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.

*Corresponding author
Email: nardus.mollentze@glasgow.ac.uk (NM)

Short title:
Candidate zoonoses identified from viral genomes

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://doi.org/10.5281/zenodo.4271479

Abstract

Determining which animal viruses may be capable of infecting humans is currently intractable at the time of their discovery, precluding prioritization of high-risk viruses for early investigation and outbreak preparedness. Given the increasing use of genomics in virus discovery and the otherwise sparse knowledge of the biology of newly-discovered viruses, we developed machine learning models that identify candidate zoonoses solely using signatures of host range encoded in viral genomes. Within a dataset of 861 viral species with known zoonotic status, our approach outperformed models based on the phylogenetic relatedness of viruses to known human-infecting viruses (AUC = 0.773), distinguishing high-risk viruses within families that contain a minority of human-infecting species and identifying putatively undetected or so far unrealized zoonoses. Analyses of the underpinnings of model predictions suggested the existence of generalisable features of viral genomes that are independent of virus taxonomic relationships and that may preadapt viruses to infect humans. Our model reduced a second set of 645 animal-associated viruses that were excluded from training to 272 high and 41 very high-risk candidate zoonoses and showed significantly elevated predicted zoonotic risk in viruses from non-human primates, but not other mammalian or avian host groups. A second application showed that our models could have identified SARS-CoV-2 as a relatively high-risk coronavirus strain and that this prediction required no prior knowledge of zoonotic SARS-related coronaviruses. Genome-based zoonotic risk assessment provides a rapid, low-cost approach to enable evidence-driven virus surveillance and increases the feasibility of downstream biological and ecological characterisation of viruses.

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 [1–3]. Existing models of human infection risk rely on viral phenotypic information that is unknown for newly discovered viruses (e.g., the diversity of species a virus can infect) or that vary insufficiently to discriminate risk at the virus species or strain level (e.g., replication in the cytoplasm), limiting their predictive value before the virus in question has been characterised [4–6]. Since most viruses are now discovered using untargeted genomic sequencing, often involving many simultaneous discoveries with limited phenotypic data, an ideal approach would quantify the relative risk of human infectivity upon relevant exposure 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-characterised human-infecting viruses from genomic sequences [7,8]. 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 generalise across viruses. Consequently, predictions may be highly sensitive to substantial biases in current knowledge of viral diversity [3,9].

Empirical and theoretical evidence suggests that generalisable signals of human infectivity might exist within viral genomes. Viruses associated with broad taxonomic groups of animal reservoirs (e.g., primates versus rodents) can be distinguished using aspects of their

genome composition, including dinucleotide, codon, and amino acid biases [10]. Whether such measures of viral genome composition are specific enough to distinguish host range at the species level remains unclear, but their specificity might arise through several commonly hypothesised mechanisms. First, aspects of antiviral immunity that target nucleotide motifs in viral genomes might select for common mutations in diverse human-associated viruses [11,12]. 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 [12]. 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 [13]. Second, the frequencies of specific codons in virus genomes often resemble those of their reservoir hosts, possibly owing to increased efficiency and/or accuracy of mRNA translation [14]. By driving genome compositional similarity to human-adapted viruses or to the human genome, such processes may preadapt viruses for human infection [15,16]. Finally, even in the absence of mechanisms that assert common selective pressures on divergent viral genomes, the phylogenetic relatedness of viruses could allow prediction of the potential for human infectivity since closely related viruses are generally assumed to share common phenotypes and host range. However, despite being a common rule of thumb for virus risk assessment, to our knowledge, whether evolutionary proximity to viruses with known human infection ability predicts zoonotic status remains untested.

We aimed to develop machine learning models 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 (here, zoonotic potential). Using a large dataset of viruses which had previously been assessed for human infection ability based

on published reports, we first build models which assess zoonotic potential based on virus taxonomy and/or phylogenetic relatedness to known human-infecting viruses and contrast these models to alternatives based on hypothesized selective pressures on viral genome composition which favour human infectivity. We then apply the best-performing model to explore patterns in the predicted zoonotic potential of additional virus genomes sampled from a range of species.

Results

We collected a single representative genome sequence from 861 RNA and DNA virus species spanning 36 viral families that contain animal-infecting species (S1 Fig). We labelled each virus as being capable of infecting humans or not using published reports as ground truth, and trained models to classify viruses accordingly. These classifications of human infectivity were obtained by merging three previously published datasets which reported data at the virus species level and therefore did not consider potential for variation in host range within virus species [5,9,17]. Importantly, given diagnostic limitations and the likelihood that not all viruses capable of human infection have had opportunities to emerge and be detected, viruses not reported to infect humans may represent unrealized, undocumented, or genuinely non-zoonotic species. Identifying potential or undocumented zoonoses within our data was an *a priori* goal of our analysis.

We first evaluated whether phylogenetic proximity to human-infecting viruses predictably elevates zoonotic potential. Gradient boosted machine (GBM) classifiers trained on virus taxonomy or the frequency of human-infecting viruses among close relatives identified by sequence similarity searches ("phylogenetic neighbourhood", defined using nucleotide BLAST [10]) outperformed chance (median area under the receiver-operating characteristic curve [AUC_m] = 0.604 and 0.558, respectively), but were no better than

manually ranking novel viruses by the proportion of human-infecting viruses in each family ("taxonomy-based heuristic", $AUC_m = 0.596$, Fig 1A). This indicates that relatedness-based models were not only unable to identify novel zoonoses which are not close relatives of known human-infecting viruses, but were also largely unable to accurately distinguish risk among closely related viruses (S2 Fig). Moreover, the performance of these models depended on the data available for model training, sometimes performing worse than chance, making them highly sensitive to current knowledge of viral diversity.

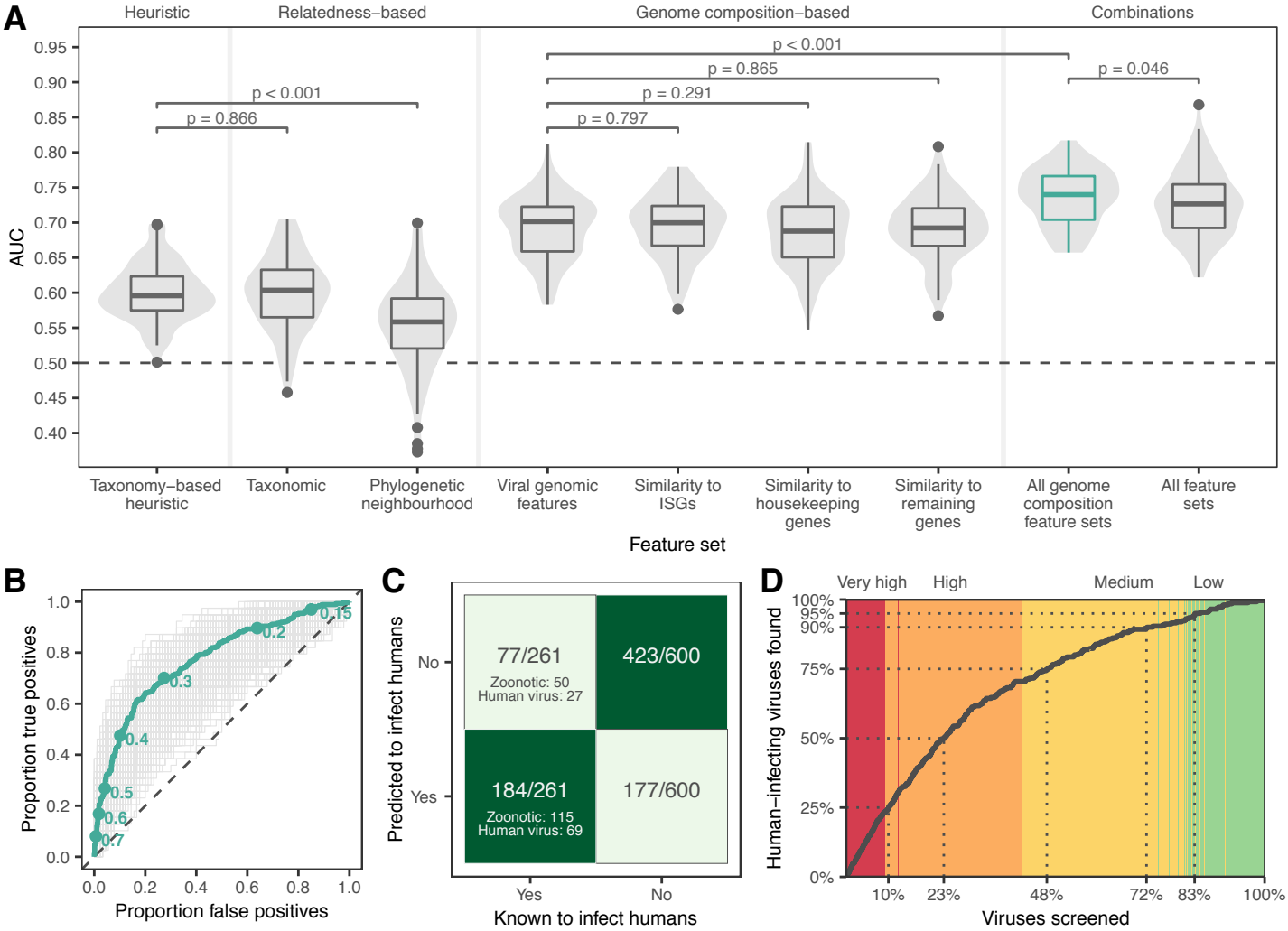


Fig 1. Machine learning prediction of human infectivity from viral genomes. (A) Violins and boxplots show the distribution of AUC scores across 100 replicate test sets. (B) Receiver-operating characteristic curves showing the performance of the model trained on all genome composition feature sets across 1000 iterations (grey) and performance of the bagged model

derived from the top 10% of iterations (green). Points indicate discrete probability cut-offs for categorizing viruses as human-infecting. (C – D) show binary predictions and discrete zoonotic potential categories from the bagged model, using the cut-off which balanced sensitivity and specificity (0.293). (C) Heatmap showing the proportion of predicted viruses in each category. (D) Cumulative discovery of human-infecting species when viruses are prioritised for downstream confirmation in the order suggested by the bagged model. Dotted lines highlight the proportion of all viruses in the training and evaluation data which need to be screened to detect a given proportion of known human-infecting viruses. Background colour highlights the assigned zoonotic potential categories of individual viruses encountered (red: very high, orange: high, yellow: medium, green: low).

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 the similarity of viral genome composition to that of three distinct sets of human gene transcripts ("human similarity features"): interferon-stimulated genes (ISGs), housekeeping genes, and all other genes. We hypothesized that if viruses need to adapt to either evade innate immune surveillance for foreign nucleic acids or to optimise gene expression in humans, they should resemble ISGs since both tend to be expressed concomitantly in virus-infected cells. We selected two additional sets comprising non-ISG housekeeping genes and all remaining genes to explore whether signals were specific to ISGs. GBMs trained using genome composition feature sets performed similarly when tested separately (AUC_m = 0.688–0.701) and consistently outperformed models based on relatedness alone (both the taxonomy-based heuristic and machine learning models trained on virus taxonomy or phylogenetic neighbourhood, Fig 1A). Combining all four genome composition feature sets further improved, and reduced variance

150 in, performance ($AUC_m = 0.740$, Fig 1A), suggesting that measures of similarity to human
151 transcripts contained information unavailable from viral genomic features alone. In contrast,
152 adding relatedness features to this combined model reduced accuracy ($AUC_m = 0.726$) and
153 increased variance (Fig 1A). Averaging output probabilities over the best 100 out of 1000
154 iterations of training on random test/train splits of the data (a process akin to bagging, using
155 ranking performance on non-target viruses to select high performing models) further
156 improved the combined genome feature-based model (AUC = 0.773, Fig 1B).

157 To estimate model sensitivity and specificity, we converted the mean of predicted
158 probabilities of human infection from the bagged model into binary classifications (i.e.,
159 human infecting or not), predicting viruses with predicted probabilities > 0.293 as human-
160 infecting. This cut-off balanced sensitivity and specificity (both 0.705, Fig 1C), though in
161 principle, higher or lower cut-offs could be selected to prioritize reduction of false positives
162 or false negatives, respectively (Fig 1B). These binary predictions correctly identified 71.9%
163 of viruses which predominately or exclusively infect humans and 69.7% of zoonotic viruses
164 as human infecting, though performance varied among viral families (Fig 1C, S3 Fig). Since
165 binary classifications ignore both the variability between iterations and the rank of viruses
166 relative to each other, we further converted predicted probabilities of zoonotic potential into
167 four zoonotic potential categories, describing the overlap of confidence intervals with the
168 0.293 cut-off from above (low: entire 95% confidence interval [CI] of predicted probability ≤
169 cut-off; medium: mean prediction ≤ cut-off, but CI crosses it; high: mean prediction > cut-
170 off, but CI crosses it; very high: entire CI > cut-off). Under this scheme, the majority (92%)
171 of known human-infecting viruses were predicted to have either medium (21.5%), high
172 (47.1%) or very high (23.4%) zoonotic potential, while only 8% (N=21) had low zoonotic
173 potential (S4 Fig, S1 Table). Fourteen viruses not currently considered to infect humans by
174 our criteria were predicted to have very high zoonotic potential (S5 Fig), although at least

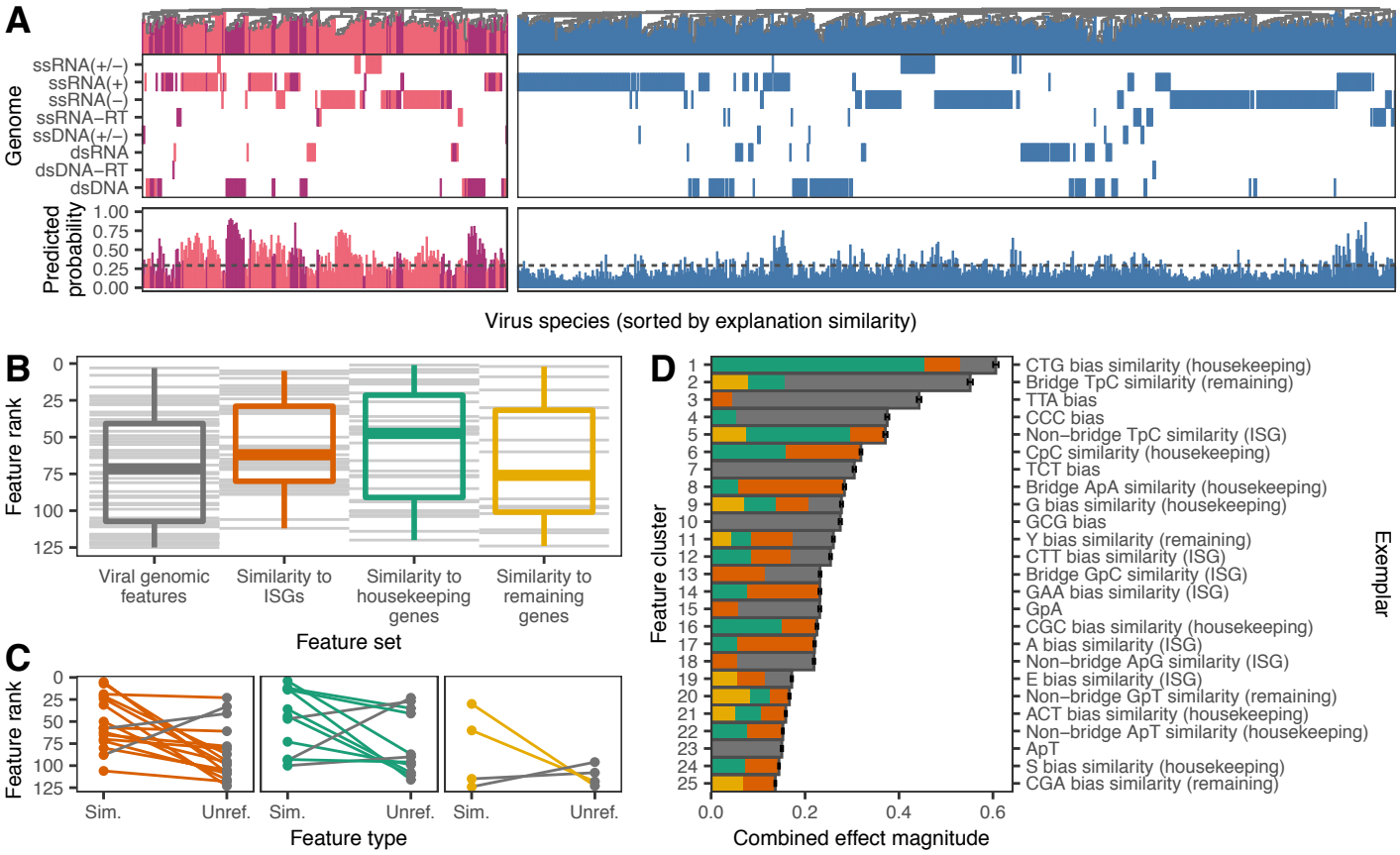
175 three of these (*Aura virus*, *Ndumu virus* and *Uganda S virus*) have serological evidence of
176 human infection [5,17], suggesting they may be valid zoonoses rather than model
177 misclassifications. Across the full dataset, 77.2% of viruses predicted to have very high
178 zoonotic potential were known to infect humans (S1 Table). Consequently, studies aimed at
179 confirming human infectivity (e.g., by attempting to infect human-derived cell lines or by
180 serological testing of humans in high-risk populations) while screening viruses in the order
181 suggested by our ranking would have found 23.4% of all known human-infecting viruses in
182 this dataset after screening just the very high zoonotic potential viruses (9.2% of all viruses).
183 More generally, 50% of known human-infecting viruses would have been found after
184 screening the top-ranked 23.3% of viruses, and 75% after screening the top 48% of viruses
185 (Fig 1D). In contrast, if relying only on relatedness to known zoonoses, confirming the first
186 50% of currently-known zoonoses would have required screening either 40.2% (taxonomy-
187 based model) or 41.5% (phylogenetic neighbourhood-based model) of viruses; a 1.7 to 1.8-
188 fold increase in effort compared to our best model (S6 Fig).

189 Since genome composition features partly track viral evolutionary history [10], it is
190 conceivable that our models made predictions by reconstructing taxonomy more accurately
191 than the phylogenetic neighbourhood estimator or in more detail than available to the
192 taxonomy-based model. We therefore compared dendrograms that clustered viruses by either
193 taxonomy, raw genomic features, or the relative influence of each genomic feature on the
194 model prediction for each virus (here estimated using the SHAP algorithm of [18], and hence
195 termed SHAP values in the rest of this manuscript). As high levels of similarity in SHAP
196 values between viruses would indicate that they share similar predictions for the same
197 reasons, we therefore analysed to what extent such similar uses of the same genomic features
198 followed known taxonomic patterns. While dendrograms using raw feature values closely
199 correlated with virus taxonomy for both human-infecting and other viruses (Baker's [19]

200 $\gamma = 0.617$ and 0.492 respectively, $p < 0.001$), dendrograms of SHAP similarity had 10.28-
201 and 2.07-fold reduced correlations with virus taxonomy ($\gamma = 0.060$ and 0.238, although this
202 was still more correlated than expected by chance, $p \leq 0.008$; S7 Fig). Among human-
203 infecting viruses, correlations between SHAP similarity-based clustering and virus taxonomy
204 weakened at deeper taxonomic levels, even though the input genomic features provided
205 sufficient information to partially reconstruct virus taxonomy at the realm, kingdom and
206 phylum levels (S8 Fig). These results indicate that more taxonomic information was available
207 than was utilised by the trained model to predict human infection ability. Interestingly,
208 dendrograms of SHAP similarity showed that even viruses with different genome types –
209 indicating ancient evolutionary divergence or separate origins – clustered together (Fig 2A).
210 Alongside earlier observations on classifier performance (Fig 1A), this suggests that the
211 genome composition-based model outperformed relatedness-based approaches because it
212 found common viral genome features that increase the capacity for human infection across
213 diverse viruses.

214 Although our analysis was not designed to conclusively identify biological
215 mechanisms underlying genomic predictors of human infection, we nevertheless were able to
216 explore emergent patterns relating to how specific genome composition features and groups
217 of features relate to human infectivity. We first compared the relative influence of features
218 from different genome composition categories (i.e., genomic features versus the three sets of
219 human similarity features). Representatives of all genome composition categories were
220 retained in the final model, though we found some evidence that compositional similarity to
221 human housekeeping genes and ISGs influenced predictions more strongly than unreferenced
222 viral genomic features (Fig 2B–C; S1 Text). We next explored the influence of individual
223 features on model predictions in more detail. Unsurprisingly, given that GBMs are designed
224 to make predictions from large numbers of weakly informative features [20], no single

225 feature stood out as the driving force and many features formed correlated clusters (Fig 2D &
226 S9 Fig). More interestingly, many features had complex, non-linear relationships with human
227 infection (S10 Fig), such that increased similarity to human gene transcripts did not always
228 increase the likelihood of infecting humans (S1 Text). We speculate this might reflect trade-
229 offs between different features within viral genomes or context dependencies whereby both
230 mimicry of human transcripts (e.g., for improved translation efficiency) or divergence from
231 human transcripts (e.g., for evasion of nucleotide motif-targeting defences) may occur for
232 different features (S1 Text).



233 **Fig 2. Genomic determinants of human-infecting viruses.** (A) SHAP value clustering of
234 viruses known to infect humans (primarily human-associated, dark purple, and zoonotic,
235 pink) and those with no known history of human infection (blue) shows that similar features
236 predicted human infection across viruses with different genome types (rows). A second set of
237 panels shows the predicted probability of infecting humans for each virus, with the dashed

238 line indicating the cut-off which balances sensitivity and specificity. (B) Relative importance
239 of individual features in shaping predictions, determined by ranking features by the mean of
240 absolute SHAP values across all viruses. Grey lines represent individual features; boxplots
241 show the median, $25^{th}/75^{th}$ percentiles and range of ranks for each feature set. (C) Difference
242 in ranks of features when both unreferenced ("Unref.") and similarity to human genomes
243 ("Sim.") forms were retained in the final model. Lines are coloured according to the highest
244 ranked representation in each pairwise comparison; colours as in B. (D) Composition of the
245 top 25 most important clusters of correlated features shaping predictions. Discrete clusters of
246 correlated features were identified by affinity propagation clustering. Clusters are shown
247 ranked by the combined effect magnitude of constituent features, defined as the sum of mean
248 absolute SHAP values for all features in the cluster, and the exemplar feature of each cluster
249 is provided on the right axis. Bars represent means (± SEM) across 1000 iterations and are
250 shaded by the proportion of the cluster from each feature set; colours as in B.

251

252 Finally, we carried out two case studies to illustrate the utility of our prediction
253 framework. First, we used the combined genome feature-based model to rank 758 virus
254 species which were not present in our training data. We included all species in the most
255 recent ICTV taxonomy release (#35, 24 April 2020) belonging to animal-infecting virus
256 families and which were originally discovered or sequenced from mammals (including
257 humans), birds, two insect orders containing common virus vectors (Diptera and Ixodida), or
258 where the sampled host was not reported. This dataset contained representatives from 38 viral
259 families, including 2 (*Anelloviridae* and *Genomoviridae*) which were not present in data used
260 to train our model. In total, 70.8% of viruses sampled from humans were correctly identified
261 as having either very high (N=36) or high zoonotic potential (N=44; Fig 3A). The remaining
262 human-associated viruses were primarily classified as medium zoonotic potential (N=30),

263 with 3 species predicted to have low zoonotic potential (*Mammalian orthoreovirus* and
264 *Human associated gemykibivirus 2* and *3*; Fig 3A). Within the viral families never previously
265 seen by our model, the majority of human-associated anelloviruses (39/45, 86.6%) were
266 correctly identified as having either very high or high zoonotic potential, consistent with the
267 conclusion that viral genomic features which enhance human infectivity can generalise across
268 viral families. In contrast, all 6 human-associated genomoviruses were classified as either
269 medium or low zoonotic potential. The lower performance on genomoviruses may reflect the
270 unusual genomic structure of this family (circular, single stranded DNA) which was poorly
271 represented in training (only 2 representatives from the *Circoviridae* family; S1 Fig) and may
272 impose different selective forces. Further, the small genome sizes of genomoviruses (2.2–
273 2.4kb) may complicate calculation of genomic features due to the low number of nucleotides,
274 dinucleotides and codons available (cf. S3 Fig). Among the 645 viruses with unknown human
275 infectivity that were sequenced from non-human animal or potential vector samples, 45.0%
276 were predicted to have either very high (N=41) or high zoonotic potential (N=272; S11 Fig &
277 S1 Table). The very high zoonotic potential category was dominated by *Papillomaviridae*
278 (34.1%) and *Peribunyaviridae* (19.5%).

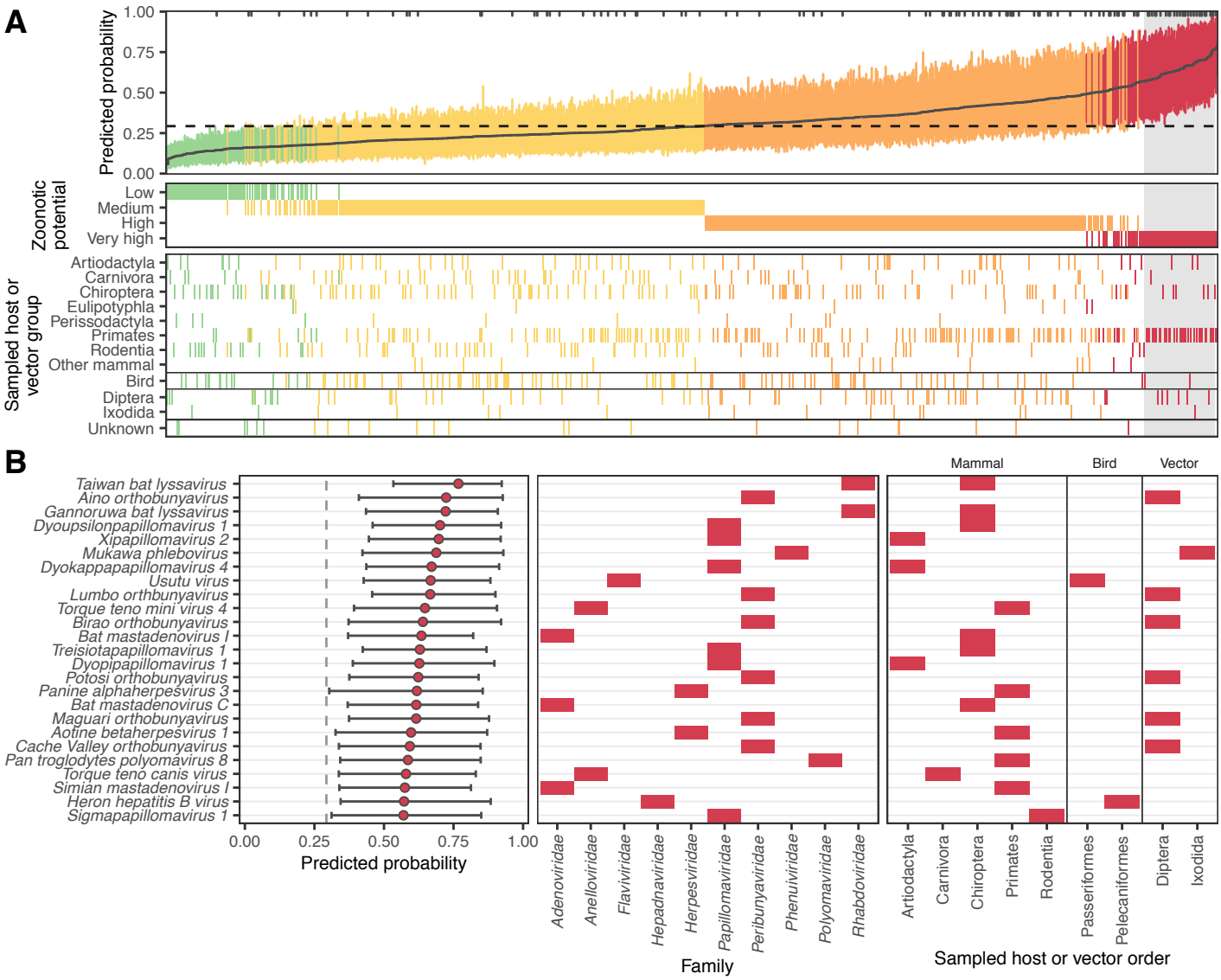


Fig 3. Probability of human infection predicted from holdout viral genomes. (A) Predicted probability of human infection for 758 virus species which were not in the training data. Colours show the assigned zoonotic potential categories, with an additional panel showing the host or vector group each virus genome was sampled from. Tick marks along the top edge of the first panel show the location of virus genomes sampled from humans, while a dashed line shows the cut-off which balanced sensitivity and specificity in the training data. The top 25 viruses which were not sampled from humans (contained within the grey box) are illustrated in more detail in (B). Bars show the 95% interquantile range of predicted probabilities across the best-performing 10% of iterations (based on the training

288 data), while a solid line (A) or circles (B) show the mean predicted probability from these
289 iterations.
290
291 We next used a beta regression model to explore how predictions of zoonotic
292 potential varied among host and viral groups. As expected given the performance on our
293 training and evaluation data (Fig 1), the 113 virus species that were sequenced from human
294 samples scored consistently higher than those detected in other hosts ($p < 0.001$; figs. 3A &
295 4D). Although viruses from putatively high risk host groups including bats, rodents and
296 artiodactyls formed a large fraction of our holdout data (with viruses from bats outnumbering
297 even those from humans, S11 Fig), they did not have elevated predicted probabilities of being
298 zoonotic (Fig 4C), and no differences were detected at higher host taxonomic levels (Fig 4A
299 – B). This highlights a potential disparity between current sampling efforts for virus
300 discovery/reporting and the distribution of zoonotic risk. In contrast, viruses linked to
301 primates had higher predicted probabilities of infecting humans, even after accounting for
302 human-associated viruses and the effects of virus family (Figs 3 – 4 & S11 Fig). That genome
303 composition-based models predicted elevated zoonotic potential in non-human primate-
304 associated viruses despite receiving no information on sampled host further supports host-
305 mediated selective processes as a biological basis for our model's predictions. In addition to
306 relatively rare and small host effects, we observed more pervasive positive and negative
307 effects of virus family on predicted zoonotic status (Fig 4E). Taken together, our results are
308 consistent with the expectation that the relatively close phylogenetic proximity of non-human
309 primates may facilitate virus sharing with humans and suggest that this may in part reflect
310 common selective pressures on viral genome composition in both humans and non-human
311 primates. However, broad differences among other animal groups appear to have less
312 influence on zoonotic potential than virus characteristics [9].

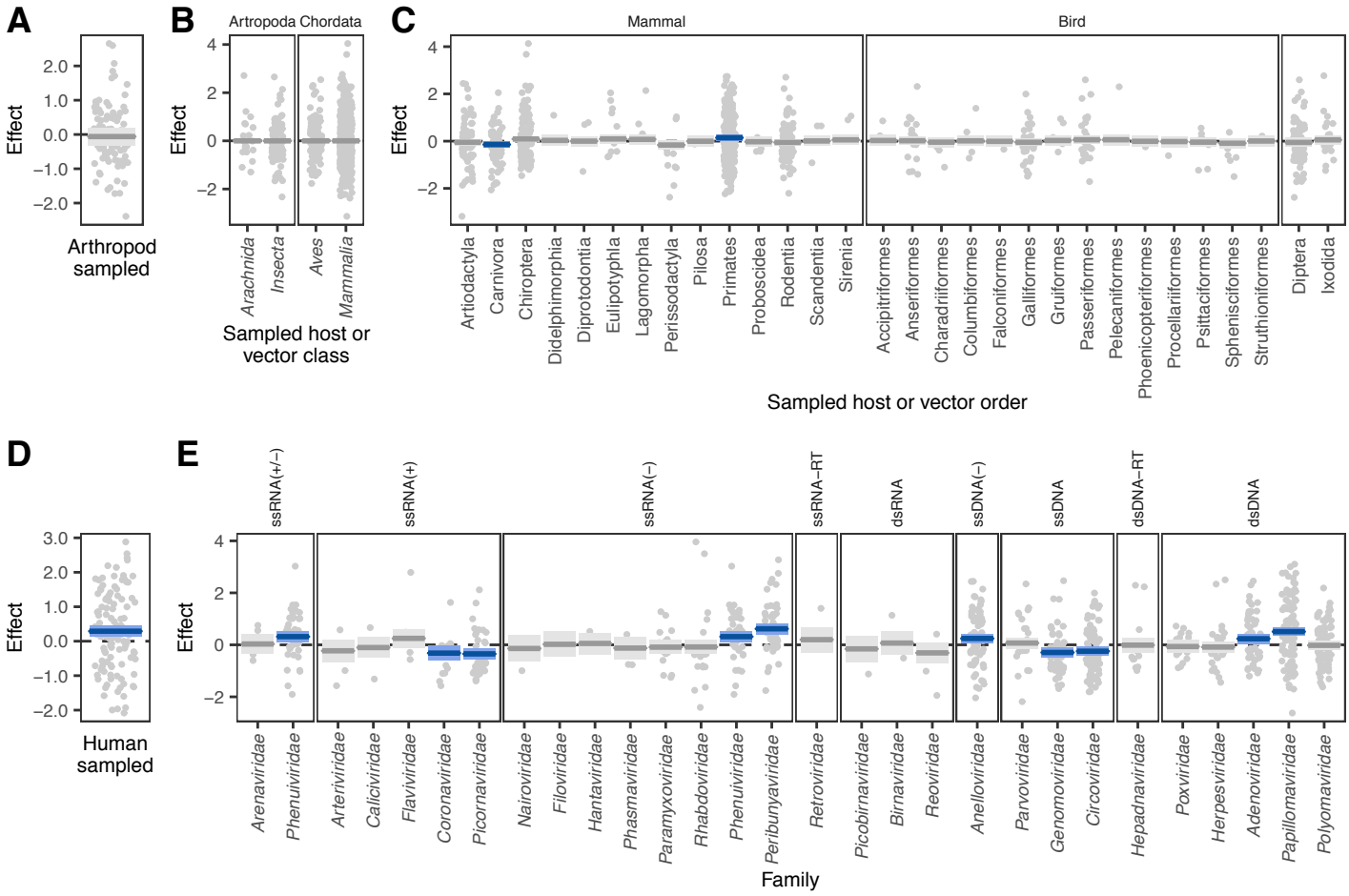


Fig 4. Factors correlated with the probability of human infection predicted from holdout viral genomes. Partial effects plots are shown for a beta regression model attempting to explain the mean probability assigned by the bagged model to all viruses in Fig 3A, accounting for whether or not the genome predicted was sequenced from arthropods (as opposed to chordates, A), random effects for the taxonomic class and order of sampled hosts (B – C), whether the sequence derived from a human sample (D), and a random effect for the virus family represented (E). Points indicate partial residuals, while lines and shaded areas respectively show the maximum likelihood and 95% confidence interval of partial effects. Confidence intervals which do not include zero are highlighted in blue.

Our second case study used coronaviruses to explore the ability of our combined genome feature-based model to distinguish different virus species within the same family and

325 different genomes within a single virus species. Specifically, we predicted the zoonotic
326 potential of all currently recognized coronavirus species, along with 62 human and animal-
327 derived genomes of *Severe acute respiratory syndrome-related coronavirus* (SARS
328 coronavirus), a species which includes many well-known viruses, including SARS-CoV,
329 SARS-CoV-2, and related sarbecoviruses. All known human-infecting coronaviruses were
330 classified as either medium or high zoonotic potential (Fig 5A). We also identified two
331 additional animal-associated coronaviruses – *Alphacoronavirus 1* and the recently described
332 *Sorex araneus coronavirus T14* – as being at least as, or more likely to be capable of
333 infecting humans than known, high ranking, human-infecting coronaviruses; these should be
334 considered high priority for further research. While this manuscript was in revision, a
335 recombinant *Alphacoronavirus 1* was detected in nasopharyngeal swabs from pneumonia
336 patients, further strengthening the case that this species may be zoonotic [21]. We further
337 observed variation in predicted zoonotic potential within coronavirus genera which was
338 consistent with our current understanding of these viruses. *Alphacoronavirus* and
339 *Betacoronavirus* (the genera which contain known human-infecting species) also contained
340 non-zoonotic species that were correctly predicted to have low zoonotic potential, while the
341 majority of delta- and gammacoronaviruses received relatively low predictions (Fig 5A).
342 These findings further illustrate the capacity of our models to discriminate risk below the
343 virus family or genus levels.

344 Within SARS coronavirus, most genomes (85.5%) were classified as having medium
345 zoonotic potential, including the causal agent of the 2003 outbreak (Fig 5B). Interestingly
346 however, SARS-CoV-2 (the causative agent of the current COVID-19 pandemic), the closely
347 related RaTG13 from a rhinolophid bat, and all 5 closely-related pangolin-associated isolates
348 tested were predicted to have high zoonotic potential (although confidence intervals between
349 all sarbecoviruses tested overlapped, Fig 5B). Importantly, these predictions were made using

iterations of our model which excluded the 2003 SARS-CoV genome or any other sarbecovirus from training. This finding, together with our observation that relatively few other animal-infecting, allegedly non-zoonotic coronaviruses had similarly high scores, suggests that the elevated risk of SARS-CoV-2 and closely-related genomes discovered in animals could have been anticipated via sequencing-based surveillance and might have led to actionable research or surveillance prior to the zoonotic emergence of any sarbecovirus (Fig 5).

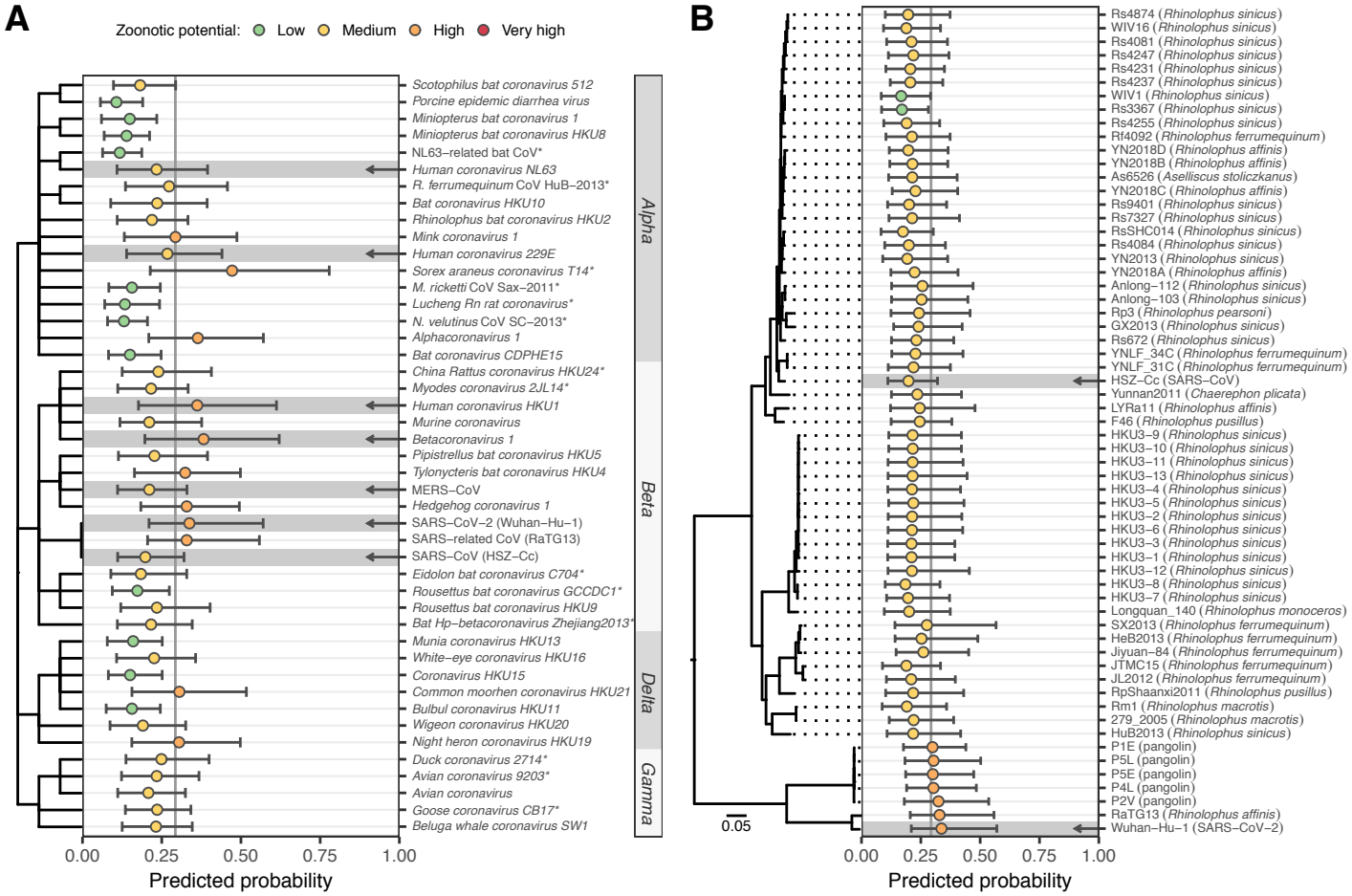


Fig 5. Probability of human infection predicted from Coronavirus genomes. (A) Predictions for currently recognized *Coronaviridae* species and for 3 variants of *Severe Acute Respiratory Syndrome-related coronavirus*: SARS-CoV (isolate HSZ-Cc, sampled early in the 2003 pandemic), SARS-CoV-2 (isolate Wuhan-Hu-1, sampled early in the current pandemic), and the closely related RaTG13 (sampled from *Rhinolophus affinis* in 2013). A

362 dendrogram illustrates taxonomic relationships, with abbreviated genus names annotated on
363 the right. Arrows highlight known human-infecting species. Asterisks indicate species absent
364 from the training data, also present in Fig 3A. (B) Predictions for different representatives of
365 *Severe acute respiratory syndrome-related coronavirus*. The isolation source of animal-
366 associated genomes is indicated in parentheses. A maximum likelihood phylogeny illustrates
367 relationships and was created as described in [6]. The outgroup, BtKy72 (sampled in Kenya
368 in 2007), is not shown. MERS-CoV = *Middle East respiratory syndrome-related*
369 *coronavirus*; *M. ricketti* CoV Sax-2011 = *Myotis ricketti alphacoronavirus Sax-2011*; NL63-
370 related bat CoV = *NL63-related bat coronavirus strain BtKYNL63-9b*; *N. velutinus* CoV SC-
371 2013 = *Nyctalus velutinus alphacoronavirus SC-2013*; *R. ferrumequinum* CoV HuB-2013 =
372 *Rhinolophus ferrumequinum alphacoronavirus HuB-2013*. In both panels, bars show the 95%
373 interquantile range of predicted probabilities across the best-performing 10% of iterations
374 excluding the species being predicted, while circles show the mean predicted probability
375 from these iterations.

376 Discussion

377 In an age of rapid, genomic-based virus discovery, rational prioritization of research
378 and surveillance activities has been an unresolved challenge. While approaches to prioritise
379 known, relatively well-characterised viruses based on a range of common risk factors have
380 been developed [4–6], the large number of viruses still being discovered presents a bottleneck
381 for the very characterisation needed to apply such prioritisation schemes, necessitating use of
382 expert opinion or surrogate data from related species [6]. Our findings show that the zoonotic
383 potential of viruses can be inferred to a surprisingly large extent from their genome sequence,
384 outperforming current alternatives. Indeed, our results suggest that routine proxies of
385 zoonotic risk which can be applied to poorly characterised viruses including virus taxonomy

386 and relative phylogenetic proximity to human-infecting species [5,9,22] have limited
387 discriminatory power. This has far-reaching implications for how risk is perceived – while it
388 is intuitive to assume that novel viruses which are closely related to known human-infecting
389 viruses are a threat, to our knowledge this assumption had never been tested. Worryingly,
390 with some training datasets such relatedness-based models performed worse than random
391 guessing (AUC < 0.5, Fig 1A), suggesting that the current incomplete knowledge of virus
392 diversity could lead to entirely incorrect priorities under such approaches. In contrast, models
393 that exploited features of viral genomes that were at least partly independent of virus
394 taxonomy both generalised predictions across divergent viruses and provided capacity to
395 discriminate risk among closely related virus species.

396 In requiring only a genome sequence, our approach has quantitative and qualitative
397 advantages over alternative models for zoonotic risk assessment. The most comprehensive
398 alternative model requires virus species-level information on publication count (a proxy for
399 study effort), the diversity of hosts infected, whether or not the virus is vector borne, and the
400 ability to replicate in the cytoplasm [5]. We estimate similar predictive performance for this
401 model ($AUC_m = 0.770$) using a subset of only mammalian viruses (see methods). However,
402 neither study effort nor knowledge of host range are available for novel viruses, and
403 restricting this model to factors which might reasonably be inferred from virus taxonomy
404 (vector borne status and ability to replicate in the cytoplasm) performs considerably worse
405 ($AUC_m = 0.647$) than our approach. We were unable to compare the performance of our
406 model to a more recently developed prioritisation system based on ecological variables
407 weighted by expert opinion, as metrics of performance were not provided and could not be
408 calculated for a comparable set of viruses with known zoonotic status [6]. Although we
409 emphasize that the viruses included and study objectives differed, that genome-based ranking
410 seems to perform comparably to or better than currently-available alternatives that require far

411 more, and often unavailable, data highlights the surprisingly informative signals of human
412 infection-ability contained within viral genomes. Crucially, from the perspective of virus risk
413 assessment, models based purely on genome sequences can be applied much earlier to
414 identify many potential zoonoses immediately after virus discovery and genome sequencing,
415 when data on most other risk factors are still unknown. Ultimately, genome-based rankings
416 could be combined with data on additional known risk factors as they become available [5,6].

417 By highlighting viruses with the greatest potential to become zoonotic, genome-based
418 ranking allows further ecological and virological characterisation to be targeted more
419 effectively. Indeed, studying viruses in the order suggested by genome-based ranking would
420 find many zoonotic viruses much earlier than current taxonomic or phylogenetically-
421 informed approaches (Fig 1D & S6 Fig). Nevertheless, we acknowledge that even after
422 applying our models, considerable numbers of viruses may need to undergo confirmatory
423 testing (e.g. infectivity experiments on human-derived cell lines [23]) before significant
424 further research investments, and this need will only increase with ongoing virus discovery.
425 Although these numbers are more manageable considering that experimental validation will
426 be dispersed across virus taxonomic groups which will be studied by different experts, efforts
427 to increase the success rates of virus isolation (a prerequisite for current *in vitro* validation
428 methods) and to create systems for high-throughput virus host range testing are clearly
429 needed to improve the efficiency of this process [23,24]. Such efforts could further generate
430 valuable feedback data; iteratively improving model performance and consequently reducing
431 the relative proportion of new viruses requiring additional laboratory testing.

432 Several lines of evidence – including SHAP clustering of viruses with different
433 genome organizations, accurate prediction of human-infecting viruses from families withheld
434 from training, and the prediction of the zoonotic risk of SARS-CoV-2 when withholding data
435 from other zoonotic sarbecoviruses – suggested that our models make predictions using

436 genomic features that predict human infection across divergent virus taxa. From a practical
437 standpoint, this is a major advantage since it means that our model borrows information
438 across families and might therefore anticipate the zoonotic potential of viruses which, due to
439 their rarity or lack of historical precedent, would not otherwise be considered high risk
440 (sometimes referred to as Disease X) [25]. From a broader evolutionary standpoint, the
441 putative existence of convergently evolved features in viral genomes which seem to
442 predispose human infection is a discovery which deserves further mechanistic study.
443 Encouragingly, a substantial literature in vaccine development facilitates genome-wide
444 synonymous recoding of viral genome composition and has established that these changes
445 can dramatically affect viral fitness [26,27]. Our results provide a path by which analogous
446 approaches could test how the features we identified affect viral host range in general and
447 human infection ability in particular. Doing so may reveal novel mechanisms of viral
448 adaptation to humans, which might represent both biologically-verified risk factors for the
449 improvement of future models of zoonotic potential and potential therapeutic targets.

450 We used single exemplar genomes from each virus species to maximise the likelihood
451 of discovering generalisable signatures of human infection while avoiding performance
452 measures which would be over-optimistic for novel viruses. A potential drawback of this
453 approach was that we omitted substantial viral diversity which is not yet formally recognized
454 by the ICTV [3]. However, we contend that including currently unrecognized viruses is
455 unlikely to improve the predictions of our models because: (a) most will be non-human
456 infecting (an already over-represented class) and hence provide little additional information,
457 (b) those which do infect humans will not generally be known to do so due to a lack of
458 historic testing, adding misleading signals, and (c) the predictive features identified often
459 span across families, reducing the impacts of taxonomic gaps. The use of single genomes
460 does however mean that the ranks produced here pertain only to the specific genomes tested

461 (in most cases, the NCBI reference sequence, S1 Table), and may not apply equally across all
462 strains within a species (Fig 5B). We also stress that our model predicts baseline zoonotic
463 *potential* (i.e., ability to infect humans), which ultimately will be modulated by ecological
464 opportunities for emergence [28,29]. Further, the societal impact of emergence will depend
465 on capacity for human to human spread and on the severity of human disease, which likely
466 require additional non-genomic data to anticipate [28,30].

467 In summary, we have constructed a genomic model that can retrospectively or
468 prospectively predict the probability that viruses will be able to infect humans. The success of
469 our models required aspects of genome composition calculated both directly from viral
470 genomes and in units of similarity to human transcripts, and some viruses were predicted to
471 be zoonotic due to common genomic traits despite ancient evolutionary divergence. This
472 highlights the potential existence of currently unknown phenotypic consequences of viral
473 genome composition that appear to influence viral host range across divergent viral families.
474 Independently of the mechanisms involved, the performance of our models shows how
475 increasingly ubiquitous and low-cost genome sequence data can inform decisions on virus
476 research and surveillance priorities at the earliest stage of virus discovery with virtually no
477 extra financial or time investment.

478 Materials and methods

479 Data

480 Although our primary interest was in zoonotic transmission, we trained models to predict the
481 ability to infect humans in general, reasoning that patterns found in viruses predominately
482 maintained by human-to-human transmission may contain genomic signals which also apply
483 to zoonotic viruses. Data on the ability to infect humans were obtained by merging the data of
484 [5], [9], and [17], which contain species-level records of reported human infections, resulting

485 in a final dataset of 861 virus species from 36 families. In all cases, only viruses detected in
486 humans by either PCR or sequencing were considered to have proven ability to infect
487 humans. All viruses for which no such reports were found were considered to not infect
488 humans (as long as they were assessed for potential human infection by at least one of the
489 studies above), although we emphasise that many of these viruses are poorly characterised
490 and could therefore be unrecognized or unreported zoonoses. We therefore expect our models
491 to further improve as these and new viruses become better characterised. For figures, human-
492 infecting viruses were further separated into primarily human-transmitted viruses and
493 zoonotic viruses, based on virus reservoirs recorded in [9]. Human-infecting viruses for
494 which the reservoir remains unknown were assumed to be zoonotic, while viruses with both a
495 human and non-human reservoir cycle (e.g., *Dengue virus*) were recorded as primarily
496 human-transmitted, reflecting the primary source of human infection. A representative
497 genome was selected for each virus species, giving preference to sequences from the RefSeq
498 database wherever possible. RefSeq sequences that had annotation issues, represented
499 extensively passaged isolates, or were otherwise not judged to be representative of the
500 naturally circulating virus were replaced with alternative genomes.

501 Features

502 We compared the predictive power of all classifiers to that which could be obtained through
503 knowledge of a virus' taxonomic position alone. This captures the intuitive expectation that
504 viruses can be risk-assessed based on knowledge of the human-infection abilities of their
505 closest known relatives. To formalise this idea, we first created a simple heuristic which
506 ranks viruses based on the proportion of other viruses in the same family which are known to
507 infect humans ('*taxonomy-based heuristic*' in Fig 1). Viral family was chosen as the level of
508 comparison because not all viruses are classified in a scheme which includes subfamilies,
509 while lower taxonomic levels suffer from limited sample size. To further characterise the

510 predictive power of virus taxonomy, we also included potential predictor variables (here
511 termed *features*) describing virus taxonomy when training classifiers. These included the
512 proportion of human infecting viruses in each family (calculated from species in the training
513 data only) along with categorical features describing the phylum, subphylum, class, subclass,
514 order, suborder, and family to which each species was assigned (‘*taxonomic feature set’*, 8
515 features). This information was taken from version 2018b of the ICTV master species list
516 (https://talk.ictvonline.org/files/master-species-lists/). To capture taxonomic effects at finer
517 resolution, we summarised the human infection ability of the closest relatives of each virus in
518 the training data (following [10], here termed the ‘*phylogenetic neighbourhood feature set*’, 2
519 features). To calculate these features, the genome (or genome segments, where applicable) of
520 each virus species was nucleotide BLASTed against a database containing genomes or
521 genome segments for all species in the training data. All BLAST matches with e-
522 value ≤ 0.001 were retained and used to calculate the proportion of human-infecting viruses
523 in the phylogenetic neighbourhood of each virus (excluding the current species). We also
524 calculated a ‘distance-corrected’ version of this proportion by re-weighting matches
525 according to the proportion of nucleotides p matching the genome of the focal virus:

526 $$C = \frac{\sum_{i=1}^{N} p_i \cdot x_i}{N},$$

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

535 Various features summarising the compositional biases in each virus genome were
536 calculated as described in [10]. These included codon usage biases, amino acid biases,
537 dinucleotide biases across the entire genome, dinucleotide biases across coding regions only,
538 dinucleotide biases spanning the bridges between codons (i.e. across base 3 of the preceding
539 codon and base 1 of the current codon), and dinucleotide biases at non-bridge positions
540 ('*viral genomic features*', 146 features). Similarity features to human RNA transcripts were
541 obtained by first calculating the above compositional biases for human genes. For each gene,
542 the sequence of the canonical transcript was obtained from version 96 of Ensembl [31].
543 Genes were divided into three mutually exclusive sets, encompassing interferon stimulated
544 genes (ISGs, taken from [13]; N = 2054), non-ISG housekeeping genes ([32]; N=3172), and
545 remaining genes (N=9565). The distribution of observed values for each genome feature was
546 summarised across all genes in a set by calculating an empirical probability density function
547 using version 2.3.1 of the EnvStats library in R version 3.5.1 [33]. The final similarity score
548 for each genome feature of each virus was then calculated by evaluating this density function
549 at the value observed in the virus genome, giving the probability of observing this value
550 among the transcripts of the set of human genes in question (S12 Fig). This yielded three
551 feature sets termed '*similarity to ISGs*', '*similarity to housekeeping genes*' and '*similarity to*
552 *remaining genes*', each containing 146 features.

553 # Training

554 Gradient boosted classification trees were trained using the xgboost and caret libraries
555 (versions 0.90 and 6.0-85, respectively) in R [34,35]. Note that while we separate primarily
556 human-transmitted viruses and zoonotic viruses in some figures, these viruses were
557 considered a single class during training. Thus, models were trained to distinguish viruses
558 known to infect humans from those with no reports of human infection.

559 A range of models were trained using different combinations of the individual feature
560 sets described above. To reduce run-times and potential overfitting in subsequent steps,
561 features were subjected to a pre-screening step to remove those with little or no predictive
562 value. During this pre-screening step, models were trained on a random selection of 70% of
563 the data using all features within the respective feature group (e.g., all ISG similarity
564 features). Training sets were selected using stratified sampling (i.e., selecting positive and
565 negative examples separately) to retain the observed frequencies of positive (known human-
566 infecting) and negative (not known to infect humans) virus species. We were unable to
567 additionally stratify training set selection by virus family due to the small numbers of species
568 in many families. All hyperparameters were kept at their default values, except for the
569 number of training rounds, which was fixed at 150. The importance of each feature was
570 summarised across 100 iterations in which the same features were used to train a model on
571 different samples of the full dataset, and the N most predictive features were retained. A
572 range of possible values for N was evaluated by combining all feature sets and using the
573 selected features to optimize and train a final set of model iterations as described below. The
574 final value of $N = 125$ was chosen as the point at which additional features provided no
575 further improvement in performance, measured as the area under the receiver-operating
576 characteristic curve (AUC; S13 Fig). Here, AUC measures the probability that a randomly
577 chosen human-infecting virus would be ranked higher than a randomly chosen virus which
578 has not been reported to infect humans. When a given feature set or combination of feature
579 sets comprised < 125 features, all features were retained.
580 Final models were trained using reduced feature sets. To assess the variability in
581 accuracy across different training sets, training was repeated 100 times [10]. In each iteration,
582 training was performed on a random, class-stratified selection of 70% of the available data
583 (here, the *training set*). Output probabilities were calibrated using half the remaining data

584 (*calibration set*, again selected randomly and stratified by human infection status), leaving
585 15% of the full dataset for evaluation of model predictions (*test set*). In each iteration,
586 hyperparameters were selected using 5-fold cross-validation on the training set, searching
587 across a random grid of 500 hyperparameter combinations. This cross-validation was
588 adaptive, evaluating each parameter combination on a minimum of 3 folds before continuing
589 cross-validation only with the most promising candidates [35]. The parameter combination
590 maximising AUC across folds was selected and used to train a final model on the entire
591 training set. This model was then used to produce quantitative scores for each species in the
592 calibration and test sets.

593 Next, outputs were calibrated to allow interpretation as probabilities using the beta
594 calibration method of [36]. A calibration model was fit to scores obtained for the calibration
595 set using version 0.1.0 of the betacal R package. The fitted calibration model was used to
596 produce final output probabilities for virus species in the test set. Finally, to summarise
597 predicted probabilities from the same model trained on different training sets, we averaged
598 the calibrated probabilities across the best-performing iterations (a process with similarities to
599 bagging [10]; 1000 iterations performed). Bagging of virus species from the training data
600 relied on the best 10% of iterations in which each virus occurred in the test set. As such, the
601 focal virus had no influence on the training or calibration of the iterations used in bagging.
602 Further, the performance of each model iteration was re-calculated while excluding the focal
603 species from the test set, to prevent accurate prediction of the focal virus from influencing the
604 choice of iterations used for bagging. When predicting the probability of human infectivity
605 for viruses that were completely separated from training (i.e., those in our first case study),
606 we used the best 10% of iterations overall.

607 Calculating phylogenetic neighbourhood features represented a bottleneck in the
608 iterative training strategy described above, because each iteration had a different reference

BLAST database, corresponding to the specific training set selected. This would have required repeating BLAST searches at every iteration. 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 [37]:

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

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.

Feature importance and clustering

To assess the variability in feature importance while accounting for all viruses, feature importance was assessed across all 1000 iterations produced for bagging above. In each iteration, the influence of features was assessed using SHAP values, an approximation of Shapley values which here describe the change in the predicted log odds of infecting humans attributable to each genome composition feature used in the final model [18]. In each iteration, this produced a SHAP value for each virus–feature combination. 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 [38].

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:

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

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

635 $$I_{c,i} = \frac{\sum_{j=1}^{N_j} I_{c,j,i}}{N_j}.$$

636 Feature clusters were obtained by affinity propagation clustering, which seeks to identify
637 discrete clusters of features centred around a representative feature (the *exemplar* feature)
638 [39]. Features were clustered using pairwise Spearman correlations as the similarity measure,
639 using version 1.4.8 of the apcluster library in R [40].

640 To further explore patterns in feature importance across virus species, we followed a
641 strategy similar to [38], clustering viruses based on the average SHAP values assigned to
642 individual features for each virus across all iterations:

643 $$\bar{S}_{f,j} = \frac{\sum S_{f,i,j}}{N_i}.$$

644 These values were used to calculate the pairwise Euclidean distances between all virus
645 species using version 2.1.0 of the cluster library in R [41]. Viruses were then clustered using
646 agglomerative hierarchical clustering, calculating distances between clusters as the mean
647 distance between all points in the respective clusters (i.e., UPGMA clustering). To explore
648 patterns common to viruses from each class, clustering was performed separately for known
649 human infecting and other viruses.

650 To compare this explanation-based clustering with virus taxonomy, we also constructed
651 a dendrogram based on taxonomic assignments as recorded in version 2018b of the ICTV
652 master species list, using all taxonomic levels from phylum to subgenus. Since some levels of
653 the ICTV taxonomy are not used consistently across all viruses, missing taxonomic levels
654 were interpolated to ensure accurate representation of the underlying taxonomy. For example,

655 for viruses which are not classified in a scheme which includes subfamilies, the next level
656 downstream – genus – was repeated, thereby treating each genus as belonging to a distinct
657 subfamily. Categorical taxonomic assignments were used to calculate pairwise Gower
658 distances between virus species [42], before performing agglomerative hierarchical clustering
659 as described above. We also assessed the ability of underlying genome feature values to
660 reconstruct virus taxonomy by performing hierarchical clustering on a Euclidean distance
661 matrix calculated directly from all genome composition features (i.e., the unreferenced
662 genome, ISG similarity, housekeeping gene similarity and remaining gene similarity feature
663 sets). The similarity between dendrograms was assessed using the gamma correlation index
664 of [19], as implemented in dendextend version 1.12.0 in R [43]. A null distribution for this
665 statistic was calculated by randomly shuffling the labels (i.e., virus species names) of both
666 dendrograms 1000 times. To assess the taxonomic depth at which dendrograms were
667 concordant, the Fowlkes-Mallows index was calculated at each possible cut-point in the
668 dendrograms being compared [44], again using the dendextend library. As before, a null
669 distribution was generated by randomly shuffling the labels of both dendrograms 1000 times.

670 Ranking holdout viruses

671 To illustrate the use of our models in practice, the best performing model (i.e. the bagged
672 model trained using the best 125 features selected from among all genome composition-based
673 feature sets, here termed the '*combined genome feature-based model*') was used to generate
674 predictions for a set of held out viruses. We included all virus species recognized in the latest
675 version of the ICTV taxonomy (release #35, 2019; https://talk.ictvonline.org/taxonomy)
676 which were from families known to contain species that infect animals but which did not
677 occur in our training data because they were absent from the previously described databases
678 of human infection ability used to form the training data [5,9,17]. These included all 36
679 families represented in the training data, plus *Anelloviridae* and *Genomoviridae*. Names of

viruses in the training data were updated to the latest taxonomy and checked for matches in the corresponding ICTV master species list before extracting non-matching species. For each 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 composition features as described above. The host from which each virus genome was generated was obtained from either the corresponding GenBank entry, the publication first describing the sequence, or the ArboCat database. This host information was used to further subset viruses to include only those sampled from birds, mammals, Diptera (which includes common vectors such as mosquitos and sandflies) and Ixodida (ticks), or for which the sampled host could not be identified. Scoring was performed using each of the top 10% out of 1000 iterations, 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. Tests for the effects of virus family and sampled host on the predicted probability of infecting humans were performed by fitting a beta regression model using version 1.8-27 of the mgcv library in R. This model fitted the mean predicted probability for each virus as a function of whether a human host was sampled, whether the sample derived from arthropods (both binary fixed effects), random effects for the taxonomic order and class of sampled hosts, and a random effect for virus family. Partial effects plots were generated using code from [9].

Evaluating existing model performance

To compare the performance of our models to previously published ecological models, we downloaded the fitted models of [5]. The best viral traits model fitted while excluding serological detections (termed “stringent data” in [5], N = 408) was then subjected to a testing regime similar to that used for our models. Across 100 iterations, models were re-fit using a randomly selected subset consisting of 85% of the virus species used in [5]. Each

fitted model was then used to predict probabilities of being zoonotic for the remaining 15% of species. This matched the evaluation strategy used for our own models (as reported in Fig 1A), with 85% of the data used during training and calibration, leaving 15% of the data to test model accuracy. Predictions from each iteration were used to calculate AUC using version 1.2.0 of the ModelMetrics package in R.

Acknowledgments

We thank Laura Bergner, Richard Orton, Andrew Shaw, Sam Wilson and David Robertson for helpful discussions and suggestions. 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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data and materials availability

Data and code related to this paper are available at: https://doi.org/10.5281/zenodo.4271479

References

1. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, et al. The global virome project. Science. 2018;359: 872–874. doi:10.1126/science.aap7463

2. Holmes EC, Rambaut A, Andersen KG. Pandemics: spend on surveillance, not prediction. Nature. 2018;558: 180–182. doi:10.1038/d41586-018-05373-w

3. Wille M, Geoghegan JL, Holmes EC. How accurately can we assess zoonotic risk? PLOS Biology. 2021;19: e3001135. doi:10.1371/journal.pbio.3001135

4. Pulliam JRC, Dushoff J. Ability to replicate in the cytoplasm predicts zoonotic transmission of livestock viruses. The Journal of Infectious Diseases. 2009;199: 565–568. doi:10.1086/596510

730 5. Olival KJ, Hosseini PR, Zambrana-Torrelio C, Ross N, Bogich TL, Daszak P. Host and
731 viral traits predict zoonotic spillover from mammals. Nature. 2017;546: 646–650.
732 doi:10.1038/nature22975

733 6. Grange ZL, Goldstein T, Johnson CK, Anthony S, Gilardi K, Daszak P, et al. Ranking
734 the risk of animal-to-human spillover for newly discovered viruses. PNAS. 2021;118.
735 doi:10.1073/pnas.2002324118

736 7. Zhang Z, Cai Z, Tan Z, Lu C, Jiang T, Zhang G, et al. Rapid identification of human-
737 infecting viruses. Transboundary and Emerging Diseases. 2019;66: 2517–2522.
738 doi:10.1111/tbed.13314

739 8. Bartoszewicz JM, Seidel A, Renard BY. Interpretable detection of novel human viruses
740 from genome sequencing data. bioRxiv. 2020; 2020.01.29.925354.
741 doi:10.1101/2020.01.29.925354

742 9. Mollentze N, Streicker DG. Viral zoonotic risk is homogenous among taxonomic orders
743 of mammalian and avian reservoir hosts. Proceedings of the National Academy of
744 Sciences. 2020;117: 9423–9430. doi:10.1073/pnas.1919176117

745 10. Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors
746 from evolutionary signatures in RNA virus genomes. Science. 2018;362: 577–580.
747 doi:10.1126/science.aap9072

748 11. Rothenburg S, Brennan G. Species-Specific Host–Virus Interactions: Implications for
749 Viral Host Range and Virulence. Trends in Microbiology. 2020;28: 46–56.
750 doi:10.1016/j.tim.2019.08.007

751 12. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al.
752 CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature.
753 2017;550: 124–127. doi:10.1038/nature24039

754 13. Shaw AE, Hughes J, Gu Q, Behdenna A, Singer JB, Dennis T, et al. Fundamental
755 properties of the mammalian innate immune system revealed by multispecies
756 comparison of type I interferon responses. PLoS Biology. 2017;15: 1–23.
757 doi:10.1371/journal.pbio.2004086

758 14. Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its
759 evolutionary origin. Virus Research. 2003;92: 1–7. doi:10.1016/S0168-1702(02)00309-
760 X

761 15. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of Evolution and Host Gene
762 Mimicry in Influenza and Other RNA Viruses. PLOS Pathogens. 2008;4: e1000079.
763 doi:10.1371/journal.ppat.1000079

764 16. Shackelton LA, Parrish CR, Holmes EC. Evolutionary Basis of Codon Usage and
765 Nucleotide Composition Bias in Vertebrate DNA Viruses. J Mol Evol. 2006;62: 551–
766 563. doi:10.1007/s00239-005-0221-1

767 17. Woolhouse MEJ, Brierley L. Epidemiological characteristics of human-infective RNA
768 viruses. Scientific Data. 2018;5: 180017. doi:10.1038/sdata.2018.17

18. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. 2017. pp. 4765–4774.

19. Baker FB. Stability of Two Hierarchical Grouping Techniques Case 1: Sensitivity to Data Errors. Journal of the American Statistical Association. 1974;69: 440–445. doi:10.2307/2285675

20. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001;29: 1189–1232. doi:10.1214/aos/1013203451

21. Vlasova AN, Diaz A, Damtie D, Xiu L, Toh T-H, Lee JS-Y, et al. Novel Canine Coronavirus Isolated from a Hospitalized Pneumonia Patient, East Malaysia. Clinical Infectious Diseases. 2021 [cited 31 May 2021]. doi:10.1093/cid/ciab456

22. Washburne AD, Crowley DE, Becker DJ, Olival KJ, Taylor M, Munster VJ, et al. Taxonomic patterns in the zoonotic potential of mammalian viruses. PeerJ. 2018;6: e5979. doi:10.7717/peerj.5979

23. Letko M, Seifert SN, Olival KJ, Plowright RK, Munster VJ. Bat-borne virus diversity, spillover and emergence. Nature Reviews Microbiology. 2020;18: 461–471. doi:10.1038/s41579-020-0394-z

24. Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. Nature Microbiology. 2020;5: 562–569. doi:10.1038/s41564-020-0688-y

25. World Health Organization. 2018 Annual review of diseases prioritized under the Research and Development Blueprint. Geneva, Switzerland: World Health Organization; 2018. Available: https://www.who.int/news-room/events/detail/2018/02/06/default-calendar/2018-annual-review-of-diseases-prioritized-under-the-research-anddevelopment-blueprint

26. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. Science. 2008;320: 1784–1787. doi:10.1126/science.1155761

27. Martínez MA, Jordan-Paiz A, Franco S, Nevot M. Synonymous Virus Genome Recoding as a Tool to Impact Viral Fitness. Trends in Microbiology. 2016;24: 134–147. doi:10.1016/j.tim.2015.11.002

28. Morse SS, Mazet JAK, Woolhouse M, Parrish CR, Carroll D, Karesh WB, et al. Prediction and prevention of the next pandemic zoonosis. The Lancet. 2012;380: 1956–1965. doi:10.1016/S0140-6736(12)61684-5

29. Plowright RK, Parrish CR, McCallum H, Hudson PJ, Ko AI, Graham AL, et al. Pathways to zoonotic spillover. Nature Reviews Microbiology. 2017;15: 502–510. doi:10.1038/nrmicro.2017.45

30. Geoghegan JL, Senior AM, Giallonardo FD, Holmes EC. Virological factors that increase the transmissibility of emerging human viruses. PNAS. 2016;113: 4170–4175. doi:10.1073/pnas.1521582113

31. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019;47: D745–D751. doi:10.1093/nar/gky1113

32. Eisenberg E, Levanon EY. Human housekeeping genes , revisited. Trends in Genetics. 2013;29: 569–574. doi:10.1016/j.tig.2013.05.010

33. Millard SP. EnvStats: an R package for environmental statistics. New York: Springer; 2013.

34. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery; 2016. pp. 785–794. doi:10.1145/2939672.2939785

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

36. Kull M, Filho TMS, Flach P. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electron J Statist. 2017;11: 5052–5080. doi:10.1214/17-EJS1338SI

37. Altschul S. The Statistics of Sequence Similarity Scores. In: BLAST documentation [Internet]. [cited 2 Sep 2020]. Available: https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

38. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2: 56–67. doi:10.1038/s42256-019-0138-9

39. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science. 2007;315: 972–976. doi:10.1126/science.1136800

40. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. Bioinformatics. 2011;27: 2463–2464. doi:10.1093/bioinformatics/btr406

41. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster analysis basics and extensions. 2019.

42. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;27: 857. doi:10.2307/2528823

43. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015;31: 3718–3720. doi:10.1093/bioinformatics/btv428

44. Fowlkes EB, Mallows CL. A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association. 1983;78: 553. doi:10.2307/2288117

45. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44: 837. doi:10.2307/2531595

46. Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Processing Letters. 2014;21: 1389–1393. doi:10.1109/LSP.2014.2337313