

1 Vorpal: A novel RNA virus feature- 2 extraction algorithm demonstrated 3 through interpretable genotype-to- 4 phenotype linear models

5
6 Phillip Davis¹, John Bagnoli¹, David Yarmosh¹, Alan Shteyman¹, Lance Presser¹, Sharon
7 Altmann¹, Shelton Bradrick², Joseph A. Russell¹

8
9 1. MRIGlobal – 65 West Watkins Mill Rd., Gaithersburg, MD, USA

10 2. MRIGlobal – 425 Volker Blvd., Kansas City, MO, USA

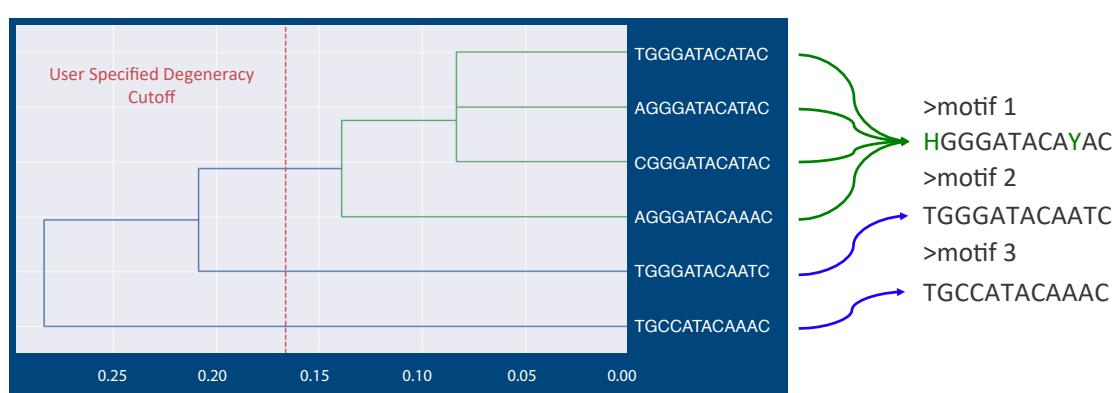
11 SUMMARY

12 In the analysis of genomic sequence data, so-called “alignment free” approaches are often
13 selected for their relative speed compared to alignment-based approaches, especially in the
14 application of distance comparisons and taxonomic classification^{1,2,3,4}. These methods are
15 typically reliant on excising K-length substrings of the input sequence, called K-mers⁵. In
16 the context of machine learning, K-mer based feature vectors have been used in
17 applications ranging from amplicon sequencing classification to predictive modeling for
18 antimicrobial resistance genes^{6,7,8}. This can be seen as an analogy of the “bag-of-words”
19 model successfully employed in natural language processing and computer vision for
20 document and image classification^{9,10}. Feature extraction techniques from natural language
21 processing have previously been analogized to genomics data¹¹; however, the “bag-of-
22 words” approach is brittle in the RNA virus space due to the high intersequence variance
23 and the exact matching requirement of K-mers. To reconcile the simplicity of “bag-of-
24 words” methods with the complications presented by the intrinsic variance of RNA virus
25 space, a method to resolve the fragility of extracted K-mers in a way that faithfully reflects
26 an underlying biological phenomenon was devised. Our algorithm, *Vorpal*, allows the
27 construction of interpretable linear models with clustered, representative ‘degenerate’ K-
28 mers as the input vector and, through regularization, sparse predictors of binary
29 phenotypes as the output. Here, we demonstrate the utility of *Vorpal* by identifying
30 nucleotide-level genomic motif predictors for binary phenotypes in three separate RNA
31 virus clades; human pathogen vs. non-human pathogen in *Orthocoronavirinae*,
32 hemorrhagic fever causing vs. non-hemorrhagic fever causing in *Ebolavirus*, and human-
33 host vs. non-human host in *Influenza A*. The capacity of this approach for *in silico*
34 identification of hypotheses which can be validated by direct experimentation, as well as
35 identification of genomic targets for preemptive biosurveillance of emerging viruses, is
36 discussed. The code is available for download at <https://github.com/mriglobal/vorpal>.

37 38 Feature Extraction Algorithm Overview

39 In the quasispecies model, the virus organism is represented by the “cloud” of genotypes that can
40 be maintained by the virus within the allowable fitness parameters¹². In the method proposed

41 here, the frame of reference for the quasispecies “cloud” is reduced to the level of K-length
42 motifs. In order to estimate the connectedness of these K-mers across the input assemblies, a
43 distance matrix between all of the unique K-mers observed across the designated virus genome
44 assemblies is established using hamming distance. Hierarchical clustering is then performed on
45 the resulting distance matrix using an average linkage function, corresponding to the ultrametric
46 assumption used in Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
47 phylogenies, and flat clusters are extracted using a hyperparameter for the distance cutoff of
48 cluster membership. The constituents of these clusters are then aligned and their positional
49 variants represented using the International Union of Pure and Applied Chemistry (IUPAC)
50 nucleic acid notation with degenerate base symbols. These degenerate motifs are mapped back to
51 their respective assemblies. This approach facilitates interpretation of model features in a
52 functional profiling and hypothesis generating context. To demonstrate the effectiveness of this
53 new feature extraction technique, genotype-to-phenotype linear models were trained on various
54 RNA virus groups. A description of the Python implementation of the algorithm is detailed in
55 Methods and the code is available for download at <https://github.com/mriglobal/vorpal>, along
56 with persistent versions of the models described here-in. A simplified example of the
57 agglomerative clustering step is depicted in Figure 1.
58



59
60 **Figure 1. Hierarchical K-mer Clustering.** A simplified example of K-mer clustering to
61 produce degenerate motifs. After K-mer counting and filtering on frequency, K-mers are
62 clustered using an average linkage function with hamming distance, or positional
63 agreement, as the metric. The resulting alignments, after tree-cutting at a user specified
64 cutoff, are collapsed into their IUPAC character representation.
65

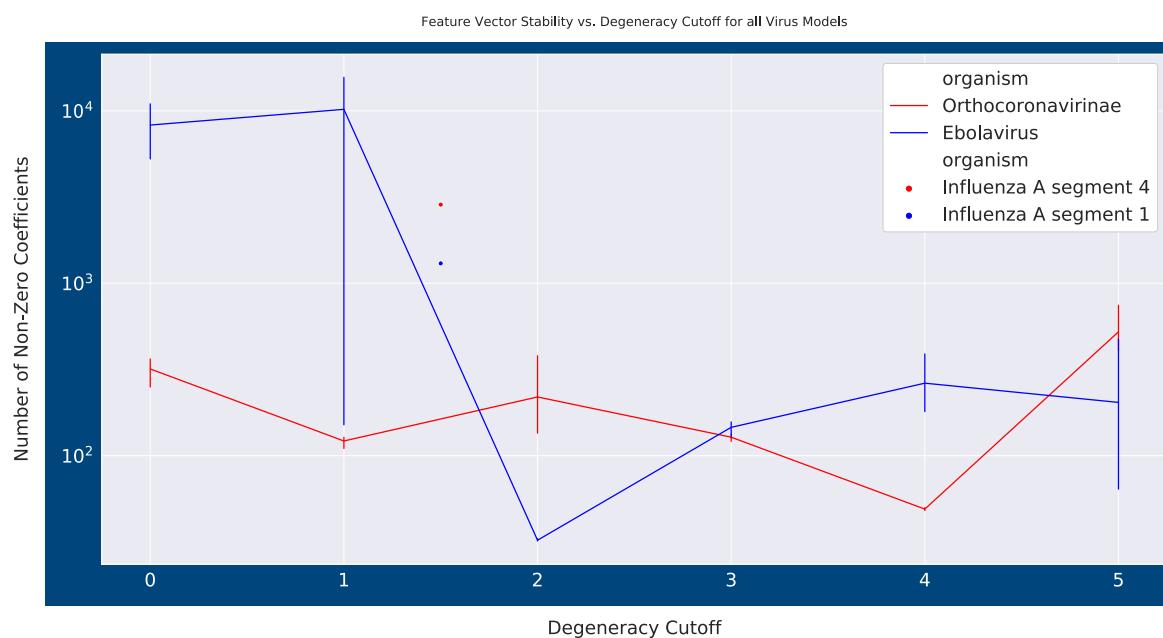
66 By their nature, feature extraction methods make either explicit or implicit hypotheses about
67 what the learner can discover about the data. For instance, in the Natural Language Processing
68 (NLP) domain, the famous “distributional hypothesis” is what forms the theoretical framework
69 for word embedding algorithms such as Word2Vec^{13,14}. The hypothesis central to the Vorpal
70 algorithm makes the following predictions about the types of phenomena that could be learned
71 from RNA virus genomics data, if they are relevant to the output label:
72

1. The predictive motifs are positionally independent
2. The frequency of occurrence of a motif is predictive
3. There are predictive motifs observable only at the nucleic acid level, i.e. in non-coding regions or not observable in the translated product

77 The strongest predictors for the output phenotypes in the models discussed in this paper
78 demonstrate each of these phenomena.
79 Three RNA virus groups were chosen to evaluate the methodology, due to their relevance as
80 important human pathogens – Orthocoronavirinae at the sub-family level, Ebolavirus at the
81 genus level, and Influenza A at the species level. The phenotypes for these virus groups were
82 binary output variables corresponding to human pathogen (vs. non-pathogen), human-
83 hemorrhagic-fever-causing (vs. not human-hemorrhagic-fever-causing), and human-host isolate
84 (vs. non-human-host isolate), respectively. The procedure for labeling these phenotypes is
85 detailed in Methods.
86 This entire algorithm was developed and implemented using Biopython, skbio, and the scipy
87 computing stack contained in the open-source Anaconda Distribution.
88

89 Results

90 Logistic regression models were fit, in triplicate, for the binary phenotypes described above,
91 across different degeneracy cutoffs for the Ebolavirus and Orthocoronavirinae groups. Due to the
92 training time for the Influenza A models (around 72 hours), instead of exploring different
93 degeneracy cutoffs to find the sparsest feature vector, all Influenza A segment models, which
94 were fit independently, were evaluated with a 1.5 degeneracy cutoff for clustering. Model
95 parameter selection for degeneracy cutoff is visualized in Figure 2. All models were highly
96 accurate on both the training and test sets. Selected models are summarized in Table 1.
97 Construction of the training and test sets is described in the Methods section.



98
99 Figure 2. **Degeneracy cutoff parameter search.** Range of feature vector sizes across
100 different degeneracy cutoff levels. Ebolavirus and Orthocoronavirinae find the least
101 number of non-zero coefficients in the weights vector at 2.0 and 4.0 average degeneracy
102 respectively. They also find very high numerical stability at these cutoffs, with repeated
103 fitting returning almost identical motif set membership. Error bars correspond to
104 standard error of the mean.
105

106

Organism	Training Instances (n)	Features (p)	Model degeneracy cutoff	Quantile	Training Set accuracy	Regularization Method	NNzs	Test Set accuracy
Orthocoronavirinae	2278	120444	4.0	.95	1.0	LASSO	48	1.0
Ebolavirus	542	92109	2.0	0.0	1.0	Elastic Net	33	1.0
Influenza A – Segment 1	35184	11435	1.5	.95	.9937	LASSO	1304	.9797
Influenza A – Segment 2	35252	10159	1.5	.95	.9917	LASSO	1330	.9832
Influenza A – Segment 3	35359	9985	1.5	.95	.9969	LASSO	1693	.9769
Influenza A – Segment 4	79882	41285	1.5	.95	.9969	LASSO	2858	.9768
Influenza A – Segment 5	35492	6558	1.5	.95	.9903	LASSO	1104	.9807
Influenza A – Segment 6	57525	27435	1.5	.95	.9897	LASSO	1749	.9833
Influenza A – Segment 7	46343	3489	1.5	.95	.9816	LASSO	997	.9759
Influenza A – Segment 8	36586	5816	1.5	.95	.9836	LASSO	938	.9802

107 **Table 1. Models Summary.** A summary of the attributes for the models built for each RNA
108 virus group that are discussed. NNzs indicate number of non-zero coefficients in the
109 weights vector after regularization.

110

111 Explanatory Modeling through Feature Selection

112 Tables containing the motif identity and corresponding coefficients for the selected models,
113 along with a list of the accession numbers used for training and test sets, are provided as part of
114 the Supplementary materials. We encourage researchers to explore the contents of these models.
115 Below, we analyze a handful of properties of the models to explain their utility in interpretation.
116

117 Orthocoronavirinae

118 The model for the Orthocoronavirinae sub-family was built around the phenotype of human
119 pathogen. The motif with the highest coefficient for the human pathogen phenotype,
120 AKRATGKTGTTAATMAA, is an example of the positional independence phenomena that the
121 Vorpal algorithm could learn if it contains information about the response variable. The motif
122 also appears across both Alphacoronavirus and Betacoronavirus group species that infect
123 humans. Interestingly its pattern of appearance in those groups varies in a way not predicated on
124 this taxonomic organization. In the Alphacoronavirus examples that it appears in, namely 229E
125 and NL63, this motif is located in the same reading frame within the spike S2 glycoprotein
126 protein and encodes a conserved QDVVNQ amino acid sequence. However, when it appears in
127 Severe Acute Respiratory Syndrome (SARS), it remains in the same reading frame, coding for a
128 YNVVNK amino acid sequence, but instead occurs in the polyprotein in the N-terminus of non-
129 structural protein (NSP) 15. The other Betacoronavirus member it appears in, OC43, presents
130 this motif in the same reading frame but it has returned to the spike protein as QDGVNK. This
131 motif serves as a signal for human pathogenicity whose importance is based at least partially on
132 its translation, though the domain itself can appear in completely different protein products. It
133 was also recognized that another positive predictor in the model was a motif related to this one,
134 KGATGTTGTTARWCAAY, offset by a single nucleotide. This related motif sometimes co-
135 occurred at the same position as the one mentioned above, and other times appears at a different
136 position in the genome, which suggests this is part of a larger, repetitive motif.

137 This is summarized in Table 2.

138

139

140

141

Predictor motif	Amino acid motif	COV species	Genome position	Protein Product	Model Coefficient
AKRATGKTGTTAATMAA	YNVVNK	SARS	19569	nsp15	4.54
	QDGVNK	OC43	24096	Spike S2	
	QDVVNQQ	NL63	23514	Spike S2	
	QDVVNQQ	229E	23069	Spike S2	
KGATGTTGTTARWCAAY	FDVVRQC	SARS	10865	nsp5	1.67
	LDVVVKQF	COV JC34	16559	nsp13	
	FDVVRQC	Bat SARS-like	10865	nsp5	
	SDVVKQP	MERS	20064	nsp15	
	QDVVNQQ	NL63	23515	Spike S2	
	QDVVNQQ	229E	23070	Spike S2	
	FDVVRQC	2019-nCoV*	10935	nsp5	

142 **Table 2. Positive Coefficient Coronavirus (COV) motifs of interest.** Organism, genome
 143 locations, and corresponding translated products for selected predictors in the
 144 Orthocoronavirinae model. Bolded examples are instances labeled Non-human-pathogens
 145 in the training set, all others are members of the Human pathogen class. Note: 2019-
 146 nCoV was not part of the training set when these models were developed.
 147

148 Ebolavirus

149 The model for the genus Ebolavirus was specified for a phenotype corresponding to human-
 150 hemorrhagic-fever causing, i.e. the African Ebolavirus constituents, and non-human-
 151 hemorrhagic-fever causing, i.e. Reston ebolavirus (EBOV). The recently discovered Bombali
 152 EBOV, was excluded due to its ambiguity as a human pathogen¹⁵.
 153 The Ebola model demonstrates the utility of the assumption in the Vorpal algorithm that the
 154 feature vector contains information about the frequency of genomic motifs. The preservation of
 155 repeated motifs in the 5' untranslated region (UTR), especially of those in the overlapping UTRs
 156 in the Ebola genome, are the predictors of primary importance in differentiating the phenotypes.
 157 These repeating motifs, or “motif blocks”, and their corresponding coefficients in the model, are
 158 summarized in Table 3 and visualized in Figure 3. These motifs in the 5' UTRs, specifically in
 159 the leading sequence of the L protein, have been previously established as being functionally
 160 important to growth kinetics in cell culture¹⁶. The presence and location of this motif across the
 161 Reston and African constituents of the Ebola genus forms an obvious distinguishing factor. The
 162 contiguous block of overlapping motifs identified in Table 3, appear across all known Ebolavirus
 163 genomes. However, in the Reston version, this block appears only in the 5' UTR of VP40 and L,
 164 which is one of the several genome locations in Reston containing overlapping 3' and 5' UTRs.
 165 When this motif block occurs in the African-derived constituents of the Ebola genus, it appears
 166 in the 5' UTR of VP40, VP30, and L. The VP40 and VP30 5' UTRs are characterized by
 167 overlapping transcriptional units in African ebolaviruses. In *Zaire ebolavirus*, there is an
 168 intergenic region between VP24 and L protein. However, despite the insertion of an intergenic
 169 region at this location in *Zaire ebolavirus*, this representation of the motif block is still preserved.
 170 Comparison of the transcriptional start and stop signals between Reston and Zaire ebolavirus has
 171 been performed before¹⁷, but the conservation of this motif and this pattern of appearance across
 172 the genus has not been established to our knowledge.
 173
 174

Motif	Coefficient
NTGAKGAAGATTAAGAA	0.048876
YGAKGAAGATTAAGAAA	0.061316
GAKGAAGATTAAGAAAAA	0.063464
AKGAAGATTAAGAAAAAA	0.063464
KGAAGATTAAGAAAAAS	0.061316
GAAGATTAAGAAAAASN	0.051440

Table 3. Ebolavirus overlapping UTR “motif block”. Contiguous motifs that form the 5'UTR overlap conserved at varying frequency across the entire Ebola genus. Identical coefficients represent completely colinear predictors.



181
182 **Figure 3. Ebolavirus UTR overlap mapping.** Visual comparison of the UTR overlap motifs
183 specified in Table 3. A) Mapping of motifs on the Zaire Ebolavirus genome. The motifs
184 occur three times in the African constituents of Ebolavirus. B) Mapping of motifs on the
185 Reston Ebolavirus genome. The motifs occur only twice in Reston ebolavirus, with the
186 UTR overlap between VP30 and VP24 replaced by an intergenic spacer.

188 Influenza A

189 The Influenza A model was trained using isolation host as the output variable. As illustrated in
190 Table 1 above, an independent model was built around each segment of Influenza A's genome.
191 Therefore, the model is trying to find signals of host conformational changes on each segment.
192 However, within the constraints of this paper, only results derived from the segment 4 model will
193 be discussed in detail.

194

195 Influenza A's fourth segment contains the HA gene from which Influenza A strains derive their
196 H subtype designation. In the corresponding model, a pattern was observed in the motif
197 distributions that was common to all of the Influenza A segment models examined. This pattern
198 aligns with the third assumption associated with the Vorpal feature extraction method – some
199 degenerate predictors encode only silent mutations. In other words, the signal for the output label
200 is observed only at the nucleotide level for many explanatory variables. For example, one of the
201 highest coefficient predictors for the human-isolate phenotype, GTCTCTACARTGTAGAA,
202 appeared to be related to one of the motifs amongst the most negative predictors,
203 GGTCTYTACARTGTAGA. These motifs correspond to a location towards the end of the C
204 terminus of the HA2 protein, at the location of a conserved, H1-subtype, N-linked glycosylation
205 site following the transmembrane region¹⁸. The pattern of appearance for these motifs is
206 described in Table 4A.

207

Motif	Coefficient	Amino acid sequence	# Human Instances	# Swine Instances	# Avian Instances
-GTCTCTACARTGTAGAA	5.21	SLQCR	12915	828	12*
GGTCTYTACARTGTAGA-	-5.79	SLQCR	13001	1373	12*
GGTCWTTGCAATGCAGA-	N/A	SLQCR	14	759	427

208 **Table 4A. Influenza A HA2 motifs.** Shows three overlapping segments where the addition
209 of a degeneracy allowing for the TTA codon for leucine is an important predictor for the
210 non-human conformation for the H1 subtype. The third motif with no coefficient was
211 identified by looking in avian isolates at the same genomic position. This motif was not
212 used by the model but provides additional interpretation of the phenomenon in effect.
213 The only avian flu examples in the model predictors that these motifs appear in are North
214 American Turkey isolates. No other avian examples of any HA gene subtypes contain
215 these motifs utilizing rare leucine codons. This serine at the beginning of this amino acid
216 sequence is the tail constituent of a N-x-S/T glycosylation motif.

217

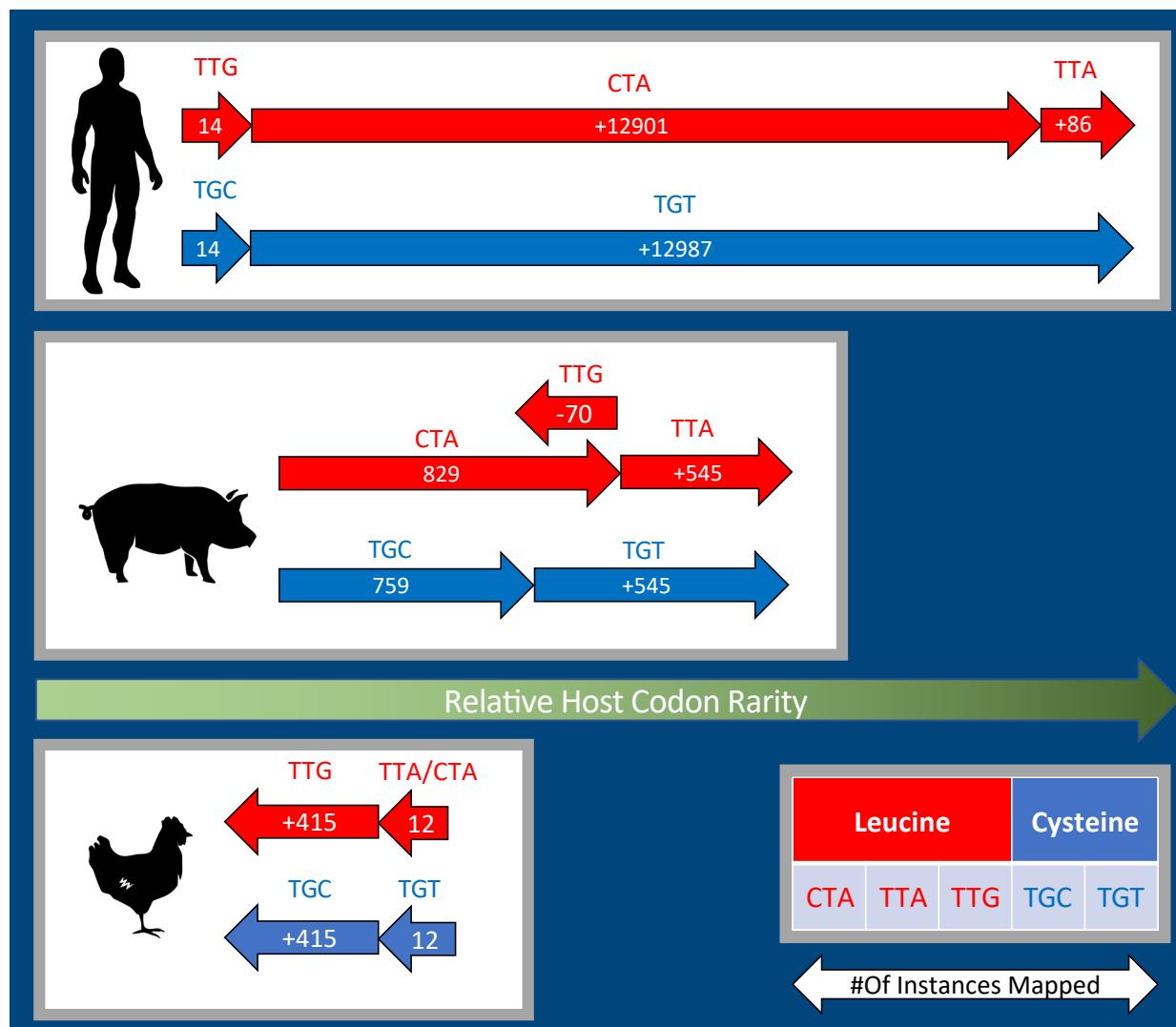
218

Organism	Leucine Codons			Cysteine Codons	
	CTA	TTA	TTG	TGC	TGT
Human	0.07	0.07	0.13	0.55	0.45
Avian	0.06	0.06	0.12	0.6	0.4
Swine	0.13	0.06	0.1	0.61	0.39

219 **Table 4B. Relative Host Codon Frequencies for HA2 motifs.** Shows three overlapping
220 segments where the addition of a degeneracy allowing for the TTA codon for leucine is an
221 important predictor for the non-human conformation for the H1 subtype. The third motif
222 with no coefficient was identified by looking in avian isolates at the same genomic
223 position.

224

225 Examination of the constituent K-mers of these motifs demonstrated that the allowance of the
226 negative predictor to map to the TTA leucine codon introduced, almost exclusively, swine
227 isolates. The conservation of the CTA leucine codon in the human-isolate predictor is
228 noteworthy because this codon is one of the rare leucine codons in the human genome, with a
229 relative frequency of 7%. Alternatively, the TTA codon being more predictive for swine isolates
230 is notable because while TTA also only has a 7% relative abundance in humans, its abundance in
231 pigs is 6% while the CTA codon is less rare (13% relative abundance)¹⁹. This mammalian
232 adaptation separates it almost entirely from any avian examples and there appears to be a fitness
233 gradient. When it appears in mammals, there is a higher incidence of the uncommon leucine
234 codon at this location. As previously mentioned, the SLQCR motif is canonical across all H1
235 subtype examples, including those of chicken and duck. A degenerate motif that mapped to the
236 corresponding position in avian examples was determined to be GGTCWTTGCAATGCAGA.
237 The underlying nucleotide conformations appear to be strictly enforced where the use of the TTG
238 codon for leucine, along with the TGC codon for cysteine, produces 427 avian examples and
239 only 14 human examples. Curiously, the preference for these codons in the avian examples are
240 not correlated with their rarity in those hosts. The TTG codon for leucine has a relative
241 frequency of 13% in mallards, while the CTA and TTA codons are both 6%. A table of the
242 relative codon frequencies by host are noted in Table 4B, and this relationship between motif
243 mapping frequency following a codon rarity gradient in mammals, and the inverse in birds, is
244 visualized in Figure 4.
245



246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262

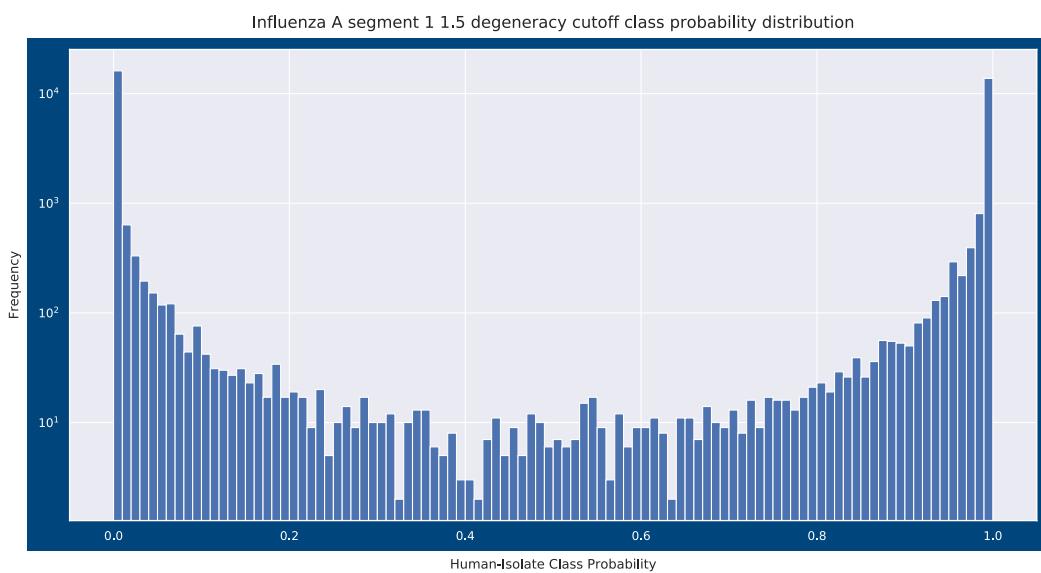
Figure 4. Host Codon Optimizations for H1 subtype. The justification for the coefficients assigned by the model for the motifs specified in Table 4A are demonstrated by the clear role that the TTA codon for leucine plays in increasing the probability of a Swine isolate classification for the H1 subtype. Arrows indicating the increase or decrease in total number of motif mappings point in a direction along the relative host codon frequency gradient where rightwards movement indicates optimization towards the lower frequency rank for the corresponding amino acid. Table 4B shows the relative frequencies for these codons across these animal clades. Magnitude of arrows expressing change in number of reference mappings are not drawn to scale.

The predictor variable with the highest coefficient from the segment 4 model is another, more dramatic example, of the phenomenon described above. The identity of the motif, AATGTRACAGTAACACA, and its translated product, NVTVTH, again demonstrate a preference for rare human codons - in this case, valine. Like the example discussed above, this motif is present almost exclusively in human (N=15913) and swine (N=3885) examples of the H1 subtype. The associated NVTVTH amino acid sequence is also completely conserved across

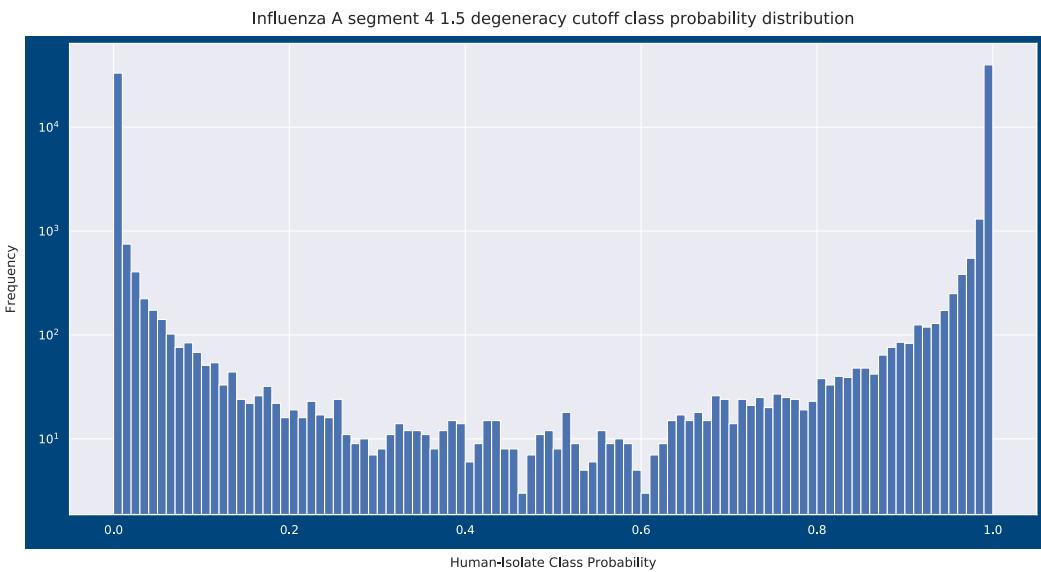
263 all the examples, avian included. The valine codons in the human-isolate versions, almost
264 exclusively GTA, have a relative frequency of only 11% in the human genome. While in the
265 avian examples of segment 4, those codons are switched to GTG, which are the most common
266 Valine codons with a 46% relative frequency in mallards. A motif for the avian version of this
267 was developed using a multiple sequence alignment of the non-human and non-swine isolates of
268 the H1 subtype assemblies in the training set. This motif was established as
269 AAYGTRACYGTGACYCA and mapped back to the training set sequences. When mapped, this
270 new motif was resolved to 480 avian isolates, 33 swine isolates, and nothing else. Unfortunately,
271 unlike the above-mentioned Influenza A motif, the constituent K-mers for this motif were below
272 the quantile cutoff for clustering, and thus, were unable to become a directly observed feature of
273 the model. The use of rare codons, and their tendency to cluster, has been observed across both
274 eukaryotes and prokaryotes²⁰. This NVTVTH amino acid motif is also, like the SLQCR
275 sequence described above, an experimentally validated N-linked glycosylation site on the HA
276 gene in H1N1²¹. Rare-codon clusters in association with N-linked glycosylation sites in human
277 pathogens have previously observed in HIV-1 envelope glycoprotein gp120, where the
278 conservation of the rare-codon RNA sequence conferred increased glycosylation efficiency
279 compared to gp120 mutants²². Codon optimization efforts for lentivirus envelope protein have
280 also induced non-functional proteins, hypothesized to be related to glycosylation disruption²³.
281 The fidelity of conformational change in mammalian isolates to these rare codon identities is
282 extremely high. The oscillation between these conformational states is suggestive of another
283 dimension of interpretation that these logistic regression models offer, outside of the examination
284 of the genomic motifs themselves.
285

286 Other Dimensions of Interpretation

287 The fragility of the phenotype for the Influenza A model resulted in a model with higher
288 complexity than the other RNA viruses studied. However, this provides another avenue for
289 model analysis. Logistic regression classifiers offer not only an output label, but also a
290 probability assignment to the corresponding label. Thus, additional information can be encoded
291 in this output. Figure 5 presents a graphical representation of the distribution of these class
292 probabilities for the training sets for the segments described.



293



294

295 **Figure 5. Influenza A training set class probability histograms.** The class probability
296 distributions for the Influenza A segment 1 and segment 4 models discussed in this
297 section. The frequency is presented in log scale so it can be observed that the vast
298 majority of class predictions belong to the highest and lowest probability bins. We
299 explore the possibility that instances with class probabilities in the middle of the
300 distribution are in transition between host-isolate states as a results of recent zoonosis
301 events.

302

303 The highest coefficient predictor in the model for Influenza A segment 1, which codes for the
304 PB2 polymerase gene, is a motif which represents a mammalian amino acid substitution
305 experimentally observed in a mouse model²⁴. This mammalian adaptation was identified as
306 relevant to the temperature sensitivity of the polymerase in H5N1. The reversion of the avian
307 conformation containing the glutamic acid residue, to the mammalian conformation containing

308 lysine, was observed to be approximately six days. By chance, some subset of viral isolates could
309 have been sampled during this window while “in transit” between host-signature genotypes.
310 Thus, the misclassified examples from the training set invite further scrutiny. Of the 79892
311 instances in the training set, 274 were misclassified, and approximately 10 of these
312 misclassifications were discovered to be mislabeling due to erroneous formatting of the WHO
313 nomenclature. The remainder are examples where the model has, in some cases with a high
314 probability, assigned a classification that disagreed with the class labeling.
315

316 One particularly interesting example of this can be seen in a pair of swine isolates (KM289087.1,
317 KM289089.1) misclassified by the model as human, which were attributed to human-to-pig
318 H1N1 transmission events in backyard farms in Peru²⁵. A third isolate from this study
319 (KM289088.1) was classified correctly but also expressed some ambiguity in the class
320 designation from the perspective of the class probability. Fortunately, this study included in the
321 publication the sampling dates for the pigs at a central processing facility, allowing the Vorpal
322 algorithm to detect a trend in the data as demonstrated in Table 5. Transition from the human
323 conformation of the virus (from the perspective of the model) to the non-human conformation
324 follows the progression of the calendar date. The original authors had previously speculated
325 about the simultaneous exposure of two of these swine isolates based on phylogeny.
326

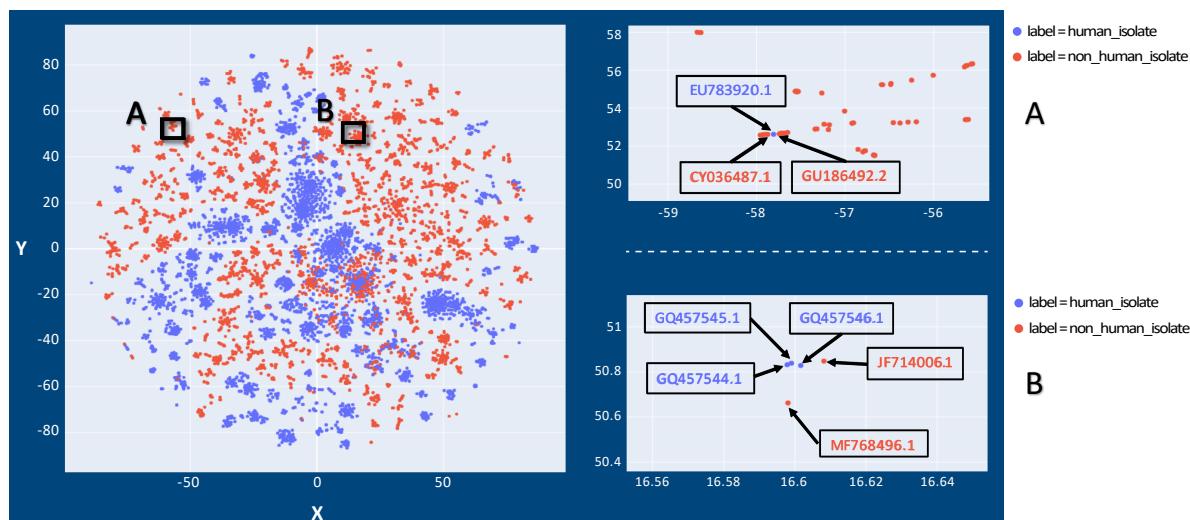
Accession	Host	Human Isolate Class Probability	Sample Date
KM289089.1	Swine	.99	10/15/2009
KM289087.1	Swine	.791	10/17/2009
KM289088.1	Swine	.145	10/19/2009

327 **Table 5. Notable H1N1 Swine Isolates.** Transition from the human conformation of the
328 H1N1 virus to the swine conformation from samples in Peru.
329
330

331 A second case where a training sample was misclassified as a human isolate from the model was
332 an Influenza A H1N1 instance (KF277197.1) isolated from a giant panda at the Conservation and
333 Research Center for the Giant Panda in Ya'an City, China. There are several plausible
334 hypotheses that could explain the consistent misclassification of this isolate from the model,
335 including the most obvious, that the Giant Panda conformation of the virus is only represented by
336 this distinct example, and thus, the model could not learn the features that may distinguish it
337 from a human-isolated example. However, this assembly was accompanied by a publication
338 which points to a different explanation for model confusion. The paper's authors, through
339 phylogenetic analysis, suggest that this case was an example of pandemic H1N1 transmitted
340 directly from humans to the pandas²⁶. Similar examples are abundant. A pair of misclassified
341 swine isolates were identified in a 2009 publication studying triple-reassortment swine Influenza
342 A infections in people from 2005 – 2009²⁸. Both of these human infections were linked to direct
343 contact with sick pigs presented at a county fair within a 3 to 4 day window of sampling. The
344 findings regarding these examples are contained in Tables 6A and 6B. Model prediction
345 probabilities that disagree with the known host source may be useful as a way to infer spill-over
346 events.
347

348 If the misclassified Giant Panda isolate is observed in context in a two-dimensional embedded
349 space, where the motif feature vectors are used as the input space, then its nearest-neighbor in the
350 lower dimensional representation is a human H1N1 isolate, also from Sichuan, in 2009. In the

351 case of the misclassified swine isolates, they are surrounded in the local neighborhood by H1N1
 352 Swine instances from Ohio and Iowa in 2007 and 2008. This proximity in embedded spaces
 353 offers another angle for interpretation, especially in regards to identifying possible spill-over or
 354 re-assortment events and is depicted in Figure 6. Neighbors in the local embedding are often
 355 temporally and geographically proximal, in addition to sharing host isolate membership.
 356 Comingling of class labels in the embedded space potentially offers the opportunity for
 357 identification of zoonosis events.
 358



359
 360 **Figure 6. Embedding of Influenza A segment four train set data.** Two-dimensional t-SNE
 361 embedding of the feature vector for the Influenza A segment four (HA gene). Many
 362 clusters can be observed to segregate with respect to the human isolate class label (blue)
 363 vs. non-human isolates (red). Close inspection of region (A) identifies linkage of H1N1
 364 isolates from swine and humans likely infected from the same swine population, with the
 365 swine-conformation shifting towards a human-conformation. Region A corresponds with
 366 data in Table 6A. Close inspection of region (B) identifies linkage of human-conformation
 367 H1N1 isolates from humans in Sichuan, China with those from pandas believed to have
 368 been infected by direct human contact at a conservation center in the same locale.
 369 Region B corresponds with data in Table 6B. Note: Axes in t-SNE plots have no intrinsic
 370 meaning except to represent pair-wise distances between points.
 371

Accession	Year	Location	Human-Isolate Class Probability	Host	Subtype
FJ986620.1	2007	Ohio	.420	Human	H1N1
FJ986621.1	2007	Ohio	.420	Human	H1N1
EU604589.1	2007	Ohio	.310	Swine	H1N1
HQ833582.1	2007	Ohio	.069	Swine	H1N1
HM461778.1	2008	Ohio	.016	Swine	H1N1
HQ378729.1	2007	Iowa	.010	Swine	H1N1

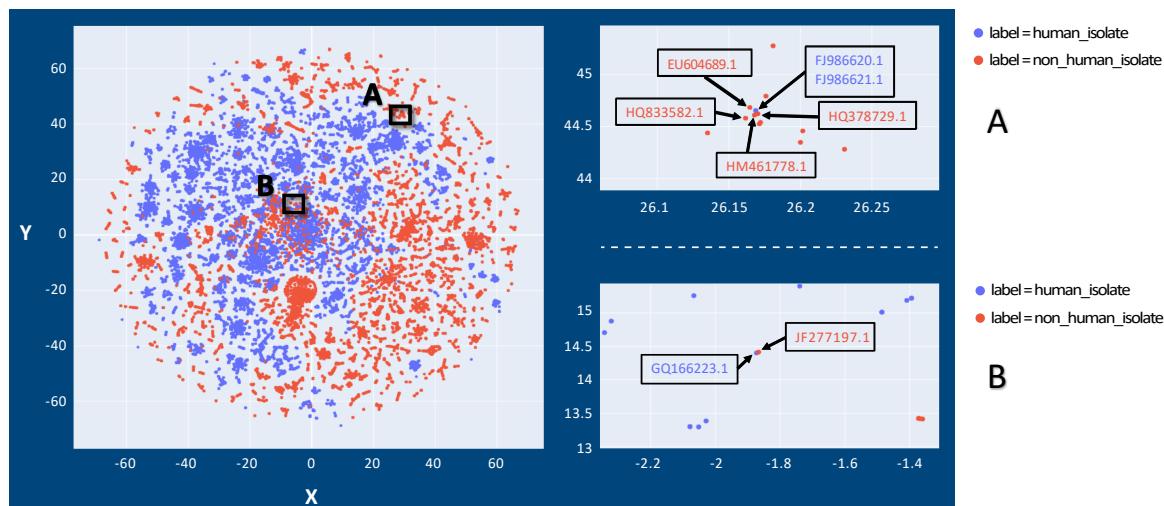
372 **Table 6A. Midwest, US Influenza A segment four isolates.** The local neighbors of
 373 A/Ohio/01/2007 (H1N1) and A/Ohio/02/2007 (H1N1) identified in Shinde et. al. 2009²⁷ as
 374 swine influenza virus infections of human hosts at a county fair in 2007. The estimated
 375 incubation period for these misclassified training examples was 3-4 days.

376

Accession	Year	Location	Human-Isolate Class Probability	Host	Subtype
JF277197.1	2009	Sichuan, China	.980	Giant Panda	H1N1
GQ166223.1	2009	Sichuan, China	.991	Human	H1N1

377 **Table 6B. Sichuan, CN Influenza A segment four isolates.** The Giant Panda isolate
 378 misclassified in the training set and its nearest neighbor in the embedding space.
 379

380 Inspection of the embedded space makes it possible to identify candidate events, even if the
 381 model has not made a classification error. Examples of these are summarized in Figure 7 and
 382 Table 7A and 7B where Influenza A segment 1 (PB2) sequences are embedded into a two-
 383 dimensional field. Further experimentation may also help develop models that incorporate a
 384 velocity to the conformational changes of host-predictor motifs and estimate temporal distance
 385 from a prospective zoonotic event, in a segment-specific manner.



386
 387 **Figure 7. Embedding of Influenza A segment one train set data.** Two-dimensional t-SNE
 388 embedding of the feature vector for the Influenza A segment one (PB2 gene). Many
 389 clusters can be observed to segregate with respect to the human isolate class label (blue)
 390 vs. non-human isolates (red). Close inspection of region (A) identifies linkage of H7N2
 391 isolates from birds and a human likely infected from the same avian population at a live
 392 bird market within the same locality. Region A corresponds with data in Table 7A. Close
 393 inspection of region (B) identifies co-mingling of human-isolated and swine-isolated H1N1
 394 and H3N2 strains from Saskatchewan, Canada. Region B corresponds with data in Table
 395 7B. Note: Axes in t-SNE plots have no intrinsic meaning except to represent pair-wise
 396 distances between points.
 397
 398
 399
 400
 401
 402
 403
 404

Accession	Location	Subtype	Host	Class Probability	Sample Date
EU783920.1	New York	H7N2	Human	0.620	2003
GU186492.2	New York	H7N2	Avian	0.006	2003
CY036487.1	New York	H7N2	Avian	0.013	2003

405 Table 7A. **New York Influenza A segment one isolates.** A human isolate collocated
406 amongst avian isolates in a cluster of H7N2 subtype Influenza examples from New York in
407 2003. The nearest-neighbor for the human isolate (GU186492.2) is an environmental
408 sample from a live-bird market. The model also encodes the ambiguity of the
409 classification in class probability for the human-isolate phenotype.
410

Accession	Location	Subtype	Host	Class Probability	Onset Date	Sample Date
GQ457546.1	Saskatchewan, CA	H1N1	Human	0.742	6/16/09	6/18/09
GQ457544.1	Saskatchewan, CA	H1N1	Human	0.922	6/17/09	6/19/09
GQ457545.1	Saskatchewan, CA	H1N1	Human	0.769	6/15/09	6/18/09
JF714006.1	Saskatchewan, CA	H1N1	Swine	0.396	Unknown	7/20/09
MF768496.1	Saskatchewan, CA	H3N2	Swine	0.001	Unknown	1/15/15

411 Table 7B. **Saskatchewan, CA Influenza A segment one isolates.** A Saskatchewan specific
412 sub-cluster that belongs to a larger cluster of PB2 genes isolated from Swine and co-
413 assortes with H1N1, and H3N2 subtypes in circulation in North America. The human
414 isolates represented in this group belong to pig farm workers who all contracted swine
415 influenza virus, it is presumed, through their place of work²⁷. Interestingly, this
416 distinctive genotype of PB2 seems to be preserved across long time frames (2009 to
417 2015) and is free to re-assort with different Influenza A subtypes (H1N1 and H3N2). In
418 addition, these same isolates had corresponding HA gene sequences published, but the
419 ambiguities seen in the class probabilities for PB2 segment were not observed in the HA
420 gene (i.e. they were all 99% probability human-isolate).
421

422 Discussion

423 The observations presented in this paper represent a fraction of the information potentially
424 contained in the developed models using the Vorpal feature extraction algorithm. Efforts to build
425 robust metanalysis tools based on the model outputs is a focus of further development. While we
426 also think the discoveries mentioned herein make a compelling argument for the power of these
427 models in automatically generating hypotheses to direct experiments, we acknowledge the
428 inherent difficulty in leveraging these models for predictive analytics, where, due to the role of
429 evolution, extrapolation to data unsupported at training time is inevitable.
430

431 To emphasize the hazard of using these models to predict on new data, the emerging Wuhan
432 pneumonia coronavirus and Bombali ebolavirus provide illustrative examples. The Wuhan COV
433 (MN908947.1) and Bombali ebolavirus (NC_039245.1) assemblies were predicted on using the
434 models denoted in Table 1. The model classified Wuhan COV as 0.004% probable for the

435 Human pathogen phenotype and Bombali ebolavirus as 90.2% probable for the Human-
436 hemorrhagic-fever phenotype. Both of these classifications, especially the Wuhan COV
437 designation, are out-of-step with what is known, or in the case of Bombali, suspected, about
438 these viruses. However, it is possible to imagine these functional profiles leading to a more
439 deterministic understanding of function with which to build a predictive frame work.
440 Nonetheless, improvements in data structure and metadata association may yield better abilities
441 to estimate the probability of future events. Certain observations seen in the models thus far may
442 themselves be predictive of the respective phenotype before it is observed, rather than an effect
443 of it already having occurred. The primary example of this is the predictor identified in the
444 Orthocoronavirinae model. As described in the Methods section, certain assumptions were built
445 into labeling for the human-pathogen phenotype that incorporated theories about the zoonotic
446 provenance of SARS and Middle East Respiratory Syndrome-related (MERS) from civets and
447 camels respectively. Observing human-pathogen predictors occurring in SARS and MERS
448 viruses from non-human hosts could suggest the ability to predict the potential of a virus as a
449 human pathogen in advance of a spill-over event. This is observed in the data. The
450 AKRATGKTGTTAATMAA motif appears in all five of the civet SARS assemblies in the
451 dataset. In the case of the camel isolates, the motif KGATGTTGTTARWCAAY, which is also
452 related to the one mentioned above, is another high coefficient predictor for human pathogenicity
453 and it appears in 231 of the 232 Camel-MERS instances in the training set. This motif also
454 appears in the emerging 2019-nCoV as noted in Table 2.
455

456 As for the obstacles for predictive efforts, there are many opportunities for improvements in the
457 collection and annotation of viral genomic data. In Table 1, a slight drift can be observed in the
458 Influenza A model accuracies between the training and test sets. Because the test set represented
459 the most recently isolated viruses, it is attractive to explain this drift as real, i.e. due to evolution.
460 However, there are other factors to control for since the underlying process generating the data
461 has changed over the time period of data collection. The use of cell lines and PCR based
462 amplification of signal for genome assembly, as well as the use of different sequencing
463 technologies suggest other variables to account for. To demonstrated this, a search through the
464 Genbank records for the Influenza A training set members for “passage” annotations revealed
465 that 42.3% of the instances in that set contained such annotations for cell passage. In contrast, the
466 Influenza A test set members, which represents more recently generated data, only contained
467 “passage” annotations in 29.0% of those records.
468

469 Lastly, we hope that this analysis demonstrates that the utility of a Global Virome Project is not
470 ambiguous. Controversy about the value of such a project has been described²⁹ and this thinking
471 has been reflected in policymakers’ decision to end funding to USAID Predict. If recent
472 estimates of mammalian viral diversity hold true³⁰, then marginal increases in monitoring
473 infrastructure combined with new and developing analysis methods, such as Vorpal, might
474 finally deliver the long sought preemptive strategies for emergent diseases, and enable us to
475 more effectively battle those from which we are already suffering.
476

477 Conclusion

478 The use of this algorithm for genotype-to-phenotype models is just one of the potential
479 applications. Automated molecular assay design and degenerate-motif based phylogenetics are
480 examples of the downstream uses already being investigated. The ability to make use of the

481 latent data that is accumulating in databases, as well as novel surveillance data, is made more
482 tangible with this algorithm. Well-curated and richly annotated metadata promises to allow
483 machine learning and other data science techniques to unleash a torrent of discovery in genomics
484 at large. The mantra we are positing for the infectious and emergent diseases surveillance
485 community is “*More data, Better data, Metadata.*” The techniques to unlock the potential of
486 data-driven genomic science are gathering momentum.
487

488 References

- 489 1. Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence
490 comparison: benefits, applications, and tools. *Genome Biology* **18**, (2017).
- 491 2. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using
492 MinHash. *Genome Biology* **17**, (2016).
- 493 3. Koslicki, D. & Falush, D. MetaPalette: a k-mer Painting Approach for Metagenomic
494 Taxonomic Profiling and Quantification of Novel Strain Variation. *mSystems* **1**, (2016).
- 495 4. Déraspe, M. *et al.* Phenetic Comparison of Prokaryotic Genomes Using k-mers.
496 *Molecular Biology and Evolution* **34**, 2716–2729 (2017).
- 497 5. Rizk, G., Lavenier, D. & Chikhi, R. DSK: k-mer counting with very low memory usage.
498 *Bioinformatics* **29**, 652–653 (2013).
- 499 6. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian Classifier for Rapid
500 Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and*
501 *Environmental Microbiology* **73**, 5261–5267 (2007).
- 502 7. Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A. & Sharma, V. K. 16S Classifier: A
503 Tool for Fast and Accurate Taxonomic Classification of 16S rRNA Hypervariable
504 Regions in Metagenomic Datasets. *Plos One* **10**, (2015).
- 505 8. Drouin, A. *et al.* Interpretable genotype-to-phenotype classifiers with performance
506 guarantees. *Scientific Reports* **9**, (2019).
- 507 9. Fei-Fei, L. & Perona, P. A Bayesian Hierarchical Model for Learning Natural Scene
508 Categories. *IEEE CVPR*. (2005)
- 509 10. Pu, W. *et al.* Local Word Bag Model for Text Categorization. *Seventh IEEE International*
510 *Conference on Data Mining (ICDM 2007)* (2007). doi:10.1109/icdm.2007.69
- 511 11. Asgari, E. & Mofrad, M. R. K. Continuous Distributed Representation of Biological
512 Sequences for Deep Proteomics and Genomics. *Plos One* **10**, (2015).
- 513 12. Domingo, E., Sheldon, J. & Perales, C. Viral Quasispecies Evolution. *Microbiology and*
514 *Molecular Biology Reviews* **76**, 159–216 (2012).
- 515 13. Sahlgren, M. The Distributional Hypothesis. *Rivista di Linguistica (Italian Journal of*
516 *Linguistics)* **20**, 33-53 (2008)
- 517 14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations
518 of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs.CL]*. (2013)
- 519 15. Martell, H. J., Masterson, S. G., Mcgreig, J. E., Michaelis, M. & Wass, M. N. Is the
520 Bombali virus pathogenic in humans? *Bioinformatics* **35**, 3553–3558 (2019).
- 521 16. Shabman, R. S. *et al.* An Upstream Open Reading Frame Modulates Ebola Virus
522 Polymerase Translation and Virus Replication. *PLoS Pathogens* **9**, (2013).
- 523 17. Ikegami, T. *et al.* Genome structure of Ebola virus subtype Reston: differences among
524 Ebola subtypes. *Archives of Virology* **146**, 2021–2027 (2001).

- 525 18. Sriwilaijaroen, N. & Suzuki, Y. Molecular basis of the structure and function of H1
526 hemagglutinin of influenza virus. *Proceedings of the Japan Academy, Series B* **88**, 226–
527 249 (2012).
- 528 19. Nakamura, Y. Codon usage tabulated from international DNA sequence databases: status
529 for the year 2000. *Nucleic Acids Research* **28**, 292–292 (2000).
- 530 20. Clarke, T. F. & Clark, P. L. Rare Codons Cluster. *PLoS ONE* **3**, (2008).
- 531 21. Cruz, E., Cain, J., Crossett, B. & Kayser, V. Site-specific glycosylation profile of
532 influenza A (H1N1) hemagglutinin through tandem mass spectrometry. *Human Vaccines
& Immunotherapeutics* **14**, 508–517 (2017).
- 533 22. Ebrahimi, K. H., West, G. M. & Flefil, R. Mass Spectrometry Approach and ELISA
534 Reveal the Effect of Codon Optimization on N-Linked Glycosylation of HIV-1 gp120.
535 *Journal of Proteome Research* **13**, 5801–5811 (2014).
- 536 23. Zucchelli, E. et al. Codon Optimization Leads to Functional Impairment of RD114-TR
537 Envelope Glycoprotein. *Molecular Therapy - Methods & Clinical Development* **4**, 102–
538 114 (2017).
- 539 24. Min, J.-Y. et al. Mammalian Adaptation in the PB2 Gene of Avian H5N1 Influenza
540 Virus. *Journal of Virology* **87**, 10884–10888 (2013).
- 541 25. Tinoco, Y. O. et al. Transmission dynamics of pandemic influenza A(H1N1)pdm09 virus
542 in humans and swine in backyard farms in Tumbes, Peru. *Influenza and Other
543 Respiratory Viruses* **10**, 47–56 (2015).
- 544 26. Li, D. et al. Influenza A(H1N1)pdm09 Virus Infection in Giant Pandas, China. *Emerging
545 Infectious Diseases* **20**, 480–483 (2014).
- 546 27. Shinde, V. et al. Triple-Reassortant Swine Influenza A (H1) in Humans in the United
547 States, 2005–2009. *New England Journal of Medicine* **360**, 2616–2625 (2009).
- 548 28. Bastien, N. et al. Human Infection with a Triple-Reassortant Swine Influenza A(H1N1)
549 Virus Containing the Hemagglutinin and Neuraminidase Genes of Seasonal Influenza
550 Virus. *The Journal of Infectious Diseases* **201**, 1178–1182 (2010).
- 551 29. Jonas, O. & Seifman, R. Do we need a Global Virome Project? *The Lancet Global Health*
552 7, (2019).
- 553 30. Carlson, C. J., Zipfel, C. M., Garnier, R. & Bansal, S. Global estimates of mammalian
554 viral diversity accounting for host sharing. *Nature Ecology & Evolution* **3**, 1070–1075
555 (2019).
- 556

557 Methods

558 Algorithm

559 The Vorpal algorithm for feature extraction was developed using the libraries and versions
560 delineated in the requirements.txt document located on the Github. The Vorpal feature
561 extraction algorithm has 3 steps, each corresponding to a script that becomes the Vorpal
562 workflow.

563 1. kmercountouter_sparse.py

564 a. Input:

- 565 i. a reference genome in FASTA format
566 ii. a folder containing complete assemblies for the viral group of interest
567 iii. a parameter for K-mer size
568 iv. a percent variance argument for filtering out assemblies that are divergent
569 from the reference genome in terms of length

- 571 b. Output:
572 i. a pickled sparse dataframe object containing K-mer counts across every
573 input instance
574 2. hammingclusters_fast.py
575 a. Input:
576 i. A pickled sparse dataframe produced by kmercountouter_sparse.py
577 ii. The average number of allowed degenerate bases for clustering. This is
578 converted to the equivalent hamming distance cutoff by
579
$$distance\ cutoff = \frac{Ave.\ number\ of\ positional\ degeneracies}{K\ length}$$

580 iii. The quantile cutoff for high frequency K-mer filtering
581 iv. The number of chunks to split the count data into when calculating K-mer
582 frequency. This allows for processing of the K-mer counts table in a
583 memory constrained environment (optional)
584 v. A temp folder directory to memory map the distance matrix to, again to
585 allow for more memory overhead to be available at the linkage step.
586 (optional)
587 vi. A memory allocation argument for the development of the distance matrix
588 in chunks. This can be used in conjunction with memory mapping or
589 without it. Uses the sci-kit learn `pairwise_distances_chunked`
590 function instead of the scipy `pdist` function (optional)
591 b. Output:
592 i. A multi-FASTA file with degenerate motifs of K length.
593 3. referencemapping_mp.py
594 a. Input:
595 i. A multi-FASTA with all of the assemblies to map to
596 ii. The multi-FASTA file of degenerate motifs produced by
597 hammingclusters_fast.py
598 iii. A threads argument for parallel processing
599 b. Output:
600 i. A series of BED files with the following column specifications:
601

Chr	Start	End	Name	Score
Accession Number	Start Index	End Index	Motif Identity	$S = \frac{M \text{ instances that motif } i \text{ aligned to}}{\text{Total instances } N} \times 1000$

602
603 Wrapper scripts for reproducing the models with the parameters described below are also
604 provided as `binary_vorpal_model.py` and `binary_vorpal_model_ElasticNet.py`.
605

606 Model Parameters

607 All models were built around binary output variables using a logistic regression classifier. The
608 models were regularized using either ℓ_1 or ElasticNet methods, using the liblinear³¹ solver or
609 Stochastic Gradient Descent estimators^{32,33,34} in scikit-learn, respectively. The parameters
610 evaluated for optimization for both approaches were kept uniform for every model fit, with the
611 parameter values searched over listed in the Table 8.

612

Regularization	Term	Search values	Cross Validation Folds
LASSO	lambda	1.0e ⁻⁴ , 7.742e ⁻⁴ , 5.995e ⁻³ , 4.642e ⁻² , 3.594e ⁻¹ , 2.783, 2.154e ¹ , 1.668e ² , 1.292e ³ , 1.0e ⁴	5
ElasticNet (.15 ℓ_1 ratio)	alpha	1.0e ⁻¹ , 1.0e ⁻² , 1.0e ⁻³ , 1.0e ⁻⁴ , 1.0e ⁻⁵ , 1.0e ⁻⁶	5

613

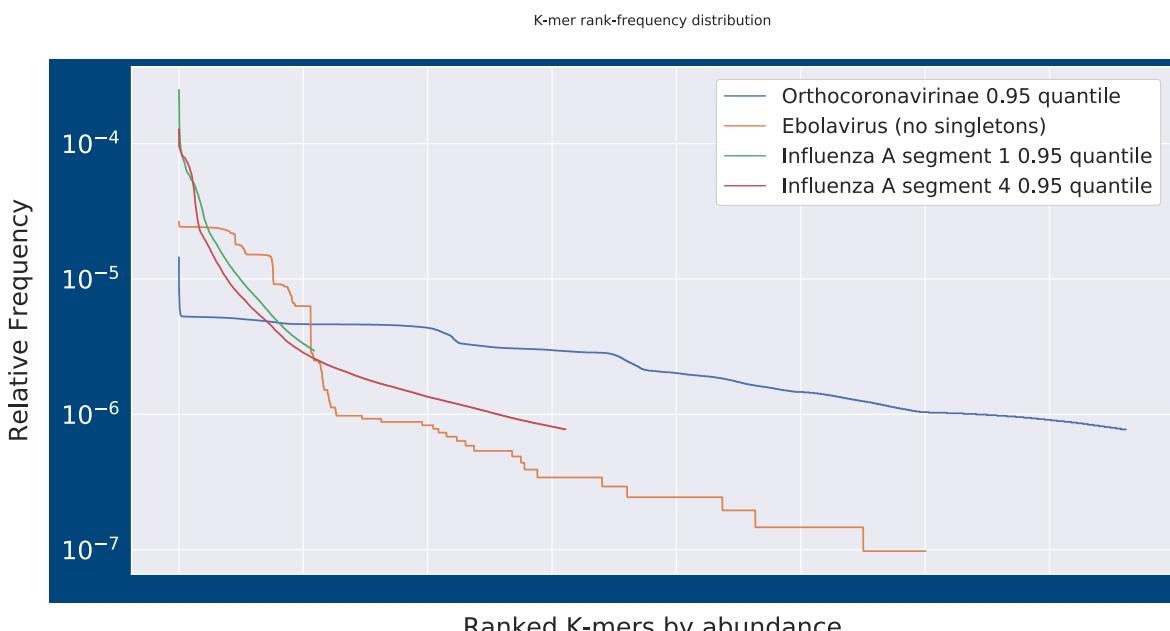
614 Table 8. **Grid Search Parameters.** Optimization search parameters for regularization
615 methods. Lambda in the LASSO method corresponds to the constraint on the ℓ_1 norm of
616 the feature vector while alpha in ElasticNet corresponds to the constraint on the vector
617 magnitude as well as the learning rate for Stochastic Gradient Descent.
618

619 All of the input parameters for feature extraction and the rationale behind the use and tuning of
620 each parameter, and their relation to the corresponding model discussed above is provided here.
621

622 Feature Extraction Parameters

623 The first parameter, K length, can be a variable input, but in the development of these methods
624 was fixed at 17. The decision to set the k-length at 17 had many facets. The first is that the
625 feature space should be large enough, that the introduction of degenerate positions does not
626 cause a complete collapse of feature structures. Evaluation of optimal K length for specific tasks
627 has been performed in many contexts. For phylogenetic representations of viruses, an optimal
628 range of 9 to 13 has been proposed³⁵, for the optimal uniqueness ratio in plant genomes a K
629 length of 20 has been identified³⁶, and in phenetic analysis of bacteria a K length of 31 has been
630 demonstrated to yield the best balance between sensitivity and specificity in intra- and
631 interspecies distance analysis⁴. However, defining a subspace that lends itself to genotype-to-
632 phenotype model interpretability should have the following desiderata:

- 633 1. K-size motifs should map to mostly unique genomic loci. In other words, sparsity in the
634 weights vector is influenced by sparsity in the input vector.
635 2. K-size should be small enough, that the feature space inflation is not catastrophic to
636 memory constraints.



637
638 **Figure 8. Post-quantile filtering K-mer distribution.** K-mer rank-frequency distribution
639 plots for Influenza segments 1 and 4, Ebolavirus, and Orthocoronavirinae at the quantile
640 used in the models discussed. Frequency is calculated as number of instances the ranked
641 K-mer appears in.
642
643 This method implements canonical K-mer counting, where the reverse complement of a K-mer is
644 counted as the same time as the forward oriented K-mer, because of uncertainty about strand
645 orientation in the input data. It was known that there were example assemblies in GenBank for
646 Lassa virus where different instances had inconsistent strand reporting. This assumption seems
647 to be unwarranted for the viruses selected for this study and could be removed for future
648 implementations. It should be pointed out that, while maintaining this assumption seems wasteful
649 from a memory overhead perspective, certain features could only be revealed through this
650 canonical approach, such as hair-pin complements in RNA secondary structures, where the
651 resolution of this structural motif is only possible when compared to the K-mer produced by the
652 complementary region. Other dimensionality reduction techniques, namely high-frequency K-
653 mer filtering, allowing the feature extraction to remain tractable, given the computing resources
654 available for this study. The effect of this canonical approach, and the information it potentially
655 encodes in the feature space, is demonstrated in Table 9A and 9B.
656
657
658
659
660
661
662
663

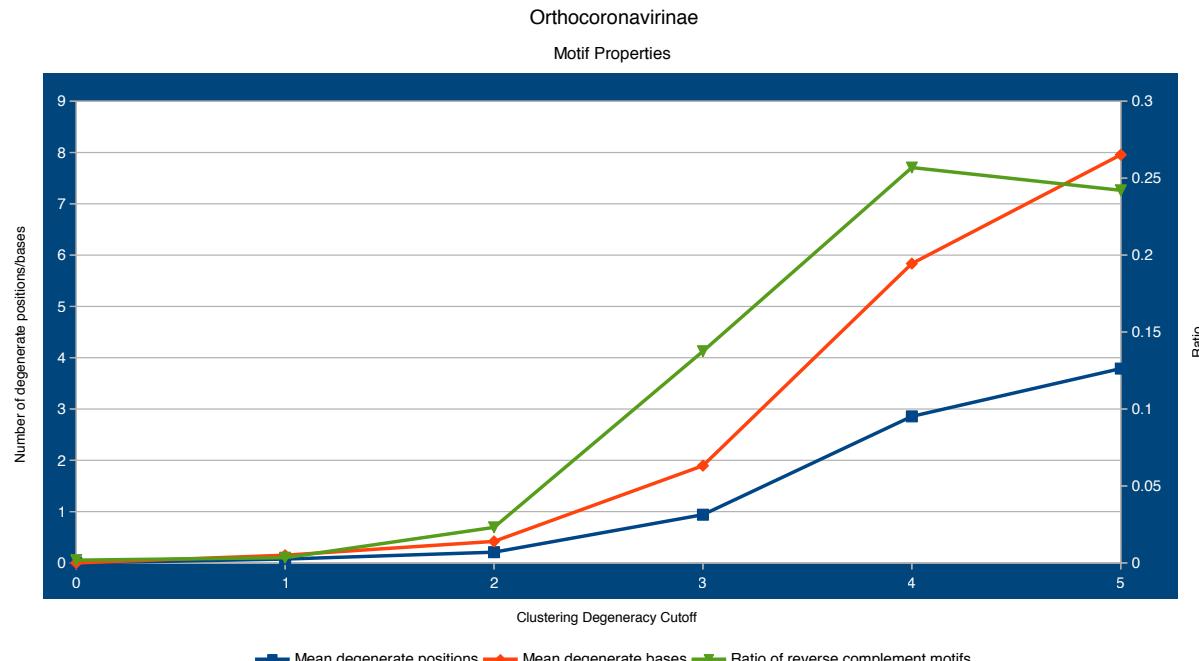
664

Degeneracy Cutoff	# Motifs	Average degenerate positions	Average degenerate bases represented	Feature space	Motif/Feature ratio	Reverse Complement Motif Ratio
0.0	380726	0	0	190363	0.5	0.00177
1.0	354138	0.074744873	0.14982295	177375	0.50086407	0.00339
2.0	331512	0.20877374	0.419245759	167825	0.506241101	0.02309
3.0	266349	0.939684399	1.894649501	146501	0.550033978	0.13751
4.0	183004	2.855790037	5.835227645	120444	0.658149549	0.26588
5.0	150418	3.787837892	7.956893457	106462	0.707774336	0.24210

665

Table 9A. Orthocoronavirinae Feature Extraction Summary (0.95 quantile). Summary statistics for feature extraction for Orthocoronavirinae from 0.0 to 5.0 degeneracy cutoff for clustering. Feature space tracks the dimensionality reduction introduced by degeneracy to motifs that map back to training set. Initially, since no odd-length K-mer can be a reverse complement of itself, canonical K-mers counted compared to those mapped should be half. As degeneracy is introduced, the Motif/Feature ratio is expected to converge to 1.0, which describes a single motif of all "N" symbols. This ratio tracks the amount of previously distinct motifs now represented as a single feature. The final column, shows the phenomena of motifs that are now reverse complements of themselves as a result of degeneracy, contributing to the inflation of the Motif/Feature ratio. Of note in the Orthocoronavirinae features, is while dimensionality reduction continues with the allowance of more degeneracy, the fraction of those resulting features that have corresponding reverse complements in the feature set does not increase past 4.0 degeneracy.

679



680

681

682

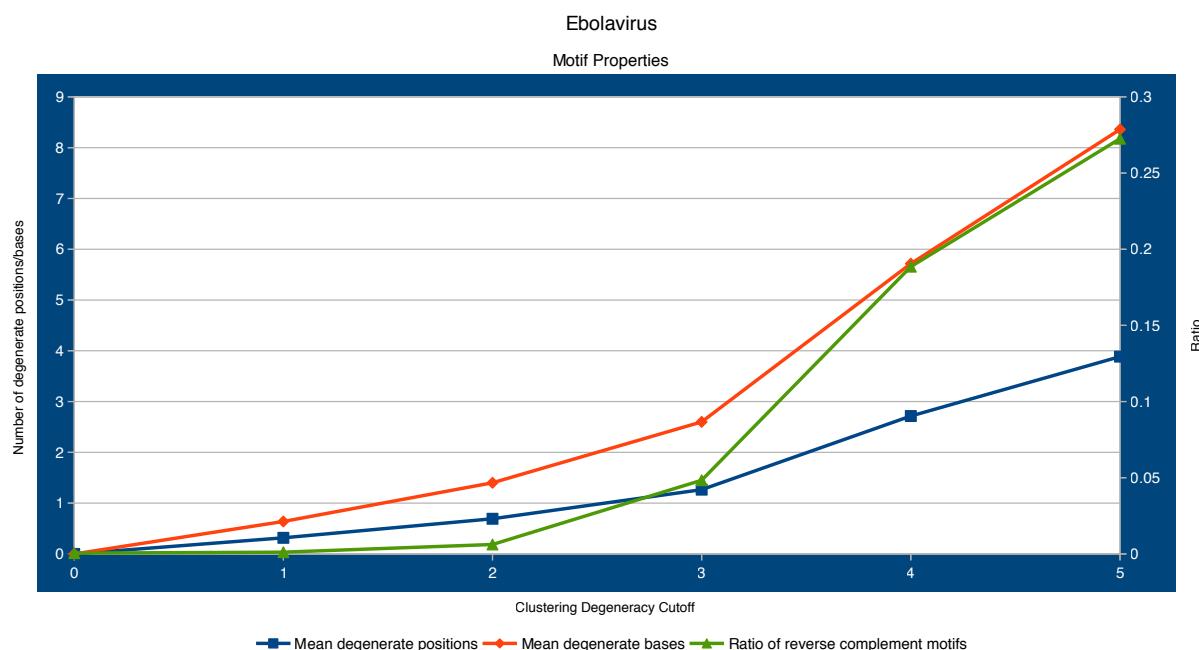
683

Figure 9A. Line plot for three of the selected columns from Table 10A. The plateau reached at 4.0 degeneracy for the ratio of reverse complement motifs is clearly evident.

684

Degeneracy Cutoff	# Motifs	Average degenerate positions	Average degenerate bases represented	Feature space	Motif/Feature ratio	Reverse Complement Motif Ratio
0.0	300196	0	0	150098	0.5	0.00052
1.0	227131	0.316856792	0.638543396	113622	0.500248755	0.00114
2.0	183592	0.692203364	1.401003312	92109	0.501704867	0.00620
3.0	151326	1.265717722	2.599824221	77855	0.514485283	0.04837
4.0	108073	2.715136991	5.716635978	61648	0.570429247	0.18857
5.0	86031	3.884692727	8.358347572	53537	0.622298939	0.27256

685 Table 9B. **Ebolavirus Feature Extraction Summary (0.0 quantile)**. Summary statistics for
686 feature extraction for Ebolavirus from 0.0 to 5.0 degeneracy cutoff for clustering. A
687 larger fraction of the features at the highest degeneracy allowance produced contain
688 corresponding reverse complement motifs in the feature set in the Ebolavirus data than
689 in the Orthocoronavirinae data. This could be attributable to the high frequency K-mer
690 quantile cutoff utilized in the Coronavirus group, or it could allude to generally higher
691 fraction of the Ebolavirus genomes having self-complementation than Coronavirus
692 genomes.



693

694 Figure 9B. **Line plot for three of the selected columns from Table 9b.** The relationship
695 between mean number of degenerate positions in the motifs and the mean number of
696 degenerate bases represented is very similar between Ebolavirus and Coronavirus.
697

698 Applying a filter to the K-mers that are allowed to proceed to the clustering step has two
699 purposes. The first is to denoise the data by removing low abundance features that could be the
700 result of error or other transient sources of variance. The removal of these K-mers is achieved
701 through a parameter specified at the clustering stage, the K-mer quantile. Singletons, or K-mer
702 that are unique to a single instance, are always removed no matter the quantile specified. It was
703

704 discovered that allowing the singletons to form motifs through agglomerative clustering
705 introduced instability into the model parameter estimation (data not shown). Contribution to
706 frequency is determined not by cumulative sum of count across every instance but rather
707 frequency of presence across the sample instances. This is identical to the way the “TopN” score
708 is calculated for K-mers in PriMux primer design software³⁷. Using a K-mer frequency filter
709 selects for a conserved variance signal. This is a reasonable heuristic to introduce, especially for
710 predictive models, where these high-frequency K-mer derived motifs are the those with the
711 presumed highest probability of appearing in a novel example of a related organism in nature.
712 The second function of this feature extraction parameter, made reference to above, is as a
713 dimensionality reduction technique to make K-mer clustering more tractable in the current
714 algorithm implementation, given limitations in computational resources. Memory constraints
715 during the tree building step represents the primary bottleneck with the scipy implementation of
716 the nearest-neighbors chain algorithm for average linkage using $\mathcal{O}(n^2)$ memory^{38,39}.
717 The user specifies an average number of degenerate bases to apply when flat clustering. This
718 number is then divided by the K length specified to estimate the corresponding hamming
719 distance to provide as the max distance for flat clustering. After flat clusters are grouped into
720 alignments and a degenerate motif of the alignment is generated by collapsing each position in
721 the K length alignment into the IUPAC symbol matching the bases seen at that position.
722 This clustering of K-mers, and subsequent representation as degenerate motifs, is another layer
723 of dimensionality reduction similar to lemmatization of words in a Natural Language Processing
724 (NLP) feature extraction technique⁴⁰. Much of this approach could be described as modifications
725 of equivalent NLP feature extraction and modeling strategies. It should be noted however, that
726 data preparation techniques such as term frequency-inverse document frequency (tf-idv), were
727 considered inappropriate to apply in this circumstance for multiple reasons. First, “document”
728 length was invariant in the sense that complete assemblies were the only instances allowed in the
729 training data, and differences in genome sizes within the taxonomies considered were considered
730 irrelevant. Second, document terms, in this case K-mer motifs, that follow a frequency pattern
731 similar to the word “the” in the English language are not present. Additionally, for this reason,
732 the data was not normalized, however to improve convergence speed this could be a future
733 improvement.

734

735 [Phenotype Labeling](#)

736 Phenotype labels for the different organisms modeled were applied using a variety of strategies
737 with some specific assumptions introduced for labeling of the Orthocoronavirinae group. In the
738 cases of Ebolavirus and Coronavirus, taxonomy was used as a guide for phenotype labels, where
739 knowledge about the phenotype of interest was usually easily delineated along taxonomic
740 boundaries. For Influenza A, the World Health Organization nomenclature for Influenza strain
741 identification, which is encoded in the FASTA header, was parsed for labeling of human
742 isolate⁴¹. For those FASTA headers which contained malformed strain identifiers, an ambiguous
743 labeling was applied and removed from the training set.

744

745 The following explicit assumptions were applied when labeling viral instances for the human
746 pathogen phenotype in the Orthocoronavirinae model. First, since most transmissions of Middle
747 East Respiratory Syndrome-related (MERS) betacoronavirus to humans have been zoonic events
748 traced to dromedary camels, all camel isolates for MERS coronavirus were labeled in the
749 positive class corresponding to human pathogen. Likewise, in the cases for Severe Acute

750 Respiratory Syndrome-related betacoronavirus, since the initial outbreak had been theorized to
751 begin from a zoonic event from infected palm civets at a market in Guangdong, China, along
752 with a specific civet spill over event documented in a waitress and a customer in a restaurant in
753 Guangzhou⁴², all civet SARS-like isolates were also labeled as belonging to the positive class.
754 However, since there is no clear evidence of bat-Coronavirus-to-human transmissions, the
755 assumption was built-in that bat isolates of both MERS-like and SARS-like betacoronaviruses
756 were not part of the human pathogen class. In the instance of MERS-like bat isolates, examples
757 have been found across wide geographic ranges, such as South Africa, while human cases appear
758 to restricted to areas where Saudi Arabian dromedary camels are present⁴³ or hospital acquired
759 infections. The same is true of SARS-like bat isolates discovered in caves in China, where
760 assemblies from these isolates show varying similarities to the strain from the 2003-2004
761 outbreak but not the sum of them⁴⁴.

762
763 Training sets were developed from the un-clustered Reference Viral Database⁴⁵ (RVDB) version
764 14 published October 1st, 2018. Accessions for designated taxonomic groups were derived from
765 National Center for Biotechnology Virus⁴⁶ and then used to extract the associated assemblies
766 from RVDB. Test sets were developed from RVDB version 15 published February 6th, 2019
767 using the references for the modeled organisms that had been added between version releases.
768

769 Embedding Visualization

770 The same feature vectors used to produce the logistic regression models were topic modeled
771 similarly to Latent Semantic Analysis⁴⁷ (LSA) using a truncated Singular Value Decomposition
772 (SVD) to a 500 component subspace, which was then subjected to a t-distributed Stochastic
773 Neighbor Embedding^{48,49} (t-SNE) to a two-dimensional space to observe the local structure of
774 the Influenza A viral assemblies. Both of these methods were employed using the associated
775 classes in Scikit-learn. Visualization and exploration of the embedded space was facilitated by
776 Plotly⁵⁰.
777

778 Method References

- 779 31. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A
780 Library for Large Linear Classification. *Journal of Machine Learning Research* **9**, 1871–
781 1874 (2008).
- 782 32. Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings
783 of COMPSTAT2010* 177–186 (2010). doi:10.1007/978-3-7908-2604-3_16
- 784 33. Shalev-Shwartz, S., Singer, Y. & Srebro, N. Pegasos. *Proceedings of the 24th
785 international conference on Machine learning - ICML 07* (2007).
786 doi:10.1145/1273496.1273598
- 787 34. Tsuruoka, Y., Tsujii, J. & Ananiadou, S. Stochastic gradient descent training for L1-
788 regularized log-linear models with cumulative penalty. *Proceedings of the Joint
789 Conference of the 47th Annual Meeting of the ACL and the 4th International Joint
790 Conference on Natural Language Processing of the AFNLP: Volume 1 - ACL-IJCNLP 09*
791 (2009). doi:10.3115/1687878.1687946
- 792 35. Zhang, Q., Jun, S.-R., Leuze, M., Ussery, D. & Nookaew, I. Viral Phylogenomics Using
793 an Alignment-Free Method: A Three-Step Approach to Determine Optimal Length of k-
794 mer. *Scientific Reports* **7**, (2017).

- 795 36. Kurtz, S., Narechania, A., Stein, J. C. & Ware, D. A new method to compute K-mer
796 frequencies and its application to annotate large repetitive plant genomes. *BMC*
797 *Genomics* **9**, 517 (2008).
- 798 37. Hysom, D. A. *et al.* Skip the Alignment: Degenerate, Multiplex Primer and Probe Design
799 Using K-mer Matching Instead of Alignments. *PLoS ONE* **7**, (2012).
- 800 38. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. (2011). doi:
801 arXiv:1109.2378v1 [stat.ML]
- 802 39. Müllner, D. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines
803 for RandPython. *Journal of Statistical Software* **53**, (2013).
- 804 40. May, C., Cotterell, R. & Van Durme, B. An Analysis of Lemmatization on TopicModels
805 of Morphologically Rich Language. (2016). doi: arXiv:1608.03995v2 [cs.CL]
- 806 41. A revision of the system of nomenclature for influenza viruses: a WHO memorandum.
807 *Bulletin of the World Health Organization* **58**, 585–591 (1980).
- 808 42. Wang, M. *et al.* SARS-CoV Infection in a Restaurant from Palm Civet. *Emerging*
809 *Infectious Diseases* **11**, 1860–1865 (2005).
- 810 43. Younan, M., Bornstein, S. & Gluecks, I. V. MERS and the dromedary camel trade
811 between Africa and the Middle East. *Tropical Animal Health and Production* **48**, 1277–
812 1282 (2016).
- 813 44. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides
814 new insights into the origin of SARS coronavirus. *PLOS Pathogens* **13**, (2017).
- 815 45. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A. S. A Reference
816 Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput
817 Sequencing for Novel Virus Detection. *mSphere* **3**, (2018).
- 818 46. Hatcher, E. L. *et al.* Virus Variation Resource – improved response to emergent viral
819 outbreaks. *Nucleic Acids Research* **45**, (2016).
- 820 47. Manning, C. D. in *Introduction to Information Retrieval* 403–419 (Cambridge University
821 Press, 2008).
- 822 48. van der Maaten, L. & Hinton, G. Visualizing High-Dimensional Data Using t-SNE.
823 *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
- 824 49. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *Journal of*
825 *Machine Learning Research* **15**, 3221–3245 (2014).
- 826 50. Plotly Technologies Inc. *Collaborative data science*. 2015. <https://plot.ly>.
- 827
- 828