



Synthetic DNA spike-ins (SDSIs) enable sample tracking and detection of inter-sample contamination in SARS-CoV-2 sequencing workflows

Kim A. Lagerborg^{1,2,17}, Erica Normandin^{1,3,17}, Matthew R. Bauer^{1,2,17}, Gordon Adams¹, Katherine Figueroa¹, Christine Loreth¹, Adrienne Gladden-Young¹, Bennett M. Shaw^{1,4}, Leah R. Pearlman¹, Daniel Berenzy⁵, Hannah B. Dewey⁵, Susan Kales⁵, Sabrina T. Dobbins¹, Erica S. Shenoy⁴, David Hooper⁴, Virginia M. Pierce^{6,7,8}, Kimon C. Zachary^{4,9,10}, Daniel J. Park¹, Bronwyn L. MacInnis^{1,11,12}, Ryan Tewhey^{5,13,14}, Jacob E. Lemieux^{1,4}, Pardis C. Sabeti^{1,3,11,12,15,18}, Steven K. Reilly^{1,16,18} ✉ and Katherine J. Siddle^{1,3,18}

The global spread and continued evolution of SARS-CoV-2 has driven an unprecedented surge in viral genomic surveillance. Amplicon-based sequencing methods provide a sensitive, low-cost and rapid approach but suffer a high potential for contamination, which can undermine laboratory processes and results. This challenge will increase with the expanding global production of sequences across a variety of laboratories for epidemiological and clinical interpretation, as well as for genomic surveillance of emerging diseases in future outbreaks. We present SDSI + AmpSeq, an approach that uses 96 synthetic DNA spike-ins (SDSIs) to track samples and detect inter-sample contamination throughout the sequencing workflow. We apply SDSIs to the ARTIC Consortium's amplicon design, demonstrate their utility and efficiency in a real-time investigation of a suspected hospital cluster of SARS-CoV-2 cases and validate them across 6,676 diagnostic samples at multiple laboratories. We establish that SDSI + AmpSeq provides increased confidence in genomic data by detecting and correcting for relatively common, yet previously unobserved modes of error, including spillover and sample swaps, without impacting genome recovery.

The coronavirus disease 2019 (COVID-19) pandemic has again demonstrated the crucial role of genomic sequencing in combatting infectious disease outbreaks globally. Monitoring the emergence of pathogens and the spread of variants of concern has become commonplace in government, academic and private laboratories^{1,2}. Genomics data provide insights into the diversity, evolution and transmission of a virus, and serve as a critical guide for public health interventions ranging from contact tracing, identifying cases of reinfection or understanding the basis of resistance to clinical interventions^{3–6}. Laboratories are now performing viral genomic sequencing of SARS-CoV-2 at an unprecedented scale^{7,8}, highlighting the pressing need to ensure the integrity of genomes.

Multiplexed amplicon-based genome sequencing methods have accelerated the massive scale of SARS-CoV-2 genomic surveillance due to their improved sensitivity, cost and speed over other, lower-amplification RNA sequencing approaches, such as unbiased metagenomic sequencing⁹. However, the risk for contamination

during the amplification stage is especially high as the 35+ cycles of virus-specific PCR produces trillions of SARS-CoV-2 amplicons in a single reaction. Other high-risk modes of contamination, including sample swaps, cross-contamination of samples, or aerosolization, can occur throughout the sample processing pipeline. With many laboratories performing viral sequencing by processing multiple large batches in parallel, the potential for contamination increases¹⁰. Even small amounts of sample mixing or contamination of amplicons could potentially confound studies where detection is sensitive to tens of molecules^{10,11}. Moreover, as SARS-CoV-2 has relatively low genetic diversity and often spreads in local outbreaks or clusters^{11,12}, genomes are often identical at the consensus level^{11,15–17}, a pattern that could also occur with contamination. The risk of contamination and the challenges in detecting it can confound a wide array of genomic analyses, including estimates of the frequencies of variants, lineage dynamics and transmission events. Additionally, methods to address the critical risk of sample

¹Broad Institute of Harvard and MIT, Cambridge, MA, USA. ²Harvard Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. ³Department of Systems Biology, Harvard Medical School, Boston, MA, USA. ⁴Division of Infectious Diseases, Massachusetts General Hospital, Boston, MA, USA. ⁵The Jackson Laboratory, Bar Harbor, ME, USA. ⁶Department of Pathology, Massachusetts General Hospital, Boston, MA, USA.

⁷Pediatric Infectious Disease Unit, MassGeneral Hospital for Children, Boston, MA, USA. ⁸Department of Pathology, Harvard Medical School, Boston, MA, USA. ⁹Department of Medicine, Harvard Medical School, Boston, MA, USA. ¹⁰Infection Control Unit, Massachusetts General Hospital, Boston, MA, USA.

¹¹Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Harvard University, Boston, MA, USA. ¹²Massachusetts Consortium on Pathogen Readiness, Boston, MA, USA. ¹³Graduate School of Biomedical Sciences and Engineering, University of Maine, Orono, ME, USA.

¹⁴Graduate School of Biomedical Sciences, Tufts University School of Medicine, Boston, MA, USA. ¹⁵Howard Hughes Medical Institute, Chevy Chase, MD, USA. ¹⁶Department of Genetics, Yale School of Medicine, New Haven, CT, USA. ¹⁷These authors contributed equally: Kim A. Lagerborg, Erica Normandin, Matthew R. Bauer. ¹⁸These authors jointly supervised this work: Pardis C. Sabeti, Steven K. Reilly, Katherine J. Siddle. ✉e-mail: steven.k.reilly@yale.edu

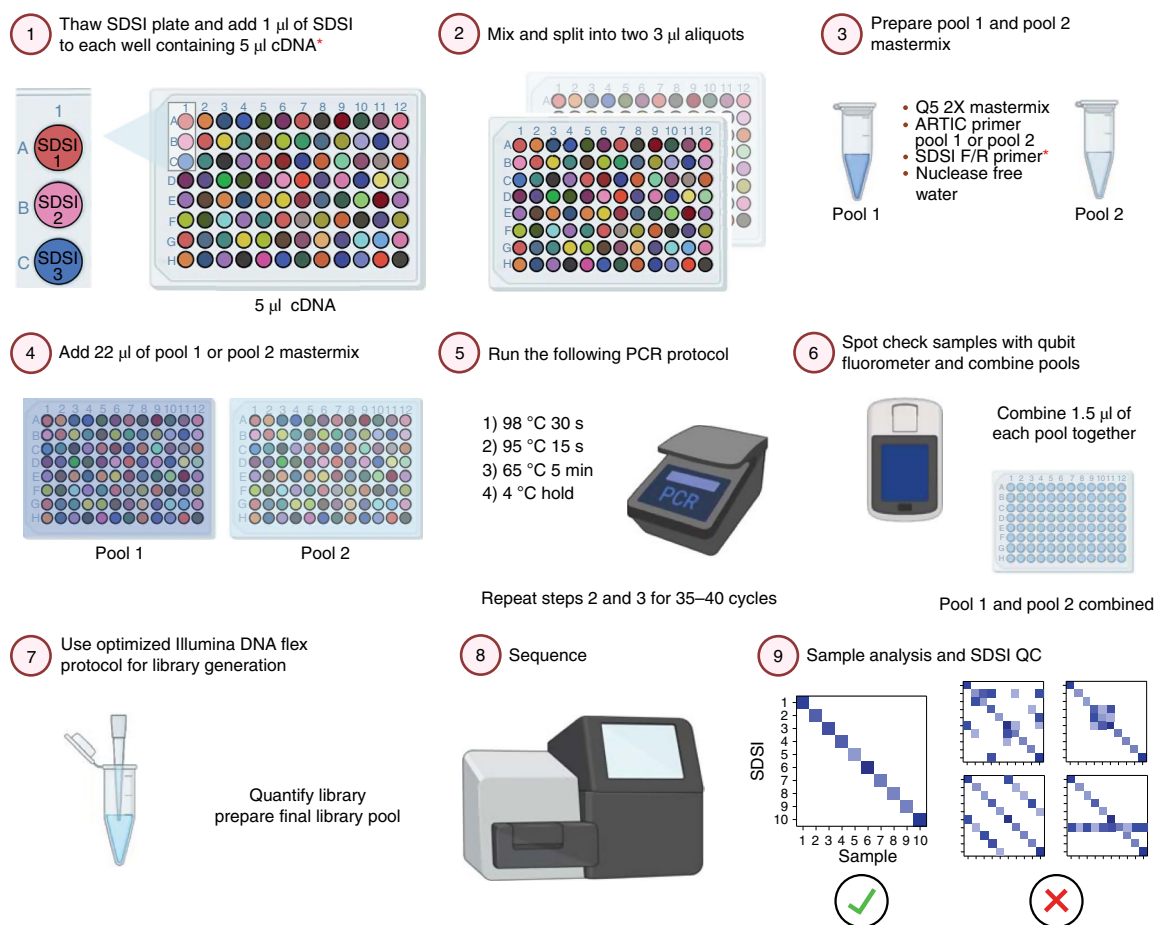


Fig. 1 | SDSI + AmpSeq protocol. Illustrative workflow for 96 samples through the SDSI + AmpSeq amplicon-sequencing pipeline. A unique, synthetic DNA spike-in (SDSI) will be added to each cDNA sample to allow for contamination tracking and accurate sample identification in analysis. Red asterisks indicate additional steps to the standard ARTIC pipeline.

processing errors in clinical sequencing could enable its use more widely in clinical decision-making.

While inclusion of internal standards is commonplace in many experimental approaches^{13–15} and some technical assay controls exist for DNA sequencing^{16–18}, the use of internal controls is currently rare in amplicon-based genomic surveillance. Here we developed and extensively tested a sample identification method using 96 synthetic DNA spike-ins (SDSIs) for amplicon-based sequencing approaches. Using the widely used open-access ARTIC tiled primer design (<https://artic.network/>), we implemented these SDSIs for SARS-CoV-2 genomic sequencing in thousands of residual diagnostic (clinical) samples. The resulting user-friendly and highly versatile SDSI + AmpSeq protocol can be easily implemented to improve the quality of genomic data generated for epidemiological and clinical investigations of human pathogens (Fig. 1 and Supplementary Data Table 1).

Results

Design and characterization of SDSIs for amplicon sequencing.

We designed a simple, flexible system for sample tracking and contamination tracing using a uniquely identifiable core DNA sequence flanked by constant priming regions satisfying several design requirements. This design allows in-sample tracking through the addition of a different SDSI to each sample during sample processing. Following sequencing, the data can be analysed for both the presence of the expected or any other SDSI, illuminating sample misassignment or contamination with high resolution and accuracy (Fig. 1). Our

design focused on highly stable DNA oligos that could be added to sample complementary DNA (cDNA) to capture contamination at or after the critical viral amplification step, including contamination generated during amplification and in handling amplified material. In contrast to a short barcode system, SDSIs are compatible with both tagmentation and ligation-based sequencing approaches¹⁸. The constant priming regions mean that only a single primer pair needs to be added into the existing multiplexed PCR step to co-amplify all SDSIs with the primary reaction target(s) (Fig. 2a).

The core portion of each SDSI constitutes 96 distinct DNA sequences from the genomes of diverse uncommon Archaea, precluding false detection and cross-identification (Supplementary Data File 1 and Methods). By using extremophilic Archaea, our designs maximized evolutionary distance from common human pathogens. To avoid false positive results, the core SDSI sequences should be sufficiently distinct from one another, as well as from sequences commonly found in laboratories and clinical samples. A permissive BLASTn search performed against the entire National Center for Biotechnology Information (NCBI) database confirmed that the SDSI core sequences had limited homology outside the domain Archaea, specifically to genera unlikely to be found in laboratories (Extended Data Fig. 1a). While this limited homology outside of the domain Archaea maximized the potential for broad applications, we also confirmed that none of the core sequences shared homology (defined as >90% identity over 50 base pairs (bps)) with *Homo sapiens* or known viral genomes (Methods). Similarly, we confirmed that all SDSIs were substantially different from each other to prevent

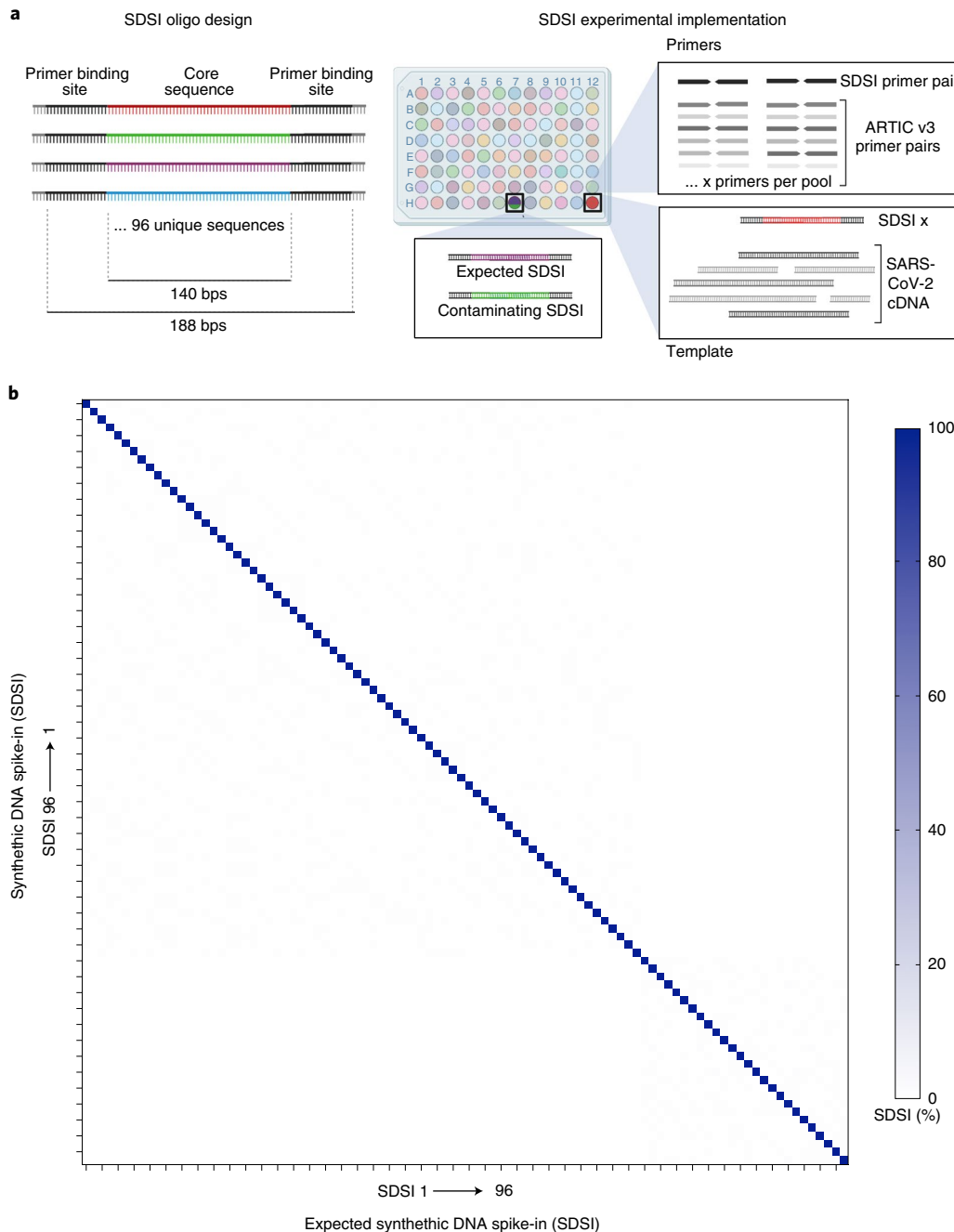


Fig. 2 | Synthetic DNA oligos spiked into amp-seq reactions designed to flag contamination and sample swaps. **a**, Schematic of SDSI design. Each oligo contains 140 bp of unique sequence flanked by common primer binding sites. Primers designed to amplify all SDSIs are added to ARTIC primer pools, and a unique SDSI is added to each clinical sample. Identification of multiple SDSIs in the same sample indicates contamination. **b**, Percent of SDSI reads mapping for each of the 96 SDSIs (horizontal axis) were quantified for each of the 96 SDSIs (vertical axis). Any off-diagonal signal would indicate non-specific identification of SDSIs.

misidentification (Extended Data Fig. 1b). Sequencing of the SDSIs confirmed that each of the 96 constructs resulted in a robust and specific signal of mapped reads (Fig. 2b).

We selected a pair of primers and priming regions that are highly specific and show reliable amplification across SDSIs and under standard PCR conditions. Using Primer-BLAST, we predicted that these sequences had limited homology to common organisms and thus were unlikely to amplify off-target sequences. Experimentally we confirmed that the SDSI primers did not produce any non-specific amplification, including in the presence of cDNA from a nasopharyngeal

(NP) swab sample (Extended Data Fig. 2a). The primer pair also had a common length (24 bps), guanine-cytosine (GC) content (45.8%) and melting temperature (62 °C and 63 °C, respectively, in our SDSI + AmpSeq protocol), ensuring their compatibility with many multiplexed PCR reactions, including the most widely used SARS-CoV-2 amplicon sequencing strategy (<https://artic.network/>) (Extended Data Fig. 2b). Each SDSI was identically sized, shared a priming region and had similar GC content (range: 33–65%) to promote similar amplification rates across different SDSIs and to viral amplicons (for example, the SARS-CoV-2 genome is 37 ± 5%

GC)¹⁹ (Extended Data Fig. 2c). All SDSIs amplified in an ARTIC SARS-CoV-2 PCR reaction with SDSI primers included, in each case yielding a single clean product of the expected size (Extended Data Fig. 2d). Furthermore, we observed that GC content did not significantly bias the number of SDSI reads detected in clinical samples (Extended Data Fig. 2e).

Validation of an SDSI + AmpSeq SARS-CoV-2 sequencing approach. The addition of SDSIs into the ARTIC PCR did not detrimentally alter SARS-CoV-2 cDNA amplification from clinical samples. To prevent SDSIs from overtaking the amplification and sequencing of SARS-CoV-2 amplicons, we titrated the amount of SDSI added to the reaction. We found that the highest concentration tested, 600 copies per μl , resulted in reliable SDSI detection with >96% of reads mapping to SARS-CoV-2 and no apparent alteration in coverage across the genome (Extended Data Fig. 3a,b). We assessed the specificity of the 96 SDSIs in clinical samples to confirm that there was no unpredicted cross-mapping, misidentification or large differences in amplification rate (Fig. 3a). We processed 14 samples spanning a range of viral cycle threshold (CT) values (CT range: 25–33), with both the standard ARTIC and SDSI + AmpSeq methods. For each amplicon, across all samples, there was no significant difference in coverage between the ARTIC and SDSI + AmpSeq conditions (Fig. 3b). Even in samples with low viral loads (CT > 30), we found that there were no significant differences in amplicon coverage (Extended Data Fig. 4a). Additionally, within the 14 samples processed \pm SDSI, we see a 100% genome concordance rate, illustrating that the addition of the SDSIs does not impact the accuracy of assembled genomes.

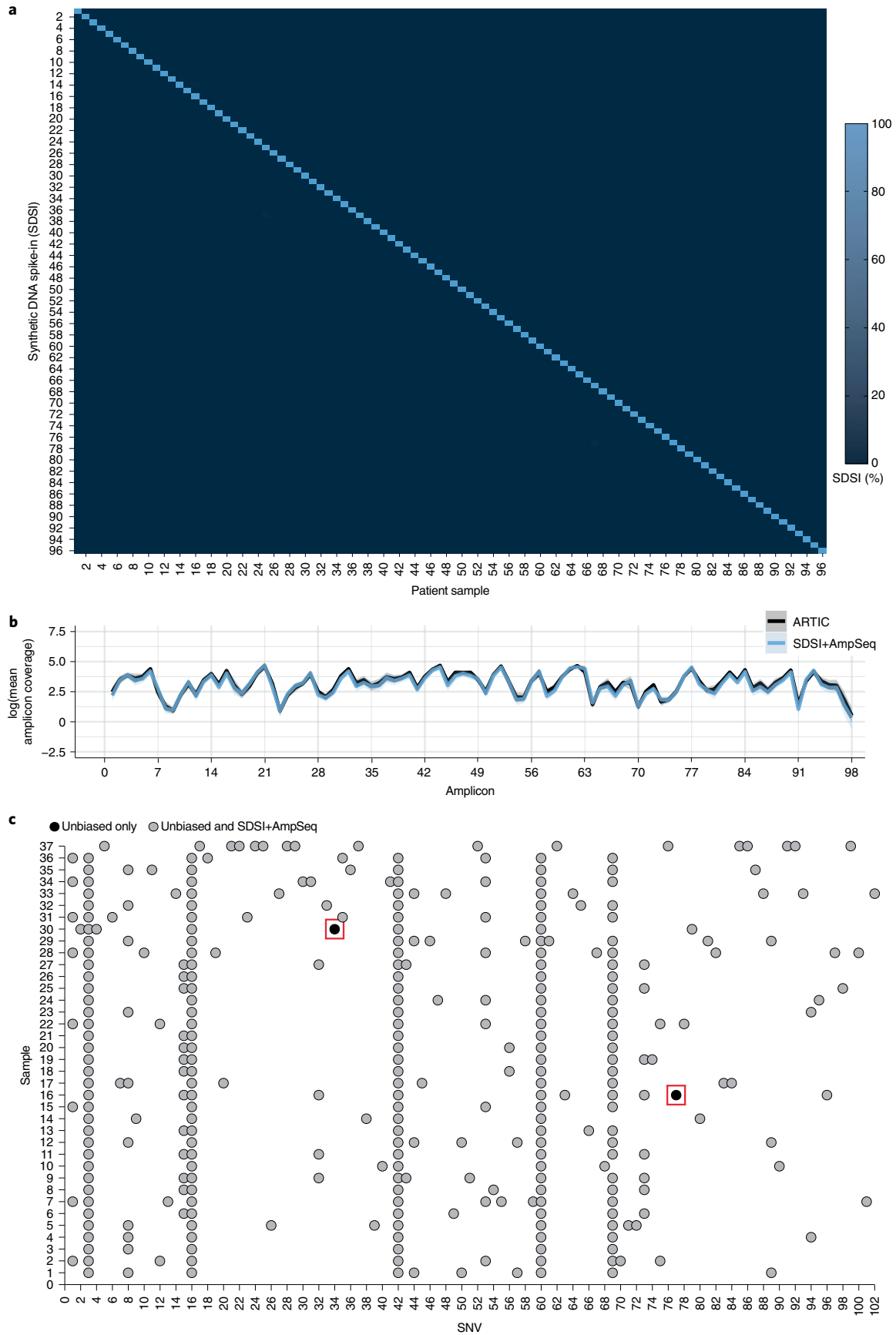
As extensive PCR can result in the propagation of numerous types of errors, such as DNA polymerase base substitution errors, PCR recombination events and thermocycling induced DNA damage, we further compared SARS-CoV-2 genome concordance between SDSI + AmpSeq and unbiased, metagenomic sequencing on a batch of 89 unique patient samples^{9,10,20,21}. The samples consisted of diverse viral lineages and loads (CT range: 11.9–37.4), with the more sensitive amplicon sequencing method generating more complete genomes at higher CTs (Extended Data Fig. 5a–d). We assessed the coverage uniformity, as increasing uniformity reduces the sequencing depth required to generate reliable genomes²². We found that unbiased sequencing had more uniform coverage up to a CT of 25 ($n = 31$, Gini coefficient = 0.240 ± 0.046 (unbiased) vs 0.428 ± 0.026 (SDSI + AmpSeq)), while SDSI + AmpSeq generated more uniform coverage for samples above a CT of 25 ($n = 39$, Gini coefficient = 0.766 ± 0.265 (unbiased) vs 0.554 ± 0.124 (SDSI + AmpSeq)) (Extended Data Fig. 5e). For the 37 samples that assembled a full genome in both methods, only two out of 332 total single nucleotide variants (SNVs) were divergently identified by SDSI + AmpSeq (Fig. 3c). Each SNV was observed in only one sample and both fell within an ARTIC primer region, despite primer trimming during analysis; for one SNV, this suggests that PCR error from the ARTIC protocol contributed to the discrepancy²³, whereas manual inspection of the other (C9565T) indicated the presence of intra-host variation in both methods, with a variant allele frequency of 39.4% (SDSI + AmpSeq) and 59.2% (unbiased sequencing). Overall, the discordance SNV calling rate was 0.6%, consistent with observed circulating diversity and sequencing-based errors. Consistent with

previous reports from other groups, ARTIC amplicon sequencing maintains a high level of concordance at the consensus genome level¹⁰, even with the addition of SDSIs.

We explored other technical modifications to the ARTIC amplicon sequencing protocol to improve genome recovery, limit contamination points and enhance reproducibility of our SDSI approach. Foremost, increasing cDNA length by using more processive reverse transcriptases improved amplicon coverage (Extended Data Fig. 6a,b). Amplification of ARTIC amplicons and SDSIs by Q5 Hot-Start High-Fidelity 2x Master Mix resulted in higher amplification (Extended Data Fig. 6c and Supplementary Data Table 2). We found that increasing (2 \times) primer concentrations (20.8 nM final concentration) for poorly performing amplicons increased coverage in these amplicons, even enabling whole genome recovery for multiple samples, supporting the observation that primer rebalancing can enable greater coverage^{24,25} (Extended Data Figs. 6d and 7 and Supplementary Data Table 3). We explored the effects of different numbers of PCR cycles, DNA-hybridization steps, and temperature ramp speeds. Both DNA-hybridization steps and temperature ramping provided no substantial changes in amplification (Extended Data Fig. 6e,f). Additional PCR cycles can increase genome coverage uniformity of low-viral-load samples, although it may lead to increased erroneous SNV calls²³ (Extended Data Fig. 6g). Using a standardized cDNA input, we found that the DNA Flex library workflow resulted in increased genome coverage depth across all CT values tested, compared to Nextera XT (Extended Data Fig. 6h). To further mitigate the risk of contamination from highly amplified libraries, we took advantage of the self-normalizing feature of the DNA flex kit and found that limiting the tagmentation beads by scaling down (0.5 \times) all components of the DNA Flex library construction reagents restricted library over-amplification. Notably, this did not impact final library size distributions or SDSI amplification, while generating final sequencing libraries at half their original concentrations and halving the library construction cost per sample (Methods and Supplementary Data Table 4). We summarized the results of our optimizations (<https://benchling.com/sprt-R95g0tCxKOeCAqn8IAk3>) and show that SDSIs can be easily integrated with numerous protocol alterations.

Implementation of SDSIs to clinical samples at scale. The SDSI + AmpSeq method is compatible with a range of viral CTs, SARS-CoV-2 lineages, origin of the patient sample, and laboratory in which the pipeline is implemented, demonstrating that this is a robust and flexible approach that can be readily implemented for surveillance. A half plate of SDSIs were used at two large-scale sequencing facilities, the Broad Institute and Jackson Laboratories (JAX), for SDSI + AmpSeq SARS-CoV-2 surveillance across a total of 6,676 clinical samples and controls (JAX: $n = 3,773$; Broad: $n = 2,903$). Individual batches typically consisted of 92 clinical samples with 4 designated water controls. Clinical samples were largely from Maine, Massachusetts, and Rhode Island from December 2020 to July 2021 and covered a wide range of viral CT values (8.4–39.9) and pango lineages (77 total lineages) (Fig. 4a). The SDSI + AmpSeq method worked robustly despite minor implementation differences in protocols between the two laboratories, including different cDNA synthesis enzymes, CT normalization, and library construction approaches (0.5X Illumina DNA Flex vs Illumina COVID-Seq) (Methods).

Fig. 3 | SDSI + AmpSeq amplicon coverage and genome concordance. **a**, Percent of SDSI for SDSI 1–96 in patient samples. **b**, log of the mean amplicon coverage for the same clinical samples run with and without an SDSI ($n = 14$). A unique SDSI was used in each sample. The solid blue line represents SDSI + AmpSeq and the solid black line is ARTIC only with no SDSI. Blue and black shading around the solid lines represents the 95% confidence interval. There were no statistical differences ($P > 0.05$) in the mean amplicon coverage for each amplicon between the groups (two-tailed Mann Whitney t -test and multiple comparison two-stage step-up Benjamini, Krieger and Yekutieli test with FDR set to 5%). **c**, SNV concordance plot between SDSI + AmpSeq and unbiased consensus sequences. Two discordant SNVs, outlined in a red box, were found. Grey dots represent SNVs found in both the unbiased and SDSI + AmpSeq method, whereas black dots indicate that the SNV was only present in unbiased sequencing.



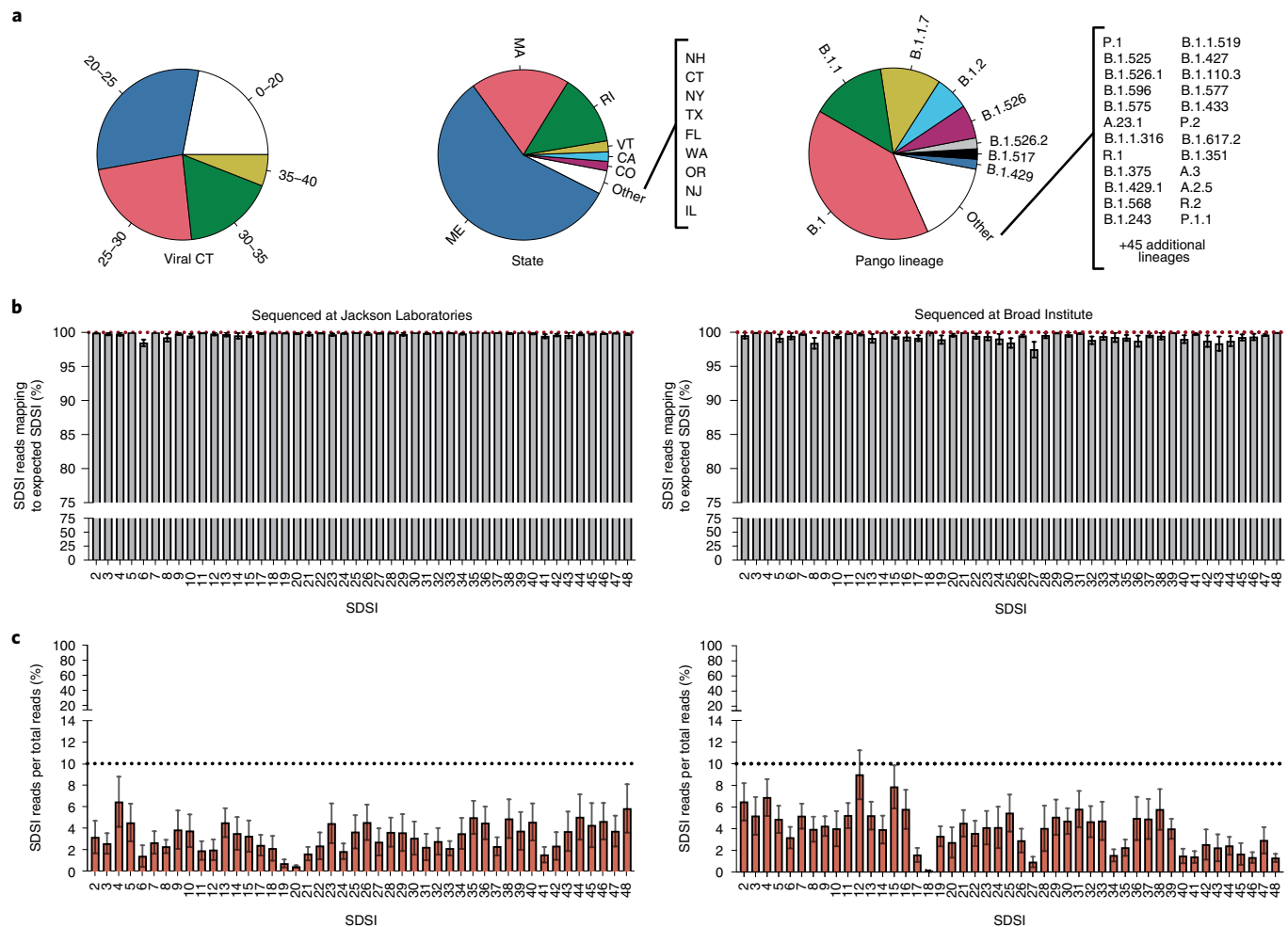


Fig. 4 | SDSI + AmpSeq performs well across thousands of samples. a, Sample diversity from two different institutions representing a range of CTs, viral lineages, and states of sample collection from samples where the data was available. **b**, Percentage of SDSI reads out of the sum of all SDSI reads that map to the correct spike-in. Left: JAX, $n=3,773$ biologically independent samples. Right: Broad, $n=2,903$ biologically independent samples. Data are presented as mean \pm s.e.m. Individual data points are displayed when $n \leq 10$. **c**, Percentage of SDSI reads over the total of all sequenced reads for all SARS-CoV-2-positive samples. Left: JAX, $n=3,045$ biologically independent samples. Right: Broad, $n=2,670$ biologically independent samples. Data are presented as mean \pm s.e.m. Individual data points are displayed when $n \leq 10$. The dashed line represents 10% SDSI reads.

SDSI + AmpSeq is a tractable and easily-implemented method for genome quality control when applied to high-throughput processing of clinical samples. Across thousands of clinical samples, the SDSIs performed consistently and reliably (Fig. 4b,c). The mean percentage of SDSI reads that mapped to the expected SDSI was above 95% for all SDSIs in both laboratories (Fig. 4b). This demonstrated that across a large set of highly variable clinical samples, there were no systemic issues of misidentification for specific SDSIs. Additionally, across 5,715 samples from both institutions, the percentage of all SDSI reads in SARS-CoV-2-positive samples averaged 3.72% (90% of samples fell between 0.0003–9.9887%) (Fig. 4c). Importantly, this low, but consistent percentage of reads mapping to SDSIs allows for their implementation without necessitating increased sequencing depth. Notably, the SDSIs performed well with and without previous normalization of cDNA based on CT; however, normalizing increased the percentage of SDSI reads (Extended Data Fig. 4, Fig. 4b(left) and Methods). Normalization of viral CT may provide an additional level of quality control that is especially important for labs with limited sequencing capacities.

Confident genome sequencing and analysis with SDSI + AmpSeq. SDSIs enable detection of sample swaps and contamination events

that occur in large-scale batch processing, which may otherwise go undetected. To demonstrate that SDSI + AmpSeq can accurately detect contamination, we mixed two SDSIs at various ratios before the ARTIC PCR and found that SDSI ratios were reflected in the sequencing output (Fig. 5a). Across 6,676 SARS-CoV-2 samples processed in large batches, SDSIs allowed for identification of multiple key modes of error (Fig. 5b). As plotted, a plate without contaminating events or sample swaps should display a simple diagonal pattern with 1:1 matching of expected and observed SDSIs. In some cases, off-diagonal events occur in clear patterns, enabling speculation on the nature of the contamination and demonstrating the utility of SDSIs as an internal control and in-sample label. We observed cases where a plate was accidentally inverted when the post-PCR reaction pool 1 was mixed with pool 2 (Fig. 5b). SDSI + AmpSeq detects entire flawed batches that may not have been flagged with standard controls (as in the case with the plate inversion where water controls in plate corners would not have been affected). In another example, SDSIs were detected unexpectedly throughout a batch, indicating that an SDSI (and possibly SARS-CoV-2) contaminated a common reagent.

SDSI + AmpSeq also enables fine-resolution insight into sample processing errors with high specificity. In one example, SDSI counts

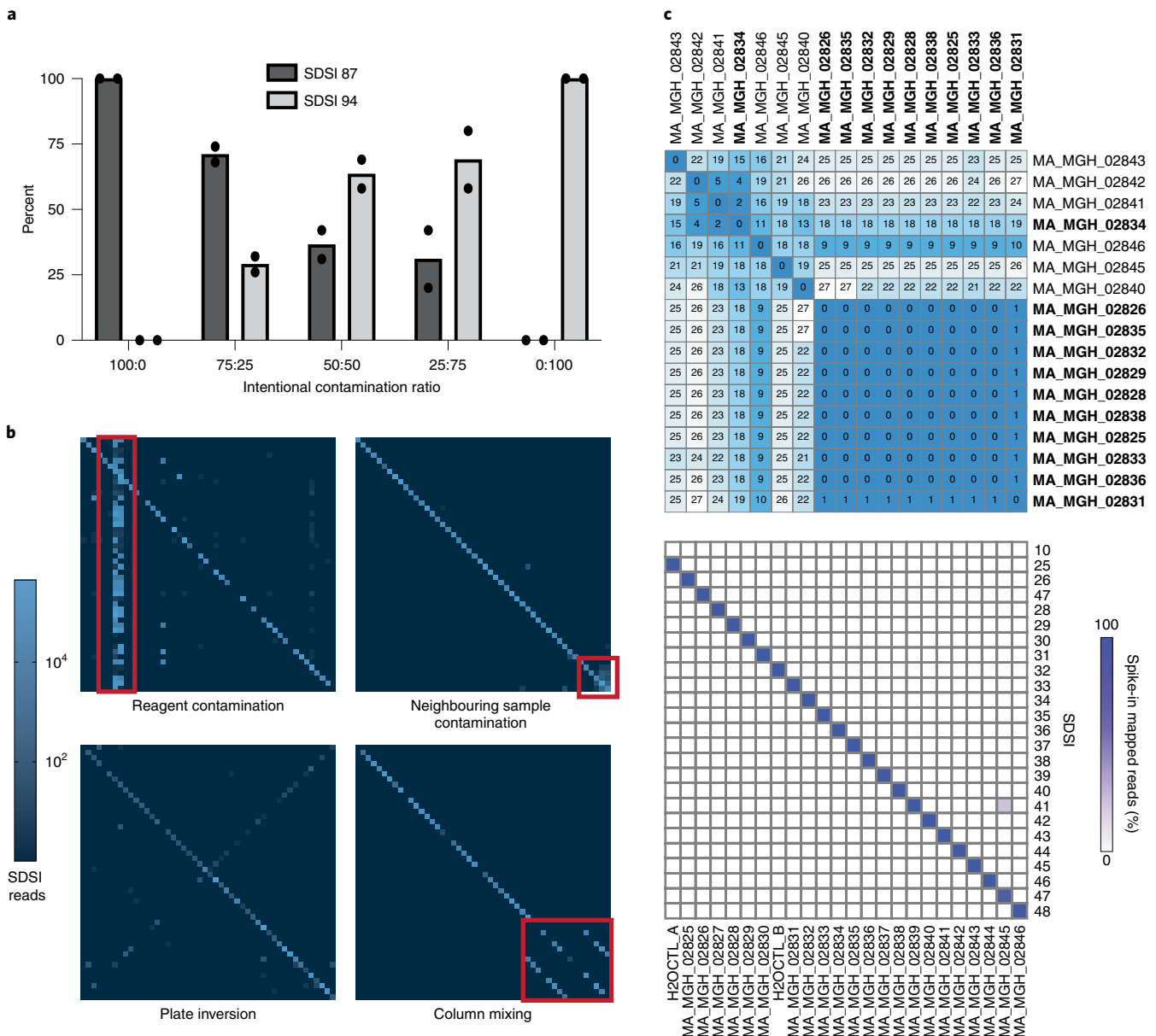


Fig. 5 | SDSI + AmpSeq is used to identify sample swaps and contamination. a, Intentional SDSI contamination experiment (run in duplicate) assessing whether different ratios of contamination between SDSI 87 and SDSI 94 (SDSI 87:SDSI 94) were detectable with the SDSI + AmpSeq method. **b**, Examples of experimental errors that were caught using the SDSI + AmpSeq method. **c**, Top: distance matrix showing pairwise differences between the 17 complete genomes assembled from this sample set. Putative cluster samples are in bold. Bottom: spike-in counts for each of the 24 samples and water controls in this sequencing batch.

indicated that columns were unintentionally mixed together (Fig. 5b). Here, in-sample labelling allowed researchers to confidently move forward with analyses on unaffected samples. In other cases, samples were associated with both the expected SDSI and SDSIs from neighbouring samples, indicating potential spillover or pipetting errors. Again, genomes generated from samples with suspicious SDSI profiles can be investigated further, and removed from analyses and/or reprocessed. We recommend manual curation of genomes assembled from any samples with <95% of SDSI reads mapping to the expected SDSI. This level of impurity is probably attributable to sample processing contamination, given minimal baseline crosstalk from sources such as indexing primer or oligo synthesis observed (Methods and Extended Data Fig. 8). Moreover, these patterns of contamination events identified via use of SDSI + AmpSeq

illuminated key sources of error in processing pipelines and provided an opportunity to improve processing fidelity in subsequent batches.

To demonstrate the ability of SDSIs to enable confident interpretation of sequencing data in an outbreak investigation, we used SDSI + AmpSeq to investigate a putative SARS-CoV-2 cluster from Massachusetts General Hospital (MGH). The Infection Control Unit suspected nosocomial transmission, yet sample swaps and contamination could easily undermine findings. We sequenced 24 samples with SDSI + AmpSeq (including 14 suspected cluster-related samples, 8 unlinked contextual controls, and 2 water controls). Out of 11 assembled genomes that were suspected to be part of the cluster, 10 were genetically highly similar (0–1 consensus nucleotide difference) (Fig. 5c) and distinct from other samples from Massachusetts around the same time (Extended Data Fig. 9),

strongly suggesting that this cluster arose from nosocomial transmission. Analysis of the SDSIs confirmed that genome sequence similarity among cluster-associated samples was not the result of cross-contamination (Fig. 5c). Indeed, 23/24 libraries contained >95% SDSI-mapped reads corresponding to the expected SDSI. One sample that was not part of the cluster (MA_MGH_02845) showed 18% of reads from a second SDSI, which was added to a different sample in the batch (MA_MGH_02839). We resequenced both samples implicated in the contamination event. We confirmed that the two genome sequences for MA_MGH_02845 were 100% concordant, and no genome was assembled for MA_MGH_02839 in either attempt, probably due to its very low viral load (CT=37). This example illustrates how SDSIs can be used to isolate and validate only those samples implicated in contamination events and altogether increase confidence in cluster investigations.

To further increase our ability to comprehensively identify sources of error in AmpSeq methods for viral genomics, we sought to capture contamination and sample swaps that might occur before cDNA synthesis. We explored the feasibility of modifying the SDSI approach to use synthetic RNA spike-ins (SRSIs) from the same constructs, which could be added to clinical sample RNA. For a subset of SDSIs, we included a T7 promoter site to produce RNAs in vitro. For two clinical samples representing low (20) and mid (26) CTs, we detected reads from the SRSIs added directly to extracted viral RNA as a proof-of-principle (Extended Data Fig. 10). Notably, this approach did not require any additional protocol modifications, and we therefore expect it to be a highly versatile and user-friendly method when deployed at scale for complete end-to-end sample tracking.

Discussion

Amplicon-based sequencing methods enable rapid, full genome recovery for emerging SARS-CoV-2 variant surveillance; however, robust tools are needed to ensure accuracy in genomic data. SDSI + AmpSeq is an invaluable technique for detecting key modes of contamination, addressing this critical gap in standard controls and practices. SDSIs do not compromise genome quality, have been successfully deployed in thousands of clinical samples, and are in use across multiple laboratories with differing protocols. These SDSIs revealed numerous instances of sample swaps and contamination, many of which would go unnoticed with standard batch-level controls. SDSIs further provide critical confidence in the interpretation of clusters of identical genomes, a renewed challenge in the surveillance of more transmissible variants. The common primer design enables the applicability of SDSIs to multiple short amplicon designs and sequencing strategies, adding only minor changes to existing protocols and minimal additional cost.

SDSIs overcome multiple modes of error in the production of amplicon-based genomic sequencing data, and are a critical component of quality control measures. The approach is most effective when adopted fully within a laboratory setting and thus we propose routine use of the SDSI + AmpSeq method to flag laboratory-wide contamination. We provide an extensively tested protocol with ARTIC v3 and Illumina-based tagmentation. It can also be applied to other sequencing pipelines, although this potentially requires further optimization. Our pathogen-exclusion design criteria allow our 96 validated SDSIs to be immediately incorporated into other tiled amplicon panels, such as existing ones for Zika, Ebola and other viruses of epidemic potential^{26,27}.

The SDSI-labelling paradigm is broadly applicable to many amplicon-based needs: it is amenable to various technical enhancements, flexible to remaining error modes and expandable to additional targets. One could apply the same design parameters to expand the set of SDSIs, such as to 384-well formats. To design larger panels of SDSIs, one could use artificial core sequences, rather than excerpting from Archaea. Primer sites could also be

easily adapted for integration with new advancements in amplicon sequencing, such as tailed primer approaches or new primer schemes^{28–32}. In its current implementation, our SDSIs detect contamination or workflow errors that occur during and after amplification, but not issues arising at the RNA or cDNA generation stage. While we expect DNA spike-ins would be more robust to variation in samples and protocols and are applicable to both RNA and DNA viruses, further refinement of the RNA spike-in approach could address other modes of contamination, enabling end-to-end sample tracking for RNA viruses at scale. Specifically, SRSIs could improve the tracking of high-viral-load samples that are more likely to have well-to-well contamination than low-viral-load samples. Future work improving quantification and SDSI analysis pipelines may enable them to serve as within-sample controls, since samples or batches with outlier SDSI read counts may reveal experimental error.

The integration of SDSIs can mitigate a critical vulnerability of amplicon-based sequencing while preserving its many advantages, increasing the robustness of its use across laboratory and clinical settings. Adoption of controls across the viral surveillance community would increase accuracy and integrity of genomic data worldwide. Looking forward, SDSIs could serve as a crucial component in improving data integrity in amplicon-based genomic sequencing beyond infectious disease surveillance, such as for food safety, species identification and environmental sampling.

Methods

We have provided our SDSI implementation protocol on Benchling (<https://benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3>) in addition to the methods detailed below.

Sample collection and study design. Research was conducted at the Broad Institute with an exempt determination from the Broad Office of Research Subjects Protections and with approval from the MIT Institutional Review Board under protocol #1612793224. Samples were obtained from MGH, Massachusetts Department of Public Health, the Rhode Island Department of Public Health and the Broad Institute Clinical Research Sequencing Platform. Samples from MGH fall under Partners Institutional Review Board under protocol #2019P003305. Samples were secondary-use or residual clinical and diagnostic specimens (referred to collectively throughout as clinical samples), obtained by researchers under a waiver of consent. All samples were nasopharyngeal or anterior nares swabs in a stabilizing medium (for example, molecular transport media (MTM), viral transport media (VTM) or universal transport media (UTM)). These unique biological materials are not available to other researchers as they are human patient samples from clinical excess material and are thus of limited volume. Samples sequenced at JAX were approved under protocol 2020-NHSR-019-BH.

SDSI design and in silico validation. We designed synthetic DNA fragments that each contained a 140 bp unique sequence and constant priming regions. Core SDSI sequence homology to sequences from various organisms was predicted by a permissive BLAST search (blastn; 5,000 maximum targets; E = 10; word size = 11; no mask for low complexity). We considered homologies identified with this BLASTn search described above that were additionally >50 bps (>35% query cover) and >90% sequence identity to be above our homology threshold. For all 96 selected SDSIs, there were no such homologies when results were filtered to all *Homo sapiens* (taxid:9606) or viral (taxid:10239) sequences in the NCBI database. For homologies above the threshold to bacterial or eukaryotic sequences in the NCBI database (excluding Archaea: taxid:2157), we report both the SDSI and the genus it mapped to in each case (Extended Data Fig. 1a). Using the same BLASTn (BLASTN 2.12.0+) parameters, we also mapped SDSIs against a custom database including SDSI core sequences, and found no homologies above our threshold between SDSIs. As there were no homologies below our threshold between SDSIs and human, virus or other SDSI sequences, we noted the maximum alignment scores for any homology below our threshold identified, and reported these scores (Extended Data Fig. 1b).

We confirmed that SDSI primers and amplicons were predicted to amplify specifically and consistently with ARTIC v3 amplicons. We used Primer-BLAST to predict 50–5,000 bp amplicons produced on templates in the entire non-redundant nucleotide database; no amplicons were identified. We calculated the length and GC content of SDSI primers and full SDSI amplicon sequences and ARTIC v3 primers and amplicons using Geneious Prime (2019.2.1) and compared their distributions (Extended Data Fig. 2b,c). ARTIC and SDSI primer melting temperatures were matched and calculated using the New England Biolabs online calculator (version 1.13.0; <https://tmcalculator.neb.com>).

SDSI experimental validation. We sought to validate *in silico* predictions for the performance of our SDSI primers and amplicons. We ordered primers (from IDT; oligo sequences in Supplementary Data File 1) and performed quantitative PCR (qPCR) using the Q5 Hot-Start 2x Master Mix, with 500 nM SDSI primers and 0.17X SYBR Gold (ThermoFisher S11494), and without ARTIC primer pools. We performed this assay in triplicate in 10 μ l reactions on a QuantStudio 6 with the following cycling conditions: 95 °C for 30 s, followed by 35 cycles of 95 °C for 15 s and 65 °C for 5 min. We tested 4 conditions: (1) 0.5 μ l of an SDSI gene block (IDT) (1 pM), (2) 0.5 μ l of an SDSI gene block + 0.5 μ l of cDNA from an NP swab, (3) 0.5 μ l of cDNA from an NP swab, and (4) no template to detect any non-specific amplification of the primers (Extended Data Fig. 2a). We performed PCR on each SDSI oligo, using the standard SDSI + AmpSeq PCR conditions (<https://benchling.com/s/prt-R95g0tCxKOeCAqn8lAk3>), then ran the PCR products on a 2.2% agarose gel to confirm that these primers amplified the SDSIs and that the product was clean and of the expected size (Extended Data Fig. 2d).

We ordered unique oligos as TruGrade ultramers (IDT), then resuspended and stored them at 10 μ M in water (oligo sequences in Supplementary Data File 1). Further characterization for identification of 96 SDSIs was achieved by direct PCR amplification with primers containing the constant SDSI handle and an Illumina P5/P7 adapter, followed by sequencing with a MiSeq Nano 2 \times 150 bp kit (Illumina #MS-102-2002). SDSI reads were quantified by mapping each SDSI against other SDSIs with the 'align_and_count_multiple_report' workflow implemented in Terra, as described below, and purity and sequence fidelity of SDSIs were achieved by calculating the percentage of reads mapping to each SDSI out of total SDSI reads (Fig. 2b). Given these same data, we explored our SDSI mapping stringency threshold. We determined whether each SDSI was uniquely identified over a range of SDSI stringency thresholds (0.01%–50% of SDSI reads mapping, with a step size of 0.01%) (Extended Data Fig. 8). We tested 142 total unique SDSIs; all SDSIs amplified successfully with high sequence fidelity and purity (>95% of reads mapped to the expected SDSI in the experiment described above). The final set of 96 SDSIs was chosen after first-pass validation in a combination of clinical sample amplification tests, GC cut-offs and sequence homology cut-offs. SDSIs excluded because of poor amplification or impurity in clinical sample processing were not retested to determine whether error was technical or biological.

Viral CT determination. Viral CT values for all samples sequenced at the Broad Institute were obtained using the Centers for Disease Control and Prevention (CDC) reverse transcription qPCR (RT-qPCR) assay with the N1 probe as previously described²¹. Viral CTs for samples sequenced at JAX were obtained from various providers and the RT-qPCR assays used were thus variable.

CT normalization. CT normalization was performed by first setting a desired mock viral CT and calculating the difference between this desired mock viral CT and the measured viral CT of a given sample, rounding to the nearest whole number. We next calculated the number of doublings required for the mock viral CT (assuming 100% PCR efficiency), and multiplied this by the volume of cDNA input to be used for the normalization. The final volume of water used to dilute the cDNA was the doubling factor minus the volume of cDNA input. An example calculation is illustrated below:

CT normalization calculation:

N = Difference between actual and mock

X = Volume (μ l) of cDNA to use for normalization

DF = Doubling factor is $X(2^N)$

Volume water for dilution (μ l) = DF – X

This CT normalization was done for certain method development samples which are described throughout the manuscript as being 'mock diluted' or 'normalized to CT X'. The nosocomial cluster was normalized to CT 27. The majority of batch data generated at the Broad Institute underwent CT normalization to CT 25. Batch data from JAX did not undergo CT normalization.

cDNA generation and ARTIC amplification optimization. *Reverse transcriptase.* We tested reverse transcriptase enzymes using extracted RNA from four SARS-CoV-2-positive clinical samples (CTs = 13.9, 23.9, 29.6, 33.6) (Extended Data Fig. 6a,b). We added 2 μ l purified DNase-treated RNA as input into SuperScript III (Thermo 18080093), SuperScript IV (Thermo 18091050) or SuperScript IV VILO (Thermo 11756500). Superscript IV (SSIV) reactions were incubated at room temperature for 10 min, followed by 50 °C for 60 min and an inactivation step at 80 °C for 10 min. Superscript IV VILO shared the same protocol, but with a temperature of 85 °C for the inactivation step. We input 2.5 μ l of cDNA for ARTIC pool 1 PCR under standard conditions for 40 cycles. We then tested the resulting pool 1 using the scaled-down Illumina DNA Flex library construction (as described in Methods below) and sequenced it on Illumina Miseq (V2 reagent kit) with 2 \times 150 bp paired-end sequencing.

ARTIC PCR enzyme. We tested PCR enzyme efficiency using extracted RNA from SARS-CoV-2-positive clinical samples, followed by cDNA generation using SuperScript IV, and diluted the resulting cDNA to a mock CT value of 35 for standardization across all PCR enzyme tests. We set up the standard ARTIC PCR pool 1 and pool 2 using an input of 2.5 μ l, altering only the PCR enzyme and

corresponding buffer. We tested NEB Q5 Hot-Start High-fidelity 2x Master Mix (Q5 2X MM) (NEB M0494L), NEB Q5 Hot-Start High-fidelity 2x Master Mix plus 0.01% SDS, NEB Q5 Ultra II Master Mix (NEB M0544L), KAPA HiFi HotStart (Roche KK2601), and KOD Hot-Start DNA polymerase (Sigma–Aldrich 71842) (Extended Data Fig. 6c). We quantified the resulting ARTIC PCR amplicons using a High-Sensitivity DNA Qubit kit, then input 25 ng from each pool (50 ng total) into scaled-down Illumina DNA Flex library construction. The resulting libraries (except Q5 plus 0.01% SDS, which had no visible product using the TapeStation D1000 High-Sensitivity Kit) were quantified and pooled on Illumina Miseq (V2 reagent kit) with 2 \times 150 bp paired-end sequencing.

Rehybridization PCR. We optimized PCR cycling conditions on mock CT 35 cDNA (generated as described above) using standard ARTIC PCR primer conditions. We performed a catch-up/rehybridization PCR under the following conditions: 98 °C for 30 s, 95 °C for 15 s then 65 °C for 5 min (10 cycles); 95 °C for 15 s, 80 °C for 30 s then 65 °C for 5 min (2 cycles); 95 °C for 15 s then 65 °C for 5 min (8 cycles); then 4 °C hold (Extended Data Fig. 6e). We quantified the resulting ARTIC PCR amplicons using a High-Sensitivity DNA Qubit kit, then input 25 ng from each pool (50 ng total) into scaled-down Illumina DNA Flex library construction. We then quantified these libraries and pooled them on Illumina Miseq (V2 reagent kit) with 2 \times 150 bp paired-end sequencing.

Cycle test. We further optimized ARTIC PCR by modifying PCR cycle numbers. Extracted RNA from six SARS-CoV-2-positive clinical samples ranging from CT 27–37 were converted to cDNA with Superscript IV and amplified under standard ARTIC PCR reaction components (with Q5 2X MM), modifying the final number of cycles of PCR from 35, 40 and 45 (Extended Data Fig. 6g). We quantified cDNA and used a standard 50 ng of input for scaled-down Illumina DNA Flex library construction, then quantified the resulting libraries and pooled them on Illumina Miseq (V2 reagent kit) with 2 \times 150 bp paired-end sequencing.

Ramp test. We used mock CT 35 cDNA to test the effect of decreased ramp speed on genome recovery and coverage. ARTIC PCR conditions for this experiment were 98 °C for 30 s, followed by 40 cycles of 95 °C for 15 s and 65 °C for 5 min, with a cooling and heating ramping speed of 3 °C s⁻¹. We tested a slow ramp PCR protocol with the ramp speed reduced to 1.5 °C s⁻¹ (Extended Data Fig. 6f). Libraries were constructed with Illumina DNA Flex and sequenced on Illumina Miseq (V2 reagent kit) with 2 \times 150 bp paired-end sequencing.

Primer concentration optimization. Under standard ARTIC protocol conditions, we ordered lyophilized ARTIC v3 primers from IDT and resuspended them in water at 100 μ M each. Pool 1 primers consisted of all odd numbered amplicons, whereas pool 2 primers consisted of all even numbered amplicons. To generate the 100 μ M pool 1 primer stock, we combined 5 μ l of each 100 μ M pool 1 primer, and repeated this protocol for the even numbered primers to give a 100 μ M pool 2 primer stock. We selected a total of 20 amplicons as regions of low coverage from previous sequencing data (Supplementary Table 3). Low-coverage amplicons were present in both pools, with 11 coming from pool 1 and 9 coming from pool 2. For the primer 2X pools, we spiked in primers for the corresponding amplicons at 2 \times the concentration (20.8 nM final) of the other primers in the pool. For these low-coverage primers, we used 10 μ l of the 100 μ M stock rather than 5 μ l. We diluted both the original and 2X primer pools 1:10 in nuclease-free water to generate a 10 μ M working stock. We then selected 8 samples with varying CT values to determine whether selectively increasing primer concentrations reduced amplicon dropout (Extended Data Fig. 6d). We used the SDSI + AmpSeq protocol (without the SDSI or SDSI primers) and processed each sample with both the original primer pool, as well as the 2X primer pool, then sequenced these 16 samples on an Illumina Miseq (V2 reagent kit) with 2 \times 150 bp paired-end sequencing. Only 6 of the 8 samples generated complete genomes (>98%) in both conditions and were used for further analysis.

CT normalization experiment. The CT normalization experiment was performed by taking four individual clinical samples (CT = 18–25) with four randomly selected SDSIs and either not normalizing the cDNA or normalizing to CT 25, 26 or 27 before the ARTIC PCR (Extended Data Fig. 4b). Samples were processed with the standard SDSI + AmpSeq protocol described below and sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 cycles).

Illumina DNA flex. We performed a head-to-head comparison of standard Illumina Nextera DNA Flex and Nextera XT (Illumina FC-131-1096) library construction kits (Extended Data Fig. 6h). The Nextera XT protocol was performed as previously described^{21,33}. Both library construction methods were compared on post ARTIC v1 PCR amplicons from clinical samples. Briefly, we amplified samples with a range of SARS-CoV-2 viral CT values (CT = 22.9, 26.2, 30.3) with ARTIC v1 primers, producing 400-bp-size fragments. We quantified amplicons from each ARTIC primer pool and pooled in equal molar concentrations. Standard Nextera DNA Flex input was 100 ng (50 ng from each pool) and 1 ng (0.5 ng from each pool) for Nextera XT. We quantified and pooled the resulting libraries before sequencing on an Illumina Miseq (V2 reagent kit) with 2 \times 150 bp paired-end sequencing.

We optimized Illumina DNA Flex library construction (Illumina 20018705) with the goal of reducing normalization steps, cost and increasing throughput. We scaled down (0.5X) Illumina DNA Flex throughout the standard Illumina sequencing protocol, also scaling down sample input for a total of 50 ng (25 ng from each primer pool). Due to the CT normalization step, we removed the pre-DNA Flex DNA concentration and pooling step. We used 1–2 µl post ARTIC PCR amplicon as input into the scaled-down DNA Flex library construction, and performed post library construction quantification and pooling with more uniform library size and concentration, further reducing time and cost of pooling libraries for sequencing. This protocol was used for all method development experiments, the cluster investigation, and a portion of the batch data generated from both the Broad Institute and JAX.

SDSI + AmpSeq SDSI titration in ARTIC SARS-CoV-2 sequencing. To determine an optimal concentration for SDSIs in ARTIC SARS-CoV-2 sequencing, we diluted SDSI 49 to 0.6, 6, 60 and 600 copies per µl (1, 0.1, 0.01 and 0.001 fM); 1 µl SDSI 49 was added to 5 µl cDNA, to be split to 2 × 3 µl for each ARTIC pool (Extended Data Fig. 3 and Supplementary Data File 1). SDSI primers were added to each ARTIC pool with a final concentration of 40 nM. The cDNA from one clinical sample (MA_MGH_00195; CT = 16) was mock diluted to CT 20, 25, 30 and 35 for this experiment using the protocol described within the CT normalization section. On the basis of the results of this experiment, SDSIs were used at 6×10^2 copies per µl (1 fM) for all method development data. Batch processing modifications to this approach from the Broad Institute and JAX are detailed below.

SDSI + AmpSeq protocol. Full protocol details can be found at <https://benchling.com/s/prt-R95g0tCkKOCaQn8lAk3> (Fig. 1). Briefly, cDNA synthesis was performed on 2.5 µl DNase-treated viral RNA with SSIV following the manufacturer's protocol, with an extension of the 50 °C incubation from 10 min to 60 min. An additional cDNA normalization step could be performed (see above) or one could move directly into the ARTIC PCR by taking 5 µl cDNA and mixing this with 1 µl of a 1 fM SDSI (equal to 600 copies per µl). After mixing, the mixture was split into 2 × 3 µl aliquots and ARTIC primer pool 1 or pool 2 was added, as well as 1 µM of the spike-in forward and reverse primers (40 nM final concentration in the ARTIC pool). The ARTIC PCR conditions were 98 °C for 30 s, followed by 40 cycles of 95 °C for 15 s and 65 °C for 5 min. Pool 1 and pool 2 PCR reactions were combined and taken through library construction with scaled-down Illumina DNA Flex.

Broad Institute sample processing. The batch data from the Broad Institute was generated using SDSI + AmpSeq with minor modifications (Fig. 4). Briefly, SSIV was used for cDNA synthesis. Q5 2X MM was used for the ARTIC PCR, which was run for 35 cycles. The SDSIs were spiked in at 6×10^3 copies per µl and the SDSI-specific primers were added to each ARTIC pool at a final concentration of 40 nM. Library construction was performed either with the scaled-down Illumina DNA Flex (previously described) or COVID-seq (Illumina 20043675). Samples were sequenced on a NovaSeq 6000 SP Reagent Kit v1 (300 cycles), v1.5 kits (300 cycles) or NextSeq 500 v2 kit (300 cycles).

The correlation between the GC percentage for each SDSI and the percentage SDSI reads over total reads for SDSI 2–48 was performed with the samples sequenced at the Broad Institute ($n = 2,903$) (Extended Data Fig. 2e). A linear regression was used to evaluate significance (P value = 0.8160).

Jackson Laboratory sample processing. Data generated at Jackson Laboratory used two different protocols publicly available at <https://github.com/tewhey-lab/SARS-CoV-2-Consensus> (Fig. 4). All samples included 6×10^3 copies per µl of SDSIs and the SDSI-specific primers were added to each ARTIC pool at a final concentration of 4 nM. Samples processed from December 2020 to April 2021 used Lunascript (NEB E3010) for cDNA synthesis and Q5 2X MM for the ARTIC PCR, which was run for 35 cycles. These samples used scaled-down Illumina DNA Flex for library construction. Samples sequenced after April 2021 used the standard COVID-seq protocol. All samples were sequenced on a NextSeq 500 using paired 75 bp reads by the Genome Technology group on Jackson Laboratory's Bar Harbor campus. The validation of all SDSIs in clinical samples (Fig. 3a) was performed with this protocol and is presented as the percentage of SDSI reads over the total of all reads for each sample. Of note, the SDSIs (used at the lowest recommended concentration of 6×10^2 copies per µl) were reliably detected in the samples sequenced at JAX. This reliable detection, however, is also dependent on the sequencing depth used by the institution.

SDSI impact on genome recovery. For +/- SDSI experiments testing the impact on recovery of viral genomes, 14 clinical samples spanning a range of CTs (17.6–30) were selected (Fig. 3b and Extended Data Fig. 4a). Samples were CT normalized and split after cDNA synthesis into 2 × 5 µl aliquots. Samples below CT 20 were normalized to CT 25 and samples between CT 20–25 were normalized to CT 26. Fourteen randomly selected SDSIs were used, with each sample receiving either an SDSI (600 copies per µl) and the SDSI-specific primers (40 nM final concentration in the ARTIC pool), or just the ARTIC pool 1 and pool 2 mastermix with additional nuclease-free water and no SDSI primers. Samples were processed

according to the SDSI + AmpSeq protocol using scaled-down Illumina DNA Flex for library construction, sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles) and analysed as described below.

Statistical analysis for the +/- SDSI experiment involved analysis of the mean coverage for all 98 amplicons for the full sample set with a two-tailed Mann Whitney t -test and multiple comparison two-stage step-up Benjamini, Krieger and Yekutieli test, with the false discovery rate (FDR) set to 5%. All 98 amplicons were found to be not significantly different ($P > 0.05$) between the plus and minus SDSI group. Samples were also separated into three CT bins (CT < 27 ($n = 4$), 27–29 ($n = 6$), > 30 ($n = 4$)) and the test repeated for each CT bin. This analysis also revealed that there was no significant difference ($P > 0.05$) in the mean coverage across any amplicon for any CT bin.

Intentional SDSI contamination experiment. The intentional contamination experiment used SDSI 87 and SDSI 94. The SDSIs were mixed at five different proportions (SDSI 87:SDSI 94; 100:0, 75:25, 50:50, 25:75 and 0:100) (Fig. 5a). Each condition was performed in duplicate. All validation experiment samples were processed according to the SDSI + AmpSeq protocol using scaled-down Illumina DNA Flex for library construction. Samples were processed with the standard SDSI + AmpSeq protocol and sequenced on a NextSeq 500 Mid Output Kit v2.5 (300 Cycles).

Production and application of synthetic RNA spike-ins (SRSI). We ordered SDSI oligos with minor modifications to enable in vitro transcription of RNAs (including a T7 promoter upstream of the SDSI amplicon, as well as 17 bps of constant sequence within the primer region) (Twist Bioscience) (Supplementary Data File 1). For two SDSIs (SDSI 1 and SDSI 4), we in vitro transcribed RNA using a T7 transcription kit (NEB E2050), quantified RNA using an RNA screen tape (Agilent 5067-5579 and 5067-5580), then diluted RNA in water to 10 fM (6,000 copies per µl), 1 fM (600 copies per µl), 100 aM (60 copies per µl) and 10 aM (6 copies per µl). We added 1 µl SRSI at each concentration directly to 5 µl RNA from two patient samples with high and intermediate viral loads, respectively, and prepared sequencing libraries using the SDSI + AmpSeq protocol (without the SDSI addition step at the cDNA stage). For the sample with a high viral load, we performed a dilution at the cDNA stage (diluting 32-fold for a mock Ct of 25 rather than 20). Reads mapping to unique SDSI sequences and SARS-CoV-2 were quantified using the 'align_and_count_multiple_report' and 'assemble_refbased' wdfs, respectively, and % SDSI per combined reads was reported (Extended Data Fig. 10).

Computational analysis workflow. We analysed sequencing data on the Terra platform (app.terra.bio) using viral-ngs 2.1.28 with workflows that are publicly available on the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs).

Samples were demultiplexed using the 'demux_plus' workflow with a spike-in database file for the SDSIs. We performed any separate analyses to quantify read counts, including those for SDSIs, with the 'align_and_count_multiple_report' workflow with the relevant database. For most analyses involving direct comparisons between samples, we performed downsampling to the lowest number of reads passing filter with the 'downsample' workflow. We performed assembly using the 'assemble_refbased' workflow to the following reference fasta: https://www.ncbi.nlm.nih.gov/nucleotide/NC_045512.2?report=fasta. We used iVar version 1.2.1 for primer trimming on all samples, followed by assembly with minimap2 set to a minimum coverage of either 3, 10 or 20, skipping deduplication procedures. The computational pipeline for all samples sequenced at JAX is publicly available at <https://github.com/tewhey-lab/SARS-CoV-2-Consensus>.

Samples from the batch data were subset in the following way for analysis. All samples with a present SDSI were used for the percentage of SDSI reads out of the sum of all SDSI reads analysis (JAX: $n = 3,773$; Broad: $n = 2,903$). Samples with known experimental contamination errors or where the dominant (>50%) SDSI was not the correct SDSI were removed. For the percentage of SDSI reads over the total of all sequenced reads analysis (JAX: $n = 3,045$; Broad: $n = 2,670$), non-template controls (waters) and clinical samples with no detectable viral load (CT > 40 or not detected via qPCR as described above) were removed from analysis.

Metagenomic sequencing and comparison. Metagenomic sequencing data and genome assemblies used for the comparison of amplicon-based sequencing were prepared, sequenced, analysed as described previously²¹, and the data are made publicly available at NCBI's GenBank and Sequence Read Archive databases under BioProject [PRJNA622837](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA622837). We prepared amplicon sequencing libraries from the sample RNA extract following our SDSI + AmpSeq protocol (Fig. 1). We normalized cDNA samples that had a high viral load (CT < 27) to a CT of 27. To prepare for the ARTIC PCR, we transferred 5 µl of the normalized cDNA to a new plate and added 1 µl of an SDSI (600 copies per µl). After mixing, we transferred 3 µl to a new plate, added ARTIC PCR pool 1 mastermix and pool 2 mastermix to the respective plates, and incubated the mixture on a thermal cycler at 98 °C for 30 s, followed by 40 cycles of 95 °C for 15 s and 65 °C for 5 min. We then combined equal molar amounts of amplified samples for a total of 50 ng and processed the mixture through a 0.5X Illumina Flex library construction pipeline. We sequenced

the concordance dataset on a NovaSeq 6000 SP Reagent Kit v1 (300 cycles) and analysed the data as detailed in the methods below. For SNV analysis, the coverage depth over each divergent SNV was greater than 1,000× for both platforms, and both SNV calls persisted at relaxed ($n=3$) and conservative ($n=20$) minimum coverage thresholds. Primer trimming using iVar version 1.2.1 was manually confirmed.

Suspected nosocomial cluster investigation. We received NP swab samples in UTM and extracted RNA from 200 µl of biosample as previously described⁸. We prepared amplicon sequencing libraries as described above and analysed them as detailed in the methods below. A pairwise distance was calculated between all partial genomes (>80% complete), excluding gaps, to determine whether samples were likely to be the result of nosocomial transmission (Fig. 5c). We calculated the proportion of reads that mapped to a given SDSI out of all reads that mapped to any SDSI. Data has been made available in both the Short Read Archive and the NCBI GenBank under Bioprojects PRJNA622837 and PRJNA648098. GenBank accessions for SARS-CoV-2 genomes from this set of samples are MW454553–MW454562.

For phylogenetic tree reconstruction, we placed the suspected nosocomial cluster in a broader genomic context by performing a subsampling of the genome sequences available in GISAID³⁴ (Extended Data Fig. 9). We used the 'sarscov2_nextstrain' workflow to perform a Massachusetts-weighted subsampling of samples from 1 January 2020 to 1 November 2020. Our subsampled dataset included 3,146 sequences; 1,449 samples from Massachusetts, 1,425 samples from elsewhere in the United States and 283 from other countries. We constructed a maximum likelihood tree using iqtree with a General Time Reversible substitution model, and edited and interpreted the tree in Figtree v1.4.4.

Data presentation. Data analysis and graphing were performed using R Statistical Software (version 1.3.959; R Foundation for Statistical Computing), GraphPad PRISM (version 9.0.2; GraphPad Software, www.graphpad.com) and Python (version 3.7). We created original figures using BioRender (BioRender.com).

Statistics and reproducibility. The aim of our study was to establish the potential utility of our method, and study sample sets were largely dictated by the availability of clinical samples. No statistical method was used to predetermine sample size and experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment. We performed some experiments, such as RT-qPCRs, with technical replicates and Fig. 5a was run in duplicate.

Samples that failed sequencing due to technical reasons or where insufficient genome coverage was obtained for a given analysis (see Methods) were excluded. Data for Fig. 4b and Extended Data Fig. 2e were subsetted to exclude samples where SDSIs showed evidence of substantial contamination (expected SDSI is not the dominant (>50%) SDSI), the addition of no SDSIs, or failed sequencing. This dataset was further subsetted for Fig. 4c to exclude non-template controls and samples with no detectable CT for SARS-CoV-2. Analysis for 46 SDSIs that we did not proceed with further (due to suspected overlap with lab contaminants or contamination during synthesis or preparation) was omitted from this manuscript. Data from one sample that we added SRSI to was omitted from reported data.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequences and genome assembly data are publicly available on NCBI's Genbank and SRA databases under BioProjects PRJNA715749, PRJNA622837 and PRJNA648098. GenBank accessions for SARS-CoV-2 genomes newly reported in this study are MW454553–MW454562. SDSI sequences are reported in Supplementary Data File 1. We performed SDSI homology analyses using NCBI databases (*Homo sapiens* (taxid:9606), viral (taxid:10239)). We performed assembly using the 'assemble_refbased' workflow to the following reference fasta: https://www.ncbi.nlm.nih.gov/nucleotide/NC_045512.2?report=fasta. Additional data are available upon request. Source data are provided with this paper.

Code availability

Viral genomes were processed using the Terra platform (app.terra.bio), using viral-ngs 2.1.28 with workflows that are publicly available on the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs). Downstream analyses were performed using Geneious Prime 2020.1.2 or standard R (Version 1.3.959) packages. Custom scripts used to generate figures are available upon request.

Received: 31 March 2021; Accepted: 10 November 2021;
Published online: 14 December 2021

References

1. Washington, N. L. et al. Emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *Cell* **184**, 2587–2594.e7 (2021).

- Walensky, R. P., Walke, H. T. & Fauci, A. S. SARS-CoV-2 variants of concern in the United States—challenges and opportunities. *JAMA* **325**, 1037–1038 (2021).
- Wang, P. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).
- Focosi, D., Tuccori, M., Baj, A. & Maggi, F. SARS-CoV-2 variants: a synopsis of in vitro efficacy data of convalescent plasma, currently marketed vaccines, and monoclonal antibodies. *Viruses* **13**, 7, <https://doi.org/10.3390/v13071211> (2021).
- Wang, P. et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. *Cell Host Microbe* **29**, 747–751.e4 (2021).
- Naveca, F. et al. SARS-CoV-2 Reinfection by the New Variant of Concern (VOC) P.1 in Amazonas, Brazil <https://virological.org/t/sars-cov-2-reinfection-by-the-new-variant-of-concern-voc-p-1-in-amazonas-brazil/596> (2021).
- Genomic Sequencing of SARS-CoV-2: A Guide to Implementation for Maximum Impact on Public Health <https://www.who.int/publications/item/9789240018440> (WHO, 2021).
- COVID-19 Genomics U. K. (COG-UK) consortiumcontact@cogconsortium.uk. An integrated national scale SARS-CoV-2 genomic surveillance network. *Lancet Microbe* **1**, e99–e100 (2020).
- Chiara, M. et al. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Brief. Bioinform.* <https://doi.org/10.1093/bib/bbaa297> (2020).
- Charre, C. et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol.* **6**, veaa075 (2020).
- Rausch, J. W., Capoferri, A. A., Katusiime, M. G., Patro, S. C. & Kearney, M. F. Low genetic diversity may be an Achilles heel of SARS-CoV-2. *Proc. Natl Acad. Sci. USA* **117**, 24614–24616 (2020).
- Endo, A. Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott, S., Kucharski, A. J. & Funk, S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67, <https://doi.org/10.12688/wellcomeopenres.15842.3> (2020).
- Lagerborg, K. A., Watrous, J. D., Cheng, S. & Jain, M. High-throughput measure of bioactive lipids using non-targeted mass spectrometry. *Methods Mol. Biol.* **1862**, 17–35 (2019).
- Boja, E. S. & Rodriguez, H. Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins. *Proteomics* **12**, 1093–1110 (2012).
- Chen, K. et al. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Mol. Cell. Biol.* **36**, 662–667 (2016).
- illumina: COVIDSeq Test <https://www.illumina.com/products/by-type/ivd-products/covidseq.html> (2020).
- Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
- Quail, M. A. et al. SASI-Seq: sample assurance spike-ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics* **15**, 110, <https://doi.org/10.1186/1471-2164-15-110> (2014).
- Dilucca, M., Forcelloni, S., Georgakilas, A. G., Giansanti, A., & Pavlopoulou, A. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. *Viruses* **12**, 5 (2020).
- Potapov, V. & Ong, J. L. Examining sources of error in PCR by single-molecule sequencing. *PLoS ONE* **12**, e0169774 (2017).
- Lemieux, J. E. et al. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, 6529, <https://doi.org/10.1126/science.abe3261> (2021).
- So, A. P. et al. A robust targeted sequencing approach for low input and variable quality DNA from clinical samples. *NPJ Genom. Med.* **3**, 2, <https://doi.org/10.1038/s41525-017-0041-4> (2018).
- Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8, <https://doi.org/10.1186/s13059-018-1618-7> (2019).
- Pipelines R&D, et al. COVID-19 ARTIC v3 Illumina Library Construction and Sequencing Protocol v5 <https://doi.org/10.17504/protocols.io.bibtann> (2020).
- Lam, C. et al. SARS-CoV-2 Genome Sequencing Methods Differ in Their Abilities To Detect Variants from Low-Viral-Load Samples. *J. Clin. Microbiol.* **59**, 11, <https://doi.org/10.1128/JCM.01046-21> (2021).
- Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
- Metsky, H. C. et al. Zika virus evolution and spread in the Americas. *Nature* **546**, 411–415 (2017).
- Gohl, D. M. et al. A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. *BMC Genomics* **21**, 863, <https://doi.org/10.1186/s12864-020-07283-6> (2020).
- Itokawa, K., Sekizuka, T., Hashino, M., Tanaka, R. & Kuroda, M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by multiplex tiling PCR. *PLoS ONE* **15**, e0239403 (2020).
- Tyson, J. R. et al. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.09.04.283077> (2020).

31. *VarSkip: VarSkip Multiplex PCR Designs for SARS-CoV-2 Sequencing* <https://github.com/nebiolabs/VarSkip> (2021).
32. ARTIC: nanopore protocol for nCoV2019 novel coronavirus. <https://github.com/artic-network/artic-ncov2019> (2020).
33. Matranga, C. B. et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014).
34. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017).

Acknowledgements

We thank the microbiology laboratory staff and infection control personnel at Massachusetts General Hospital, Massachusetts Department of Public Health, the Rhode Island Department of Health, the Maine CDC and the Broad Clinical Research Sequencing Platform; all members of the regional COVID-19 emergency response efforts; P. Rodgers and the entire Broad Flow Core Team for sharing laboratory space and equipment; K. Barnes, S. Raju and S. Siddiqui for valuable feedback and helpful discussions; and R. Lynch, M. Barter and the Genome Technologies Service at The Jackson Laboratory. This work was funded by the National Institute of Allergy and Infectious Diseases (U19AI110818 and U01AI151812 to P.C.S.), Centers for Disease Control (75D30120C09605 to B.L.M. and baseline surveillance contract 75D30121C1050 to the Clinical Research Sequencing Platform, LLC), Maine CDC (contract to R.T. funded by the Centers for Disease Control ELC NU50CK000523), the Bill and Melinda Gates Foundation (Broad Institute), and the US Food and Drug Administration (HHSF223201810172C), with in-kind support from Illumina, Inc., as well as support from the Herchel Smith Fellowship (K.A.L.), the Doris Duke Charitable Foundation (J.E.L.), the Howard Hughes Medical Institute (P.C.S.), and the National Human Genome Research Institute (K99HG010669 to S.K.R. and R00HG008179 to R.T.). This work was made possible by support from the Flu Lab and a cohort of generous donors through TED's Audacious Project, including the ELMA Foundation, MacKenzie Scott, the Skoll Foundation, and Open Philanthropy. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We also thank the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, which

were used for Extended Data Fig. 9 (Supplementary Data File 2). We created Figs. 1 and 2b, as well as Extended Data Figs. 1, 2b,c, 8 and 10 with the aid of Biorender.com.

Author contributions

G.A., M.R.B., K.A.L., E.N., L.R.P., S.K.R., A.G.-Y., R.T., S.K., S.T.D. and D.B. performed laboratory experiments. K.F., C.L., K.A.L., E.N., H.B.D., S.K.R. and K.J.S. designed and performed data analysis. J.E.L., D.H., V.M.P., B.M.S. and E.S.S. identified and provided samples. R.T., S.K.R. and K.J.S. conceived the study. M.R.B., K.A.L., E.N., S.K.R. and K.J.S. drafted the manuscript. B.L.M., D.J.P., K.J.S. and P.C.S. secured resources and oversaw study implementation. All authors reviewed and approved the manuscript.

Competing interests

J.E.L. has received consulting fees from Sherlock Biosciences. P.C.S. is a co-founder of, shareholder in, and scientific advisor to Sherlock Biosciences, Inc., as well as a Board member of and shareholder in Danaher Corporation. M.R.B., K.A.L., E.N., S.K.R., K.J.S., B.L.M. and P.C.S. are co-inventors on a patent application (U.S. Provisional Patent Application No. 63/273,117) filed by the Broad Institute relating to methods of this manuscript. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41564-021-01019-2>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41564-021-01019-2>.

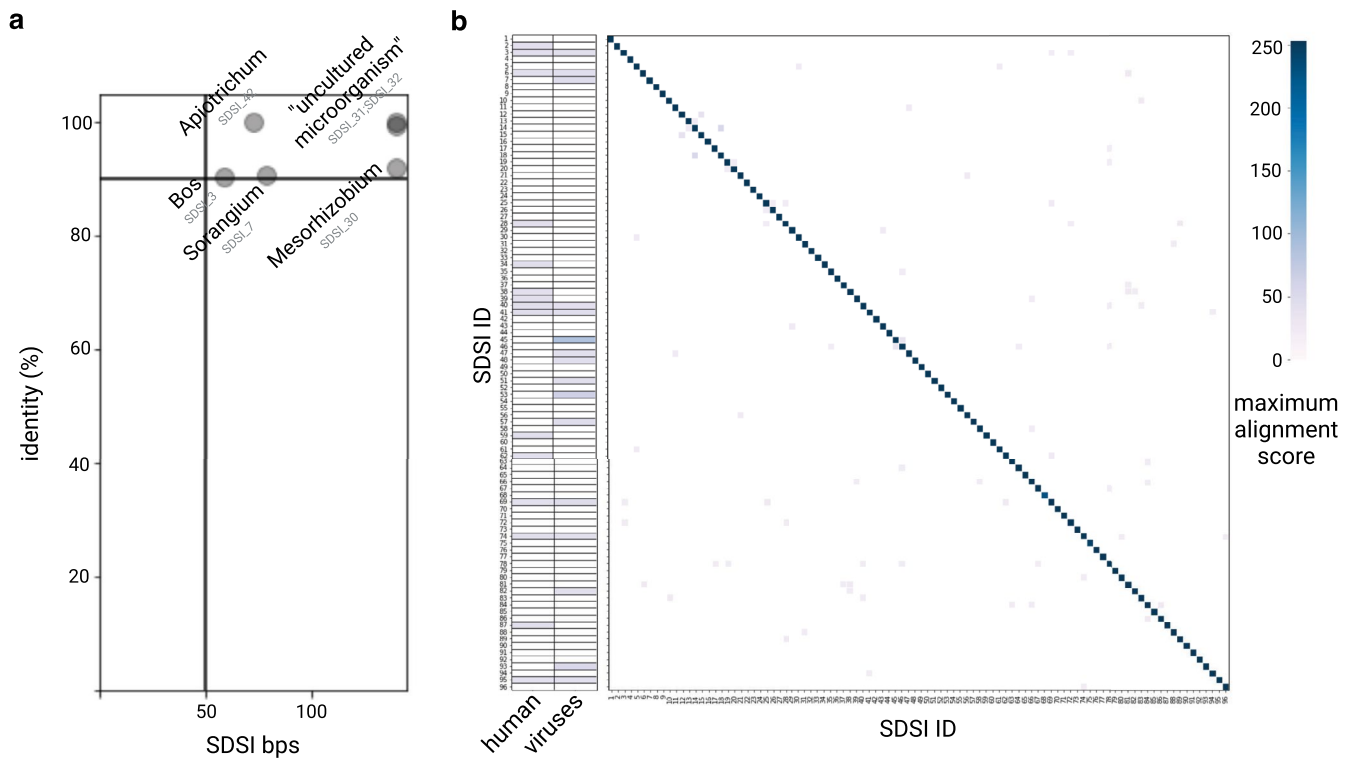
Correspondence and requests for materials should be addressed to Steven K. Reilly.

Peer review information *Nature Microbiology* thanks Rory Bowden, Charles Chiu, Nick Redshaw and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

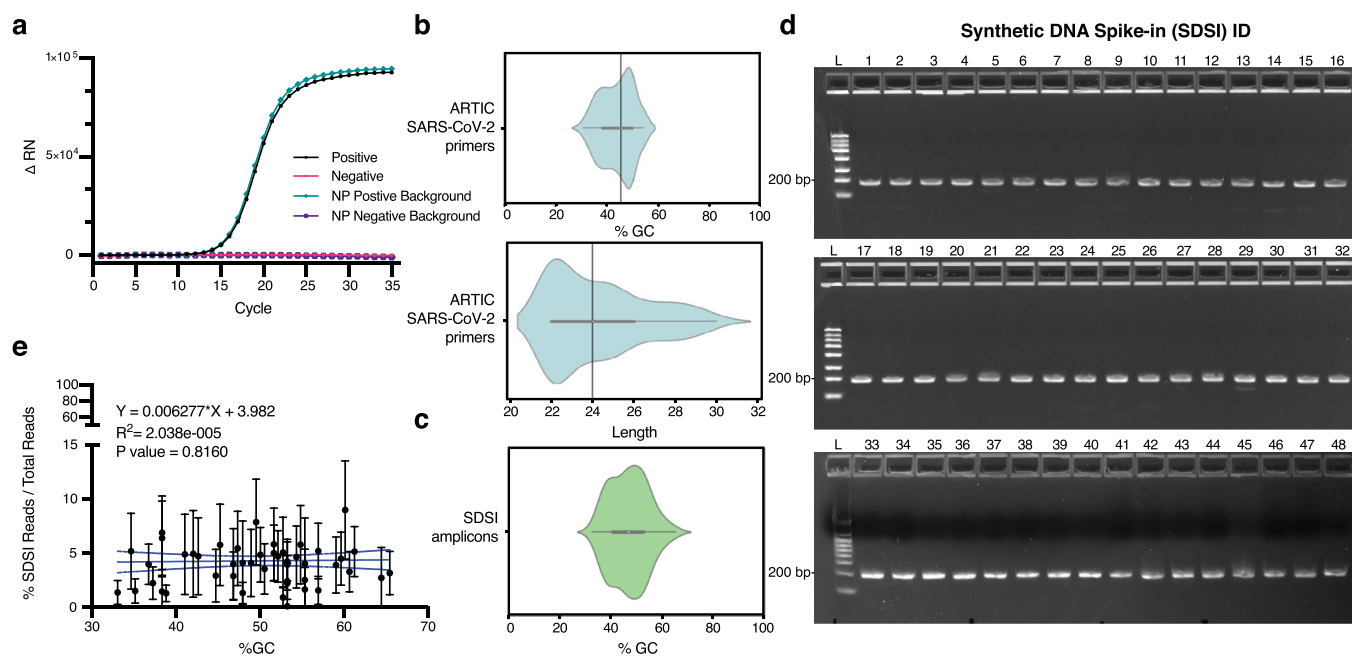
Reprints and permissions information is available at www.nature.com/reprints.

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

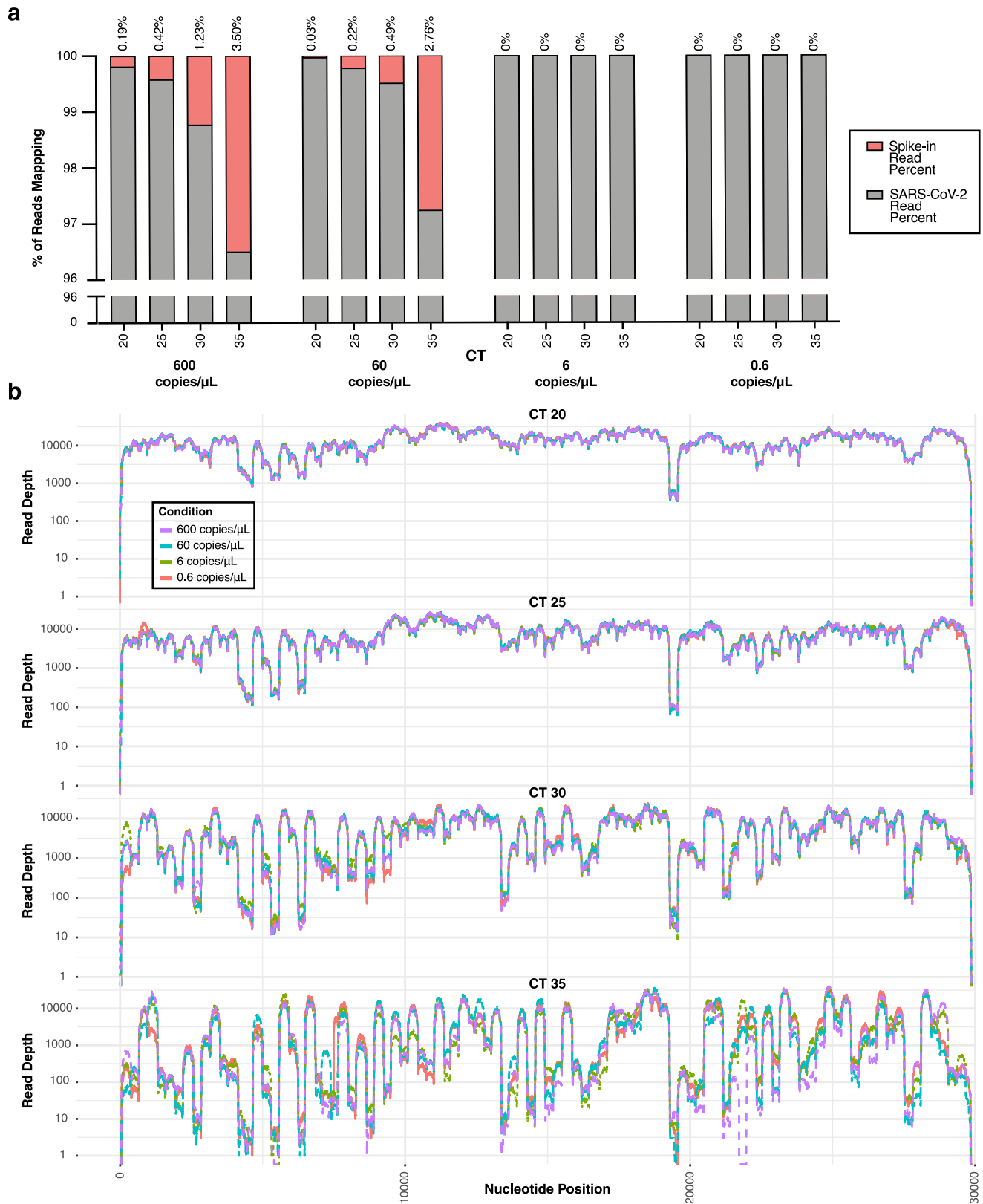
© The Author(s), under exclusive licence to Springer Nature Limited 2021



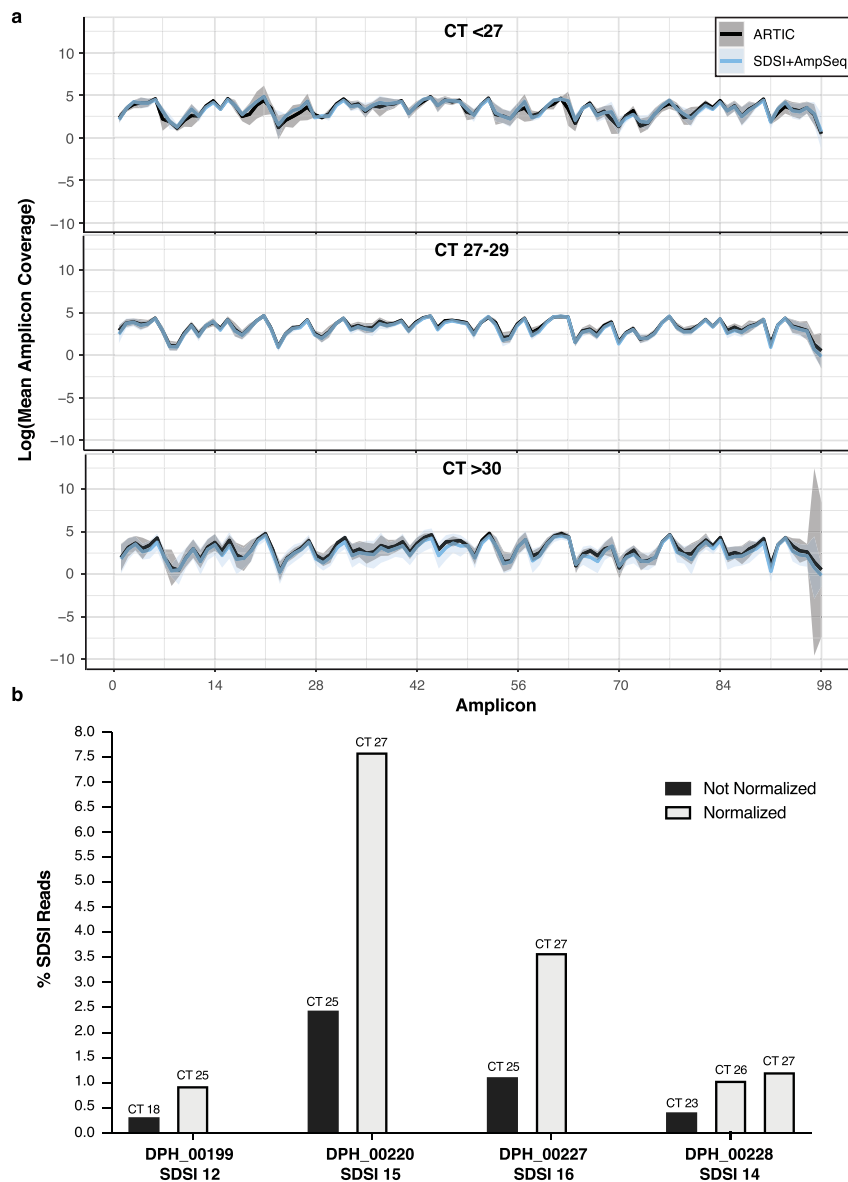
Extended Data Fig. 1 | SDSI core sequence in silico validation. We surveyed the core SDSI sequences by BLASTn to identify homology. **a**, Homology above our threshold between SDSIs and anything in the NCBI database outside the domain archaea was identified and the SDSI and genus were plotted if identity (y-axis) was greater than 90% and query cover (x-axis) was greater than 50 bps. **b**, For each SDSI, we identified and plotted (see color scale) the maximum alignment score for a homology above our threshold to human (taxid:9606) and viral (taxid:10239) sequences in the NCBI database. We also identified and plotted the alignment score for each pairwise combination of SDSIs.



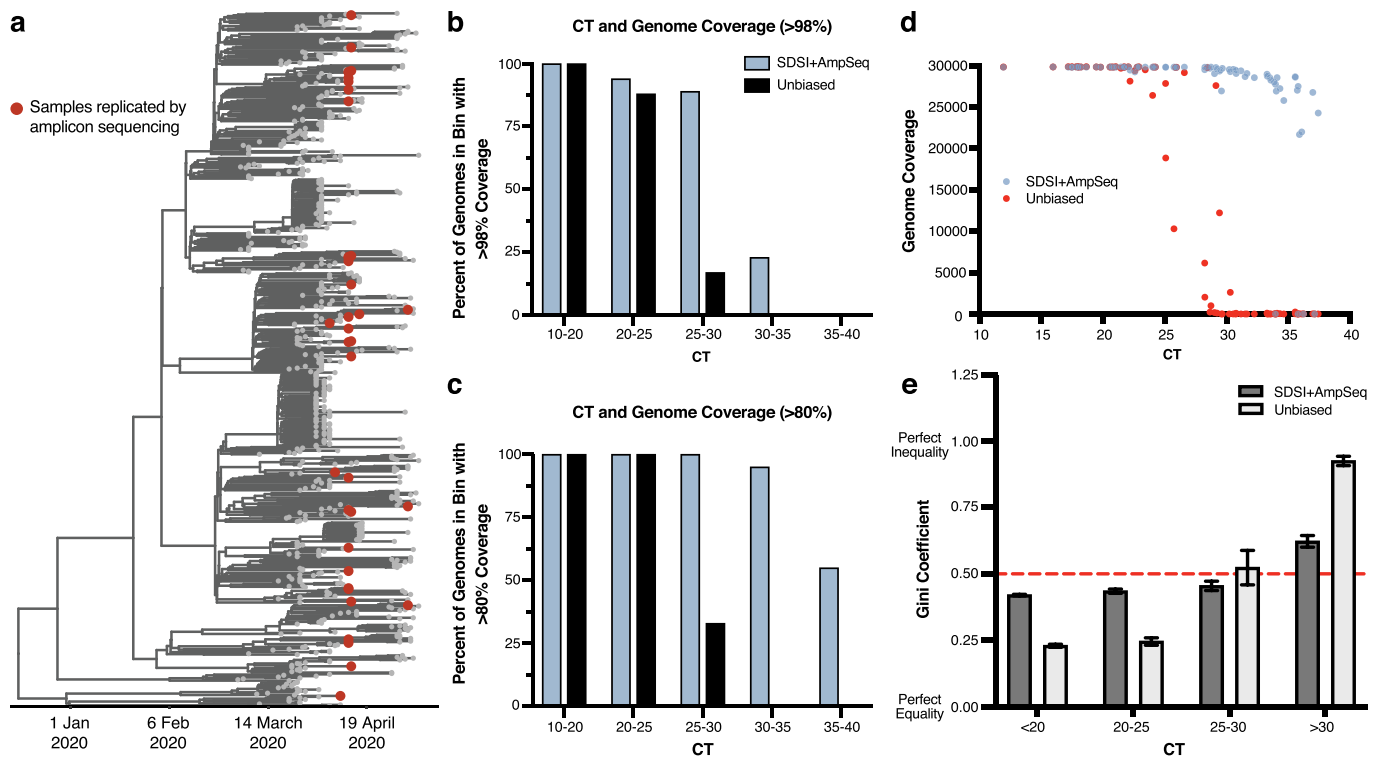
Extended Data Fig. 2 | Spike-in validation. **a**, RT-PCR for an SDSI in water and a SARS-CoV-2 positive clinical sample background. Mastermix and SDSI specific primers were added to all samples. SARS-CoV-2 positive clinical sample is cDNA generated from a nasopharyngeal (NP) swab. **b**, The distribution of GC content and length for ARTIC v3 primers. **c**, The distribution of GC content of SDSI amplicons ($n=96$). **d**, 100fmol DNA spike-in amplified under standard ARTIC PCR conditions for 40 cycles run on 2.2% agarose gel image with 188 bp amplified spike-in (SDSI 1-48) ($n=2$, representative image shown). **e**, % SDSI reads over total reads for SDSI (2-48) over a range of SDSI GC% (33%-65.4%) showed no significant read depth bias. Data are presented as mean values \pm 95% CI. Linear regression: p -value=0.8160, $R^2=2.038e-005$ (Broad, $n=2,903$ biologically independent samples).



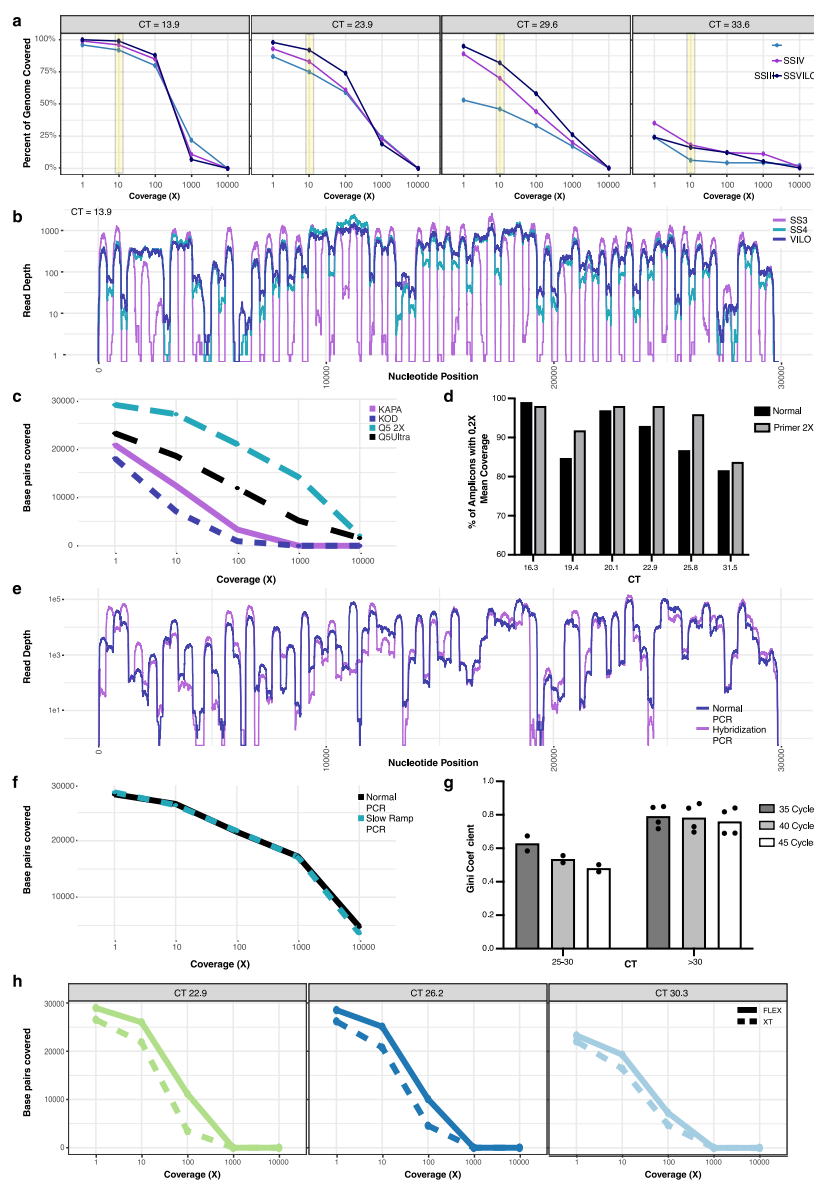
Extended Data Fig. 3 | SDSI Titration. **a**, In a titration of SDSI 49 across one clinical sample (CT = 16) mock diluted to various CTs (CT = 20, 25, 30, 35), the number of reads mapping to both SARS-CoV-2 and the SDSI were quantified, and the percentage of each was calculated. SDSI 49 was tested at 600, 60, 6, and 0.6 copies/μL in each mock diluted sample. **b**, Coverage plots for the SDSI 49 titration experiment.



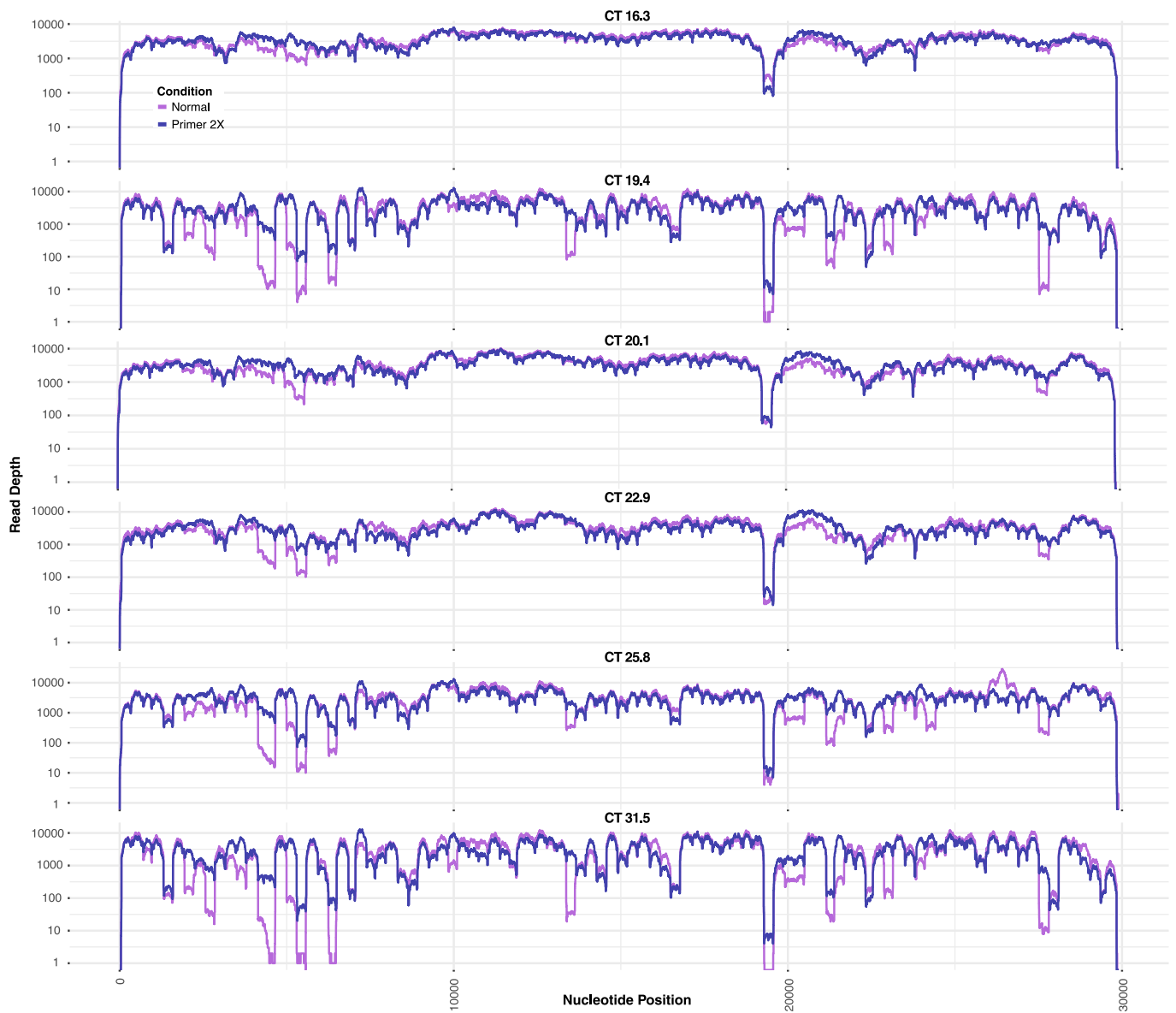
Extended Data Fig. 4 | ARTIC SARS-CoV-2 amplicon sequencing with and without SDSI and normalization. a, In three different CT bins, we show coverage plots with 95% confidence intervals for multiple samples sequenced with and without SDSIs (CT < 27, n = 4; CT 27-29, n = 6; CT > 30, n = 4). The solid blue line represents SDSI + AmpSeq and the solid black line is ARTIC only with no SDSI. Blue and black shading around the solid lines represents the 95% confidence interval. There were no significant differences (p-value > 0.05) between the with and without SDSI group for the mean coverage at any of the amplicons (two-tailed Mann Whitney t-test and multiple comparison two-stage step-up Benjamini, Krieger, and Yekutieli test with FDR set to 5%). **b**, The percentage of SDSI reads for 4 different SDSIs was assessed within 4 clinical samples that were run with and without CT normalization of the cDNA prior to the ARTIC PCR.



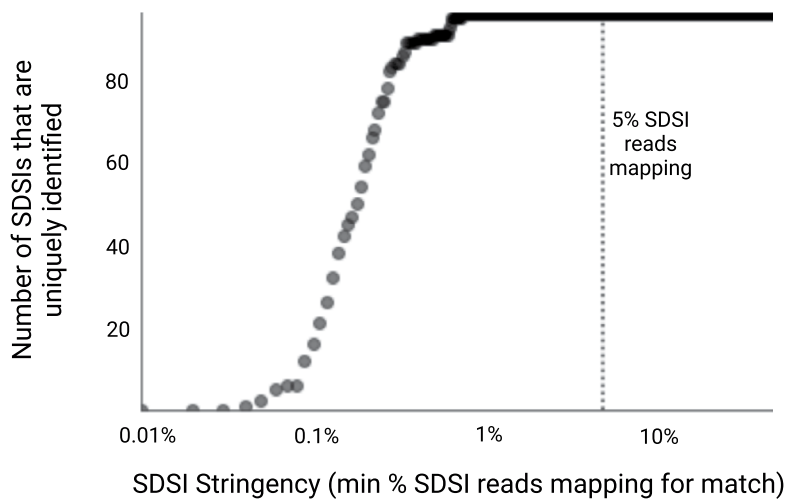
Extended Data Fig. 5 | SDSI + AmpSeq over a diverse set of samples has superior genome recovery and more coverage uniformity at higher CTs. a, Time-measured maximum clade credibility tree of 772 genomes from Massachusetts, reported in Lemieux et al., 2021. The 89 samples compared for metagenomic and amplicon sequencing are shown with red dots. **b**, Percent of assemblies with greater than 98% or **c**, 80% coverage in different CT bins (n = 81 biologically independent samples, excluded samples had no detectable CT) (downsampled to 975,000 reads). **d**, Genome coverage for unbiased metagenomic sequencing versus SDSI + AmpSeq amplicon sequencing pipeline (n = 81 biologically independent samples, excluded samples had no detectable CT). All samples downsampled to 975,000 reads. **e**, Gini coefficients grouped by CT (n = 70 biologically independent samples, excluded samples that did not generate assemblies in either one or both methods). Dashed red line represents a Gini coefficient of 0.50. Data are presented as mean values +/- SEM.



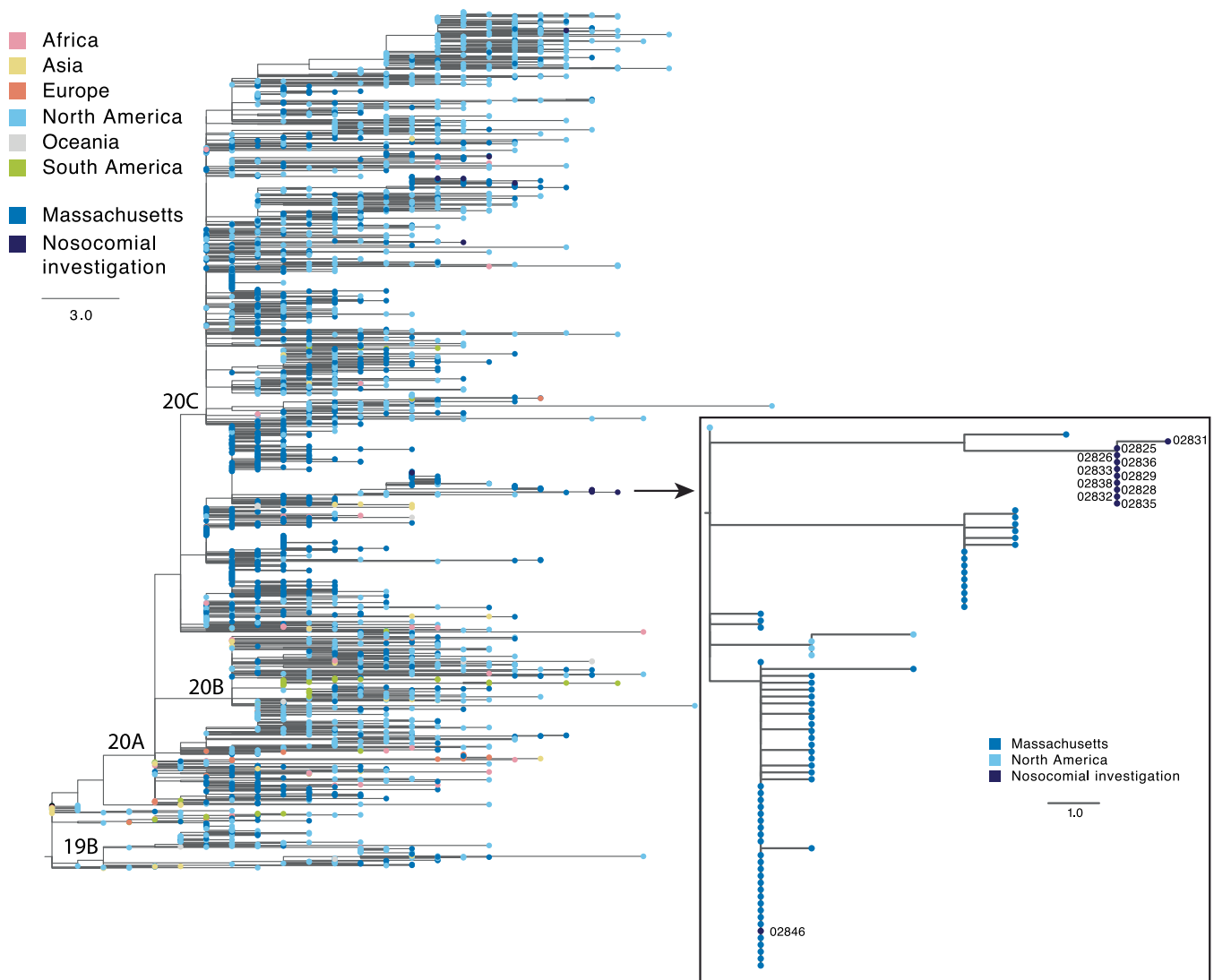
Extended Data Fig. 6 | Maximizing Genome Recovery and Coverage with SDSI + AmpSeq. **a**, The percent of the target genome covered at various depths of coverage for four individual samples (CT = 13.9, 23.9, 29.6, 33.6), with each undergoing cDNA with three different reverse transcriptases (SSIII, SSIV, or SSVILO). Yellow bar highlights comparison between the reverse transcriptases at a coverage depth of 10X. **b**, Read depth across each nucleotide position for the same sample (CT = 13.9) when using these reverse transcriptases. **c**, Base pairs of the SARS-CoV-2 genome covered at various depths when using different enzymes for the ARTIC PCR (n = 1). **d**, Amplicons with at least 0.2X of the mean amplicon coverage with the normal ARTIC v3 primer pools or with a modified primer pool with a 2X concentration of 20 poor-performing ARTIC primer pairs. Six individual samples with different CTs were used. **e**, Read depth across each nucleotide position for normal ARTIC PCR vs an alternate hybridization PCR (n = 1). **f**, Base pairs of the SARS-CoV-2 genome covered at various depths when using either normal ramping (3°C/s) or reduced ramping (1.5°C/s) speed for the ARTIC PCR (n = 1). **g**, Mean Gini coefficients for two mid-high CT samples (n = 2 biologically independent samples over 3 independent experiments) and four high CT samples (n = 4 biologically independent samples over 3 independent experiments) when using either 35, 40, or 45 cycles for the ARTIC PCR. **h**, Comparison of Nextera DNA Flex and Nextera XT on the number of SARS-CoV-2 base pairs covered at various depths of coverage for three samples with different CTs.



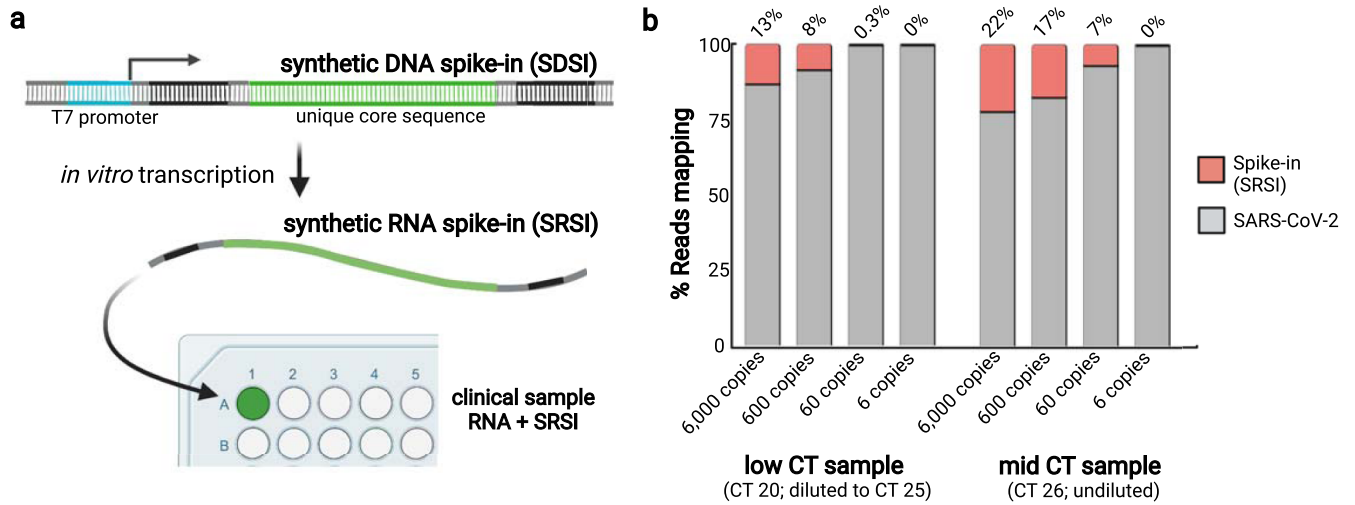
Extended Data Fig. 7 | Increasing primer concentration 2-fold in regions of low amplicon coverage. Data represents 6 individual samples at different CTs.



Extended Data Fig. 8 | Unique identification of SDSIs given varying thresholds of SDSI mapping stringency. We considered a range of cutoffs of the percentage of all SDSI-mapped reads mapping to a given SDSI (0.01%-50%, with a step size of 0.01). For an experiment where we sequenced SDSIs without any clinical sample, we calculated, at each cutoff, the number of SDSIs (y-axis) in the set we present (96 total) for which only the expected SDSI had a proportion of mapped reads that exceeded the cutoff (x-axis). Assuming no contamination, all 96 SDSIs should be identified uniquely, that is no other SDSI should have a proportion of mapped reads that exceeds the cutoff. The dotted line at $x=5\%$ represents the stringency cutoff that we recommend in practice to detect contamination events.



Extended Data Fig. 9 | Deployment of SDSI + AmpSeq to assess for possible nosocomial transmission. Phylogenetic tree showing the location of the putative cluster sequences in the context of a global subset of circulating SARS-CoV-2 diversity. Zoom box shows the 10 highly similar cluster genomes and one putative cluster sample that was excluded from the cluster based on genome sequence.



Extended Data Fig. 10 | Modification enables addition of spike-ins to RNA. **a**, A schematic of how to design, produce, and apply synthetic RNA spike-ins (SRSIs). **b**, A limited titration experiment where SRSIs of varying concentrations were added to two clinical samples with low and intermediate SARS-CoV-2 CTs. SRSIs were added to the sample at the RNA stage; the sample with a low CT (20) was then normalized to CT 25 at the cDNA stage, whereas the sample with mid CT (26) was not normalized.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection qPCR data was collected using QuantStudio Software v1.7.1.

Data analysis

The computational pipeline for all samples sequenced at JAX is publicly available at the following: <https://github.com/tewhey-lab/SARS-CoV-2-Consensus>. Viral genomes sequenced at Broad were processed using the Terra platform (app.terra.bio) using viral-ngs 2.1.28 with workflows that are publicly available on the Dockstore Tool Repository Service (dockstore.org/organizations/BroadInstitute/collections/pgs). This analysis utilized iVar (version 1.2.1). Downstream analyses were performed using Geneious Prime (2019.2.1) or standard R (Version 1.3.959) packages. Maximum likelihood phylogenetic tree was created using iqtree with a GTR substitution model and edited and interpreted the tree in Figtree v1.4.4. We used both the NCBI web interface and the command line interface (2.12.0+) of BLASTn. The New England Biolabs online calculator (version 1.13.0; <https://tmcaculator.neb.com>) to calculate primer melting temperatures. Data analysis and graphing was performed using R Statistical Software (version 1.3.959; R Foundation for Statistical Computing, Vienna, Austria), GraphPad PRISM (version 9.0.2; GraphPad Software, La Jolla California USA, www.graphpad.com) and Python (version 3.7). We created original figures using BioRender (BioRender.com). Custom scripts used to generate figures are available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Sequences and genome assembly data are publicly available on NCBI's Genbank and SRA databases under BioProjects PRJNA715749, PRJNA622837 and PRJNA648098. GenBank accessions for SARS-CoV-2 genomes newly reported in this study are MW454553 - MW454562. SDSI sequences are reported in Extended Data File 1. We performed SDSI homology analyses using NCBI databases [Homo sapiens (taxid:9606), viral (taxid:10239)]. We performed assembly using the assemble_refbased workflow to the following reference fasta: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. Sample size was limited to the maximum number of available clinical samples.
Data exclusions	Analysis for 46 SDSIs that we did not proceed with further (due to suspected overlap with lab contaminants or contamination during synthesis or preparation) was omitted from this manuscript. Data from one sample that we added SRSI to was omitted from reported data. Any data exclusions or data sub-setting that was performed for a specific analysis has been described in the methods. Samples that failed sequencing due to technical reasons or where insufficient genome coverage was obtained for a given analysis (see Methods) were excluded. Data for Figure 4b and Extended Data Figure 2e was subsetted to exclude samples where SDSIs showed evidence of substantial contamination (expected SDSI is not the dominant (>50%) SDSI), the addition of no SDSIs, or failed sequencing. This dataset was further subsetted for Figure 4c to exclude non-template controls and samples with no detectable CT for SARS-CoV-2.
Replication	qPCR replication was performed in a minimum of triplicate, which is a field standard. Figure 5a data was run in duplicate. Extended Data Fig. 2d was run twice with one representative image shown. Data from Fig 5c was re-run without SDSIs and with metagenomic sequencing to look at the samples implicated in the contamination event. Certain batches of samples for Figure 4 were re-run when a clear technical or experimental error was noted.
Randomization	Samples were randomly assigned to SDSIs, batched according to their collection timing.
Blinding	This study did not compare cases/controls or utilize a study design where blinding was required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Participants were clinically diagnosed with COVID-19.
Recruitment	Samples were secondary-use or residual clinical and diagnostic specimens, obtained by researchers under a waiver of consent.
Ethics oversight	Research was conducted at the Broad Institute with an exempt determination from the Broad Office of Research Subjects Protections and with approval from the MIT Institutional Review Board under protocol #1612793224. Samples were obtained from Massachusetts General Hospital (MGH), Massachusetts Department of Public Health, the Rhode Island Department of Public Health and the Broad Institute Clinical Research Sequencing Platform. Samples from Massachusetts General Hospital (MGH) fall under Partners Institutional Review Board under protocol #2019P003305. Samples sequenced at Jackson Laboratories (JAX) were approved under protocol 2020-NHSR-019-BH.

Note that full information on the approval of the study protocol must also be provided in the manuscript.