



Deep learning based on biologically interpretable genome representation predicts two types of human adaptation of SARS-CoV-2 variants

Jing Li ^{††}, Ya-Nan Wu[†], Sen Zhang[†], Xiao-Ping Kang and Tao Jiang 

Corresponding authors: Jing Li, State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, AMMS, Beijing 100071, China. E-mail: lj-pbs@163.com; Tao Jiang, State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, AMMS, Beijing 100071, China. E-mail: jiangtao@bmi.ac.cn

[†]Jing Li, Ya-Nan Wu and Sen Zhang contributed equally to this work.

^{††}Leading contact.

Abstract

Explosively emerging SARS-CoV-2 variants challenge current nomenclature schemes based on genetic diversity and biological significance. Genomic composition-based machine learning methods have recently performed well in identifying phenotype-genotype relationships. We introduced a framework involving dinucleotide (DNT) composition representation (DCR) to parse the general human adaptation of RNA viruses and applied a three-dimensional convolutional neural network (3D CNN) analysis to learn the human adaptation of other existing coronaviruses (CoVs) and predict the adaptation of SARS-CoV-2 variants of concern (VOCs). A markedly separable, linear DCR distribution was observed in two major genes—receptor-binding glycoprotein and RNA-dependent RNA polymerase (RdRp)—of six families of single-stranded (ssRNA) viruses. Additionally, there was a general host-specific distribution of both the spike proteins and RdRps of CoVs. The 3D CNN based on spike DCR predicted a dominant type II adaptation of most Beta, Delta and Omicron VOCs, with high transmissibility and low pathogenicity. Type I adaptation with opposite transmissibility and pathogenicity was predicted for SARS-CoV-2 Alpha VOCs (77%) and Kappa variants of interest (58%). The identified adaptive determinants included D1118H and A570D mutations and local DNTs. Thus, the 3D CNN model based on DCR features predicts SARS-CoV-2, a major type II human adaptation and is qualified to predict variant adaptation in real time, facilitating the risk-assessment of emerging SARS-CoV-2 variants and COVID-19 control.

Keywords: dinucleotide composition representation, 3D convolutional neural networks, SARS-CoV-2, variants of concern, human adaptation

Introduction

The sporadic zoonotic transfer of a pathogen may cause human disease and even death but does not necessarily cause sustained human-to-human transmissibility (e.g. Ebola and Hanta viruses); only human-adapted pathogens, such as human-transmissible coronaviruses (CoVs) and influenza A viruses (IAVs), cause sustained transmission in populations and pandemics [1, 2]. Human-infective, bat-originating CoVs [3], such as the highly pathogenic severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS)

CoVs, only resulted in regional, passing outbreaks [4], indicating limited human adaptation. In contrast, HCoV-229E, HCoV-NL63, HCoV-OC43 and HCoV-HKU1 are globally transmissible, causing 15–30% of common cold cases every year [5] thus, they are more human-adapted (type II human adaptation). Taking the transmissibility and pathogenicity of CoVs into account, we defined two types of human adaptation: type I adaptation (SARS and MERS), characterized by higher pathogenicity and lower transmissibility in the population; and type II adaptation (HCoV-229E, HCoV-NL63, HCoV-OC43 and

Jing Li is interested in the application of deep learning and natural language processing to parse viral genomes and worked at the State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, AMMS, Beijing, China.

Ya-Nan Wu is interested in respiratory pathogens and worked at the State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, AMMS, Beijing, China.

Sen Zhang is interested in respiratory pathogens and worked at the State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, AMMS, Beijing, China.

Xiao-Ping Kang is interested in insect-borne pathogens and worked at the State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, AMMS, Beijing, China.

Tao Jiang is interested in molecular virology and the surveillance of viral infectious diseases and worked at the State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, AMMS, Beijing, China.

Received: November 25, 2021. **Revised:** January 23, 2022. **Accepted:** January 25, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

HCoV-HKU1), characterized by lower pathogenicity but higher transmissibility. Since 2019, SARS-CoV-2 [6] has caused a global COVID-19 pandemic (<https://covid19.who.int>) characterized by high human adaptation [7–11]. Emerging SARS-CoV-2 variants have presented more worrisome epidemic-promoting advantages [7, 10–12], possibly due to faster optimization under vaccination selection pressure [13, 14]. Thus, there is a need to address fundamental questions related to CoV adaptation and discriminate adaptive types of emerging SARS-CoV-2 variants.

Two current nomenclature schemes facilitate the recognition of epidemiological outbreak links based on genetic diversity [15], and reliably indicate the biological significance of specific variants, which are designated as either variant of interest (VOI) or variant of concern (VOC), based on the expert monitoring of significant amino acid substitutions by the WHO (<https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>). However, neither scheme can rapidly assess the risk of any emerging SARS-CoV-2 variant. Research has indicated a contribution of specialized virus–host receptor binding and avidity to interspecies transmission, mainly related to the CoV spike (S) glycoprotein [16]. Many VOC mutations in SARS-CoV-2 S have been shown to promote transmission, enhance receptor binding affinity [9] and antagonize immune escape [11]. However, these *ex post*-experimental methods are not qualified to assess the adaptation of each SARS-CoV-2 variant in real time.

Sequence compositions of nucleic acids and proteins are significantly associated with genome evolution and adaptation across all kingdoms of life [17]. Machine/deep learning methods have worked well in predicting viral hosts based on the amino acid [18] or dinucleotide (DNT) [19] composition in the sequence alignment of large datasets [20]. Moreover, language representation methods have learned the language of viral evolution and escape based on represented amino acids [21] or statistically represented proteins [22]. Here, we aimed to design a novel method of dinucleotide composition representation (DCR) to learn the general human adaptation of CoVs and build an adaptation predictor for SARS-CoV-2 variants. DCR was first assessed as a classification trait based on the two major viral proteins—receptor-binding glycoprotein (Gp; also named as S for CoVs) and RNA-dependent RNA polymerase (RdRp)—of six single-stranded (ssRNA) orders/families. We evaluated the possible general adaptation of both Gp and RdRp and the representativeness of DCR as a general genomic nucleotide composition trait for DNTs, codons, codon pairs and amino acids (AA). Finally, we built a three-dimensional convolutional neural network (3D-CNN) predictor based on DCR to predict the human adaptation of SARS-CoV-2 variants. Our study provides a novel, simplified and reliable genome representation as well as a framework to discern the general human adaptation of SARS-CoV-2, CoVs in general and other emerging viruses.

Results

Pipeline of DCR and 3D-CNN prediction

As shown in the architecture diagram (Figure 1), six ssRNA virus families, Bunyaviridae, Orthomyxoviridae, Filoviridae, Flaviviridae, Togaviridae and Coronaviridae, were included in the analysis of the general separability and linearity of the nucleotide d-traits of DNT, DCR, AA, codons and codon pairs in the viral genome (Figure 1A). Gp and RdRp were targeted to decompose these traits. Coronaviridae sequences, excluding SARS-CoV-2, were randomly split into training and validation datasets to build an adaptation classifier to predict the human adaptation of the test dataset of SARS-CoV-2 variants. The DCR algorithm was a fine-grained version of compositional DNT embedding [19], representing the local nucleotide context in a sequence (Figure 1B). Six channels of DCR were set to represent six types of DNT pairs (Figure 1C). The unsupervised projection methods of t-distributed stochastic neighbour embedding (t-SNE) [23] and principal component analysis (PCA) [24] were utilized to learn the multigrained separability and linearity of randomly sampled CoVs and other RNA viruses (Figure 1D). A 3D-CNN framework for three-category adaptation classification—inadaptation, type I human adaptation and type II human adaptation—was built with three layers of convolution + ReLU, two average pooling layers + one maximum pooling layer, two fully connected layers and a final softmax layer, to predict the human adaptation of each SARS-CoV-2 variant (Figure 1E). Temporal and special shifts of SARS-CoV-2 adaptation were further analysed (Figure 1F); DCR features that are important for the human adaptation of SARS-CoV-2 were assessed by orthogonality in the DCR vector (Figure 1G). Synthetic Minority Over-sampling Technique resampling was implemented to rectify the amount of imbalance among different host-originated samples (Additional file 1: [Supplementary Tables S1 and S2](#), Additional file 2: [Supplementary Figure S1](#) available online at <http://bib.oxfordjournals.org/>).

General separability and linearity of the DCR of ssRNA viruses

To assess the host adaptation of coronaviruses as well as the possibility of predicting human adaptation based on DCR, we first analysed the general separability and linearity of DNT, DCR and other compositional traits in the six families of ssRNA viruses. The two-dimensional t-SNE or PCA projection of these features of 200 randomly sampled Gp and RdRp sequences ([Supplementary Table S2](#) available online at <http://bib.oxfordjournals.org/>) indicated clear interfamily separation of Gp among the six virus families in the two reduced t-SNE components of the compositional DNT (Figure 2A) or DCR (Figure 2A) of Gp. Such interviridae separation was generally observed in the reduced t-SNE components of other compositional traits and in the reduced-PCA components of the compositional DNTs, DCRs, codons, codon pairs

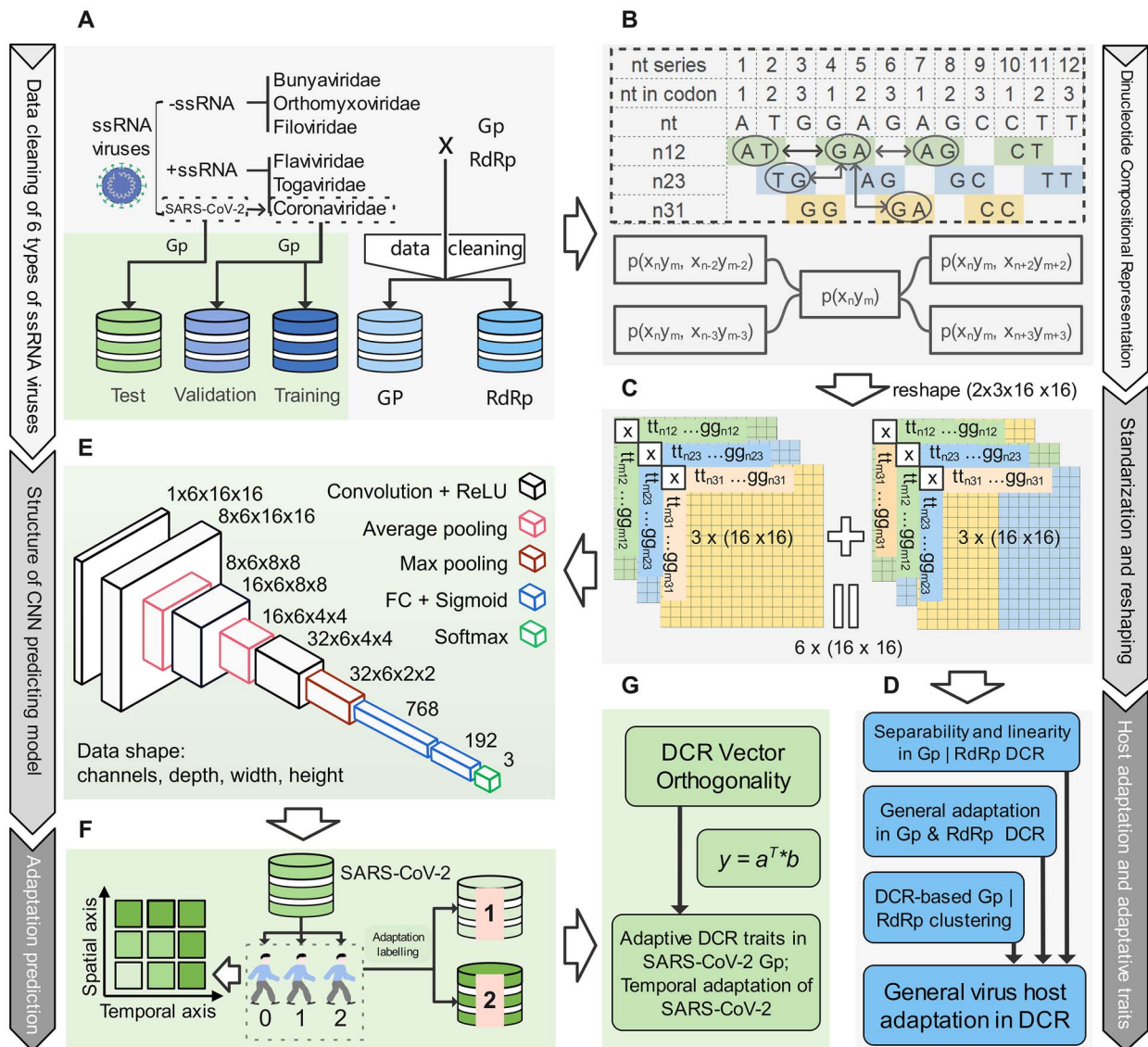


Figure 1. Workflow of nucleotide composition representation and adaptation prediction. The workflow was designed from the top to bottom as follows: data clearing (A), demonstration of the DCR algorithm (B) and data dimension settings (C) unsupervised learning of the separability, linearity and general adaptation of ssRNA viruses (D), design of the 3D-CNN model (E), adaptation prediction (F) and assessment of DCR features for the human adaptation of SARS-CoV-2 (G).

and AA of Gp (Additional file 2: Supplementary Figure S2 available online at <http://bib.oxfordjournals.org/>) and RdRp (Additional file 2: Supplementary Figure S3 available online at <http://bib.oxfordjournals.org/>). Further evaluation of data linearity was performed to reflect its continuity and differentiability, and support the machine/deep learning classification of these samples [25]. The linearity feature was designed as the ratio of the data range of PCA1 to the data range of PCA2 based on the orthogonal distribution between PCA1 and PCA2. A higher relative linear distribution (linearity value of $PCA1/PCA2 > 1$) was obtained for DNTs, codons and DCRs of the six families of viruses for both RdRp (Figure 2C) and Gp (Figure 2D). In particular, DCR linearity was highest for Gp and the second highest for RdRp among the CoVs, indicating an obvious intraviridae linear distribution. A regression plot reconfirmed the

linearity of DNTs (Figure 2E) and DCRs (Figure 2F) for Gp between affined PCA2 and affined PCA1, with a PCA1-PCA2 R^2 value as high as 0.51 for the DCR of CoV Gp S. Similar linear regression was observed to varying degrees for DNTs, DCRs, codons, codon pairs and AA of both RdRp (Additional file 2: Supplementary Figure S4A–E available online at <http://bib.oxfordjournals.org/>) and Gp (Additional file 2: Supplementary Figure S4F–J available online at <http://bib.oxfordjournals.org/>). More detailed analysis showed a larger distribution interval within CoV S sequences with the labels of seven types of human-infective CoVs and other CoV-infected hosts for the five traits of RdRp (Additional file 2: Supplementary Figure S5A–E available online at <http://bib.oxfordjournals.org/>) and S (Additional file 2: Supplementary Figure S5F–J available online at <http://bib.oxfordjournals.org/>). Linearity was also observed in these traits in both genes,

particularly for the DCR or codon pairs of S for most CoVs (Additional file 2: [Supplementary Figure S5F and I](#) available online at <http://bib.oxfordjournals.org/>). Taken together, the results indicated a general separability and linearity of the DCR, other compositional traits of coronaviruses and other RNA virus families.

General host adaptation of the DCR of coronavirus proteins

Host-specific or host adaptation-related compositional features in the genome have been observed for various types of RNA viruses at both the nucleic acid [19] and protein levels [18, 21]. We assumed that a genomic compositional adaptation feature would be generally projected in all coding regions or genes of a virus if it existed. Surprisingly, the PCA1 values of DNTs, DCRs (Figure 3A and B), codons and codon pairs (Additional file 2: [Supplementary Figure S6A and B](#) available online at <http://bib.oxfordjournals.org/>) of S and RdRp were similarly distributed in a two-dimensional view, with clear separation of coronaviruses and different labels of hosts or lineages for sampled human coronaviruses, but this was not observed for the AA trait (Figure 3C). A general, strong and positive Spearman correlation was observed in the PCA-reduced component of DNTs (Figure 3A) and DCRs (Figure 3B) between CoV S and RdRp among five randomly sampled coronaviruses (Figure 3D), but this was not found for AA (Figure 3C). A similar correlation was observed for the traits of codons and codon pairs (Figure 3D). Additionally, the representativeness of the DCR for the other four traits was evaluated; using the R^2 score for the linear regression of other traits against DCR, it was shown that DNTs, codons, codon pairs and AA were highly dependent on DCRs for both S and RdRp (Figure 3E and F). Thus, we selected DCR as the most representative compositional feature for parsing host adaptation in the coronavirus genome.

DCR-based 3D-CNN prediction of the human adaptation of SARS-CoV-2 VOCs

Based on the separability and linearity in the DCR of S for CoVs and the representativeness of the DCR of genomic compositional traits, we built a CNN classifier to identify human adaptation within the S genes (three adaptation labels, 1, 2 and 0, indicating type I, type II adaptation and inadaptive, respectively) of SARS-CoV-2 variants based on 3D DCRs (Figure 1C). CoVs with type I adaptation (SARS and MERS), CoVs with type II adaptation (HCoV-229E, HCoV-NL63, HCoV-OC43 and HCoV-HKU1) and inadaptive CoVs (Suiformes CoVs) were randomly sampled to learn the model parameters. To visualize the learning curves, 15, 40 or 50 training epochs were performed. Higher error rates were indicated by lower values of the confusion matrix (Figure 4A) and the receiver operating characteristic (ROC)-area under the curve (AUC)-ROC ratio (Figure 4B) after 15 epochs, along with a sustained high-training loss value (Figure 4C). The pair plotting of PCA1 and PCA2 reduced via the PCA method from

768 fully connected layers only indicated a separation between the S samples with type II adaptation and those with type I adaptation/inadaptation (Figure 4D). Learning with 40 epochs improved model performance, with a lower error rate for the confusion matrix and ROC/AUC-ROC curve (Figure 4E and F), and a significant decline in training loss (Figure 4G), but without clear separation between inadaptation and type I adaptation (Figure 4H). A balanced training result of the separation among three adaptation types (Figure 4I-K) indicated separation among three types of reduced data and the theoretical temporality of inadaptation, type I adaptation and type II adaptation (Figure 4L). Therefore, the trained CNN classifier obtained after 50 epochs was utilized to predict the adaptation of SARS-CoV-2.

Human adaptation was predicted with the 3D-CNN model based on DCRs. All 1 457 628 SARS-CoV-2 S sequences were predicted to show human adaptation of either type I (39.04%) or type II (60.96%). A total of 1 376 088 of the S dataset sequences with complete collection-date information were analysed for temporal and special adaptation shifts. After the observation of a number of sporadic S sequences per month up to November 2020 worldwide, a steep rise in sequences with type I adaptation appeared as of December 2020, peaking at 145 851 in March 2021, predominantly in Europe and North America (Figure 5A). The number of sequences with type II adaptation rose to 32 392 in March 2020, which were widely observed in North America, Europe, Asia and Oceania. This number remained fairly steady until September 2020 (approximately 30 000 sequences per month), followed by a steep rising period between October 2020 and March 2021, mainly in North America and Europe (Figure 5B). Moreover, sequences with type I adaptation with ratios ranging from 0.15 to 0.44 were labelled VOCs/VOIs from January to May 2020. Starting from October 2020, approximately 10–50% of these sequences were labelled VOCs/VOIs until June 2021. Beginning in December 2020, this set of sequences was increasingly labelled as VOCs/VOIs (Figure 5C). A total of 67% of VOC/VOI sequences—mainly Alpha and Kappa—were predicted to show type I adaptation, while 91% of VOIs and 95% of variants other than VOIs and VOCs were predicted to show type II adaptation, indicating consistency in the SARS-CoV-2 risk assessment of the VOC label and type I adaptation (Figure 5D). The differences between the other proposed dynamic nomenclature and our model were also analysed. Our results showed that 77% of alpha and 58% of kappa sequences were predicted to show type I adaptation; 88–100% of beta, delta and other sequences were predicted to show type II adaptation (Figure 5E). Additionally, the difference in the probability of predicting type I and II adaptation (Δ -probability, absolute value of probability I minus probability II) was much lower for the S samples showing type I adaptation than for the type II samples (Figure 5F). The sampled months of February 2020, December 2020 and June

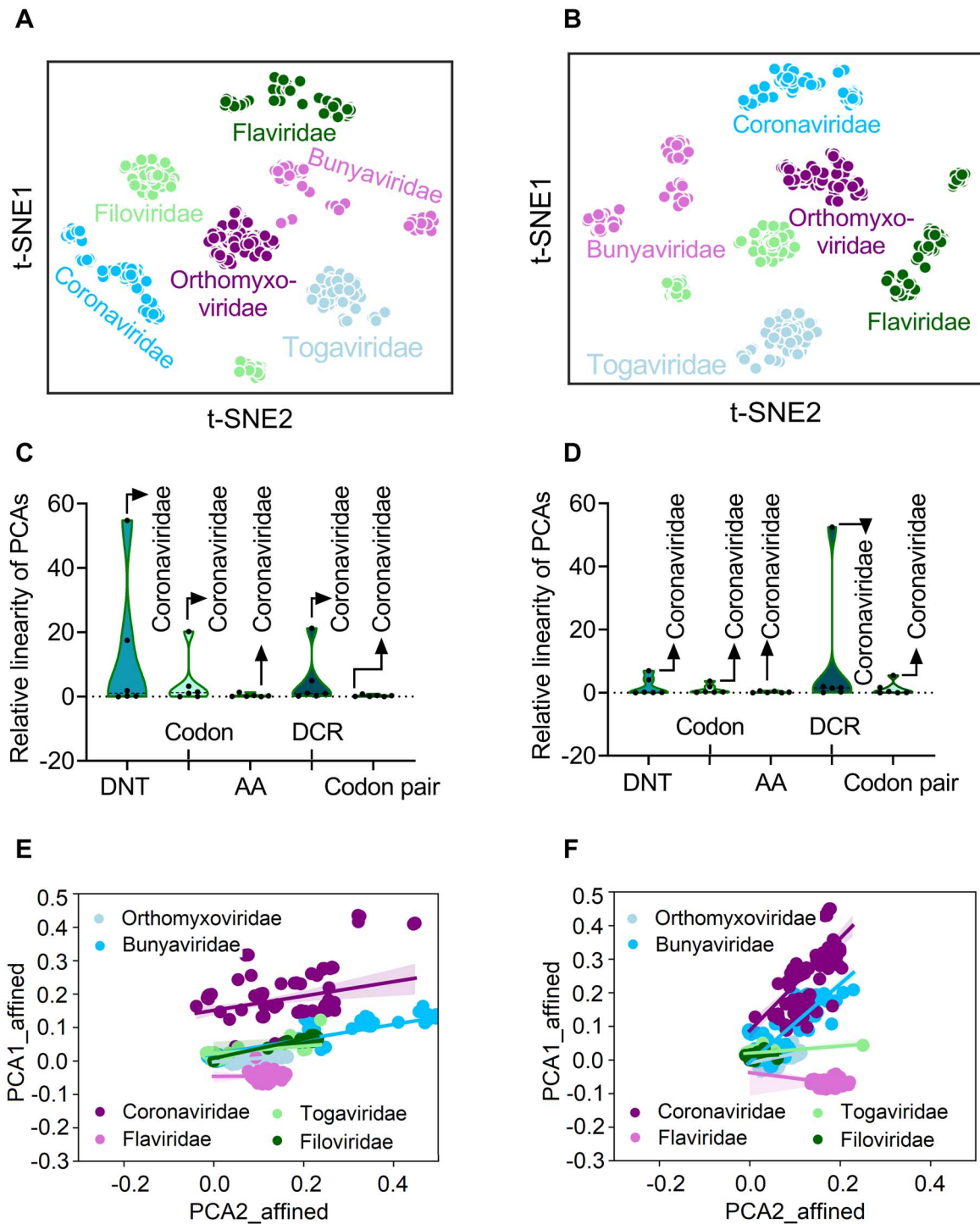


Figure 2. Projection of nucleotide compositional traits with t-SNE and PCA. Plot of two components reduced by t-SNE from 48 DNT features (A) or 1536 DCR features (B) for six ssRNA virus families, DCR and codon pairs of RdRp (C) or Gp (D) for the six virus families. The plot represents PCA1/PCA2 ratio of the absolute difference value for truncated data between the 20% and 80% quantiles. PCA1 and PCA2 were affined in space with the data point of least PCA1 value as the coordinate origin and were then calculated for simple linear regression for RdRp (E) and Gp (F), respectively.

2021 showed a significant temporal increase in the Δ -probability for the type II samples, compared to the Δ -probability for the type I samples (Figure 5F). Taken together, the DCR-based 3D-CNN results predicted a dominant type I human adaptation of SARS-CoV-2 Alpha VOCs, and a dominant type II human adaptation of Beta, Delta and other SARS-CoV-2 variants.

Important genomic features of SARS-CoV-2 VOCs

The AA mutations of each SARS-CoV-2 variant have been listed by GISAID (<https://www.gisaid.org/>). We aimed to analyse the differences in the DNT of each nucleotide site in the S sequence between VOCs and other VOC and VOI variants. A dot product was generated for each DNT vector between the two groups of variants (Additional

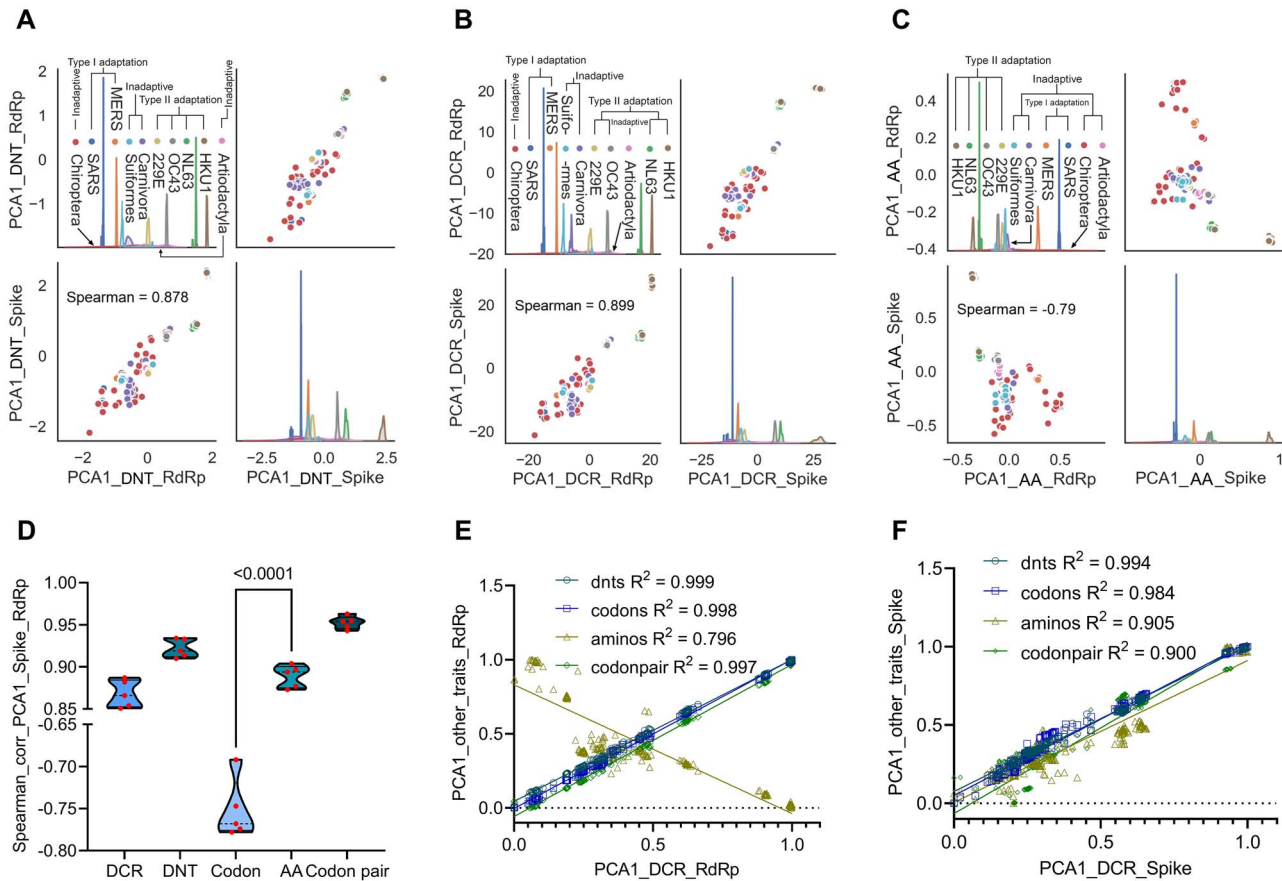


Figure 3. General adaptation and representativeness of DCR for nucleotide compositional traits of CoVs. (A–C) Simple linear regression between the PCA1 component reduced from the features of RdRp and S for the trait of DNT (A), DCR (B) or AA (C) for randomly sampled CoVs. (D) Spearman correlation of the PCA1 values of RdRp and Spike for all five compositional traits for five sampled CoVs. (E and F) Representativeness was calculated as regression dependence on each of the five traits based on the remaining four traits of CoV RdRp (E) and Spike (F).

file 1: [Supplementary Tables S3](#) and [S4](#)). A list of 23 significantly different nucleotide sites was obtained with a normalized dot product value less than 0.1 (Figure 6A) or among all sites (Additional file 1: [Supplementary Table S5](#) available online at <http://bib.oxfordjournals.org/>). The numbers for each type of DNTs in both groups were plotted and marked DNT bias was indicated for each site (Figure 6B). Biased DNTs encoding AAs mutations revealed the dominant mutations D1118H, A570D, P681H, S982A, T716I and N501Y (Figure 6C). Thus, the 3D-CNN based on DCRs could discriminate the differential DNTs, codons or AAs of sequences with both types of adaptation labels.

Adaptation prediction of SARS-CoV-2 omicron VOCs

The most recent VOC of SARS-CoV-2 B.1.1.529 (Omicron) was first reported to the WHO by South Africa on 24 November 2021 ([https://www.who.int/news/item/26-11-2021-classification-of-omicron-\(b.1.1.529\)-sars-cov-2-variant-of-concern](https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern)) [26]. Little is known about the transmissibility and pathogenicity of Omicron, which contains many mutations. We performed adaptation prediction of Omicron VOCs to assess their potential transmissibility. A total of 148 of 157 total

Omicron S sequences of high quality reported up to 6 December 2021, without any ambiguous nucleotides, were predicted to show type II adaptation (94.27%; Figure 7A; Additional file 1: [Supplementary Table S6](#) available online at <http://bib.oxfordjournals.org/>). The Omicron VOC with type II adaptation was first collected (high-quality sequence) on 9 November 2021, in South Africa (HCoV-19/South Africa/NICD-N21437/2021, EPI_ISL_6913991) and has since been found worldwide (Figure 7B). The margin score of type II over type I was only 0.0072, although the score difference was statistically significant (Figure 7C).

Discussion

Numerous artificial intelligence frameworks have been developed in different areas of biological inquiry to predict phenotypes from genomic traits. However, the causal relationship between genotype and phenotype has not been seriously considered as a key point for predictor building. Thus, it is vital to find interpretable traits in the genome for phenotype prediction. Several models have been utilized to predict SARS-CoV-2 variants based on viral protein sequences, with a particular focus on key mutant AAs related to receptor binding

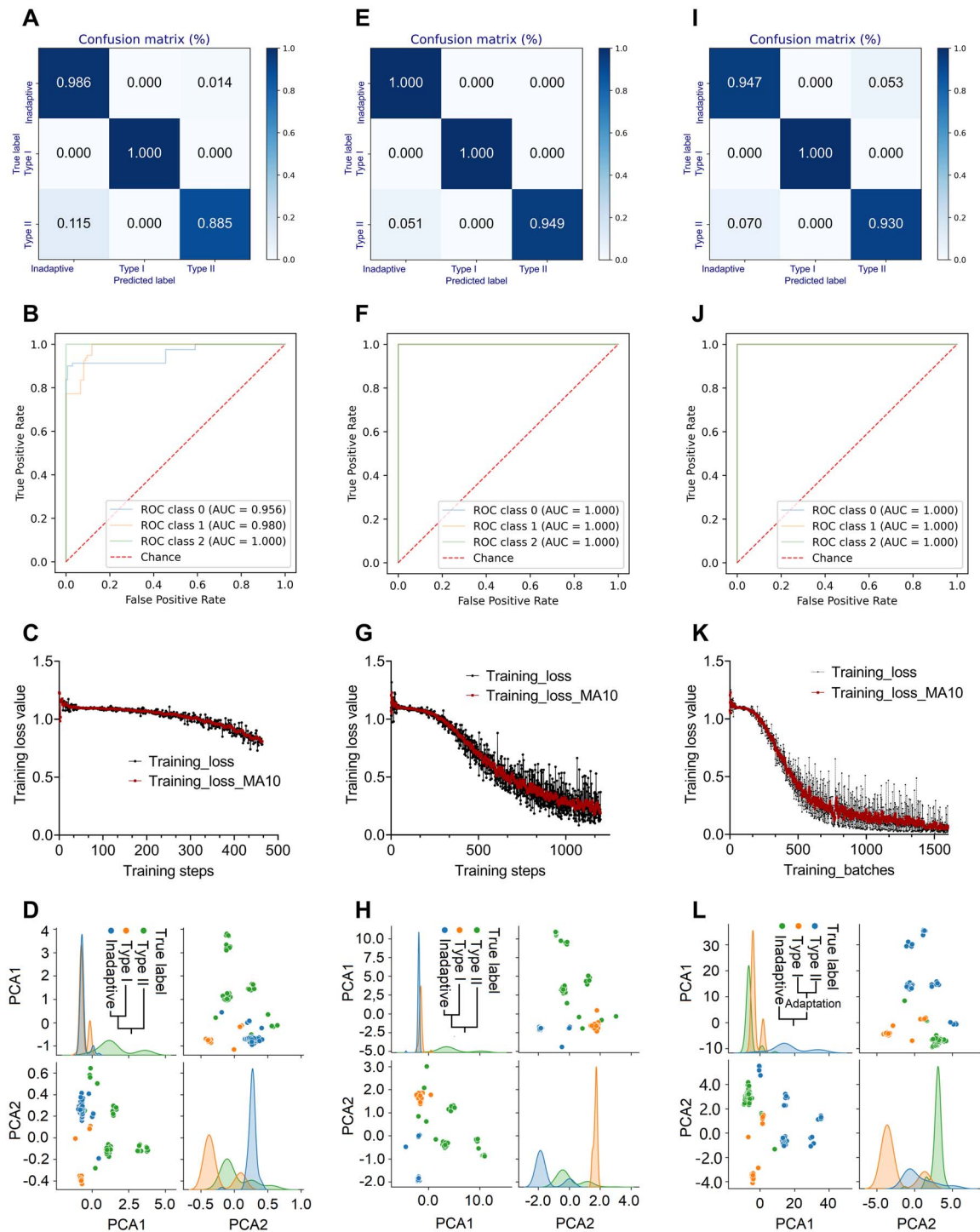


Figure 4. Performance of a CNN predictor based on DCR for human adaptation prediction. Confusion matrices (A), ROC (B), temporal training loss for training steps (C) and the pair plot of PCA1 and PCA2 of the full connection layer after the last convolution and pooling of CoV samples (D) for 15, 40 (E-H) or 50 (I-L) training epochs. The training loss was plotted as the loss value set of all steps and its moving average value.

[27–29]. However, these types of models easily fall into the trap of overfitting and are not suitable for predictions of future possible CoVs other than SARS-CoV-2. The present study aimed to build a more robust model based on general genome features at the viral RNA level. Adaptive determinants have recently been widely identified at the nucleic acid level (genomic DNA, RNA or mRNA) among pathogens such as parasites

[30], bacteria [31] and viruses [18, 19, 32, 33]. The dynamic homeostasis of genomic RNA sequences shapes the transcription, translation and decay of mRNA [34], particularly for RNA viruses. These determinants regulate the replication of pathogens in hosts via the machinery related to codon usage bias [29–32], the dinucleotide composition [19, 35], tRNA abundance [31, 33], mRNA decay [36], the translation elongation

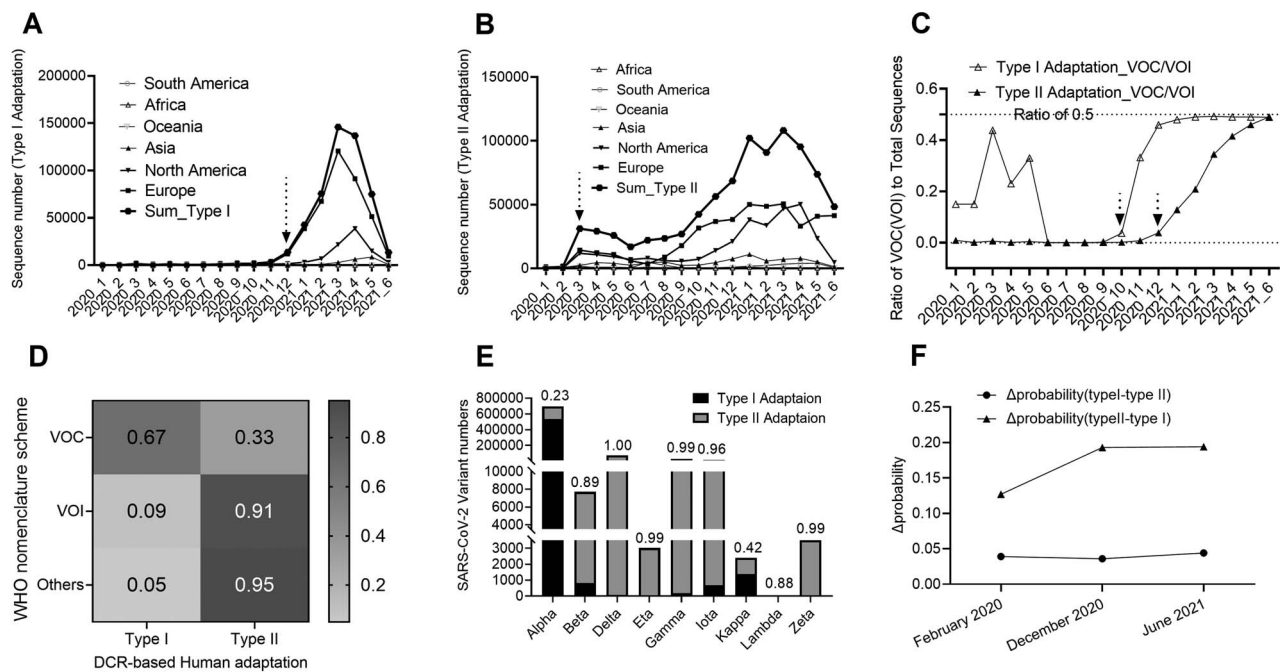


Figure 5. Prediction of the adaptation of SARS-CoV-2 by CNN based on DCR. (A–C) Monthly numbers from January 2020 to June 2021 of SARS-CoV-2 variants of type I (A) or type II (B) human adaptation in each continent, or the ratio of VOCs/VOIs to total variants of type I or type II adaptation (C). (D) Confusion matrix of VOCs, VOIs and other variants under WHO nomenclature scheme and the variants with type I and type II human adaptation, predicted by the DCR-based 3D-CNN. (E) Numbers of VOC/VOI variants with type I or type II human adaptation, the ratio of type II to total variants was annotated, respectively on each column. (F) Temporary net probability (difference for the two probability value) for the two types of variant.

speed [37] and translation efficiency [38]. Thus, the RNA sequence-based nucleotide composition is biologically meaningful and is closely related to the causal inference of virus phenotype. Family-, genus-, species- and even variant-specific determinants of the host adaptation of SARS-CoV-2 [9, 11] have been explored. Sporadic studies have described significant features of the CpG dinucleotides [39, 40] or codon usages [41] of SARS-CoV-2. Natural language processing methods have recently been shown to perform well in genomic sequence embedding [42, 43], particularly in the identification of contextual meaning in viral genomes. DNA viruses are another group of viruses posing a potential threat to human population. It is not clear if adaptation also occurs in DNA virus genomes. However, the markedly lower diversity among DNA virus genes poses a challenge in the modulation of their adaptation space.

In response to questions about CoV adaptation, the present study focused on the identification of a general genomic trait of DCR at the viral RNA level in two major genes, Gp and RdRp, of ssRNA viruses. General separability and linearity of DCR and other compositional traits were shown among the six virus families. Interviridae separation and intraviridae clustering of Gp or RdRp samples were observed. The continuity and differentiability of these traits supported the potential of these for machine/deep learning classification [25]. A high linear distribution of each of these traits was also found for both genes among the six virus families, particularly in DCR. Moreover, we confirmed the assumed general

adaptation of DCR and other compositional traits, with a strong Gp–RdRp correlation in DCR and a high representativeness value of DCR against any of the other four types of compositional traits. Thus, DCR played its role in genomic sequence embedding well, representing local contextual semantics to project genes in a nucleotide compositional space with the same vector length, with a higher sequence length comparability.

The current WHO nomenclature scheme, which is based on the use of significant AA substitutions to facilitate the risk assessment of SARS-CoV-2 variants, is hysteretic due to the monitoring period. As of June 2021, 43.11% of SARS-CoV-2 samples in GISAID were not labelled, although D614G, N501Y and other mutations were widely distributed among these variants. Based on biologically interpretable meaning, the DCR-based 3D-CNN predictor was promising for the prediction of CoV adaptation. The SARS-CoV-2-excluded training data, after intra-DCR-type convolution without inter-DCR-type convolution, clearly classified CoVs as showing inadaptation, type I or type II human adaptation, as validated by randomly sampled validation data. Surprisingly, 60.96% of SARS-CoV-2 variants recorded since December 2019 were predicted to show type II human adaptation, although SARS-CoV-2 is phylogenetically most closely related to SARS CoVs [6, 44], which are defined as showing type I human adaptation. According to the WHO nomenclature scheme, SARS-CoV-2 VOCs pose a greater pandemic risk. Interestingly, the DCR-based 3D-CNN predictor recognized 67% of VOCs/VOIs

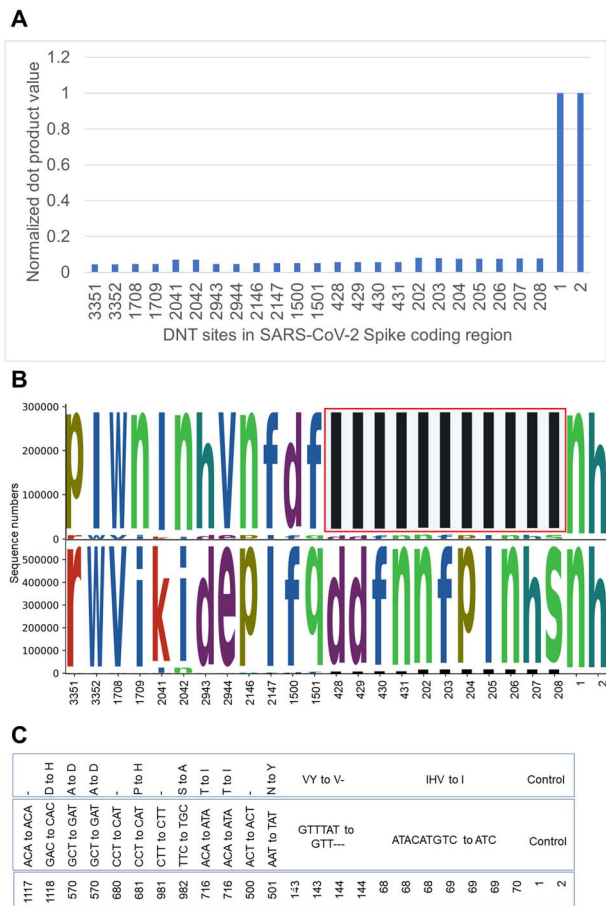


Figure 6. DNT, Codons and AAs important for type I human adaptation of SARS-CoV-2 VOCs. **(A)** Normalized dot product value of the first 25 AA sites markedly different between type I SARS-CoV-2 VOCs and variants other than VOCs/VOIs of type II human adaptation. **(B)** Logo plot of the 25 sorted DNTs markedly different between the two groups of SARS-CoV-2 variants (above and below x axis, respectively; {TT: d, TC: e, TA: f, TG: h, T:-, CT: i, CC: k, CA: l, CG: m, C:-, AT: n, AC: p, AA: q, AG: r, A:-, GT: s, GC: v, GA: w, GG: y, G:-, -T: -, -C: -, -A: -, -G: -, -:-}). **(C)** Mutation of AAs and codons in the abovementioned 25 corresponding sites for type I human adaptation.

(mainly Alpha and Kappa) as showing type I adaptation but recognized 91% VOIs and 95% of variants other than VOCs and VOIs as showing type II adaptation. According to our predictions, it is reasonable to deduce that SARS-CoV-2 VOCs present relatively higher pathogenicity, while most SARS-CoV-2 variants, SARS-CoV-2 VOIs and other variants exhibit typical ‘common cold’-like CoV adaptation. Considering the obvious timeliness advantage, DCR-based 3D-CNN shows strong potential to assess the risk of any emerging variant with novel mutations in real time, without a monitoring period. Most variants with type I adaptation were found from January to April 2021 in Europe and America, whereas the variants with type II adaptation were distributed over a wider temporal and spatial range. Additionally, there was only a marginal probability advantage of these variants exhibiting type I human adaptation, while a markedly higher probability advantage was observed for type II adaptation. Taken together, the data indicate that the SARS-CoV-2 variant population is mainly characterized

by higher transmissibility and lower pathogenicity than other ‘common cold’-type CoVs. This model also predicted the most recent Omicron VOC to show type II adaptation, indicating potentially high transmissibility and low pathogenicity of this new VOC member.

In addition, since compositional DCR or DNT traits could not identify the specific determinant mutations contributing to type I or II adaptation, it was reasonable to perform a comparative analysis of different features of the two groups of variants. Dot products for each pair of DNT vectors at each nucleotide site identified the significantly different mutations D1118H, A570D, P681H, S982A, T716I and N501Y, implying their contribution to SARS-CoV-2 VOCs and other variants with type I adaptation. Such performance in discriminating key mutations and classifying SARS-CoV-2 variants based on the general CoV-adaptive DCR-based predictor is encouraging, as it implies the potential to assess the risk of emerging CoVs by predicting adaptation.

Conclusions

In summary, our study provides a strategy for parsing the general genomic traits of RNA viruses and accordingly building a deep-learning predictor to assess the risk of emerging viruses. The DCR-based 3D-CNN predictor provides real-time predictions of emerging SARS-CoV-2 variants, facilitating the risk assessment of SARS-CoV-2 variants and the control of the current COVID-19 pandemic.

Methods

Data preprocessing and genomic compositional trait parsing of ssRNA viruses

Full sequences of six families/orders of RNA viruses were parsed from GenBank files with the script `DCR_scripts`. `Data_parsing` (<https://github.com/Jamalijama/Dinucleotide-Composition-Representation-DCR->) or randomly sampled from past influenza viruses (Orthomyxoviridae) [19, 45]. Detailed information on all six families of viruses involved in this study was provided under the environmental variable `Supporting_data_Full_DCR_6Viridae` under the data of `DCR_scripts`. CoV coding sequences were labelled with host information of Primates, Chiroptera, Carnivora, Artiodactyla or the Suiformes sub-order for viruses originating from humans, bats, canines/felines, bovines and other Artiodactyla [46–48] or swine, respectively. Sequence length filtering was performed with down and up thresholds of 27 000 and 32 000 bp, respectively. After sequence length filtering, sequences were counted for compositional DCR, DNT, AAs, codons and codon pairs for each sequence sample with the script `DCR_scripts.DCR_counting_sampling`. For each compositional trait, every type of feature was calculated as a frequency relative to the total. DNT and DCR were counted depending on each nucleotide in a codon. Each of the 16 types of DNTs [19] for nucleotide

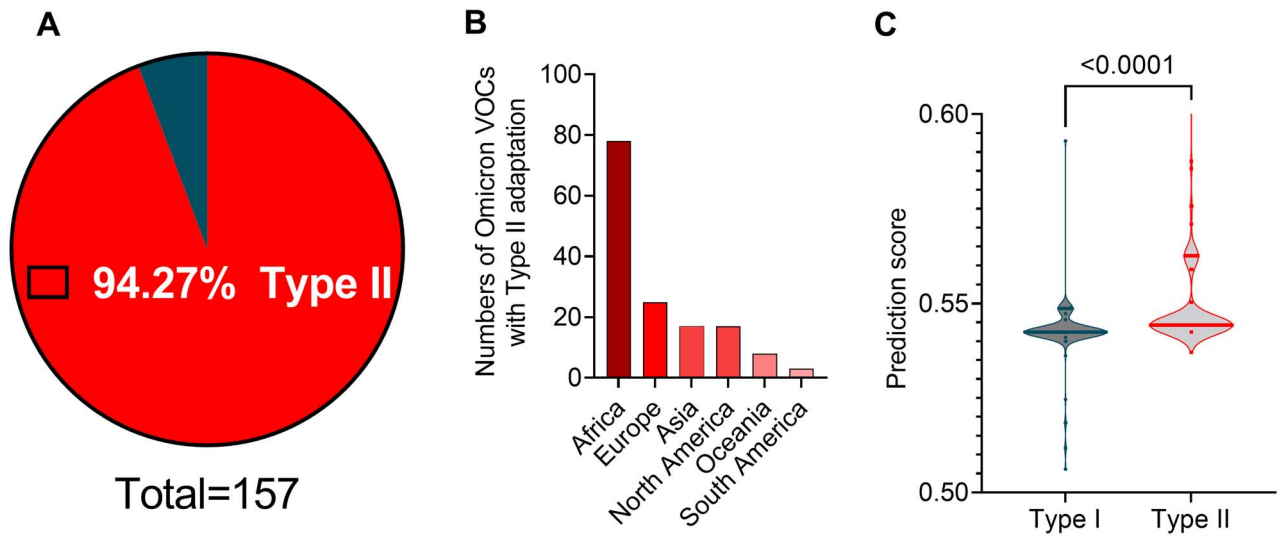


Figure 7. Human adaptation of SARS-CoV-2 Omicron VOCs. (A) Numbers and ratio of SARS-CoV-2 Omicron VOCs with type I and type II adaptation. (B) World distribution of Omicron VOCs with type II adaptation. (C) Prediction score for type I and type II adaptation (score for the label of inadaptive not shown) of Omicron VOCs with type II adaptation.

(nt) 1, 2 or 3 of a codon was counted as follows:

$$\text{Freq (DNT}_{ij}) = \frac{\sum (\text{DNT}_{ij})}{\sum_{i=1}^{16} \text{DNT}_{ij}}, i = 1 \text{ to } 16 \text{ for DNT1} - 16, j$$

$$= 1 \text{ to } 3 \text{ for nt1-3 in codon}$$

DCR methods were designed to embed contextual codon nucleotide-dependent DNTs into a compositional space as three types of paired DNTs at the meta position, with a starting site for the first DNT of 1, 2 or 3 in the codon, and the other three types of paired DNTs at the ortho position (one nucleotide apart). The former DCR set represented the tetranucleotide with its first nucleotide starting from codon site 1, 2 or 3. The last DCR set represented two pairs of DNTs starting at site 1 or 2 in two neighbouring codons and the bridge DNT of the two neighbouring codons. DCR was counted as a frequency value as follows:

$$\text{Freq (DCR}_{n,m}) = \frac{\sum \text{DCR}_{n,m}}{\sum_{m=1}^{256} \text{DCR}_n}, n = 1 \text{ to } 6 \text{ for DCR set}$$

$$1 - 6, m = 1 \text{ to } 256 \text{ for DCR1} - 256$$

The composition of each AA, codon or codon pair was counted as a frequency value for the 20 types of AAs, 64 types of codons and 3721 types of codon pairs (61 paired codons without the three stop codons). Five compositional traits of each sequence were calculated with the script under the environment variable of DCR_scripts/DCR_counting_sampling, according to the GitHub script document (<https://github.com/JamaliJama/Dinucleotide-Composition-Representation->

DCR-). Two hundred Gp and RdRp samples of composition data were randomly sampled with the sampling script.

Unsupervised learning of DCR and other compositional traits

PCA and t-SNE were performed to observe the distribution of compositional traits. Two components (PCA1 and PCA2 or t-SNE1 and t-SNE2) were reduced from the features of 48 DNTs, 1536 DCRs, 20 AA, 64 codons or 3971 codon pairs for Gp or RdRp for the six families of ssRNA viruses. For correlations between the PCA values of Gp and RdRp in each of the five traits or between the PCA values of features between one and any of the other traits, PCA was performed with one component. Reduced PCA (PCA1 and PCA2) or t-SNE (t-SNE1 and t-SNE2) values were visualized with the Python-Seaborn model. Affine transformation was performed to project PCA1 and PAC2 values in space with a coordinate point of the minimum PCA1 value of features for each trait and its PCA2.

$$\text{Affine function : Aff}(x) = x - x_{\min}$$

The general adaptation of the compositional traits in both CoV genes was evaluated to determine whether there was a similar feature distribution of a trait between the two genes. Thus, the reduced PCA1 values for both S and RdRp were plotted in pairs, with host information labelled for each data point. Spearman's correlation was used to quantify such adaptation, with five repeats of randomly sampled data. The representativeness of DCR for other traits was examined by linear regression for both RdRp and S genes based on the sampled data, with the script under the environment variable of Analysis/DCR_scripts. Analysis details are available

at GitHub (<https://github.com/JamaliJama/Dinucleotide-Composition-Representation-DCR->).

Model architecture of a 3D-CNN

To avoid the influence of number-imbalance among sequence samples with different host labels, down- and up-sampling were performed with the Python scripts `pandas.DataFrame.sample` and `python imblearn.over_sampling.SMOTE`, respectively. *S* samples of CoVs (excluding SARS-CoV-2) were randomly divided into a training-set for supervised classifier training and a validation-set for classification validation (test size=0.3). SARS-CoV-2 *S* was subjected to adaptation prediction with the trained 3D CNN model. Both the training and validation data were reshaped into a (6, 16, 16) array with one channel. The 3D CNN was constructed with three layers of convolution calculations with 8, 16 and 32 out channels, with a stride of one and a padding of one. A $6 \times 16 \times 16$ matrix was convoluted with a ReLU function into a flattened vector of 768, which was activated with a sigmoid function and reduced to a vector of 192. Finally, the reduced vector was linearized into three predicted values, and adaptation was predicted with the function `max` (value 1, value 2, value 3) for the adaptation label set of inadaptation, type I adaptation and type II adaptation. For validation, the host with the greatest probability was considered the adaptation host. To validate the performance of the 3D-CNN, confusion matrices and micro-average ROC curves with AUCs were drawn [49, 50]. The detailed scripts are under the variables of `3D_CNN_training` and `3D_CNN_predicting` of `DCR_scripts`.

Sigmoid function : $f(x) = 1 / (1 + e^{-x})$

ReLU function : $f(x) = \max(0, w^T x + b)$

Temporal and spatial host adaptation prediction

To assess the temporal and spatial adaptation of SARS-CoV-2 VOCs, VOIs and other variants, we extracted data on the collection year and month and continent information from the spatial labels of country and area. The *S* sequence number of SARS-CoV-2 samples with their predicted adaptation label (type I or II adaptation) was temporally and spatially plotted. The ratios of SARS-CoV-2 samples of VOCs/VOIs and total SARS-CoV-2 variants were also temporally plotted. A confusion matrix calculation was performed to compare the adaptation predictions (type I and II adaptation) to the WHO nomenclature of VOCs, VOIs and other variants. Greek alphabet variants with a predicted adaptation label were stack plotted with the ratio of variants showing type II adaptation to the total annotated variants. Statistics of probability were calculated for the two types of variant adaptation; the probability distribution was plotted for the two groups

of samples in the sampling months of February 2020, August 2020 and April 2021.

Analysis of adaptation-related DNTs and amino acids

To evaluate the specific DNT or amino acid differences between the type I and II adaptation groups, all SARS-CoV-2 *S* (Gp) cDNA sequences were transformed into DNT sequences according to a defined transformation table (Supplementary file 'DNT_transforming_dict.txt' available online at <http://bib.oxfordjournals.org/>) or were translated to AA sequences. For DNT sequences with 3821 DNTs (the last nucleotide was not transformed for a cDNA sequence with 3822 nts) or protein sequences with 1273 amino acids, each DNT/AA was calculated for each site for sequences with a label of type I or type II adaptation. A vector with serial 17 (16 types of DNTs and one deletion sign of '-') or 21 (20 types of DNTs and one deletion sign of '-'), count numbers were normalized with a normalization function as follows:

$$f(x) = \left(x - x_{\min(\text{axis}=0)} \right) / \left(x_{\max(\text{axis}=0)} - x_{\min(\text{axis}=0)} \right)$$

To compare the DNT or AA distribution at each site for the two groups of sequences, a dot product was calculated for each pair of 3821 DNT counting vectors or of 1273 amino acid counting vectors for the two groups of *S* sequences, and the product value was used as an index of vector similarity (DNT/AA distribution similarity). A logo plot of the top 23 significantly different DNTs was drawn using Logomaker [51] according to the DNT frequency. Significantly different AAs at the AA sites corresponding to the 23 DNT sites are also listed. The scripts are under the environmental variable `DCR_scripts\Feature_importance`.

Key Points

- Six families of ssRNA viruses show general separability and linearity, and both the CoV spike and RdRp genes show host specificity according to DCR.
- 3D CNNs based on DCR of the spike gene of other CoVs predict two types of human adaptation of SARS-CoV-2 variants.
- Alpha SARS-CoV-2 VOCs present SARS-CoV-like human adaptation, while Delta, Beta and Omicron VOCs present 'common cold' CoV-like human adaptation.

Supplementary data

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

Acknowledgement

We gratefully acknowledge the scientific community on the GISAID EpiCoV platform and all contributing experts

in SARS-CoV-2 sequences; detailed sequence information is listed in the Supplementary Data available online at <http://bib.oxfordjournals.org/>.

Author contributions

J.L. conceived the study and performed model building. J.L., S.Z., Y.N.W. and X.P.K. collected data and performed data analysis. J.L. and T.J. supervised the project. J.L. wrote the manuscript with help from S.Z., Y.N.W. and T.J.

Data and code availability

The full genomic sequences and other detailed information of three -ssRNA order/families of Bunyavirales and Filoviridae and three +ssRNA virus families Flaviviridae, Togaviridae and Coronaviridae available up to November 2019 were downloaded from the nucleotide database (<https://www.ncbi.nlm.nih.gov/nucleotide>). Orthomyxoviridae sequences of IAVs were downloaded from the Influenza Research Database (IRD) [45]. Coding sequences of Gp and RdRp were parsed by using Python scripts. SARS-CoV-2 sequences from December 2019 to June 2021 were downloaded from GISAID [52]; audacity multiple sequence alignment (unmasked) (<https://www.epicov.org/epi3/cfrontend>); SARS-CoV-2S (Gp) was parsed with a fixed length of 3822 nts. Key scripts for nucleotide composition counting, data analysis, visualization and model training models that support the findings of this study have been deposited in GitHub (<https://github.com/Jamalijama/Dinucleotide-Composition-Representation-DCR->).

Funding

National Natural Science Foundation of China (grant no. 32070166) and the Capital's Funds for Health Improvement and Research (grant no. 2021-1G-4302).

References

1. Taubenberger JK, Kash JC. Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe* 2010;**7**:440–51.
2. Simmonds P, Aiweisakun P, Katzourakis A. Prisoners of war—host adaptation and its constraints on virus evolution. *Nat Rev Microbiol* 2019;**17**:321–8.
3. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* 2019;**17**:181–92.
4. Fung TS, Liu DX. Human coronavirus: host-pathogen interaction. *Annu Rev Microbiol* 2019;**73**:529–57.
5. Lim YX, Ng YL, Tam JP, et al. Human coronaviruses: a review of virus-host interactions. *Diseases* 2016;**4**:26.
6. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**:270–3.
7. Neuzil KM. Interplay between emerging SARS-CoV-2 variants and pandemic control. *N Engl J Med* 2021;**384**:1952–4.
8. Hu J, Peng P, Wang K, et al. Emerging SARS-CoV-2 variants reduce neutralization sensitivity to convalescent sera and monoclonal antibodies. *Cell Mol Immunol* 2021;**18**:1061–3.
9. Ozono S, Zhang Y, Ode H, et al. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat Commun* 2021;**12**:848.
10. Arora P, Pohlmann S, Hoffmann M. Mutation D614G increases SARS-CoV-2 transmission. *Signal Transduct Target Ther* 2021;**6**:101.
11. Cele S, Gazy I, Jackson L, et al. Escape of SARS-CoV-2 501Y.V2 from neutralization by convalescent plasma. *Nature* 2021;**593**:142–6.
12. Gomez CE, Perdiguero B, Esteban M. Emerging SARS-CoV-2 variants and impact in global vaccination programs against SARS-CoV-2/COVID-19. *Vaccines (Basel)* 2021;**9**:243.
13. Hacisuleyman E, Hale C, Saito Y, et al. Vaccine breakthrough infections with SARS-CoV-2 variants. *N Engl J Med* 2021;**384**:2212–8.
14. Wang R, Chen J, Hozumi Y, et al. Emerging vaccine-breakthrough SARS-CoV-2 variants. *ArXiv* 2021;**9**:1–15, 2109.04509v1.
15. Rambaut A, Holmes EC, O'Toole A, et al. Addendum: a dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2021;**6**:415.
16. Zhang J, Cai Y, Xiao T, et al. Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* 2021;**372**:525–30.
17. Jiang S, Du Q, Feng C, et al. CompoDynamics: a comprehensive database for characterizing sequence composition dynamics. *Nucleic Acids Res* 2022;**50**:D962–9.
18. Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science* 2018;**362**:577–80.
19. Li J, Zhang S, Li B, et al. Machine learning methods for predicting human-adaptive influenza A viruses based on viral nucleotide compositions. *Mol Biol Evol* 2020;**37**:1224–36.
20. Hu J, Chen M, Zhou X. Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. *Nucleic Acids Res* 2021;**1**–16, gkab1147, <https://doi.org/10.1093/nar/gkab1147>.
21. Hie B, Zhong ED, Berger B, et al. Learning the language of viral evolution and escape. *Science* 2021;**371**:284–8.
22. Alley EC, Khimulya G, Biswas S, et al. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315–22.
23. Van Der ML, HG. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**86**:2579–605.
24. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* 2016;**374**:20150202.
25. Kim D, Kim SH, Kim T, et al. Review of machine learning methods in soft robotics. *PLoS One* 2021;**16**:e246102.
26. CDC COVID-19 Response Team. SARS-CoV-2 B.1.1.529 (Omicron) variant—United States, December 1–8, 2021. *Morb Mortal Wkly Rep* 2021;**70**:1731–4.
27. Zahradnik J, Marciano S, Shemesh M, et al. SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution. *Nat Microbiol* 2021;**6**:1188–98.
28. Pucci F, Rooman M. Prediction and evolution of the molecular fitness of SARS-CoV-2 variants: introducing SpikePro. *Viruses* 2021;**13**:935.
29. Chen J, Gao K, Wang R, et al. Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. *Chem Sci* 2021;**12**:6929–48.
30. Forsberg R, Christiansen FB. A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Mol Biol Evol* 2003;**20**:1252–9.
31. Charles H, Calevro F, Vinuelas J, et al. Codon usage bias and tRNA over-expression in *Buchnera aphidicola* after aromatic amino acid

- nutritional stress on its host *Acyrtosiphon pisum*. *Nucleic Acids Res* 2006;**34**:4583–92.
32. Bahir I, Fromer M, Prat Y, et al. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol* 2009;**5**:311.
 33. Chen F, Wu P, Deng S, et al. Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat Ecol Evol* 2020;**4**:589–600.
 34. Hausser J, Mayo A, Keren L, et al. Central dogma rates and the trade-off between precision and economy in gene expression. *Nat Commun* 2019;**10**:68.
 35. Upadhyay M, Samal J, Kandpal M, et al. CpG dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. *J Virol* 2013;**87**:13816–24.
 36. Contu L, Balistreri G, Domanski M, et al. Characterisation of the Semliki Forest virus-host cell interactome reveals the viral capsid protein as an inhibitor of nonsense-mediated mRNA decay. *PLoS Pathog* 2021;**17**:e1009603.
 37. Hershberg R, Petrov DA. Selection on codon bias. *Annu Rev Genet* 2008;**42**:287–99.
 38. Duret L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* 2000;**16**:287–9.
 39. Xia X. Extreme genomic CpG deficiency in SARS-CoV-2 and evasion of host antiviral defense. *Mol Biol Evol* 2020;**37**:2699–705.
 40. Pollock DD, Castoe TA, Perry BW, et al. Viral CpG deficiency provides no evidence that dogs were intermediate hosts for SARS-CoV-2. *Mol Biol Evol* 2020;**37**:2706–10.
 41. Roy A, Guo F, Singh B, et al. Base composition and host adaptation of the SARS-CoV-2: insight from the codon usage perspective. *Front Microbiol* 2021;**12**:548275.
 42. Alley EC, Khimulya G, Biswas S, et al. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315–22.
 43. Rao R, Bhattacharya N, Thomas N, et al. Evaluating protein transfer learning with TAPE. *Adv Neural Inf Process Syst* 2019;**32**:9689–701.
 44. Andersen KG, Rambaut A, Lipkin WI, et al. The proximal origin of SARS-CoV-2. *Nat Med* 2020;**26**:450–2.
 45. Zhang Y, Aevermann BD, Anderson TK, et al. Influenza research database: an integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res* 2017;**45**:D466–74.
 46. Watzinger F, Mayr B, Haring E, et al. High sequence similarity within ras exons 1 and 2 in different mammalian species and phylogenetic divergence of the ras gene family. *Mamm Genome* 1998;**9**:214–9.
 47. Eckerle I, Corman VM, Müller MA, et al. Replicative capacity of MERS coronavirus in livestock cell lines. *Emerg Infect Dis* 2014;**20**:276–9.
 48. Gafer JA, Thanaa HK, Madboly M, Salem HA. Genetic detection and pathological finding of BVDV and BHV-1 in camel calves. *Assiut Vet Med J* 2015;**61**:34–45.
 49. Townsend JT. Theoretical analysis of an alphabetic confusion matrix. *Percept Psychophys* 1971;**9**:40–50.
 50. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2005;**27**:861–74.
 51. Ammar T, Kinney JB. Logomaker: beautiful sequence logos in python. *Bioinformatics* 2019;**7**:2272–4.
 52. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 2017;**1**:33–46.