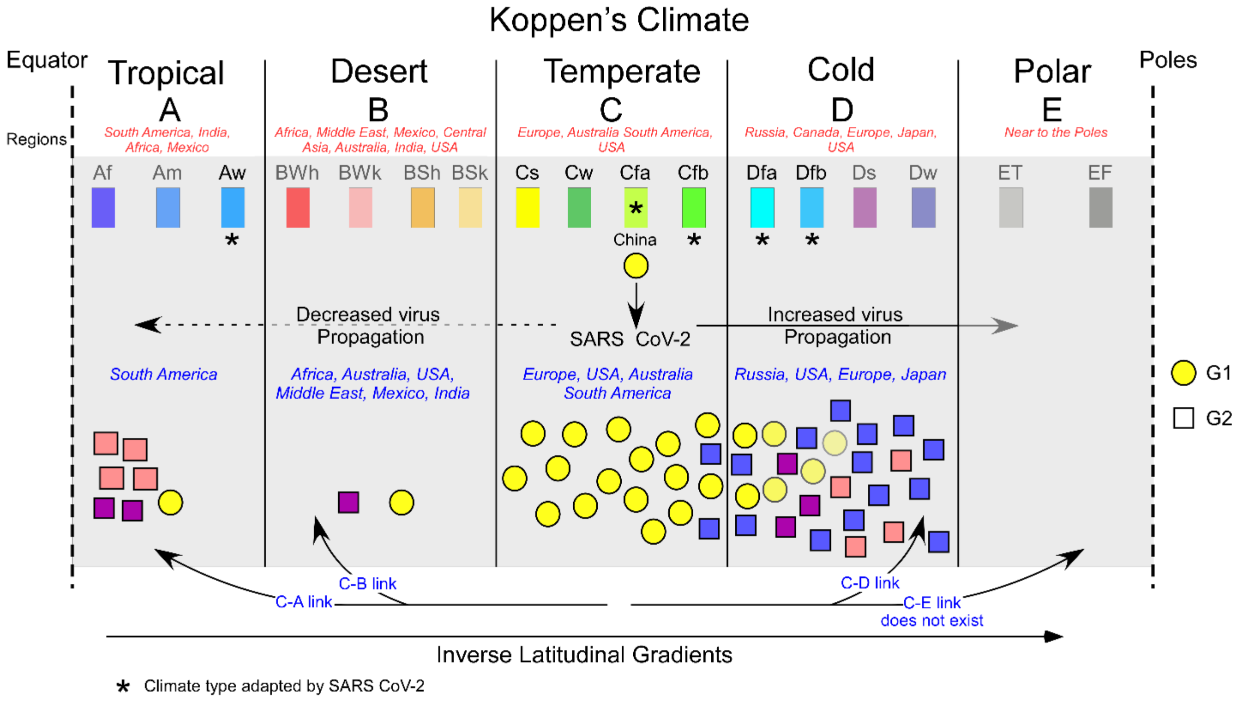
Climatic-niche evolution of SARS-CoV-2

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

Graphical Abstract


Koppen's Climate
Equator
Poles
Regions
Tropical A
South America, India, Africa, Mexico
Desert B
Africa, Middle East, Mexico, Central Asia, Australia, India, USA
Temperate C
Europe, Australia South America, USA
Cold D
Russia, Canada, Europe, Japan, USA
Polar E
Near to the Poles
Af
Am
Aw
BWh
BWk
BSh
BSk
Cs
Cw
Cfa
Cfb
Dfa
Dfb
Ds
Dw
ET
EF
China
Decreased virus Propagation
SARS CoV-2
Increased virus Propagation
South America
Africa, Australia, USA, Middle East, Mexico, India
Europe, USA, Australia South America
Russia, USA, Europe, Japan
G1
G2
C-A link
C-B link
C-D link
C-E link does not exist
Inverse Latitudinal Gradients
* Climate type adapted by SARS CoV-2

In Brief:

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

Highlights:

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

Climatic-niche evolution of SARS-CoV-2

Priyanka Bajaj[1#*] & Prakash Chandra 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 equally contributed
**Correspondence: prakasha@iisc.ac.in
*Correspondence: priyankaba@iisc.ac.in

Abstract

COVID-19 pandemic is studied by several field experts. However, it is still unclear why it was restricted to higher latitudes during the initial days & later cascaded in the tropics. Here, we analyzed 176 SARS-CoV-2 genomes across different latitudes & climate (Koppen's climate) that provided insights about within species virus evolution & its relation to abiotic factors. Two genetically variant groups, named as G1 & G2 were identified, well defined by four mutations. The G1 group (ancestor), is mainly restricted to warm & moist, temperate climate (Koppen's C climate) while its descendent G2 group surpasses the climatic restrictions of G1, initially cascading into neighboring cold climate (D) of higher latitudes & later into hot climate of the tropics (A). It appears that the gradation of temperate climate (Cfa-Cfb) to "cold climate" (Dfa-Dfb) climate drives the evolution of G1 into G2 variant group which later adapted to tropical climate (A) as well. It seems this virus follows inverse latitudinal gradient in the beginning due to its preference towards temperate (C) & cold climate (D). Nevertheless, due to the uncertainty of COVID-19 data, the results must be cautiously interpreted & should not be extrapolated to climate types and climatic conditions other than those analyzed here for the early evolution period. Our work elucidates virus evolutionary studies combined with

climatic studies can provide crucial information about the pathogenesis & natural spreading pathways in such outbreaks which is hard to achieve through individual studies.

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

Introduction

The first case of COVID-19 pandemic caused by SARS-CoV-2 pathogen was first reported from Wuhan China[1]. In spite of various precautions such as lockdown, social distancing, wearing mask & sanitization, the disease was able to reach to almost every part of the world infecting nearly 6 million people worldwide & putting an end to nearly 370000 lives[2]. This zoonotic virus is like SARS coronavirus (79% similarity) & MERS (50% similarity) & is closely related to bat derived coronaviruses[1]. The SARS-CoV-2 can survive up to 3, 4, & 24 hours on aerosols, copper, & cardboard respectively & up to 2-3 days on stainless-steel or plastic[1]. These results provide vital information about the survival of the virus in its external environment, few surfaces tend to be relatively favorable than others. It has also been observed that SARS-CoV-2 transmits faster than its two ancestors SARS-CoV & MERS-CoV[1,3]. It is well understood that the SARS-CoV-2 has concurred a larger geographical region & hosted a larger population. Since the social behaviour & travelling of humans have not changed much, what makes few respiratory viruses confined locally & others spread globally is still unclear. Due to its unique pattern of spread, the outbreak led to a big discussion, that does climate have a role in the spread of the disease. The ancestor SARS-CoV-1 losses its viability at higher temperature (38°C) & relatively higher humidity (>95%)[4]. Experiments support that the virus is highly stable at 4°C but is sensitive to heat[5]. The effect of climate on COVID-19 transmission has been discussed by several authors. Studies both in favour & against have been published & the topic is still debatable[6,7]. A recent review compiles 61 studies relating COVID-19 with climate[1]. Several climatic factors such as humidity, precipitation, radiation, temperature, & wind speed affecting this virus spread have been incorporated. Both positive & negative association of COVID-19 with temperature & humidity have been published[1]. Recently, Carlson et al. mentioned that COVID-19 transmission could be affected by climate but discouraged the use of SDMs for COVID-19 transmission due to their limitations, which generally takes only climatic parameters as input, as they may not be appropriate tools[7]. This study was challenged by Araujo et al. mentioning strengths & limitations of the tools & reasoned that R_0 of COVID-19 depends on several factors it may also be affected by climate[6]. Since only climatic parameters are insufficient to capture climatic signatures of COVID-19 spread, such patterns can be recognized by combining phylogenetic & climatic studies. Such approach enables to probe the similarities & differences in virus genome across similar & different climate types present all over the world. To understand such a behaviour we have attempted to study the genomic sequence of the SARS-CoV-2 across different latitudes & climate (Koppen's climate).

The plant & animal diversity generally decrease from equator to pole[8]. This pattern is known as the latitudinal biodiversity gradient, identified & discussed by several authors[8,9] with few exceptions[9]. Unlike free living plants & animals, pathogens are poorly mapped & very little is known about their underlying ecological & evolutionary causes[10]. Nucleotide substitution has been proposed to be one of the most important mechanisms of viral evolution in nature[11]. However, factors responsible for the generation of these mutations are not well understood. One of the possible factors is adaptation to new environments dictated by natural selection that discriminates among genetic variations & favours survival of the fittest[12]. Virus evolution as a consequence of climate change is poorly understood. SARS-CoV-2 consists of large single-stranded ~30kb long positive-sense RNA. These viruses majorly have a conserved genomic organization, consisting of a unique 265bp long leader sequence, ORF1ab polyprotein, & structural proteins like S (spike glycoprotein), E (Envelope), M (Membrane), & N (Nucleocapsid). ORF1ab encodes replicase, transcriptase & helicase, essential enzymes required for replication, along with non-structural & accessory proteins. Expression of non-structural proteins is facilitated by ribosomal frameshifting[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 & entry & is the main target for neutralizing antibodies & development of vaccines[15]. Membrane protein is also antigenic as it stimulates a humoral immune response[16]. E protein is responsible for virus assembly & release of virion particles[17]. Nucleocapsid protein packages RNA genome into a helical ribonucleocapsid protein (RNP) complex during virion assembly & is capable of eliciting an immune response[18]. Since it is still not clear whether SARS-CoV-2 evolution & spread have relation with climate, our study may act as a missing link between genomic sequence, climate & COVID-19 severity. If SARS-CoV-2 is responding towards external climate it can be delineated by superimposing its genomic variants across different latitudes & Koppen's climate. The earliest & 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) & South Frigid Zone (SFZ)[19]. Wladimir Koppen presented a modified classification of Earth's climate based on the precipitation & temperature[20]. He divided Earth's climate into five major climates, A (Tropical), B (Arid), C (Temperate), D (Cold or Continental) & E (Polar) which are further subdivided into 30 climate types[20]. To understand the effect of climate on SARS-CoV-2 evolution, the present study comprises of three parts, (1) latitudinal distribution of COVID-19 cases, (2) sequence analysis of SARS-CoV-2 strains, (3) mapping SARS-CoV-2 strains across different climates. These combined studies can provide insights on within species evolution & preferential distribution of SARS-CoV-2 across different climatic zones which might be difficult to probe through individual studies. These results will provide useful information to design efficacious vaccines which can be stored and transported in a wide range of temperature and humid conditions, thereby minimizing cold storage costs.

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 (Figure1a). 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) & negligible number of cases (<0.5%) are from remaining 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 majority of terrestrial taxa prefers to stay near tropical region, suggesting that SARS-CoV-2 follows inverse latitudinal gradient in early stages of pandemic. Since a majority of the cases lie in the NTZ, 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 & 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 & 10% in 58°N-66.5°N (Figure1b). 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) & 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 & ~< 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. The latitudes have 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 & their distribution in different climates

To probe genomic similarities between SARS-CoV-2 virus isolates, a phylogenetic tree was constructed by aligning 176 virus genomes to the reference genome[21] retrieved from GISAID. Interestingly, our Multiple Sequence Alignment (MSA) results reveal sixty virus cluster Single Nucleotide Polymorphisms (SNPs) (see methodology). Table1 comprises of SNPs of virus clusters across different climatic zones, Koppen's climate & climate type. Climatic parameters (temperature & precipitation) for each virus strain is mentioned in TableS2. Based on phylogenetic clustering, 176 SARS-CoV-2 strains are majorly divided into two groups, we named them as G1 (1-58) & G2 (59-176) (Figure2). Predominantly four mutations distinguish G2 from G1 group, i.e., 1) a synonymous mutation (C241T) appeared in the unique leader sequence, 2) F924 (C3037T) appeared in nsp3, encoding for papain-like proteinase[22], 3) a non-synonymous mutation, P214L (C14408T) arose in ORF1b, that codes for four putative non-structural proteins (nsp13, nsp14, nsp15 & nsp16), functionally involved in replication-transcription complex[23], & 4) D614G (A23403G) arose in S gene, encoding spike glycoprotein[14] (Figure3a). Among four mutations in G2, the D614G mutation, lying in spike glycoprotein was widely studied due to its higher infectivity & involvement in entering the host cell through hACE2 receptors[24–27]. The other three mutations in G2 have co-evolved with D614G making it distinguishable from G1. We explored the extent of genome-wide divergence of G1 & G2 group across different climate zones & Koppen's climate (Figure3b). 59% of G1 viruses fall in NTZ, 14% in NSTZ, 12% in TZ, 10% in SSTZ & 5% in STZ. 76% of the virus isolates in G2 group are present in the NTZ, 13.5% in TZ, 7.6% in STZ & remaining 2% is equally distributed in NSTZ & 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 & G2 strains have a strong preference towards higher latitudes i.e., NTZ, which agrees with the analyzed worldometer data (Figure3c). It also supports that in the initial stages of the pandemic, the virus isolates follow inverse latitudinal gradient. Mapping viral strains on Koppen's map (thoroughly discussed in the next section) reveal their prevalence majorly in the C & D climate (Figure3d). 71% of G1 lie in C climate, 17% in D & the remaining is equally distributed in the A & B climate. 54% of G2 lie in C climate, 36% in D, 9% in A & 1% in B climate pointing towards a preferential shift of the novel coronavirus towards D climate (Figure3b), alluding G2 is climatically & genomically more diverse than G1. The analysis suggests that the G1 group is mostly restricted to temperate climate (C) & G2 is climatically & geography widely distributed, it is possible that these mutations were acquired by G1 to stabilize itself in different climates hence allowing it to spread globally. Similar climatic concordance with the temperate climate (C) was also observed for SARS-CoV that was responsible for 2002-2004 epidemic as it prevailed in regions of Australia, Europe, Canada and, China[28], having Koppen's C climate. Such similar occurrence of SARS-CoV and G1 group of SARS-CoV-2 hints towards, why initially G1 variant group (consisting of the reference genome NC_045512[21]) that has 79% similarity to SARS-CoV[1] was majorly located in the temperate climate and latter it evolved into G2 variant group that allowed it to extend its climatic boundaries into temperate, cold and tropical climate. These results suggest these four SNPs could be the key factors in increasing the virulence, transmission & sustainability of the virus in humans.

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 & 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; G28883C, G204R), 116-125 (A20268G, L2167), 126-176 (G25563T, Q57H) & 62-79 (cluster, acquired genomic mutation & its corresponding amino acid mutation). In Figure3e, looking at the distribution of the viruses in different climate zones, no such preference was observed as the virus evolved. Virus cluster 58-61, linking G1 & G2 has an equal distribution of virus strains in C & D climate. The virus cluster 80-115 of G2 more closely related to G1, is widely distributed in A, C & D climates. Within 80-115 virus cluster, 106-115 cluster shows distribution in C & A climate. A trend was observed that virus clusters in G2 group gradually evolved to sustain in Koppen's D climate which supports our previous observation. These analyses led us interpret G1 group (ancestor), is mainly restricted to warm & moist, temperate climate (C) while its descendent G2 group surpasses the climatic restrictions of G1, initially cascading into neighboring cold climate (D) of higher latitudes & later into hot climate of the tropics (A). Within these major virus clusters, small clusters also exist as shown in Table1 with their mutations along with their climatic distribution.

We have examined whether climatic conditions exhibit any selective pressure on each gene (Figure3f). Since, the present picture of the data appears that SARS-CoV-2 is following inverse latitudinal gradient, as expected all genes are having mutations in NTZ, suggesting the virus is probably using varied mechanisms to adapt to the two main climates of NTZ i.e. temperate (C) & cold or continental (D). Mutations in the M gene are only pertaining to NTZ & NSTZ & are present in C & 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 & 20% in the D climate. Overall, the distribution of virus cluster SNPs of ORF1ab, S, ORF3a, & N gene follows a similar pattern across all the climatic zones & Koppen's climate, implying no difference in selective pressure of the climate in generating mutations in these genes. S, M, & N proteins are immunogenic[14,16,18], 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[29] & through mechanisms such as ribosome stalling[30] & mRNA secondary structure formation[31]. We probed the frequency of derived synonymous versus non-synonymous mutations & observed a very similar distribution pattern of the derived synonymous versus missense mutations across all climate zones & Koppen's climate (Figure3g). These analyses suggest novel coronavirus is using varied mechanisms both at the transcriptional as well as translational level to adapt, survive, & increase infectivity in all types of climates. These findings unequivocally bolster a requirement for further prompt, comprehensive studies that join genomic information, epidemiological information, & climatic distribution with COVID-19 severity.

Distribution of strains across Koppen's climate

To probe the relation between climate & SARS-CoV-2 strains, we superimposed genomic information along with their geolocations on the climate map of Wladimir Koppen (Figure4). We carefully examined the distribution of strains on Koppen's map & an overview of the map shows, the distribution of 176 strains are mainly concentrated in the western coasts of Europe & North America, & eastern coasts of China, North America, Australia & South America (Figure4). Throughout the text Koppen's climate type is marked within quotations & its standard symbol is written within brackets e.g., "*humid-subtropical" (Cfa)*. List of Koppen's symbol of each climate type is given in Supplementary TableS3 & its criteria for classification is given in Supplementary TableS4. Mostly the SARS-CoV-2 strains are distributed in the "*humid-subtropical" (Cfa)* & "*marine-temperate" (Cfb)* & "*humid-continental" (Dfa-Dfb)* climate &, two strains from virus clusters (80-115 & 126-176) belonging to South America, are found in "tropical-savanna" *(Aw)* of 'A' climate (Supplementary TableS5). The map displays ~86% (n=176) of virus isolates are distributed in the coastal regions & the remaining in continental region (Chi-square test, *P*<.001, Figure5a). Around ~73.86% of the total strains are distributed in "humid-subtropical" (Cfa) & "marine-temperate" (Cfb) climate type of C climate & "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 & non 'Dfa-Dfb' of D climate (Figure5b). It seems that spread of COVID-19 is maximally in areas with 'Cfa' & 'Cfb' climate type. The climatic parameters (temperature & precipitation) in which these strains were found were analyzed. Statistically, significant difference was found in the latitudes of G1 ~24.14±3.5 (mean±s.e.) & G2 ~34.03±2.7 (one-way ANOVA, *P*=.03251, Figure6a). Statistically, significant difference was observed in the temperatures of G1 (15.82±0.75 ^{o}C (mean±s.e.) & G2 (11.67±0.68 ^{o}C) strains (one-way ANOVA, *P*<.001, Figure6b). However, the difference in precipitation for G1 (1046.95±80 mm) & G2 (896.64±35.48 mm) strains is statistically not significant (one-way ANOVA, *P*=.06118, Figure6c). The latitudes & temperature are inversely related to each other (r = -0.6649, Supplementary Figure1a & Figure6d), which explains the occurrence of G1 strains in lower & G2 strains in higher latitudes. Such relation between latitude & precipitation has not been observed (r = -0.3064, Supplementary FigureS1b & Figure6e). A mesh plot simultaneously evaluates all climatic parameters for both G1 & G2 strains, the results agrees to the limited temperature & wider precipitation range of G1 group & interestingly the G2 group appears in a wider temperature & shows a preferential shift towards lower temperature which is evident from the fact that it initially appeared more in higher latitudes (Supplementary FigureS2). A complete description of the distribution of

G1 & G2 strains lying in different countries &/or continents of the world is provided in Supplementary FigureS3 & Supplementary Material.

Discussion

Unlike majority of terrestrial organisms, it appears that SARS-CoV-2 is following an inverse latitudinal gradient, a possible explanation from our analysis is that evolution from G1 to G2 might help to sustain this virus in temperate & cold climates which might change with time as it appears it is adapting to different climate. Mainly four mutations in leader sequence, ORF1ab, & S gene were identified that led G1 to evolve into G2. The leader sequence & ORF1ab is involved in replication & transcription, & the S gene is involved in binding to the host cell through hACE2 receptors. Substitutions in the ORF1ab gene may increase the synthesis of replicase-transcriptase complex, thus, increasing the replication rate of the virus & 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[32]. 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 & greater incorporation of the S protein into the pseudovirion[24]. Several studies reported D614G mutation is increasing at an alarming rate[25,26]. Few observed that this alteration correlated with increased viral loads in COVID-19 patients[25]. 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[27]. This substitution coevolved with substitution in the leader sequence, nsp3 & 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 & infective. Also, Forster et al. observed in his phylogenetic analysis the preferential geographical spread of SARS-CoV-2 & provided a plausible cause which could be founders effect or immunological or environmental effect[33]. Although there is a possibility that the stable variant might have appeared because of host innate immune response or some unknown reason, in such a case it would not show any close association with climate. Since our analysis shows largely G1 is restricted to temperate climate & G2 spreads in temperate & adjoining climates, a pattern that is consistently observed all over the world, led us doubt that climate could be one of the possible selective pressures towards which SARS-CoV-2 responds by altering its genome. Through our analysis we are inclined to say that climate does not display any selective pressure on each gene of SARS-CoV-2. Our genomic analysis of virus strains show that the novel coronavirus undergoes both synonymous as well as non-synonymous mutations throughout its genome in various climates, suggesting the novel coronavirus uses multiple mechanisms both at the transcriptional & translational level for evading the immune response, developing drug resistance & increasing pathogenesis. However, the actual role of these mutations is not yet determined, & these studies need to be further enlightened by biophysical & biochemical studies. Such mutational insights will aid the design of efficacious vaccines that can be stored and transported in a wide range of temperature and conditions, thereby minimizing cold storage costs.

To delineate the signatures of underlying abiotic factors (temperature, precipitation, & latitude) responsible for evolution of SARS-CoV-2 (n=176), spreading patterns of G1 & G2 strains were carefully examined on Koppen's climate map. Figure4 shows an elevated spread of COVID-19 in the western & eastern coasts of the continents & a diminished spread in the hot & cold deserts. The G1 strains are majorly present in the eastern & western coasts of the continents & G2 strains lie in both the coastal regions & continent's interior. On a closer inspection, the eastern coasts of continents consist of "humid-subtropical" (Cfa) climate while the western coasts of continents consist of "marine-temperate" (Cfb) commonly known as east & west coast climate, respectively. These two climates are very similar to each other & belongs to temperate climate also known as C type climate of Koppen's classification scheme. A very large portion (~94%) of habitable China consists of temperate climate (C), i.e., humid subtropical climate (Cfa), which explains presence of only G1 strains in China & one strain of G1 is present in cold climate (D) present near the transition of temperate (C) to cold climate (D), thus probably temperate climate was suitable for G1. A similar association of G1 with temperate climate (C) was found in eastern & western coast of North America, eastern coast of South America, western coast of Europe & eastern & western coast of Australia. Statistically, distribution of G1 strains all over the globe is in concordance with the temperate climate & strongly favor C climate (Chi-square test, *P*<.001) as compared to any other climate. If climate does not have any role in the evolution & preferential spread of coronavirus, in such a case G1 would have been evenly distributed in all climate types which is not the case. Few exceptions of G1 seen in other climate types are most probably because of travel as they remained subsided in that climate, implying their inability to sustain in other climate types. Such a significant difference in association of climate with G1 & G2 population could not merely arise due to human transportation. It appears that the G1 strains existed in temperate climate all over the world but could not extend their geographical territories beyond temperate climate. Contrastingly, the evolved G2 strains can sustain in temperate (C), cold (D) & tropical (A) climate. It appears that G2 strains enters the continent's interior through D climate (e.g., North America & Russia). Temperate climate (C) generally grades into cold climate (D) & deserts (B) in the northern hemisphere (e.g., C to D: Europe to Russia, & USA to Canada; C to B: China, & USA). In southern hemisphere, gradation of temperate climate (C) into tropical climate (A) & deserts (B) exists (e.g., C to A, Brazil; C to B, Australia), C to A transition is identified by virus cluster 105-115 in phylogenetic tree. In Russia, 91.3% (21/23) of the strains belong to G2 (Figure4), are mainly present in the ~8500 km long & 600-1700 km wide D climate belt ('Dfa-Dfb-Dw'), suggesting the G2 strains might have adapted to the D climate (Chi-square test, $P<.001$). Similar observations are seen for North America, South America & Australia. The eastern & western coasts of North America have temperate climate & are connected by cold climate along USA-Canada boundary (i.e., having humid subtropical (Cfa) in eastern coast & marine temperate climate (Cfb) in western coast) (Figure4). The G2 strains follows this cold climate (Dfa-Dfb) belt which is ~3800 km long & ~600 to 1000 km wide. The dominance of G2 & nearly absence of G1 population in cold climate of North America is similar to the observations of Russia. Our analysis suggests that a fall of temperature from temperate to cold climate might have dictated the evolution of G1 into G2 variant group (Chi-square test, $P<.001$). Similarly, a change in climate from C to A probably made the strains stable in tropical regions. Throughout the world, the G1 strains have expanded only in temperate climate, suggesting a close relationship between G1 with temperate climate, however the evolved G2 strains were able to infect temperate, cold, & tropical climate. There might be other factors controlling virus evolution, but the possibility of virus evolving to sustain in different climates of the Earth cannot be neglected as several other studies also verified that viruses do respond towards temperature[4,5]. Our analysis also shows that it is highly possible SARS-CoV-2 has evolved to sustain in different climate. Studies combining genetic information with climate can provides useful information about virus evolution & possible climatic pathways during an outbreak.

Conclusion

It is reasonable to assume COVID-19 transmission pathway & evolution is influenced by climate. Phylogenetic network classified 176 SARS-CoV-2 strains into two variant groups G1 & 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 & D614G in S gene), plausibly extended its climatic boundaries from C to D climate, displaying role of natural selection on virus evolution. In our analysis SARS-CoV-2, were found resistive to desert climate (B). 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. There is a high possibility that the evolutionary pathway adopted by the virus is temperate climate to cold climate for higher latitude & temperate climate to tropical climate for lower latitudes. Nevertheless, due to the uncertainty of COVID-19 data, the results should be carefully interpreted & should not be extrapolated to climate types and climatic conditions other than those analyzed here for the early evolution period. The study agrees that viruses are sensitive to their environment & respond towards naturally occurring abiotic factors such as temperature, latitude & 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 & 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 & provides vital information about the direction of the spread & highlights vulnerable regions of Earth. Such inter-disciplinary studies will play an imperative role in designing antiviral strategies & 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[34]. '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 & 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), & (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 & compared.

Molecular phylogenetic analysis

185 full-length SARS-CoV-2 genomic sequences from countries across the globe, with genome length more than 29 kb & high coverage were obtained from Global Initiative on Sharing Avian Influenza Data (GISAID) database, accessed till 2 May 2020 & the reference genome was retrieved from GenBank24 (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[21] by using Biomanager & Seqinr packages of R (version 3.6.3). Among 185 genomes, some partial genomes were discarded. NC_045512 genome sequence was used as reference & 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)[21]. To ensure comparability, we trimmed the flanks of all sequences. The aligned sequences were used to construct a phylogenetic tree using MEGA X[35]. The evolutionary history was inferred using the Neighbor-Joining method (500 bootstrap tests)[36]. 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[37] & 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 & 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 (Figure2). 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[38]. The Gieger-Koppen’s climate map is used for climatic studies[20]. The Koppen climate type, temperature, precipitation of each strain is assessed from weatherbase[39] & CLIMATE.ORG[40]. The map is georeferenced by using ‘Arc-GIS 10.1’[41]. The locations of all strains (n=176) were transferred to the georeferenced map[41]. On the map, the G1 strains were symbolized as ‘Yellow-circle’, & G2 as ‘Square’ (Figure4). Each strain in the map is labelled as per their Strain ID (1 to 176) (Figure4), the map combines information of the phylogeny, climate, & global distribution of SARS-CoV-2. These locations were classified into coastal & continental region, we define the coastal region as land region < 500 km from the ocean/sea & the continental region as land lying >500 km from the coastline measured through google maps.

Statistical analysis

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

Data accessibility

The full-length genomic sequences were downloaded from GIS-AID website (https://www.gisaid.org/), an open source database for influenza viruses. The data is downloaded as FASTA file along with the acknowledgement. The location of each strain is accessed from its METADATA file. The Koppen's Climate map is taken from the published article[20] (Peel et al., 2007). The Koppen climate type, temperature & precipitation for each strain is taken from weatherbase (https://www.weatherbase.com/) & CLIMATEDATA.ORG (https://en.climate-data.org/). The total cases of COVID-19 all over the globe is retrieved from worldometer website (https://www.worldometers.info/coronavirus/). Refer Supplementary Tables S1-S5. The code is available from the corresponding author on request.

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, a large fraction of 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. Data from different individual locations without travel history & large gaps in genomic sequences have been incorporated in this study. To overcome this, the inverse latitude gradients were studied based on the total number of COVID-19 cases all over the globe. Due to the uncertainty of COVID-19 data, these results should be carefully interpreted & should not be extrapolated to climate types and climatic conditions other than those analyzed here for the early evolution period.

References

1. Briz-Redón, Á. & Serrano-Aroca, Á. The effect of climate on the spread of the COVID-19 pandemic: A review of findings, and statistical and modelling techniques. *Prog. Phys. Geogr.* (2020). doi:10.1177/0309133320946302

2. WHO. *Coronavirus disease (COVID-19)*. (2020).

3. Vellingiri, B. *et al.* COVID-19: A promising cure for the global panic. *Science of the Total Environment* (2020). doi:10.1016/j.scitotenv.2020.138277

4. Chan, K. H. *et al.* The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Adv. Virol.* (2011). doi:10.1155/2011/734690

5. Chin, A. W. H. *et al.* Stability of SARS-CoV-2 in different environmental conditions. *The Lancet Microbe* (2020). doi:10.1016/s2666-5247(20)30003-3

6. Araújo, M. B., Mestre, F. & Naimi, B. Ecological and epidemiological models are both useful for SARS-CoV-2. *Nature Ecology and Evolution* (2020). doi:10.1038/s41559-020-1246-y

7. Carlson, C. J., Chipperfield, J. D., Benito, B. M., Telford, R. J. & O'Hara, R. B. Species distribution models are inappropriate for COVID-19. *Nature Ecology and Evolution* (2020). doi:10.1038/s41559-020-1212-8

8. 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).

9. Kindlmann, P., Schödelbauerová, I. & Dixon, A. F. G. Inverse latitudinal gradients in species diversity. *Scaling Biodivers.* 246–257 (2012). doi:10.1017/cbo9780511814938.014

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

11. Lauring, A. S. & Andino, R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathogens* (2010). doi:10.1371/journal.ppat.1001005

12. 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

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* (2012). doi:10.3390/v4030363

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. Allaby, M. *Atmosphere. A scientific history of air, weather, and climate*. *Advances in Space Research* (2009).

20. 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).

21. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020). doi:10.1038/s41586-020-2008-3

22. 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

23. Snijder, E. J., Decroly, E. & Ziebuhr, J. The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. in *Advances in Virus Research* (2016). doi:10.1016/bs.aivir.2016.08.008

24. 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

25. 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

26. 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

27. 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

28. WHO | SARS (Severe Acute Respiratory Syndrome). Available at: https://www.who.int/ith/diseases/sars/en/. (Accessed: 10th September 2020)

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

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

31. 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

32. 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).

33. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U. S. A.* (2020). doi:10.1073/pnas.2004999117

34. Coronavirus Update (Live): 8,522,724 Cases and 453,714 Deaths from COVID-19 Virus Pandemic - Worldometer.

35. 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).

36. 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).

37. 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).

38. EarthExplorer.

39. Travel Weather Averages (Weatherbase).

40. Climate data for cities worldwide - Climate-Data.org.

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

Authors Contribution

Phylogenetic study is carried out by P.B. GIS study & Koppen's climate map interpretations is done by P.C.A. Worldometer data analysis is carried out by both the authors. Both authors have written, reviewed & edited the manuscript.

Acknowledgement

We gratefully acknowledge the authors & originating & 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 & Chetankumar Jalihal of Indian Institute of Science for proofreading the article.

Conflict of Interest

Authors declare no conflict of interest.

TABLE

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

Virus cluster	Nucleotide mutation	Amino acid mutation	Gene	Climate Zone	KCT	KC
1-22	C8782T T28144C	S2839 L84S	ORF1a ORF8	NTZ	Cfa	C
5-6	C29095T	F274	N	NTZ	Cfa	C
8-9	T9477A G25979T C28657T C28863T	F3071Y G196V D128 S197L	ORF1a ORF3a N N	NTZ, TZ	Mix	C-A
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 T26729C G28077C	N824 A69 V62L	S M ORF8	NTZ	Cfa	C
21-22	T490A C3177T T18736C	D75 P971L F1657L	ORF1a ORF1a ORF1b	NTZ, NTSZ	Cfa	C
23-25	C6312A C13730T C23929T C28311T	T2016K L4489 Y789 P13L	ORF1a ORF1a S N	NTZ, TZ, SSTZ	Mix	Mix
28-32	G1397A T28688C	D392G L139	ORF1a N	NTZ	Mix	Mix
33-43	G26144T	G251V	ORF3a	NTZ	Cfa-Cfb	C
37-39	A2480G C2558T	I739V P765S	ORF1a ORF1a	NTZ	Mix	Mix
37-43	C14805T	Y346	ORF1b	NTZ	Cfa	C
42-43	T17247C	R1160	ORF1b	NTZ	Cfb	B
44-47	G1440A G2891A	G392D A876T	ORF1a ORF1a	NTZ	Cfb	C
58-61	C15324T	N519	ORF1b	NTZ	Cfa-Dfb	C-D
59-176	C3037T A23403G C14408T	F924 D614G P214L	ORF1a S ORF1b	NTZ	Cfa-Cfb-Dfb-Aw	C-D -A
59-125, 127-176	C241T C241T		Leader seq.	NTZ	Cfa-Cfb-Dfa-Dfb	C-D-A 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		
	G28882A	R203K	N			
	G28883C	G204R	N		Cfa-Cfb-Dfb-Aw	C-D-A
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 tree (Supplementary table S1 & S2). Genomic coordinates in this study is based on reference genome[21]. The SNP 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 Climate zone column we have mentioned the major climate zone for the corresponding virus cluster[19]. KCT is Koppen's Climate Type & KC is Koppen's Climate columns display the main Koppen's climate in which most of the virus isolates of the corresponding virus cluster lie. 'Mix' implies no particular climate type is favored[20].

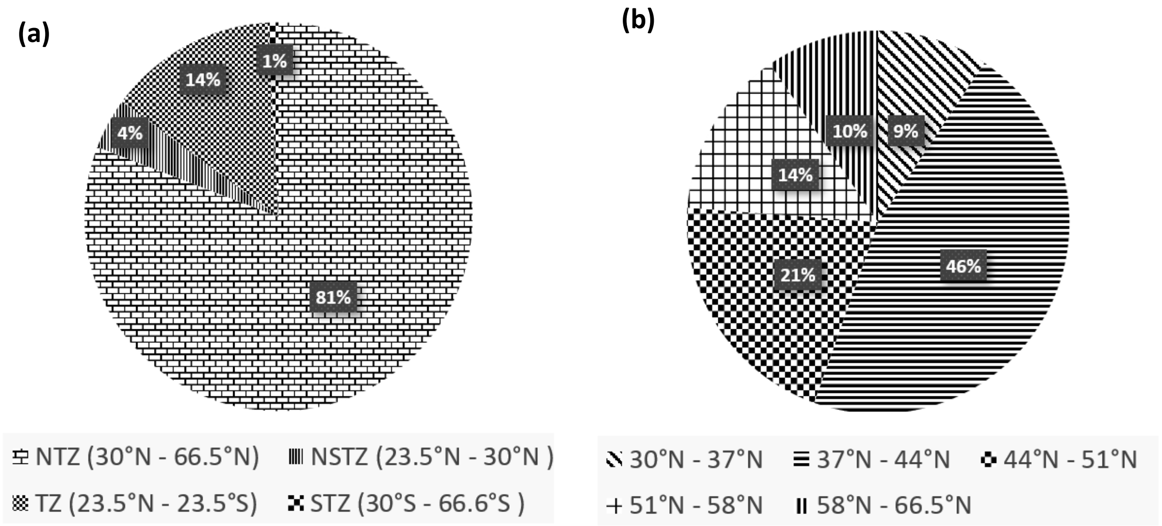


Figure1: Distribution of COVID-19 cases across different climate zones based on latitudes. (a) Area of the pie-chart covered by a climate zone is proportional to the percentage of COVID-19 cases occurring in their respective climate zones as depicted by black squares. The percentage of COVID-19 cases for NFZ & 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 in black squares.

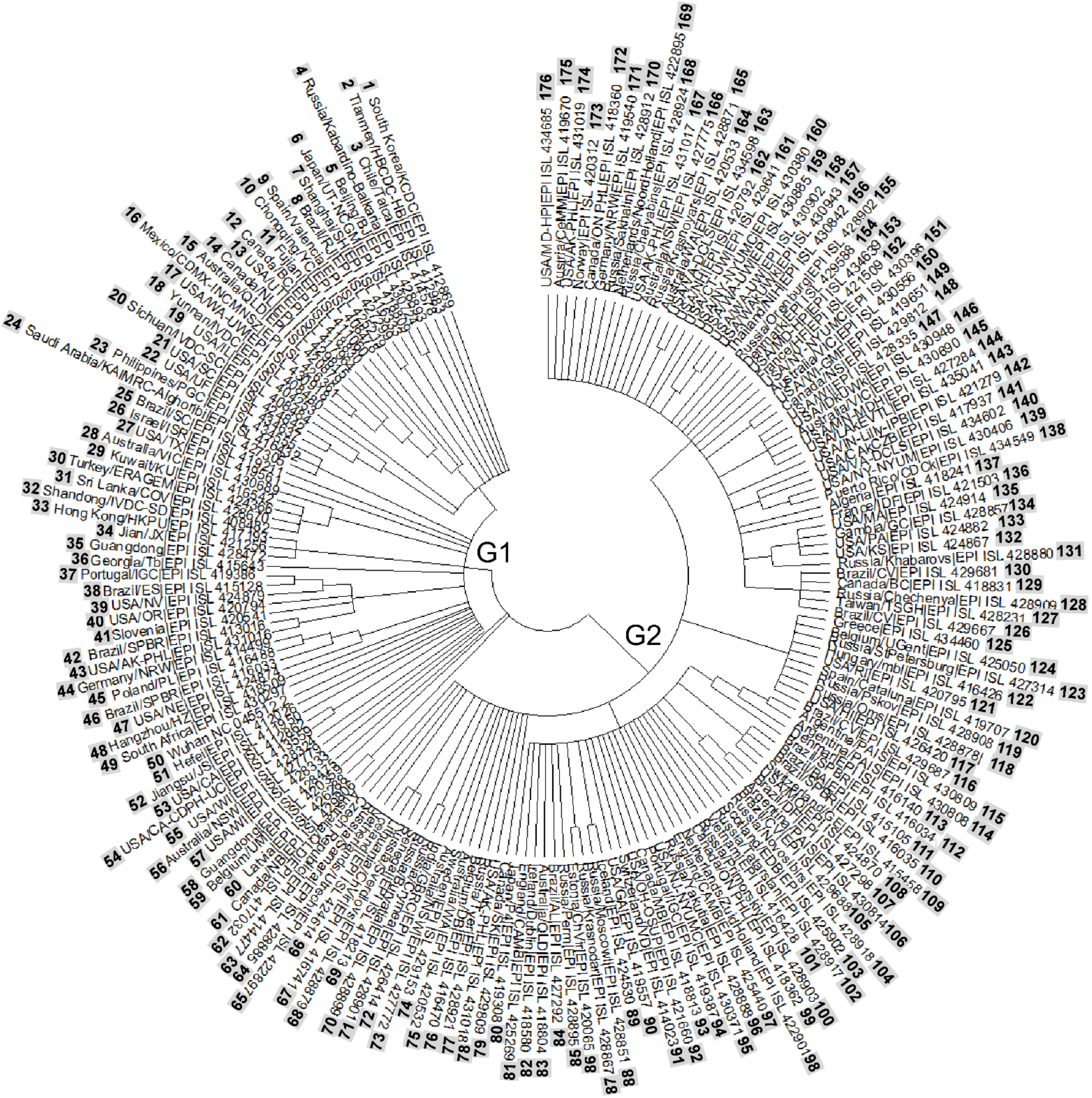
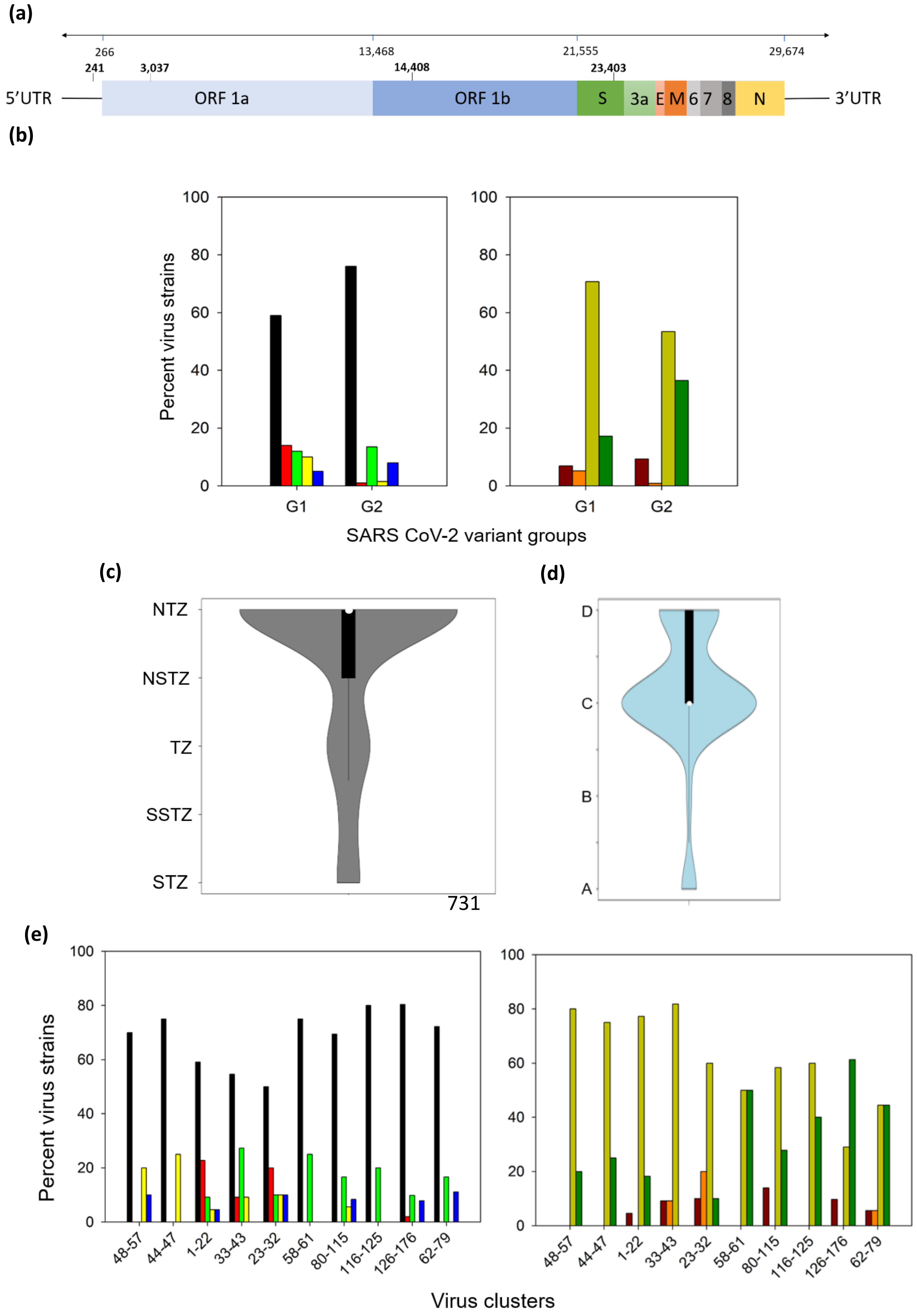


Figure2: Phylogenetic network divides 176 SARS-CoV-2 strains into two variant groups. Largely, the left side of the tree (1 to 58) constitute the G1 group & the right side of the tree constitutes the G2 group (59 to 176). Branch length is proportional to the genomic relatedness of the viral isolates. Closely related virus isolates comprise the same SNP with respect to the reference genome (Strain ID: 50) & form a cluster. The evolutionary history of 176 taxa was inferred using the Neighbor-Joining method[36] (500 bootstrap tests). A total of 29408 positions were analysed with nucleotide position numbering according to the reference sequence[21].


(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)
Percent virus strains
0
20
40
60
80
100
G1
G2
SARS CoV-2 variant groups
(c)
NTZ
NSTZ
TZ
SSTZ
STZ
(d)
D
C
B
A
(e)
Percent virus strains
48-57
44-47
1-22
33-43
23-32
58-61
80-115
116-125
126-176
62-79
Virus clusters

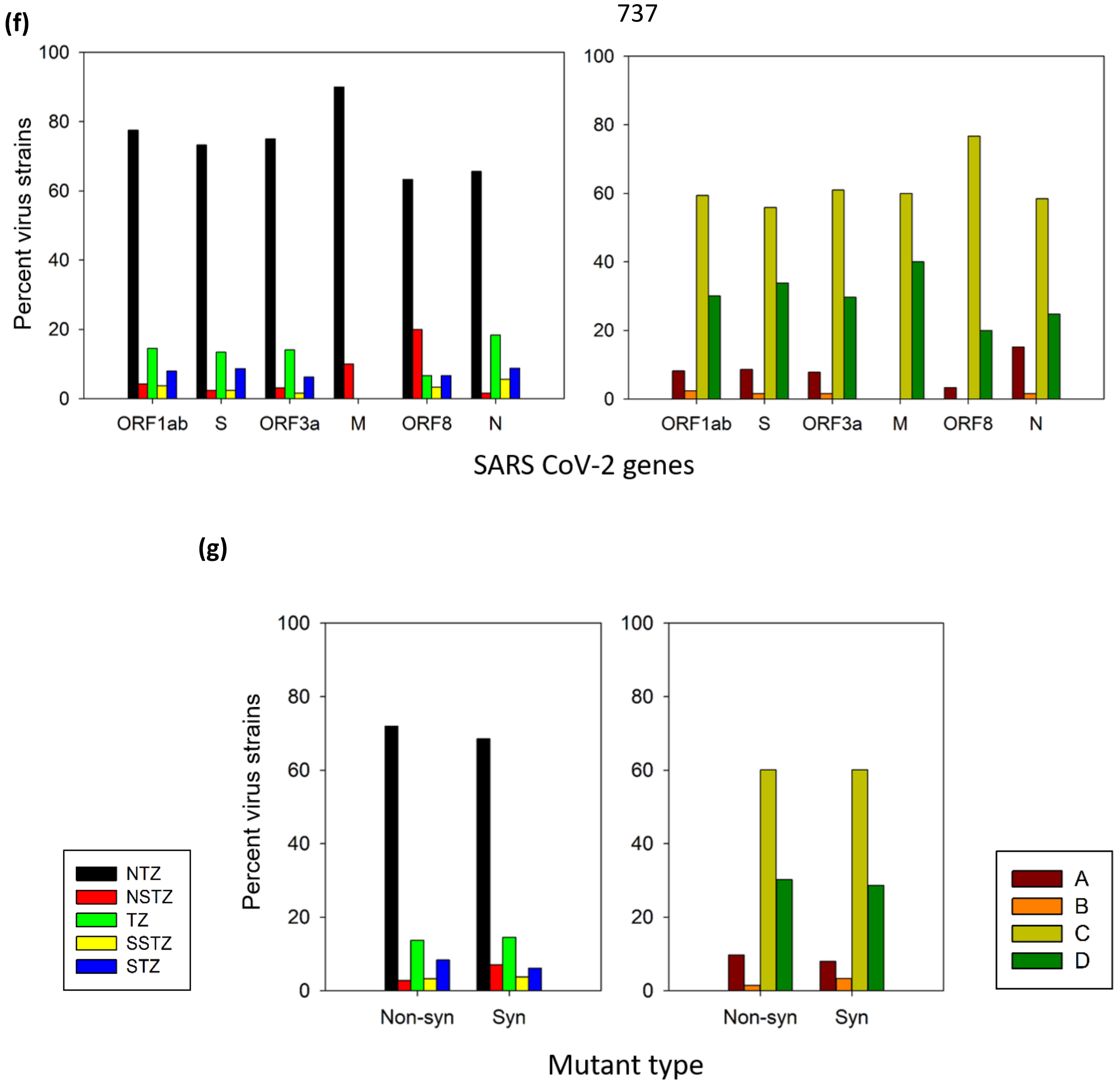


Figure3: Molecular phylogeny analysis to infer genomic similarities of SARS-CoV-2 & their distribution across different climate zones[19] **& Koppen's climate types**[20]. (a) Genomic architecture of SARS-CoV-2 genome highlighting four positions, substitutions on these positions enabled evolution of G1 into G2. (b, e-g) Strains found within a virus cluster (as shown in the phylogenetic tree & 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 & 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 & 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 & (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.

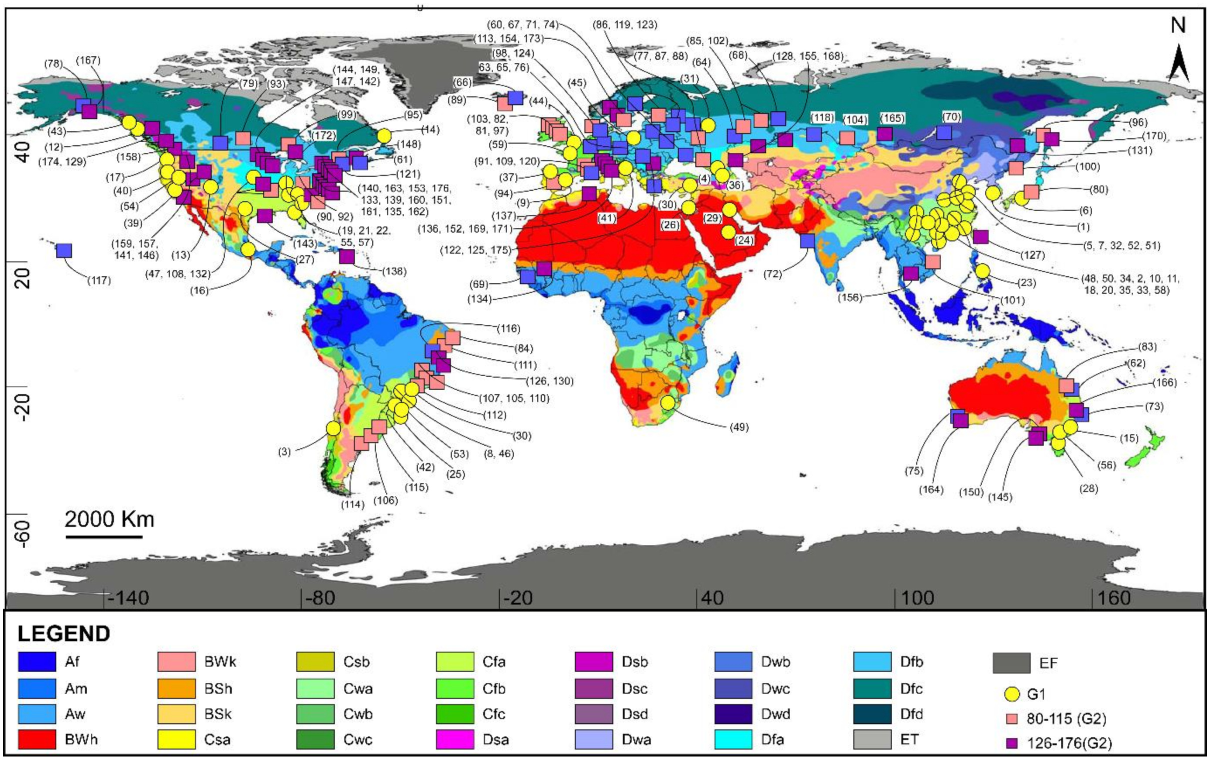


Figure4: Global distribution of SARS-CoV-2 strains on the Gieger-Koppen's map displaying different climate types[20]. Each strain is labelled as per the strain ID (1 to 176) within parenthesis. The G1 strains were symbolized as 'Yellow-circle', & G2 as 'Square', pink square denotes strain clusters (80-115) stable across C, D & A climate, purple square represents strain cluster (126-176) stable majorly in D climate, the remaining G2 strains (blue squares) are stable across C & 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 & 'D' to cold & 'E' to polar climate. The shades of blue on the map, in North America & Russia belongs to D climate. Shades of yellow & green belongs to C climate, shades of red, orange & pink belongs to Desert climate.

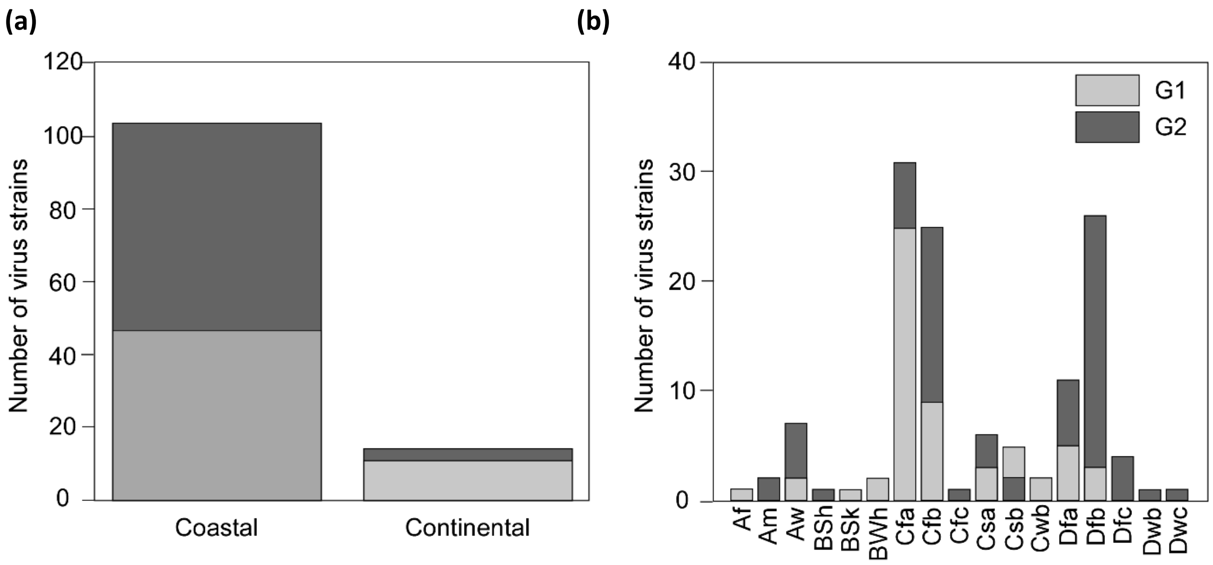


Figure5: Global distribution of SARS-CoV-2 strains (n=176) (a) in the coastal & continental region (b) & in different Koppen's climate types[20]. Number of virus strains in G1 population is represented by light grey color & of virus strains in G2 population is represented by dark grey color.

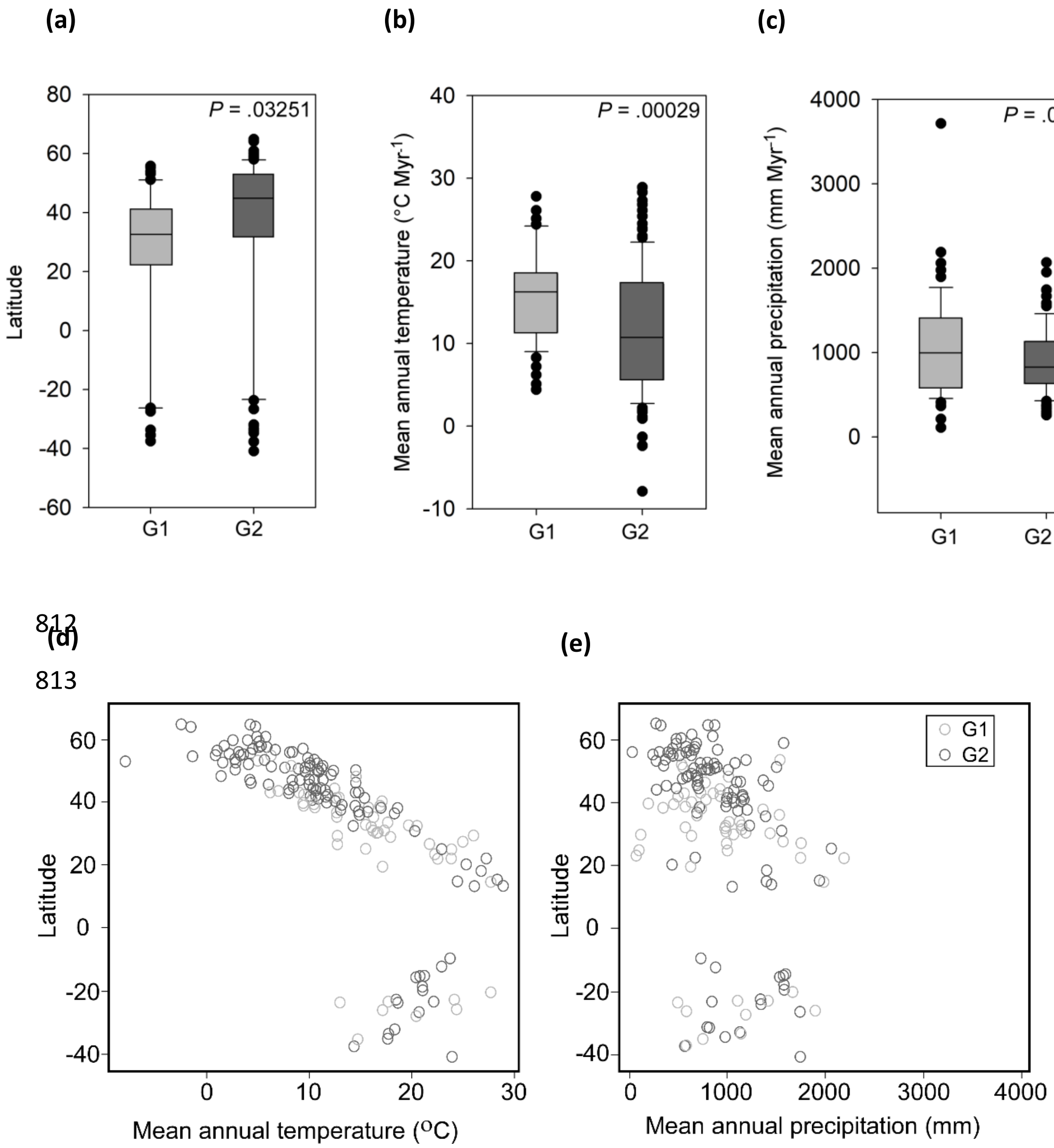


Figure6: Comparative analysis of different climatic parameters such as latitude, temperature & precipitation between G1 & G2 variant groups. (a) Positive values represent the latitude range falling in Northern Hemisphere & negative values represent latitude range falling in Southern Hemisphere. The G2 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 & 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 & lower limits of the box indicate first & third quartile. Black dots represent outliers. P values is based on one-way ANOVA. Scatter plot for (d) latitude & annual temperature & (e) latitude & precipitation for each SARS-CoV-2 strain (n=176) belonging to G1 group (n=58, shown in light grey) & G2 group (n=118, shown in dark grey).