
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

International Journal of Infectious Diseases 100 (2020) 216–223

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

# A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic

Qingtian Guan[a], Mukhtar Sadykov[a], Sara Mfarrej[a,1], Sharif Hala[a,b,c,1], Raeece Naeem[a,1], Raushan Nugmanova[a], Awad Al-Omari[d,e], Samer Salih[e], Abbas Al Mutair[e], Michael J. Carr[f,g], William W. Hall[f,g,h], Stefan T. Arold[i,j], Arnab Pain[a,g,k,*]

[a] King Abdullah University of Science and Technology (KAUST), Pathogen Genomics Laboratory, Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia
[b] Clinical Microbiology Department, King Abdullah International Medical Research Centre, Ministry of National Guard Health Affairs, Jeddah, Saudi Arabia
[c] King Saud bin Abdulaziz University for Health Sciences, Jeddah, Saudi Arabia
[d] School of Medicine, Alfaisal University, Riyadh, Saudi Arabia
[e] Dr.Suliman Al-Habib Medical Group, Riyadh, Saudi Arabia
[f] National Virus Reference Laboratory (NVRL), School of Medicine, University College Dublin, Belfield, D04 V1W8, Dublin, Ireland
[g] Research Center for Zoonosis Control, Global Institution for Collaborative Research and Education (GI-CoRE), Hokkaido University, N20 W10 Kita-ku, Sapporo, 001-0020 Japan
[h] Global Virus Network (GVN), 801 W. Baltimore St., Baltimore, MD, 21201, USA
[i] King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia
[j] Centre de Biochimie Structurale, CNRS, INSERM, Université de Montpellier, 34090 Montpellier, France
[k] Nuffield Division of Clinical Laboratory Sciences (NDCLS), The John Radcliffe Hospital, University of Oxford, Headington, Oxford, OX3 9DU, United Kingdom

ARTICLE INFO

*Article history:*
Received 4 June 2020
Received in revised form 15 August 2020
Accepted 18 August 2020

*Keywords:*
SARS-CoV-2
Genetic surveillance
Barcoding
Genome variation

ABSTRACT

*Objective:* The SARS-CoV-2 pathogen has established endemicity in humans. This necessitates the development of rapid genetic surveillance methodologies to serve as an adjunct to existing comprehensive, albeit though slower, genome sequencing-driven approaches.
*Methods:* A total of 21,789 complete genomes were downloaded from GISAID on May 28, 2020, for analyses. We have defined the major clades and subclades of circulating SARS-CoV-2 genomes. A rapid sequencing-based genotyping protocol was developed and tested on SARS-CoV-2-positive RNA samples by next-generation sequencing.
*Results:* We describe eleven major mutation events that defined five major clades ($G_{614}$, $S_{84}$, $V_{251}$, $I_{378}$, and $D_{392}$) of globally-circulating viral populations. The clades can be specifically identified using an 11-nucleotide genetic barcode. An analysis of amino acid variation in SARS-CoV-2 proteins provided evidence of substitution events in the viral proteins involved in both host entry and genome replication.
*Conclusion:* Globally-circulating SARS-CoV-2 genomes could be classified into five major clades based on mutational profiles defined by an 11-nucleotide barcode. We have successfully developed a multiplexed sequencing-based, rapid genotyping protocol for high-throughput classification of major clade types of SARS-CoV-2 in clinical samples. This barcoding strategy will be required to monitor genetic diversity decrease as treatment and vaccine approaches become widely available.
© 2020 The Authors. Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## Introduction

The SARS-CoV-2 infection has now reached 188 countries across all continents, except Antarctica. A pandemic was declared by the World Health Organization (WHO) on March 11, 2020, and as of August 11, 2020, more than 20 million cases have been confirmed globally (Dong et al. 2020). SARS-CoV-2 belongs to the

* Corresponding author at: King Abdullah University of Science and Technology (KAUST), Pathogen Genomics Laboratory, Biological and Environmental Science and Engineering (BESE), Thuwal-Jeddah, 23955-6900, Saudi Arabia.
*E-mail address:* arnab.pain@kaust.edu.sa (A. Pain).
[1] S.M, S.H, and R.Na contributed equally to this work.

https://doi.org/10.1016/j.ijid.2020.08.052
1201-9712/© 2020 The Authors. Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

*Coronaviridae* family, genus *Betacoronavirus*, which are enveloped, positive-sense, single-stranded RNA viruses of zoonotic origin. Among human RNA viruses, coronaviruses have the largest known genome (~30 kb), which encodes the structural proteins (spike, envelope, membrane, and nucleocapsid), nonstructural proteins (nsp1–16), and accessory proteins (ORF3a, ORF6, ORF7a and b, ORF8, ORF10). Structural proteins are required for host cell entry, viral assembly and exit (Bárcena et al. 2009; Hoffmann et al. 2020). Nonstructural proteins are involved in genome replication-transcription and formation of vesicles (Hagemeijer et al. 2014), whereas accessory proteins interfere with host innate defense mechanisms (Wu et al. 2020a; Wu et al. 2020b; Liu et al. 2014). During genome replication within the host, the virus acquires genome mutations, which can be passed on to virus progeny in subsequently infected individuals.

More than 52,600 complete and high coverage genomes are available on GISAID as of August 11. Systematically tracking of mutations in SARS-CoV-2 genomes is, therefore, important as it allows monitoring of the molecular epidemiology of circulating viral sequences nationally and internationally. Here, we have investigated the nucleotide variation landscape of a large set of globally derived SARS-CoV-2 genomes and defined major mutation events. This analysis allowed us to produce a first-generation genetic classifier or 'barcode' defining the major clades of the virus circulating up to May 28, 2020. Notably, this barcode allowed reliable tracking of the spatial distribution and prevalence of these viral clades over time. While most of the nonsynonymous mutation events appear neutral with respect to protein function and stability, we also found evidence of mutations in the spike protein that may modulate interactions between SARS-CoV-2 and the human host.

## Materials and methods

### *Phylogenomic analysis*

1427 coronavirus genomes were downloaded from Virus Pathogen Database and Analysis Resource (ViPR)(Pickett et al. 2012) on February 14, 2020, including 329 SARS, 35 SARS-CoV-2, 61 HCoV-NL63, 521 MERS, 52 HCoV-HKU1, 170 HCoV−OC43, 97 bovine coronaviruses, and 61 mouse hepatitis viruses. Sequence alignment was performed using MAFFT (version 7.407) (Katoh and Standley 2013) software and then trimmed by trimAL (version 1.4.1) (Capella-Gutiérrez et al. 2009). Phylogenetic analyses of the complete genomes were done with FastTree (version 2.1.10) (Price et al. 2009) software with default parameters, and iTOL (version 5) (Letunic and Bork 2007) was used for phylogenetic tree visualization.

21,789 complete SARS-CoV2 genomes were downloaded from GISAID (Shu and McCauley 2017) (May 28, 2020). Sequence alignment was performed with MAFFT (version 7.407) (Katoh and Standley 2013) and trimmed by trimAL (version 1.4.1) (Capella-Gutiérrez et al. 2009). Phylogenetic analyses of the complete genomes were performed with RAxML (version 8.2.12) (Stamatakis 2014) with 10,000 bootstrap replicates, employing the GTRCAT model. A random set of genomes from each defined subclade was used to generate a simplified phylogenetic tree using the same method. iTOL (version 5) (Letunic and Bork 2007) was used for the phylogenetic tree visualization. Single nucleotide polymorphisms (SNPs) from each of the genomes were called by Parsnp (version 1.2) from the Harvest suite (Treangen et al. 2014) using MN908947.3 as the reference genome, and the SNPs were further annotated by SnpEff (version 4.3 m) (Cingolani et al. 2012). To compare the lineages recently defined by Rambaut and colleagues (Rambaut et al. 2020), Pangolin software (version 1.1.2) was applied to assign the assemblies belonging to each clade in our study.

### *Sequencing-based multiplexed genotyping*

Sequencing-based multiplexed genotyping primers (Table 1) were designed and tested on SARS-CoV-2-positive RNA samples collected from nasopharyngeal swabs of critical patients in Sulaiman Alhabib Hospital in Riyadh, Saudi Arabia. The swab was placed in 1 mL of TRIzol (Ambion, USA) and transported to King Abdullah University of Science and Technology (KAUST) for further downstream applications. RNA was extracted using the Direct-zol RNA Miniprep kit (Zymo Research, USA) following the manufacturer's instructions. RT-PCR was done using the one-step SuperScript III with Platinum Taq DNA Polymerase (Thermo Fisher, USA) on the 7900 H T Fast Real-Time PCR instrument (Applied Biosystems, USA). We chose 24 SARS-CoV-2 positive samples and generated cDNA using the VILO™ Superscript™ III kit (Thermo-Fisher, USA), and amplicons encompassing clade-specific SNPs were produced with the Q5 Hot Start High-Fidelity Master Mix (NEB, USA). The thermocycling conditions employed in this study were as follows: one cycle of 98 °C for 30 s followed by 35 cycles of 98 °C for 10 s, 64 °C for 30 s and 72 °C for 30 s, followed by the final extension at 72 °C for two minutes. The PCR products were electrophoresed on a 2% (w/v) agarose gel. The Illumina library preparation was performed using the ARTIC V3 PCR tiling protocol (https://artic.network/ncov-2019) with 20% PhiX control and a loading concentration of 8 pmol. The prepared libraries were sequenced on a MiSeq platform using a 600 cycle V3 kit with paired-end sequencing. To facilities the virus clade typing from the data generated by MiSeq, a Python script was developed using the mapped. bam file as the input. The barcode and clade information can be obtained using the script from the. vcf files generated by bcftools outputs (https://github.com/raeece/sarscov2barcode).

**Table 1**
Primer sets for targeted multiplex PCR.

| Clade | Position in the genome | Primer sequence | Amplicon size (bp) | Melting temperature $T_m$ (°C) |
|---|---|---|---|---|
| $G_{614}$ | 241 | GF-1: 5'- TGTCGTTGACAGGACACGAG-3' GR-1: 5'- TCCTCCACGGAGTCTCCAAA-3' | 228 | 60.94 59.33 |
| $G_{614}$ | 3037 | GF-2: 5'- ATGAGTTCGCCTGTGTTGTG-3' GR-2: 5'- TGTCTGATTGTCCTCACTGCC-3' | 392 | 58.77 60.00 |
| $G_{614}$ | 14,408 | GF-3: 5' - TGGGATCAGACATACCACCCA - 3' GR-3: 5' - GTGCAGCTACTGAAAAGCACG - 3' | 334 | 60.27 60.40 |
| $G_{614}$ | 23,403 | GF-4: 5' -CTGATGCTGTCCGTGATCCA - 3' GR-4: 5' - ACTAGCGCATATACCTGCACC - 3' | 302 | 59.82 60.00 |
| $S_{84}$ | 8782 | SF-1: 5'- GCGTCATATTAATGCGCAGGT-3' SR-1: 5'- GCAGCCAAAACACAAGCTGA-3' | 663 | 59.47 59.90 |
| $S_{84}$ | 28,144 | SF-2: 5' - CGTGGATGAGGCTGGTTCTA - 3' SR-2: 5' - CCCACTGCGTTCTCCATTCT - 3' | 300 | 59.18 60.04 |
| $V_{251}$ | 26,144 | VF-1: 5'- TCAGGTGATGGCACAACAAGT-3' VR-1: 5'- GTACGCACACAATCGAAGCG-3' | 468 | 60.13 60.25 |
| $I_{378}$ | 1397 | IF-1: 5' – GAAACTTCATGGCAGACGGG-3' IR-1: 5' – GCTAGCACGTGGAACCCAAT-3' | 303 | 59.20 60.68 |
| $I_{378}$ | 28,688 | IF-2: 5' - ACCGCTCTCACTCAACATGG - 3' IR-2: 5' - GCAGTACGTTTTTGCCGAGG - 3' | 632 | 60.04 60.11 |
| $D_{392}$ | 1440 | DF-1: 5' - AGGTGCCACTACTTGTGGTT - 3' DR-1: 5' - AGTTTCAAGAGTGCGGGAGA - 3' | 607 | 59.16 58.95 |
| $D_{392}$ | 2891 | DF-2: 5' - CGGTGCACCAACAAAGGTTAC - 3' DR-2: 5' - GCAGAAGTGGCACCAAATTCC - 3' | 450 | 60.27 60.34 |

We calculated the maximum throughput of samples for genotyping on the MiSeq run based on the following assumptions: (i) targeted sequencing depth of at least 100X, pair-end reads across ten positions per sample except the position 28,688 for which the targeted depth was 30X, (ii) MiSeq V3 kit generates 44–50 million reads and 70% of the total reads are higher than Q30 as specified by Illumina (https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html).

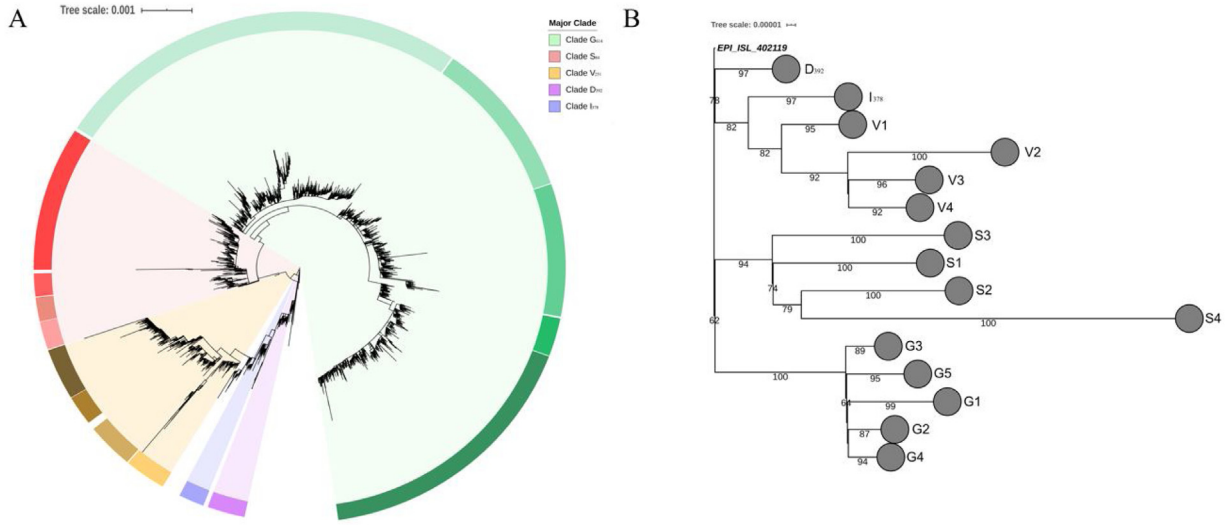
*Protein structure analysis*

Experimentally determined protein structures were obtained from the Protein Data Bank (PDB). SwissModel (Arnold et al. 2006), I-Tasser (Yang et al. 2014), RaptorX (Källberg et al. 2014), and an in-house modeling pipeline was used to produce protein structure homology models. Phobius (Käll et al. 2007) was used for the prediction of transmembrane regions. RaptorX was also used for predicting secondary structure, protein disorder, and solvent exposure of amino acids. Pymol (version 1.8.6.2) was used for visualization.

## Results

*Five clades of SARS-CoV-2 are characterized by 11 major mutational events across the globe*

The SARS-CoV-2 genome is genetically most closely related to horseshoe bat SARS-related coronaviruses (Lu et al. 2020; Zhu et al. 2020) (SARSr-CoV), 96% to RaTG13, 93% to RmYN02 (Zhou et al. 2020), and 88% to bat-SL-CoVZC45 and bat-SL-CoVZXC21 and also to SARSr-CoVs from pangolins (Tang et al. 2020) (Fig. S1). We studied the chronological occurrences of SARS-CoV-2 genomes during the early stages of the pandemic and global spread of the SARS-CoV-2 across human populations, which allowed us to define several clades of the virus that share unique SNPs. We observed 8427 SNPs in the current dataset (May 28, 2020), and we defined five major clades that contain 15 subclades (Figure1, Fig. S2) compared to the prototype (MN908947.3). The clades were named by their respective amino acid mutations: $S_{84}$ (Orf8, **L84S**), $V_{251}$ (Orf3a, **G251V**), $I_{378}$ (Orf1ab, **V378I**), $D_{392}$ (Orf1ab, **G392D**), and $G_{614}$




C

| Clade | Subclade | 241 | 3037 | 14408 | 23403 | 28881 | 28882 | 28883 | 25563 | 1059 | 20268 | 15324 | 25429 | 8782 | 28144 | 17747 | 17858 | 18060 | 24034 | 28077 | 3177 | 26144 | 11083 | 14805 | 17247 | 2480 | 2558 | 28077 | 1397 | 28688 | 1440 | 2891 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reference | | C | C | C | A | G | G | G | G | C | A | C | G | C | T | C | A | C | C | G | C | G | G | C | T | A | C | G | G | T | G | G |
| $G_{614}$ | G1 | T | T | T | G | A | A | C | | | | | | | | | | | | | | | | | | | | | | | | |
| | G2 | T | T | T | G | | | | T | T | | | | | | | | | | | | | | | | | | | | | | |
| | G3 | T | T | T | G | | | | | | G | | | | | | | | | | | | | | | | | | | | | |
| | G4 | T | T | T | G | | | | | | | T | | | | | | | | | | | | | | | | | | | | |
| | G5 | T | T | T | G | | | | | | | | T | | | | | | | | | | | | | | | | | | | |
| $S_{84}$ | S1 | | | | | | | | | | | | | T | C | | | | | | | | | | | | | | | | | |
| | S2 | | | | | | | | | | | | | T | C | | | | | | | | | T | | | | C | | | | |
| | S3 | | | | | | | | | | | | | T | C | T | G | T | | | | | | | | | | | | | | |
| | S4 | | | | | | | | | | | | | T | C | | | | T | C | T | | | | | | | | | | | |
| $V_{251}$ | V1 | | | | | | | | | | | | | | | | | | | | | T | | | | | | | | | | |
| | V2 | | | | | | | | | | | | | | | | | | | | | T | T | T | | | | | | | | |
| | V3 | | | | | | | | | | | | | | | | | | | | | T | T | T | C | | | | | | | |
| | V4 | | | | | | | | | | | | | | | | | | | | | T | T | T | | G | T | | | | | |
| $I_{378}$ | | | | | | | | | | | | | | | | | | | | | | | T | | | | | | A | C | | |
| $D_{392}$ | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | A | A |

D

| Clade | Subclade | Pangolin lineage | GISAID Clade |
|---|---|---|---|
| $G_{614}$ | G1 | B.1 | GR |
| | G2 | B.1 | GH |
| | G3 | B.1.5 | G |
| | G4 | B.1 | G |
| | G5 | B.1 | G |
| $S_{84}$ | S1 | A, A.3,A.5,A.6 | S |
| | S2 | A.2 | S |
| | S3 | A.1 | S |
| | S4 | A.1 | S |
| $V_{251}$ | V1 | B.2,B.7, B.10 | V |
| | V2 | B.2 | V |
| | V3 | B.2 | V |
| | V4 | B.2.1 | V |
| $I_{378}$ | | B.4 | O |
| $D_{392}$ | | B.3 | L |

**Figure 1.** Clades of SARS-CoV-2. (A) A global SNP-based radial phylogeny of SARS-CoV-2 genomes defining five major clades ($G_{614}$, $S_{84}$, $V_{251}$, $I_{378}$, and $D_{392}$) and several subclades based on nucleotide substitution events. (B) A simplified phylogenetic tree to illustrate the evolutionary relationship of the clades/subclades based on a random sampling of complete genomes from each subclade. (C) The clade and subclade-defining SNPs for each clade and subclade. *These SNPs are developed independently in more than one clade hence are not clade-defining SNPs (refer to Fig. S2). (D) A comparative guide to clades defined by our study and the lineages recently defined by Rambaut et al. (Rambaut et al. 2020) and GISAID (Shu and McCauley 2017).

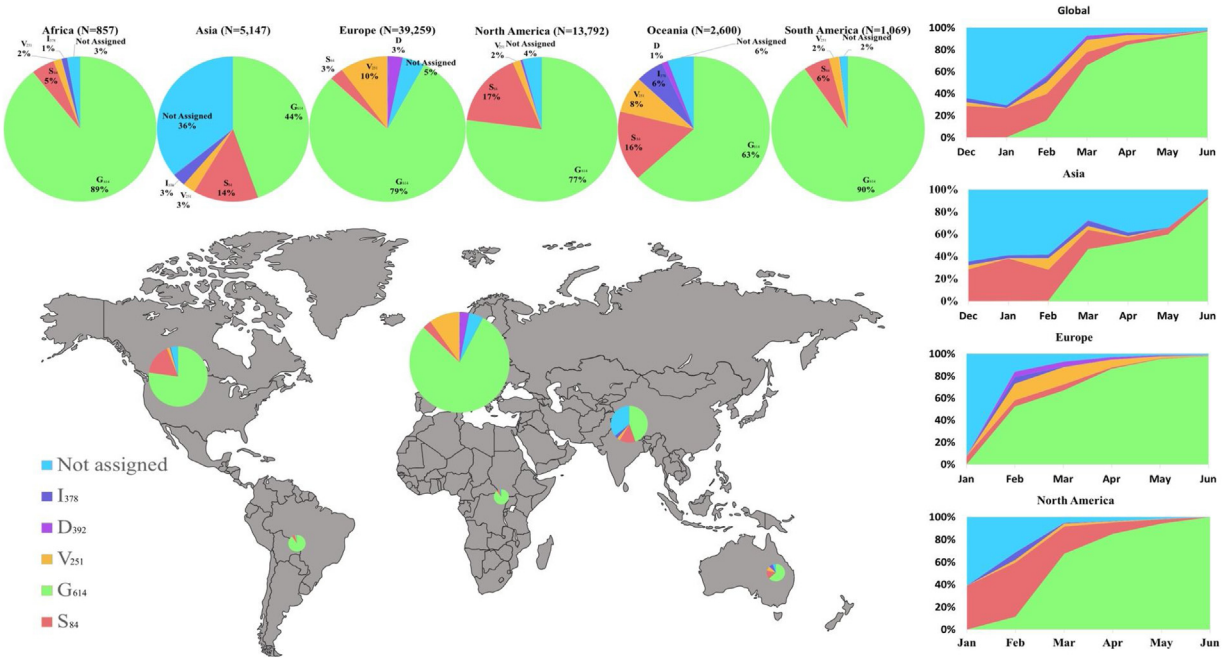



**Figure 2.** Global distribution of various major and minor clades of SARS-CoV2 genomes and their relative prevalence over six months from December 24, 2019, to June 30, 2020, from the outbreak and early stages of the pandemic. The size of each pie chart is proportional to the numbers within each respective clade. The cumulative trend of the clades is shown on the right, and the span of time indicates the first and last observed cases in each particular clade.

(S, **D614G**). A collapsed simplified phylogenetic tree to illustrate the subclades' correspondence is shown in Figure 1B along with the clade-defining mutations (Figure 1C). Recent work has defined (Rambaut et al. 2020) two viral lineages - A and B (analogous to clades L and S described in (Tang et al. 2020)). The major clades proposed here correspond to lineage A (Clade $S_{84}$), lineage B.2, B.7, and B.10 (Clade $V_{251}$), lineage B.1 (Clade $G_{614}$), lineage B.4 (Clade $I_{378}$), B.3 (Clade $D_{392}$) (Figure 1D). Due to intrinsic virus mutation rates, further refinements in nomenclature will likely need to be defined in the future; however, a robust, high throughput barcoding scheme will be required to breach the surveillance gap and lag time between molecular diagnostics and full-genome sequencing to provide actionable genetic surveillance data.

To obtain a global picture of the clades' regional distribution during the six months, December 24, 2019 to June 30, 2020, we plotted the relative proportions of the major clades and their cumulative trends (Figure 2). We observed major differences in the apparent spread for individual clades: Clade $G_{614}$ represents 71.14% of all the sequenced viral sequences, followed by $S_{84}$ (10.85%), $V_{251}$ (7.66%), $D_{392}$ (1.03%), and $I_{378}$ (1.70%). The remaining 7.6% were not assigned to a major clade. Clade $G_{614}$ is widely distributed globally; whereas Clade $S_{84}$ represents 17% of North American sampled genomes, and 16% of those from Oceania. Asia has the greatest number of unassigned genomes (36%). The global and regional cumulative trends were plotted over time, revealing the increasing dominance of few clades in each geographic region in the sampled genomes (Figure 2). This is likely attributable to founder effects during the early phases of the seeding of the local epidemics from imported cases and subsequent dissemination in the regions.

*Newly sequenced SARS-CoV-2 genomes could be assigned to clades using a rapid NGS-based genotyping protocol*

We have validated a set of oligonucleotide primers (Fig.S3) that can be utilized for rapid NGS sequencing-based genotyping of the major clades of SARS-CoV-2 using an Illumina MiSeq 600 V3 kit (Figure 3A). We have tested this protocol on 24 RNA samples from COVID-19 patients and generated the barcode information from the publicly accessible Python script (https://github.com/raeece/sarscov2barcode). The sequencing results show a high coverage of reads in the amplified regions across all 24 samples except for the position 28,688, which accompanies a G1397A mutation (Figure 3B). We observed the C241 T, C3037 T, C14408 T, and G23403A substitutions in all 24 tested samples, indicating that they all belong to Clade $G_{614}$.

Based on our calculation, roughly 78 K reads are needed for a given sample to obtain >800X genome coverage in ten clade-defining sites (Table S1), except position 28,688 (30X coverage). Based on the commercially-available multiplexing solutions available for Illumina library preparations, up to 384 samples can be processed with a MiSeq V3 600 Cycles kit in a single run to obtain reliable genotype information on the major clades. A higher number of samples could be genotyped with higher levels of multiplexing with custom-made index sequences for significant cost reduction in the future.

The robustness of our 11-site genomic barcoding method was verified by assigning ~93.68% of the 79,486 SARS-CoV-2 genomes that became available in GISAID until August 3, 2020, to one of the five major clades. This barcode represents a snapshot of the early phases of the genetic diversity of the virus during the first six months of its global spread and is expected to change over time; however, a barcoding strategy to monitor the progress of virus elimination after vaccines become widely available will be strategically useful to monitor decreases in viral genetic diversity over time. Our barcode could also serve as a reference for setting the baseline for global genomic diversity analysis at the beginning of the pandemic.

*Mutations are not equally distributed across the SARS-CoV-2 genome*

Based on our analysis, we observed that the genes Orf3a, Orf6, Orf7, Orf8, N, and Orf10 accumulated markedly more mutations than expected solely by random drift (Fig. S4) (e.g., Real/Expected ratio: N: 1.43; ORF3a: 1.52; E: 1.17). This mutation rate may indicate

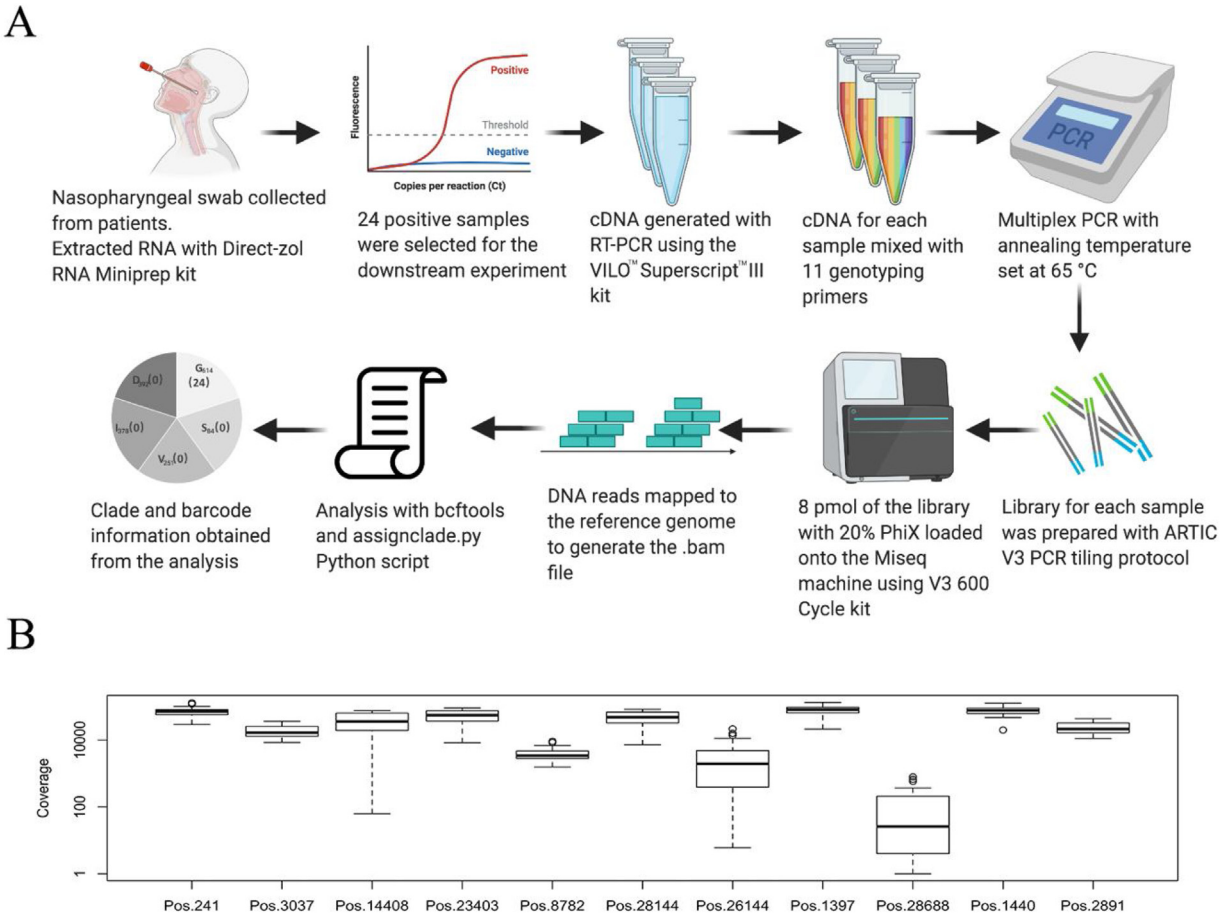



**Figure 3.** Workflow from clinical sample collection, next-generation sequencing, and SARS-CoV-2 clade assignment. (A) Schematic representation of the genotyping method described in this study. Positive samples were subjected to RNA extraction and multiplex RT-PCR. The amplicons were purified and prepared for the Illumina library. The sequencing was performed using the MiSeq 600 cycles V3 kit, and results were analyzed using our clade-defining script. **(B)** Boxplot of the coverage showing the log fold depth of the eleven clade-defining positions across the 24 SARS-CoV-2 genomes in multiplex sequencing-based genotyping. The primer sequences and PCR products of each pair of the primer are shown in Table 1 and Fig.S3.

adaptation to the human host following recent spillover from an, as yet, unknown animal reservoir. Conversely, several nonstructural proteins showed a lower-than-expected mutation rate (e.g., Real/Expected ratio: nsp5: 0.81; nsp10: 0.74). These proteins are predicted to be involved in evading host immune defenses, in enhancing viral expression, and in cleavage of the replicase polyprotein, based on prior studies of related beta coronaviruses (Báez-Santos et al. 2015; Posthuma et al. 2017). Hence, this lower mutation rate may indicate purifying selection to maintain these functions essential for efficient immune evasion and subsequent viral dissemination. Indeed, structural proteins in coronaviruses undergo a greater degree of antigenic variation, which increases the fitness of the virus, utilizing adaptation to the host and by facilitating immune escape (Walls et al. 2020).

Structural protein modeling confirmed that most of the nonsynonymous mutations in the nonstructural proteins were neutral (Figs. S5–S11). Conversely, several nonsynonymous mutations in the spike protein might have functional consequences; notably, the $G_{614}$ clade–defining mutation D614 G is located in subdomain one (SD1; Figure 4, Fig. S5). In the trimeric S, D614 engages stabilizing interactions within SD1 (R646 or the backbone of F592, depending on the chain) and with the S1 of the adjacent chain (T859 and K854). Replacement of D614 with a glycine would entail losing these stabilizing electrostatic interactions and increase the dynamics in this region. Recent studies have shown D614 G disrupts a critical interprotomer contact and that this dramatically shifts the S protein trimer conformation toward an ACE2-binding and fusion-competent state (Pascal et al. 2020). Hospitalized patients with G614 had higher viral titers in upper respiratory tract specimens, but no discernible differences in disease outcome (Korber et al., 2020). Notably, V483A, V367 F, and G476S are localized in the receptor-binding domain (RBD) of the spike protein, which mediates binding to the host receptor angiotensin-converting enzyme two (ACE2) (Figure 4) (Hoffmann et al. 2020). All of the viral genomes harboring V483A and G476S mutations belong to Clade $S_{84}$. Interestingly, the V367 F mutation has appeared independently in Clade $V_{251}$ and Clade $S_{84}$, suggesting that this mutation is under positive selection and contributes to viral fitness.

We found the V483A substitution has an equivalent amino acid substitution located in a similar position within the RBD in the MERS-CoV spike protein, which reduces its binding to its cognate receptor DPP4/CD26 (Kleine-Weber et al. 2018). However, V483 is more than 10 Å away from ACE2 and could affect receptor binding by SARS-CoV-2 only indirectly by altering the structural dynamics of the RBD loop it is a part of. The V367 F mutation is located at an even greater distance from ACE2 (Figure 4, Fig. S5). The exchange of the small hydrophobic residue valine with bulky hydrophobic phenylalanine might influence the efficiency of glycosylation of the nearby N343 or the positioning of the sugars. The substitution G476S would lead to possible clashes with predicted interacting ACE2 residues and with the RBD residue N487. However, minor reorientation of the side chains might allow an additional hydrogen bond to be formed between S476 and ACE2 Q24 and

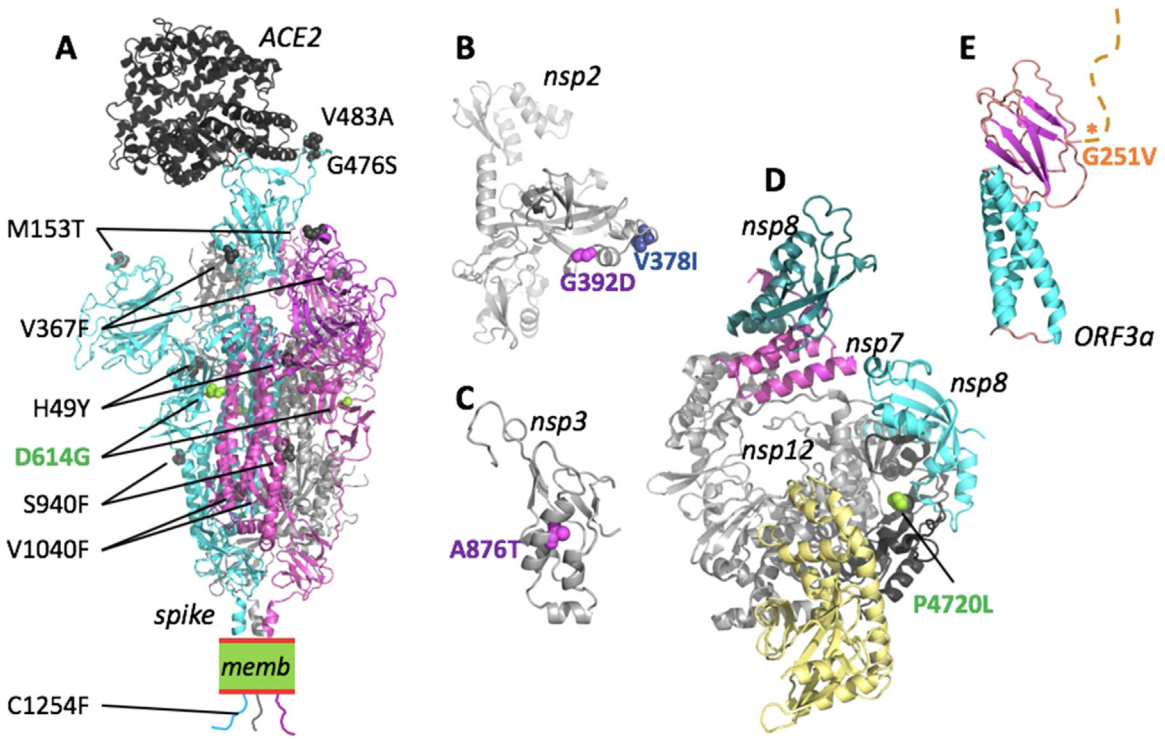



**Figure 4.** Mapping of SARS-CoV-2 clade-defining mutations onto the proteins. Nonsynonymous mutations for proteins where the 3D structure was experimentally determined (spike, nsp12/7/8) or can be inferred with reasonable confidence. Mutations are color-coded for the corresponding clades in Fig.3(D: magenta; G: light green; I: blue; V: orange). For a detailed analysis, see Fig.S5-11. (A) The structure of the SARS-CoV-2 spike trimer in its open conformation (chains are cyan, magenta, and grey) bound to the human receptor ACE2 (black) modeled based on PDB accessions 6m17 and 6vyb. Identified nonsynonymous mutations are shown as spheres in the model. For reasons of visibility, only mutations of two of the three spike chains are labeled. memb. indicates the plasma membrane. (B) Fragment comprising residues 180-534 of nsp2, modeled by AlphaFold35. Both clade-defining mutations are located in solvent-exposed regions and would not lead to steric clashes. (C) The substitution A876 T (corresponding to residue A58 in the nsp3 cleavage product numbering) is situated in the N-terminal ubiquitin-like domain of nsp3. The structure of this domain can be inferred based on the 79% identical structure of residues 1-112 from SARS-CoV (PDB id 2idy). The substitution A876 T can be accommodated with only minor structural adjustments and is not expected to substantially influence protein stability or function. (D) The structure shows the nsp12 in complex with nsp7 (magenta) and nsp8 (cyan and teal), based on PDB 7btf. P4720 (P323 in nsp12 numbering) is located in the 'interface domain' (black). In this position, the P323 L substitution is not predicted to disrupt the folding or protein interactions and hence is not expected to have strong effects. (E) A theoretical model for the Orf3a monomer has been proposed by AlphaFold36. The structure-function relationship of this protein remains to be clarified. The mutation G251 V is located C-terminal to the ®-sandwich domain and the tail (marked by an asterisk).

E23, thus enhancing the affinity to ACE2 (Fig.S5). An equivalent amino acid substitution located in an analogous position within the RBD in the SARS-CoV spike protein was associated with neutralization escape from monoclonal antibodies, together with other mutations observed previously (Rockx et al. 2010).

Given that V483A and V367F are solvent-exposed and markedly alter the surface characteristics of the RBD, they might also facilitate antibody escape. Indeed, V483A mutation was shown to give a selective advantage to the virus by escaping several neutralizing antibodies (Li et al., 2020). Escape mutations in the RBD of the SARS-CoV S protein (T332I, F460C, and L443R) were identified previously (Rockx et al. 2010). These mutations negatively impact viral fitness by reducing the affinity to the host receptor (Tang et al. 2014). The nonsynonymous mutations in the N protein, which have key roles in viral assembly, might also have functional implications. The hotspot mutations S202N, R203K, and G204R all cluster in a linker region where they might potentially enhance RNA binding and alter the response to serine phosphorylation events (Fig.S6). The clade $I_{378}$–defining mutation (T28,688C) in the nucleocapsid protein, which has key roles in viral assembly, is synonymous. However, we observed nonsynonymous mutations in the nucleocapsid protein that are predicted to have functional implications. The hotspot mutations S202N, R203K, and G204R all cluster in a linker region where they might potentially enhance RNA binding and alter the response to serine phosphorylation events (Fig. S6).

## Discussion

This study has defined five major clades ($G_{614}$, $S_{84}$, $V_{251}$, $I_{378}$, and $D_{392}$) of fully-sequenced SARS-CoV-2 genomes available until May 28, 2020, in the GISAID database. The clustering of these genomes revealed the spread of clades to diverse geographical regions (Figures 1, 2). This pattern contrasts with those observed for other epidemic coronaviruses, such as MERS-CoV, which display distinct geographical clustering (Kim et al. 2016). For example, clade $G_{614}$, which was first detected on January 28, is now dominant in the sampled genomes globally, and its percentage is increasing globally from Europe to North America, Oceania, and then to Asia (Korber et al. 2020). This pattern demonstrates efficient viral transmission through frequent intercontinental travel during the period when international travel restrictions were only present sporadically and which has enabled the virus to spread to multiple distant locations within a short period of time. This observation reinforces the importance of curtailing international travel and imposing restrictions early in pandemics and imposing social distancing to contain the global spread of viruses.

We have observed a distinct distribution of the major clades in different parts of the world (Figure 2). Most of the viral genomes that have not been assigned to a major clade are found in Asia and have earlier detection dates in January and February around the start of the epidemic in China. We observed a decrease in the genetic diversity of the virus over time following dissemination

from China, especially in Europe and North America that each notably now has clade $G_{614}$ as the predominant clade type, which is likely due to the increased fitness advantage that this mutation brings to the virus as was reported recently (Korber et al. 2020). An important caveat of the present study is that the current sampling of available public genomes likely does not represent the extant genetic diversity of virus populations in circulation. This is due to biases of genome data deposits from the sequencing laboratories based mainly in the Northern hemisphere, and new datasets may define new clades or subclades in the near future from other regions, including Africa, the Indian subcontinent, and Latin America from where there are comparably few genomes available at present. In this case, additional identifiers within an evolving barcode scheme can be added to track and monitor future emerging clades with higher resolution.

On the other hand, the genetic stability of SARS-CoV-2 (Jia et al. 2020) may result in the continuing circulation of a limited number of clades until such time as mitigation measures, including the isolation of vulnerable populations and the availability of efficacious antivirals and vaccines might reduce the genetic diversity in circulation. This molecular genotyping approach has been demonstrated for other viruses (e.g., measles, poliovirus, rotavirus, and human papillomaviruses) with herd-immunity vaccination programs working to eliminate pathogens from endemic circulation in humans (Brown et al., 2019; Grassly, 2013; Mankertz et al., 2011; Soares-Weiser et al., 2012). The availability of a barcoding scheme that rapidly generates SNPs allowing clade assignment will be critical in this elimination phase when widespread availability of vaccines permits eradication of SARS-CoV-2 from endemicity in humans.

In this study, we have designed and tested the first iteration of a multiplexed sequencing-based SARS-CoV-2 genotyping protocol for major clade assignment (Figure 3). This will facilitate genetic surveillance to be performed at a high-throughput level from SARS-CoV-2 detection to clade assignment. Future versions of the multiplexed primer sets can be optimized to cover the relatively low coverage regions (position 26,144 and 28,688). We concluded that the relatively low coverage at these two positions was a result of the position of mutation relative to the amplicon sites and not related to the PCR conditions used (Fig.S12). Our approach is also compatible with genotyping by other platforms, such as the MinION nanopore sequencer from Oxford Nanopore Technologies (ONT). Different sets of primers can also be designed to meet the requirements of different sequencing platforms. For example, the amplicon sizes could be optimized for the Illumina iSeq100 for rapid genotyping of a large number of samples without compromising the robustness of the genotyping calls at a fraction of current costs using a MiSeq. Moving forward, further studies may also include subclade-defining sites to have a higher resolution of each sequenced sample to assist in tracking and surveillance of the virus. The U.S. Food and Drug Administration (USFDA) has authorized the Illumina COVIDSeq test, which is intended for the detection of SARS-CoV-2 virus RNA for research use (https://www.illumina.com/products/by-type/ivd-products/covidseq.html). According to the COVIDseq instructions, up to 384 results per lane can be processed on the NovaSeq 6000 System for whole-genome sequencing. Our methodology is complementary to the COVIDSeq test and aimed at revealing the major clades of the virus in a given sample.

Our work provides a baseline genomic epidemiology of SARS-CoV-2 prior to the introduction of therapeutic and prophylactic approaches. The mutational landscape of global populations of over 21,789 SARS-CoV-2 genomes provides an evidence-based framework for tracking the clades that comprise the pandemic on different continents. However, due to the biases in the representation of countries depositing SARS-CoV-2 genomes, with over-representation of North American and European genomes, the available genome data represents only a minute proportion of the total COVID-19 positive cases from each of these regions. Therefore, this may require that the genetic barcode described here may need to be updated to be globally representative, once sufficient numbers of genomes covering less represented parts of the world are eventually sequenced and deposited in publicly-available databases. As viral genetic diversity is anticipated to decrease with the advent of widespread vaccination, the barcoding approach will allow more rapid discrimination of autochthonous and imported cases to monitor interruption of viral transmission.

## Conclusion

This study provides a baseline reference of genomic diversity at the early stages of the COVID-19 pandemic and will prove useful for monitoring changes in circulating clades of SARS-CoV-2 in different geographic regions over time. An 11-nucleotide genetic barcode of SARS-CoV-2 is presented, which identifies the five major clades of circulating viral genomes. The robustness of our 11-site genomic barcoding approach was validated by correctly assigning ~94% of the 79,486 SARS-CoV-2 genomes available in GISAID until August 3, 2020, to one of the five major clades. We have designed the first version of a high-throughput and robust genotyping protocol that can be readily applied for SARS-CoV-2 detection and clade assignment. This barcoding strategy will be important to target genomic sequencing efforts and monitor decreases in viral genetic diversity as intervention approaches become widely available.

## Conflict of interest

The authors declare no conflict of interest.

## Funding

Work in AP's laboratory is supported by the KAUSTfaculty baseline fund (BAS/1/1020-01- 01) and research grants from the Office for Sponsored Research (OSR-2015-CRG4−2610, OCRF-2014-CRG3−2267). This work was also supported by funding from King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR), under award number FCC/1/1976−25-01 to STA.

## Ethical approval

This study is covered by IRB approval 20IBEC14 and HAP-01-R-082 to work on the SARS-CoV-2 positive RNA samples derived from COVID-19 positive patients at the Dr.Suliman Al-Habib Medical group in KSA.

## Author contributions

A.P. conceived the study and supervised the work; A.P., W.W.H., S.T.A., and Q.G. designed the research; Q.G., M.S., R.Na and S.T.A. analyzed data; Q.G., M.S., S.T.A., and M.J.C. wrote the initial draft of the manuscript, followed by edits from A.P., M.J.C., W.W.H and R. Nu; R.Nu. and S.M. contributed new reagents; Q.G., S.H., S.M., R.Nu., A.A.O., S.S, and A.A.M provided samples, designed and performed the genotyping experiment. All authors have commented and edited various sections of the draft manuscript.

## Acknowledgments

We are deeply grateful to all laboratories contributing genomic data and metadata to GISAID and nextstrain.org databases. We

thank the KAUST Rapid Research Response Team (R3T) for supporting our research. We thank Olga Douvropoulou for her support during the research. We also thank Richard Culleton (Nagasaki University, Japan) and Gabo Gonzalez (UCD, Ireland) for their critical comments on the draft manuscript.

## Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijid.2020.08.052.

## References

Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. Bioinformatics 2006;.
Báez-Santos YM, St. John SE, Mesecar AD. The SARS-coronavirus papain-like protease: Structure, function and inhibition by designed antiviral compounds. Antiviral Research 2015;.
Bárcena M, Oostergetel GT, Bartelink W, Faas FGA, Verkleij A, Rottier PJM, et al. Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. Proc Natl Acad Sci U S A 2009;.
Brown KE, Rota PA, Goodson JL, Williams D, Abernathy E, Takeda M, et al. Genetic characterization of measles and rubella viruses detected through global measles and rubella elimination surveillance, 2016-2018. Morb Mortal Wkly Rep 2019;.
Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25(15):1972–3.
Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;.
Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis [Internet] 2020;3099(20)19–20, doi:http://dx.doi.org/10.1016/S1473-3099(20)30120-1 Available from:.
Grassly NC. The final stages of the global eradication of poliomyelitis. Philos Trans R Soc B: Biol Sci 2013;.
Hagemeijer MC, Monastyrska I, Griffith J, van der Sluijs P, Voortman J, van Bergen en Henegouwen PM, et al. Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. Virology 2014;.
Hoffmann M, Kleine-Weber H, Schroeder S, Mü MA, Drosten C, Pö S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor Article SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell [Internet] 2020;181:1–10, doi:http://dx.doi.org/10.1016/j.cell.2020.02.052 Available from:.
Jia Y, Shen G, Zhang Y, Huang K, Ho H, Hor W, et al. Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity. bioRxiv 2020;.
Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res 2007;35(SUPPL.2).
Källberg M, Margaryan G, Wang S, Ma J, Xu J. Raptorx server: A resource for template-based protein structure modeling. Methods Mol Biol 2014;.
Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 2013;.
Kim Il YJ, Kim YJ, Lemey P, Lee I, Park S, Bae JY, et al. The recent ancestry of Middle East respiratory syndrome coronavirus in Korea has been shaped by recombination. Sci Rep 2016;.
Kleine-Weber H, Elzayat MT, Wang L, Graham BS, Müller MA, Drosten C, et al. Mutations in the Spike Protein of Middle East Respiratory Syndrome Coronavirus Transmitted in Korea Increase Resistance to Antibody-Mediated Neutralization. J Virol 2018;.
Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 2020;.
Letunic I, Bork P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics 2007;.
Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 2020;.
Liu DX, Fung TS, Chong KKL, Shukla A, Hilgenfeld R. Accessory proteins of SARS-CoV and other coronaviruses. Antiviral Research 2014;.
Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;.
Mankertz A, Mihneva ZR, Gold H, Baumgarte S, Baillot A, Helble R, et al. Spread of measles virus D4-Hamburg, Europe, 2008-2011. Emerg Infect Dis 2011;.
Pascal KE, Veinotte K, Egri SB, Schaffner SF, Jacob E. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. bioRxiv 2020;.
Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: An open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 2012;.
Posthuma CC, te Velthuis AJW, Snijder EJ. Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes. Virus Research 2017;.
Price MN, Dehal PS, Arkin AP. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009;.
Rambaut A, Holmes EC, Toole ÁO, Hill V, Mccrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;.
Rockx B, Donaldson E, Frieman M, Sheahan T, Corti D, Lanzavecchia A, et al. Escape from Human Monoclonal Antibody Neutralization Affects In Vitro and In Vivo Fitness of Severe Acute Respiratory Syndrome Coronavirus. J Infect Dis 2010;.
Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance 2017;.
Soares-Weiser K, MacLehose H, Bergman H, Ben-Aharon I, Nagpal S, Goldberg E, et al. Vaccines for preventing rotavirus diarrhoea: vaccines in use. Cochrane Database of Systematic Reviews. .
Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30(9):1312–3.
Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 2020;6(3).
Tang XC, Agnihothram SS, Jiao Y, Stanhope J, Graham RL, Peterson EC, et al. Identification of human neutralizing antibodies against MERS-CoV and their role in virus adaptive evolution. Proc Natl Acad Sci U S A 2014;.
Treangen TJ, Ondov BD, Koren S, Phillippy AM. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 2014;15(11).
Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020;.
Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe 2020a;.
Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020b;.
Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: Protein structure and function prediction. Nature Methods 2014;.
Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Curr Biol 2020;.
Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;.