RSV-GenoScan: An automated pipeline for whole-genome human respiratory syncytial virus (RSV) sequence analysis

Alexandre Dosbaa, Romane Guilbaud, Anna-Maria Franco Yusti, Valentine Marie Ferré, Charlotte Charpentier, Diane Descamps, Quentin Le Hingrat, Romain Coppée

▶ **To cite this version:**

Alexandre Dosbaa, Romane Guilbaud, Anna-Maria Franco Yusti, Valentine Marie Ferré, Charlotte Charpentier, et al.. RSV-GenoScan: An automated pipeline for whole-genome human respiratory syncytial virus (RSV) sequence analysis. Journal of Virological Methods, 2024, 327, pp.114938. 10.1016/j.jviromet.2024.114938 . hal-04600904

HAL Id: hal-04600904
https://u-paris.hal.science/hal-04600904v1

Submitted on 4 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Journal of Virological Methods 327 (2024) 114938

Contents lists available at ScienceDirect

Journal of Virological Methods

journal homepage: www.elsevier.com/locate/jviromet

Protocols

RSV-GenoScan: An automated pipeline for whole-genome human respiratory syncytial virus (RSV) sequence analysis

Alexandre Dosbaa [a], Romane Guilbaud [a,b], Anna-Maria Franco Yusti [a], Valentine Marie Ferré [a,b], Charlotte Charpentier [a,b], Diane Descamps [a,b], Quentin Le Hingrat [a,b], Romain Coppée [a,1,*]

[a] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, IAME, Paris F-75018, France
[b] Service de Virologie, AP-HP, Hôpital Bichat – Claude Bernard, Paris F-75018, France

ARTICLE INFO

Keywords:
Respiratory syncytial virus
Software
Bioinformatics
Monoclonal antibody resistance
Next-generation sequencing

ABSTRACT

Background: Advances in high-throughput sequencing (HTS) technologies and reductions in sequencing costs have revolutionised the study of genomics and molecular biology by making whole-genome sequencing (WGS) accessible to many laboratories. However, the analysis of WGS data requires significant computational effort, which is the major drawback in implementing WGS as a routine laboratory technique.
Objective: Automated pipelines have been developed to overcome this issue, but they do not exist for all organisms. This is the case for human respiratory syncytial virus (RSV), which is a leading cause of lower respiratory tract infections in infants, the elderly, and immunocompromised adults.
Results: We present RSV-GenoScan, a fast and easy-to-use pipeline for WGS analysis of RSV generated by HTS on Illumina or Nanopore platforms. RSV-GenoScan automates the WGS analysis steps directly from the raw sequence data. The pipeline filters the sequence data, maps the reads to the RSV reference genomes, generates a consensus sequence, identifies the RSV subgroup, and lists amino acid mutations, insertions and deletions in the F and G viral genes. This enables the rapid identification of mutations in these coding genes that are known to confer resistance to monoclonal antibodies.
Availability: RSV-GenoScan is freely available at https://github.com/AlexandreD-bio/RSV-GenoScan.

1. Introduction

The use of high-throughput sequencing (HTS) technologies and the decreasing cost of sequencing have transformed the way viruses are studied, making whole-genome sequencing (WGS) accessible to many laboratories. The ability to routinely generate viral genome sequences has many applications, including clinical and laboratory strain and mutant analysis, pathogen surveillance and outbreak detection (Houldcroft et al., 2017). The latest HTS platforms enable the generation of whole viral genome sequences in just a few hours, providing a detailed understanding of the genomic content of the target virus that can inform public health decisions in various ways (Beerenwinkel et al., 2012; Stockdale et al., 2022). Despite the increasing accessibility and affordability of WGS for laboratories, analysing the large amount of data generated requires advanced computational skills that can be difficult and time-consuming to learn for microbiologists or clinicians. Therefore, WGS analysis software are needed to facilitate genomic investigations for specific viruses.

Currently, there is no automated tool to simplify genomics analyses of the human respiratory syncytial virus (RSV), which is a leading cause of lower respiratory tract infections in infants, the elderly, and immunocompromised adults (Falsey et al., 2005; Shi et al., 2017). RSV is a non-segmented, single-stranded, negative-sense RNA human orthopneumovirus of ~15.2 kb in length, encoding 11 viral proteins. The two main proteins of clinical interest are the G and F surface glycoproteins, which mediate viral entry and are major targets of the human immune response (Levine et al., 1987; Connors et al., 1991; Yin et al., 2006; McLellan et al., 2013). RSV infections are classified into two co-circulating groups, named A and B, that diverged about 350 years ago (Mufson et al., 1985), and each of them is further divided into several subgroups. Until now, the therapeutic armamentarium was very limited. Ribavirin was the only approved drug, and it only has a modest effect

* Corresponding author.
E-mail address: romain.coppee@univ-rouen.fr (R. Coppée).
[1] Current affiliation: Université de Rouen Normandie, Laboratoire de parasitologie-mycologie, UR 7510 ESCAPE, Centre Hospitalier Universitaire de Rouen, F-76000 Rouen, France.

https://doi.org/10.1016/j.jviromet.2024.114938
Received 28 November 2023; Received in revised form 17 March 2024; Accepted 5 April 2024
Available online 6 April 2024
0166-0934/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

when used to treat severe RSV infections. In 2023, the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have approved vaccines targeting the F protein in its pre-fusion form for pregnant women and adults over 60 years of age. In addition, a new monoclonal antibody, nirsevimab, was approved for infants under two years of age. These therapies aim to prevent severe lower respiratory tract infections in two vulnerable populations: neonates and the elderly. However, the use of new treatments may put selective pressure on RSV, potentially leading to the selection of resistance mutations, as seen with SARS-CoV-2 after the widespread use of vaccines. Consequently, monitoring the evolution of the RSV genome, and in particular the G and F glycoproteins, is essential to track novel resistance mutations to therapeutic solutions. Several techniques are being developed to sequence the entire RSV genome, which will help to track the emergence and spread of such mutations.

Here we present RSV-GenoScan, an easy-to-use pipeline for WGS analysis of RSV sequences generated by HTS on Illumina or Nanopore platforms. Designed primarily for microbiologists and clinicians, RSV-GenoScan automates the steps required for WGS analysis directly from the raw sequence data. The pipeline first filters the sequence data, then maps the reads to the RSV reference genomes and outputs genome coverage and depth statistics. It then generates the fasta consensus sequence and identifies the RSV subgroup using a phylogenetic approach. The pipeline also provides a complete list of mutations on the F and G glycoproteins – including those already known to confer resistance to monoclonal antibodies. We demonstrate the performances and usefulness of this software using two RSV sequence datasets generated with Illumina or Nanopore platforms.

2. Methods

RSV-GenoScan is a pipeline for WGS analysis of RSV from raw sequences obtained from Illumina or Nanopore platforms (Fig. 1). RSV-GenoScan requires third-party software (Supplementary Table S1), but can be easily installed using a script provided by us. Once installed, RSV-GenoScan prompts for the fastq files (either single-end or paired-end reads) to be deposited in a specified folder. The algorithm then aligns the reads to two RSV reference genomes (GenBank accession: NC_001803.1, group A; AY353550.1, group B) using the bwa mem algorithm (default parameters) for Illumina data (Li and Durbin, 2009) or Minimap2 (option: -ax map-ont) for Nanopore data (Li, 2018). Aligned reads are then sorted and indexed using SAMtools (Danecek et al., 2021), then a pileup file containing information on matches, mismatches and indels is generated using the *mpileup* function of SAMtools (parameters: -a -B). The pileup is then used as input to Python scripts to perform a variety of analyses. The first step is to identify the RSV group for each sample by calculating which of the two genomes (RSV group A or B) is best covered. The whole-genome sequence in fasta is then generated if at least 70% of the genome is covered with at least 10 reads. All coverage and depth statistics (especially for F and G glycoproteins) are summarised in a table and in a graph showing the number of reads along the genome (Fig. 1). As F and G proteins are important therapeutic targets, their consensus sequences are also generated in separate files, even if the whole-genome does not meet the coverage and depth requirements. The genome sequences are then compared with previously published sequences belonging to different RSV subgroups to determine the most likely subgroup (Fig. 1 and Supplementary Table S2) according to a phylogenetic approach using IQ-TREE (Nguyen et al., 2015). We used the nomenclature proposed by Goya and collaborators (Goya et al., 2020), which to our knowledge is the most accurate RSV classification. Of note, this database can be enriched with additional sequences by the user before the analysis. Finally, the pipeline reports all mutations detected in F and G glycoproteins (Fig. 1). Mutations known to confer reduced susceptibility to a monoclonal antibody are indicated (note that

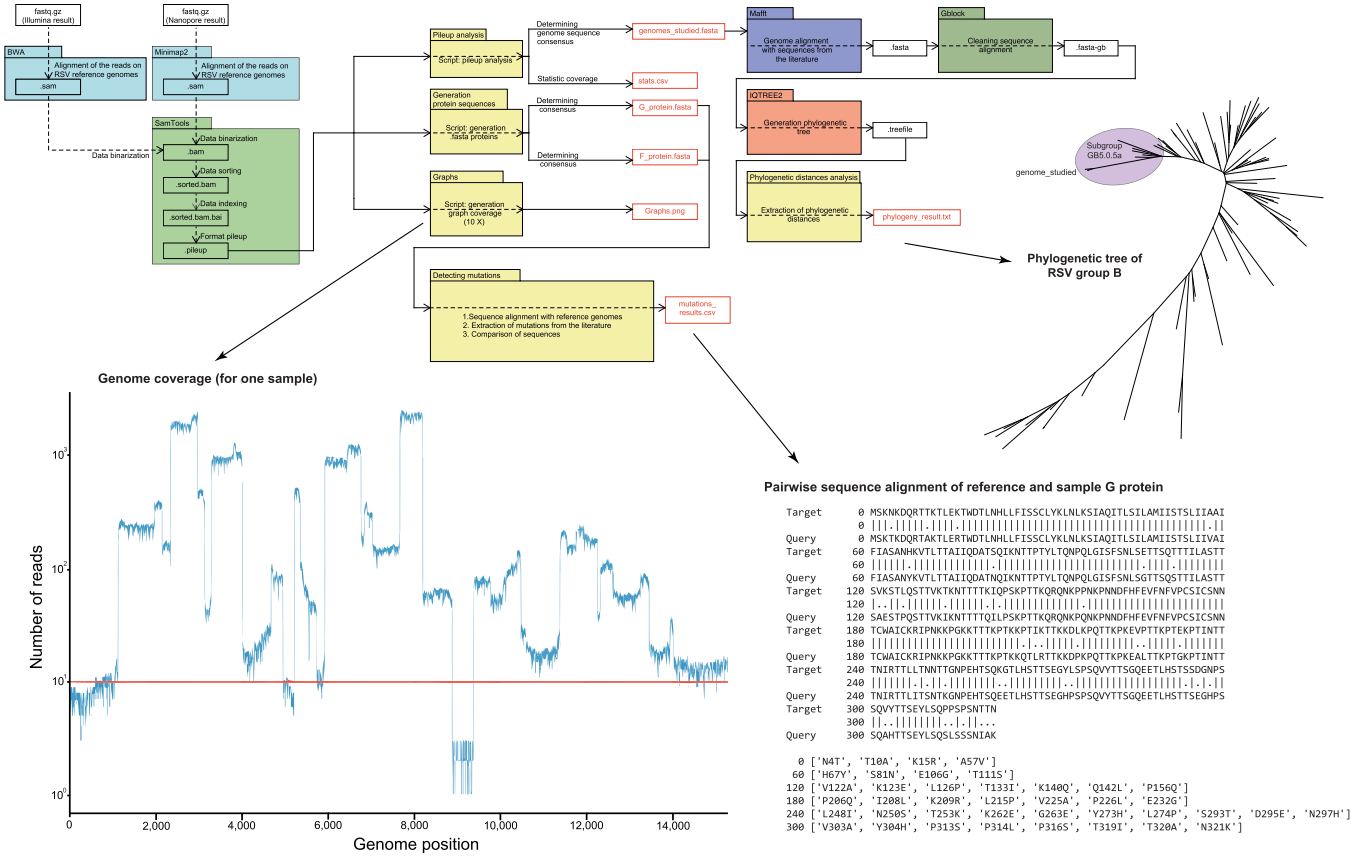

Fig. 1. – RSV-GenoScan workflow.

this list will be updated periodically) (Zhu et al., 2018; Wilkins et al., 2023). All files generated during the WGS analysis are stored in different directories according to the different steps of the analysis. A video tutorial was made to easily install and use RSV-GenoScan: https://youtu.be/8LQlHOGjkfI.

RSV-GenoScan was developed under Linux Ubuntu 20.04 system and Python v.3.4.3 and can be run on any computer with a Linux system. In this study, a machine consisting of 32 Go random access memory (RAM) and an Intel Core I7–6820HQ was used. The RSV-GenoScan software and instructions for its use can be found at https://github.com/AlexandreD-bio/RSV-GenoScan. To demonstrate the usefulness of RSV-GenoScan, the software was tested with two different datasets, with RSV sequencing performed using either Illumina or Oxford Nanopore technology (list of accession numbers for fastq files in Supplementary Table S3).

3. Results

RSV-GenoScan was initially tested on RSVs detected in 14 nasopharyngeal swabs from immunocompromised patients diagnosed at the Foch Hospital (Suresnes, France) between January 1 and March 31, 2021. The viral genome sequences were generated using a hybrid capture-based approach and a Miseq system (Illumina, Inc.) in a paired-end run (2 ×150 bp). The full procedure is described elsewhere (Coppée et al., 2022). Raw fastq reads were submitted directly to the pipeline without further bioinformatic processing. The analysis was completed in 18 minutes on 8-core CPUs. The whole-genome sequence was generated for all samples, with a genome coverage at 10× ranging from 99.48% to 99.89% and a median coverage depth between 545 and 7054× (Supplementary Table S3). Ten samples belonged to group A and were classified as subgroup GA2.3.5, and four samples belonged to group B and were classified as subgroup GB5.0.5a. The sequences of the F and G glycoproteins were fully interpretable for all samples. From eight to nine mutations were found in the F glycoprotein, including the A103V, L172Q and S173L mutations, which were present in all group B samples but are not predicted to confer resistance to either palivizumab or nirsevimab.

We then applied RSV-GenoScan to HTS reads obtained from 46 clinical specimens (40 nasopharyngeal swabs and 6 bronchoalveolar lavages collected during the 2022–2023 and 2023–2024 winter seasons) that were positive for RSV using the FilmArray RP2.1plus assay (BioMerieux). All specimens were obtained from patients hospitalized at the Bichat Claude-Bernard University Hospital (Paris, France). DNA extraction, library preparation and quality control are described in Supplementary Method S1. Viral genome sequences were generated on a GridION system (Nanopore Oxford Technology) using a Flowcell R9.4.1 (for the samples collected in the 2022–2023 winter season; n = 25) or R10.4.1 (for the samples collected in the 2023–2024 winter season; n = 21). A negative control (water only) was also included as a negative control. After sequencing, the raw fastq reads were submitted to the pipeline without any treatment. The analysis was completed in 37 minutes on 8-core CPUs. The whole-genome sequence of RSV was generated for 42/46 samples, with a genome coverage at 10× ranging from 35.90% to 100% and a median coverage depth between 7 and 3886× (Supplementary Table S3). The negative control had only 54 reads mapping to the RSV genome and was logically uninterpretable. Out of the 42 samples, 10 samples belonged to group A and were classified as subgroup GA2.3.5. The 32 remaining samples belonged to group B and were classified as subgroup GB5.0.5a. For the samples collected during the 2022–2023 winter season, the F glycoprotein sequence was complete in 24/25 samples and contained between 8 and 15 mutations compared to the reference sequence. All the samples had the mutations A103V+L172Q+S173L. Three of these samples had the additional mutation I206M, while 21 of them had the additional mutations I206M+Q209R. Regarding the samples from the 2023–2024 winter season, all group B samples carried the A103V+L172Q+S173L mutations. They also harboured the I206M+Q209R mutations, except one that had only the I206M mutation. As these mutations are not associated with resistance, all samples were predicted to be sensitive to nirsevimab and palivizumab.

Overall, using reads obtained either with the Illumina or Oxford Nanopore Technology platforms, our pipeline was able *i*) to identify the RSV group and subgroup; *ii*) to generate the consensus sequence for the whole-genome and for the F and G glycoproteins; *iii*) to list all mutations within the F and G glycoproteins; and *iv*) to predict the sensitivity to drugs within a few minutes.

4. Discussion

HTS has transformed several fields, including virology. The development of rapid microbial genome sequencing, coupled with a steady decline in sequencing prices, has led to its implementation in many microbiology laboratories and it may become the gold standard in a few years (Houldcroft et al., 2017). However, several drawbacks limit the use of WGS as a routine laboratory technique. Despite the increasing availability of sequencing data, its processing and interpretation requires scientists trained in bioinformatics. To overcome this bottleneck, easy-to-use software and pipelines are being developed to facilitate genomic analyses by individuals without bioinformatics training.

Here, we have developed RSV-GenoScan, an open-source bioinformatics pipeline for direct analysis of RSV raw reads obtained with the Illumina or Nanopore HTS platforms. The software generates whole-genome RSV sequences, summarises coverage and depth statistics, identifies subgroups, and lists all mutations present in the F and G glycoproteins. All result files are generated in a tabular format that can be easily viewed and manipulated in any spreadsheet software. The tool can be run using simple commands and could be a method of choice for researchers without a strong bioinformatics background. The data obtained can be used to guide public health decisions by monitoring the evolution of RSV genomes in a population and tracking the potential emergence of mutations at positions associated with reduced susceptibility to existing monoclonal antibodies or at amino acids located in the epitopes of the F or G proteins targeted by vaccines. Furthermore, antivirals targeting other viral proteins (nucleoprotein, polymerase) are being developed. Mutations conferring reduced susceptibility to these inhibitors could also emerge and, if they are detected in a patient, this could lead to a change in therapy.

We aim to regularly update this software with user feedback and/or new subgroups and resistance mutations identified after the release of this tool.

Consent for publication

There are no case presentations in this study that require the disclosure of confidential data or information.

Funding

This work was supported by the Agence Nationale de la Recherche sur le SIDA et les Maladies Infectieuses Emergentes (ANRS MIE), ANRS MIE Medical Virology network.

CRediT authorship contribution statement

Anna-Maria Franco Yusti: Writing – review & editing, Resources, Investigation, Data curation. **Valentine Marie Ferré:** Writing – review & editing, Resources, Project administration. **Quentin Le Hingrat:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Romain Coppée:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project

administration, Methodology, Formal analysis, Data curation, Conceptualization. **Charlotte Charpentier:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Diane Descamps:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Alexandre Dosbaa:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Romane Guilbaud:** Writing – review & editing, Visualization, Validation, Software, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no competing interests regarding the current work.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jviromet.2024.114938 View High-Res Image.

References

Beerenwinkel, N., Günthard, H.F., Roth, V., Metzner, K.J., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front. Microbiol. 3, 329.

Connors, M., Collins, P.L., Firestone, C.Y., Murphy, B.R., 1991. Respiratory syncytial virus (RSV) F, G, M2 (22K), and N proteins each induce resistance to RSV challenge, but resistance induced by M2 and N proteins is relatively short-lived. J. Virol. 65, 1634–1637.

Coppée, R., Chenane, H.R., Bridier-Nahmias, A., et al., 2022. Temporal dynamics of RSV shedding and genetic diversity in adults during the COVID-19 pandemic in a French hospital, early 2021. Virus Res. 323, 198950.

Danecek, P., Bonfield, J.K., Liddle, J., et al., 2021. Twelve years of SAMtools and BCFtools. Gigascience 10, giab008.

Falsey, A.R., Hennessey, P.A., Formica, M.A., Cox, C., Walsh, E.E., 2005. Respiratory syncytial virus infection in elderly and high-risk adults. N. Engl. J. Med. 352, 1749–1759.

Goya, S., Galiano, M., Nauwelaers, I., et al., 2020. Toward unified molecular surveillance of RSV: a proposal for genotype definition. Influenza Other Respir. Virus 14, 274–285.

Houldcroft, C.J., Beale, M.A., Breuer, J., 2017. Clinical and biological insights from viral genome sequencing. Nat. Rev. Microbiol. 15, 183–192.

Levine, S., Klaiber-Franco, R., Paradiso, P.R., 1987. Demonstration that glycoprotein G is the attachment protein of respiratory syncytial virus. J. Gen. Virol. 68 (Pt 9), 2521–2524.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

McLellan, J.S., Chen, M., Leung, S., et al., 2013. Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody. Science 340, 1113–1117.

Mufson, M.A., Orvell, C., Rafnar, B., Norrby, E., 1985. Two distinct subtypes of human respiratory syncytial virus. J. Gen. Virol. 66 (Pt 10), 2111–2124.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.

Shi, T., McAllister, D.A., O'Brien, K.L., et al., 2017. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. Lancet 390, 946–958.

Stockdale, J.E., Liu, P., Colijn, C., 2022. The potential of genomics for infectious disease forecasting. Nat. Microbiol. 7, 1736–1743.

Wilkins, D., Langedijk, A.C., Lebbink, R.J., et al., 2023. Nirsevimab binding-site conservation in respiratory syncytial virus fusion glycoprotein worldwide between 1956 and 2021: an analysis of observational study sequencing data. Lancet Infect. Dis. 23, 856–866.

Yin, H.-S., Wen, X., Paterson, R.G., Lamb, R.A., Jardetzky, T.S., 2006. Structure of the parainfluenza virus 5 F protein in its metastable, prefusion conformation. Nature 439, 38–44.

Zhu, Q., Lu, B., McTamney, P., et al., 2018. Prevalence and significance of substitutions in the fusion protein of respiratory syncytial virus resulting in neutralization escape from antibody MEDI8897. J. Infect. Dis. 218, 572–580.