Spread of SARS-CoV-2 Genomes on Genomic Index Maps of Hierarchy - Compared with B.1.1.7 Lineage on BLAST

Jeffrey Zheng (conjugatelogic@yahoo.com)
Yunnan University https://orcid.org/0000-0003-4225-7077
Yang Zhou
Yunnan University
Minghan Zhu
Yunnan University
Mu Qiao
Yunnan University
Zhigang Zhang
Yunnan University

Research Article

Keywords: genomic index, visual maps, phylogeny, projection, information entropy, diversity measure, global invariant, hierarchical projection, optimization

Posted Date: February 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-31883/v4

Spread of SARS-CoV-2 Genomes on Genomic Index Maps of Hierarchy - Compared with B.1.1.7 Lineage on BLAST

Jeffrey Zheng, Yang Zhou, Minghan Zhu, Mu Qiao, Zhigang Zhang

Abstract COVID-19 patients worldwide are conveniently described by position information to collect samples, and modern GIS maps are useful to show influenced flows and numbers of patients on various regions of a pendamic. From an analysis viewpoint, it is more interesting to organize genomic information into a phylogenic tree with multiple branches and leaves in representations. Clusters of genomes are organized as phylogenic trees to represent intrinsic information of genomes. However, there are structural difficulties in projecting phylogenetic information into 2D distributions as GIS maps naturally.

Considering advanced generating schemes of phylogenetic trees, information entropy provides ultra optimal properties in the minimum computational complexity, superior flexibility, better stability, improved reliability and higher quality on global constructions.

In this paper, a novel projection is proposed to arrange SARS-CoV-2 genomes by genomic indexes to make a structural organization as 2D GIS maps. For any genome, there is a unique invariant under certain conditions to provide an absolute position on a specific region. In this hierarchical framework, it is possible to use a visual tool to represent any selected region for clustering genomes on refined effects. Applied diversity measure to a given set of genomes, equivalent clusters and complementary visual effects are provided between genomic index maps and phylogenetic trees.

Sample genomes of three UK new lineages are aligned by BLAST as a basis on both RNA-dependent RNA polymerase RDRP segments and whole genomes. Selected

Jeffrey Zheng[1,2]
[1] Key Laboratory of Quantum Information of Yunnan; [2] Key Laboratory of Software Engineering of Yunnan; Yunnan University, Kunming, e-mail: conjugatelogic@yahoo.com

Yang Zhou, Minghan Zhu, Mu Qiao
Yunnan University

Zhigan Zhang
School of Life Sciences & Technology, Yunnan University

This work was supported by the NSFC (62041213), the Key Project on Electric Information and Next Generation IT Technology of Yunnan (2018ZI002).

regions and various projections show spread effects of five thousand SARS-CoV-2 genomes in 72 countries on both RDRP and whole genomes, and six special countries/regions are selected on genomic index maps.
Based on genomic index maps, one SNV of two genomes on B.1.1.7 lineage can be identified from a unit of 10^{-4} probability measure to a unit of 10^{-6} difference for genomic indexes on a special 'G' projection to extract the finest variation.
Further exploration on optimal classification and phylogenetic analysis of genomic index maps and phylogenetic trees on SARS-CoV-2 genomes worldwide are discussed.
Keywords genomic index, visual maps, phylogeny, projection, information entropy, diversity measure, global invariant, hierarchical projection, optimization

Introduction

The outbreak of SARS-CoV-2 caused COVID-19 to start in Dec. 2019 and is now pendamic. To the date of 26 January 2021, there are more than 100 million confirmed cases and 2.15 million deaths worldwide. An understanding of the prevalence and contagiousness of the disease and of whether the strategies used to contain it to date have been successful is important for understanding future containment strategies.

One excellent strategy for containment of SARS-CoV-2 is to collect sample genomes globally into the GISAID genetic database [1] for infected viruses. Based on this effective activity, Nextstrain provides Phylogenetic tree [2] to organize sample datasets from different places to categorize them as clusters under the maximal likelihood relationship to view intrinsic variations among SARS-CoV-2 genomes. Based on phylogenetic information, a dynamic simulation system provides flexible illustrations on selected branches [4] to support medical doctors, virological experts, biomedical specialists and psychologic doctors for detailed treatments on COVID-19 patients.

Advanced Researches in Phylogenetic Analysis

The NCBI developed Basic Local Alignment Search Tool BLAST [3] in 1990s to provide powerful software tools for generating phylogenetic trees under a list of optimal inference conditions [6]-[30]: maximum likelihood [6, 7, 24], probability [8], Bayers [9], stochastic search [10], unalignable sequences [11], best fit model [12], tradition and Bayers [13], reconstruction [14], multiple alternative phylogenies [15], phylogenetic diversity measures [16], entropy approach [17], Shannon entropy and mutual information [18], viral phylogenomics [19], phylogenetic tree building [20], IQ-TREE [21], neural network [29], and deep learning classifier [30].

Viral phylogentics using an alignment-free method [19] provide optimal length of k-mer on N genomes of a phylogenetic tree to have computational complexity

applying cumulative relative entropy and Shannon entropy on $O(N)$, significantly faster than miminum likelihood or Bayers alignment on $O(N^2)$.

Useful technologies to build phylogenetic trees are viewed in [20]. Special problems in data collection of the world for SARS-CoV-2 are discussed in [22]. The key difficulties of phylogenetic analysis of SARS-CoV-2 are described in [23]. On wider researches on SARS-CoV-2 of phylogenetic analysis, a list of researches are carried out: phylogenetic supertree [25], informative subtype marker ISM [26], CG dinucleotide [27], CpG deficiency [28], classification and geographical analysis [29], light-weight classifier [30], and phylogenetic structure [31], S protein [32], and stability [33].

In [26], informative subtype marker ISM applied entropy analysis and ISM extraction to simulate Nextstrain through GISAID clades of SARS-CoV-2 genomes in details. In this scheme, key positions of relevant open reader frames ORFs associated with probability measures on time variations to describe viral evolutionary information from historic datasets.

For N number of unique sequences, L width of alignment, a size of alphabet, three executable complexity as follows.

Scheme	Complexity	Description
Minimum likelihood Bayers alignment	$O(N^2 \times L \times a)$	Common optimal schemes
Fasttree	$O(N^{1.25} \times \log(N) \times L \times a)$	Nextstrain's phylogenetic trees
Information entropy	$O(N \times L \times a)$	Fastest optimal scheme

Limitations of Phylogenetic Representations

Further arrangement may not be a direct approach. Regular zoom operators in GIS could be simulated along deeper or upper movement along branching nodes in a phylogenetic tree. Since phylogenetic trees correspond to neither 1D nor 2D structures, it is difficult to rearrange various subtrees [25] as visual objects. Using BLAST or MEGA packages, expensive computational complexity may be required to process N genomes to handle a set of phylogenetic trees.

In general, effective projections for a subset of phylogenetic trees provide a natural projection, and other forms of visual representations could not be directly supported.

Phylogenetic Trees in Nextstrain

The phylogenetic tree of Nextstrain is based on the maximal likelihood relationship to organize genomic datasets as hierarchical clusters under differential information. After a sample genome of SARS-CoV-2 compared with root node and following branch nodes recursively, it is possible to push it into the most likelihood node that contains the most similar genomes to be a target group. Since a genome contains a

long sequence, there are multiple relationships among various clusters in the phylogenetic tree shown in Fig. 1. Using GIS maps, it is useful to see various genomes distributed worldwide.

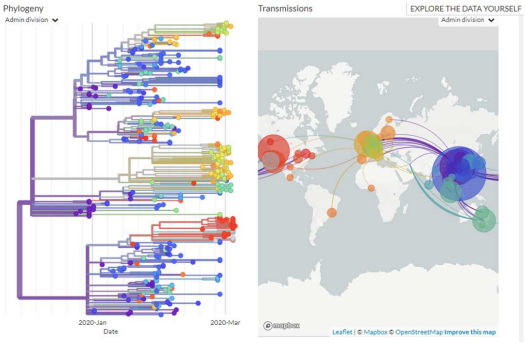


Fig. 1 The phylogenetic tree of real cases over global on Nextstrain

Difficulties in Phylogenetic Analysis of SARS-CoV-2 Data

A list of difficulties are discussed in [23], phylogenetic analysis of SARS-CoV-2 data is challenging due to numeric difficulties and the rugged likelihood surface.

Larger taxa on a low number of distinct site patterns have large topologic variability.

Signal is weak, it is difficult for standard phylogenetic significant tests.

Bayersian tree interferences use a plausible tree set for computing summary statistics on trees ...

Since old phylogenetic trees were generated from original genomes, there may not contain invariant structures to support new variations and mutations emergent from larger numbers of genomes everywhere.

In general, huge number of new genomes collected over the world makes extensive structural difficulties to use phylogenetic trees constructed for update and extensions.

Combination, Matrix and Thermodynamics

In modern mathematics and physics, there are many theoretical constructions to handle invariant and variation problems for entropy issues [35]-[65] such as combinatorial mathematics, combinatorial theory, combinatorics, multiple variable complex theory, statistical physics, thermodynamics, thermostatistics, statistical mechanics et al.

Variant Construction

In this direction, vectors, matrices and invariant measurements are described relevant to wider applications [67]-[70] on variant construction [71]-[74].

The genomic index provides unique identification for each genome to be an invariant under given conditions. Based on these types of global quantitative characteristics, it is convenient for large numbers of genomes to be located in a certain geometric region to be collected as clusters.

Different entropy quantities were discussed in separate papers: Visualization of SARS-CoV-2 Genomes on Genomic Index Maps [75], Visualizations of Topological Entropy on SARS-CoV-2 Genomes in Multiple Regions [76], Visual Variations between Pairs of SARS-CoV-2 Genomes on Integrated Density Matrix [77], Visualizations of Combinatorial Entropy Index on Whole SARS-CoV-2 Genomes [78].

Considering this is an extremely important research direction, it is necessary to handle this topic from a foundation level to provide additional information to explore hidden structures among this type of multiple levels of hierarchical constructions from a visual representation viewpoint.

Materials and Methods

Input on Four Meta Symbols

For genomes, each element of input sequences is composed of four meta symbols: {A,C,G,T}.

The First Order of Combinations

From a combinatorial viewpoint, the first order of combinations from the four symbols is composed of sixteen states as a lattice of hierarchy, as shown in Fig. 2.

The sixteen states $SS = \{\emptyset, A, C, G, T, AC, AG, AT, CG, CT, GT, CGT, AGT, ACT, ACG, ACGT\}$ can be mapped into the sixteen numbers $SI = \{0, 1, 2, \cdots, 15\}$ to represent a 1D linear structure with 16 distinct positions. For a segment of a genome with m elements, there are four meta measures: $\{m_A, m_C, m_G, m_T\} = \{m_1, m_2, m_3, m_4\}$ and sixteen combinatorial measures: $\{m_i\}, 0 \leq i \leq 15$ to correspond a meta measuring vector with four elements and a combinatorial measuring vector with sixteen elements, respectively.

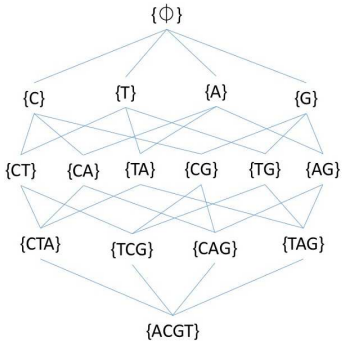


Fig. 2 Sixteen combinations of four meta-symbols in a hierarchy of a lattice

Multiple Probability Measures

When a genome contains m elements, the numbers of four Meta symbols can be counted. Let $m_s, s \in SS$ be a number of symbols s and p_s be a probability measure. We have the following equations for multiple probability measures.

$$\begin{aligned} m &= m_A + m_C + m_G + m_T \\ p_s &= \frac{m_s}{m}, s \in SS \\ 1 &= p_A + p_C + p_G + p_T \end{aligned}$$

Under multiple probability conditions, there are sixteen distinct probability measures $\{p_i\}_{i=0}^{15}, 0 \leq p_i \leq 1, i \in SI$ respectively.

Two Workflows from Input to Output

Two workflows (1) and (2) can be identified by the type of output.

(1) Vector of Genome $\rightarrow$ Probability $\rightarrow$ Sixteen Probability Vectors
$\rightarrow$ Entropy $\rightarrow$ Sixteen Indexes
$\rightarrow$ Selection $\rightarrow$ An Index
(2) $\{$Pair of Indexes$\}$ $\rightarrow$ Mapping $\rightarrow$ A Genomic Index Map

Genomic Index Projection and Genomic Index Map

Three workflows are described in three parts as input, output and process.

In Step (1), one index of 16 Combinatorial Entropies can be generated.

$$
\begin{aligned}
\textit{Input}: &\ N \text{ elements in a genome, } N = m \times M \\
\textit{Output}: &\ 1 \text{ (an index} \in [0,\ \log_2(m+1)]) \\
\textit{Process}: &\ N \overset{\text{Segment}}{\rightarrow} m \times M \overset{\text{Meta-Measure}}{\rightarrow} 4 \times M \overset{\text{Combination}}{\rightarrow} 16 \times (m+1) \\
&\overset{\text{Entropy}}{\rightarrow} 16 \overset{\text{Selection}}{\rightarrow} 1 \\
CF: &\ 16 \text{ (Total number of selections)}
\end{aligned}
$$

In Step (2), a genomic index map can be generated from multiple sets of sixteen indexes.

$$
\begin{aligned}
\textit{Input}: &\ \forall(x,y) \in \text{Multiple sets of sixteen indexes}, x, y \in [0,\ \log_2(m+1)] \\
\textit{Output}: &\ \text{An Index Map on } [0,\ \log_2(m+1)] \times [0,\ \log_2(m+1)] \text{ Region} \\
\textit{Process}: &\ \forall(x,y) \overset{Projection}{\rightarrow} [0,\ \log_2(m+1)] \times [0,\ \log_2(m+1)] \\
CF: &\ 256 \text{ (Total number of selections)}
\end{aligned}
$$

Combinatorial Entropy Measurement

Let a vector Z with $(m+1)$ elements, $Z = (Z_0, Z_1, \cdots, Z_j, \cdots, Z_m), 0 \le Z_j \le M$ and $M = \sum_{j=0}^{m} Z_j$. Under this condition, let $P_j = \frac{Z_j}{M}$ be the j-th probability measurement, and a relevant information entropy eZ can be determined and restricted in a $[0,\ \log_2(m+1)]$ region.

$$eZ = -\sum_{j=0}^{m} P_j \log_2(P_j),\ eZ \in [0,\ \log_2(m+1)] \tag{1}$$

$$1 = \sum_{j=0}^{m} P_j,\ 0 \le j \le m \tag{2}$$

For sixteen combinations of the first order, sixteen entropy measurements of eZ correspond to $\{eZ_i\}, 0 \le i \le 15$.

2D Combinatorial Entropies

Extending this construction to higher orders, the second order of combinations are composed of 2D 16×16 pairs of states or a 2D square with 256 positions.

Under this condition for a segment with m elements on a genome Z with $N = m \times M$ elements, sixteen entropies $\{eZ_i\}, 0 \le i \le 15, ZE_i \in [0,\ \log_2(m+1)]$ are determined.

$$(eZ_{i,j}) = \begin{pmatrix} eZ_{0,0} & \cdots & eZ_{i,0} & \cdots & eZ_{15,0} \\ \cdots & & \cdots & & \cdots \\ eZ_{0,j} & \cdots & eZ_{i,j} & \cdots & eZ_{15,j} \\ \cdots & & \cdots & & \cdots \\ eZ_{0,15} & \cdots & eZ_{i,15} & \cdots & eZ_{15,15} \end{pmatrix} \quad i,j \in SI$$

A pair of indexes corresponds to : $eZ_{i,j} = (eZ_i, eZ_j), 0 \leq i,j \leq 15$. There are a total of 256 pairs of 2D positions determined by the genome Z in the square on the $[0,\ \log_2(m+1)] \times [0,\ \log_2(m+1)]$ region.

Multiple Genomes

For multiple genomes $\{Z^t\}, 1 \leq t \leq T$ on maximal T members of each (i,j) projection, a total number of T positions can be collected on 2D square of $\forall(eZ_i^t, eZ_j^t), 1 \leq t \leq T$. This provides a special distribution for whole genomes of T members on (i,j) projection based on combinatorial entropy measurements.

$$\begin{aligned} EZ_{i,j} &= \sum_{t=1}^{T} eZ_{i,j}^t \\ &= \sum_{t=1}^{T} (eZ_i^t, eZ_j^t), 0 \leq i,j \leq 15 \end{aligned}$$

Each $EZ_{i,j}$ represents an index map corresponding to a $[0,\ \log_2(m+1)] \times [0,\ \log_2(m+1)]$ region.

Genomic Index Maps

Different from a genome, it has a relative position in a phylogenic tree on the maximal likelihood relationship. A genomic index is an absolute invariant to correspond a genome into a quantitative measurement under information entropy based on variant construction. Visual representations of multiple projections are illustrated.

Diversity Measures between BLAST Phylogenetics and Genomic Index Maps

Differences between phylogenetic trees on given levels and whole genomes on genomic index maps can be systematically measured by diversity measure for N genomes. In [16], a list of phylogenetic diversity measures are discussed on phylogenetic trees. Using information theory, this type of diversity measures is restricted in $[0, \log_2 N]$.

For a selected segment R such as an ORF area, if all R segments of N genomes are transferred as a genomic index map with at most M distinguished positions $1 \leq M \leq N$ (or a certain level of a relevant phylogenetic tree with M branches), let $E_R(N)$ be a diversity entropy of genomic index maps on R areas, then the diversity measure is defined as

$$E_R(N) = \log_2(M) \tag{3}$$

Difference and Error Margin

For two genomes Z_1, Z_2, let $s, s \in SS$ be a projection on s direction, and a difference $\Delta(eZ_s(Z_1), eZ_s(Z_2))$ of two genomic indexes is

$$\Delta(eZ_s(Z_1), eZ_s(Z_2)) = \max(eZ_s(Z_1), eZ_s(Z_2)) - \min(eZ_s(Z_1), eZ_s(Z_2)) \tag{4}$$

Let Δe be a given error margin, e.g. $\Delta e = 0.001$. If $\Delta(eZ_s(Z_1), eZ_s(Z_2)) > \Delta e$ is true, then two genomes can be distiguished in a genomic index map. Otherwise, two genomes cannot be separated in a cluster on the genomic index map.

If all R segments of N genomes contain in the same content as the same genomic index, then there is $E_R(N) = \log_2(1) = 0$ to be the minimalist diversity measure for the system configuration. However, if all N genomic indexes of R segments can be distinguished without any equal genomic index, then there is $E_R(N) = \log_2(N)$ to be the maximalist diversity measure for the R segments.

Equivalent Condition between Phylogenetic Trees and Genomic Index Maps

From diversity measures for N genomes, there is a natural correspondence via equivalent diversity measures between clusters of a certain level of a phylogenetic subtree and an enlarged region of a genomic index map in general.

Datasets

From a collection of more than 30K genomes from the GISAID genetic database before July 2020, more than 5K genomes were selected without any uncertain element of 'N' in whole sequences. Approximately 25K genomes contain at least one 'N'. There are 72 countries involved that contain more than one genome.

Based on COG-UK's report of SARS-CoV-2 spike mutations [34], three groups of datasets on UK new variations $\{$B.1.1.7, B.1.177, B.1.258$\}$ are selected, and each group contains 10 genomes.

Selecting a total of 5336 genomes, both RNA-dependent RNA polymerase RDRP and whole genomes are involved for further analysis in corresponding genomic index maps.

RDRP and S Protein Alignments, and Phylogenetic Trees

Three selected datasets of 30 variation genomes and their processed results from BLAST are shown in Fig. 3(a)-(c). Two phylogenetic trees for three groups are constructed by BLAST using neighbor-joining (NJ) for DNA on maximal likelihood (ML) shown in Fig. 3(a). Three groups are separated as three branches. Variations on RDRP are listed in Fig. 3(b) and variations on S protein are listed in Fig. 3(c).

Variations of BLAST on RDRP

From Fig. 3(b), all variations of B.1.1.7 RDRP sequences have the same contents with $\{C \leftrightarrow T\}$ exchanged positions, two positions with $\{A \rightarrow C, C \rightarrow T\}$ are in B.1.177 sequences, and there are multiple variations on $\{C \rightarrow A, G \rightarrow A, C \leftrightarrow T, \cdots\}$ in B.1.258 respectively.

Variations of BLAST on S Protein

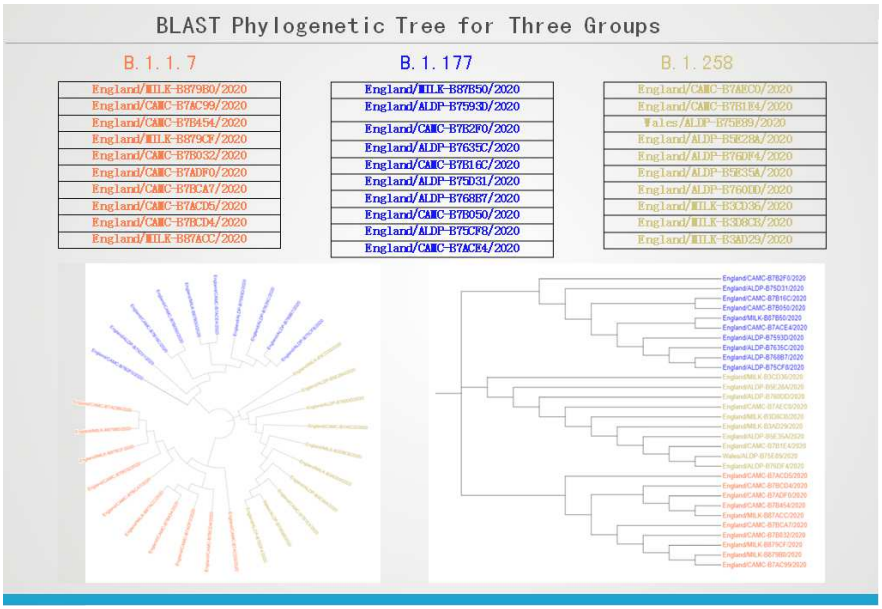
From Fig. 3(c), B.1.1.7 S protein sequences have at least five significant variations on $\{A \rightarrow T, C \rightarrow A, C \rightarrow T, G \rightarrow C\}$, B.1.177 sequences have multiple variations on $\{C \rightarrow T, G \rightarrow T, G \rightarrow C\}$ and B.1.258 sequences contain multiple variations on $\{C \rightarrow A, C \leftrightarrow T, G \rightarrow A\}$ respectively.

Visual Tool - Plotly

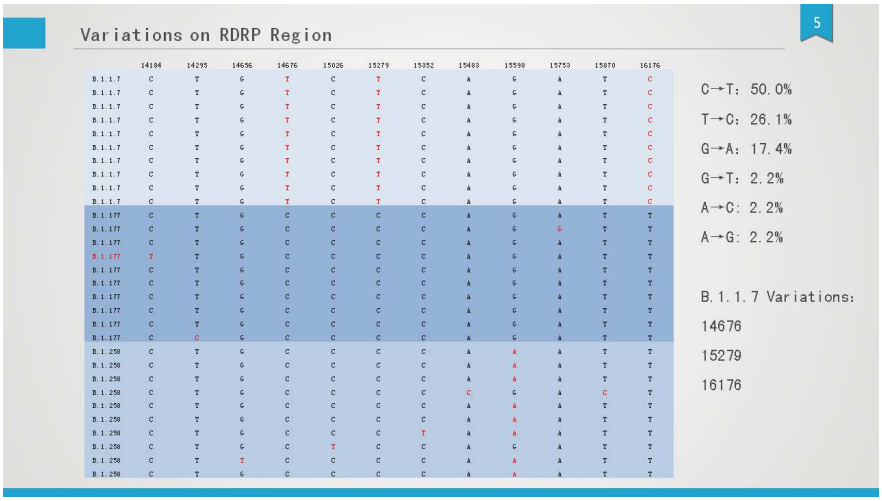
Plotly is a visual tool [66] of open-source visualization libraries for R, Python and JavaScript. In this project, we use this visual tool to illustrate hierarchical distributions for multiple genomes on selected regions of $EZ_{i,j}$ maps.

Clustering on Genomic Index Maps

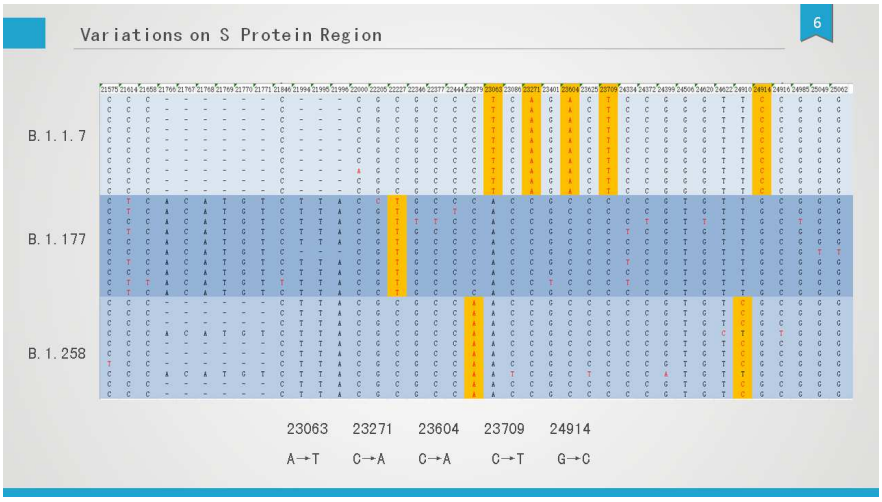
Since all genomic indexes are associated with absolute invariants, this makes it possible to apply 1D or 2D distributions to represent complicated clusters for multiple genomes in hierarchical structures.



(a)



(b)



(c)

Fig. 3 Three groups of variation genomes on BLAST (a) Selected genomes and two Phylogenetic trees (b) Variations in RDRP (c) Variations in S Protein

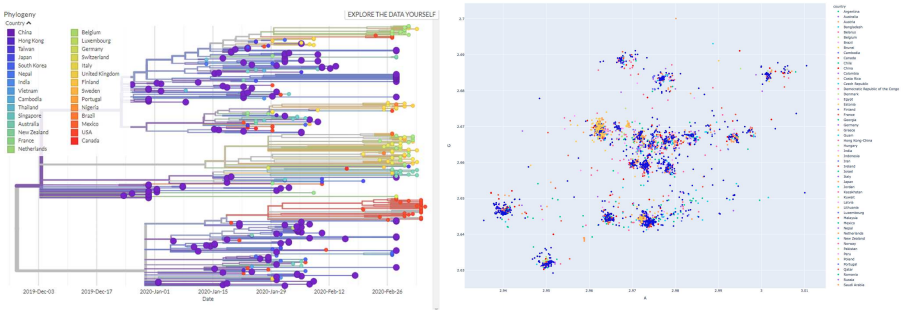


Fig. 4 The phylogenetic tree of real cases over global on Nextstrain and Global Genomic Index Map

Two distinct schemes are shown in Fig 4 for both the phylogenetic tree of Nextstrain and a global genomic index map on five thousand genomes in 72 countries. Different colors are applied to distinguish relevant countries. Various clusters of genomes are clearly visualized by distinct color points for relevant countries on the genomic index map. Refined maps are shown in the next section.

Results

Relevant results are included in two separated files: 5306-RDRP16-A-G.html (for 30 + 5306 RDRP segments) and 5306-Whole16-A-G.html (30 + 5306 Whole genomes) that can be visualized by an HTML browser in the newest version for Plotly libraries.

For RDRP sequences of five thousand genomes, a global genomic index map and various projection maps for three variations and six regions: {Australia, Chile, China, Taiwan, UK, USA} were selected to show relevant projections of results in Fig 5(a)-(c), and enlarged parts of selected regions are shown in Fig 6(a)-(h). Two special projections are shown in Fig 7(a)-(b) to illustrate six selected regions and three variation on RDRP.

For whole sequences of five thousand genomes, a global genomic index map and various projection maps for three variations and six regions: {Australia, Chile, China, Taiwan, UK, USA} were selected to show relevant projections of results in Fig 8(a)-(c), and enlarged parts of selected regions are shown in Fig 9(a)-(h). Two special projections are shown in Fig 10(a)-(b) to illustrate six selected regions and three variation on whole genomes.

Projections of three variations from RDRP to whole genomes are illustrated in Fig 11(a)-(b), 100 times of enlarged projections of ten B.1.1.7 genomes from RDRP to whole genomes are illustrated in Fig 12(a)-(b), and projections of six regions and three variations from RDRP to whole genomes are illustrated in Fig 13(a)-(b)

Discussion

Since there is an autoscale function in the Plotly package, visual regions for selected datasets may not be a fixed one with slight differences for each selected region.

Projections for RDRP

In Fig. 5(a)-(c), three genomic index maps are represented for all genomes of three variations, 72 countries and selected six regions: Australia, Chile, China, Taiwan, the UK and the USA.

Initial Maps for RDRP

In Fig. 5(a), all genomic indexes of three variations and 72 countries are restricted to a region of $A \in [2.70, 2.95] \times G \in [2.40, 2.75], (\Delta A = 0.25, \Delta G = 0.35)$ with an error margin $\Delta e = 0.01$ and multiple clusters could be identified by visual clustering technologies. The central point of this map is located on $(A = 2.825, G = 2.675)$.

In Fig. 5(b), all genomic indexes of three variations and six regions: Australia, Chile, China, Taiwan, the UK and the USA are selected in a region of $[2.70, 2.95] \times [2.43, 2.72], (\Delta A = 0.25, \Delta G = 0.29)$ and a selected region of $[2.75, 2.90] \times [2.50, 2.65]$, $(\Delta A = 0.25, \Delta G = 0.15)$ with an error margin $\Delta e = 0.01$.

In Fig. 5(c), the selected region of Fig. 5(b) has expanded as a full frame restricted in the region of $[2.75, 2.90] \times [2.50, 2.65], (\Delta A = 0.25, \Delta G = 0.15)$. Three variations and six selected regions will be projected as further selection in an enlarged map.

Selected Areas for RDRP

Eight maps are shown in Fig. 6(a)-(h) in the region of $[2.75, 2.90] \times [2.50, 2.65], (\Delta A = 0.25, \Delta G = 0.15)$.

In Fig. 6(a), RDRP of three variations and six selected regions are illustrated.

In Fig. 6(b), RDRP of three variations are illustrated around the center part of the frame. There are three color points distinguished one in red (B.1.177), three (two connected) in green (B.1.258), and one in green is a common point to be shared with red (B.1.177) and blue (B.1.1.7) overlapped at the same position: $(A = 2.824042, G = 2.569091)$. Three color points in a triangle can be restricted in a rectangle on $[2.8236, 2.8277] \times [2.5687, 2.5756]$ of $(\Delta A = 0.0041, \Delta G = 0.0069)$ differences with an error margin $\Delta e = 0.0001$.

In Fig. 6(c), RDRP collected from Australia and three variations are illustrated. One pink point is covered on the same position of B.1.1.7, other pink points are located on the east and north-east direction far away from the three variations.

In Fig. 6(d), RDRP collected from Chile and three variations are illustrated. One purple point is covered on the same position of B.1.1.7, other purple points are located on from the north, north-east, east to south-east directions far away from the three variations.

In Fig. 6(e), RDRP collected from China and three variations are illustrated. One yellow point is covered on the same position of B.1.1.7, other yellow points are located on from north, north-east, east, south-east to south directions far away from the three variations.

In Fig. 6(f), RDRP collected from Taiwan and three variations are illustrated. One green point is covered on the same position of B.1.1.7, other green points are located on from the north, north-east, east, south-east to south directions far away from the three variations.

In Fig. 6(g), RDRP collected from the UK and three variations are illustrated. One light-blue point is covered on the same position of B.1.1.7, other light-blue points are located on from the north to north-east directions far away from the three variations.

In Fig. 6(h), RDRP collected from the USA and three variations are illustrated. One blue point is covered on the same position of B.1.1.7, other blue points are located on from north-west, north, north-east, east, south-east to south directions far away from the three variations.

It is interesting to notice that at least one genome in each region has covered RDRP of B.1.1.7 on A-G projections.

Enlarged Maps for RDRP

In Fig. 7(a)-(b), two enlarged genomic index maps are represented for six regions and three variations on RDRP. Fig. 7(a) is an enlarged map of Fig. 6(a) to show refined genomic indexes for the six regions selected. Fig. 7(b) is an enlarged map of Fig. 6(b) to show refined genomic indexes for three variations. There is a clear triangle shape in the middle area of the map.

Projections for Whole Genomes

In Fig. 8(a)-(h), eight genomic index maps are represented for all whole genomes of three variations, 72 countries and selected six regions: Australia, Chile, China, Taiwan, the UK and the USA.

Initial Maps for Whole Genomes

In Fig. 8(a), all genomic indexes of three variations and 72 countries are restricted to a region of $A \in [2.93, 3.02] \times G \in [2.62, 2.70], (\Delta A = 0.09, \Delta G = 0.08)$ and multiple

clusters could be identified by visual clustering technologies. The center of this map is located on $(A = 2.97, G = 2.66)$.

In Fig. 8(b), all genomic indexes of three variations and six regions: Australia, Chile, China, Taiwan, the UK and the USA are selected in a region of $[2.93, 3.02] \times [2.62, 2.70], (\Delta A = 0.09, \Delta G = 0.08)$ and a selected region of $[2.97, 3.01] \times [2.64, 2.67]$, $(\Delta A = 0.04, \Delta G = 0.03)$.

In Fig. 8(c), the selected region of Fig. 8(b) has expanded as a full frame restricted in the region of $[2.97, 3.01] \times [2.64, 2.67], (\Delta A = 0.04, \Delta G = 0.03)$. Three variations and six selected regions will be projected as further selection in an enlarged map.

Selected Areas for Whole Genomes

Eight maps are shown in Fig. 6(a)-(h) in the region of $[2.97, 3.01] \times [2.64, 2.67], (\Delta A = 0.04, \Delta G = 0.03)$.

In Fig. 9(a), whole genomes of three variations and six selected regions are illustrated. At least, three clusters in green (B.1.258) and blue (B.1.1.7) are located on the right-top and middle-bottom positions as two edge parts. A larger cluster with multiple color points is located on the left-bottom of the map.

In Fig. 9(b), whole genomes of three variations are illustrated on the bottom and north-east parts of the map. There are three types of color points distinguished. One cluster in red (B.1.177) is located on left-bottom corner and a single one is located on the center far away from 1/2 map in a region of $[2.9723, 2.9865] \times [2.6442, 2.6563]$, $(\Delta A = 0.0142, \Delta G = 0.0121)$. Eight clusters in green (B.1.258) are located on left-bottom and right-top of the frame in diagonal directions in a region of $[2.9716, 3.0024] \times [2.6422, 2.6672]$, $(\Delta A = 0.0308, \Delta G = 0.0250)$. And one cluster in blue(B.1.177) is located on middle-bottom part of the map in a region of $[2.9861, 2.9869] \times [2.6440, 2.6451]$, $(\Delta A = 0.0008, \Delta G = 0.0006)$ with an error margin $\Delta e = 0.0001$.Three variations are restricted in three areas with $18 \sim 40$ times respectively. For the most points compared with RDRP maps, each genomic index can be identified with less overlaps.

In Fig. 9(c), whole genomes collected from Australia and three variations are illustrated. One pink point is located between two clusters of B.1.177, other pink points are located on the right-top direction far away from the three variations.

In Fig. 9(d), whole genomes collected from Chile and three variations are illustrated. Multiple purple points are located between clusters of B.1.177 and B.1.1.7, other purple points are located on from the north-west, north, north-east to south-west directions far away from the three variations.

In Fig. 9(e), whole genomes collected from China and three variations are illustrated. Multiple yellow points are located between clusters of B.1.177 and B.1.1.7, other yellow points are located on from the north-west, north, north-east to south-west directions far away from the three variations.

In Fig. 9(f), whole genomes collected from Taiwan and three variations are illustrated. Multiple green points are closed to clusters of B.1.258, B.1.177 and B.1.1.7,

other green points are located on from the north-west, north, north-east to south-west directions far away from the three variations.

In Fig. 9(g), whole genomes collected from the UK and three variations are illustrated. Multiple light-blue points are closed to clusters of B.1.177 and B.1.258, other light-blue points are separated from the north-west, north to south-west directions far away from the three variations.

In Fig. 9(h), whole genomes collected from the USA and three variations are illustrated. Main cluster of blue points are located on B.1.177 and a few points are closed to B.1.1.7, other blue points are located mainly from the north-west, north, north-east, south-east, south to south-west directions far away from the three variations.

Enlarged Maps for Whole Genomes

In Fig. 10(a)-(b), two enlarged genomic index maps are represented for six regions and three variations on whole genomes. Fig. 10(a) is an enlarged map of Fig. 9(a) to show refined genomic indexes for the six regions selected. Fig. 10(b) is an enlarged map of Fig. 9(b) to show refined genomic indexes for three variations. It is interesting to see B.1.1.7 located as an edge cluster in the middle bottom of the map.

Projections from RDRP to Whole Genomes

Pair of genomic index maps for three variations from RDRP to whole genomes are compared in Fig. 11(a)-(b). In Fig. 11(a), clusters of 30 variations are expanded from a triangle shape in a smaller area of RDRP to at least seven clusters of whole genomes from east, south-east and south directions as brushes. For one point of B.1.1.7 on RDRP, a unique blue cluster of whole genomes was developed on the edge part of south-east direction. Ten genomes of B.1.1.7 could be separated under enlarged genomic index maps.

Ten Genomes of B.1.1.7

Pair of genomic index maps for ten B.1.1.7 genomes from RDRP to whole genomes are compared in Fig. 12(a)-(b). In Fig. 12(a), a single blue cluster of 10 genomes on $(A = 2.824042, G = 2.569091)$ position is still as a point, even both vertical and horizontal axises have been magnified more than 100 times. Nine clusters of B.1.1.7 can be clearly observed under enlarged genomic index map in Fig. 12(b). Only the second cluster on the west direction is composed of two separable points on $(A = 2.986233, G = 2.644524)$ and $(A = 2.986233, G = 2.644521)$ with $(\Delta A = 0, \Delta G = 0.000003)$ differences with an error margin $\Delta e = 0.000001$ to show at least

one G variation between the two whole genomes. Further 1000 times of enlarged operation can effectively separate two points on the vertical direction.

Verification on Two SNVs of Two Whole Genomes for B.1.1.7 Lineage

Based on the 5336-Whole16-A-G.html package, it is convenient to identify the two genomes {England/CAMC-B7B454/2020, England/MILK-B87ACC/2020} from enlarged genomic index maps on two visual screens. The two genomes are aligned by BLAST to extract the finest variations as follows.

Sequence	Lineage	18252	25437	T	G	Sample date
England/CAMC-B7B454/2020	B.1.1.7	T	T	2.981561197	2.644520837	2020-11-12
England/MILK-B87ACC/2020	B.1.1.7	C	G	2.981326383	2.644524394	2020-11-13
$\Delta T =$				0.000234814		
$\Delta G =$					0.000003557	

For the two whole genomes, only two SNV sites can be identified different at 18252 and 25437 positions. Two SNVs of the England/CAMC-B7B454/2020 genome contain two 'T' symbols, but two SNVs of the England/MILK-B87ACC/2020 genome change 'T' to 'C' and 'G' symbols respectively. The pair of differences on bi-pairs of genomic indexes is $(\Delta T = 0.000234814, \Delta G = 0.000003557)$.

From this pair of differences, the statement in previous section has been verified. There is merely one 'G' partial variation to be a SNV at the 25437 site from $T \to G$ shown in Fig. 12(b).

Since a SARS-CoV-2 genome has 30K nucleotides, a unit of probability measure on its nucleotides is $\sim O(10^{-4})$. In the above 'G' SNV projection, the unit of genomic indexes is $\Delta G = 0.000003557 \sim O(10^{-6})$ significantly enlarged visual maps at least 100 times than original region. Selected an error margin properly, different topologic configurations can be illustrated similar to separate distinct numbers of branches on given levels of a phylogenetic tree.

5633 Genomes

Pair of genomic index maps for 5603 genomes and three variations from RDRP to whole genomes are compared in Fig. 13(a)-(b). In Fig. 13(a), compact clusters of 5603 genomes and three variations on RDRP are developed mainly in the middle areas and two larger clusters in north-east and west directions. Compared with Fig. 11(a), it is feasible to identify the three UK variation locations on the map as a reference.

In Fig. 13(b), expanded clusters of 5603 genomes and three variations on whole genomes are distributed in the middle areas and at least four larger clusters are distributed in north-east and south-west directions. Compared with Fig. 11(b), it is feasible to identify the three UK variation locations on the map as a reference, especially for the B.1.1.7 cluster.

Differences between RDRP and Whole Genomes in Genomic Index Maps

The corresponding relationships of three variations are transformed from RDRP in Fig. 11(a) to whole genomes in Fig. 11(b) to illustrate characteristic distributions with significantly visual diffussions.

For genomic index maps of whole genomes in Fig.11(b), B.1.1.7 retains one cluster, both B.1.177 and B.1.258 separated as three clusters with larger distances more than $0.01 \sim 0.1$ differences among clusters on genomic index maps.

Since the 5306 genomes were collected before July 2020 from GISAID over the world, no B.1.1.7 variations were identified on this dataset.

From listed comparisons on genomic index maps, larger clusters have significant differences in the six selected regions shown in Fig.6(a)-(h) and Fig. 9(a)-(h) . Different from RDRP maps at least one genome has covered B.1.1.7 position. In relation to whole genomes, there are only two regions (Taiwan, USA) contained a few genomes located nearby the B.1.1.7 cluster shown in Fig. 9(f) and (h). Other four regions of selected whole genomes were far away from the B.1.1.7 cluster.

Larger Clusters of Whole Genomes

For all 5336 genomes on RDRP, multiple clusters may have higher compacted degrees. Hundreds of distinguished color points can be identified in Fig. 13(a). This density indicates at least 30 genomes with the same RDRP content may be mapped in one position.

For all 5336 genomes on whole genomes, different types of distributions were shown in Fig. 13(b). More than $20 \sim 50$ larger clusters can be identified with multiple color points connected as distinguished areas, and many separated single points on the map. Thousands of distinguished color points could be visualized in Fig. 13(b) as larger connected areas. From a statistical viewpoint, each cluster could be distributed as Gaussian normal distributions with central symmetry. This type of clusters could collect huge number of genomes especially in central areas to be generated as multiple normal distributions of the statistical probability for larger number of whole genomes.

Optimal Properties of BLAST Results and Genomic Index Maps

Significant differences between RDRP and whole genomes on genomic index maps can be systematically compared by diversity measures for N genomes. Using diversity measures, this type of diversity measures is restricted in $[0, \log_2(N)]$.

Diversity Measures between RDRP Segments and Whole Genomes

Using BLAST operations, multiple RDRP segments are processed to make alignments one by one on selected N genomes.

Let $E_{RDRP}(N)$ be a diversity measure of genomic index maps on RDRP, and $E_{WG}(N)$ be a diversity measure of genomic index map on whole genomes.

If all RDRP segments of N genomes contain in the same content being the same genomic index, then there is $E_{RDRP}(N) = \log_2(1) = 0$ to provide the minimalist diversity measure for the system configuration. However, if all RDRP segments of N genomes can be distinguished without any equal genomic index, then there is $E_{RDRP}(N) = E_{WG}(N) = \log_2(N)$ to provide the maximalist diversity measure for the system configuration.

In Fig.12(a)-(b), $N = 10$ only a single cluster of ten B.1.1.7 RDRP segments can be identified in Fig. 12(a), and so the diversity measure of Fig. 12(a) is $E_{RDRP}(10) = 0$. However, nine clusters of ten B.1.1.7 genomes are separated with an error margin $\Delta e = 0.00001$ on (A,P) genomic index map in Fig. 12(b) and the diversity measure of Fig. 12(b) is $E_{WG}(10) = \log_2(9)$. If further enlargement has performed and an error margin $\Delta e \leq 0.000001$, then ten clusters can be distinguished and $E_{WG}(10) = \log_2(10)$.

Under those conditions, both the minimal and maximal borders of diversity measures can be obtained.

If no BLAST operations were performed to align RDRP segments, then the diversity measure satisfies $E_{RDRP}(N) > 0$. It is extremely hard for anyone to obtain a better result if at least two distinguished genomes are selected from different countries over the world.

In general, N genomes collected from different places, a diversity measure on $0 < E_{RDRP}(N) \leq E_{WG}(N) < \log_2(N)$ will be observed.

Due to this structural restriction, traditional BLAST operations provide a necessary condition for genomic index in system optimization. Under BLAST supports, genomic index maps provide an optimal scheme in genomic analysis to visualize multiple genomes in one genomic index map.

Without BLAST operations, genomic index maps for multiple genomes cannot have the minimal configuration of the diversity measure systematically at all.

Optimal Solution for Multiple ORFs

Twenty nine ORFs are identified from SARS-CoV-2 genomes, in a natural condition, each ORF may bring some random variations. The local alignment can be effectively performed on one selected ORF. It is difficult to make alignment same time more than one ORFs in general.

It is necessary for multiple ORFs to make multiple alignments of relevant ORF first, aligned ORF segments can be processed in further calculation.

If each aligned ORF segment has transferred into a genomic index, multiple aligned ORF recombination will provide the minimal diversity measures smaller than directly calculated from whole unalignable genomes.

From an optimal viewpoint, neither RDRP nor whole genomes provides an optimal solution to explore complex-inner structures of whole SARS-CoV-2 genomes. A better solution is to apply multiple ORFs of alignments separately to create an optimal solution of the diversity measure for future explorations.

Phylogenetic Trees and Genomic Index Maps

Using diversity measures, it is convenient for both phylogenetic trees and genomic index maps to be compared consistently. This measurable mechanism is confirmed on the equivalent diversity measures between viral genomes under certain levels of a phylogenetic tree and various enlarged regions of genomic index maps.

In principle on any genomic index map, the enlarging operations can be repeatedly applied to selected regions to recursively detailed regions via a series of proper error margins $\Delta e \in \{1, 0.1, 0.01, \cdots, 0.0\cdots01\}$ from a rough gap to the finest margin respectively.

In the most conditions, if two genomic indexes are different, then two positions can be visually separated when a larger fold magnification has been applied and proper error margin selected.

In application levels, the diversity measure provides conveniently classified effects for medical doctors and researchers to treat COVID-19 patients with similar genomic indexes as one group of genomes.

Conclusion

Using combinatorial entropy as 2D genomic index maps, there are 256 projections to support multiple genomes in representations. Various computational measurements are described to cover from local to global statistics properties. Richness of both phylogenetic trees and genomic index maps can be measured on diversity measures with equivalent effects.

Applying thirty genomes of UK new variations, and five thousand genomes of SARS-CoV-2 on 72 countries and special selections on six countries based on Plotly libraries, a list of genomic index maps selected for both RDRP segments and whole genomes are shown in significant different distributions on each country to illustrate complicated contagiousness patterns among various regions.

From the ten genomes of B.1.1.7 lineage, it is feasible to distinguish one SNV from genomic index maps, and different magnifications on selected areas to provide better effects of visualization on selected samples for the finest analysis under the minimum optimal condition. Diversity measures provide numeric quantities as clade

information to be compared with both RDRP and whole genomes on the ten whole genomes of B.1.1.7 samples consistently.

It is a challenging task to generate optimal phylogenetic trees of SARS-CoV-2 genomes accurate with stability to support huge number of update genomes over the global, to make twenty nine ORFs such as {S protein,M,N,E} gradually in optimal conditions to simulate GISAID clades and Nextstrain phylogenetic trees in higher stability under huge updates of mutations and variations of SARS-CoV-2 genomes worldwide.

Using genomic index maps, further refined classifications and categories of genomes could be visually and numerically explored, and this powerful optimal-measure tool would be useful in refined medical treatments for COVID-19 patients worldwide in near future.

Conflict Interest

No conflict of interest has been claimed.

Acknowledgements The authors would like to thank NCBI, GISAID, CNGBdb, and Nextstrain for providing invaluable information on the newest dataset collections of SARS-CoV-2 and other coronavirus genomes to support this project working smoothly. The NSFC (62041213) provided financial support for the project.

References

1. GISAID: Open access to influenza virus data https://gisaid.org
2. Nextstrain Real time tracking of pathogen evolution https://nextstrain.org
3. Basic Local Alignment Search Tool BLAST https://blast.ncbi.nlm.nih.gov/Blast.cgi
4. HP Yao, XY Lu, ⋯, LJ Li, Patient-drived mutations impact pathogenicity of SARS-CoV-2, DOI: https://doi.org/10.1101/2020.04.14.20060160 https://www.medrxiv.org/10.1101/2020.04.14.20060160v2
5. Li C, Yang Y, Ren L. Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. Infect Genet Evol. 2020 Mar 10;82:104285. doi: 10.1016/j.meegid.2020.104285. [Epub ahead of print] PMID: 32169673
6. Thorne, J. L., Kishino, H. & Felsenstein, J. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33, 114124 (1991).
7. Lewis, P. O. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. Mol. Biol. Evol. 15, 277283 (1998).
8. Mitchison, G. J. A probabilistic treatment of phylogeny and sequence alignment. J. Mol. Evol. 49, 1122 (1999).
9. Holmes, I. & Bruno, W. J. Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics 17, 803820 (2001).
10. Salter, L. A. & Pearl, D. K. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. Syst. Biol. 50, 717 (2001).
11. Lee, M. S. Y. Unalignable sequences and molecular evolution. Trends Ecol. Evol. 16, 681685 (2001).

12. Posada, D. & Crandall, K. A. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50, 580601 (2001).
13. Holder, M., Lewis, P. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet 4, 275284 (2003). https://doi.org/10.1038/nrg1044
14. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6, 361375 (2005). https://doi.org/10.1038/nrg1603
15. Nye TMW. Trees of Trees: An Approach to Comparing Multiple Alternative Phylogenies. Systematic Biology. 2008. pp. 785794. https://doi.org/10.1080/10635150802424072 PMID: 18853364
16. Chao A, Chiu CH, Jost L. Phylogenetic diversity measures based on Hill numbers. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 2010 Nov;365(1558):3599-3609. DOI: 10.1098/rstb.2010.0272.
17. Batista MVA, Ferreira TAE, Freitas AC, Balbino VQ. An entropy-based approach for the identification of phylogenetically informative genomic regions of Papillomavirus. Infection, Genetics and Evolution.2011; 11(8):2026—2033. https://doi.org/10.1016/j.meegid.2011.09.013 PMID: 21964599
18. Arellano-Valle, R.B.; Contreras-Reyes, J.E.; Genton, M.G. Shannon entropy and mutual information for multivariate skew-elliptical distributions. Scand. J. Stat. 2013, 40, 4262.
19. Zhang, Q. et al. Viral Phylogenomics Using an Alignment-Free Method: A Three-Step Approach to Determine Optimal Length of k-mer. Sci. Rep. 7, 40712; doi: 10.1038/srep40712 (2017)
20. Kapli, P., Yang, Z. & Telford, M.J. Phylogenetic tree building in the genomic age. Nat Rev Genet 21, 428444 (2020). https://doi.org/10.1038/s41576-020-0233-0
21. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2:New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020; 37:15301534. https://doi.org/10.1093/molbev/msaa015 PMID: 32011700
22. DeMaio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. Issues with SARS-CoV-2sequencing data. In: Virological [Internet]. 5 May 2020 [cited 13 May 2020]. Available: http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473
23. Morel B, Barbera P, Czech L, Bettisworth B, H ü bner L, Lutteropp S, Serdari D, Kostaki EG, Mamais I, Kozlov AM, Pavlidis P, Paraskevis D, Stamatakis A. Phylogenetic analysis of SARS-CoV-2 data is difficult. Mol Biol Evol, 15 Dec 2020
24. Shen, XX., Li, Y., Hittinger, C.T. et al. An investigation of irreproducibility in maximum likelihood phylogenetic inference. Nat Commun 11, 6096 (2020). https://doi.org/10.1038/s41467-020-20005-6
25. Li, T., Liu, D., Yang, Y. et al. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. Sci Rep 10, 22366 (2020). https://doi.org/10.1038/s41598-020-79484-8
26. Zhao Z, Sokhansanj BA, Malhotra C, Zheng K, Rosen GL (2020) Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization. PLoS Comput Biol 16(9): e1008269. https://doi.org/10.1371/journal.pcbi.1008269
27. Wang Y, Mao J-M, Wang G-D, Qiu Z, Yao Q, Chen K-P. Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames. https://doi.org/10.1038/s41598-020-69342-y PMID:32704018
28. Xia X. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. Mol Biol Evol. 2020. https://doi.org/10.1093/molbev/msaa094 PMID: 32289821
29. Yazar S. SARS-CoV-2 virus RNA sequence classification and geographical analysis with convolutional neural networks approach arXiv, 09 Jul 2020 PPR: PPR269967
30. Acera Mateos P, Balboa RF, Easteal S, Eyras E, Patel HR. PACIFIC: a lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses. Sci Rep, 11(1):3209, 05 Feb 2021
31. Fountain-Jones NM, Appaw RC, Carver S, Didelot X, Volz EM, Charleston M. Emerging phylogenetic structure of the SARS-CoV-2 pandemic. bioRxiv. 2020. p. 2020.05.19.103846. https://doi.org/10.1101/2020.05.19.103846
32. Banerjee AK, Begum F, Ray U. Mutation Hot Spots in Spike Protein of COVID-19. https://doi.org/10.20944/preprints202004.0281.v1

33. Turakhia Y, De Maio N, Thornlow B, et al. Stability of SARS-CoV-2 phylogenies. Plos Genetics. 2020 Nov;16(11):e1009175. DOI: 10.1371/journal.pgen.1009175.
34. COG-UK, COG-UK Update on SARS-CoV-2 Spike Mutations of Special Interest Report 1, 2020 https://www.attogene.com/wp-content/uploads/2020/12/Report-1_OOG-UK_19-December-2020_SARS-COV-2-Mutations.pdf
35. P. J. Cameron, *Combinatorics: Topics, Techniques, Algorithms*, Cambridge University Press, 1994.
36. J. R. Chen, *Combinatorial Mathematics*, Harbin Institute of Technology Press, 2012 (in Chinese).
37. H. W. Gould, Some Generalizations of Vandermonde's Convolution, *The American Mathematical Monthly*, Vol. 63, No.2 84-91, 1956.
38. H. W. Gould, *Combinatorial identities*, Morganton, 1972.
39. M. Hall, *Combinatorial Theory*, 2nd edition, Blaisdell, 1986.
40. L. K. Hua, *Loo-Keng Hua Selected Papers*, Springer, 1982.
41. L. K. Hua, *Selected Work of Hua Loo-Keng on Popular Sciences*, Shanghai Education Press, 1984 (in Chinese).
42. D. E. Knuth, *The Art of Computer Programming*, Vol. 1, 3rd edition, Addison-Wesley, 1998.
43. D. E. Knuth, *The Art of Computer Programming*, Vol. 4A: Combinatorial Algorithms, Part 1, Addison-Wesley, 2011.
44. F. Morgan, *Geometric Measure Theory*, 4th edition, Elsevier 2009.
45. G. Polya, R. Tarjan and D. Woods, *Notes on Introductory Combinatorics*, Birkhauser, 1983.
46. R. P. Stanley, *Enumerative Combinatorics*, Vol. 1, 2nd edition, Cambridge University Press, 1997.
47. D. Stanton, R. Stanton and D. White, *Constructive Combinatorics*, Springer-Verlag, 1986.
48. G.Z. Tu, *Combinatorial Enumeration Methods & Applications*, Science Press,1981 (in Chinese).
49. A. Tucker, *Applied Combinatorics*, John Wiley & Sons, 2007.
50. J. H. van Lint and R. M. Wilson, *A Course in Combinatorics*, 2nd edition, Cambridge University Press, 2001.
51. L. X. Wang, *An Elementary Treatise on Combinations*, Harbin Institute of Technology Press, 2012 (in Chinese).
52. L. Z. Xu, M. S. Jiang and Z. Q. Zhu, *Combinatorial Mathematics of Computation*, Shanghai Science & Technology Press, 1983 (in Chinese).
53. L D Landau, E M Lifshitz. Statistical Physics, 3rd edition, Part 1, Pergamon Press 1986.
54. HB Callen. Thermodynamics and an Introduction to Thermostatistics, 2nd Edition. John Wiley & Sons 1985.
55. W Greiner, L Neise, H Stocker. Thermodynamics and Statistical Mechanics. Springer-Verlag 1995.
56. T L Hill, Introduction to Statistical Thermodynamics, Addison-Wesley, Reading Mass. 1960.
57. R Kubo. Thermodynamics. North-Holland Pub. Co. 1968.
58. R P Freynman, Statistical Mechanics, Benjamin Reading Mass. 1972.
59. B Widom, in Foundamental Problems in Statistical Mechanics, Vol. III edi. by Cohen, Horth-Holland, 1975 1-45.
60. K G Wilson, Rev., Mod. Phys. 55, 583 1983
61. J Beatte, I Oppenheim. Thermodynamics, Elsevier Scientific 1979.
62. S K Ma, Statistical Mechanics, World Scientific 1985.
63. P W Atkins, The Second Law, Scientific American Books and W H Freeman and Co. 1984.
64. D Chandler. Introduction to Modern Statistical Mechanics, 1st edition. Oxford University Press Inc. 1987.
65. K Huang. Statistical Mechanics, 2nd edition John Wiley & Sons 1987.
66. Plotly: The front-end for ML and data science models. https://plotly.com
67. Z. J. Zheng, A. Maeder, The The conjugate classification of the kernel form of the hexagonal grid, *Modern Geometric Computing for Visualization*, Springer-Verlag, 73-89, 1992.
68. Z. J. Zheng,*Conjugate transformation of regular plan lattices for binary images*, PhD Thesis, Monash University, 1994.

69. Jeffrey Z. J. Zheng, Christian H. H. Zheng, A framework to express variant and invariant functional spaces for binary logic, *Frontiers of Electrical and Electronic Engineering in China*, 5(2):163-172, Higher Educational Press and Springer-Verlag, 2010.
70. Jeffrey Z.J. Zheng, Christian H.H. Zheng and Tosiyasu L. Kunii. A Framework of Variant Logic Construction for Cellular Automata, *Cellular Automata - Innovative Modeling for Science and Engineering*, Dr. Alejandro Salcido (Ed.), InTech Press, 2011.
71. Jeffrey Zheng, Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019 https://www.springer.com/in/book/9789811322815
72. Jeffrey Zheng, Variant Construction Theory and Applications, Vol. 1: Theoretical Foundation and Applications, Science Press 2021 (Chinese, Formal Publishing Soon).
73. Jeffrey Zheng, ResearchGate: http://researchgate.net/pprofile/Jeffrey_Zheng
74. Jeffrey Zheng, Chris Zheng, Biometrics and Knowledge Management Information Systems, Chapter 11: Variant Construction from Theoretical Foundation to Applications, Springer Nature 2019, 193-202 https://link.springer.com/chapter/10.1007/978-981-13-2282-2_11
Selected in Research of COVID-19 for PubMed Central PMC and the World Health Organization WHO by Springer-Nature on Free Access of all scientific researchers worldwide.
75. Jeffrey Zheng, Minghan Zhu, Mu Qiao, Yang Zhou. Visualizations of SARS-CoV-2 Genomes on Genomic Index Maps, EC Neurology, SI-02(2021): 206-221 https://www.ecronicon/specialissue21_neurology.php
76. Mu Qiao, Renyang Liu, Zhenhui Wang, Xinmei Li, Jeffrey Zheng. Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions, EC Neurology, SI-02(2021): 86-93 https://www.ecronicon/specialissue21_neurology.php
77. Minghan Zhu, Jeffrey Zheng. Visual Variations between Pairs of SARS-CoV-2 Genomes on Integrated Density Matrix, EC Neurology,SI-02(2021): 94-100 https://www.ecronicon/specialissue21_neurology.php
78. Yang Zhou, Jeffrey Zheng. Visualizations of Combinatorial Entropy Index on Whole SARS-CoV-2 Genomes, EC Neurology, SI-02(2021):101-109 https://www.ecronicon/specialissue21_neurology.php

Appendix

Three Variation Lineages Selected: {B.1.1.7, B.1.177.B.1.258}

Three lineages of thirty variation genomes are briefly listed by names as follows.

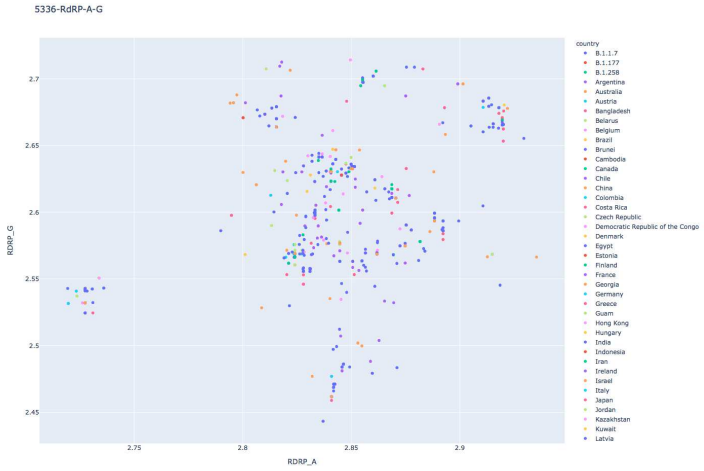
B.1.1.7	B.1.177	B.1.258
England/MILK-B879B0/2020	England/MILK-B87B50/2020	England/CAMC-B7AEC0/2020
England/CAMC-B7AC99/2020	England/ALDP-B7593D/2020	England/CAMC-B7B1E4/2020
England/CAMC-B7B454/2020	England/CAMC-B7B2F0/2020	Wales/ALDP-B75E89/2020
England/MILK-B879CF/2020	England/ALDP-B7635C/2020	England/ALDP-B5E28A/2020
England/CAMC-B7B032/2020	England/CAMC-B7B16C/2020	England/ALDP-B76DF4/2020
England/CAMC-B7ADF0/2020	England/ALDP-B75D31/2020	England/ALDP-B5E35A/2020
England/CAMC-B7BCA7/2020	England/ALDP-B768B7/2020	England/ALDP-B760DD/2020
England/CAMC-B7ACD5/2020	England/CAMC-B7B050/2020	England/MILK-B3CD36/2020
England/CAMC-B7BCD4/2020	England/ALDP-B75CF8/2020	England/MILK-B3D8CB/2020
England/MILK-B87ACC/2020	England/CAMC-B7ACE4/2020	England/MILK-B3AD29/2020

5336 Genomes and Genomic Indexes on (A,G)

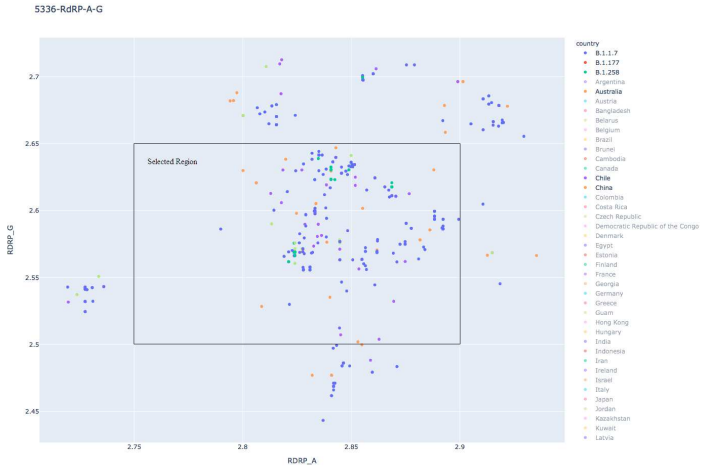
Two information files in two formats contain detailed information for each selected genome: (Genome Name, Location, Time, Type, Clade, ..., Genomic Indexes on whole genome and RDRP)
5306 genome information: {5306-Infor.xlsx, 5306-Infor.tsv}
30 UK genome information: {UK30-Infor.xlsx, UK30-Infor.tsv}

Two Excutable Packages

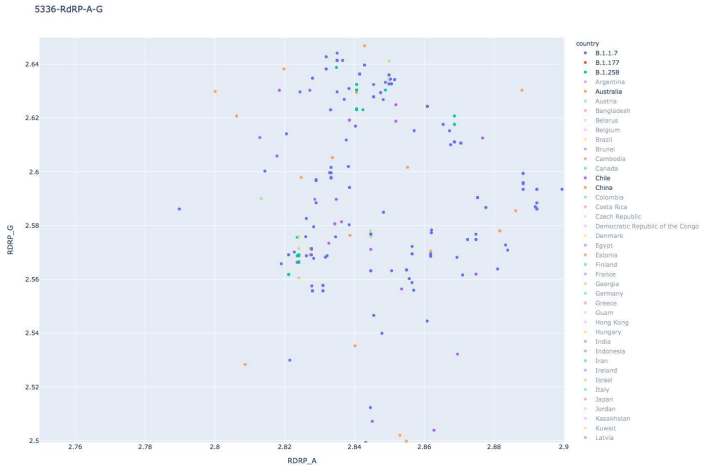
Two interactive visual packages: RDRP and whole genomes
5336-RDRP16-A-G.html (30 + 5306 RDRP segments in 72 countries/regions)
5336-Whole16-A-G.html (30 + 5306 Whole genomes in 72 countries/regions)



(a)

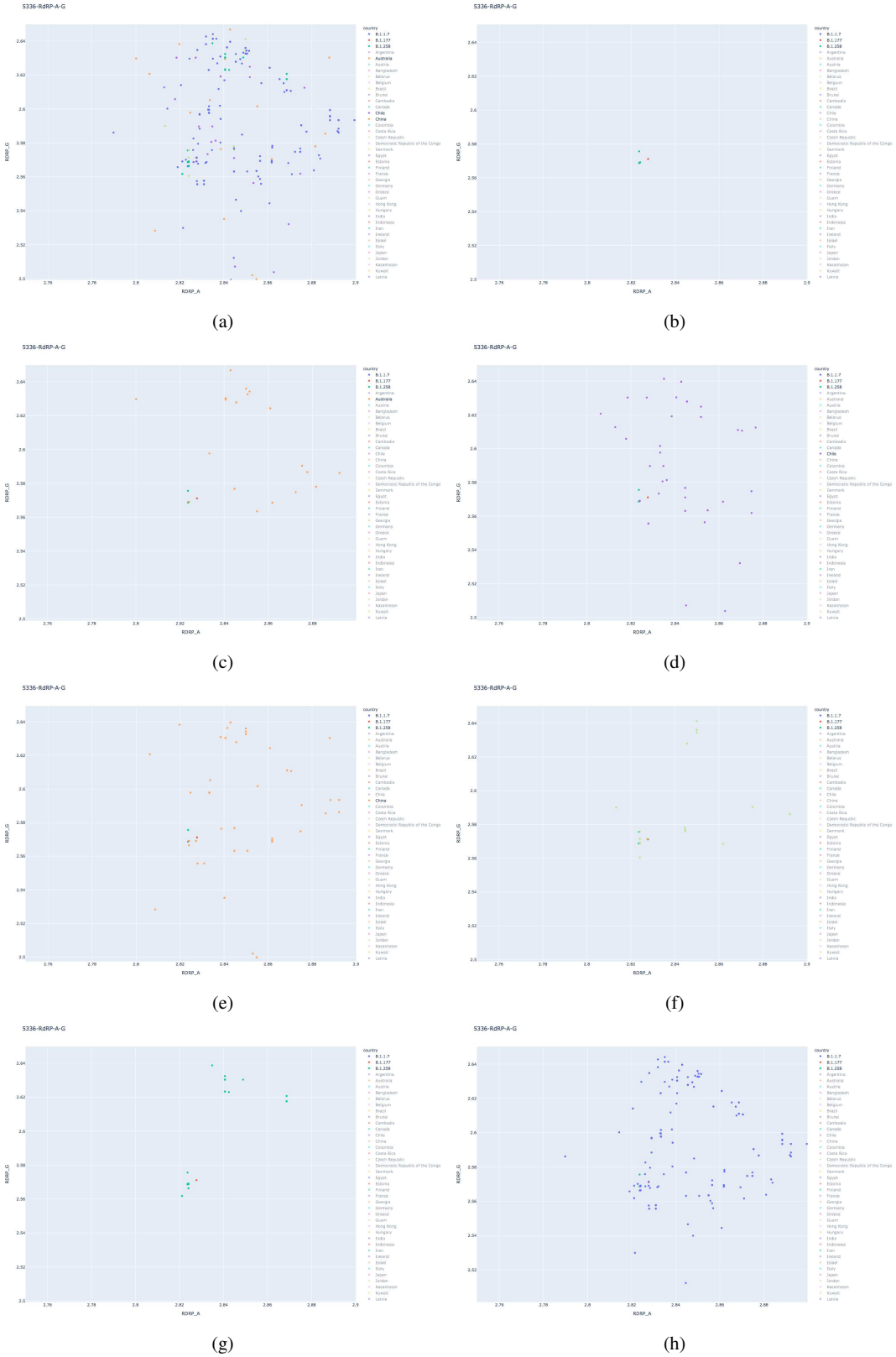


(b)



(c)

Fig. 5 Five thousands of RDRP genomes on genomic index maps (a) Global (b) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (c) An enlarged region selected from (b)



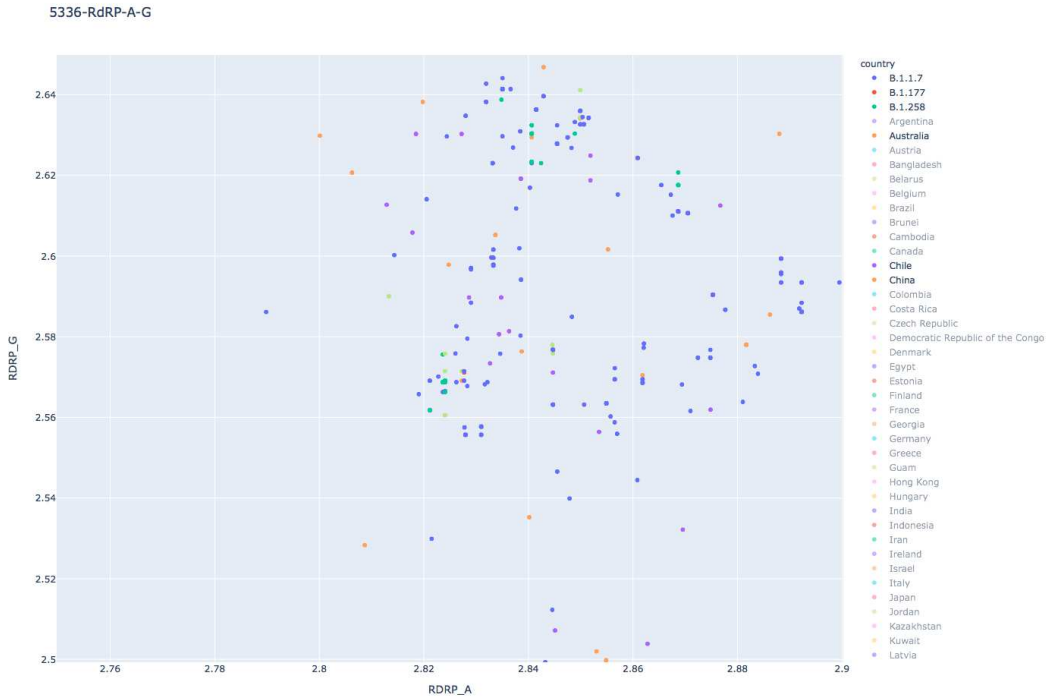
(a) (b)

(c) (d)

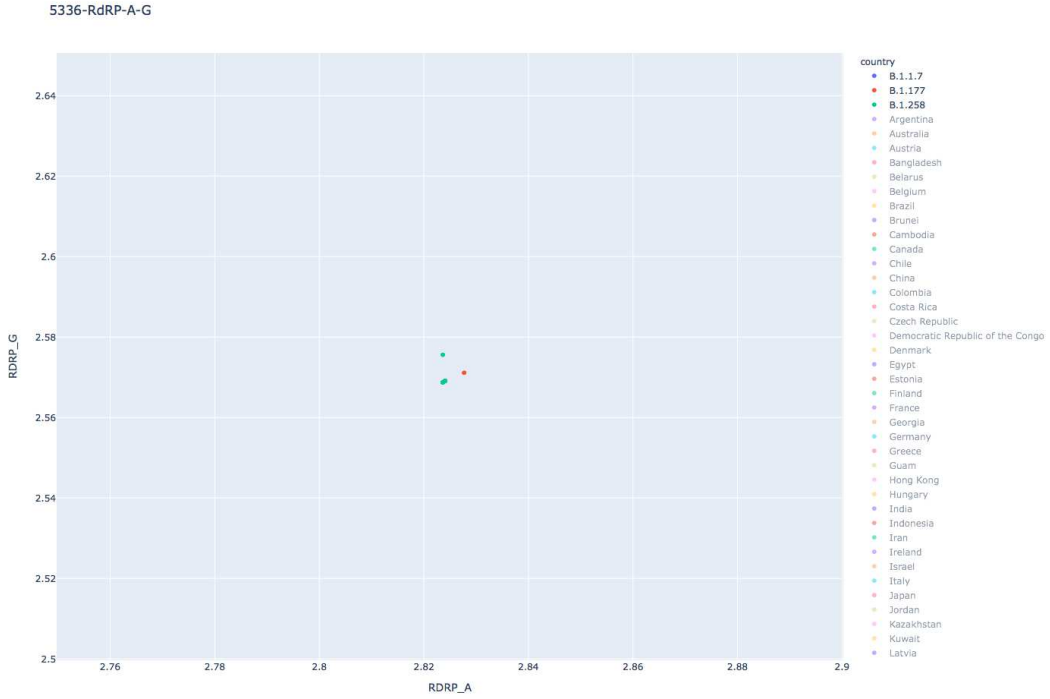
(e) (f)

(g) (h)

Fig. 6 An enlarged region of RDRP on genomic index maps with three groups of variations (a) Six regions: Australia + Chile + China + Taiwan + UK + USA (b) Three groups: B.1.1.7 + B.1.177+B.1.258 (c) Australia (d) Chile (e) China (f) Taiwan (g) UK (h) USA

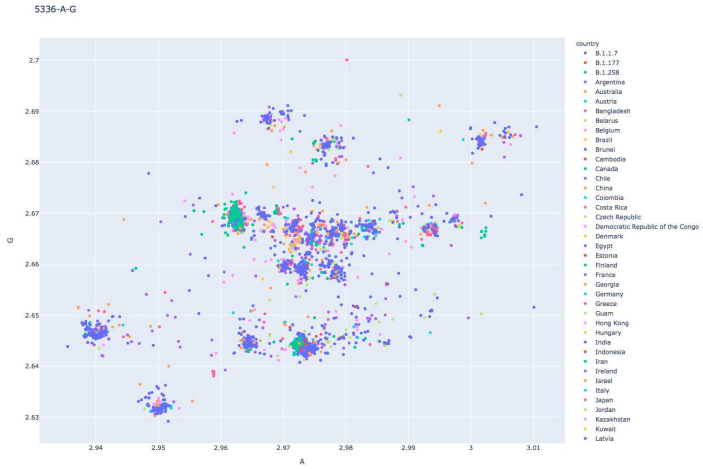


(a)

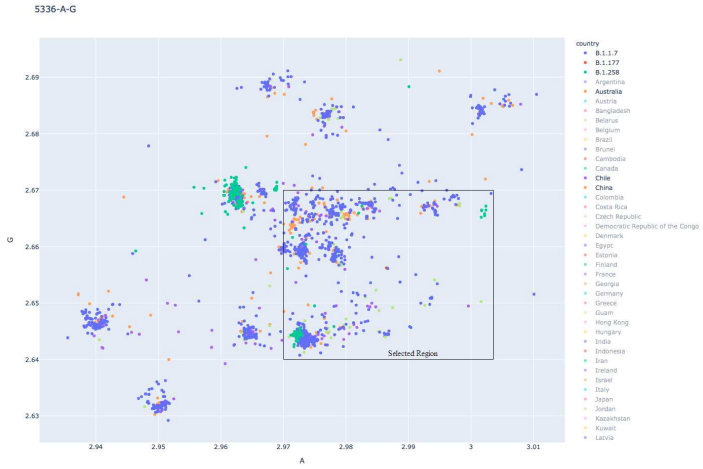


(b)

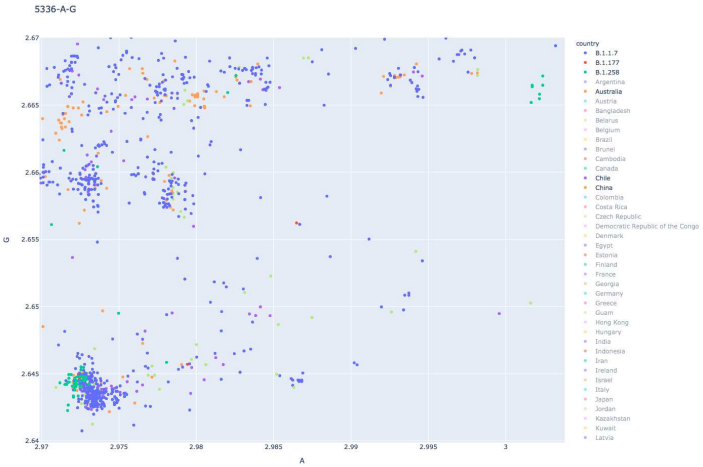
Fig. 7 Enlarged Region of RDRP genomes on genomic index maps (a) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (b) Three variations



(a)



(b)



(c)

Fig. 8 Five thousands of whole genomes on genomic index maps (a) Global (b) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (c) An enlarged region selected from (b)

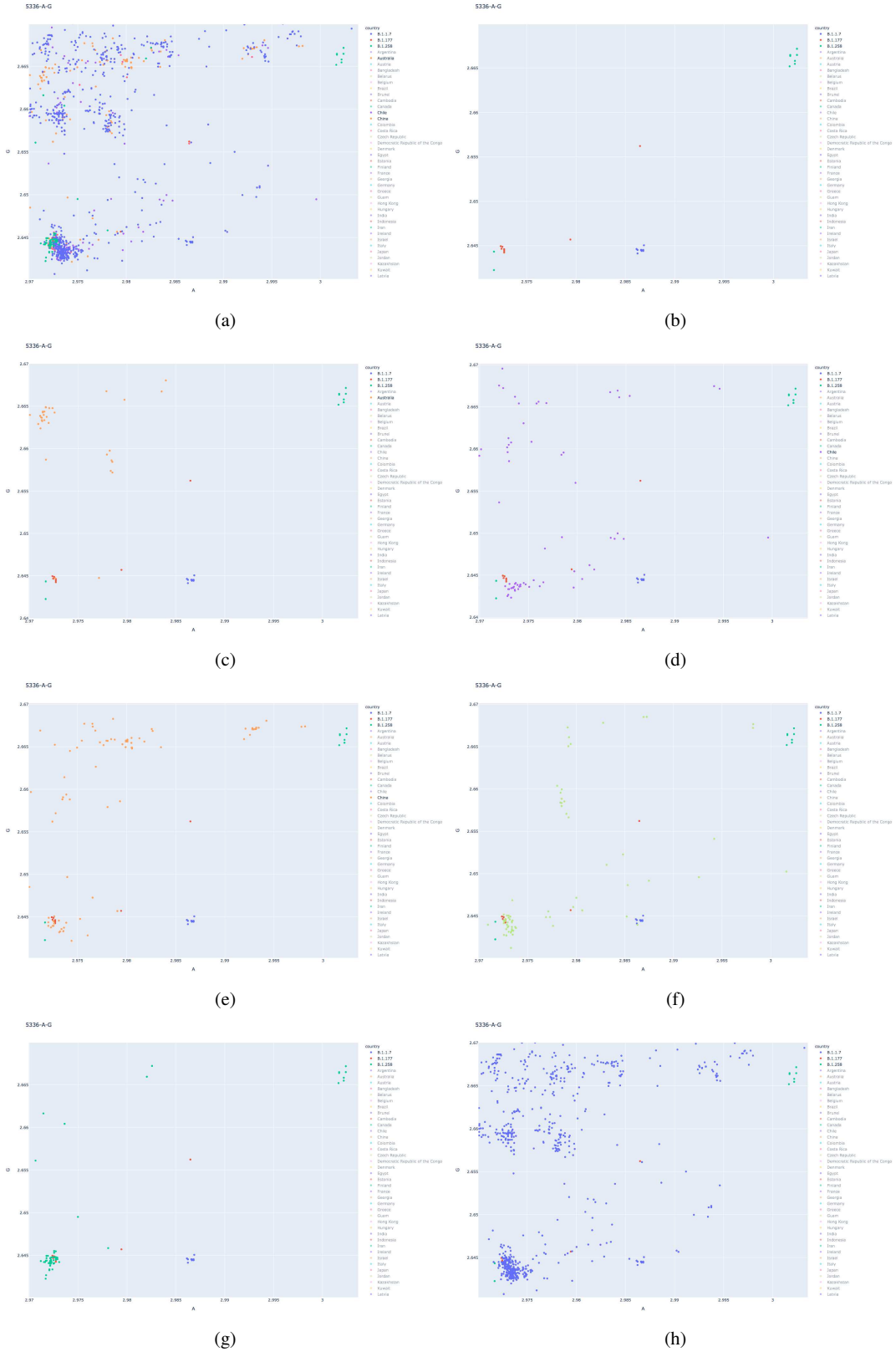
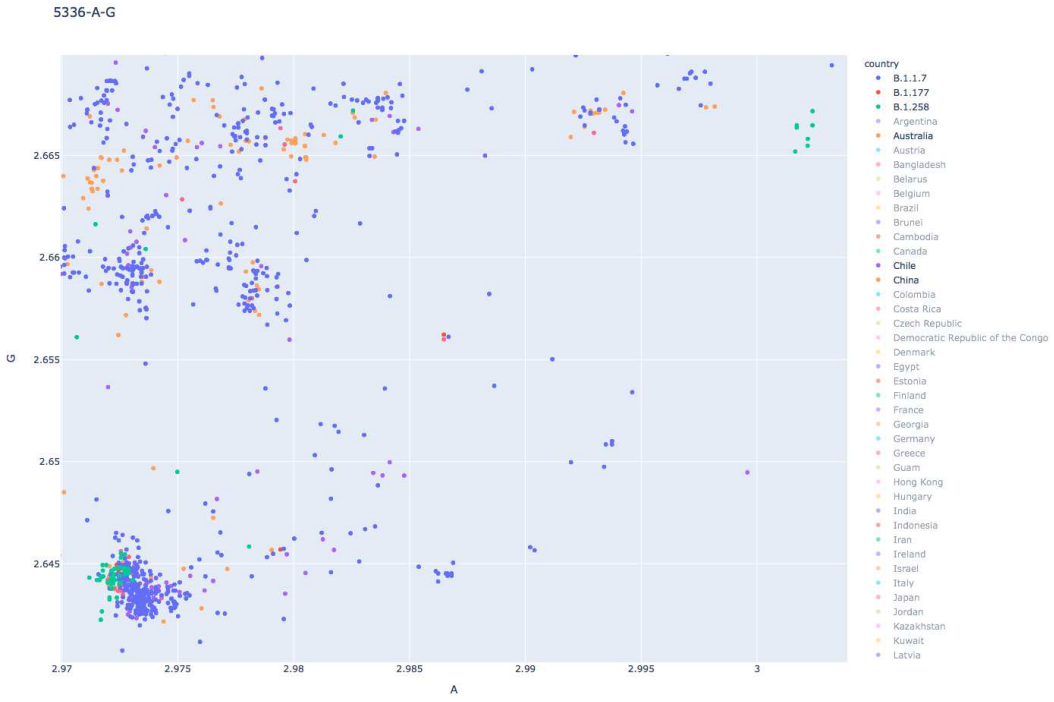
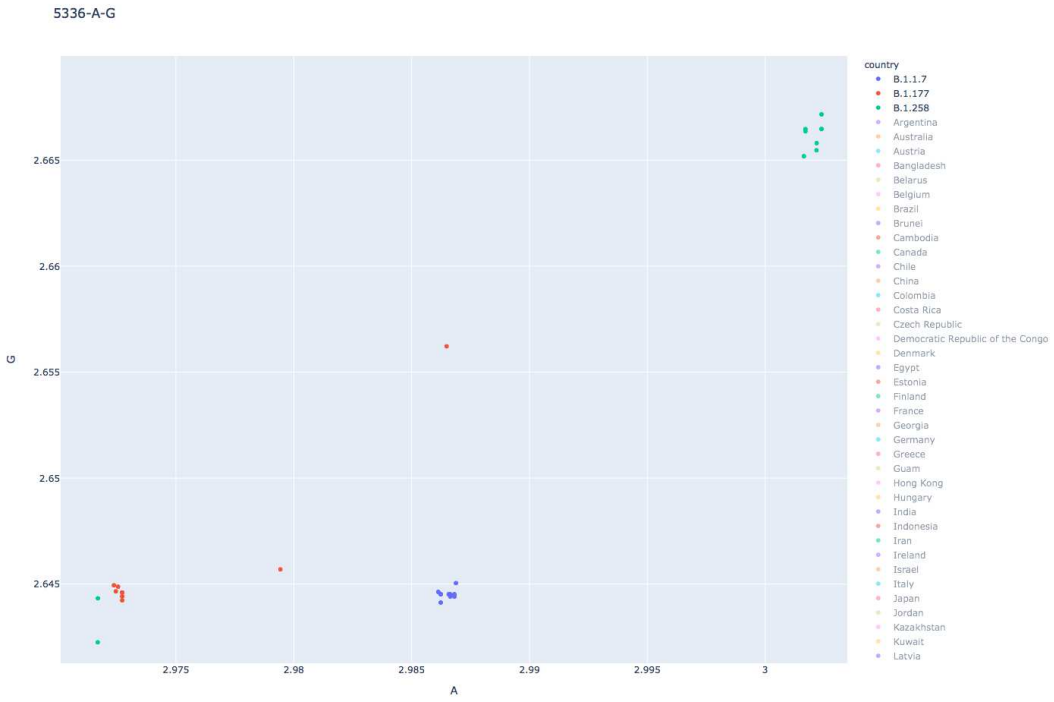


Fig. 9 An enlarged region of whole genomes on genomic index maps with three groups of variations (a) Six regions: Australia + Chile + China + Taiwan + UK + USA (b) Three groups: B.1.1.7 + B.1.177+B.1.258 (c) Australia (d) Chile (e) China (f) Taiwan (g) UK (h) USA

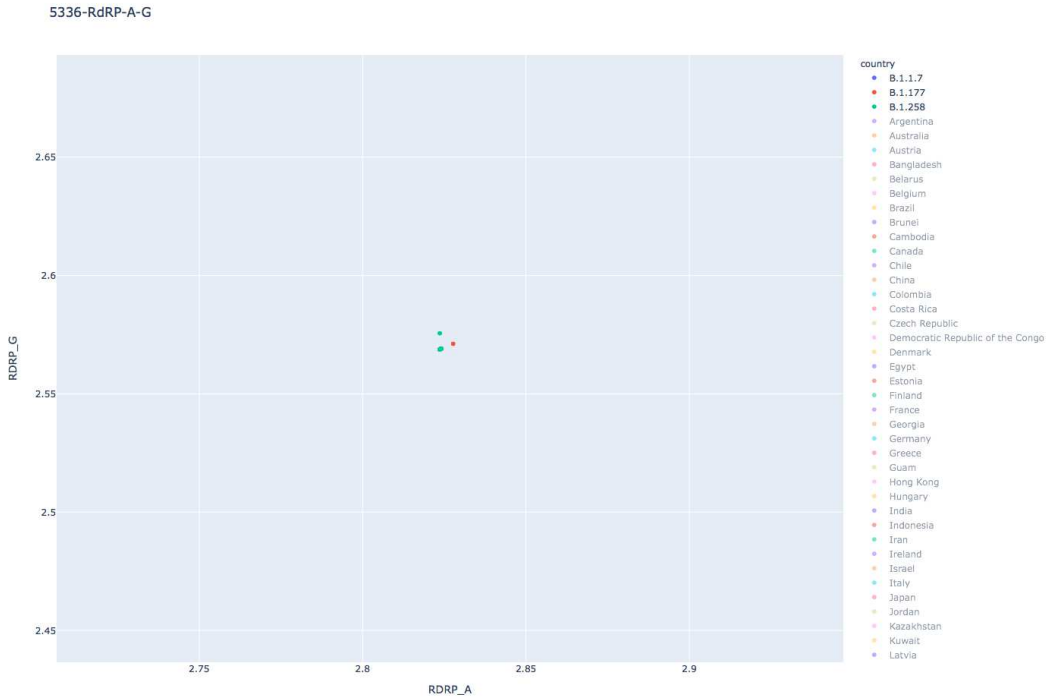


(a)

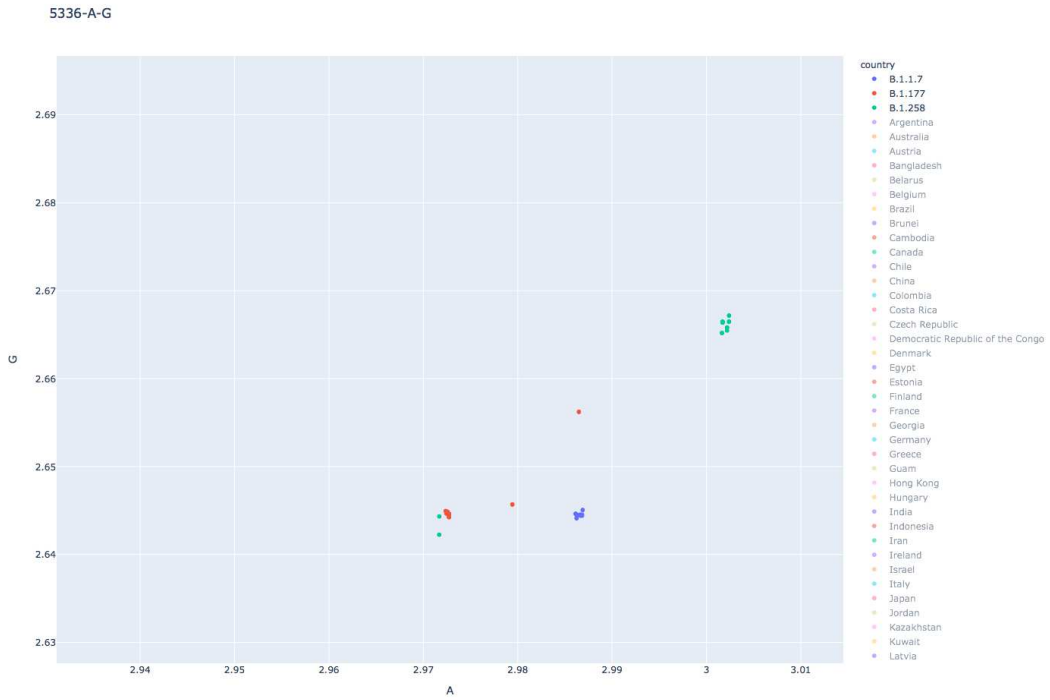


(b)

Fig. 10 Enlarged Region of whole genomes on genomic index maps (a) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (b) Three variations

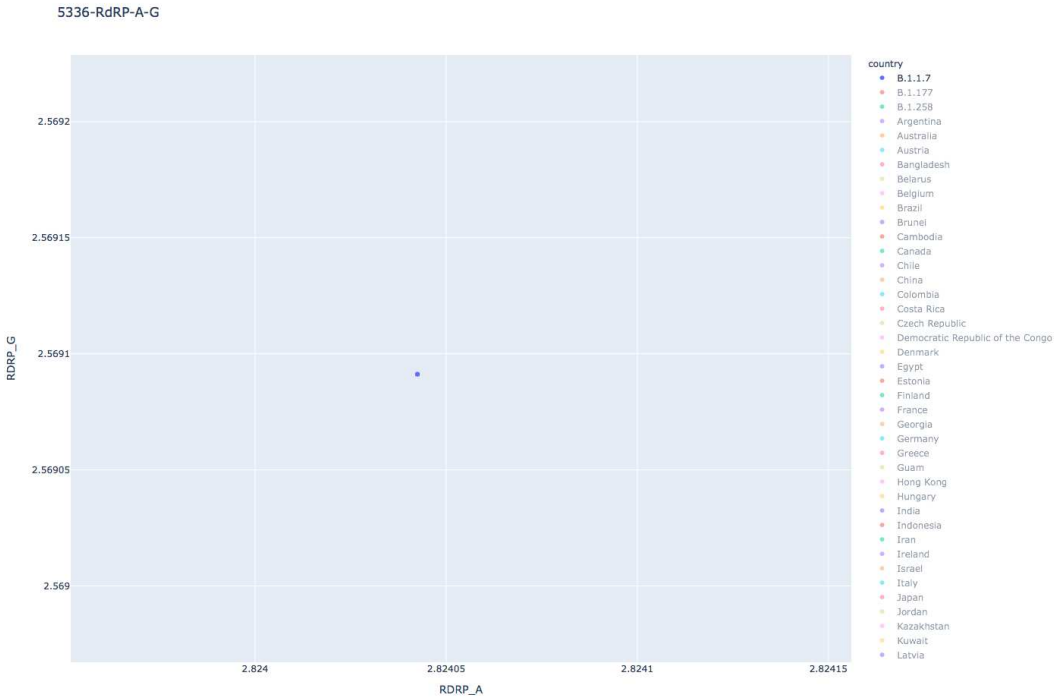


(a)

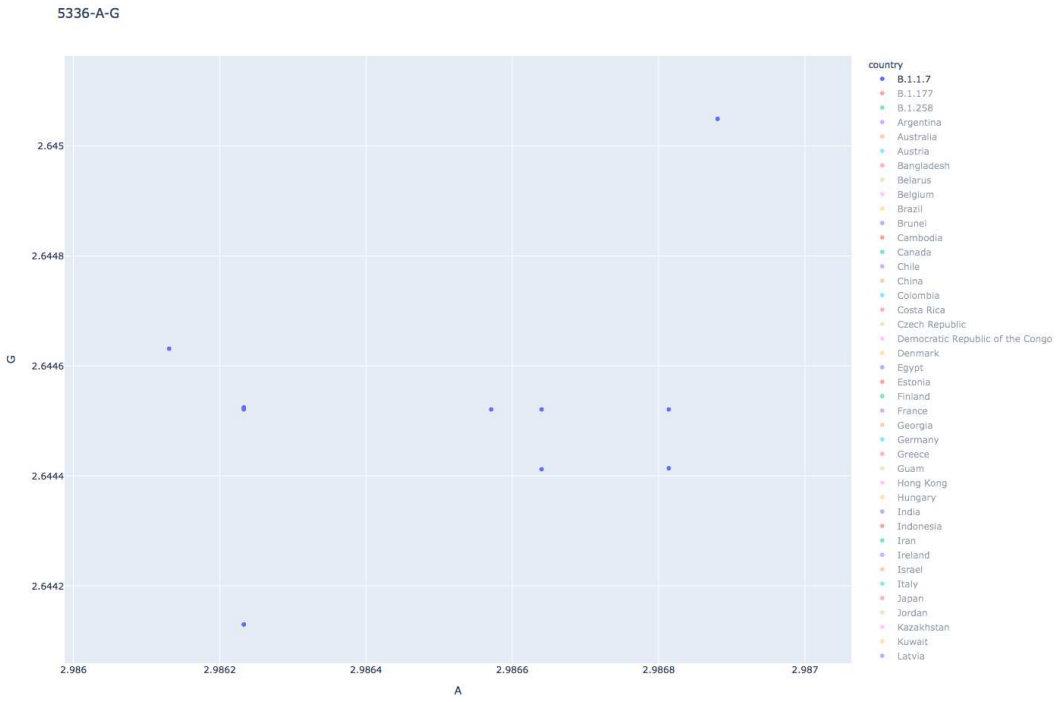


(b)

Fig. 11 Three variations of RDRP and whole genomes on genomic index maps (a) RDRP (b) Whole genomes



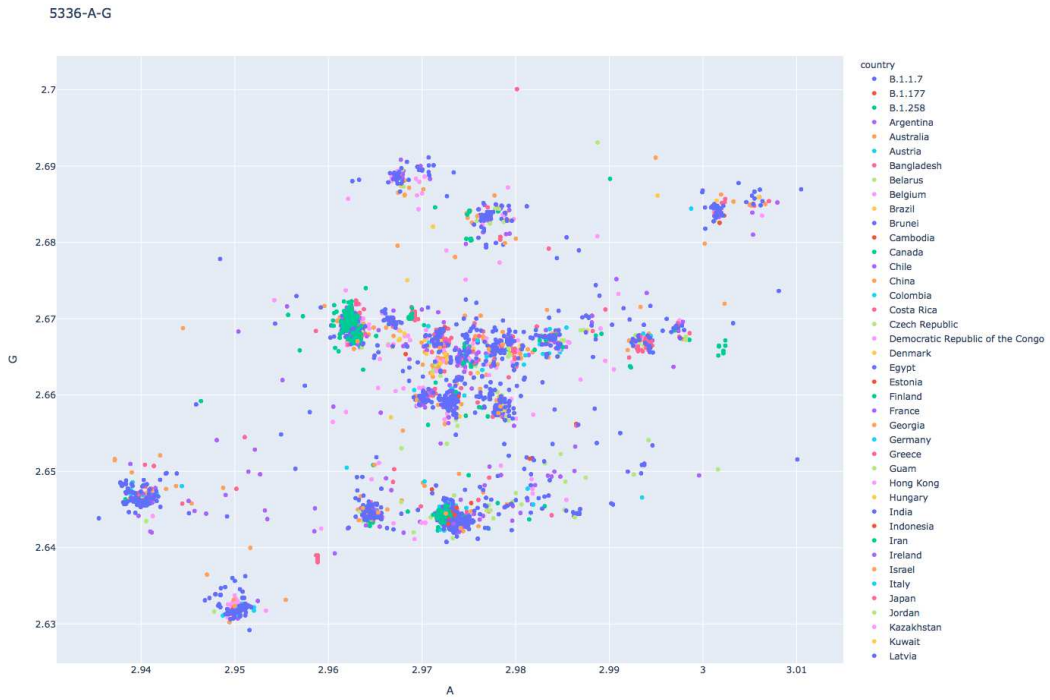
(a)



(b)

Fig. 12 Three variations of RDRP and whole genomes on 100 times of enlarged genomic index maps (a) RDRP (b) Whole genomes

(a)



(b)

Fig. 13 Three variations and six selected regions of RDRP and whole genomes on genomic index maps (a) RDRP (b) Whole genomes

Figures

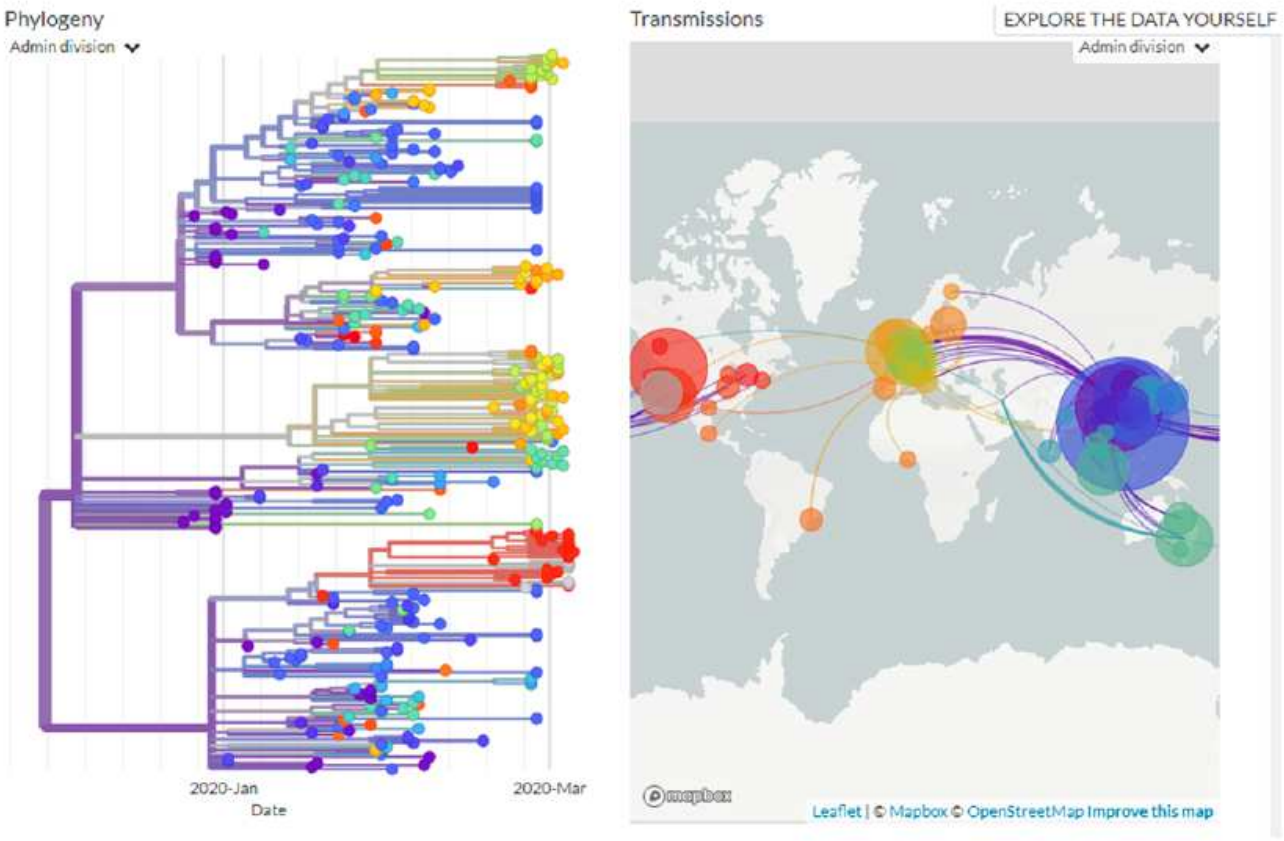


Figure 1

The phylogenetic tree of real cases over global on Nextstrain. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

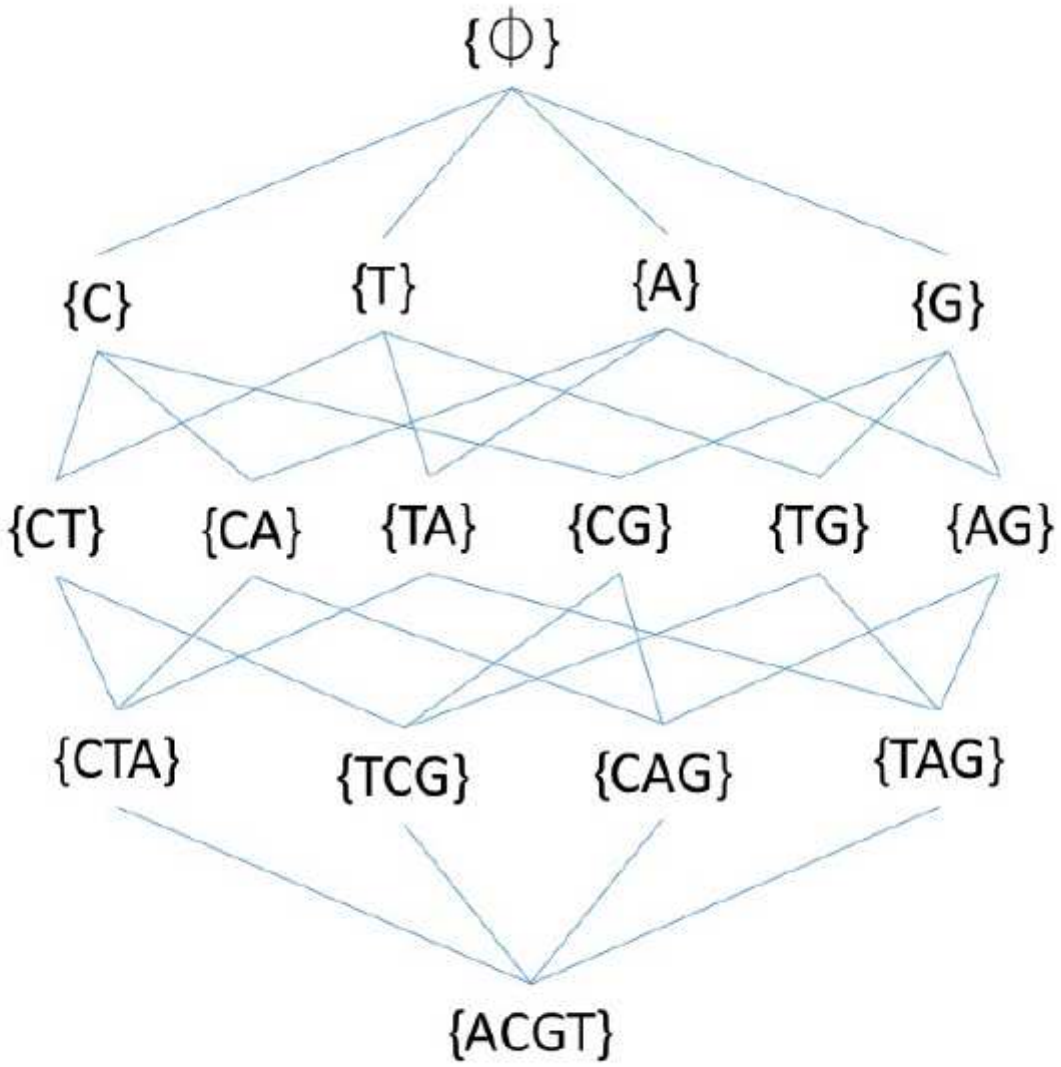


Figure 2

Sixteen combinations of four meta-symbols in a hierarchy of a lattice

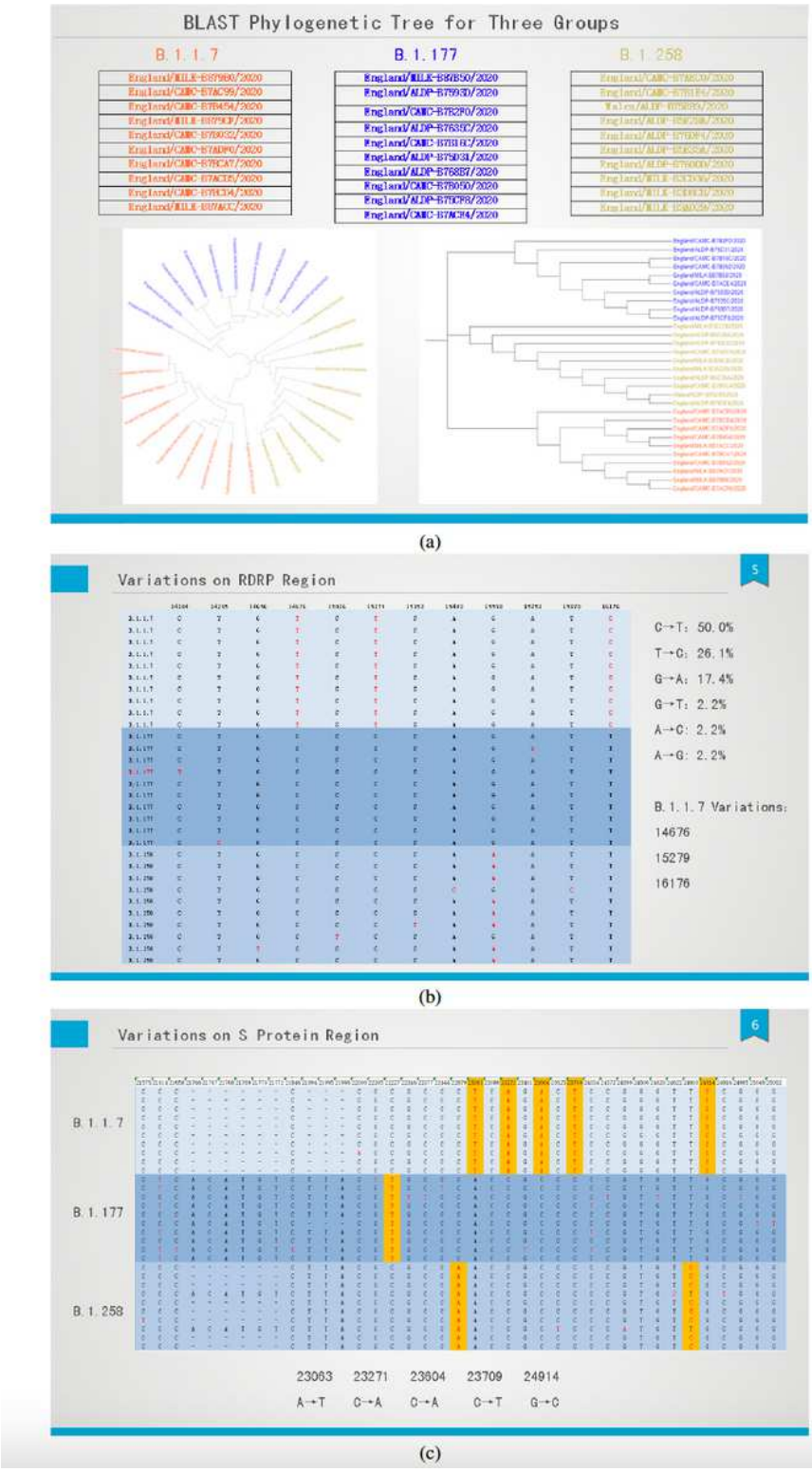


Figure 3

Three groups of variation genomes on BLAST (a) Selected genomes and two Phylogenetic trees (b) Variations in RDRP (c) Variations in S Protein

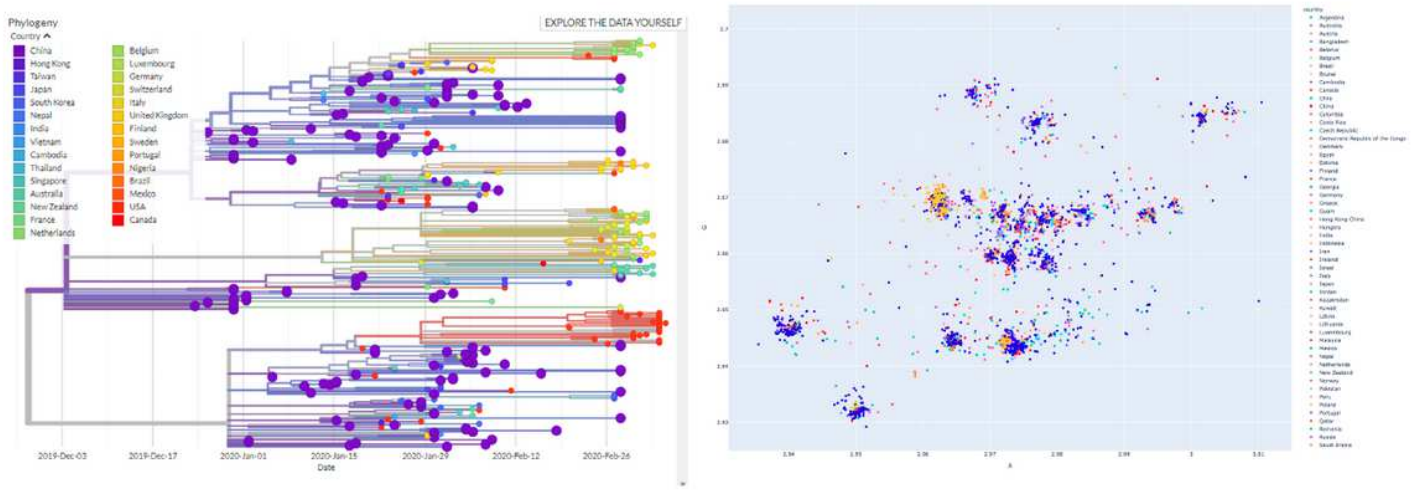


Figure 4

The phylogenetic tree of real cases over global on Nextstrain and Global Genomic Index Map

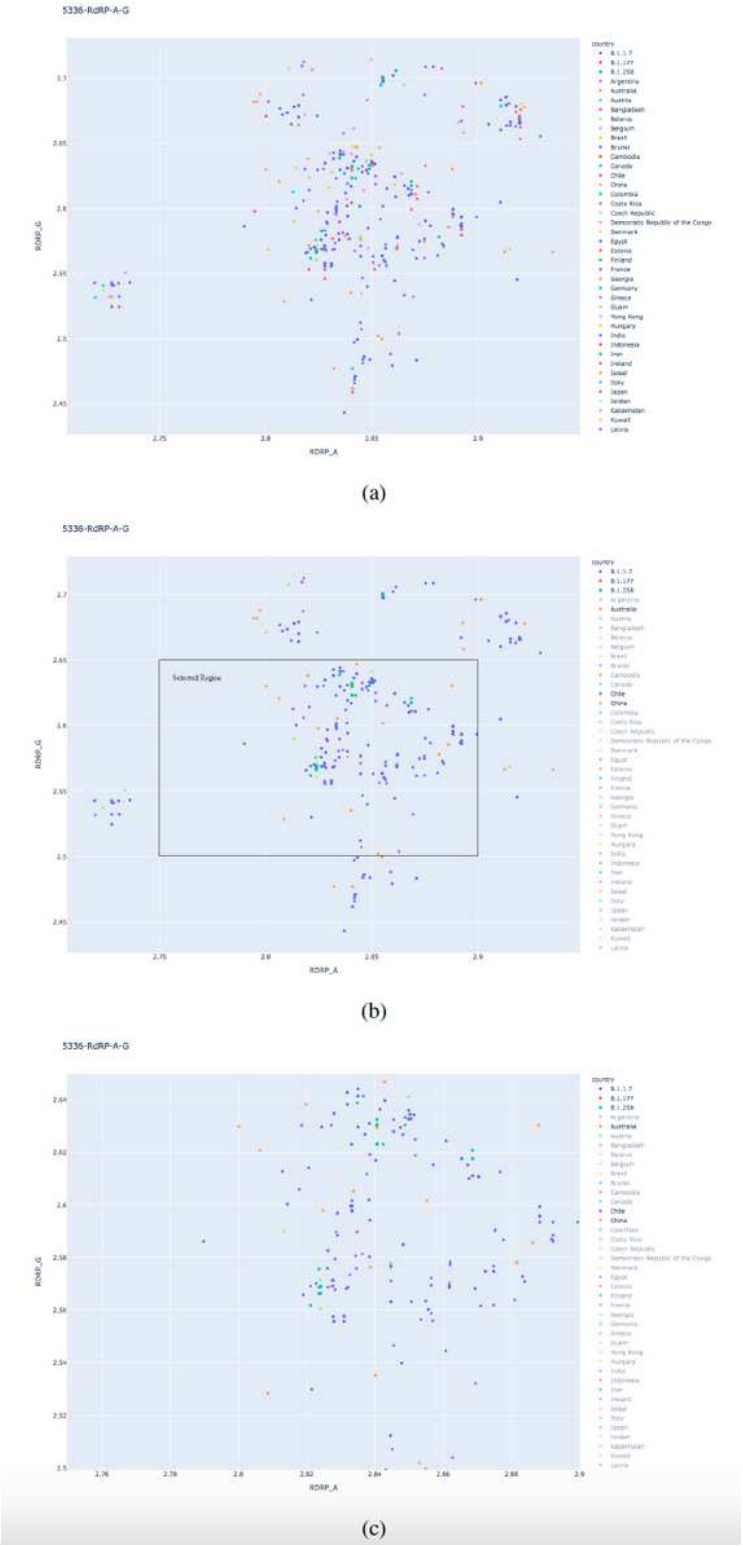


Figure 5

Five thousands of RDRP genomes on genomic index maps (a) Global (b) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (c) An enlarged region selected from (b)

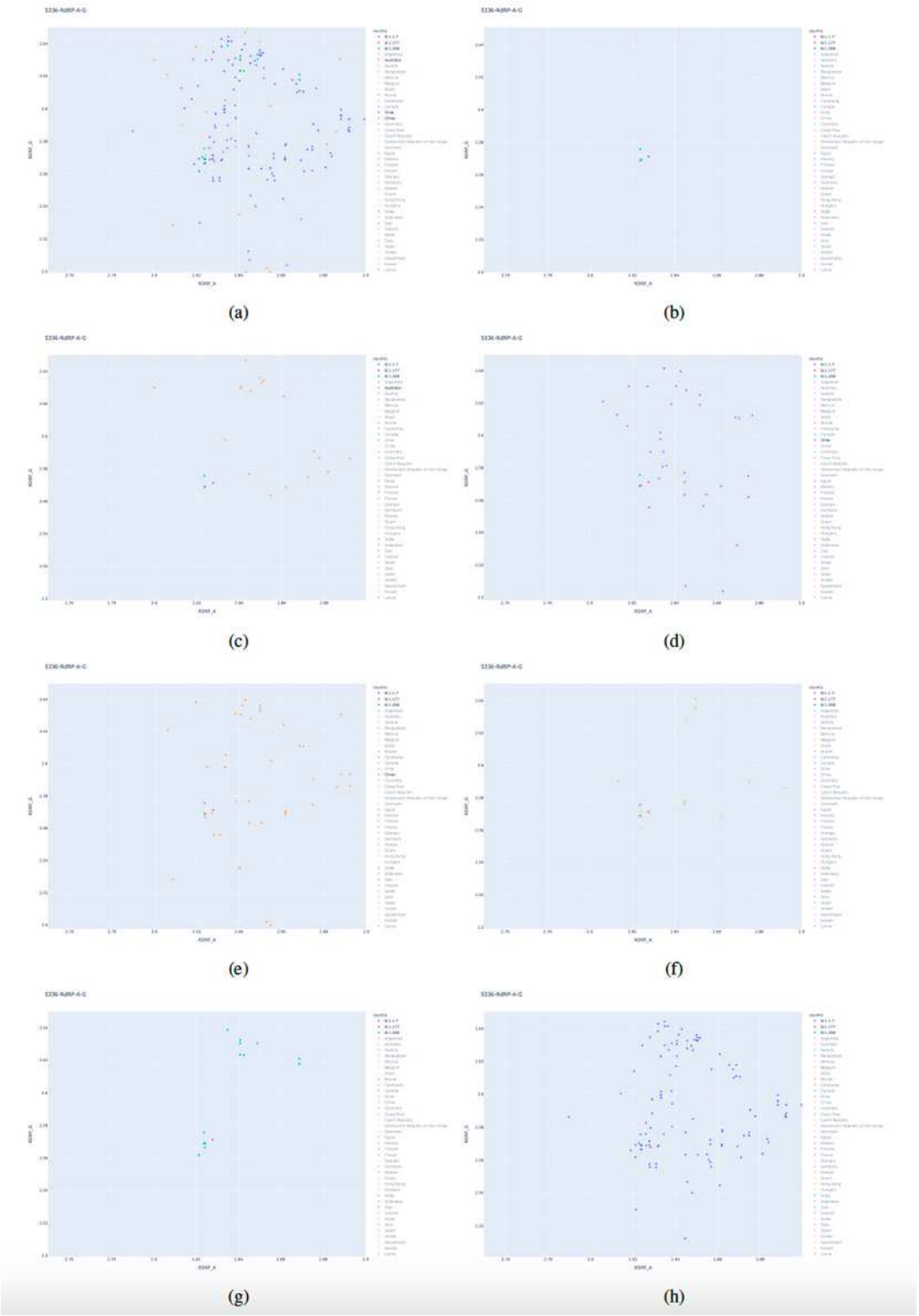


Figure 6

An enlarged region of RDRP on genomic index maps with three groups of variations (a) Six regions: Australia + Chile + China + Taiwan + UK + USA (b) Three groups: B.1.1.7 + B.1.177+B.1.258 (c) Australia (d) Chile (e) China (f) Taiwan (g) UK (h) USA

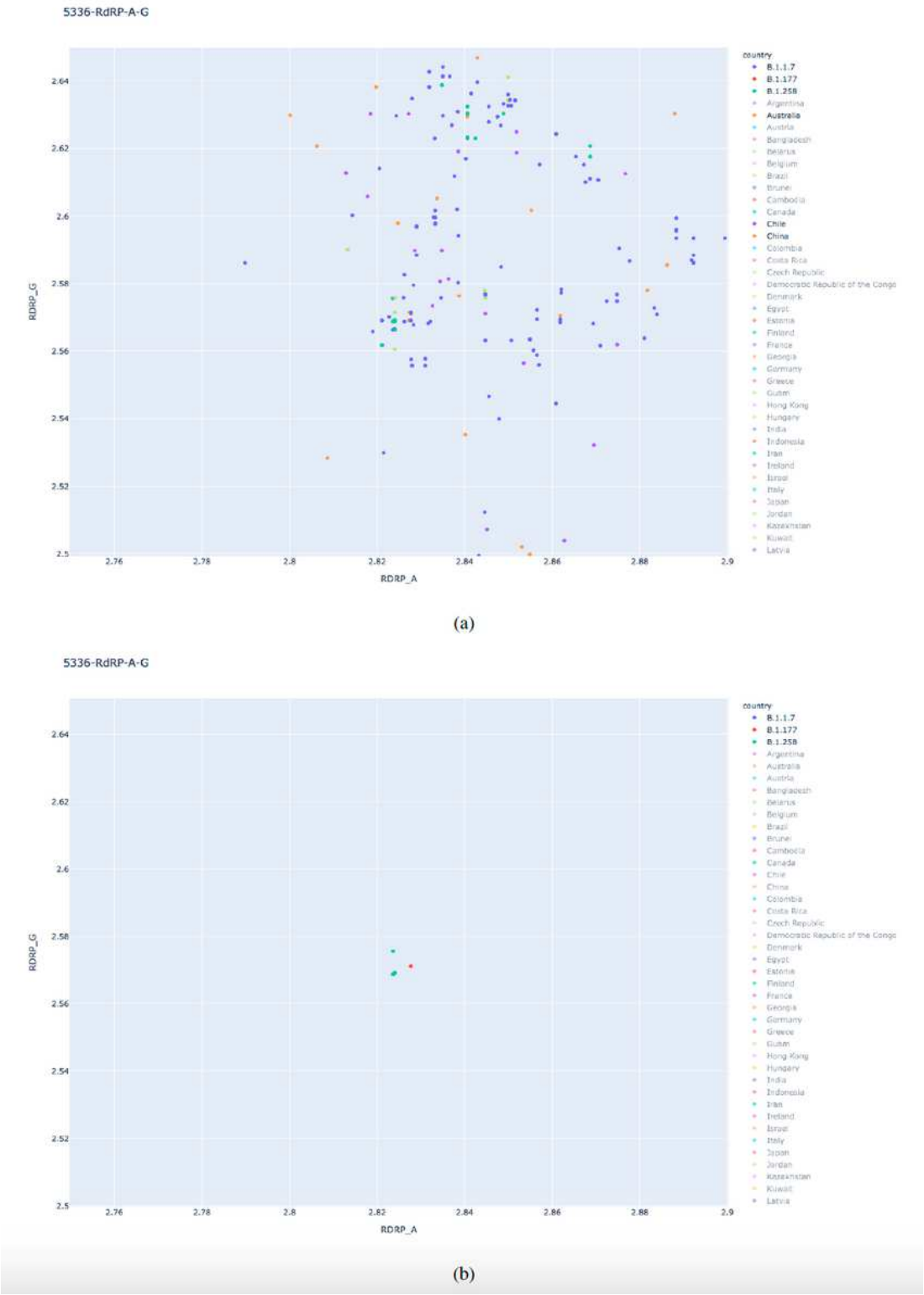


Figure 7

Enlarged Region of RDRP genomes on genomic index maps (a) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (b) Three variations

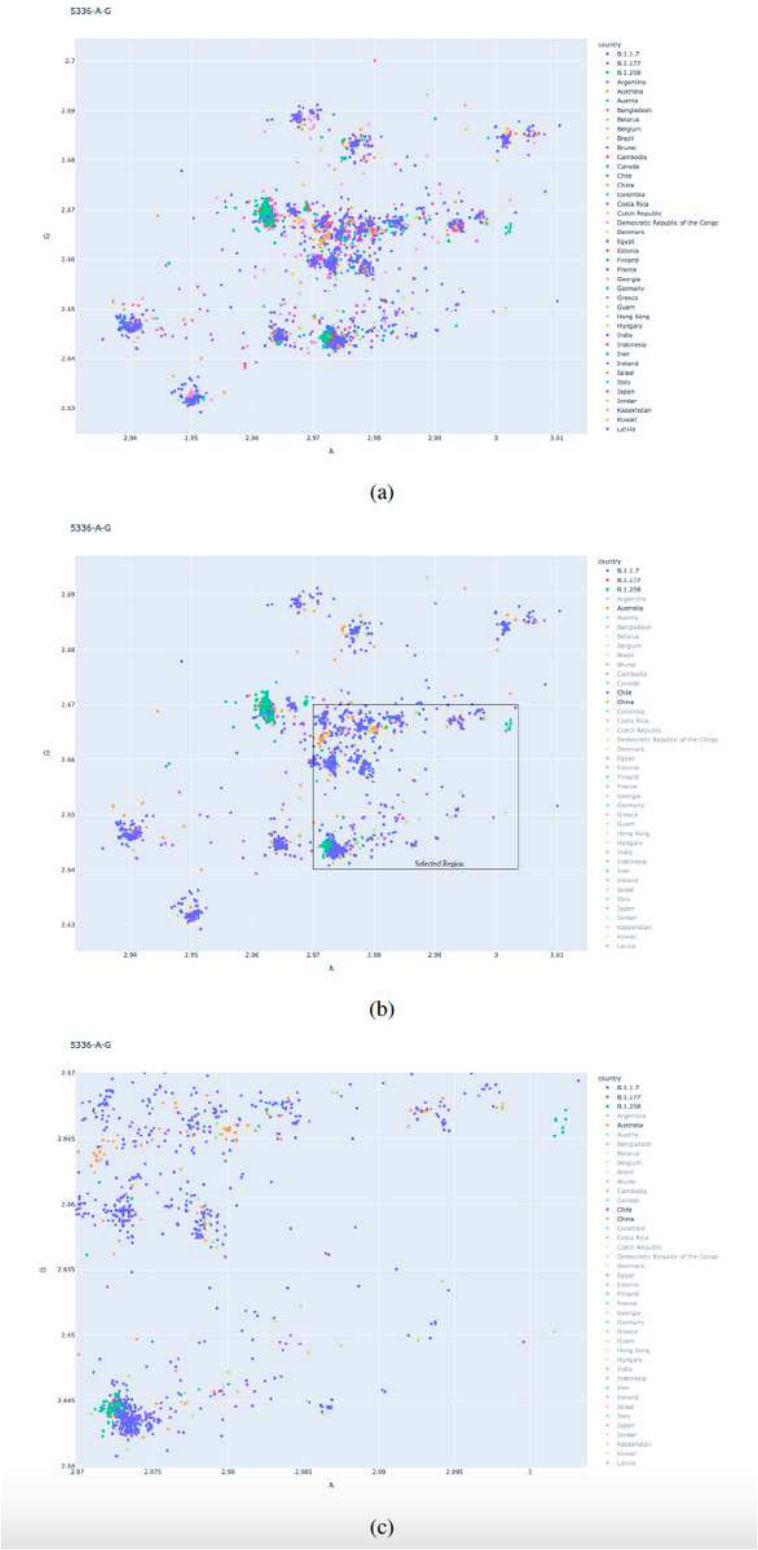


Figure 8

Five thousands of whole genomes on genomic index maps (a) Global (b) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (c) An enlarged region selected from (b)

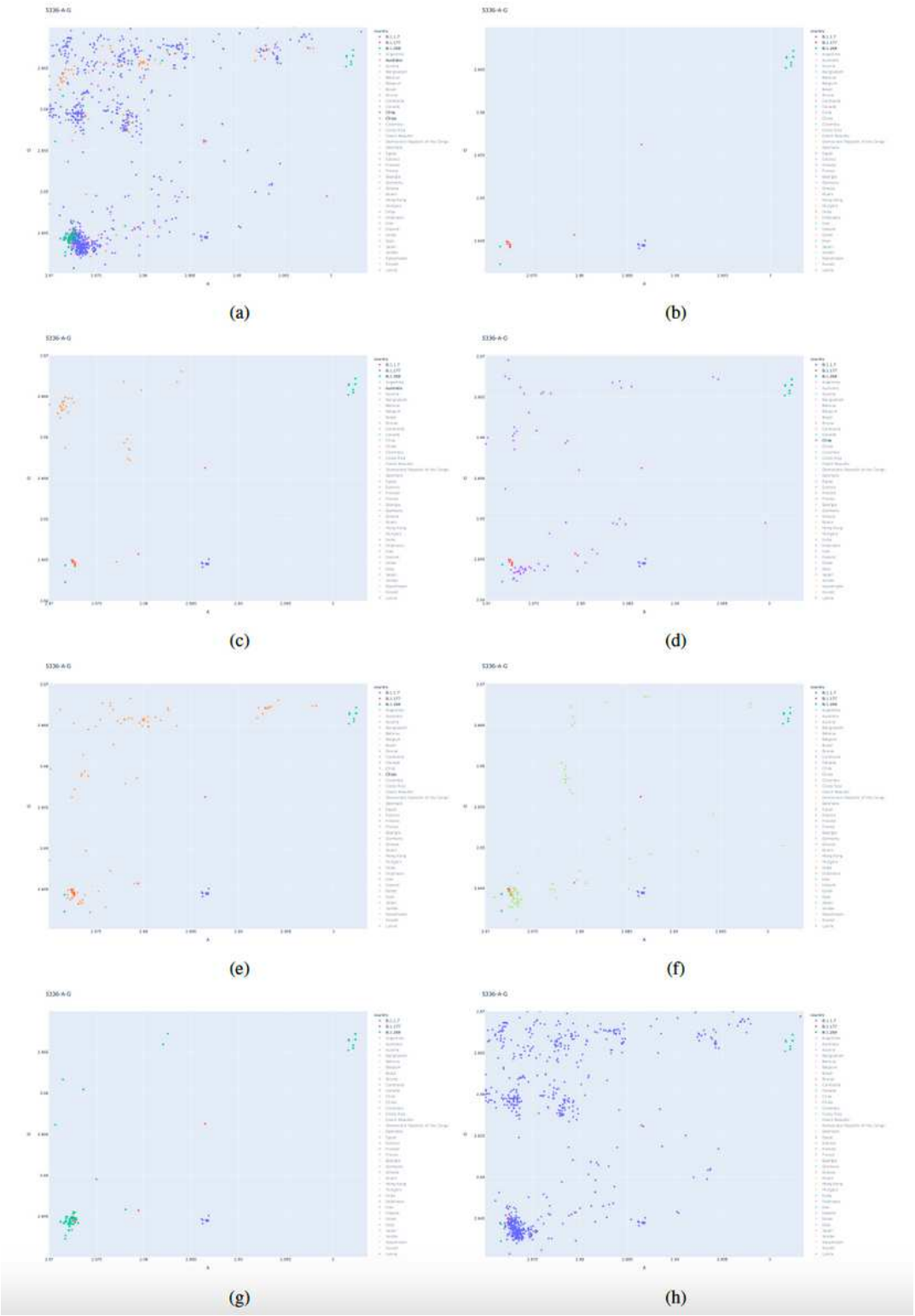


Figure 9

An enlarged region of whole genomes on genomic index maps with three groups of variations (a) Six regions: Australia + Chile + China + Taiwan + UK + USA (b) Three groups: B.1.1.7 + B.1.177+B.1.258 (c) Australia (d) Chile (e) China (f) Taiwan (g) UK (h) USA

Figure 10

Enlarged Region of whole genomes on genomic index maps (a) Six selected regions: Australia + Chile + China + Taiwan + UK + USA (b) Three variations

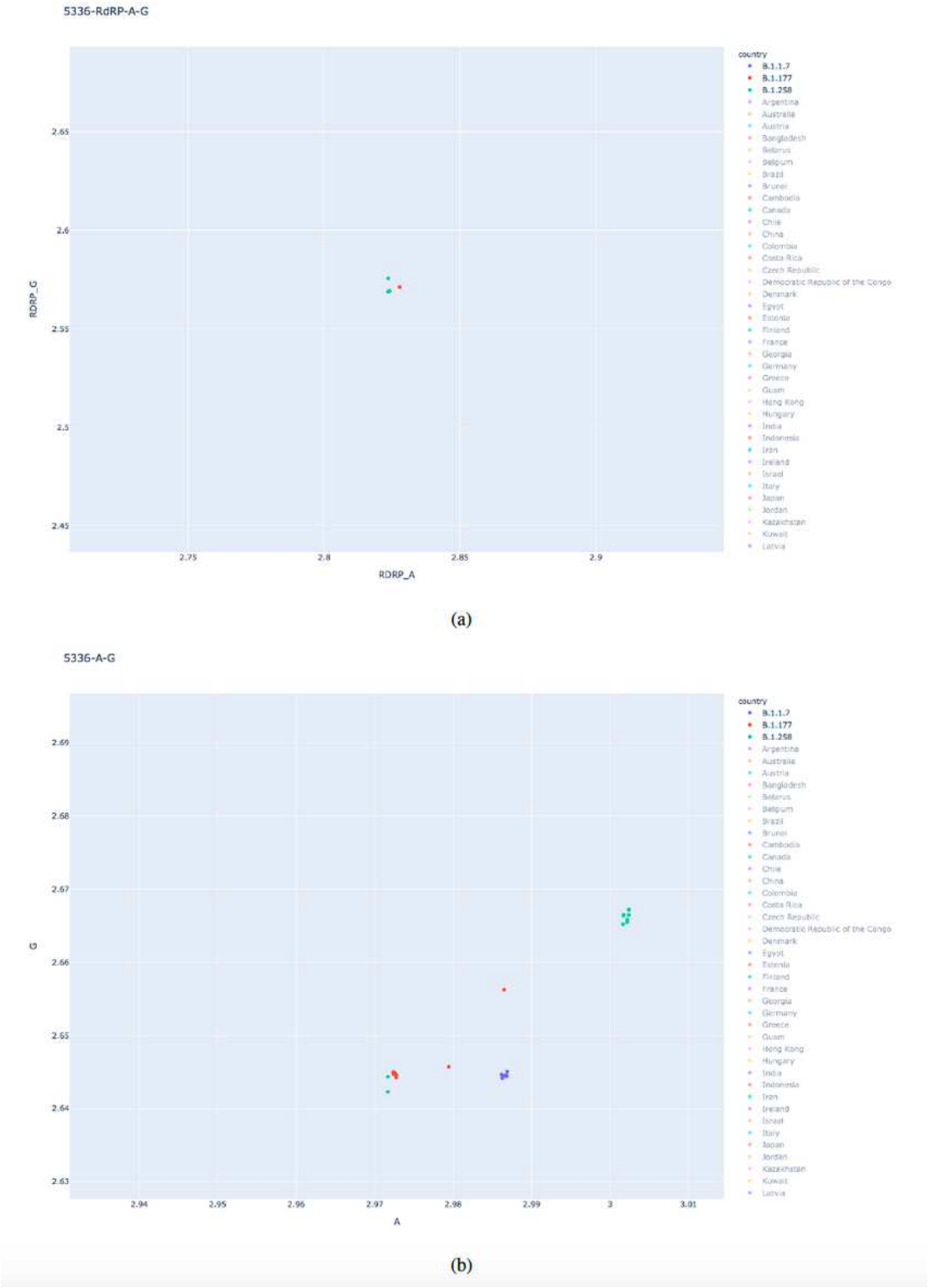


Figure 11

Three variations of RDRP and whole genomes on genomic index maps (a) RDRP (b) Whole genomes

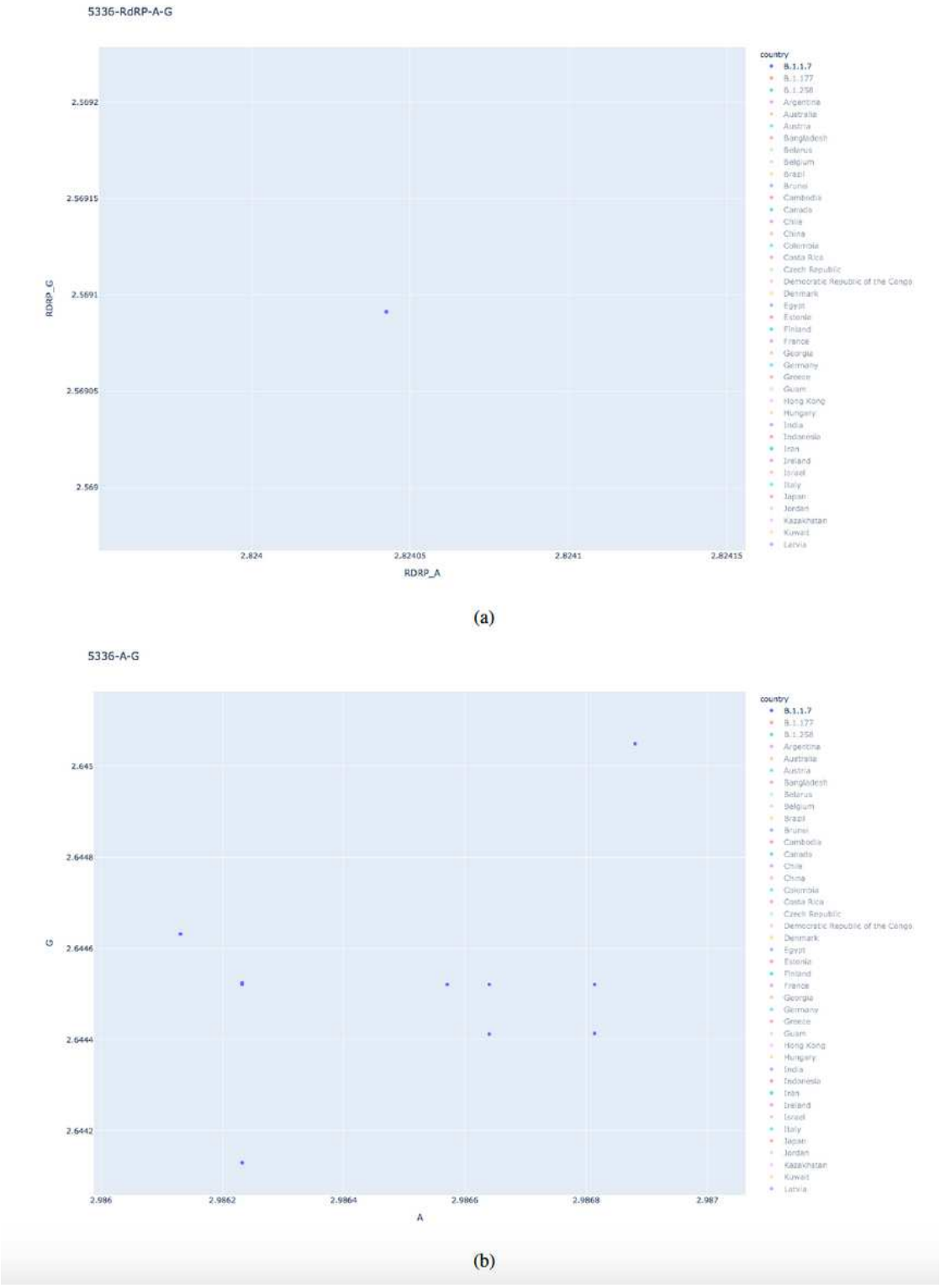


Figure 12

Three variations of RDRP and whole genomes on 100 times of enlarged genomic index maps (a) RDRP (b) Whole genomes

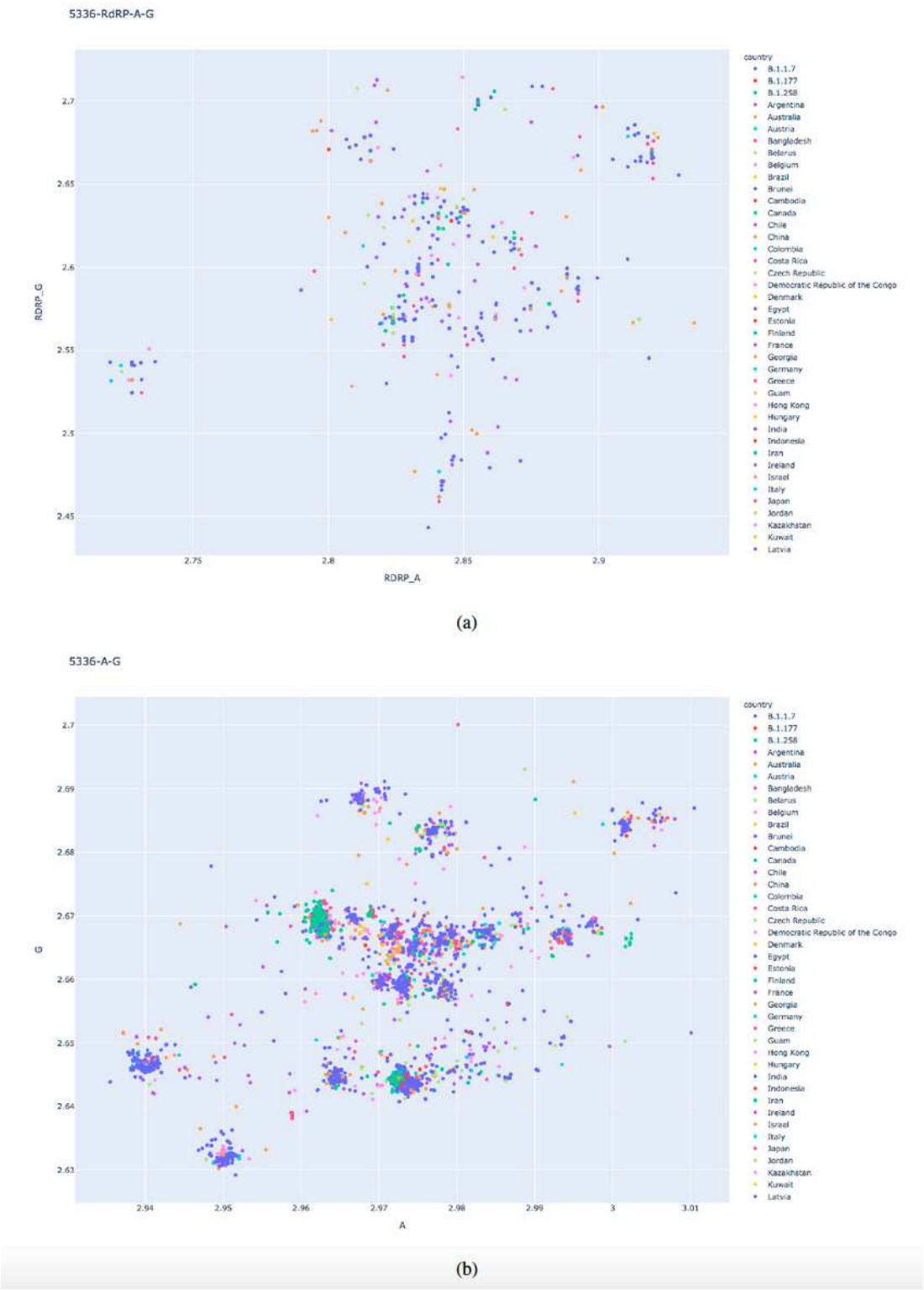


Figure 13

Three variations and six selected regions of RDRP and whole genomes on genomic index maps (a) RDRP (b) Whole genomes

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 5306Infor.xlsx
- UK30Infor.tsv
- 5306Infor.tsv
- UK30Infor.xlsx
- 5336Whole16AG.html
- 5336RdRP16AG.html