Interpretable detection of novel human viruses from genome sequencing data

Jakub M. Bartoszewicz [1 2 3 4] **Anja Seidel** [1 2 5] **Bernhard Y. Renard** [1 3 4]

Abstract

Viruses evolve extremely quickly, so reliable methods for viral host prediction are necessary to safeguard biosecurity and biosafety alike. Novel human-infecting viruses are difficult to detect with standard bioinformatics workflows. Here, we predict whether a virus can infect humans directly from next-generation sequencing reads. We show that deep neural architectures significantly outperform both shallow machine learning and standard, homology-based algorithms, cutting the error rates in half and generalizing to taxonomic units distant from those presented during training. We propose a new approach for convolutional filter visualization to disentangle the information content of each nucleotide from its contribution to the final classification decision. Nucleotide-resolution maps of the learned associations between pathogen genomes and the infectious phenotype can be used to detect virulence-related genes in novel agents, as we show here for the SARS-CoV-2 coronavirus, unknown before it caused a COVID-19 pandemic in 2020.

1. Introduction

1.1. Background

Within a globally interconnected and densely populated world, pathogens can spread more easily than they ever did before. As the recent outbreaks of Ebola and Zika viruses have shown, the risks posed even by these previously known agents remain unpredictable and their expansion hard to control (Calvignac-Spencer et al., 2014). What is more, it is almost certain that more unknown pathogen species and strains are yet to be discovered, given their constant, extremely fast-paced evolution and unexplored biodiversity, as well as increasing human exposure (Vouga & Greub, 2016; Trappe et al., 2016). Some of those novel pathogens may cause epidemics (similar to the SARS and MERS coronavirus outbreaks in 2002 and 2012) or even pandemics (e.g. SARS-CoV-2 and the "swine flu" H1N1/09 strain). Many have more than one host or vector, which makes assessing and predicting the risks even more difficult. For example, Ebola has its natural reservoir most likely in fruit bats (Leendertz et al., 2016), but causes deadly epidemics in both humans and chimpanzees. As the state-of-the art approach for the open-view detection of pathogens is genome sequencing (Lecuit & Eloit, 2014; Calistri & Palù, 2015), it is crucial to develop automated pipelines for characterizing the infectious potential of currently unidentifiable sequences.

Screening against potentially dangerous subsequences before their synthesis may also be used as a way of ensuring responsible research in synthetic biology. While potentially useful in some applications, engineering of viral genomes could also pose a biosecurity and biosafety threat. Two controversial studies modified the influenza A/H5N1 ("bird flu") virus to be airborne transmissible in mammals (Herfst et al., 2012; Imai et al., 2012). A possibility of modifying coronaviruses to enhance their virulence triggered calls for a moratorium on this kind of research (Lipsitch & Inglesby, 2014). Synthesis of an infectious horsepox virus closely related to the smallpox-causing *Variola* virus (Noyce et al., 2018) caused a public uproar and calls for intensified discussion on risk control in synthetic biology (Thiel, 2018).

1.2. Current tools for host range prediction

Several computational, genome-based methods exist that allow to predict the host-range of a bacteriophage (a bacteria-infecting virus). A selection of composition-based and alignment-based approaches has been presented in an extensive review by Edwards et al. (2016). Prediction of eukariotic host tropism (including humans) based on known protein sequences was shown for the influenza A virus (Eng et al., 2014). Two recent studies employ k-mer based, k-NN classifiers (Li & Sun, 2018) and deep learning (Mock et al., 2019) to predict host range for a small set of three well-studied species directly from viral sequences. While

[1]Bioinformatics (MF1), Department of Methodology and Research Infrastructure, Robert Koch Institute, Berlin, Germany [2]Department of Mathematics and Computer Science, Free University of Berlin, Berlin, Germany [3]Hasso Plattner Institute for Digital Engineering, Potsdam, Brandenburg, Germany [4]Digital Engineering Faculty, University of Postdam, Potsdam, Brandenburg, Germany [5]Currently at: Central Research Institute of Ambulatory Health Care, Berlin, Germany. Correspondence to: Jakub M. Bartoszewicz <jakub.bartoszewicz@hpi.de>, Bernhard Y. Renard <bernhard.renard@hpi.de>.

those approaches are limited to those particular species and do not scale to viral host-range prediction in general, the Host Taxon Predictor (HTP) (Gałan et al., 2019) uses logistic regression and support vector machines to predict if a novel virus infects bacteria, plants, vertebrates or arthropods. Yet, the authors argue that it is not possible to use HTP in a read-based manner; it requires long sequences of at least 3,000 nucleotides. This is incompatible with modern metagenomic next-generation sequencing workflows, where the DNA reads obtained are at least 10-20 times shorter. Another study used gradient boosting machines to predict reservoir hosts and transmission via arthropod vectors for known human-infecting viruses (Babayan et al., 2018).

Zhang et al. (2019) designed several classifiers explicitly predicting whether a new virus can potentially infect humans. Their best model, a k-NN classifier, uses k-mer frequencies as features representing the query sequence and can yield predictions for sequences as short as 500 base pairs (bp). It worked also with 150bp-long reads from real DNA sequencing runs, although in this case the reads originated also from the viruses present in the training set (and were therefore not "novel").

1.3. Deep Learning for DNA sequences

While DNA sequences mapped to a reference genome may be represented as images (Poplin et al., 2018), a majority of studies uses a distributed orthographic representation, where each nucleotide $\{A, C, G, T\}$ in a sequence is represented by a one-hot encoded vector of length 4. An "unknown" nucleotide (N) can be represented as an all-zero vector. CNNs and LSTMs have been successfully used for a variety of DNA-based prediction tasks. Early works focused mainly on regulation of gene expression in humans (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015; Zeng et al., 2016; Quang & Xie, 2016; Kelley et al., 2016), which is still an area of active research (Greenside et al., 2018; Nair et al., 2019; Avsec et al., 2019). In the field of pathogen genomics, deep learning models trained directly on DNA sequences were developed to predict host ranges of three multi-host viral species (Mock et al., 2019) and to predict pathogenic potentials of novel bacteria (Bartoszewicz et al., 2019). DeepVirFinder (Ren et al., 2018) and ViraMiner (Tampuu et al., 2019) can detect viral sequences in metagenomic samples, but they cannot predict the host and focus on previously known species. For a broader view on deep learning in genomics we refer to a recent review by Eraslan et al. (2019).

Interpretability and explainability of deep learning models for genomics is crucial for their wide-spread adoption, as it is necessary for delivering trustworthy and actionable results. Convolutional filters can be visualized by forward-passing multiple sequences through the network and extracting the most-activating subsequences (Alipanahi et al., 2015) to create a position weight matrix (PWM) which can be visualized as a sequence logo (Schneider & Stephens, 1990; Crooks et al., 2004). Direct optimization of input sequences is problematic, as it results in generating a dense matrix even though the input sequences are one-hot encoded (Lanchantin et al., 2016; 2017). This problem can be alleviated with Integrated Gradients (Sundararajan et al., 2016; Jha et al., 2019) or DeepLIFT, which propagates activation differences relative to a selected reference back to the input, reducing the computational overhead of obtaining accurate gradients (Shrikumar et al., 2019a). If a reference of all-zeros is used, the method is analogous to Layer-wise Relevance Propagation (Bach et al., 2015). DeepLIFT is an additive feature attribution method, and may used to approximate Shapley values if the input features are independent (Lundberg & Lee, 2017). TF-MoDISco (Shrikumar et al., 2019b) uses DeepLIFT to discover consolidated, biologically meaningful DNA motifs (transcription factor binding sites).

1.4. Contributions

In this paper, we first improve the performance of read-based predictions of the viral host (human or non-human) from next-generation sequencing reads. We show that reverse-complement (RC) neural networks (Bartoszewicz et al., 2019) significantly outperform both the previous state-of-the-art (Zhang et al., 2019) and the traditional, alignment-based algorithm – BLAST (Altschul et al., 1990), which constitutes a gold standard in homology-based bioinformatics analyses. We show that defining the negative (non-human) class is non-trivial and compare different ways of constructing the training set. Strikingly, a model trained to distinguish between viruses infecting humans and viruses infecting other chordates (a phylum of animals including vertebrates) generalizes well to evolutionarily distant non-human hosts, including even bacteria. This suggests that the host-related signal is strong and the learned decision boundary separates human viruses from other DNA sequences surprisingly well.

Next, we propose a new approach for convolutional filter visualization using partial Shapley values to differentiate between simple nucleotide information content and the contribution of each sequence position to the final classification score. To test the biological plausibility of our models, we generate genome-wide maps of "infectious potential" and nucleotide contributions. We show that those maps can be used to visualize and detect virulence-related regions of interest (e.g. genes) in novel genomes. Finally, we analyze a recently discovered SARS-CoV-2 coronavirus, which caused a pandemic in 2020 (Wu et al., 2020).

2. Methods

2.1. Data collection and preprocessing

2.1.1. VHDB DATASET

We accessed the Virus-Host Database (Mihara et al., 2016) on July 31, 2019 and downloaded all the available data. The original dataset contained 14,380 records comprising RefSeq IDs for viral sequences and associated metadata. Some viruses are divided into discontiguous segments, which are represented as separate records in VHDB; in those cases the segments were treated as contigs of a single genome in the further analysis. We removed records with unspecified host information and those confusing the highly pathogenic Variola virus with a similarly named genus of fish. Further, we filtered out viroids and satellites. Human-infecting viruses were extracted by searching for records containing "Homo sapiens" in the "host name" field. Note that VHDB contains information about multiple possible hosts for a given virus where appropriate. Any virus infecting humans was assigned to the positive class, also if other, non-human hosts exist. In total, the dataset contained 9,496 viruses, including 1,309 human viruses. We considered both DNA and RNA viruses; RNA sequences were encoded in the DNA alphabet, as in RefSeq.

2.1.2. DEFINING THE NEGATIVE CLASS

While defining a human-infecting class is relatively straightforward, the reference negative class may be conceptualized in a variety of ways. The broadest definition takes all non-human viruses into account, including bacteriophages (bacterial viruses). This is especially important, as most of known bacteriophages are DNA viruses, while many important human (and animal) viruses are RNA viruses. One could expect that the multitude of available bacteriophage genomes dominating the negative class could lower the prediction performance on viruses similar to those infecting humans. This offers an open-view approach covering a wider part of the sequence space, but may lead to misclassification of potentially dangerous mammalian or avian viruses. As they are often involved in clinically relevant host-switching events, a stricter approach must also be considered. In this case, the negative class comprises only viruses infecting Chordata (a group containing vertebrates and closely related taxa). Two intermediate approaches consider all eukaryotic viruses (including plant and fungi viruses), or only animal-infecting viruses. This amounts to four nested host sets: "All" (8,187 non-human viruses), "Eukaryota" (5,114 viruses), "Metazoa" (2,942 viruses) and "Chordata" (2,078 viruses). Auxiliary sets containing only non-eukaryotic viruses ("non-Eukaryota"), non-animal eukaryotic viruses ("non-Metazoa Eukaryota") etc. can be easily constructed by set subtraction.

For the positive class, we generated a training set containing 80% of the genomes, and validation and test sets with 10% of the genomes each. Importantly, the nested structure was kept also during the training-validation-test split: for example, the species assigned to the smallest test set ("Chordata") were also present in all the bigger test sets. The same applied to other taxonomic levels, as well as the training and validation sets wherever applicable.

2.1.3. READ SIMULATION

We simulated 250bp long Illumina reads following a modification of a previously described protocol (Bartoszewicz et al., 2019) and using the Mason read simulator (Holtgrewe, 2010). First, we only generated the reads from the genomes of human-infecting viruses. Then, the same steps were applied to each of the four negative class sets. Finally, we also generated a fifth set, "Stratified", containing an equal number of reads drawn from genomes of the following disjunct host classes: "Chordata" (25%), "non-Chordata Metazoa" (25%), "non-Metazoa Eukaryota" (25%) and "non-Eukaryota" (25%).

In each of the evaluated settings, we used a total of 20 million (80%) reads for training, 2.5 million (10%) reads for validation and 2.5 million (10%) paired reads as the held-out test set. Read number per genome was proportional to genome length, keeping the coverage uniform on average. While the original datasets are heavily imbalanced, we generated the same number of negative and positive data points (reads) regardless of the negative class definition used.

This protocol allowed us to test the impact of defining the negative class, while using the exactly same data as representatives of the positive class. We used three training and validation sets ("All", "Stratified", and "Chordata"), representing the fully open-view setting, a setting more balanced with regard to the host taxonomy, and a setting focused on cases most likely to be clinically relevant. In each setting, the validation set matched the composition of the training set. The evaluation was performed using all five test sets to gain a more detailed insight on the effects of negative class definition on the prediction performance.

2.1.4. HUMAN BLOOD VIROME DATASET

Similarily to Zhang et al. (2019), we used the human blood DNA virome dataset (Moustafa et al., 2017) to test the selected classifiers on real data. We obtained 14,242,329 reads of 150bp and searched all of VHDB using blastn (with default parameters) to obtain high-quality reference labels. If a read's best hit was a human-infecting virus, we assigned it to a positive class; the negative class was assigned if this was not the case. This procedure yielded 14,012,665 "positive" and 229,664 "negative" reads.

2.2. Training

We used the DeePaC package (Bartoszewicz et al., 2019) to investigate RC-CNN and RC-LSTM architectures, previously shown to accurately predict bacterial pathogenicity. Here, we employ a CNN with two convolutional layers with 512 filters of size 15 each, average pooling and 2 fully connected layers with 256 units each. The LSTM used has 384 units. We use dropout regularization in both cases, together with aggressive input dropout at the rate of 0.2 or 0.25 (tuned for each model). Input dropout may be interpreted as a special case of noise injection, where a fraction of input nucleotides is turned to Ns. Representations of forward and reverse-complement strands are summed before the fully connected layers. As two mates in a read pair should originate from the same virus, predictions obtained for them can be averaged for a boost in performance. If a contig or genome is available, averaging predictions for constituting reads yields a prediction for the whole sequence. We used Tesla P100 GPUs for training and an RTX 2080 Ti for visualizations.

We wanted the networks to yield accurate predictions for both 250bp (our data, modelling a sequencing run of an Illumina MiSeq device) and 150bp long reads (as in the Human Blood Virome dataset). As shorter reads are padded with zeros, we expected the CNNs trained using average pooling to misclassify many of them. Therefore, we prepared a modified version of the "Stratified" dataset, in which the last 100bp of each read were turned to zeros, mocking a shorter sequencing run while preserving the error model. Then, we retrained the CNN which had performed best on the original dataset. Since in principle, the Human Blood Virome dataset should not contain viruses infecting non-human Chordata, a "Chordata"-trained classifier was not used in this setting.

2.3. Benchmarking

We compare the networks to the k-NN classifier proposed by Zhang et al. (2019) and used by them for read-based predictions. We trained the classifier on the "All" dataset as described by the authors, i.e. using non-overlapping, 500bp-long contigs generated from the training genomes (retraining on simulated reads is computationally prohibitive). We also tested the performance of using BLAST to search against an indexed database of labeled genomes. We constructed the database from the "All" training set and used discontiguous megablast to achieve high inter-species sensitivity.

Note that both BLAST and k-NN can yield conflicting predictions for the individual mates in a read pair. What is more, BLAST yields no prediction at all if no match is found. Therefore, similarly to Bartoszewicz et al. (2019), we used the *accept anything* operator to integrate binary predictions for read pairs and genomes. At least one match is needed to predict a label, and conflicting predictions are treated as if no match was found at all. Missing predictions lower both true positive and true negative rates.

2.4. Filter visualization

2.4.1. Substring extraction

In order to visualize the learned convolutional filters, we downsample a matching test set to 125,000 reads and pass it through the network. This is modelled after the method presented by Alipanahi et al. (2015). For each filter and each input sequence, the authors extracted a subsequence leading to the highest activation, and created sequence logos from the obtained sequence sets ("max-activation"). We used DeepLIFT (Shrikumar et al., 2019a) to extract score-weighted subsequences with the highest contribution score ("max-contrib") or all subsequences with non-zero contributions ("all-contrib"). Computing the latter was costly and did not yield better quality logos.

We use an all-zero reference. As reads from real sequencing runs are usually not equally long, shorter reads must be padded with Ns; the "unknown" nucleotide is also called whenever there is not enough evidence to assign any other to the raw sequencing signal. Therefore, Ns are "null" nucleotides and are a natural candidate for the reference input. We do not consider alternative solutions based on GC content or dinucleotide shuffling, as the input reads originate from multiple different species, and the sequence composition may itself be a strong marker of both virus and host taxonomy.

However, some of the training sequences contain Ns themselves. It is therefore possible that a filter will learn only negative weights at a given position, even though there is no biological justification for that. This may lead to assigning only negative contributions to all four possible nucleotides at a given position if the filter's contribution is positive (and positive nucleotide contributions if the filter's contribution is negative). To solve the problem, we first normalize the weight matrices position-wise, as described by Shrikumar et al. (2019a). Finally, we calculate average filter contributions to obtain a crude ranking of feature importance with regard to both the positive and negative class.

2.4.2. Partial Shapley values

Building sequence logos involves calculating information content (IC) of each nucleotide at each position in a prospective DNA motif. This can be then interpreted as measure of evolutionary sequence conservation. However, high IC does not necessarily imply that a given nucleotide is relevant in terms of its contribution to the classifier's output. Some sub-motifs may be present in the sequences used to build the logo, even if they do not contribute to the final prediction (or even a given filter's activation).

To test this hypothesis, we use partial Shapley values. Intuitively speaking, we capture the contributions of a nucleotide to the network's output, but only in the context of a given intermediate neuron of the convolutional layer. More precisely, for any given feature x_i, intermediate neuron y_j and the output neuron z, we aim to measure how x_i contributes to z while regarding only the fraction of the total contribution of x_i that influences how y_j contributes to z.

Using the formalism of DeepLIFT's multipliers (Shrikumar et al., 2019a) and their reinterpretation in SHAP (Lundberg & Lee, 2017), we backpropagate the activation differences only along the paths "passing through" y_j. In Eq. 1, we define partial multipliers $\mu_{x_i z}^{(y_j)}$ and express them in terms of Shapley values ϕ and activation differences w.r.t. the expected activation values (reference activation). Calculating partial multipliers is equivalent to zeroing out the multipliers $m_{y_k z}$ for all $k \neq j$ before backpropagating $m_{y_j z}$ further.

$$\mu_{x_i z}^{(y_j)} = m_{x_i y_j} m_{y_j z} = \frac{\phi_i(y_j, x)\phi_j(z, y)}{(x_i - E[x_i])(y_j - E[y_j])} \quad (1)$$

We define partial Shapley values $\varphi_i^{(y_j)}(z, x)$ analogously to how Shapley values can be approximated by a product of multipliers and input differences w.r.t. the reference (Eq. 2):

$$\varphi_i^{(y_j)}(z, x) = \mu_{x_i z}^{(y_j)}(x_i - E[x_i]) = \frac{\phi_i(y_j, x)\phi_j(z, y)}{y_j - E[y_j]} \quad (2)$$

From the chain rule for multipliers (Shrikumar et al., 2019a), it follows that standard multipliers are a sum over all partial multipliers for a given layer y. Therefore, Shapley values as approximated by DeepLIFT are a sum of partial Shapley values for the layer y.

Once we calculate the contributions of convolutional filters for the first layer, $\varphi_i^{(y_j)}(z, x)$ for the first convolutional layer of a network with one-hot encoded inputs and an all-zero reference can be efficiently calculated using weight matrices and filter activation differences (Eq. 3-4). First, in this case we do not traverse any non-linearities and can directly use the linear rule (Shrikumar et al., 2019a) to calculate the contributions of x_i to y_j as a product of the weight w_i and the input x_i. Second, the input values may only be 0 or 1.

$$\phi_i(y_j, x) = w_i x_i = \begin{cases} w_i & \text{if } x_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$\varphi_i^{(y_j)}(z, x) = \frac{w_i \phi_j(z, y)}{y_j - E[y_j]} \quad (4)$$

Resulting partial contributions can be visualized along the IC of each nucleotide of a convolutional kernel. To this end, we design extended sequence logos, where each nucleotide is colored according to its contribution. Positive contributions are shown in red, negative contributions are blue, and near-zero contributions are gray. Therefore, no information is lost compared to standard sequence logos, but the relevance of individual nucleotides and the filter as a whole can be easily seen.

2.5. Genome-wide phenotype analysis

We create genome-wide phenotype analysis (GWPA) plots to analyse which parts of a viral genome are associated with the infectious phenotype. We scramble the genome into overlapping, 250bp long subsequences (pseudo-reads) without adding any sequencing noise. For the highest resolution, we use a stride of one nucleotide. We predict the infectious potential of each pseudo-read and average the obtained values at each position of the genome. Analogously, we calculate average contributions of each nucleotide to the final prediction of the convolutional network. We visualize the resulting nucleotide-resolution maps with IGV (Thorvaldsdóttir et al., 2013).

For well-annotated genomes, we compile a ranking of genes (or other genomic features) sorted by the average infectious potential within a given region. In addition to that, we scan the genome with the learned filters of the first convolutional layer to find genes enriched in subsequences yielding non-zero filter activations. We use Gene Ontology to connect the identified genes of interest with their molecular functions and biological processes they are engaged in.

As a proof of concept, we analyze one of the viruses randomly assigned to the test set – the Taï Forest ebolavirus, which has a history of host-switching and can cause a serious disease. To show that the method can also be used for other biological problems, we investigated the networks trained by Bartoszewicz et al. (2019) and their predictions on a genome of a pathogenic bacterium *Staphylococcus aureus*. The authors used this particular species to assess the performance of their method on real sequencing data. In this case, we used a stride of 125bp to generate the pseudo-reads. Finally, we analyzed the SARS-CoV-2 coronavirus, which emerged in December 2019, causing a pneumonia outbreak (Wu et al., 2020).

3. Results

3.1. Negative class definition

Choosing which viruses should constitute the negative class is application dependent and influences the performance of the trained models. Table 1 summarizes the prediction accuracy for different combinations of the training and test

set composition. The models trained only on human and Chordata-infecting viruses maintain similar, or even better performance when evaluated on viruses infecting a much broader host range, including bacteria. This suggests that the learned decision boundary separates human viruses from all the others surprisingly well. We hypothesize that the human host signal must be relatively strong and contained within the Chordata host signal. Dropout rate of 0.2 resulted in the highest validation accuracy for $CNN_{Str\text{-}150}$ and $LSTM_{Str}$. A rate of 0.25 was selected for the other models.

Adding more diversity to the negative class may still boost performance on more diverse test sets, as in the case of CNN trained on the "All" dataset (CNN_{All}). This model performs a bit worse on viruses infecting hosts related to humans, but achieves higher accuracy than the "Chordata"-trained models and the best recall overall. Rebalancing the negative class using the "Stratified" dataset helps to achieve higher performance on animal viruses while maintaing high overall accuracy. The LSTMs are outperformed by the CNNs, but they can be used for shorter reads without retraining (see Sections 2.2 and 3.2).

3.2. Prediction performance

We selected $LSTM_{All}$ and CNN_{All} for further evaluation. Table 2 presents the results of a benchmark using the "All" test set. Low performance of the k-NN classifier (Zhang et al., 2019) is caused by frequent conflicting predictions for each read in a read pair (in a single-read setting it achieves 75.5% accuracy, while our best model – 87.8%). Although BLAST achieves the highest precision, it yields no predictions for over 10% of the samples. CNN_{All} is the most sensitive and accurate (Table 3).

We benchmarked our models against the human blood virome dataset used by Zhang et al. (2019). Our models outperform their k-NN classifier. As the positive class massively outnumbers the negative class, all models achieve over 99% precision. $CNN_{All\text{-}150}$ performs best (Table. 4). However, the positive class is dominated by viruses which are not necessarily novel. The CNN was more accurate on training data, so we expected it to detect those viruses easily.

3.3. Filter visualization

In the Fig. 1 we present example filters, visualized as "max-contrib" sequence logos based on mean partial Shapley values for each nucelotide at each position. All nucleotides of the filters with the highest (Fig. 1a) or lowest (Fig. 1b) score have relatively strong contributions in accordance with the filters' own contributions. However, we observe that sometimes, there is a "conserved" nucleotide which consistently appears in the activating subsequences, but the sign of its contributions is opposite to the filter's (Fig. 1c). Those "counter-contributions" may arise if a nucleotide with a negative weight forms a frequent motif with others with positive weights strong enough to activate the filter. We comment on this fact in the Section 4.2. Some filters seem to learn gapped motifs resembling a codon structure (Fig. 1d). We extracted this filter from a network predicting bacterial pathogenicity trained by Bartoszewicz et al. (2019), but we find similar filters in our networks as well. We scanned a genome of *S. aureus* with this filter and discovered that the learned motif is indeed significantly enriched in coding sequences (Fisher exact test with Benjamini-Hochberg correction, $q < 10^{-15}$). It is also enriched in a number of specific genes. The one with the most hits (sraP, $q < 10^{-15}$) is associated with virulence in endovascular infection.

Table 1. Classification accuracy depending on the negative class definition, read pairs. Euk. – Eukaryota dataset; Met. – Metazoa dataset, Cho. – Chordata dataset, Str. – Stratified dataset, X-150 – first 150 bp of each read in X. Training set in subscript of the model name; test set in column headers. Recall (Rec.) is identical in all cases, as the positive class remains unchanged. Best performance in bold. CNN_{All} achieves best overall accuracy.

	ALL	EUK.	MET.	CHO.	STR.	REC.
CNN_{ALL}	**89.9**	85.9	83.3	78.6	**88.1**	**85.4**
CNN_{CHO}	84.9	84.2	83.6	82.4	84.6	71.7
CNN_{STR}	88.2	**86.4**	**85.1**	**82.7**	87.4	78.8
$CNN_{ALL\text{-}150}$	89.4	85.5	82.9	78.4	87.7	83.2
$CNN_{STR\text{-}150}$	88.2	86.3	84.9	82.5	87.3	78.3
$LSTM_{ALL}$	86.4	78.2	74.1	65.5	82.6	83.0
$LSTM_{CHO}$	82.8	81.9	80.8	80.0	82.4	70.6
$LSTM_{STR}$	85.8	82.1	79.6	75.2	84.2	76.3

Table 2. Classification performance in the fully open-view setting (all virus hosts), read pairs. Acc. – accuracy, Prec. – precision, Rec. – recall. BLAST yields no predictions for over 10% of the samples. Best performance in bold.

	ACC.	PREC.	REC.
ZHANG ET AL. (2019)	57.1	57.8	52.1
BLAST	80.6	**98.4**	79.1
CNN_{ALL} (OURS)	**89.9**	93.9	**85.4**
$LSTM_{ALL}$ (OURS)	86.4	89.0	83.0

3.4. Genome-wide phenotype analysis

We created a GWPA plot for the Taï Forest ebolavirus genome (Fig. 2a). Most genes can be detected by finding peaks of elevated pathogenic potential score predicted by at least one of the models. Intergenic regions are characterized by lower mean scores.

We ran a similar analysis of *S. aureus* using the built-in DeePaC models (Bartoszewicz et al., 2019) and our interpretation workflow. While a viral genome contains usually

Table 3. Classification performance, all hosts. Whole available genomes. Negative class is the majority class. Rec. – recall, Spec. – specificity, BAcc. – balanced accuracy.

	AUPR	REC.	SPEC.	BACC.
BLAST	N/A	85.5	**95.1**	90.3
CNN_{ALL} (OURS)	**91.2**	89.3	94.2	**91.7**
$LSTM_{ALL}$ (OURS)	85.8	**96.2**	76.4	86.3

Table 4. Classification performance on the human blood virome dataset. Positive class is the majority class. Rec. – recall, Spec. – specificity, BAcc. – balanced accuracy.

	AUPR	REC.	SPEC.	BACC.
ZHANG ET AL. (2019)	99.7	80.9	85.4	83.1
$CNN_{ALL\text{-}150}$ (OURS)	**>99.9**	**97.3**	**96.2**	**96.8**
$LSTM_{ALL}$ (OURS)	**>99.9**	88.2	95.5	91.8

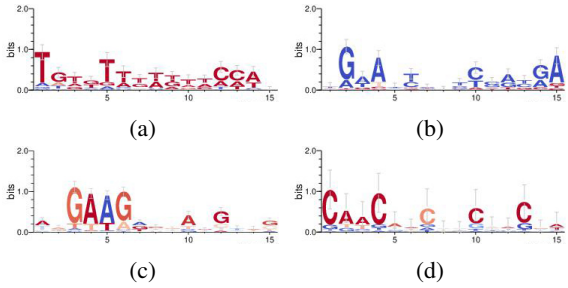


Figure 1. Nucleotide contribution logos of example filters. 1a: Highest mean contribution score (CNN_{All}). 1b: Lowest mean contribution score (CNN_{All}). 1c: Local counter-contributions ($CNN_{Str\text{-}150}$). 1d: Gaps resembling a codon structure, extracted from Bartoszewicz et al. (2019)

Table 5. Gene ranking for *S. aureus* (top 3 out of 870). hupB is indirectly engaged in virulence. Our method detects functionally relevant genes using the model of Bartoszewicz et al. (2019).

RANK	GENE	SCORE	BIOLOGICAL PROCESS
1	SARR	0.644	**VIRULENCE**
2	HUPB	0.642	DNA CONDENSATION
3	SSPB	0.637	**VIRULENCE**

only a handful of genes, by compiling a ranking of 870 annotated genes of the analyzed *S. aureus* strain we could test if the high-ranking regions are indeed associated with pathogenicity. Indeed, out of three top-ranking genes, sarR and sspB are directly engaged in virulence, while hupB regulates expression of virulence-involved genes in many pathogens (Stojkova et al., 2019).

Fig. 2b presents a GWPA plot for the whole genome of the SARS-CoV-2 coronavirus. We highlighted the score peaks aligning with the gene encoding the spike protein, which plays a significant role in host entry (Li, 2016), as well as the E and N genes, which were scored the highest (apart from an unconfirmed ORF10 of just 38aa downstream of N) by the CNN and the LSTM, respectively. Fig. 2c shows the nucleotide-level contributions in a small peak within the receptor-binding domain of the S protein, crucial for recognizing the host cell. The domain location was predicted with CD-search (Marchler-Bauer et al., 2017). While this could suggest a host adaptation, more research is needed.

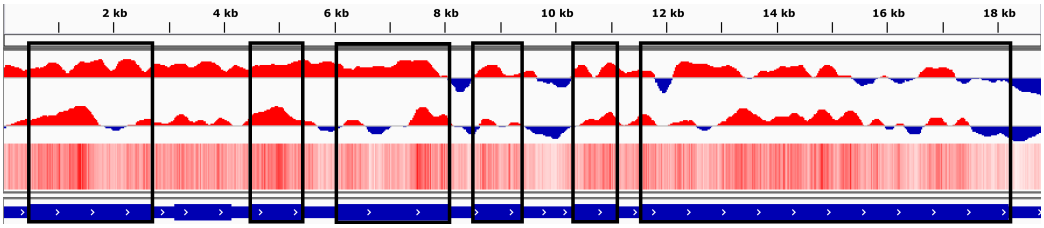
4. Discussion

4.1. Accurate predictions from short DNA reads

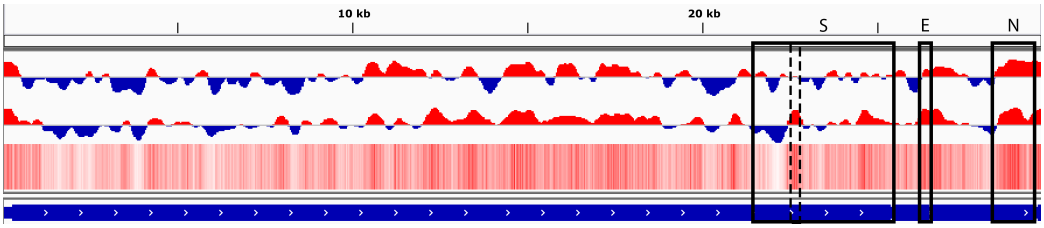
Compared to the previous state-of-the-art in viral host prediction directly from next-generation sequencing reads (Zhang et al., 2019), our models drastically reduce the error rates. This holds also for novel viruses not present in the training set. In the paired read scenario, the previously described method fails, and standard, alignment-based homology testing algorithm cannot find any matches in more than 10% of the cases, resulting in relatively low accuracy. On a real human virome sample, where a main source of negative class reads is most likely contamination (Moustafa et al., 2017), our method filters out non-human viruses with high specificity. In this scenario, the BLAST-derived ground-truth labels were mined using the complete database (as opposed to just a training set). In all cases, our results are only as good as the training data used; high quality labels and sequences are needed to develop trustworthy models. Ideally, sources of error should be investigated with an in-depth analysis of a model's performance on multiple genomes covering a wide selection of taxonomic units. This is especially important as the method assumes no mechanistic link between an input sequence and the phenotype of interest, and the input sequence constitutes only a small fraction of the target genome without a wider biological context. Still, it is possible to predict a label even from those small, local fragments. A similar effect was also observed for image classification with CNNs (Brendel & Bethge, 2019).

4.2. Nucleotide contribution logos

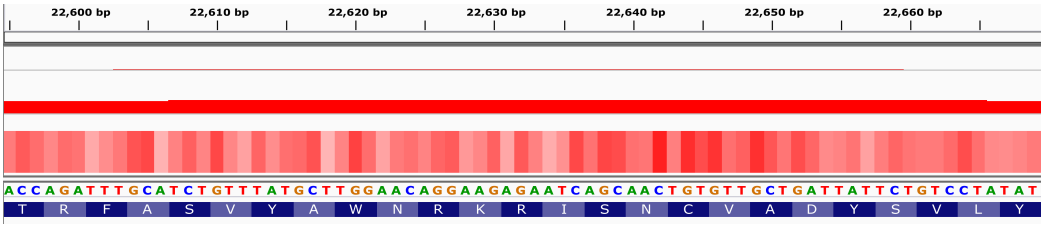
Visualizing convolutional filters may help to identify potential problems. If the input data contains Ns, a ReLU network may learn only negative weights at some positions, resulting in counter-contributions for all possible nucleotides at those positions. In our case, as the filters were weight-normalized, the counter-contributions suggest that the information content and the contribution of a nucleotide are not necessarily correlated. Visualizing learned motifs by aligning the activating sequences (Alipanahi et al., 2015) would not fully



(a) Taï Forest ebolavirus. Genes that can be detected by at least one model are highlighted in black.



(b) Whole genome and sequences encoding the spike protein (S), envelope protein (E) and nucleocapsid protein (N).



(c) Spike protein gene, a small peak (positions 22,595-22,669, dashed line in Fig. 2b) within the receptor-binding domain (predictied by CD-search). Binding to the receptor is crucial for entry to the host cell. Local host adaptation could help switch hosts between the animal reservoir and humans.

Figure 2. Taï Forest ebolavirus and SARS-CoV-2 coronavirus genomes. Top: score predicted by LSTM_{All}. Middle: score predicted by CNN_{All}. Heatmap: nucleotide contributions of CNN_{All}. Bottom, in blue: reference sequence.

describe how the filter reacts to presented data. It seems that the assumption of nucleotide independence – which is crucial for treating DeepLIFT as a method of estimating Shapley values for input nucleotides – is broken. Indeed, k-mer distribution profiles are frequently used features for modelling DNA sequences (as shown also by the dimer-shuffling method of generating reference sequences proposed by Shrikumar et al. (2019a)). However, DeepLIFT's multiple successful applications in genomics indicate that the assumption probably holds approximately. We see information content and DeepLIFT's contribution values as two complementary channels that can be jointly visualized for better interpretability and explainability of CNNs in genomics.

4.3. Genome-scale interpretability

Mapping predictions back to a target genome can be used both as a way of investigating a given model's performance and as a method of genome analysis. GWPA plots of well-annotated genomes highlight the sequences with erroneous and correct phenotype predictions at both genome and gene level, and nucleotide-resolution contribution maps help track those regions down to individual amino-acids. On the other hand, once a trusted model is developed, it can be used on newly emerging pathogens, as the SARS-CoV-2 virus briefly analyzed in this work. The methods presented here may also be applied to other biological problems outside of the field of pathogen genomics. However, experimental work and traditional sequence analysis are required to truly understand the biology behind host adaptation and distinguish true hits from false positives.

4.4. Conclusion

We presented a new approach for predicting a host of a novel virus based on a single DNA read or a read pair, cutting the error rates in half compared to the previous state-of-the-art. For convolutional filters, we jointly visualize nucleotide contributions and information content. Finally, we use GWPA plots to gain insights into the models' behavior and analyze a recently emerged SARS-CoV-2 virus. Data is available at https://doi.org/10.5281/zenodo.3630803 and the code is submitted in Supplementary Material.

Acknowledgements

We gratefully acknowledge Yong-Zhen Zhang and the scientists at the Shanghai Public Health Clinical Center & School of Public Health, Fudan University, who shared the sequence of the 2019-nCoV virus ahead of publication. We thank Melania Nowicka (Max Plank Institute for Molecular Genetics) for inspiring discussions on efficient calculations of partial Shapley values and Lothar H. Wieler (Robert Koch Institute) for useful comments on the first draft of the manuscript. This work was supported by the German Academic Scholarship Foundation (JMB) and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B).

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015. ISSN 1546-1696. doi: 10.1038/nbt.3300.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

Avsec, Ž., Weilert, M., Shrikumar, A., Alexandari, A., Krueger, S., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *bioRxiv*, pp. 737981, August 2019. doi: 10.1101/737981.

Babayan, S. A., Orton, R. J., and Streicker, D. G. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science*, 362(6414):577–580, November 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aap9072.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140.

Bartoszewicz, J. M., Seidel, A., Rentzsch, R., and Renard, B. Y. DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, 36(1):81–89, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz541.

Brendel, W. and Bethge, M. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations*, 2019.

Calistri, A. and Palù, G. Editorial commentary: Unbiased next-generation sequencing and new pathogen discovery: undeniable advantages and still-existing drawbacks. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 60(6):889–891, 2015. ISSN 1537-6591. doi: 10.1093/cid/ciu913.

Calvignac-Spencer, S., Schulze, J. M., Zickmann, F., and Renard, B. Y. Clock rooting further demonstrates that guinea 2014 ebov is a member of the zaïre lineage. *PLoS currents*, 6, 2014.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, June 2004. ISSN 1088-9051. doi: 10.1101/gr.849004.

Edwards, R. A., McNair, K., Faust, K., Raes, J., and Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *FEMS microbiology reviews*, 40(2):258–272, 2016. ISSN 1574-6976. doi: 10.1093/femsre/fuv048.

Eng, C. L., Tong, J. C., and Tan, T. W. Predicting host tropism of influenza a virus proteins using random forest. *BMC Medical Genomics*, 7(3):S1, 2014. ISSN 1755-8794. doi: 10.1186/1755-8794-7-S3-S1.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403, July 2019. ISSN 1471-0064. doi: 10.1038/s41576-019-0122-6.

Gałan, W., Bąk, M., and Jakubowska, M. Host taxon predictor - a tool for predicting taxon of the host of a newly discovered virus. *Scientific Reports*, 9(1):3436, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-39847-2.

Greenside, P., Shimko, T., Fordyce, P., and Kundaje, A. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics*, 34(17):i629–i637, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty575.

Herfst, S., Schrauwen, E. J. A., Linster, M., Chutinimitkul, S., Wit, E. d., Munster, V. J., Sorrell, E. M., Bestebroer, T. M., Burke, D. F., Smith, D. J., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. Airborne Transmission of Influenza A/H5n1 Virus Between Ferrets. *Science*, 336(6088):1534–1541, June 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1213362.

Holtgrewe, M. Mason – a read simulator for second generation sequencing data. *Technical Report FU Berlin*, 2010.

Imai, M., Watanabe, T., Hatta, M., Das, S. C., Ozawa, M., Shinya, K., Zhong, G., Hanson, A., Katsura, H., Watanabe, S., Li, C., Kawakami, E., Yamada, S., Kiso, M., Suzuki, Y., Maher, E. A., Neumann, G., and Kawaoka, Y. Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1n1 virus in ferrets. *Nature*, 486(7403):420–428, June 2012. ISSN 1476-4687. doi: 10.1038/nature10831.

Jha, A., Aicher, J. K., Singh, D., and Barash, Y. Improving interpretability of deep learning models: splicing codes as a case study. *bioRxiv*, 2019. doi: 10.1101/700096.

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26 (7):990–999, 2016. ISSN 1088-9051. doi: 10.1101/gr.200535.115.

Lanchantin, J., Singh, R., Lin, Z., and Qi, Y. Deep Motif: Visualizing Genomic Sequence Classifications. *CoRR*, abs/1605.01133, 2016.

Lanchantin, J., Singh, R., Wang, B., and Qi, Y. Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22:254–265, 2017. ISSN 2335-6936. doi: 10.1142/9789813207813_0025.

Lecuit, M. and Eloit, M. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Frontiers in Cellular and Infection Microbiology*, 4:25, 2014. ISSN 2235-2988. doi: 10.3389/fcimb.2014.00025.

Leendertz, S. A. J., Gogarten, J. F., Düx, A., Calvignac-Spencer, S., and Leendertz, F. H. Assessing the evidence supporting fruit bats as the primary reservoirs for ebola viruses. *EcoHealth*, 13(1):18–25, Mar 2016. ISSN 1612-9210. doi: 10.1007/s10393-015-1053-0.

Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annual Review of Virology*, 3(1):237–261, 2016. doi: 10.1146/annurev-virology-110615-042301.

Li, H. and Sun, F. Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. *Scientific Reports*, 8(1):10032, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-28308-x.

Lipsitch, M. and Inglesby, T. V. Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens. *mBio*, 5(6), December 2014. ISSN 2150-7511. doi: 10.1128/mBio.02366-14.

Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y., and Bryant, S. H. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, 45(D1):D200–D203, 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw1129.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., and Ogata, H. Linking virus genomes with host taxonomy. *Viruses*, 8(3):66, 2016. ISSN 1999-4915. doi: 10.3390/v8030066.

Mock, F., Viehweger, A., Barth, E., and Marz, M. Viral host prediction with deep learning. *bioRxiv*, pp. 575571, 2019. doi: 10.1101/575571.

Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K. E., Venter, J. C., and Telenti, A. The blood DNA virome in 8,000 humans. *PLOS Pathogens*, 13(3):e1006292, March 2017. ISSN 1553-7374. doi: 10.1371/journal.ppat.1006292.

Nair, S., Kim, D. S., Perricone, J., and Kundaje, A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics*, 35(14):i108–i116, July 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz352.

Noyce, R. S., Lederman, S., and Evans, D. H. Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. *PLOS ONE*, 13 (1):e0188453, January 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0188453.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., and DePristo, M. A. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, 2018. ISSN 1546-1696. doi: 10.1038/nbt.4235.

Quang, D. and Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107–e107, 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw226.

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., and Sun, F. Identifying viruses from metagenomic data by deep learning. *arXiv:1806.07810 [q-bio]*, June 2018. arXiv: 1806.07810.

Schneider, T. D. and Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, October 1990. ISSN 0305-1048. doi: 10.1093/nar/18.20.6097.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv:1704.02685 [cs]*, October 2019a. arXiv: 1704.02685.

Shrikumar, A., Tian, K., Shcherbina, A., Avsec, Ž., Banerjee, A., Sharmin, M., Nair, S., and Kundaje, A. TF-MoDISco v0.4.2.2-alpha: Technical Note. *arXiv:1811.00416 [cs, q-bio, stat]*, March 2019b. arXiv: 1811.00416.

Stojkova, P., Spidlova, P., and Stulik, J. Nucleoid-associated protein hu: A lilliputian in gene regulation of bacterial virulence. *Frontiers in Cellular and Infection Microbiology*, 9:159, 2019. ISSN 2235-2988. doi: 10.3389/fcimb.2019.00159.

Sundararajan, M., Taly, A., and Yan, Q. Gradients of Counterfactuals. *CoRR*, abs/1611.02639, 2016.

Tampuu, A., Bzhalava, Z., Dillner, J., and Vicente, R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLOS ONE*, 14 (9):e0222271, September 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0222271.

Thiel, V. Synthetic viruses-Anything new? *PLoS pathogens*, 14(10):e1007019, 2018. ISSN 1553-7374. doi: 10.1371/journal.ppat.1007019.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2):178–192, March 2013. ISSN 1467-5463. doi: 10.1093/bib/bbs017.

Trappe, K., Marschall, T., and Renard, B. Y. Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. *Bioinformatics*, 32(17):i595–i604, September 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw423.

Vouga, M. and Greub, G. Emerging bacterial pathogens: the past and beyond. *Clinical Microbiology and Infection*, 22(1):12–21, January 2016. ISSN 1198-743X. doi: 10.1016/j.cmi.2015.10.010.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Hu, Y., Song, Z.-G., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., and Zhang, Y.-Z. Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. *bioRxiv*, pp. 2020.01.24.919183, January 2020. doi: 10.1101/2020.01.24.919183.

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics*, 32(12):i121–i127, 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw255.

Zhang, Z., Cai, Z., Tan, Z., Lu, C., Jiang, T., Zhang, G., and Peng, Y. Rapid identification of human-infecting viruses. *Transboundary and Emerging Diseases*, August 2019. ISSN 1865-1682. doi: 10.1111/tbed.13314.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3547.