VODKA2: An accurate method to detect copy-back and deletion viral genomes from next-generation sequencing data

Emna Achouri[a,b], Sébastien A. Felt[a,b], Matthew Hackbart[a,b], and Carolina B. López[a#]

[a] Department of Molecular Microbiology and Center for Women Infectious Disease Research, Washington University School of Medicine, St Louis, MO, USA

[b] These authors contributed equally.

Address correspondence to Carolina B. López (clopezzalaquett@wustl.edu)

ABSTRACT

Summary: During viral replication, RNA viruses produce non-standard viral genomes (nsVGs), including copy-back viral genomes (cbVGs) and deletion viral genomes (delVGs), that play a crucial role in regulating viral replication. The sequences of these nsVGs are highly similar to the standard genome, making identification and quantification of the species generated during viral infections a challenging task. Here, we present our next-generation sequencing data analysis pipeline, Viral Opensource DVG Key Algorithm 2 (VODKA2), that identifies nsVGs with high accuracy from NGS data.

Availability and implementation: VODKA2 is freely available at GitHub (https://github.com/lopezlab-washu/VODKA2)

Contact: clopezzalaquett@wustl.edu

INTRODUCTION

Viral RNA populations consist of standard full-length viral genomes as well as truncated and/or rearranged non-standard viral genomes (nsVGs) (Gonzalez Aparicio et al., 2022). Two of the most studied nsVGs are copy-back viral genomes (cbVGs) and internal deletion viral genomes (delVGs). CbVGs are generated when the polymerase detaches from the template and resumes elongation downstream by copying the nascent strand generating viral genomes with complementary ends (Lazzarini et al., 1981; Perrault and Leavitt, 1978; Re et al., 1983; Vignuzzi and Lopez, 2019) (Fig. 1A). Internal delVGs have truncations either due to the viral polymerase skipping portions of the genome or the viral genome recombining to eliminate portions of the genome (Davis et al., 1980; Nomoto et al., 1979; O'Hara et al., 1984; Perrault and Semler, 1979; Vignuzzi and Lopez, 2019) (Fig. 1A). To date, nsVGs have three well characterized functions that shape virus infection outcome: interference with standard viral genome replication, immunostimulation, and establishment of viral persistence (Vignuzzi and Lopez, 2019). Due to their biological significance, it is of utmost importance to develop tools that allow us to determine the nsVG content of RNA virus populations. However, the discovery of nsVGs has been limited by their vast diversity and the limited presence of unique sequences that differentiate nsVGs from the standard viral genome. To facilitate the characterization and quantification of nsVGs, we previously developed the Viral Opensource DVG Key Algorithm (VODKA), a bioinformatics pipeline that identifies cbVGs from RNA-sequencing data (Felt et al., 2021; Sun et al., 2019). Here we present a substantially updated version of this pipeline (VODKA2) that can now detect both cbVGs and delVGs and has significantly improved sensitivity and specificity.

PIPELINE OVERVIEW

VODKA2 detects cbVG junctions from specific regions of the viral genome and quantifies the number of RNA-sequencing reads that contain such cbVG junction with enhanced accuracy. The VODKA2 pipeline proceeds in the following 5 steps (Fig. 1B):

1. NsVG database generation: a database of all potential nsVG junctions is created based on a reference standard viral genome selected by the user. A nsVG junction is defined as the connected nucleotides where the viral polymerase breaks from the viral genome and then subsequently rejoins at either the nascent strand (copy-back genomes) or further downstream of the template (deletion genomes) to generate a continuous break/rejoin sequence. To generate the database of all potential cbVGs or delVGs, all possible combinations of break and rejoin positions together with *N* nucleotides upstream of

break and downstream of rejoin are first extracted from the reference genome (*N* should be at least equal to the sequencing read size of the dataset to be analyzed). The sequence downstream of rejoin is maintained for delVG junction sequences (region A of theoretical delVG shown in Fig. 1A) or reverse complemented for cbVG junction sequences (region C' of theoretical cbVG in Fig. 1A) before being concatenated with the sequence upstream of break (region C in Fig. 1A). Further description of the generation of the nsVG databases are detailed in the Suppl. Fig. 1. The theoretical nsVG junction sequences have a length of 2*N, allowing for full mapping of potential nsVG junction reads during step 2 described next. The databases generated for the analysis presented in this paper are described in Suppl. Table 1.

2. Mapping reads to nsVG database with Bowtie2: Before mapping RNA-sequencing reads to the database, we recommend trimming sequencing adapters and removing host and standard viral genome reads from raw sequencing data (see raw data preprocessing description in Fig. 1C). Using only non-host, non-viral reads for the nsVG mapping step limits the number of false positives, such as host and viral sequences that are similar to a theoretical nsVG junction. Mapping is performed with Bowtie2 set to lenient settings (option -mp 0,0) that minimize mismatch penalties due to viral mutations (e.g. nucleotide substitutions, insertions and/or deletions). We designed the algorithm to report any read that would match exactly 30 nucleotides around the junction (15 on each side), but any mismatch outside of that region is allowed in order to maximize the chances of detecting all nsVG junction reads and avoid false negative. This approach assures high nsVG detection sensitivity and generates a comprehensive preliminary list of possible reads mapped to nsVG junctions for each dataset.

3. Filtering possible junction reads with BLAST: To ensure that no false positive reads are part of the output, VODKA2 runs an additional BLAST alignment of all possible nsVG reads from step 2 against the viral reference genome. Any potential nsVG read that does not meet the following criteria is discarded: (i) the read must match exactly 2 separate regions of the reference viral genome (reported as "alignment ranges" in BLAST output). Sequences mapped to the VODKA2 database during step 2 but showing ambiguous BLAST alignment (more than 2 ranges) are removed. (ii) The break/rejoin positions shown by the 2 BLAST ranges must be consistent with step 2 output (i.e. within 5 nucleotides). (iii) The last requirement is nsVG subtype specific. For delVGs, the 2 ranges must map to the same strand of the viral genome. For cbVGs, the 2 ranges must match opposite strands of the genome. Upon fulfillment of the three requirements, the read is designated as a true nsVG junction read.

4. NsVG species aggregation: Among the confirmed nsVG reads, some nsVG junctions share similar sequences, but the reported break/rejoin junction is shifted to a nearby position. This shift often occurs if the break/rejoin junction contains identical sequences flanking the break and rejoin positions as the junction can be mapped to any position within these sequences. This phenomenon was already reported by others (Beauclair et al., 2018; Routh and Johnson, 2014) and can potentially bias quantification of nsVGs junction reads. To provide more accurate nsVG read counts at these shifted junctions, similar reads are grouped into nsVG "species" based on two criteria: (i) the predicted nsVG size that is calculated based on the junction break and rejoin positions in the reference genome must be identical for all nsVG reads within the same group, (ii) break position shifts must be of 5 nucleotides or less. A new "Species_ID" is then generated to represent each group of similar nsVG reads. The Species_ID is called as the most represented break (and its corresponding rejoin) in the group (Suppl. Fig. 2). While this aggregation process provides a more accurate representation and quantification of nsVG junctions present in a sample, the aggregation settings can be altered depending on the question asked.

5. Output Table and Visualizations: VODKA2 provides abundant output data for the detected nsVG species (Fig. 1D). For both cbVGs and delVGs, the output file includes the read ID, the sequence of the read, the break point, the rejoin point, the strandedness of the read (SAM flag), the predicted size of the nsVG and the aggregated Species_ID. For cbVGs only, the output file also includes the theoretical length of the cbVG loops and stems. For delVGs only, the output file includes the deletion size, the viral gene where the break occurs, and the viral gene where the rejoin occurs. Using this output, multiple types of analyses can be performed to study the abundance and diversity of cbVGs and delVGs.

RESULTS

To determine the accuracy of VODKA2 for the detection of cbVGs, we utilized artificial copy-back sequences created *in silico* based on the RSV genome and simulated Novaseq 150 bp paired end reads with InSilicoSeq (Gourle et al., 2019)(data described in Suppl. Table 2.). We selected a strain of RSV known to be circulating in nature (Choudhary et al., 2013) and focused on the G gene. The G gene contains a 72 nt repetition within its sequence, posing an additional challenge for the pipeline as it is expected to impact alignments. The VODKA2 analysis resulted in detection of all the expected copy-back junction regions, including species presenting a junction within the repetition region, while no false positive was generated (Fig. 1E).

With the goal of showing that VODKA2 can detect delVGs and confirm that results are consistent with previously published delVG data, we next analyzed published data on delVGs from IAV for the presence of delVGs in the PB1 and PB2 segments (Alnaji et al., 2021). In this published work, the authors reported that most IAV delVGs derive from the PB1 and PB2 segments, while none were found in the M and NS segments. Using a database of all potential delVGs for each of these four segments, VODKA2 reproduced these data and confirmed that more delVGs are detected in PB1 than in PB2, and minimal amounts in M and NS (Fig. 1F and data not shown). We report similar proportions of delVGs, demonstrating that VODKA2 can now be used as a tool to not only reliably detect cbVGs but also delVG species (Fig. 1F).

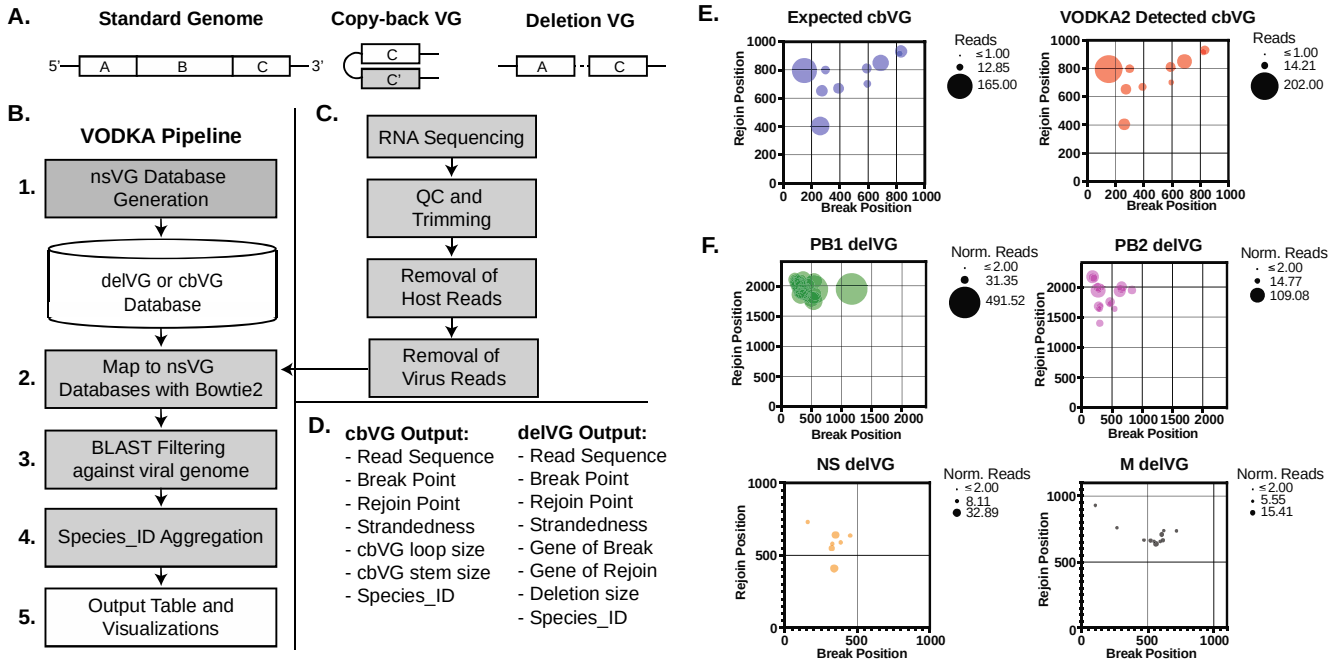

Figure 1. A) Schematic of a standard viral genome (VG) containing 3 genes: A, B, and C, a copy-back VG containing gene C and the complement C' end, and a deletion VG that has the B gene removed. (B-D) Brief descriptions of the steps and output files of the VODKA bioinformatics pathway. Gray boxes indicate bioinformatics programs/processes and white boxes indicate output files. B) The steps of the VODKA pipeline that are discussed in the text. C) Pre-processing steps of RNA-sequencing reads to obtain non-Host Reads. D) Example of Output Data from VODKA pipeline analysis. E+F) Representation of E) RSV G cbVGs and F) IAV delVG break and rejoin points. Size of each point is dependent on the number of normalized reads at each break/rejoin species.

DISCUSSION AND CONCLUSION

It has become clear over the past decade that nsVGs play a crucial role in virus pathogenesis (Felt et al., 2021; Penn et al., 2022; Vasilijevic et al., 2017; Vignuzzi and Lopez, 2019; Zhou et al., 2022). As a consequence, there has been a big push in the nsVG field to develop NGS/bioinformatics approaches that are more sensitive, less biased, and more accurate, such as DI-tector (Beauclair et al., 2018), DVG-Profiler (Bosma et al., 2019), DG-Seq (Boussier et al., 2020), ViReMa (Jaworski and Routh, 2017; Routh and Johnson, 2014; Sotcheff et al., 2023) and DVGfinder (Olmo-Uceda et al., 2022). Our own group has previously developed a pipeline, VODKA, which is the first tool focusing on the detection of cbVGs within NGS data (Sun et al., 2019). We have updated and improved this pipeline, now called VODKA2, to detect not only cbVGs but also delVGs. The pipeline can analyze NGS datasets of non-segmented or segmented viruses and generates accurate and detailed outputs for further analysis of cbVGs and delVGs. We show that the new filtering process reduces the risk of detecting false positives, increasing the specificity of VODKA2. Indeed, the analysis of simulated data showed that the pipeline did not generate any junction that were not present in the raw sequences. Furthermore, although we reported similar amounts of delVG species as Alnaji et al for IAV PB1 and PB2 segments, VODKA2 noticeably detected more species in M and N, suggesting an increased sensitivity of this method. Taken together, VODKA2 is a valuable tool for the detection of nsVGs species providing data that set the stage for understanding better the biology of nsVG amongst different RNA virus populations.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGEMENTS

We want to thank Gregory R Grant and Eun Ji Kim for initial VODKA.

FUNDING INFORMATION

This works was supported by the US National Institutes of Health National Institute of Allergy and Infectious Diseases [AI137062 and AI134862 to CLB] and [T32-007317 to MH].

REFERENCES

Alnaji, F.G., Reiser, W.K., Rivera-Cardona, J., Te Velthuis, A.J.W., Brooke, C.B., 2021. Influenza A Virus Defective Viral Genomes Are Inefficiently Packaged into Virions Relative to Wild-Type Genomic RNAs. mBio, e0295921.

Beauclair, G., Mura, M., Combredet, C., Tangy, F., Jouvenet, N., Komarova, A.V., 2018. DI-tector: defective interfering viral genomes' detector for next-generation sequencing data. RNA 24, 1285-1296.

Boussier, J., Munier, S., Achouri, E., Meyer, B., Crescenzo-Chaigne, B., Behillil, S., Enouf, V., Vignuzzi, M., van der Werf, S., Naffakh, N., 2020. RNA-seq accuracy and reproducibility for the mapping and quantification of influenza defective viral genomes. RNA 26, 1905-1918.

Choudhary, M.L., Wadhwa, B.S., Jadhav, S.M., Chadha, M.S., 2013. Complete Genome Sequences of Two Human Respiratory Syncytial Virus Genotype A Strains from India, RSV-A/NIV1114046/11 and RSV-A/NIV1114073/11. Genome Announc 1.

Davis, A.R., Hiti, A.L., Nayak, D.P., 1980. Influenza defective interfering viral RNA is formed by internal deletion of genomic RNA. Proc Natl Acad Sci U S A 77, 215-219.

Felt, S.A., Sun, Y., Jozwik, A., Paras, A., Habibi, M.S., Nickle, D., Anderson, L., Achouri, E., Feemster, K.A., Cardenas, A.M., Turi, K.N., Chang, M., Hartert, T.V., Sengupta, S., Chiu, C., Lopez, C.B., 2021. Detection of respiratory syncytial virus defective genomes in nasal secretions is associated with distinct clinical outcomes. Nat Microbiol 6, 672-681.

Gonzalez Aparicio, L.J., Lopez, C.B., Felt, S.A., 2022. A Virus Is a Community: Diversity within Negative-Sense RNA Virus Populations. Microbiol Mol Biol Rev 86, e0008621.

Gourle, H., Karlsson-Lindsjo, O., Hayer, J., Bongcam-Rudloff, E., 2019. Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics 35, 521-522.

Jaworski, E., Routh, A., 2017. Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. PLoS Pathog 13, e1006365.

Lazzarini, R.A., Keene, J.D., Schubert, M., 1981. The origins of defective interfering particles of the negative-strand RNA viruses. Cell 26, 145-154.

Nomoto, A., Jacobson, A., Lee, Y.F., Dunn, J., Wimmer, E., 1979. Defective interfering particles of poliovirus: mapping of the deletion and evidence that the deletions in the genomes of DI(1), (2) and (3) are located in the same region. J Mol Biol 128, 179-196.

O'Hara, P.J., Nichol, S.T., Horodyski, F.M., Holland, J.J., 1984. Vesicular stomatitis virus defective interfering particles can contain extensive genomic sequence rearrangements and base substitutions. Cell 36, 915-924.

Olmo-Uceda, M.J., Munoz-Sanchez, J.C., Lasso-Giraldo, W., Arnau, V., Diaz-Villanueva, W., Elena, S.F., 2022. DVGfinder: A Metasearch Tool for Identifying Defective Viral Genomes in RNA-Seq Data. Viruses 14.

Penn, R., Tregoning, J.S., Flight, K.E., Baillon, L., Frise, R., Goldhill, D.H., Johansson, C., Barclay, W.S., 2022. Levels of Influenza A Virus Defective Viral Genomes Determine Pathogenesis in the BALB/c Mouse Model. J Virol 96, e0117822.

Perrault, J., Leavitt, R.W., 1978. Inverted complementary terminal sequences in single-stranded RNAs and snap-back RNAs from vesicular stomatitis defective interfering particles. J Gen Virol 38, 35-50.

Perrault, J., Semler, B.L., 1979. Internal genome deletions in two distinct classes of defective interfering particles of vesicular stomatitis virus. Proc Natl Acad Sci U S A 76, 6191-6195.

Re, G.G., Gupta, K.C., Kingsbury, D.W., 1983. Genomic and copy-back 3' termini in Sendai virus defective interfering RNA species. J Virol 45, 659-664.

Routh, A., Johnson, J.E., 2014. Discovery of functional genomic motifs in viruses with ViReMa-a Virus Recombination Mapper-for analysis of next-generation sequencing data. Nucleic Acids Res 42, e11.

Sotcheff, S., Zhou, Y., Yeung, J., Sun, Y., Johnson, J.E., Torbett, B.E., Routh, A.L., 2023. ViReMa: a virus recombination mapper of next-generation sequencing data characterizes diverse recombinant viral nucleic acids. Gigascience 12.

Sun, Y., Kim, E.J., Felt, S.A., Taylor, L.J., Agarwal, D., Grant, G.R., Lopez, C.B., 2019. A specific sequence in the genome of respiratory syncytial virus regulates the generation of copy-back defective viral genomes. PLoS Pathog 15, e1007707.

Vasilijevic, J., Zamarreno, N., Oliveros, J.C., Rodriguez-Frandsen, A., Gomez, G., Rodriguez, G., Perez-Ruiz, M., Rey, S., Barba, I., Pozo, F., Casas, I., Nieto, A., Falcon, A., 2017. Reduced accumulation of defective viral genomes contributes to severe outcome in influenza virus infected patients. PLoS Pathog 13, e1006650.

Vignuzzi, M., Lopez, C.B., 2019. Defective viral genomes are key drivers of the virus-host interaction. Nat Microbiol 4, 1075-1087.

Zhou, T., Gilliam, N.J., Li, S., Spaudau, S., Osborn, R.M., Anderson, C.S., Mariani, T.J., Thakar, J., Dewhurst, S., Mathews, D.H., Huang, L., Sun, Y., 2022. Generation and functional analysis of defective viral genomes during SARS-CoV-2 infection. bioRxiv.