




1 Resvidex: An R package for molecular classification of 2 Respiratory Syncytial Virus (HRSV) sequences

3 Marco Cacciabue ^{1,2}¶, Nahuel Axel Fenoglio ³, Melina Obregón ³,
4 Stephanie Goya ⁴, María Inés Gismondi ^{1,2}¶, and Mariana Viegas ^{5,1}

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

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

10 Summary

11 Resvidex aims to facilitate the classification of human respiratory syncytial virus (HRSV)
12 sequences at the lower levels of resolution. It can handle both whole genome and partial
13 sequences (three classification models). Resvidex comes with its own shiny app for an
14 user-friendly option.

15 Statement of need

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

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

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

35 The overall classification algorithm that Resvidex uses is divided into three majors steps.
36 In the initial phase, the user data is loaded in a multifasta format, and the k-mer counting
37 operation is executed utilizing the k-mer package ([Wilkinson, 2018](#)). Each count of k-mers
38 undergoes normalization based on both the k-mer size ($k = 6$) and the length of the sequence.
39 Alternatively, the user can copy and paste the query sequence directly to the app. In the second
40 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 three classification models: one for whole genome sequences ("FULL_GENOME", 15000 nt), one for sequences that cover the G coding region ("G", 900 nt) and one for sequences that cover the G+F coding region ("G_F", 2800 nt). The HRSV classification comprises 41 clades or genetic groups: 25 for subgroup A and 16 for subgroup B.

Examples

The main functions of resvidex are the following:

- `kcounter()` : count and normalize the k-mers present in each sequence.
- `prediction_caller()` : perform the classification based on the pretrained classification model.
- `quality_control()` and `quality_filter()` : add the corresponding quality FLAGS.

Additionally, `classify()` acts like a wrapper function, enabling the handling of all the above functions in one simple step, for example:

```
#load the library
library(resvidex)
```

```
# In this example, we use a test file provided with the package.
```

```
file_path<-system.file("extdata","test_dataset.fasta",package="resvidex")
```

```
# Use the wrapper function. You can change the classification model and pass other argum
classify(inputFile=file_path,model=FULL_GENOME)
```

Alternatively, the user can fire up the resvidex shiny app using the `run_shiny_app()` function.

Other examples are available as vignettes: 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*,

- 77 17(1), e13096. <https://doi.org/10.1111/irv.13096>
- 78 Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson,
79 J., Dipert, A., & Borges, B. (2023). *Shiny: Web application framework for r*. [https://](https://CRAN.R-project.org/package=shiny)
80 CRAN.R-project.org/package=shiny
- 81 Goya, S., Ruis, C., Neher, R. A., Meijer, A., Aziz, A., Hinrichs, A. S., Von Gottberg, A.,
82 Roemer, C., Amoako, D. G., Acuña, D., McBroome, J., Otieno, J. R., Bhiman, J. N.,
83 Everatt, J., Muñoz-Escalante, J. C., Ramaekers, K., Duggan, K., Presser, L. D., Urbanska,
84 L., ... Viegas, M. (2024). *The unified proposal for classification of human respiratory*
85 *syncytial virus below the subgroup level* [Preprint]. *Infectious Diseases (except HIV/AIDS)*.
86 <https://doi.org/10.1101/2024.02.13.24302237>
- 87 Humayun, F., Khan, F., Fawad, N., Shamas, S., Fazal, S., Khan, A., Ali, A., Farhan, A., &
88 Wei, D.-Q. (2021). Computational Method for Classification of Avian Influenza A Virus
89 Using DNA Sequence Information and Physicochemical Properties. *Frontiers in Genetics*,
90 12, 599321. <https://doi.org/10.3389/fgene.2021.599321>
- 91 R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation
92 for Statistical Computing. <https://www.r-project.org/>
- 93 Wang Y, D. J., Bao J. (2020, January). Rapid detection and prediction of influenza a
94 subtype using deep convolutional neural network based ensemble learning. *Proceedings of*
95 *the 2024 14th International Conference on Bioscience, Biochemistry and Bioinformatics*.
96 ISBN: 9798400716768
- 97 Wilkinson, S. (2018). Kmer: An r package for fast alignment-free clustering of biological
98 sequences. In *GitHub repository*. <https://doi.org/10.5281/zenodo.1227690>
- 99 Wright, M. N., & Ziegler, A. (2015). *Ranger: A Fast Implementation of Random Forests for*
100 *High Dimensional Data in C++ and R*. <https://doi.org/10.48550/ARXIV.1508.04409>