MosViR - Classification of mosquito viruses in R

Amanda Araújo Serrão de Andrade[1]
Otávio Brustolini[1]
Marco Grivet[2]
Carlos Eduardo Guerra Schrago[3]
Ana Tereza Ribeiro de Vasconcelos[1] (Correspondence: atrv@lncc.br)

[1] LNCC, Laboratory for Scientific Computing - Bioinformatics Laboratory (LABINFO), Rio de Janeiro, Brazil

[2] PUC, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil

[3] UFRJ, Federal University of Rio de Janeiro - Genetics department, Rio de Janeiro, Brazil

Abstract

The MosViR package offers an an easy way to accurately identify complex discriminatory patterns in viral contigs and predict their potential host range The package enhances our ability to capture the diverse genomic landscape of 1) mosquito-associated viruses, 2) Other viruses, 1a) Mosquito-specific viruses, and 1b) Arboviruses.

Introduction to the MosViR package and it's applications

The MosViR package is a computational pipeline designed to analyze metatranscriptomic mosquito data, providing accurate host-based classifications for mosquito-associated viruses from contigs as short as 500 bp. Using multiple predictive models, MosViR classifies mosquito-associated viruses into four distinct categories: Mosquito-associated viruses, Other viruses, Mosquito-specific viruses, and Arboviruses, with high statistical power.

We employ two sequential binary classifications within MosViR. The initial classification step distinguishes between Mosquito-associated viruses and Other viruses, while the subsequent classification step focuses on Arboviruses and Mosquito-specific viruses. A secondary classifier exclusively considers data classified as Mosquito-associated viruses (Figure 1), where positive classes include Mosquito-associated viruses and Arboviruses, while negative classes encompass Mosquito-specific viruses and Other viruses (Figure 1).

The MosViR pipeline demonstrates versatility in classifying mosquito-associated viral sequences, offering several applications:

1) **Exploring the Discriminatory Patterns:** Use MosViR's alignment-free complex discriminatory patterns to explore host-range associations by classifying previously described viruses from various taxonomic organizations and comparing the frequency of specific (1,2)-mers.
2) **Monitoring Host Shifts:** Classify previously described mosquito-associated viruses to monitor potential host shifts, such as transitions from Mosquito-specific viruses to Arboviruses, or even shifts from non-mosquito viruses (e.g., plant viruses) to mosquito-associated viruses.
3) **Classifying Novel Viruses:** Employ the MosViR pipeline to classify unknown viral sequences, providing unique insights into novel viruses and facilitating their classification.

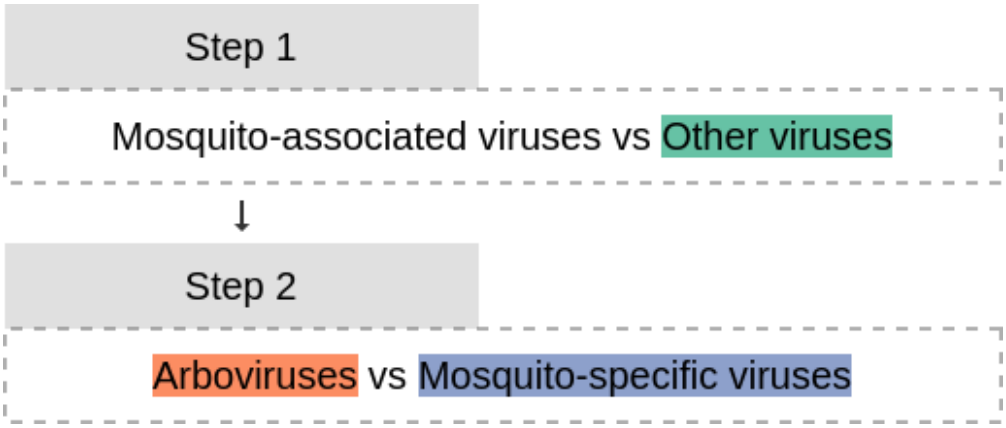


Figure 1: Flowchart illustrating the link between classifiers

MosViR serves as a groundbreaking tool for researchers, enabling the classification of unknown contigs through host-based classification of previously unexplored sequences. This has practical implications, including the identification of novel pathogens, detection of potential host shifts, and guidance for conducting wet lab experiments.

For detailed methodological insights, performance comparisons, and biological implications, please refer to Andrade et al., 2024 (in press).

The MosViR as an R package

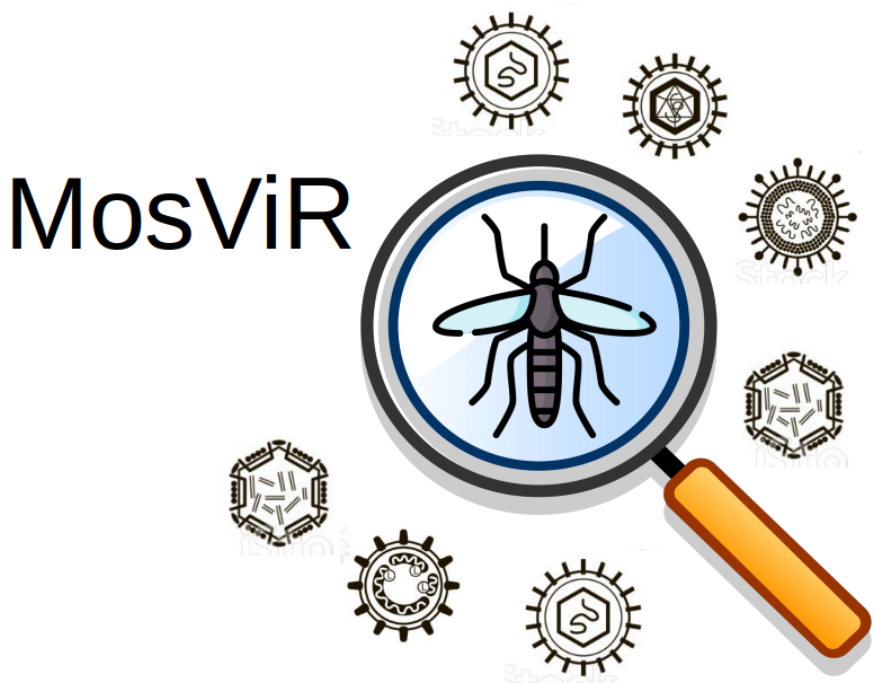
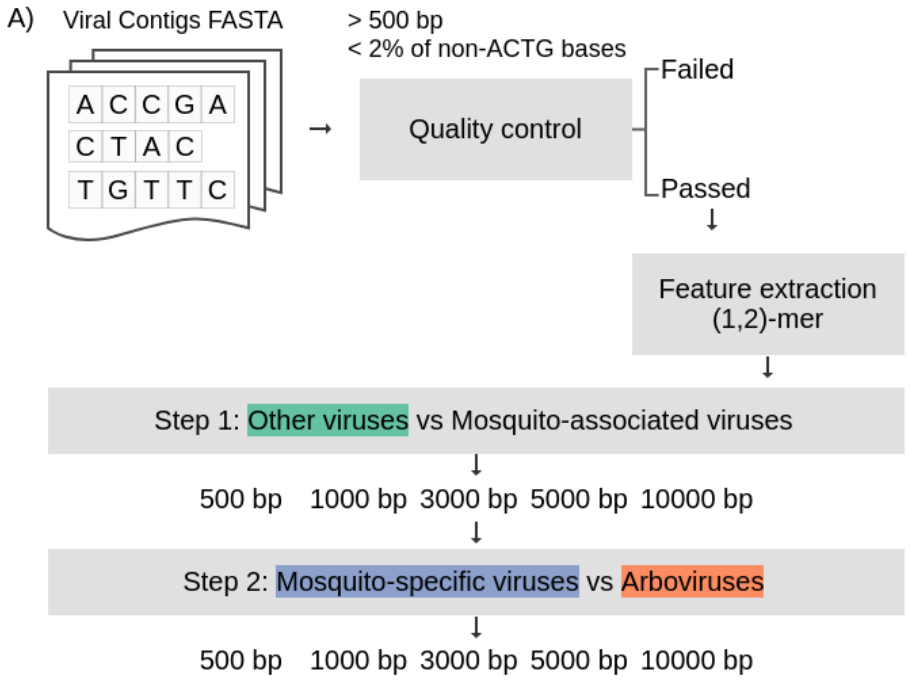


Figure 2: Package logo

The pipeline receives as an input viral contigs, which are obtained by assembling reads from metatranscriptomic sequencing of environmental samples or tissues from the vertebrate host or mosquito vectors. The MosViR

pipeline is described in Figure 3.



B)

Contig ID	Step 1		Step 2	
	Other viruses	Mosquito-associated viruses	Mosquito-specific viruses	Arboviruses
Contig1	0.99	0.01	-	-
Contig2	0.10	0.90	1	0
Contig3	0.92	0.08	0.15	0.85

Figure 3: MosViR pipeline. A) Analytical steps included in the pipeline, B) Example of the pipeline output containing the probability scores

Dependencies

The package needs R(>= 4.2.0) and mnmer(>= 1.0.0).

Instalation

The user should install the package from the Zenodo repository. Please note that the MosViR package has 4,19G due to the multiple predictive models. It may take a while to download.

```
library(devtools)
devtools::install_github("labinfo-lncc/mnmer", ref="main")
devtools::install_url ("https://zenodo.org/records/10950999/files/MosViR_0.99.1.tar.gz")

OR

In the command line:
wget https://zenodo.org/records/10950999/files/MosViR_0.99.1.tar.gz
```

```
In R:
library(devtools)
devtools::install_github("labinfo-lncc/mnmer", ref="main")
install.packages("MosViR_0.99.1.tar.gz", repos = NULL, type="source")
```

The *predict_sequences* function

The main function of our pipeline, *predict_sequences*, is designed to handle a FASTA file containing viral contigs. It anticipates the prior usage of Biostrings for importing FASTA files into the R environment. This function divides the sequences from the DNAStringSet object into segments of varying nucleotide lengths. Subsequently, it uses the mnmer package to generate feature matrices for testing against previously trained predictive models.

Key Features:

Customization: Users can tailor the function by specifying the number of sequences to load, choosing sequences randomly, and setting a probability score cutoff.

The parameters are:

`seqs` = DNAStringObject.

`cutoff` = Adjust the probability score threshold. Higher cutoff values can result in accurate positive classifications (Mosquito-associated viruses and Arboviruses) while inflating the False Negative results. Default 0.50

`all.data` = Select sequences randomly or not. Set TRUE or FALSE

The *predict_sequences* function returns a data frame containing the classification outputs, as well as the probability scores for all sequences.

The *save_fasta* function

The *save_fasta* function filters the output generated by the predict_sequences function and subsequently saves the sequences into a FASTA file. This function empowers users to selectively choose the sequences to retain. For instance, users may opt to save only the sequences classified as Arboviruses.

`seqs` = DNAStringSet object.

`res` = The output from the predict_sequences.

`file` = Name of the FASTA file that will be saved.

`category` = Specifies the classes of sequences to save. Options include All, Mosquito, or Arboviruses.

Work in progress

If you have any queries or find a bug, please submit an issue on GitHub or email atrv@lncc.br.

Credits

This pipeline was developed by Andrade, AAS (aandradebio@gmail.com) and Brustolini, Otávio at the National Laboratory for Scientific Computing - Bioinformatic Laboratory (LABINFO), with contributions from Grivet, Marco., Schrago, Carlos G, and Vasconcelos, ATR.

Andrade AAS, Brustolini O, Grivet M, Schrago C, and Vasconcelos, ATR. Predicting novel mosquito-associated viruses from metatranscriptomic dark matter. (in press)

SessionInfo

```
## R version 4.3.3 (2024-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.6 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3;  LAPACK version 3.9.0
##
## locale:
##  [1] LC_CTYPE=pt_BR.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=pt_BR.UTF-8        LC_COLLATE=pt_BR.UTF-8
##  [5] LC_MONETARY=pt_BR.UTF-8    LC_MESSAGES=pt_BR.UTF-8
##  [7] LC_PAPER=pt_BR.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=pt_BR.UTF-8 LC_IDENTIFICATION=C
##
## time zone: America/Sao_Paulo
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.3.3    fastmap_1.1.1     cli_3.6.1         tools_4.3.3
##  [5] htmltools_0.5.6   rstudioapi_0.15.0 yaml_2.3.7        rmarkdown_2.24
##  [9] knitr_1.43        xfun_0.40         digest_0.6.33     rlang_1.1.1
## [13] evaluate_0.21
```