BMC Research Notes

RESEARCH NOTE Open Access

Sequence-matching adapter trimmers generate consistent quality and assembly metrics for Illumina sequencing of RNA viruses

Grace Nabakooza[1,2*], Darlene D. Wagner[2], Nehalraza Momin[2], Rachel L. Marine[3], William C. Weldon[3] and M. Steven Oberste[3]

Abstract

Trimming adapters and low-quality bases from next-generation sequencing (NGS) data is crucial for optimal analysis. We evaluated six trimming programs, implementing five different algorithms, for their effectiveness in trimming adapters and improving quality, contig assembly, and single-nucleotide polymorphism (SNP) quality and concordance for poliovirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and norovirus paired data sequenced on Illumina iSeq and MiSeq platforms. Trimmomatic and BBDuk effectively removed adapters from all datasets, unlike FastP, AdapterRemoval, SeqPurge, and Skewer. All trimmers improved read quality (Q ≥ 30, 87.8 – 96.1%) compared to raw reads (83.6 – 93.2%). Trimmers implementing traditional sequence-matching (Trimmomatic and AdapterRemoval) and overlapping algorithm (FastP) retained the highest-quality reads. While all trimmers improved the maximum contig length and genome coverage for iSeq and MiSeq viral assemblies, BBDuk-trimmed reads assembled the shortest contigs. SNP concordance was consistently high (> 97.7 – 100%) across trimmers. However, BBDuk-trimmed reads had the lowest quality SNPs. Overall, the two adapter trimmers that utilized the traditional sequence-matching algorithm performed consistently across the viral datasets analyzed. Our findings guide software selection and inform future versatile trimmer development for viral genome analysis.

Keywords Next-generation sequencing, Illumina, Quality control, Adapter trimming, RNA viruses, De novo assembly

Introduction

Next-generation sequencing (NGS) has revolutionized infectious disease research and public health, enabling faster pathogen discovery, surveillance, and response [1–4] at a lower cost and higher throughput than traditional Sanger sequencing [5]. NGS sample preparation involves attaching adapters and unique barcodes to the target genomic DNA or cDNA. These sequences are vital on the Illumina NGS platform for flow cell binding, cluster generation, and demultiplexing of target genome reads [6, 7]. When target DNA fragments are shorter than the sequencing run cycle, sequencing may extend into the adapters, resulting in adapter-contaminated reads [8].

*Correspondence:
Grace Nabakooza
sxv8@cdc.gov
[1] Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee, United States, assigned to Centers for Disease Control and Prevention, Atlanta, GA, USA
[2] Eagle Global Scientific LLC, contracting agency to the Centers for Disease Control and Prevention, Atlanta, GA, USA
[3] Division of Viral Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Effective adapter trimming is essential for accurate reference mapping, *de novo* assembly, and SNP calling.

This study used poliovirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and norovirus paired reads sequenced on Illumina iSeq and MiSeq platforms [8] to evaluate the performance of six adapter and quality trimming tools, implementing five algorithms: (i) sequence-matching using global alignment with no gaps, (ii) sequence overlapping with mismatches, (iii) probabilistic overlapping, (iv) *k*-mer based sequence matching, and (v) bit-masked *k*-difference algorithm with mismatches, gaps, and indels (Supplementary Material Sect. 1.1).

Methods

Published adapter trimming software programs were selected based on their unique algorithms, sensitivity, specificity, positive and negative predictive values, and speed. These trimmers included Trimmomatic v0.39 [9] and AdapterRemoval v2.2.2 [10] for sequence-matching, FastP v0.20.1 [11] for sequence-overlapping, SeqPurge v2022_07 [12] for probabilistic overlapping, BBDuk v38.90, a tool included in the BBMap package (https://sourceforge.net/projects/bbmap/) for *k*mer-based, and Skewer v0.2.2 [13] for *k*-difference matching algorithm (Fig. S1). Parameter thresholds for adapter identification and quality trimming, as well as allowed mismatches for read alignments were standardized across trimmers (Supplementary Methods Sect. 2.1 and Table S2). Libraries prepared from random cDNA of 13 poliovirus clinical isolates and amplicons generated from eight SARS-CoV-2-positive nasopharyngeal swabs and seven norovirus-positive stool samples were sequenced using Illumina 300-cycle (2×150 bp, paired-end) MiSeq v2 Micro and iSeq i1 kits following standardized protocols (Supplementary Materials Sect. 2.2). Raw MiSeq and iSeq data were demultiplexed onboard the instrument without adapter trimming. The sequenced viral reads were then processed through the selected trimmers.

Trimmer performance was evaluated by comparing read statistics for raw versus trimmed datasets, including percent residual adapters, read count, length, and base quality (Q≥30). Assembly statistics compared include N50, which is the length of the shortest contig spanning at least 50% of the complete (reference) genome analyzed, maximum contig length (maxContig), "genome coverage" calculated as Maximum contig length (bps)/Viral reference genome Length (bps)×100, adapted from Illumina [14], and single nucleotide polymorphism (SNP) quality and SNP concordance.

Sequence read statistics were calculated using SeqKit v.0.10.1 [15], and quality was assessed using FastQC v0.11.5 [16] and MultiQC v1.9 [17]. Raw and trimmed reads per trimmer were separately assembled *de novo* using SPAdes v3.15.3 [18]. SNP calling was performed using BCFtools v1.10.2 [19] with appropriate viral-specific references described in Supplementary Methods Sect. 2.3–2.5. Results between trimmers were statistically compared using the Wilcoxon signed-rank test with Bonferroni correction and data visualized using ggplot2 in R v4.0.2 (https://www.r-project.org/). The average runtime and memory usage were compared across trimmers for the largest poliovirus dataset.

Results

Residual adapters

Compared to MiSeq, iSeq raw reads had significantly more adapters for all viral datasets analyzed ($p \leq 1.35 \times 10^{-3}$) (Fig. 1 and Table S3). After trimming, residual adapters were detected in AdapterRemoval, FastP, and SeqPurge-trimmed single and Skewer-trimmed paired reads, with FastP retaining the most adapters for poliovirus (0.038−12.54%), SARS-CoV-2 (0.043−13.06%), and norovirus trimmed reads (0.32−3.51%) (Fig. 1). AdapterRemoval left more adapters in MiSeq than iSeq poliovirus and SARS-CoV-2 trimmed reads ($p < 0.015$). SeqPurge only left detectable adapters in SARS-CoV-2 single reads.

Differences in raw versus trimmed read statistics

Overall, iSeq and MiSeq raw reads showed similar mean total read (paired and single) counts, paired read counts, base counts, and read lengths, except MiSeq generated more SARS-CoV-2 raw reads and bases ($p = 0.035$, Table S4). The iSeq generated more high-quality raw reads for poliovirus and SARS-CoV-2 than MiSeq ($p \leq 1.09 \times 10^{-3}$), while no differences were observed for noroviruses.

All trimmers output similar counts of total reads, read pairs, and bases for poliovirus, SARS-CoV-2, and norovirus (Table S5-S7), except BBDuk, had significantly fewer bases for SARS-CoV-2 ($p < 0.028$, Table S6). BBDuk also retained the shortest trimmed viral reads compared to other trimmers ($p \leq 3.12 \times 10^{-5}$, Fig. 2, Table S5-S7). SeqPurge and Skewer consistently output longer trimmed reads than Trimmomatic, AdapterRemoval, and FastP across viruses and sequencers (Fig. S8-S10, panels D and J).

The iSeq poliovirus and SARS-CoV-2 trimmed datasets had significantly fewer paired reads compared to the raw datasets ($p < 0.012$, Tables S4−S6, Fig. S8B and S9B), with Trimmomatic, AdapterRemoval, FastP, and BBDuk consistently retaining fewer trimmed read pairs than raw reads ($p < 0.027$) for both poliovirus and SARS-CoV-2. Also, poliovirus and SARS-CoV-2 trimmed datasets had significantly fewer bases than raw datasets ($p < 5.44 \times 10^{-4}$). Overall, trimmed reads were shorter but with higher quality bases (82.41–96.2% with Q≥30) than raw reads (77.74–93.61%) for poliovirus, SARS-CoV-2,

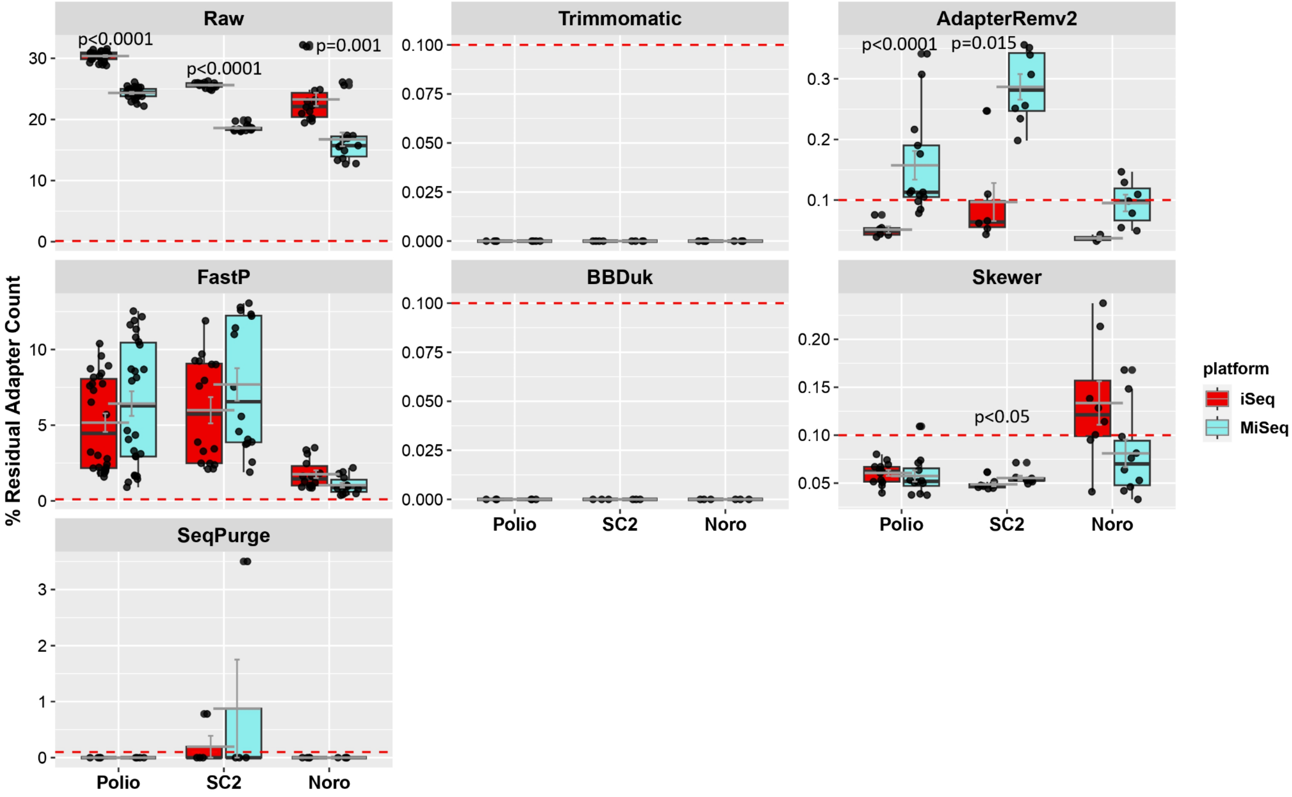

Fig. 1 Differences in percentage residual adapters between iSeq vs. MiSeq poliovirus, severe acute respiratory syndrome coronavirus 2, and norovirus reads, grouped by the trimmer. The median is shown as black bars, while the gray bars show the mean and standard error interval. Only statistically significant differences are indicated with corresponding p-values. Abbreviations: Polio is for poliovirus, Noro for norovirus, and SC2 for severe acute respiratory syndrome coronavirus 2

and noroviruses ($p<3.75\times10^{-3}$, Tables S5–S7, Fig. S8–S10, panels E, F, K and L). Additionally, trimmers preserved longer MiSeq poliovirus and SARS-CoV-2 reads than iSeq ($p\leq5.59\times10^{-3}$, Fig. 2, Table S4), and higher-quality iSeq than MiSeq reads for all three viruses ($p\leq0.035$) (Table S4).

Differences in trimmed read quality

AdapterRemoval, Trimmomatic, and FastP consistently output reads with a higher percentage of quality bases (Q≥30, 93.15–96.7%) than SeqPurge, BBDuk, and Skewer (87.73–95.72%) (Tables S5-S7 and S11, Fig. S8-S10, panels E, F, K and L). Specifically, BBDuk, SeqPurge, and Skewer retained significantly fewer quality iSeq reads across all viruses ($p<7.9\times10^{-3}$) and MiSeq norovirus reads ($p<0.024$) compared to other trimmers. Only AdapterRemoval retained significantly more quality MiSeq SARS-CoV-2 reads than BBDuk and SeqPurge ($p<0.016$), and no quality differences were observed for MiSeq poliovirus reads ($p>0.088$).

Overall, trimmers output more high-quality (Q≥30) iSeq than MiSeq SARS-CoV-2 and norovirus reads ($p<0.035$), with no platform-specific differences for poliovirus reads (Table S4).

De novo assembly statistics

All trimmers except BBDuk improved N50 and maxContig for assemblies across viral datasets compared to raw reads. Notably, BBDuk-trimmed poliovirus and SARS-CoV-2 read assemblies resulted in the lowest N50 ($p<0.037$, Table S12) and maxContig ($p<7.83\times10^{-3}$, Table S13), achieving only 8–39.9% genome coverage compared to raw reads (8.8–87.5%) and other trimmers (54.8–98.9%) (Table 1). Trimmed poliovirus reads assembled into long contigs, significantly improving genome coverage, compared to raw read assemblies, from 35.7 to 98.9% for iSeq FastP-trimmed reads and from 87.5 to 95.6% for MiSeq AdapterRemoval-trimmed reads (Table 1). Assemblies from trimmed norovirus reads showed no significant differences.

MiSeq and iSeq showed comparable mean N50 and maxContig for SARS-CoV-2 and norovirus trimmed reads. However, FastP-trimmed iSeq poliovirus reads assembled longer contigs than MiSeq reads ($p=0.014$, Table S14).

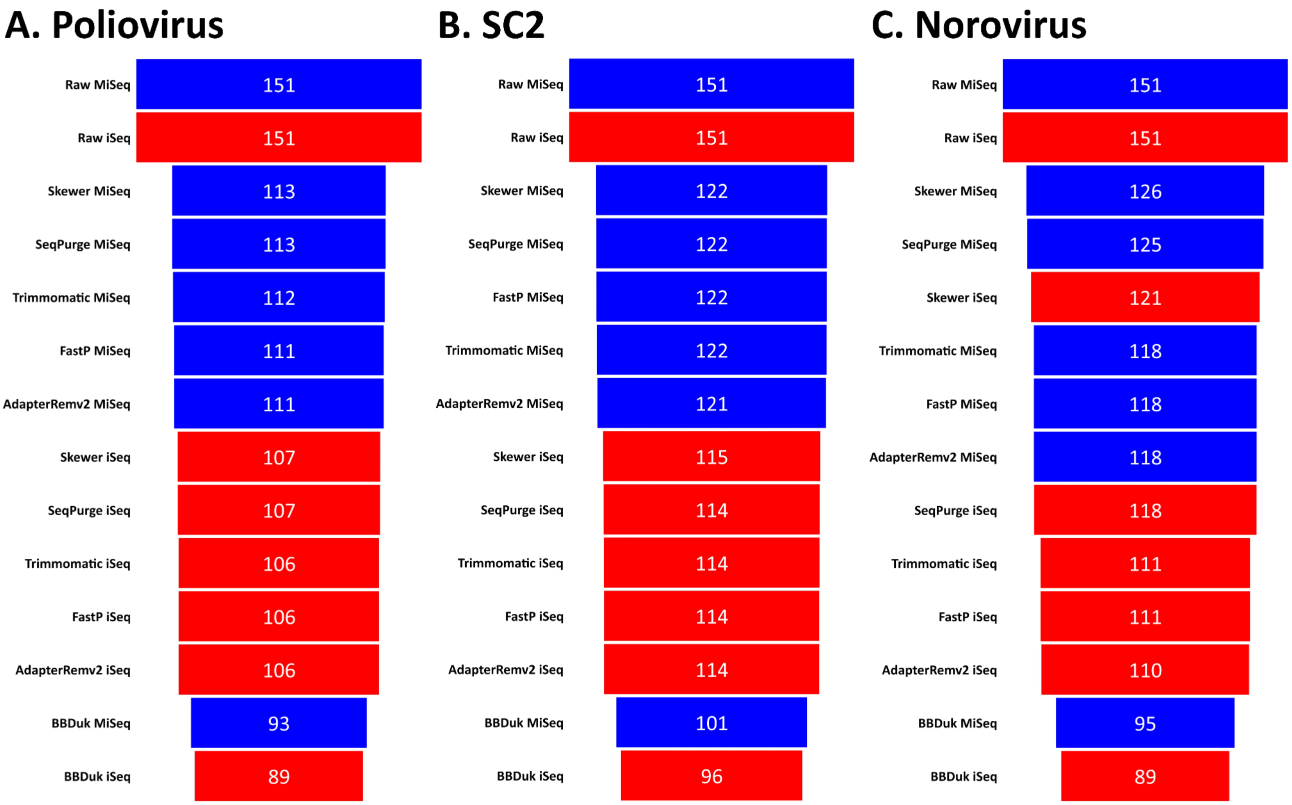

Fig. 2 Mean (average) read lengths of raw and trimmed iSeq (red) and MiSeq (blue) reads for poliovirus **(A)**, severe acute respiratory syndrome coronavirus 2 **(B)**, and norovirus **(C)** datasets. Abbreviations: SC2 for severe acute respiratory syndrome coronavirus 2

Single nucleotide polymorphism (SNP) quality and concordance

There were no differences in SNP quality for SARS-CoV-2 and norovirus datasets across the trimmers. However, BBDuk-trimmed poliovirus read assemblies had lower mean SNP quality than other trimmers (Table S15).

Illumina iSeq and MiSeq read assemblies exhibited SNPs with similar quality, ranging from 3 to 228 for all viruses (Table S14). SNP concordance across trimmers was high (>97.7–100%) for both iSeq and MiSeq viral datasets; however, BBDuk-trimmed read assemblies had 2–8 unique SNPs relative to other trimmers (Fig. S16).

Discussion

We tested six trimming software programs on viral sequencing data generated using Illumina iSeq and MiSeq platforms. Trimmomatic and BBDuk produced the cleanest trimmed reads with the least residual adapters for poliovirus, SARS-CoV-2, and norovirus datasets. Viral reads trimmed using FastP, AdapterRemoval, SeqPurge (SARS-CoV-2 single-reads only), and Skewer exhibited varying levels of residual adapters, with FastP-trimmed reads retaining the highest percentage (0.038–13.06%) across viral datasets. Our results align with a previous study reporting low levels of residual adapters in human cancer data trimmed using AdapterRemoval (0.4%), Skewer (0.1%), and Trimmomatic ($<1.0\times10^{-5}$ percent) [12]. In contrast to our study, high numbers of residual adapters were reported in ChIP-seq human H3K4me1 data trimmed using BBDuk (37.2%) and Trimmomatic (57.7%) [20]. These differences in adapter trimming performance likely depend on specimen type, adapter contamination levels, and trimmer settings. For instance, AdapterRemoval v2.2.2 showed less specificity in trimming single reads with multiple or short (<12 bp) adapters [21]. FastP-trimmed data showed increased residual adapters when allowed mismatches during read alignment exceed four [22], possibly due to overlooking multiple or interweaved adapters, as FastP assumes a single adapter at read tails [11].

All trimmers retained a similar number of total reads, paired reads, and bases for poliovirus, SARS-CoV-2 (except BBDuk), and norovirus datasets. This aligns with a previous analysis of human cancer genes, where SeqPurge, AdapterRemoval, Trimmomatic, and Skewer retained a similarly high percentage (99.9%) of input read pairs [12]. However, analysis of RNA-Seq reads from *Drosophila simulans* gonads and carcasses showed that Skewer retained more usable RNA-Seq read pairs (20% of input reads) than Trimmomatic (14%) and AdapterRemoval (13%) [13]. All trimmers significantly improved data quality ($Q\geq30=87.73-96.07\%$) compared to raw

Table 1 Mean N50 and mean maxContig before and after trimming, grouped by virus and sequencing platform. The value in parentheses is the percent genome coverage represented by the N50 or maxContig. Data in bold highlights the highest mean N50 and mean maxContig values for a given virus/platform and data in italic indicates the lowest values

Starting material	Virus (Genome size)	Platform	Contig statistic	Raw Mean (% genome coverage)	Trimmomatic	AdapterRemoval	FastP	SeqPurge	BBDuk	Skewer
Isolates	**Poliovirus (7,433 bp)**	**iSeq**	N50	2,021 (27.2)	7,196 (96.8)	7,196 (96.8)	**7,354 (98.9)**	6,873 (92.5)	*1,249 (16.8)*	6,873 (92.5)
			maxContig	2,656 (35.7)	7,196 (96.8)	7,196 (96.8)	**7,354 (98.9)**	7,105 (95.6)	*2,203 (29.6)*	7,105 (95.6)
		MiSeq	N50	6,273 (84.4)	7,087 (95.3)	**7,107 (95.6)**	6,579 (88.5)	5,681 (76.4)	*2,087 (28.1)*	6,172 (83)
			maxContig	6,505 (87.5)	7,087 (95.3)	**7,107 (95.6)**	6,812 (91.6)	6,001 (80.7)	*2,952 (39.7)*	6,610 (88.9)
Clinical samples	**SARS-CoV-2 (29,903 bp)**	**iSeq**	N50	1,191 (3.9)	15,122 (50.6)	14,734 (49.3)	17,039 (57)	16,999 (56.8)	*1,736 (5.8)*	**17,383 (58.1)**
			maxContig	2,621 (8.8)	16,388 (54.8)	16,929 (56.6)	18,549 (62)	18,553 (62)	*3,331 (11.1)*	**18,957 (63.4)**
		MiSeq	N50	6,632 (22.2)	19,315 (64.6)	**19,683 (65.8)**	6,669 (22.3)	19,274 (64.5)	*2,402 (8)*	19,269 (64.4)
			maxContig	8,243 (27.6)	19,963 (66.8)	20,384 (68.2)	19,963 (66.8)	**21,920 (73.3)**	*3,890 (13)*	19,957 (66.7)
	Norovirus (7,563 bp)	**iSeq**	N50	2,538 (33.6)	5,118 (67.7)	5,326 (70.4)	**5,330 (70.5)**	3,362 (44.5)	*2,611 (34.5)*	3,705 (49)
			maxContig	4,195 (55.5)	**5,685 (75.2)**	5,613 (74.2)	5,609 (74.2)	4,290 (56.7)	*3,017 (39.9)*	4,703 (62.2)
		MiSeq	N50	2,345 (31)	4,176 (55.2)	**5,335 (70.5)**	4,063 (53.7)	3,612 (47.8)	*1,466 (19.4)*	4,355 (57.6)
			maxContig	2,901 (38.4)	5,017 (66.3)	**5,703 (75.4)**	4,940 (65.3)	4,696 (62.1)	*2,214 (29.3)*	4,871 (64.4)

reads (83.55−93.17%), with AdapterRemoval and Trimmomatic (traditional sequence-matching algorithm) and FastP (overlapping algorithm) producing reads with the highest quality. These tools' better performance could stem from their simultaneous comparisons of read-to-read and adapter-to-read alignments [9–11], effectively removing poor-quality bases. The variation in adapter trimming outcomes observed across studies is likely due to differences in the type of data sequenced (human versus virus) and trimming parameters used.

For raw reads, the iSeq had more detectable adapters than MiSeq ($p \leq 0.001$), likely due to differences in their chemistry, workflow, and flow cell mechanisms, which may bias the average insert length [21]. Despite no platform-specific differences in the number of trimmed reads and bases, trimmers retained longer MiSeq reads but higher-quality iSeq reads, possibly because iSeq reads required more trimming to remove adapters. Differences in assembly metrics were observed between sequencing platforms only for poliovirus, where raw and FastP-trimmed iSeq read assemblies had higher N50 and maxContig values than MiSeq reads. The most pronounced differences were observed between trimmers, with BBDuk-trimmed read assemblies resulting in the lowest N50, maxContig, and genome coverage relative to other trimmers. Trimming poliovirus reads with Trimmomatic improved genome coverage breadth by up to 71.8%, aligning with results showing Trimmomatic increasing N50 and maxContig values for *Escherichia coli* genomes by 58–77% and 28–55%, respectively [9]. In our study, poliovirus assemblies, sequenced from isolates (non-targeted), exhibited higher genome coverage (35.7–98.9%) compared to SARS-CoV-2 (8.7–67.9%) and noroviruses (29.3–75.6%), which were amplified from clinical samples before sequencing. Poliovirus reads may have assembled into longer contigs due to the virus' ability to inhibit host cell RNA synthesis during enterovirus infection, increasing the viral RNA proportion [22].

Identification of high-quality SNPs is crucial for comprehensive genome analysis. Our study found 97.7−100% concordant SNPs per virus across all six trimmers. Sturm et al. also reported high SNP concordance when benchmarking SeqPurge performance on breast and ovarian cancer exon sequences [12]. Notably, BBDuk-trimmed read assemblies had 2−8 additional unique SNPs, possibly due to low read coverage or false-positive SNP calls [12]. Poliovirus assemblies using BBDuk-trimmed reads had the lowest SNP quality compared to other trimmers.

When choosing a trimmer, researchers should consider factors like throughput, speed, and memory usage [11–13]. All trimmers except SeqPurge took < 50 s to process 13 poliovirus datasets. FastP was the fastest, using 12.75 s, while SeqPurge took 3.9 min. Trimmomatic used the most memory (4.4GB). Our findings confirm previous studies showing Trimmomatic and AdapterRemoval offer the highest throughput [10], FastP provides the fastest processing [11], and AdapterRemoval, SeqPurge, and Skewer require less memory [12, 13]. However, with larger datasets, these differences in performance between trimmers may become significant.

Limitations

This study analyzed poliovirus, SARS-CoV-2, and norovirus samples prepared using metagenomic or targeted genome sequencing strategies. These viruses were chosen because they are of significant public health importance and were readily available through collaborating laboratories. However, the small sample size analyzed may not be representative of the complete genomic diversity of these virus types/strains and results may differ for more complex viruses.

Conclusion

This study found that sequence-matching trimmers, Trimmomatic and AdapterRemoval, consistently performed well for viral iSeq and MiSeq data. Overall, current trimming tools demonstrate a trade-off between speed/memory and accuracy/consistency across viral datasets. There is a need for new adapter and quality trimming tools that balance speed and accuracy without compromising on quality.

Abbreviations

SNP	Single-nucleotide polymorphism
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-024-06951-0.

Supplementary Material 1

Acknowledgements

We thank Anna Kelleher and Yan Li from the Coronavirus and Other Respiratory Viruses Division for the SARS-CoV-2 samples, Anna Montmayeur from the Viral Gastroenteritis Branch, Division of Viral Diseases (DVD) for the Norovirus samples, and Ann Frolov and James Bullows from Polio and Picornavirus Branch in DVD for the poliovirus samples analyzed in this study.

Author contributions

G.N. wrote the main manuscript text, analyzed the data, and prepared Fig. 1; Table 1. D.D.W. prepared Fig. 2 and analyzed the SNPs data. N. M. and R.L.M generated the sequence data and wrote the laboratory sequencing methods. W.C.W, M. S. O. and all authors reviewed the manuscript.

Funding

This project was supported in part by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Centers for Disease Control and Prevention. The findings and conclusions are those of the authors and do not represent the official position of the CDC. Evaluation of products does not imply an endorsement by CDC.

Data availability

Viral datasets analyzed in this study were processed to mask human reads and submitted to the NCBI Sequence Read Archive under BioProject ID PRJNA1041361.

Declarations

Ethics approval and consent to participate

Not Applicable.

Competing interests

The authors declare no competing interests.

Conflict of interests

The authors declare no conflicting interests.

Received: 10 April 2024 / Accepted: 19 September 2024
Published online: 14 October 2024

References

1. Gargis AS, Kalman L, Lubin IM. Assuring the quality of Next-Generation Sequencing in Clinical Microbiology and Public Health Laboratories. J Clin Microbiol. 2016;54(12):2857–65.
2. Kanzi AM, San JE, Chimukangara B, Wilkinson E, Fish M, Ramsuran V et al. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Front Genet [Internet]. 2020 [cited 2023 Jul 17];11. https://www.frontiersin.org/articles/https://doi.org/10.3389/fgene.2020.544162
3. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, et al. Next generation sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: approaches, applications, and considerations for development of Laboratory Capacity. J Infect Dis. 2020;221(Supplement3):S292–307.
4. Nabakooza G, Owuor DC, de Laurent ZR, Galiwango R, Owor N, Kayiwa JT, et al. Phylogenomic analysis uncovers a 9-year variation of Uganda influenza type-A strains from the WHO-recommended vaccines and other Africa strains. Sci Rep. 2023;13(1):5516.
5. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. BioMed Res Int. 2012;2012:e251364.
6. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: advances and applications. Biochim Biophys Acta BBA - Mol Basis Dis. 2014;1842(10):1932–41.
7. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014;56(2):61–77.
8. Illumina. How short inserts affect sequencing performance [Internet]. 2023 [cited 2023 Jul 3]. https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000003874
9. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
10. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes [Internet]. 2016 [cited 2022 Dec 21];9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751634/
11. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
12. Sturm M, Schroeder C, Bauer P. SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. BMC Bioinformatics. 2016;17:208.
13. Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics. 2014;15(1):182.
14. Illumina. De Novo Assembly Using Illumina Reads. [cited 2024 Feb 20]; Available from: https://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf
15. Shen W, Le S, Li Y, Hu F, SeqKit:. A cross-platform and Ultrafast Toolkit for FASTA/Q file manipulation. PLoS ONE. 2016;11(10):e0163962.
16. Andrews S, FastQC. A Quality Control tool for High Throughput Sequence Data [Internet]. 2010 [cited 2020 Mar 21]. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

17. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.
18. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo Assembler. Curr Protoc Bioinforma. 2020;70(1):e102.
19. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):giab008.
20. Guzman C, D'Orso I. CIPHER: a flexible and extensive workflow platform for integrative next-generation sequencing data analysis and genomic regulatory element prediction. BMC Bioinformatics. 2017;18(1):363.
21. Illumina. Calculating Percent Passing Filter for Patterned and Non-Patterned Flow Cells. 2017.
22. Lloyd RE. Enterovirus Control of translation and RNA granule stress responses. Viruses. 2016;8(4):93.

Publisher's note

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