replicating enzyme 3CL[Pro] targeting repurposed drug candidates

Nitin Chitranshi[1*], Vivek K. Gupta[1*], Rashi Rajput[1], Angela Godinez[1], Kanishka Pushpitha[1], Ting Shen[1], Mehdi Mirzaei[3,4], Yuyi You[1], Devaraj Basavarajappa[1], Veer Gupta[2], Stuart L. Graham[1,5]

[1] Faculty of Medicine, Health and Human Sciences, Macquarie University, F10A, 2 Technology Place, North Ryde, NSW 2109, Australia

[2] School of Medicine, Deakin University, Melbourne, VIC, Australia.

[3] Department of Molecular Science, Macquarie University, North Ryde, NSW 2109, Australia

[4]Australian Proteome Analysis Facility, Macquarie University, North Ryde, NSW 2109, Australia

[5] Save Sight Institute, Sydney University, Sydney NSW 2000, Australia

Corresponding Author: Nitin Chitranshi, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, 2109, Australia, nitin.chitranshi@mq.edu.au

Vivek Gupta, Faculty of Medicine and Health Sciences, Macquarie University, Sydney, 2109, Australia, vivek.gupta@mq.edu.au

Abstract

Background: Severe acute respiratory syndrome (SARS) has been initiating pandemics since the beginning of this century. In December 2019, the world was hit again by a devastating SARS episode that has so far infected almost four million individuals worldwide with over 200,000 fatalities having already occurred by mid-April 2020, and the infection rate continues to grow exponentially. SARS coronavirus 2 (SARS-CoV-2) is a single stranded RNA pathogen which is characterised by a high mutation rate. It is vital to explore mutagenic capability of the viral genome that enables SARS-CoV-2 to rapidly jump from one host immunity to another and adapt to genetic pool of the local populations.

Methods: For this study, we have analysed 1,921 complete viral sequences reported from SARS-CoV-2 infected patients. SARS-CoV-2 host genomes were collected from The Global Initiative on Sharing All Influenza Data (GISAID) database containing 9 genomes from pangolin-CoV origin and 3 genomes from bat-CoV origin, Wuhan SARS-CoV2 reference genome was collected from GeneBank database. The Multiple sequence alignment tool, Clustal Omega was used for genomic sequence alignment. The viral replicating enzyme, 3-chymotrypsin-like cysteine protease ($3CL^{pro}$) that plays a key role in its pathogenicity was used to assess its affinity with pharmacological inhibitors and repurposed drugs such as anti-viral flavones, biflavanoids, anti-malarial drugs and vitamin supplements.

Results: Our results demonstrate that bat-CoV shares >96% similar identity, while pangolin-CoV shares 85.98% identity with Wuhan SARS-CoV-2 genome. This in-depth analysis has identified 12 novel recurrent mutations in South American and African viral genomes out of which 3 were unique in South America, 4 unique in Africa and 5 were present in-patient isolates from both populations. The study further investigated the interaction of repurposed drugs with SARS-CoV-2 $3CL^{pro}$ enzyme which regulates viral replication machinery, using state of the art *in silico* approaches.

Conclusions: Overall, this study provides insights into the evolving mutations with implications to understand viral pathogenicity and possible new strategies for repurposing compounds to combat nCovid-19 pandemic.

Background

In early January 2020, the World Health Organisation (WHO) reported cases of pneumonia of unknown cause in Wuhan City, Hubei Province of China and by 30 January 2020, WHO escalated the warning to public health emergency of international concern. By 12 March 2020, novel coronavirus (nCoV) outbreak achieved a global pandemic status and was recognised as novel Covid-19 disease (nCovid-19) [1]. The present coronavirus outbreak is associated with severe acute respiratory syndrome 2 (SARS-CoV-2), phylogeny and taxonomy designated [2]. Worldometer reported total SARS-CoV-2 infected cases on 28 April 2020, 3,136,505 and deaths 233,824 worldwide (https://www.worldometers.info/coronavirus/#countries). The pathogen has been established to transmit from human to human contact and has quickly spread to more than 187 countries across the globe (https://gisanddata.maps.arcgis.com/).

Coronaviruses are single and positive stranded RNA viruses belonging to the genus *Coronavirus* of the family Coronaviridae that can cause acute and chronic respiratory and central nervous system illnesses in animals, including in humans [3, 4]. The infection can also cause mild episodes of follicular conjunctivitis in certain patients. In animal models, the infection has been shown to induce anterior uveitis, retinitis, and optic neuritis like symptoms [5]. The disease is also shown to affect sense of smell and taste bud sensitivity in patients [6].

All coronaviruses have minimal 3 basic viral proteins (i) an envelope protein (E), which is a highly hydrophobic protein involved in several aspects of the virus life cycle such as assembly and envelope formation [7] (ii) a spike protein (S), a glycoprotein involved in receptor recognition and membrane fusion [8] and (iii) a membrane protein (M), which plays a key role in virion assembly [9] (Fig. 1). The viral genome also encodes two open reading frames (ORF), ORFa and ORFb that activate intracellular pathways and triggers the host innate immune response [10]. The polyprotein encoded by the virus are initially processed by two main viral proteases, which include a papain-like cysteine protease (PL^{pro}) and chymotrypsin-like cysteine protease known as 3C-like protease ($3CL^{pro}$), into intermediate and mature non-structural proteins [11].

$3CL^{pro}$, the main proteinase is one of the primary targets for development of antiviral drug therapies. It plays a critical role in the viral replication [12]. K11777, camostat and EST, is a cysteine protease inhibitor, has been shown to inhibit SARS-CoV $3CL^{pro}$ replication in cell culture conditions [13, 14]. Recent release of the high-resolution crystal structure of main proteinase $3CL^{pro}$ (Protein Data Bank, PDB ID: 6Y2G) describing additional amide bond with a α-ketoamide inhibitor pyridone ring to enhance the half-life of the compound in plasma [15]

is suggested to accelerate the targeted drug discovery efforts. Two HIV-1 proteinase inhibitors, lopinavir and ritonavir, have been considered to target SARS-CoV [16]. Interestingly, the substrate binding cleft is located between domains I and II of both SARS-CoV $3CL^{pro}$ and SARS-CoV-2 $3CL^{pro}$ enzymes [15, 17].

Since the initial stages of the SARS-CoV-2 outbreak, laboratories and hospitals around the world have sequenced viral genome data with unprecedented speed, enabling real-time understanding of this novel disease process and which will hopefully contribute to the development of novel candidate drugs. The complete genomes of SARS-Cov-2 from all over the world have been deposited at The Global Initiative on Sharing Avian Influenza Data (GISAID) [18] database and more sequences continue to be deposited with the passage of time. Development of a novel vaccine against SARS-CoV-2 so far remains elusive and requires a thorough understanding of molecular changes in viral genetics. This may be attained by freely accessing GISAID database and processing the data to enhance our understanding of the fine biochemical and genetic differences that differentiate this virus from the previously known strains [19].

It is well known that viruses are non-living and that they require host cells to survive and to reproduce, with the sole aim to perpetuate themselves. When a virus jumps from animal to humans, it is termed a zoonotic virus. That happened with SARS in 2002, when a new coronavirus spread around the world and resulted in death of hundreds of people [20]. In 2012, Middle East respiratory syndrome (MERS), a novel coronavirus outbreak caused over 400 fatalities spread over 20 different countries [21]. There are many circulating viruses but why SARS-CoV-2 has achieved such a devastating pandemic status and whether this pandemic will subside remain unanswered.

The purpose of this study is to characterise known viral variants that have spread across different countries, especially hot-spot regions, with a focus on recurrent mutations in South American and African geographical regions. We also focused on SARS-CoV-2 main proteinase, $3CL^{pro}$ which is highly conserved in most of the coronaviruses and has been suggested to be a potential drug target to fight against nCovid-19. Repurposed drugs like flavonoids and biflavanoids, known anti-malarial and anti-viral drugs and inhibitory effects of vitamins could selectively inhibit this enzyme and can be used either alone or in combination with other disease management approaches to suppress the virulence of SARS-CoV-2. These bioinformatics, computational modelling and molecular docking approaches using repurposed drugs could be particularly useful in the current nCovid-19 outbreak.

Methods

Collection of SARS-Cov-2 genome

The Global Initiative on Sharing Avian Influenza Data (GISAID) is headquartered in Munich, Germany and is a public-private partnership project between German government and the non-profit organization founded by leading medical researchers in 2006. Since December 2019, GISAID has become a repository storage database for nCovid-19 genome. The genome analysis was carried out for data deposited up to 15 April 2020 (https://www.gisaid.org/). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) Wuhan genome was collected from NCBI, NC_045512.2.

Multiple sequence alignment and Phylogenetic tree construction

Multiple sequence alignment (MSA) of all nucleotide sequences was carried out in the EMBL-EBI Clustal Omega server to investigate sequence conservation [22]. The Newick format for the multiple align sequence was used to generate phylogeny [23]. The phylogenetic tree was constructed in the Interactive Tree of Life (iTOL) online tool [24]. The iTOL server generate phylogeny trees in a circular (radial) and normal standard trees. The circular trees can be rooted and displayed in different arc sizes [25-27].

Structure analysis SARS and SARS-CoV-2 $3CL^{Pro}$

Crystal structure of SARS and SARS-CoV-2 $3CL^{Pro}$ with bound inhibitors were collected from protein data bank (PDB) [28]. PDB ID: 3TNT, SARS main protease was selected as reference to analyse the variants in SARS-CoV-2 $3CL^{Pro}$ (PDB ID: 6Y2G). All the PDBs were visualised using UCSF Chimera software [29]. Multiple alignment, ribbon, surface and superimposition module in Chimera software were used for analysis and image generation [30, 31].

Computer aided molecular modelling

Collection and preparation of SARS-CoV-2 protease inhibitors

The dataset comprises of flavones and biflavanoids, anti-viral, anti-malarial and vitamins as SARS-CoV-2 $3CL^{pro}$ inhibitor [15]. In total 17 repurposed drugs were collected from Pubchem database [32]. Two-dimensional (2D) structure were downloaded from Pubchem database in .sdf format. The inhibitors energy were minimized using the Austin Model-1 (AM1) until the root mean square (RMS) gradient value become smaller than 0.100 kcal/mol Å and later re-optimization was done by MOPAC (Molecular Orbital Package) method [33, 34]. Later, all

the inhibitors were converted to .pdb format in Open Babel software [35] submitting to docking studies.

Selection and preparation of SARS-Cov-2 main protease protein (3CL ^{pro})

Crystal structure of the SARS-CoV-2 3CL^{pro} was retrieved from PDB (PDB ID: 6Y2G). The protein macromolecule (SARS-CoV-2 3CL^{pro}) optimization was carried out in UCSF Chimera software [29, 36] by adding polar hydrogen atoms, removing water molecules, implying amber parameters, followed by minimization with MMTK method in 500 steps with a step size of 0.02 Å. SARS-CoV-2 3CL^{pro} contains chain A and B of 306 amino acids sequence length. Chain A of PDB ID: 6Y2G containing alpha-ketoamide (O6K) inhibitor was used for identification of substrate binding site.

SARS-Cov-2 main protease protein (3CL ^{pro}) inhibitors docking studies

The docking of SARS-CoV-2 3CL^{pro} specific pharmacological inhibitors into the catalytic site was performed by AutoDock 4.2 program [37]. Alpha-ketoamide (O6K) inhibitor was extracted from the SARS-CoV-2 3CL^{pro} protein. The polar hydrogen atoms were added, the non-polar hydrogen atoms were merged, Gasteiger charges were assigned and solvation parameters were added to the protease, SARS-CoV-2 3CL^{pro}. The protonation state for all inhibitors and O6K were set to physiological pH and rotatable bonds of the ligands were set to be free. AutoGrid program was used to generate grid maps. Cys145 residue in SARS-CoV-2 3CL^{pro} protein was selected and the grid box dimensions of 40 × 40 × 40 Å was formed around the Cys145 protease residue which is present in the substrate binding site. Protein rigid docking was performed using the empirical free energy function together with the Lamarckian genetic algorithm (LGA) [38]. LGA default parameters were used in each docking procedure and 10 different docking poses were calculated. Chimera and Discovery Studio (DS) Visualizer2.5 [29] software were used for visualization and calculation of protein–ligand interactions.

Results

Distribution analysis of SARS-CoV-2 in different geographic regions

A total of 9,174 SARS-CoV-2 genomes were retrieved from The GISAID database (https://www.gisaid.org/) that contain 3 sequences from bat (*Betacoronavirus*) and 9 sequences from Malayan Pangolin (*Manis javanica*) (Data Supplementary Table 1). Out of 9,174 genome sequences, 1,921 complete genome sequences of SARS-CoV-2 were selected randomly, aligned and compared with Wuhan SARS-CoV-2 (NC_045512.2) reference genome. We have divided our dataset into 6 different geographic areas: Europe (20.31%), North America (21.13%), Asia (35.37%), Oceania (20.86%), South America (2.87%) and Africa (3.52%). The European group comprises of SARS-CoV-2 infected patient data from the following: Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, , Slovakia, Slovenia, Spain, Sweden, Switzerland, and United Kingdom. The North American group contains genomes from the United States and Canada. The Asian group comprises genomes obtained from patients located in China, Indonesia, Pakistan, Philippines, Taiwan, Turkey, Kuwait, Georgia, South Korea, Japan, Iran, India, Thailand, Hong Kong, Malaysia, Singapore, Vietnam. The Oceanian group comprises genomes from Australia and New Zealand. South America includes Brazil, Peru, Chile, Colombia, Argentina, Ecuador (Fig. 2A – C).

Sequences from bat-SARS-CoV and Pangolin-SARS-CoV were aligned and compared to the Wuhan SARS-CoV-2 (NC_045512.2) as a reference genome. To determine the evolutionary relationship among bat-CoV, Pangolin-CoV and SARS-CoV-2, we estimated a phylogenetic tree based on the nucleotide sequences of the whole-genome sequence. Bat-SARS-CoV and SARS-CoV-2 were grouped together and were observed to share >96% similarity, whereas the Pangolin-SARS-CoV was closest evolutionary ancestor (Fig. 2D). Isolate of human Wuhan SARS-CoV-2 (NC_045512.2) shared 85.98% identity with Pangolin-SARS-CoV which suggests that Pangolin may be associated with SARS-CoV-2 evolution or subsequent outbreak [39, 40].

Identification of hotspot mutations in SARS-Cov-2 complete genome from South American and African regions and analysis of main protease (3CL ^{Pro}) sequence

Recently, Pachetti et al [41] has reported eight novel recurrent mutations of SARS-Cov-2 that have been identified in positions 1397, 2891, 14408, 17746, 17857, 18060, 23403 and 28881

in Asian, Oceanic, European and North American outbreaks. However, SARS-CoV-2 mutations from South American and African patient isolates are not yet reported. We confirmed the occurrence of these mutations in South Americans and Africans located at positions 3036, 8782, 11083,14408, 23403, 28144 and 28881 as reported in previous literature [41]. Our study highlights the presence of additional "conserved mutations" in the South American and African communities, taking into account only those occurring ≥ 2 times in our database. We report here 12 new mutations that have evolved in the SARS-Cov2 sequence in South American and African populations. These are located at positions 14805, 25563, 26144, 28882, 28883, 9477, 28657, 28863, 1059, 15324, 28878 and 29742 sites. The high tendency of the virus to demonstrate genetic variability is evident from the fact that even within these variants, three variations 9477 (nsp4), 28657 and 28863 (ORF9, structural protein) were uniquely identified in isolates from South American patients while four novel mutations viz. 1059 (nsp2), 15324 (RdRp), 28878 (ORF9, structural protein), and 29742 (stem-loop II-like motif) were detected only in isolates from African patient samples (Figure 3B). Interestingly, some mutations were identified to be common between these two separate sets of sequences that have been reported from the two distinct geographical locations viz. 14805, 25563, 26144, 28882 and 28883, belonging to gene ORF1ab (14805 RNA-dependent RNA polymerase (RdRp), ORF3a (25563 and 26144 ORF3a protein) and ORF9, N gene (28882 and 28883 nucleocapsid phosphoprotein) sequences, respectively (Figure 3A). An interesting finding of this analysis is the concurrence of 14805 mutation with 14808 mutation in the same locus. This double point mutation was observed in RdRp genome from isolates of both South American and African patients. In contrast, 28882/ 28883 mutation locus corresponded with another previously reported mutation 28881, and this triple point mutation was also present in both the South American and African genomic sequences. Identification of point mutations at the same locus indicates the high susceptibility of these genetic regions to change as the virus evolves.

For its actions, single-stranded SARS-CoV-2 RNA viral genome encodes two protease polyproteins (i) papain-like cysteine-protease (PL^{pro}) and (ii) chymotrypsin-like cysteine protease known as 3C-like protease ($3CL^{pro}$). $3CL^{pro}$ is a main protease and therefore it is important to examine the incidence of any mutation in the SARS-CoV-2 $3CL^{Pro}$. Multiple sequence alignment of SARS-CoV-2 genome collected from patients from six different geographical locations exhibited 100% similarity and no discernible variations in sequences obtained from diverse geographical regions, for this enzyme.

SARS-CoV and SARS-CoV-2 similarity

SARS and SARS-CoV-2 complete genomes were collected from NCBI, GenBank (NC_004718 and NC_045512). Protease nucleotide sequences were extracted from SARS (NC_004718) and were aligned with SARS-CoV-2 (NC_045512). Clustal Omega alignment of 918 SARS nucleotides showed around 95% similarity with SARS-CoV-2 (Supplementary Table 2). Higher amino acid sequence identity was also observed in SARS-CoV and SARS-CoV-2 main protease ($3CL^{Pro}$) derived from Wuhan and US patients. SARS-CoV and SARS-CoV-2 $3CL^{Pro}$ showed highly conserved region in both the catalytic sites, His41 and Cys145 [42] and substrate binding region of the enzyme (163-167 and 187-192) [43] (Fig. 4A), inferring that these proteases exhibit high similarities. Furthermore, 12 variant positions (Thr35Val, Ala46Ser, Ser65Asn, Leu86Val, Arg88Lys, Ser94Ala, His134Phe, Lys180Asn, Leu202Val, Ala267Ser, Thr285Ala and Ile286Leu) were observed in SARS-CoV-2 $3CL^{Pro}$ (Fig. 4B, C). The effects of mutations and potential resultant amino acids on SARS-CoV-2 $3CL^{Pro}$ structure are expected to conserve the polarity and hydrophobicity, except when the resulting amino acid is Leucine at 286 position. However, it is important to mention that these 12 variants are not present in catalytic and substrate binding regions which are involved in critical proteolytic activity of the molecule.

Docking study of SARS-CoV-2 $3CL^{Pro}$ inhibitors

SARS-CoV-2 $3CL^{Pro}$ receptor binding pocket was determined by superimposing SARS and SARS-CoV-2 $3CL^{Pro}$ with their respective inhibitors (Fig. 4). Interestingly, Needleman-Wunsch alignment algorithm and BLOSUM-62 matrix analysis revealed 94.44 % sequence identity between SARS (Fig. 5A, grey) and SARS-CoV-2 $3CL^{Pro}$ (Fig. 5A, Cyan). Cys-His catalytic dyad (Cys145 and His41) comprises the active catalytic binding site in SARS-CoV-2 $3CL^{Pro}$ (Fig. 5A' and B) and indicated the strong possibility that intended pharmacological inhibitors of SARS-CoV-2 $3CL^{Pro}$ may also suppress the activity of SARS-CoV-2 $3CL^{Pro}$ viral enzyme. Docking protocol of Autodock 4.2 program was optimized by extracting and re-docking of alpha-ketoamide inhibitor named O6K in the binding pocket of SARS-CoV-2 $3CL^{Pro}$. Lowest binding energy -6.45 kcal/mol and 18.72 µM inhibitory constant (K*i*) was predicted for alpha-ketoamide inhibitor (shown in Table 1). Re-docking of O6K inhibitor occupied the similar docking pose in the SARS-CoV-2 $3CL^{Pro}$ catalytic dyad active site as previously reported in the crystal structure (PDB ID: 6Y2G) (Fig. 5C, D).

Seven flavonoids and biflavonoid, three anti-malarial compounds, seven anti-viral drugs and three vitamin molecules were subjected to automated docking with the active site of SARS-CoV-2 $3CL^{Pro}$ catalytic-dyad. The superimposition of all docked flavones and biflavones (Fig. 6a), anti-malarial drugs (Fig. 6b), anti-viral drugs (Fig. 6c) and vitamins (Fig. 6d) is shown in Fig. 6 and various binding parameter have been detailed in Table 2.

Amentaflavone, a biflavonoid showed the highest binding energy (-8.49 kcal/mol) implicating a strong affinity with SARS-CoV-2 $3CL^{Pro}$. This corresponded with previously reported enzyme inhibitory assay with amentaflavone that showed highest IC_{50} value at low concentration of the molecule, 8.3±1.2 µM [44]. However, bilobetin demonstrated the lowest IC_{50} value at higher concentration of 72.3±4.5 µM in SARS-CoV enzyme activity assay [44]. In contrast, our docking studies revealed that bilobetin, predicted almost comparable binding energy with that of amentaflavone (-8.29 kcal/mol) suggesting that mutation in SARS-CoV-2 $3CL^{Pro}$ could potentially disrupt hydrogen bonding or induce some conformational change that could result in alterations in the binding site thus affecting inhibitor interactions with the enzyme active site residues. Amentaflavone showed H-bond interactions with the catalytic dyad residues (Cys145 and His41) as well as significant interactions with the SARS-CoV-2 $3CL^{Pro}$ residues Thr26, Ser46, Ser144 and Glu166 whereas His164, and Gln189 amino acid contributed to the hydrophobic interactions for the SARS-CoV-2 $3CL^{Pro}$ inhibitors (Fig. 7A).
Three antimalarial drugs were then selected to study their inhibitory actions on SARS-CoV-2 $3CL^{Pro}$. We found, Artemisinin, a natural compound derived from Chinese herb Artemisia *annua* produces the highest docking score (-6.40 kcal/mol) as compared to O6K, chloroquine (-4.95 kcal/mol) and hydroxychloroquine (-5.77 kcal/mol) molecules. Importantly, Artemisinin has demonstrated broad anti-viral activity against human cytomegalovirus, herpes simplex virus type 1, Epstein-Barr virus, hepatitis B virus, hepatitis C virus, and bovine viral diarrhea virus [45]. Artemisinin was shown to exhibit hydrogen bonding with His41, Leu141, Asn142, Gly143, Ser144 and Glu166 SARS-CoV-2 $3CL^{Pro}$ amino-acid residues (Fig. 7B).
Amongst the seven antiviral drugs, Ritonavir showed the highest binding energy (-7.45 kcal/mol) and lowest inhibitory constant K*i* value (3.49 µM). Ritonavir produced hydrogen bond formation with Thr26, His41 and Cys145 SARS-CoV-2 amino acids (Fig. 7C). A combination of two HIV-1 protease inhibitors, lopinavir and ritonavir, has been given clinically to critically ill patients infected with SARS-CoV 2 [46]. However, the combination therapy of lopinavir and ritonavir was also stopped early in 13 patients (total recruitment 99 patients) due to associated gastrointestinal adverse events [46].

The severity of antiviral therapy adverse events has led to researchers exploring potential macro-, micro- and phytonutrients that can potentially promote immune response and suppress the viral effects. Vitamins are previously known to modulate the host immune functions by providing anti-oxidants and anti-inflammatory activity [47, 48]. Therefore, we selected vitamins, ascorbic acid (vitamin C), cholecalciferol (vitamin D) and alpha-tocopherol (vitamin E) to investigate their potential interactions with the enzyme SARS-CoV-2 $3CL^{Pro}$. Our docking results interestingly, showed that vitamin D has the lowest binding energy and K*i* (-7.75 kcal/ mol and 2.08 µM respectively) as compared to vitamin C and vitamin E. Amino acid residues Thr24, Thr26, His41 and Cys145 of SARS-CoV-2 $3CL^{Pro}$ showed hydrogen bond formation with vitamin D structure (Fig. 7D). Amino acid Thr is extensively involved in intracellular signalling changes through phosphorylation changes, and here we observed that vitamin D formed a strong hydrogen bond with Thr residues and this could potentially block the phosphorylation of this residue in SARS-CoV-2 $3CL^{Pro}$ enzyme. There is evidence that serious SARS-CoV-2 infected cases have reported severe vitamin D deficiency and thus therapeutic concentrations of this molecule can be have been used clinically with SARS-CoV-2 [49, 50].

Discussions

The novel coronavirus termed "nCovid-19" is now known as the third large-scale epidemic coronavirus introduced into the human population in the twenty-first century. At the time of writing, more than 3.67 million confirmed cases globally, with nearly 250,000 deaths had been reported by WHO. Clinically, nCovid-19 is similar to SARS regarding its presentation, however the sheer capacity and speed of which nCovid-19 has spread to global pandemic levels have left researchers asking what makes this outbreak so similar in presentation, yet so different in its virulence to previous coronaviruses. Genome sequence analysis has looked to investigate similarities in the phylogeny of SARS-CoV-2, which like SARS and MERS, have now placed it in the betacoronavirus genus [51]. The known severe and often fatal pathogenicity of betacoronaviruses has been highlighted in these previous epidemics and has reported higher transmission and pathogenicity than the milder and lesser known a CoVs, which are often compared to the common cold [52]. Our study further compares the similarities between SARS-CoV and SARS-CoV-2 using Clustal Omega alignment to show that of 918 SARS nucleotides, there was a similarity of approximately 95%. Furthermore, we report high amino acid sequence identity in both SARS-CoV and SARS-CoV-2 main protease $3CL^{Pro}$, which

regulates coronavirus replication complexes [53]. Such highly conserved regions in both catalytic sites and the substrate binding regions of the enzymes has also been validated previously in studies by Huang *et al* and Muramatsu *et al* [42, 43]. While this region provides an attractive target for anti-viral drug design, it also can begin to elucidate on viral origins and uncover its ease in transmission.

Based on more recent virus genome sequencing results and evolutionary analysis, the origins and transmission of nCovid-19 have uncovered bats as the natural host of the virus origins [40]. As such, studies earlier this year queried the unknown intermediate host between bats and humans, and recent studies have pointed this to pangolins [40, 54]. To determine the extent of the evolutionary relationship between bat-CoV, Pangolin-CoV and SARS-CoV-2, we corroborate that based on the nucleotide sequences of the whole-genome sequence, bat-SARS-CoV and SARS-CoV-2 are grouped together and share >96% similarity, with Pangolin-SARS-CoV as the closest evolutionary ancestor [40, 54]. Furthermore, we report that in isolates of human Wuhan SARS-CoV-2 there is an 85.98% similarity in identity to Pangolin-SARS-CoV, which suggests that Pangolin may be associated with the evolution of subsequent outbreak of COVID-19.

Regarding nCovid-19 and its similarity in transmission to SARS-CoV, recent studies have also demonstrated that transmission occurs via the receptor angiotensin-converting enzyme 2 (ACE2) [40]. This may indicate why SARS-CoV-2 has often led to severe and in many cases fatal respiratory tract infections, like its two SAR-CoV predecessors. Since the SARS-CoV epidemic of 2002 is was also known to use the ACE2 receptor to infect humans [55]. Bronchoalveolar lavage fluid taken from nCovid-19 patients have shown that ACE2 is widely distributed in the lower respiratory tracts of humans [40]. Furthermore, the virion S-glycoproteins expressed on the surface of coronaviruses adhere to ACE2 receptors on human cells [56]. This location provides a target for uncovering the mechanistic insights into the severity of the disease and how this region has assisted in the zoonosis of SARS-CoV-2 specifically. Additionally, mutations in the genomic structure of SARS-CoV-2 also might elucidate on the aggressiveness and pathogenicity of the viruses, which may also help to explain why some strains are evolutionarily much more virulent and contagious. Angeletti *et al* have described mutations in the endosome-associated-protein-like domain of the nsp2 and nsp3 proteins, the former possible accounting for the high virulence and contagion, while the latter suggesting a mechanism that differentiates nCovid-19 to SARS-CoV [57]. Our studies build on this knowledge and assist to begin to identify the sub-clinical causes for the virulence and unique pandemic pattern of this outbreak by identifying the evolving mutations from region

to region. Additionally, previous studies by Pachetti et al have reported novel recurrent mutations of the SARS-Cov-2, and our study corroborates these mutations in South America and Africa regions [41].

Drug discovery and vaccine development against SARS-CoV-2 infection require time and lengthy process, however drug repurposed represents an alternative strategy in current scenario. Some of the antivirals are being used clinically in SARS-CoV-2 treatment, including lopinavir [58], ritonavir [59], remdesivir [60], and oseltamivir [61]. In the clinical setting, lopinavir/ ritonavir, a $3CL^{Pro}$ and RdRp inhibitors, showed no benefit in Covid-19 adult patients [46]. The double point mutation in RdRp gene identified in our study can potentially lead to drug-resistance event. Moreover, other class of drugs, like chloroquine and hydroxychloroquine shown antiviral property by blocking viral entry into the cells by inhibiting glycosylation of host receptor [62]. We observed no differences in the SARS-CoV-2 main proteinase, $3CL^{Pro}$ genome sequences but important differences in SARS-CoV-2 $3CL^{Pro}$ with SARs-CoV protein underline the extreme need for identification of inhibitors to target the viral life cycle.

Conclusions

Various theories have been proposed regarding the origin of highly virulent SARS-CoV-2 particle. Our analysis shows that Bat-SARS-CoV shares >90% similarity with the SARS-CoV-2, however it is possible that bat coronavirus infected another "intermediate host", such as Pangolin which subsequently transmitted the virus to humans. Pangolin isolates do share sequence identity with SARS-CoV-2 genome and could be an intermediate host. We identified novel mutation hotspot regions from South American and African isolates of SARS-CoV-2 genome sequences. Interestingly, double point mutations in RdRp at position 14805 and 14808 and triple point mutations in nucleocapsid protein at position 28881, 28882 and 28883 were identified in both South American and African genomic sequences, suggesting the vulnerability of these genetic loci to undergo change. In addition, a novel mutation pattern specifically oriented towards nucleocapsid phosphoprotein in both South American and African sequences was noted while novel ORF3a and RdRp specific variants were observed particularly from African genomic sequences. The potential effects of double and triple point mutations on translated proteins and virulence of SARS-CoV-2 requires further investigations. SARS-CoV-2 main proteinase, $3CL^{Pro}$ genome was observed to be conserved across all collected genomic sequences. Despite significant similarities in the SARS-CoV $3CL^{Pro}$ structure with SARS-CoV protein, SARS-CoV-2 $3CL^{Pro}$ revealed certain key differences, which highlight the extreme need for identification of novel mechanism-based drugs to target the virus processing.

Repurposed drugs including natural flavonoids and bioflavonoids, antimalarial, antiviral and vitamins-based compounds have previously been shown to be beneficial in several viral infections and outbreaks. The novel data generated from this study enhances our knowledge of the fine molecular differences that differentiate SARS-CoV-2 virus SARS-CoV. It also highlights the emerging variations in viral genome across different populations as the virus evolves to local genetic and environmental factors. These findings will likely play a key role in development of mechanism-based targeted therapeutic strategies to treat SARS-CoV-2 infection and reduce its virulence.

Abbreviations

SARS-CoV-2: Severe acute respiratory syndrome coronavirus (nCovid-19); 3 CLPro: 3-chymotrypsin-like cysteine protease; WHO: World Health Organisation; nCoV: novel coronavirus; ORF: open reading frame; PDB: Protein Data Bank; GISAID: The Global Initiative on Sharing Avian Influenza Data; MERS: Middle East respiratory syndrome; RBD: Receptor binding domain; RdRP: RNA-dependent RNA polymerase; nt: nucleotide; MSA: Multiple sequence alignment; iTOL: Interactive Tree of Life; AM1: Austin Model-1; RMS: Root mean square; MOPAC: Molecular Orbital Package; DS: Discovery studio; nsp: Non-structural protein

Acknowledgements

Not applicable

Authors' contributions

NC and VKG designed and conducted the research. RR, AG, TS and KP helped in data collection and analysis. NC and VKG wrote the paper. VG, DB, MM, YY and SLG supervised the data analysis and contributed to scientific discussion. All authors read and approved the final manuscript.

Funding

No funding was used to conduct this research.

Availability of data and materials

GISAID database (https://www.gisaid.org/), SARS-CoV-2 isolate Wuhan-Hu-1, complete genome (https://www.ncbi.nlm.nih.gov/nuccore/1798174254)

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests

References

1. Eurosurveillance Editorial T: **Note from the editors: World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern.** *Euro Surveill* 2020, **25**.
2. Coronaviridae Study Group of the International Committee on Taxonomy of V: **The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2.** *Nat Microbiol* 2020, **5:**536-544.
3. To KK, Hung IF, Chan JF, Yuen KY: **From SARS coronavirus to novel animal and human coronaviruses.** *J Thorac Dis* 2013, **5 Suppl 2:**S103-108.
4. Pillaiyar T, Manickam M, Namasivayam V, Hayashi Y, Jung SH: **An Overview of Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy.** *J Med Chem* 2016, **59:**6595-6628.
5. Seah I, Agrawal R: **Can the Coronavirus Disease 2019 (COVID-19) Affect the Eyes? A Review of Coronaviruses and Ocular Implications in Humans and Animals.** *Ocul Immunol Inflamm* 2020, **28:**391-395.
6. Giacomelli A, Pezzati L, Conti F, Bernacchia D, Siano M, Oreni L, Rusconi S, Gervasoni C, Ridolfo AL, Rizzardini G, et al: **Self-reported olfactory and taste disorders in SARS-CoV-2 patients: a cross-sectional study.** *Clin Infect Dis* 2020.
7. Schoeman D, Fielding BC: **Coronavirus envelope protein: current knowledge.** *Virol J* 2019, **16:**69.
8. Li F: **Structure, Function, and Evolution of Coronavirus Spike Proteins.** *Annu Rev Virol* 2016, **3:**237-261.
9. de Haan CA, Smeets M, Vernooij F, Vennema H, Rottier PJ: **Mapping of the coronavirus membrane protein domains involved in interaction with the spike protein.** *J Virol* 1999, **73:**7441-7452.
10. Shi CS, Nabar NR, Huang NN, Kehrl JH: **SARS-Coronavirus Open Reading Frame-8b triggers intracellular stress pathways and activates NLRP3 inflammasomes.** *Cell Death Discov* 2019, **5:**101.
11. Gadlage MJ, Denison MR: **Exchange of the coronavirus replicase polyprotein cleavage sites alters protease specificity and processing.** *J Virol* 2010, **84:**6894-6898.
12. Zumla A, Chan JF, Azhar EI, Hui DS, Yuen KY: **Coronaviruses - drug discovery and therapeutic options.** *Nat Rev Drug Discov* 2016, **15:**327-347.
13. Zhou Y, Vedantham P, Lu K, Agudelo J, Carrion R, Jr., Nunneley JW, Barnard D, Pohlmann S, McKerrow JH, Renslo AR, Simmons G: **Protease inhibitors targeting coronavirus and filovirus entry.** *Antiviral Res* 2015, **116:**76-84.
14. Kawase M, Shirato K, van der Hoek L, Taguchi F, Matsuyama S: **Simultaneous treatment of human bronchial epithelial cells with serine and cysteine protease inhibitors prevents severe acute respiratory syndrome coronavirus entry.** *J Virol* 2012, **86:**6537-6545.
15. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, Becker S, Rox K, Hilgenfeld R: **Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors.** *Science* 2020, **368:**409-412.
16. Nukoolkarn V, Lee VS, Malaisree M, Aruksakulwong O, Hannongbua S: **Molecular dynamic simulations analysis of ritonavir and lopinavir as SARS-CoV 3CL(pro) inhibitors.** *J Theor Biol* 2008, **254:**861-867.
17. Xue X, Yu H, Yang H, Xue F, Wu Z, Shen W, Li J, Zhou Z, Ding Y, Zhao Q, et al: **Structures of two coronavirus main proteases: implications for substrate binding and antiviral drug design.** *J Virol* 2008, **82:**2515-2527.
18. Shu Y, McCauley J: **GISAID: Global initiative on sharing all influenza data - from vision to reality.** *Euro Surveill* 2017, **22**.
19. Elbe S, Buckland-Merrett G: **Data, disease and diplomacy: GISAID's innovative contribution to global health.** *Glob Chall* 2017, **1:**33-46.

20. Hung LS: **The SARS epidemic in Hong Kong: what lessons have we learned?** *J R Soc Med* 2003, **96:**374-378.
21. Hui DS, Azhar EI, Kim YJ, Memish ZA, Oh MD, Zumla A: **Middle East respiratory syndrome coronavirus: risk factors and determinants of primary, household, and nosocomial transmission.** *Lancet Infect Dis* 2018, **18:**e217-e227.
22. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol Syst Biol* 2011, **7:**539.
23. Subramanian S, Ramasamy U, Chen D: **VCF2PopTree: a client-side software to construct population phylogeny from genome-wide SNPs.** *PeerJ* 2019, **7:**e8213.
24. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v4: recent updates and new developments.** *Nucleic Acids Res* 2019, **47:**W256-W259.
25. Letunic I, Bork P: **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.** *Bioinformatics* 2007, **23:**127-128.
26. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39:**W475-478.
27. Letunic I, Bork P: **Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.** *Nucleic Acids Res* 2016, **44:**W242-245.
28. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
29. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera--a visualization system for exploratory research and analysis.** *J Comput Chem* 2004, **25:**1605-1612.
30. Chitranshi N, Gupta V, Dheer Y, Gupta V, Vander Wall R, Graham S: **Molecular determinants and interaction data of cyclic peptide inhibitor with the extracellular domain of TrkB receptor.** *Data in Brief* 2016, **6:**776-782.
31. Gupta VK, Gowda LR: **Alpha-1-proteinase inhibitor is a heparin binding serpin: molecular interactions with the Lys rich cluster of helix-F domain.** *Biochimie* 2008, **90:**749-761.
32. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al: **PubChem 2019 update: improved access to chemical data.** *Nucleic Acids Res* 2019, **47:**D1102-D1109.
33. Chitranshi N, Gupta V, Kumar S, Graham SL: **Exploring the Molecular Interactions of 7,8-Dihydroxyflavone and Its Derivatives with TrkB and VEGFR2 Proteins.** *International Journal of Molecular Sciences* 2015, **16:**21087-21108.
34. Chitranshi N, Dheer Y, Vander Wall R, Gupta V, Abbasi M, Graham SL, Gupta V: **Computational analysis unravels novel destructive single nucleotide polymorphisms in the non-synonymous region of human caveolin gene.** *Gene Reports* 2017, **6:**142-157.
35. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: An open chemical toolbox.** *J Cheminform* 2011, **3:**33.
36. Chitranshi N, Dheer Y, Kumar S, Graham SL, Gupta V: **Molecular docking, dynamics, and pharmacology studies on bexarotene as an agonist of ligand-activated transcription factors, retinoid X receptors.** *J Cell Biochem* 2019.
37. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ: **AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.** *J Comput Chem* 2009, **30:**2785-2791.
38. Chitranshi N, Gupta S, Tripathi PK, Seth PK: **New molecular scaffolds for the design of Alzheimer's acetylcholinesterase inhibitors identified using ligand- and receptor-based virtual screening.** *Medicinal Chemistry Research* 2013, **22:**2328-2345.
39. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, et al: **A new coronavirus associated with human respiratory disease in China.** *Nature* 2020, **579:**265-269.

40. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al: **A pneumonia outbreak associated with a new coronavirus of probable bat origin.** *Nature* 2020, **579:**270-273.
41. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, et al: **Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant.** *J Transl Med* 2020, **18:**179.
42. Huang C, Wei P, Fan K, Liu Y, Lai L: **3C-like proteinase from SARS coronavirus catalyzes substrate hydrolysis by a general base mechanism.** *Biochemistry* 2004, **43:**4568-4574.
43. Muramatsu T, Takemoto C, Kim YT, Wang H, Nishii W, Terada T, Shirouzu M, Yokoyama S: **SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity.** *Proc Natl Acad Sci U S A* 2016, **113:**12997-13002.
44. Ryu YB, Jeong HJ, Kim JH, Kim YM, Park JY, Kim D, Nguyen TT, Park SJ, Chang JS, Park KH, et al: **Biflavonoids from Torreya nucifera displaying SARS-CoV 3CL(pro) inhibition.** *Bioorg Med Chem* 2010, **18:**7940-7947.
45. Efferth T, Romero MR, Wolf DG, Stamminger T, Marin JJ, Marschall M: **The antiviral activities of artemisinin and artesunate.** *Clin Infect Dis* 2008, **47:**804-811.
46. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, Ruan L, Song B, Cai Y, Wei M, et al: **A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19.** *N Engl J Med* 2020.
47. Zhang L, Liu Y: **Potential interventions for novel coronavirus in China: A systematic review.** *J Med Virol* 2020, **92:**479-490.
48. Conti P, Ronconi G, Caraffa A, Gallenga CE, Ross R, Frydas I, Kritas SK: **Induction of pro-inflammatory cytokines (IL-1 and IL-6) and lung inflammation by Coronavirus-19 (COVI-19 or SARS-CoV-2): anti-inflammatory strategies.** *J Biol Regul Homeost Agents* 2020, **34**.
49. Grant WB, Lahore H, McDonnell SL, Baggerly CA, French CB, Aliano JL, Bhattoa HP: **Evidence that Vitamin D Supplementation Could Reduce Risk of Influenza and COVID-19 Infections and Deaths.** *Nutrients* 2020, **12**.
50. Marik PE, Kory P, Varon J: **Does vitamin D status impact mortality from SARS-CoV-2 infection?** *Med Drug Discov* 2020**:**100041.
51. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, et al: **A Novel Coronavirus from Patients with Pneumonia in China, 2019.** *New England Journal of Medicine* 2020, **382:**727-733.
52. Yin Y, Wunderink RG: **MERS, SARS and other coronaviruses as causes of pneumonia.** *Respirology* 2018, **23:**130-137.
53. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R: **Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs.** *Science* 2003, **300:**1763-1767.
54. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, et al: **A new coronavirus associated with human respiratory disease in China.** *Nature* 2020, **579:**265-269.
55. Jia HP, Look DC, Shi L, Hickey M, Pewe L, Netland J, Farzan M, Wohlford-Lenane C, Perlman S, McCray PB, Jr.: **ACE2 receptor expression and severe acute respiratory syndrome coronavirus infection depend on differentiation of human airway epithelia.** *J Virol* 2005, **79:**14614-14621.
56. Tortorici MA, Veesler D: **Structural insights into coronavirus entry.** *Adv Virus Res* 2019, **105:**93-116.
57. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M: **COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis.** *J Med Virol* 2020.
58. Yao TT, Qian JD, Zhu WY, Wang Y, Wang GQ: **A systematic review of lopinavir therapy for SARS coronavirus and MERS coronavirus-A possible reference for coronavirus disease-19 treatment option.** *J Med Virol* 2020.
59. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, Ruan L, Song B, Cai Y, Wei M, et al: **A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19.** *N Engl J Med* 2020, **382:**1787-1799.

60. Ko WC, Rolain JM, Lee NY, Chen PL, Huang CT, Lee PI, Hsueh PR: **Arguments in favour of remdesivir for treating SARS-CoV-2 infections.** *Int J Antimicrob Agents* 2020, **55:**105933.
61. Pavone P, Ceccarelli M, Taibi R, La Rocca G, Nunnari G: **Outbreak of COVID-19 infection in children: fear and serenity.** *Eur Rev Med Pharmacol Sci* 2020, **24:**4572-4575.
62. Zhou D, Dai SM, Tong Q: **COVID-19: a recommendation to examine the effect of hydroxychloroquine in preventing infection and progression.** *J Antimicrob Chemother* 2020.
63. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA: **Nextstrain: real-time tracking of pathogen evolution.** *Bioinformatics* 2018, **34:**4121-4123.

Figure legends

Fig. 1 Schematic representation of SARS-CoV-2 structure showing single stranded RNA viral genomic assembly of 29,674 nucleotide base pair which encodes open reading frame 1a (ORF1a, nt 266-13468), yellow colour, open reading frame 1b (ORF1b, nt 13468-21563), blue colour, Spike (S, nt 21563-25384), Envelope (E, nt 26245-26472), Membrane (M, nt 26523-27191) and Nucleocapsid (N, nt 28274-29533) proteins in green. ORF1a gene encodes papain-like protease and 3CL protease, ORF1b gene encodes RNA-dependent RNA polymerase, helicase and endo ribo-nuclease, S, E, M and N gene encodes spike, membrane glycoprotein and nucleocapsid phosphoprotein respectively. Three-dimensional crystal structure of 3CL-protese, endoribonuclease and SARS-Cov-2 spike protein receptor binding domain (RBD) engaged human angiotensin converting enzyme 2 (ACE2) receptor were collected from protein data bank.

Fig. 2 (a, b) Phylogenetic evolutionary relationships of SARS-CoV-2 virus showing an initial emergence in Wuhan, China, in Nov-Dec 2019 followed by continued human-to-human transmission. SARS-CoV-2 patient genome sequences deposited in GISAID database from more than 60 different countries (a) radial and (b) unrooted phylogeny created by Nextstrain program [63] (c) Pie chart representation of number of SARS-CoV-2 patient genomes deposited in GISAID till 15th April 2020 from six different regions; Asia (orange, 35.37%), Oceania (purple, 20.85%), North America (brown, 21.13%), Europe (blue, 20.31%), Africa (red, 3.52%) and South America (black, 2.87%). (d) Phylogenetic relationship of CoVs based on whole genome nucleotide sequences from bat, pangolin, and Wuhan SARS-CoV-2 (NC_045512.2) confirms that SARS-CoV-2 share >90% similarity with bat SARS-CoV while pangolin could be the closest ancestral.

Fig. 3 Graphical representation of SARS-CoV-2 mutation frequency in South American and African patient isolates. (a) Five novel recurrent hotspots mutations (namely 14805, 25563, 26144, 28882 and 28883) were subdivided into 2 geographical areas: South America (n = 53) and Africa (n=65). Previously confirmed mutations at positions nt3036, nt8782, nt11083, nt14408, nt23403, nt28144 and nt28881 were also present in South American and African populations. We normalize the mutation frequency percentage by estimating the frequency of genomes carrying mutation and comparing it with the overall number of collected genomes per geographical area. The graph shows the cumulative mutation frequency of all given mutations

present in South American and African regions. Mutation localisation in viral genes are reported in the legend as well as the proteins (i.e. non-structural protein, nsp) presenting these mutations. (b) It is also evident that South American and African clusters show a differential pattern of novel mutations: mutation 9477 (pink), 28657 (green) and 28878 (red) in South American, whereas mutation 1059 (black), 15324 (orange), 28878 (yellow) and 29742 (magenta) are present with greater frequency in African patients.

Fig. 4 SARS-CoV-2, Main proteinase $3CL^{pro}$ analysis (a) Cartoon representation structure of the SARS-CoV-2 $3CL^{pro}$ homodimer with inhibitor (green) in greyish black colour. Variant positions of amino acids in $3CL^{pro}$ (Thr35Val, Ala46Ser, Ser65Asn, Leu86Val, Arg88Lys, Ser94Ala, His134Phe, Lys180Asn, Leu202Val, Ala267Ser, Thr285Ala and Ile286Leu) are shown in yellow colours (b) Multiple sequence alignment between SARS-CoV and SARS-CoV-2 $3CL^{pro}$ from Wuhan (Wu) and United States of America (US) patients sharing more than 90% sequence identity. (c) Surface view representation of SARS-CoV-2 $3CL^{pro}$ (PDB ID: 6Y2G) showing muted amino acid residues in yellow and alpha-ketoamide inhibitor (green) in the substrate binding region. Images are generated by UCSF Chimera software.

Fig. 5 (a) Cartoon representation of superimposed structures from SARS-CoV $3CL^{Pro}$ (PDB ID: 3TNT, grey) and SARS-CoV-2 $3CL^{Pro}$ (PDB ID: 6Y2G, cyan) showing 94.44 % sequence identity. Two different $3CL^{Pro}$ inhibitors represented in red and green colour in the substrate binding region (a') magnified view of substrate binding region. (b) The residues of the catalytic dyad (His41 and Cys145) are shown in surface view. Autodock 4.2 docking protocol was validated by re-docking of O6K inhibitor in SARS-CoV-2 $3CL^{Pro}$, original O6K inhibitor is shown in red colour and re-docked pose in cyan colour (c) three-dimensional and (d) surface view representation.

Fig. 6 Binding modes of different repurposed drugs in the substrate binding region of SARS-CoV-2 $3CL^{Pro}$ (A) flavanoids and biflavonoid (7,8 DHF (purple), apigenin (pink), luteolin (orange), quercitin (green), amentoflavone (grey), bilobetin (brown) and ginkgetin (white) (B) anti-malarial (chloroquine (green), hydroxychloroquine (golden yellow) and artemisinin (blue) (C) anti-viral (remdesivir (yellow), darunavir (blue), lopinavir (red), galidesivir (dark pink), favipiravir (light blue), ritonavir (light pink) and umifenovir (green) and (D) vitamins (vitamin C (cyan), vitamin D (golden orange) and vitamin E (pink)).

Fig. 7 The binding model of repurposed drugs against SARS-CoV-2 $3CL^{Pro}$ (a) Binding mode of the Amentaflavone drug (red) in SARS-CoV-2 $3CL^{Pro}$ (green) substrate binding pocket (b) Artemisinin drug (blue) binding mode in SARS-CoV-2 $3CL^{Pro}$ (pink) substrate binding pocket (c) Binding mode of the Ritonavir drug (magenta) in SARS-CoV-2 $3CL^{Pro}$ (cyan) substrate binding pocket (d) Binding mode of the Vitamin D (green) in SARS-CoV-2 $3CL^{Pro}$ (red) substrate binding pocket. Protein-ligand interaction (hydrogen bond) are shown with red dotted lines.

Figure 1

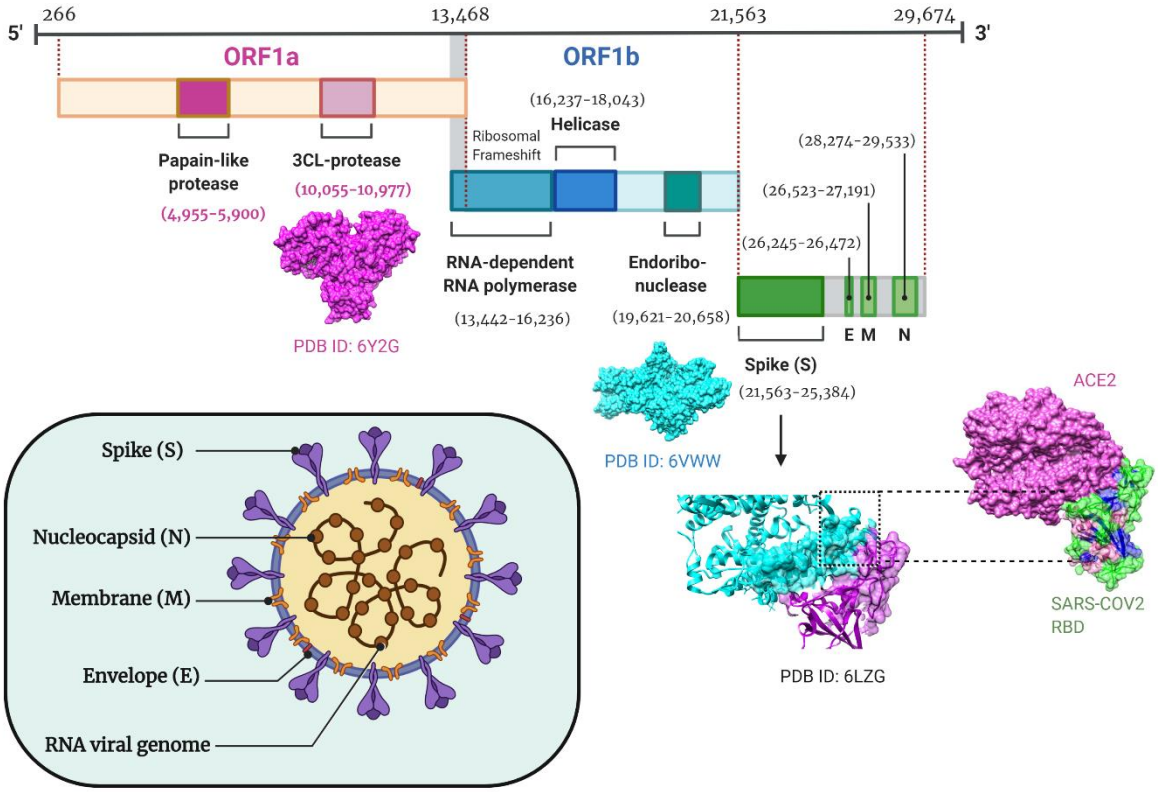
5'
266
13,468
21,563
29,674
3'
ORF1a
ORF1b
Papain-like protease
(4,955-5,900)
3CL-protease
(10,055-10,977)
PDB ID: 6Y2G
Ribosomal Frameshift
(16,237-18,043)
Helicase
RNA-dependent RNA polymerase
(13,442-16,236)
Endoribo-nuclease
(19,621-20,658)
(28,274-29,533)
(26,523-27,191)
(26,245-26,472)
E M N
Spike (S)
(21,563-25,384)
PDB ID: 6VWW
ACE2
SARS-COV2 RBD
PDB ID: 6LZG
Spike (S)
Nucleocapsid (N)
Membrane (M)
Envelope (E)
RNA viral genome

Figure 2

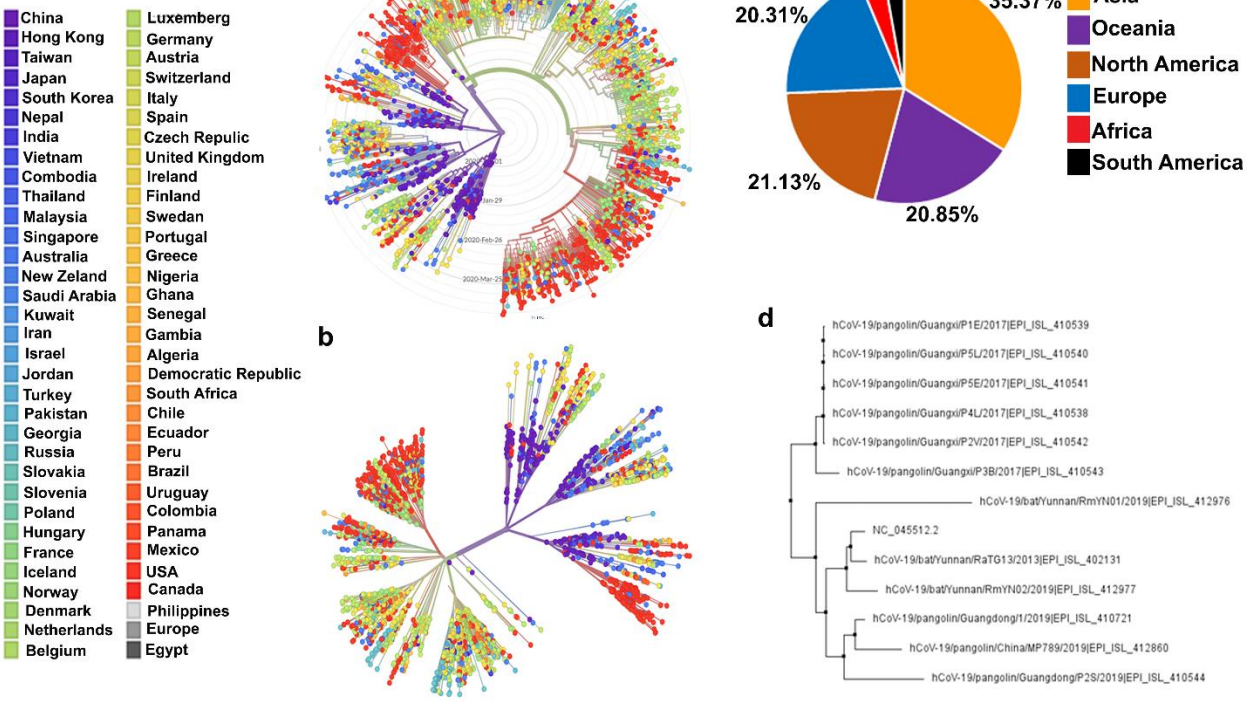
China
Hong Kong
Taiwan
Japan
South Korea
Nepal
India
Vietnam
Combodia
Thailand
Malaysia
Singapore
Australia
New Zeland
Saudi Arabia
Kuwait
Iran
Israel
Jordan
Turkey
Pakistan
Georgia
Russia
Slovakia
Slovenia
Poland
Hungary
France
Iceland
Norway
Denmark
Netherlands
Belgium
Luxemberg
Germany
Austria
Switzerland
Italy
Spain
Czech Repulic
United Kingdom
Ireland
Finland
Swedan
Portugal
Greece
Nigeria
Ghana
Senegal
Gambia
Algeria
Democratic Republic
South Africa
Chile
Ecuador
Peru
Brazil
Uruguay
Colombia
Panama
Mexico
USA
Canada
Philippines
Europe
Egypt
a
b
c
3.52%
2.87%
35.37%
20.31%
21.13%
20.85%
Asia
Oceania
North America
Europe
Africa
South America
d
hCoV-19/pangolin/Guangxi/P1E/2017|EPI_ISL_410539
hCoV-19/pangolin/Guangxi/P5L/2017|EPI_ISL_410540
hCoV-19/pangolin/Guangxi/P5E/2017|EPI_ISL_410541
hCoV-19/pangolin/Guangxi/P4L/2017|EPI_ISL_410538
hCoV-19/pangolin/Guangxi/P2V/2017|EPI_ISL_410542
hCoV-19/pangolin/Guangxi/P3B/2017|EPI_ISL_410543
hCoV-19/bat/Yunnan/RmYN01/2019|EPI_ISL_412976
NC_045512.2
hCoV-19/bat/Yunnan/RaTG13/2013|EPI_ISL_402131
hCoV-19/bat/Yunnan/RmYN02/2019|EPI_ISL_412977
hCoV-19/pangolin/Guangdong/1/2019|EPI_ISL_410721
hCoV-19/pangolin/China/MP789/2019|EPI_ISL_412860
hCoV-19/pangolin/Guangdong/P2S/2019|EPI_ISL_410544

Figure 3

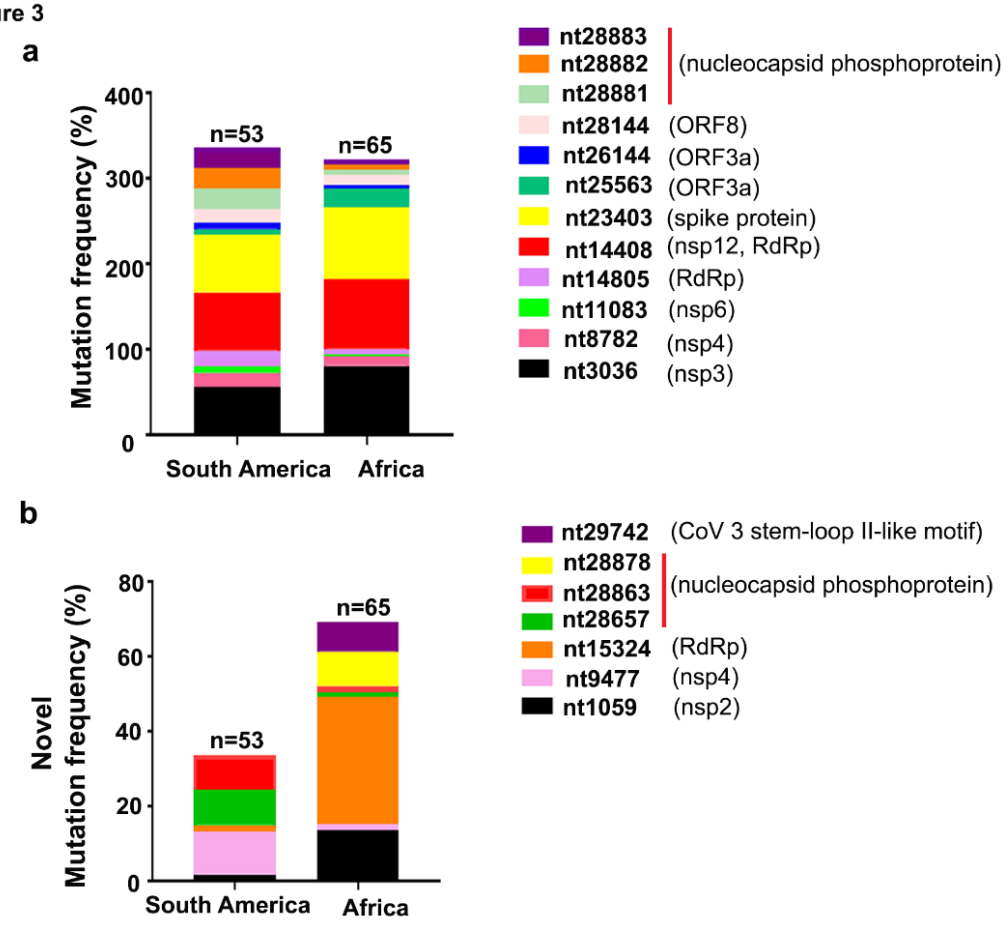
a
Mutation frequency (%)
400
300
200
100
0
n=53
n=65
South America
Africa
nt28883
nt28882
nt28881
(nucleocapsid phosphoprotein)
nt28144 (ORF8)
nt26144 (ORF3a)
nt25563 (ORF3a)
nt23403 (spike protein)
nt14408 (nsp12, RdRp)
nt14805 (RdRp)
nt11083 (nsp6)
nt8782 (nsp4)
nt3036 (nsp3)
b
Novel Mutation frequency (%)
80
60
40
20
0
n=53
n=65
South America
Africa
nt29742 (CoV 3 stem-loop II-like motif)
nt28878
nt28863
nt28657
(nucleocapsid phosphoprotein)
nt15324 (RdRp)
nt9477 (nsp4)
nt1059 (nsp2)

Figure 4

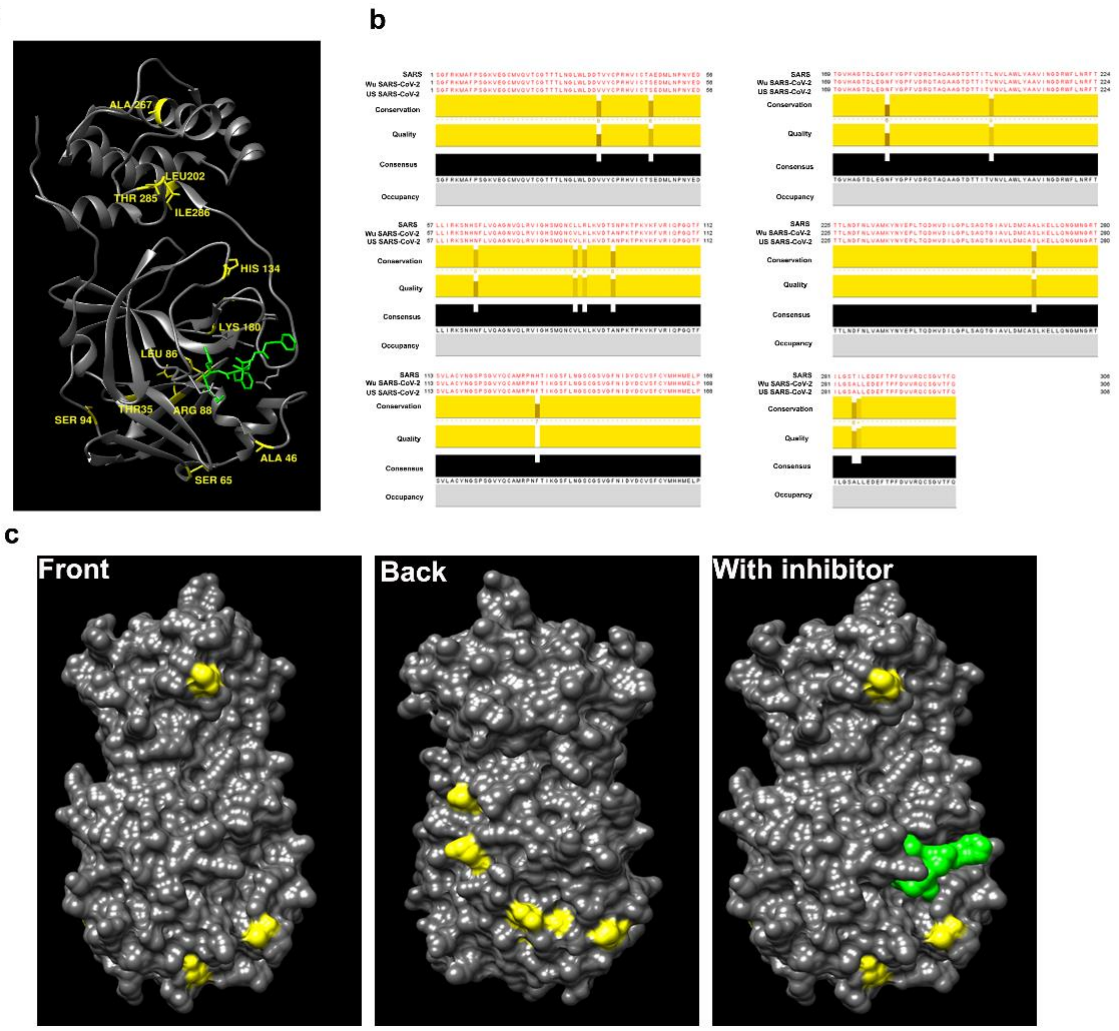
a
ALA 267
LEU202
THR 285
ILE286
HIS 134
LYS 180
LEU 86
SER 94
THR35
ARG 88
ALA 46
SER 65
b
SARS
Wu SARS-CoV-2
US SARS-CoV-2
Conservation
Quality
Consensus
Occupancy
c
Front
Back
With inhibitor

Figure 5

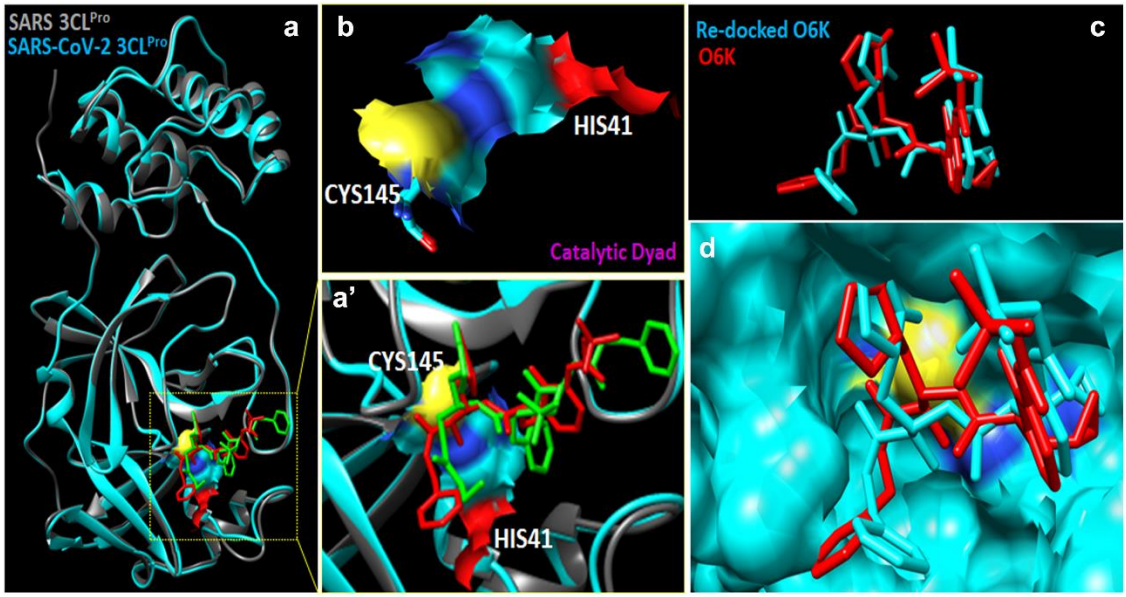
SARS 3CLPro
SARS-CoV-2 3CLPro
a
b
HIS41
CYS145
Catalytic Dyad
a'
CYS145
HIS41
Re-docked O6K
O6K
c
d

Figure 6

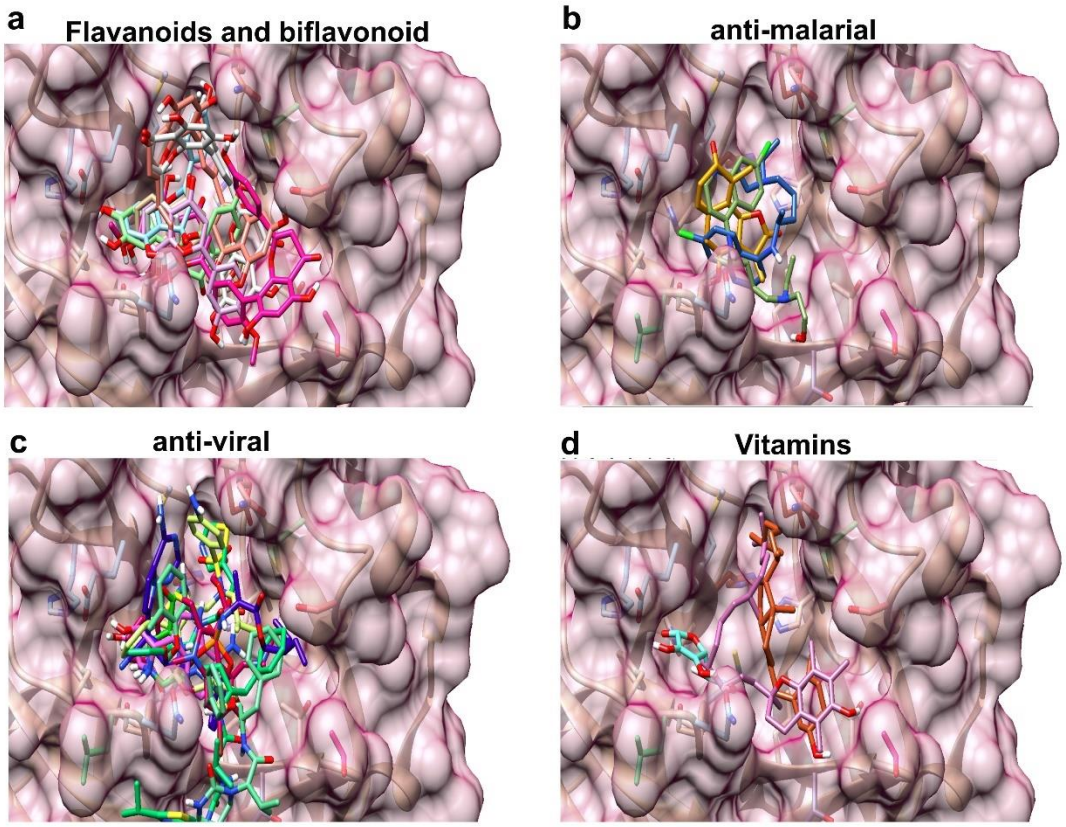
a
Flavanoids and biflavonoid
b
anti-malarial
c
anti-viral
d
Vitamins

Figure 7

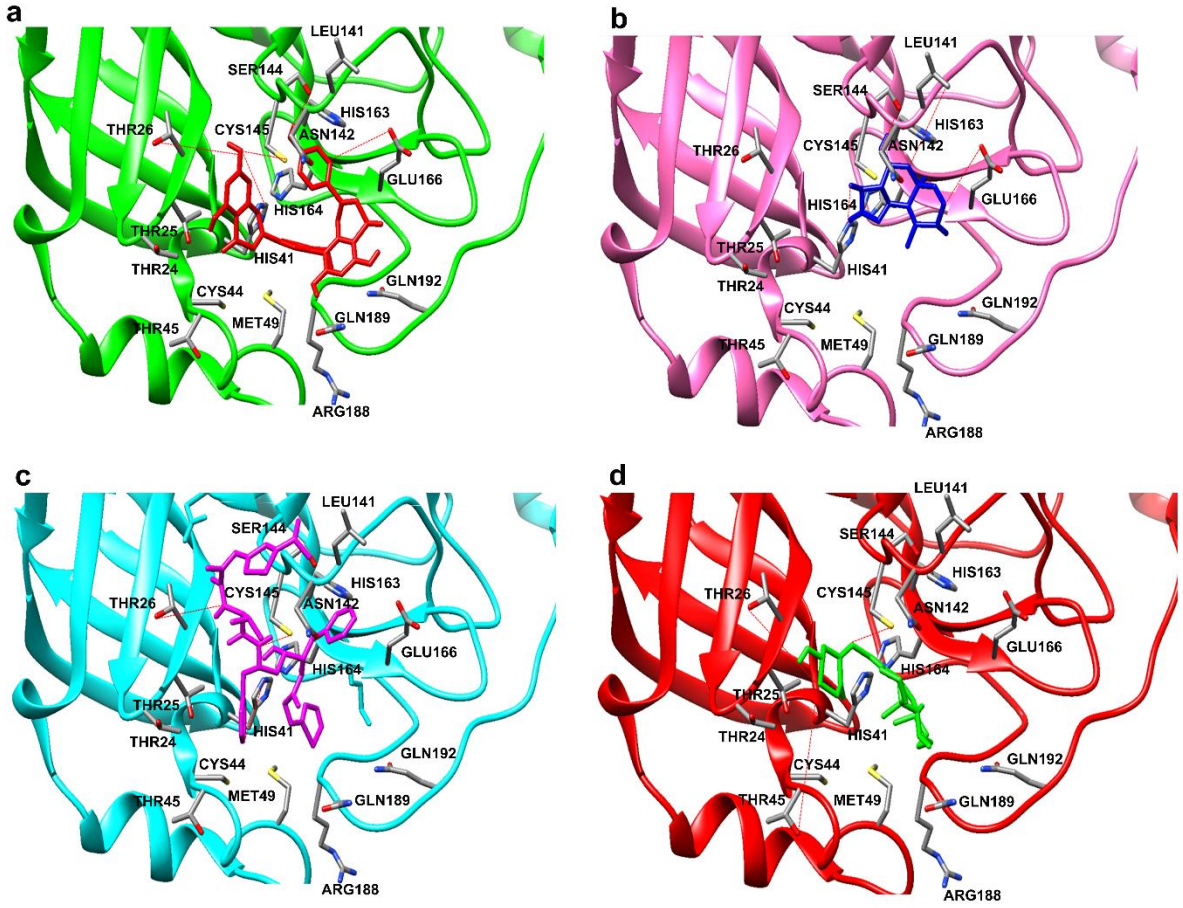
a
LEU141
SER144
HIS163
THR26
CYS145
ASN142
GLU166
HIS164
THR25
HIS41
THR24
CYS44
GLN192
MET49
GLN189
THR45
ARG188
b
LEU141
SER144
HIS163
THR26
CYS145
ASN142
HIS164
GLU166
THR25
HIS41
THR24
CYS44
GLN192
THR45
MET49
GLN189
ARG188
c
LEU141
SER144
HIS163
THR26
CYS145
ASN142
GLU166
HIS164
THR25
HIS41
THR24
GLN192
CYS44
MET49
GLN189
THR45
ARG188
d
LEU141
SER144
HIS163
THR26
CYS145
ASN142
GLU166
HIS164
THR25
THR24
HIS41
GLN192
CYS44
GLN189
THR45
MET49
ARG188

Table 1. Dataset of 20 repurposed drugs and reference ligand (O6K) corresponding energies obtained from the docking test performed using AutoDock 4.2 program. *BE*^{e} Estimated binding free energy in kcal mol^{-1}; *Ki* Inhibitory constant in micro-molar; *IME*^{e} Final Intermolecular Energy in kcal mol^{-1}; V_{dw}–H_b–D_s Van der waals-hydrogen bond-desolvation energy component of binding free energy in kcal mol^{-1}; *E*^{e} Electrostatic energy in kcal mol^{-1}; *IE*^{e} Final total internal energy in kcal mol^{-1}; *TFE*^{e} Torsional free energy in kcal mol^{-1}; *USE* Unbound system's energy; *RMSD* Root mean square deviation (Å)

S. No	C.Name	*BE*^{e} *(kcal/ mol)*	*Ki (µM)*	*IME*^{e} *(kcal/mol)*	V_{dw}-H_b-D_s *(kcal/mol)*	*E*^{e} *(kcal/mol)*	*IE*^{e} *(kcal/mol)*	*TFE*^{e} *(kcal/mol)*	*USE*	*RMSD (Å)*
	O6K	-6.45	18.72	-10.33	-10.36	+0.03	-2.10	+3.88	-2.10	5.59
1	7, DHF	-6.24	26.49	-7.14	-6.83	-0.31	-1.45	+0.89	-1.45	32.85
2	Apigenin	-7.52	3.05	-8.75	-8.52	-0.20	+9.72	+1.19	+9.72	33.64
3	Luteolin	-5.59	80.53	-7.08	-6.78	-0.29	-1.84	+1.49	-184	33.26
4	Quercetin	-6.16	30.49	-7.95	-7.56	-0.39	-1.96	+1.79	-1.96	32.79
5	Amentoflavone	-8.49	0.59	-11.18	-10.97	-0.21	-3.34	+2.68	-3.34	35.22
6	Bilobetin	-8.29	0.83	-10.98	-10.68	-0.30	-4.30	+2.68	-4.30	35.32
7	Ginkgetin	-8.14	1.09	-10.82	-10.52	-0.30	-3.51	+2.68	-3.51	34.77
8	Chloroquine	-4.95	233.3	-7.34	-6.96	-0.38	-1.34	+2.39	-1.34	30.78
9	Hy-chloroquine	-5.77	58.47	-8.76	-8.53	-0.23	-0.68	+2.98	-0.68	30.56
10	Artemisinin	-6.40	20.22	-6.70	-6.50	-0.20	+0.06	+0.30	+0.06	31.94
11	Remdesivir	-6.40	28.28	-11.47	-11.55	+0.07	-4.52	+5.07	-4.52	31.05
12	Darunavir	-7.16	5.64	-11.34	-11.26	-0.08	-3.75	+4.18	-375	33.98
13	Lopinavir	-6.98	7.68	-11.75	-11.57	-0.18	-4.92	+4.77	-4.92	29.98
14	Galidesivir	-4.69	362.43	-6.48	-5.79	-0.70	-2.02	+1.79	-2.02	32.04
15	Favipiravir	-4.15	905.42	-4.45	-4.31	-0.14	-0.09	+0.30	-0.09	32.27
16	Ritonavir	-7.45	3.49	-13.11	-13.15	+0.03	-3.46	+5.67	-5.67	28.61
17	Umifenovir	-5.71	65.42	-8.39	-8.13	-0.26	-1.81	+2.68	-1.81	31.86
18	Vitamin C	-4.22	805.2	-6.01	-5.72	-0.29	-2.03	+17.79	-2.03	29.14
19	Vitamin D	-7.75	2.08	-9.84	-9.78	-0.06	-1.75	+2.09	-1.75	33.21
20	Vitamin E	-7.59	2.75	-11.46	-11.40	-0.06	-1.57	+3.88	-1.57	25.17

Table 2: Repurposed drugs (20) and O6K inhibitor interactions with the SARS-Cov-2 $3CL^{Pro}$ key amino acid residues involved in replication activity

S.No	Compound name/PubChem CID	H-bond residue
	O6K	His41, Phe140, Cys145, His163, His164, Glu166
1	7, DHF (PubChem CID:1880)	Ser144, His163
2	Apigenin (PubChem CID:5280443)	Cys145, His163
3	Luteolin (PubChem CID:5280445)	Thr26, His163, Glu166
4	Quercetin (PubChem CID:5280343)	Phe140, Leu141, Cys145
5	Amentoflavone (PubChem CID:5281600)	Thr26, His41, Ser46, Ser144, Cys145, Glu166
6	Bilobetin (PubChem CID:5315459)	Thr25, Leu141, Ser144, Cys145
7	Ginkgetin (PubChem CID:5271805)	Thr26, Gly143, His163
8	Chloroquine (PubChem CID:2719)	His41, Gly143, Cys145
9	Hydroxychloroquine (PubChem CID:3652)	Thr26, Asn142, Glu166
10	Artemisinin (PubChem CID:68827)	His41, Leu141, Asn142, Gly143, Ser144, Glu166
11	Remdesivir (PubChem CID:121304016)	Ser46, Ser144, Cys145, His163
12	Darunavir (PubChem CID:213039)	Gly143, Ser144, Cys145, His164
13	Lopinavir (PubChem CID:92727)	Thr26, His41, Gly143, Ser144, Cys145
14	Galidesivir (PubChem CID:10445549)	Phe140, Leu141, Gly143, Ser144, Cys145, His163, Glu166
15	Favipiravir (PubChem CID:492405)	Ser144, His163, His164, Glu166
16	Ritonavir (PubChem CID:392622)	Thr26, His41, Cys145
17	Umifenovir (PubChem CID:131411)	Thr26
18	Vitamin C (PubChem CID:54670067)	Asn142, Ser144, Cys145, His163, Gln166
19	Vitamin D (PubChem CID:5280795)	Thr24, Thr26, His41, Cys145
20	Vitamin E (PubChem CID:14985)	Thr25, Cys145