PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

In silico analyses identify sequence contamination thresholds for Nanopore-generated SARS-CoV-2 sequences

Ayooluwa J. Bolaji[1,2], **Ana T. Duggan**[1]*

1 Public Health Agency of Canada, National Microbiology Laboratory, Winnipeg, Canada, **2** Cadham Provincial Laboratory, Winnipeg, Canada

* ana.duggan@phac-aspc.gc.ca

OPEN ACCESS

Citation: Bolaji AJ, Duggan AT (2024) *In silico* analyses identify sequence contamination thresholds for Nanopore-generated SARS-CoV-2 sequences. PLoS Comput Biol 20(8): e1011539. https://doi.org/10.1371/journal.pcbi.1011539

Editor: Jinyan Li, Chinese Academy of Sciences Shenzhen Institutes of Advanced Technology, CHINA

Received: September 25, 2023

Accepted: July 14, 2024

Published: August 19, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1011539

Data Availability Statement: The data used in this publication has been made publicly available on Zenodo, https://doi.org/10.5281/zenodo.8206455.

Abstract

The SARS-CoV-2 pandemic has brought molecular biology and genomic sequencing into the public consciousness and lexicon. With an emphasis on rapid turnaround, genomic data informed both diagnostic and surveillance decisions for the current pandemic at a previously unheard-of scale. The surge in the submission of genomic data to publicly available databases proved essential as comparing different genome sequences offers a wealth of knowledge, including phylogenetic links, modes of transmission, rates of evolution, and the impact of mutations on infection and disease severity. However, the scale of the pandemic has meant that sequencing runs are rarely repeated due to limited sample material and/or the availability of sequencing resources, resulting in the upload of some imperfect runs to public repositories. As a result, it is crucial to investigate the data obtained from these imperfect runs to determine whether the results are reliable prior to depositing them in a public database. Numerous studies have identified a variety of sources of contamination in public next-generation sequencing (NGS) data as the number of NGS studies increases along with the diversity of sequencing technologies and procedures. For this study, we conducted an *in silico* experiment with known SARS-CoV-2 sequences produced from Oxford Nanopore Technologies sequencing to investigate the effect of contamination on lineage calls and single nucleotide variants (SNVs). A contamination threshold below which runs are expected to generate accurate lineage calls and maintain genome-relatedness and integrity was identified. Together, these findings provide a benchmark below which imperfect runs may be considered robust for reporting results to both stakeholders and public repositories and reduce the need for repeat or wasted runs.

Author summary

Large-scale genomic comparisons provide a wealth of knowledge, including modes of transmission, rates of evolution, and the impact of mutations on infection, disease severity, and treatment effectiveness. As a result, the public release of genomic data has proven crucial to response of the SARS-CoV-2 pandemic. However, studies continue to show that

Funding: This study was funded/supported by the Public Health Agency of Canada and Genome Canada through the Applied Genomics Innovation for Laboratory Excellence (AGILE) program (formerly the Canadian Public Health Laboratory Network COVID-19 Genomics Program (CCGP)) and the Canadian COVID-19 Genome Network (CanCOGeN), respectively. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

some of the genomic data in public repositories are contaminated from a variety of sources. For instance, the rapid response to the SARS-CoV-2 pandemic prevented many sequencing runs from being repeated, resulting in the occasional upload of imperfect runs to public repositories. It is of note that when genomic data is contaminated, both scientific decisions/studies and public health measures may be compromised. To identify genome contamination threshold(s) for SARS-CoV-2 sequences generated by Nanopore sequencing, computational biology techniques were used to generate artificially subsampled and contaminated libraries. This is the first study of its kind and we hope that the results obtained provide a starting point to investigate and report contamination for groups producing and analyzing NGS data.

Introduction

Genomics and whole genome sequencing of pathogens provide vital information for disease transmission, identification of novel outbreaks, and vaccine candidate selection [1]. Numerous investigations have shown that in the early days of the COVID-19 pandemic, results from genomic monitoring were not only equivalent to epidemiological contact tracing data, [1] but also capable of tracing previously unidentified linked transmissions [2]. It is noteworthy that public health decisions were guided by genomic investigations in some jurisdictions to stop the spread of SARS-CoV-2, including travel bans and stay-at-home orders [1–3]. Thus, rapid whole genome sequencing for SARS-CoV-2 is essential for public health intervention.

Since the SARS-CoV outbreak in 2002–2003, the importance of genomic information for addressing outbreaks brought on by pathogenic coronaviruses has grown. Indeed, progress regarding the studies of this virus shifted dramatically as the complete viral genome was sequenced [4]. However, due to the technology available and the lag in data sharing, it took about 3 months to complete the sequencing of the first complete genome of the SARS-CoV virus [5,6]. Complete genomes were generated in 2002–2003 by first propagating the virus in cell lines, extracting viral RNA from these cell lines, and using a Sanger sequencing approach to produce complete and partial genomes [7]. It is worth noting that advances in genomics have significantly improved sequencing methodologies and timelines in less than two decades, owing to the development of third generation NGS and long-read sequencing technologies. Thus, in late December 2019, the first whole genome sequences of the novel beta coronaviruses, now known as SARS-CoV-2, was obtained using metagenomics and NGS approaches—supplemented with PCR and Sanger sequencing [8–10] and made available online within days. The availability of the SARS-CoV-2 reference whole genome sequences facilitated the development of real-time PCR-based diagnostic assays that helped to understand the transmission patterns and epidemiology of the virus [11]. Both partial and whole genome sequences of SARS-CoV-2 genomes have been reported from around the world and used to monitor the global spread of the virus.

Prior to the 2019–2020 SARS-CoV-2 pandemic, there were approximately 1200 complete betacoronavirus genomes in GenBank. As of July 2023, there were over 15.8 million SARS--CoV-2 genomes available in the Global Initiative on Sharing Avian Influenza Data (GISAID) (https://www.gisaid.org) platform, reflecting a significant increase in the number of available genomes throughout the pandemic. These genomic sequences are generated on different next-generation sequencing (NGS) devices, namely Illumina, Ion Torrent, Oxford Nanopore, and PacBio SMRT platforms. While sequencing technologies have error rates of varying degrees [12,13] genome sequence contamination may also occur during sample preparation and

sample processing at both wet and dry lab steps of the workflow. Also, contamination in reference databases is more concerning than contamination in individual sequencing studies and, according to a few studies, human DNA contamination has been found in non-primate reference genomes [14,15]. GenBank has been reported to contain millions of contaminated sequences, and human contamination in bacterial reference genomes has resulted in thousands of false protein sequences [16]. Therefore, even if researchers properly decontaminated or controlled for contaminants, contamination in reference databases runs the risk of tainting the results of many genomic studies. Further, numerous studies have identified a variety of sources of contamination in public NGS databases and these studies have discovered widespread cross-contamination between samples as well as contamination in sequencing kits and laboratory reagents [16–19].

While NGS is widely used for the rapid detection and characterization of positive COVID-19 cases, one of the drawbacks is that NGS runs are rarely repeated for reasons including limited funds to repeat expensive library preparation reactions even when samples are multiplexed. This has meant that in some cases, the results of imperfect runs are uploaded to public repositories and used to drive public health decisions. Most studies, with few exceptions, do not clearly define the quality control metrics used to include or exclude genomic data from public repositories. Thus, contamination can seriously affect the results of genomic analyses of organisms and viruses leading to spurious alignments and incorrect downstream variant calls.

For this study, we conducted an *in silico* experiment using known SARS-CoV-2 genomes produced from Nanopore sequencing. We assessed the effect of contamination on lineage calls and single nucleotide variants (as a measure of assembly accuracy) using sequences from the same variants and sequences from different variants. The effect of sequencing depth on contamination detection was further investigated using three different bins of reads (12,500 reads, 25,000 reads, and 50,000 reads) as a proxy for sequencing depth. For each sequencing depth, 14 artificially subsampled and contaminated genomes were generated. These samples were generated by mixing clinical SARS-CoV-2 samples *in silico* at different proportions to represent low (1% to 9%) and high (10%, 20%, 30%, 40%, and 50%) contamination levels. Results obtained in this study should help establish internal quality controls and contamination thresholds for SARS-CoV-2 sequences that can be extrapolated from the amount of reads identified within negative controls. Our goal is to provide groups generating the sequencing data with benchmarks to guide their decisions of which runs are of sufficient quality for submission to public repositories and to offer researchers a standard by which results obtained from contaminated SARS-CoV-2 runs can be trusted for variant calling and other downstream reporting.

Methods

SARS-CoV-2 genome sequencing and generation of the *in silico* contaminated libraries

Due to low quantities of viral genomic materials in clinical swab specimens, most SARS-CoV-2 genomes are generated from a tiled amplicon sequencing approach. In this study, amplicons were generated with the Freed primer scheme [20], using tiling PCR and prepared for Oxford Nanopore Technologies sequencing using the ONT Ligation Sequencing Kit (SQK-LSK109) as per the manufacturer's guidelines. The resulting reads were basecalled using the Guppy high accuracy model (5.0.7) with default settings. The average number of reads generated from 812 clinical SARS-CoV-2 samples (representing Alpha, Delta and Omicron lineages) sequenced on MinION and GridION devices were determined using NanoStat (https://github.com/wdecoster/nanostat). Since ONT does not run for a fixed number of cycles, there is no

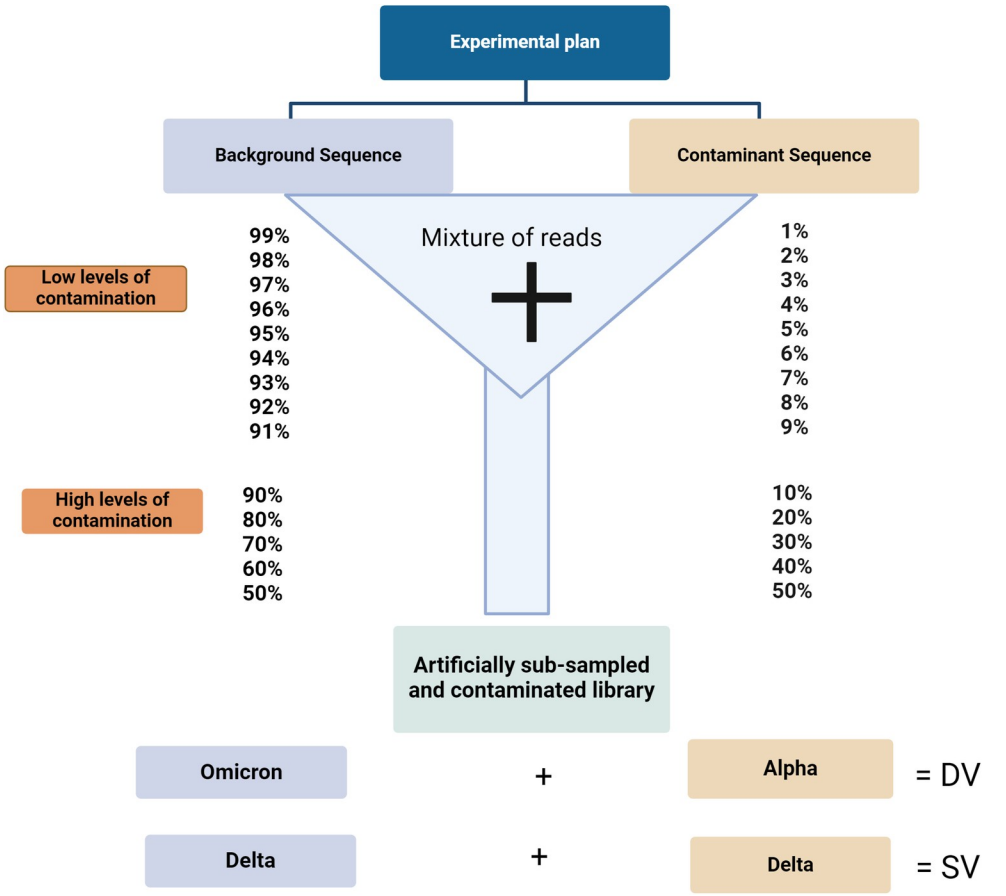

Fig 1. Experimental design of the artificially subsampled and contaminated genomes for the 14 levels of contamination (low and high levels) at three sequencing depths (low—12,500 reads, medium -25,000 reads, and high—50,000 reads). The controlled datasets were generated from known clinical SARS-CoV-2 samples. The sequence combinations were either known omicron sequences mixed with known alpha sequences or known delta sequences mixed with known delta sequences. Textual representation of experimental plan is depicted in Table 1. Created with BioRender.com.

https://doi.org/10.1371/journal.pcbi.1011539.g001

chemistry-induced constraint to measure average run sizes. The results obtained were used as a guide for the selection of the read depths as well as experimental design for the generation of the artificial genomes, where low, medium, and high sequencing depths were represented by 12,500 reads, 25,000 reads, and 50,000 reads, respectively (Fig 1). Artificially subsampled and contaminated reads were generated with seqtk (https://github.com/lh3/seqtk) from 4 clinical sample libraries to represent the contributions of both the known background and contaminant genomes to the artificially contaminated libraries (Fig 1 and Table 1).

Data processing

The artificially generated libraries were processed using a nextflow implementation of the ARTIC pipeline (https://github.com/connor-lab/ncov2019-artic-nf). Variant candidates were identified using Nanopolish (https://github.com/jts/nanopolish). Output files generated from the ARTIC pipeline were further processed using ncov-tools to perform quality control on sequencing results (https://github.com/jts/ncov-tools). Reads were mapped to the reference SARS-CoV-2 genome NCBI GenBank accession (MN908947.3) to further detect the major mutations and single nucleotide variants (SNVs) of consensus sequences for each sample.

Table 1. Standardized terms and parameters of the artificially subsampled and contaminated genomes.

Artificially subsampled and contaminated genomes	Standardized term	Background genome	Contaminant genome	Sequencing depth
Low sequencing depth sample with contaminants from similar variant	LSD_SV	Delta—AY.25.1 genome	Delta–AY.27 genome	Low– 12,500 reads
Medium sequencing depth sample with contaminants from similar variant	MSD_SV	Delta—AY.25.1 genome	Delta–AY.27 genome	Medium– 25,000 reads
High sequencing depth sample with contaminant from similar variant	HSD_SV	Delta—AY.25.1 genome	Delta–AY.27 genome	High– 50,000 reads
Low sequencing depth sample with contaminants from different variant	LSD_DV	Omicron–BA. 1 genome	Alpha– B.1.1.7 genome	Low– 12,500 reads
Medium sequencing depth sample with contaminants from different variant	MSD_DV	Omicron–BA. 1 genome	Alpha– B.1.1.7 genome	Medium– 25,000 reads
High sequencing depth sample with contaminants from different variant	HSD_DV	Omicron–BA. 1 genome	Alpha– B.1.1.7 genome	High– 50,000 reads

https://doi.org/10.1371/journal.pcbi.1011539.t001

Number of consensus SNVs was determined by the number of variants found in the consensus file while the number of variants SNVs is the number of variants found in the iVar variants/VCF files. Number of consensus 'N' was determined by the number of Ns (missing data) in the consensus sequences. Lineages were assigned using Pangolin (version 4.0.3: https://github.com/cov-lineages/pango-designation, pangoLEARN–version 1.2.333: https://github.com/cov-lineages/pangoLEARN). Scorpio call is the output of the constellation assigned to a query by Scorpio–a tool for classifying, haplotyping and defining Variants of Concern or Variants of Interest for a species. As all data used in this study originated from known clinical sample genomes, the metrics of the artificially contaminated libraries were compared against their "true" genome sequence.

The artificially generated datasets (raw reads) as well as their corresponding consensus sequences have been deposited to Zenodo: https://doi.org/10.5281/zenodo.8206455.

Genome pairwise comparison and heat map

Aligned nucleotide consensus genome sequences of both the background and contaminant source clinical samples and the artificially generated genomes were imported to MEGA11 software to calculate pairwise distance. The observed p-distance option was chosen as input for the Model/Method setting while the default options were chosen for the other settings. The pairwise distances output table was imported as a text-delimited file into R v.4.1.1 and the ggplot2 v3.3.1 package was used to generate heat maps for data visualization.

Results

Global nucleotide comparison at different levels of contamination for different sequencing depths

By subsampling the sequences of a known clinical delta sample (AY.25.1) contaminated with reads from another known clinical delta sample (AY.27), we simulated 14 different scenarios to quantify the effect of contamination (Fig 1). To investigate the effect of both low and high levels of contamination on lineage calls and SNVs as a measure of assembly accuracy, we performed a series of global nucleotide comparisons using pairwise p-distance analyses. The distance (proportion) of variant nucleotide sites was compared and plotted as a heat map for all artificially generated samples at the three sequencing depths–low (12,500 reads), medium (25,000 reads), and high (50,000 reads) (Fig 2). These results show that, regardless of sequencing depth and contamination type (i.e., similar (Fig 2A) or different variant contaminants

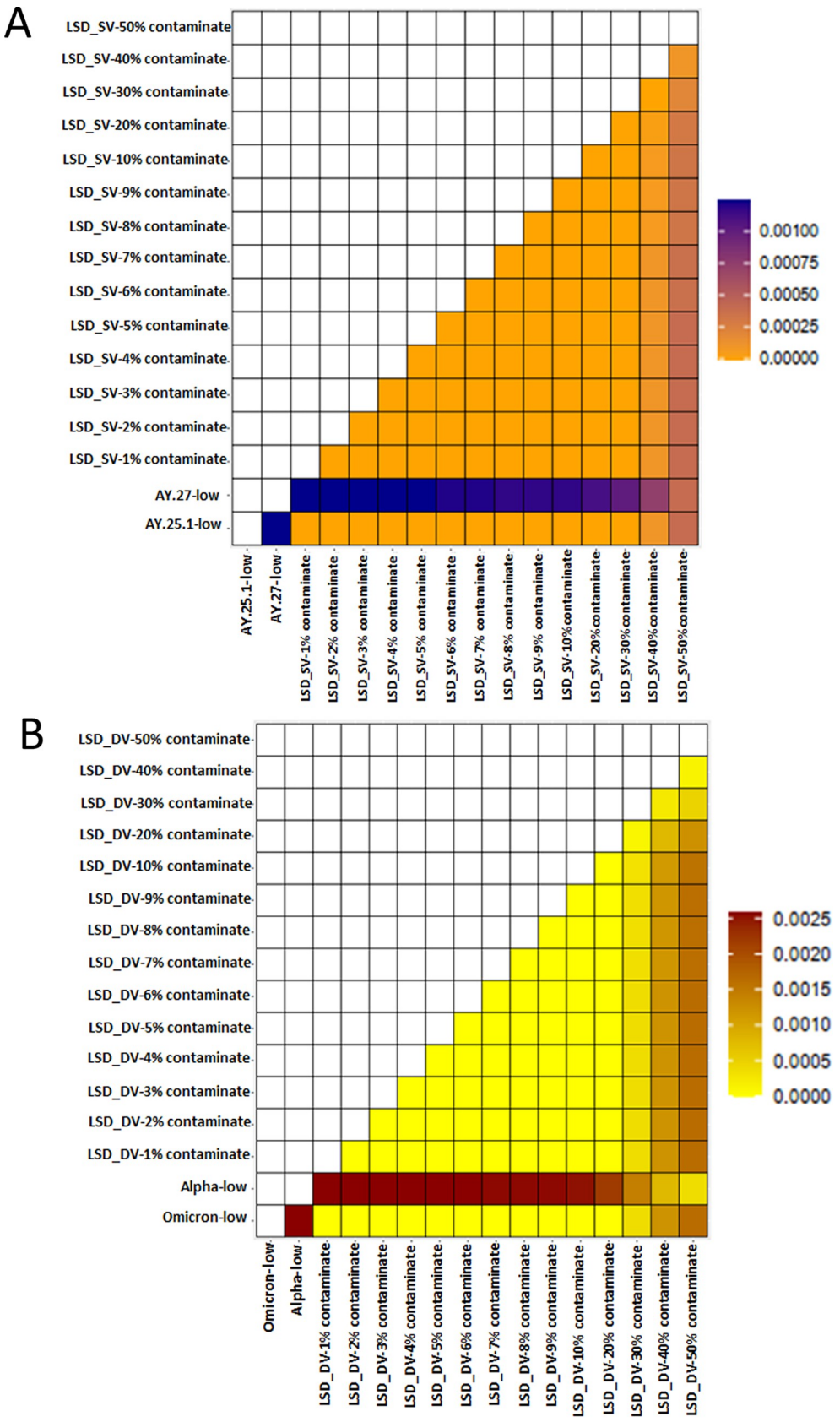
A
LSD_SV-50% contaminate
LSD_SV-40% contaminate
LSD_SV-30% contaminate
LSD_SV-20% contaminate
LSD_SV-10% contaminate
LSD_SV-9% contaminate
LSD_SV-8% contaminate
LSD_SV-7% contaminate
LSD_SV-6% contaminate
LSD_SV-5% contaminate
LSD_SV-4% contaminate
LSD_SV-3% contaminate
LSD_SV-2% contaminate
LSD_SV-1% contaminate
AY.27-low
AY.25.1-low
0.00100
0.00075
0.00050
0.00025
0.00000
B
LSD_DV-50% contaminate
LSD_DV-40% contaminate
LSD_DV-30% contaminate
LSD_DV-20% contaminate
LSD_DV-10% contaminate
LSD_DV-9% contaminate
LSD_DV-8% contaminate
LSD_DV-7% contaminate
LSD_DV-6% contaminate
LSD_DV-5% contaminate
LSD_DV-4% contaminate
LSD_DV-3% contaminate
LSD_DV-2% contaminate
LSD_DV-1% contaminate
Alpha-low
Omicron-low
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000

Fig 2. Global nucleotide comparison of artificially generated contaminated samples and their corresponding clinical SARS-CoV-2 reads as the background samples at a low sequencing depth. A) A heatmap of the pairwise p-distance comparison of the LSD_SV samples—a delta background sequence (AY.25.1) contaminated with a similar delta contaminant sequence (AY.27). B) A heatmap of the pairwise p-distance comparison of the LSD_DV samples–an omicron background sequence (BA.1) contaminated with an alpha contaminant sequence (B.1.1.7).

https://doi.org/10.1371/journal.pcbi.1011539.g002

(Fig 2B)), differences observed for global nucleotide composition is relatively minor for contamination levels less than 50% (see Fig 2 for the low sequencing depth, S1 and S2 Figs for medium and high sequencing depths respectively).

The effect of contamination from similar variants on assembly accuracy and lineage calls

The impacts of contamination on SNVs and lineage call outputs for the SARS-CoV-2 genome were assessed by subsampling and mixing the reads from clinical libraries to simulate contaminated genomes at predetermined sequencing depths. Phylogenetic trees were constructed to examine the impact of SNVs found within each subsampled dataset and sequences from the clinical SARS-CoV-2 reads were used as controls. The identified SNVs were plotted with an associated single nucleotide polymorphism (SNP) matrix (Figs 3, S3, and S4). We have highlighted seven quality control metrics (QC metrics) as important metrics in determining contamination thresholds and the effect(s) of sequence contamination on genome completeness and assembly accuracy. These metrics include the number of consensus SNVs obtained from the number of variants in the consensus sequence, the number of consensus 'N', the number of variants SNVs–obtained from the number of gene sequence variations, the number of variants indels, genome completeness, lineage, and Scorpio call, all compared to the reference SARS-CoV-2 strain.

Changes in both the numbers of consensus SNVs and consensus 'N' (number of undetermined data sites) were investigated as essential determinants of assembly accuracy and completeness. For the LSD_SV genomes (12,500 reads), differences in the two aforementioned metrics were observed for the genomes with contamination levels greater than 5% (625 reads) (Fig 3A and Table 2). For instance, the number of consensus SNVs for the clinical delta AY.25.1 relative to the Wuhan reference strain (GenBank accession MN908947.3) was 40, the number of consensus 'N' was 190 and the number of variants SNVs was 45 (Table 2). As the levels of contamination increased, a decrease in the number of SNVs and an increase in the number of consensus 'N's were observed in the contaminated genomes compared to the clinical control samples (Fig 3A and Table 2). Since these are artificially contaminated datasets, we can compare the known source genome sequence. Further, LSD_SV genomes were assigned incorrect lineage calls at contamination levels greater than 30% (3,750 reads). Thus, for LSD_SV genomes, we conclude the contamination threshold for preserving assembly accuracy is 5% while the identified threshold for lineage calls is 30% (Fig 3A and Table 2). See S3 Fig and S1 Table for results obtained for MSD and HSD samples.

The effect of contamination on assembly accuracy and lineage calls for SARS-CoV-2 sequences for different variants

We investigated the effect of different levels of contamination on SARS-CoV-2 sequences contaminated by different strains (i.e. omicron clinical sample (BA.1), contaminated with an alpha clinical sample (B.1.1.7)). A contamination threshold was identified for changes in SNVs and the number of consensus 'N'—a measure of assembly accuracy and lineage calls at the

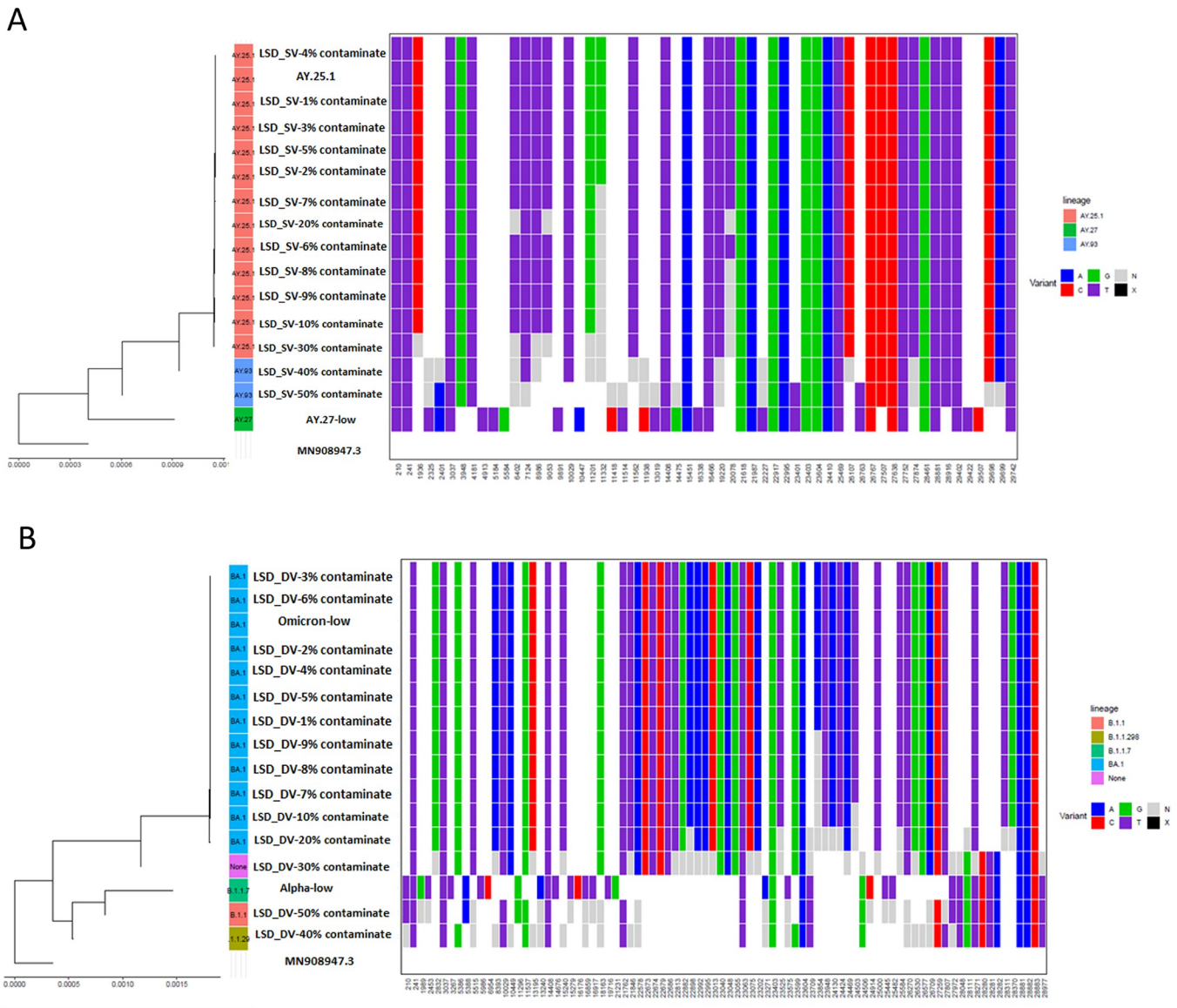

Fig 3. Phylogenetic tree and heatmap showing single nucleotide variants (SNVs) at different positions of the SARS-CoV-2 genome for (A) AY.25.1 (delta variant) contaminated with an AY.27 (delta variant) sequence at contamination levels 1–10%, 20%, 30%, 40%, and 50% at low sequencing depth sequence. (B) BA.1 (omicron variant) contaminated with a B.1.1.29 (alpha variant) sequence at contamination levels 1–10%, 20%, 30%, 40%, and 50% at low sequencing depth sequence.

https://doi.org/10.1371/journal.pcbi.1011539.g003

three sequencing depths beyond which the integrity of the data is altered from the source or "true" clinical sample.

At a low sequencing depth (12,500 reads), the number of consensus SNVs for the clinical omicron BA.1 sample relative to the Wuhan reference strain was 56, the number of consensus 'N' was 189, and the number of variants SNVs was 61 (Table 3). Therefore, we investigated the differences in these QC metrics for each of the artificially generated genomes. For the LSD_DV at a contamination level of 7%, it was observed that the number of consensus SNVs changed from 56 to 55, the number of consensus 'N' increased to 190 while the number of variant SNVs decreased at 60 and other QC metrics remained unchanged at this contamination level (Table 3). However, at 30%, the assigned lineage calls for the artificially generated genome

Table 2. Quality control metrics comparison for artificially subsampled and contaminated genomes of contamination by similar variants at a low sequencing depth–all LSD_SV genomes.

Genome	Num of consensus SNVs	Number of consensus 'N'	Number of variants SNVs	Number of variants indel	Mean sequencing depth	Genome completeness	Lineage	Scorpio calls	Watch mutations
AY.25.1_low	40	190	45	2	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-1% contaminate	40	190	45	2	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-2% contaminate	40	190	45	2	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-3% contaminate	40	190	45	2	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-4% contaminate	40	190	45	2	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-5% contaminate	40	190	45	2	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-6% contaminate	39	190	44	3	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-7% contaminate	39	190	44	3	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-8% contaminate	38	191	43	3	472.1	0.9936	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-9% contaminate	38	191	43	3	472.2	0.9935	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-10% contaminate	38	191	43	3	472.1	0.9934	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-20% contaminate	36	194	41	2	472.2	0.9935	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-30% contaminate	33	196	38	3	472.4	0.9934	AY.25.1	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-40% contaminate	29	202	34	2	472.3	0.9932	AY.93	Delta (B.1.617.2-like)	S:G142D,S:L452R
LSD_SV-50% contaminate	28	203	32	2	472	0.9932	AY.93	Delta (B.1.617.2-like)	S:G142D,S:L452R
AY.27_low	39	189	43	3	471.7	0.9937	AY.27	Delta (B.1.617.2-like)	S:G142D,S:L452R

https://doi.org/10.1371/journal.pcbi.1011539.t002

(LSD_DV) changed from BA.1 to none (Table 3), and this held true for artificial genomes with 40% and 50% contamination. Taken together, for low sequencing depth (LSD_DV), 6% level of contamination (750 reads) was identified as the contamination threshold for the preservation of assembly accuracy while a 20% level of contamination (2,500 reads) was identified as the threshold for accurate lineage call (Fig 3B and Table 3). Results for the MSD and HSD samples can be found in S4 Fig and S2 Table.

To identify the effect of mixture on the artificially subsampled and contaminated genomes, we generated an amino acid mutation heatmap. Mutational profiles and other host-modulating factors have been reported as major contributors to disease severity in COVID-19 [21]. As such, there is a critical need to evaluate the effect of contamination on mutational profiles that may be of clinical importance. The mutational profile compared all defining mutations of the artificially mixed genomes to the clinical SARS-CoV-2 samples and also identified the type/nature of the mutations (conservative in-frame deletion, disruptive in-frame deletion, missense variant, stop gained, and synonymous variant) (Fig 4 and Tables 2 and 3). The amino acid mutation plots reveal the similarity in the mutation profile of each artificially subsampled

Table 3. Quality control metrics comparison for artificially subsampled and contaminated genomes of contamination by different variants at a low sequencing depth–for all LSD_DV genomes.

Genome	Num. of consensus_snvs	Number of consensus 'N'	Number of variants SNVs	Number of variants indel	Mean sequencing depth	Genome completeness	Lineage	Scorpio calls	Watch mutations
BA.1_low	56	189	61	7	470.4	0.9937	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H11
LSD_DV-1% contaminate	56	189	61	4	470.4	0.9937	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-2% contaminate	56	189	61	4	470.4	0.9937	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-3% contaminate	56	189	61	4	470.4	0.9937	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-4% contaminate	56	189	61	4	470.4	0.9937	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-5% contaminate	56	189	61	4	470.4	0.9937	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-6% contaminate	56	189	61	4	470.4	0.9937	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-7% contaminate	55	190	60	4	470.4	0.9936	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-8% contaminate	55	190	60	4	470.4	0.9936	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-9% contaminate	55	190	60	4	470.4	0.9935	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-10% contaminate	54	191	59	4	470.4	0.9936	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-20% contaminate	46	210	51	4	470.6	0.993	BA.1	Omicron (BA.1-like)	S:del69-70,S:K417N,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-30% contaminate	34	214	38	4	470.7	0.9928	None	Probable Omicron (Unassigned)	S:del69-70,S:Q493R,S:N501Y,S:P681H,S:P681H
LSD_DV-40% contaminate	25	214	28	4	470.7	0.9928	B.1.1.298		S:del69-70,S:N501Y,S:P681H,S:P681H,S:T716I,S:S982A
LSD_DV-50% contaminate	26	210	28	4	471	0.993	B.1.1		S:del69-70,S:N501Y,S:P681H,S:P681H,S:T716I,S:S982A
B.1.1.7 low	40	192	44	4	471.1	0.9936	B.1.1.7	Alpha (B.1.1.7-like)	S:del69-70,S:del144,S:N501Y,S:A570D,S:P681H,S:P681H,S:T716I,S:S982A,S:D1118H

https://doi.org/10.1371/journal.pcbi.1011539.t003

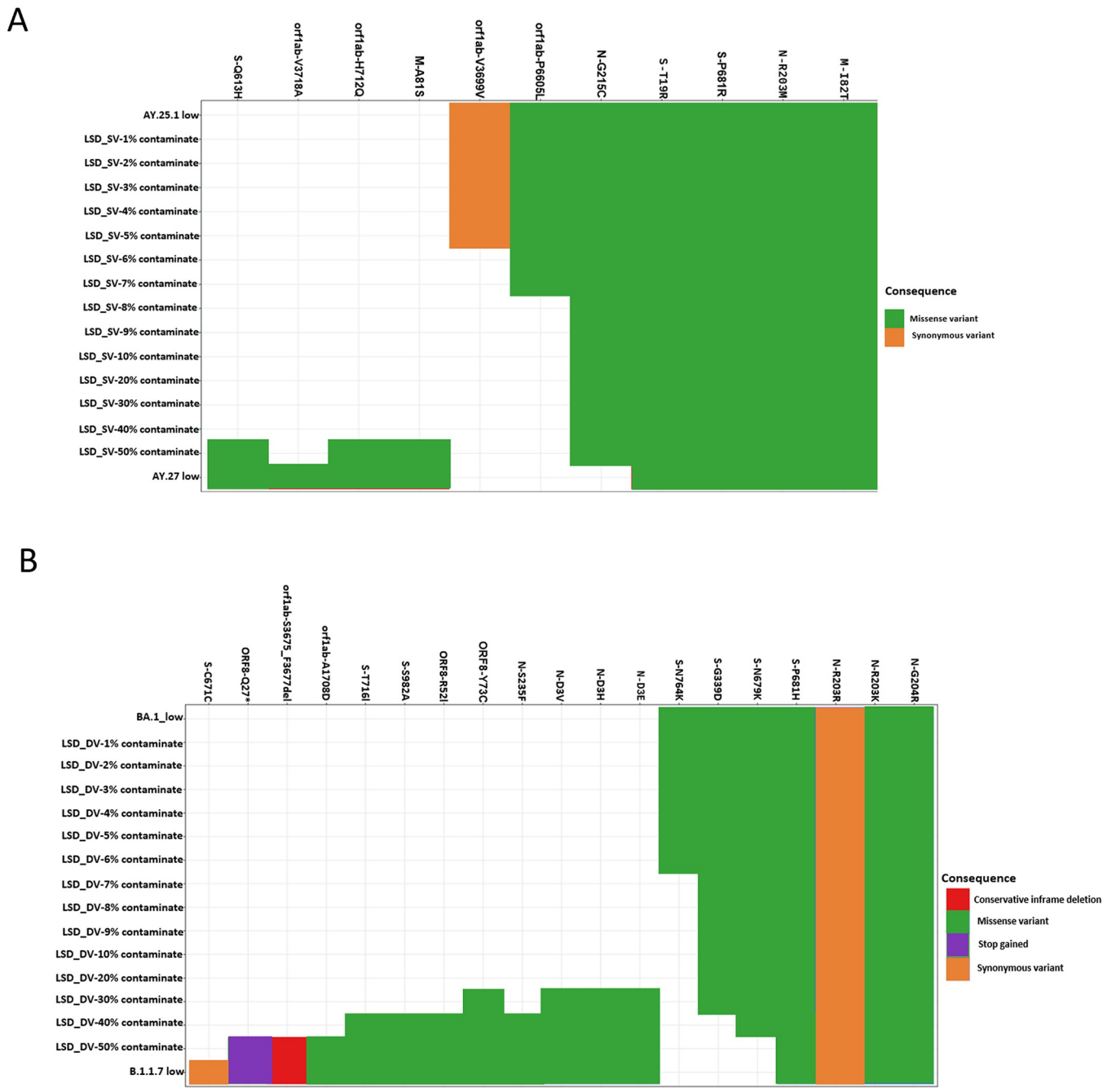

Fig 4. Mutational profile comparison of SARS-CoV-2 genome for the clinical genomes to the artificially generated genomes for (A) LSD_SV (AY.25.1 contaminated with an AY.27 variant) sequence at contamination levels 1–10%, 20%, 30%, 40%, and 50%. (B) LSD_DV (BA.1 contaminated with a B.1.1.29 variant) at contamination levels 1–10%, 20%, 30%, 40%, and 50%.

https://doi.org/10.1371/journal.pcbi.1011539.g004

and contaminated genome and the clinical control samples. Samples contaminated by a similar variant (LSD_SV, MSD_SV, and HSD_SV) also had similar mutational profiles while samples with contaminants from different variants (LSD_DV, MSD_DV, and HSD_DV) had different mutational profiles (Tables 1 and 4). It is noteworthy that the artificially subsampled and contaminated genomes mixed from substrains of the same variant had similar mutational

Table 4. A summary of the identified threshold for the artificially subsampled and contaminated reads as well as the origin of both background and contaminant samples.

Standardized term	The identified threshold for assembly accuracy	The identified threshold for lineage call
LSD_SV	5 percent	30 percent
MSD_SV	4 percent	30 percent
HSD_SV	10 percent	50 percent
LSD_DV	6 percent	20 percent
MSD_DV	7 percent	20 percent
HSD_DV	7 percent	20 percent

https://doi.org/10.1371/journal.pcbi.1011539.t004

profiles to the clinical SARS-CoV-2 samples up to and including 5% contamination (Fig 4 and Tables 2 and 4). While the artificially generated subsamples contaminated from a divergent variant had similar mutational profiles to the corresponding clinical SARS-CoV-2 up to and including 6% contamination (Fig 4 and Tables 3 and 4).

In conclusion, for artificial genomes generated by mixing different SARS-CoV-2 variants (e.g., an omicron sample contaminated by an alpha sample), we identified the following thresholds for data integrity: assembly accuracy is maintained at contamination levels up to 6% for LSD (12,500 reads) and 7% for both MSD (25,000 reads) and HSD (50,000 reads) depths. The integrity of lineage calls is more robust, remaining unaltered up to contamination rates of 20% at all sequencing depths.

Discussion

The scale of sequencing data available in public repositories over the course of the SARS-CoV-2 pandemic is unprecedented. Due to the rapidly evolving nature of the SARS-CoV-2 genome, routine monitoring and public health warnings were crucial in controlling the pandemic. Continuous monitoring and genomic sequencing during the SARS-CoV-2 coronavirus pandemic also hastened the development of the most effective vaccines [22]. However, recurrent mutations in the SARS-CoV-2 genome have tested the efficacy of the vaccines and point to the need for routine updates to both the vaccine targets and vaccination schedules [22,23]. The importance of routine monitoring of SARS-CoV-2 mutations for public health applications cannot be overstated, therefore it is critical that we maintain confidence in the sequences both submitted and pulled from public repositories lest erroneous variants affect major public health decisions [24].

Contaminant-induced mutations have been found and documented in other large-scale genomic studies and it was concluded that these contaminated sequences can spread into and across databases over time [14]. This issue cannot be ignored since genome sequences in public repositories are frequently obtained for comparative studies, forecasting, and decision-making purposes. Therefore, researchers interested in a particular pathogen can collect hundreds of sequences for comparative genomic or phylogenomic investigations in this manner. Lupo et al. demonstrated the presence of mis-affiliated genomes in NCBI RefSeq [25]. While these genomes may not be contaminated in the strictest sense, the dominant organism was not what was expected in the study, leading to problems for downstream analyses and reporting [25]. Despite the findings of these studies, sequences submitted to public repositories/databases are rarely checked for contamination [25]. To further validate the effect of contamination on sequencing data and demonstrate the need for contaminant investigation before data are uploaded to public repositories, this study aimed to identify a contamination threshold for which runs can be considered ideal for upload to public repositories while also offering practical guidelines. While it can be very difficult to identify contamination in a clinical setting, we

completed this study such that the thresholds for contamination identified through the *in silico* experiments can be extrapolated to negative controls in a clinical setting giving the sequence-generating laboratories a guideline to assess before reporting results or uploading data to public repositories.

Since there is no consensus within the scientific community on how to validate assembly accuracy, we investigated the amino acid mutational profile, genome completeness, number of SNVs, number of consensus N, number of variant SNVs, and indels for all samples as a measure of assembly accuracy for this study (Tables 1, 2, S1, and S2)

We further investigated the effect of contamination on the phylogenetic placement and sample relatedness of the artificially subsampled and contaminated genomes (Figs 2, S3, and S4). The results obtained from the phylogenetic analyses are in agreement with the identified contamination thresholds for mutation profile as a measure of assembly accuracy, wherein the artificially subsampled and contaminated genomes with contaminants of less than 5% for LSD_SV and 6% for LSD_DV clustered in the same branch with the corresponding clinical samples. Similar results were also obtained for both MSD_SV and HSD_SV as well as for MSD_DV and HSD_DV. With this observation, we showed that at contamination levels of less than 6%, at all sequencing depths, the artificially subsampled and contaminated genomes were very similar to the genomes generated by the original clinical samples from which they were derived. Thus, we concluded that contamination levels of 5% and below do not affect genome-relatedness and integrity.

By performing a global nucleotide comparison, varying both the levels of simulated contamination and the sequencing depth, we investigated the effect of contamination on the artificially subsampled and contaminated genomes. According to the results obtained from the p-distance pairwise comparison analysis, differences observed for global nucleotide composition among the samples were negligible for contamination levels less than 20% when the metric of interest is simply the lineage assignment irrespective of the sequencing depth and the contamination type. Since p-distance is the proportion of nucleotide sites at which two sequences are different, this result is expected. The analysis performed considers all nucleotides present in the samples compared without any regard for the origin of the nucleotide (i.e., background or contaminant). However, it is noteworthy that with contamination levels greater than 20%, differences were observed at the global nucleotide levels when compared to the original clinical samples at all sequencing depths for both types of contaminants (Figs 2, S1, and S2).

Studies have identified the importance of lineage tracking and its role in providing answers to evolutionary questions about the SARS-CoV-2 genome [26,27]. The extensive recombination between SARS-CoV-2 strains, first identified by so-called "deltacron" lineages with diagnostic mutations associated with both the delta and omicron variants have become identified with increasing frequency since late 2021, and the emergence of the omicron variant [28,29]. Thus, the accurate assignment of lineage calls for SARS-CoV-2 lineages is important, as these lineages also offer insights for clinicians and public health personnel during an outbreak of infection. Based on the above notion, we investigated the effect that the different levels and types of contamination had on the accuracy of lineage calls (Tables 1, 2, S1, and S2). Our results showed that regardless of the type of contaminant (similar or different sequences), a 20% contamination threshold was the maximum amount permissible for accurate lineage calls (Tables 1, 2, S1, and S2).

Foreign sequences can be introduced at many different stages of the sequencing process, from organism culture to data processing [14]. Here, we offer some practical guidelines on how to track contaminants during sequencing experiments and offer recommendations to researchers on steps to take before uploading reads to public repositories. We recommend that researchers include a negative control in the following steps: (i) nucleic acid extraction, (ii)

nucleic acid amplification (if applicable), and (iii) library preparation steps. By having multiple negative controls introduced at different stages of the sequencing experiment, the source of contamination may be identified. We also recommend that these negative controls be carried forward to the data processing steps so that if contamination occurs, the amount of sequenced data present in the negative controls could be investigated and used to determine the appropriate contamination threshold based on the objective(s) of the sequencing experiment in question. Upon the investigation of the negative controls, it is advised that the number of reads mapped in the negative controls should be checked to ensure that the percentage of reads mapped do not exceed the recommended threshold (determined by the average read depth of samples in the run). This practice will also be useful to determine if samples were contaminated by multiple strains.

As this study is the first of its kind, we are aware that these identified thresholds may change as more sequence data become available and as more studies expand on and investigate the parameters required for assembly accuracy and lineage calls. However, we hope that having a standardized method for determining the integrity of genomes and lineage calls will provide a benchmark below which imperfect runs may be considered robust for reporting results to both stakeholders and public repositories, thereby reducing the need for repeat or wasted runs. In this study, we investigated contamination thresholds for SARS-CoV-2 samples generated by Nanopore sequencing by conducting *in silico* analyses. We identified a contamination threshold of 5% that does not compromise the integrity of the genome and a contamination threshold of 20% for lineage calls. Our results suggest the establishment of a stricter threshold if the preservation of assembly accuracy is of utmost importance. Future larger-scale studies are warranted to systematically investigate the effects of contamination on both SARS-CoV-2 reads and other viral and bacterial sequences to serve as a check step for sequencing upload.

Supporting information

S1 Fig. Global nucleotide comparison of artificially generated contaminated samples and their corresponding clinical SARS-CoV-2 reads as the background samples at a medium sequencing depth. A) A heatmap of the pairwise p-distance comparison of the MSD_SV samples—a delta background sequence (AY.25.1) contaminated with a similar delta contaminant sequence (AY.27). B) A heatmap of the pairwise p-distance comparison of the MSD_DV samples–an omicron background sequence (BA.1) contaminated with an alpha contaminant sequence (B.1.1.7).
(TIF)

S2 Fig. Global nucleotide comparison of artificially generated contaminated samples and their corresponding clinical SARS-CoV-2 reads as the background samples at a high sequencing depth. A) A heatmap of the pairwise p-distance comparison of the HSD_SV samples—a delta background sequence (AY.25.1) contaminated with a similar delta contaminant sequence (AY.27). B) A heatmap of the pairwise p-distance comparison of the HSD_DV samples–an omicron background sequence (BA.1) contaminated with an alpha contaminant sequence (B.1.1.7).
(TIF)

S3 Fig. Phylogenetic tree and heatmaps showing single nucleotide variation at different positions of the SARS-CoV-2 genome for (A) a delta variant (AY.25.1) contaminated with another delta variant (AY.27) sequence at contamination levels 1–10%, 20%, 30%, 40%, and 50% for medium sequencing depth and (B) an omicron variant (BA1) contaminated with an alpha contaminant sequence (B.1.1.7) at contamination levels 1–10%, 20%, 30%, 40%, and

50% for medium sequencing depth (25,000 reads).
(TIF)

S4 Fig. Phylogenetic tree and heatmaps showing single nucleotide variation at different positions of the SARS-CoV-2 genome for (A) a delta variant (AY.25.1) contaminated with another delta variant (AY.27) sequence at contamination levels 1–10%, 20%, 30%, 40%, and 50% for high sequencing depth and (B) an omicron variant (BA1) contaminated with an alpha contaminant sequence (B.1.1.7) at contamination levels 1–10%, 20%, 30%, 40%, and 50% for high sequencing depth (50,000 reads).
(TIF)

S5 Fig. Mutational profile comparison of SARS-CoV-2 genome for the clinical genomes to the artificially generated genomes for (A) MSD_SV and (B) (AY.25.1 contaminated with an AY.27 variant) sequence at contamination levels 1–10%, 20%, 30%, 40%, and 50%. (C) MSD_DV and (D) HSD_DV (BA.1 contaminated with a B.1.1.29 variant) at contamination levels 1–10%, 20%, 30%, 40%, and 50%.
(TIF)

S1 Table. A. Quality control metrics comparison for artificially subsampled and contaminated genomes of contamination by similar variants at a medium sequencing depth–for all MSD_SV genomes. **B.** Quality control metrics comparison for artificially subsampled and contaminated genomes of contamination by different variants at a medium sequencing depth–for all MSD_DV genomes.
(DOCX)

S2 Table. A. Quality control metrics comparison for artificially subsampled and contaminated genomes of contamination by similar variants at a high sequencing depth–for all HSD_SV genomes. **B.** Quality control metrics comparison for artificially subsampled and contaminated genomes of contamination by different variants at a high sequencing depth–for all HSD_DV genomes.
(DOCX)

Acknowledgments

We owe a debt of gratitude to Dr. Anna Majer and the DNA core team at the National Microbiology Laboratory for sequencing the clinical samples utilized in this study. We sincerely thank Dr. Andrea Tyler for her contributions to the experimental design of this study and we appreciate the insightful discussions had with Dr. David Alexander and Dr. Kerry Dust of Cadham Provincial Laboratory.

Author Contributions

Conceptualization: Ayooluwa J. Bolaji, Ana T. Duggan.

Data curation: Ayooluwa J. Bolaji.

Formal analysis: Ayooluwa J. Bolaji.

Investigation: Ayooluwa J. Bolaji.

Methodology: Ayooluwa J. Bolaji, Ana T. Duggan.

Project administration: Ayooluwa J. Bolaji.

Resources: Ana T. Duggan.

Supervision: Ayooluwa J. Bolaji, Ana T. Duggan.

Validation: Ayooluwa J. Bolaji.

Visualization: Ayooluwa J. Bolaji.

Writing – original draft: Ayooluwa J. Bolaji.

Writing – review & editing: Ayooluwa J. Bolaji, Ana T. Duggan.

References

1. Park SY, Faraci G, Ward PM, Emerson JF, Lee HY. High-precision and cost-efficient sequencing for real-time COVID-19 surveillance. Scientific Reports |. 2021; 11: 13669. https://doi.org/10.1038/s41598-021-93145-4 PMID: 34211026
2. Geoghegan JL, Douglas J, Ren X, Storey M, Hadfield J, Silander OK, Freed NE, Jelley L, Jefferies S, Sherwood J, Paine S, Huang S, Sporte, A, Baker MG, Murdoch DR, Drummond AJ, Welch D, Simpson CR, French N, Homes EC, de Ligt J. Use of Genomics to Track Coronavirus Disease Outbreaks, New Zealand. Emerg Infect Dis. 2021; 27: 1317. https://doi.org/10.3201/EID2705.204579 PMID: 33900175
3. Magalis BR, Ramirez-Mata A, Zhukova A, Mavian C, Marini S, Lemoine F, Prosperi M, Gascuel O, Salemi M. Differing impacts of global and regional responses on SARS-CoV-2 transmission cluster dynamics. bioRxiv. 2020; 2020.11.06.370999. https://doi.org/10.1101/2020.11.06.370999 PMID: 33173870
4. McLaughlin A, Montoya V, Miller RL, Mordecai GJ, Worobey M, Poon AFY, Joy J. Genomic epidemiology of the first two waves of SARS-CoV-2 in Canada. Elife. 2022; 11. https://doi.org/10.7554/eLife.73896 PMID: 35916373
5. Zhu Y, Liu M, Zhao W, Zhang J, Zhang X, Wang K, Gu c, Wu K, Li Y, Zheng C, Xiao G, Yan H, Zhang J, Guo D, Tien P, Wu J. Isolation of Virus from a SARS Patient and Genome-wide Analysis of Genetic Mutations Related to Pathogenesis and Epidemiology from 47 SARS-CoV Isolates. Virus Genes 2005 30:1. 2005; 30: 93–102. https://doi.org/10.1007/s11262-004-4586-9 PMID: 15744567
6. Yang Y, Peng F, Wang R, Guan K, Jiang T, Xu G, Sun J, Chang C. The deadly coronaviruses: The 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. J Autoimmun. 2020; 109: 102434. https://doi.org/10.1016/j.jaut.2020.102434 PMID: 32143990
7. Zhong NS, Zeng GQ. Our Strategies for Fighting Severe Acute Respiratory Syndrome (SARS). 2003; 168: 7–9. https://doi.org/10.1164/rccm.200305-707OE PMID: 12773318
8. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu W, Wang D, Xu W, Holmes E, Gao G, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. www.thelancet.com. 2020; 395: 565. https://doi.org/10.1016/S0140-6736(20)30251-8 PMID: 32007145
9. Zhou H, Chen X, Hughes AC, Bi Y, Shi W. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Current Biology. 2020; 30: 2196–2203.e3. https://doi.org/10.1016/j.cub.2020.05.023 PMID: 32416074
10. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao Z, Tian J, Pei Y, Yuan M, Zhang Y, Dai F, Liu Y, Wang Q, Zheng J, Xu L, Holmes E, Zhang Y. A new coronavirus associated with human respiratory disease in China. Nature 2020 579:7798. 2020; 579: 265–269. https://doi.org/10.1038/s41586-020-2008-3 PMID: 32015508
11. Zhu Z, Lian X, Su X, Wu W, Marraro GA, Zeng Y. From SARS and MERS to COVID-19: A brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses. Respir Res. 2020; 21: 1–14. https://doi.org/10.1186/S12931-020-01479-W/TABLES/4
12. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genom Bioinform. 2021;3. https://doi.org/10.1093/nargab/lqab019 PMID: 33817639
13. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. PLoS One. 2021; 16: e0257521. https://doi.org/10.1371/journal.pone.0257521 PMID: 34597327
14. Cornet L, Baurain D. Contamination detection in genomic data: more is not enough. Genome Biol. 2022; 23. https://doi.org/10.1186/S13059-022-02619-9 PMID: 35189924
15. Longo MS, O'Neill MJ, O'Neill RJ. Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. PLoS One. 2011; 6: e16410. https://doi.org/10.1371/journal.pone.0016410 PMID: 21358816

16. Breitwieser FP, Pertea M, Zimin A V., Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res. 2019; 29: 954–960. https://doi.org/10.1101/GR.245373.118/-/DC1
17. Lu J, Salzberg SL. Removing contaminants from databases of draft genomes. PLoS Comput Biol. 2018; 14: e1006277. https://doi.org/10.1371/journal.pcbi.1006277 PMID: 29939994
18. Goig GA, Blanco S, Garcia-Basteiro AL, Comas I. Contaminant DNA in bacteriARal sequencing experiments is a major source of false genetic variability. BMC Biol. 2020; 18: 1–15. https://doi.org/10.1186/S12915-020-0748-Z/FIGURES/5
19. Bagheri H, Severin AJ, Rajan H. Detecting and correcting misclassified sequences in the large-scale public databases. Bioinformatics. 2020; 36: 4699–4705. https://doi.org/10.1093/bioinformatics/btaa586 PMID: 32579213
20. Freed NE, Vlková M, Faisal MB, Silander OK. Rapid and inexpensive whole-genome sequencing of SARS-CoV-2 using 1200 bp tiled amplicons and Oxford Nanopore Rapid Barcoding. Biol Methods Protoc. 2020; 5. https://doi.org/10.1093/biomethods/bpaa014 PMID: 33029559
21. Maurya R, Shamim U, Chattopadhyay P, Mehta P, Mishra P, Devi P, Swaminathan A, Saifi S, Khare K, Yadav A, Parveen S, Sharma P, A V, Tyagi A, Jha V, Tarai B, Jha S, Faruq M, Budhiraja S, Pandey R. Human-host transcriptomic analysis reveals unique early innate immune responses in different subphenotypes of COVID-19. Clin Transl Med. 2022; 12. https://doi.org/10.1002/ctm2.856 PMID: 35696527
22. Malik P, Patel K, Pinto C, Jaiswal R, Tirupathi R, Pillai S, Patal U. Post-acute COVID-19 syndrome (PCS) and health-related quality of life (HRQoL)—A systematic review and meta-analysis. J Med Virol. 2022; 94: 253–262. https://doi.org/10.1002/jmv.27309 PMID: 34463956
23. Ramesh S, Govindarajulu M, Parise RS, Neel L, Shankar T, Patel S, Lowery P, Smith F, Dhanasekaran M, Moore T. Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy. Vaccines (Basel). 2021; 9. https://doi.org/10.3390/vaccines9101195 PMID: 34696303
24. David Nelson AR, Hazzouri KM, Lauersen KJ, Lomas MW, Amiri KM, Salehi-Ashtiani K. Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. 2021. https://doi.org/10.1016/j.chom.2020.12.005 PMID: 33434515
25. Lupo V, Van Vlierberghe M, Vanderschuren H, Kerff F, Baurain D, Cornet L. Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. Front Microbiol. 2021; 12. https://doi.org/10.3389/fmicb.2021.755101 PMID: 34745061
26. Boni MF, Lemey P, Jiang X, Lam TTY, Perry BW, Castoe TA, Rambaut A, Robertson D. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature Microbiology 2020 5:11. 2020; 5: 1408–1417. https://doi.org/10.1038/s41564-020-0771-4 PMID: 32724171
27. Singh D, Yi S V. On the origin and evolution of SARS-CoV-2. Experimental & Molecular Medicine 2021 53:4. 2021; 53: 537–547. https://doi.org/10.1038/s12276-021-00604-z PMID: 33864026
28. Rahimi F, Talebi Bezmin Abadi A. Emergence of the Omicron SARS-CoV-2 subvariants during the COVID-19 pandemic. Int J Surg. 2022; 108: 106994. https://doi.org/10.1016/j.ijsu.2022.106994 PMID: 36375791
29. Markov P V, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, Katzourakis A. The evolution of SARS-CoV-2. Nature Reviews Microbiology |. 2023; 21: 361–379. https://doi.org/10.1038/s41579-023-00878-2 PMID: 37020110