C. J. *et al.* Synonymous Mutations and Ribosome Stalling Can Lead to Altered Folding Pathways and Distinct Minima. *Journal of Molecular Biology* vol. 383 281–291 (2008).

38. Shabalina, S. A., Ogurtsov, A. Y. & Spiridonov, N. A. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res.* (2006) doi:10.1093/nar/gkl287.

39. Ren, L.-L. *et al.* Identification of a novel coronavirus causing severe pneumonia in human. *Chin. Med. J. (Engl).* **133**, 1015–1024 (2020).

40. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358–369 (2006).

41. Zhang, L. *et al.* The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. *bioRxiv* 2020.06.12.148726 (2020) doi:10.1101/2020.06.12.148726.

42. Korber, B. *et al.* Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* (2020) doi:10.1101/2020.04.29.069054.

43. Junior, I. J. M. *et al.* The global population of SARS-CoV-2 is composed of six major subtypes. *bioRxiv* (2020) doi:10.1101/2020.04.14.040782.

44. Becerra-Flores, M. & Cardozo, T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int. J. Clin. Pract.* (2020) doi:10.1111/ijcp.13525.

TABLE AND FIGURES

Table 1: SNPs representing virus cluster and their distribution across varied climates.

Virus cluster	Nucleotide mutation	Amino acid mutation	Gene	LAT	KCT	KC
1-22	C8782T	S2839	ORF1a	NTZ	Cfa	C
	T28144C	L84S	ORF8			
5-6	C29095T	F274	N	NTZ	Cfa	C
8-9	T9477A	F3071Y	ORF1a	NTZ, TZ	Mix	C-A
	G25979T	G196V	ORF3a			
	C28657T	D128	N			
	C28863T	S197L	N			
10-17	C18060T	L1431	ORF1b	NTZ	Cfa-Cfb	C
12-17	A17858G	Y1364C	ORF1b	NTZ	Cfa-Cfb	C
13-17	C17747T	P1327L	ORF1b	NTZ	Cfa	C
20-22	C24034T	N824	S	NTZ	Cfa	C
	T26729C	A69	M			
	G28077C	V62L	ORF8			
21-22	T490A	D75	ORF1a	NTZ, NTSZ	Cfa	C
	C3177T	P971L	ORF1a			
	T18736C	F1657L	ORF1b			
23-25	C6312A	T2016K	ORF1a	NTZ, TZ, SSTZ	Mix	Mix
	C13730T	L4489	ORF1a			
	C23929T	Y789	S			
	C28311T	P13L	N			
28-32	G1397A	D392G	ORF1a	NTZ	Mix	Mix
	T28688C	L139	N			
33-43	G26144T	G251V	ORF3a	NTZ	Cfa-Cfb	C
37-39	A2480G	I739V	ORF1a	NTZ	Mix	Mix
	C2558T	P765S	ORF1a			
37-43	C14805T	Y346	ORF1b	NTZ	Cfa	C
42-43	T17247C	R1160	ORF1b	NTZ	Cfb	B
44-47	G1440A	G392D	ORF1a	NTZ	Cfb	C
	G2891A	A876T	ORF1a			
58-61	C15324T	N519	ORF1b	NTZ	Cfa-Dfb	C-D
59-176	C3037T	F924	ORF1a	NTZ	Cfa-Cfb-Dfb-Aw	C-D -A
	A23403G	D614G	S			
	C14408T	P214L	ORF1b			
59-125,	C241T		Leader	NTZ	Cfa-Cfb-Dfa-Dfb	C-D-A
127-176	C241T		seq.			C-D
66-68	A26530G	D3G	M	NTZ	Cfc-Dfb	C-D
70-71	G4201T	M1312I	ORF1a	NTZ	Cfa-Dwc	C-D
	C26527T	A2V	M			
80-115	G28881A	R203K	N	NTZ	Cfa-Cfb-Dfb-Aw	C-D-A

	G28882A	R203K	N			
	G28883C	G204R	N			
86-87	C27046T	T175M	M	NTZ	Cfa-Dfb	C-D
88-89	C3373A	D1036E	ORF1a	NTZ	Dfb-Cfb	C-D
105-107	T29148C	I292T	N	TZ, STZ	Cfa-Aw	C-A
106-107	A27299C	I33T	ORF6	NTZ, TZ	Cfa-Aw	C-A
108-111	C313T	L16	ORF1a	NTZ, TZ	Cfa-Cfb-Aw	C
113-115	C4002T	T1246I	ORF1a	STZ	Cfa-Cfb-Am	C-A
	G10097A	G3278S	ORF1a			
	C13536T	T4424I	ORF1a			
	C23731T	T723	S			
116-125	A20268G	L2167	ORF1b	NTZ	Cfa-Cfb-Dfa-Dfb	C-D
126-176	G25563T	Q57H	ORF3a	NTZ	Cfa-Cfb-Dfa-Dfb	C-D
126-130	C18877T	L1704	ORF1b	NTZ	Cfa-Dfa-Dcb	C
131-135	C2416T	Y717	ORF1a	NTZ	Cfa-Dfa-Aw	D
136-176	C1059T	T265I	ORF1a	NTZ	Cfa-Cfb-Dfa-Dfb	C-D
138-139	C18998T	A1744V	ORF1b	NTZ, TZ	Cfa-Am	C-A
	G29540A					
138-141	C11916T	S3884L	ORF1a	NTZ	Cfa-Csb-Am	C
143-147	C27964T	S24L	ORF8	NTZ	Cfa-Cfb-Dfa-Dfb	C-D
148-149	C11224T	V3653	ORF1a	NTZ	Dfa-Dfb	D
157-159	G29553A			NTZ	Cfa	C

NOTE: Virus clusters are named by Strain ID as depicted on the phylogenetic tree. Virus clusters are named by Strain ID as depicted on the tree. Genomic coordinates in this study is based on reference genome[24]. The SNP mutation positions are based on the reference genome. Nucleotide T represents nucleotide U in the SARS-CoV-2 RNA genome. Mutation at the protein level is not mentioned for the SNPs arising in the non-coding region. The amino acid position numbering is according to its position within the specified gene (CDS). In LATs column we have mentioned the major climate zone for the corresponding virus cluster[4]. KCT is Koppen's Climate Type and KC is Koppen's Climate columns display the main Koppen's climate in which the virus isolates of the corresponding virus cluster lie. 'Mix' implies no particular climate type is favored[5].

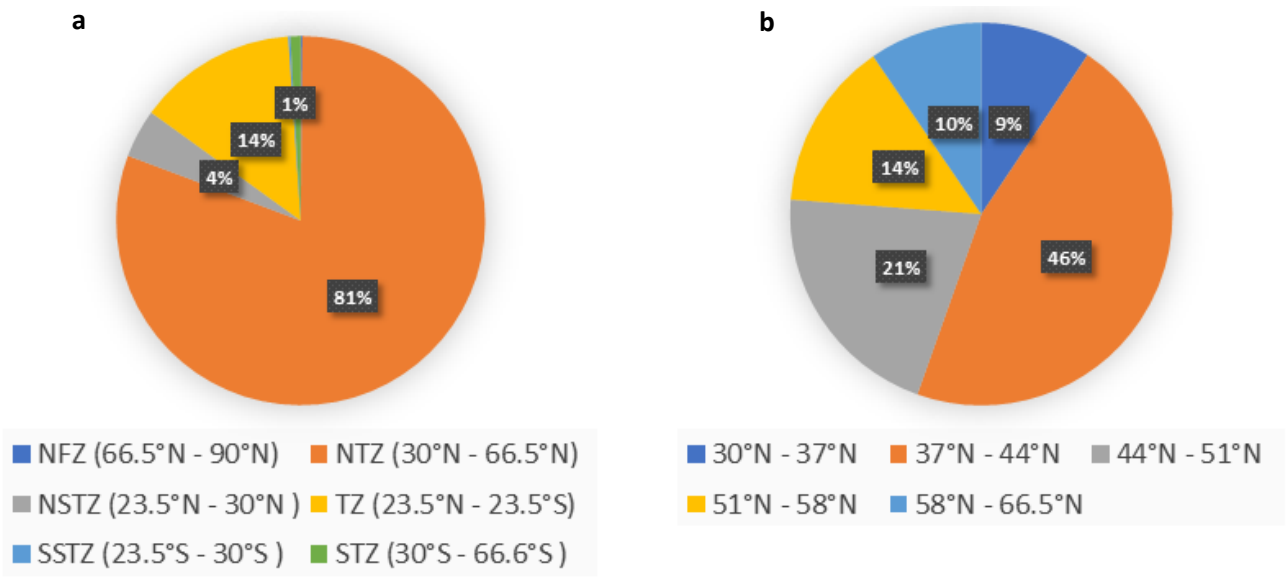


Figure 1: Distribution of COVID-19 cases across different climate zones based on latitudes[5]. (a) Area of the pie-chart covered by a climate zone is proportional to the percentage of COVID-19 cases (https://www.worldometers.info/coronavirus/) occurring in their respective climate zones as depicted by black squares. The percentage of COVID-19 cases for NFZ and SSTZ is extremely low, therefore, it is not mentioned in the pie-chart. (b) The North Temperate Zone is divided into an interval of 7° latitude. The area of the pie-chart covered is directly proportional to the percentage of COVID-19 cases occurring in their respective latitude range as depicted by black squares.

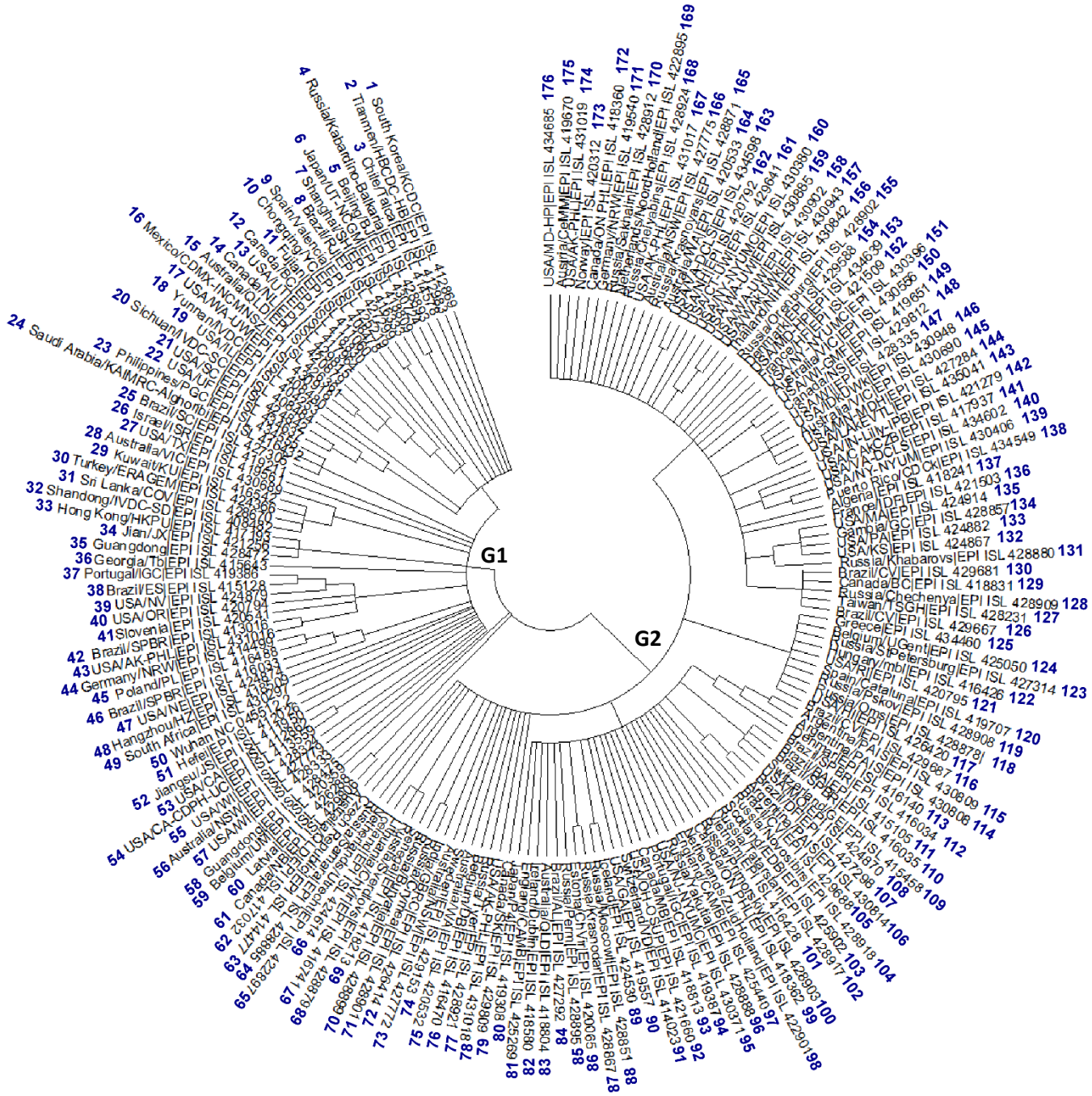
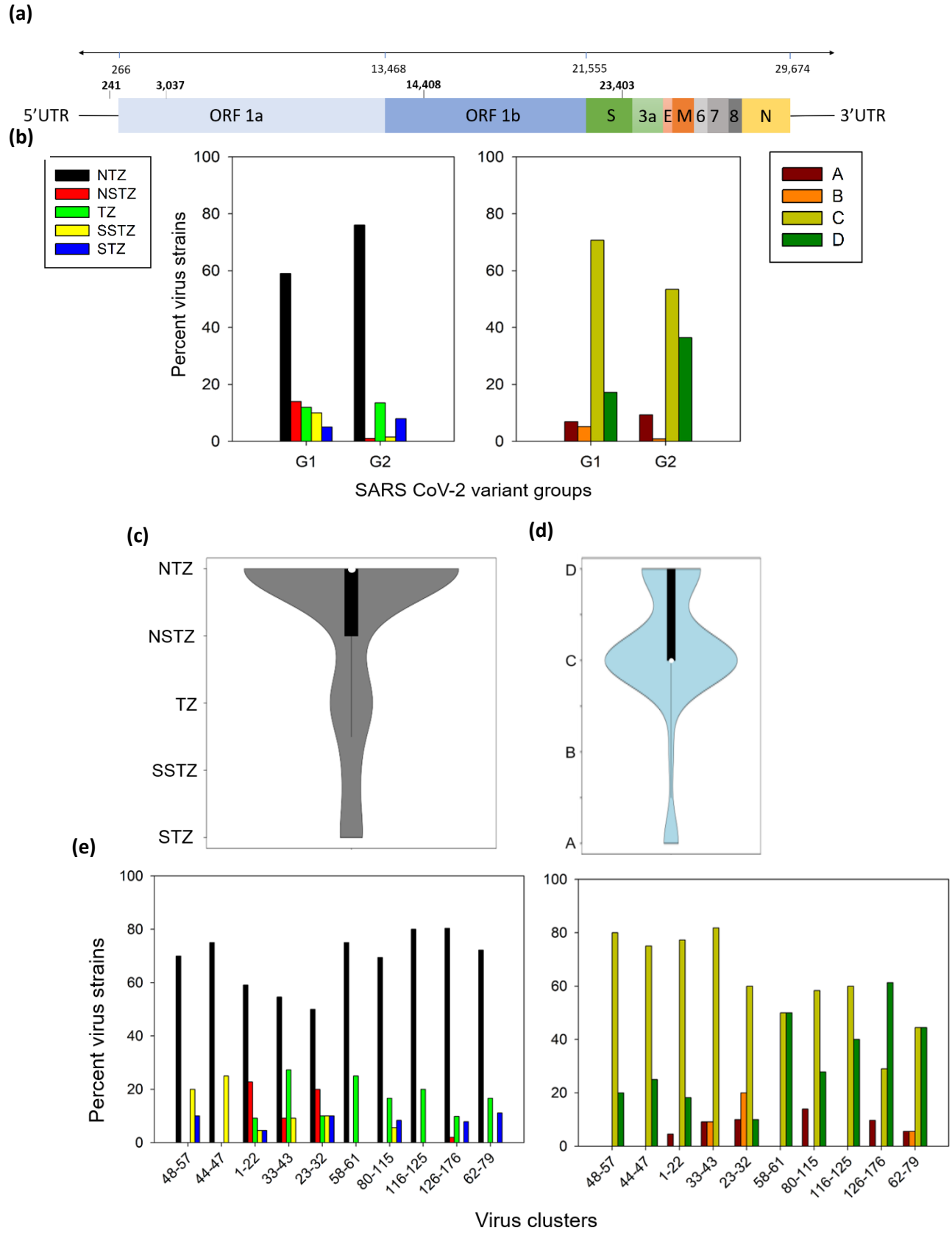


Figure 2: Phylogenetic network divides 176 SARS CoV-2 strains into two variant groups. Largely, the left side of the tree (1 to 58) constitute the G1 group and the right side of the tree constitutes the G2 group (59 to 176). Branch length is proportional to the genomic relatedness of the viral isolates. Closely related virus isolates comprise the same SNP with respect to the reference genome (Strain ID: 50) and form a cluster. The evolutionary history of 176 taxa was inferred using the Neighbor-Joining method[24] (500 bootstrap tests). A total of 29408 positions were analysed, with nucleotide position numbering according to the reference sequence[24].


(a)
266
13,468
21,555
29,674
241
3,037
14,408
23,403
5'UTR
ORF 1a
ORF 1b
S
3a
E
M
6
7
8
N
3'UTR
(b)
NTZ
NSTZ
TZ
SSTZ
STZ
Percent virus strains
G1
G2
SARS CoV-2 variant groups
A
B
C
D
(c)
(d)
(e)
48-57
44-47
1-22
33-43
23-32
58-61
80-115
116-125
126-176
62-79
Virus clusters

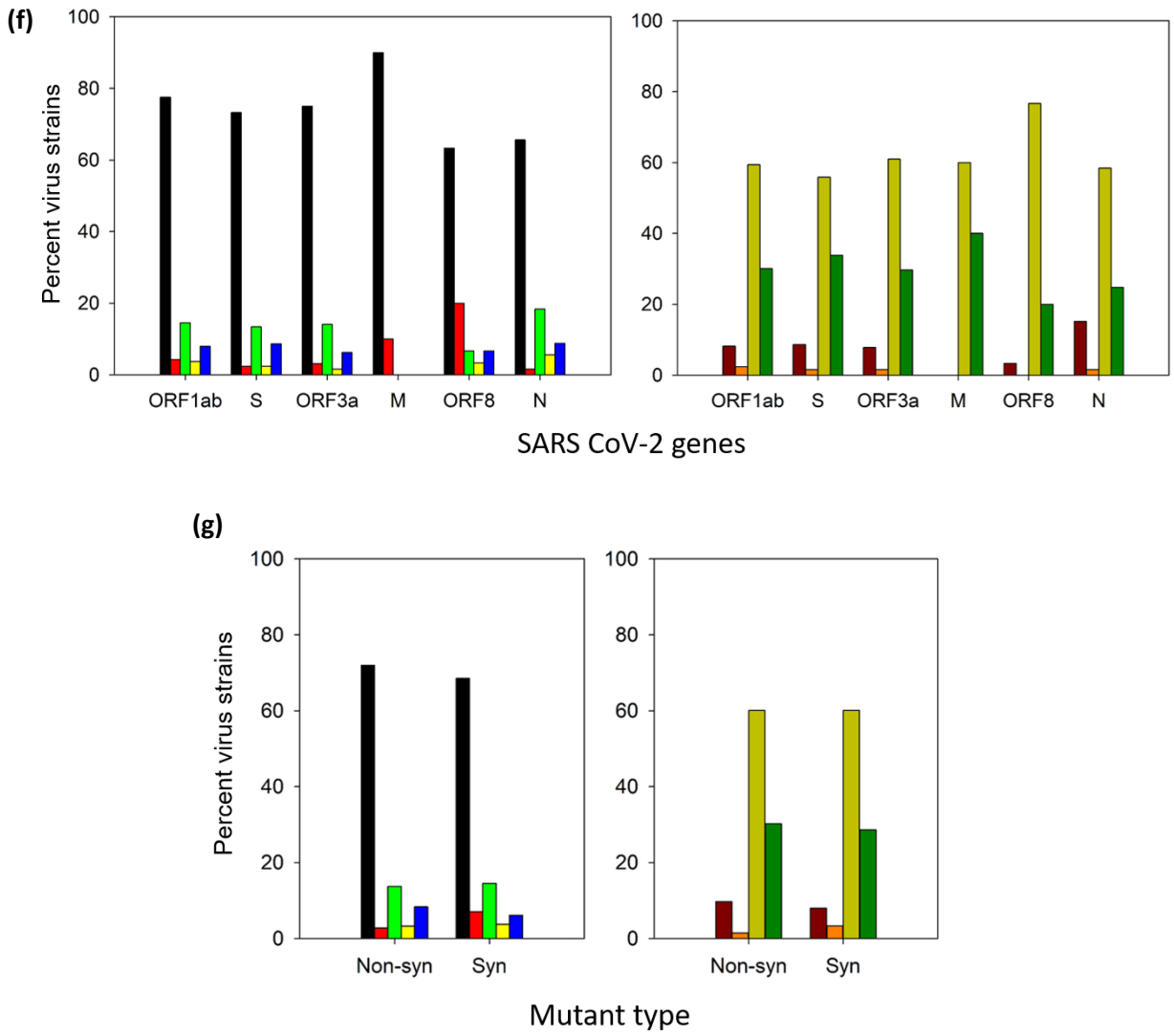


Figure 3: Molecular phylogeny analysis to infer genomic similarities of SARS CoV-2 and their distribution across different climate zones[4] and Koppen's climate types[5]. (a) Genomic architecture of SARS CoV-2 genome highlighting four positions, substitutions on these positions enabled evolution of G1 into G2. (b, e-g) Strains found within a virus cluster (as shown in the phylogenetic tree and mentioned in Table 1) were analysed for significant mutations that may have arisen due to climatic pressure. Hence, percentage of such virus strains is plotted according to the geographical location of the climate zone from where they were isolated. The height of the bar is proportional to percent virus strain occurring in the specified condition i.e., labelled on the x-axis. Box in the left panel consist of color code for each climate zone and box in the right panel consist of color code for Koppen's climate. Left panel shows distribution of percent virus strains in different climate zones and right panel shows distribution of percent virus strain in Koppen's climate (b) Percent virus strains prevailing in different climate zones, stratified by SARS CoV-2 variant groups. Width of curves of violin plot is proportional to the number of SARS CoV-2 strains (n=176) in varied (c) climate zones and (d) Koppen's climate. (e) Abiotic factors influencing evolutionary dynamics of phylogenetic virus clusters. (f) Percent of virus strains with high frequency SNPs in each gene. (g) Type of mutation i.e. non-synonymous or synonymous exhibited by viruses.

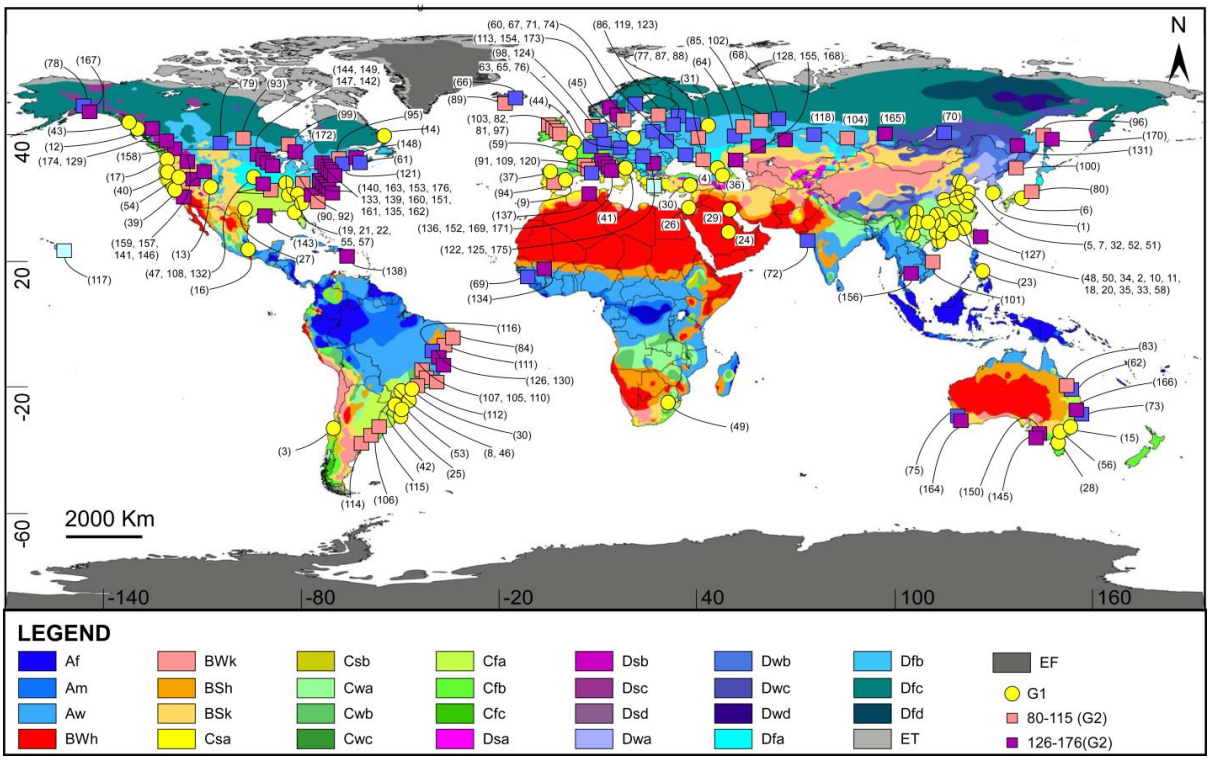


Figure 4: Global distribution of SARS CoV-2 strains on the Gieger-Koppen's map displaying different climate types[5]. Each strain is labelled as per the strain ID (1 to 176) within parenthesis. The G1 strains were symbolized as 'Yellow-circle', and G2 as 'Square', pink square denotes strain clusters (80-115) stable across C, D and A climate, purple square represents strain cluster (126-176) stable majorly in D climate, the remaining G2 strains (blue squares) are stable across C and D climate. Standard Koppen's climate-type symbols are mentioned in the legend, the criteria for distinguishing these climate types is mentioned in Table S3. Table S4 contains full form of these symbols. All symbols with initials 'A' (Af, Am, Aw) are of tropical climate, initials with 'B' belong to desert climate, 'C' to temperate and 'D' to cold and 'E' to polar climate. The shades of blue on the map, in North America and Russia belongs to D climate. Shades of yellow and green belongs to C climate, shades of red, orange and pink belongs to Desert climate.

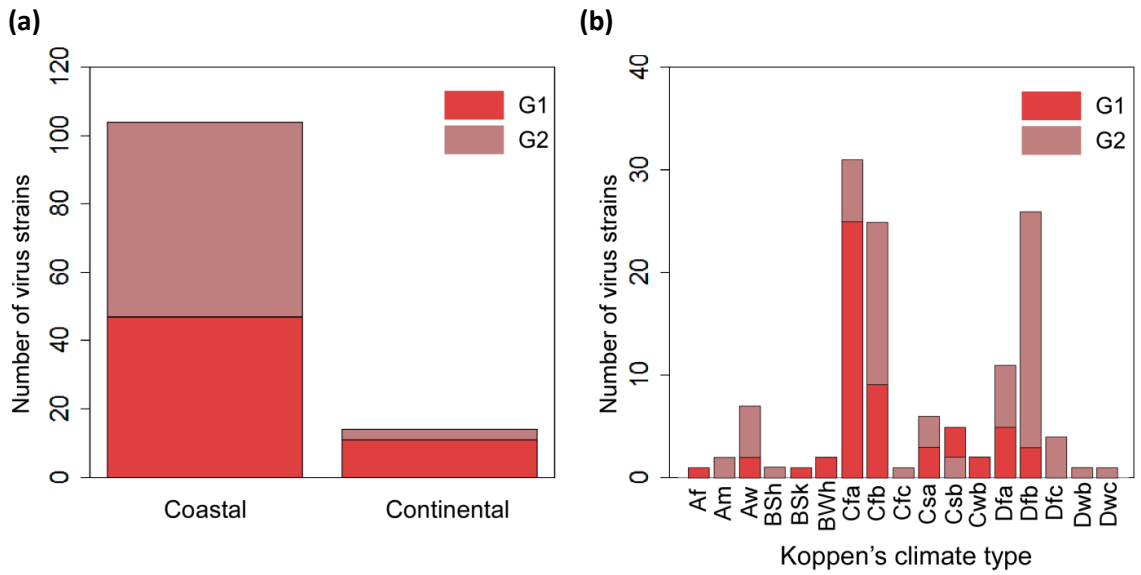


Figure 5: Global distribution of SARS CoV-2 strains (n=176) (a) in the coastal and continental region (b) and in different Koppen's climate types[5]. Number of virus strains in G1 population is represented by maroon color and of virus strains in G2 population is represented by grey color.

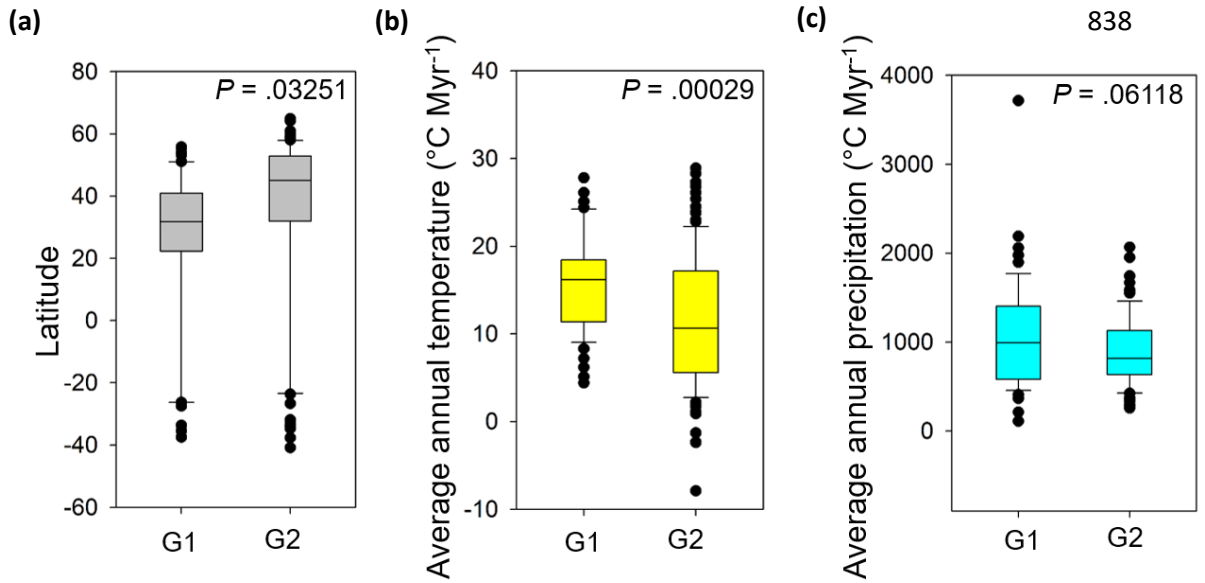


Figure 6: Comparative analysis of different climatic parameters such as latitude, temperature and precipitation between G1 and G2 variant groups. (a) Positive values represent the latitude range falling in Northern Hemisphere and negative values represent latitude range falling in Southern Hemisphere. The G2 strains preferentially occur towards the higher latitudes than G1 (P = .032; 95% CI 17.12-31.12 for G1; 95% CI 28.67-68.06 for G2). (b) The mean annual temperature of G2 is significantly lower than the G1 strains (P<.001; 95% CI 17.32-14.32 for G1; 95% CI 13.02-10.33 for G2) (c) Mean annual precipitation of G1 and G2 strains is nearly same (P=.061; 95% CI 1207.16-886.75 for G1; 95% CI 966.91-826.37 for G2). (a-c) Black horizontal line in the middle of the box is median, upper and lower limits of the box indicate first and third quartile. Black dots represent outliers. P values is based on one-way ANOVA.

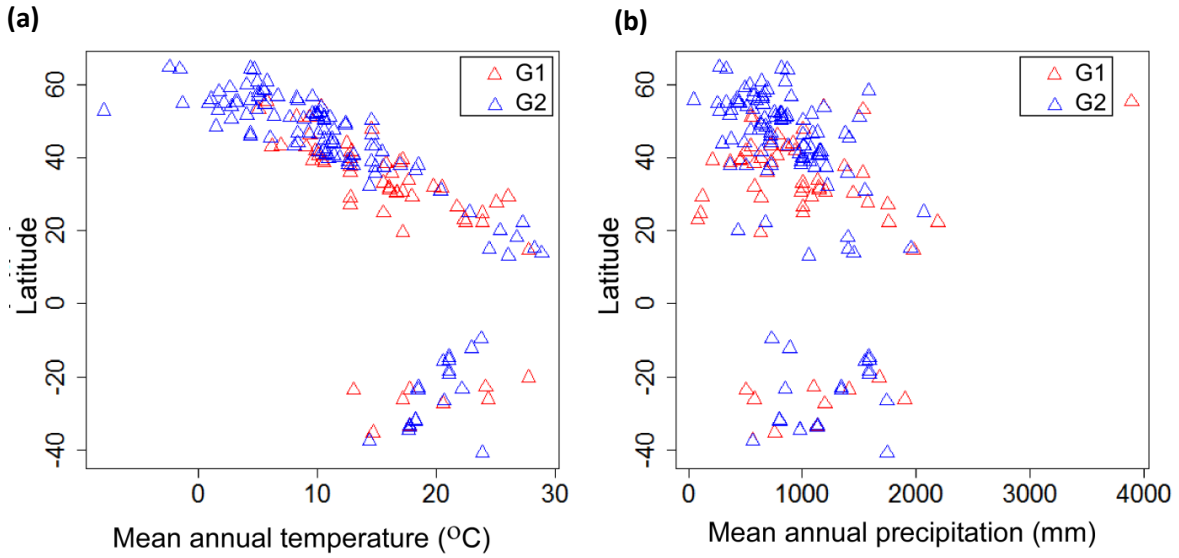


Figure 7: **Relationship between latitude and other two climatic variables, temperature and precipitation.** Scatter plot for (a) latitude and annual temperature and (b) latitude and precipitation for each SARS CoV-2 strain (n=176) belonging to G1 group (n=58, shown in red) and G2 group (n=118, shown in blue).