Supplementary Information

Alignment free sequence comparison methods and reservoir host prediction

Bill Lee[1], David K. Smith[1]*, Yi Guan[12]*

[1] State Key Laboratory of Emerging Infectious Diseases, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Rd., Pok Fu Lam, Hong Kong

[2] Joint Institute of Virology (Shantou University and The University of Hong Kong), Guangdong-Hongkong Joint Laboratory of Emerging Infectious Diseases, Shantou University, Shantou, P. R. China.

* Correspondence to DKS dsmith@hku.hk Orcid Id 0000-0003-3608-0586, or YG yguan@hku.hk Orcid-Id 0000-0001-6057-9243

The MLDSP algorithm (Randhawa, et al., 2019; Randhawa, et al., 2020a) was conducted in three steps: (1) Create a numerical representation of the R/DNA sequences, each nucleotide in a viral sequence was converted into a number, or a Chaos Game Representation (CGR) with a k-tuple size of 7 (Jeffrey, 1990; Karamichalis, et al., 2015; Randhawa, et al., 2020b) of the sequence was constructed; (2) Perform a Discrete Fourier Transform (DFT) of each (now numerical) sequence, and (3) Calculate the Pearson Correlation Coefficient (PCC) of each pair of sequences.

For each of n sequence $S_i = (S_i(0), S_i(1), \ldots, S_i(p-1))$, where $1 \leqslant i \leqslant n$, $S_i(k) \in \{A, C, G, T\}$, $0 \leqslant k \leqslant p-1$, the one-dimensional discrete numerical representation (discrete digital signal) N_i, defined as $N_i = f(S_i(0)), f(S_i(1)), \ldots, f(S_i(p-1))$, was calculated, where for each $0 \leqslant k \leqslant p-1$, the quantity $f(Si(k))$ is the value under the numerical representation f of the nucleotide in the position k of the viral sequence S_i. The two dimensional CGR numerical form of the sequence is an ordered (by CGR convention) frequency of all k-tuples (Karamichalis, et al., 2015) of (in this case) size 7, so that $S_i = (S_i(0), S_i(1), \ldots, S_i(p-1))$ where $p = 4^7 - 1$ (the number of 7-tuples, rather than the length of the sequence) and $f(Si(k))$ is the frequency of 7-tuple 'k' in sequence i. This representation is frequently presented graphically as a $2^7 \times 2^7$ plot/heat map.

Then the DFT of the signal N_i was computed as the vector F_i, where for $0 \leqslant k \leqslant p-1$,

$$F_i(k) = \sum_{j=0}^{p-1} f\left(S_i(j)\right) \cdot e^{(-2\pi i/p)kj}$$

For each $0 \leqslant k \leqslant p-1$, the magnitude vector, M_i, representing the signal N_i, has $M_i(k)$ as the absolute value of $F_i(k)$, i.e. $M_i(k) = |F_i(k)|$. The PCC was calculated between two magnitude vectors. The PCC between X and Y is defined as:

$$r_{XY} = \frac{\sum_{i=0}^{p-1}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=0}^{p-1}\left(X_i - \overline{X}\right)^2} \times \sqrt{\sum_{i=0}^{p-1}\left(Y_i - \overline{Y}\right)^2}}$$

where the average $\overline{X}$ is defined as:

$$\overline{X} = \left(\sum_{i=0}^{p-1} X_i\right)/p$$

The same rule applies to the average $\overline{Y}$. PCC values were normalized by taking (1 - Υ_{XY})/2 to obtain distance values between 0 and 1, where 0 corresponds to completely correlated signals and 1 to completely negatively correlated signals.

• The viral genomic traits GT described in (Babayan, et al., 2018) contain codon pair score (CPS), dinucleotide biases, codon biases, and amino acid biases.

• The CPS was calculated as:

$$CPS = \ln\left(\frac{AB}{\frac{A\times B}{X\times Y}\times XY}\right)$$

where *A* and *B* represent the observed counts of codons *A* and *B*, *AB* represents the observed count of codon pair *AB*, *X* and *Y* represent the observed counts of the corresponding amino acids *X* and *Y*, and *XY* represents the observed count of amino acid pair *XY* across all coding sequences of the selected species isolate.

• Dinucleotide bias for each of the 16 possible dinucleotides was obtained using the following formula:

$$\frac{\left(\frac{N_{XY}}{DN_{tot}}\right)}{\left(\frac{N_X}{N_{tot}}\times\frac{N_Y}{N_{tot}}\right)}$$

where N_X and N_Y are the total numbers of nucleotides *X* and *Y*, N_{tot} is the total number of nucleotides, N_{XY} is the total number of the dinucleotide *XY*, and DN_{tot} is the total number of dinucleotides across all coding sequences of the selected isolate. Dinucleotide biases at "bridge" (between adjacent codons) and "non-bridge" codon positions were also calculated.

• Codon bias for each of the 64 codons was calculated by dividing the total count of each codon by the total count of all codons that encode the corresponding amino acid or stop codon across all coding sequences of the selected isolate.

• Amino acid bias for each of the 21 amino acids (including stop codon) was calculated by dividing the total count of each amino acid by the total number of amino acids in the isolate.

• Bat taxonomy followed (Hutcheon and Kirsch, 2006) - *Pteropodiformes* (families *Pteropodidae, Rhinolophidae, Hipposideridae, Megadermatidae, and Rhinopomatidae*, also referred to as Pterobats) and *Vespertilioniformes* (the other microbat families, also referred to as Vespbats) as also used in (Babayan, et al., 2018).

• Examination of Chaos Game Representation (CGR): The role of a two-dimensional (CGR) vs one-dimensional (P/P) representation was examined, using the Dugbe virus isolate ArD 44313 segment S, (complete sequence) and Calicivirus pig/NC-WGP93C/USA/2009 polyprotein and minor basic structural protein VP2 genes (complete cds) sequences from the virus dataset and 3 artificial random deletions from each sequence to make two sets of related sequences. Applying the MLDSP procedure and converting to "distances" yielded Supplementary Table 6, which shows generally superior performance of the 2D-CGR method at in-group and out-group recognition. This performance gain was compensated for with the GBM model and bagging in small scale testing. The detailed analysis was performed using the P/P encoding.

Supplementary References

Babayan, S.A., Orton, R.J. and Streicker, D.G. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science* 2018;362(6414):577-580.

Hutcheon, J.M. and Kirsch, J.A.W. A moveable face: deconstructing the Microchiroptera and a new classification of extant bats. *Acta Chiropterol* 2006;8(1):1-10.

Jeffrey, H.J. Chaos Game Representation of Gene Structure. *Nucleic Acids Res* 1990;18(8):2163-2170.

Karamichalis, R.*, et al.* An investigation into inter- and intragenomic variations of graphic genomic signatures. *Bmc Bioinformatics* 2015;16:246.

Randhawa, G.S., Hill, K.A. and Kari, L. ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. *BMC Genomics* 2019;20(1):267.

Randhawa, G.S., Hill, K.A. and Kari, L. MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis. *Bioinformatics* 2020a;36(7):2258-2259.

Randhawa, G.S.*, et al.* Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS One* 2020b;15(4):e0232391.

Supplementary Table 1. Description of the deep learning architecture

Layer (type)	Output Shape	Param #[a]
dense_1 (Dense)	(None, 32)	1664
dense_2 (Dense)	(None, 32)	1056
dense_3 (Dense)	(None, 16)	528
dense_4 (Dense)	(None, 16)	272
dense_5 (Dense)	(None, 8)	136
dropout_1 (Dropout)	(None, 8)	0
dense_6 (Dense)	(None, 1)	9

Total params:	3665
Trainable param:	3665
Non-trainable params:	0

[a] A total of 50 models with the same parameters (default values within Keras) were trained. The best ten models were selected for bat subgroup predictions following the bagging strategy. Bat subgroups include *Pteropodiformes* (families *Pteropodidae*, *Rhinolophidae*, *Hipposideridae*, *Megadermatidae*, and *Rhinopomatidae*) and *Vespertilioniformes* (remaining microbat families). An average accuracy of 76.67% was achieved.

Supplementary Table 2. Accuracy of virus-family predictions using MLDSP traits

Model	Accuracy (%)
LinearDiscriminant	94.9
LinearSVM	87.8
QuadraticSVM	89.3
FineKNN	83.2
SubspaceDiscriminant	93.9
SubspaceKNN	93.4
AverageAccuracy	90.4

Supplementary Table 3. The best 10 trained models of the first 50 trainings including or without M-SP

Bagged Model	Model [a]	Accuracy	Artiodactyl	Bat	Bird	Carnivore	Fish	Insect	Plant	Primate	Rodent
Including M-SP	*47*	0.830	0.800	0.888	0.818	0.666	1.000	1.000	1.000	0.625	0.916
	39	0.800	0.700	0.555	1.000	0.833	1.000	0.666	1.000	0.750	0.833
	40	0.769	0.800	0.888	0.818	0.333	0.666	0.666	1.000	0.625	0.916
	45	0.769	0.600	0.777	1.000	0.833	0.666	0.000	1.000	0.625	0.916
	50	0.769	1.000	0.777	0.818	0.500	0.666	0.666	1.000	0.500	0.833
	12	0.769	0.600	0.444	0.818	1.000	0.666	1.000	1.000	0.750	0.916
	16	0.753	0.900	0.888	0.909	0.666	0.333	1.000	1.000	0.375	0.666
	11	0.753	0.700	0.666	0.909	0.666	1.000	0.666	0.666	0.750	0.750
	32	0.738	0.700	0.888	0.818	0.666	0.666	1.000	1.000	0.375	0.750
	01	0.738	0.700	0.666	0.818	0.666	0.666	0.666	1.000	0.875	0.666
Without M-SP	*12*	0.800	0.600	0.777	0.818	1.000	0.666	1.000	1.000	0.750	0.833
	47	0.784	0.700	0.666	0.818	0.666	1.000	1.000	1.000	0.625	0.916
	39	0.769	0.600	0.444	0.909	0.833	1.000	0.666	1.000	0.750	0.916
	45	0.769	0.600	0.666	1.000	0.833	0.666	0.333	1.000	0.750	0.833
	02	0.738	0.600	0.777	0.909	0.333	1.000	1.000	1.000	0.750	0.666
	40	0.738	0.600	0.888	0.818	0.666	0.666	0.666	0.666	0.625	0.833
	03	0.723	0.800	0.444	1.000	0.833	0.666	1.000	1.000	0.500	0.583
	21	0.723	0.600	0.666	0.636	0.833	0.666	1.000	1.000	0.500	0.916
	32	0.723	0.700	0.777	0.818	0.666	0.666	1.000	1.000	0.500	0.666
	50	0.723	0.800	0.888	0.818	0.500	0.666	1.000	0.666	0.500	0.666

[a] Individual model IDs are listed.

Supplementary Table 4. Please refer to Excel file.

Supplementary Table 5. Please refer to Excel file.

Supplementary Table 6. Comparison of sequence distance calculated by MLDSP using 1-D P/P encoding (upper triangle) and 2-D CGR encoding (lower triangle)

PP/CGR	1a	1b	1c	1d	2a	2b	2c	2d
1a		0.3386	0.3626	0.4060	0.4985	0.4896	0.4521	0.4870
1b	0.0677		0.3835	0.4267	0.4930	0.4620	0.4842	0.4809
1c	0.0257	0.0947		0.4495	0.4692	0.4755	0.4921	0.4947
1d	0.1694	0.2490	0.1783		0.4932	0.4743	0.4773	0.4926
2a	0.2626	0.2760	0.2653	0.3114		0.4091	0.4245	0.4292
2b	0.2890	0.2997	0.2909	0.3331	0.0978		0.3783	0.3289
2c	0.3219	0.3333	0.3254	0.3561	0.1844	0.2220		0.3006
2d	0.2943	0.3083	0.2964	0.3379	0.1082	0.1546	0.1858	

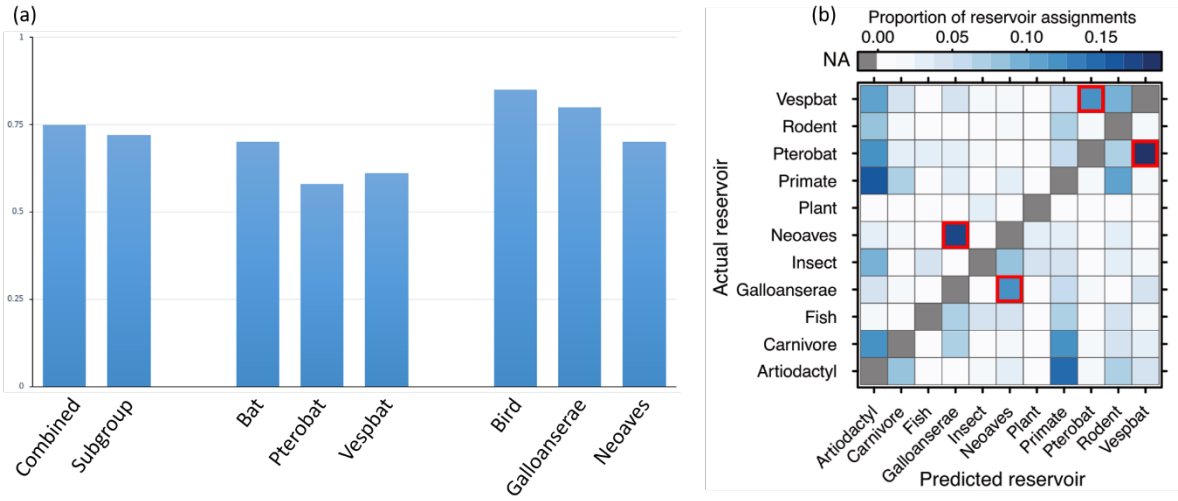
Sequence 1a Dugbe virus isolate ArD 44313 segment S, (complete sequence).

Sequence 2a Calicivirus pig/NC-WGP93C/USA/2009 polyprotein and minor basic structural protein VP2 genes (complete cds)

Sequences “b”, “c” and “d” contain random deletions of their corresponding “a” sequence.

Values were calculated with the procedure described above.

Supplementary Figure 1

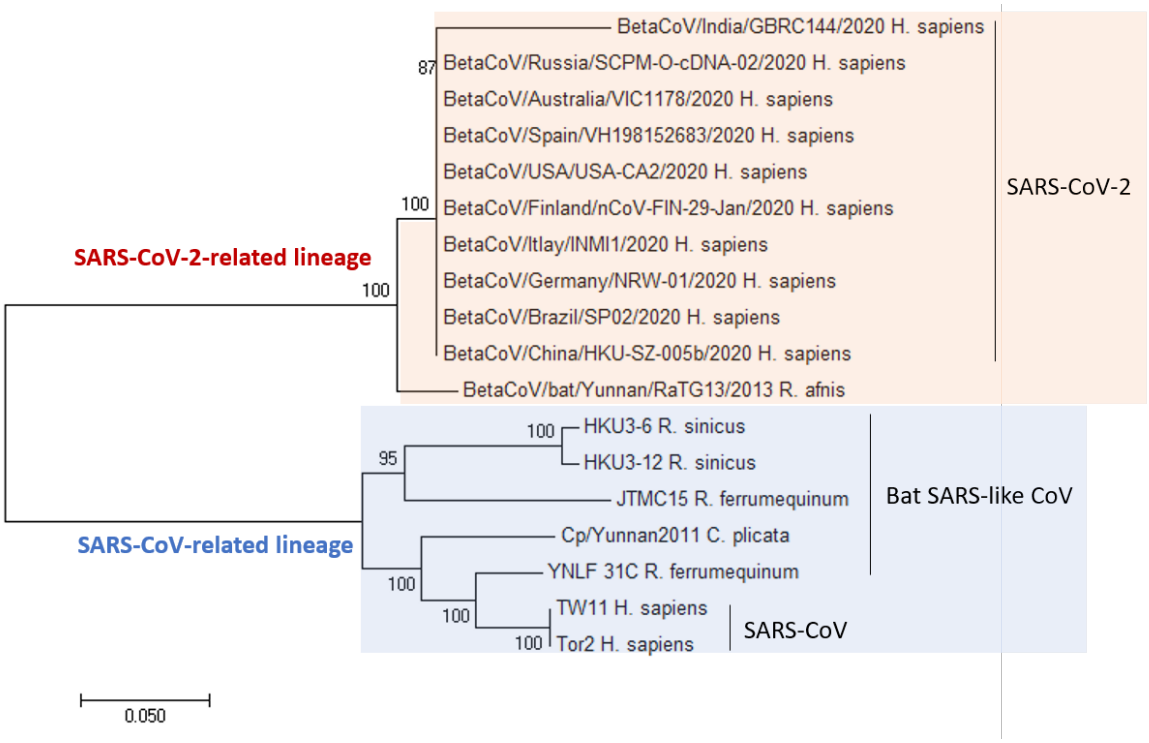


Accuracy of the GBMs model with and without the combination of the two bat and two bird subgroups.

(a) Left: Overall accuracy. Middle: Accuracies for bats combined or as two separate groups. Right: Accuracies for birds combined or as two separate groups. Pterobat = Pteropodiformes, and Vespbat = Vespertilioniformes.

(b) Error analysis of the model in (Babayan, et al., 2018) indicates that predictions between Galloanserae and Neoaves, Pterobat and Vespbat categories led to the most errors, as marked by the red boxes. Integration of these categories could result in better prediction accuracy.

Supplementary Figure 2



Phylogeny of ten SARS-CoV-2 viruses estimated from the concatenated ORF1ab, S, E, M, and N genes. A maximum likelihood approach using the GTRGAMMA nucleotide substitution model was used with 100 bootstrap replicates. Scientific names are indicated at the end of the sequence names, Genus abbreviations are: R. Rhinolophus; C. Chaerephon, H. Homo.