September 13th, 2022

Epi-Clock: A sensitive platform to help understand pathogenic disease outbreaks and facilitate the response to future outbreaks of concern.

Cong Ji[1*], Junbin (Jack) Shao[1*]
1 Liferiver Science and Technology Institute, Shanghai ZJ Bio-Tech Co., Ltd., Shanghai, China

*Corresponding author
Email: c_ji@liferiver.com.cn; junbin_shao@liferiver.com.cn

Abstract

Determining how viruses originated and diverged and how infectious diseases are transmitted are serious challenges. Surprisingly, some inter-species genetic distances among Coronaviridae were shorter than intra-species distances, possibly representing the intermediate states of different species or subspecies in the evolutionary history of Coronaviridae. The indels of whole genome sequences between different hosts were separately associated with new functions or turning points, clearly indicating their important roles in the host transmission and shifts of Coronaviridae. Furthermore, we believe that non-nucleosomal DNA may play dominant roles in the divergence of different lineages of SARS-CoV-2 in different regions of the world because of the lack of nucleosome protection. We suggest that strong selective variation among different lineages of SARS-CoV-2 is required to produce strong codon usage bias. Interestingly, we found that an increasing number of other types of substitutions, such as those resulting from the hitchhiking effect, have accumulated, especially in the pre-breakout phase. To predict potential epidemic outbreaks, we tested our strategy, Epi-Clock, which applies the ZHU algorithm on different SARS-CoV-2 datasets to search for real significant mutational accumulation patterns correlated with the outbreak events. We could accurately predict the potential pre-phase of outbreaks with a median interval of 5 days before the outbreaks. Using our pipeline, users may review updated information on the website https://bioinfo.liferiver.com.cn with easy registration. Therefore, we propose Epi-Clock, a sensitive platform similar to a clock that can signal the need to assist individuals at focal locations by using diagnostics, isolation control, vaccines or therapy at any time.

Keywords

Divergence, Host bias, Mutation type, Codon usage bias, ZHU algorithm

Introduction

Determining how viruses originated and diverged and how a disease was transmitted are extremely important for understanding viral disease outbreaks and facilitating responses to future outbreaks. Only 13.7% of the areas on Earth are suitable for human use, and the remaining areas are enriched with over millions of species of viruses, with an extremely large population scale and accumulations of genetic mutations, especially in the ocean. Viruses are the communication bridges between humans and nature, regardless of whether the effect is positive, so each communication process should be co-evolutionary. Since viruses lack fossilization, preventing the availability of any references or ancestors for inferring evolutionary processes, some theories about the origin of viruses have been reported, such as the degeneracy theory, DNA escape from plasmids or transposons, and viroid or satellite viruses[1]. The origin of new genes with novel functions creates genetic and phenotypic diversity in organisms, which gradually acquire increasing numbers of gene partners[2]. A non-parametric approach for estimating the date of origin of genetic variants in large-scale sequencing datasets has been developed, with allele age estimates providing a rapid approach for inferring the ancestry shared between individual genomes, determining genealogical relationships at different points in the past, and describing and exploring the evolutionary history of modern human populations[3]. However, viruses, agents on the edge of life, have the same genes as living organisms and undergo evolution by natural selection and reproduction by self-assembly but have no cell structure, metabolism, or space to reproduce except in the host cells. To study the macroevolution of viruses, several methods have been proposed, i.e., phylogenetic[4], neutral selection[5], and divergence[6] approaches. Most high-pathogenicity mutations go extinct in the population, but adaptive mutations can be fixed in the population, such that most intra-species mutations should be under neutral selection[5]. To answer the above crucial questions about outbreaks, sequencing of viral samples and supplementation of epidemiological methods could play an important role in providing nucleotide-level resolution data for outbreak-causing pathogens[7].

How does viral evolution occur across hosts? Most emerging infectious diseases in humans originate from animal reservoirs, and it is necessary to understand how and why some pathogens become capable of crossing host species barriers. Population genetic analyses provide estimates suggesting that the putative introduced genetic sequence within the RBD is evolving under directional selection, which may result in the adaptation of the virus to hosts. Unsurprisingly, the putative recombination region in the S protein was found to be highly diverse among strains from bats[8]. Emerging viral diseases are often the product of a host shift, where a pathogen jumps from its original host into a novel species. Host shifts are frequent in the evolution of most pathogens, but why pathogens successfully jump between some host species but not others is only just becoming clear[9]. Interestingly, the same sequence changes are often seen each time a virus infects a particular host. These changes may come at a cost to other aspects of the pathogen's fitness, and this may sometimes prevent host shifts from occurring[9]. Therefore, evolution could help viruses overcome species barriers,

principally by affecting virus–host interactions; however, evolving the capability for sustained transmission in a new host species represents a major adaptive challenge because the number of mutations required is often large[10]. Truly understanding disease emergence requires a new mechanistic and integrated view of the factors that allow or prevent viruses from spreading in novel hosts, suggesting that both ecological and genetic aspects of virus emergence could be placed within a simple population genetic framework[11]. The continuous struggle between viruses and their hosts to maintain at least a constant fitness level leads to the development of an unceasing arms race, where weapons are often shuttled between the participants[12]. Variables including genome length, genome type, and recombination frequency have little predictive power[13]. The emergence of these and many other human diseases occur when an established animal virus switches to human hosts and is subsequently transmitted within human populations, while transfers between different animal hosts lead to the analogous emergence of epizootic diseases. The importance of viral host switching was underscored by the recent avian epizootics of high-pathogenicity strains of H5N1 influenza A, in which hundreds of spillover human cases and deaths have been documented[14]. The movement of EBOV within the region and viral evolution during prolonged human-to-human transmission have been reported[15]. The specific amino acid substitutions in the EBOV GP have increased tropism for human cells while reducing tropism for bat cells. Such increased infectivity may have enhanced the ability of EBOV to transmit among humans and contributed to the wide geographic distribution of some viral lineages[16]. Hence, it seems reasonable to conclude that if there is a positive relationship between mutations in viruses and the possibility of adaptation to the host, the association between the divergence of viruses and adaptation allowing host transfer is under strong selection.

Remarkably, it has been reported that nucleosomes suppress spontaneous mutations in a base-specific manner in eukaryotes such that nucleosome occupancy nearly eliminated cytosine deamination, resulting in an ~50% decrease in the C->T mutation rate in nucleosomal DNA[17]. However, the organization of genomic DNA into defined nucleosomes, which has long been viewed as a hallmark of eukaryotes, has been challenged by the identification of “minimalist” histones in archaea and, more recently, by the discovery of genes that encode fused remote homologues of the four eukaryotic histones in Marseilleviridae, a subfamily of giant viruses that infect amoebae. That is, viral doublet histones are essential for viral infectivity, localize to cytoplasmic viral factories after virus infection and ultimately are found in mature virions[18]. It has also been reported that giant viruses belonging to the nucleocytoplasmic large DNA virus (NCLDV) group have histone-like genes in their genomes. One possibility is histone doublets in MVs, and a nucleosomal structure may have been acquired to function in the infected cell for protection of the viral genome from detection and degradation and/or as a means of interfering with host cell regulation during infection, which seems reasonable given that host–virus arms races are a powerful source of adaptation, and most genes gained by NCLDVs from their different hosts are likely linked to viral defences[19].

At present, the phylogenetic network faithfully traces routes of infections for documented coronavirus disease 2019 cases, indicating that such networks could likewise be successfully used to help trace undocumented SARS-CoV-2 infection sources, which could then be quarantined to prevent recurrent spread of the disease worldwide[20]. To compare virus genomes before epidemic and after epidemics, it is necessary to search for the accumulation pattern of pathogenic mutations contributing to the outbreak. Recently, it has been reported that the Global Epidemic and Mobility Model can provide an estimate of seasonal transmission potential by running a Monte Carlo maximum likelihood analysis to generate stochastic simulations of epidemic spread worldwide[21]. Epi-Factors, a manually curated database providing information about epigenetic regulators and their complexes, targets and products, contains information on 815 proteins, including 95 histones and protamines[22]. Some of these mutations have presumably arisen as a result of the virus evolving under immune system selection pressure in infected individuals, and at least one, lineage B.1.1.7, potentially resulted from a chronic infection[23]. It has been reported that SARS-CoV-2 Lambda, a new variant of interest, is now spreading in some South American countries. The spike protein of the Lambda variant is more infectious and was attributed to the T76I and L452Q mutations[24]. Because emerging pathogens have the potential to impose substantial mortality, morbidity and economic burdens on human populations, tracking the spread of infectious disease to assist in their control has traditionally relied on the analysis of case data gathered as the outbreak proceeds[25]. Based on the requirements and challenges, including the need for flexible platforms that generate sequence data in real time and for these data to be shared as widely and openly as possible[25], here, we demonstrate our strategy, with the aim of developing a sensitive process for predicting the potential trigger of outbreaks to facilitate the response to future outbreaks of concern.

Results

Divergence of the whole Coronaviridae family within different populations and adaptation to cross host species barriers.

To comprehensively explore the evolution of the whole family of Coronaviridae, we performed a genomic comparison of Coronaviridae to search for new age patterns related to virulence or transmissibility, as shown in Figure 1. In the analysis of intra-species and inter-species genetic distances of whole genome sequences of Coronaviridae, it was obvious that the intra-species genetic distances were lower, as shown in Figure 1a. For instance, the sequencing similarities of SARS-CoV-2 and bat-SL-CovZC45 and SARS-CoV-2and bat-SL-CovZXC21 were nearly 88%; however, the similarities of SARS-CoV-2and SARS-CoV and SARS-CoV-2and MERS were 79% and 50%, respectively. According to a comparison of the genetic distances of H1N1 and H3N2 with those of SARS-CoV-2and SARS-CoV, we found that the p-distance of SARS-CoV-2and Bat SARS-like CoV was nearly 0.13 and the p-distance of SARS-CoV-2and SARS-CoV was 0.24 in MEGA[26]. Both of these values were lower than the p-distances of H1N1 and H3N2 (0.8) and H1N1 and H7N9 (0.73). As shown in

Figure 1a and Supplementary Figure 1, it was obvious that most inter-species genetic distances of Coronaviridae were longer than intra-species distances, i.e., SARS-CoV-2, SARS-CoV, SADS, NL63, MERS, London1, HKU5, HKU4, HKU3, HKU2, and BATS. Conversely, some inter-species genetic distances of Coronaviridae were shorter than the intra-species genetic distances, such as those for OC43, HKU9, HKU5, HKU1, and 229E, which may be the intermediate states of different species or subspecies in the whole evolutionary history of Coronaviridae. The relationship between variants and frequency in different geographical regions has been characterized, and the value of using age information in interpreting variants of functional and selective importance, such as using allele age estimates to power a rapid approach for inferring the ancestry shared between individual genomes in order to quantify genealogical relationships, has been demonstrated[3]. Here, we explored the divergence time of species or sub-species of Coronaviridae, which presented the diversity of such ages within populations separately associated with new functions or important turning points. Interestingly, we explored the dated variants of different species or subspecies and generated an atlas of new ages with SARS-CoV-2 as the reference genome in Figure 1b and Supplementary Figure 2. It was demonstrated that SARS-CoV, HKU3, and BATS were very close to SARS-CoV-2 and far away from OC43 and HKU1 on the S protein, although there were 5 kb insertions of BATS in Orf7a, Orf7b, and Orf8 and 15 kb deletions of OC43, HKU9, and HKU1 in Orf7a, Orf7b, Orf8 and N. All populations show evidence of insertions, deletions or substitutions that have driven the divergence of the whole family of Coronaviridae in response to natural selection.

Most strikingly, as shown in Figure 1c and Supplementary Table 1, the insertions and deletions of whole genome sequences between different hosts should play an important role in host transmission and the shift of Coronaviridae presented by Circos. For instance, human coronavirus 229E in *Camelus* and *Vicugna pacos* have S protein sequences similar to those observed in *Homo sapiens*, while having 500 bp deletions in *Hipposideros* and *Macronycteris vittata*. When used as the reference genome, Rhinacovirus in *Sus scrofa* was similar to that in *Rhinolophus ferrumequinum* and divergent from that in *Rhinolophus affinis* in that insertions and deletions all appeared in the S protein, which may be related to host bias. It is likely that accumulations of insertions and deletions of alphacoronavirus, human coronavirus OC43, Middle East respiratory syndrome-related coronavirus, severe acute respiratory syndrome-related coronavirus, Deltacoronavirus, and avian coronavirus emerged on the S protein. The similarity of SARS-CoV-2 S proteins in *Rhinolophus sinicus*, *Rhinolophus macrotis*, *Canis lupus familiaris*, and *Panthera tigris jacksoni* was very high with the SARS-CoV-2 S protein in *Homo sapiens*, which means that the ancestral hosts of SARS-CoV-2 in *Homo sapiens* should be close related to the species as detailed in Supplementary Figures 3-10. Wildlife host species richness has been proposed as an important predictor of disease emergence. Likewise, host populations with low biodiversity might harbour fewer viruses and an increased risk of emergence. Conversely, high host biodiversity has also been linked to a decrease in disease risk through the 'dilution effect'[11]. Advances in deep sequencing have made it possible to measure within-host genetic

diversity in both acute and chronic influenza infections[27]. Factors such as antigenic selection, antiviral treatment, tissue specificity, spatial structure, and multiplicity of infection may affect how influenza viruses evolve within human hosts[27]. In summary, the more mutations viruses have, the greater the possibility of adaptation allowing them to cross host species barriers, which indicates that increased host diversity will protect against the emergence of pathogens.

The distribution of rates of different mutation types and codon usage bias in different lineages of SARS-CoV-2.

Pathogens have always been a major cause of human mortality, so they impose strong selection pressure on the human genome[28]. Across different lineages of SARS-CoV-2, most nucleotide substitutions were found in the regions of the S gene and N gene as presented in Figure 2a and Supplementary Figure 11, which clearly demonstrates that accumulations of mutations drove the evolutionary history of different lineages of SARS-CoV-2. Then, we demonstrated the richness distribution of different mutation types across the whole genomes, in which mutation rates of C->T, G->T and T->C were dominant in driving the divergence of different lineages of SARS-CoV-2, as shown in Figure 2b and Supplementary Figures 12-17. In particular, the mutation rate distributions of C->T were highly enriched in all lineages of SARS-CoV-2, and the mutation rate distributions of G->T were relatively rich in B.1.1.318, B.1.1.529 (Omicron), B.1.617.1 (Kappa), B.1.617.3, B.1.621 (Mu), B.1.640.1, B.1.640.2, C.1.2, C.37 (Lambda), and P.3 (Theta). Conversely, the patterns of A->T, T->G, G->C were relatively rare, and it was clear that some points with high mutation rates in the A->T distribution came from B.1.1.7 (Alpha) and B.1.621 (Mu); several points with high mutation rates in the T->G distribution came from B.1.526 (Lota), B.1.617.1 (Kappa), B.1.617.2 (Delta), and C.36.3; and some points in the G->C distribution came from B.1.525 (Eta), B.1.617.2 (Delta), B.1.617.3, B.1.621 (Mu), B.1.640.1 and B.1.640.2. Therefore, we believe that nucleosomal DNA undergoes fewer C->T mutations in SARS-CoV-2. This DNA is protected by the nucleosomal structure, with viral doublet histones essential for viral infectivity and viral defences leading to multiple host adaptation/transmission. In contrast, non-nucleosomal DNA may play dominant roles in the divergence of various lineages of SARS-CoV-2 in different regions of the world without the protection of nucleosomes.

Furthermore, we presented the recent amino acid substitutions among different SARS-CoV-2 lineages, including B.1.617.2 (Delta), B.1.1.529 (Omicron), B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), C.37 (Lambda), B.1.621 (Mu), C.1.2, B.1.525 (Eta), B.1.526 (Lota), B.1.617.1 (Kappa), B.1.617.3, P.3 (Theta), B.1.1.318, C.36.3, B.1.640.1, and B.1.640.2, with a focus on NSP2, NSP3, NSP5, NSP12, NSP13, Spike, NS3, NS8, and N proteins in Figure 2c and Supplementary Table 1, which have strong effects on the divergence and evolutionary history of different lineages. This is especially true for the Spike protein, which contains a receptor-binding domain (RBD), a fusion domain and a transmembrane domain, and the NSP3 protein that is the N-terminus of coronavirus SARS-CoV non-structural protein 3 (Nsp3) and related proteins.

Interestingly, compared with the dominant amino acid substitutions, i.e., NSP12:P323L and S:D614G, there seem to be unique and specific amino acid substitutions in one lineage relative to another lineage. Examples include the NSP2:P129L, NSP2:E272G, NSP2:L462F, S:E154Q, S:I210T, and S:D936H substitutions in the B.1.640.1 lineage; NSP13:V371A, S:P129L, S:D1139H, and NS8:C37R substituted in the B.1.640.2 lineage; NSP2:S358L, and NSP3:E378V substitutions in the B.1.1.318 lineage; NSP3:A41V, NSP3:D821Y, NSP13:D105Y, S:S12F, S:G212V, and S:A899S substitutions in the C.36.3 lineage; S:K2Q, S:L280F, S:G313S, S:A368V, S:D736G, S:E1092K, and S:H1101Y substitutions in the P.3 (Theta) lineage; NSP3:A1T, S:Q779K, S:E1072K, NS8:T26I, and N:P67S substitutions in the B.1.617.3 lineage; NSP13:G206C, NSP13:M429I, S:Q1071H, and NS3:L15F in the B.1.617.1 (Kappa) lineage; NSP13:Q88H, S:D253G, NS3:P42L, N:P199L, N:M234I, S:D253G, NS3:P42L, N:P199L, and N:M234I substitutions in the B.1.526 (Lota) lineage; NSP3:T1198I, S:Q52R, S:P323F, S:Q677H, NS3:T89I, NS3:S92L, NS3:G172C, and N:A12G substitutions in the B.1.525 (Eta) lineage; NSP3:P822S, S:P25L, S:C136F, S:A879T, NS3:T223I, and NS3:D238Y substitutions in the C.1.2 lineage; NSP3:T237A, NSP3:A562T, NSP3:T720I, NSP13:P419S, S:Y146S, S:H147T, S:T205I, S:R346K, NS8:P38S, and NS8:S67F substitutions in the B.1.621 (Mu) lineage; S:P13L, S:G75V, S:V76I, and S:T428I substitutions in the C.37 (Lambda) lineage; S:L18F, S:T20N, S:P26S, S:D138Y, S:S247G, S:E341D, S:S371L, S:K977Q, and S:T1027I substitutions in the P.1 (Gamma) lineage; NSP3:K837N, S:D80A, NS3:L52F, and NS3:S171L substitutions in the B.1.351 (Beta) lineage; NSP3:T183I, NSP3:A890D, S:A570D, S:S982A, S:D1118H, NS8:R52I, NS8:K68Stop, NS8:Y73C, and N:S235F substitutions in the B.1.1.7 (Alpha) lineage; NSP3:K38R, NSP3:M84V, NSP3:V1069I, S:G339D, S:S371L, S:S373P, S:S375F, S:N440K, S:G446S, S:S477N, S:Q493R, S:G496S, S:Q498R, S:Y505H, S:T547K, and S:N764K substitutions in the B.1.1.529 (Omicron) lineage; and NSP3:A488S, NSP12:P227S, and S:P77L substitutions in the B.1.617.2 (Delta) lineage. Provided that the amino acid composition of the proteomes reflects the action of natural selection to enhance metabolic efficiency, synonymous codon usage bias as a measure of translation rates and shows increases in the abundance of less energetically costly amino acids in highly expressed proteins[31]. Comparison of the relative fixation rates of synonymous (silent) and nonsynonymous (amino acid-altering) mutations provides a means for understanding the mechanisms of molecular sequence evolution[29], and positive selection might be expected to follow viral emergence in a new host species because this event would obviously entail a major boost in the number of susceptible hosts and a concomitant increase in fitness[30]. The compiled codon usage data for different lineages of SARS-CoV-2 are presented in Supplementary Table 1. We summarize the synonymous codon usage of all different lineages of SARS-CoV-2 in Figure 2d and Supplementary Table 1. In the B.1.640.2 lineage, codon AGU was obviously biased for Ser, codon UUU was stronger for Phe in the B.1.640.1 and P.3 (Theta) lineages, and codon GCU was clearly biased for Ala in the B.1.617.2 (Delta) lineage. That is, a large selective difference between lineages of SARS-CoV-2 was required for strong codon usage bias.

Clock-like prediction of focal outbreak points worldwide to provide warnings.
To explain the epidemic outbreak points related to key mutations, we set the Epi-Clock device to predict potential epidemics and assist in the presentation of detailed mutation information for severely affected areas. Therefore, we analysed the whole evolutionary pathway in the timeline of lineages of SARS-CoV-2 presented in Figure 3a. and Figure 3b; i.e., each lineage followed the same principle in which accumulations of mutations were the driving force of the phylogenetic relationships among different lineages of SARS-CoV, showing information for the earliest publicly collected samples in different regions of the world. At the same time, we summarized the smoothed distribution of new cases per million cases of SARS-CoV-2 in different severely affected areas, such as Africa, Asia, Europe, North America, Oceania, and South America, along the timeline from Our World in Data (OWID) [31] in Figure 3c and Supplementary Figures 18-24. Interestingly, we found the pattern of an increasing number of other types of substitutions (shown with an asterisk) as the hitchhiking effect progressed, especially in the pre-breakout phase, even though some substitutions were replaced by other dominant genotypes, as shown in the red box in Figure 3d, Supplementary Table 2 and Supplementary Figures 25-32.

Hypothesized that specific amino acid substitutions would be triggers for outbreaks. To predict potential epidemic outbreaks, we proposed the ZHU algorithm presented in Figure 4a and tested it on different true sets of SARS-CoV-2 data before outbreaks to search for real significant mutational accumulation patterns related to the outbreak events. To avoid system errors, that is, errors caused by regions varying in other epi-factors such as population density, geographical environment, vaccination coverage rates and national/social rules and norms, we extracted the true sets by the control group in different individual countries or regions as the individual related values of the new cases per million. Then, we separately performed generalized linear model (GLM) analysis on individual countries/regions to find the optimal mutational patterns as in the training phase. Very interestingly, it was obvious that amino acid substitution type X13.NSP3_T183I.I was a potential epi-factor in Iran, the Netherlands and Norway. X74.S_S371L.L was significantly correlated with numbers of new cases in Finland, Canada and Brazil, as shown in Figure 4b. Across N generations of training by GLM and reordering, we found 171 significant substitutions as potential epi-factors within 55 different countries and regions, as demonstrated in Supplementary Table 3. Finally, we successfully presented the performance of the ZHU algorithm on 42 validation sets in Figure 4c. From most validations, we could accurately predict the potential pre-phase of the outbreak by ZHU prediction. Nevertheless, in only four countries, our model could not be fitted because of unavailable sequencing samples in the outbreak phase, such as in Jamaica, Kyrgyzstan, Libya, and Palestine. By counting the number of true instances, we summarized the precision, recall, and accuracy of Epi-Clock with the true validation sets as presented in Figure 4d and Supplementary Table 3, where the median interval before outbreaks was 5 days.

Materials and Methods

To explore divergence of the whole family of Coronaviridae, we computed intra-species and inter-species p-distances of whole genome sequences by MEGA11 and plotted the distances by the boxplot function of R. To illustrate the atlas of new ages within different populations, we applied a genome alignment-based pipeline to infer the origin time of a given genomic region by a 6 bp sliding window with numbers of dated variants, including insertions in red and deletions in blue. We scanned the whole genome sequences with 6 bp sliding windows and summarized the mean values of mutation rates of different mutation types. According to the reference genomes among different hosts, we demonstrated host shifts within 6 bp sliding windows in the evolution of Coronaviridae and presented them by Circos.

Similarly, we demonstrated nucleotide mutations and amino acid substitutions among different SARS-CoV-2 lineages. We computed mutation rates by sliding windows for different mutation types. Then, we presented the 144 amino acid substitution distributions of different lineages by GeneWise. The codon usage numbers were converted into relative synonymous-codon usage (RSCU) values[28], which was simply the observed frequency of a codon divided by the frequency expected under the assumption of equal usage of the synonymous codons for an amino acid[28] [32].

$$\mathrm{RSCU_{ij}} = \frac{\mathrm{Xij}}{1/\mathrm{ni}} \sum_{j=1}^{ni} Xij$$

A neighbour-joining phylogenetic tree of separate lineages of SARS-CoV-2 was constructed by MEGA v11.0.11. To predict the potential epidemic mutation patterns in the severely affected areas, we summarized testing data from OWID until Feb 7th in 2022 that were continually collated from official government sources worldwide. We plotted the frequency distribution of confirmed cases of SARS-CoV-2 in different severely affected areas, such as Africa, Asia, Europe, North America, Oceania, and South America, along the timeline. At the same time, we set the baseline of new cases per millions of substitutions as the internal control group and the widespread types of substitutions as the external control group so that we could exclude other effects on epidemic outbreak events. We excluded areas without any outbreak time, as labelled by the red box, and searched the mutation patterns appearing to be related to the outbreak until the appearance of the Omicron variant because this corresponded with the time of universal immunization. To ensure the true values of the outbreak of new cases of SARS-CoV-2, we carefully detected peaks following strict standards; that is, the values of new cases per million had to be above 30. Then, we separately split the sets of population samples from 1 day to 30 days before outbreaks based on the same location according to 144 amino acid substitutions or deletions shown in Figure 2c and Supplementary Table 2. The dataset contained 13740300 observations, including 6300 true sets and 2181 features, which amounted to 117 different countries or regions on all five continents, divided into the 75 training sets and 42 validation sets listed in Supplementary Table 3. Then, we performed GLM analysis on individual countries/regions to find the optional mutational patterns as in the training phase.

Across N generations of training by GLM and reordering, we found 171 significant substitutions as potential epi-factors within 55 different countries and regions. Finally, we assessed the ZHU prediction by 42 validation sets with precision, recall, and accuracy.

Discussion

The epidemic viruses did not spread into neighbouring countries showing susceptibility to significant outbreaks but a lower risk of introduction, which revealed the large epidemic to be a heterogeneous and spatially dissociated collection of transmission clusters of varying size, duration and connectivity[33]. Through the analysis of over 900 full genomes from an epidemiological collection of more than 26000 strains from Australia and New Zealand, fundamental differences in the phylodynamics of two co-circulating lineages of influenza B virus were determined by examining the complex relationships between virus transmission, age of infection and receptor binding preference, which are new factors that are important determinants of influenza B evolution and epidemiology[34]. In short, parallel adaptation occurs in the context of several viral behaviours, including cross-species transmission, drug resistance, and host immune escape, and its existence suggests that at least some aspects of virus evolution and emergence are repeatable and predictable[35]. The development of a combination of pathogenic epidemic prediction and pathogenic testing strategies not only improved the containment of COVID-19 but also contributed to the control of other infectious diseases, such as HIV and human papillomavirus, and non-communicable diseases[36].

Rarely, viruses gain the ability to spread efficiently within a new host that was not previously exposed or susceptible, which involves either increased exposure or the acquisition of variations that allow them to overcome barriers to infection of the new host. It is well known that host switching leads to viral emergence, considering the evolutionary mechanisms, virus–host interactions, host range barriers to infection, and processes that allow efficient host-to-host transmission in the new host population[14]. For instance, at least 49 human-to-swine transmission events occurred globally during 2009-2011, highlighting the ability of the H1N1pdm09 virus to transmit repeatedly from humans to swine, even following adaptive evolution in humans[37]. Whether hosts and their symbionts speciate in parallel, by co-speciation, or through host shifts is a key issue in host-symbiont evolution. Convincing cases of co-speciation are rare (7%), and examining the relationships between short-term co-evolutionary dynamics and long-term patterns of diversification in host-symbiont associations revealed that these events could occur following host shifts and do not necessarily involve co-speciation. Therefore, there is now substantial evidence that the co-evolutionary dynamics of hosts and parasites do not favour long-term co-speciation[38]. Above all, host range is a viral property reflecting natural hosts that are infected either as part of a principal transmission cycle or, less commonly, via spillover infections into alternative hosts. Wildlife host species richness has been proposed as an important predictor of disease emergence[11].

Host immunity is a major driver of pathogen evolution and a major determinant of pathogen diversity; however, the complexity of many host–pathogen interactions is dynamically important, broadening the definition of a pathogen's immunological phenotype or what can be thought of as its immunological niche[39]. Recently, it was reported that a higher rate of severe outcomes and considerable mortality exist in unvaccinated people, especially older adults. That is, the strict and comprehensive pandemic control strategies implemented in Shanghai were able to reduce the number of people infected and to support early diagnosis and appropriate treatment for severe COVID-19 so that the case fatality rate could be minimized and to buy time for full vaccination coverage[40]. The advent of high-content single-cell technologies has facilitated a greater understanding of the properties of host cells harbouring infection, the host's pathogen-specific immune responses, and the mechanisms pathogens have evolved to escape host control[41] [42]. It is important to identify natural genetic variants that underlie variation in the host innate immune response to infection and analyse the mechanisms by which such variants alter these responses[43]. This work provided important insights into how natural selection has shaped immunity and host defence genes in specific human populations and in the human species as a whole, information that is helping delineate genes that were important for host defence and increase our understanding of how past selection had an impact on disease susceptibility in modern populations[28].

Conclusions

With the accumulation of sequencing datasets, the performance of our strategy will improve, and the approach will become more sensitive in its ability to predict the potential trigger of epidemic outbreaks, which will ultimately facilitate responses to future outbreaks of concern.

Acknowledgements
Appreciation goes to Chen Si, Liu Yan, and Zhang Hanyan for their valuable suggestions and comments on this work. Many facets of the user-interface design benefited from help by Niu Xingsheng, Pan Yajie and Lu Wang. Great thanks to Elizabeth Sung for flow and language corrections. All the other data supporting the findings of this study and the computational code used in this study are available from the corresponding authors upon reasonable request. The other authors declare no competing interests.

FIGURE AND TABLE LEGENDS

Fig. 1. Demonstration within different populations in Coronaviridae. 1a. Divergence of the whole family of Coronaviridae and intra-species and inter-species genetic distances. **1b.** Atlas of new ages within different populations. **1c.** Host shifts in the evolution of Coronaviridae.

Fig. 2. The distribution of mutation rates for different mutation types and codon usage bias of different lineages of SARS-CoV-2. 2a. The percentage nucleotide identity of different lineages in the whole genome of SARS-CoV-2. **2b.** Different lineages are represented by different mutation types. **2c.** Amino acid substitutions among different SARS-CoV-2 lineages. **2d.** RSCU values in different lineages of SARS-CoV-2.

Fig. 3. Illustration of mutation information for different lineages of SARS-CoV-2 in severely affected areas. 3a. Phylogenetic tree of separate lineages of SARS-CoV-2. **3b.** Global distribution of lineages of SARS-CoV-2. **3c.** Distribution of confirmed cases of SARS-CoV-2 in different severely affected areas, such as Africa, Asia, Europe, North America, Oceania, and South America, along the timeline. **3d.** Frequency distribution of types of amino acid substitutions in the pre-breakout phase.

Fig. 4. Algorithm and performance of Epi-Clock. 4a. ZHU algorithm for inferring the timing of an epidemic of a new lineage related to increasing confirmed cases. **4b.** Prediction of 22 significant substitutions as the potential trigger for the outbreak from 75 training sets. **4c.** Performance of ZHU prediction measured by the relationship between observed confirmed numbers of new cases per million and normalized predicted numbers in different regions. **4d.** Performance of Epi-Clock evaluated by 42 validation sets.

Supplementary Figure 1. Intra-species and inter-species genetic distances of Coronaviridae. Coronaviridae included SARS-CoV-2, SARS-CoV, SADS, OC43, NL63, MERS, London1, HKU9, HKU5, HKU4, HKU3, HKU2, HKU1, BATS, and 229E.

Supplementary Figure 2. Atlas of new ages within different populations. Here, we

separately present the divergence of species or subspecies from SARS-CoV-2, i.e., SARS-CoV, SADS, OC43, NL63, MERS, London1, HKU9, HKU5, HKU4, HKU3, HKU2, HKU1, BATS and 229E. Red represents insertions, and blue represents deletions.

Supplementary Figure 3. The indels of whole genome sequences of human coronavirus 229E between different hosts. Several hosts are shown, such as (from outside to inside) *Camelus*, *Hipposideros*, *Macronycteris vittata*, and *Vicugna pacos*, while the genome of *Homo sapiens* was used as the reference genome.

Supplementary Figure 4. The indels of whole genome sequences of Rhinacovirus between different hosts. Several hosts are shown, such as (from outside to inside) *Rhinolophus ferrumequinum* and *Rhinolophus affinis*, while the genome of *Sus scrofa* was used as the reference genome.

Supplementary Figure 5. The indels of whole genome sequences of Tegacovirus between different hosts. Several hosts are shown, such as (from outside to inside) *Felis catus*, *Canis lupus familiaris*, and Feliformia, while the genome of *Sus scrofa* was used as the reference genome.

Supplementary Figure 6. The indels of whole genome sequences of human coronavirus OC43 between different hosts. Several hosts are shown, such as (from outside to inside)*Bos taurus*, *Bos grunniens*, Canis lupus familiaris, *Pan troglodytes verus*, *Sus scrofa*, *Hydropotes inermis*, *Bubalus bubalis*, *Camelus bactrianus*, *Equus caballus*, *Kobus ellipsiprymnus*, *Odocoileus virginianus*, *Rusa unicolor*, Bovidae, *Giraffa camelopardalis*, and *Vicugna pacos*, while the genome of *Homo sapiens* was used as the reference genome.

Supplementary Figure 7. The indels of whole genome sequences of Middle East respiratory syndrome-related coronavirus between different hosts. Several hosts are shown, such as (from outside to inside)*Camelus dromedaries*, *Lama glama*, *Neoromicia capensis*, *Hypsugo savii*, *Pipistrellus kuhlii*, and *Vespertilio sinensis*, while the genome of *Homo sapiens* was used as the reference genome.

Supplementary Figure 8. The indels of whole genome sequences of severe acute respiratory syndrome coronavirus 2 between different hosts. Several hosts are shown, such as (from outside to inside)*Mus musculus*, *Rhinolophus sinicus*, *Mustela lutreola*, *Rhinolophus macrotis*, *Canis lupus familiaris*, *Panthera tigris jacksoni*, *Rhinolophus affinis*, *Aselliscus stoliczkanus*, *Rhinolophus ferrumequinum*, Chiroptera, *Rhinolophus pusillus*, *Chaerephon plicatus*, *Chlorocebus aethiops*, *Paradoxurus hermaphroditus*, Viverridae, *Paradoxurus hermaphroditus*, and *Paguma larvata*, while the genome of *Homo sapiens* was used as the reference genome.

Supplementary Figure 9. The indels of whole genome sequences of

Deltacoronavirus between different hosts. Several hosts are shown, such as (from outside to inside) *Pycnonotus jocosus*, *Mareca*, *Lonchura striata*, Zosteropidae, *Falco*, *Chlamydotis*, Columbidae, Galliformes, Passeridae, Muscicapidae, *Turdus hortulorum*, Ardeidae, *Gallinula chloropus*, and *Coturnix japonica*, while the genome of *Sus scrofa* was used as the reference genome.

Supplementary Figure 10. The indels of whole genome sequences of avian coronavirus between different hosts. Several hosts are shown, such as (from outside to inside) Phasianinae, Anatidae, and *Meleagris gallopavo*, while the genome of *Gallus gallus* was used as the reference genome.

Supplementary Figure 11. The whole-genome percentage nucleotide identity of different lineages of SARS-CoV-2. There were several lineages such as P.3(Theta), P.1(Gamma), C.37(Lambda), C.36.3, C.1.2, B.1.640.2, B.1.640.1, B.1.621(Mu), B.1.617.3, B.1.617.2(Delta), B.1.617.1(Kappa), B.1.526(Lota), B.1.525(Eta), B.1.351(Beta), B.1.1.7(Alpha), B.1.1.529(Omicron), and B.1.1.318.

Supplementary Figure 12. The richness distribution of the mutatiion rate of C->T across the whole SARS-CoV-2 genome in different lineages. There were several lineages such as P.3(Theta), P.1(Gamma), C.37(Lambda), C.36.3, C.1.2, B.1.640.2, B.1.640.1, B.1.621(Mu), B.1.617.3, B.1.617.2(Delta), B.1.617.1(Kappa), B.1.526(Lota), B.1.525(Eta), B.1.351(Beta), B.1.1.7(Alpha), B.1.1.529(Omicron), and B.1.1.318.

Supplementary Figure 13. The richness distribution of the mutation rate of G->T across the whole SARS-CoV-2 genome in different lineages. The same as above.

Supplementary Figure 14. The richness distribution of the mutation rate of A->T across the whole SARS-CoV-2 genome in different lineages. The same as above.

Supplementary Figure 15. The richness distribution of the mutation rate of T->C across the whole SARS-CoV-2 genome in different lineages. The same as above.

Supplementary Figure 16. The richness distribution of the mutation rate of T->G across the whole SARS-CoV-2 genome in different lineages. The same as above.

Supplementary Figure 17. The richness distribution of the mutation rate of G->C across the whole SARS-CoV-2 genome in different lineages. The same as above.

Supplementary Figure 18. Smoothed distribution of new cases per million cases of SARS-CoV-2 in different severely affected areas, such as Africa, Asia, Europe, North America, Oceania, and South America, along the timeline.

Supplementary Figure 19. Smoothed distribution of new cases per million cases of

SARS-CoV-2 in Africa along the timeline.

Supplementary Figure 20. Smoothed distribution of new cases per million cases of SARS-CoV-2 in Asia along the timeline.

Supplementary Figure 21. Smoothed distribution of new cases per million cases of SARS-CoV-2 in South America along the timeline.

Supplementary Figure 22. Smoothed distribution of new cases per million cases of SARS-CoV-2 in Europe along the timeline.

Supplementary Figure 23. Smoothed distribution of new cases per million cases of SARS-CoV-2 in North America along the timeline.

Supplementary Figure 24. Smoothed distribution of new cases per million cases of SARS-CoV-2 in Oceania along the timeline.

Supplementary Figure 25. Frequency distribution of different types for substitutions 1-20.

Supplementary Figure 26. Frequency distribution of different types for substitutions 21-40.

Supplementary Figure 27. Frequency distribution of different types for substitutions 41-60.

Supplementary Figure 28. Frequency distribution of different types for substitutions 61-80.

Supplementary Figure 29 Frequency distribution of different types for m substitutions 81-100.

Supplementary Figure 30. Frequency distribution of different types for substitutions 101-120.

Supplementary Figure 31. Frequency distribution of different types for substitutions 121-140.

Supplementary Figure 32. Frequency distribution of different types for substitutions 141-158.

Supplementary Table 1. Comprehensive information on Coronaviridae, especially different lineages of SARS-CoV-2. Sheet "Plague" contains data on the outbreak of epidemic disease in human history. Sheet "Cov" contains the taxonomy of

Coronaviridae according to the NCBI. Sheet "branch" contains the studied groups related to the outgroup on the evolutionary branch of Coronaviridae. Sheet "host" contains the basic information of hosts of Coronaviridae. Sheet "AA substitutes" contains specific amino acid substitutions with different mutation types in different lineages of SARS-CoV-2. Sheet "Codon Usage" contains the codon numbers of different lineages of SARS-CoV-2. Sheet "RSCU" contains RSCU values for different lineages of SARS-CoV-2. Sheet "lineages" contains the global distribution of lineages of SARS-CoV-2.

Supplementary Table 2. The comprehensive information of true sets. Sheets include the distribution of a total of 144 amino acid substitutions in different lineages, new cases per million from different countries/regions on different continents, and the statistics of different mutation types of amino acid substitutes.

Supplementary Table 3. Comprehensive information of ZHU prediction. Sheets include 75 training sets and 43 validation sets, the information on the input data, GLM coefficient estimates and AIC values, 171 significant substitutions, prediction results for the training sets, GLM and reordered processing data, the ZHU prediction model, prediction results for the validation sets, and performance parameters.

References

1. Forterre, P. and D. Prangishvili, *The origin of viruses.* Research in Microbiology, 2009. **160**(7): p. 466-472.
2. Zhang, W., et al., *New genes drive the evolution of gene interaction networks in the human and mouse genomes.* Genome Biology, 2015. **16**(1): p. 202.
3. Albers, P.K. and G. McVean, *Dating genomic variants and shared ancestry in population-scale sequencing data.* PLoS Biol, 2020. **18**(1): p. e3000586.
4. Kitchen, A., L.A. Shackelton, and E.C. Holmes, *Family level phylogenies reveal modes of macroevolution in RNA viruses.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(1): p. 238-243.
5. Vieira, M.C., D. Zinder, and S. Cobey, *Selection and Neutral Mutations Drive Pervasive Mutability Losses in Long-Lived Anti-HIV B-Cell Lineages.* Molecular biology and evolution, 2018. **35**(5): p. 1135-1146.
6. Simmonds, P. and P. Aiewsakun, *Virus classification - where do you draw the line?* Archives of virology, 2018. **163**(8): p. 2037-2046.
7. Wohl, S., S.F. Schaffner, and P.C. Sabeti, *Genomic Analysis of Viral Outbreaks.* Annual review of virology, 2016. **3**(1): p. 173-195.
8. Zhu, Z., K. Meng, and G. Meng, *Genomic recombination events may reveal the evolution of coronavirus and the origin of SARS-CoV-2.* Scientific Reports, 2020. **10**(1): p. 21617.
9. Longdon, B., et al., *The evolution and genetics of virus host shifts.* PLoS Pathog, 2014. **10**(11): p. e1004395.
10. Kuiken, T., et al., *Host species barriers to influenza virus infections.* Science, 2006. **312**(5772): p. 394-7.
11. Geoghegan, J.L. and E.C. Holmes, *Predicting virus emergence amid evolutionary noise.* Open biology, 2017. **7**(10): p. 170189.
12. Kaján, G.L., et al., *Virus–Host Coevolution with a Focus on Animal and Human DNA Viruses.* Journal of Molecular Evolution, 2020. **88**(1): p. 41-56.
13. Geoghegan, J.L., et al., *Virological factors that increase the transmissibility of emerging human viruses.* Proceedings of the National Academy of Sciences, 2016. **113**(15): p. 4170-4175.
14. Parrish, C.R., et al., *Cross-species virus transmission and the emergence of new epidemic diseases.* Microbiol Mol Biol Rev, 2008. **72**(3): p. 457-70.
15. Park, D.J., et al., *Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone.* Cell, 2015. **161**(7): p. 1516-1526.
16. Urbanowicz, R.A., et al., *Human Adaptation of Ebola Virus during the West African Outbreak.* Cell, 2016. **167**(4): p. 1079-1087.e5.
17. Chen, X., et al., *Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes.* Science, 2012. **335**(6073): p. 1235-8.
18. Liu, Y., et al., *Virus-encoded histone doublets are essential and form nucleosome-like structures.* Cell, 2021. **184**(16): p. 4237-4250 e19.
19. Vannini, A. and I. Marazzi, *A small nucleosome from a weird virus with a fat genome.* Mol Cell, 2021. **81**(17): p. 3447-3448.
20. Forster, P., et al., *Phylogenetic network analysis of SARS-CoV-2 genomes.* Proceedings of the National Academy of Sciences, 2020. **117**(17): p. 9241.

21. Tizzoni, M., et al., *Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm.* BMC medicine, 2012. **10**: p. 165-165.
22. Medvedeva, Y.A., et al., *EpiFactors: a comprehensive database of human epigenetic factors and complexes.* Database (Oxford), 2015(bav067).
23. Kemp, S.A., et al., *Recurrent emergence and transmission of a SARS-CoV-2 spike deletion H69/V70.* bioRxiv, 2021: p. 2020.12.14.422555.
24. Kimura, I., et al., *SARS-CoV-2 Lambda variant exhibits higher infectivity and immune resistance.* bioRxiv, 2021: p. 2021.07.28.454085.
25. Grubaugh, N.D., et al., *Tracking virus outbreaks in the twenty-first century.* Nature Microbiology, 2019. **4**: p. 10-19.
26. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Molecular biology and evolution, 2018. **35**(6): p. 1547-1549.
27. Xue, K.S., et al., *Within-Host Evolution of Human Influenza Virus.* Trends Microbiol, 2018. **26**(9): p. 781-793.
28. Barreiro, L.B. and L. Quintana-Murci, *From evolutionary genetics to human immunology: how selection shapes host defence genes.* Nat Rev Genet, 2010. **11**(1): p. 17-30.
29. Yang, Z., et al., *Codon-substitution models for heterogeneous selection pressure at amino acid sites.* Genetics, 2000. **155**(1): p. 431-49.
30. Holmes, E.C., *The evolution of viral emergence.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(13): p. 4803.
31. Hasell, J., et al., *A cross-country database of COVID-19 testing.* Scientific Data, 2020. **7**(1): p. 345.
32. Sharp Pm Fau - Li, W.H. and W.H. Li, *An evolutionary perspective on synonymous codon usage in unicellular organisms.* J Mol Evol., 1986. **24**(1-2): p. 28-38.
33. Dudas, G., et al., *Virus genomes reveal factors that spread and sustained the Ebola epidemic.* Nature, 2017. **544**(7650): p. 309-315.
34. Vijaykrishna, D., et al., *The contrasting phylodynamics of human influenza B viruses.* Elife., 2015. **4**: p. e05055.
35. Gutierrez, B., M. Escalera-Zamudio, and O.G. Pybus, *Parallel molecular evolution and adaptation in viruses.* Curr Opin Virol., 2019. **34**: p. 90-96.
36. Han, X., et al., *SARS-CoV-2 nucleic acid testing is China's key pillar of COVID-19 containment.* The Lancet, 2022. **399**(10336): p. 1690-1691.
37. Nelson, M.I., et al., *Global transmission of influenza viruses from humans to swine.* The Journal of general virology, 2012. **93**(Pt 10): p. 2195-2203.
38. de Vienne, D.M., et al., *Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution.* New Phytol, 2013. **198**(2): p. 347-385.
39. Cobey, S., *Pathogen evolution and the immunological niche.* Annals of the New York Academy of Sciences, 2014. **1320**(1): p. 1-15.
40. Zhang, X., W. Zhang, and S. Chen, *Shanghai's life-saving efforts against the current omicron wave of the COVID-19 pandemic.* The Lancet, 2022.
41. Chattopadhyay, P.K., M. Roederer, and D.L. Bolton, *A deadly dance: the choreography of host–pathogen interactions, as revealed by single-cell technologies.* Nature Communications, 2018. **9**(1): p. 4638.

42. Kazer, S.A.-O., et al., *Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection.* Nature Medicine, 2020. **26**(4): p. 511-518.
43. Lee, M.N., et al., *Common genetic variants modulate pathogen-sensing responses in human dendritic cells.* Science, 2014. **343**(6175): p. 1246980.

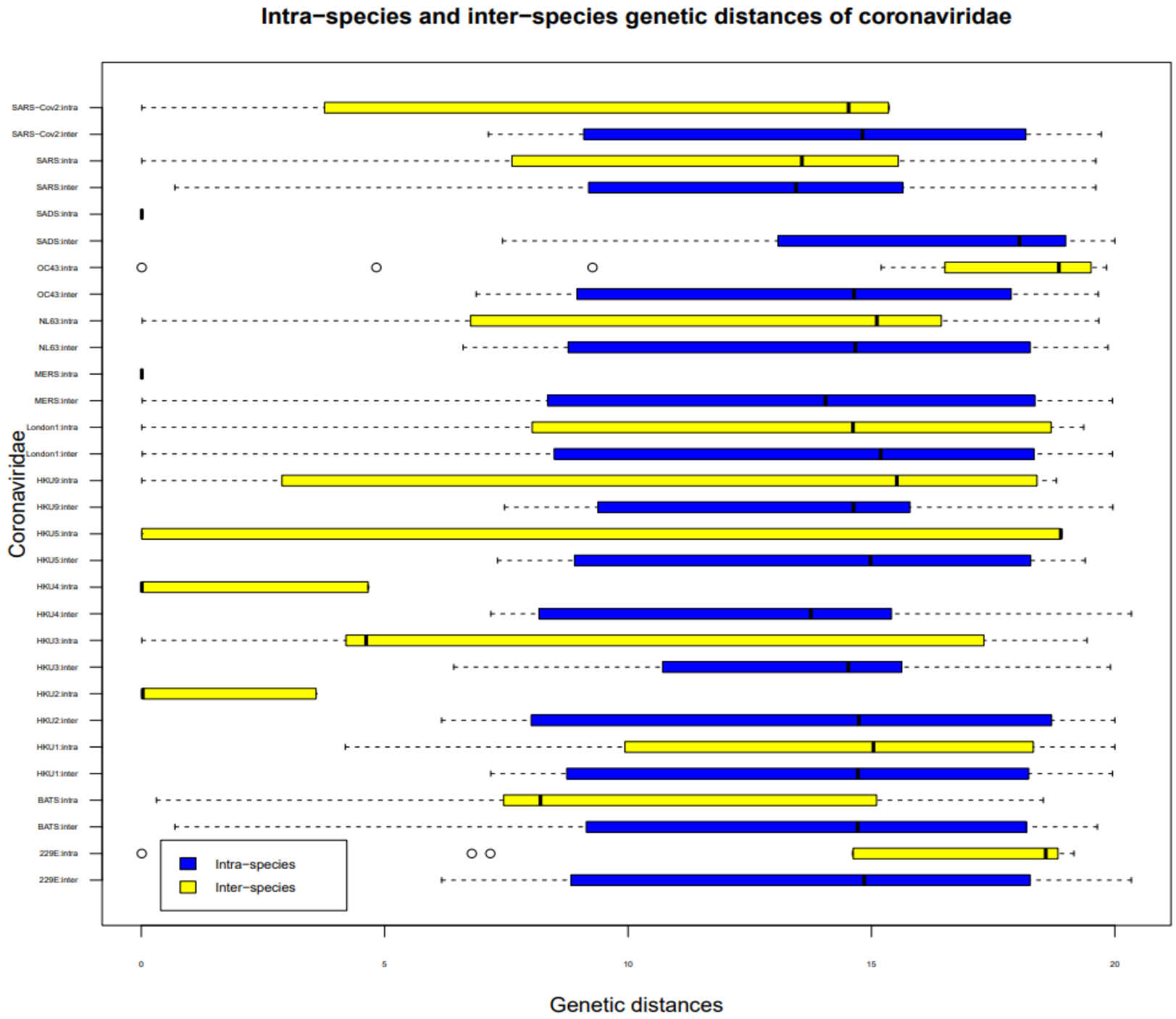

Intra−species and inter−species genetic distances of coronaviridae
SARS−Cov2:intra
SARS−Cov2:inter
SARS:intra
SARS:inter
SADS:intra
SADS:inter
OC43:intra
OC43:inter
NL63:intra
NL63:inter
MERS:intra
MERS:inter
London1:intra
London1:inter
HKU9:intra
HKU9:inter
HKU5:intra
HKU5:inter
HKU4:intra
HKU4:inter
HKU3:intra
HKU3:inter
HKU2:intra
HKU2:inter
HKU1:intra
HKU1:inter
BATS:intra
BATS:inter
229E:intra
229E:inter
Coronaviridae
Intra−species
Inter−species
0
5
10
15
20
Genetic distances

Fig.1a

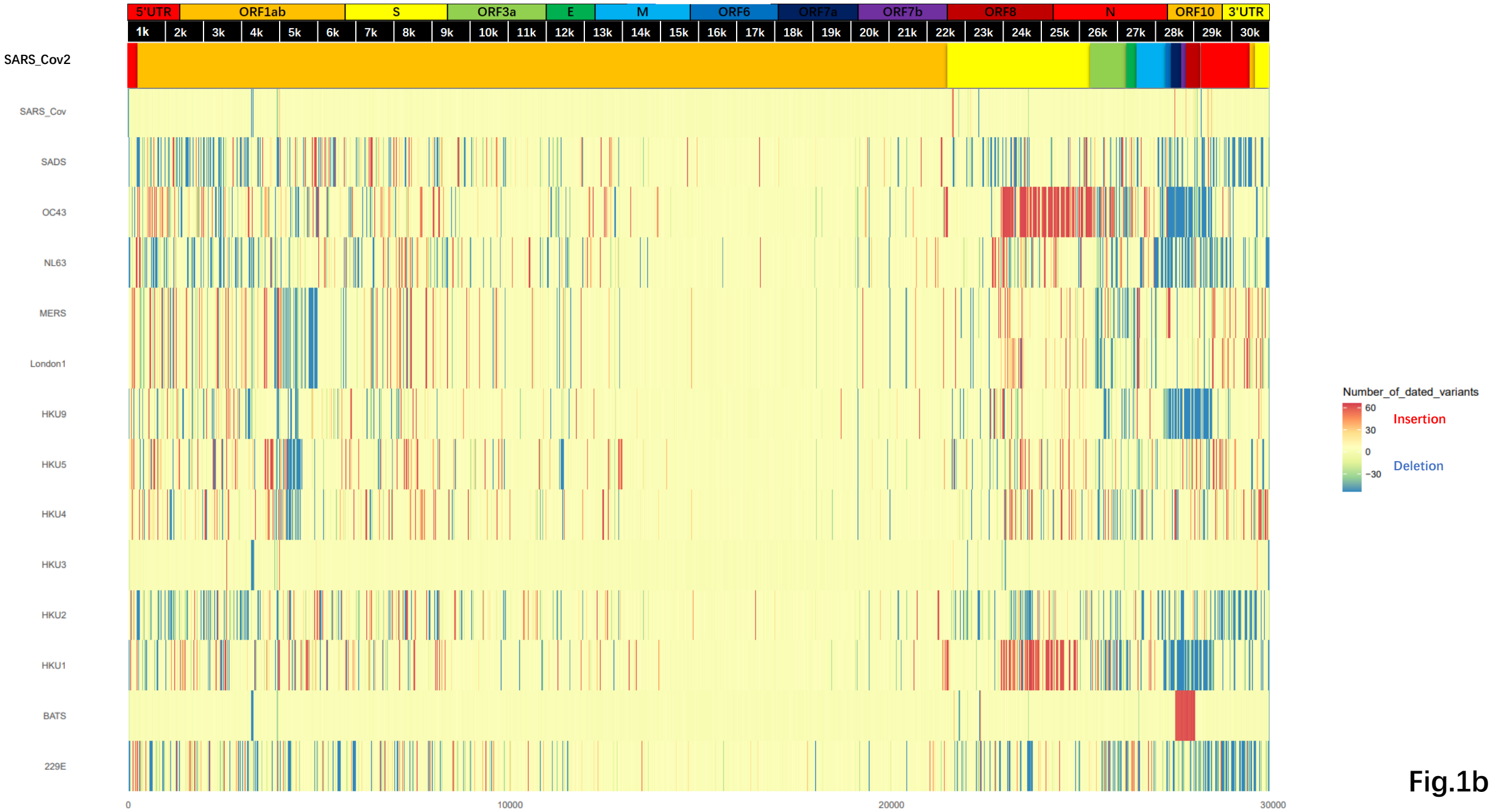
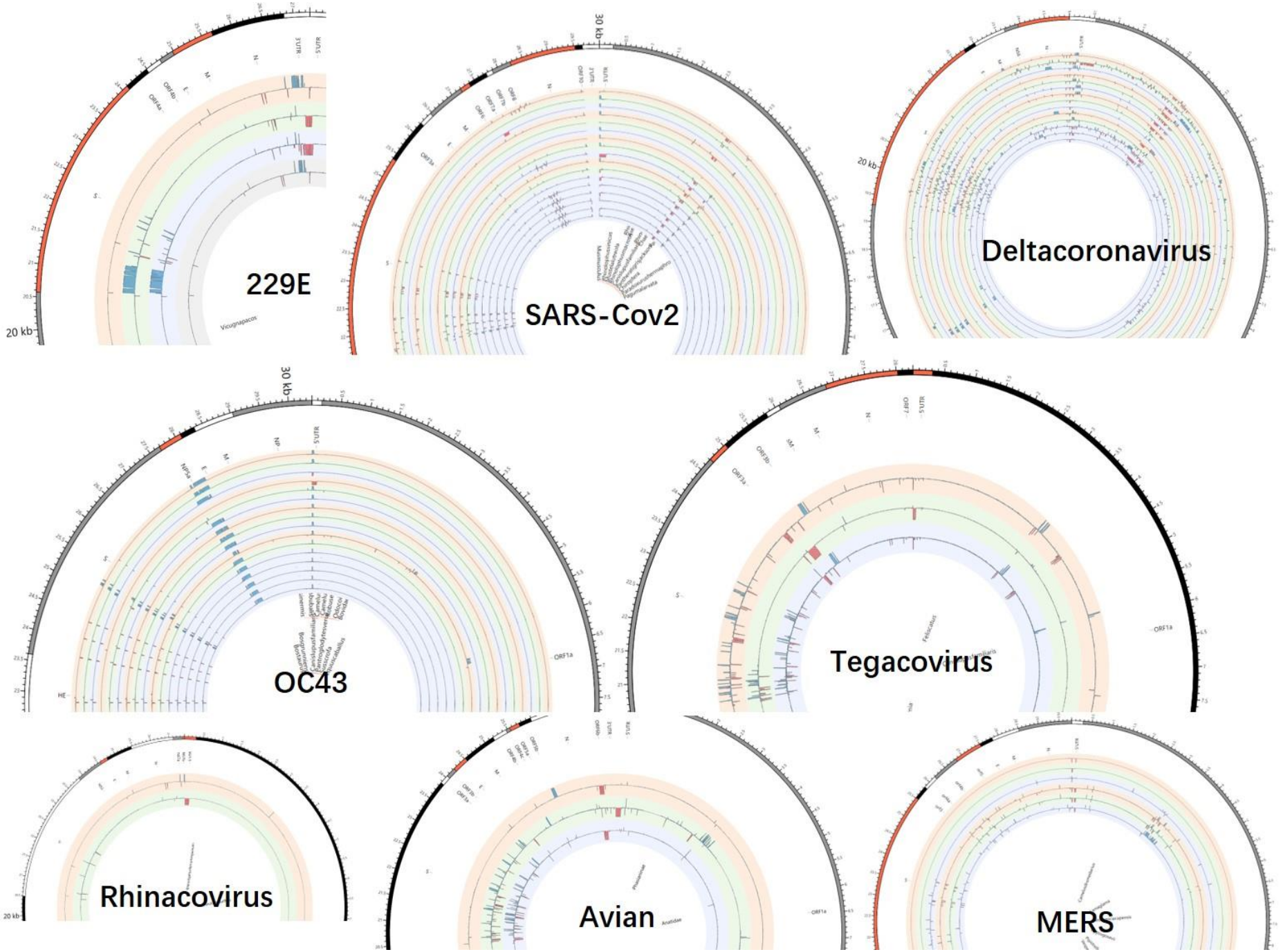

5'UTR
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
3'UTR
1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k 12k 13k 14k 15k 16k 17k 18k 19k 20k 21k 22k 23k 24k 25k 26k 27k 28k 29k 30k
SARS_Cov2
SARS_Cov
SADS
OC43
NL63
MERS
London1
HKU9
HKU5
HKU4
HKU3
HKU2
HKU1
BATS
229E
0
10000
20000
30000
Number_of_dated_variants
60
30
0
−30
Insertion
Deletion

Fig.1b



Coronaviridae	Host
Human coronavirus 229E	Homo sapiens
	Camelus
	Hipposideros
	Macronycteris vittata
	Vicugna pacos
Rhinacovirus	Sus scrofa
	Rhinolophus ferrumequinum
	Rhinolophus affinis
Alphacoronavirus	Sus scrofa
	Felis catus
	Canis lupus familiaris
	Feliformia
Human coronavirus OC43	Homo sapiens
	Bos taurus
	Bos grunniens
	Canis lupus familiaris
	Pan troglodytes verus
	Sus scrofa
	Hydropotes inermis
	Bubalus bubalis
	Camelus dromedarius
	Camelus bactrianus
	Equus caballus
	Kobus ellipsiprymnus
	Odocoileus virginianus
	Rusa unicolor
	Bovidae
	Giraffa camelopardalis
	Vicugna pacos
MERS	Homo sapiens
	Camelus dromedarius
	Lama glama
	Neoromicia capensis
	Hypsugo savii
	Pipistrellus kuhlii
	Vespertilio sinensis
SARS-Cov2	Homo sapiens
	Mus musculus
	Rhinolophus sinicus
	Mustela lutreola
	Rhinolophus macrotis
	Canis lupus familiaris
	Panthera tigris jacksoni
	Rhinolophus affinis
	Aselliscus stoliczkanus
	Rhinolophus ferrumequinum
	Chiroptera
	Rhinolophus pusillus
	Chaerephon plicatus
	Chlorocebus aethiops
	Paradoxurus hermaphroditus
	Viverridae
	Paradoxurus hermaphroditus
	Paguma larvata
Deltacoronavirus	Pycnonotus jocosus
	Mareca
	Lonchura striata
	Zosteropidae
	Falco
	Chlamydotis
	Columbidae
	Galliformes
	Passeridae
	Muscicapidae
	Turdus hortulorum
	Ardeidae
	Gallinula chloropus
	Coturnix japonica
Avian	Gallus gallus
	Phasianinae
	Anatidae
	Meleagris gallopavo

Fig.1c

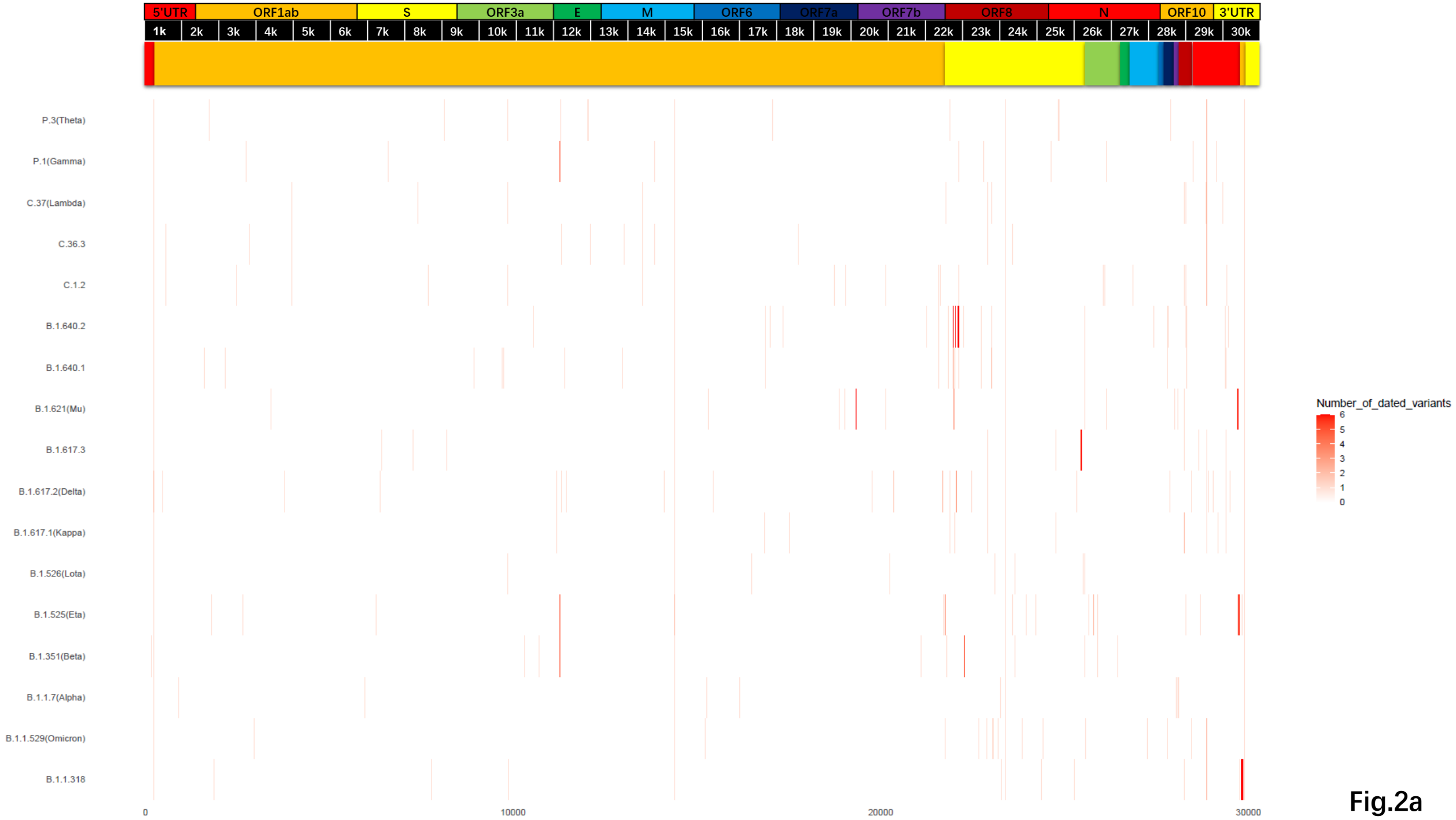
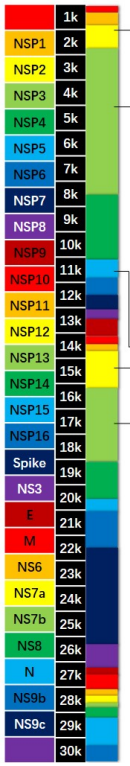

5'UTR
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
3'UTR
1k
2k
3k
4k
5k
6k
7k
8k
9k
10k
11k
12k
13k
14k
15k
16k
17k
18k
19k
20k
21k
22k
23k
24k
25k
26k
27k
28k
29k
30k
P.3(Theta)
P.1(Gamma)
C.37(Lambda)
C.36.3
C.1.2
B.1.640.2
B.1.640.1
B.1.621(Mu)
B.1.617.3
B.1.617.2(Delta)
B.1.617.1(Kappa)
B.1.526(Lota)
B.1.525(Eta)
B.1.351(Beta)
B.1.1.7(Alpha)
B.1.1.529(Omicron)
B.1.1.318
0
10000
20000
30000
Number_of_dated_variants
6
5
4
3
2
1
0

Fig.2a

B.1.1.318
B.1.1.529(Omicron)
B.1.1.7(Alpha)
B.1.351(Beta)
B.1.525(Eta)
B.1.526(Lota)
B.1.617.1(Kappa)
B.1.617.2(Delta)
B.1.617.3
B.1.621(Mu)
B.1.640.1
B.1.640.2
C.1.2
C.36.3
C.37(Lambda)
P.1(Gamma)
P.3(Theta)
mutation_rate_C_T
mutation_rate_G_T
mutation_rate_A_T
mutation_rate_T_C
mutation_rate_T_G
mutation_rate_G_C
Genomic_coordinate
C->T
G->T
A->T
T->C
T->G
G->C

Fig.2b



Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
85	NSP2:T85I				I						I							
129	NSP2:P129L																L	
272	NSP2:E272G																G	
358	NSP2:S358L														L			
462	NSP2:L462F																F	
1	NSP3:A1T												T					
38	NSP3:K38R		R															
41	NSP3:A41V															V		
84	NSP3:M84V		V															
88	NSP3:Y88H								S		H					S		
183	NSP3:T183I			I														
237	NSP3:T237A							A										
378	NSP3:E378V														V			
428	NSP3:T428I								I							I		
488	NSP3:A488S	S																
562	NSP3:A562T							T										
720	NSP3:T720I							I										
821	NSP3:D821Y															Y		
822	NSP3:P822S								S									
837	NSP3:K837N				N													
890	NSP3:A890D			D														
1069	NSP3:V1069I		I															
1198	NSP3:T1198I									I								
15	NSP5:G15S								S							S		
227	NSP12:P227S	S																
323	NSP12:P323L	L	L	L	L	L	L	L	L	F	L	L	L	L	L	L	L	L
88	NSP13:Q88H										H							
105	NSP13:D105Y															Y		
206	NSP13:G206C											C						
371	NSP13:V371A																	A
419	NSP13:P419S							S										
429	NSP13:M429I											I						

Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
15	NS3:L15F											F						
26	NS3:S26L	L										L						
32	NS3:T32I																I	I
42	NS3:P42L										L							
52	NS3:L52F				F													
57	NS3:Q57H				H			H			H						H	H
89	NS3:T89I									I								
92	NS3:S92L									L								
171	NS3:S171L				L													
172	NS3:G172C									C								
223	NS3:T223I								I									
238	NS3:D238Y								Y									
11	NS8:T11IK							K			I							
23	NS8:Q23Stop			Stop													Stop	Stop
26	NS8:T26I												I					
27	NS8:Q27stop			stop													stop	stop
37	NS8:C37R																	R
38	NS8:P38S							S										
52	NS8:R52I			I														
67	NS8:S67F							F										
68	NS8:K68Stop			Stop														
73	NS8:Y73C			C														
12	N:A12G									G								
13	N:P13L						L		L									
22	N:D22Y																Y	Y
63	N:D63G	G	G															
67	N:P67S												S					
199	N:P199L										L							
203	N:R203KM	M	K	K		K	K		K			M	M	K	K	K		
204	N:G204R		R	L		R	R		R					R	R	R		
205	N:T205I				I					I							I	I
215	N:G215C	C					C											
234	N:M234I										I							
235	N:S235F			F														

Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
5	S:L5F										F	F					F	
9	S:P9L									L								L	L
12	S:S12F																F		
18	S:L18F					F													
19	S:T19R	R													R				
20	S:T20N					N													
25	S:P25L									L									
26	S:P26S					S													
52	S:Q52R										R								
67	S:A67V		V								V								
69	S:H69del			del							del								
75	S:G75V						V												
80	S:D80A				A														
95	S:T95I		I					I			I	I			I				
96	S:E96Q																	Q	Q
136	S:C136F									F									
138	S:D138Y					Y													
142	S:G142D	D													D				
145	S:Y145del			del				del		del									
146	S:Y146S							S											
147	S:H147T							T											
152	S:W152R									R									
154	S:E154Q																	Q	
190	S:R190S					S				S								S	
210	S:I210T																	T	
215	S:D215GH				G					G									H
247	S:S247G					G													
253	S:D253G										G								
323	S:P323F											F							
339	S:G339D		D																
341	S:E341D					D													
346	S:R346KS							K									S	S	S
370	S:S371L		L			L													
373	S:S373P		P																
375	S:S375F		F																

Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
394	S:N394S																S	S
417	S:K417NT		N		N	T												
428	S:T428I						I											
440	S:N440K		K															
446	S:G446S		S															
449	S:Y449HN								H								N	N
452	S:L452RQ	R					Q					R	R			R		
477	S:S477N		N															
478	S:T478K	K	K						K									
484	S:E484KQ		A		K	K		K	K	K	K	Q	Q	K	K			K
490	S:F490RS						S										R	S
493	S:Q493R		R															
496	S:G496S		S															
498	S:Q498R		R															
501	S:N501Y		Y	Y	Y	Y		Y	Y					Y			Y	Y
505	S:Y505H		H															
547	S:T547K		K															
570	S:A570D			D														
614	S:D614G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
655	S:H655Y		Y			Y			Y									
677	S:Q677H									H						H		
679	S:N679K		K						K									
681	S:P681HR	R	H	H				H				R	R	H	H		H	H
701	S:A701V				V						V							
716	S:T716I			I					I									
736	S:D736G													G				
764	S:N764K		K															
779	S:Q779K												K					
796	S:D796HY		Y												H			
859	S:T859N						N		N								N	N
879	S:A879T								T									
899	S:A899S															S		
936	S:D936H																H	
950	S:D950N	N						N					N					
982	S:S982A			A														
1027	S:T1027I					I												
1071	S:Q1071H											H						
1072	S:E1072K												K					
1092	S:E1092K													K				
1101	S:H1101Y													Y				
1118	S:D1118H			H														
1139	S:D1139H																	H
1176	S:V1176F					F								F				

Fig.2c

AA	AA	Codon	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	B.1.621 (Mu)	B.1.640.1	B.1.640.2	C.1.2	C.36.3	C.37 (Lambda)	P.1 (Gamma)
Phe	F	UUU	2									12			2		3
		UUC	2									4			2		6
Leu	L	UUA	3					2				2	3			2	
		UUG					6	2						6		2	
		CUU					6					2	6	6			
		CUC															
		CUA	6				6										
		CUG												6			
Ile	I	AUU	16			6					8		6	12			
		AUC	8	2							16	6	6	4		2	
		AUA	8	2		3					8	3	6			2	
Met	M	AUG															
Val	V	GUU	6	6						2				2			
		GUC		6													
		GUA		6						2				2			
		GUG	3														
Tyr	Y	UAU				2				2					2		
		UAC				2				2					2		
Ter	Stop	UAA															
		UAG															
		UGA															
His	H	CAU															
		CAC															
Gln	Q	CAA															
		CAG															
Asn	N	AAU															
		AAC															
Lys	K	AAA															
		AAG															
Asp	D	GAU	2														6
		GAC	2														3
Glu	E	GAA															
		GAG															
Ser	S	UCU	2										12				
		UCC													6		
		UCA													6		
		UCG											12				
		AGU	2										24		6		
		AGC											12				
Pro	P	CCU															
		CCC															2
		CCA															
		CCG															2
Thr	T	ACU	3												2		
		ACC	6		2										2		
		ACA															
		ACG			2												
Ala	A	GCU	12						6								
		GCC															
		GCA							3								
		GCG	4														
Cys	C	UGU															
		UGC															
Trp	W	UGG															
Arg	R	CGU				6											
		CGC															
		CGA															
		CGG															
		AGA				6											
		AGG				6											
Gly	G	GGU					2										
		GGC															
		GGA															

Fig.2d

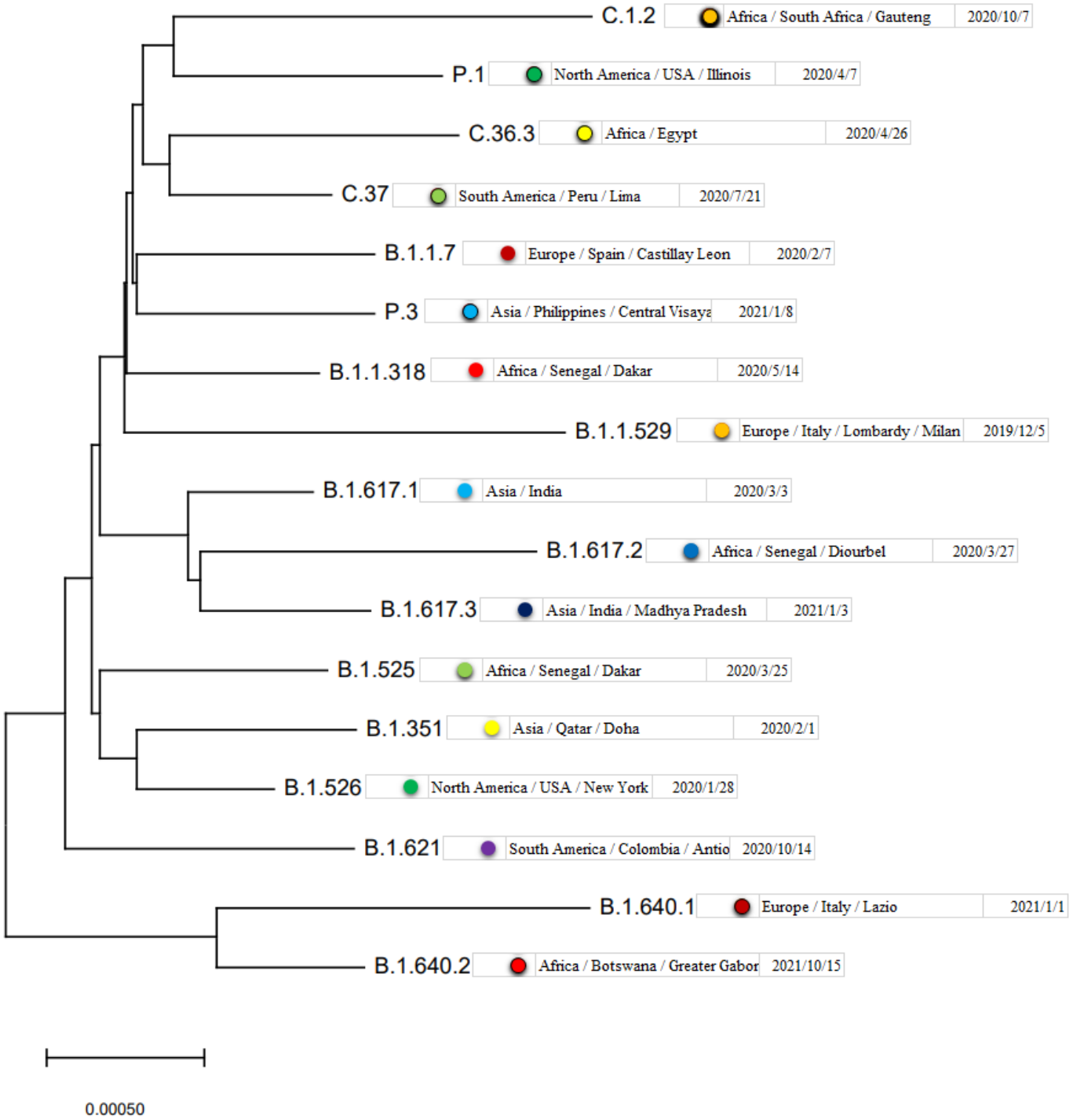
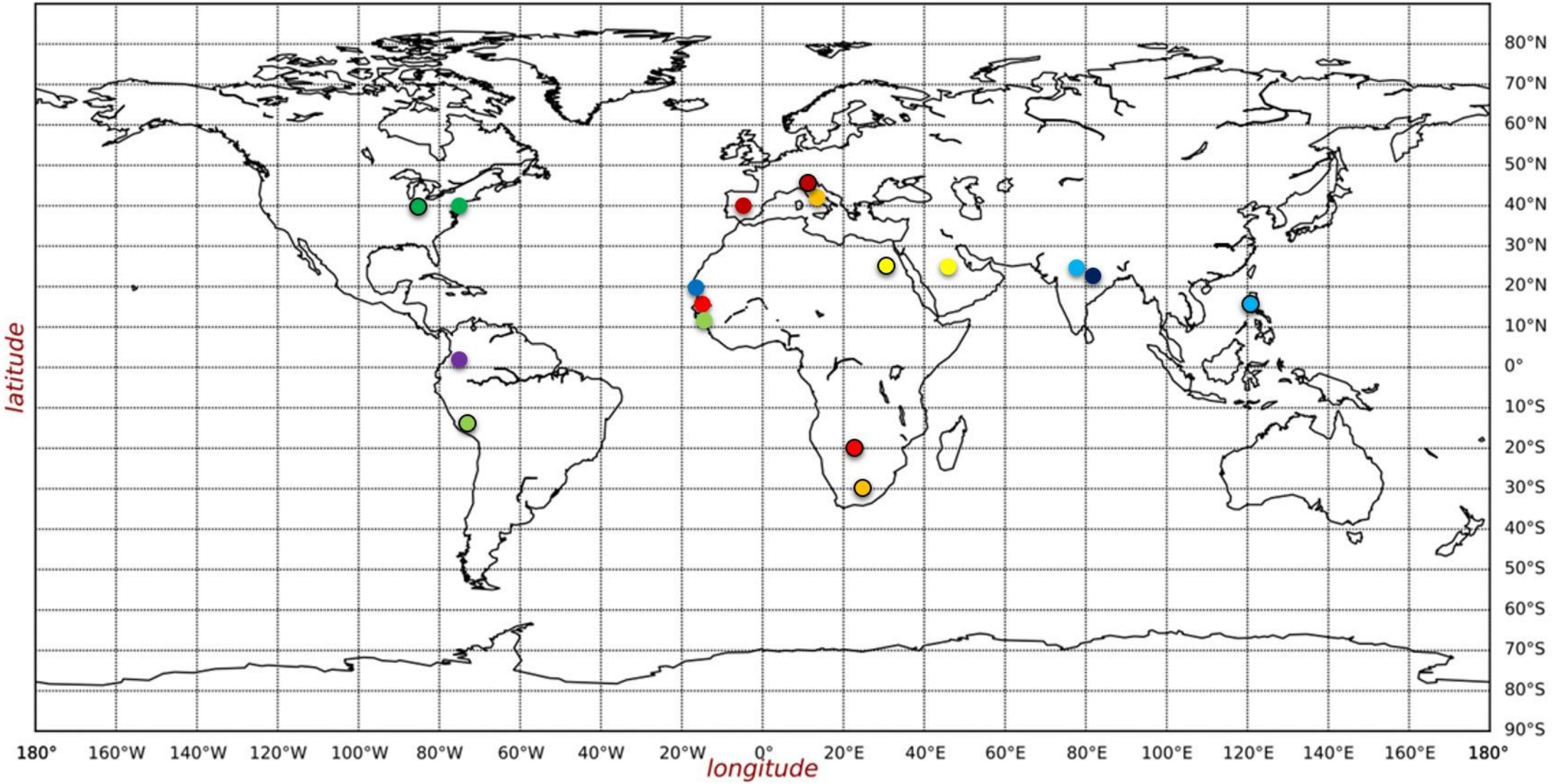

C.1.2
Africa / South Africa / Gauteng
2020/10/7
P.1
North America / USA / Illinois
2020/4/7
C.36.3
Africa / Egypt
2020/4/26
C.37
South America / Peru / Lima
2020/7/21
B.1.1.7
Europe / Spain / Castillay Leon
2020/2/7
P.3
Asia / Philippines / Central Visaya
2021/1/8
B.1.1.318
Africa / Senegal / Dakar
2020/5/14
B.1.1.529
Europe / Italy / Lombardy / Milan
2019/12/5
B.1.617.1
Asia / India
2020/3/3
B.1.617.2
Africa / Senegal / Diourbel
2020/3/27
B.1.617.3
Asia / India / Madhya Pradesh
2021/1/3
B.1.525
Africa / Senegal / Dakar
2020/3/25
B.1.351
Asia / Qatar / Doha
2020/2/1
B.1.526
North America / USA / New York
2020/1/28
B.1.621
South America / Colombia / Antio
2020/10/14
B.1.640.1
Europe / Italy / Lazio
2021/1/1
B.1.640.2
Africa / Botswana / Greater Gabor
2021/10/15
0.00050

Fig.3a

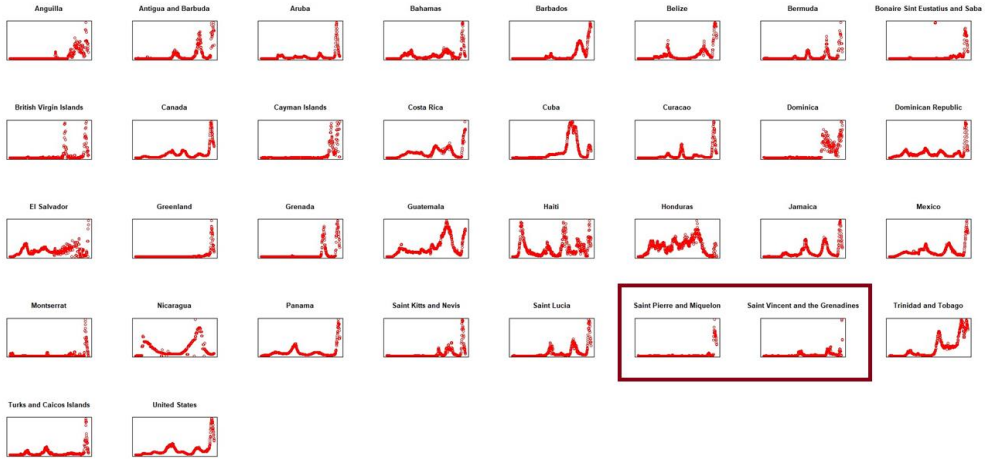


	B.1.1.7	Europe / Spain / Castilla y Leon	2020/2/7		**B.1.617.1(Kappa)**	Asia / India	2020/3/3		**C.1.2**	Africa / South Africa / Gauteng	2020/10/7
	B.1.1.318	Africa / Senegal / Dakar	2020/5/14		**B.1.617.2(Delta)**	Africa / Senegal / Diourbel	2020/3/27		**C.36.3**	Africa / Egypt	2020/4/26
	B.1.1.529(Omicron)	Europe / Italy / Lombardy / Milan	2019/12/5		**B.1.617.3**	Asia / India / Madhya Pradesh	2021/1/3		**C.37(Lambda)**	South America / Peru / Lima	2020/7/21
	B.1.351(Beta)	Asia / Qatar / Doha	2020/2/1		**B.1.621(Mu)**	South America / Colombia / Antioquia / Apartado	2020/10/14		**P.1(Gamma)**	North America / USA / Illinois	2020/4/7
	B.1.525(Eta)	Africa / Senegal / Dakar	2020/3/25		**B.1.640.1**	Europe / Italy / Lazio	2021/1/1		**P.3(Theta)**	Asia / Philippines / Central Visayas	2021/1/8
	B.1.526(Lota)	North America / USA / New York	2020/1/28		**B.1.640.2**	Africa / Botswana / Greater Gaborone	2021/10/15				

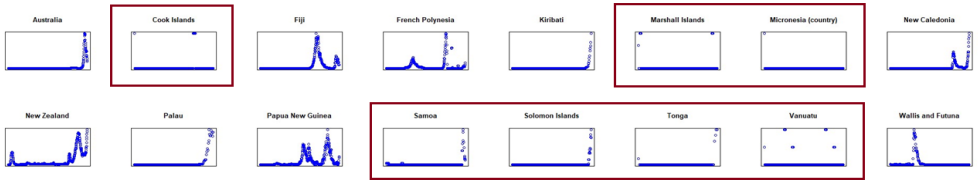
Fig.3b

Distribution of new_cases_smoothed_per_million of SARS-Cov2 in Afria along the timeline.

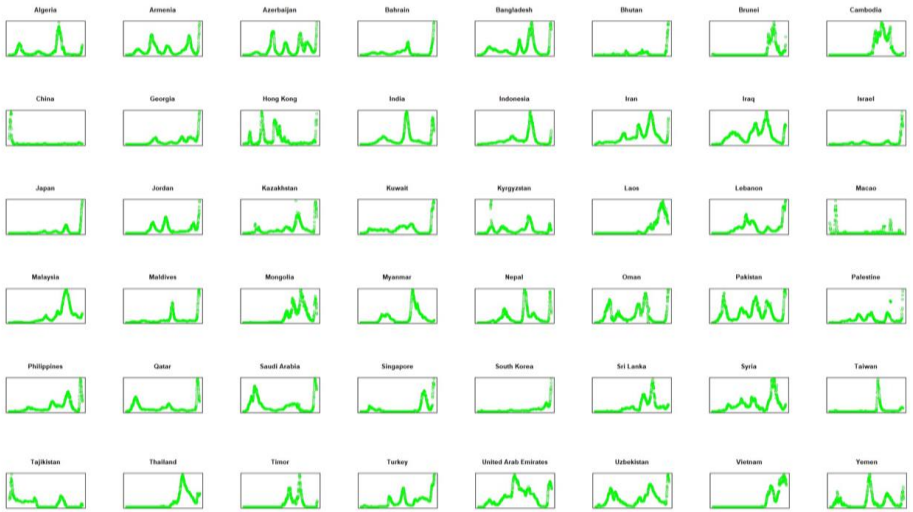
Distribution of new_cases_smoothed_per_million of SARS-Cov2 in North America along the timeline.



Distribution of new_cases_smoothed_per_million of SARS-Cov2 in Oceania along the timeline.



Distribution of new_cases_smoothed_per_million of SARS-Cov2 in Asia along the timeline.



Distribution of new_cases_smoothed_per_million of SARS-Cov2 in Europe along the timeline.

Distribution of new_cases_smoothed_per_million of SARS-Cov2 in South America along the timeline.

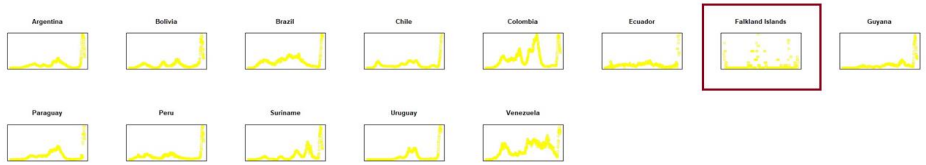


Fig.3c

Fig.3d

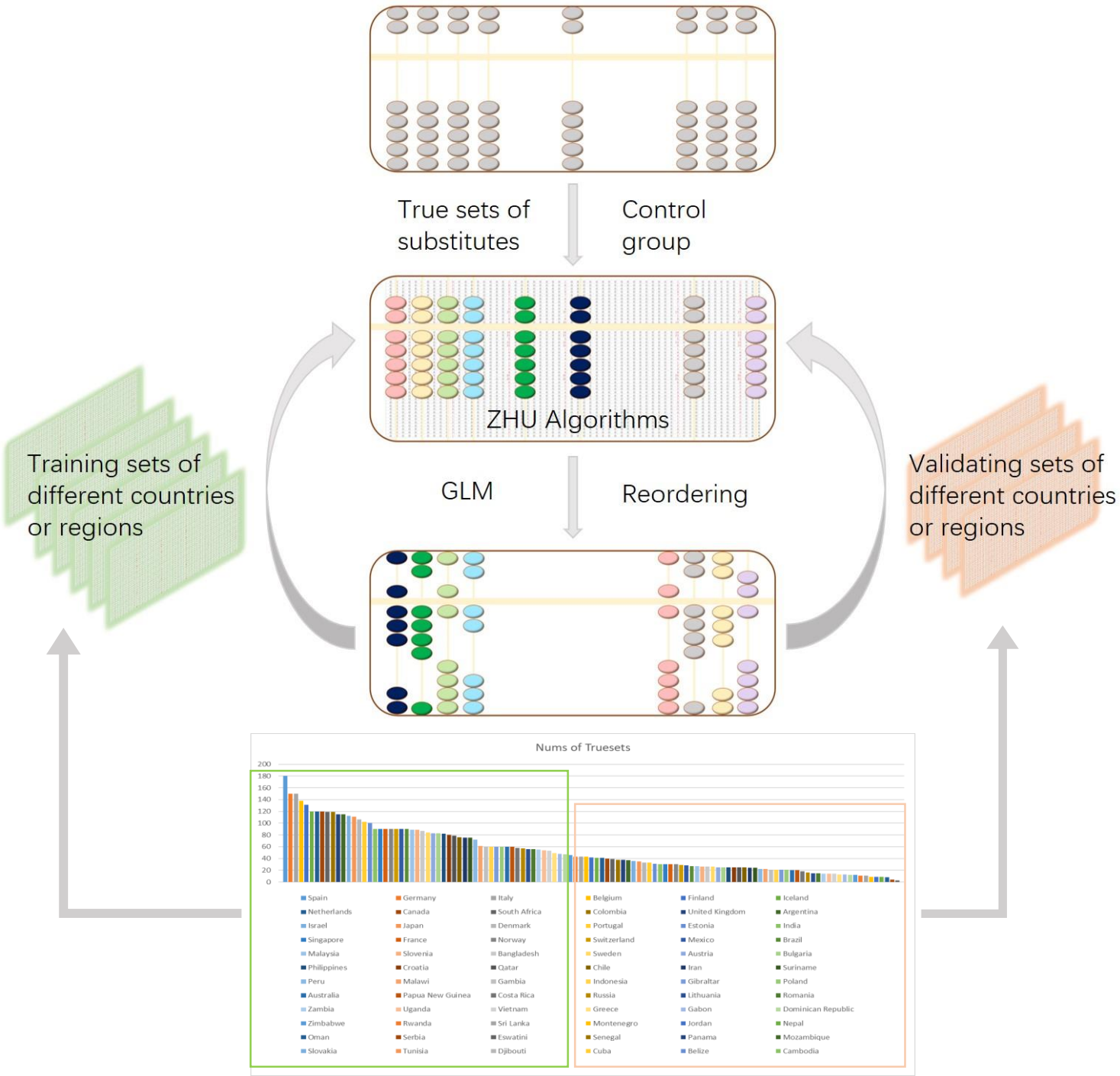
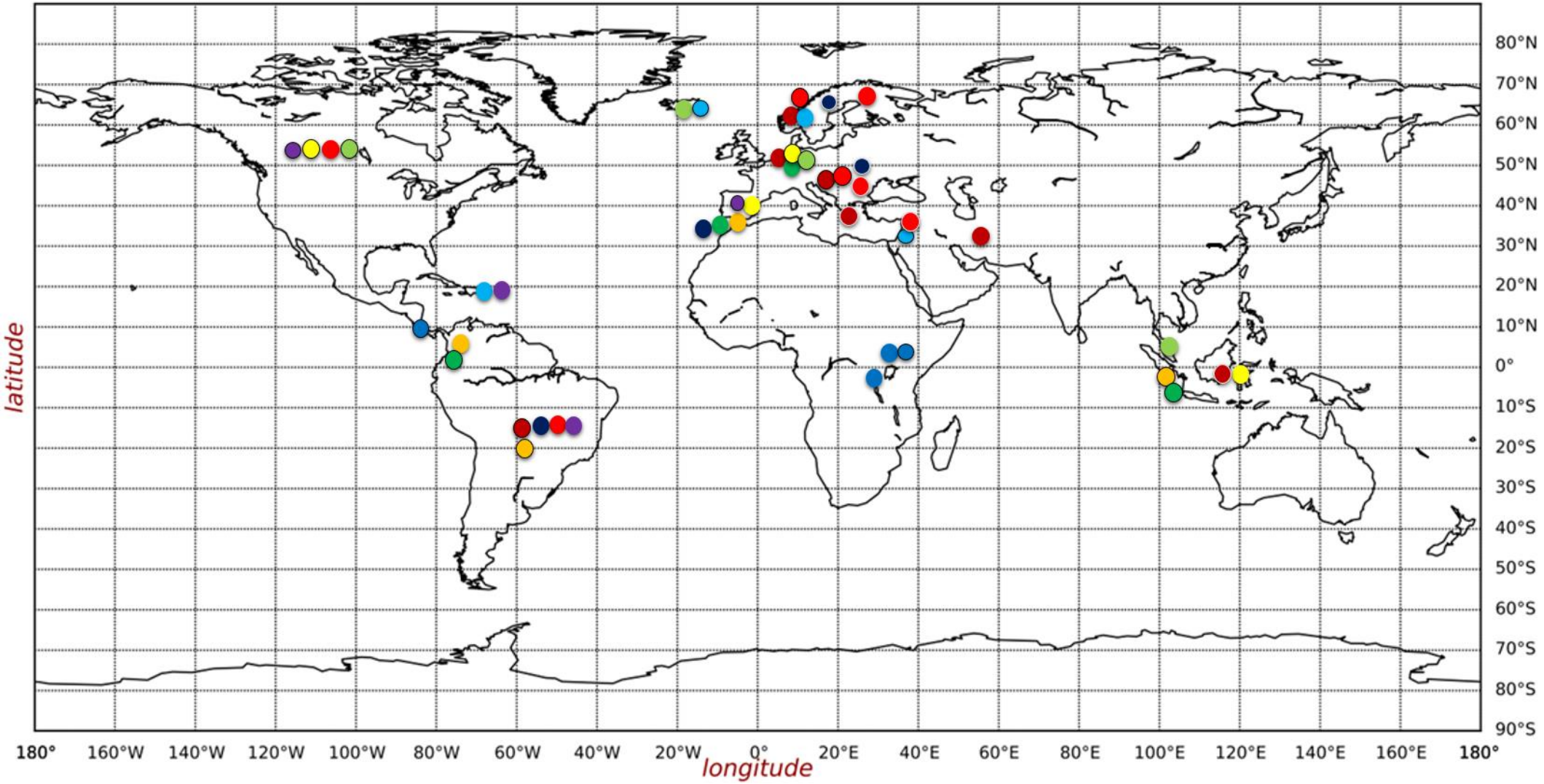
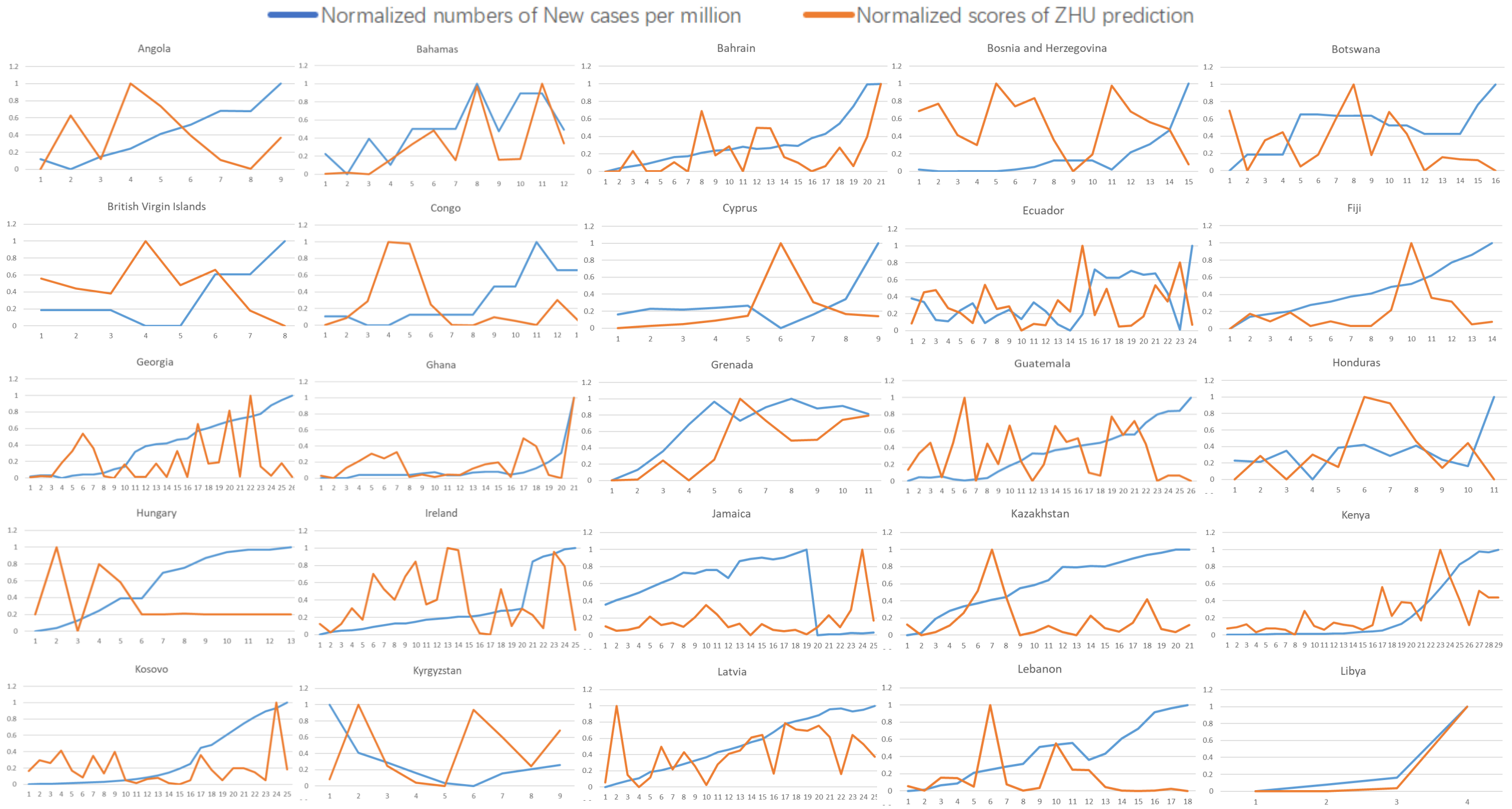

True sets of substitutes
Control group
ZHU Algorithms
GLM
Reordering
Training sets of different countries or regions
Validating sets of different countries or regions
Nums of Truesets
200
180
160
140
120
100
80
60
40
20
0
Spain
Germany
Italy
Belgium
Finland
Iceland
Netherlands
Canada
South Africa
Colombia
United Kingdom
Argentina
Israel
Japan
Denmark
Portugal
Estonia
India
Singapore
France
Norway
Switzerland
Mexico
Brazil
Malaysia
Slovenia
Bangladesh
Sweden
Austria
Bulgaria
Philippines
Croatia
Qatar
Chile
Iran
Suriname
Peru
Malawi
Gambia
Indonesia
Gibraltar
Poland
Australia
Papua New Guinea
Costa Rica
Russia
Lithuania
Romania
Zambia
Uganda
Vietnam
Greece
Gabon
Dominican Republic
Zimbabwe
Rwanda
Sri Lanka
Montenegro
Jordan
Nepal
Oman
Serbia
Eswatini
Senegal
Panama
Mozambique
Slovakia
Tunisia
Djibouti
Cuba
Belize
Cambodia

Fig.4a



	X13.NSP3_T183I.I	2020/03 - 2021/08 Iran: 352.54	2020/03 - 2021/11 Netherlands: 817.422	2020/10 - 2021/09 Norway: 41.525		**X69.S_D253G.G**	2020/10 - 2021/09 Norway: 67.733	2021/08 - 2022/02 Slovenia: 387
	X74.S_S371L.L	2020/09 - 2022/01 Finland: 65.9662	2020/03 - 2022/01 Canada: 206.464	2020/06 - 2022/01 Brazil: 221.244		**X89.S_E484KQ.V**	2020/03 - 2022/02 Singapore: 361.997	2020/06 - 2022/01 Brazil: 1104.313
	X11.NSP3_M84V.I	2020/12 - 2021/12 Gibraltar: 5590.54	2020/12 - 2022/01 Colombia: 63.5			**X101.S_Q677H.H**	2020/03 - 2021/11 Netherlands: 1051.28	2020/03 - 2022/01 Canada: 21.4259
	X29.NSP12_P227S.P	2021/01 - 2021/07 Indonesia: 2.9861	2020/03 - 2022/01 Spain: 250.599			**X101.S_Q677H.Q**	2020/03 - 2021/11 Netherlands: 1099.87	2020/03 - 2022/01 Canada: 24.4729
	X31.NSP13_Q88H.Q	2021/01 - 2021/08 Malaysia: 1998.7116	2020/03 - 2021/08 Iceland: 205.8784			**X110.S_D796HY.D**	2020/03 - 2022/02 Singapore: 940.076	2020/12 - 2022/01 Colombia: 41.1
	X42.S_L18F.F	2020/12 - 2021/12 Gibraltar: 546.44	2020/03 - 2021/11 Netherlands: 89.456			**X129.NS3_P42L.P**	2020/03 - 2022/01 Israel: 1200	2020/03 - 2021/08 Iceland: 1293.6818
	X53.S_D80A.D	2020/10 - 2021/09 Norway: 5.148	2020/12 - 2022/01 Dominican Republic: 10.58			**X56.S_C136F.C**	2021/05 - 2022/01 Uganda: 56.381	2021/04 - 2022/01 Costa Rica: 701.2643
	X56.S_C136F.stop	2021/06 - 2022/01 Rwanda: 30.347	2021/05 - 2022/01 Uganda: 56.452			**X76.S_S375F.S**	2020/11 - 2021/11 Slovakia: 1857.245	2020/05 - 2022/01 Sweden: 1290
	X62.S_W152R.W	2020/12 - 2021/12 Gibraltar: 1272.88	2020/06 - 2022/01 Brazil: 276.801			**X80.S_N440K.K**	2020/03 - 2022/01 Spain: 292.031	2020/03 - 2022/01 Canada: 117.2041
	X68.S_S247G.S	2020/12 - 2022/01 Dominican Republic: 113.174	2020/06 - 2022/01 Brazil: 306.985			**X81.S_G446S.S**	2021/01 - 2021/07 Indonesia: 113.0921	2020/10 - 2022/01 Greece: 3060.4
	X69.S_D253G.D	2021/08 - 2022/02 Slovenia: 245	2020/06 - 2022/01 Brazil: 28.109			**X86.S_S477N.N**	2020/10 - 2021/12 Jordan: 140.499703	2020/07 - 2020/12 Serbia: 316.05

Fig.4b


Normalized numbers of New cases per million
Normalized scores of ZHU prediction
Angola
Bahamas
Bahrain
Bosnia and Herzegovina
Botswana
British Virgin Islands
Congo
Cyprus
Ecuador
Fiji
Georgia
Ghana
Grenada
Guatemala
Honduras
Hungary
Ireland
Jamaica
Kazakhstan
Kenya
Kosovo
Kyrgyzstan
Latvia
Lebanon
Libya

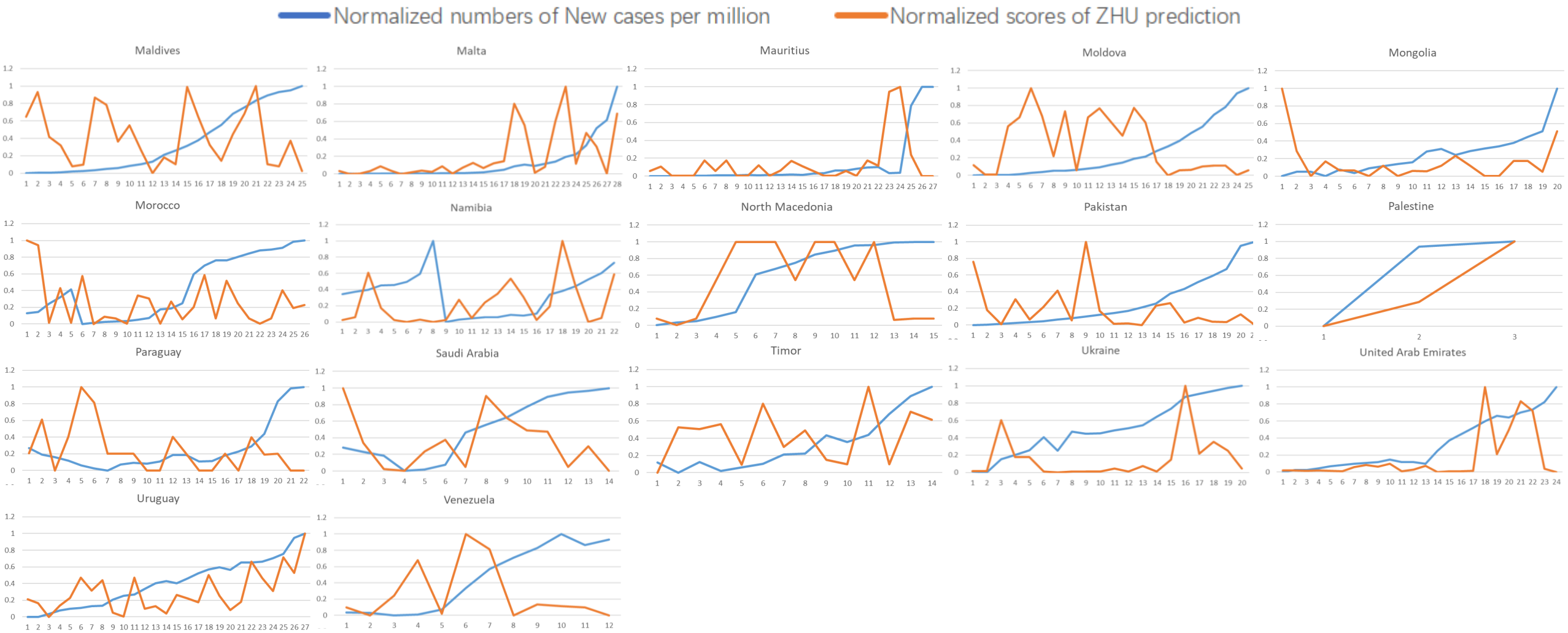

Normalized numbers of New cases per million
Normalized scores of ZHU prediction
Maldives
Malta
Mauritius
Moldova
Mongolia
Morocco
Namibia
North Macedonia
Pakistan
Palestine
Paraguay
Saudi Arabia
Timor
Ukraine
United Arab Emirates
Uruguay
Venezuela

Fig.4c

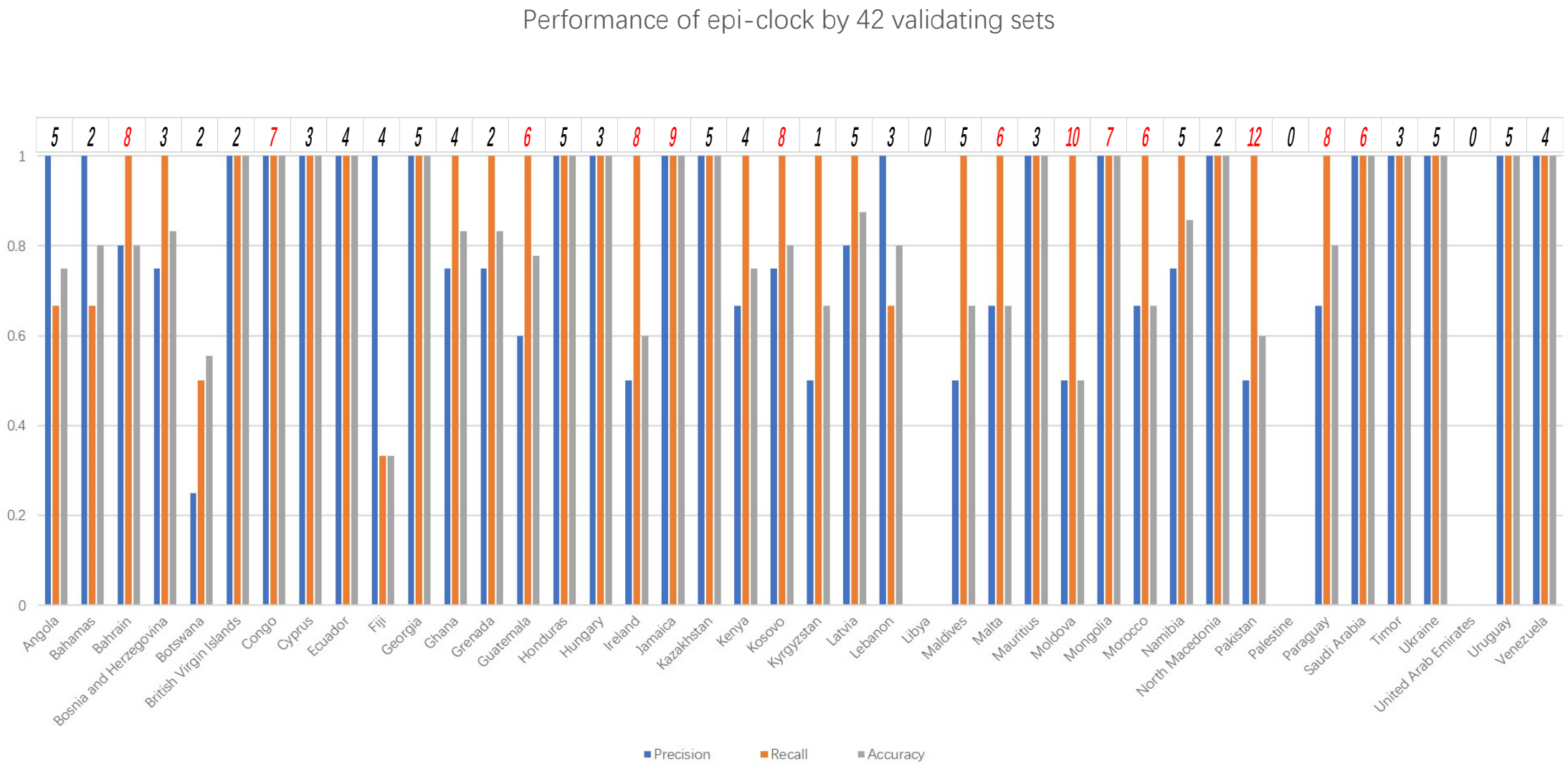

Performance of epi-clock by 42 validating sets
5 2 8 3 2 2 7 3 4 4 5 4 2 6 5 3 8 9 5 4 8 1 5 3 0 5 6 3 10 7 6 5 2 12 0 8 6 3 5 0 5 4
1
0.8
0.6
0.4
0.2
0
Angola
Bahamas
Bahrain
Bosnia and Herzegovina
Botswana
British Virgin Islands
Congo
Cyprus
Ecuador
Fiji
Georgia
Ghana
Grenada
Guatemala
Honduras
Hungary
Ireland
Jamaica
Kazakhstan
Kenya
Kosovo
Kyrgyzstan
Latvia
Lebanon
Libya
Maldives
Malta
Mauritius
Moldova
Mongolia
Morocco
Namibia
North Macedonia
Pakistan
Palestine
Paraguay
Saudi Arabia
Timor
Ukraine
United Arab Emirates
Uruguay
Venezuela
Precision
Recall
Accuracy

Fig.4d