Castro *et al. BMC Genomics* (2020) 21:421
https://doi.org/10.1186/s12864-020-06801-w

BMC Genomics

RESEARCH ARTICLE Open Access

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

Christina J. Castro[1,2], Rachel L. Marine[1], Edward Ramos[3] and Terry Fei Fan Ng[1*]

Abstract

Background: Viruses have high mutation rates and generally exist as a mixture of variants in biological samples. Next-generation sequencing (NGS) approaches have surpassed Sanger for generating long viral sequences, yet how variants affect NGS de novo assembly remains largely unexplored.

Results: Our results from > 15,000 simulated experiments showed that presence of variants can turn an assembly of one genome into tens to thousands of contigs. This "variant interference" (VI) is highly consistent and reproducible by ten commonly-used de novo assemblers, and occurs over a range of genome length, read length, and GC content. The main driver of VI is pairwise identities between viral variants. These findings were further supported by in silico simulations, where selective removal of minor variant reads from clinical datasets allow the "rescue" of full viral genomes from fragmented contigs.

Conclusions: These results call for careful interpretation of contigs and contig numbers from de novo assembly in viral deep sequencing.

Keywords: De novo assembly, Variant, Quasispecies, Virus, Microbe

Background

For many years, Sanger sequencing has been used to complement classical epidemiological and laboratory methods for investigating viral infections [1]. As technologies have evolved, the emergence of next-generation sequencing (NGS), which drastically reduced the cost per base to generate sequence data for complete viral genomes, has allowed scientists to apply viral sequencing on a grander scale [2–4]. Genomic sequencing is ideal for elucidating viral transmission pathways, characterizing emerging viruses, and locating genomic regions which are functionally important for evading the host immune system or antivirals [2, 5].

Genomic surveillance of viruses is particularly important in light of their rapid rate of evolution. Viruses have higher mutation rates than cellular-based taxa, with RNA viruses having mutation rates as high as 1.5×10^{-3} mutations per nucleotide, per genomic replication cycle [6]. Due to this high mutation rate, it is well established that most RNA viruses exist as a swarm of quasispecies, [7] with each quasispecies containing unique single nucleotide polymorphisms (SNPs). The presence of these variants plays a key role in viral adaptation.

Due to viruses' rapid evolution, a single clinical sample often contains a mixture of many closely related viruses. Viral quasispecies are mainly derived from intra-host evolution, with RNA viruses such as poliovirus, human immunodeficiency virus (HIV), hepatitis C (HCV), influenza, dengue, and West Nile viruses maintaining diverse quasispecies populations within a host [8–15]. Conversely, the term "viral strains" often refers to different lineages of viruses found in separate hosts, or a co-infection of viruses in the same host due to multiple

* Correspondence: ylz9@cdc.gov
[1]Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA
Full list of author information is available at the end of the article

infection events. As a result, sequence divergence is usually higher when comparing viral strains compared to quasispecies. In this study, we use the term “variant” to encompass both quasispecies and strains regardless of how the variants originated in the biological samples.

Since many sequencing technologies produce reads that are significantly shorter than the target genome size, a process to construct contigs, scaffolds, and full-length genomes is needed. Reference-mapping and de novo assembly are the two primary bioinformatic strategies for genome assembly. Reference-mapping requires a closely-related genome as input to align reads, while de novo assembly generates contigs without the use of a reference genome. Therefore, de novo assembly is the most suitable strategy for analyzing underexplored taxa [16] or for viruses with high mutation and/or recombination rates.

The two most common graph algorithms employed by de novo assembly programs are: overlap graphs for overlap-layout-consensus (OLC) methods, and k-mer based graphs for de Bruijn graph (DBG) methods. OLC methods involve determining overlaps by performing a series of pair-wise sequence alignments. Such assemblies may be computationally expensive (especially for large datasets), and generally work better with longer reads [17, 18]. Conversely, DBG assemblers split reads into smaller k-mers, with k-mers connected when they share a common prefix and suffix of length k – 1. DBG methods are usually faster to run than OLC methods, but this strategy is known to be sensitive to repeats, sequencing error, and the presence of variants, which increase the k-mer complexity and ambiguity during sequence reconstruction [19, 20]. These challenges could lead to fragmented contigs when analyzing viral assemblies from clinical or environmental samples [21].

In this study, we first examined how often NGS and de novo assembly were applied for viral sequences deposited in the GenBank nucleotide database (www.ncbi.nlm.nih.gov/nucleotide/). Then, we investigated how the presence of variants affected assembly results - simulated and clinical NGS datasets were analyzed using multiple assembly programs to explore the effects of genome variant relatedness, read length, genome GC and genome length on the resulting contig distribution. As viruses in different taxa vary in length and GC content, these experiments demonstrate how assembly of viral variants is impacted by basic genome structure characteristics, as well as by the nucleotide similarity between variants and sequencing read length.

Results

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

GenBank viral entries from 1982 to 2019 were collected and analyzed, with extensive analyses performed to evaluate technologies and bioinformatics programs cited in records deposited between 2011 and 2019. Through 2019, there were over 2.7 million viral entries in GenBank; however, over 70% (1.9 million) do not specify a sequencing technology (Supplement Table S1) due to the looser data requirement in earlier years. When looking at recently deposited records (2014–2019), the Illumina sequencing platform was the most common NGS platform used for viral sequencing, with over a 2-fold increase over the next most popular NGS platform (Fig. 1d & e). When long sequences (≥2000 nt) are considered, NGS technologies surpassed Sanger in 2017 as the dominant strategy for sequencing, comprising 53.8% (14,653/27,217) of entries compared to 46.2% of entries (12,564/27,217) for Sanger. This trend held true in 2018 and 2019 as well (Fig. 1f and Supplement Table S2).

Hybrid sequencing approaches, where researchers use more than one sequencing technology to generate complete viral sequences, have also become more common over the past several years. The most common combination observed was 454 and Sanger (18,124 entries), likely due to the early emergence of the 454 technology compared to other NGS platforms (Fig. 1c and Supplement Table S3). However, combining Illumina with various other sequencing platforms is quite commonplace (> 19,000 entries).

De novo assembly programs (ABySS, BWA, Canu, Cap3, IDBA, MIRA, Newbler, SOAPdenovo, SPAdes, Trinity, and Velvet) have increased from less than 1% of viral sequence entries ≥2000 nt in 2012, to 20% of all viral sequence entries in 2019 (Fig. 1h & i). A similar increase was observed for reference-mapping programs (i.e., Bowtie and Bowtie2), from 0.03% in 2012 to 12.5% in 2019. Multifunctional programs that offer both assembly options were the most common programs cited for the years 2013–2019, but since the exact sequence assembly strategy used for these records is unknown (Tables S1-S5), the contributions of de novo assembly are likely underestimated. An expanded summary of the sequencing technologies and assembly approaches used for viral GenBank records is available in the Supplement text and Supplement Tables S1-S6.

Effect of variant assembly using popular de novo assemblers

After establishing the growing use of NGS technologies for viral sequencing, we next focused on understanding how the presence of viral variants may influence de novo assembly output. We generated 247 simulated viral NGS datasets representing a continuum of pairwise identity (PID) between two viral variants, from 75% PID (one nucleotide difference every 4 nucleotides), to 99.6% PID (one nucleotide difference every 250 nucleotides) (Fig. 2). For Experiment 1, these datasets were assembled using

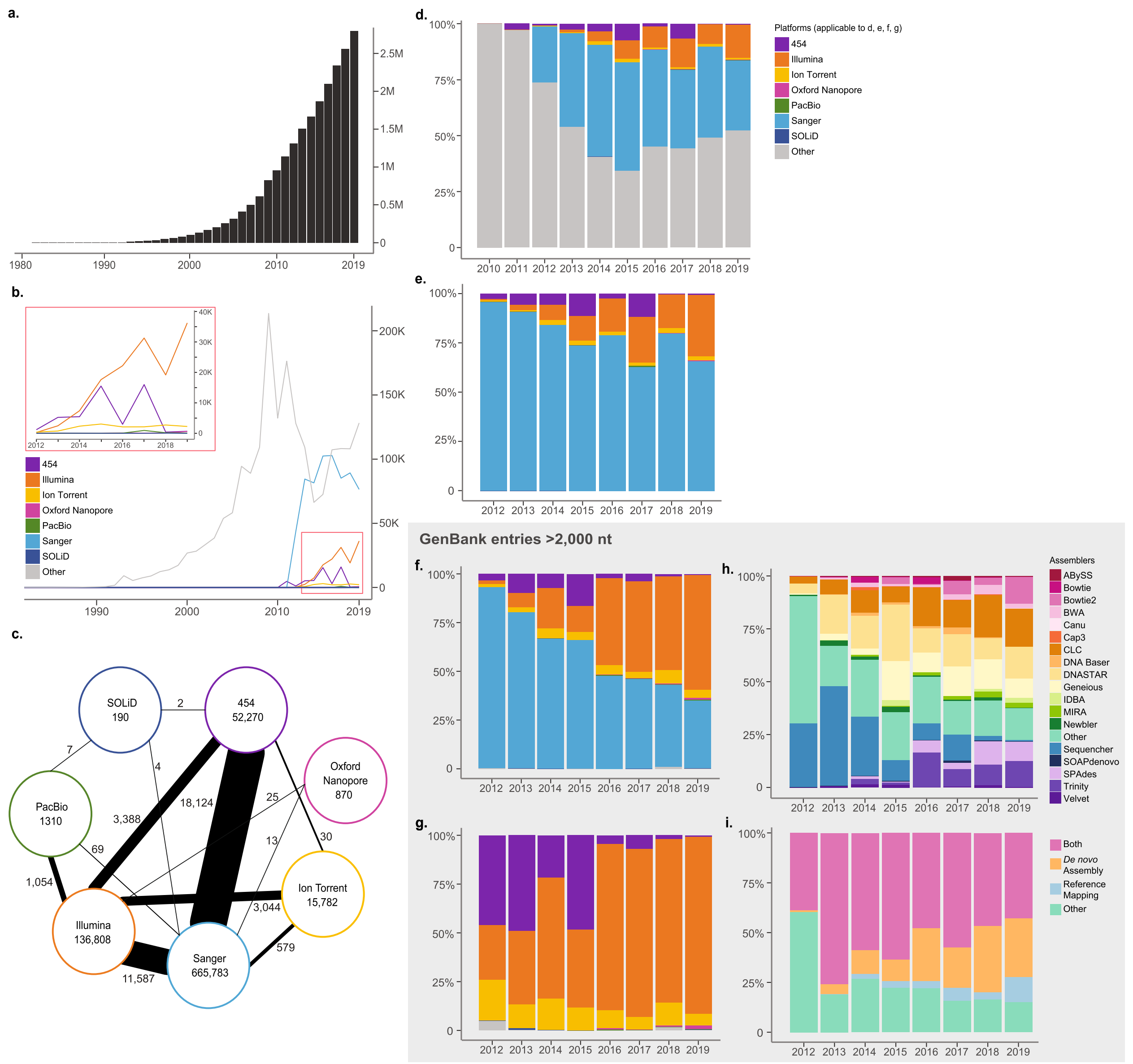

Fig. 1 Trends and patterns of sequencing technology and assembly methods of viral entries in the GenBank database. **a** Cumulative frequency histogram of all viral entries in GenBank from Jan. 1, 1982 through Dec. 31, 2019 (total = 2,793,810 entries). **b** Count of all viral entries with at least one *Sequencing Technology* documented for the years 1982–2019. For panels (**b**) and (**d**), the "Other" category denotes entries with the *Sequencing Technology* field omitted or mis-assigned. **c** Relationship between viral entries listing one or two *Sequencing Technologies* during 1982–2019. The number inside the circle indicates viral entries with only one *Sequencing Technology* listed; the number adjacent to the line indicates entries combining two *Sequencing Technologies*. The thicker the connection line, the stronger the relationship. **d** and **e** Percentage ratio graph of all viral entries with *Sequencing Technology* documented for the years 2010–2019, with (**d**) and without (**e**) the *Other* category. The majority of entries in earlier years include omissions classified under the *Other* category, which is detailed in Supplement Table S1. **f** Percentage ratio graph of viral entries with length greater than 2000 nt that have been documented with one of the seven *Sequencing Technologies* for the years 2012–2019. The seven technologies include Sanger (*n* = 1) and NGS technologies (*n* = 6). **g** Percentage ratio graph of viral entries with length greater than 2000 nt and that have been documented with one of the six NGS as the *Sequencing Technology* for the years 2012–2019. Compared to panel (**f**), Sanger is excluded in this graph. **h** Assembly method of viral entries greater than 2000 nt, showing percentage ratio graph of entries with at least one *Assembly Method*. For (**h**) and (**i**), the *Other* category describes assembly methods outside of the 18 most popular programs investigated. **i** Reclassification of panel (**h**) by the nature of the assembly methods. The programs can be grouped into de novo assembler, reference-mapping, and software that can perform both

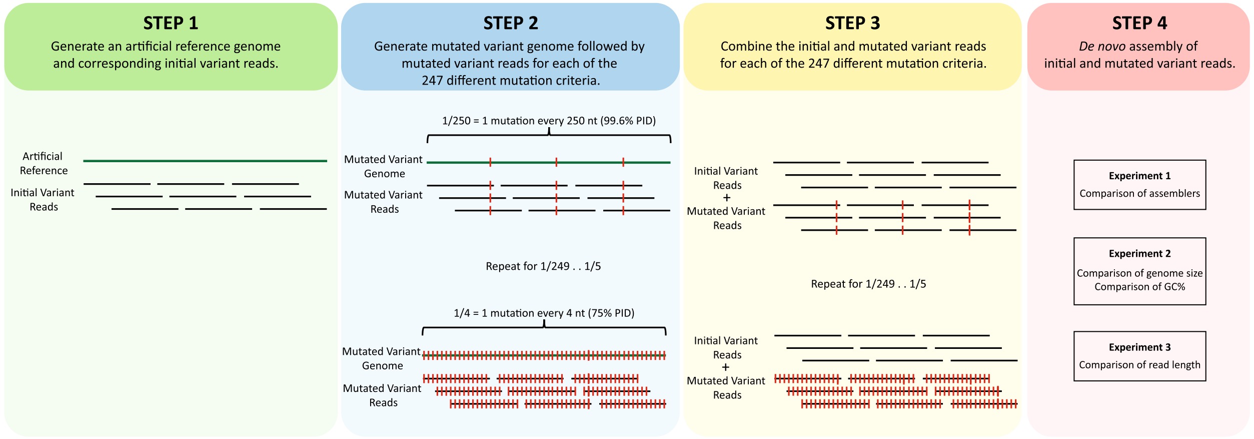
STEP 1
Generate an artificial reference genome and corresponding initial variant reads.
Artificial Reference
Initial Variant Reads
STEP 2
Generate mutated variant genome followed by mutated variant reads for each of the 247 different mutation criteria.
1/250 = 1 mutation every 250 nt (99.6% PID)
Mutated Variant Genome
Mutated Variant Reads
Repeat for 1/249 . . 1/5
1/4 = 1 mutation every 4 nt (75% PID)
Mutated Variant Genome
Mutated Variant Reads
STEP 3
Combine the initial and mutated variant reads for each of the 247 different mutation criteria.
Initial Variant Reads + Mutated Variant Reads
Repeat for 1/249 . . 1/5
Initial Variant Reads + Mutated Variant Reads
STEP 4
De novo assembly of initial and mutated variant reads.
Experiment 1
Comparison of assemblers
Experiment 2
Comparison of genome size
Comparison of GC%
Experiment 3
Comparison of read length

Fig. 2 Workflow diagram of the investigation of variant simulated NGS reads through de novo assembly. First, in step 1, an artificial reference genome and corresponding initial variant reads were created with varying constraints such as genome length, GC content, read length, and assemblers, according to the experiment types as detailed in Supplement Figure S1. In the second step, an artificial mutated variant genome was created. The process is repeated to generate 247 different mutated variants with controlled mutation parameters— starting with 1 mutation every 4 nucleotides (75% PID) and ending with 1 mutation in every 250 nucleotides (99.6% PID). Mutated variant reads are also generated for each of the mutation parameters. In the third and fourth steps, the initial and mutated variants were then combined and used as input for de novo assembly for the three experiments, as detailed in Supplement Figure S1

10 of the most used de novo assembly programs (Fig. 2 and Supplement Figure S1a) to evaluate their ability to assemble the two variants into their own respective contigs as the PID between the variants increases.

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

Slight differences in the variant interference patterns (relative to the canonical variant interference model) were observed for the 10 assemblers investigated. VD was observed for SPAdes, MetaSPAdes, and ABySS assemblers. While it was not observed with Cap3 and IDBA with the current simulated data parameters, we speculate that VD may occur at a lower PID level for these assemblers than tested in this study. The PID range where VI was observed was distinct for each de novo assembler (Fig. 3). During VI, SPAdes produced as many as 134 contigs and ABySS produced 3076 contigs, while MetaSPAdes, Cap3, and IDBA produced up to 10.

A different pattern was observed for Mira, Trinity, and SOAPdenovo2 assemblers. The average number of contigs generated by Mira, Trinity, and SOAPdenovo2 was 5, 36, and 283, respectively across all variant PIDs from 75 to 99.96%. Specifically, Mira and Trinity generated fewer contigs at low PID, but produced many contigs when the two variants reach 97.1% PID and 96.0% PID, respectively. For SOAPdenovo2, a larger number of contigs were produced regardless of the PID. This indicates that these assemblers generally have major challenges producing a single genome; this has been observed in previous studies comparing assembly performance [22].

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

Effect of GC content and genome length on variant assembly

For Experiment 2, we focused our study on evaluating whether VI observed in SPAdes de novo assembly is influenced by the GC content or genome length of the pathogen. SPAdes was chosen because it produced a

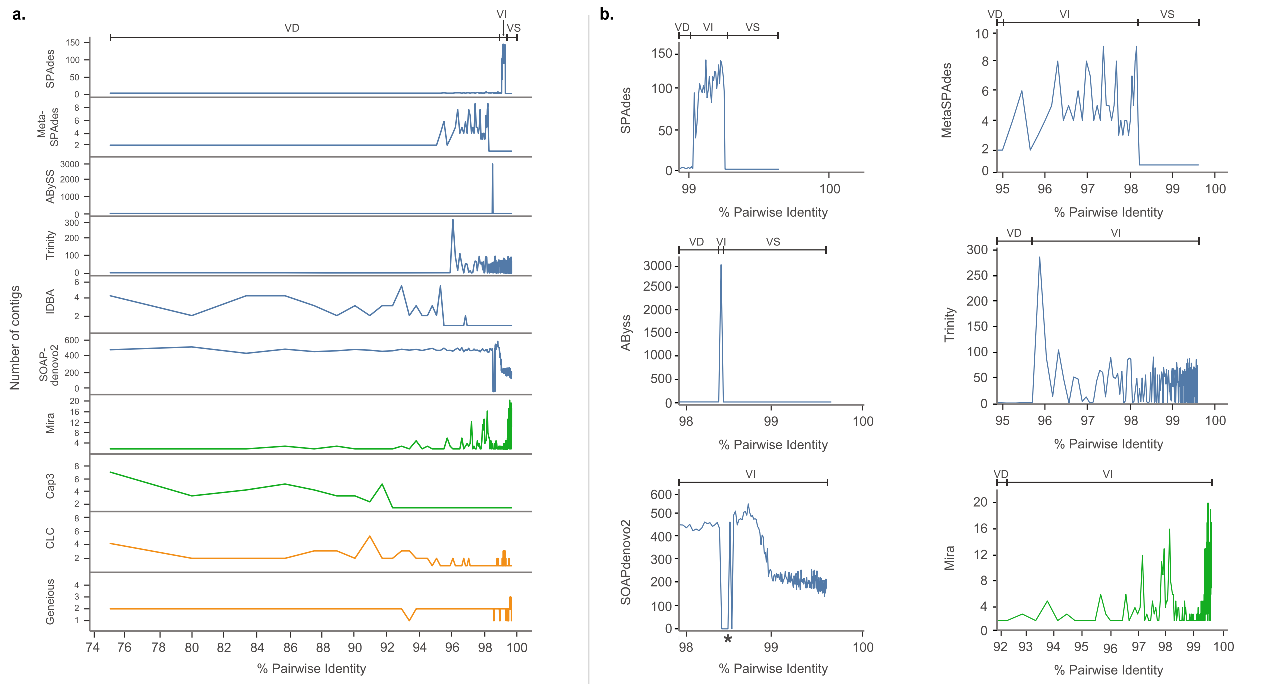

Fig. 3 . Variant interference in 10 de novo assemblers. **a** Schematic diagram depicting concepts of the VD, VI, and VS, and their relationship to PID. **b** Comparison of output from 10 different assemblers. The number of contigs produced by each de novo assemblers at different variant PID ranges (75–99.6%) were shown. **c** Close-up of PID ranges where variant interference is the most apparent. Blue denotes de Bruijn graph assemblers (DBG); green denotes overlap-layout-consensus assemblers (OLC); orange denotes commercialized proprietary algorithms. Variant distinction, VD; variant interference, VI; variant singularity, VS. *For SOAPdenovo2, several data points returned zero contigs due to a well-documented segmentation fault error. The y-axis denotes the number of contigs

well-defined variant interference that closely resembled the conceptual model (Fig. 2). It is also one of the leading assemblers for viral assembly (Fig. 1), possibly due to its ability to assemble viral variants without variant interference in most PID. Two datasets were used for the evaluation: reads generated from four artificial genomes ranging in length from 2 Kb to 1 Mb, as well as from genome sequences of poliovirus (NC_002058; 7440 nt in length) and coronavirus (NC_002645; 27,317 nt in length). No discernable correlation was observed between the GC content of variant genomes and the degree of VI for any of the simulated datasets (Supplemental Figure S1, $p < 0.0001$). Therefore, for subsequent analyses examining the effects of genome length on VI, the number of contigs at each PID level was obtained by averaging the 13 GC simulations.

Notably, no matter the genome length, SPAdes produced vastly more contigs (i.e., VI) in a constant, narrow range of PID (99–99.21%; Fig. 4a & b). The effect of variants on assembly was characterized by the three distinct intervals described previously: VD at lower PIDs, VI (Fig. 4b), and VS at higher PIDs for all genome lengths. For example, during VS, a single contig was generated when the two variants shared ≥99.22% PID, but tens to thousands of contigs were generated at a slightly lower PID of 99.21%. This PID threshold, 99.21%, marked the drastic transition from VI to VS, whereas the transition from VD to VI (i.e., the VD threshold) occurred at 98.99% PID (Fig. 4b). A correlation was observed between genome length and the number of contigs produced during VI, where longer genomes returned proportionally more contigs as expected as total VI occurrence should increase with length ($r^2 = 0.967$; $p <$ 0.0001 Fig. 4b and c).

Effect of read length on variant assembly

The read length of a given NGS dataset will vary depending on the sequencing platform and kits utilized to generate the data. Since read length is an important factor for de novo assembly success, [23] we hypothesized that it may also influence the ability to distinguish viral variants. For Experiment 3, using SPAdes we investigated assemblies with four typical read lengths: 50, 100, 150, and 250 nt. At longer read lengths, the VD threshold occurred at higher PIDs (Fig. 4d & e). Also, with increasing read length, the width of the PID window where VI occurs gradually decreased from a 1.52% spread to a 0.21% spread (Fig. 4e). This indicates that longer reads are better for distinguishing viral variants with high PIDs.

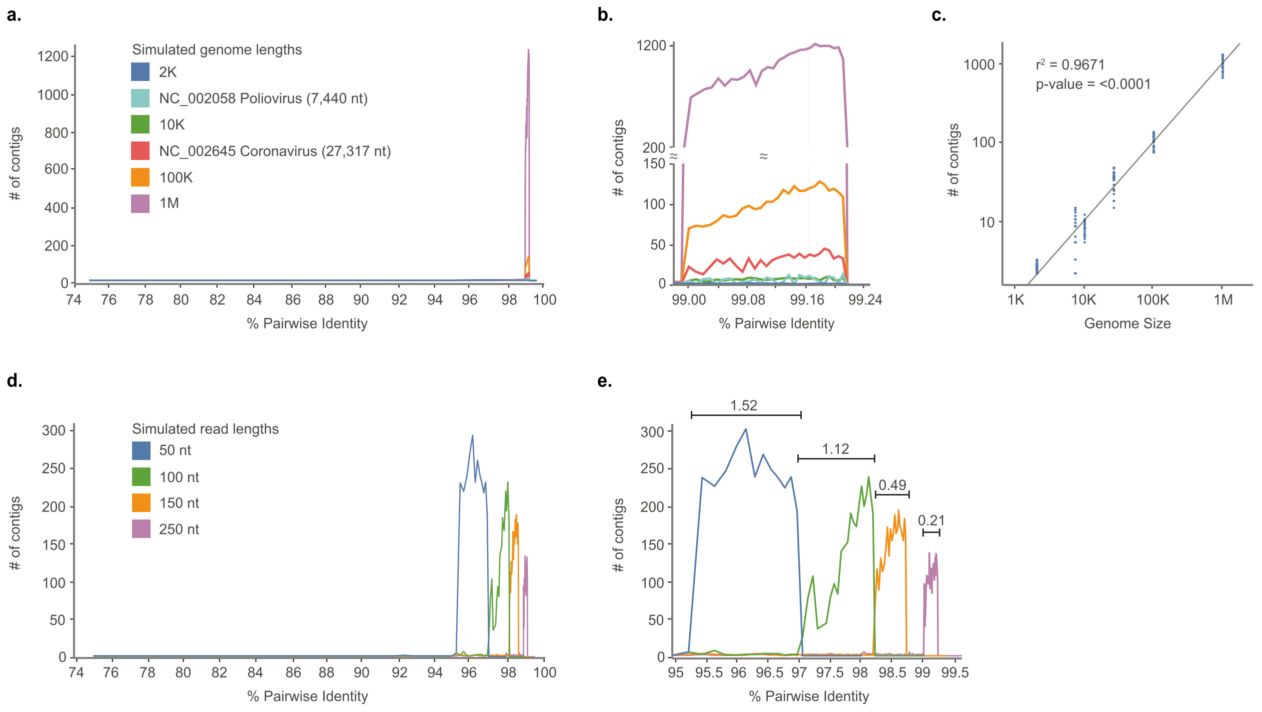

Fig. 4 The effect of genome length and read length on de novo assembly of simulated variants across a range of percentage identities (PID). **a** & **b** Comparison of genome lengths. Six different genome lengths were assembled and the final contig counts were tallied across varying PID thresholds (75–99.6%).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

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

For clinical samples, assembly of viral genomes is affected by multiple factors other than the presence of variants, including sequencing error rate, host background reads, depth of genome coverage, and the distribution (i.e., pattern) of genome coverage. We next utilized viral NGS data generated from four picornavirus-positive clinical samples (one coxsackievirus B5, one enterovirus A71, and two parechovirus A3) to explore VI in datasets representative of data that may be encountered during routine NGS. The NGS data for each sample was partitioned into four bins of read data: (1) total reads after quality control (T); (2) major variants only (M); (3) major and minor variants only (Mm); and (4) major variants and background non-viral reads only (MB) (Fig. 5). These binned datasets were then assembled separately using three assembly programs: SPAdes, Cap3, and Geneious. These programs were chosen as representatives of different assembly algorithms: SPAdes is a leading de Bruijn graph (DBG) assembler, Cap3 is a leading overlap-layout-consensus (OLC) assembler, and Geneious is a proprietary software. By comparing these manipulations, we aimed to test the hypothesis that minor variants directly affect the performance of assembly through VI in real clinical NGS data.

Even with an adequate depth of coverage for genome reconstruction, assembly of total reads (T) in 11/12 experiments resulted in unresolved genome construction – resulting in numerous fragmented viral contigs (Fig. 6). The only exception was one experiment where one single PeV-A3 (S1) genome was assembled using Cap3. When only reads from the major variant were assembled (M), full genomes were obtained for all datasets using SPAdes and Cap3, and for the CV-B5 sample using Geneious. Conversely, assembly of the read bins containing major and minor variants (Mm) resulted in an increased number of contigs for 9 of the 12 sample and assembly software combinations tested (Fig. 6), indicating that VI due to the addition of the minor variant reads likely

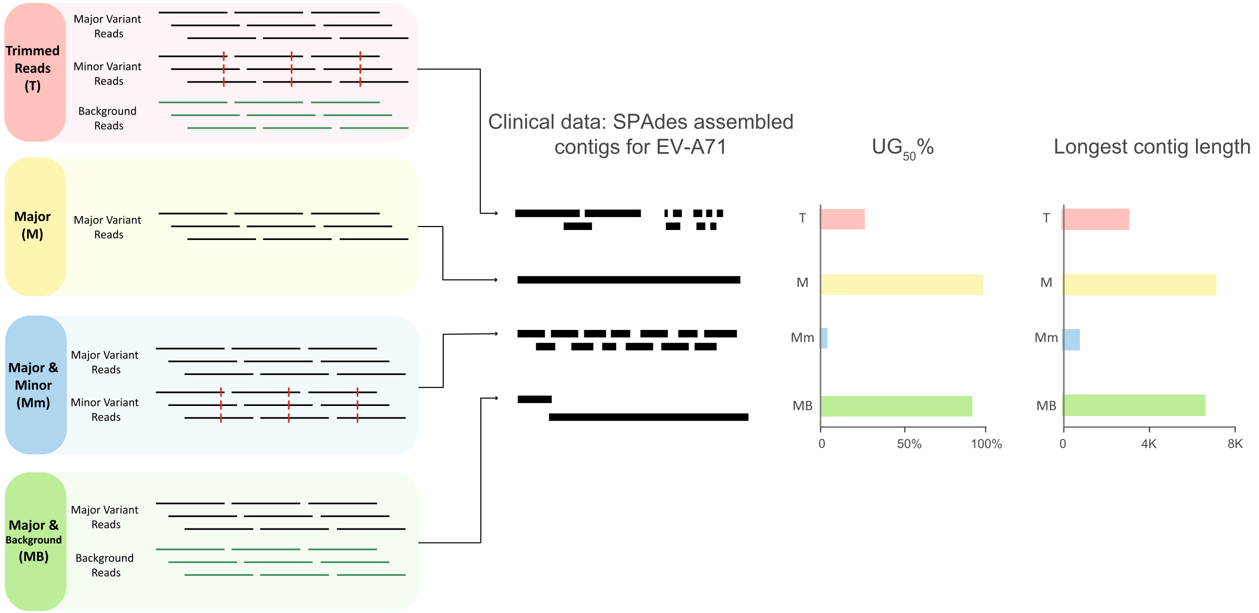

Fig. 5 The effect of variant interference in a real dataset from a clinical sample containing enterovirus A71 (EV-A71) and its variants. Fastq reads were partitioned into four components: trimmed reads after quality control (T), major variant (M), minor variant (m), and background (B). These reads were then combined into four different experiments: T, M, Mm, and MB and assembled using SPAdes. The contig representation schematic showing the abundance and length of the generated contigs reveals the impact of variant interference on de novo assembly. The bar graphs show the UG_{50}% metric and the length of the longest contig. UG_{50}% is a percentage-based metric that estimates length of the unique, non-overlapping contigs as proportional to the length of the reference genome [24]. Unlike N_{50}, UG_{50}% is suitable for comparisons across different platforms or samples/viruses. More clinical samples and viruses are analyzed similarly in Fig. 6

adversely affected the assembly. The presence of background reads with major variant reads (MB) did not appear to affect viral genome assembly, as the UG_{50}% value, a performance metric which only considers unique, non-overlapping contigs for target viruses [24], was similar between M and MB datasets.

Discussion

Our analysis of the GenBank entries quantified the decade-long expansion of NGS technologies and de novo assembly for viral sequencing (Fig. 1). As the number of viral sequences in public databases continues to grow, an important question that naturally arises is how well

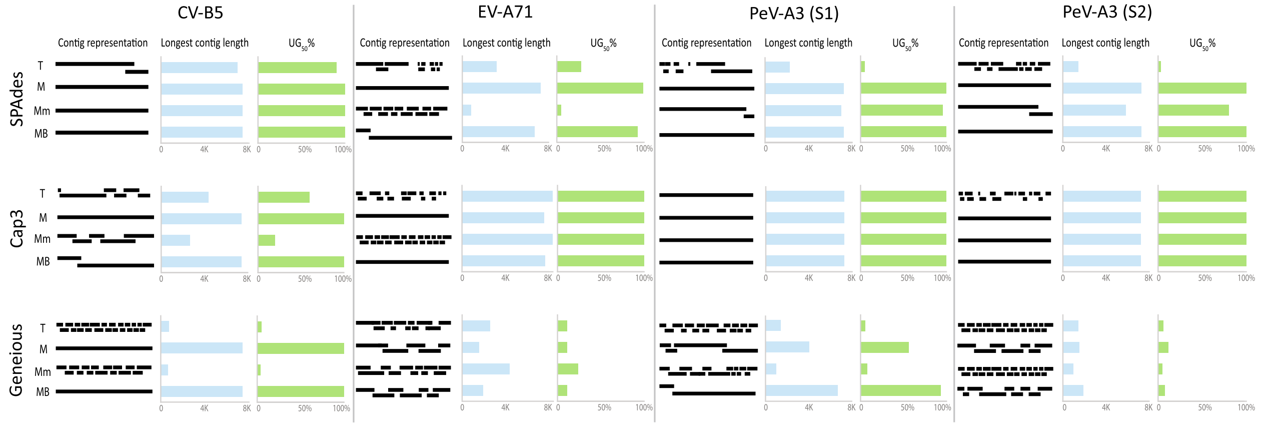

Fig. 6 The effect of variant interference on the assembly of four clinical datasets using three assembly programs. Fastq reads were partitioned into four categories: total reads (T), major variant (M), minor variant (m), and background (B). These reads were then combined into four different categories: T, M, major and minor variants (Mm), and major variant and background (MB). Datasets were assembled using SPAdes, Cap3, and Geneious. The bar graphs show the UG_{50}% metric and the length of the longest contig. Coxsackievirus B5, CV-B5; Enterovirus A71, EV-A71; Parechovirus A3 (Sample 1), PeV-A3 (S1); Parechovirus A3 (Sample 2), PeV-A3 (S2)

current de novo assembly programs perform for datasets with viral variants. Viral variants are expected in biological samples, with the number of variants and the extent of the sequence divergence between variants related to the mutation rate of the virus and the types of specimens that are being investigated. For example, samples containing rapidly evolving RNA viruses, such as poliovirus, HIV, and HCV [9, 11, 25], environmental samples, [26] and clinical samples from immunosuppressed individuals [27, 28] usually harbor many variants. The ability to accurately distinguish variants is imperative to inform treatments (in the case of HIV and HCV), or determine whether a subpopulation of a more virulent variant is present.

Several experiments using simulated and clinical sample NGS data were performed to evaluate the ability of genome assembly programs to distinguish genome variants. All assemblers investigated generated fragmented assemblies when the data contained reads from two closely related variants due to “variant interference” (VI). Changes in pairwise identity (PID) as small as 0.01% between the two variants triggered an assembler to change from producing one or two contigs to producing hundreds of contigs. A quintessential example of this phenomenon was the SPAdes assembly of EV-A71 sequences during the in silico experiments with clinical NGS data. Assembly of major variant reads resulted in one full length contig (Fig. 5), whereas assembly of datasets containing the major and minor variant reads (Mm and T) were characterized by a number of contigs, resulting in “cobwebs” of contig fragments when visualized using Bandage (Supplement Figure S2) [29]. Even though the de novo assembly graph linked the different contig fragments, the assembly could not differentiate the multiple routes of possible contig construction. We speculate this is the main reason why VI occurs in the context of de Bruijn graph assemblers.

The simulated experiments suggested that genome length and read length influence VI; A longer genome length will produce proportionally more contigs during VI, whereas a longer read length decreases the PID range where VI occurs (Fig. 4). While longer read length improves assembly, unfortunately, platforms that produce long reads such as Oxford Nanopore and PacBio have higher error rates [30]. Until long reads can be produced at high fidelity, researchers must continue to rely on combining long- and short-read NGS datasets, and genome polishing techniques [30].

The large number of contigs generated due to VI may be overwhelming for most researchers, and for viral ecology studies, could lead to over-estimation of species richness for methods that use contig spectra to infer richness, such as PHACCS or CatchAll [31–33]. This phenomenon may also impact studies differently depending on the overall goal for generating viral sequence data. For example, some researchers may only be concerned with generating a single major consensus genome, even when variants are detected in the data. This is common during outbreak responses for pathogens such as Ebola virus or Middle East respiratory syndrome coronavirus, where detection of SNPs (indicative of minor variants) is not immediately important. On the other hand, some investigations could favor distinguishing variants, such as for investigating the presence of vaccine-derived poliovirus, where a small number of SNPs may distinguish a vaccine-derived strain from a normal vaccine strain genome [28].

The effects of VI could potentially be mitigated by running multiple assembly programs. A previous study testing bioinformatics strategies for assembling viral NGS data found that employing sequential use of de Bruijn graph and overlap-layout-consensus assemblers produced better assemblies [22]. We speculate that this “ensemble strategy” [22] may perform better because the multiple assemblers complement one another by having different VI PID thresholds. Future assembly approaches could also consider resolving the VI problem by possibly discriminating the major and minor variant reads first (perhaps by coverage or SNP analysis), and then assembling major and minor variant reads separately.

Since we observed VI occurring in simulated data from 2 Kb to 1 Mb genome lengths, we speculate that it may not only affect viral data but also larger draft contigs of bacteria and other microorganisms. Even though bacterial mutation rates are much lower than those of most viruses, bacterial variants are common. For environmental studies, bacterial metagenomes are known to contain many related taxa and variants [34–37], and in clinical investigations, minor bacterial variants can harbor SNPs that provide resistance against antimicrobials. This warrants future investigation into how the presence of variants may impact the assembly of other microbial datasets.

Conclusion

This study aimed to understand how variants affect assembly. As an initial investigation, many confounding factors were simplified for experimentation. Simulated variants studied here only depicted periodic mutations, set at regular intervals. However, in real viral data, SNPs are never evenly distributed across the genome, with zones of divergence and similarity [38, 39]. Other important factors which influence genome assembly include sequencing error rates, presence of repetitive regions, and coverage depth. We limited our experiments to keep these factors constant in order to investigate the sole effect of VI. A pilot experiment analyzing three variants demonstrated the VI theory generally holds true even with multiple variants with varying PID (Supplement Figure S3). Introducing a third variant within the VI threshold (Set B) destabilized assembly and caused the contig number to inflate. Through this study, we demonstrated that reads from related genome variants adversely affect de novo assembly. As NGS and de novo assembly have become essential for generating full-length viral genomes, future studies should investigate the combined effects of the number and relative proportion of minor variants, as well as additional assembly factors (e.g., error rates) to supplement this work.

Methods

Analyzing NGS and assembler usage in the virus nucleotide collection in GenBank

Viral sequence entries from the GenBank non-redundant nucleotide collection were obtained by downloading all sequences under the virus taxonomy through the end of 2019. A total of 2,793,810 GenBank entries were investigated.

The total number of viral sequences submitted annually in GenBank through December 2019 was calculated by filtering GenBank submissions by "virus," followed by application of the following additional filtering steps: "genomic DNA/RNA" was selected and a "release date: Jan 1 through Dec 31" was applied to find the total number of viruses for a given year. A custom script was used to filter and count all documented sequencing technologies and assembly methods used for each GenBank entry.

Creation of simulated variant genomes and reads

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

Following the generation of initial and mutated variant genomes, high-quality fastq reads were generated using ART, [40] simulating Illumina MiSeq paired-end runs at 50X coverage with 250 nt reads, DNA/RNA mean fragments size of 500, and quality score of 93. Fastq reads were combined in equal numbers for the initial and mutated variants and used as input for subsequent de novo assembly experiments (Supplement Figure S1). The same process was utilized to generate the artificial genomes, initial and mutated variant genomes, and reads for each of the experiments.

Experiment 1: analyzing simulated reads from variants using different de novo assembly programs

The simulated datasets containing reads from two variant genomes with nucleotide pairwise identity ranging from 75 to 99.6% were analyzed using 10 different genome assembly programs using default parameters. The de novo assembly algorithms used were either overlap-layout-consensus (OLC) [Cap [41] and Mira [42, 43]], de Bruijn graph (DBG) [ABySS [44], IDBA [45], MetaSPAdes [46], SOAPdenovo2 [47], SPAdes [48], and Trinity [49]], or commercial software packages [CLC (https://www.qiagenbioinformatics.com/) and Geneious [50]] whose assembly algorithms are proprietary (Supplement Table S6). The simulation settings for the reads were paired-end reads, 250 nt read length, and 50X coverage. A total of 2470 assemblies (247 datasets per genome X 10 assemblers) were analyzed (Supplement Figure S1a).

Experiment 2: simulated data by varying genome length and GC content

Artificial genomes were constructed for four genome lengths: 2 Kb, 10 Kb, 100 Kb, and 1 Mb, with varying GC content from 20 to 80%, in 5% increments (Supplement Figure S1b). Datasets derived using one poliovirus genome (NC_002058) and one coronavirus genome (NC_002645) were also included in this analysis, representing the lower and upper genome length range typical of RNA viruses. The original GC content was kept constant for the poliovirus and coronavirus genomes. For all of these genomes, simulated reads for initial and mutated variants were generated as above.

A total of 13,338 SPAdes assemblies were generated, which included 12,844 assemblies for the four artificial genomes (247 datasets per genome X 4 artificial genome lengths X 13 GC content proportions X 1 assembler) and 494 assemblies for the poliovirus and coronavirus datasets (247 datasets per genome X 2 genomes X 1 assembler) (Supplement Figure S1b). JMP v13.0.0 (www.sas.com) was used to calculate Pearson's correlation and Spearman's ρ values to compare the association between percent GC levels and the number of contigs produced at each PID level. Since there was little statistical difference when comparing the contig numbers generated at varying percent GC for each of the four genome length datasets (Spearman's $\rho = 0.8299$ to 0.9801, $p < 0.001$) (Supplement Excel file), the final contig number was averaged across the 13 GC percentages at a given PID. The average contig number was used for plotting the contig assembly results vs percent PID for each simulated genome length (Fig. 4a-b).

Experiment 3: simulated data by varying read length

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

Evaluation of NGS datasets from clinical samples

Four datasets derived from clinical samples containing picornaviruses (one enterovirus A71 [EV-A71], one coxsackievirus B5 [CV-B5] and two parechovirus A3 [PeV-A3]) were analyzed for this experiment, as previous sequencing analysis using Geneious indicated the presence of genome variants. The datasets were analyzed using an in-house pipeline (VPipe), [25] which performs various quality control (QC) steps and de novo assembly using SPAdes. The post-QC reads were considered total reads (T) and mapped to their respective reference genome in order to determine the major and minor variants present in each sample. Total reads which mapped with high similarity (≥99%) to the major variant were categorized as reads representing the major variant (M). Unbinned reads from the major variant reference recruitment were used to construct the minor variant consensus using a second round of reference recruitment, and these reads were categorized as the minor variant (m). Remaining reads from the previous two steps were considered background (B) reads.

De novo assembly for each of the four clinical samples was performed for the following binned NGS datasets: (1) total reads only (T); (2) major variants only (M); (3) major and minor variants only (Mm); and (4) major variants and background reads only (MB). This was repeated with three assembly programs: SPAdes, Cap3, and Geneious. The length of the longest contig produced from each assembly and the performance metric UG_{50}% [24] were calculated to compare the results for these 48 assemblies (4 experiments X 4 viruses X 3 assemblers).

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-06801-w.

Additional file 1. Supplemental Text: Analysis of viral GenBank records demonstrated the advent of NGS in viral sequencing in the last two decades. Supplement **Figure S1.** Workflow diagrams of simulated data from data creation through de novo assembly. Supplement **Figure S2.** Analysis of the final contig assembly graphs for a clinical sample containing enterovirus A71 (EV-A71) variants using Bandage. **Supplement Figure S3.** Assembly with three simulated variants. **Supplement Table S1.** Total counts from NCBI's GenBank non-redundant nucleotide database. **Supplement Table S2.** Total count of sequencing technologies for sequences >2000 nt in the NCBI GenBank non-redundant nucleotide database for years 2012–2019. **Supplement Table S3.** Total counts from NCBI's GenBank non-redundant nucleotide database with multiple sequencing technologies listed per entry. **Supplement Table S4.** Total counts from NCBI's GenBank non-redundant nucleotide database of all entries with three and four sequencing technologies listed. **Supplement Table S5.** Total count of assembly programs used to generate sequences >2000 nt in the NCBI GenBank non-redundant nucleotide database. **Supplement Table S6.** The 10 de novo assemblers used for analysis of the simulated data, as categorized by their underlying assembly algorithms.

Additional file 2. Number of contigs generated by SPades using the simulated data of two variants differed from 75% PID to 99.6% PID, across 20%-80% GC content.

Abbreviations

CV-B5: Coxsackievirus B5; DBG: de Bruijn graph; EV-A71: Enterovirus A71; HCV: Hepatitis C; HIV: Human immunodeficiency virus; M: Major variant reads; MB: Major variant and background non-viral reads; Mm: Major and minor variants reads; NGS: Next-generation sequencing; nt: Nucleotide; OLC: Overlap-layout-consensus; PeV-A3: Parechovirus A3; PID: Pairwise identity; SNP: Single nucleotide polymorphism; T: Total reads after quality control; VD: Variant distinction; VI: Variant interference; VS: Variant singularity

Acknowledgements

The authors thank Dr. Steve Oberste for thoughtful suggestions on this work. We also thank the CDC Office of Advanced Molecular Detection Scientific Computing Team for their computing support.

Authors' contributions

All authors contributed to the conceptualization, data analysis, preparation, and review of this manuscript. C.J.C, R.L.M., and T.F.F.N. wrote this manuscript. The authors read and approved the final manuscript.

Funding

C.J.C, R.L.M., and T.F.F.N. were supported by Federal appropriations to the Centers for Disease Control and Prevention.

Availability of data and materials

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

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Division of Viral Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA. [2]Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA. [3]General Dynamics Information Technology, Inc., contracting agency to the Office of Informatics, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Falls Church, VA, USA.

Received: 2 March 2020 Accepted: 2 June 2020
Published online: 22 June 2020

References

1. Rasmussen AL, Katze MG. Genomic signatures of emerging viruses: a new era of systems epidemiology. Cell Host Microbe. 2016;19(5):611–8.
2. Leung P, Eltahla AA, Lloyd AR, Bull RA, Luciani F. Understanding the complex evolution of rapidly mutating viruses with deep sequencing: beyond the analysis of viral diversity. Virus Res. 2017;239:43–54.
3. Huang SW, Hung SJ, Wang JR. Application of deep sequencing methods for inferring viral population diversity. J Virol Methods. 2019;266:95–102.
4. Perez-Losada M, Arenas M, Galan JC, Bracho MA, Hillung J, Garcia-Gonzalez N, Gonzalez-Candelas F. High-throughput sequencing (HTS) for the analysis of viral populations. Infect Genet Evol. 2020;80:104208.
5. Pierce BG, Keck ZY, Foung SK. Viral evasion and challenges of hepatitis C virus vaccine development. Curr Opin Virol. 2016;20:55–63.
6. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9(4):267–76.
7. Andino R, Domingo E. Viral quasispecies. Virology. 2015;479-480:46–51.
8. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S, et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 2012;8(3): e1002529.
9. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H, Wong K, Stoddard JN, Raugi D, Sorensen S, et al. Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. J Virol. 2011;85(15):7523–34.
10. Jerzak G, Bernard KA, Kramer LD, Ebel GD. Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. The Journal of general virology. 2005;86(Pt 8):2175–83.
11. Lauck M, Alvarado-Mora MV, Becker EA, Bhattacharya D, Striker R, Hughes AL, Carrilho FJ, O'Connor DH, Pinho JRR. Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing. J Virol. 2012;86(7):3952–60.
12. Lin S-R, Hsieh S-C, Yueh Y-Y, Lin T-H, Chao D-Y, Chen W-J, King C-C, Wang W-K. Study of sequence variation of dengue type 3 virus in naturally infected mosquitoes and human hosts: implications for transmission and evolution. J Virol. 2004;78(22):12717–21.
13. Murcia PR, Baillie GJ, Daly J, Elton D, Jervis C, Mumford JA, Newton R, Parrish CR, Hoelzer K, Dougan G, et al. Intra- and interhost evolutionary dynamics of equine influenza virus. J Virol. 2010;84(14):6943–54.
14. Vignuzzi M, Stone JK, Arnold JJ, Cameron CE, Andino R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature. 2006;439(7074):344–8.
15. Thai KTD, Henn MR, Zody MC, Tricou V, Nguyet NM, Charlebois P, Lennon NJ, Green L, de Vries PJ, Hien TT, et al. High-resolution analysis of intrahost genetic diversity in dengue virus serotype 1 infection identifies mixed infections. J Virol. 2012;86(2):835–43.
16. Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ, Qu J, Ryan EM, Zody MC, Henn MR. De novo assembly of highly diverse viral populations. BMC Genomics. 2012;13:475.
17. Khan AR, Pervez MT, Babar ME, Naveed N, Shoaib M. A comprehensive study of De novo genome assemblers: current challenges and future prospective. Evol Bioinformatics Online. 2018;14:1176934318758650.
18. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95(6):315–27.
19. Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, Pop M. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinform. 2019;20(4):1140–50.
20. Rizzi R, Beretta S, Patterson M, Pirola Y, Previtali M, Della Vedova G, Bonizzoni P. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. Quantitative Biology. 2019;7(4):278–92.
21. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. Choice of assembly software has a critical impact on virome characterisation. Microbiome. 2019;7(1):12.
22. Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, Delwart EL. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res. 2015;43(7):e46.
23. Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. Appl Environ Microbiol. 2008;74(5):1453–63.
24. Castro CJ, Ng TFF. U50: a new metric for measuring assembly output based on non-overlapping, Target-Specific Contigs. J Comput Biol. 2017; 24(11):1071–80.
25. Montmayeur AM, Ng TF, Schmidt A, Zhao K, Magana L, Iber J, Castro CJ, Chen Q, Henderson E, Ramos E, et al. High-throughput next-generation sequencing of polioviruses. J Clin Microbiol. 2017;55(2):606–15.
26. Ng TFF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J Virol. 2012;86(22):12161.
27. Ma S, Du Z, Feng M, Che Y, Li Q. A severe case of co-infection with Enterovirus 71 and vaccine-derived poliovirus type II. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology. 2015;72:25–9.
28. Jorba J, Diop OM, Iber J, Henderson E, Zhao K, Sutter RW, Wassilak SGF, Burns CC. Update on vaccine-derived polioviruses - worldwide, January 2017-June 2018. MMWR Morb Mortal Wkly Rep. 2018;67(42):1189–94.
29. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics (Oxford, England). 2015; 31(20):3350–2.
30. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION sequencing and genome assembly. Genomics, Proteomics & Bioinformatics. 2016;14(5):265–79.
31. Herath D, Jayasundara D, Ackland D, Saeed I, Tang SL, Halgamuge S. Assessing species diversity using Metavirome data: methods and challenges. Comput Struct Biotechnol J. 2017;15:447–55.
32. Bunge J, Woodard L, Bohning D, Foster JA, Connolly S, Allen HK. Estimating population diversity with CatchAll. Bioinformatics. 2012;28(7):1045–7.
33. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics. 2005;6:41.
34. Wang NF, Zhang T, Yang X, Wang S, Yu Y, Dong LL, Guo YD, Ma YX, Zang JY. Diversity and composition of bacterial Community in Soils and Lake Sediments from an Arctic Lake area. Front Microbiol. 2016;7:1170.
35. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, et al. The sorcerer II Global Ocean sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007;5(3):e77.
36. The Human Microbiome Project C, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207.
37. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449:804.
38. Schneider WL, Roossinck MJ. Genetic diversity in RNA virus Quasispecies is controlled by host-virus interactions. J Virol. 2001;75(14):6566.
39. Gregori J, Perales C, Rodriguez-Frias F, Esteban JI, Quer J, Domingo E. Viral quasispecies complexity measures. Virology. 2016;493:227–37.
40. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics (Oxford, England). 2012;28(4):593–4.
41. Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res. 1999;9(9):868–77.
42. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 2004;14(6): 1147–59.
43. Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. German conference on bioinformatics. 1999;99(1):45–56.
44. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. ABySS 2.0: resource-efficient assembly of large genomes using a bloom filter. Genome Res. 2017;27(5):768–77.
45. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. In: Research in Computational Molecular Biology: 2010// 2010; Berlin, Heidelberg. Berlin Heidelberg: Springer; 2010. p. 426–40.

46. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27(5):824–34.
47. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1(1):18.
48. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
49. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52.
50. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.