Improving Zika Virus sequence validation, classification and annotation using VADR software

EB Dickinson, Eric P. Nawrocki

Computational Biology Branch, Department of Intramural Research, NLM, NIH

Introduction

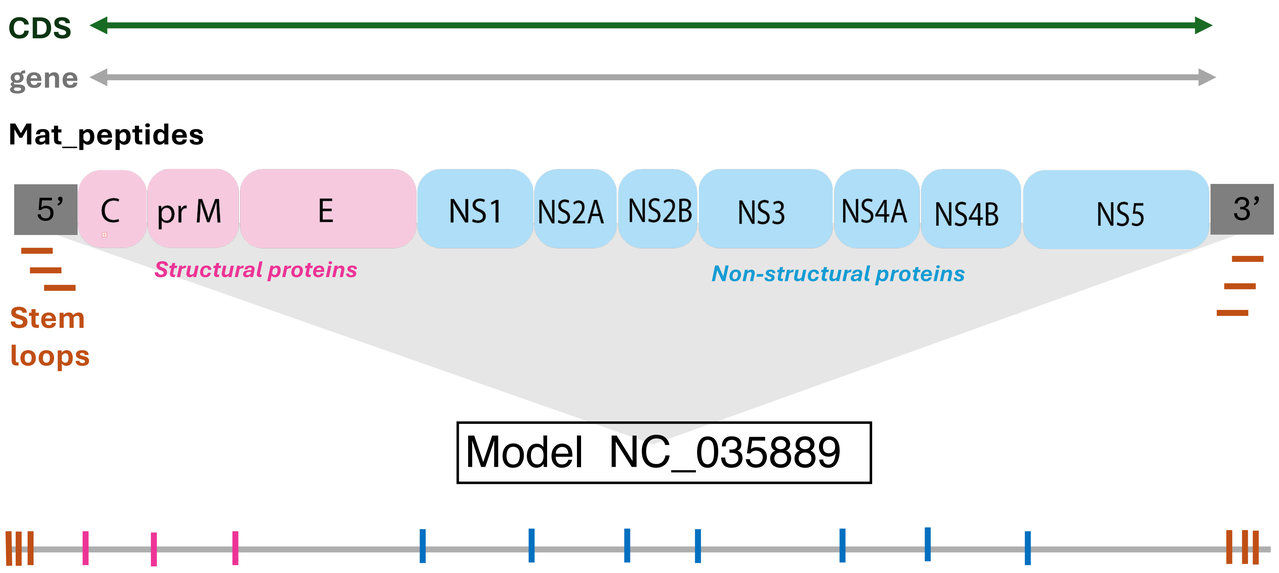
VADR (Viral Annotation DefineR) is a software that classifies and annotates viral sequences based on a Reference Sequence (RefSeq) annotation.

VADR can also check sequences for problems and report alerts before submission to GenBank.

How does VADR work?

VADR builds a reference model along with its features like coding sequences (CDS), gene, mat peptides, and non-coding RNAs.

ZIKV Feature Annotation



V-annotate.pl

Classification

Assigned based on each sequence's highest Hidden Markov Model (HMM) score

	Models		
	AY632535	MG807646	KU509998
OK571913.1	9571.7	11562.4	11659.2
MW123921.1	9412.9	11371.9	11487.7
OK054369.1	9485.3	11459.3	11581.9

Table 1. Each sequence is scored against each model, the highest hmm score means the best likely fit. .

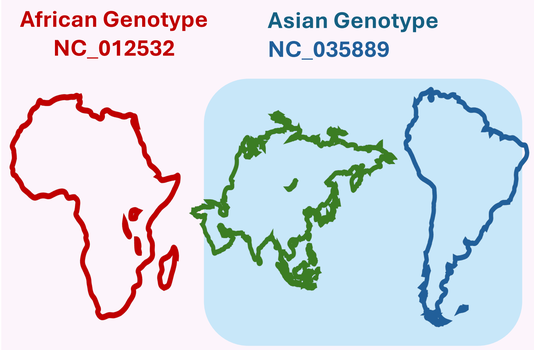
Classifying, Annotating, And Reporting Alerts For ZIKV

Zika virus (ZIKV) is a newly emerged mosquito-borne disease associated with neurological symptoms. ZIKV can be transmitted from mother to fetus, causing severe birth defects such as microcephaly, which can lead to abnormal brain development.

We want to demonstrate that VADR is useful for ZIKV with annotation, classification, and reporting problems (alerts).

Results

Proposing a new RefSeq model for the African sequences



model	subgroup	seqs	pass	fail
NC_012532	**ZA-african**	**52**	**3 (5.8%)**	**49 (94%)**
NC_035889	ZB1-asian	448	378 (84.4%)	70 (15.6%)

Table 2. Out of 500 random ZIKV sequences, 94% of African sequences were failing.

There are 2,575 ZIKV sequences contained within GenBank, and we ran 1,199 non-synthetic and non-patent sequences through our **new African model**, and the **current Asian RefSeq**. To find a candidate model, our criteria was that the model was full length and had the representative frameshift.

model	subgroup	seqs	pass	fail
KY989511	**ZA-african**	**128**	**120 (93.8%)**	**8(6.25%)**
NC_035889	ZB1-asian	1071	1057 (98.96%)	14 (1.3%)

After using the new model, KY989511, we saw an improvement in the African sequences that pass.

Genotyping ZIKV Sequences Using VADR

For ZIKV sequences, and there is a gap in accessible taxonomic tools that can accurately genotype ZIKV. Using a curated and genotyped dataset of 759 ZIKV sequences from Seabra et al. 2022, we selected clade-level phylogenetically-informed VADR models that could accurately sort sequences. We think users will benefit from a three-genotype classification: **African, Asian,** and **American.**

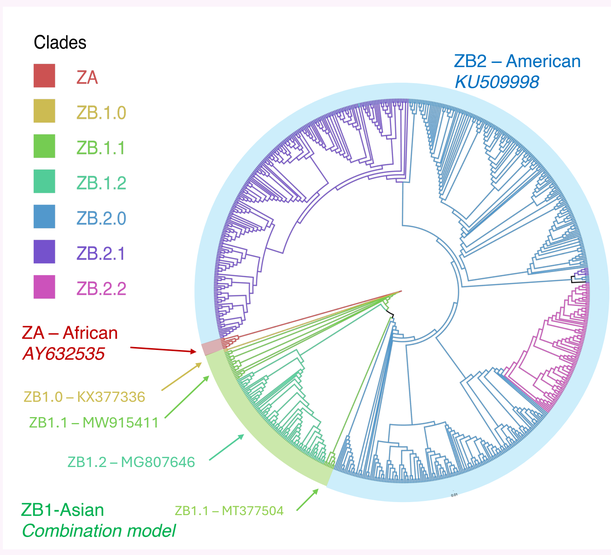


Figure 1. We selected three total models: **ZA-African-*AY632535*, a combination model of four ZB1–Asian sequences** (ZB1.0-*KX377336*, top branch of ZB1.1-*MW915411*, bottom branch of ZB1.1- *MT377504*, and ZB1.2 *MG807646*), and **ZB2 – American, *KU509998*** from the Seabra dataset.

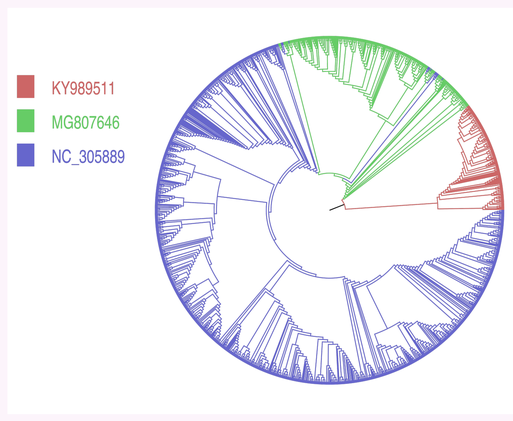


Figure 2. With the curated models, VADR can accurately genotype African (KY989511), Asian (MG807646), and American (NC_035889) groups.

Summary

- We identified **a new African reference model , KY989511**, that better represents the diversity of the ZA-African genotype.
- We found that VADR can **accurately sort ZIKV sequences into African, Asian, and American genotypes** using curated models.

References

Seabra, Sofia G., Pieter J. K. Libin, Kristof Theys, Anna Zhukova, Barney I. Potter, Hanna Nebenzahl-Guimaraes, Alexander E. Gorbalenya, et al. 2022. "Genome-Wide Diversity of Zika Virus: Exploring Spatio-Temporal Dynamics to Guide a New Nomenclature Proposal." *Virus Evolution* 8 (1): veac029.

Acknowledgements

Thank you to Dr. Alvin Crespo-Bellido for generating trees. This research was supported by the Division of Intramural Research (DIR) of the National Library of Medicine (NLM), National Institutes of Health