OXFORD

Briefings in Bioinformatics, 00(00), 2020, 1–9

doi: 10.1093/bib/bbaa208
Advance Access Publication Date: 00 Month 0000
Problem Solving Protocol

Origin-independent analysis links SARS-CoV-2 local genomes with COVID-19 incidence and mortality

Wenzhong Yang and Guangxu Jin

Corresponding author: Guangxu Jin, Department of Cancer Biology, Wake Forest School of Medicine Wake Forest Baptist Comprehensive Cancer Center Center for Precision Medicine, Wake Forest School of Medicine Medical Center Boulevard, Winston-Salem, NC 27157, USA.Tel: +(336) 713-7515; Fax: +(336) 713-7544. E-mail: gjin@wakehealth.edu

Abstract

There is an urgent public health need to better understand Severe Acute Respiratory Syndrome (SARS)-CoV-2/COVID-19, particularly how sequences of the viruses could lead to diverse incidence and mortality of COVID-19 in different countries. However, because of its unknown ancestors and hosts, elucidating the genetic variations of the novel coronavirus, SARS-CoV-2, has been difficult. Without needing to know ancestors, we identified an uneven distribution of local genome similarities among the viruses categorized by geographic regions, and it was strongly correlated with incidence and mortality. To ensure unbiased and origin-independent analyses, we used a pairwise comparison of local genome sequences of virus genomes by Basic Local Alignment Search Tool (BLAST). We found a strong statistical correlation between dominance of the SARS-CoV-2 in distributions of uneven similarities and the incidence and mortality of illness. Genomic annotation of the BLAST hits also showed that viruses from geographic regions with severe infections tended to have more dynamic genomic regions in the SARS-CoV-2 receptor-binding domain (RBD) and receptor-binding motif (RBM) of the spike protein (S protein). Dynamic domains in the S protein were also confirmed by a canyon region of mismatches coincident with RBM and RBD, without hits of alignments of 100% matching. Thus, our origin-independent analysis suggests that the dynamic and unstable SARS-CoV-2-RBD could be the main reason for diverse incidence and mortality of COVID-19 infection.

Key words: SARS-CoV-2; origin-independent analysis; uneven local genomic similarities; bioinformatics; S protein; dynamic RBD and RBM

Introduction

As of 4 June 2020, the SARS-CoV-2 virus had infected over 1.9 million people in the United States and >6 million worldwide. SARS-CoV-2 is a novel coronavirus recently identified as the causative agent of COVID-19. To understand the genetic characteristics of SARS-CoV-2, human samples from different countries have been sequenced [1–11]. These analyses found 60–96% alignment of genome sequences between sequenced SARS-CoV-2 viruses and coronaviruses from the most likely hosts, bat species [1, 12, 13], in which a bat coronavirus (CoV RaTG13) has a genetic similarity of 96.3%. In turn, the sequenced genome of CoV RaTG13 has been used as reference in recent mutational analyses [14, 15]. For example, a mutation, D614G, was identified by using Bat Coronavirus RaTG13 [15]. Unfortunately, D614G

Wenzhong Yang is a Research Fellow at Department of Internal Medicine, Wake Forest School of Medicine. He has bioinformatics interest in next-generation sequencing data modeling and analysis.

Guangxu Jin is an Assistant Professor at Department of Cancer Biology, Wake Forest School of Medicine; Senior Bioinformatician, Faculty Member of Bioinformatics Shared Resource (BISR), Wake Forest Baptist Comprehensive Cancer Center (WFBCCC); Faculty Member of Center for Precision Medicine, Wake Forest School of Medicine, the Wake Forest Clinical and Translational Science Institute (CTSI) 2020-2022 Translational Research Academy (TRA) Scholar. He has a long-term interest in developing novel computational methods for sequencing analysis and their application to translational and clinical research.

Submitted: 17 June 2020; **Received (in revised form):** 24 July 2020

status was not statistically associated with hospitalization status. Thus, one major challenge is to develop a genetic comparative method independent from unclear ancestors of SARS-CoV-2. The second challenge is to associate sequencing mechanism with the incidence and mortality of COVID-19.

To address these challenges, we used the Basic Local Alignment Search Tool (BLAST) and identified local sequencing similarities among the SARS-CoV-2 genomes (Figure 1A). For the first time, our novel analyses showed a strong statistical association of the uneven local genomic similarity mapping with the incidence and mortality of COVID-19. Furthermore, our findings in the receptor-binding domain (RBD)/receptor-binding motif (RBM) of spike (S) protein reveal, for the first time, the possible genomic mechanism causing the diverse COVID-19 incidence and mortality.

Results and discussion

We collected all available raw sequencing data for SARS-CoV-2 from the four sequencing platforms (Illumina, BGI, Ion-Torrent and Nanopore) deposited in the National Center for Biotechnology Information (NCBI) Short Reads Archive (SRA) database as of 15 April 2020 (Supplementary Table S1). BLAST alignments were implemented between the raw sequencing data of SARS-CoV-2, and the assembled genomes in the NCBI Betacoronavirus database and our customized SARS-CoV-2 database (Figure 1A). The total sequencing read number from 32 data sets in the four sequencing platforms is 1,584,295,725. To ensure the data quality while searching for local sequence similarity, we used the Q30 ratio—the percentage of bases in a read with a Phred score [16] or called Q score >30. The 95% percentile of Q30 ratios of all reads in a sequenced sample is the threshold for the high-quality reads used for BLAST alignments. Thus, we set the Q30 ratio as 0.9 for Illumina and BGI platforms and 0.2 for Ion-Torrent and Nanopore platforms (Figure 1B). Our sequencing analysis strategy is distinct from the most existing methods considering the assembled genomes only (one virus one sequence). The utility of high-quality sequencing raw data (from total of 1,584,295,725) into our analysis ensured the searching of the similarities by local genomic regions and the statistical requirements for genetic comparative analyses.

Firstly, we identified the uneven distribution of the local genomic similarities among the SARS-CoV-2 viruses and its strong association with the incidence and mortality of the illness. We defined two quantitative metrics to facilitate the comparison of the BLAST results across the genomes in the considered coronavirus databases. The BLAST hit score (BHS) is a metric based on the similarity between a high-quality read of a SARS-CoV-2 sample and each assembled genome in the coronavirus databases (Methods). BHS evaluates the similarity of two SARS-CoV-2 viruses by examining the local genomic regions (Methods). Surprisingly, we found that BHS is not evenly disxtributed but dominated by viruses from certain countries (Figure 2, Supplementary Figure S1), notably Spain and the United States. The geographic information was derived from target data sets with the assembled genomes. The results were consistent in analyses using the NCBI Betacoronavirus database and our customized SARS-CoV-2 genome database (Supplementary Figure S1). Another hit score, BLAST total hit score (BTHS), was also defined, evaluating which viruses categorized by geographic regions or the existing coronaviruses are more likely to have high values of BHS in the BLAST alignments (Methods). BTHS excluded bat and pangolin coronaviruses for the similarity analysis because of the low number of hits (Supplementary Figure S2). The BTHS values were strongly correlated to the incidence and mortality of illness in those countries caused by SARS-CoV-2 viruses (Figure 3, Supplementary Table S2, infection cases: $r = 0.468$, $P = 0.009$ and deaths: $r = 0.455$, $P = 0.01$, Pearson's product–moment correlation).

Secondly, we explored why the local genomes of the viruses with high BTHS show the uneven similarity distribution and which genomic regions are crucial for these viruses that may lead to the distinct infection incidence and mortality in different countries. Using the sample collection dates (Supplementary Table S3), we identified that BHS scores of early outbreak samples from China and the United States show a significant difference from other samples. BHS scores displayed two clusters, 4I.C1 and 4I.C2 (Figure 4A, Supplementary Figures S3 and S4, Supplementary Tables S4 and S5). The geographic information was derived from target data sets with the assembled genomes. The cluster of 4I.C1 includes the early outbreak samples from China collected and sequenced as of January 2020 (Figure 4A, shown in columns). 4I.C2 cluster includes the early outbreak samples from both China and the United States (Figure 4A, Supplementary Table S5). We annotated the identified hit sequences of the viruses aligned to the clusters of 4I.C1 and 4I.C2. The top hits from these two clusters have different favorite genomic regions, that is, the SARS-CoV-2 S protein RBD (SARS-CoV-2-RBD) (Figure 4B, Supplementary Tables S6 and S9). Growing evidence showed that the SARS-CoV-2 S protein plays the most important role in viral attachment, fusion and entry and is a target for the development of antibodies, entry inhibitors and vaccines [1, 17–20].

Thirdly, we tested if the viruses in the 4I.C1 and 4I.C2 have specific genome sequences that determine the alignments in SARS-CoV-2-RBD. 4I.C1 does not have any hits with 100% match in the RBD region of S protein, indicating mismatches in the alignments (Figure 4B, right panel). We reduced the match from 100 to 97.5% and identified a few hits that fall in the RBD region including gaps. As an instance, 4I.C1 includes an alignment of 99.4% match with two gaps in the coding region of S protein between MT334549 (SARS-CoV-2/human/USA/UT-00087/2020, complete cds) and a China-Wuhan SARS-CoV-2 sample (raw data from SRR11313271), involving the loci of 22,767–22,956 within the RBD (Figure 4D). Distinctly, 4I.C2 identified another coding region of S protein in the MT334562 genome, which shows 100% match in the alignment to RBM (Figure 4E). These two genomic regions are encoding the RBD/RBM of SARS-CoV-2.

SARS-CoV-2 initiates its infection by the recognition of human angiotensin-converting enzyme 2 (ACE2), which is mediated by the S protein on the virion surface [21–25]. The S protein is cleaved by the human protease into S1 and S2 subunits. S1 subunit contains the RBD that directly binds to ACE2 [21–25]. The RBM is the core component of the RBD [21–25]. RBD and RBM are critical for S protein to bind to human ACE2, in which RBM accounts for the interaction domain of S protein. We call the mismatches from the alignments for 4I.C1 as 'hidden mutations' because we cannot tell which sequence in the alignments has a mutated locus (or loci). These mutations indicate that the binding between RBM and ACE2 is dynamic. To investigate the dynamic binding between RBM and ACE2, we implemented protein structure dockings between the modeled S proteins (with and without the 'hidden mutations' in the alignments of 4I.C1) and human ACE2 protein using HDOCK [26] (Figure 5, Supplementary Figure S5). The mismatches in the alignments for 4I.C1 (Figure 4D) or a deletion of amino acids of R403 and G404 in the RBD region of S protein resulted in the change of binding affinities (ligand rsmd) from 0.90 to 1.64 Å (Methods).

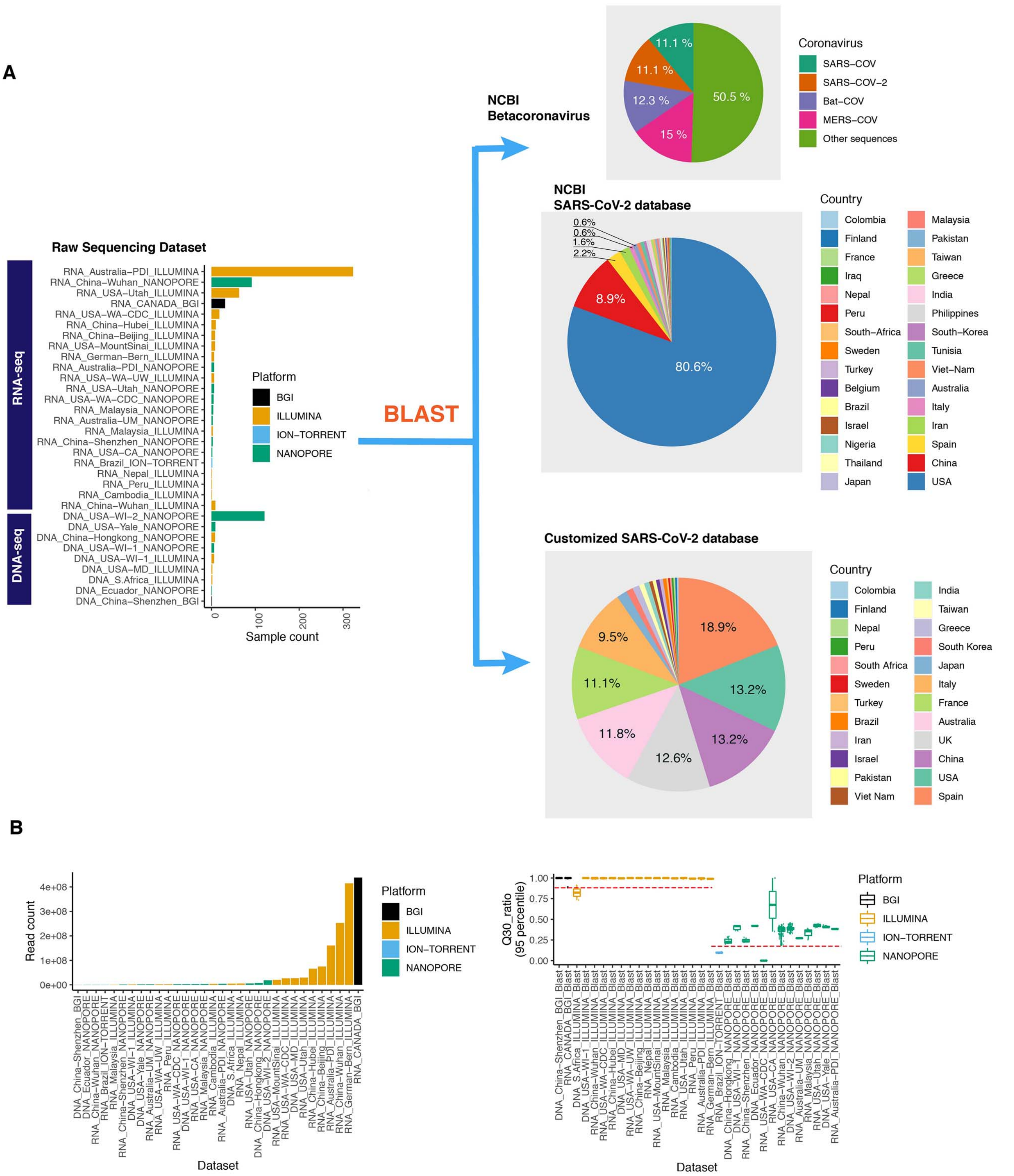


Figure 1. Origin-independent analysis flowchart. (**A**) BLAST alignments were implemented between the high-quality sequencing reads from the data sets identified from four sequencing platforms and the assembled genomes in the coronavirus databases—the NCBI Betacoronavirus database and our customized database. Left: the sample numbers from the 32 raw RNA-seq and DNA-seq data sets collected from SRA; and right: the genome distributions of coronaviruses and SARS-CoV-2 viruses in each of the two databases used. (**B**) The numbers of raw sequencing reads in the 32 data sets. (**C**) The Q30 ratio (at 95% percentile) distribution in the 32 data sets. The threshold for the high-quality reads from BGI and Illumina platforms is 0.9 (marked by a dash line) and that for Ion-Torrent and Nanopore is 0.2 (marked by a dash line).

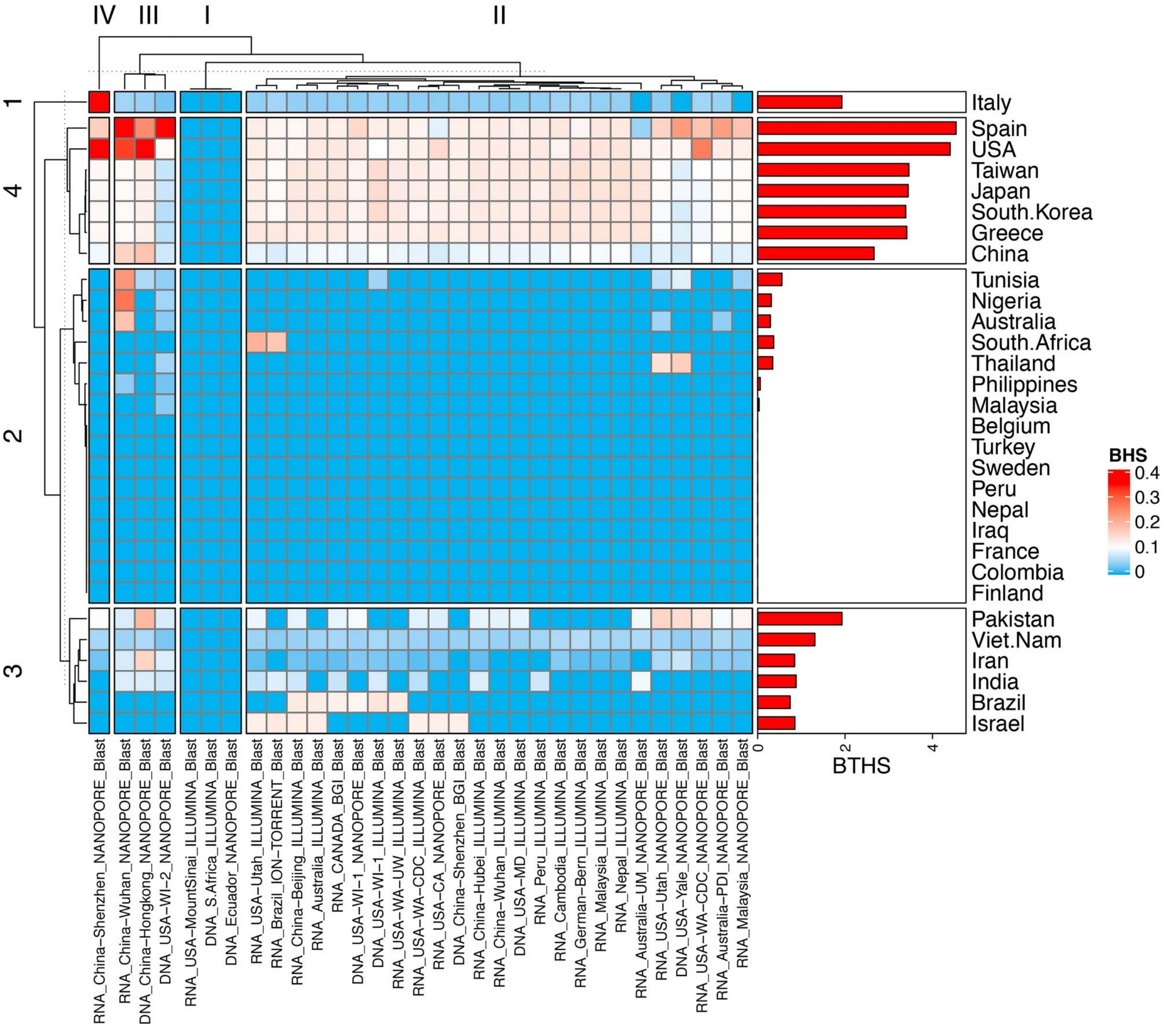


Figure 2. Local genomic similarity distribution by origin-independent analysis. Local genomic similarity mapping by the BHS values of the BLAST alignments between the raw sequencing reads from each data set and the genome sequence in the NCBI Betacoronavirus database. The virus genomes identified from the same country were categorized by the country name, as shown in rows. Results from our customized database are shown in Supplementary Figure S1. The geographic information was derived from target data sets with the assembled genomes.

Lastly, we further identified the dynamic RBM in the SARS-CoV-2 genomes. We used BLAST to align the RBM regions of 380 genomes in our customized SARS-CoV-2 database. A longer region with 'hidden' mutations is elucidated, which we called 'canyon' (Figure 4C). The canyon covers ~80% of RBM and carries hidden genomic mutations that prevent BLAST alignments. The canyon is more likely to appear in the genomes of the viruses from Spain but not the United States, which may explain why the pandemic is more severe in the United States. It may be due to the conservation of RBM in the viruses from the United States.

The difference between the alignments of 4I.C1 and 4I.C2 is in the alignments to the viruses from Spain (Figure 4A). 4I.C1 has low BHS in the alignments to Spain but 4I.C2 has high BHS, meaning that the viruses in 4I.C2 have more hits aligned to the virus genomes from Spain. The hit domains in the S protein can explain the difference (Figure 4B, Chi-square test, $\chi^2 = 15.3$, df = 2, $P = 0.0005$). The hit domains for the alignments in the 4I.C1 (Figure 4B, right panel) and the low BHS scores for 4I.C1 aligned to the viruses from Spain suggest that the dynamic of the RBM/RBD is more complex than expected. The early outbreak samples in 4I.C1 may have their unique mutations (Figure 4D) that prevent the BLAST alignment hits with 100% match.

In summary, the uneven similarity distribution among the SARS-CoV-2 viruses is related to the dynamic genomes of these viruses during the pandemic. Our analysis avoided using the assumed ancestors and hosts for SARS-CoV-2 but instead used local BLAST genomic alignments to study the genomic similarities among the SARS-CoV-2 viruses worldwide. The merit of our local genomic analysis is its elegant capability in revealing the important local genomic sequences for comparison of the SARS-CoV-2 viruses worldwide, e.g. the unstable canyon region in RBD. Distinct from the traditional mutation analyses, we are the first to identify the strong statistical association between the uneven similarity distribution and the infection incidence and mortality. The dynamic regions identified from the RBD/RBM of S protein also explain why the viruses can lead to diverse incidence and

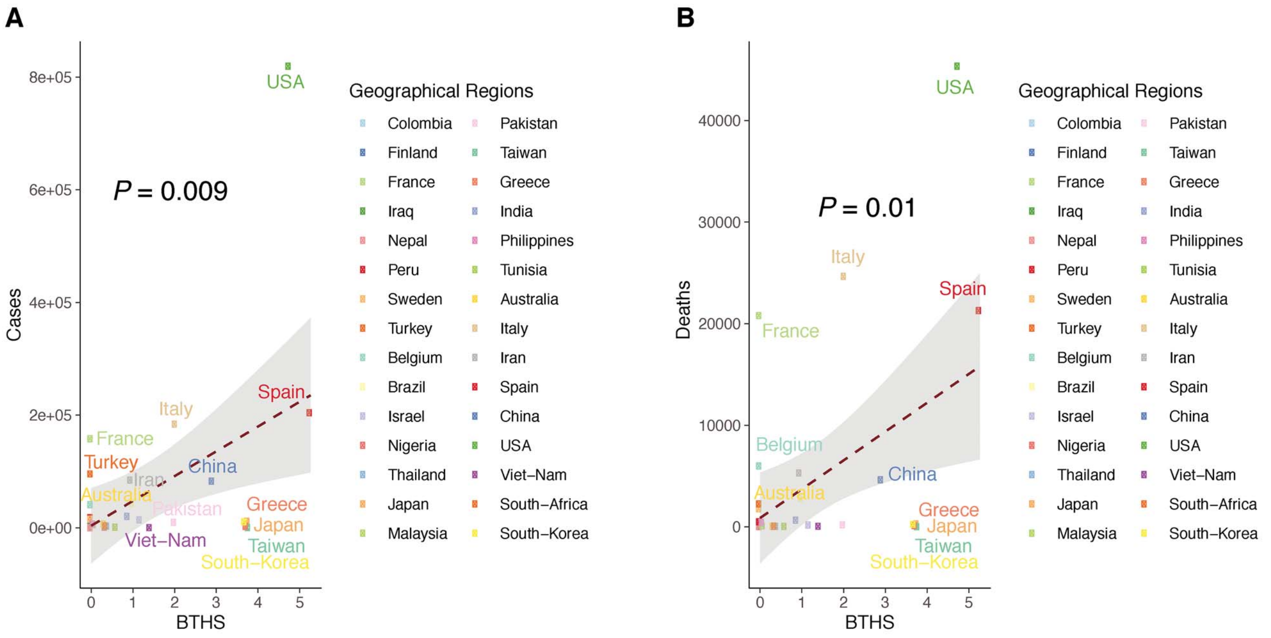


Figure 3. Statistical correlation of BTHS scores of the countries with the infection cases and deaths. (**A**) Association between BTHS scores and infections. (**B**) Association between the BTHS scores and deaths. Statistical data for infections and deaths were collected as of 22 April 2020 (Supplementary Table S2).

mortality of illness. The current severity information of different strains is incomplete yet. If the severity information of each strain is available, additional characterization of protein coding potentials and host–virus protein interactions should be further evaluated. The dynamic genome regions in the viruses should be explored by experimental strategies in the future, which will be more promising to discover potential genomic targets for neutralizing antibodies and vaccine development.

Methods

SARS-CoV-2 raw sequencing data sets

We searched and collected raw data through the SRA (https://www.ncbi.nlm.nih.gov/sra) and identified 32 RNA sequencing (RNA-seq) and DNA sequencing (DNA-seq) data sets for SARS-CoV-2 (Supplementary Table S1). Our data were downloaded by 15 April 2020, and samples were collected from 30 December 2019 to 5 April 2020. The 12 countries covered by the raw sequencing data are worldwide (Asia, Australia, North America, South America, Europe and Africa). The sequencing platforms for generating the raw sequencing data are Illumina, Oxford Nanopore Technologies, BGI and Ion-Torrent. The number of sequence reads is 1,584,295,725 (Figure 1B, Supplementary Table S1). We calculated the Q30 ratio for each sample and used the 95% percentile of the Q30 ratios as the threshold for selecting high-quality reads for local genomic similarity analysis. The threshold for Q30 ratios is 0.9 for Illumina and BGI platforms and 0.2 for Ion-Torrent and Nanopore.

SARS-CoV-2 genomes in the searching databases

To search for similarities among SARS-CoV-2 viruses and known coronaviruses, such as SARS, Middle East Respiratory Syndrome (MERS) and other coronaviruses from bats and other species, we used the NCBI Betacoronavirus database, which includes 1083 SARS-CoV-2 genomes and 8840 sequences (81 375 785 total bases, the longest sequence: 37 971 bases, version: 5, time: 10:29 AM 17 April 2020) (Supplementary Table S10). We also constructed another SARS-CoV-2 genome database with (i) randomly sampled 50 genomes each from the United States and China (from NCBI's database); (ii) keeping genomes from countries other than the United States and China in NCBI's database; and (iii) plus downloaded genomes from the global initiative on sharing all influenza data database for Australia, France, Italy, the United Kingdom and Spain (sample numbers ranging from 40 to 70) (Supplementary Tables S11 and S12). This customized SARS-CoV-2 genome database was designed to exclude the bias of the genome numbers in the NCBI database and test the robustness of our local genomic similarity mapping network. Distinct from the analyses for genomic mutations, we did not exclude the partial genomes from the NCBI genome database in our local genomic similarity analyses. Notably, the names of countries and regions were provided by data submitters.

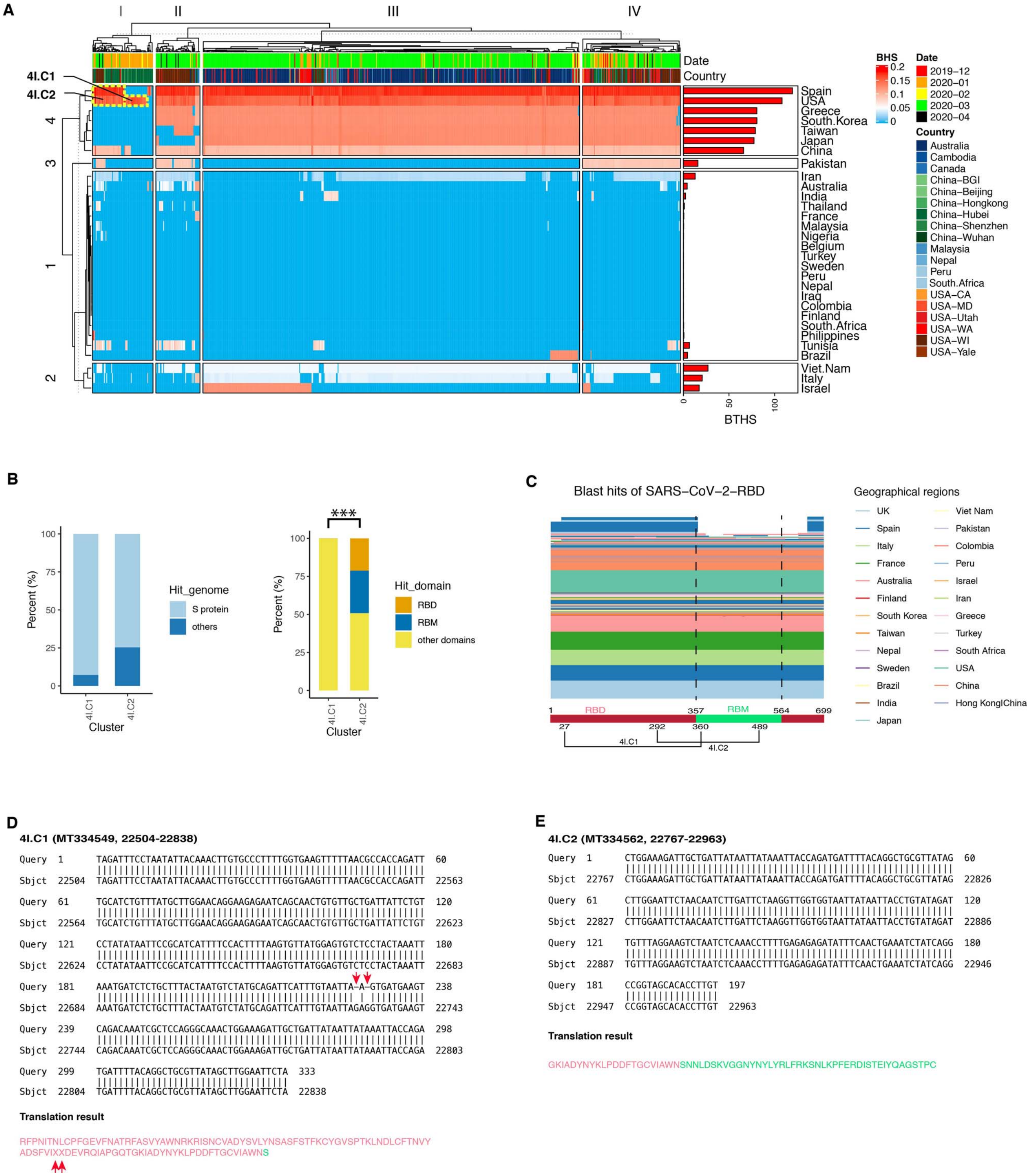
BLAST alignment

We used BLAST for nuclei (blastn) as the alignment tool to align raw sequencing reads to assembled genomes in the NCBI database and our customized database (Supplementary Tables S10 and S12). The newest version of BLASTN was downloaded from the National Institutes of Health BLAST website (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/), which was ncbi-blast-2.10.0+. The parameters used in the alignment are as follows and other parameters were used as default, -perc_identity 97.5 -max_target_seqs 1000 -outfmt "6 qseqid sseqid sallacc stitle pident ppos gaps gapopen mismatch length qstart qend sstart send evalue bitscore score"

We implemented large-scale high-performance computing using Google Cloud Computing.

BLAST hit metrics: BHS and BTHS

To evaluate similarities among the SARS-CoV-2 genomes based on BLAST alignments, we defined the BHS for distribution of



4I.C1 (MT334549, 22504-22838)

```
Query  1      TAGATTTCCTAATATTACAAACTTGTGCCCTTTTGGTGAAGTTTTTAACGCCACCAGATT  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22504  TAGATTTCCTAATATTACAAACTTGTGCCCTTTTGGTGAAGTTTTTAACGCCACCAGATT  22563

Query  61     TGCATCTGTTTATGCTTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGT  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22564  TGCATCTGTTTATGCTTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGT  22623

Query  121    CCTATATAATTCCGCATCATTTTCCACTTTTAAGTGTTATGGAGTGTCTCCTACTAAATT  180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22624  CCTATATAATTCCGCATCATTTTCCACTTTTAAGTGTTATGGAGTGTCTCCTACTAAATT  22683

Query  181    AAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTA-A-GTGATGAAGT  238
              ||||||||||||||||||||||||||||||||||||||||||||||| | ||||||||||
Sbjct  22684  AAATGATCTCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGATGAAGT  22743

Query  239    CAGACAAATCGCTCCAGGGCAAACTGGAAAGATTGCTGATTATAATTATAAATTACCAGA  298
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22744  CAGACAAATCGCTCCAGGGCAAACTGGAAAGATTGCTGATTATAATTATAAATTACCAGA  22803

Query  299    TGATTTTACAGGCTGCGTTATAGCTTGGAATTCTA  333
              |||||||||||||||||||||||||||||||||||
Sbjct  22804  TGATTTTACAGGCTGCGTTATAGCTTGGAATTCTA  22838
```

Translation result

```
RFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVY
ADSFVIXXDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNS
```

4I.C2 (MT334562, 22767-22963)

```
Query  1      CTGGAAAGATTGCTGATTATAATTATAAATTACCAGATGATTTTACAGGCTGCGTTATAG  60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22767  CTGGAAAGATTGCTGATTATAATTATAAATTACCAGATGATTTTACAGGCTGCGTTATAG  22826

Query  61     CTTGGAATTCTAACAATCTTGATTCTAAGGTTGGTGGTAATTATAATTACCTGTATAGAT  120
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22827  CTTGGAATTCTAACAATCTTGATTCTAAGGTTGGTGGTAATTATAATTACCTGTATAGAT  22886

Query  121    TGTTTAGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGG  180
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  22887  TGTTTAGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAACTGAAATCTATCAGG  22946

Query  181    CCGGTAGCACACCTTGT  197
              |||||||||||||||||
Sbjct  22947  CCGGTAGCACACCTTGT  22963
```

Translation result

```
GKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC
```

Figure 4. Dynamic SARS-CoV-2-RBD/RBM. (**A**) BHS scores displayed two clusters, 4I.C1 and 4I.C2. The cluster of 4I.C1 includes the early outbreak samples from China collected and sequenced as of January 2020 (shown in columns), whereas the 4I.C2 cluster includes the early outbreak samples from both China and the United States. The geographic information was derived from target data sets with the assembled genomes. (**B**) Distribution of BLAST hits in the genomic regions for S protein and other genes (left) and distribution of BLAST hits in the specific genomic regions of S protein (right). ***$P < 0.005$ (Chi-square test). (**C**) Alignment of SARS-CoV-2-RBD. (**D**) The alignment of the loci 22838–22504 of the MT334549 genome top hit in 4I.C1 and the corresponding amino acid sequence. Note that two gaps (upper) and amino acids (bottom) are indicated by arrows. The translated amino acids in RBM and RBD other than RBM region are shown in red and green, respectively. (**E**) The alignment of the loci 22767–22963 of the MT334562 genome top hit in 4I.C2 and the corresponding amino acid sequence. The translated amino acids in RBM and RBD other than RBM region are shown in red and green, respectively.

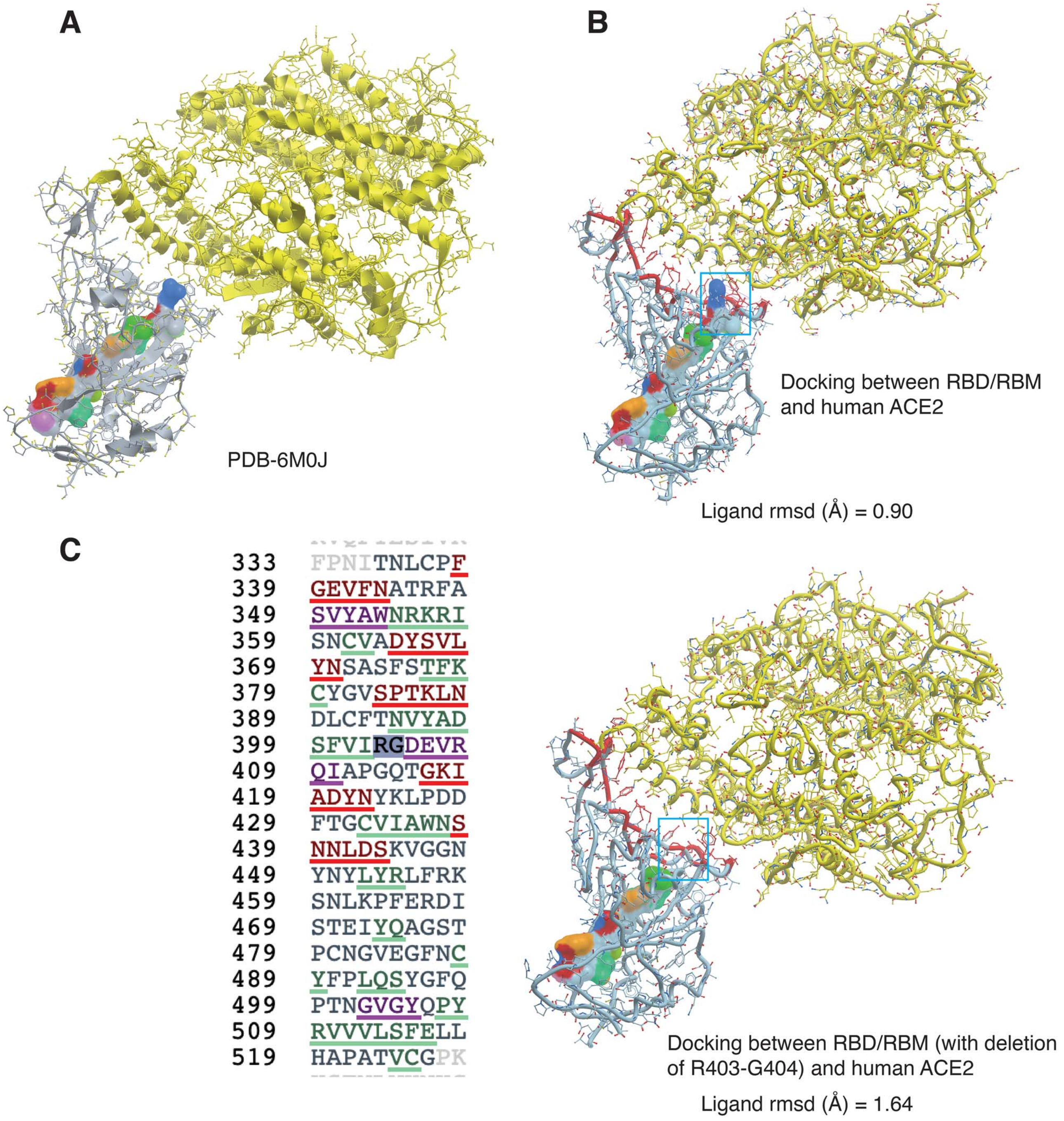


Figure 5. Protein docking for Spike and ACE2 proteins. (**A**) Crystal structure of SARS-CoV-2 spike RBD bound with ACE2 (PDB ID: 6M0J). (**B**) Predicted binding between S protein with ACE2. Meshes are for the backbone from N394 to G404. (**C**) Predicted binding between S protein with a deletion (R403–G4040) and ACE2. Meshes are for the backbone from N394 to I402.

alignments of one sample to the genomes in the searching database categorized by countries. The BHS metric is calculated by the number, N, of total reads aligned to all the SARS-CoV-2 genomes in the database, the number, n, of reads aligned to the genomes from one country, c, and the number of the genomes from this country, g, in the searching database. We defined

$$\mathrm{BHS}_c = \frac{n}{(\mathrm{N} * g)}$$

Based on the BHS, we also defined another metric, called BTHS, to integrate the BHS scores across data sets. For the data sets, 1, 2, ..., d, we defined the BTHS for a country, c, as

$$\mathrm{BTHS}_c = \sum_{i=1}^{d} \mathrm{BHS}_i$$

We evaluated the BHS and BTHS scores for the data sets and the countries by two thresholds for BLAST alignments, that is, 97.5 and 100 for perc_identity in blastn. All the results shown in Figures 2 and 3 were taken from the BLAST alignments with perc_identity = 100.

Structure modeling of binding between S and ACE2 proteins

To investigate the deletion of R403-G404 in RBM of S protein in the alignments for 4I.C1 (Figure 4D), we predicted protein structures of S protein with and without the deletion and implemented protein dockings between the proteins and human ACE2 protein using HDOCK [26]. Predicted models were sorted by docking score and ligand rmsd. The models selected for S protein/ACE2 binding were those with lowest docking score and ligand rmsd. To compare binding affinities of these two dockings, we used the metric of DockRMSD [27]. The protein sequence for human ACE2 was derived from Protein Data Bank (PDB) (6M0J) [28]. The binding between S protein and human ACE2 was visualized and annotated using the ICM software [29].

Data visualization

The heat maps in Figures 2 and 3 were generated by the complexheatmap package in R (https://jokergoo.github.io/ComplexHeatmap-reference).

Statistical analysis

Pearson's product–moment correlations and Welch two-sample *t*-tests (two sided) were implemented by R software.

Data availability

Raw sequencing data set information was included in Supplementary Table S1.

Code availability

Data and codes used to create the figures can be found in the supplemental files. The raw data, results and analyses can be found at https://github.com/guangxujin/COVID-19.

Key Points

- Without needing to know ancestors, we developed origin-independent analyses using a pairwise comparison of local genome sequences of SARS-CoV-2 genomes by BLAST.
- We identified an uneven distribution of local genome similarities among the SARS-CoV-2 viruses categorized by geographic regions.
- For the first time, our novel analyses showed a strong statistical association of the uneven local genomic similarity mapping with the incidence and mortality of COVID-19.
- Our findings in the RBD/RBM of S protein reveal, for the first time, the possible genomic mechanism causing the diverse COVID-19 incidence and mortality.
- Our methods reveal a new strategy to explore S protein function and its impact on COVID-19 incidence and mortality and infection patterns.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

The authors would like to acknowledge all the researchers who openly shared their genomic data on GenBank, SRA and GISAID.

Funding

The Wake Forest Baptist Comprehensive Cancer Center Bioinformatics Shared Resource (P30CA012197); Wake Forest Start-up funds (PI: G.J.)

Disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute.

Conflict of Interest

There is no conflict of interest declared.

Author contributions

G.J. was responsible for conception and design of the study and designed the computational analysis software and codes. W.Y. collected data and implemented the analyses. W.Y. and G.J. wrote the paper.

References

1. Zhou P, Yang XL, Wang XG, *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**(7798):270–3.
2. Manning E, Bohl JA, Lay S, *et al*. Rapid metagenomic characterization of a case of imported COVID-19 in Cambodia. 2020; *bioRxiv*. doi: 10.1101/2020.03.02.968818.
3. Blanco-Melo D, Nilsson-Payant BE, Liu W, *et al*. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. 2020; *bioRxiv*. 10.1101/2020.03.24.004655.
4. Shen Z, Xiao Y, Kang L, *et al*. Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients. *Clin Infect Dis* 2020; ciaa203.
5. Harcourt J, Tamin A, Lu X, *et al*. Severe acute respiratory syndrome coronavirus 2 from patient with 2019 novel coronavirus disease United States. *Emerg Infect Dis* 2020;**26**(6): 1266–73.
6. To KK, Tsang OT, Leung W, *et al*. Temporal profiles of viral load in posterior oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. *Lancet Infect Dis* 2020;**26**(6):565–74.
7. Sah R, Rodriguez-Morales AJ, Jha R, *et al*. Complete genome sequence of a 2019 novel coronavirus (SARS-CoV-2) strain isolated in Nepal. *Microbiol Resour Announc* 2020;**9**(11): e00169–20.

8. Bedford T, Greninger AL, Roychoudhury P, *et al*. Cryptic transmission of SARS-CoV-2 in Washington State. *medRxiv* 2020. doi: 10.1101/2020.04.02.20051417.
9. Chu H, Chan JF, Yuen TT, *et al*. Comparative tropism, replication kinetics, and cell damage profiling of SARS-CoV-2 and SARS-CoV with implications for clinical manifestations, ransmissibility, and laboratory studies of COVID-19: an observational study. *Lancet Microbe* 2020;**1**(1):e14–e23.
10. Paules CI, Marston HD, Fauci AS. Coronavirus infections-more than just the common cold. *JAMA* 2020;**323**(8): 707–8.
11. Deng X, Gu W, Federman S, *et al*. A genomic survey of SARS-CoV-2 reveals multiple introductions into Northern California without a predominant lineage. *medRxiv* 2020. doi: 10.1101/2020.03.27.20044925.
12. Shereen MA, Khan S, Kazmi A, *et al*. COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. *J Adv Res* 2020;**24**:91–8.
13. Khan S, Siddique R, Shereen MA, *et al*. Emergence of a novel coronavirus, severe acute respiratory syndrome coronavirus 2: biology and therapeutic options. *J Clin Microbiol* 2020;**58**(5): e00187–20.
14. Forster P, Forster L, Renfrew C, Forster M Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* 2020;**117**(17):9241–3.
15. Korber B , Fischer WM, Gnanakaran S, *et al*. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020. doi: 10.1101/2020.04.29.069054.
16. Liao P, Satten GA, Hu YJ. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genet Epidemiol* 2017;**41**(5):375–87.
17. Wrapp D, Wang N, Corbett KS, *et al*. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020;**367**(6483):1260–3.
18. Ou X, Liu Y, Lei X, *et al*. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun* 2020;**11**(1):1620.
19. Tai W, He L, Zhang X, *et al*. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell Mol Immunol* 2020;**17**(6):613–20.
20. Andersen KG, Rambaut A, Lipkin WI, *et al*. The proximal origin of SARS-CoV-2. *Nat Med* 2020;**26**(4):450–2.
21. Lan J, Ge J, Yu J, *et al*. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* 2020;**581**(7807):215–20.
22. Yan R, Zhang Y, Li Y, *et al*. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 2020;**367**(6485):1444–8.
23. Wang Q, Zhang Q, Wu L, *et al*. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* 2020; **181**(4):894–904.
24. Walls AC, Park YJ, Tortorici MA, *et al*. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 2020;**181**(2):281–92.
25. Wan Y, Shang J, Graham R, *et al*. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J Virol* 2020;**94**(7):e00127–20.
26. Yan Y, *et al*. The HDOCK server for integrated protein-protein docking. *Nat Protoc* 2020;**15**(5):1829–52.
27. Bell EW, Zhang Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *J Chem* 2019;**11**(1):40.
28. Berman HM, Westbrook J, Feng Z, *et al*. The protein data Bank. *Nucleic Acids Res* 2000;**28**(1):235–42.
29. Katritch V, Rueda M, Abagyan R. Ligand-guided receptor optimization. *Methods Mol Biol* 2012;**857**:189–205.