Data-driven identification of potential Zika virus vectors

Supplement I. Comparison Model Trained on Virus Isolation Data

Michelle V. Evans Tad A. Dallas Barbara A. Han
Courtney C. Murdock John M. Drake

The primary model in Evans et al. (2016) is trained on vector-virus pairs for which the full transmission cycle has been observed. However, many sources, such as the Global Infectious Diseases and Epidemiology Network database (GIDEON), interpret isolation of a virus in wild-caught mosquitoes as evidence of a mosquito's role as vector. In order to investigate the robustness of our findings, we conducted a supplementary analysis in which any evidence for association, including isolation of the virus, is used as the basis for a link in the vector-virus network.

1 Data Collection

As in the primary model, the mosquito-virus pair matrix was constructed based on the Global Infectious Diseases and Epidemiology Network database (GIDEON, 2016), the International Catalog of Arboviruses Including Certain Other Viruses of Vertebrates (ArboCat) (Karabatsos, 1985), *The Encyclopedia of Medical and Veterinary Entomology* (Russell et al., 2013) and Mackenzie et al. (2012). This resulted in a dataset containing 180 mosquito species and 37 viruses, for a total of 334 vector-virus pairs. The vector and virus trait datasets were identical to those used in the primary model (see Supp. II for lists of traits).

2 Predictive Model

We used boosted regression trees (Friedman, 2001) to fit a logistic-like predictive model relating the status of all possible virus-vector pairs (0: not associated, 1:associated) to a predictor matrix comprising the traits of the mosquito and virus traits in each pair. We fit a total of 25 models, applying different training and testing datasets to each, to reduce the dependence dependent on the split between training and testing data. Prior to the analysis of each model, we randomly split the data into training (70%) and test (30%) sets while preserving the proportion of positive labels in each of the training and test sets. Models were trained using the `gbm` package in *R* (Ridgeway, 2015), with the maximum

number of trees set to 25,000 and a learning rate of 0.001. To correct for optimistic bias (Smith et al., 2014), we performed 10-fold cross validation and bagged 50% of the training data for each iteration of the model. These methods are identical to those used to train the primary model. We quantified variable importance by permutation (Breiman, 2001) to assess the relative contribution of virus and vector traits to the propensity for a virus and vector to form a pair. Each of our twenty-five trained models was then used to predict novel mosquito vectors of Zika over the whole virus-vector pair dataset, resulting in twenty-five propensity values assigned to each mosquito species, of which we took the mean. Our prediction dataset, therefore, consisted of the common virus traits of Zika paired with the common traits of all mosquitoes in our flavivirus dataset, for a total of 180 species. The output of this model was a propensity score ranging from 0 to 1. In our case, the final propensity score for each vector was the mean propensity score assigned by the twenty-five models. To label unobserved edges, we thresholded propensity at the value of lowest ranked known vector (Liu et al., 2013).

3 Results

Boosted regression models trained on the weakest evidence of association accurately predicted mosquito vector-virus associations in the test dataset ($AUC = 0.84 \pm 0.02$). When thresholded at the value of the lowest ranked known vector, the model predicted 66 potential vectors of ZIKV, including 42 unknown vectors 1. The majority of predicted vectors were *Aedes* species (39 species), with *Culex* as the second most predicted genus (15 species). It included all but three of the vectors predicted by the main model (*Ae. occidentalis*, *Ru. frontosa*, *Cx. rubinotus*).

4 Model Comparisons

Our supplementary and primary models, trained on virus isolation and above and full transmission cycle, respectively, generally concur. The models are fairly correlated (Spearman's coefficient, $\rho = 0.508$) when considering the propensities of all 180 species 1.However, when only comparing correlation of propensities between those vectors above the threshold of lowest ranked known vector, the models become much more correlated ($\rho = 0.693$). This suggests that our model has a higher sensitivity than specificity, and is better able to predict those vectors that are competent for ZIKV than those that are not. The predictive accuracy of our supplementary model was slightly lower than our primary model. However, this may be an indirect effect of a lower positive-negative label ratio in the dataset used in the primary model, which can artificially inflate AUC values (Lobo et al., 2008).

The models differ in their ability to differentiate between vectors and non-vectors. The distribution of propensities for our main model is more skewed

towards lower propensity values than is the supplementary model 2. This is logical, as the dataset used to train the main model contains a higher proportion of zeros (e.g. vector-virus pairs with no known association) = than the supplementary model. The difference in distributions is accounted for by a similar discrepancy in threshold propensity values based on the lowest ranked known vector. The main model, which has a higher frequency of near-zero propensities, uses a lower threshold value than the supplementary model, however both thresholds qualitatively lie above the majority of the distributions.

5 Conclusion

In summary, our supplementary model predicts which mosquito species may test positive for ZIKV through isolation in wild-caught individuals. As isolation can be understood as evidence of a vector's role in transmission of a disease, our supplementary model may also be interpreted as a ranking of potential vectors of ZIKV, similar to our main model. In fact, both models are well correlated in their ranking of species, although the main model, which trains on fewer vector-virus links, predicts fewer vectors than the supplementary model. Those species predicted by both models, such as *Cx. quinquefasciatus* and *Ae. vexans*, should be prioritized for further research on their competency to transmit ZIKV. Furthermore, as suggested by the main model, the current geographic range at risk for ZIKV transmission in the United States should be expanded to include the range of these species ranked highly by both our main and supplementary models.

References

Breiman, L. 2001. Random Forests. Machine learning **45**:5–32.

Friedman, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics **29**:1189–1232.

GIDEON, 2016. Global Infectious Diseases and Epidemiology Network.

Karabatsos, N. 1985. International Catalog of Arboviruses Including Certain Other Viruses of Vertebrates. The American Journal of Tropical Medicine and Hygiene **27**:372–440.

Liu, C., M. White, and G. Newell. 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. Journal of Biogeography **40**:778–789.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography **17**:145–151.

Mackenzie, J., A. D. T. Barrett, and V. Deubel. 2012. Japanese Encephalitis and West Nile Viruses. Springer Science & Business Media.

Ridgeway, G., 2015. gbm: Generalized Boosted Regression Models.

Russell, R. C., D. Otranto, and R. L. Wall. 2013. The Encyclopedia of Medical and Veterinary Entomology. CABI.

Smith, G. C. S., S. R. Seaman, A. M. Wood, P. Royston, and I. R. White. 2014. Correcting for Optimistic Prediction in Small Data Sets. American Journal of Epidemiology **180**:318–324.

6 Tables

Table 1: Vector Predictions by the Supplementary Model

Vector	**GBM Prediction**	***SD***
Aedes aegypti	0.84	0.06
Aedes albopictus	0.81	0.07
Aedes vittatus	0.76	0.10
Aedes africanus	0.70	0.11
Aedes taylori	0.65	0.14
Aedes furcifer	0.65	0.14
Aedes luteocephalus	0.59	0.12
Aedes metallicus	0.59	0.13
Aedes opok	0.58	0.13
Culex quinquefasciatus	0.56	0.13
Aedes tarsalis	0.56	0.12
Aedes scutellaris	0.56	0.11
Aedes minutus	0.55	0.12
Aedes polynesiensis	0.53	0.11
Mansonia uniformis	0.52	0.12
Aedes fowleri	0.48	0.14
Aedes vexans	0.46	0.11
Aedes dalzieli	0.45	0.13
Culex annulirostris	0.45	0.08
Mansonia africana	0.42	0.12
Psorophora ferox	0.39	0.14
Culex tarsalis	0.38	0.09
Culex tritaeniorhynchus	0.37	0.08
Culex pipiens	0.37	0.13
Culex neavei	0.34	0.06
Aedes vigilax	0.34	0.07
Aedes flavicollis	0.33	0.14
Aedes scapularis	0.31	0.07
Aedes taeniarostris	0.31	0.13
Aedes jamoti	0.31	0.13
Aedes circumluteolus	0.30	0.13
Eretmapodites inornatus	0.30	0.15
Aedes cumminsii	0.29	0.11
Culex vishnui	0.28	0.05
Aedes lineatopennis	0.28	0.11
Aedes neoafricanus	0.27	0.11
Aedes bromeliae	0.26	0.10
Culex guiarti	0.26	0.06
Culex perfuscus	0.26	0.06

Continued on next page

Table 1 – continued from previous page

Vector	GBM Prediction	*SD*
Aedes stokesi	0.26	0.12
Culex telesilla	0.25	0.06
Anopheles gambiae	0.24	0.11
Sabethes chloropterus	0.24	0.11
Aedes hensilli	0.24	0.09
Aedes serratus	0.23	0.06
Aedes chemulpoensis	0.23	0.08
Aedes normanensis	0.23	0.06
Culex bitaeniorhynchus	0.22	0.09
Culex pseudovishnui	0.22	0.05
Aedes argenteopunctatus	0.21	0.06
Wyeomyia vanduzeei	0.21	0.15
Culex p. molestus	0.21	0.06
Culex salinarius	0.20	0.04
Aedes grahami	0.19	0.15
Anopheles coustani	0.19	0.08
Aedes longipalpis	0.18	0.18
Uranotaenia sapphirina	0.17	0.08
Aedes domesticus	0.17	0.06
Aedes abnormalis	0.17	0.06
Aedes natronius	0.17	0.06
Eretmapodites chrysogaster	0.17	0.08
Aedes mcintoshi	0.17	0.06
Aedes ochraceus	0.16	0.06
Culex fatigans	0.16	0.07
Anopheles amictus	0.16	0.06
Eretmapodites quinquevittatus	0.16	0.08

7 Figures

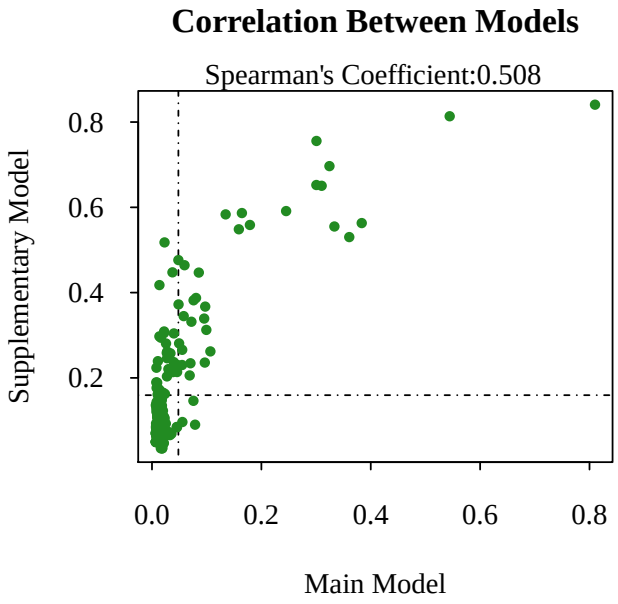

Figure 1: **Propensity values of the main and supplementary models.** Dashed lines represent corresponding threshold values for each model based on lowest ranked known vector propensities.

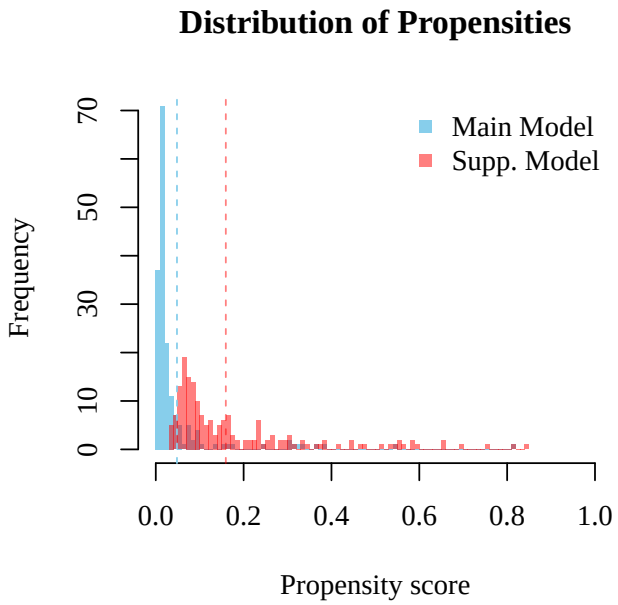

Figure 2: **Distribution of propensity values for the main and supplementary models.** Dashed lines represent corresponding threshold values for each model based on lowest ranked known vector propensities.