bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PACIFIC: A lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses

Pablo Acera Mateos[1,2]*, Renzo F. Balboa[1,3]*, Simon Easteal[1,3], Eduardo Eyras[1,2,4,5†], and Hardip R. Patel[1,3†]

[1]John Curtin School of Medical Research, Australian National University, Canberra, Australian Capital Territory 2600, Australia.

[2]EMBL Australia Partner Laboratory Network at the Australian National University, Canberra, Australian Capital Territory 2601, Australia.

[3]National Centre for Indigenous Genomics, Australian National University, Canberra, Australian Capital Territory 2600, Australia.

[4]IMIM - Hospital del Mar Medical Research Institute. E08003 Barcelona, Spain.

[5]Catalan Institution for Research and Advanced Studies. E08010 Barcelona, Spain.

Email addresses: pablo.aceramateos@anu.edu.au (PAM), renzo.balboa@anu.edu.au (RFB), simon.easteal@anu.edu.au (SE), eduardo.eyras@anu.edu.au (EE), hardip.patel@anu.edu.au (HRP).

*These authors contributed equally to this work

†Correspondence: Hardip R. Patel (hardip.patel@anu.edu.au) and Eduardo Eyras (eduardo.eyras@anu.edu.au)

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Viral co-infections occur in COVID-19 patients, potentially impacting disease progression and severity. However, there is currently no dedicated method to identify viral co-infections in patient RNA-seq data. We developed PACIFIC, a deep-learning algorithm that accurately detects SARS-CoV-2 and other common RNA respiratory viruses from RNA-seq data. Using *in silico* data, PACIFIC recovers the presence and relative concentrations of viruses with >99% precision and recall. PACIFIC accurately detects SARS-CoV-2 and other viral infections in 63 independent *in vitro* cell culture and patient datasets. PACIFIC is an end-to-end tool that enables the systematic monitoring of viral infections in the current global pandemic.

Keywords

SARS-CoV-2; sequence classification; co-infection; deep learning; respiratory virus; machine learning; Rhinovirus; Metapneumovirus; Influenza; COVID-19

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

Acute respiratory tract infections are the third largest global cause of death, infecting 545 million people and claiming 4 million lives every year (1–3). RNA viruses such as influenza, parainfluenza virus, respiratory syncytial virus, metapneumovirus, rhinovirus, and coronavirus are amongst the top pathogens causing respiratory infections and disease (4,5). Novel respiratory diseases, including coronaviruses, cross species boundaries repeatedly. Since December 2019, millions of people have been affected by COVID-19, an infectious zoonotic disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, NCBI Taxonomy ID: 2697049). Novel zoonotic coronaviruses also caused the 2002-2003 outbreak of SARS-CoV respiratory disease with at least 8,098 known cases and the ongoing 2012-2020 outbreaks of Middle East respiratory syndrome coronavirus (MERS-CoV) with at least 2,519 known cases (5–8). This recurrent emergence of respiratory viruses warrants increased surveillance and highlights the need for rapid, accurate, and timely diagnostic tests.

Diagnostic testing, treatment, and disease severity are complicated by the occurrence of respiratory co-infections. Up to 40% of individuals infected with respiratory viruses test positive for co-infections with up to three different pathogens (9,10), and recent studies have reported that ~20% of SARS-CoV-2 positive individuals had a co-infection with other respiratory viruses (11). Viral co-infections can alter the severity of disease and modify survival rates, and while COVID-19 remains poorly understood, early studies indicate a potential for increased mortality associated with influenza co-infections (12). Further studies are required to investigate the relationship between SARS-CoV-2 co-infections with prognosis and mortality rate (12,13).

Current diagnostic tests for respiratory infections are often limited in their capacity to detect co-infections. The current standard of viral detection for COVID-19 is based on polymerase

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

chain reaction (PCR) assays directed towards SARS-CoV-2 (14), which do not detect co-infecting viruses. Multiple virus identification in clinical settings is typically performed by using multiplexed PCR assays with primers specifically designed to target known respiratory pathogens (15). However, this approach is generally used only with pathogens that are expected *a priori* and the range of pathogen detection is limited by the probe design. Additionally, these protocols must be updated as new species or strains are identified as clinically relevant.

High-throughput RNA sequencing (RNA-seq) provides an unbiased measurement of the RNA molecules present in a sample and can potentially enable the systematic detection of SARS-CoV-2 infections and co-infections. Multiple species identification has been effectively performed in sequence data in the context of metagenomics studies (16). Programs such as Kraken (17–19) use k-mers to taxonomically classify sequencing reads into species from metagenomic samples. However, these tools use large databases of species sequences to compare against, resulting in considerable storage and computing requirements. In contrast, machine learning based tools have the advantage of extracting required features and encapsulating the necessary information for sequence classification in a computationally efficient model. This approach has been successfully used in the past for sequence classification problems (20). For example, DeepMicrobes (21) uses deep learning for genus and species level classification of metagenomic DNA sequencing reads from human gut bacteria. Similarly, ViraMiner (22) uses a deep learning binary classifier to identify DNA viruses from human microbiome metagenomic reads. Despite these advances, there is currently no equivalent deep learning classifier for the detection of SARS-CoV-2 and possible co-infections by RNA viruses.

To address this limitation, we have developed PACIFIC, a deep learning model to detect the presence of SARS-CoV-2 and other common respiratory RNA viruses in RNA-seq data from

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

patient samples. PACIFIC is an easy-to-use, streamlined tool useful for clinical and epidemiological applications in the context of the COVID-19 pandemic. Our tool accurately identifies and discriminates reads into five distinct classes: SARS-CoV-2, influenza (representing H1N1, H2N2, H3N2, H5N1, H7N9, H9N2, and Influenza B), metapneumovirus (representing 5 distinct assemblies), rhinoviruses (representing rhinovirus A and A1, B, C1, C2, C10 and other enteroviruses) and other coronaviruses (representing alpha, beta, gamma, and other unclassified coronaviruses). Extensive *in silico* tests show that PACIFIC achieves >99% precision, accuracy and recall. In addition, predictions in 63 infected human cell-lines and human primary samples demonstrate greater performance using PACIFIC for the detection of each virus class in comparison with alignment (BWA-MEM) and k-mer (Kraken2) based methods.

To the best of our knowledge, PACIFIC is the first software that uses deep learning to classify different RNA viruses from RNA-seq reads. By enabling the systematic identification of co-infections, we anticipate that PACIFIC can aid the clinical management of COVID-19 patients during the current pandemic and the surveillance of respiratory infections in future epidemiological studies.

Results

PACIFIC model

PACIFIC is a deep learning method designed to classify RNA-seq reads into five distinct respiratory virus classes and a human class (Figure 1a). The model architecture for PACIFIC is composed of an embedding layer, a convolutional neural network, and a bi-directional long short-term memory (BiLSTM) network that ends in a fully connected layer (Figure 1b). One of the main advantages of deep neural networks compared to other machine learning models

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in the context of sequence classification is the ability to extract relevant complex classification features from DNA or RNA sequences without having to explicitly define them *a priori*. However, the strategy to encode nucleotide sequences must be carefully considered, as it can dramatically affect the performance of the classifier (23). PACIFIC implements an embedding layer, which boosts the performance of the model in comparison with other encoding approaches (24). PACIFIC first converts nucleotide sequences into k-mers, assigns them to numerical tokens and converts these tokens into dense representations using a continuous vector space.

The use of a convolutional neural network adds several advantageous properties to the model. One advantage is location invariance (25), which allows the model to identify combinations of features with predictive value regardless of their relative position along the sequence. In addition, each filter used in the convolution layers can capture the predictive value of specific regions or combinations of k-mers.

After the convolution layers, PACIFIC uses a pooling layer to decrease the dimensionality of the feature space while maintaining essential information. PACIFIC uses a BiLSTM to model long-range dependencies in nucleotides, which provides the capacity to incorporate complex relationships in the input sequence that are sometimes ignored by single LSTMs (26). PACIFIC then implements a dense layer to estimate posterior probabilities for each of the five classes considered: Coronaviridae, Influenza, Metapneumovirus, Rhinovirus, and SARS-CoV-2. We included a sixth class in the model (Human) to classify RNA-seq reads derived from the human host. Finally, PACIFIC takes the reverse-complement of each predicted read, and only assigns the read to a particular class if the posterior probabilities of both the forward and reverse-complemented versions of the read for that class are ≥ 0.95.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

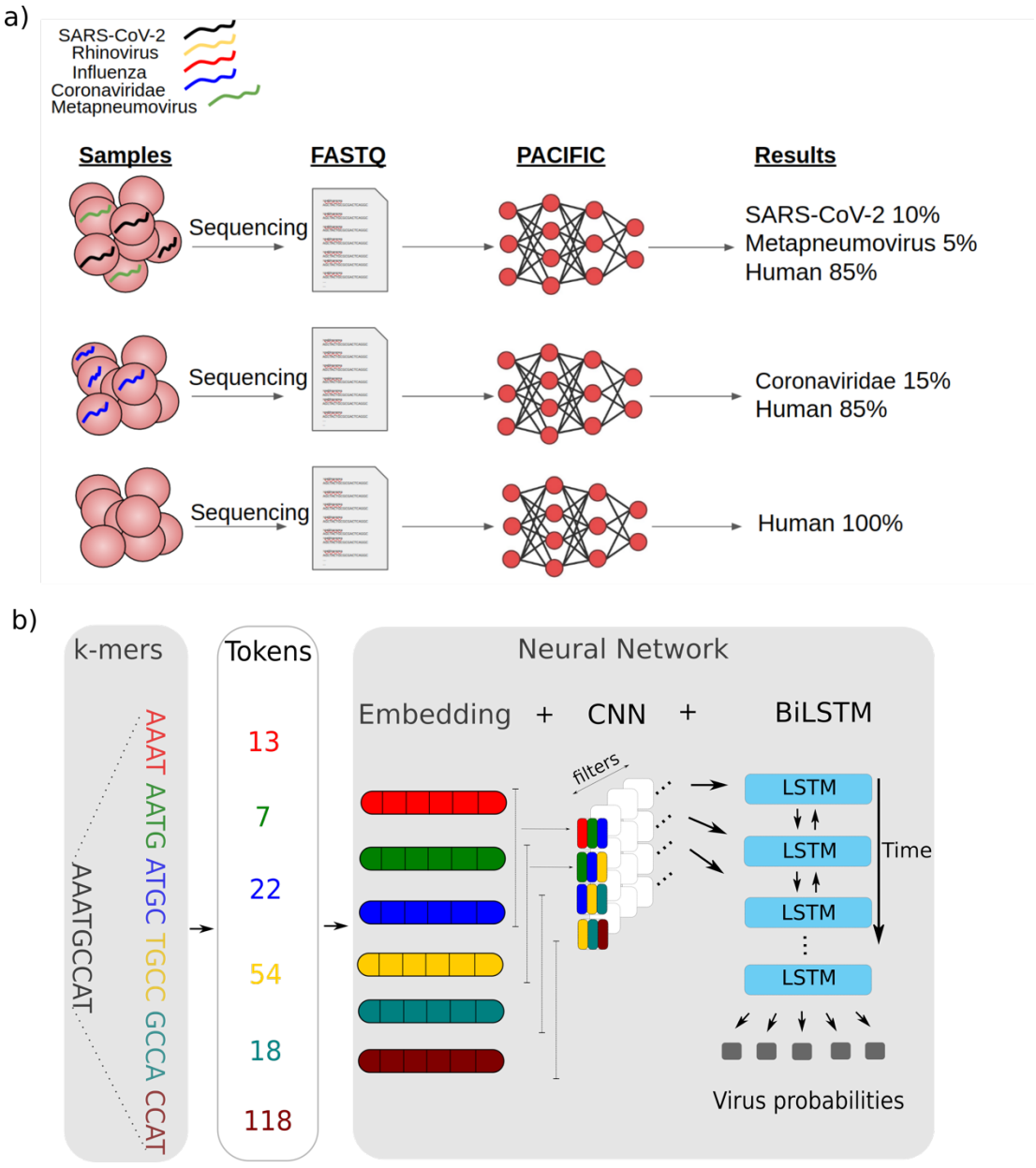


Figure 1. Overview of PACIFIC and its model architecture. **a)** PACIFIC uses FASTQ or FASTA files as inputs to make read-level predictions and report the relative percentage of RNA virus and human reads in a sample. **b)** Schematic view of PACIFIC deep neuronal network architecture. PACIFIC uses embedding, convolutional neural network and BiLSTM layers. The model is trained using *in silico* generated sequences from RNA virus genomes and the human transcriptome.

Properties of PACIFIC training data

PACIFIC was trained using 7.9 million 150nt long random fragments from 362 viral genome assemblies belonging to one of five viral classes (SARS-CoV-2, Influenza, Metapneumovirus,

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Rhinovirus and Coronaviridae) and the human transcriptome (Additional file 1: Table S1) (Methods). *In silico* fragments from both strands were generated without errors to accommodate paired-end sequenced reads and to retain the natural variation between genomes in each class. We used 90% of the data for training and 10% to tune the hyperparameters and network architecture.

The selection and grouping of virus classes were based on several considerations. First, we wanted PACIFIC to accurately detect SARS-CoV-2 as an independent class and to discriminate it from other coronaviruses. Second, we selected viruses that have been recently reported to appear as co-infections with SARS-CoV-2 (11). Third, we restricted our selection to viruses for which humans have been defined as one of the host species in NCBI Taxonomy database (27). Fourth, as the majority of reads in a sample are expected to be derived from human RNAs, we included an independent human class representing the human transcriptome to avoid the misclassification of human reads as viral origin.

K-mer length selection and sequence divergence

As input reads are divided into k-mers within the model, we investigated appropriate virus and human k-mer properties. A k-mer length of 9 was previously reported to be the optimal k-mer length for the phylogenetic separation of viral genomes (28). However, 9-mer profiles of SARS-CoV-2 and the human transcriptome have not been previously explored. We computed all-vs-all Jensen-Shannon divergence (JSD) scores using 9-mers to confirm that k=9 is the effective *k*-length to distinguish between the six PACIFIC classes. JSD is a symmetric measure of (dis)similarity that accounts for shared k-mer frequency distributions between a pair of sequences (29). JSD values range between 0 for identical sequences and 1 for two sequences that do not share any k-mer. Overall, inter-class JSD values were higher compared to the intra-class JSD values for 9-mers, which confirm that 9-mers are effective at separating sequences

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

belonging to different viral classes and human transcripts (Additional file 2: Figure S1). Specifically, the average JSD between the SARS-CoV-2 class and the Coronaviridae class was 0.786, which was greater than 0.767 intra-class JSD for the Coronaviridae and 0.002 for the SARS-CoV-2 class, thus indicating sufficient divergence for their separation into distinct classes. Given these results, we decided to encode input sequences as 9-mers with a stride of 1.

PACIFIC testing shows high precision and recall for simulated data

Performance metrics (false positive rates (FPRs), false negative rates (FNRs), precision, recall, and accuracy) were calculated using *in silico* generated reads that modelled sequencing-induced substitution and indel errors from sample mixtures with known class labels. We generated 100 independent datasets of 150nt single end reads with Illumina HiSeq2500 errors using ART (30). Each dataset contained ~700,000 reads and was comprised of approximately 100,000 reads from each of the 6 classes in the PACIFIC model, plus ~100,000 reads from unrelated viral genomes (Methods).

First, we compared the performance metrics between predictions using only the forward strand of a read (*single prediction*) and predictions using both the forward and reverse-complemented strands (*double prediction*) (Figure 2). In *single prediction*, a read was assigned the class label with the highest posterior probability for the forward sequence if the posterior probability for a class was ≥ 0.95. For *double prediction*, the read was predicted to be of a given class if the posterior probability of the predicted forward and reverse-complemented read was ≥0.95 and predictions agreed on the same class. The average FNR across 100 datasets in *double prediction* relative to *single prediction* increased by 1.45× for Coronaviridae, 1.49× for Influenza, 1.41× for Metapneumovirus, 1.40× for Rhinovirus and 1.62× for SARS-CoV-2 (Figure 2). However,

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

we observed a large decrease in the FPR in *double prediction*. The average FPR in the 100 datasets decreased by 4.60× for Coronaviridae, 11.02× for Influenza, 16.55× for Metapneumovirus, 3.92× for Rhinovirus and 16.47× for SARS-CoV-2 class in *double prediction* relative to *single prediction* (Figure 2). Concomitantly, the average precision for all viral classes increased and the average recall decreased by a small margin. Due to these observations, *double prediction* was implemented as the standard classification approach in PACIFIC.

Overall, PACIFIC achieved high precision (average ≥ 0.9995), recall (average ≥ 0.9966) and accuracy (average ≥ 0.9995) for each of the virus classes (Table 1). For the human class, the average precision was lower at 0.50140 compared to the viral classes; this is attributed to the large number of reads (>99%) from unrelated viral genomes being assigned to the human class. As a result, sequences that do not belong to any of the viruses in the model are unlikely to be mislabelled as one of the virus classes.

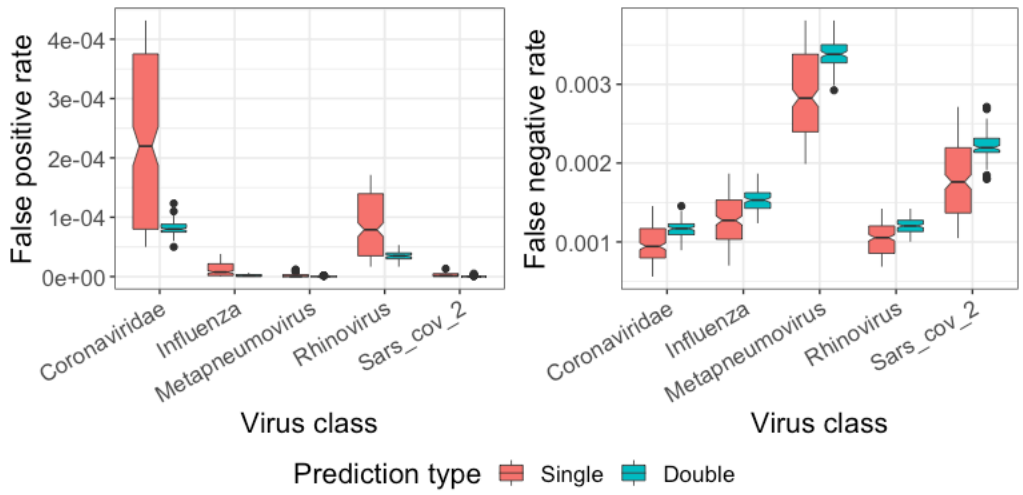


Figure 2. Comparison of false positive and false negative rates between *single* and *double predictions.* *Single prediction* (red) results in relatively higher false positives and lower false negatives compared to *double prediction* (green) where predictions are made on the forward strand of a sequence and its reverse complement.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

225 **Table 1.** PACIFIC performance metrics for each class in 100 independent simulated datasets.

Class	Average FNR (±95% CI)	Average FPR (±95% CI)	Average Precision (±95% CI)	Average Recall (±95% CI)	Average Accuracy (±95% CI)
Coronaviridae	0.001162 (2.14e-05)	8.16e-05 (2.38e-06)	0.999518 (1.41e-05)	0.9988 (2.14e-05)	0.9998 (3.67e-06)
Influenza	0.001530 (2.65e-05)	1.91e-06 (3.18e-07)	0.999988 (1.95e-06)	0.9985 (2.65e-05)	0.9998 (3.80e-06)
Metapneumovirus	0.003397 (3.62e-05)	1.83e-07 (1.04e-07)	0.999999 (6.23e-07)	0.9966 (3.62e-05)	0.9995 (5.20e-06)
Rhinovirus	0.001209 (1.90e-05)	3.55e-05 (1.41e-06)	0.999788 (8.44e-06)	0.9988 (1.90e-05)	0.9998 (2.79e-06)
SARS-CoV-2	0.002220 (3.27e-05)	2.49e-07 (1.43e-07)	0.999999 (8.63e-07)	0.9978 (3.27e-05)	0.9997 (4.64e-06)
Human	0.000366 (1.32e-05)	1.66e-01 (1.03e-05)	0.501400 (1.45e-05)	0.9996 (1.32e-05)	0.8581 (8.94e-06)

226 CI = confidence interval, FPR = False positive rate, FNR = False negative rate

227

228 Using the same *in silico* datasets, we then assessed the effect of mismatches on FPR and FNR.
229 All 100 controlled datasets contained ~22% of reads with substitutions and indel errors relative
230 to the reference genomes. FNRs were higher for mismatch-containing reads relative to exact
231 reads for all viral classes, increasing 5 to 64-fold (Figure 3). In contrast, FPRs increased 0.98
232 to 2-fold for mismatch-containing reads for all classes. These results suggest that FPRs are
233 relatively less affected by the presence of mismatches compared to FNRs for viral classes.
234 Despite relative differences of FPR and FNR in mismatch-containing reads compared to exact
235 reads, PACIFIC achieved high precision (0.9994), high recall (0.9909) and high accuracy
236 (0.9982) for mismatch-containing reads for all five viral classes (Additional file 1: Table S2).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

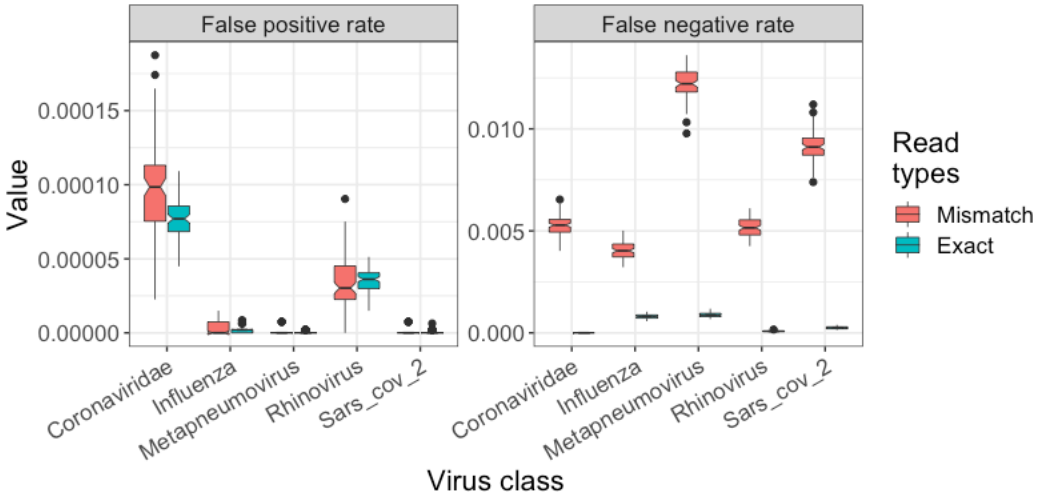


Figure 3. False positive (left panel) and false negative (right panel) rates for reads identical to the corresponding reference genome (Exact, green), and for reads with mismatches with respect to their reference genome (Mismatch, red).

Reads derived from unrelated virus genomes contributed most of the false positives for all predicted classes (Figure 4). Of note, there was negligible cross contamination between viral class labels. The largest inter-viral class misclassification was from SARS-CoV-2 to Coronaviridae, where out of 100,000 SARS-CoV-2 reads, ~2.9 were misclassified as Coronaviridae. Other inter-viral class misclassifications were between 0-0.2 reads. The majority of false negatives for each viral class were either discarded because of the *double prediction* criteria (rc_discarded in Figure 4), or because they did not meet the minimum 0.95 posterior probability criteria (pDiscarded in Figure 4). Taken together, our results demonstrate that PACIFIC is highly specific and sensitive for all five viral classes, with negligible false positive and false negative rates.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

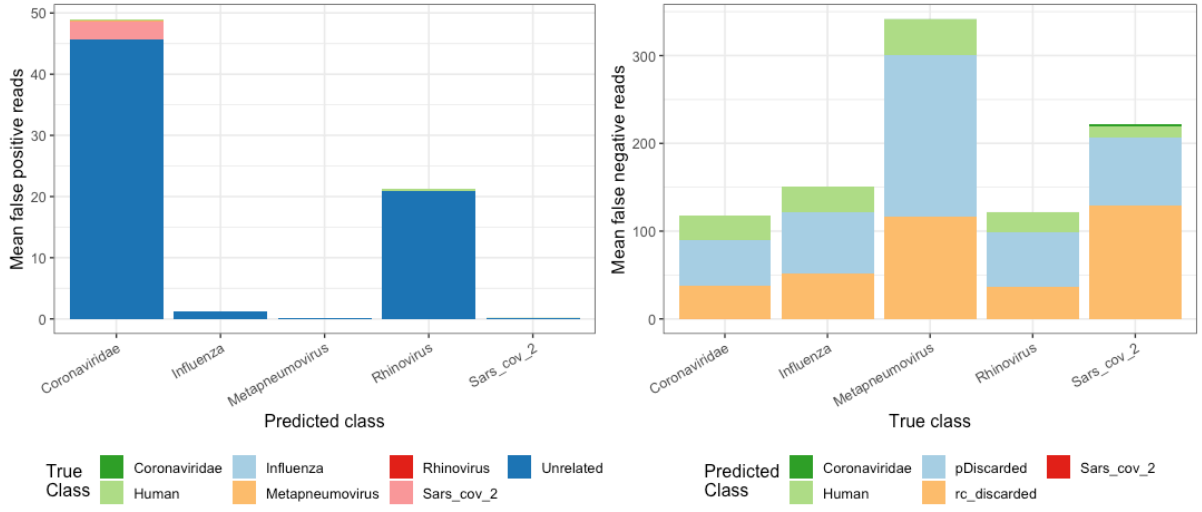


Figure 4. False positive and false negative assignments for 100 independently simulated datasets. Left panel - Average number of false positives (y axis). True labels are indicated by colour for each class and predicted labels are given in the x-axis. Right panel - Average number of false negatives (y axis). Predicted labels are indicated by colour for each class and true labels as indicated in the x-axis. pDiscarded - reads that do not reach the 0.95 posterior probability cut-off; rc_discarded - reads that were discordantly predicted using *double prediction*.

Establishing virus detection thresholds

In the previous section, we showed that PACIFIC displayed low FPR in balanced datasets with similar proportions of reads from each class. However, incorrect predictions about the presence or absence of a virus in a sample could lead to misguided follow-ups and the unnecessary use of valuable clinical resources. Therefore, we decided to establish the minimum percentage of reads for each viral class required to confidently predict the presence of a virus in a sample.

In practice, RNA-seq data will be unbalanced with almost all reads originating from the human transcriptome mixed with variable proportions of viral reads. To model this imbalance in class proportions, we simulated 500 independent datasets (100 for each class), each containing 500,000 150nt long reads using ART (30) with Illumina HiSeq2500 error profiles. Each dataset contained variable proportions of simulated reads for 4 of the 5 viral classes, plus human and

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

unrelated viral genomes. One of the five viral classes was intentionally excluded, and the excluded class was considered to be the test class. All reads assigned to this test class were counted as false positives. PACIFIC achieved similar average FPRs to the benchmarking experiments using balanced datasets (Table 2).

Assessment of the distribution of false positive rates for each viral class and skewness-kurtosis plots indicated that the percentage of false positives observed in unbalanced datasets followed a Beta distribution. Therefore, we used moment matching to estimate the shape parameters for the quantile function of the Beta distribution and determined the numeric threshold above which 99% of false positive samples were excluded. Using these thresholds, a sample would be classified as positive for Coronaviridae if >0.0405% reads were labelled as Coronaviridae by PACIFIC. Similarly, these limits were >0.000807% for Influenza, >0.000154% for Metapneumovirus, >0.0418% for the Rhinovirus, and >0.000213% for the SARS-CoV-2 class (Table 2).

Table 2. Average false positive reads across 100 experiments for each viral class.

Class	**Average FP% (Balanced)**	**Average FP% (Unbalanced)**	**FP % threshold***
Coronaviridae	0.00816	0.006664	0.0405
Influenza	0.000191	0.000062	0.000807
Metapneumovirus	0.0000183	0.000006	0.000154
Rhinovirus	0.00355	0.005392	0.0418
SARS-CoV-2	0.0000249	0.000010	0.000213

FP = False positive, Balanced = 100 experiments with equal proportion of reads from each class, Unbalanced = Variable proportion of reads from all classes and no reads from the test class. * These thresholds represent 0.99 quantile for FP% for that class.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PACIFIC accurately detects viruses in human RNA-seq samples

Next, we assessed PACIFIC's performance in classifying viral reads in RNA-seq data derived from human biological samples and compared its output with alignment-based (BWA-MEM) and k-mer based (Kraken2) approaches. To reduce bias in the comparisons, we built the BWA-MEM index and Kraken2 database using the same virus genome assemblies and the human transcriptome that were used for PACIFIC training (Additional file 1: Table S1). Additionally, we used the same percentage thresholds determined in the previous section (Table 2) for all three methods to assign the presence of a virus in a sample. All three methods were applied to 63 human RNA-seq datasets from independent research studies, with five of them known to contain SARS-CoV-2 (31,32). Four RNA-seq datasets were derived from primary human lung epithelium cells (NHBE) infected with SARS-CoV-2 *in vitro* (NCBI SRA accession: SRX7990869) (31) and one dataset was from a patient bronchoalveolar lavage fluid sample (NCBI SRA accession: SRR10971381) that was positive for SARS-CoV-2 (32). In addition, we analysed RNA-seq datasets for 48 airway epithelial cell samples from the GALA II cohort study, of which 22 were reported to contain respiratory infection viruses (33,34), and 10 were reportedly devoid of infections by the viral classes studied here, as indicated by the sample metadata and the corresponding publications (Additional file 1: Table S3).

For the five samples that were positive for SARS-CoV-2, PACIFIC assigned 0.047%-0.048% of all reads to the SARS-CoV-2 class for the *in vitro* infected cells and 0.19% of all reads in the patient sample; all five samples were above the established detection threshold for that class (>0.000213%) (Figure 5). Similarly, BWA-MEM and Kraken2 successfully identified SARS-CoV-2 reads above the detection threshold in these five samples (Figure 5a). All three methods accurately predicted the presence of SARS-CoV-2 and the absence of other virus classes in these samples as per class specific detection thresholds.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

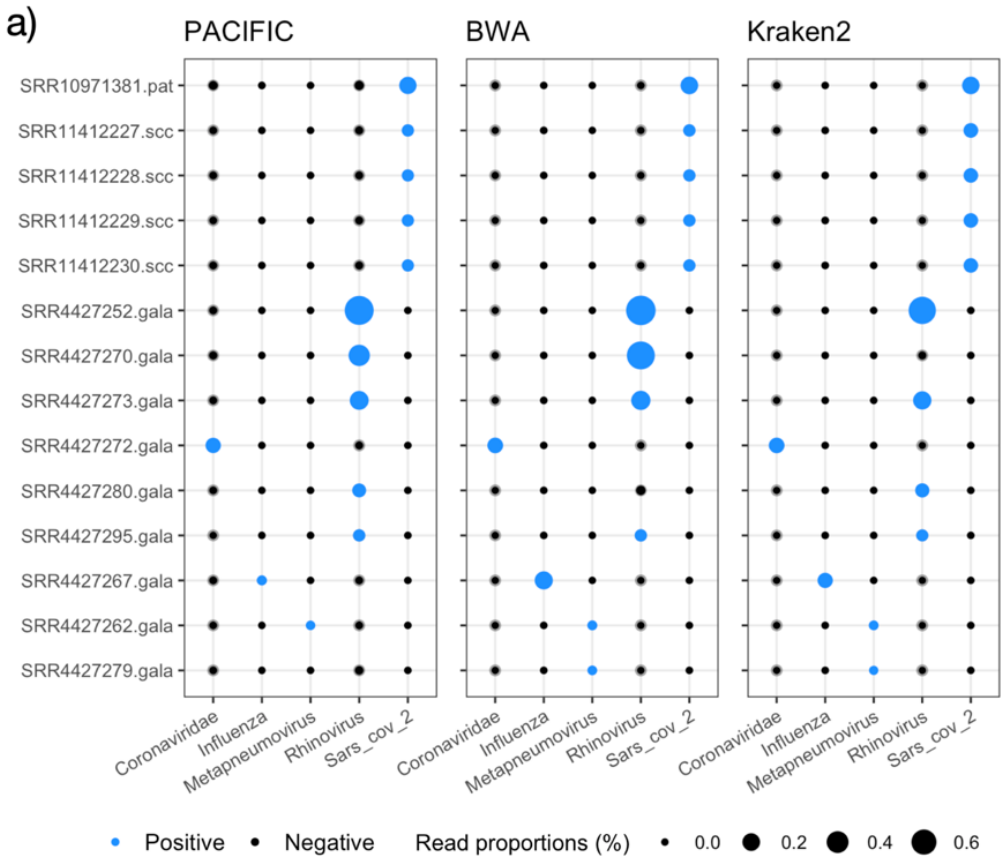
We subsequently tested the 48 samples from the GALA II cohort (33,34). Of these, 22 were reported to contain between 4 and 164,870 reads from respiratory viruses (human rhinovirus, respiratory syncytial virus, human metapneumovirus or human parainfluenza viruses I, II and III). PACIFIC, BWA-MEM and Kraken2 identified 9 samples as positive for one of the five virus classes considered, and 39 samples as negative for the same viral classes (Figure 5a; Additional file 2: Figure S2, Additional file 1: Table S3). The discrepancy between these results and the original study (34) could be partly explained by the exclusion of the Respiratory Syncytial Virus class in our analyses. However, further verifications could not be performed because sample labels provided in the manuscript were mismatched with the submitted sequence data (see correction (35)).

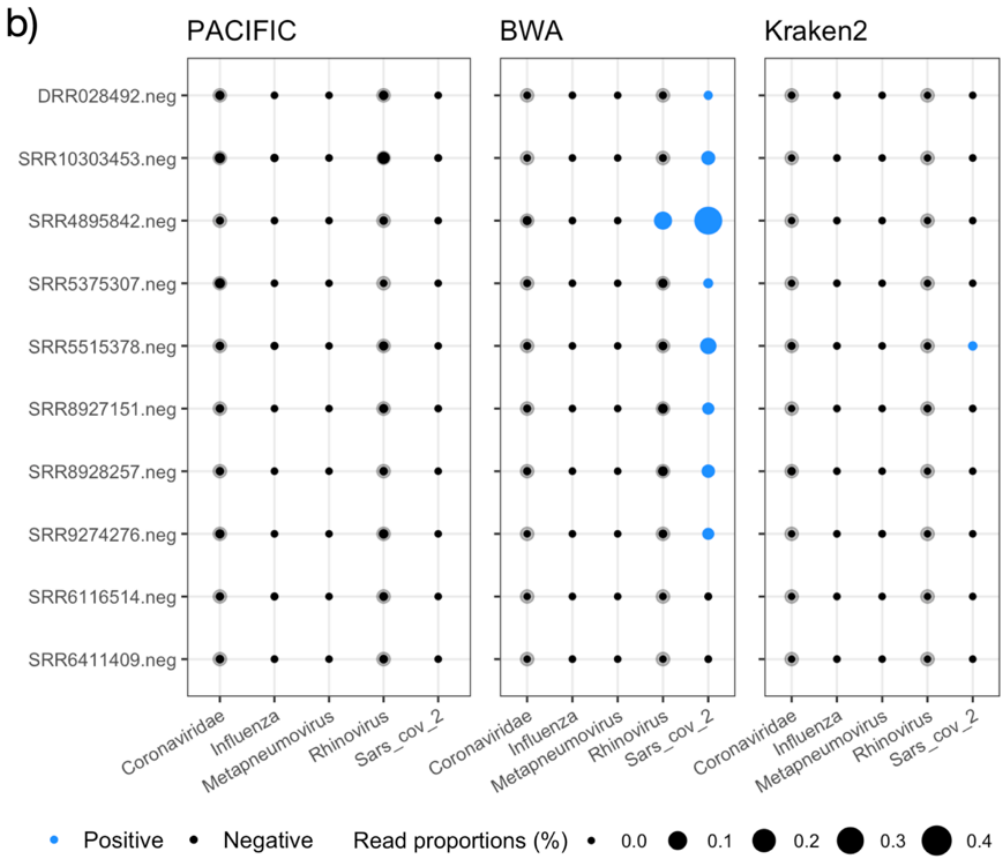
From the 9 positive samples, six were concordantly labelled positive for the same virus class by all three methods: three samples were positive for Rhinovirus, and one for Coronaviridae, Influenza, and Metapneumovirus, respectively (Figure 5a). In contrast, the other three samples were discordantly labelled by one of the three methods. One sample (SRR4427279) was classified as positive for Metapneumovirus by BWA-MEM and Kraken2 but not by PACIFIC. BWA-MEM and Kraken2 collectively assigned 28 reads to the Metapneumovirus class as opposed to 0 reads by PACIFIC. To investigate the origin of these reads, we used BLASTN searches against the NCBI nucleotide *(nt)* database encompassing sequences from all domains of life and extracted the best hit for each read (Methods). All 28 reads had their best hits to human respiratory syncytial virus A sequences (E-values $\leq$ 4.07e-68; bit-scores $\geq$ 267). Further analysis showed that metapneumovirus was not identified in any of the top 10 significant hits (Additional file 2: *Extension of BLAST analysis*). Another discordant sample (SRR4427270) was positive for the Rhinovirus class by PACIFIC (1,065 reads) and BWA-MEM (2,338 reads) but not by Kraken2 (3 reads). BLASTN searches showed that best hits (3,145/3,150 total BLAST alignments) were sequences from one of Enterovirus C105, Enterovirus C or Human

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

enterovirus C105 (E-values ≤ 9.46e-41, bit-scores ≥178). Rhinoviruses and other enteroviruses are taxonomically part of the Enterovirus genus (36,37). The final discordant sample (SRR4427280) was labelled positive for the Rhinovirus class by PACIFIC (355 reads) and Kraken2 (387 reads) but not by BWA-MEM (29 reads) using our thresholds (Figure 5a). BLASTN searches revealed that the majority of reads (385/390 collectively between PACIFIC and Kraken2) had their best hits to Rhinovirus C (E-values ≤ 4.47e-53; bit-scores ≥ 219). BWA-MEM therefore failed to assign most Rhinovirus reads classified by the other two methods.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.





347

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 5. PACIFIC, BWA-MEM and Kraken2 virus predictions in RNA-seq data. Transparent grey filled circles represent detection thresholds for each class, overlaid with black filled circles representing the percentage of predicted reads using PACIFIC (left panel), BWA-MEM (centre panel) and Kraken2 (right panel). Circles are filled blue when the percentage of reads for a class are above detection thresholds described in Table 2. RNA-seq samples (y-axis) are labelled with NCBI SRA run accessions and abbreviations for sample type. **(a)** RNA-seq samples predicted to be positive for at least one viral class by PACIFIC, BWA-MEM, or Kraken2. Samples include a SARS-CoV-2-infected human patient bronchoalveolar lavage fluid sample (.pat), four *in vitro* SARS-CoV-2-infected NHBE cell lines (.scc) and 9 samples from the GALA II cohort (.gala). **(b)** Human RNA-seq samples without expected viral infections (.neg; n=10). Samples were selected from the NCBI SRA database without any evidence of any infection.

In addition to the 53 samples tested above, we analysed 10 publicly available human RNA-seq datasets without expected viral infections using all three methods. In particular, all ten samples were registered in the NCBI SRA database on or before 17 October 2019 and therefore were unlikely to contain SARS-CoV-2 (Additional file 1: Table S3). Of note, two samples (SRR8927151 and SRR8928257) were published in February 2020 (Additional file 1: Table S3) (38). However, the two NCBI BioProjects for these samples were registered on 18 April 2019.

PACIFIC accurately predicted all 10 samples as negative for viral infections using our detection thresholds. In contrast, Kraken2 assigned one sample (SRR5515378) as positive for SARS-CoV-2 with 256 reads assigned (Figure 5b). Further verification with BLASTN searches confirmed that 242 out of the 256 reads mislabelled by Kraken2 aligned to the Mycoplasma bacterial genus (E-values ≤2.76e-23, bit-scores ≥121), indicating false positive assignments by Kraken2 for these reads.

BWA-MEM performed relatively worse in these 10 datasets, with 8 samples classified as positive for SARS-CoV-2 (Figure 5b). A total of 15,395 reads were aligned to SARS-CoV-2 genomes from these 8 samples, ranging from 86 to 6,901 reads in any given sample. BLASTN searches of these SARS-CoV-2 assigned reads showed *Homo sapiens* as the best hit (E-values ≤ 5.78e-07, bit-scores ≥ 65.8) for 4,352 (28%) reads. Further examination revealed that 53%

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

379 of all 9-mers in these reads were poly-A or poly-T derived, suggesting low-complexity
380 sequences. In addition, SRR4895842 was also assigned as co-positive for the Rhinovirus class
381 in addition to SARS-CoV-2 by BWA-MEM (Figure 5b). BLASTN searches of 406 reads
382 assigned to Rhinovirus within this sample revealed that 303 reads had best hits to *Homo*
383 *sapiens,* and 48 reads had their best hits to *Pan paniscus* (E-value ≤ 6.06e-10; bit-scores ≥
384 76.8). These reads had 19% of their 9-mers derived from poly-A or poly-T sequences,
385 suggesting low-complexity sequences.

386 Overall, these results show that PACIFIC can accurately identify viral reads and the use of our
387 detection thresholds assisted in correctly establishing the presence or absence of viral classes
388 in RNA-seq data from biological samples with better accuracy than existing methods.

389

390 **Discussion**

391 We have developed PACIFIC, a deep learning-based tool for the detection of SARS-CoV-2
392 and other common respiratory viruses from RNA-seq data. To the best of our knowledge,
393 PACIFIC is the first deep learning model that performs detection of SARS-CoV-2 and different
394 RNA virus groups using short-read sequence data with >0.99 precision, recall and accuracy. A
395 recent analysis of 4,909 scientific articles identified 47 models for detecting COVID-19, 34 of
396 which were based on medical images (39). This study concluded that these predictive models
397 were, in general, poorly described and contained multiple biases, likely resulting in unreliable
398 predictions when applied in practice. To overcome these potential limitations, we used multiple
399 diverse and independent simulated datasets reflecting realistic scenarios to validate the
400 performance of PACIFIC. Importantly, PACIFIC was successfully applied to 63 RNA-seq
401 datasets derived from infected cell cultures and patient samples for the detection of viral

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

infections, demonstrating that PACIFIC can be applied to human-derived RNA-seq datasets and assist in clinical settings.

In 2013, the World Health Organisation launched the Battle against Respiratory Viruses (BRaVe) initiative, which identified six research strategies to tackle and mitigate risks of death due to respiratory tract infections. One of the proposed strategies was to "*improve severe acute respiratory infection diagnosis and diagnostic tests amongst others"* (40). High-throughput sequencing-based approaches can provide immense diagnostic potential and facilitate molecular epidemiological studies, thereby contributing towards the BRaVe initiative's goals (41,42). It is more important than ever to explore and determine the diagnostic potential of RNA-seq for the SARS-CoV-2 pandemic.

A comprehensive study using multiplex RT-PCR and a sequencing-based metagenomic approach revealed that RNA-seq has sufficient sensitivity and specificity to be applicable in the clinic for respiratory viruses (42). However, the use of RNA-seq in diagnostic settings is often complicated due to complex analytical workflows (34,42). A typical workflow for virus detection in high-throughput sequencing data involves quality assessment and filtering of raw data, removal of host sequences, *de novo* assembly of remaining reads, and lastly, the alignment and annotation of the generated contigs (43). Implementation of these workflows require expert knowledge of bioinformatics software and databases and often dedicated computing facilities. PACIFIC overcomes these limitations by modelling the differences in k-mer content of respiratory viruses and human sequences in a model that is efficient in compute and storage requirements, easy to use, and therefore applicable in contexts with minimal resources. Specifically, we have designed PACIFIC to be run as a single command using raw RNA-seq data as the only required input to obtain quantified predictions about viral classes within a sample.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Despite the higher costs of sequencing compared to PCR-based experiments, multiplexing, block-testing or pooling strategies (44) could be implemented for unbiased cost-effective testing. For example, sequencing with Illumina platforms could be done with 96 samples per lane using multiplexing, reducing the sequencing cost per sample. In this scenario, the number of reads obtained per sample could be approximately 200,000 or higher. We have demonstrated the accuracy of PACIFIC in a variety of sample sizes, which suggests the potential value of this approach.

One of the major challenges in the identification of virus classes is the high rate of natural sequence variation for RNA viruses (45,46), in addition to high-throughput sequencing induced errors and artefacts, and the presence of low-complexity A-rich sequences common to the host transcriptome. We showed that 22% of reads containing mismatches and indel errors were accurately assigned to a virus class by PACIFIC with negligible loss in sensitivity at a sample level. Given the ability of PACIFIC to accurately assign error-containing reads, we speculate that PACIFIC is applicable to cases where viruses present natural sequence variation. In such cases, or when new species are required to be added to the model, strategies like transfer learning can be used to update the model without the need to retrain the entire model, with low computational cost (47). Future versions of PACIFIC could focus on training a model that incorporates class specific mutation rates and sequence diversity to reduce the need for regular updates as new viral mutations emerge.

PACIFIC is intentionally focused on the identification of viral classes reported to be co-infecting along with SARS-CoV-2 (12). Therefore, samples containing other viruses and bacterial infections may require additional analysis. Future versions of PACIFIC could include the classification of a broader range of virus and bacterial classes at a species level, and variable input read lengths to increase PACIFIC's utility in other contexts.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Conclusions

PACIFIC is a powerful end-to-end and easy to use tool that predicts the presence of SARS-CoV-2, Influenza, Metapneumovirus, Rhinovirus and other Coronavirus class-derived sequences directly from RNA-seq data with high sensitivity and specificity. PACIFIC will enable effective monitoring and tracking of viral infections and co-infections in the population in the context of the COVID-19 global pandemic and allow for the development of new strategies in molecular epidemiology of co-infections to understand variable host responses and improve the management of infectious diseases caused by viruses.

Methods

PACIFIC and other associated software written for this manuscript is available at https://github.com/pacific-2020/pacific. We have used Python (version 3), scipy (v1.4.1), numpy (v1.18.1), scikit (v0.23.1), pandas (v1.0.1), tensorflow (v2.2.0), keras (v2.3.1), R (v3.6), tidyverse (v1.3.0), Biobase (v2.46.0) and Perl (v5.26) in our analysis.

Training data

We downloaded 362 virus genomes from the NCBI assembly database corresponding to five classes of single stranded RNA viruses (Table 4, Additional file 1: Table S1). GenBank assembly identifiers and assembly versions with other metadata are listed in Additional file 1: Table S1. Since our focus was to detect co-infections with SARS-CoV-2, we made a separate class for SARS-CoV-2 containing 87 different assemblies (Table 4). The *Coronaviridae* class contained 12 genomes of alpha, beta, gamma and unclassified coronaviruses. The Influenza class contained assemblies of influenza A (H1N1, H2N2, H3N2, H5N1, H7N9, and H9N2 strains) and influenza B viruses. For the Rhinovirus class, assemblies of rhinovirus A (including A1 strain), B, C (including C1, C2, and C10 strains), and unlabelled enterovirus

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

were grouped together. There were five distinct assemblies for metapneumovirus which were grouped into a single class. We included Human GENCODE (48) canonical transcript sequences (downloaded from Ensembl v99 database (49)) as an additional class to distinguish sequencing reads derived from the human transcriptome. We generated between 0.44 and 3.5 million 150nt-long fragments *in silico* for each class using a custom Perl script available at https://github.com/pacific-2020/pacific (generatetestdata.pl, Table 4). These training sequences were randomly sampled without any base substitutions and were derived from both strands of the genome assemblies.

Table 4. Summary of training classes used for PACIFIC.

Class	Total reads	Number of genome assemblies	Number of taxonomic units	Included species/genus groups
Coronaviridae [ssRNA(+)]	644,483	12	12	Alpha, beta, gamma and unclassified coronaviruses
Influenza [ssRNA(-)]	1,073,237	128	125	Influenza A (H1N1, H2N2, H3N2, H5N1, H7N9, H9N2), Influenza B
Metapneumovirus [ssRNA(-)]	443,974	5	1	Metapneumovirus
Rhinovirus [ssRNA(+)]	1,339,435	130	107	Rhinovirus A and A1, B, C (C1, C2, C10), other enterovirus
SARS-CoV-2 [ssRNA(+)]	865,303	87	1	SARS-CoV-2
Human	3,531,425	1*	1	Human transcriptome
Total	7,897,857	363	247	

"*": GENCODE canonical transcripts were used to represent human reads in RNA-seq data.

Model architecture

PACIFIC was implemented using the Keras API with a TensorFlow backend. Input reads were converted into 9-mers with a stride of 1, forming a vocabulary size of $4^9 = 262{,}144$ k-mers. Each of these k-mers is assigned a number using the Tokenize API from Keras (50) from 1 to 262,144. The first index position of 0 is reserved to denote zero-padding for variable length sequences. Tokens are fed into the first hidden layer of the neural network and transformed into continuous vectors of length 100. After the embedding, a convolutional layer takes the previous numerical vectors and uses 128 convolution filters with a kernel size of 3.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A pooling layer is used after the convolution, using *max pooling* with a kernel size of 3. A bidirectional long-short term memory (BiLSTM) layer then follows, which uses two traditional LSTMs; one starts 'reading' the input sequence from one of the two flanks, and the other from the opposite end. The output of the two LSTMs is then combined and passed to the next layer. Finally, PACIFIC has a fully connected layer using a *softmax* function to calculate posterior probabilities for each of the six classes. To reduce overfitting, we used 20% dropout at each hidden layer.

Cross-entropy was used as the loss function and ADAM (51) was used as the optimizer. The final configuration of the network, hyperparameter tuning and the number and configurations of layers was obtained after several iterations between training and validation data. The final model is implemented *as double prediction* on both strands of the input sequence, whereby the forward and reverse-complement of the input sequence are predicted for class assignment. Classes for both predictions were required to match. The threshold of posterior probability for the assigned class was ≥0.95.

PACIFIC training

NVIDIA GeForce RTX2080Ti was used to accelerate training. We trained two LSTM implementations, one using the fast LSTM implementation backed by CuDNN, supported only with NVIDIA Graphical Processing Unit (GPU). The other model was built using the regular implementation of LSTM. Both models achieved the same results. We started the training by shuffling the training sequences, using chunks of 200,000 reads to avoid loading all reads into memory. 90% of the data was used for training and 10% for optimization of parameters. After 15 chunks, the model converged on the validation set and training was halted. During training, we used binary accuracy (1), categorical accuracy (2) and cross-entropy loss from the optimization set to monitor the training.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

519 1. $Categorical\ accuracy = \frac{\#\ correct\ predictions}{\#\ total\ predictions}$

520 2. $Binary\ accuracy = \frac{\#\ correct\ predictions}{\#\ total\ predictions}$ if highest output probability > 0.5

521 Training was completed when the model converged, obtaining final categorical and binary
522 accuracy values of 0.99, and 0.003 for optimization loss.

523 *PACIFIC test datasets*

524 We generated 100 independent test datasets using the ART sequence simulation software
525 (version 2.5.8, (30)) with default error models for substitutions, insertions and deletions using
526 the Illumina® HiSeq 2500 sequencing platform. For each dataset, we set seeds starting from
527 2021 to 2120 using a random number generator for reproducibility. Synthetic data contained
528 150nt single end reads derived from seven classes; the five model virus classes, a human class,
529 and an "unrelated" class composed of 32,550 distinct virus genomes downloaded from the
530 NCBI Assembly database. We sampled ~100,000 reads per class using a class-specific fold-
531 coverage parameter to generate ~700,000 reads per test data (Table 5). Approximately 22% of
532 reads contained mismatches, insertions or deletions relative to their respective reference
533 sequences, reflecting error profiles of the Illumina sequencing platform. This process was
534 automated using a custom script (generatebenchmarkdata.pl).

535

536

537

538

539

540

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 5. Summary of benchmark datasets

Sequence class	Total bases	Fold coverage	Number of reads	Reads with mismatches or indels
Coronaviridae	323274	47.5	102076	22295-22865
Human	75434059	0.227	100208-100209	21768-22425
Influenza	1684539	9.65	100038-100290	21825-22570
Metapneumovirus	66596	228	100548	21890-22559
Rhinovirus	925412	16.7	101940-101953	22196-22919
SARS-CoV-2	2599395	5.8	100337-100349	21831-22507
Unrelated viruses	1038794620	0.01738	100175-100196	21890-22479

PACIFIC performance tests

PACIFIC was used to assign class labels to reads in the test data, and performance metrics were calculated by comparing known and predicted labels for each read. A read was assigned a class if the maximum posterior probability score for a class was ≥0.95. A true positive (TP) was defined when the true label and the predicted label were the same for a read. A true negative (TN) was defined when a read that did not belong to the true class was correctly predicted as a class different from the true class. False positives (FP) were reads which were predicted to be as the true class, although they originated from a different class. False negatives (FN) were all reads belonging to the true class but were predicted as a different class. An example confusion matrix for SARS-CoV-2 is described in Table 6. Precision, recall, accuracy, false positive rate and false negative rate were calculated using equations 3-7 below.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 6. Confusion matrix using SARS-CoV-2 as an example of a positive class.

	True/ Actual condition		
Predicted condition		***Positive*** *SARS-CoV-2 +*	***Negative*** *All other classes*
	SARS-CoV-2 +	True positive (TP)	False positive (FP)
	All other classes	False negative (FN)	True negative (TN)

3. $Precision = \frac{TP}{TP+FP}$
4. $Recall = \frac{TP}{TP+FN}$
5. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
6. $FPR = \frac{FP}{FP+TN}$
7. $FNR = \frac{FN}{FN+TP}$

where TP = True positive, FP = False positive, TN = True negative, FN = False negative, FPR = False positive rate, FNR = False negative rate.

Establishing false positive rate thresholds for each class

This experiment was performed to quantify the impact of variable proportions of reads from each class on the percentage of false positives and to establish the detection threshold for each virus class in RNA-seq data. For each viral class in PACIFIC, we generated 100 datasets containing 500,000 reads derived from 4 out of the 5 viral classes, the human transcriptome and unrelated viral genomes in variable proportions. Reads were simulated using the ART software (30) with Illumina® HiSeq2500 error profiles that were 150nt long and modelled single end experiments. One of the five viral classes that was excluded was considered as the test class. This process was automated using a custom script (generatefprdata.pl).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Subsequently, PACIFIC was run in *double prediction* to assign classes to each read. To calculate the percentage of false positives in each experiment, we counted the number of reads predicted as the absent test class and divided by the total number of reads.

Detecting viruses in human datasets and comparison with other tools

We downloaded 63 RNA-seq experiments from NCBI SRA database. Run accessions and other metadata details are supplied in Additional file 1: Table S2. All data were downloaded from the NCBI database using the SRA Toolkit *prefetch* and *fastq-dump* commands and applying the *--gzip* and *--fasta* options (52). For the GALA II cohort study with 48 RNA-seq datasets and read lengths 18-390nt, we discarded reads <150nt long. We then used PACIFIC to assign the presence/absence of each virus class in all 63 samples using the detection thresholds established in the previous section. We compared PACIFIC's predictions with two alternative methods for virus detection: an alignment-based approach using BWA-MEM (53), and a k-mer based approach using Kraken2 (19), described below.

For BWA-MEM (53), all reads were mapped using default parameters to a combined reference containing assembly sequences for the five viral classes and the human transcriptome used for training PACIFIC (Table 4). Reads were assigned to a virus class based on the class membership of the genome assembly as described in Table 4 and Additional file 1: Table S1. For Kraken2, we first downloaded the Kraken taxonomy database and built a k-mer database using the same genomes used to train PACIFIC (Table 4). Kraken2 was then run using the *–use-names* flag, and output reads were parsed using species scientific names and reads were assigned a class based on the class membership of the genome assembly (Additional file 1: Table S1, Table 4). To fairly compare all three methods, we applied class detection thresholds as determined for and used in PACIFIC (Table 2) for the presence or absence of a virus class within a sample.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

599 To investigate the origin of reads for all reads in samples that were discordantly predicted for
600 the presence of a virus class by PACIFIC, BWA-MEM or Kraken2, we used the BLAST suite
601 (v2.10.1+) (54,55) to align reads to the NCBI nucleotide (*nt*) database, which includes
602 sequences from all domains of life. We took the best hit from the pairwise alignment for each
603 read, filtering for alignments with an E-value <1e-6. BLASTN was used with the following
604 parameters: *-task 'megablast' -max_target_seqs 1 -max_hsps 1 -evalue 1e-6* to query
605 discordant viral class assignments between PACIFIC, BWA-MEM and Kraken2.

606

607 ## Declarations

608 ### Ethics approval and consent to participate

609 Not applicable.

610 ### Consent for publication

611 Not applicable.

612 ### Availability of data and materials

613 Source code is available in the PACIFIC Github repository [https://github.com/pacific-
614 2020/pacific/] under the MIT License. Publicly available data generated or analysed during
615 this study are included in the following published articles (31,32,34), accession numbers and
616 other metadata are described in [https://github.com/pacific-
617 2020/pacific/tree/master/metadata]. Other supplementary information and test data are
618 available and can be downloaded from
619 [https://cloudstor.aarnet.edu.au/plus/s/sRLwF3IJQ12pNGQ].

620 ### Competing interests

621 The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Funding

PAM is supported by the John Curtin School of Medical Research PhD scholarship and by EMBL Australia. RFB is supported by the Australian Government Research Training Program PhD scholarship and the Australian Genome Health Alliance PhD Award. EE is supported by EMBL Australia. HRP is supported by the Australian National University Research Fellowship. Computational resources were provided by the Australian Government through the National Computational Infrastructure (NCI) under the National Computational Merit Allocation Scheme awarded to HRP and SE. This research was also supported by allocations awarded to HRP under the National Computational Infrastructure by use of the Nectar Research Cloud. The Nectar Research Cloud is a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS).

Authors' contributions

PAM, RFB, EE and HRP conceptualised the project. PAM contributed to the training and initial testing of the model. PAM, RFB and HRP performed all data analyses with input from EE. HRP, SE and EE supervised the project. PAM, RFB, EE and HRP drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to dedicate this study to Smt. Mukta Sheladiya and all other unfortunate souls who lost their battle against the COVID-19 infection. We would like to offer our sincere gratitude to all healthcare workers who have supported millions of people during these trying times across the world. We thank Dr. Cheng Soon Ong for his input in the development of the model. Finally, we thank Dr. Saul Newman and Dr. Teresa (Terry) Neeman for their valuable feedback and comments in preparation of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

646 ## References

647 1. World Health Organization. WHO - The top 10 causes of death [Internet]. 24 Maggio.
648 2018 [cited 2020 Jun 17]. p. 1–7. Available from: https://www.who.int/news-
649 room/fact-sheets/detail/the-top-10-causes-of-death

650 2. Legand A, Briand S, Shindo N, Brooks WA, De Jong MD, Farrar J, et al. Addressing
651 the public health burden of respiratory viruses: the Battle against Respiratory Viruses
652 (BRaVe) Initiative. Vol. 8, Future Virology. Future Medicine Ltd.; 2013. p. 953–68.

653 3. Soriano JB, Kendrick PJ, Paulson KR, Gupta V, Abrams EM, Adedoyin RA, et al.
654 Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017:
655 a systematic analysis for the Global Burden of Disease Study 2017. Lancet Respir
656 Med. 2020 Jun 1;8(6):585–96.

657 4. Tang JW, Lam TT, Zaraket H, Lipkin WI, Drews SJ, Hatchette TF, et al. Global
658 epidemiology of non-influenza RNA respiratory viruses: data gaps and a growing need
659 for surveillance. Vol. 17, The Lancet Infectious Diseases. Lancet Publishing Group;
660 2017. p. e320–6.

661 5. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Vol. 17, Nature
662 Reviews Microbiology. Nature Publishing Group; 2019. p. 181–92.

663 6. Al-Omari A, Rabaan AA, Salih S, Al-Tawfiq JA, Memish ZA. MERS coronavirus
664 outbreak: Implications for emerging viral infections. Vol. 93, Diagnostic Microbiology
665 and Infectious Disease. Elsevier Inc.; 2019. p. 265–85.

666 7. Centers for Disease Control and Prevention (CDC). Revised U.S. surveillance case
667 definition for severe acute respiratory syndrome (SARS) and update on SARS cases--
668 United States and worldwide, December 2003. MMWR Morb Mortal Wkly Rep.
669 2003;52(49):1202–6.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

8. WHO EMRO | MERS situation update, January 2020 | MERS-CoV | Epidemic and pandemic diseases [Internet]. [cited 2020 Jul 13]. Available from: http://www.emro.who.int/pandemic-epidemic-diseases/mers-cov/mers-situation-update-january-2020.html

9. Bezerra PGM, Britto MCA, Correia JB, Duarte M do CMB, Fonceca AM, Rose K, et al. Viral and atypical bacterial detection in acute respiratory infection in children under five years. PLoS One. 2011;6(4).

10. May L, Tatro G, Poltavskiy E, Mooso B, Hon S, Bang H, et al. Rapid Multiplex Testing for Upper Respiratory Pathogens in the Emergency Department: A Randomized Controlled Trial. Open Forum Infect Dis. 2019 Nov 5;6(12).

11. Kim D, Quinn J, Pinsky B, Shah NH, Brown I. Rates of Co-infection between SARS-CoV-2 and Other Respiratory Pathogens. Vol. 323, JAMA - Journal of the American Medical Association. American Medical Association; 2020. p. 2085–6.

12. Tong X, Xu X, Lv G, Wang H, Cheng A, Wang D, et al. Clinical characteristics and outcome of influenza virus infection among adults hospitalized with severe COVID-19: A retrospective cohort study from Wuhan, China. Research Square; 2020.

13. Wang G, Xie M, Ma J, Guan J, Song Y, Wen Y, et al. Is Co-Infection with Influenza Virus a Protective Factor of COVID-19? SSRN Electron J. 2020 May 6;

14. Rockett RJ, Arnott A, Lam C, Sadsad R, Timms V, Gray K-A, et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat Med. 2020 Jul 9;1–7.

15. Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE. Multiplex PCR: Optimization and application in diagnostic virology. Vol. 13, Clinical Microbiology Reviews. American Society for Microbiology (ASM); 2000. p. 559–70.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

16. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 2019 Jul 19;20(4):1125–36.

17. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar 3;15(3):R46.

18. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 2018 Nov 16;19(1):198.

19. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257.

20. Bzhalava Z, Tampuu A, Bała P, Vicente R, Dillner J. Machine Learning for detection of viral sequences in human metagenomic datasets. BMC Bioinformatics. 2018 Sep 24;19(1):336.

21. Liang Q, Bible PW, Liu Y, Zou B, Wei L. DeepMicrobes: taxonomic classification for metagenomics with deep learning. NAR Genomics Bioinforma. 2020 Mar 1;2(1).

22. Tampuu A, Bzhalava Z, Dillner J, Vicente R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. Melcher U, editor. PLoS One. 2019 Sep 11;14(9):e0222271.

23. Li H, Li X, Caragea D, Caragea C. Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks. Proc ISCRAM Asian Pacific 2018 Conf. 2018;(November):1–13.

24. Trabelsi A, Chaabane M, Ben-Hur A. Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. Bioinformatics. 2019 Jul 5;35(14):i269–77.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

718 25. Gong Y, Wang L, Guo R, Lazebnik S. Multi-scale orderless pooling of deep
719 convolutional activation features. In: Lecture Notes in Computer Science (including
720 subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
721 Springer Verlag; 2014. p. 392–407.

722 26. Siami-Namini S, Tavakoli N, Namin AS. The Performance of LSTM and BiLSTM in
723 Forecasting Time Series. In: 2019 IEEE International Conference on Big Data (Big
724 Data). 2019. p. 3285–92.

725 27. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic
726 Acids Res. 2015/11/20. 2016 Jan 4;44(D1):D67–72.

727 28. Zhang Q, Jun SR, Leuze M, Ussery D, Nookaew I. Viral phylogenomics using an
728 alignment-free method: A three-step approach to determine optimal length of k-mer.
729 Sci Rep. 2017 Jan 19;7(1):1–13.

730 29. Lin J. Divergence Measures Based on the Shannon Entropy. Vol. 37, IEEE
731 TRANSACTIONS ON INFORMATION THEORY. 1991.

732 30. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read
733 simulator. Bioinformatics. 2011 Dec 23;28(4):593–4.

734 31. Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S, Hoagland D, Møller R, et al.
735 Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. Cell.
736 2020 May 28;181(5):1036-1045.e9.

737 32. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus
738 associated with human respiratory disease in China. Nature. 2020 Mar
739 12;579(7798):265–9.

740 33. Kumar R, Nguyen EA, Roth LA, Oh SS, Gignoux CR, Huntsman S, et al. Factors
741 associated with degree of atopy in Latino children in a nationwide pediatric sample:

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The Genes-environments and Admixture in Latino Asthmatics (GALA II) study. J Allergy Clin Immunol. 2013;132(4).

34. Wesolowska-Andersen A, Everman JL, Davidson R, Rios C, Herrin R, Eng C, et al. Dual RNA-seq reveals viral infections in asthmatic children without respiratory illness which are associated with changes in the airway transcriptome. Genome Biol. 2017 Jan 19;18(1):12.

35. Wesolowska-Andersen A, Everman JL, Davidson R, Rios C, Herrin R, Eng C, et al. Correction: Dual RNA-seq reveals viral infections in asthmatic children without respiratory illness which are associated with changes in the airway transcriptome [Genome Biol., 18, (2017) (12)] DOI: 10.1186/s13059-016-1140-8. Vol. 19, Genome Biology. BioMed Central Ltd.; 2018. p. 49.

36. Tapparel C, Junier T, Gerlach D, Cordey S, Van Belle S, Perrin L, et al. New complete genome sequences of human rhinoviruses shed light on their phylogeny and genomic features. BMC Genomics. 2007 Jul 10;8(1):224.

37. Tapparel C, Junier T, Gerlach D, Van Belle S, Turin L, Cordey S, et al. New respiratory enterovirus and recombinant rhinoviruses among circulating picornaviruses. Emerg Infect Dis. 2009 May;15(5):719–26.

38. Aynaud MM, Mirabeau O, Gruel N, Grossetête S, Boeva V, Durand S, et al. Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. Cell Rep. 2020 Feb 11;30(6):1767-1779.e6.

39. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. BMJ. 2020 Apr 7;369.

40. Who. Research needs for the Battle against Respiratory Viruses (BRaVe). Future

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Virol. 2013;1–35.

41. Langelier C, Kalantar KL, Moazed F, Wilson MR, Crawford ED, Deiss T, et al. Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. Proc Natl Acad Sci U S A. 2018 Dec 26;115(52):E12353–62.

42. Graf EH, Simmon KE, Tardif KD, Hymas W, Flygare S, Eilbeck K, et al. Unbiased detection of respiratory viruses by use of RNA sequencing-based metagenomics: A systematic comparison to a commercial PCR panel. J Clin Microbiol. 2016 Apr 1;54(4):1000–7.

43. Brinkmann A, Andrusch A, Belka A, Wylezich C, Höper D, Pohlmann A, et al. Proficiency testing of virus diagnostics based on bioinformatics analysis of simulated in silico high-throughput sequencing data sets. J Clin Microbiol. 2019 Aug 1;57(8):466–85.

44. Hogan CA, Sahoo MK, Pinsky BA. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. Vol. 323, JAMA - Journal of the American Medical Association. American Medical Association; 2020. p. 1967–9.

45. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral Mutation Rates. J Virol. 2010 Oct 1;84(19):9733–48.

46. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Vol. 73, Cellular and Molecular Life Sciences. Birkhauser Verlag AG; 2016. p. 4433–48.

47. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data. 2016 Dec 1;3(1):1–40.

48. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Res. 2018 Oct 24;47(D1):D766–73.

49. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2019 Nov 6;48(D1):D682–8.

50. Charles PWD. Project Title. GitHub repository. GitHub; 2013.

51. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR; 2015.

52. SRA Toolkit Development Team. SRA Toolkit [Internet]. Vol. 10. 2018 [cited 2020 Jun 17]. p. 2017–9. Available from: http://ncbi.github.io/sra-tools/

53. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 Mar 16;

54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

55. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008 Jun 21;24(16):1757–64.

Supplementary Information

Additional file 1. Supplementary Tables S1, S2, and S3.

- **Supplementary Table S1**: Summary table of genomes and assemblies used to train PACIFIC.
- **Supplementary Table S2**: PACIFIC testing metrics.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.24.219097; this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- **Supplementary Table S3.** Publicly available samples used to run PACIFIC, BWA and Kraken2.

Additional file 2. Supplementary Figures S1 and S2, Details of the BLAST analysis.