

1 **The effect of variant interference on *de novo* assembly for viral deep sequencing**

2

3 **Short title: Variant interference in *de novo* assembly**

4

5 Christina J. Castro<sup>1,2</sup>, Rachel L. Marine<sup>1</sup>, Edward Ramos<sup>3</sup>, Terry Fei Fan Ng<sup>1#</sup>

6

7 <sup>1</sup>Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease  
8 Control and Prevention, Atlanta, Georgia, USA

9

10 <sup>2</sup>Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, USA

11

12 <sup>3</sup>General Dynamics Information Technology, Inc., contracting agency to the Office of Informatics, National  
13 Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Falls Church,  
14 VA, USA

15

16 # To whom correspondence should be addressed:

17 Terry Fei Fan Ng

18 Division of Viral Diseases

19 Centers for Disease Control and Prevention

20 1600 Clifton Rd. NE, Mailstop H17-6

21 Atlanta, GA 30329

22 ylz9@cdc.gov, Phone: 404.639.4880, FAX: 404.639.4011

23 **Abstract**

24 Viruses have high mutation rates and generally exist as a mixture of variants in biological samples. Next-  
25 generation sequencing (NGS) approach has surpassed Sanger for generating long viral sequences, yet how  
26 variants affect NGS *de novo* assembly remains largely unexplored. Our results from >15,000 simulated  
27 experiments showed that presence of variants can turn an assembly of one genome into tens to thousands of  
28 contigs. This “variant interference” (VI) is highly consistent and reproducible by ten most used *de novo*  
29 assemblers, and occurs independent of genome length, read length, and GC content. The main driver of VI is  
30 pairwise identities between viral variants. These findings were further supported by *in silico* simulations,  
31 where selective removal of minor variant reads from clinical datasets allow the “rescue” of full viral genomes  
32 from fragmented contigs. These results call for careful interpretation of contigs and contig numbers from *de*  
33 *novo* assembly in viral deep sequencing.

34

## 35 Introduction

36

37 For many years, Sanger sequencing has been used to complement classical epidemiological and  
38 laboratory methods for investigating viral infections.<sup>1</sup> As technologies have evolved, the emergence of next-  
39 generation sequencing (NGS), which drastically reduced the cost per base to generate sequence data for  
40 complete viral genomes, has allowed scientists to apply viral sequencing on a grander scale.<sup>2</sup> Genomic  
41 sequencing is ideal for elucidating viral transmission pathways, characterizing emerging viruses, and locating  
42 genomic regions which are functionally important for evading the host immune system or antivirals.<sup>3</sup>

43

44 Genomic surveillance of viruses is particularly important in light of their rapid rate of evolution. Viruses  
45 have higher mutation rates than cellular-based taxa, with RNA viruses having mutation rates as high as  $1.5 \times$   
46  $10^{-3}$  mutations per nucleotide, per genomic replication cycle.<sup>4</sup> Due to this high mutation rate, it is well  
47 established that most RNA viruses exist as a swarm of quasispecies,<sup>5</sup> with each quasispecies containing unique  
48 single nucleotide polymorphisms (SNPs). The presence of these variants plays a key role in viral adaptation.

49

50 Due to viruses' rapid evolution, a single clinical sample often contains a mixture of many closely related  
51 viruses. Viral quasispecies are mainly derived from intra-host evolution, with RNA viruses such as poliovirus,  
52 human immunodeficiency virus (HIV), hepatitis C (HCV), influenza, dengue, and West Nile viruses maintaining  
53 diverse quasispecies populations within a host.<sup>6, 7, 8, 9, 10, 11, 12, 13</sup> Conversely, the term "viral strains" often refers  
54 to different lineages of viruses found in separate hosts, or a co-infection of viruses in the same host due to  
55 multiple infection events. As a result, sequence divergence is usually higher when comparing viral strains  
56 compared to quasispecies. In this study, we use the term "variant" to encompass both quasispecies and  
57 strains regardless of how the variants originated in the biological samples.

58

59 Since many sequencing technologies produce reads that are significantly shorter than the target  
60 genome size, a process to construct contigs, scaffolds, and full-length genomes is needed. Reference-mapping  
61 and *de novo* assembly are the two primary bioinformatic strategies for genome assembly. Reference-mapping  
62 requires a closely-related genome as input to align reads, while *de novo* assembly generates contigs without

63 the use of a reference genome, and therefore is the most suitable strategy for analyzing underexplored taxa<sup>14</sup>  
64 or for viruses with high mutation and/or recombination rates.

65

66 In this study, we first examined how often NGS and *de novo* assembly were applied in viral sequencing  
67 in GenBank Nucleotide entries ([www.ncbi.nlm.nih.gov/nucleotide/](http://www.ncbi.nlm.nih.gov/nucleotide/)). Then we investigated how the presence  
68 of variants affected assembly results - simulated and clinical NGS datasets were analyzed using multiple  
69 assembly programs to explore the effects of genome variant relatedness, read length, and genome length on  
70 the resulting contig distribution.

## 71 Results

72

### 73 The rise of NGS and *de novo* assembler use in GenBank viral sequences

74

75 GenBank viral entries from 1982-2017 were collected and analyzed, with extensive analyses performed  
76 to evaluate technologies and bioinformatics programs cited in records deposited between 2011 and 2017.  
77 Through 2017, there were over 2.3 million viral entries in GenBank; however, over 70% (1.7 million) do not  
78 specify a sequencing technology [Supplement Table S1] due to the looser data requirement in earlier years.  
79 When looking at recently deposited records (2014-2017), the Illumina sequencing platform was the most  
80 common NGS platform used for viral sequencing, with about a 2-fold increase over the next most popular NGS  
81 platform [Figure 1d & e]. When long sequences ( $\geq 2,000$  nt) are considered, NGS technologies surpassed  
82 Sanger in 2017 as the dominant strategy for sequencing, comprising 53.8% (14,653/27,217) of entries  
83 compared to 46.2% of entries (12,564/27,217) for Sanger [Figure 1f and Supplement Table S2].

84

85 Hybrid sequencing approaches, where researchers use more than one sequencing technology to  
86 generate complete viral sequences, have also become more common over the past several years. The most  
87 common combination observed was 454 and Sanger (18,002 entries), likely due to the early emergence of the  
88 454 technology compared to other NGS platforms [Figure 1c and Supplement Table S3]. However, combining  
89 Illumina with various other sequencing platforms is quite commonplace ( $>10,000$  entries).

90

91 *De novo* assembly programs (ABYSS, BWA, Canu, Cap3, IDBA, MIRA, Newbler, SOAPdenovo, SPAdes,  
92 Trinity, and Velvet) have increased from less than 1% of viral sequence entries in 2012, to 20% of all viral  
93 sequence entries in 2017 [Figure 1h & i]. A similar increase was observed for reference-mapping programs  
94 (i.e., Bowtie and Bowtie2), from 0.03% in 2012 to 6.5% in 2017. Multifunctional programs (Suppl. Information)  
95 that offer both assembly options were the most common programs cited for the years 2013-2017, but since  
96 the exact sequence assembly strategy used for these records is unknown, the contributions of *de novo*  
97 assembly are likely underestimated. An expanded summary of the sequencing technologies and assembly  
98 approaches used for viral GenBank records is available in Supplement Tables S1-S6.

99

## 100 **Effect of variant assembly using popular *de novo* assemblers**

101

102 After establishing the growing use of NGS technologies for viral sequencing, we next focused on  
103 understanding how the presence of viral variants may influence *de novo* assembly output. We generated 247  
104 simulated viral NGS datasets representing a continuum of pairwise identity (PID) between two viral variants,  
105 from 75% PID (one nucleotide difference every 4 nucleotides), to 99.6% PID (one nucleotide difference every  
106 250 nucleotides) [Figure 2]. For Experiment 1, these datasets were assembled using 10 of the most used *de*  
107 *nov*o assembly programs [Figure 2 and Supplement Figure S1a] to evaluate their ability to assemble the two  
108 variants into their own respective contigs as the PID between the variants increases.

109

110 One key observation is that the assembly result can change from two (correct) contigs to many  
111 (unresolvable) contigs simply by having variant reads; the presence of viral variants affected the contig  
112 assembly output of all 10 assemblers tested. The output of the SPAdes, MetaSPAdes, ABySS, Cap3, and IDBA  
113 assemblers shared a few commonalities, demonstrated by a conceptual model in Figure 3A. First, below a  
114 certain PID, when viral variants have enough distinct nucleotides to resolve the two variant contigs, the *de*  
115 *nov*o assemblers produced two contigs correctly [Figure 3]. We refer to this as “variant distinction” (VD), with  
116 the highest pairwise identity where this occurs as the VD threshold. Above this threshold, the assemblers  
117 produced tens to thousands of contigs [Figure 3], a phenomenon we define as “variant interference” (VI). As  
118 PID between the variants continue to increase, the *de novo* assemblers can no longer distinguish between the  
119 variants and assembled all the reads into a single contig, a phenomenon we define as “variant singularity”  
120 (VS). [Figure 3]. The lowest pairwise identity where a single contig is assembled is the VS threshold.

121

122 Slight differences in the variant interference patterns (relative to the canonical variant interference  
123 model) were observed for the 10 assemblers investigated. VD was observed for SPAdes, MetaSPAdes, and  
124 ABySS assemblers. While it was not observed with Cap3 and IDBA with the current simulated data parameters,  
125 we speculate that VD may occur at a lower PID level for these assemblers than tested in this study. The PID  
126 range where VI was observed was distinct for each *de novo* assembler [Figure 3]. During VI, SPAdes produced  
127 as many as 134 contigs and ABySS produced 3,076 contigs, while MetaSPAdes, Cap3, and IDBA produced up to  
128 10.

129

130 A different pattern was observed for Mira, Trinity, and SOAPdenovo2 assemblers. The average number  
131 of contigs generated by Mira, Trinity, and SOAPdenovo2 was 5, 36, and 283, respectively across all variant PIDs  
132 from 75%–99.96%. Specifically, Mira and Trinity generated fewer contigs at low PID, but produced many  
133 contigs when the two variants reach 97.1% PID and 96.0% PID, respectively. For SOAPdenovo2, a larger  
134 number of contigs were produced regardless of the PID. This indicates that these assemblers generally have  
135 major challenges producing a single genome; this has been observed in previous studies comparing assembly  
136 performance.<sup>15</sup>

137

138 Finally, Geneious and CLC were the least affected by VI in the simulated datasets tested, returning only  
139 1–5 contigs for all pairwise identities. CLC’s assembly algorithm primarily returned a single contig over the  
140 range of PIDs tested (218/247 simulations; 88.3%), thus favoring VS. In comparison, Geneious predominantly  
141 distinguished the two variants (234/247 simulations; 94.7%), favoring VD.

142

### 143 **Effect of GC content and genome length on variant assembly**

144

145 For Experiment 2, we focused our study on evaluating whether VI observed in SPAdes *de novo*  
146 assembly is influenced by the GC content or genome length of the pathogen. Two datasets were used for the  
147 evaluation: reads generated from four artificial genomes ranging in length from 2 Kb to 1 Mb, as well as from  
148 genome sequences of poliovirus (NC\_002058; 7,440 nt in length) and coronavirus (NC\_002645; 27,317 nt in  
149 length). No discernable correlation was observed between the GC content of variant genomes and the degree  
150 of VI for any of the simulated datasets [[Supplemental Dataset S2](#),  $p < 0.0001$ ]. Therefore, for subsequent  
151 analyses examining the effects of genome length on VI, the number of contigs at each PID level was obtained  
152 by averaging the 13 GC simulations.

153

154 Notably, no matter the genome length, SPAdes produced vastly more contigs (i.e., VI) in a constant,  
155 narrow range of PID [99%–99.21% ; [Figure 4a & b](#)]. The effect of variants on assembly was characterized by  
156 the three distinct intervals described previously: VD at lower PIDs, VI [[Figure 4b](#)], and VS at higher PIDs for all  
157 genome lengths. For example, during VS, a single contig was generated when the two variants shared  $\geq 99.22\%$

158 PID, but tens to thousands of contigs were generated at a slightly lower PID of 99.21%. This PID threshold,  
159 99.21%, marked the drastic transition from VS to VI, whereas the transition from VI to VD (i.e., the VD  
160 threshold) occurred at 98.99% PID [Figure 4b]. A correlation was observed between genome length and the  
161 number of contigs produced during VI, where longer genomes returned proportionally more contigs as  
162 expected as total VI occurrence should increase with length [ $r^2 = 0.967$ ;  $p < 0.0001$  Figure 4b and 4c].

163

#### 164 **Effect of read length on variant assembly**

165

166 The read length of a given NGS dataset will vary depending on the sequencing platform and kits utilized  
167 to generate the data. Since read length is an important factor for *de novo* assembly success,<sup>16</sup> we  
168 hypothesized that it may also influence the ability to distinguish viral variants. For Experiment 3, using SPAdes  
169 we investigated assemblies with four typical read lengths: 50, 100, 150, and 250 nt. At longer read lengths,  
170 the VD threshold occurred at higher PIDs [Figure 4d & e]. Also, with increasing read length, the width of the  
171 PID window where VI occurs gradually decreased from a 1.52% spread to a 0.21% spread [Figure 4e]. This  
172 indicates that longer reads are better for distinguishing viral variants with high PIDs.

173

#### 174 ***In silico* experiments examining variant assembly with NGS data derived from clinical samples**

175

176 For clinical samples, assembly of viral genomes is affected by multiple factors other than the presence  
177 of variants, including sequencing error rate, host background reads, depth of genome coverage, and the  
178 distribution (i.e., pattern) of genome coverage. We next utilized viral NGS data generated from four  
179 picornavirus-positive clinical samples (one coxsackievirus B5, one enterovirus A71, and two parechovirus A3)  
180 to explore VI in datasets representative of data that may be encountered during routine NGS. The NGS data  
181 for each sample was partitioned into four bins of read data: (1) total reads after quality control (**T**); (2) major  
182 variants only (**M**); (3) major and minor variants only (**Mm**); and (4) major variants and background non-viral  
183 reads only (**MB**) [Figure 5]. These binned datasets were then assembled separately using three assembly  
184 programs: SPAdes, Cap3, and Geneious. By comparing these manipulations, we aimed to test the hypothesis  
185 that minor variants directly affect the performance of assembly through VI in real clinical NGS data.

186



187 Even with an adequate depth of coverage for genome reconstruction, assembly of total reads (**T**) in  
188 11/12 experiments resulted in unresolved genome construction – resulting in numerous fragmented viral  
189 contigs [Figure 6]. The only exception was one experiment where one single PeV-A3 (S1) genome was  
190 assembled using Cap3. When only reads from the major variant were assembled (**M**), full genomes were  
191 obtained for all datasets using SPAdes and Cap3, and for the CV-B5 sample using Geneious. Conversely,  
192 assembly of the read bins containing major and minor variants (**Mm**) resulted in an increased number of  
193 contigs for 9 of the 12 sample and assembly software combinations tested [Figure 6], indicating that VI due to  
194 the addition of the minor variant reads likely adversely affected the assembly. The presence of background  
195 reads with major variant reads (**MB**) did not appear to affect viral genome assembly, as the  $UG_{50\%}$  value, a  
196 performance metric which only considers unique, non-overlapping contigs for target viruses<sup>17</sup>, was similar  
197 between **M** and **MB** datasets.

198

## 199 Discussion

200

201 Our analysis of the GenBank quantified the decade-long expansion of NGS technologies and *de novo*  
202 assembly for viral sequencing [Figure 1]. As the number of viral sequences in public databases continues to  
203 grow, an important question that naturally arises is how well current *de novo* assembly programs perform for  
204 datasets with viral variants. Viral variants are expected in biological samples, with the number of variants and  
205 the extent of the sequence divergence between variants related to the mutation rate of the virus and the  
206 types of specimens that are being investigated. For example, samples containing rapidly evolving RNA viruses,  
207 such as poliovirus, HIV, and HCV<sup>7, 9, 18</sup>, environmental samples,<sup>19</sup> and clinical samples from immunosuppressed  
208 individuals<sup>20, 21</sup> usually harbor many variants. The ability to accurately distinguish variants is imperative to  
209 inform treatments (in the case of HIV and HCV), or determine whether a subpopulation of a more virulent  
210 variant is present.

211

212 Several experiments using simulated and clinical sample NGS data were performed to evaluate the  
213 ability of genome assembly programs to distinguish genome variants. All assemblers investigated generated  
214 fragmented assemblies when the data contained reads from two closely related variants due to “variant  
215 interference” (VI). Changes in pairwise identity (PID) as small as 0.01% between the two variants triggered an

216 assembler to change from producing one or two contigs to producing hundreds of contigs. A quintessential  
217 example of this phenomenon was the SPAdes assembly of EV-A71 sequences during the *in silico* experiments  
218 with clinical NGS data. Assembly of major variant reads resulted in one full length contig [Figure 5], whereas  
219 assembly of datasets containing the major and minor variant reads (**Mm** and **T**) were characterized by a  
220 number of contigs, resulting in “cobwebs” of contig fragments when visualized using Bandage [Supplement  
221 Figure S2].<sup>22</sup> Even though the *de novo* assembly graph linked the different contig fragments, the assembly  
222 could not differentiate the multiple routes of possible contig construction. We speculate this is the main  
223 reason why VI occurs in the context of de Bruijn graph assemblers.

224

225 The simulated experiments suggested that genome length and read length influence VI; A longer  
226 genome length will produce proportionally more contigs during VI, whereas a longer read length decreases  
227 the PID range where VI occurs [Figure 4]. While longer read length improves assembly, unfortunately,  
228 platforms that produce long reads such as Oxford Nanopore and PacBio have higher error rates.<sup>23</sup> Until long  
229 reads can be produced at high fidelity, researchers must continue to rely on combining long- and short-read  
230 NGS datasets, and genome polishing techniques.<sup>23</sup>

231

232 The large number of contigs generated due to VI may be overwhelming for most researchers, and for  
233 viral ecology studies, could lead to over-estimation of species richness for methods that use contig spectra to  
234 infer richness, such as PHACCS or CatchAll.<sup>24, 25, 26</sup> This phenomenon may also impact studies differently  
235 depending on the overall goal for generating viral sequence data. For example, some researchers may only be  
236 concerned with generating a single major consensus genome, even when variants are detected in the data.  
237 This is common during outbreak responses for pathogens such as Ebola virus or Middle East respiratory  
238 syndrome coronavirus, where detection of SNPs (indicative of minor variants) is not immediately important.  
239 On the other hand, some investigations could favor distinguishing variants, such as for investigating the  
240 presence of vaccine-derived poliovirus, where a small number of SNPs may distinguish a vaccine-derived strain  
241 from a normal vaccine strain genome.<sup>21</sup>

242

243 The effects of VI could potentially be mitigated by running multiple assembly programs. A previous  
244 study testing bioinformatics strategies for assembling viral NGS data found that employing sequential use of

245 de Bruijn graph and overlap-layout-consensus assemblers produced better assemblies.<sup>15</sup> We speculate that  
246 this “ensemble strategy”<sup>15</sup> may perform better because the multiple assemblers complement one another by  
247 having different VI PID thresholds. Future assembly approaches could also consider resolving the VI problem  
248 by possibly discriminating the major and minor variant reads first (perhaps by coverage or SNP analysis), and  
249 then assembling major and minor variant reads separately.

250

251 Since we observed VI occurring in simulated data from 2 Kb to 1 Mb genome lengths, we speculate  
252 that it may not only affect viral data but also larger draft contigs of bacteria and other microorganisms. Even  
253 though bacterial mutation rates are much lower than those of most viruses, bacterial variants are common.  
254 For environmental studies, bacterial metagenomes are known to contain many related taxa and variants<sup>27, 28,</sup>  
255 <sup>29, 30</sup>, and in clinical investigations, minor bacterial variants can harbor SNPs that provide resistance against  
256 antimicrobials. This warrants future investigation into how the presence of variants may impact the assembly  
257 of other microbial datasets.

258

259 This study aimed to understand how variants affect assembly. As an initial investigation, many  
260 confounding factors were simplified for experimentation. Simulated variants studied here only depicted  
261 periodic mutations, set at regular intervals. However, in real viral data, SNPs are never evenly distributed  
262 across the genome, with zones of divergence and similarity.<sup>31, 32</sup> Other important factors which influence  
263 genome assembly include sequencing error rates, presence of repetitive regions, and coverage depth. We  
264 limited our experiments to keep these factors constant in order to investigate the sole effect of VI. Through  
265 this exploration, we demonstrated that reads from related genome variants adversely affect *de novo*  
266 assembly. As NGS and *de novo* assembly have become essential for generating full-length viral genomes,  
267 future studies should investigate the combined effects of the number and relative proportion of minor  
268 variants, as well as additional assembly factors (e.g., error rates) to supplement this work.

269

## 270 **Methods**

271

### 272 **Analyzing NGS and assembler usage in the virus nucleotide collection in GenBank**

273           Viral sequence entries from the GenBank non-redundant nucleotide collection were obtained by  
274 downloading all sequences under the virus taxonomy through the end of 2017. A total of 2,338,775 GenBank  
275 entries were investigated.

276

277           The total number of viral sequences submitted annually in GenBank through December 2017 was  
278 calculated by filtering GenBank submissions by “virus,” followed by application of the following additional  
279 filtering steps: “genomic DNA/RNA” was selected and a “release date: Jan 1 through Dec 31” was applied to  
280 find the total number of viruses for a given year. A custom script was used to filter and count all documented  
281 sequencing technologies and assembly methods used for each GenBank entry.

282

### 283 **Creation of simulated variant genomes and reads**

284           Simulated genomes were generated using custom scripts that randomly assign each nucleotide over a  
285 designated genome length with a weighted distribution dependent on the GC content [Supplement Figure S1].  
286 The random genomes were then screened using NCBI BLAST to insure no similarity/identity existed to any  
287 classified organism (i.e., no BLAST hits). These simulated genomes served as the initial variant genome (variant  
288 1). To generate the mutated variant genomes (variant 2), a custom script was used to systematically introduce  
289 evenly distributed random mutations at rates from 1 mutation in every 4 nucleotides (75% PID) to 1 mutation  
290 in every 250 nucleotides (99.6% PID), incrementing by 1 nucleotide.

291

292           Following the generation of initial and mutated variant genomes, high-quality fastq reads were  
293 generated using ART,<sup>33</sup> simulating Illumina MiSeq paired-end runs at 50X coverage with 250 nt reads,  
294 DNA/RNA mean fragments size of 500, and quality score of 93. Fastq reads were combined in equal numbers  
295 for the initial and mutated variants, and used as input for subsequent *de novo* assembly experiments  
296 [Supplement Figure S1]. The same process was utilized to generate the artificial genomes, initial and mutated  
297 variant genomes, and reads for each of the experiments.

## 298 **Experiment 1: Analyzing simulated reads from variants using different *de novo* assembly programs**

299

300 The simulated datasets containing reads from two variant genomes with nucleotide pairwise identity  
301 ranging from 75%–99.6% were analyzed using 10 different genome assembly programs. The *de novo* assembly  
302 algorithms used were either overlap-layout-consensus (OLC) [Cap<sup>34</sup> and Mira<sup>35, 36</sup>], de Bruijn graph (DBG)  
303 [ABYSS<sup>37</sup>, IDBA<sup>38</sup>, MetaSPAdes<sup>39</sup>, SOAPdenovo2<sup>40</sup>, SPAdes<sup>41</sup>, and Trinity<sup>42</sup>], or commercial software packages  
304 [CLC (<https://www.qiagenbioinformatics.com/>) and Geneious<sup>43</sup>] whose assembly algorithms are proprietary  
305 [Supplement Table S6]. The simulation settings for the reads were single-end reads, 250 nt read length, and  
306 50X coverage. A total of 2,470 assemblies (247 datasets per genome X 10 assemblers) were analyzed  
307 [Supplement Figure S1a].

308

## 309 **Experiment 2: Simulated data by varying genome length and GC content**

310

311 Artificial genomes were constructed for four genome lengths: 2 Kb, 10 Kb, 100 Kb, and 1 Mb, with  
312 varying GC content from 20%–80%, in 5% increments [Supplement Figure S1b]. Datasets derived using one  
313 poliovirus genome (NC\_002058) and one coronavirus genome (NC\_002645) were also included in this analysis,  
314 representing the lower and upper genome length range typical of RNA viruses. The original GC content was  
315 kept constant for the poliovirus and coronavirus genomes. For all of these genomes, simulated reads for initial  
316 and mutated variants were generated as above.

317

318 A total of 13,338 SPAdes assemblies were generated, which included 12,844 assemblies for the four  
319 artificial genomes (247 datasets per genome X 4 artificial genome lengths X 13 GC content proportions X 1  
320 assembler) and 494 assemblies for the poliovirus and coronavirus datasets (247 datasets per genome X 2  
321 genomes X 1 assembler) [Supplement Figure S1b]. JMP v13.0.0 ([www.sas.com](http://www.sas.com)) was used to calculate  
322 Pearson's correlation and Spearman's  $\rho$  values to compare the association between percent GC levels and the  
323 number of contigs produced at each PID level. Since there was little statistical difference when comparing the  
324 contig numbers generated at varying percent GC for each of the four genome length datasets (Spearman's  $\rho$  =  
325 0.8299 to 0.9801,  $p < 0.001$ ) [Supplement Excel file], the final contig number was averaged across the 13 GC

326 percentages at a given PID. The average contig number was used for plotting the contig assembly results vs  
327 percent PID for each simulated genome length [Figures 4a-b].

328

### 329 **Experiment 3: Simulated data by varying read length**

330

331 Genome variants were generated as described above (“Creation of simulated variant genomes and  
332 reads”) for a genome of size 100 Kb with 50% GC; this was the starting initial variant genome. In this  
333 simulation, initial and mutation variant reads at four sequencing read lengths (50, 100, 150, and 250 nt) were  
334 created using ART. A total of 538 SPAdes assemblies were generated (47, 97, 147, and 247 datasets for the 50,  
335 100, 150 and 250 nt read lengths, respectively) [[Supplement Figure S1c](#)].

336

### 337 **Evaluation of NGS datasets from clinical samples**

338

339 Four datasets derived from clinical samples containing picornaviruses (one enterovirus A71 [EV-A71],  
340 one coxsackievirus B5 [CV-B5] and two parechovirus A3 [PeV-A3]) were analyzed for this experiment, as  
341 previous sequencing analysis using Geneious indicated the presence of genome variants. The datasets were  
342 analyzed using an in-house pipeline (VPipe),<sup>18</sup> which performs various quality control (QC) steps and *de novo*  
343 assembly using SPAdes. The post-QC reads were considered total reads (**T**) and mapped to their respective  
344 reference genome in order to determine the major and minor variants present in each sample. Total reads  
345 which mapped with high similarity ( $\geq 99\%$ ) to the major variant were categorized as reads representing the  
346 major variant (**M**). Unbinned reads from the major variant reference recruitment were used to construct the  
347 minor variant consensus using a second round of reference recruitment, and these reads were categorized as  
348 the minor variant (**m**). Remaining reads from the previous two steps were considered background (**B**) reads.

349

350 *De novo* assembly for each of the four clinical samples was performed for the following binned NGS  
351 datasets: (1) total reads only (**T**); (2) major variants only (**M**); (3) major and minor variants only (**Mm**); and (4)  
352 major variants and background reads only (**MB**). This was repeated with three assembly programs: SPAdes,  
353 Cap3, and Geneious. The length of the longest contig produced from each assembly and the performance

354 metric  $UG_{50}\%$ .<sup>17</sup> were calculated to compare the results for these 48 assemblies (4 experiments X 4 viruses X 3  
355 assemblers).

356

### 357 **Data Availability**

358 Sequencing reads for the experiments conducted using clinical specimens are available through the NCBI  
359 Sequence Read Archive (SRA) accession PRJNA577924. Reads from simulated datasets (Experiments 1-3) are  
360 available upon request.

361

### 362 **Funding Information**

363 This work was supported in part by Federal appropriations to the Centers for Disease Control and Prevention,  
364 through the Advanced Molecular Detection Initiative line item.

365

### 366 **Acknowledgements**

367 We thank Dr. Steve Oberste for thoughtful suggestions on this work.

368

### 369 **Author contributions**

370 All authors contributed to the conceptualization, data analysis, preparation, and review of this manuscript.

371 C.J.C, R.L.M., and T.F.F.N. wrote this manuscript.

372

### 373 **Competing interests**

374 The authors declare no competing interests.

375 **References**

376

377 1. Rasmussen AL, Katze MG. Genomic Signatures of Emerging Viruses: A New Era of Systems Epidemiology. *Cell*  
378 *Host Microbe* **19**, 611-618 (2016).

379

380 2. Leung P, Eltahla AA, Lloyd AR, Bull RA, Luciani F. Understanding the complex evolution of rapidly mutating  
381 viruses with deep sequencing: Beyond the analysis of viral diversity. *Virus Res* **239**, 43-54 (2017).

382

383 3. Pierce BG, Keck ZY, Fong SK. Viral evasion and challenges of hepatitis C virus vaccine development. *Curr Opin*  
384 *Viro* **20**, 55-63 (2016).

385

386 4. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev*  
387 *Genet* **9**, 267-276 (2008).

388

389 5. Andino R, Domingo E. Viral quasispecies. *Virology* **479-480**, 46-51 (2015).

390

391 6. Henn MR, *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon  
392 immune recognition during acute infection. *PLoS pathogens* **8**, e1002529-e1002529 (2012).

393

394 7. Herbeck JT, *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of  
395 selective processes. *Journal of virology* **85**, 7523-7534 (2011).

396

397 8. Jerzak G, Bernard KA, Kramer LD, Ebel GD. Genetic variation in West Nile virus from naturally infected  
398 mosquitoes and birds suggests quasispecies structure and strong purifying selection. *The Journal of general*  
399 *virology* **86**, 2175-2183 (2005).

400

401 9. Lauck M, *et al.* Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep  
402 pyrosequencing. *Journal of virology* **86**, 3952-3960 (2012).

403

404 10. Lin S-R, *et al.* Study of sequence variation of dengue type 3 virus in naturally infected mosquitoes and human  
405 hosts: implications for transmission and evolution. *Journal of virology* **78**, 12717-12721 (2004).

406

407 11. Murcia PR, *et al.* Intra- and interhost evolutionary dynamics of equine influenza virus. *Journal of virology* **84**,  
408 6943-6954 (2010).

409

410 12. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through  
411 cooperative interactions in a viral population. *Nature* **439**, 344-348 (2006).

412



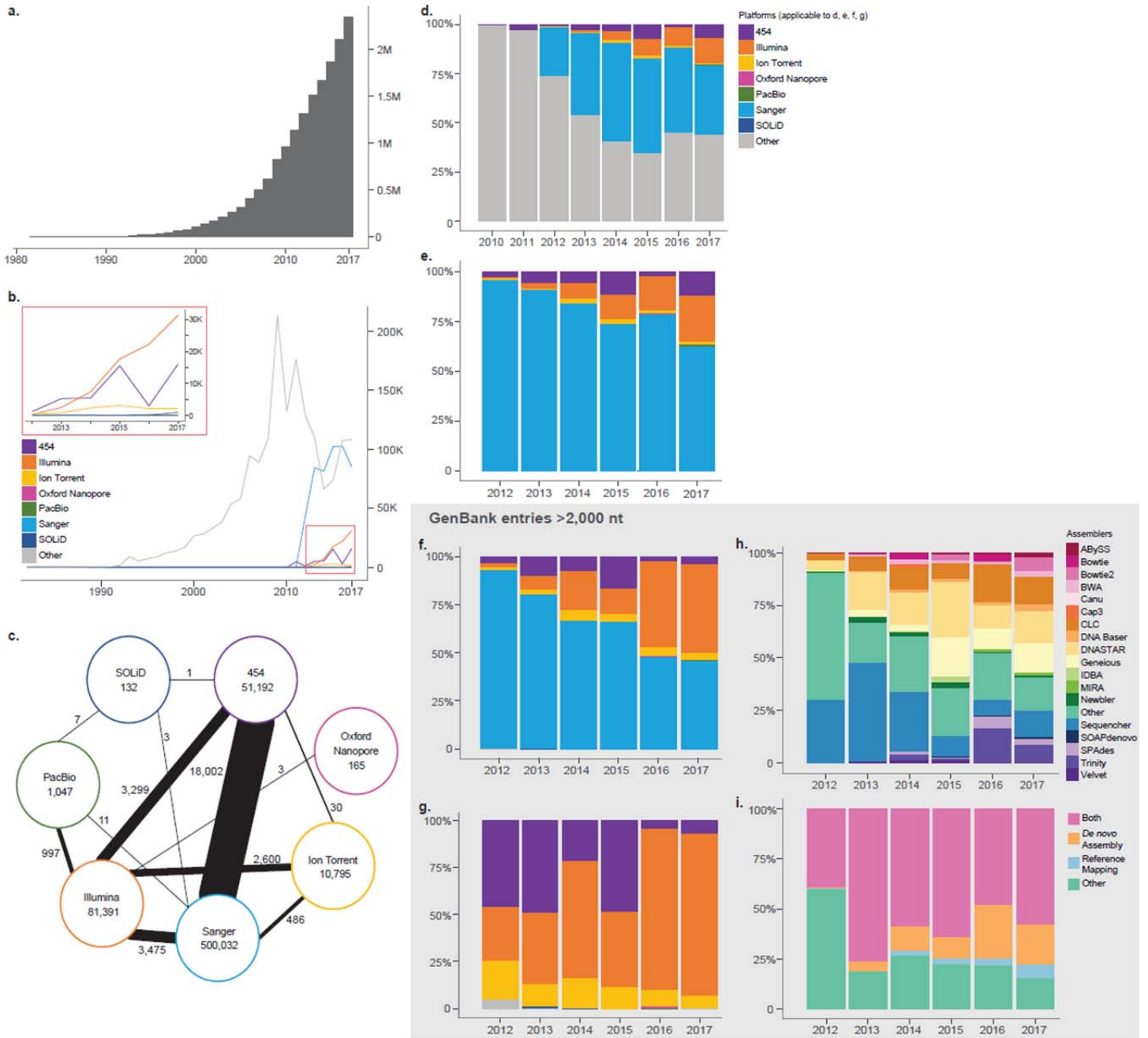
- 413 13. Thai KTD, *et al.* High-resolution analysis of intrahost genetic diversity in dengue virus serotype 1 infection  
414 identifies mixed infections. *Journal of virology* **86**, 835-843 (2012).
- 415
- 416 14. Yang X, *et al.* De novo assembly of highly diverse viral populations. *BMC Genomics* **13**, 475 (2012).
- 417
- 418 15. Deng X, *et al.* An ensemble strategy that significantly improves de novo assembly of microbial genomes from  
419 metagenomic next-generation sequencing data. *Nucleic Acids Res* **43**, e46 (2015).
- 420
- 421 16. Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. *Applied and environmental microbiology*  
422 **74**, 1453-1463 (2008).
- 423
- 424 17. Castro CJ, Ng TFF. U50: A New Metric for Measuring Assembly Output Based on Non-Overlapping, Target-  
425 Specific Contigs. *J Comput Biol* **24**, 1071-1080 (2017).
- 426
- 427 18. Montmayeur AM, *et al.* High-throughput next-generation sequencing of polioviruses. *J Clin Microbiol* **55**, 606-  
428 615 (2017).
- 429
- 430 19. Ng TFF, *et al.* High Variety of Known and New RNA and DNA Viruses of Diverse Origins in Untreated Sewage.  
431 *Journal of Virology* **86**, 12161 (2012).
- 432
- 433 20. Ma S, Du Z, Feng M, Che Y, Li Q. A severe case of co-infection with Enterovirus 71 and vaccine-derived Poliovirus  
434 type II. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* **72**,  
435 25-29 (2015).
- 436
- 437 21. Jorba J, *et al.* Update on Vaccine-Derived Polioviruses - Worldwide, January 2017-June 2018. *MMWR Morbidity*  
438 *and mortality weekly report* **67**, 1189-1194 (2018).
- 439
- 440 22. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies.  
441 *Bioinformatics (Oxford, England)* **31**, 3350-3352 (2015).
- 442
- 443 23. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics &*  
444 *Bioinformatics* **14**, 265-279 (2016).
- 445
- 446 24. Herath D, Jayasundara D, Ackland D, Saeed I, Tang SL, Halgamuge S. Assessing Species Diversity Using  
447 Metavirome Data: Methods and Challenges. *Comput Struct Biotechnol J* **15**, 447-455 (2017).
- 448
- 449 25. Bunge J, Woodard L, Bohning D, Foster JA, Connolly S, Allen HK. Estimating population diversity with CatchAll.  
450 *Bioinformatics* **28**, 1045-1047 (2012).

- 451  
452 26. Angly F, *et al.* PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities  
453 using metagenomic information. *BMC Bioinformatics* **6**, 41 (2005).
- 454  
455 27. Wang NF, *et al.* Diversity and Composition of Bacterial Community in Soils and Lake Sediments from an Arctic  
456 Lake Area. *Frontiers in microbiology* **7**, 1170-1170 (2016).
- 457  
458 28. Rusch DB, *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical  
459 Pacific. *PLoS biology* **5**, e77 (2007).
- 460  
461 29. The Human Microbiome Project C, *et al.* Structure, function and diversity of the healthy human microbiome.  
462 *Nature* **486**, 207 (2012).
- 463  
464 30. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project.  
465 *Nature* **449**, 804 (2007).
- 466  
467 31. Schneider WL, Roossinck MJ. Genetic Diversity in RNA Virus Quasispecies Is Controlled by Host-Virus  
468 Interactions. *Journal of Virology* **75**, 6566 (2001).
- 469  
470 32. Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures.  
471 *Virology* **493**, 227-237 (2016).
- 472  
473 33. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford,*  
474 *England)* **28**, 593-594 (2012).
- 475  
476 34. Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-877 (1999).
- 477  
478 35. Chevreux B, *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP  
479 detection in sequenced ESTs. *Genome Res* **14**, 1147-1159 (2004).
- 480  
481 36. Chevreux BW, T.; Suhai, S. Genome sequence assembly using trace signals and additional sequence information.  
482 *German conference on bioinformatics* **99**, 45-56 (1999).
- 483  
484 37. Jackman SD, *et al.* ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res* **27**,  
485 768-777 (2017).
- 486  
487 38. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In:  
488 *Research in Computational Molecular Biology* (ed. Berger B). Springer Berlin Heidelberg (2010).

- 489  
490 39. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler.  
491 *Genome Research* **27**, 824-834 (2017).
- 492  
493 40. Luo R, *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.  
494 *GigaScience* **1**, 18 (2012).
- 495  
496 41. Bankevich A, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J*  
497 *Comput Biol* **19**, 455-477 (2012).
- 498  
499 42. Grabherr MG, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat*  
500 *Biotechnol* **29**, 644-652 (2011).
- 501  
502 43. Kearse M, *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization  
503 and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).
- 504  
505 44. Phillippy AM. New advances in sequence assembly. *Genome research* **27**, xi-xiii (2017).
- 506  
507 45. Olivarius S, Plessy C, Carninci P. High-throughput verification of transcriptional starting sites by Deep-RACE.  
508 *BioTechniques* **46**, 130-132 (2009).
- 509  
510 46. Lagarde J, *et al.* Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput  
511 sequencing (RACE-Seq). *Nature Communications* **7**, 12339 (2016).
- 512  
513

514 **Figures and legends**

515



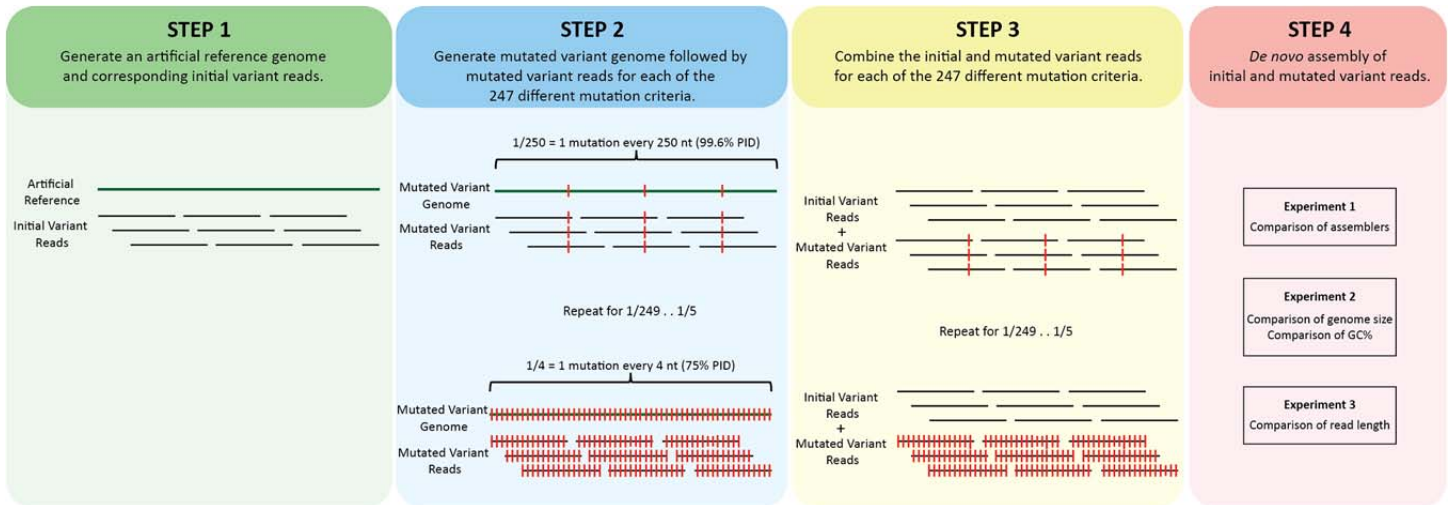
516

517

518 **Figure 1. Trends and patterns of sequencing technology and assembly methods of viral entries in the**  
519 **GenBank database. (a)** Cumulative frequency histogram of all viral entries in GenBank from Jan. 1, 1982  
520 through Dec. 31, 2017 (total=2,338,775 entries). **(b)** Count of all viral entries with at least one *Sequencing*  
521 *Technology* documented for the years 1982-2017. For panels (b) and (d), the “Other” category denotes entries  
522 with the *Sequencing Technology* field omitted or mis-assigned. **(c)** Relationship between viral entries listing  
523 one or two *Sequencing Technologies* during 1982–2017. The number inside the circle indicates viral entries  
524 with only one *Sequencing Technology* listed; the number adjacent to the line indicates entries combining two  
525 *Sequencing Technologies*. The thicker the connection line, the stronger the relationship. **(d and e)** Percentage  
526 ratio graph of all viral entries with *Sequencing Technology* documented for the years 2010–2017, with (d) and  
527 without (e) the *Other* category. The majority of entries in earlier years include omissions classified under the  
528 *Other* category, which is detailed in [Supplement Table S1](#). **(f)** Percentage ratio graph of viral entries with  
529 length greater than 2000 nt that have been documented with one of the seven *Sequencing Technologies* for  
530 the years 2012–2017. The seven technologies includes Sanger (n=1) and NGS technologies (n=6). **(g)**  
531 Percentage ratio graph of viral entries with length greater than 2000 nt and that have been documented with  
532 one of the six NGS as the *Sequencing Technology* for the years 2012–2017. Compared to panel (f), Sanger is  
533 excluded in this graph. **(h)** Assembly method of viral entries greater than 2000 nt, showing percentage ratio  
534 graph of entries with at least one *Assembly Method*. For (h) and (i), the *Other* category describes assembly  
535 methods outside of the 18 most popular programs investigated. **(i)** Reclassification of panel (h) by the nature  
536 of the assembly methods. The programs can be grouped into *de novo* assembler, reference-mapping  
537 assembler, and software that can perform both.

538

539



540

541

542

543

**Figure 2. Workflow diagram of the investigation of variant simulated NGS reads through *de novo* assembly.**

544

First, in step 1, an artificial reference genome and corresponding initial variant reads were created with

545

varying constraints such as genome length, GC content, read length, and assemblers, according to the

546

experiment types as detailed in [Supplement Figure S1](#). In the second step, an artificial mutated variant

547

genome was created. The process is repeated to generate 247 different mutated variants with controlled

548

mutation parameters— starting with 1 mutation every 4 nucleotides (75% PID) and ending with a mutated

549

variant with 1 mutation in every 250 nucleotides (99.6% PID). Mutated variant reads are also generated for

550

each of the mutation parameters. In the third and fourth steps, the initial and mutated variants were then

551

combined and used as input for *de novo* assembly for the three experiments, as detailed in [Supplement Figure](#)

552

[S1](#).

553

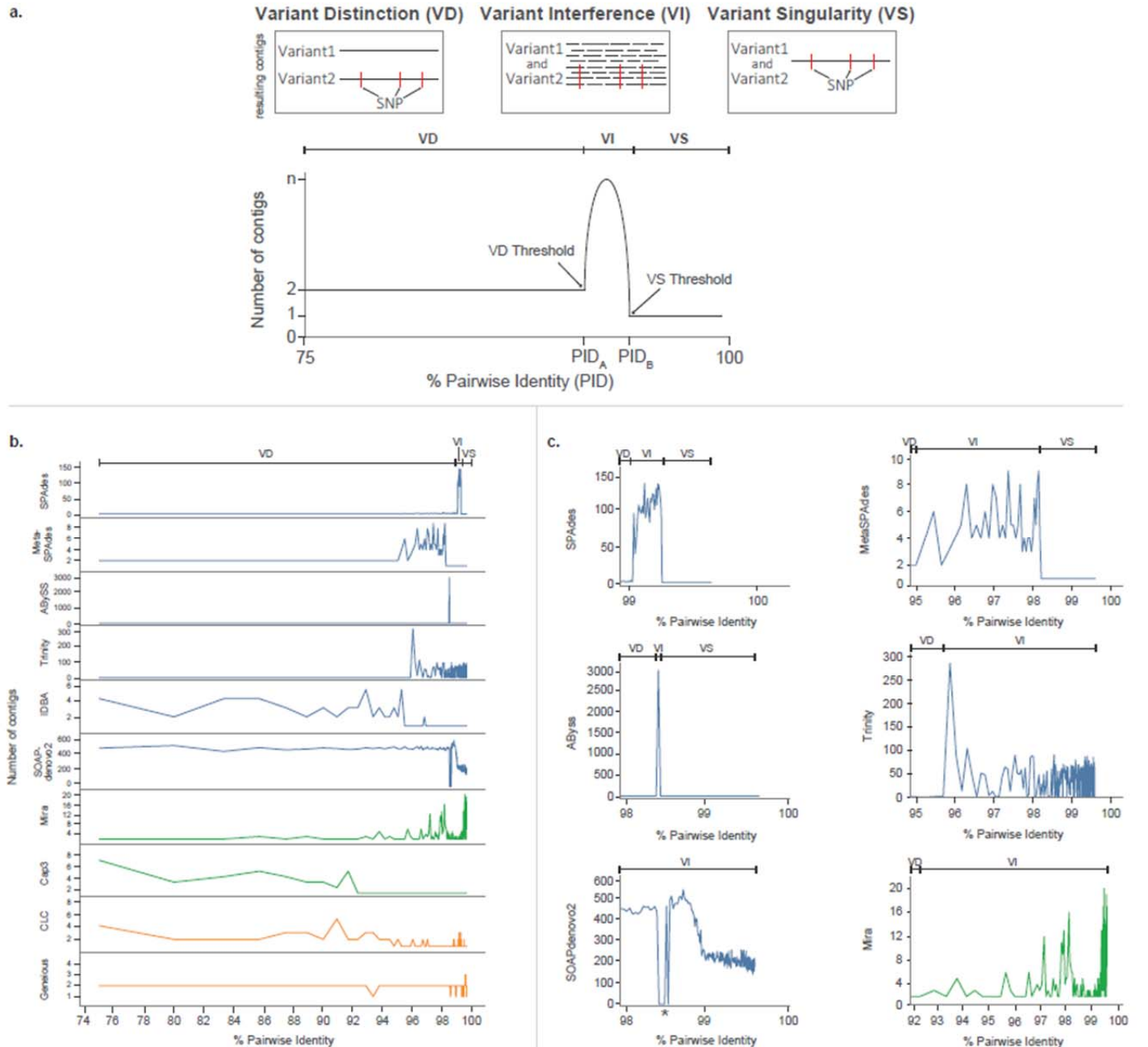
554

555

556

557

558



559

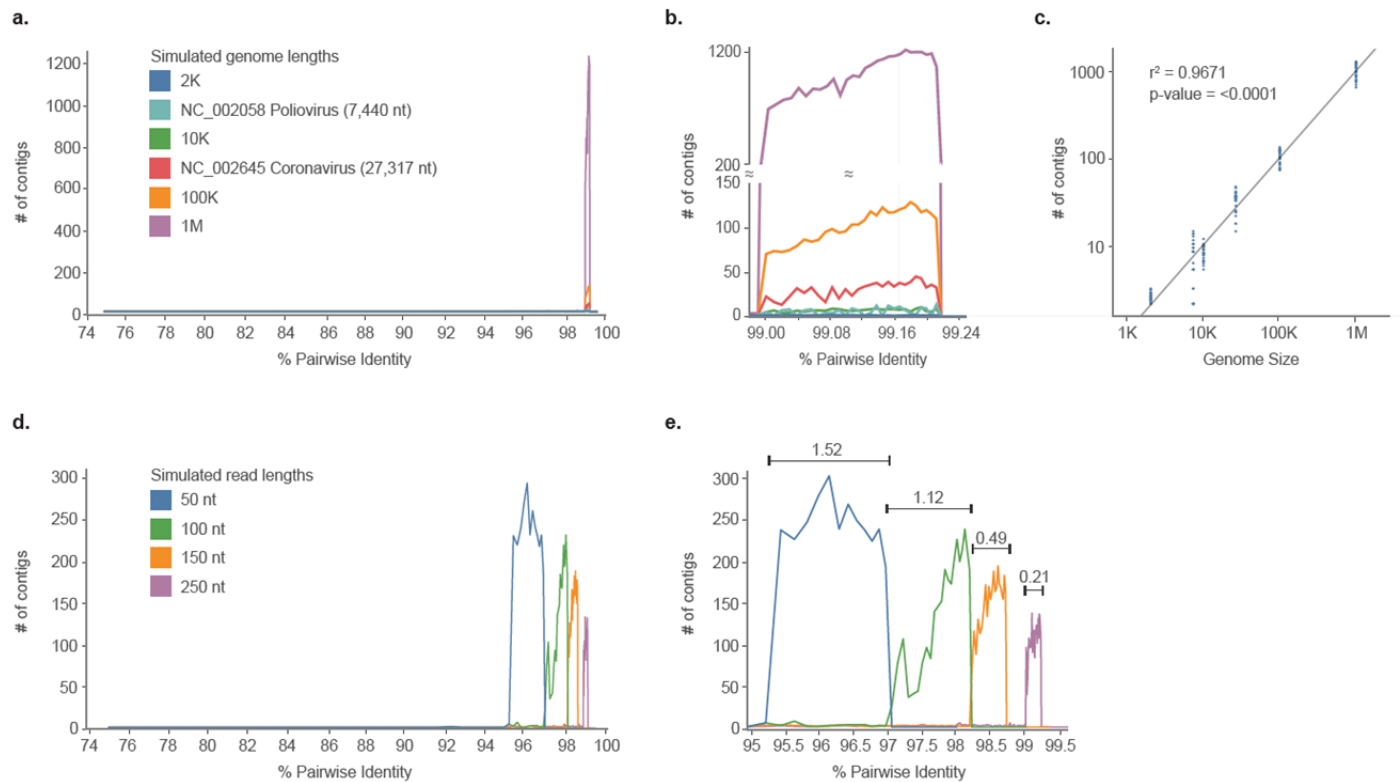
560

561 **Figure 3. The number of contigs generated by different *de novo* assemblers using simulated data containing**  
 562 **variants differed with a range of percentage identities (PID). Blue denotes de Bruijn graph assemblers (DBG);**  
 563 **green denotes overlap-layout-consensus assemblers (OLC); orange denotes commercialized proprietary**  
 564 **algorithms. Variant distinction, VD; variant interference, VI; variant singularity, VS. \*For SOAPdenovo2, several**  
 565 **data points returned zero contigs due to a well-documented segmentation fault error. (a) Schematic diagram**



566 **depicting concepts of the VD, VI, and VS, and their relationship to PID. (b) Comparison of output from 10**  
567 **different assemblers.** The number of contigs produced by each *de novo* assemblers at different variant PID  
568 ranges (75%–99.6%) were shown. **(c) Close-up of PID ranges where variant interference is the most**  
569 **apparent.** The y-axis denotes the number of contigs.  
570



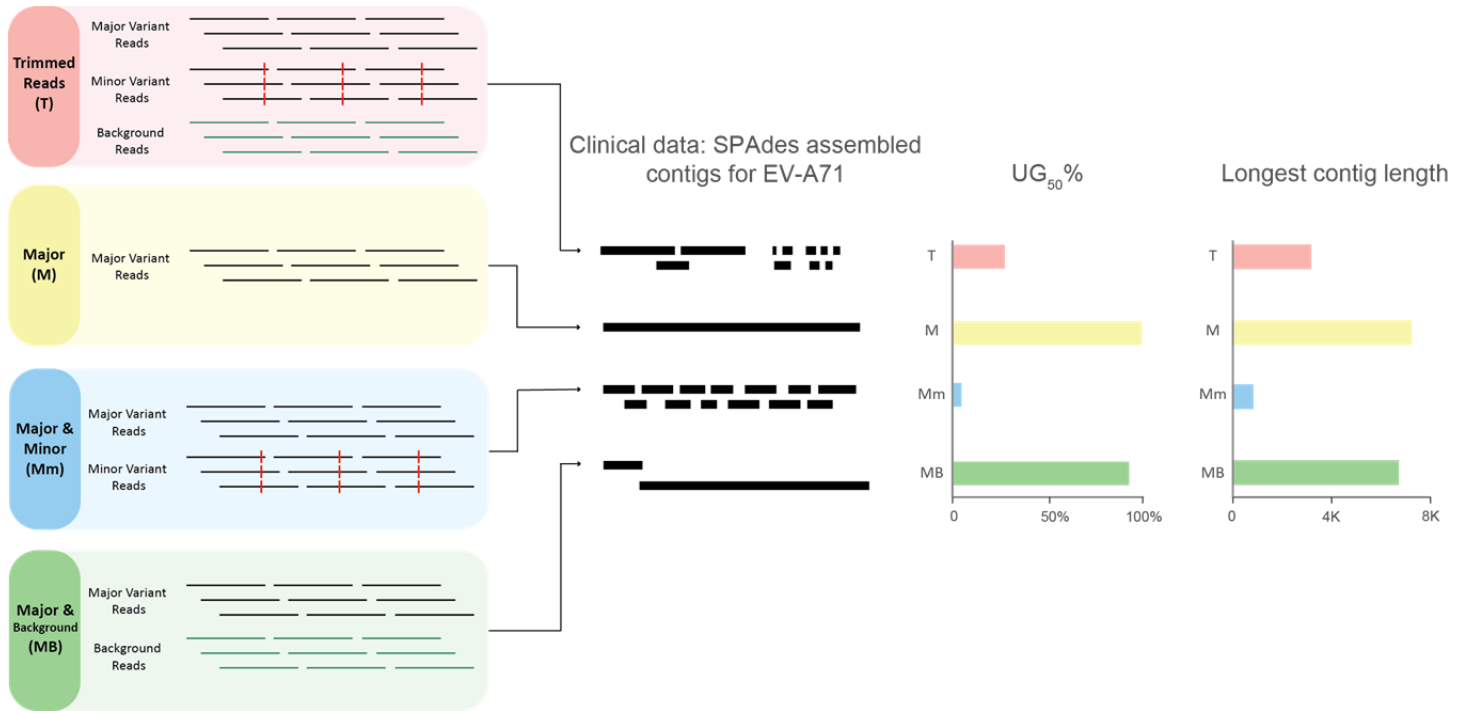


571

572

573 **Figure 4. The effect of genome length and read length on *de novo* assembly of simulated variants across a**  
 574 **range of percentage identities (PID). (a & b) Comparison of genome lengths.** Six different genome lengths  
 575 were assembled and the final contig counts were tallied across varying PID thresholds (75%–99.6%). For the  
 576 simulated genome lengths of 2Kb, 10kb, 100Kb, and 1Mb, the average of contig number at each PID was  
 577 plotted. Panel (b) shows the close-up view where interference was the most prominent. For all six genome  
 578 lengths and each of the 13 iterations, VI consistently occurred in the same range of PID (99.00%–99.24%). The  
 579 assembly makes a transition from VD to VI at the threshold of 99.00%, and it makes a transition from VI to VS  
 580 at the threshold of 99.24%. Also, the longer the genome length, the more contigs produced during VI. **(c) The**  
 581 **relationship between genome length and the total number of contigs produced.** Data from panel (a) were  
 582 plotted on a logarithmic scale. The total number of contigs produced is significantly dependent on the genome  
 583 size ( $r^2=0.967$ ;  $p\text{-value}<0.0001$ ). **(d and e) The effect of read length in variant assembly with a genome size of**  
 584 **100K.** Simulated data with four different read lengths were created and assembled, and the final contig counts  
 585 were tallied across varying PID thresholds (75%–99.6%). Panel (e) shows the close-up view where interference

586 was the most apparent. When longer read lengths were used, the variant interference PID range was much  
587 narrower than when shorter read lengths were used to build contigs.



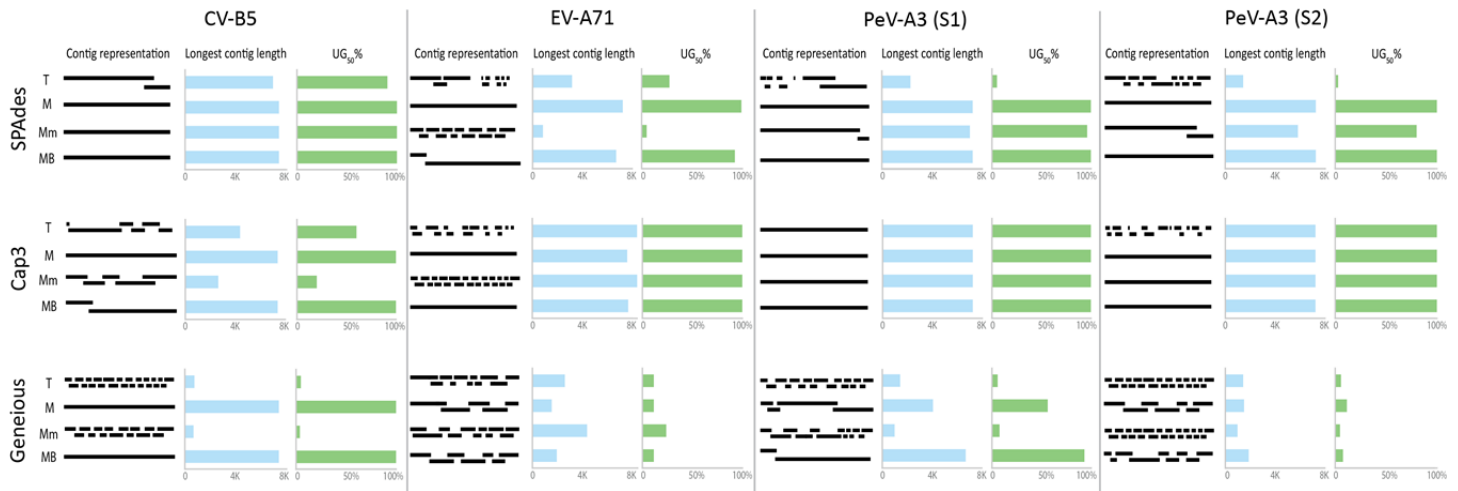
588

589

590 **Figure 5. The effect of variant interference in a real dataset from a clinical sample containing enterovirus**  
591 **A71 (EV-A71) and its variants.** Fastq reads were partitioned into four components: trimmed reads after  
592 quality control (T), major variant (M), minor variant (m), and background (B). These reads were then combined  
593 into four different experiments: T, M, Mm, and MB and assembled using SPAdes. The contig representation  
594 schematic showing the abundance and length of the generated contigs reveals the impact of variant  
595 interference on *de novo* assembly. The bar graphs show the UG<sub>50</sub>% metric and the length of the longest contig.  
596 UG<sub>50</sub>% is a percentage-based metric that estimates length of the unique, non-overlapping contigs as  
597 proportional to the length of the reference genome.<sup>17</sup> Unlike N<sub>50</sub>, UG<sub>50</sub>% is suitable for comparisons across  
598 different platforms or samples/viruses. More clinical samples and viruses are analyzed similarly in [Figure 6](#).

599

600



601

602

603 **Figure 6. The effect of variant interference on the assembly of four clinical datasets using three assembly**

604 **programs.** Fastq reads were partitioned into four categories: total reads (T), major variant (M), minor variant

605 (m), and background (B). These reads were then combined into four different categories: T, M, major and

606 minor variants (Mm), and major variant and background (MB). Datasets were assembled using SPAdes, Cap3,

607 and Geneious. The bar graphs show the UG<sub>50</sub>% metric and the length of the longest contig.

608

609 Coxsackievirus B5, CV-B5; Enterovirus A71, EV-A71; Parechovirus A3 (Sample 1), PeV-A3 (S1); Parechovirus A3

610 (Sample 2), PeV-A3 (S2).

611

612

## 613 **Supplemental Information**

### 614 **Analysis of viral GenBank records**

#### 615 **The advent of NGS fuels viral sequencing**

616

617 As of December 2017, GenBank's non-redundant nucleotide database had grown to more than 2.3  
618 million virus sequences, with the annual number of new sequences deposited increasing by 270% between  
619 2007 and 2017 [Figure 1a and Supplement Table S1]. GenBank entries started incorporating information on  
620 the sequencing technology platform used in 2011. Through 2018, 144,712 viral entries (22%) had documented  
621 utilization of NGS sequencing technology, compared to 500,027 entries (77%) utilizing Sanger methods [Figure  
622 1b and Supplement Table S1]. Illumina was the most common NGS platform used for viral sequencing, with  
623 >2-fold the number of entries compared to the next most popular NGS platform (31,000 viral entries in 2017  
624 [Figure 1d & e]). Although NGS usage has risen tremendously, Sanger sequencing still contributed the majority  
625 of all viral sequences. This is likely because Sanger is still attractive for generating short viral sequences over  
626 genotyping windows or other informative regions. If only long sequences ( $\geq 2000$  nt) are considered, NGS  
627 technologies surpassed Sanger as the dominant strategy for sequencing in 2017 [Figure 1f and Supplement  
628 Table S2].

629

630 A total of 27,217 counts of sequencing technologies were listed for the 25,344 long viral GenBank  
631 entries in 2017, as some sequences were generated using two or more sequencing technologies. NGS  
632 technologies were listed in 53.8% (14,653/27,217) of entries, versus 46.2% of entries (12,564/27,217) for  
633 Sanger. Illumina was identified as the most dominant NGS technology, accounting for 12,615/14,653 entries  
634 (86.1%) [Figure 1g and Supplement Table S2].

635

636 Multiple sequencing technologies may be used to generate viral sequence for one entry. The most  
637 common combination observed was 454 and Sanger (18,002 entries), likely due to the early emergence of the  
638 454 technology compared to other NGS platforms [Figure 1c and Supplement Table S3]. This is followed by  
639 Illumina and Sanger (3,475), Illumina and 454 (3,299), Illumina and Ion Torrent (2,600), and Illumina and  
640 PacBio (997). Interestingly, more recently released longer-read platforms like PacBio and Oxford Nanopore

641 tended to be paired with Illumina more frequently compared to traditional Sanger sequencing. A small  
642 number of studies even combined three or four different sequencing technologies (530 and 6 entries,  
643 respectively) [Supplement Table S4]. Some users employed a combined approach to circumvent the inherent  
644 flaws of one sequencing platform, particularly for genome finishing.<sup>44</sup> For example, after NGS has been used to  
645 generate the majority of a RNA virus genome, RACE (Rapid amplification of cDNA ends) is typically performed  
646 with Sanger to obtain the 5' or 3' termini.<sup>45, 46</sup>

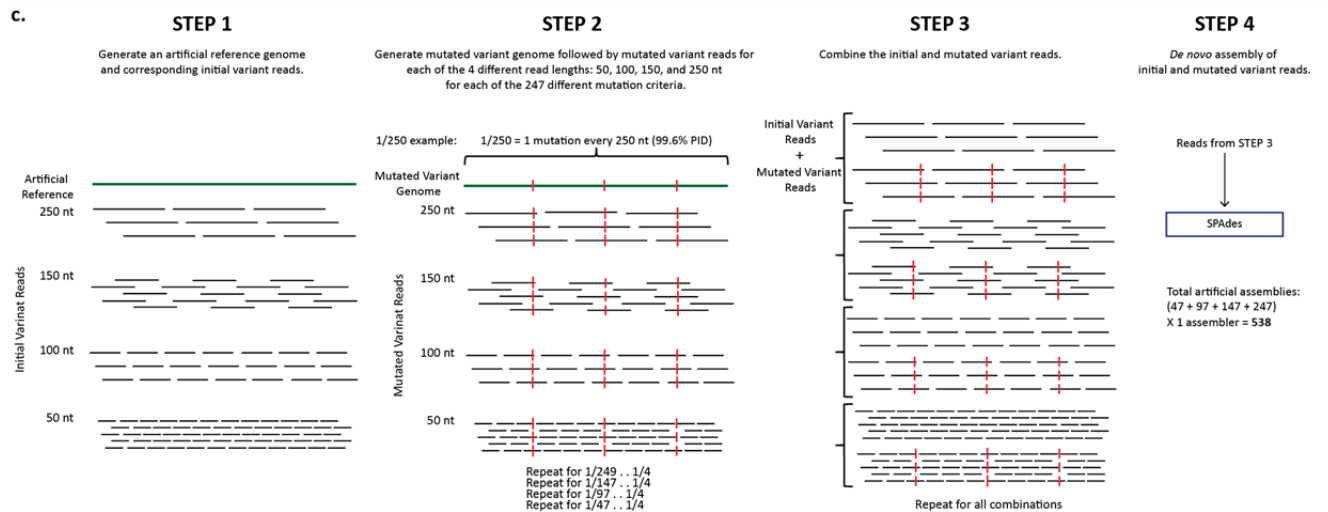
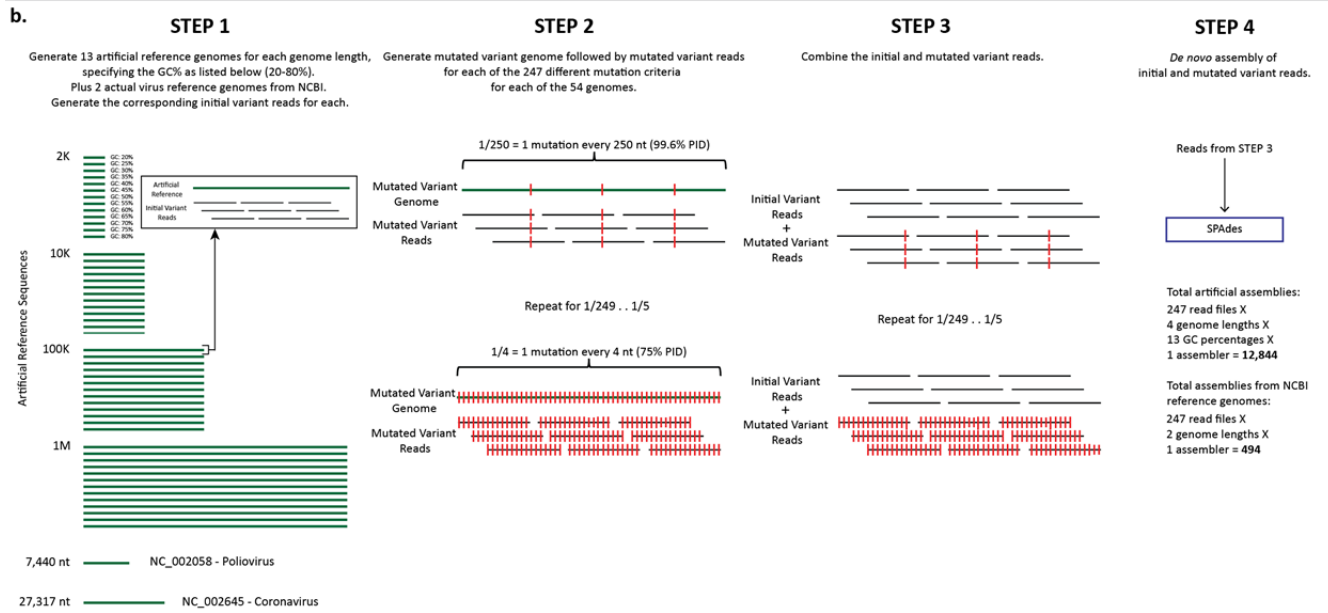
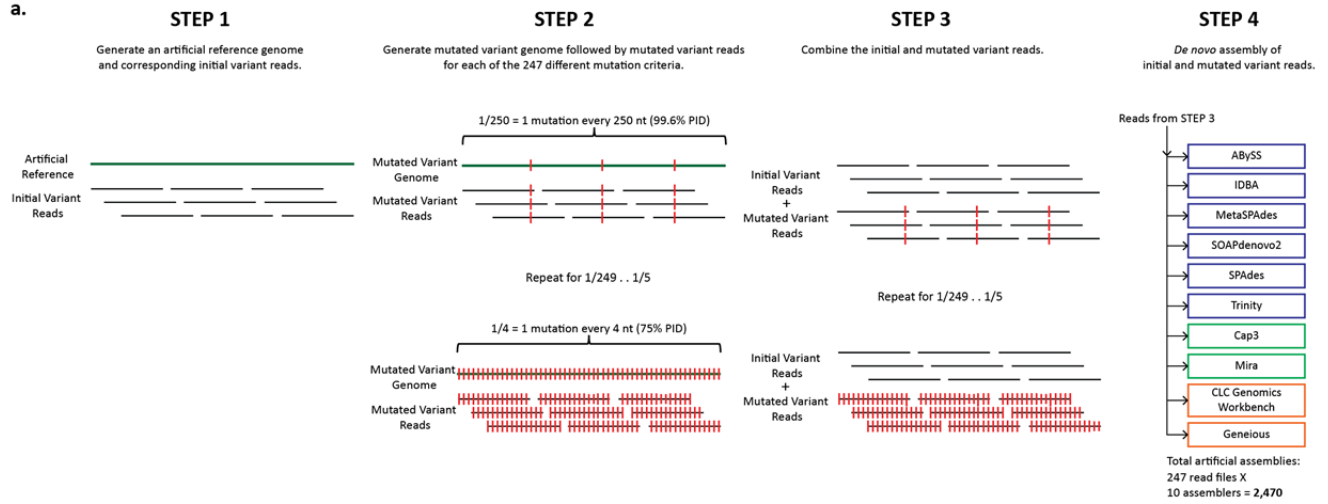
647

### 648 ***De novo* assembly plays a major role in analyzing long viral sequences**

649

650 We analyzed the assembly methods used for GenBank entries of long sequences ( $\geq 2000$  nt) from 2012 to 2017  
651 when NGS usage become relevant [Figure 1h & i and Supplement Table S5]. The number of programs used to  
652 assemble viral sequences has steadily increased over time (a  $>2$ -fold increase from 2012-2017). With new  
653 sequencing technologies emerging and computational power continually improving, the development of new  
654 and better assembly programs always follows suite. The use of specifically-designed *de novo* assembly  
655 programs (ABYSS, BWA, Canu, Cap3, IDBA, MIRA, Newbler, SOAPdenovo, SPAdes, Trinity, and Velvet) has  
656 increased from less than 1% of viral sequence entries in 2012, to 20% of all viral sequence entries in 2017. A  
657 similar increase was observed for reference-mapping software (i.e., Bowtie and Bowtie2), from 0.03% in 2012  
658 to 6.5% in 2017. Multifunctional programs that offer both assembly options, including CLC Genomics  
659 Workbench (CLC), DNA Baser, DNASTAR, Geneious, and Sequencher, were by far the most popular option for  
660 the years 2013-2017. However, since these commercial software packages can perform both *de novo* and  
661 reference-mapping assembly, the exact sequence assembly strategy used for these records is unknown, and  
662 thus the contributions of both *de novo* assembly and reference recruitment are likely underestimated.

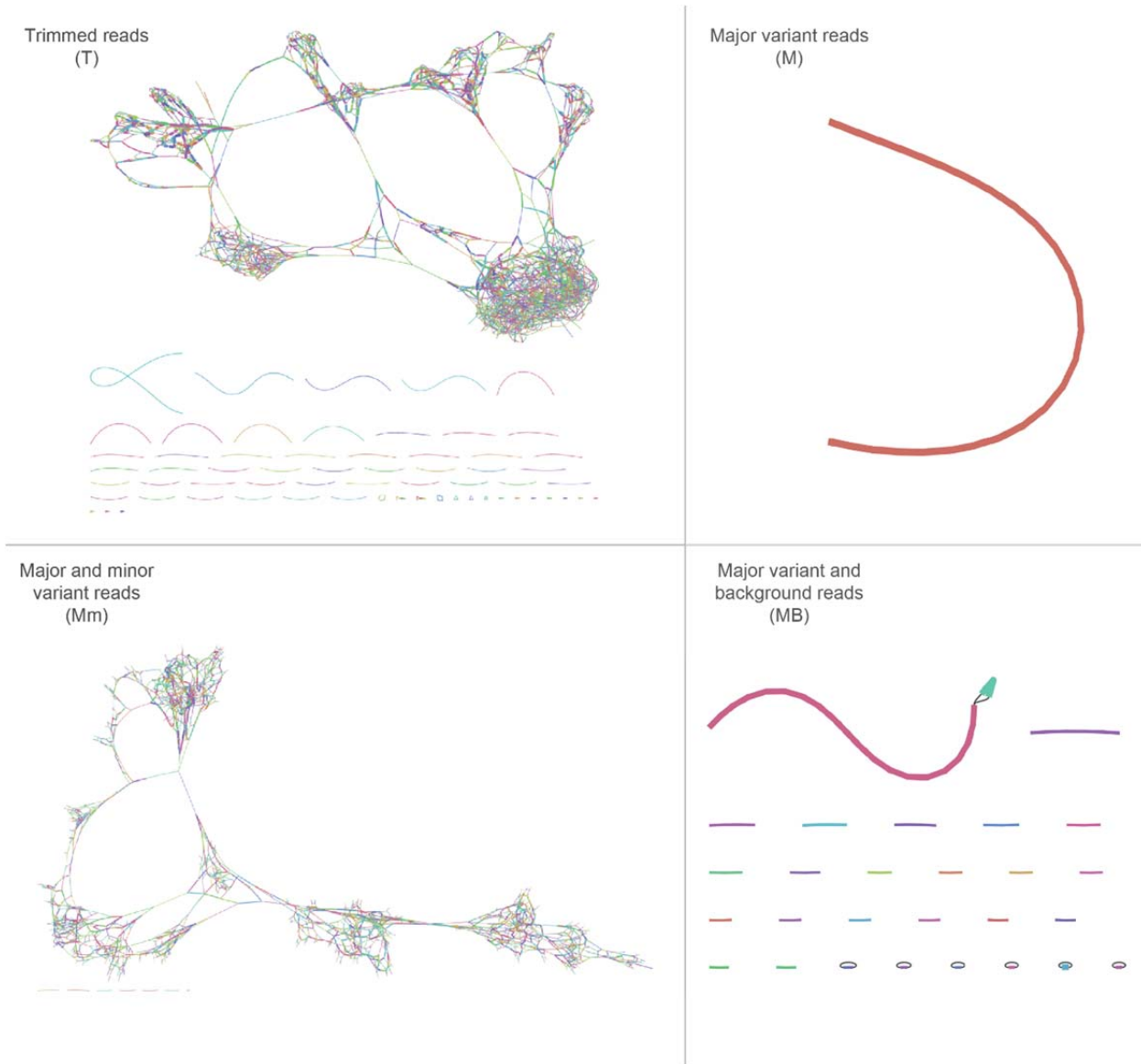
663



665 **Supplement Figure S1. Workflow diagrams of simulated data from data creation through *de novo* assembly.**  
666 **(a) Comparison of assemblers.** First, an artificial reference genome and corresponding initial variant reads  
667 were created with the following constraints: (1) reference genome length: 100K; (2) GC% of reference  
668 genome: 50%; (3) read length: 250 nt; and (4) coverage: 50X. Second, an artificial mutated variant genome and  
669 corresponding mutated variant reads were created 247 times, each with a differing pairwise percent identity  
670 ranging from 1 mutation every 4 nucleotides (75% PID) to 1 mutation in every 250 nucleotides (99.6% PID).  
671 The initial and mutated variants were then combined and used as input for 10 different *de novo* assemblers  
672 with varying underlying algorithms. A total of 2,470 assemblies were performed. **(b) Comparison of genome**  
673 **length and GC%.** First, 13 artificial reference genomes and corresponding initial variant reads were created for  
674 four different genome lengths (2Kb, 10Kb, 100Kb, and 1Mb), each specifying a different GC% ranging from  
675 20%–80%. In addition, two actual virus reference genomes from NCBI were included, NC\_002058 and  
676 NC\_002645, with genome lengths of 7,440 nt and 27,317 nt, respectively. Read lengths of 250 nt with a  
677 coverage of 50X were used for all genomes. Second, an artificial mutated variant genome and corresponding  
678 mutated variant reads were created 247 time, each with a differing pairwise percent identity ranging from 1  
679 mutation every 4 nucleotides (75% PID) to 1 mutation in every 250 nucleotides (99.6% PID). The initial and  
680 mutated variants were then combined for each and used as input for the SPAdes *de novo* assembler. A total of  
681 13,338 assemblies were performed. **(c) Comparison of read length.** First, an artificial reference genome and  
682 corresponding initial variant reads were created with the following constraints: (1) reference genome length:  
683 100K; (2) GC% of reference genome: 50%; (3) read lengths: 50 nt, 100 nt, 150 nt, or 250 nt; and (4) coverage:  
684 50X. Second, an artificial mutated variant genome and corresponding mutated variant reads were created,  
685 each with a differing pairwise percent identity ranging from 1 mutation every 4 nucleotides (75% PID) up to 1  
686 mutation in every 250 nucleotides (99.6% PID). The initial and mutated variants created for each of the four  
687 read lengths were then grouped by read length size and used as input for SPAdes *de novo* assembler. A total of  
688 538 assemblies were performed.

689

690



691

692 **Supplement Figure S2. Analysis of the final contig assembly graphs for a clinical sample containing**  
693 **enterovirus A71 (EV-A71) variants using Bandage.** Based on the four assemblies in [Figure 5](#), Bandage was  
694 used to display the contig graphs from each SPAdes output. The visualizations for T, Mm, and MB show the  
695 effects of variant interference, while M shows the ideal assembly.



| Year          | Total # of viral entries in GenBank | Total count†   | Total omitted  | Total # of entries with at least one Seq. Tech. | Sequencing Technology Breakdown |              |              |              |            |
|---------------|-------------------------------------|----------------|----------------|---|---------------------------------|--------------|--------------|--------------|------------|
|               |                                     |                |                |   | Sanger                          | 454          | Illumina     | IonTorrent   | Oxford N   |
| 2017          | 238367                              | 243849         | 108021         | 135828  | 85194                           | 15999        | 31279        | 2130         | 46         |
| 2016          | 235477                              | 237569         | 107090         | 130479  | 102837                          | 2971         | 22185        | 2111         | 119        |
| 2015          | 197440                              | 211177         | 71148          | 140029  | 102440                          | 15517        | 17625        | 3048         |            |
| 2014          | 158579                              | 163092         | 66217          | 96875   | 81515                           | 5452         | 7399         | 2345         |            |
| 2013          | 198540                              | 202232         | 108365         | 93867   | 84527                           | 5243         | 2474         | 758          |            |
| 2012          | 172850                              | 173324         | 126821         | 46503   | 43509                           | 1194         | 277          | 403          |            |
| 2011          | 181315                              | 181319         | 176355         | 4964  | 5                               | 4811         | 147          |              |            |
| 2010          | 131962                              | 131962         | 131960         | 2   | 2                               |              |              |              |            |
| 2009          | 213549                              | 213549         | 213549         |   |                                 |              |              |              |            |
| 2008          | 109265                              | 109265         | 109265         |   |                                 |              |              |              |            |
| 2007          | 88996                               | 88996          | 88996          |   |                                 |              |              |              |            |
| 2006          | 94444                               | 94444          | 94444          |   |                                 |              |              |              |            |
| 2005          | 58245                               | 58245          | 58245          |   |                                 |              |              |              |            |
| 2004          | 53841                               | 53842          | 53834          | 8   | 1                               | 4            | 2            |              |            |
| 2003          | 38578                               | 38578          | 38576          | 2   | 2                               |              |              |              |            |
| 2002          | 33412                               | 33412          | 33412          |   |                                 |              |              |              |            |
| 2001          | 28305                               | 28305          | 28304          | 1   |                                 | 1            |              |              |            |
| 2000          | 26871                               | 26871          | 26871          |   |                                 |              |              |              |            |
| 1999          | 17266                               | 17266          | 17266          |   |                                 |              |              |              |            |
| 1998          | 13840                               | 13840          | 13840          |   |                                 |              |              |              |            |
| 1997          | 12378                               | 12378          | 12378          |   |                                 |              |              |              |            |
| 1996          | 8988                                | 8988           | 8987           | 1   |                                 |              | 1            |              |            |
| 1995          | 7475                                | 7475           | 7475           |   |                                 |              |              |              |            |
| 1994          | 5449                                | 5449           | 5449           |   |                                 |              |              |              |            |
| 1993          | 9185                                | 9185           | 9184           | 1   |                                 |              | 1            |              |            |
| 1992          | 1754                                | 1754           | 1754           |   |                                 |              |              |              |            |
| 1991          | 725                                 | 725            | 725            |   |                                 |              |              |              |            |
| 1990          | 364                                 | 364            | 363            | 1   |                                 |              | 1            |              |            |
| 1989          | 424                                 | 424            | 424            |   |                                 |              |              |              |            |
| 1988          | 269                                 | 269            | 269            |   |                                 |              |              |              |            |
| 1987          | 159                                 | 159            | 159            |   |                                 |              |              |              |            |
| 1986          | 114                                 | 114            | 114            |   |                                 |              |              |              |            |
| 1985          | 130                                 | 130            | 130            |   |                                 |              |              |              |            |
| 1984          | 19                                  | 19             | 19             |   |                                 |              |              |              |            |
| 1983          | 92                                  | 92             | 92             |   |                                 |              |              |              |            |
| 1982          | 108                                 | 108            | 108            |   |                                 |              |              |              |            |
| <b>TOTALS</b> | <b>2338775</b>                      | <b>2368770</b> | <b>1720209</b> | <b>648561</b>                                   | <b>500032</b>                   | <b>51192</b> | <b>81391</b> | <b>10795</b> | <b>165</b> |

696 **Supplement Table S1. Total counts from NCBI's GenBank non-redundant nucleotide database.**

697 † *Total count* is the combination of all sequencing technologies listed for each entry plus the  
698 total number of entries with sequencing technology omitted. This number is higher than the  
699 *Total # of viral entries in GenBank* because it accounts for all entries with multiple sequencing  
700 technologies listed.

701

702 Sequencing Technology, Seq. Tech.; Oxford Nanopore, Oxford NP; Pacific Biosciences, PacBio

703

| NGS<br>Platforms  | Year         |              |              |              |              |              |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                   | 2017         | 2016         | 2015         | 2014         | 2013         | 2012         |
| <b>454</b>        | 1029         | 634          | 4987         | 1531         | 1642         | 376          |
| <b>Sanger</b>     | 12564        | 13571        | 20216        | 14294        | 13646        | 10847        |
| <b>Illumina</b>   | 12615        | 12629        | 4121         | 4414         | 1266         | 230          |
| <b>PacBio</b>     | 17           | 67           | 1            | 12           | 1            | 0            |
| <b>IonTorrent</b> | 923          | 1342         | 1217         | 1131         | 408          | 171          |
| <b>Oxford NP</b>  | 46           | 119          | 0            | 0            | 0            | 0            |
| <b>SOLiD</b>      | 8            | 0            | 0            | 13           | 29           | 1            |
| <b>Other</b>      | 15           | 4            | 1            | 5            | 10           | 41           |
| <b>TOTALS</b>     | <b>27217</b> | <b>28366</b> | <b>30543</b> | <b>21400</b> | <b>17002</b> | <b>11666</b> |

704

705 **Supplement Table S2. Total count of sequencing technologies for sequences >2000 nt in the**  
706 **NCBI GenBank non-redundant nucleotide database for years 2012–2017.**

707 These numbers were found with the following search criteria: “viruses,” “genomic RNA/DNA,”  
708 “GenBank (No RefSeq),” length: 2000 to 2000000, release date: 1/1/201X to 12/31/201X, and  
709 “sequencing technology” in any field.

710

711 Oxford Nanopore, Oxford NP; Pacific Biosciences, PacBio

712

| Year          | Total # of entries with two Seq. Techs. | Total # of entries with three Seq. Techs. | Total # of entries with four Seq. Techs. |
|---------------|---|---|--|
| 2017          | 5468                                    | 7   |  |
| 2016          | 2008                                    | 42  |  |
| 2015          | 13156                                   | 283                                       | 5  |
| 2014          | 4457                                    | 28  |  |
| 2013          | 3409                                    | 140                                       | 1  |
| 2012          | 414                                     | 30  |  |
| 2011          | 4                                       |   |  |
| 2010          |   |   |  |
| 2009          |   |   |  |
| 2008          |   |   |  |
| 2007          |   |   |  |
| 2006          |   |   |  |
| 2005          |   |   |  |
| 2004          | 1                                       |   |  |
| 2003          |   |   |  |
| 2002          |   |   |  |
| 2001          |   |   |  |
| 2000          |   |   |  |
| 1999          |   |   |  |
| 1998          |   |   |  |
| 1997          |   |   |  |
| 1996          |   |   |  |
| 1995          |   |   |  |
| 1994          |   |   |  |
| 1993          |   |   |  |
| 1992          |   |   |  |
| 1991          |   |   |  |
| 1990          |   |   |  |
| 1989          |   |   |  |
| 1988          |   |   |  |
| 1987          |   |   |  |
| 1986          |   |   |  |
| 1985          |   |   |  |
| 1984          |   |   |  |
| 1983          |   |   |  |
| 1982          |   |   |  |
| <b>TOTALS</b> | <b>28917</b>                            | <b>530</b>                                | <b>6</b>                                 |

713

714 **Supplement Table S3. Total counts from NCBI's GenBank non-redundant nucleotide database**  
 715 **with multiple sequencing technologies listed per entry.** Blank fields indicate absence of entries  
 716 for the corresponding category.

717

718 Sequencing Technologies, Seq. Techs.

719

|                 | <b>454</b> | <b>Illumina</b> | <b>IonTorrent</b> | <b>PacBio</b> | <b>SOLiD</b> |                   |
|-----------------|------------|-----------------|-------------------|---------------|--------------|-------------------|
| <b>454</b>      |            | <b>3</b>        |                   |               |              | <b>IonTorrent</b> |
| <b>454</b>      |            | <b>2</b>        |                   |               |              | <b>PacBio</b>     |
| <b>454</b>      |            | <b>452</b>      | <b>21</b>         |               | <b>1</b>     | <b>Sanger</b>     |
| <b>Illumina</b> | <b>6</b>   |                 | <b>48</b>         | <b>2</b>      | <b>1</b>     | <b>Sanger</b>     |

**IonTorrent**

720

721 **Supplement Table S4. Total counts from NCBI's GenBank non-redundant nucleotide database**  
722 **of all entries with three and four sequencing technologies listed**

723 For example, there are a total of 6 entries in GenBank that have the following sequencing  
724 technologies listed: 454, Illumina, Ion Torrent, and Sanger for one sequence entry.

725

726 Pacific Biosciences, PacBio

727

| Assembly<br>Methods | Year         |              |              |              |              |              |
|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                     | 2017         | 2016         | 2015         | 2014         | 2013         | 2012         |
| <b>ABYSS</b>        | 522          | 155          | 100          | 66           | 56           | 0            |
| <b>Bowtie</b>       | 40           | 868          | 33           | 527          | 5            | 4            |
| <b>Bowtie2</b>      | 1682         | 128          | 787          | 9            | 51           | 0            |
| <b>BWA</b>          | 671          | 294          | 281          | 440          | 148          | 1            |
| <b>Canu</b>         | 3            | 0            | 0            | 0            | 0            | 0            |
| <b>Cap3</b>         | 59           | 34           | 55           | 288          | 10           | 0            |
| <b>CLC</b>          | 3404         | 5139         | 1948         | 2186         | 1172         | 381          |
| <b>DNA Baser</b>    | 838          | 326          | 247          | 261          | 27           | 9            |
| <b>DNASTAR</b>      | 4030         | 3191         | 6897         | 3175         | 3101         | 530          |
| <b>Geneious</b>     | 3636         | 2633         | 4767         | 588          | 504          | 79           |
| <b>IDBA</b>         | 28           | 11           | 729          | 22           | 2            | 0            |
| <b>MIRA</b>         | 446          | 406          | 70           | 140          | 24           | 14           |
| <b>Newbler</b>      | 176          | 183          | 703          | 336          | 435          | 60           |
| <b>Sequencher</b>   | 3243         | 2154         | 2572         | 5727         | 7927         | 3462         |
| <b>SOAPdenovo</b>   | 258          | 67           | 105          | 24           | 9            | 1            |
| <b>SPAdes</b>       | 792          | 1632         | 89           | 266          | 0            | 0            |
| <b>Trinity</b>      | 2162         | 4576         | 301          | 509          | 4            | 0            |
| <b>Velvet</b>       | 161          | 107          | 338          | 341          | 144          | 32           |
| <b>Other</b>        | 4190         | 6220         | 5810         | 5437         | 3179         | 6950         |
| <b>TOTALS</b>       | <b>26341</b> | <b>28124</b> | <b>25832</b> | <b>20342</b> | <b>16798</b> | <b>11523</b> |

728

729

730 **Supplement Table S5. Total count of assembly programs used to generate sequences >2000 nt**

731 **in the NCBI GenBank non-redundant nucleotide database.** These numbers were found with

732 the following search criteria: “viruses,” “genomic RNA/DNA,” “GenBank (No RefSeq),” length:

733 2000 to 2000000, release date: 1/1/201X to 12/31/201X, and “sequencing technology” in any

734 field; the assembly method was then parsed out.

735

| DBG         |         | OLC     |         | Proprietary Algorithm |         |
|-------------|---------|---------|---------|-----------------------|---------|
| Program     | Version | Program | Version | Program               | Version |
| ABYSS       | 2.0.2   | Cap     | 3       | CLC Genomic Workbench | 11      |
| IDBA        | 1.1.3   | Mira    | 4.0.2   | Geneious              | 10.2.3  |
| MetaSPAdes  | 3.9.0   |         |         |                       |         |
| SOAPdenovo2 | r240    |         |         |                       |         |
| SPAdes      | 3.9.0   |         |         |                       |         |
| Trinity     | 2.1.1   |         |         |                       |         |

736

737 **Supplement Table S6. The 10 *de novo* assemblers used for analysis of the simulated data, as**

738 **categorized by their underlying assembly algorithms.** de Bruijn graph, DBG; overlap-layout-

739 consensus, OLC.

740