Identifying and prioritizing potential human-infecting viruses from their genome sequences

Nardus Mollentze[1,2]*, Simon A. Babayan[2], Daniel G. Streicker[1,2]

[1]Medical Research Council – University of Glasgow Centre for Virus Research, Glasgow, G61 1QH, United Kingdom.

[2]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, United Kingdom.

*Correspondence to: nardus.mollentze@glasgow.ac.uk

Abstract:

Rapid assessment of which animal viruses may be capable of infecting humans is currently intractable, but would allow their prioritization for further investigation and pandemic preparedness. We developed machine learning algorithms that identify candidate zoonoses using evolutionary signals of host range encoded in viral genomes. This reduces lists of hundreds of viruses with uncertain human infectivity to tractable numbers for prioritized research, generalizes to virus families excluded from model training, can distinguish high risk viruses within families that contain a minority of zoonotic species, and could have identified the exceptional risk of SARS-CoV-2 prior to its emergence. Genome-based risk assessment allows identification of high-risk viruses immediately upon discovery, increasing both the feasibility and likelihood of downstream virological and ecological characterization and allowing for evidence-driven virus surveillance.

Introduction:

Most emerging infectious diseases of humans are caused by viruses that originate from other animal species. Identifying these zoonotic threats prior to emergence is a major challenge since only a small minority of the estimated 1.67 million animal viruses may infect humans (*1*, *2*). Existing models of human infection risk rely on viral phenotypic information that is unknown for newly discovered viruses (e.g., diversity of species a virus can infect) or that vary insufficiently to discriminate risk at the species or strain level (e.g., replication in the cytoplasm), limiting their predictive value (*3*, *4*). Since most viruses are now discovered using untargeted sequencing, often involving many simultaneous discoveries with limited phenotypic data, an ideal approach would quantify the relative risk of human infectivity directly from sequence data alone. By identifying high risk viruses warranting further investigation, such predictions could alleviate the growing imbalance between the rapid pace of virus discovery and lower throughput field and laboratory research needed to comprehensively evaluate risk.

Current models can identify well-characterized human-infecting viruses from genomic sequences (*5*, *6*). However, by training algorithms on very closely related viruses (i.e., strains of the same species) and potentially omitting secondary characteristics of viral genomes linked to infection capability, such models are less likely to find signals of zoonotic status that generalize across viruses. Consequently, predictions may be highly sensitive to substantial biases in current knowledge of viral diversity (*2*, *7*). Overcoming these biases requires discovering and exploiting signals of human infectivity that generalize across unrelated viruses. Empirical and theoretical lines of evidence suggest such signals might exist (*8*, *9*). For example, the depletion of CpG dinucleotides in vertebrate-infecting RNA virus genomes may have arisen to evade zinc-finger antiviral protein (ZAP), an interferon-stimulated gene (ISG) that initiates the degradation of CpG-rich RNA molecules (*10*). While ZAP occurs widely among vertebrates, increasingly recognized lineage-specificity in vertebrate antiviral defences opens the possibility that analogous, undescribed nucleic-acid targeting defences might be human (or primate) specific (*11*). Independently, the frequencies of specific codons in virus genomes can resemble those of their reservoir hosts, possibly owing to increased efficiency and/or accuracy of mRNA translation (*12*). As such, genome compositional similarity to human-adapted viruses or to the human genome may preadapt viruses for human infection (*9*, *13*). We aimed to develop machine learning algorithms which use features engineered from viral and human genome sequences to predict the probability that any animal-infecting virus will infect humans given biologically relevant exposure.

Results and discussion:

We collected a single representative genome sequence from 861 RNA and DNA virus species spanning 36 viral families that contain animal-infecting species (fig. S1). We labelled each virus as being capable of infecting humans or not using published species-specific reports as ground truth, and trained models to classify viruses accordingly. Importantly, given diagnostic limitations and the likelihood that not all viruses capable of human infection have had opportunities to emerge, viruses not reported to infect humans may represent unrealized or undocumented zoonoses or genuinely non-zoonotic species. Identifying these potential zoonoses was an *a priori* goal of our analysis.

We first evaluated whether evolutionary proximity to human-infecting viruses predictably elevates zoonotic risk. Gradient boosted machine (GBM) classifiers trained on virus taxonomy or the frequency of human-infecting viruses among close relatives ("phylogenetic neighbourhood" (*14*)) outperformed chance (median area under the receiver-operating characteristic curve [AUC_m] = 0.604 and 0.558, respectively), but were no better than simply ranking novel viruses by the proportion of human-infecting viruses in each family ("taxonomy-based heuristic", AUC_m = 0.596, fig. 1A), indicating the inability of these relatedness-based models to distinguish risk at scales below the viral family level.

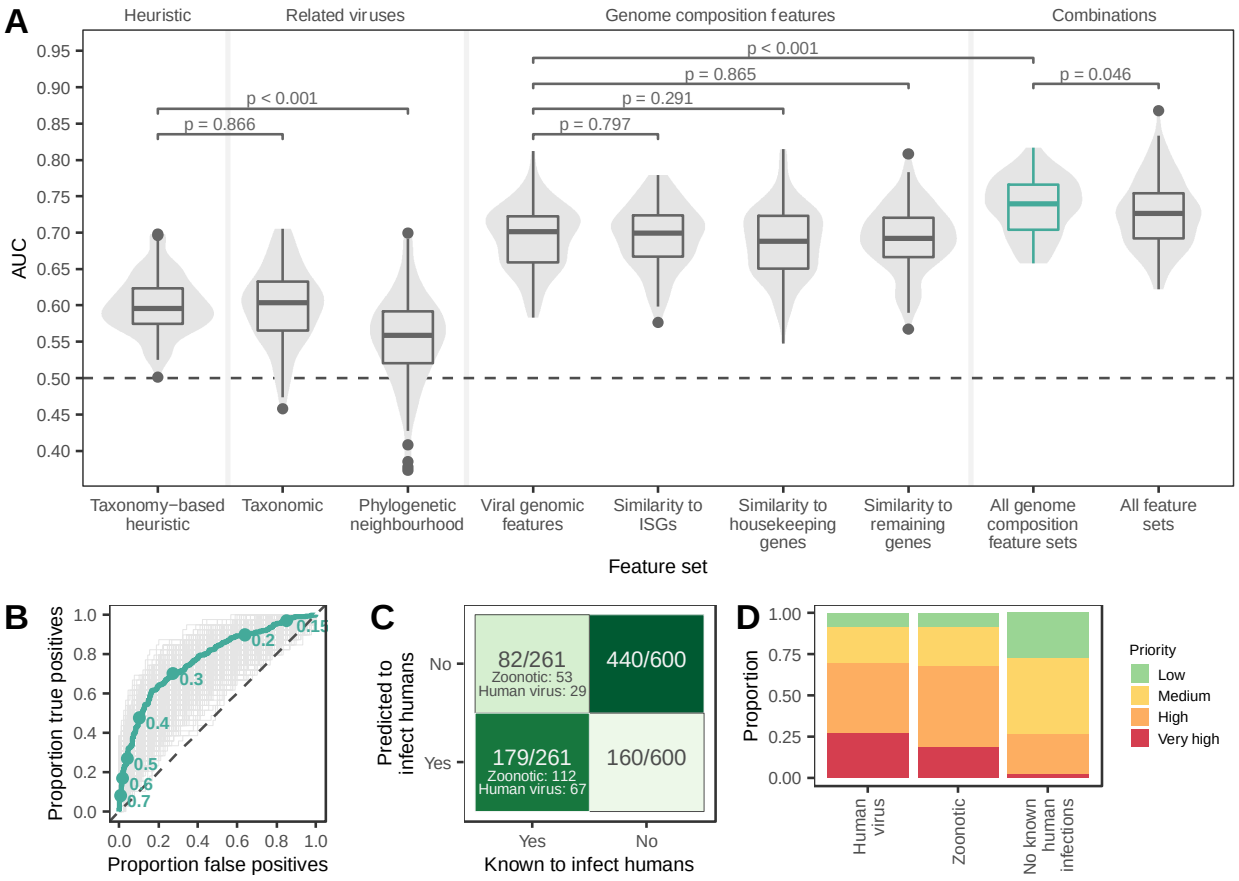

Fig. 1. Machine learning prediction of human infectivity from viral genomes. (A) Violins and boxplots show the distribution of AUC scores across 100 replicate test sets. (B) Receiver-operating characteristic curves showing the performance of 1000 models trained on all genome composition feature sets (grey) and performance of the bagged classifier derived from the top 10% of them (green). Points indicate discrete probability cut-offs for categorizing viruses as human-infecting. C and D show binary predictions of the bagged model, using the cut-off which balanced sensitivity and specificity (0.303). (C) Heatmap showing the proportion of predicted viruses in each category. (D) Proportion of viruses from the training data classified into priority categories (low: entire 95% confidence interval [CI] of predicted probability ≤ cut-off; medium: mean prediction ≤ cut-off, but CI crosses it; high: mean prediction > cut-off, but CI crosses it; very high: entire CI > cut-off.)

We next quantified the performance of GBMs trained on genome composition (i.e., codon usage biases, amino acid biases and dinucleotide biases), calculated either directly from viral genomes ("viral genomic features") or based on similarity to three alternative sets of human gene transcripts ("human similarity features"): interferon-stimulated genes (ISGs), housekeeping genes, and all other genes. We hypothesized that viruses might optimally resemble ISGs since both tend to be expressed concomitantly in virus-infected cells. However, we also included non-

ISG housekeeping genes and remaining genes to explore whether signals were specific to ISGs. Genome composition feature sets performed similarly when tested independently (AUC_m = 0.688–0.701) and consistently outperformed GBMs based on relatedness alone. Combining all genome composition features further improved, and reduced variance in, performance (AUC_m = 0.740, fig. 1A), suggesting that measures of similarity to human transcripts contained information unavailable from viral genomic features alone. In contrast, adding relatedness features to this combined model reduced accuracy (AUC_m = 0.726) and increased variance (fig. 1A). Averaging output probabilities over the best 100 out of 1000 models trained on random test/train splits (a process akin to bagging) further improved the combined genome feature-based classifier (AUC = 0.773, fig. 1B), correctly identifying 69.8% of viruses which predominately or exclusively infect humans and 67.9% of zoonotic viruses as human infecting, though performance varied among viral families (fig. 1C, fig. S2). When model outputs were translated into priority categories, 92% of known human-infecting viruses were classified as medium (23%), high (46.7%) or very high (21.8%) risk, while only 8% (N=22) were considered low risk (figs. 1D & S3 – S4, table S1). Fourteen viruses not currently known to infect humans were classified as very high priority (fig. S4), although three (*Aura virus*, *Ndumu virus* and *Uganda S virus*) have serological evidence of human infection (*4*, *15*), suggesting they may be valid zoonoses rather than model misclassifications.

Since genome composition features partly track viral evolutionary history (*14*), it is conceivable that our models made predictions by re-constructing taxonomy. We therefore compared dendrograms that clustered viruses by either taxonomy, raw genomic features, or the per-virus effect sizes of features (measured as SHAP values (*16*)). While dendrograms using raw feature values closely correlated with virus taxonomy for both human-infecting and other viruses (Baker's (*17*) $\gamma = 0.617$ and 0.492 respectively, $p < 0.001$), dendrograms of SHAP similarity had reduced taxonomic correlations ($\gamma = 0.060$ and 0.238 ; $p \leq 0.008$; fig. S5). Among human-infecting viruses, correlations between effect size-based clustering and taxonomy weakened at deeper taxonomic levels, even though the input genome features provided sufficient information to partially reconstruct virus taxonomy at the realm, kingdom and phylum levels (fig. S6). These results indicate that more taxonomic information was available than was utilised by the trained models to predict human infection ability. Alongside earlier observations on classifier performance (fig. 1A), this suggests that genome composition-based models refine taxonomic information by finding common patterns that predict human infection across diverse viruses. Consistent with this interpretation, SHAP value clustering grouped viruses with different genome types (fig. 2A). These findings point to the existence of common, but uncharacterized, viral genome features which favour human infection.

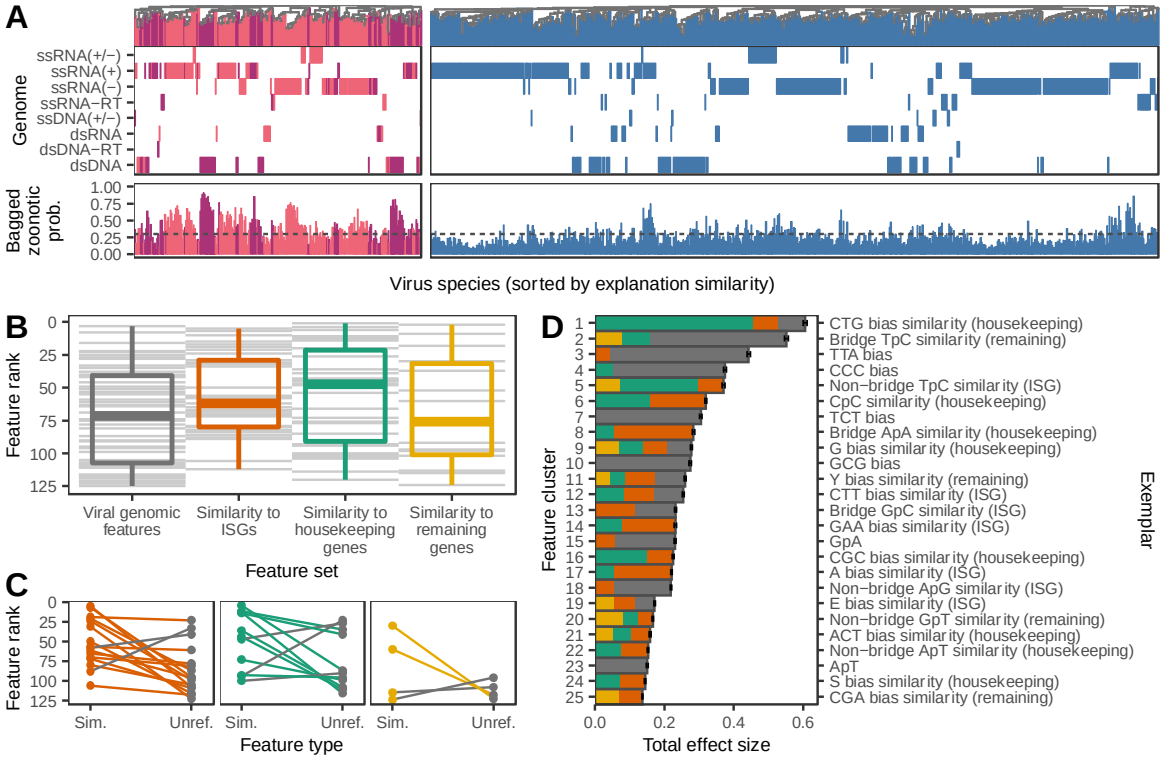

Fig. 2. Genomic determinants of human-infecting viruses. (A) SHAP value clustering of viruses known to infect humans (primarily human-associated, dark purple, and zoonotic, light purple) and those with no known history of human infection (blue) shows similar features predicted human infection across viruses with different genome types (rows). (B) Relative importance of individual features, when ranked by their mean effect size (absolute SHAP value) across all viruses. Grey lines represent individual features; boxplots show the median, 25th/75th percentiles and range of ranks for each feature set. (C) Difference in ranks of features when both unreferenced ("Unref.") and similarity to human genomes ("Sim.") forms were retained in the final set of models. Lines are coloured according to the highest ranked representation in each pairwise comparison; colours as in B. (D) Total effect size of features in the top 25 discrete clusters of correlated variables identified by affinity propagation clustering. Bars represent means (± SEM) across 1000 iterations, and are shaded by the proportion of the cluster from each feature set; colours as in B.

Although our analysis was not designed to conclusively identify biological mechanisms underlying genomic predictors of human infection, it provides a set of associations and measures of relative importance for future mechanistic research. Representatives of all genomic feature categories were retained in final models, though we found some evidence that compositional similarity to housekeeping genes and ISGs influenced predictions more strongly than unreferenced viral genomic features (fig. 2B–C; supplementary text). As expected, no single feature stood out as the driving force and many features formed correlated clusters (figs. 2D &

S7). Interestingly, many features had complex, non-linear relationships with human infection (fig. S8) such that increased similarity to human genomes did not always increase the likelihood of infecting humans. These results suggest that a complex suite of mechanisms beyond matching host genome composition are likely to underpin links between viral genomes and human infectivity.

Finally, we carried out two case studies to illustrate the utility of our prediction framework. First, we used the bagged classifier to rank 256 viruses added in the most recent ICTV taxonomy release (#35, 24 April 2020; 38 families) which had been excluded from model training. Forty-three were predicted to be unlikely to infect humans (fig. 3A). Of 19 viruses in the very high priority category, 14 were anelloviruses, 13 of which were discovered in humans (fig. 3B & table S1). Strikingly, no viruses from this family were present in the training data. The very high priority category further included two rhabdoviruses – *Menghai almendravirus* and *Taiwan bat lyssavirus*. The *Almendravirus* genus was described in 2016, and no members have been detected in humans thus far (*18*), while all lyssaviruses are assumed to be zoonotic (*19*). The remaining viruses classified as high priority were from families not currently considered zoonotic – *Birnaviridae* (*Rotifer birnavirus*, associated with zooplankton and therefore likely a false positive), *Parvoviridae* (*Rodent chaphamaparvovirus 1*), and Adenoviridae (*Bat mastadenovirus I*). Second, we predicted the priority level of all currently recognized coronavirus species, along with 55 human and animal-derived sequences of *Severe acute respiratory syndrome-related coronavirus* (SARS coronavirus). All known human-infecting coronaviruses were classified as either medium or high priority (fig. 3C). Most SARS coronavirus isolates (92.7%) were classified as medium-priority, including the causal agent of the 2002-2003 outbreak (fig. S9). Interestingly however, SARS-CoV-2 was ranked considerably higher and classified as high priority (figs. 3C & S9). Importantly, these predictions used models that excluded the SARS coronavirus genome from training. Taken together, these results demonstrate the ability of our models to generate disparate predictions within viral families and species and suggest that the elevated risk of SARS-CoV-2 could have been anticipated via sequencing-based surveillance prior to its zoonotic emergence (fig. 3C). Our model further identified two non-zoonotic coronaviruses – *Alphacoronavirus 1* and the recently described *Sorex araneus coronavirus T14* – as being at least as, or more likely to be capable of infecting humans than SARS-CoV-2; these should be considered high priority for further research.

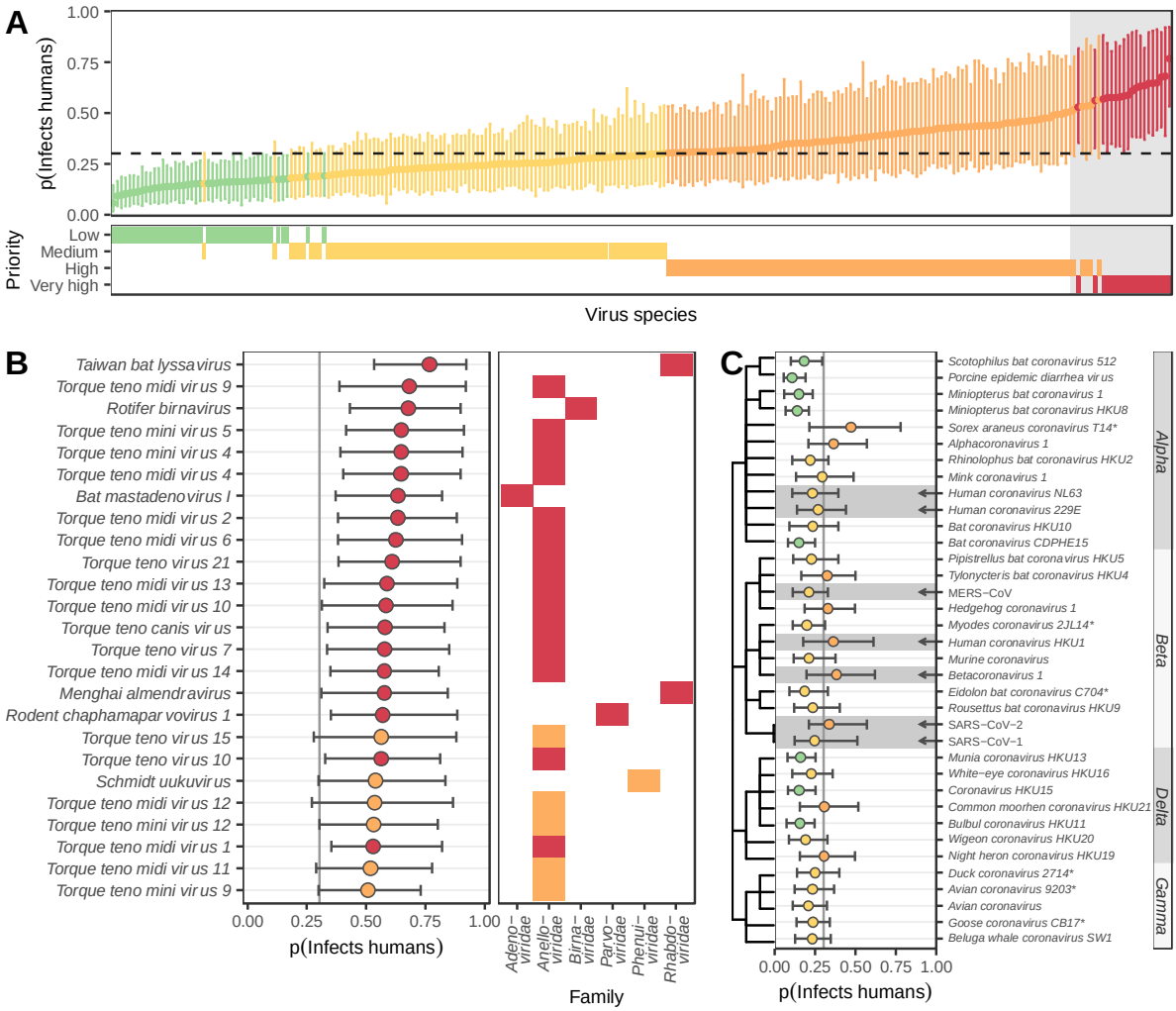

Fig. 3. Relative risk of human infection predicted from viral genomes. (A) Predicted probability of human infection (mean and 95% range across the best-performing 10% of models) for 265 newly recognized virus species. A dashed line shows the cut-off which balances sensitivity and specificity in the training data. The top 25 viruses (grey box) are illustrated in more detail in (B). (C) Predictions for currently recognized *Coronaviridae* species and for 2 strains of SARS coronavirus: SARS-CoV-1 (isolate HSZ-Cc, sampled early in the 2002-2004 pandemic) and SARS-CoV-2 (isolate Wuhan-Hu-1, sampled early in the current pandemic). A dendrogram illustrates taxonomic relationships, with abbreviated genus names annotated on the right. Arrows highlight known human-infecting species. Asterisks indicate newly recognized species, also present in A. MERS-CoV = *Middle East respiratory syndrome-related coronavirus.*

In an age of rapid, genomic-based virus discovery, rational prioritization of research and surveillance activities has been an unresolved challenge. Our findings show that routine proxies of zoonotic risk including virus taxonomy and relative phylogenetic proximity to human-infecting species (*4*, *7*, *20*) have limited discriminatory power. In contrast, models that exploited

features of viral genomes that were least partly independent of taxonomy both generalized predictions across divergent viruses and provided capacity to discriminate risk among closely related virus species or strains; characteristics which cannot be achieved by existing genomic or phenotypic models of zoonotic risk. The performance of our models, while imperfect, means that many potential zoonoses can be identified immediately after virus discovery and genome sequencing. Large-scale application of these models enables retrospective ranking-based prioritization of hundreds of recognized viruses as well as prospective ranking in parallel with virus discovery, spanning all RNA and DNA genome types (table S1). Importantly, our models predict baseline zoonotic potential, which ultimately will be modulated by ecological opportunities for emergence. Further, the societal impact of emergence will depend on capacity for human to human spread and the severity of human disease, which likely require additional non-genomic data to anticipate. Nonetheless, for both novel and recognized viruses, substantially reduced lists of candidate zoonoses heightens the feasibility of further ecological and virological characterisation.

Materials and Methods

Data

Although our primary interest was in zoonotic transmission, we trained models to predict the ability to infect humans in general, reasoning that patterns found in viruses predominately maintained by human-to-human transmission may contain genomic signals which also apply to zoonotic viruses. Data on the ability to infect humans were obtained by merging the data of (*4*) and (*15*), which contain species-level records of reported human infections. These datasets were supplemented by new literature searches for 15 species (*7*). In all cases, only viruses detected in humans by either PCR or sequencing were considered to have proven ability to infect humans. All viruses for which no such reports were found were considered to not infect humans, although we emphasise that many of these viruses are poorly characterised and could therefore be unrecognized or unreported zoonoses. We therefore expect our models to further improve as these and new viruses become better characterized. Our final dataset comprised 861 virus species from 36 families. A representative genome was selected for each virus species, giving preference to sequences from the RefSeq database wherever possible. RefSeq sequences that had annotation issues, represented extensively passaged isolates, or were otherwise not judged to be representative of the naturally circulating virus were replaced with alternate genomes. By using single exemplar genomes from each virus species, we maximised the likelihood of discovering generalizable signatures of human infection while avoiding performance measures which would be over-optimistic for novel viruses. A potential drawback of this approach was that we omitted substantial viral diversity which is not yet formally recognized by the ICTV (*2*). However, we contend that including currently unrecognized viruses is unlikely to improve the predictions of our models because: (a) most will be non-human infecting (an already over-represented class) and hence provide little additional information, (b) those which do infect humans will not generally be known to do so due to a lack of historic testing, adding misleading signals, and (c) the predictive features identified often span across families, reducing the impacts of taxonomic gaps.

Features

We compared the predictive power of all classifiers to that which could be obtained through knowledge of a virus' taxonomic position alone. This captures the intuitive expectation that viruses can be risk-assessed based on knowledge of the zoonotic abilities of their closest known relatives. To formalise this idea, we first created a simple heuristic which ranks viruses based on the proportion of other viruses in the same family which are known to infect humans. Viral family was chosen as the level of comparison because not all viruses are classified in a scheme which includes subfamilies, while lower taxonomic levels suffer from limited sample size. To further characterise the predictive power of virus taxonomy, we also included potential predictor variables (here termed *features*) describing virus taxonomy when training classifiers. These included the proportion of human infecting viruses in each family (calculated from species in the training data only) along with categorical features describing the phylum, subphylum, class, subclass, order, suborder, and family to which each species was assigned. This information was taken from version 2018b of the ICTV master species list (https://talk.ictvonline.org/files/master-species-lists/). To capture taxonomic effects at finer resolution, we summarised the human infection ability of the closest relatives of each virus in the training data (following (*14*), here termed the 'phylogenetic neighbourhood'). To calculate these features, the genome (or genome segments, where applicable) of each virus species was nucleotide BLASTed against a database containing genomes or genome segments for all species in the training data. All BLAST matches with e-value ≤ 0.001 were retained and used to calculate the proportion of human-infecting viruses in the phylogenetic neighbourhood of each virus (excluding the current species). We also calculated a 'distance-corrected' version of this proportion by re-weighting matches according to the proportion of nucleotides p matching the genome of the focal virus:

$$C = \frac{\sum_{i=1}^{N} p_i \cdot x_i}{N},$$

where N is the number of retained blast matches for the focal virus species, and $x_i = 1$ if matching species i is able to infect humans and equals 0 otherwise. Both the raw and distance-corrected proportion were used in unison to define the phylogenetic neighbourhood, allowing classifiers to pick the most informative representation or to combine both pieces of information if needed. In cases where a virus received no matches with e-value ≤ 0.001, both proportions were set to NA to reflect the fact that no information about the phylogenetic neighbourhood was available. This occurred for an average of 2% of viruses in each random training and test set (see below).

Various features summarising the compositional biases in each virus genome were calculated as described in (*14*). These included codon usage biases, amino acid biases, dinucleotide biases across the entire genome, dinucleotide biases across coding regions only, dinucleotide biases spanning the bridges between codons (i.e. across base 3 of the preceding codon and base 1 of the current codon), and dinucleotide biases at non-bridge positions. Similarity features to human RNA transcripts were calculated by first calculating the above genome features for human genes. For each gene, the sequence of the canonical transcript was obtained from version 96 of Ensembl (*21*). Genes were divided into three mutually exclusive sets, encompassing interferon stimulated genes (ISGs, taken from (*11*); N = 2054), non-ISG housekeeping genes ((*22*); N=3172), and remaining genes (N=9565). The distribution of observed values for each genome feature was summarised across all genes in a set by calculating an empirical probability density function using version 2.3.1 of the EnvStats library in R version

3.5.1 (*23*). The final similarity score for each genome feature of each virus was then calculated by evaluating this density function at the value observed in the virus genome, giving the probability of observing this value among the transcripts of the set of human genes in question (fig. S10).

Training

Gradient boosted classification trees were trained using the xgboost and caret libraries (versions 0.90 and 6.0-85, respectively) in R (*24*, *25*). Note that while we separate primarily human-transmitted viruses and zoonotic viruses in some figures, these viruses were considered a single class during training. Thus, models were trained to distinguish viruses known to infect humans from those with no reports of human infection.

A range of models were trained using different combinations of the individual feature sets described above. To reduce run-times and potential overfitting in subsequent steps, features were subjected to a pre-screening step to remove those with little or no predictive value. During this pre-screening step, classifiers were trained on a random selection of 70% of the data using all features within the respective feature group (e.g., all ISG similarity features). Training sets were class-stratified to retain the observed frequencies of positive (known human-infecting) and negative virus species. All hyperparameters were kept at their default values, except for the number of training rounds, which was fixed at 150. The importance of each feature was summarised across 100 models trained on different samples of the full dataset, and the N most predictive features were retained. A range of possible values for N was evaluated by combining all feature sets and using the selected features to optimize and train a final set of models as described below. The final value of $N = 125$ was chosen as the point at which additional features provided no further improvement in performance, measured as the area under the receiver-operating characteristic curve (AUC; fig. S11). Here, AUC measures the probability that a randomly chosen human-infecting virus would be ranked higher than a randomly chosen virus which has not been reported to infect humans. When a given feature set or combination of feature sets comprised < 125 features, all features were retained. This was the case for models trained using only taxonomy (7 features) or phylogenetic neighbourhoods (2 features).

Final models were trained using reduced feature sets. To assess the variability in accuracy across different training sets, training was repeated 100 times (*14*). In each iteration, training was performed on a random, class-stratified selection of 70% of the available data (here, the *training set*). Output probabilities were calibrated using half the remaining data (*calibration set*, again selected randomly and stratified by class), leaving 15% of the full dataset for evaluation of model predictions (*test set*). In each iteration, hyperparameters were selected using 5-fold cross-validation on the training set, searching across a random grid of 500 hyperparameter combinations. This cross validation was adaptive, evaluating each parameter combination on a minimum of 3 folds before continuing cross validation only with the most promising candidates (*25*). The parameter combination maximising AUC across folds was selected and used to train a final classifier on the entire training set. This classifier was then used to produce quantitative scores for each species in the calibration and test sets.

Next, outputs were calibrated to allow interpretation as probabilities using the beta calibration method of (*26*). A calibration model was fit to scores obtained for the calibration set using version 0.1.0 of the betacal R package. The fitted calibration model was used to produce final output probabilities for virus species in the test set. Finally, to summarise predicted probabilities from models trained on different training sets, we averaged the calibrated

probabilities across the best-performing 10% of iterations (measured in terms of AUC to a holdout set described below) in which a given virus occurred in the test set (a process with similarities to bagging (*14*); 1000 iterations performed). When predicting the probability of human infectivity for novel viruses, this involved the best 10% of models overall, while bagging of known viruses (i.e. species which occur in the training data) relied on the best 10% of models in which each virus occurred in the test set (and hence was not used to train or calibrate the model). For such 'known' virus species, model performance was re-calculated while excluding the focal species from the test set, to prevent accurate prediction of the focal virus from influencing the choice of models used for bagging.

Calculating phylogenetic neighbourhood features, which depend on the specific training set selected, represented a bottleneck in the iterative training strategy described above. To overcome this, a single all-against-all blast search was performed, with search results and e-values subsequently corrected in each iteration to emulate the result that would have been obtained when blasting only against the current training dataset. Specifically, e-values were re-calculated as described in (*27*):

$$E = \frac{mn}{2^{S'}}$$

where m is the length of the query sequence (in nucleotides), n is the total number of nucleotides in the training set (i.e., the size of the database searched), and S' is bitscore for this particular alignment in the original blast search.

<u>Feature importance and clustering</u>

To assess the variability in feature importance while accounting for all viruses, feature importance was assessed across all 1000 models trained for bagging above. In each iteration, the influence of features was assessed using SHAP values, an approximation of Shapley values which describe the change in the predicted log odds of infecting humans attributable to each genome feature (*16*). The overall importance of each feature was calculated as the mean of absolute SHAP values across all viruses in the training set of a given iteration (*28*).

Because features tended to be highly correlated, we also report importance values for clusters of correlated features, with the importance of each cluster for individual viruses calculated as the sum of absolute SHAP values across all features in a cluster:

$$I_{c,j,i} = \sum_{f=1}^{N_c} \left| S_{f,i,j} \right|,$$

where $I_{c,j,i}$ is the importance of feature cluster c in determining the output score of virus j in iteration i, N_c is the number of features in this cluster, and $S_{f,i,j}$ is the SHAP value for feature f. The overall importance of each feature cluster was then calculated as the mean of these importance values across all viruses in the training data of a given iteration:

$$I_{c,i} = \frac{\sum_{j=1}^{N_i} \left| I_{c,j,i} \right|}{N_i}.$$

Feature clusters were obtained by affinity propagation clustering, which seeks to identify discrete clusters of features centred around a representative feature (the *exemplar* feature) (*29*). Features were clustered using pairwise Spearman correlations as the similarity measure, using version 1.4.8 of the apcluster library in R (*30*).

To further explore patterns in feature importance across virus species, we followed a strategy similar to (*28*), clustering viruses based on the average SHAP values assigned to individual features for each virus across all iterations:

$$\bar{S}_{f,j} = \frac{\sum S_{f,i,j}}{N_i}.$$

These values were used to calculate the pairwise Euclidean distances between all virus species using version 2.1.0 of the cluster library in R (*31*). Viruses were then clustered using agglomerative hierarchical clustering, calculating distances between clusters as the mean distance between all points in the respective clusters (i.e. UPGMA clustering). To explore patterns common to viruses from each class, clustering was performed separately for known human infecting and other viruses.

To compare this explanation-based clustering with virus taxonomy, we also constructed a dendrogram based on taxonomic assignments as recorded in version 2018b of the ICTV master species list, using all taxonomic levels from phylum to subgenus. Since some levels of the ICTV taxonomy are not used consistently across all viruses, missing taxonomic levels were interpolated to ensure accurate representation of the underlying taxonomy. For example, for viruses which are not classified in a scheme which includes subfamilies, the next level downstream – genus – was repeated, thereby treating each genus as belonging to a distinct subfamily. Categorical taxonomic assignments were used to calculate pairwise Gower distances between virus species (*32*), before performing agglomerative hierarchical clustering as described above. We also assessed the ability of underlying genome feature values to reconstruct virus taxonomy by performing hierarchical clustering on a Euclidean distance matrix calculated directly from all genomic features (i.e. unreferenced genome, ISG similarity, housekeeping gene similarity and remaining gene similarity feature sets). The similarity between dendrograms was assessed using the gamma correlation index of (*17*), as implemented in version 1.12.0 of the dendextend library in R (*33*). A null distribution for this statistic was calculated by randomly shuffling the labels (i.e., virus species names) of both dendrograms 1000 times. To assess the taxonomic depth at which dendrograms were concordant, the Fowlkes-Mallows index was calculated at each possible cut-point in the dendrograms being compared (*34*), again using the dendextend library. As before, a null distribution was generated by randomly shuffling the labels of both dendrograms 1000 times.

Ranking novel viruses

To illustrate the use of our models in practice, the best performing set of models (i.e. those trained using the best 125 features selected from among all genome feature-based feature sets) was used to predict the probability that novel viruses are able to infect humans. Novel viruses were operationally defined as all species newly recognized in the latest version of the ICTV taxonomy (release #35, 2019; https://talk.ictvonline.org/taxonomy) from families known to contain species that infect animals. For each novel species, the genome sequence referenced in the ICTV virus metadata resource corresponding to this version of the taxonomy (https://talk.ictvonline.org/taxonomy/vmr) was retrieved and used to calculate genome features as described above. Scoring was performed separately using each of the top 10% out of 1000 best-performing classifiers, as described above, and averaged to obtain the final output probability. Confidence intervals for these mean probabilities were calculated as the 2.5% and 97.5% quantiles of probabilities output by the top 10% of classifiers.

References:

1. E. C. Holmes, A. Rambaut, K. G. Andersen, Pandemics: spend on surveillance, not prediction. *Nature*. **558**, 180–182 (2018).

2. M. Wille, J. L. Geoghegan, E. C. Holmes, *bioRxiv*, in press, doi:10.1101/2020.08.17.254961.

3. J. R. C. Pulliam, J. Dushoff, Ability to replicate in the cytoplasm predicts zoonotic transmission of livestock viruses. *The Journal of Infectious Diseases*. **199**, 565–568 (2009).

4. K. J. Olival, P. R. Hosseini, C. Zambrana-Torrelio, N. Ross, T. L. Bogich, P. Daszak, Host and viral traits predict zoonotic spillover from mammals. *Nature*. **546**, 646–650 (2017).

5. Z. Zhang, Z. Cai, Z. Tan, C. Lu, T. Jiang, G. Zhang, Y. Peng, Rapid identification of human-infecting viruses. *Transboundary and Emerging Diseases*. **66**, 2517–2522 (2019).

6. J. M. Bartoszewicz, A. Seidel, B. Y. Renard, *bioRxiv*, in press, doi:10.1101/2020.01.29.925354.

7. N. Mollentze, D. G. Streicker, Viral zoonotic risk is homogenous among taxonomic orders of mammalian and avian reservoir hosts. *Proceedings of the National Academy of Sciences*. **117**, 9423–9430 (2020).

8. S. Rothenburg, G. Brennan, Species-Specific Host–Virus Interactions: Implications for Viral Host Range and Virulence. *Trends in Microbiology*. **28**, 46–56 (2020).

9. B. D. Greenbaum, A. J. Levine, G. Bhanot, R. Rabadan, Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses. *PLOS Pathogens*. **4**, e1000079 (2008).

10. M. A. Takata, D. Gonçalves-Carneiro, T. M. Zang, S. J. Soll, A. York, D. Blanco-Melo, P. D. Bieniasz, CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature*. **550**, 124–127 (2017).

11. A. E. Shaw, J. Hughes, Q. Gu, A. Behdenna, J. B. Singer, T. Dennis, R. J. Orton, M. Varela, R. J. Gifford, S. J. Wilson, M. Palmarini, Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biology*. **15**, 1–23 (2017).

12. G. M. Jenkins, E. C. Holmes, The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research*. **92**, 1–7 (2003).

13. L. A. Shackelton, C. R. Parrish, E. C. Holmes, Evolutionary Basis of Codon Usage and Nucleotide Composition Bias in Vertebrate DNA Viruses. *J Mol Evol*. **62**, 551–563 (2006).

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

15. M. E. J. Woolhouse, L. Brierley, Epidemiological characteristics of human-infective RNA viruses. *Scientific Data*. **5**, 180017 (2018).

16. S. M. Lundberg, S.-I. Lee, in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (2017), pp. 4765–4774.

17. F. B. Baker, Stability of Two Hierarchical Grouping Techniques Case 1: Sensitivity to Data Errors. *Journal of the American Statistical Association*. **69**, 440–445 (1974).

18. M. A. Contreras, G. Eastwood, H. Guzman, V. Popov, C. Savit, S. Uribe, L. D. Kramer, T. G. Wood, S. G. Widen, D. Fish, R. B. Tesh, N. Vasilakis, P. J. Walker, Almendravirus: A proposed new genus of Rhabdoviruses isolated from mosquitoes in tropical regions of the Americas. *American Journal of Tropical Medicine and Hygiene*. **96**, 100–109 (2017).

19. C. R. Fisher, D. G. Streicker, M. J. Schnell, The spread and evolution of rabies virus: Conquering new frontiers. *Nature Reviews Microbiology*. **16**, 241–255 (2018).

20. A. D. Washburne, D. E. Crowley, D. J. Becker, K. J. Olival, M. Taylor, V. J. Munster, R. K. Plowright, Taxonomic patterns in the zoonotic potential of mammalian viruses. *PeerJ*. **6**, e5979 (2018).

21. F. Cunningham, P. Achuthan, W. Akanni, J. Allen, M. R. Amode, I. M. Armean, R. Bennett, J. Bhai, K. Billis, S. Boddu, C. Cummins, C. Davidson, K. J. Dodiya, A. Gall, C. G. Girón, L. Gil, T. Grego, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, M. Kay, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, J. C. Marugán, T. Maurel, A. C. McMahon, B. Moore, J. Morales, J. M. Mudge, M. Nuhn, D. Ogeh, A. Parker, A. Parton, M. Patricio, A. I. Abdul Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, H. Sparrow, E. Stapleton, M. Szuba, K. Taylor, G. Threadgold, A. Thormann, A. Vullo, B. Walts, A. Winterbottom, A. Zadissa, M. Chakiachvili, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, A. D. Yates, D. R. Zerbino, P. Flicek, Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2019).

22. E. Eisenberg, E. Y. Levanon, Human housekeeping genes , revisited. *Trends in Genetics*. **29**, 569–574 (2013).

23. S. P. Millard, *EnvStats: an R package for environmental statistics* (Springer, New York, 2013).

24. T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, San Francisco, California, USA, 2016; https://doi.org/10.1145/2939672.2939785), *KDD '16*, pp. 785–794.

25. M. Kuhn, Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*. **28** (2008), doi:10.18637/jss.v028.i05.

26. M. Kull, T. M. S. Filho, P. Flach, Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electron. J. Statist.* **11**, 5052–5080 (2017).

27. S. Altschul, The Statistics of Sequence Similarity Scores. *BLAST documentation*, (available at https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html).

28. S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. **2**, 56–67 (2020).

29. B. J. Frey, D. Dueck, Clustering by Passing Messages Between Data Points. *Science*. **315**, 972–976 (2007).

30. U. Bodenhofer, A. Kothmeier, S. Hochreiter, APCluster: an R package for affinity propagation clustering. *Bioinformatics*. **27**, 2463–2464 (2011).

31. M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, "cluster: Cluster analysis basics and extensions" (manual, 2019).

32. J. C. Gower, A General Coefficient of Similarity and Some of Its Properties. *Biometrics*. **27**, 857 (1971).

33. T. Galili, dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*. **31**, 3718–3720 (2015).

34. E. B. Fowlkes, C. L. Mallows, A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*. **78**, 553 (1983).

Acknowledgments: We thank Laura Bergner, Richard Orton, Andrew Shaw, Sam Wilson and David Robertson for helpful discussions and suggestions. **Funding:** D.G.S. and N.M. were supported by a Wellcome Senior Research Fellowship (217221/Z/19/Z). Additional funding was provided by the Medical Research Council through program grants MC_UU_12014/8 and

MC_UU_12014/12. **Author contributions:** Conceptualization: D.G.S., S.A.B.; Data curation: D.G.S, N.M.; Formal analysis: N.M.; Visualization: N.M.; Writing – original draft: N.M.; Writing – review & editing: D.G.S., S.A.B., N.M. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Data and code related to this paper are available at: https://github.com/nardus/zoonotic_rank