The Journal of Open Source Software

# ReSVidex: An R package for molecular classification of Respiratory Syncytial (HRSV) Virus sequences

**Marco Cacciabue** [1,2¶], **Nahuel Axel Fenoglio** [3], **Melina Obregón** [3], **Stephanie Goya** [4], **and Mariana Viegas** [5,1]

**1** Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina. **2** Departamento de Ciencias Básicas, Universidad Nacional de Luján, Luján, Argentina. **3** Universidad Nacional de Luján, Luján, Argentina. **4** Department of Laboratory Medicine and Pathology, University of Washington Medical Center, Seattle, WA, USA **5** Laboratorio de Salud Pública, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata, Buenos Aires, Argentina ¶ Corresponding author

**DOI:** 10.xxxxxx/draft

**Software**

- Review
- Repository
- Archive

**Editor:** Open Journals

**Reviewers:**

- @openjournals

**Submitted:** 01 January 1970
**Published:** unpublished



## Summary

## Statement of need

The human respiratory syncytial virus (HRSV) is one of the leading causes of acute lower respiratory tract infection in children, elderly and immunocompromised individuals. Below species level, there are two antigenic groups: HRSV subgroup A (HRSV-A) and B (HRSV-B). Within each subgroup, genotypes are defined based on statistically supported phylogenetic clades that can be inferred with the second hypervariable region (2HR) of the G gene, which encodes the attachment glycoprotein and exhibits the highest genetic and antigenic variability.

Clade classification typically involves analyzing gene sequences from current strains alongside a set of reference sequences using phylogenetic analysis. This process is usually time-consuming and requires specialized training and equipment. Alternatively, advanced machine learning methodologies have demonstrated their ability to provide accurate predictions by employing algorithms capable of uncovering intricate patterns within relevant viral datasets (Cacciabue & Marcone, 2023; Humayun et al., 2021; Wang Y, 2020).

Here we introduce `Resvidex`, an open-source R package (R Core Team, 2023), dedicated to aid researchers in classifying HRSV sequences (full genome or G gene) at the lower levels of resolution in an easy, fast and reproducible way. `Resvidex` is a tool based on alignment-free machine learning for HRSV classification into subtypes and clades. It is fast, sensitive, specific, and ready to implement, as it is available to run locally for R users. It also includes a web application (Chang et al., 2023) that has a user-friendly interface. Additionally, it can be tested on an internet connection without any installation (only for small datasets).

The overall classification algorithm that `Resvidex` uses is divided into three majors steps. In the initial phase, the user data is loaded in a multifasta format, and the k-mer counting operation is executed utilizing the k-mer package (Wilkinson, 2018). Each count of k-mers undergoes normalization based on both the k-mer size ($k = 6$) and the length of the sequence. Alternatively, the user can copy and paste the query sequence directly to the app. In the second step, the predict function from the ranger package (Wright & Ziegler, 2015) is invoked using a pre-trained random forest model. It calculates a probability score through a majority vote rule. Using this score, the application determines the classification score for each query sequence. Additionally,the app also calculates the proportion of N bases in the genome and the genome length. These values are important as divergencies from the expected values can impact notably over the classification results. On the final step, sequences are separated in two tables,

one showing the sequences that passed all the quality checks and another with sequences that did not pass at least one of the filter steps. These filters ensure that each sequence achieves a probability score of 0.4 or higher, that the sequence length aligns closely with the expected length for the classification model (with a tolerance of up to 50%), and that the proportion of ambiguous bases (N) in the sequence does not exceed 2% of the genome length. Sequences that do not meet the necessary criteria should be analyzed manually with other methodologies (i.e. alignment-dependent tools) that may shield a more robust result. Although not recommended, the app allows the user to manually tweak these filters. Additionally, a concise report can be generated, incorporating the results table, date of analysis, and model information.

`Resvidex` was designed to be used by researchers who want to classify their samples of HRSV according to the Goya et. al. proposal (Goya et al., 2024). It comes with two classification models: one for whole genome sequences (FULL_GENOME for sequence length of approximately 15000 bp) and other for the G coding sequence (G for approximately 900 bp). The HRSV classification comprises 41 clades or genetic groups: 25 for subgroup A and 16 for subgroup B.

## Examples

A few vignettes are available, these include: How to use the shiny app vignette, step-by-step explanation of a in-built example, and another example with a larger dataset.

## Acknowledgements

## References


Cacciabue, M., & Marcone, D. N. (2023). INFINITy: A fast machine learning-based application for human influenza A and B virus subtyping. *Influenza and Other Respiratory Viruses*, *17*(1), e13096. https://doi.org/10.1111/irv.13096

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *Shiny: Web application framework for r*. https://CRAN.R-project.org/package=shiny

Goya, S., Ruis, C., Neher, R. A., Meijer, A., Aziz, A., Hinrichs, A. S., Von Gottberg, A., Roemer, C., Amoako, D. G., Acuña, D., McBroome, J., Otieno, J. R., Bhiman, J. N., Everatt, J., Muñoz-Escalante, J. C., Ramaekers, K., Duggan, K., Presser, L. D., Urbanska, L., … Viegas, M. (2024). *The unified proposal for classification of human respiratory syncytial virus below the subgroup level* [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2024.02.13.24302237

Humayun, F., Khan, F., Fawad, N., Shamas, S., Fazal, S., Khan, A., Ali, A., Farhan, A., & Wei, D.-Q. (2021). Computational Method for Classification of Avian Influenza A Virus Using DNA Sequence Information and Physicochemical Properties. *Frontiers in Genetics*, *12*, 599321. https://doi.org/10.3389/fgene.2021.599321

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Wang Y, D. J., Bao J. (2020, January). Rapid detection and prediction of influenza a subtype using deep convolutional neural network based ensemble learning. *Proceedings of the 2024 14th International Conference on Bioscience, Biochemistry and Bioinformatics*. ISBN: 9798400716768

Wilkinson, S. (2018). Kmer: An r package for fast alignment-free clustering of biological sequences. In *GitHub repository*. https://doi.org/10.5281/zenodo.1227690

Wright, M. N., & Ziegler, A. (2015). *Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R*. https://doi.org/10.48550/ARXIV.1508.04409