VODKA2: A fast and accurate method to detect non-standard viral genomes from large RNA-seq datasets

Emna Achouri[a], Sébastien A. Felt[a], Matthew Hackbart[a], Nicole S. Rivera-Espinal[a] 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

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

ABSTRACT

During viral replication, viruses carrying an RNA genome 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 and pathogenesis. Because of their critical roles in determining the outcome of RNA virus infections, the study of nsVGs has flourished in recent years exposing a need for bioinformatic tools that can accurately identify them within Next-Generation Sequencing data obtained from infected samples. Here, we present our data analysis pipeline, Viral Opensource DVG Key Algorithm2 (VODKA2), that is optimized to run on a High Performance Computing (HPC) environment for fast and accurate detection of nsVGs from large data sets.

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

23 **INTRODUCTION**

24 RNA viruses contain populations of RNA genomes that include the standard full-length viral
25 genomes as well as truncated and/or rearranged non-standard viral genomes (nsVGs) (Gonzalez
26 Aparicio et al., 2022). Two of the most studied nsVGs are copy-back viral genomes (cbVGs) and internal
27 deletion viral genomes (delVGs). CbVGs are generated when during replication of the viral RNA genome
28 the polymerase detaches from the template strand and resumes elongation downstream by copying the
29 nascent strand generating viral genomes with complementary ends (Lazzarini et al., 1981; Perrault and
30 Leavitt, 1978; Re et al., 1983; Vignuzzi and Lopez, 2019) (Fig. 1A). Internal delVGs have truncations
31 either due to the viral polymerase skipping portions of the genome or the viral genome recombining to
32 eliminate portions of the genome (Davis et al., 1980; Nomoto et al., 1979; O'Hara et al., 1984; Perrault
33 and Semler, 1979; Vignuzzi and Lopez, 2019) (Fig. 1A). nsVGs have three well characterized functions
34 that shape virus infection outcome: interference with standard viral genome replication,
35 immunostimulation, and establishment of viral persistence (Vignuzzi and Lopez, 2019). Despite their
36 biological significance, the study of nsVGs has been limited by their vast diversity and the limited
37 presence of unique sequences that differentiate nsVGs from the standard viral genome. Although several
38 bioinformatic tools have been created for the identification of nsVGs from NGS data sets (Beauclair et
39 al., 2018; Bosma et al., 2019; Boussier et al., 2020; Jaworski and Routh, 2017; Olmo-Uceda et al., 2022;
40 Routh and Johnson, 2014; Timm et al., 2014), we show in this work that these tools are not optimized for
41 the identification of cbVGs and often lead to mistakes in the identity of the reported cbVG species.
42 Furthermore, these tools are not designed to efficiently handle large data sets, leading to lengthy
43 processing times. Therefore, there is a need for bioinformatics approaches that can accurately and
44 efficiently detect and characterize cbVGs from large NGS data sets. Our group previously developed and
45 utilized an algorithm specifically focused on the detection of cbVGs (Felt et al., 2022; Felt et al., 2021;
46 Sun et al., 2019). Here we present a substantially updated version of this tool, the Viral Opensource DVG
47 Key Algorithm 2 (VODKA2), that is optimized to run on a High Performance Computing (HPC)
48 environment for fast and accurate detection of nsVGs.

49 **RESULTS**

50 **Pipeline Overview**

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

54 *1. NsVG database generation*: a database of all potential nsVG junctions is created based on a reference
55 standard viral genome selected by the user. A nsVG junction is defined as the connected nucleotides
56 where the viral polymerase breaks from the viral genome and then subsequently rejoins at either the
57 nascent strand (copy-back genomes) or further downstream of the template (deletion genomes) to
58 generate a continuous break/rejoin sequence. To generate the database of all potential cbVGs or delVGs,
59 all possible combinations of break and rejoin positions together with *N* nucleotides upstream of break
60 and downstream of rejoin are first extracted from the reference genome (*N* should be at least equal to
61 the sequencing read size of the dataset to be analyzed). The sequence downstream of rejoin is
62 maintained for delVG junction sequences (region A of theoretical delVG shown in Fig. 1A) or reverse
63 complemented for cbVG junction sequences (region C' of theoretical cbVG in Fig. 1A) before being
64 concatenated with the sequence upstream of break (region C in Fig. 1A). Further description of the
65 generation of the nsVG databases are detailed in the Suppl. Fig. 1. The theoretical nsVG junction
66 sequences have a length of 2*N, allowing for full mapping of potential nsVG junction reads during step 2
67 described next.

68 *2. Mapping reads to nsVG database with Bowtie2*: Before running samples through VODKA2, we
69 recommend trimming sequencing adapters and removing host reads from the raw sequencing data (see
70 Fig. 1C). Reads corresponding to any known co-infecting virus or contaminant organism can also be
71 removed to reduce the size of input data, hence increasing the speed of analysis. The VODKA2 process
72 initiates by eliminating any viral read that aligns perfectly with the standard viral genome (using Bowtie2
73 with default parameters). These preliminary filters help reducing the data volume and minimizing false
74 positives, as some host and viral sequences could be mistakenly identified as non-standard viral genome

75 (nsVG) junctions. Mapping against the VODKA2 nsVG database, which is the key step of our method, is
76 performed with Bowtie2 set to lenient settings (option -mp 0,0) in order to minimize mismatch penalties
77 due to viral mutations (e.g. nucleotide substitutions, insertions and/or deletions). We designed the
78 algorithm to report any read that would match exactly 30 nucleotides around the junction (15 on each
79 side), but any mismatch outside of that region is allowed in order to maximize the chances of detecting
80 all nsVG junctions reads and avoid false negatives. This approach assures high nsVG detection
81 sensitivity and generates a comprehensive preliminary list of possible reads mapped to nsVG junctions
82 for each dataset.
83 *3. Filtering possible junction reads with BLAST:* To ensure that no false positive reads are part of the
84 output, VODKA2 runs an additional BLAST alignment of all possible nsVG reads from step 2 against the
85 viral reference genome. Any potential nsVG read that does not meet the following criteria is discarded:
86 (i) the read must match exactly 2 separate regions of the reference viral genome (reported as "alignment
87 ranges" in BLAST output). Sequences mapped to the VODKA2 database during step 2 but showing
88 ambiguous BLAST alignment (more than 2 ranges) are removed. (ii) The break/rejoin positions shown
89 by the 2 BLAST ranges must be consistent with step 2 output (i.e. within 5 nucleotides). (iii) The last
90 requirement is nsVG subtype specific. For delVGs, the 2 ranges must map to the same strand of the viral
91 genome. For cbVGs, the 2 ranges must match opposite strands of the genome. Upon fulfillment of the
92 three requirements, the read is designated as a true nsVG junction read.
93 *4. NsVG species aggregation:* Among the confirmed nsVG reads, some nsVG junctions share similar
94 sequences, but the reported break/rejoin junction is shifted to a nearby position. This shift often occurs if
95 the break/rejoin junction contains identical sequences flanking the break and rejoin positions. NsVG
96 reads can be mapped to any position within these sequences. This phenomenon was already reported
97 by others (Beauclair et al., 2018; Routh and Johnson, 2014) and can potentially bias quantification of
98 nsVGs junction reads. To provide more accurate nsVG read counts at these shifted junctions, similar
99 reads are grouped into nsVG "species" based on two criteria: (i) the predicted nsVG size that is calculated
100 based on the junction break and rejoin positions in the reference genome must be identical for all nsVG

101 reads within the same group, (ii) break position shifts must be of 5 nucleotides or less. A new
102 "Species_ID" is then generated to represent each group of similar nsVG reads. The Species_ID is called
103 as the most represented break (and its corresponding rejoin) in the group (Suppl. Fig. 2). While this
104 aggregation process provides a more accurate representation and quantification of nsVG junctions
105 present in a sample, the aggregation settings can be altered depending on the question asked.
106 *5. Output Table and Visualizations:* VODKA2 provides abundant output data for the detected nsVG
107 species (Fig. 1D). For both cbVGs and delVGs, the output file includes the read ID, the sequence of the
108 read, the break point, the rejoin point, the strandedness of the read (SAM flag), the predicted size of the
109 nsVG and the aggregated Species_ID. For cbVGs only, the output file also includes the theoretical length
110 of the cbVG loops and stems. For delVGs only, the output file includes the deletion size, the viral gene
111 where the break occurs, and the viral gene where the rejoin occurs. Using this output, multiple types of
112 analyses can be performed to study the abundance and diversity of cbVGs and delVGs.
113
114 **VODKA2 detects simulated cbVGs with high accuracy**
115 In order to assess the accuracy of VODKA2 in detecting cbVGs, we performed a comparative
116 analysis against state-of-the-art methodologies. We created 10 artificial cbVG sequences *in silico*, based
117 on the respiratory syncytial virus (RSV) G gene (KC731482.1), these are detailed in Table 1. This gene
118 contains a 72 nt repetition within its sequence, posing an additional challenge for the detection of nsVGs
119 as it is expected to impact sequence alignments. The simulated sample was aligned against the RSV G
120 gene reference sequence (KC731482.1) and the non-viral reads were then analyzed with VODKA2,
121 ViReMa (Sotcheff et al., 2023)and DI-tector (Fig. 2A). VODKA2 and DI-tector detected all of the artificial
122 cbVG junction regions without generating any false positive (Table 1 and Fig. 2A). ViReMa also detected
123 all of the artificial cbVGs but in addition reported a false positive species. However, the Break/Rejoin
124 positions of the cbVG junctions reported by both ViReMa and DI-tector (respectively referred as
125 Donor/Acceptor sites and Breakpoint/Reinitiation sites) were incorrect and can mislead the user. As
126 shown in Figure 2A, the Break and Rejoin positions of the cbVG species seem to be systematically

127 inverted by both tools. In the case of ViReMA, this can be explained by the fact that the software was
128 originally designed to detect recombinations between donor and acceptor genomes of the same
129 orientation. Hence, any recombination event between a positive and a negative strand and viceversa is
130 called a copy-back, without clearly identifying which of the donor or acceptor site is the break (or rejoin)
131 position of a cbVG(Sotcheff et al., 2023). Regarding DI-tector, although it also seems to be biased to the
132 opposite sense, 8 cbVG species out of 10 are reported in both orientations. This phenomenon has already
133 been observed and reported by the authors of DI-tector and suspected to be false positives (Beauclair et
134 al., 2018). The VODKA2 approach presents the advantage of avoiding such ambiguous reporting as the
135 pre-built database will constrain the putative cbVG to a specific orientation. In other terms, the Break and
136 Rejoin reported by VODKA2 are controlled by the cbVG reference sequences and cannot be distort by a
137 misinterpretation of the sequencing reads orientation.
138
139 **VODKA2 can detect delVGs**
140 In addition to cbVGs, it is possible to use VODKA2 to detect deletions by generating a database
141 of all putative delVGs of the specified reference genome. To illustrate this new feature, we analyzed 10
142 artificial delVGs based on the RSV G gene. Similarly to the results obtained with the simulated cbVG
143 data, all expected delVG species were correctly identified, showing again the specificity of this method.
144 Observably, VODKA2 reported slightly fewer reads than DI-tector and ViReMA for some particular cbVG
145 and delVG species (Fig 2B and Table 1) and it is unclear at the moment what is the couse of this.
146
147 **VODKA2 detects cbVGs from NGS data with high accuracy and efficiency**
148 With the goal of demonstrating the cbVGs detection efficiency of VODKA2 in real RNA-seq data,
149 we analyzed data from a sample of cells infected *in vitro* with Sendai virus (SeV), which is a
150 paramyxovirus of 15,384 nucleotides that is known to produce a single major cbVG species of length 546
151 nucleotides (Sun et al., 2019), with the Break and Rejoin occurring around positions 14,930 and 15,290
152 of the reference genome. VODKA2 expeditiously processed the sample and reported 7,138 validated

153 cbVG junction reads, all having a Break and Rejoin around the expected positions for cbVG 546. ViReMa
154 and DI-tector results also confirmed the presence of this major species but both algorithms also reported
155 at least 1 false positive junction and required significantly increased running time compared to VODKA2,
156 which processed the sequencing sample almost 10 times faster. This comparison demonstrates the
157 advantage of VODKA2 to fast and reliably detect cbVGs in large datasets (Fig. 3).
158

159 **DISCUSSION AND CONCLUSION**

160 It has become clear over the past decade that nsVGs play a crucial role in virus pathogenesis
161 (Felt et al., 2021; Penn et al., 2022; Vasilijevic et al., 2017; Vignuzzi and Lopez, 2019; Zhou et al., 2022).
162 As a consequence, there has been a big push in the nsVG field to develop NGS/bioinformatics
163 approaches that are more sensitive, less biased, and more accurate, such as DI-tector (Beauclair et al.,
164 2018), DVG-Profiler (Bosma et al., 2019), DG-Seq (Boussier et al., 2020), ViReMa (Jaworski and Routh,
165 2017; Routh and Johnson, 2014; Sotcheff et al., 2023) and DVGfinder (Olmo-Uceda et al., 2022). Our
166 own group has previously developed a pipeline, VODKA, which is the first tool focusing on the detection
167 of cbVGs within NGS (Sun et al., 2019). We have updated and improved this pipeline, now called
168 VODKA2, to detect not only cbVGs but also delVGs in a time effective manner.

169 The development of NGS technologies offers the capacity of deeper sequencing and increasing
170 amounts of samples within a single experiment, which raises the need for improving the running speed
171 of bioinformatics tools for more efficient analysis of large data sets. We achieved that goal in VODKA2
172 and considerably accelerated the analysis process of each sample by separating the nsVG database
173 generation step. A database is generated only once for each reference genome and remains available
174 for any further sample analysis, saving precious processing time when running larger datasets. Moreover,
175 contrary to other common methods such as ViReMa and DI-tector, VODKA2 database offers the benefit
176 of calling nsVG junctions after running a single alignment step, without having to split unmapped
177 sequences and re-align them multiple times. Furthermore, we designed VODKA2 to easily run several

178 samples in parallel by using a job scheduler when available, and take advantage of high performance
179 computing resources, offering the possibility to run effectively large-scale project.

180 Another strong advantage of VODKA2 nsVG database approach is the control of cbVG orientation
181 by unambiguously identifying the Break and Rejoin positions, whereas both ViReMa and DI-tector report
182 misleading information that can be confusing for the user, unless running further analysis of each reported
183 junction read to identify the correct sense. Yet, as cbVGs have been shown to be produced only during
184 replication of negative-sense RNA viruses, VODKA2 focuses by default on the cbVGs generated from
185 the 3' ends of the provided reference genome. If looking to detect cbVGs generated from the opposite
186 end, e.g. when working with a positive-sense RNA virus, we suggest using the reverse sequence of the
187 reference genome to generate the cbVG database and running the analysis.

188 VODKA2 can analyze large NGS datasets and generates accurate and detailed outputs for further
189 analysis of cbVGs and delVGs. We show that the new filtering process reduces the risk of detecting false
190 positives, increasing the specificity of VODKA2. Altough we observed a slight decrease of VODKA2
191 sentivity as compared to ViReMa and DI-tector, that is at the benefit of a 10 time fold increase of speed.
192 Taken together, VODKA2 is a valuable tool for the fast detection of nsVG species providing data that set
193 the stage for understanding better the biology of nsVG amongst different RNA virus populations in large-
194 scale studies.

195 **MATERIAL AND METHODS**

196 **Simulated nsVG sequencing data**

197 We used RSV reference genome (KC731482.1) and generated random numbers from a range of 15 to
198 951 to create artificial copy-back and deletion sequences from gene G. Novaseq150 bp paired end reads
199 were simulated using InSilicoSeq v1.5.3 with default parameters and random coverage values (Gourle
200 et al., 2019).

201 **RNA-Seq data pre-processing**

202 Raw sequencing reads were trimmed of Illumina adaptors using Cutadapt (v2.5) and analyzed by FastQC
203 (v0.11) to ensure quality of sequencing reads. The reads mapping the human genome (GRCh38) were
204 removed from the datasets using Bowtie2 (v2.4.1).

205 **nsVG detection with VODKA2**

206 VODKA2 nsVG databases were created using VODKA2 *genome_to_new_fasta* step with read size set
207 to 150bp. We used the whole viral reference genome size for all databases: 966nt from RSV gene G
208 reference sequence were used to generate the cbVG and delVG databases. With Sendai virus being
209 much larger, we used 400GB RAM memory to generate cbVG and delVG databases from the whole
210 reference genome (15,384). The output of this step is a multi FASTA file containing all theoretical junction
211 region sequences, which was then indexed for Bowtie2 (using Bowtie2-build v2.4.1).
212 VODKA2 analysis was performed according to the recommendations on GitHub. VODKA2, ViReMa and
213 DI-tector analyses were run using either 10GB for the simulated nsVG data or 150GB for the SeV infected
214 sample. All scatterplots shown in this work were generated using R v4.3.0.

215 **Cell line and virus**

216 A549 cells (human type II pneumocytes cells, ATCC CCL-185) were cultured at 37°C with 5% CO2 with
217 Dubelcco's modified Eagle's medium (DMEM; Invitrogen) supplemented with 10% fetal bovine serum
218 (FBS), 50 ng/ml gentamicin (Thermofisher), 2 mM L-glutamine (Invitrogen), and 1 mM sodium pyruvate
219 (Invitrogen). A549 cells were treated with mycoplasma removal agent (MP Biomedicals, 093050044) and
220 routinely tested for mycoplasma before use. Sendai Cantell (SeV) strain was grown in 10-day-old,
221 embryonated chicken eggs (Charles River) for 40 h as previously described (Yount et al., 2006).

222 **Virus infection**

223 A549 cells were washed twice with PBS. Subsequently, the cells were exposed to SeV virus in an
224 infectious medium, with a multiplicity of infection (MOI) of 1.5. The infectious medium consisted of DMEM
225 supplemented with 35% bovine serum albumin (BSA; Sigma-Aldrich), penicillin-streptomycin (Invitrogen),
226 and 5% NaHCO3 (SigmaAldrich). The virus and cells were incubated together at 37°C for 1.5 hours.
227 Following the incubation period, additional infectious medium was provided to the cells. As a control,
228 mock conditions underwent the same media change procedure but without the addition of the virus.
229 Infected and mock cells were incubated at 37°C and 5% CO_2 for 16 hours. Supernatant were collected
230 of cells and RNA was extracted by TRIzol/chloroform extraction (Thermofisher, 15596018).

231 **RNA-Sequencing**

232 Total RNA was extracted from mock and infected cells using TriZol reagent (Life Technologies). The
233 concentration of the extracted RNA was measured using a Thermo Scientific™ NanoDrop™
234 spectrophotometer. RNA quality was assessed using an Agilent TapeStation or Bioanalyzer (Agilent
235 Technologies) prior to cDNA library preparation. All samples were prepared using the Illumina TruSeq
236 Stranded Total RNA Library Prep Kit with Ribo-Zero Gold. SeV samples were run on an Illumina NextSeq
237 500 to generate 150 bp, paired-end reads, resulting in ~ 41 million reads per sample. SeV samples were
238 run on an NextSeq 550 to generate 150 bp, pairerd-end reads, resulting in ~ 64 to 91 million reads per
239 samples. Average phred quality score of samples were ~ 34.

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, NSF-GRFP DGE2139839 to NSR ,and T32-007317 to MH.

251 **REFERENCES**

252 Beauclair, G., Mura, M., Combredet, C., Tangy, F., Jouvenet, N., Komarova, A.V., 2018. DI-tector:
253 defective interfering viral genomes' detector for next-generation sequencing data. RNA 24, 1285-1296.
254 Bosma, T.J., Karagiannis, K., Santana-Quintero, L., Ilyushina, N., Zagorodnyaya, T., Petrovskaya, S.,
255 Laassri, M., Donnelly, R.P., Rubin, S., Simonyan, V., Sauder, C.J., 2019. Identification and quantification
256 of defective virus genomes in high throughput sequencing data using DVG-profiler, a novel post-sequence
257 alignment processing algorithm. PLoS One 14, e0216944.
258 Boussier, J., Munier, S., Achouri, E., Meyer, B., Crescenzo-Chaigne, B., Behillil, S., Enouf, V., Vignuzzi,
259 M., van der Werf, S., Naffakh, N., 2020. RNA-seq accuracy and reproducibility for the mapping and
260 quantification of influenza defective viral genomes. RNA 26, 1905-1918.
261 Davis, A.R., Hiti, A.L., Nayak, D.P., 1980. Influenza defective interfering viral RNA is formed by internal
262 deletion of genomic RNA. Proc Natl Acad Sci U S A 77, 215-219.
263 Felt, S.A., Achouri, E., Faber, S.R., López, C.B., 2022. Accumulation of copy-back viral genomes during
264 respiratory syncytial virus infection is preceded by diversification of the copy-back viral genome
265 population followed by selection. Virus Evol 8, veac091.
266 Felt, S.A., Sun, Y., Jozwik, A., Paras, A., Habibi, M.S., Nickle, D., Anderson, L., Achouri, E., Feemster,
267 K.A., Cardenas, A.M., Turi, K.N., Chang, M., Hartert, T.V., Sengupta, S., Chiu, C., Lopez, C.B., 2021.
268 Detection of respiratory syncytial virus defective genomes in nasal secretions is associated with distinct
269 clinical outcomes. Nat Microbiol 6, 672-681.
270 Gonzalez Aparicio, L.J., Lopez, C.B., Felt, S.A., 2022. A Virus Is a Community: Diversity within
271 Negative-Sense RNA Virus Populations. Microbiol Mol Biol Rev 86, e0008621.
272 Gourle, H., Karlsson-Lindsjo, O., Hayer, J., Bongcam-Rudloff, E., 2019. Simulating Illumina
273 metagenomic data with InSilicoSeq. Bioinformatics 35, 521-522.
274 Jaworski, E., Routh, A., 2017. Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution
275 of defective-interfering RNAs in Flock House virus. PLoS Pathog 13, e1006365.
276 Lazzarini, R.A., Keene, J.D., Schubert, M., 1981. The origins of defective interfering particles of the
277 negative-strand RNA viruses. Cell 26, 145-154.
278 Nomoto, A., Jacobson, A., Lee, Y.F., Dunn, J., Wimmer, E., 1979. Defective interfering particles of
279 poliovirus: mapping of the deletion and evidence that the deletions in the genomes of DI(1), (2) and (3)
280 are located in the same region. J Mol Biol 128, 179-196.
281 O'Hara, P.J., Nichol, S.T., Horodyski, F.M., Holland, J.J., 1984. Vesicular stomatitis virus defective
282 interfering particles can contain extensive genomic sequence rearrangements and base substitutions. Cell
283 36, 915-924.
284 Olmo-Uceda, M.J., Munoz-Sanchez, J.C., Lasso-Giraldo, W., Arnau, V., Diaz-Villanueva, W., Elena,
285 S.F., 2022. DVGfinder: A Metasearch Tool for Identifying Defective Viral Genomes in RNA-Seq Data.
286 Viruses 14.
287 Penn, R., Tregoning, J.S., Flight, K.E., Baillon, L., Frise, R., Goldhill, D.H., Johansson, C., Barclay, W.S.,
288 2022. Levels of Influenza A Virus Defective Viral Genomes Determine Pathogenesis in the BALB/c
289 Mouse Model. J Virol 96, e0117822.
290 Perrault, J., Leavitt, R.W., 1978. Inverted complementary terminal sequences in single-stranded RNAs
291 and snap-back RNAs from vesicular stomatitis defective interfering particles. J Gen Virol 38, 35-50.
292 Perrault, J., Semler, B.L., 1979. Internal genome deletions in two distinct classes of defective interfering
293 particles of vesicular stomatitis virus. Proc Natl Acad Sci U S A 76, 6191-6195.
294 Re, G.G., Gupta, K.C., Kingsbury, D.W., 1983. Genomic and copy-back 3' termini in Sendai virus
295 defective interfering RNA species. J Virol 45, 659-664.

296 Routh, A., Johnson, J.E., 2014. Discovery of functional genomic motifs in viruses with ViReMa-a Virus
297 Recombination Mapper-for analysis of next-generation sequencing data. Nucleic Acids Res 42, e11.
298 Sotcheff, S., Zhou, Y., Yeung, J., Sun, Y., Johnson, J.E., Torbett, B.E., Routh, A.L., 2023. ViReMa: a
299 virus recombination mapper of next-generation sequencing data characterizes diverse recombinant viral
300 nucleic acids. Gigascience 12.
301 Sun, Y., Kim, E.J., Felt, S.A., Taylor, L.J., Agarwal, D., Grant, G.R., Lopez, C.B., 2019. A specific
302 sequence in the genome of respiratory syncytial virus regulates the generation of copy-back defective viral
303 genomes. PLoS Pathog 15, e1007707.
304 Timm, C., Akpinar, F., Yin, J., 2014. Quantitative characterization of defective virus emergence by deep
305 sequencing. J Virol 88, 2623-2632.
306 Vasilijevic, J., Zamarreno, N., Oliveros, J.C., Rodriguez-Frandsen, A., Gomez, G., Rodriguez, G., Perez-
307 Ruiz, M., Rey, S., Barba, I., Pozo, F., Casas, I., Nieto, A., Falcon, A., 2017. Reduced accumulation of
308 defective viral genomes contributes to severe outcome in influenza virus infected patients. PLoS Pathog
309 13, e1006650.
310 Vignuzzi, M., Lopez, C.B., 2019. Defective viral genomes are key drivers of the virus-host interaction.
311 Nat Microbiol 4, 1075-1087.
312 Yount, J.S., Kraus, T.A., Horvath, C.M., Moran, T.M., Lopez, C.B., 2006. A novel role for viral-defective
313 interfering particles in enhancing dendritic cell maturation. J Immunol 177, 4503-4513.
314 Zhou, T., Gilliam, N.J., Li, S., Spaudau, S., Osborn, R.M., Anderson, C.S., Mariani, T.J., Thakar, J.,
315 Dewhurst, S., Mathews, D.H., Huang, L., Sun, Y., 2022. Generation and functional analysis of defective
316 viral genomes during SARS-CoV-2 infection. bioRxiv.

317

318

Table 1. Simulation of RSV G copy-back and deletion sequences

Type of NsVG	Break Position	Rejoin Position	Genome Coverage	Total Reads	NsVG Junction Reads	VODKA2	ViReMa	DI-tector
CB	150	795	330	1081	**234**	219	227	212
CB	261	404	180	757	**44**	30	**44**	**44**
CB	274	652	70	234	**31**	**31**	**31**	17
CB	299	799	40	111	**31**	**31**	30	30
CB	391	669	60	174	**35**	24	**35**	**35**
CB	590	810	50	98	**20**	18	**20**	**20**
CB	594	701	30	63	**17**	16	**17**	**17**
CB	689	849	140	184	**74**	**74**	**74**	**74**
CB	822	915	20	13	**6**	**6**	**6**	**6**
CB	832	930	80	46	**21**	**21**	**21**	**21**
CB	***300***	***869***	**-**	**-**	**-**	**-**	***1***	**-**
DEL	190	472	160	363	**71**	**71**	**71**	**71**
DEL	212	800	350	439	**135**	**135**	131	128
DEL	249	549	170	376	**53**	**53**	52	**53**
DEL	251	281	70	217	**26**	**26**	25	0
DEL	273	562	490	1100	**174**	53	**174**	164
DEL	368	522	100	269	**33**	**33**	31	32
DEL	441	751	110	239	**23**	**23**	22	22
DEL	470	497	110	342	**1**	0	0	0
DEL	535	672	180	495	**25**	**25**	**25**	**25**
DEL	692	747	230	695	**22**	21	20	8
DEL	***212***	***872***	**-**	**-**	**-**	**-**	***2***	**-**

Green cells indicate results that are identical to the expected. Red cells indicate false positives.

325 **FIGURE LEGENDS**
326

327 **Figure 1. Description of the VODKA2 process**. (A) Schematic of a standard viral genome (VG)
328 containing 3 genes: A, B, and C, a copy-back VG containing gene C and the complement C' end, and a
329 deletion VG that has the B gene removed. (B-D) Brief descriptions of the steps and output files of the
330 VODKA2 bioinformatics pathway. Gray boxes indicate bioinformatics programs/processes and white
331 boxes indicate output files. (B) The steps of the VODKA2 pipeline that are discussed in the text. (C) Pre-
332 processing steps of RNA-sequencing reads to obtain non-Host Reads.
333

334 **Figure 2. Detection of simulated cbVG and delVG reads**. (A+B) Representation of (A) cbVG and (B)
335 delVG break and rejoin points reported by VODKA2 (pink), ViReMa (yellow) and DI-tector (blue).
336 Expected results are shown in green. Size of dots is dependent on the number of detected read at each
337 break/rejoin junction. False positive reads are shown in red.
338

339 **Figure 3. Detection of cbVG reads in SeV**. Representation of cbVG break and rejoin points reported
340 by VODKA2 (pink), ViReMa (yellow) and DI-tector (blue). Size of each point is dependent on the number
341 of detected read at each break/rejoin species. (A) Shows an overlay of all cbVGs detected across the
342 whole genome by the three softwares. Red dot indicates a false positive from DI-tector. (B) Zooms of
343 species detected within positions 14,800 and 15,384 of the reference genome for each software. False
344 positive reads are shown in red. Processing time of each software is shown in minutes.
345
346

Figure 1

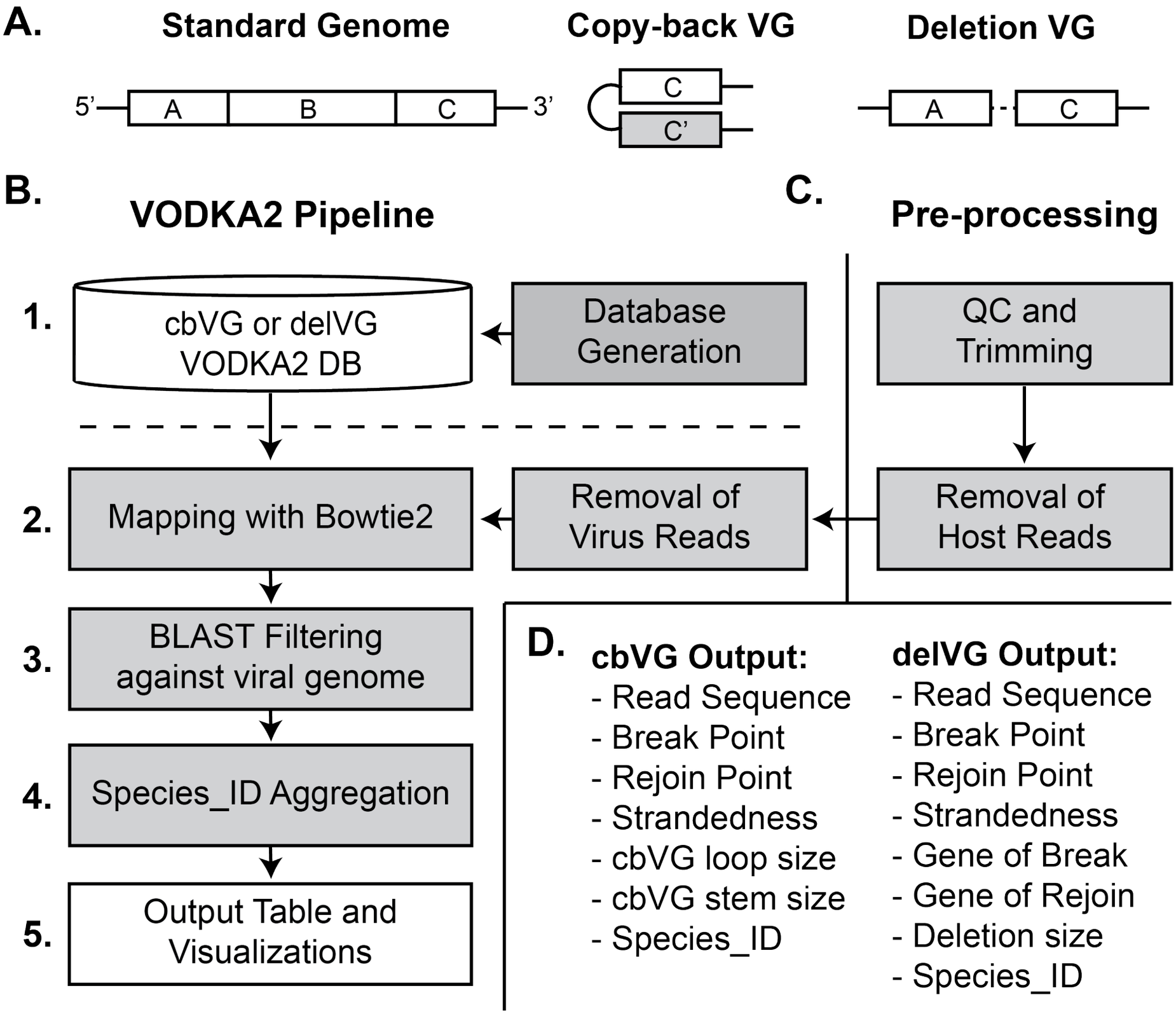

A.
Standard Genome
Copy-back VG
Deletion VG
5'
A
B
C
3'
C
C'
A
C
B.
VODKA2 Pipeline
C.
Pre-processing
1.
cbVG or delVG VODKA2 DB
Database Generation
QC and Trimming
2.
Mapping with Bowtie2
Removal of Virus Reads
Removal of Host Reads
3.
BLAST Filtering against viral genome
4.
Species_ID Aggregation
5.
Output Table and Visualizations
D.
cbVG Output:
- Read Sequence
- Break Point
- Rejoin Point
- Strandedness
- cbVG loop size
- cbVG stem size
- Species_ID
delVG Output:
- Read Sequence
- Break Point
- Rejoin Point
- Strandedness
- Gene of Break
- Gene of Rejoin
- Deletion size
- Species_ID

Figure 2

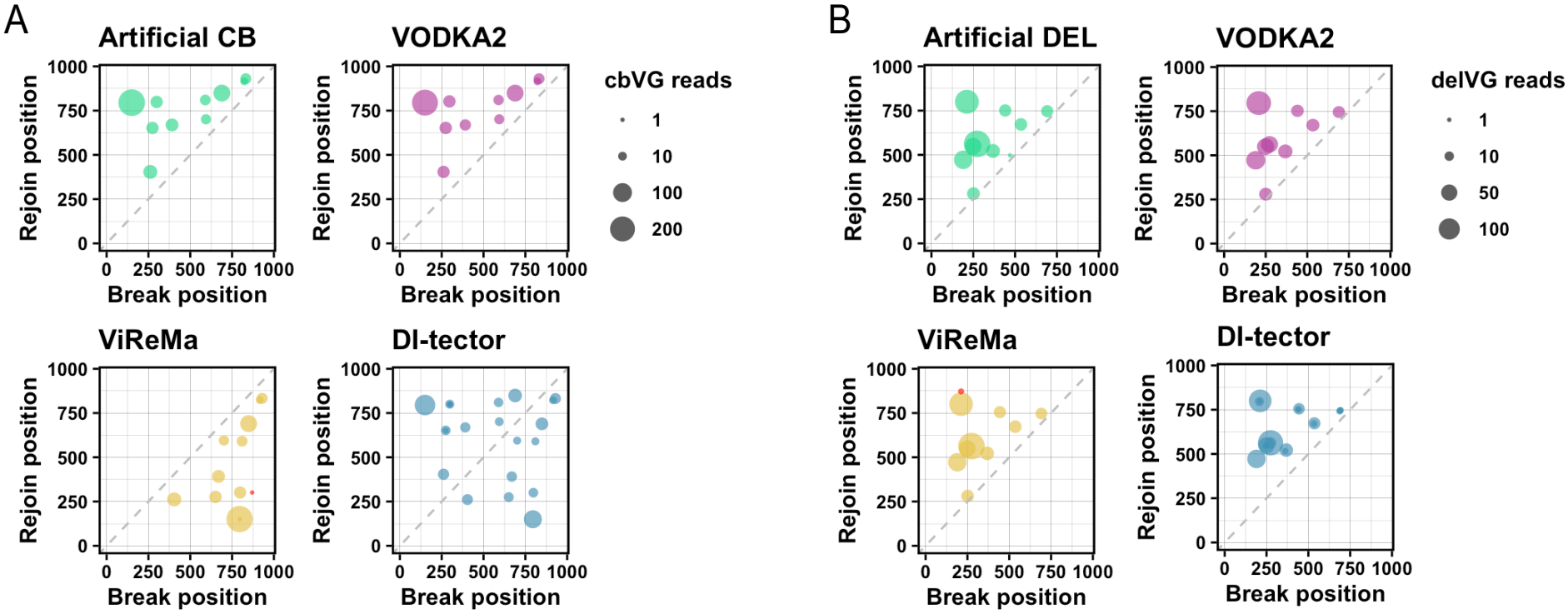
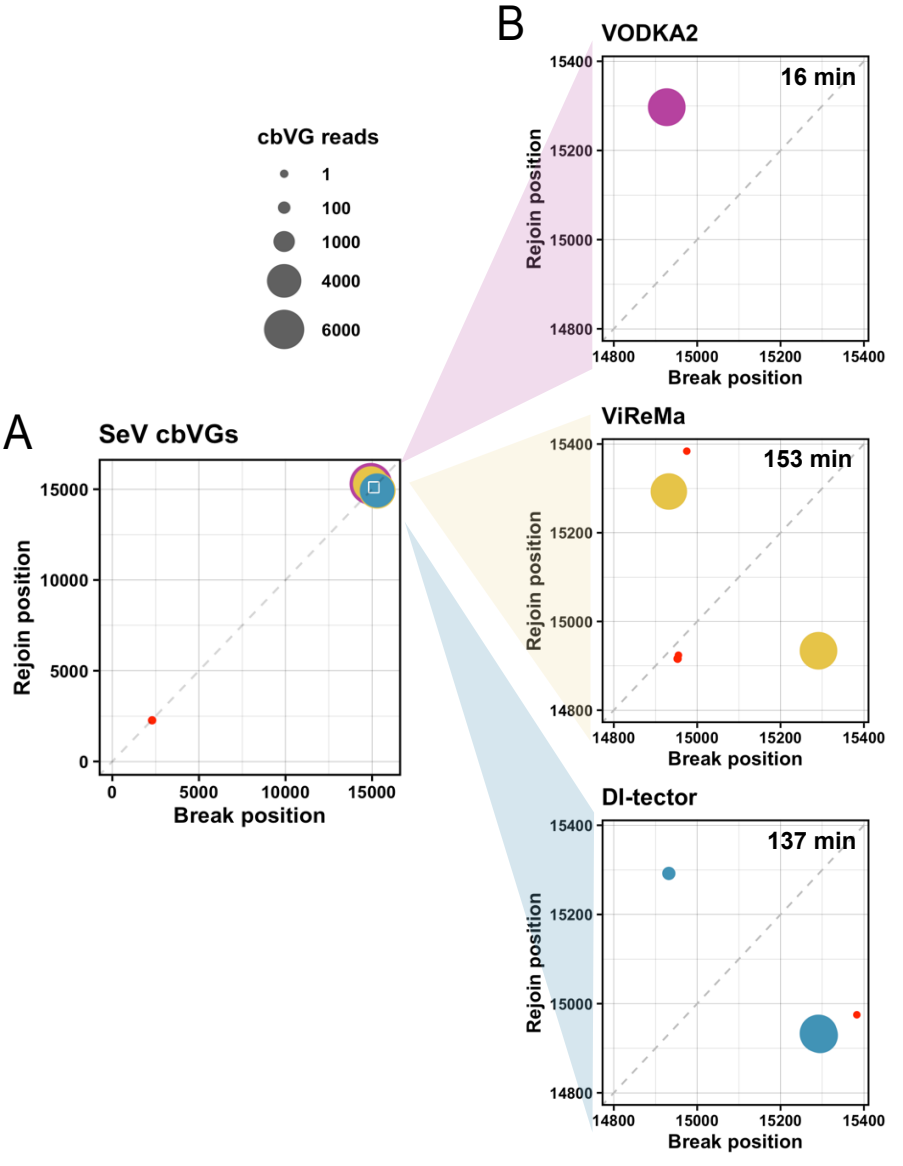

A
Artificial CB
VODKA2
ViReMa
DI-tector
Rejoin position
Break position
0
250
500
750
1000
cbVG reads
1
10
100
200
B
Artificial DEL
VODKA2
ViReMa
DI-tector
Rejoin position
Break position
delVG reads
1
10
50
100

Figure 3


A
SeV cbVGs
Rejoin position
Break position
0
5000
10000
15000
cbVG reads
1
100
1000
4000
6000
B
VODKA2
16 min
ViReMa
153 min
DI-tector
137 min
14800
15000
15200
15400