Identification of variable sites in Sars-CoV-2 and their abundance profiles in time

Francesco Comandatore[1,2,†], Alice Chiodi[2,3,†], Paolo Gabrieli[2,3] Gherard Batisti Biffignandi[4], Matteo Perini[1,2], Matteo Ramazzotti[5], Stefano Ricagno[3], Sara Giordana Rimoldi[6], Maria Rita Gismondo[6], Valeria Micheli[6], Claudio Bandi[2,3], Matteo Brilli[2,3∗]

† These authors contributed equally; [1] Dipartimento di Scienze Biomediche e Cliniche L. Sacco, Università degli Studi di Milano, Milano, Italia; [2] Centro Ricerca Pediatrica Romeo ed Enrica Invernizzi, Università degli Studi di Milano, Milano, Italia; [3] Dipartimento di Bioscienze, Università degli Studi di Milano, Milano, Italia; [4] Dipartimento di Scienze Clinico-Chirurgiche,Diagnostiche e Pediatriche, Università degli Studi di Pavia, Pavia, Italia; [5] Dipartimento di Scienze Biomediche, Sperimentali e Cliniche, Università degli Studi di Firenze, Firenze; [6] Laboratorio di Microbiologia, Virologia e Diagnostica della Bioemergenza, ASST Fatebenefratelli Sacco, Milano, Italia

∗ E-mail: matteo.brilli@unimi.it

Abstract

The Sars-Cov-2 pandemic that we are currently experiencing is one of the worst since long time in terms of cost in terms of human lives and economic impact. Besides the source of this virus, it is also important to understand how different virus-variant change their frequency in time, because this may indicate different levels of aggressiveness, or provide useful data to trace the spreading of the virus in the population. In this work we trace 6 amino acid variant that have high frequency in Italy and Europe, but were absent or very low frequency during the first explosion in China and the initial diffusion in Europe. One of these, variant D614G in the Spike protein, appears in Germany on January 28 and since then has overcome the wild-type variant arrived from China.

Introduction

The worldwide fast spread of SARS-CoV-2 virus during the first months of this year has caused 198,668 deaths with more than 2,878,196 confirmed cases since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China [1, 2] up to the 27 th April 2020 [3]. The virus belongs to the beta-coronavirus, being the seventh coronavirus known to infect humans, which causes severe respiratory and systemic disorders [4] , with a basic R0 index ranging from 1.4 to 6.49 [3]. Closest viral relatives of SARS-CoV-2 circulate in animals, either bat or pangolins [5] , suggesting that an animal virus jumped species boundaries to efficiently infect humans through multiple passage in intermediate animal hosts, that has not been yet identified to date [6]. Traces of the history of the spread are hidden in the viral genome and comparative genomics approaches has been therefore used to understand how viruses can adapt to multiple hosts, uncovering which are the key signatures of this adaptation [7], and to trace back the infection routes of the virus. At the same time, genomic studies can help tracing viral variants that might be geographically restricted and/or might account for different levels of infectivity and mortality in humans. These variants might arise during the spread of the viru as a consequence of natural selection and the use of antivirals. Viruses are indeed known for their high frequency of mutation, particularly in single stranded RNA viruses, as in the case of SARS-CoV2 [8], which has a single, positive strand RNA genome. In general, the aplotypes present at the very beginning of the infection change progressively; sometimes novel variants can overcome the ancestral ones, and this can be a consequence of different levels of aggressiveness but also of a founder effect. Indeed, if one considers different variants as identical in terms of their ability to infect and replicate in the host, switches in aplotype frequency can be explained by a mechanism whereby the major aplotype in a country depends on the very first viruses that arrive in that country. However, if aplotype frequency changes globally, then the hypothesis of differential aggressiveness becomes probable, indicating those variants that might be better adapted to infect the human hosts. Alternatively, the rising of continuous new variants, particularly in those genes coding for structural protein, might help the virus evading the immune response, challenging the human immune system with novel epitopes. Here, we performed a comprehensive study of the SARS-CoV-2 genomes isolated from patients from Italy during the entire course of the pandemy, comparing their genomes to

those isolated in the rest of the world. Italy was the first European country registering not- imported COVID-19 cases with severe outcome and requiring hospitalization: the first case of COVID-19 was registered on February 20, 2020; he was a young man in the Lombardy region in the north of Italy with an atypical pneumonia [9]. In the next 24 hours there were 36 more cases, with anyone having previous clear contact with somebody known to have COVID-19. From that moment up to 27 th April 2020, Italy registered 197,675 cases with 26,644 deaths, one of the highest in Europe and in the entire world. WHO classifies the transmission occurring in Italy as ?Community Transmission?, that is a country experiencing large outbreaks of local transmission with no clear transmission chains between the cases, and with multiple unrelated clusters of transmission [3]. To have a better insight in the COVID-19 pandemic history and spread in Italy, we identified 6 non synonymous mutations that differentiate Italian Sars-Cov-2 strains and the strains globally present in the world, by making use od the gisaid database and we traced their relative frequency in time both globally and in different countries.

Material and Methods

Extraction and editing of coding sequences

To extract coding sequences from complete and partial genomes, we performed a strategy based on tblastn comparison using as queries all the proteins from Sars-Cov-2 sequence deposited in NCBI (Accession MN908947). After the comparison, the coordinates of the alignment were used to cut from the genome the nucleotide sequences aligned with the protein queries. Nucleotide sequences were then translated, manually controlled for the presence of frameshifts, and manually curated as explained below. Sequences with a high fraction of gaps or Ns were removed (the extent depended on the length of the alignment under analysis, on average sequences with more than 5% gaps or Ns were eliminated), aligned using the translated sequence, and then partitioned into two groups: strains from Italy and all the strains (comprising italian ones). Multialignment were manually edited to remove sequences not well aligned or too partial, resulting in a different number of sequences per alignment. This manual curation resulted in alignments with a minimum number of 2262 sequences for ORF3a and a maximum of 5585 sequences for the Nucleocapsid protein, with an average of 4222 sequences per alignment. We are aware that by removing sequences with gaps we could remove part of the variability that resulted in indels. However, we did not observe indels with such a frequency to make it thinking we are in the presence of some real signal, therefore we think that we are removing more noise than genuine signal. Additional analysis with high quality genome sequences are necessary to ensure this. However, this does not change the results of our analysis as the target of this work is on point mutations.

Identification of variable and differentially evolving sites in Sars-Cov-2 coding sequences

Sequences from the multialignments were used to build amino acid frequency profiles by using the R-package seqinr [ref Lobry]. Basically, for each protein encoded in the Sars-Cov-2 we obtained two profiles, one for italian strains and one for the entire set of sequences. As the positions in the alignments for the two groups under examination are congruent, we can calculate:

$$S = max\left(\left|log_2 \frac{f(i,)^{it}}{f(i,)_{all}}\right|\right),$$

that is the log ratio of the frequency of each amino acid a for each i position in the alignment in italian sequences (^{it}) and total sequences (^{all}). Then we identify positions where $|S| > 1$ meaning a position where one of the residues have double (or half) the abundance in the denominator wrt to the denominator or more simply, positions that show a pattern of aminoacids that is different in the two groups. Variability in multi-alignments was quantified by calculating the entropy ($H^p = -\sum_{i=1}^{20} f(p,i) \times log_2 f(p,i)$) of each position of a manually curated protein multi-

alignment. Positions were considered as interesting for us if $H > 0.6$, which indicates a situation where there are more aplotypes with a significant frequency at a certain position.

Covariance Analysis

We used Mutual Information, defined as in [10] to assess if variants tend to co-occur frequently.

Time profiles

Sequences were downloaded a second time from Gisaid.org on April, 28, 2020 to produce time series containing all the information available (9075 sequences). For each variant, we integrated the information about the residue found at variant positions identified in this work and the date of sampling, as provided by the database and also indicated in the Gisaid Sequence Acknowledgment table (Supplementary Table 1). The range of time including all the sampling dates was split into 14 (10 for the analysis country) intervals of 9.5 days in average (13.5 for the analysis by country, as there are less sequences per interval). For each interval of time, we calculated the frequency of the observed variants with respect to the total number of sequences in that interval (limited to a certain country in the analysis by country). In Supplementary Table 2 we show the number of sequences used for each country for each interval. All those sequences for which the geographic origin and/or the data were not correctly defined in the Gisaid.org table (Supplementary Table 1), were removed. In the end, time profiles were built on about 8500 sequences, depending on the site taken into account.

Phylogenetic analysis

We performed a phylogenetic analysis by concatenating nucleotide alignments obtained by retro-translating manually curated amino acid alignments obtained as explained in section **Extraction and editing of coding sequences**, to avoid introducing artifacts. We selected the GTR+I+Γ model as the best phylogenetic model for our alignment using ModelTest-NG [11] and then we performed Maximum Likelihood phylogenetic estimation in RaxML8 [12]. The obtained phylogenetic tree was then visualized using iTol [13] to integrate additional information.

Results

Amino acid variants

Our analysis allowed us to identify 6 positions in three proteins where amino acid frequencies change drastically when we compare Italian sequences with all the sequences available on Gisaid.org on April, 10, 2020 (Figure 1). These positions are also the only ones that show a significant variability across all available Sars-CoV-2 sequences (Figure 2):

1. Position 614 in the Spike protein, where an Aspartate residue is found in high frequency in sequences obtained at the beginning of the pandemic. A second variant, with a Glycine residue at the same position is however very common in Europe and Italy in particular, shown in Table 1.

2. Positions 203 and 204 in Protein N, shown in Table 2.

3. Positions 265, 3606 and 4715 in the Polyprotein (Table 3)).

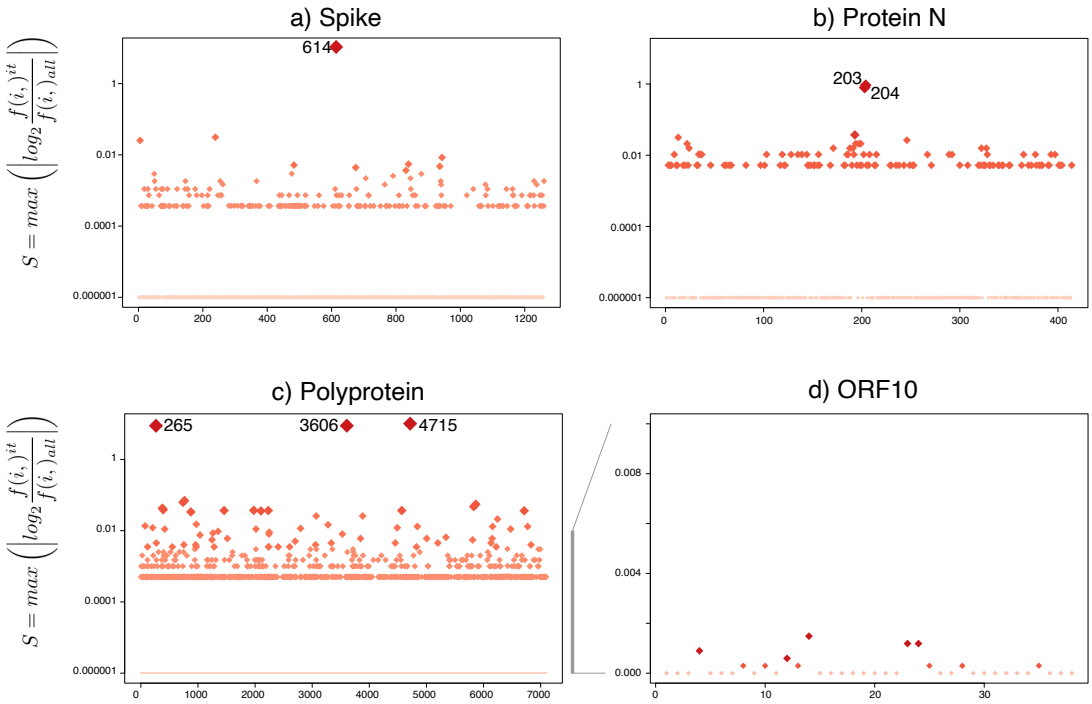


Figure 1. Positions that were identified as maximally different when contrasting amino acid frequencies in italian Vs the entire dataset. As a comparison, see panel d for ORF10, which has no differences in this comparison.

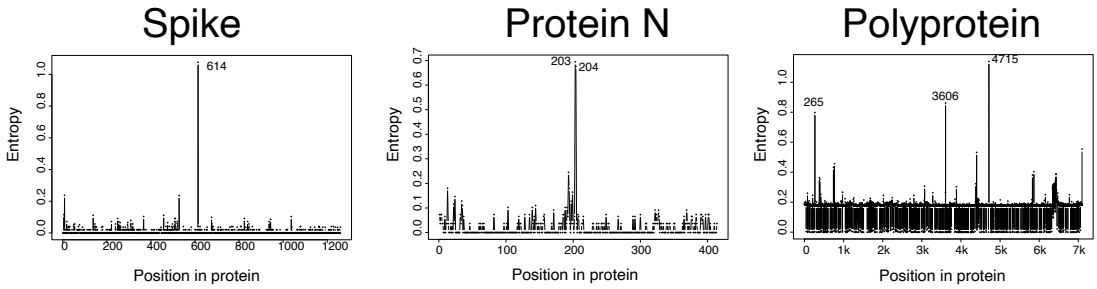


Figure 2. Entropy profiles of multialignments for Spike, protein N and Polyprotein of Sars-CoV-2 with the variable position indicated.

Table 1. Frequency of amino acids in identified positions of the Spike protein and the corresponding log-ratio.

AA	$a = f(i, AA)^{it}$	$b = f(i, AA)^{All}$	$S = log_2 \frac{a}{b}$
Position i=614			
D	0.00025	0.41973	-10.67
G	0.99510	0.57573	0.78

Table 2. Frequency of amino acids in identified positions of Protein N and the corresponding log-ratio.

AA	$a = f(i, AA)^{it}$	$b = f(i, AA)^{All}$	$S = log_2 \frac{a}{b}$
Position i=203			
K	0.28261	0.16004	0.82
R	0.65217	0.77474	0.25
Position i=204			
G	0.65217	0.78202	-0.27
R	0.28261	0.15277	0.88

Time series of variants

In Figure 3 we report the inferred time profile of the variants that clearly show how for some of them the virus experienced a shift of the major aplotype in between February and March. The Spike variant D614G was sequenced for the first time in Germany, on January 28, 2020 and since then has increased rapidly in frequency, overcoming the ancestral aplotype. It was also sampled in China on 6 and 7 February 2020 in Shangai and Wuhan, respectively. The remaining variants do not have such a clear shift in frequency, excepted for a Leucine in position 4715 in the Polyprotein that increases rapidly starting from the beginning of March after several weeks of stable frequency. Interestingly, by calculating the covariance of the variants (Figure 5) we observe a relatively high mutual information among variant Spike D614D and Polyprotein P4715L which may suggest some form of structural interdependence and therefore a functional meaning for these variants. However, a bibliographical search on known Spike proteins did not reveal any mutagenesis experiment where this position was tested. For what concern the two sites on protein N, we find a perfect agreement among the profiles, strongly suggesting that these two mutations happened together. Indeed, the observed pairs of amino acids at these sites over the total sequences used here are: KG (relative frequency, r.f.$= 0.00036$) , KR (r.f.$= 0.12500$) and RG (r.f.$= 0.87464$) indicating that the two variants at these sites behave like a single variant as there are basically not the pairs KG and RR. We reasoned that the frequencies in sequences uploaded by different countries may somehow provide an average picture of a situation that is instead variable in different countries. We therefore derived time profiles for countries represented by more than 25 sequences in Gisaid.org (Figure 4). Differently from previous analysis, the time series are based on sequences downloaded on April, 28. We show here only a selection of countries concerning the Spike variant, see Supplementary Material for complete figures concerning all countries and all variants. We also provide a movie of changes in variant frequencies day by day (Supplementary Figures 1-6). The data in the plot indicates that this variant was not sequenced in China, despite the availability of 182 sequences predating the first appearance of this variant in Germany on January 28, 2020, corresponding to the only sequence available in that time window in Germany, explaining the value 1 at the beginning of the time series). Since then D614G has raised in frequency more or less rapidly in different countries, most of the time overcoming in abundance the ancestral variant, except for China (where it is present at very low frequencies) and United Kingdom, where the two variants are coexisting at similar percentages since several weeks.

At first glance, a similar pattern might suggest that some of these variants have higher fitness than the ancestral one; for this reason, we performed a thorough bibliographical record looking for functional information associated to these positions or the domains they belong to.

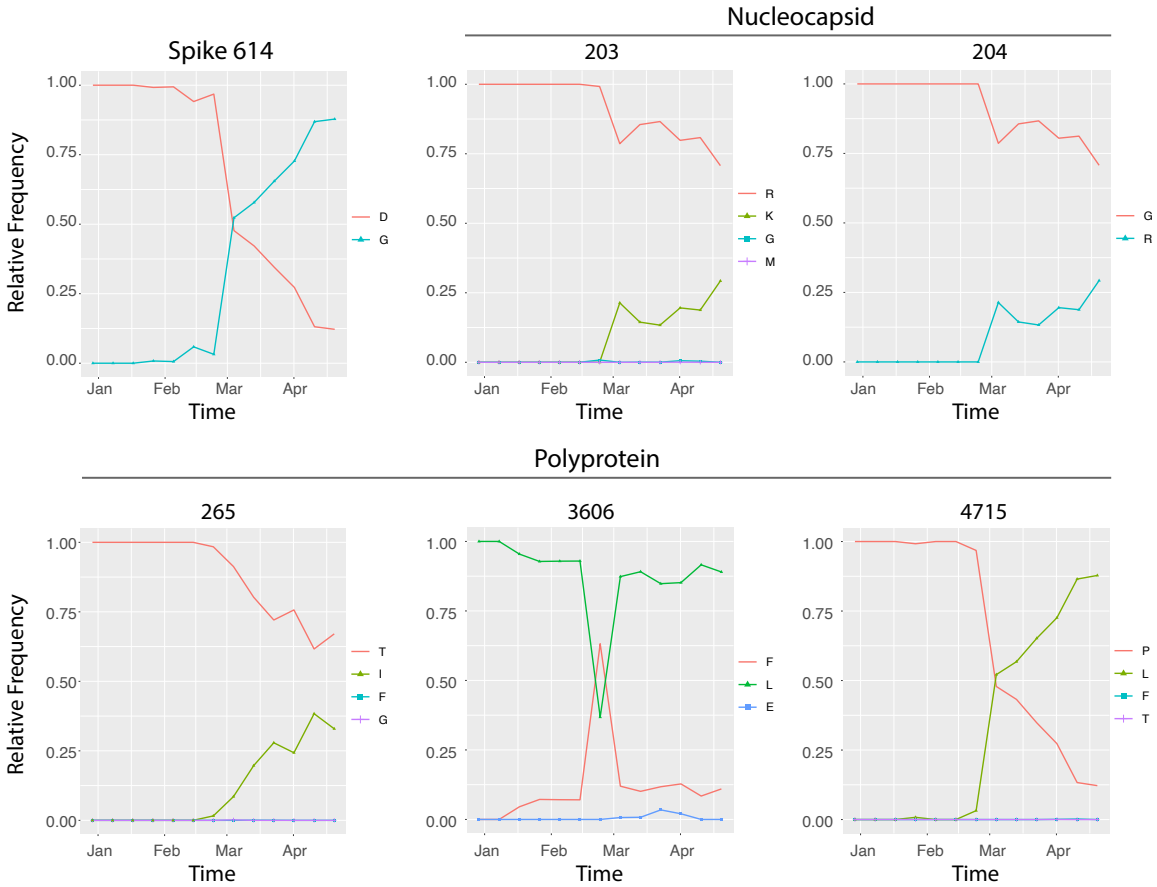


Figure 3. Time series for each variable site using sequences downloaded on April 28, 2020.

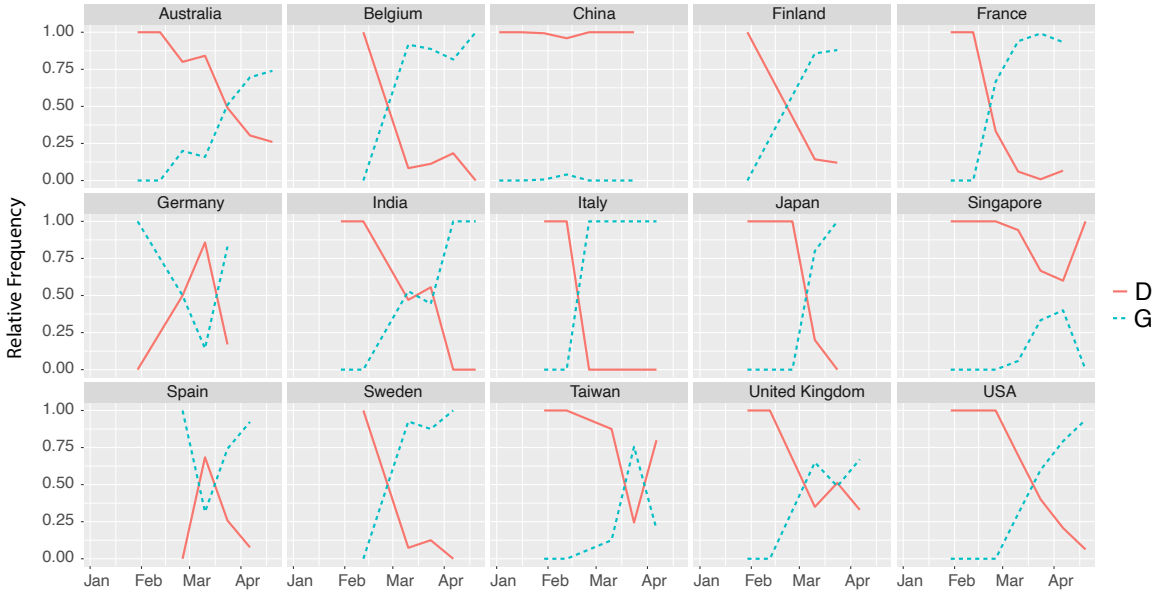


Figure 4. Spike residues at position 614, time series by country. Plotted values are relative frequencies in a time window. Values are placed at x = end of interval.

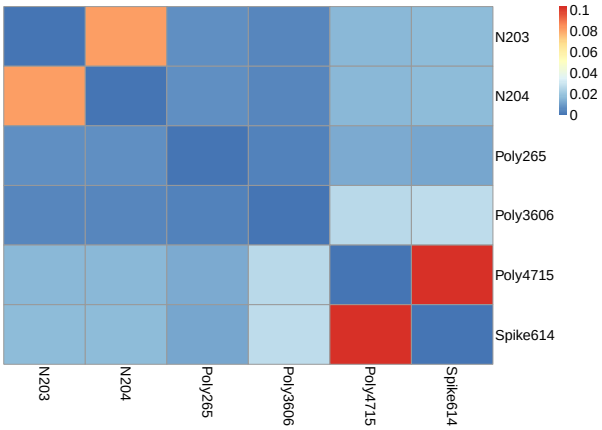


Figure 5. Mutual information calculated for all pairs of variants analysed in this work.

Table 3. Frequency of amino acids in identified positions of the Polyprotein and the corresponding log-ratio.

AA	$a = f(i, AA)^{it}$	$b = f(i, AA)^{All}$	$S = log_2 \frac{a}{b}$
Position i=265			
I	0.00035	0.15553	-8.81
T	0.99340	0.83821	0.24
Position i=3606			
F	0.00035	0.16481	-8.85
L	0.99322	0.82893	0.26
Position i=4715			
L	0.99322	0.60488	0.71
P	0.00035	0.38886	-10.09

Nucleocapsid

Positions 203 and 204 belongs to the linker region between the N- and C-terminal RNA binding domains. In particular they are positioned in the SR region named after the marked abundance in S and R residues [14]. This region plays a role in cell signalling and RNA binding. It is a thought to be a highly flexible and intrinsically disordered region (IDP) of the protein. While the mutation R203K is a conservative mutation in terms of charge and size of the residue side chain, G204R introduces a bulky charged residue in place of a G residue. However, even though the loss of a G may reduce the backbone flexibility, R residues are common in IDP regions and R204 might introduce a new charge which is compatible with the RNA binding properties of this domain.

Spike

D614 is located in the SD1 domain of the spike protein. It is positioned in a loop right after a beta- strand and close to a completely solvent exposed disordered region far from the ACE2 interaction domain [15, 16]. Even though the resolution of the three available structures is not high enough to thoroughly discuss H-bond network, it is likely that D614 side chain does not establish close interactions with neighbouring residues. Conformationally, D614 lies in the region of left-handed helices. Thus, from the structural point of view, the mutation D614G is likely to be neutral ? no interactions lost from the missing carboxylate group ? or even beneficial for the overall protein stability: given the increased conformational freedom of G, it is a perfect residue in turns following beta structures.

Polyprotein

In Coronaviruses, ORF1 is translated to yield polyproteins that are processed by proteolysis with the production of intermediate and mature nonstructural proteins (nsps). In Figure 6 we highlight the position of the variant sites on the Polyprotein with respect to known domains, as annotated using the Conserved Domain Database [17]. Position 265 belongs to non-structural protein 2 (Nsp2, corresponding in this protein to residue 85) which is characterized by four transmembrane helices (TH) [18] but whose role is still unknown. More in detail position 85 lies at the N-terminal, which protrudes from the external face of the membrane. Interaction analysis of the corresponding protein from SARS virus shows that Nsp2 can form dimeric or multimeric complexes that can also involve additional viral proteins (Nsp3, Nsp4, Nsp6, Nsp8, Nsp11, Nsp16, ORF3a) [19] suggesting that Nsp2 might be involved in the viral life cycle. Nsp2 also interact with host proteins, such as prohibitin 1 and 2 [20], that are involved in cell cycle progression, migration, differentiation, apoptosis and mitochondrial biogenesis [21, 22, 23, 24]. Murine Hepatitis Virus and SARS-CoV strains deleted in this portion of the polyprotein [25] have a strongly reduced viral growth and RNA synthesis, but not protein processing. Ectopic expression of the nucleotide sequence coding for Nsp2 in murine cell infected by strains with Nsp2 deleted, allowed to detect its recruitment in viral complexes. It is therefore possible that Nsp2 be involved in global RNA synthesis, interaction

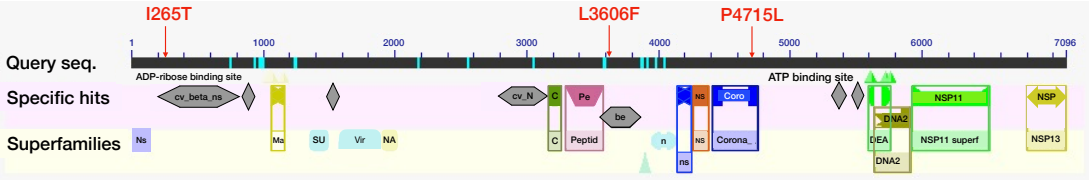


Figure 6. Position of the variant sites on the Polyprotein sequence, with indicated the domains as defined on the basis of the Conserved Domain Database (NCBI CDD, [17]).

with host proteins and maybe also pathogenesis [19]. Nsp2 seems to be present in infected cells in small vesicular foci, but in absence of additional viral proteins it localizes in cytoplasmic or nuclear membranes, without a specific target [26, 19]. Unfortunately, mutagenesis experiments are still not available, and therefore we still don't know if the variant observed at this site has any effect on *in vivo* virus properties. Position 3606 belongs to Nsp6, a protein that induces the formation of autophagosomes, as happen to cell due to starvation, but in greater number and lower diameter [27]. Autophagosomes can act as an innate defense against viral infection but they can promote the assembly of coronavirus replicase proteins [28, 29, 30]. Together with Nsp3 and Nsp4, Nsp6 moreover promotes the formation of the double membrane vesicles typically observed in SARS disease [31]. Nsp6 is a protein with 7 transmembrane domains [32] and position 3606 lays in the luminal loop between the first and second hydrophobic transmembrane domains. In the wild type Sars-CoV-2 sequence, we find 3 Phenylalanine residues just before position, where we find a Leucine. Instead, variants have a stretch of 4 Phenylalanine, which should increase hydrophobicity of this regions. Position 4715 belongs instead to the N-terminal domain of the RNA-directed RNA Polymerase (RdRp, comprised in Nsp12). By exploiting a three-dimensional structure of the catalytic domain of SARS-CoV RdRp [33] it was shown that this protein has the canonical palm and finger structure plus a N-terminal domain whose function is unclear. Through structural modelling with additional viral RdRps it was shown that the N-terminal domain bridges the fingers to from a channel through which nucleotides may diffuse. In poliovirus, the N-terminal part of the RdRp is involved in the proper folding of the fingers and the positioning of the active catalytic site and it may be supposed that the function is conserved in SARS coronaviruses [34] . This is also supported by the impossibility to correctly fold RdRp *in vitro* without expressing the N-terminal sequence [35]. Thus, despite no information is available concerning mutations at the specific sites identified in this work, the three variant sites in the polyprotein belongs to important and functional regions of the sequence.

Phylogenetic analysis and integration with variant information

The best phylogenetic model selected according to AIC and BIC was GTR+I+G, and the obtained phylogenetic tree is reported in Figure 7. The tree clearly shows the existence of well-defined groups. When integrating the information about the residues found in each sequence at each of the six sites, we demonstrate that these clades correspond to the five most frequent combinations of residues for our 6 variable positions, i.e. RGTLPD, RGTFPD, KRTLLG, RGTLLG and RGILLG (where residues are given in the order: protein Nucleocapsid sites 203, and 204, Polyprotein sites 265, 3606 and 4715, Spike site 614) and as they are perfectly correlated with the tree, in the following we will call them for simplicity Clade A, Clade B, Clade C, Clade D and Clade D, respectively. A graphical visualization of the distribution continents of each variant and of the five clades is reported in Figure 8 and Figure 9, respectively. Furthermore, the variation of the distribution of the six sites independently, and clades over time is shown in Supplementary Video S1 and S2.

Figure 7 and Figure 9 show a clear geographic pattern : while Clade A and Clade B are more associated to Asia and Oceania, the remaining three are more frequently observed in Europe, in agreement with the single variant analysis. Considering that the epidemic event began in Cina and taking into account the topology of the tree, we can infer that the Clade A clade is the ancestral one while the Clade B evolved from it in a second moment. In Europe, the Clade C and Clade D are sampled since the beginning while the Clade E evolved from a strain carrying

a configuration RGTLLG (corresponding to Clade D). In order to evaluate how the clades prevalence evenness changes over time in these macro-regions, we computed the Shannon index of the clades cumulative frequency over time (Figure 10). We found that the Shannon index values increase over time in all the continents, reaching a plateau in April, meaning that after an increase in variability in all macro-regions, the virus seems to be relatively stable at the moment. This could also be interpreted as an indication of reaching a optimized variant, but the information on variant performance will only be available in the future to confirm or confutate this hypothesis. The reached plateau index values vary among continents and the Sars-CoV-2 Oceania population resulted to be the most even in comparison to the others, indicating coexistence of several variants at comparable frequencies. Interestingly, Asia is one of the continents with the lowest final Shannon index value (in agreement with the absence of some of the clades). In Asia, the Shannon index increases in a stepwise-manner at the appearance of every new variant. We can hypothesized this low value could be due to the fact that most of the genomes are from strains isolated in China, the country that for first started the lockdown, reducing the spread of novel clades. This analysis is however based on cumulative frequencies, while it does not take into account the disappearance of variants in time. To explore this issue, we also calculated the Shannon index for sequences within the same intervals of times used to track the change in frequency of the single residue variants. Peaks indicate the emergence (from outside or by evolution of pre-existing strains) of variants and their increase in frequency. Values maintaining a high Shannon index correspond to situations where different variants coexist at comparable frequencies, while a decrease after a peak means that after the appearance of one or more new variants / clades, one of them (novel or old) takes over - reducing the variability.

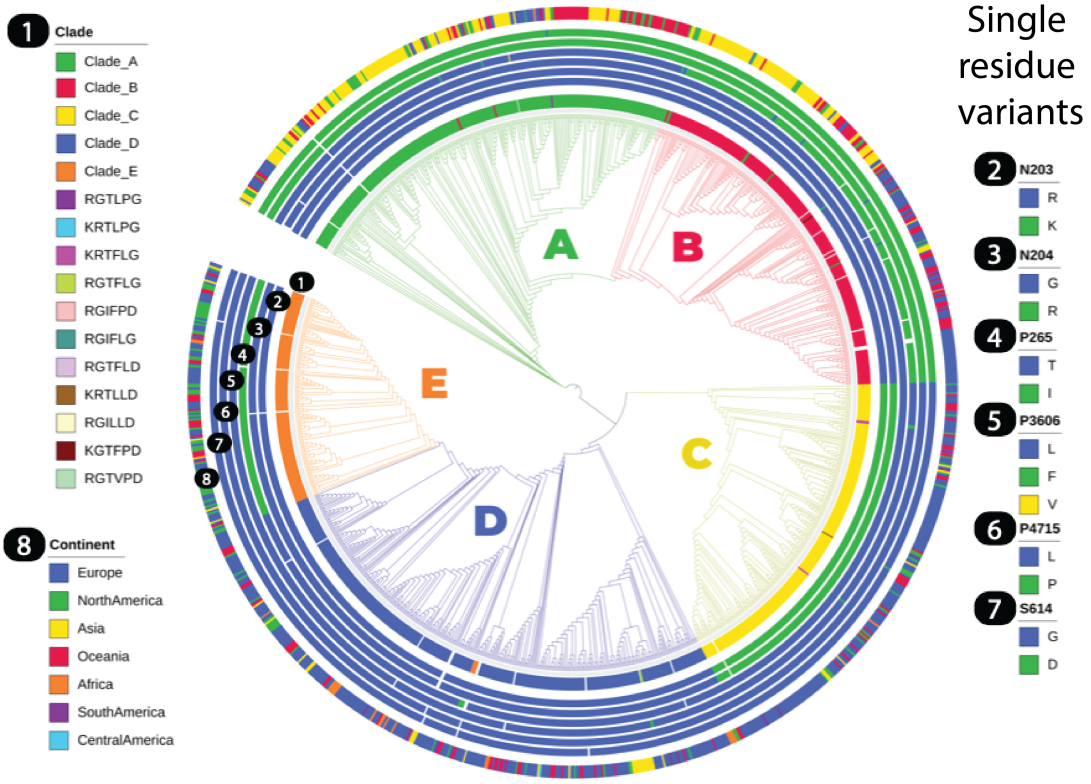


Figure 7. Phylogenetic tree using a concatenated multiple alignment of all coding sequences. Continent of origin, clades and variants are also shown.

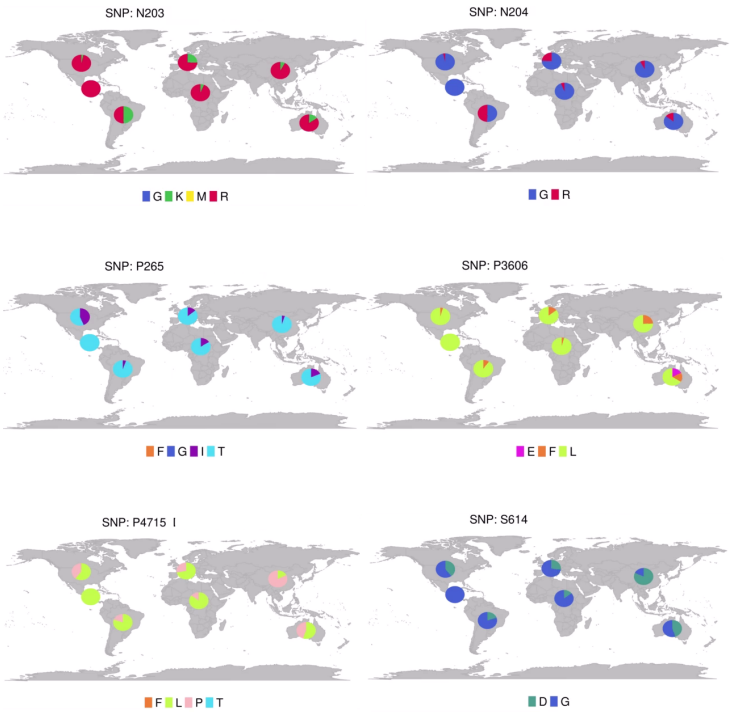


Figure 8. The abundance of the single residue variants in different geographical macro-regions considering all variants sampled up to April, 20, 2020 .

Conclusion

In this work, we identified six positions in coding sequences of the Sars-CoV-2 that are characterized by a different pattern of aminoacids when comparing italian sequences with those globally sequenced around the worlds. We found that the combination of these residuals identify five clades conserved on the global Sars-Cov-2 phylogenetic tree. This suggest that the characterization of these six positions can represent a fast and portable method for the Sars-Cov-2 typing. Using this approach we were able to follow the evolution of the virus over time among continents, showing that the different clades evolved in different moments and that their frequencies vary among continents.

These sites also seems to be the most variable across all countries, raising intriguing questions about their functional effects. One of these variant in particular, D614G on the Spike protein, is characterized by an increase in frequency since the end of January in most countries, and since then has overcome the original aplotype in most countries, suggesting it may indeed provide some advantage to the virus. However, to understand these issues, experimental data is required, for instance in the form of Tn-insertion mutagenesis, as performed on other viruses in the past [36].

Acknowledgments

All authors wish to thank the `www.gisaid.org` and all the researchers that contributed their sequences to the database for sharing fundamental data for research. We think collaboration is the only approach to counteract the spread of Sars-CoV-2 and other similar diseases. References for all sequences used in this work are in Supplementary Table 1. Authors from University of Milan wish to thank *Fondazione Romeo ed Enrica Invernizzi* for financial

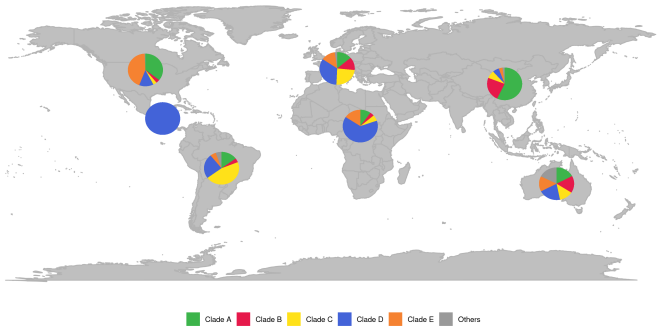


Figure 9. The abundance of all observed combinations of the amino acids at the 6 variable sites in different geographical macro-regions by using all variants sampled up to April, 20, 2020.

support.

References

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, et al. (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579: 270–273.

2. Wu F, Zhao S, Yu B, Chen YM, Wang W, et al. (2020) A new coronavirus associated with human respiratory disease in China. Nature 579: 265–269.

3. Lu Z, Kim JZ, Bassett DS (2020) Supervised chaotic source separation by a tank of water. Chaos 30.

4. Rothan HA, Byrareddy SN (2020) The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Journal of Autoimmunity 109: 102433.

5. Zhang T, Wu Q, Zhang Z (2020) Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Current Biology 30: 1346–1351.e2.

6. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nature Medicine 26: 450–452.

7. Wan Y, Shang J, Graham R, Baric RS, Li F (2020) Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. Journal of Virology 94: 1–9.

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

9. Livingston E, Bucher K (2020) Coronavirus Disease 2019 (COVID-19) in Italy. JAMA 323: 1335.

10. Buck MJ, Atchley WR (2005) Networks of Coevolving Sites in Structural and Functional Domains of Serpin Proteins. Molecular Biology and Evolution 22: 1627–1634.

11. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, et al. (2020) ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. Molecular Biology and Evolution 37: 291–294.

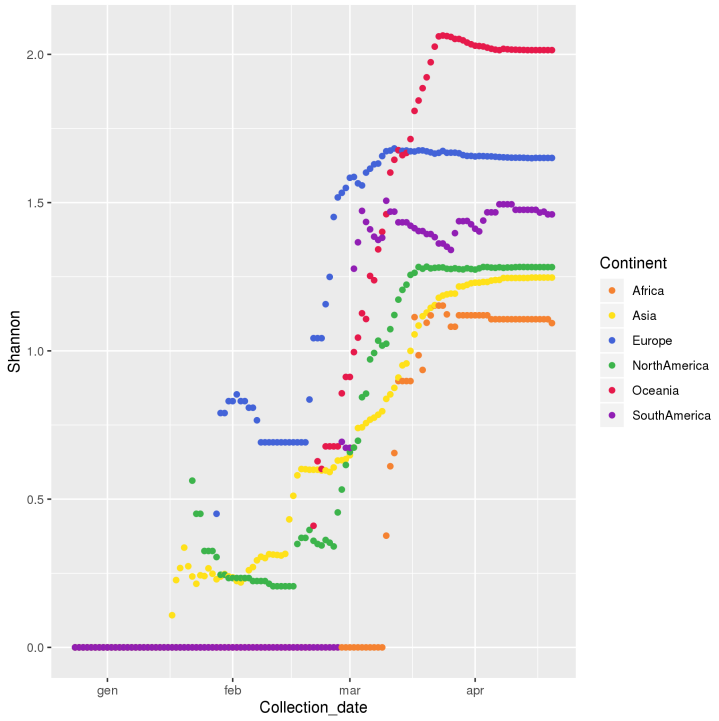


Figure 10. Shannon index as a function of time in the different macro-regions, computed on the cumulative frequencies of the clades defined using phylogenesis and variant information.

12. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

13. Letunic I, Bork P (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research 47: W256–W259.

14. McBride R, van Zyl M, Fielding BC (2014) The coronavirus nucleocapsid is a multifunctional protein. Viruses 6: 2991–3018.

15. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, et al. (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 367: 1260–1263.

16. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, et al. (2020) Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 181: 281–292.e6.

17. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, et al. (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic acids research 48: D265–D268.

18. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, et al. (2020) COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. Journal of Medical Virology : 1–5.

19. von Brunn A, Teepe C, Simpson JC, Pepperkok R, Friedel CC, et al. (2007) Analysis of intraviral protein-protein interactions of the SARS coronavirus ORFeome. PLoS ONE 2.

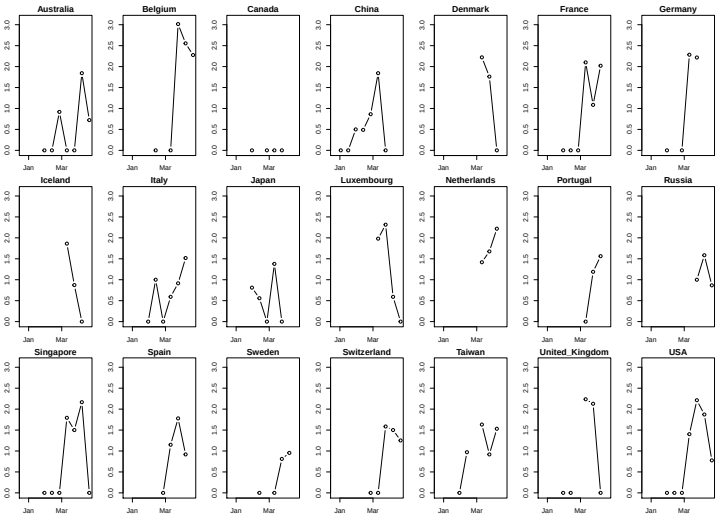


Figure 11. Shannon index as a function of time in the different countries, calculated for sequences within each time interval. Variability increases when a new variant appears, then it stays more or less stable when the existing variants maintain their relative frequencies. However, if one of the variants becomes dominant, then variability decreases again after a peak. Increasing variability in time means that the arrival of novel variants continues steadily, or that the existing ones become more homogeneous.

20. Cornillez-Ty CT, Liao L, Yates JR, Kuhn P, Buchmeier MJ (2009) Severe Acute Respiratory Syndrome Coronavirus Nonstructural Protein 2 Interacts with a Host Protein Complex Involved in Mitochondrial Biogenesis and Intracellular Signaling. Journal of Virology 83: 10314–10318.

21. Fusaro G, Dasgupta P, Rastogi S, Joshi B, Chellappan S (2003) Prohibitin Induces the Transcriptional Activity of p53 and Is Exported from the Nucleus upon Apoptotic Signaling. Journal of Biological Chemistry 278: 47853–47861.

22. Merkwirth C, Langer T (2009) Prohibitin function within mitochondria: Essential roles for cell proliferation and cristae morphogenesis. Biochimica et Biophysica Acta - Molecular Cell Research 1793: 27–32.

23. Rajalingam K, Wunder C, Brinkmann V, Churin Y, Hekman M, et al. (2005) Prohibitin is required for Ras-induced Raf-MEK-ERK activation and epithelial cell migration. Nature Cell Biology 7: 837–843.

24. Sun L, Liu L, Yang XJ, Wu Z (2004) Akt binds prohibitin 2 and relieves its repression of MyoD and muscle differentiation. Journal of Cell Science 117: 3021–3029.

25. Graham RL, Sims AC, Brockway SM, Baric RS, Denison MR (2005) The nsp2 Replicase Proteins of Murine Hepatitis Virus and Severe Acute Respiratory Syndrome Coronavirus Are Dispensable for Viral Replication. Journal of Virology 79: 13399–13411.

26. Prentice E, McAuliffe J, Lu X, Subbarao K, Denison MR (2004) Identification and Characterization of Severe Acute Respiratory Syndrome Coronavirus Replicase Proteins. Journal of Virology 78: 9977–9986.

27. Cottam EM, Whelband MC, Wileman T (2014) Coronavirus NSP6 restricts autophagosome expansion. Autophagy 10: 1426–1441.

28. Orvedahl A, Alexander D, Tallóczy Z, Sun Q, Wei Y, et al. (2007) HSV-1 ICP34.5 Confers Neurovirulence by Targeting the Beclin 1 Autophagy Protein. Cell Host and Microbe 1: 23–35.

29. Suhy DA, Giddings TH, Kirkegaard K (2000) Remodeling the Endoplasmic Reticulum by Poliovirus Infection and by Individual Viral Proteins: an Autophagy-Like Origin for Virus-Induced Vesicles. Journal of Virology 74: 8953–8965.

30. Wileman T (2006) Sites for Virus Replication. Science : 875–878.

31. Angelini MM, Akhlaghpour M, Neuman BW, Buchmeier MJ (2013) Severe Acute Respiratory Syndrome Coronavirus Nonstructural Proteins 3, 4, and 6 Induce Double-Membrane Vesicles. mBio 4: 165.

32. Oostra M, Hagemeijer MC, van Gent M, Bekker CPJ, te Lintelo EG, et al. (2008) Topology and Membrane Anchoring of the Coronavirus Replication Complex: Not All Hydrophobic Domains of nsp3 and nsp6 Are Membrane Spanning. Journal of Virology 82: 12392–12405.

33. Xu X, Liu Y, Weiss S, Arnold E, Sarafianos SG, et al. (2003) Molecular model of SARS coronavirus polymerase: Implications for biochemical functions and drug design. Nucleic Acids Research 31: 7117–7130.

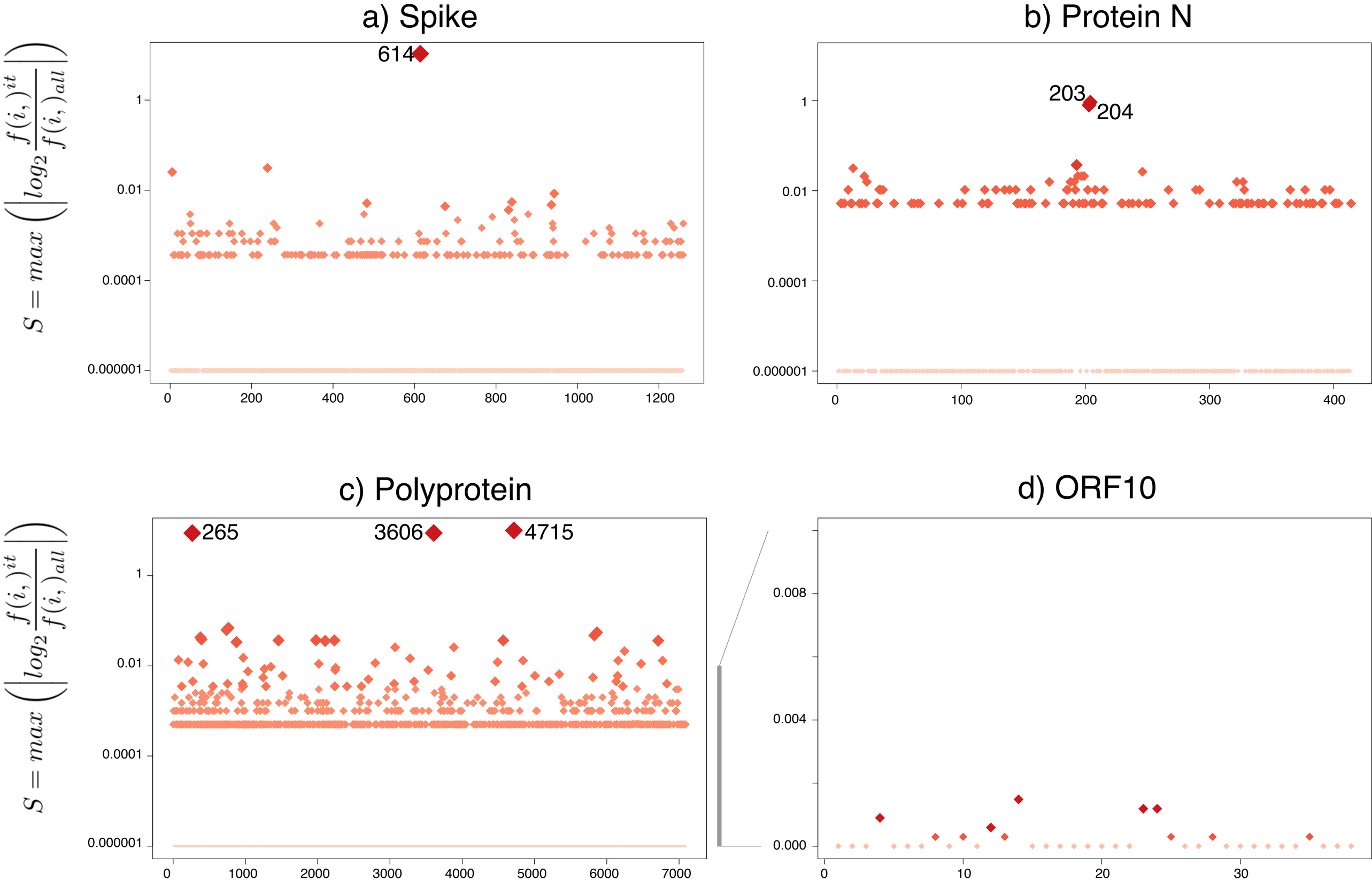
34. Thompson AA, Peersen OB (2004) Structural basis for proteolysis-dependent activation of the poliovirus RNA-dependent RNA polymerase. EMBO Journal 23: 3462–3471.

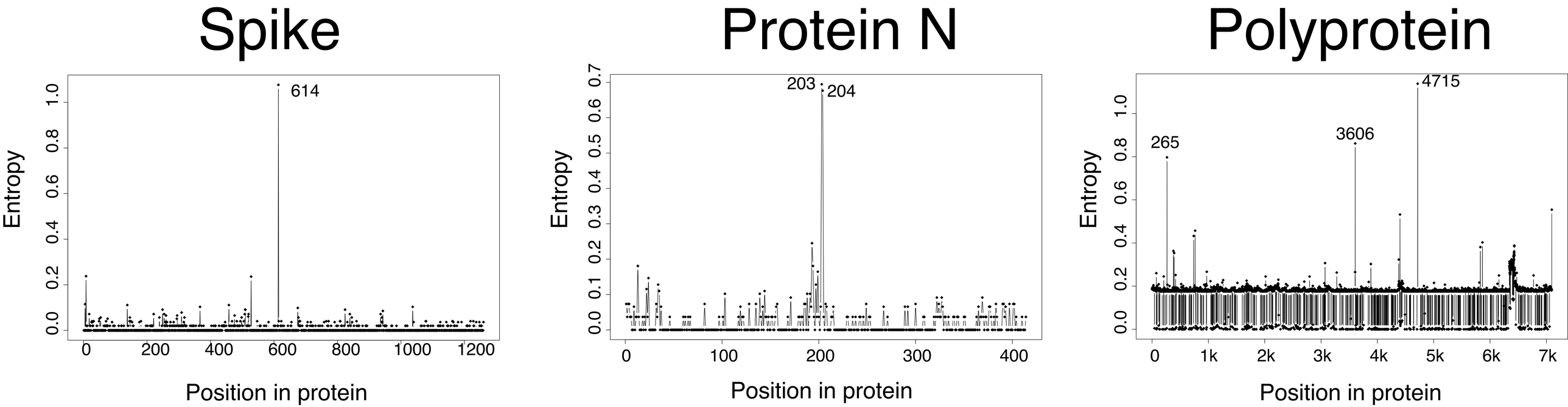
35. te Velthuis AJW, Arnold JJ, Cameron CE, van den Worm SHE, Snijder EJ (2010) The RNA polymerase activity of SARS-coronavirus nsp12 is primer dependent. Nucleic Acids Research 38: 203–214.

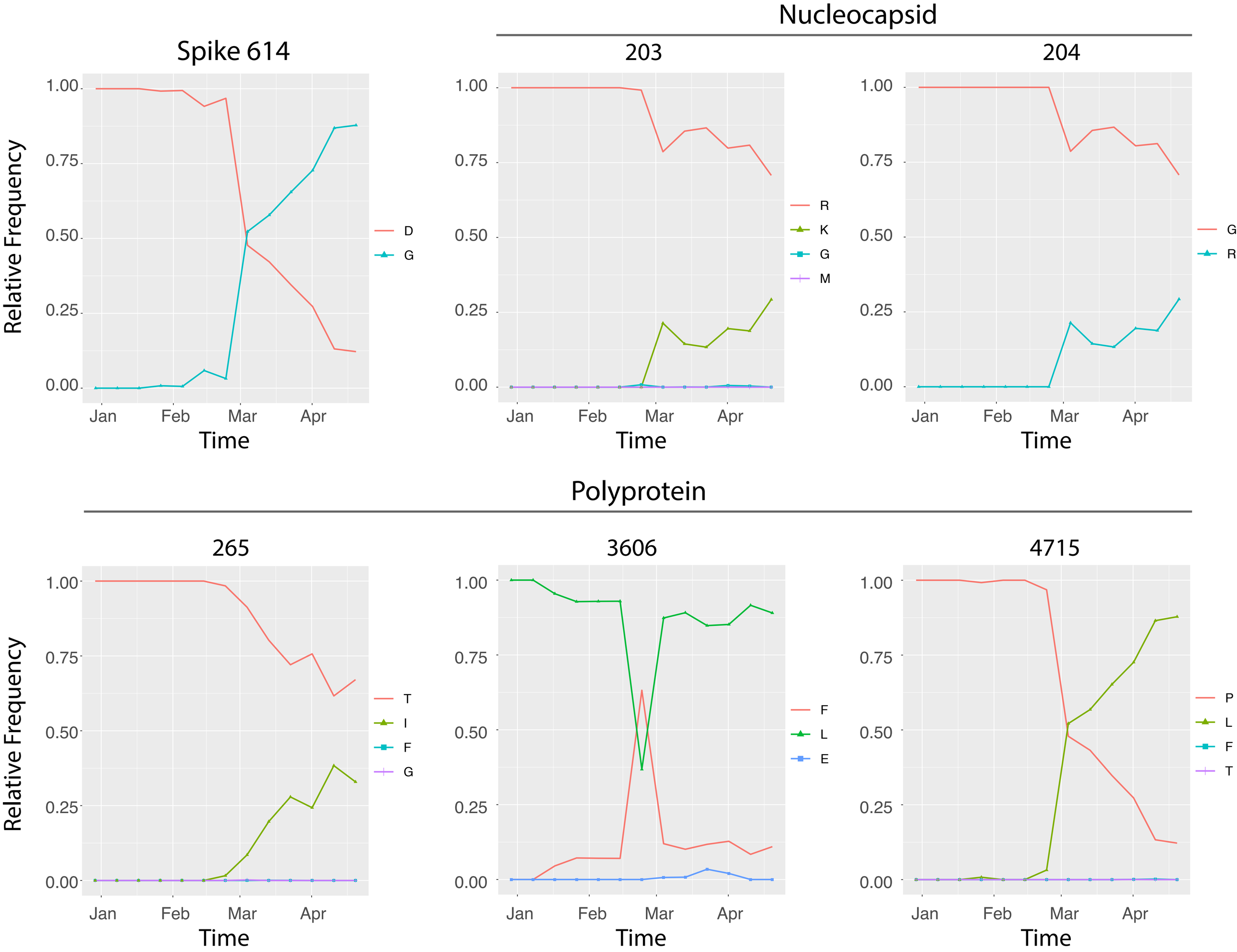
36. Fulton BO, Sachs D, Schwarz MC, Palese P, Evans MJ (2017) Transposon Mutagenesis of the Zika Virus Genome Highlights Regions Essential for RNA Replication and Restricted for Immune Evasion. Journal of Virology 91: 1–16.

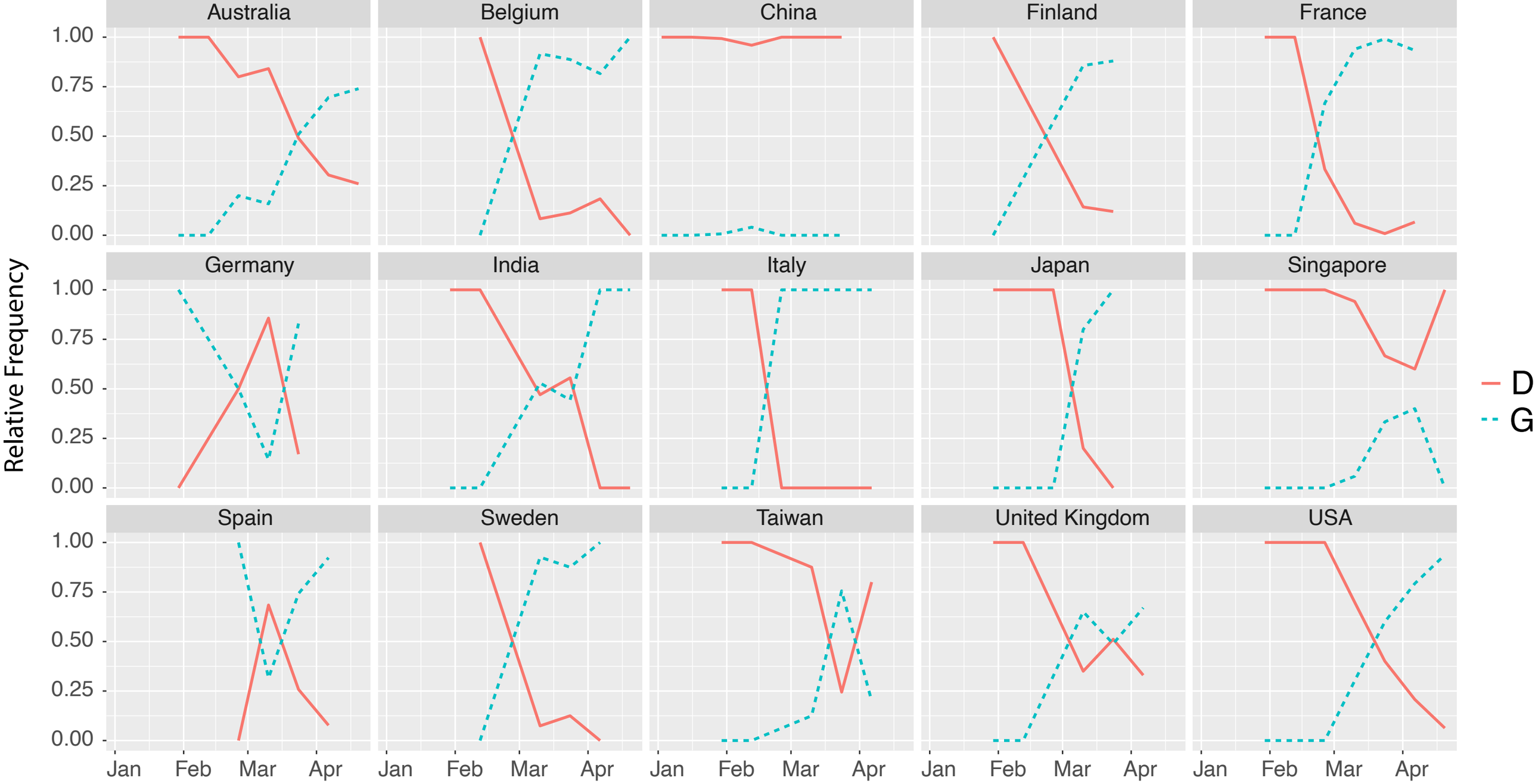
Figure Legends

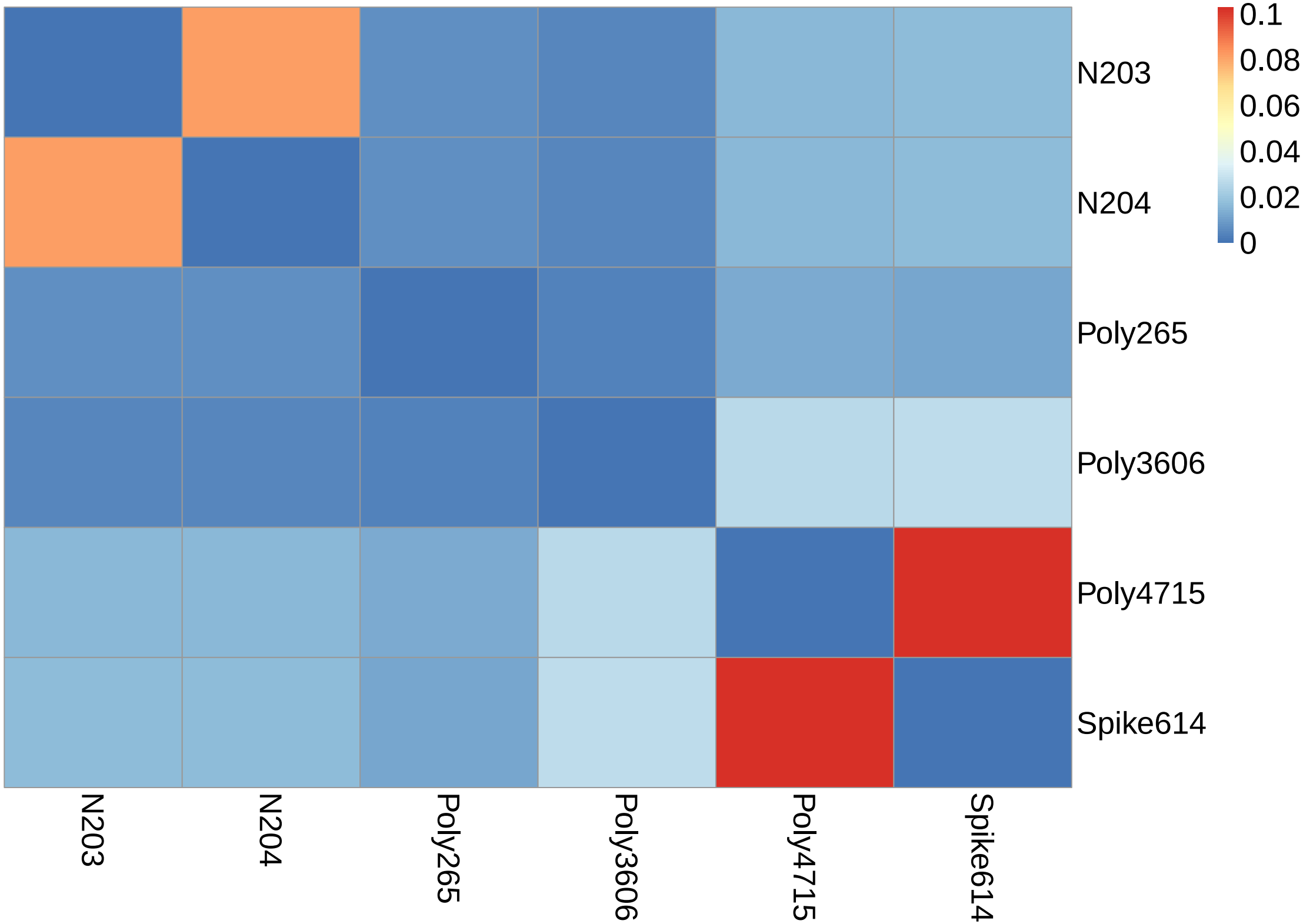
Tables

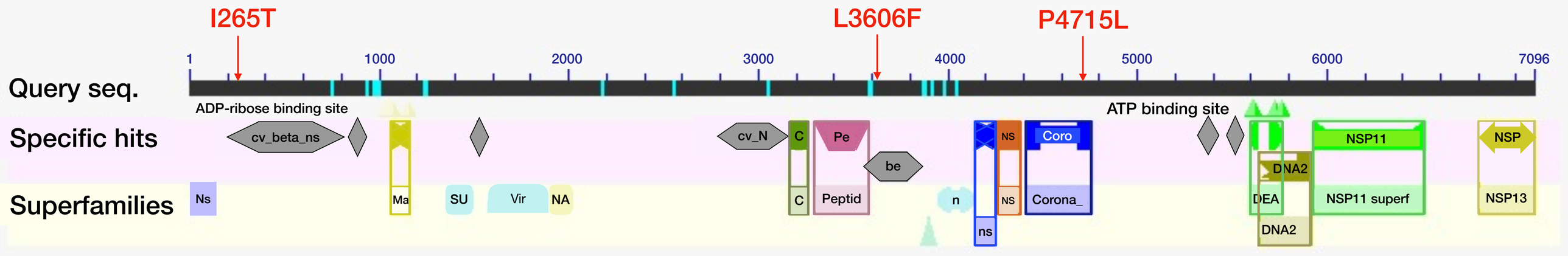

a) Spike
b) Protein N
c) Polyprotein
d) ORF10
$S = max\left(\left|log_2 \frac{f(i,)^{it}}{f(i,)_{all}}\right|\right)$
614
203
204
265
3606
4715
1
0.01
0.0001
0.000001
0.008
0.004
0.000
0
200
400
600
800
1000
1200
100
300
1000
2000
3000
4000
5000
6000
7000
10
20
30

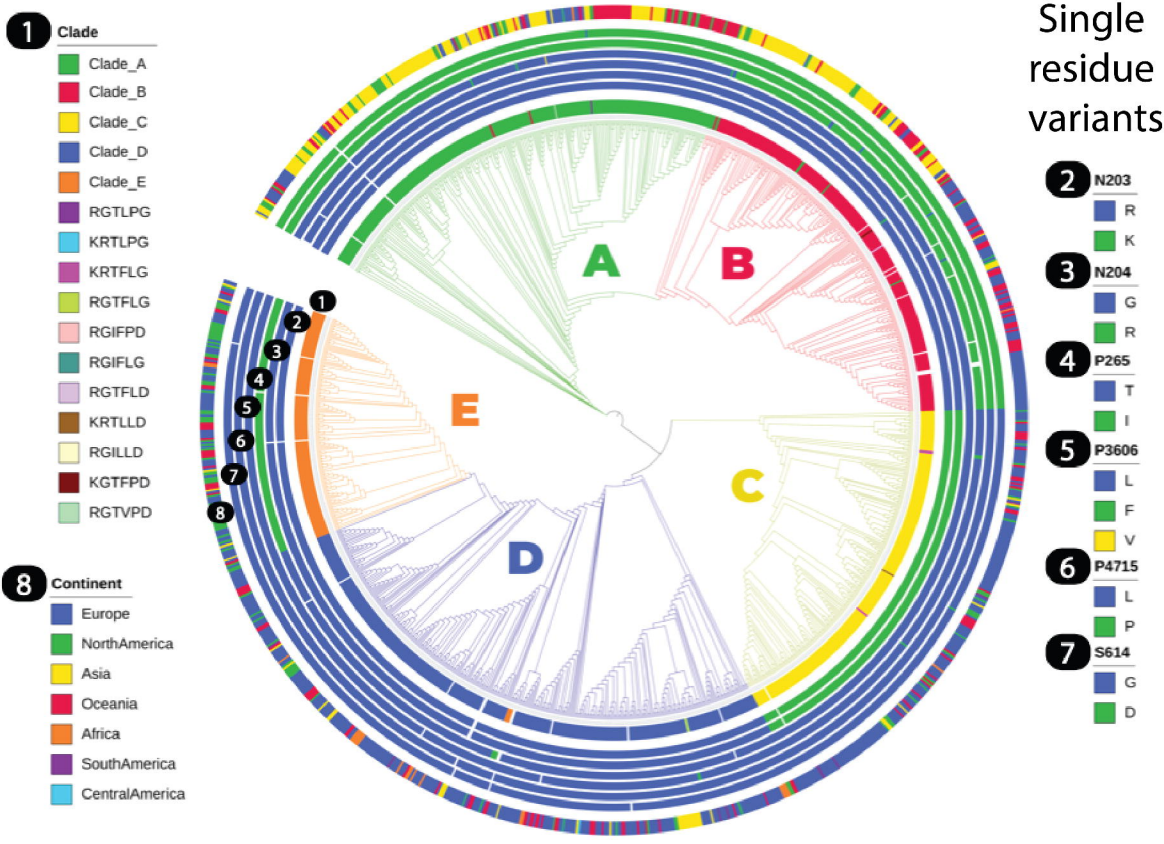

Spike
Entropy
614
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
1200
Position in protein
Protein N
Entropy
203
204
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
Position in protein
Polyprotein
Entropy
265
3606
4715
0.0
0.2
0.4
0.6
0.8
1.0
0
1k
2k
3k
4k
5k
6k
7k
Position in protein

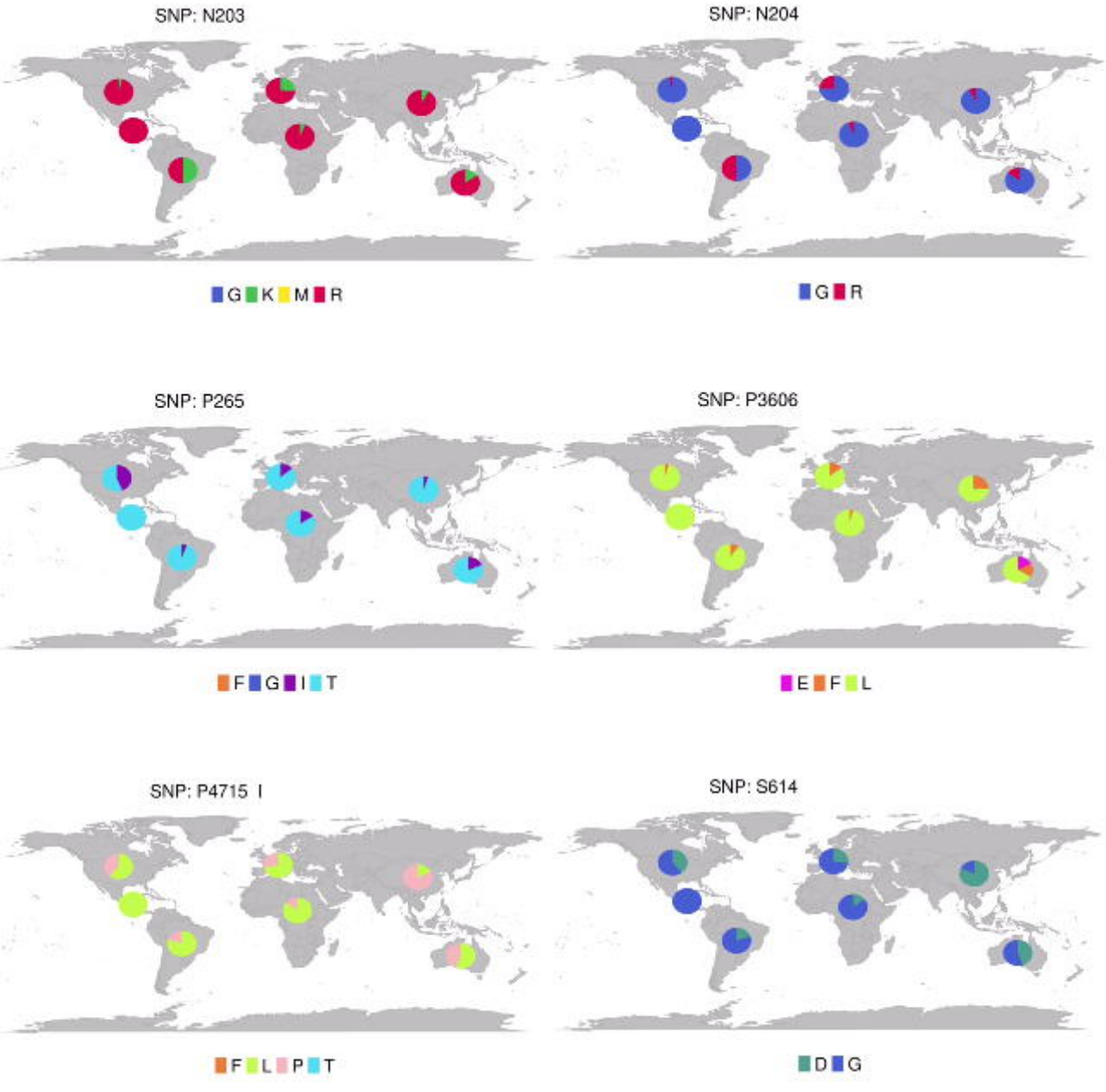

Nucleocapsid
Spike 614
203
204
Relative Frequency
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
Time
D
G
R
K
G
M
G
R
Polyprotein
265
3606
4715
T
I
F
G
F
L
E
P
L
F
T

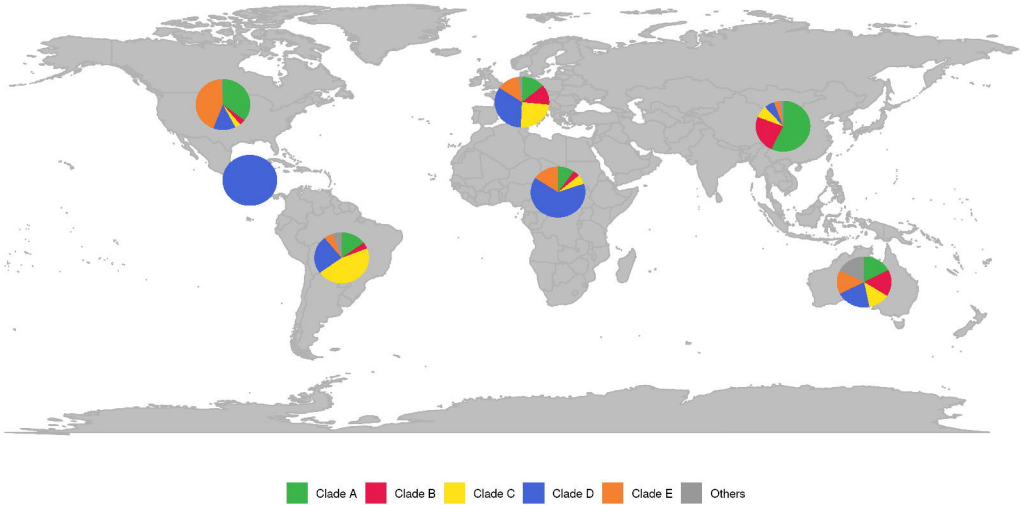

Australia
Belgium
China
Finland
France
Germany
India
Italy
Japan
Singapore
Spain
Sweden
Taiwan
United Kingdom
USA
Relative Frequency
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
D
G

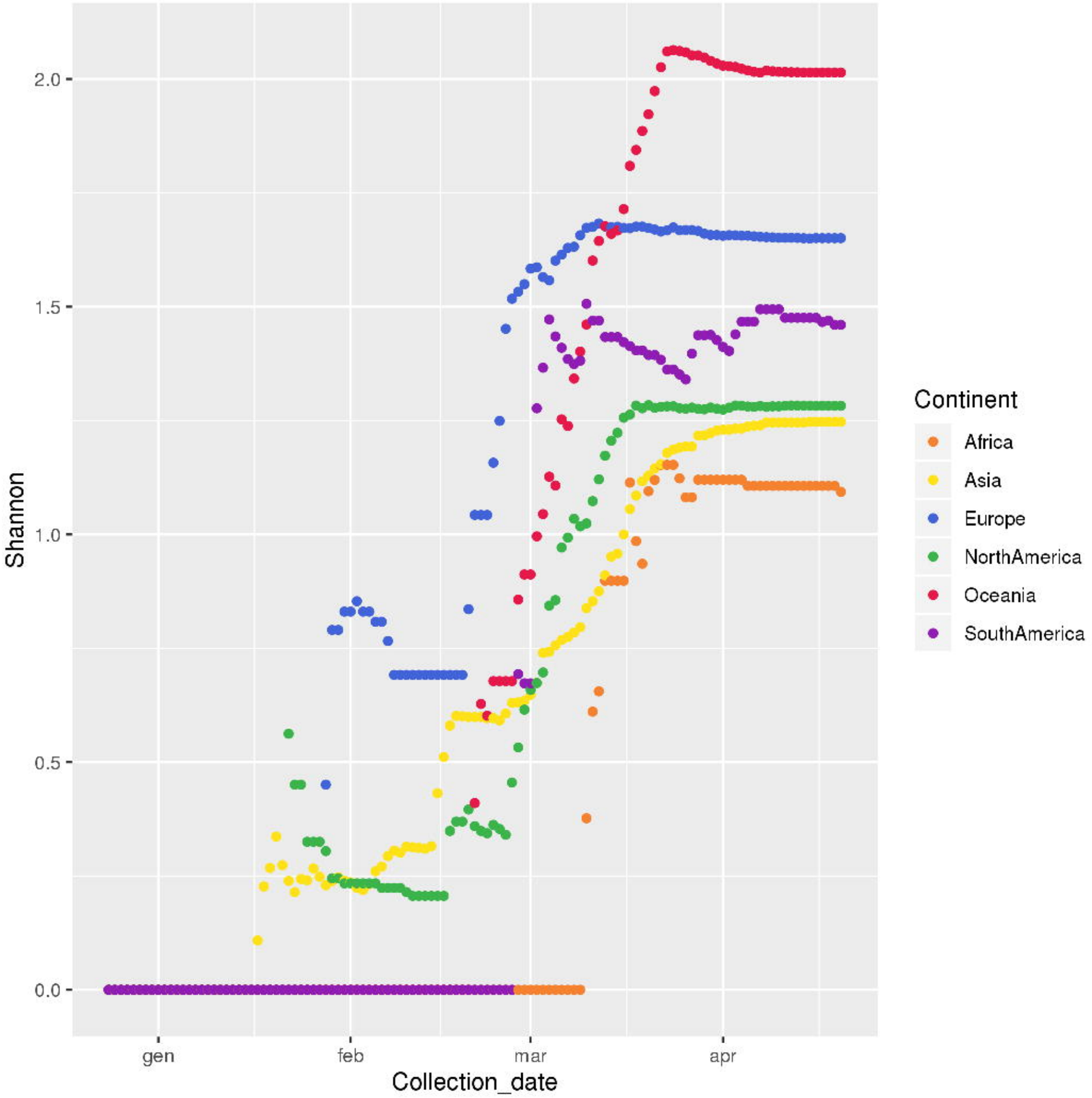

N203
N204
Poly265
Poly3606
Poly4715
Spike614
0.1
0.08
0.06
0.04
0.02
0

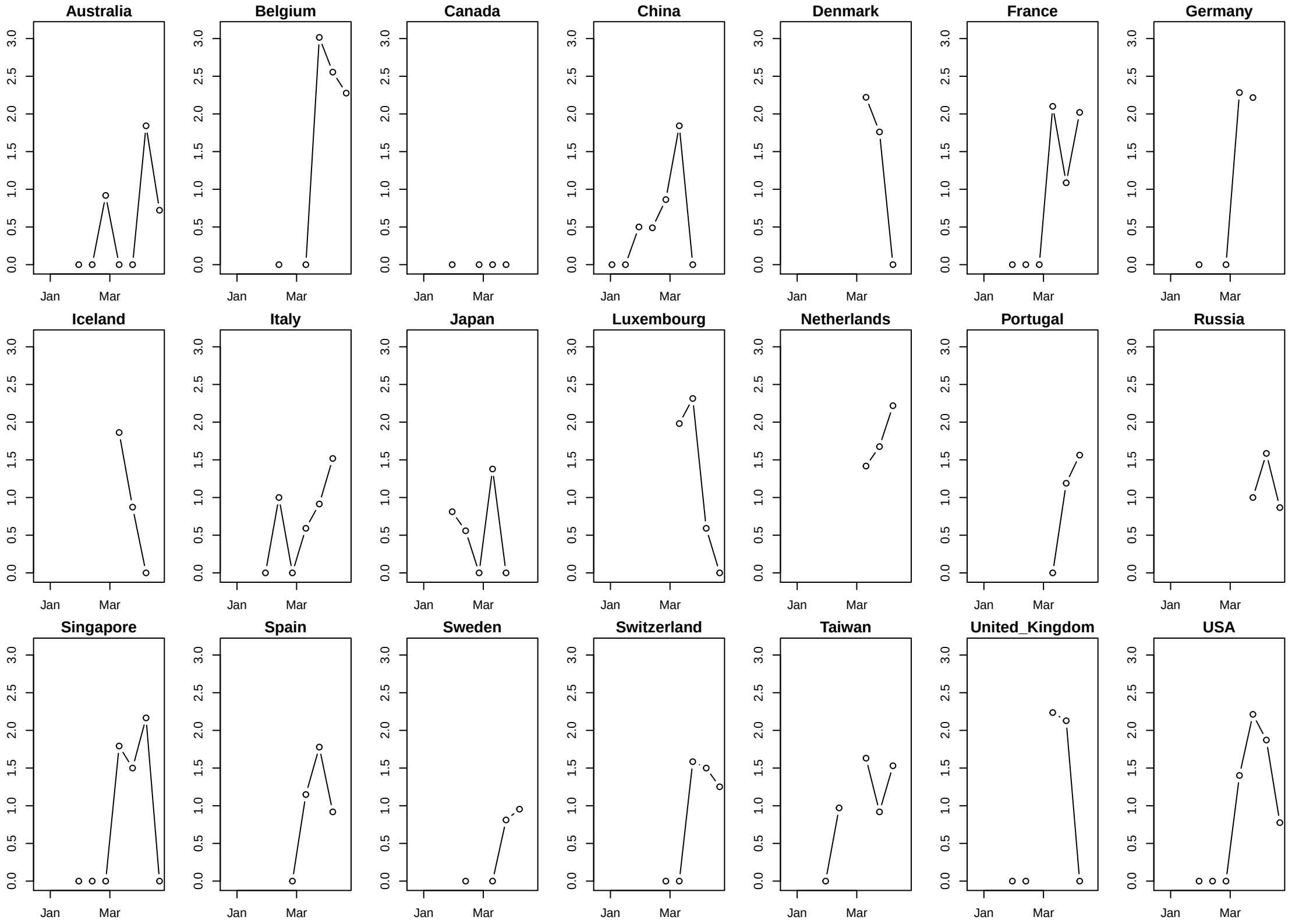

I265T
L3606F
P4715L
Query seq.
1
1000
2000
3000
4000
5000
6000
7096
ADP-ribose binding site
ATP binding site
Specific hits
cv_beta_ns
cv_N
C
Pe
be
NS
Coro
NSP11
NSP
DNA2
Superfamilies
Ns
Ma
SU
Vir
NA
C
Peptid
n
ns
NS
Corona_
DEA
DNA2
NSP11 superf
NSP13


1 Clade
Clade_A
Clade_B
Clade_C
Clade_D
Clade_E
RGTLPG
KRTLPG
KRTFLG
RGTFLG
RGIFPD
RGIFLG
RGTFLD
KRTLLD
RGILLD
KGTFPD
RGTVPD
8 Continent
Europe
NorthAmerica
Asia
Oceania
Africa
SouthAmerica
CentralAmerica
A
B
C
D
E
1
2
3
4
5
6
7
8
Single residue variants
2 N203
R
K
3 N204
G
R
4 P265
T
I
5 P3606
L
F
V
6 P4715
L
P
7 S614
G
D


SNP: N203
G K M R
SNP: N204
G R
SNP: P265
F G I T
SNP: P3606
E F L
SNP: P4715 I
F L P T
SNP: S614
D G


Clade A
Clade B
Clade C
Clade D
Clade E
Others


Shannon
2.0
1.5
1.0
0.5
0.0
gen
feb
mar
apr
Collection_date
Continent
Africa
Asia
Europe
NorthAmerica
Oceania
SouthAmerica


Australia
Belgium
Canada
China
Denmark
France
Germany
Iceland
Italy
Japan
Luxembourg
Netherlands
Portugal
Russia
Singapore
Spain
Sweden
Switzerland
Taiwan
United_Kingdom
USA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Jan
Mar