

CORONAVIRUS

Early sample tagging and pooling enables simultaneous SARS-CoV-2 detection and variant sequencing

Alon Chappleboim^{1,2†}, Daphna Joseph-Strauss^{1,2†}, Ayelet Rahat^{1,2†}, Israa Sharkia^{1,2}, Miriam Adam³, Daniel Kitsberg³, Gavriel Fialkoff^{1,2}, Matan Lotem^{1,2}, Omer Gershon^{1,2}, Anna-Kristina Schmidtner³, Esther Oiknine-Djian^{4,5}, Agnes Klochendler⁶, Ronen Sadeh^{1,2}, Yuval Dor⁶, Dana Wolf^{4,5}, Naomi Habib³, Nir Friedman^{1,2*}

Most severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) diagnostic tests have relied on RNA extraction followed by reverse transcription quantitative polymerase chain reaction (RT-qPCR) assays. Whereas automation improved logistics and different pooling strategies increased testing capacity, highly multiplexed next-generation sequencing (NGS) diagnostics remain a largely untapped resource. NGS tests have the potential to markedly increase throughput while providing crucial SARS-CoV-2 variant information. Current NGS-based detection and genotyping assays for SARS-CoV-2 are costly, mostly due to parallel sample processing through multiple steps. Here, we have established AphaSeq, in which samples are barcoded in the lysis buffer and pooled before reverse transcription. We validated this assay by applying AphaSeq to more than 500 clinical samples from the Clinical Virology Laboratory at Hadassah hospital in a robotic workflow. The assay was linear across five orders of magnitude, and the limit of detection was Ct 33 (~1000 copies/ml, 95% sensitivity) with >99.5% specificity. AphaSeq provided targeted high-confidence genotype information due to unique molecular identifiers incorporated into this method. Because of early pooling, we were able to estimate a 10- to 100-fold reduction in labor, automated liquid handling, and reagent requirements in high-throughput settings compared to current testing methods. The protocol can be tailored to assay other host or pathogen RNA targets simultaneously. These results suggest that AphaSeq can be a promising tool for current and future mass diagnostic challenges.

INTRODUCTION

Current methods for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) testing include a panel of reverse transcription quantitative polymerase chain reaction (RT-qPCR) tests, which are typically applied to nasopharyngeal swab samples (1). The swabs are mixed in a lysis buffer or heat-inactivated in a transport buffer, followed by RNA extraction and RT-qPCR. Samples with cycle threshold (Ct) lower than 35 are typically considered positive in these tests (2, 3). Although these tests are sensitive and specific, access to qualified labor and specialized equipment and reagents have limited testing capacity at different stages of the coronavirus disease 2019 (COVID-19) pandemic (4, 5). Testing capacity is also limited by tests that require samples to be treated as separate qPCRs with fixed reaction times. The most common strategy used to overcome this limitation has been sample pooling to achieve varying levels of “test compression,” but pooling strategies reduce the test sensitivity by diluting individual samples, and they rely on relatively low viral prevalence rates to be effective (6, 7).

In the past decade, next-generation sequencing (NGS) has replaced RT-qPCR and microarrays as the assay of choice for quantifying RNA molecules in research. During the COVID-19 pandemic, different NGS-based assays have been proposed to measure the presence and abundance of the viral genome in samples (8–16). In addition to detection and quantification, these assays can provide near real-time sequence information and provide epidemiologists with data on the emergence of new variants (17, 18). Similarly, by assaying the RNA from host cells, aspects of the immune response in infected individuals can be characterized (19) and can provide potential insights into the development of new treatments or vaccines.

Here, we propose an improved RNA sequencing (RNA-seq) protocol that allows for pooling of barcoded samples before RT, which we called amplicon pooling by hybridization and RNA-seq (AphaSeq). This workflow is relevant to large-scale testing by reducing labor, reagents, and overall costs by orders of magnitude in these settings.

We show that we can introduce barcoded and target-specific RT primers to the samples, allowing them to hybridize to target RNA molecules already in the lysis buffer or after an RNA cleanup step on polyT magnetic beads. Sample RNA is captured on beads after hybridization with the barcoded primers, and the primer-RNA hybrids are preserved during a subsequent wash step. The bead-bound RNA is isolated, and hundreds of samples are pooled to undergo RT from the primers that remain hybridized to their original targets. The pooled samples undergo library PCR and sequencing, and viral RNA genome counts per sample are determined by the sample-specific barcodes and unique molecular identifiers (UMIs) introduced at the beginning of the protocol. The observed molecules are

Copyright © 2021
The Authors, some
rights reserved;
exclusive licensee
American Association
for the Advancement
of Science. No claim to
original U.S. Government
Works. Distributed
under a Creative
Commons Attribution
License 4.0 (CC BY).

¹Alexander Silberman Institute of Life Science, Hebrew University of Jerusalem, Jerusalem 9190401, Israel. ²Rachel and Selim Benin School of Computer Science, Hebrew University of Jerusalem, Jerusalem 9190401, Israel. ³Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem 9190401, Israel. ⁴Hadassah Hebrew University Medical Center, Jerusalem 9112001, Israel. ⁵Lautenberg Centre for Immunology and Cancer Research, IMRIC, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem 9112001, Israel. ⁶Department of Developmental Biology and Cancer Research, IMRIC, Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem 9112001, Israel.

*Corresponding author. Email: nir.friedman@mail.huji.ac.il

†These authors contributed equally to this work.

also examined for known and unknown mutations. We validated our test on blinded synthetic samples and on a collection of ~550 clinical samples. We demonstrate that cross-sample contamination in this workflow is negligible, and we determined sensitivity to be ~800 to 1600 copies/ml, comparable to existing U.S. Food and Drug Administration (FDA)- and European Union (EU)-approved tests (20).

RESULTS

A simple and quick RNA capture step

SARS-CoV-2 nasopharyngeal swab samples typically arrive in lysis buffers that contain protein denaturation and degradation reagents. RNA extraction from lysis buffer is needed to allow for subsequent enzymatic reactions, including RT. The SARS-CoV-2 genome is a polyadenylated 30-kb RNA molecule. Thus, we tested, Solid Phase Reversible Immobilization (SPRI)- and polyT bead-based RNA capture techniques (21). These bead-based methods are inexpensive, rapid, and compatible with automation. In terms of RNA yield, the performance of both bead types was within a $\pm 50\%$ range of a widely used commercial kit (see supplementary note on RNA capture; Fig. 2A and fig. S1). Preliminary tests showed that both bead-based capture methods could be used with ApharSeq (fig. S1), highlighting the independence of the tagging step from the RNA capture technique. We focused on the polyT bead-based method, which can proceed without elution of RNA from the polyT beads, as the next steps of the protocol can be applied directly on bead-bound RNA.

Barcoded primers added to lysed samples prime RT reactions

We designed barcoded RT primers for the viral E gene (reverse), as it appears in the World Health Organization panel (22). The primer includes an interleaved 10-base pair (bp) barcode and a 10-bp UMI to allow for single-molecule counting (Fig. 1A, Materials and Methods,

and supplementary note on primer design) (23). Each sample is hybridized to primers with a different barcode (Fig. 1A), effectively identifying the source of RNA molecules for the remainder of the process. The bead-bound RNA is washed, pooled with other samples, and reverse-transcribed to generate barcode-labeled complementary DNA (cDNA) copies.

To evaluate the efficacy of primer-RNA hybrid formation and stability through the cleanup stage, we designed a qPCR targeting the generic PCR handle on the RT primer and the amplicon target sequence. This assay allowed us to quantify the hybrids that survived wash steps and generated cDNA molecules. Using the qPCR assay, we established that RT primers remain hybridized during RNA capture and initiate RT reactions (fig. S2). We used this assay to run several optimizations for the first steps of the protocol and markedly improved the cDNA yield (fig. S2).

Sequencing library preparation

The next step in the ApharSeq protocol is to generate sequencing libraries. This is achieved in a single PCR step, amplifying the target amplicons using a combination of a generic primer targeting the tail of the RT primer and amplicon-specific primers (Fig. 1B). These primers introduce Illumina-compatible sequences flanking the target amplicons (Fig. 1B). We applied this PCR to positive and negative samples and consistently obtained amplicon-specific libraries only in SARS-CoV-2-positive samples (Fig. 2B). These libraries yield highly specific results, with >95% of reads aligning to expected viral target sequences in positive samples (Fig. 2C) upon sequencing on NGS platforms, and the remaining <5% of reads were mostly primer dimers.

Cross-sample contamination is minimal

A critical concern related to pooling samples early in the protocol is that RNA molecules may be erroneously tagged due to residual free

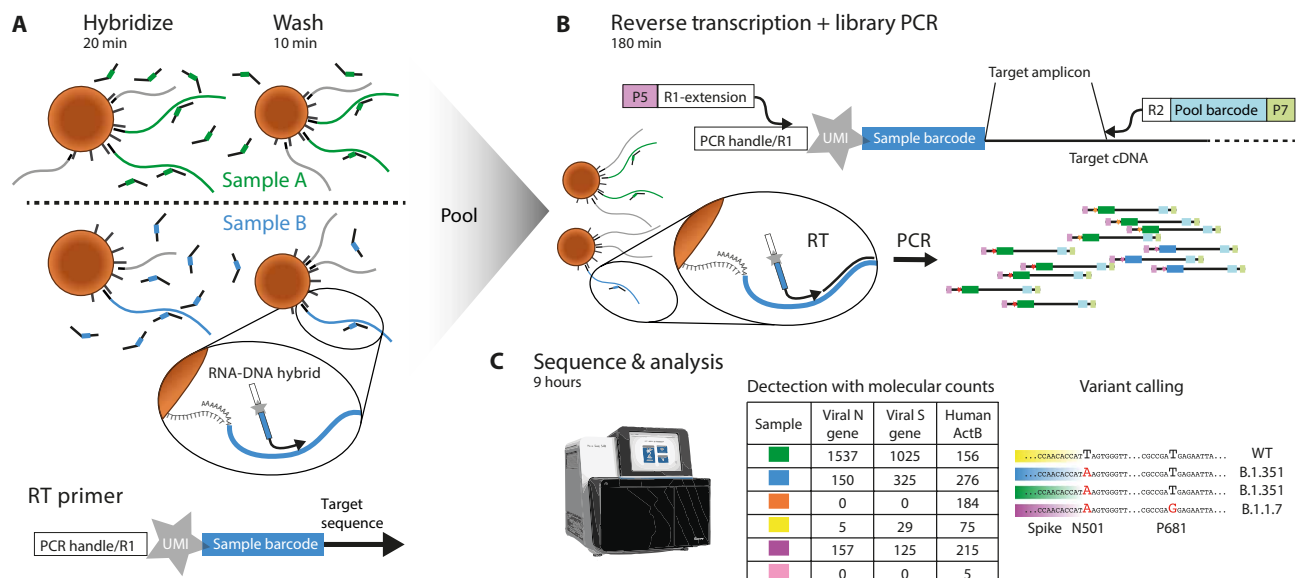


Fig. 1. ApharSeq overview. (A) Barcoded and uniquely identifiable RT primers are hybridized to samples in transport/lysis buffer, and homemade paramagnetic polyT beads are used for a quick wash and buffer replacement step. (B) Beads are pooled and RNA undergoes an RT/PCR with prehybridized target-specific barcoded primers to generate a sequencing library. (C) Libraries are sequenced and analyzed, and PCR duplicates are collapsed to molecular counts for detection and further analysis (e.g., major variant calling by sequence analysis and contact tracing by minor variations).

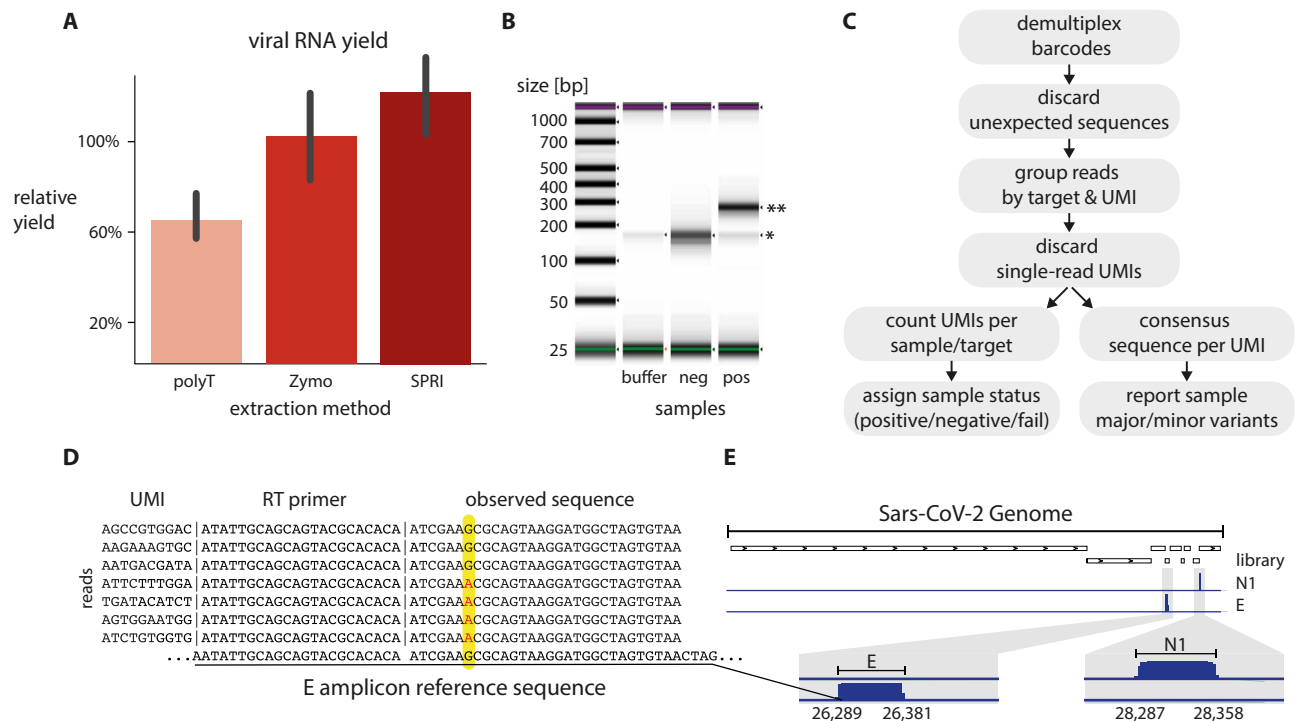


Fig. 2. The ApharSeq pipeline generates specific sequencing libraries. (A) RNA capture yields by homemade polyT and SPRI paramagnetic beads is efficient (~60–120% compared to Zymo kit) and quick (<30 min for a 96 plate on a robotic system). (B) Sequencing libraries of the viral envelope amplicon (E) are specific to positive samples (*primer dimer, **expected amplicon: 269 bp). (C) Analysis outline: sequencing reads are demultiplexed by barcodes, compared to the expected amplicon sequence, grouped by UMI, and counted to estimate viral load. Observed sequences are used to call viral variants. (D) Observed sequences conform to the reference genome in >99.9% of reads. However, in at least one sample, we observe a sequence variation in reads from the E amplicon at 26,353 (G to A). Observed single-nucleotide polymorphism (SNP) is in more than 90% of reads of each UMI shown (~200 reads each), excluding the possibility of sequencing errors. (E) Genome browser view of N1 and E amplicon libraries highlights target specificity.

primers or other artifacts during RT, PCR, or sequencing (24). To test potential cross-contamination levels at the RT stage, we hybridized positive (Ct 26) and negative samples with two differently bar-coded primers, pooled them, performed RT, and tested the amount of cross-contamination by barcode-specific qPCR (Fig. 3, A and B). We find that the pooled negative sample is indistinguishable from the unpooled negative sample, suggesting that cross-contamination is negligible.

To examine potential cross-contamination during PCR or sequencing (25), we subjected two samples with vastly different viral loads [high (Ct 18) or low (Ct 33)] and two negative controls (ddw) to ApharSeq. We hybridized the barcoded primers and pooled the samples before RT and PCR (Fig. 3C). Using the UMI sequence in each read, we were able to collapse PCR duplicates and provide an accurate and robust count of molecules captured in the assay. We found that barcodes that were hybridized to negative samples had at least 90,000-fold less observed molecules than those that were hybridized to the high Ct positive sample. These results are not unique to the polyT-based capture and were qualitatively replicated using the alternative SPRI-based RNA capture (fig. S1D). We conclude that cross-sample contamination is a minor issue in ApharSeq.

ApharSeq is quantitative and sensitive

To evaluate the dynamic range of ApharSeq, we titrated a positive sample into lysis buffer and generated samples that spanned a 64-fold range (Ct 23 to 31). We applied ApharSeq to these samples in a pool

and as individual samples (Fig. 4A). We found that the number of sequenced unique molecules scales linearly with the input ($P < 0.001$; Fig. 4B). By accounting for observed background in negative samples, we predict the limit of detection (LoD) to be ~Ct 35.3. We also compared the titration curve of the pooled and unpooled samples, which revealed minimal contamination between pooled samples (Fig. 4B).

We tested the LoD directly by performing another pooled titration experiment with highly diluted samples with a Ct range of 30 to 42 (Fig. 4C) and added samples with prequantified viral RNA (see Materials and Methods). This allowed us to estimate the end-to-end capture rate of ApharSeq at ~1.5% as we observed 33 and 14 molecules out of an input of ~2000 and ~1000 molecules, respectively. Similarly, we could calibrate the titration curve from Ct units to molecular counts and found the LoD to be 450 to 900 molecules/ml, depending on sequencing depth (Fig. 4D), threshold selection, and input volume used (see Materials and Methods).

A multiple target assay

A major advantage of sequencing-based assays is their capacity to capture and read large numbers of targets from the same sample (26). We next examined the potential for a multi-target assay by multiplexing two targets. We designed RT and PCR primers for the viral N1 amplicon, as described in a CDC (Centers for Disease Control and Prevention) panel (27), and used these primers in conjunction with the E amplicon RT and PCR primers. We applied ApharSeq to

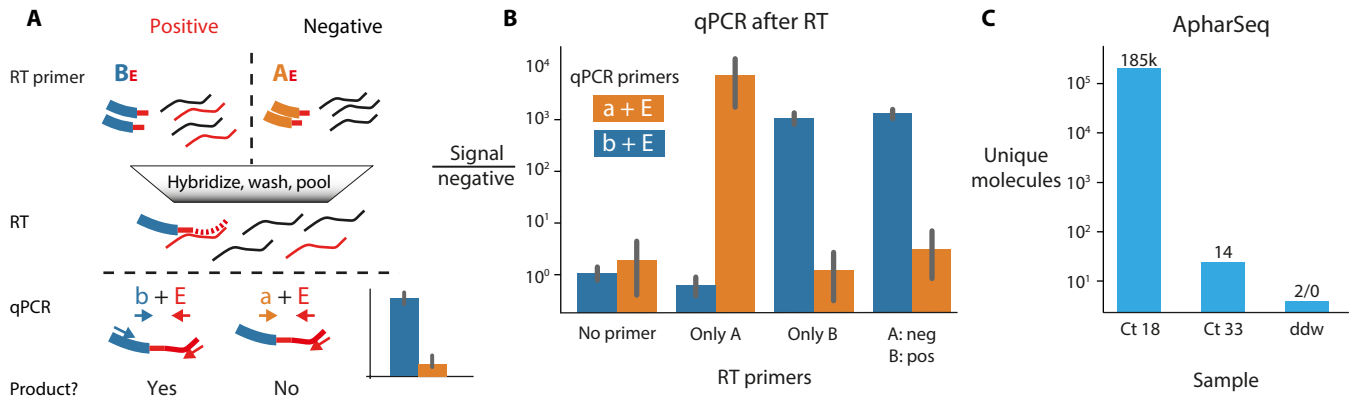


Fig. 3. Minimal cross-contamination in ApharSeq. (A) qPCR cross-contamination assay: Different RT primers (uppercase AE/BE) targeting the viral E amplicon are hybridized to a positive/negative sample, then samples are pooled and reverse-transcribed, and RT primer-specific qPCR primers (lowercase a/b) are used in qPCR to detect successful RT reactions on the viral target. (B) qPCR results in fold change relative to no template control. (C) Minimal cross-contamination in ApharSeq sequencing libraries as quantified by unique molecules detected in a pool of four samples with Ct 18, Ct 33, and two negative controls (ddw). Numbers above the bars are the counts of unique molecules in each sample.

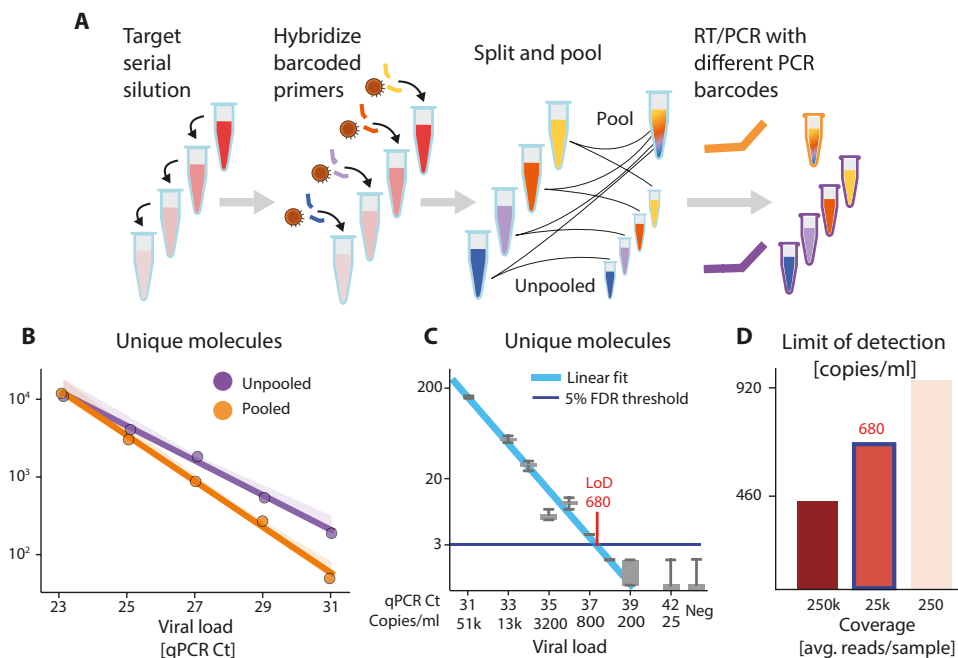


Fig. 4. ApharSeq sensitivity estimation on added samples. (A) Target titration experimental scheme. SARS-CoV-2-positive sample is diluted in lysis buffer. Diluted samples are hybridized to barcoded ApharSeq primers (interior tube color) and split, so they can be assayed separately or pooled. Samples are subjected to PCR with different barcodes (purple/orange) to distinguish their treatments. (B) Assay is linear in pooled and unpooled ApharSeq assays ($P < 0.001$). The linearly extrapolated LoD for pooled samples is ~Ct 35.7 and is four times lower than the LoD for the single samples. (C) Low target titration experiment that includes addition of a viral target control for quantification. Gray boxes are molecular counts over read down-sampling replicates that show that the LoD is ~680 molecules/ml (Ct ~37.4). (D) LoD as a function of sequencing depth as derived by down-sampling the sequencing reads. Blue outlined bar (~25,000 reads per sample) corresponds to data in (C).

a positive sample with each primer separately or with both primers together (Fig. 5A). The results of individual and multiplexed amplicons were almost identical (Fig. 5B), suggesting that the viral target sequences are amplified with minimal interactions and can be probed simultaneously to expand sequence information and improve confidence and sensitivity. The N1 amplicon yielded roughly two- to

threefold more molecules than the E amplicon, consistent with previous reports (28, 29).

As an internal control, we designed primers for several human transcripts with varying expression levels (30). After a preliminary test (Fig. 5C and fig. S3), we decided to continue with the *ACTB* amplicon as it is also used in an approved detection kit (31). We subjected positive and negative samples to the ApharSeq pipeline with primers targeting viral E and human *ACTB* amplicons to produce sequencing libraries, albeit with slightly reduced yields (Fig. 5D). qPCR tests on mixed libraries showed that decreasing the proportion of the human-specific primer in the PCR reduced the human/viral amplicon ratio accordingly, allowing for calibration of the number of reads allocated to each target in a multi-target library (fig. S3).

Evaluation of clinical samples

Last, we validated that ApharSeq can be used at scale by evaluating hundreds of samples. We developed and tested a robotic protocol on a Tecan liquid handling station. With our current unoptimized protocol, a single 96-sample plate is processed for 40 min and can be pooled into a single tube for RT-PCR. We used 96 barcoded RT primer plates for the

N1/E/ActB amplicons, with barcodes from a standard Illumina barcode collection.

We obtained positive ($n = 37$) and negative ($n = 465$) clinical nasopharyngeal swab samples. We randomly assigned samples to six 96-sample plates such that each plate had ~7.5% positivity rate (on average), and added positive and negative standards to each

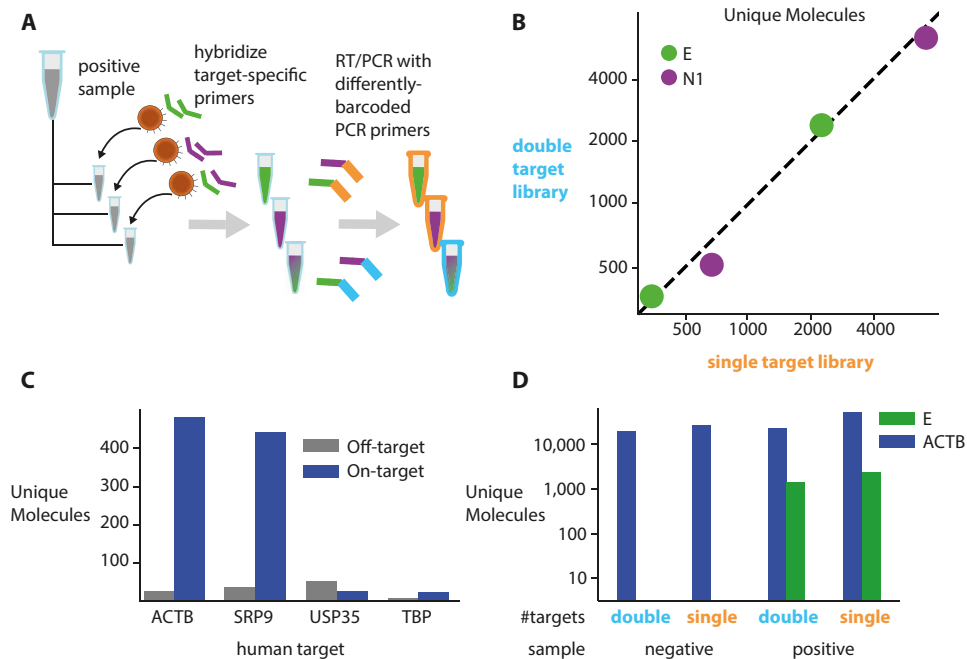


Fig. 5. Multi-target ApharSeq libraries. (A) Multiple target assay scheme. A positive sample was split and hybridized with E, N1, or E and N1 RT primers. These samples underwent ApharSeq with differently barcoded PCR primers to delineate reads from the double/single hybridization conditions. (B) ApharSeq libraries for high/medium viral load samples for the N1 and E amplicon separately (x axis) yield similar counts to a single ApharSeq library for both targets (y axis). Units are unique molecular counts. (C) Human target tests for multiple genes on a pool of negative samples highlights potential specific targets at different expression levels (see all targets tested in fig. S3). (D) Human and viral targets multiplexed in the same ApharSeq library (“double”) allow for internal control in negative/positive samples. Units are unique molecular counts.

plate (Fig. 6A). The samples in each of these plates were divided in half into two identical plates; one plate underwent standard RNA extraction and quantification by RT-qPCR, and the other plate underwent the ApharSeq protocol (Fig. 6A).

The positive controls exhibited high reproducibility between plates (Fig. 6A), and we observed quantitative agreement between the human internal control in the RT-qPCR assay and our *ACTB* amplicon reads, with a comparable number of missing values in both assays (i.e., samples that will require retesting; fig. S7). We compared the Ct value of each sample to the amount of unique molecules and observed a strong linear agreement with the N1 amplicon ($R^2 = 0.95$, $P < 10^{-38}$), which further established the quantitative nature of the assay (Fig. 6B), and we also observed a strong correlation between the N1 and E amplicons. In concordance with other reports (29), we observed more unique molecules of the N1 amplicon in virtually all the samples (Fig. 6D). To determine the LoD of the assay, we first fit the negative samples with zero-inflated Poisson to find a threshold for the number of reads at which the false-positive rate is below 0.05%. We then fit a linear model to the positive samples and find the maximal Ct where 95% of positive samples would be above the threshold. This procedure estimated a LoD of Ct 33, which is equivalent to ~1000 copies/ml (Fig. 6B and Materials and Methods).

A subsampling analysis of the data showed that sensitivity is maintained down to a sequencing depth of ~25,000 reads per sample (Fig. 6C). Overall, we conclude that the robotic ApharSeq protocol works efficiently, with minimal cross-sample contamination, and is highly quantitative.

Identifying viral variants in clinical samples with ApharSeq

ApharSeq uses sequencing for the detection of viral molecules, and hence, it is well suited to variant calling. A common hurdle in amplicon-based genotyping assays, especially in the case of viruses that might manifest multiple minor variants in the same host, is the inability to distinguish between PCR and sequencing errors from underlying genotype variants (32, 33). Unlike other common viral genome sequencing protocols (34, 35), ApharSeq incorporates a UMI at the RT step. Because PCR duplicates are produced with high fidelity, it is unlikely that sequencing or PCR errors will introduce the same variation in multiple copies of the same original molecule. ApharSeq can detect multiple reads with the same UMI and filter technical errors according to their consensus (Fig. 7A). Thus, UMIs confer increased confidence in the observed sequence and enable identification of minor genetic variants in the sample (Fig. 7B).

As a proof of concept, we designed additional primers to target the area around three mutations (N501Y, E484K, and P681H) in the spike gene of the UK (B.1.1.7/alpha) and South-African (B.1.351/beta) variants (36, 37). ApharSeq was able to detect these mutations when applied to positive samples suspected of being infected with these variants (Fig. 7C). Extrapolating from the number of observed molecules in these samples, we estimate that this amplicon around the P681 position in the spike gene will allow detection and variant calling up to Ct ~31.

DISCUSSION

RT-qPCR assays constitute the testing backbone in the COVID-19 pandemic and remain a critical tool for this constantly changing pandemic. Some sample pooling strategies have proven to be useful (6, 7), especially in light of shortages in reagents and testing equipment, but they remain limited when viral prevalence is high. RT-qPCR tests are not suited for genomic monitoring, and low- to mid-throughput NGS assays are currently used for the detection of existing and emerging SARS-CoV-2 variants of concern.

Improvements in NGS methods in the last decade have revolutionized multiple assays in research and in diagnostics, which is highlighted by several recent publications using NGS methods for large-scale SARS-CoV-2 testing (8–10). NGS-based tests can meet the needs for orders-of-magnitude scale-up and also provide ubiquitous genotyping data. Current NGS methods are reagent intensive and follow a laborious multistep protocol before sequencing. We developed ApharSeq, which is an early pooling protocol that markedly streamlines the NGS workflow and reduces reagent and labor costs. We established key properties of our approach, namely, linearity, sensitivity, low cross-reactivity, and the potential for

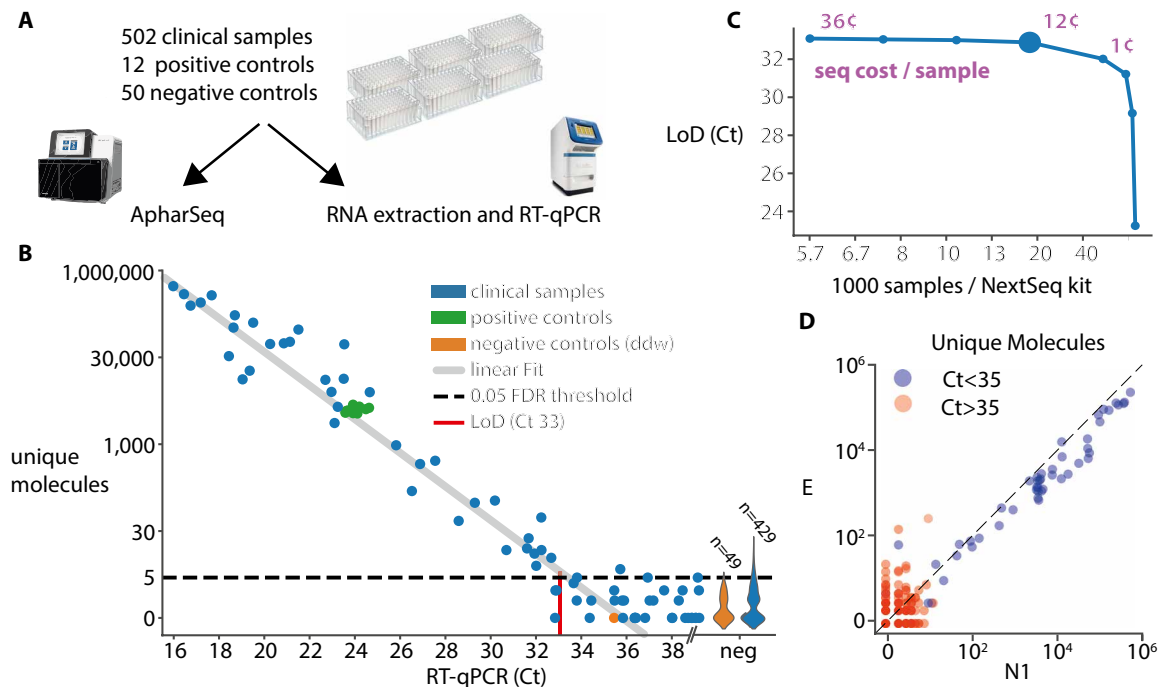


Fig. 6. Robotic test of hundreds of clinical samples. (A) Experimental design for clinical samples. Samples were randomly assigned to 96-sample plates with ddw-negative controls in each plate and two standard positive controls derived from a clinical sample that was diluted in negative samples as to suffice for multiple tests. The plates were then split and subjected in parallel to ApharSeq or to the standard clinical pipeline at Hadassah hospital (Zymo RNA extraction and BGI RT-qPCR). (B) The number of unique N1 molecules observed in each sample (y axis) is plotted against the measured Ct for the same sample. LoD was determined by controlling the false-positive rate to be 5% and assuming Poisson noise. Violin plot on the right of the x axis depicts the distribution of unique molecules observed for the samples that were not detected in the RT-qPCR. Positive controls (green) highlight reproducibility between clinically distinct and varied pools. (C) Subsampling analysis of the sequencing data demonstrates that samples are sufficiently sequenced and that there is a minimal decrease in sensitivity (y axis) when sequencing depth decreases down to ~25,000 reads per sample, which is equivalent to ~16,000 samples in a single NextSeq run (larger blue marker). Purple numbers indicate the sequencing cost per sample in selected sequencing depths. (D) Viral amplicon correlation within samples. Unique molecules of the E (y axis) and N1 (x axis) amplicons are plotted per sample, demonstrating reproducibility. Color correlates to the Ct threshold currently used in the clinically approved protocol (pos Ct < 35).

multi-target testing. We also demonstrate that ApharSeq can provide high-confidence variant calling and even detect minor sample variants because of the UMIs introduced in the first step of the protocol. These variants are crucial for a more complete understanding of the evolution process the virus is undergoing, and might help in detecting infection chains in the wild (33, 38, 39).

Multiple differences exist between recently published NGS-based diagnostic assays (8–16), including sample type (e.g., saliva and nasal swab) and reaction type [e.g., RT-PCR and loop mediated isothermal amplification (LAMP)]. Early hybridization and pooling as implemented in ApharSeq are compatible with various extraction methods, enzymatic reactions, and protocols. A concern in early pooling assays, such as ApharSeq, is the competition between pooled samples during amplification, resulting in a markedly uneven coverage of samples in the same pool, potentially reducing sensitivity and increasing sequencing costs. However, we find that ApharSeq can detect a positive sample across a large dynamic range within the same pool, with relatively shallow sequencing. Furthermore, positive controls provide an intrinsic measure for sequencing depth in each pool, allowing for deeper sequencing where needed. ApharSeq incorporates UMIs into the protocol, providing several benefits: (i) The protocol is approximately as quantitative as qPCR across a 10^5 dynamic range (Fig. 6B), (ii) the protocol provides higher confidence in genotyping even with a small number of observed molecules (Fig. 7), and (iii) erroneously assigned reads due to barcode hopping can be filtered out

(see supplementary note on barcode hopping). For future development, different features implemented in various NGS approaches can be incorporated into a single protocol. These include the introduction of synthetic internal controls in each sample (8, 9), testing for comorbid or confounding pathogens [e.g., the flu (8–10)], amplifying viral variable regions to investigate infection chains (40, 41), and monitoring the host immune response with key transcripts (42).

NGS-based methods have notable limitations. They require specialized and expensive sequencing equipment, which has a relatively slow readout process (e.g., a 55-bp run on a NextSeq requires ~7 hours). In addition, NGS methods incur substantial overhead costs due to the fixed cost of a sequencing run regardless of the number of samples tested and are therefore unfavorable in low-throughput settings. However, a rough calculation shows that equipment costs per unit throughput are in favor of sequencing (see the Supplementary Materials), and when amortizing over multiple samples, sequencing reagent costs are reasonable, ranging from \$10 per sample on a MiSeq (100 samples) to well below \$1 per sample (thousands of samples on a NextSeq/NovaSeq). Regarding the end-to-end assay duration, a 12- to 24-hour turnaround time is not suited for emergency testing but can be useful for large-scale and routine population testing as it represents a reasonable trade-off with costs and labor reduction. Further optimizations and workarounds can reduce sequencing runtimes by a factor of 2 to 3 (43), potentially reducing turnaround to under 12 hours.

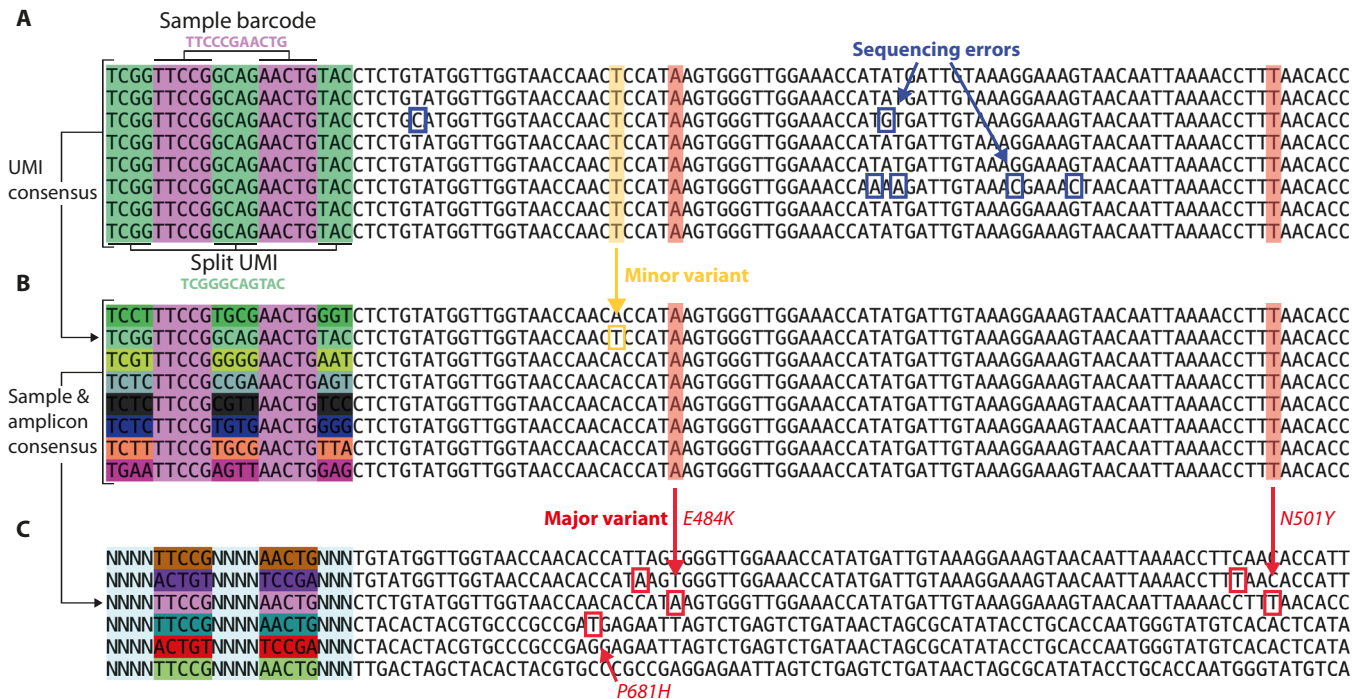


Fig. 7. Using ApharSeq for common strain classification. (A) Raw reads with the same UMI. Sequencing errors are detected (blue boxed letters) when collecting all same UMI reads emanating from the same RNA molecule. (B) Variant calling using the UMIs from the same sample (sample barcode in light purple); different molecules can be distinguished, properly counted, and used for major (red highlight/box) and minor (yellow highlight/box) variant calling per sample [UMI displayed in (A) is an outlier with an A→T variation indicated in green]. (C) Major variants per sample are called with dozens to thousands of unique molecules collapsed to a single observed consensus sequence at each assayed amplicon. Red boxes indicate major variant discrepancies from the reference sequence. Different offset primers are used to increase signal complexity during sequencing. We observe known spike protein mutations (N501Y, P681H, and E484K) in some of the shown samples (48, 49).

Each ApharSeq sample requires ~25,000 reads (see supplementary note on sequencing requirements), which means that a single Illumina NextSeq run with 400 million reads is sufficient to process 16,000 samples. The barcode design that we present here is amenable to changes, depending on the final sequencing scheme and expected throughput, but it is possible to design sufficiently distant 192 RT barcodes (pairwise edit distance >3) and 1536 pool barcodes (pairwise edit distance >2) to allow for the sequencing of ~300,000 samples in a single run. A linear increase in read length will result in a multiplicative scale-up of these numbers, effectively providing limitless pooling capability with an appropriately designed protocol.

Although increasing the size of the initial pool is beneficial in terms of labor and cost, we did encounter a slight inhibitory effect when pooling ~100 samples, suggesting that other unknown factors might limit the final pooling strategy. Last, we note that the approach developed and validated here, hybridization of barcoded primers followed by early sample pooling, is a generic protocol that can potentially be used to enhance existing protocols, including single-cell and bulk RNA-seq protocols.

MATERIALS AND METHODS

Study design

The goal of the clinical test was to find the extent of quantitative agreement between a benchmark qPCR assay and ApharSeq and allow us to estimate the LoD for our assay. Clinical samples were collected in the Clinical Virology Laboratory at Hadassah hospital. This study was part of the approved diagnosis optimization and

validation procedures at the Hadassah Medical Center, and therefore, no additional Institutional Review Board approvals were required.

Around 900 negative and ~100 positive samples were obtained. The only information that we had was that the samples were positive or negative. We randomly assigned positive samples to plates to have an average positivity rate of 7.5% per plate (i.e., final ApharSeq pool). As a negative control, we randomly assigned at least six wells in each plate to contain only water, and as a positive control, we diluted a positive sample in a pool of negative samples so that we had sufficient volume to allocate two aliquots of an identical positive control to each plate (in constant well positions). Unnamed negative samples were distributed sequentially to fill the remaining positions in each plate. Each plate was split and subjected to qPCR with a standard FDA-approved kit (BGI) and ApharSeq.

In 4 of the 10 plates, positivity rate exceeded 50% by qPCR, and together with the clinical laboratory staff, we concluded that the negative samples in these plates were likely contaminated in the laboratory. These four plates were discarded from further analysis. ApharSeq pools were subjected to an automated script on an EVOware 100 Tecan system, and the pools were stored in RNAlater for further processing following the ApharSeq protocol (detailed below).

RNA extraction benchmarking

Viral RNA at Ct ~14 was extracted from an in vitro grown virus (SARS-CoV-2 isolate USA-WA1/2020, NR-52281; obtained from BEI Resources) and serially diluted 1:25 in a negative sample. Each dilution was subjected to RNA extraction using one of three methods: (i) 400 µl of sample with Quick-RNA MagBead (Zymo Research),

following the kit manufacturer's instructions; (ii) 400 μ l of sample following the polyT capture described below; and (iii) SPRI-based capture as published (21) with modified volumes: 152 μ l of sample, 51 μ l of beads, 153 μ l of polyethylene glycol (PEG) buffer, and 122 μ l of binding buffer. Samples in the linear range were corrected for their dilution and collated to estimate the mean relative yield and error for each extraction method (Fig. 2A).

Primer design

All oligos used in this study are provided in a supplementary excel spreadsheet, and a detailed description with examples is available in the Supplementary Materials.

RT primers

RT primers consist of four main parts—from 5' to 3'—a general Illumina handle (Nextera R1), a 10-bp UMI, a 10-bp barcode, and a target-specific primer. After the first iteration of sequencing experiments, we decided to (i) interleave the UMI (U) and barcode (B) to avoid long stretches of the same nucleotide in the UMI sequence and (ii) add to each primer a variable sequence of 0 to 2 N's before the amplicon primer to increase complexity in each sequencing cycle.

PCR primers

There are two different PCR strategies that we used during the development process—one-step and two-step PCR. The two-step PCR is composed of a first step that amplifies the target molecules with an extendable handle, and in the second step, barcode and the remaining Illumina sequences are introduced. The one-step reaction performs everything in a single reaction with a single long primer (~90 bp). A one-step reaction is more convenient and is less prone to contaminations (see supplementary note on contaminations) but is less modular. The long primer contains a target-specific sequence and a barcode, which means that a barcoded primer collection must be synthesized per target. The two-step PCR decouples this dependency, which means that a single collection of barcoded primers can be used on any target, assuming that a simple target-specific primer is used in the first step. Both approaches yielded similar results, and we are currently using the one-step reaction to avoid contamination.

One-step PCR

The PCR amplifies the generic handle on the RT primer on one side and a target-specific sequence on the other side. In addition, the PCR extends the amplicons to a sequencing library by adding the relevant flanking Illumina sequences. An 8-bp barcode was included that marks the pool of samples amplified in the PCR.

Two-step PCR

The first step adds a target-specific handle on the forward side and extends the generic handle in the reverse (RT) side of the amplicon. We do this with the published Tn5-Rd1/Rd2 (Illumina FC-121-1030). The second PCR step extends the handles to a complete library with the Ad1.x and Ad2.x indexed primers as published (44).

ApharSeq protocol

The detailed and complete protocol was published separately (45). A specific instantiation of the ApharSeq protocol using polyT beads, assuming the use of 96 barcoded RT primers and 96 barcoded PCR primers, is given in the Supplementary Materials. Detailed steps are described below. Because the protocol stabilized with time, some experiments were slightly modified relative to the current protocol. The Supplementary Materials also contain a list of experimental modifications per experiment shown indexed by figure panel.

Bead preparation

We tested commercially available polyT beads (Thermo Fisher Scientific dynabeads, catalog no. 61002), or conjugated carboxylate-coated beads (GE Healthcare Sera-Mag SpeedBeads, catalog no. 65152105050250), and followed the manufacturer conjugation protocol with a 25-dT oligo.

Hybridization and RNA purification

Option 1: Purification and hybridization on polyT beads

This RNA purification protocol is based on a protocol for rapid isolation of mRNA (46) with some modifications. Briefly, polyT-conjugated beads were washed once and resuspended in binding buffer. The resuspended beads were mixed 1:1 with the sample. After a hybridization period of 10 min at room temperature with periodic mixing, the supernatant is removed and the beads are resuspended in a 50- μ l 1:1 mix of binding buffer and 10 μ M barcoded RT primers. To denature RNA secondary structures, the samples were incubated at 72°C for 2 min and immediately transferred to ice for at least 2 min. Samples were then incubated at room temperature for 10 min with periodic mixing to allow hybridization of RNA to the beads and to RT primers. Beads were resuspended in 450 μ l of wash buffer A and magnetized. The majority (380 μ l) of the supernatant was removed, and beads were resuspended in the remaining 70 μ l of buffer A and pooled. After pooling, samples are washed once in buffer A and twice in buffer B and can be kept in RNAlater until they are processed further. Preliminary tests show that RNA can be stored on the beads in RNAlater at 4°C for at least a week.

Option 2: Purification and hybridization on SPRI beads

RNA extraction with SPRI beads followed our published protocol for RNA extraction (21, 46) with several modifications. Samples in lysis/transfer buffer were mixed with barcoded RT primers, then incubated at 72°C for 2 min, and immediately transferred to ice for at least 2 min. Samples were then mixed 1:1 with binding buffer (as above) and incubated at room temperature for 10 min with periodic mixing to allow primer hybridization. Next, samples were mixed 1:0.8 with homemade SPRI beads in PEG buffer. Beads were washed twice with freshly made ethanol (80%), air-dried, and eluted in double distilled water. This was followed by a second 0.8 \times SPRI cleanup to ensure the removal of any excess primers. At this stage, samples were pooled to a PCR tube to undergo RT and PCR.

cDNA synthesis and library preparation

Twenty-five percent of pooled beads were subjected to proteinase K treatment (Lucigen), washed, and underwent RT reaction with SMARTScribe enzyme (SMARTScribe Reverse Transcriptase, Takara Bio) at 42°C for 1 hour followed by incubation at 70°C for 15 min. To elute the cDNA from the beads, the samples were incubated at 98°C for 2 min and magnetized, and the supernatant was transferred to a new tube and cleaned by SPRI beads x2 (Agencourt AMPure XP, Beckman Coulter). Illumina adaptors were added by PCR (30 cycles; KAPA HiFi HotStart ReadyMix, Kapa Biosystems), and the DNA was purified using 1 \times SPRI.

NGS data analysis

Reads were demultiplexed using bcl2fastq (version 2.20.0) and further processed by ad hoc Python scripts that are available as Jupyter Notebooks (10.5281/zenodo.5069979). For UMI analysis, we found the “uniq+” measure to be a simple and useful heuristic—we collapse unique UMIs and only count those with more than one associated

read. This strategy might result in some undercounting in under-sequenced samples, but this is a less important issue than counting spurious UMIs due to physical contamination, barcode hopping, or sequencing errors. A more detailed discussion is given in a supplementary note on UMI analysis.

Quantifying target molecules

To generate a quantitative polyA viral reference, we extracted RNA from a clinical sample and estimated the number of molecules in this RNA extract to be 6000 molecules/ μl using a synthetic viral sequence (Twist Bioscience SARS-CoV-2 RNA, MN908947.3) as a reference in a standard RT-PCR kit. We loaded two samples with 10 and 5 μl (~60,000/30,000 molecules) of this reference RNA in a total of 320 μl of lysis buffer and applied the ApharSeq protocol. Only $1/30$ of the pooled material underwent library preparation. Therefore, we expect to see ~2000 or ~1000 molecules in the 10- or 5- μl samples, respectively. After UMI clustering, we observe 33 and 14 molecules, respectively, suggesting that we capture ~1.5% of molecules. In the same experiment, a sample corresponding to cycle 29.3 had a similar UMI count (32 molecules), allowing us to roughly calibrate the Ct units to target molecules per milliliter at Ct 29.3 = $6150 \times 10/0.32 = \pm 190,000$ molecules/ml.

LoD determination

Titration LoD

For the high load titration (Fig. 4B), a linear fit (python `scipy.stats.linregress`) was performed

$$y \sim b + ax$$

where y is the $\log_{10}(\#\text{UMIs})$ and x is the calculated Ct of the sample. Given this linear fit, we can extrapolate to the UMI detection threshold, which, in this case, was set to 3 (a conservative estimate). The fit statistics are

Statistic	Singles	Pooled
Slope (a)	-0.242	-0.276
Intercept (b)	9.55	10.3
P value	2.86×10^{-4}	2.53×10^{-4}
R^2	0.99	0.99
Ct @ $\log_{10}(3)$	37.48	35.7

For the low load titration (Fig. 4, C and D), we perform resampling of the data ($\times 500$ times):

For each factor in (1, 3, 10, 30, 100, 300, 1000)

For each sample

For each UMI in sample

$\#\text{sampled-reads(UMI)} \leftarrow \text{Poisson}(\#\text{reads(UMI)}/\text{factor})$

We then count the number of UMIs per (sample, factor, replicate) as the number of UMIs with $\#\text{sampled-reads(UMI)} > 0$. Given these counts, we set the detection threshold as the minimal number of UMIs that is above 99% of replicates in the negative samples. Therefore, this number varies with sequencing depth and the UMI background in the negative samples. We fit each sampled replicate of the data with a Poisson-noised exponent

$$y \sim \text{Poisson}(2^{b-ax})$$

where y is the number of observed UMIs and x is the calculated Ct of the sample. We then set the LoD per factor to be the maximal Ct such that 95% of replicates are above the LoD.

Resampling factor	Average no. of reads per sample	UMI threshold	LoD (Ct)
$\times 1$	247,700	3	38.0
$\times 3$	82,600	3	37.8
$\times 10$	24,700	3	37.4
$\times 30$	8,260	3	37.2
$\times 100$	2,470	3	36.9
$\times 300$	826	2	37.2
$\times 1000$	248	2	36.8

Clinical test LoD

Ct 35 was used as a cutoff value for positive/negative samples, as is currently used in approved diagnostic protocols. We first determine the false discovery rate (FDR) threshold by fitting a zero-inflated Poisson mixture model (two components + zero component) to the molecule counts observed in negative samples. Using this model, we determine the theoretical threshold of detection for a given FDR. In this case, when FDR is set to 5%, the threshold is five molecules.

We fit the positive samples with a linear model (slope = -1), assuming Poisson noise. Given the linear model, the FDR threshold (five molecules), and assuming Poisson noise, the LoD is determined to be the maximal Ct in which the probability of obtaining a value lower than the threshold is lower than 5%. We subsample reads from the data and repeat this analysis to every sequencing depth to obtain the LoD as a function of sequencing depth (Fig. 6C).

Viral sequence variation analysis

As a proof of concept, we designed primers spanning the P681H mutation and the N501Y and E484K mutations, allowing us to distinguish between the alpha, beta, and gamma variants. We first cluster reads by their UMIs and then count the observed viral sequences associated with each UMI. If a UMI was observed more than two times and >50% of reads associated with that UMI have the same observed viral sequence, we call that sequence the major sequence for that UMI. There are two cases in which the major UMI sequence will have a mutation relative to the reference (or consensus) sequence: (i) An RT error had occurred and (ii) the original RNA molecule had a mutation.

The first case occurs once every ~30,000 reverse-transcribed bases (47), i.e., the probability of a single base to have a specific error (e.g., G to A) is roughly 1:90,000, which will be the (theoretic) false detection rate of a specific mutation.

To estimate the LoD, we use the samples that were tested with the P681 amplicon. We have four samples

Genotype	Ct	No. of unique molecules
P681H	20.7	32,984
P681H	20.0	25,641
P681H	22.0	15,120
P681P	20.5	21,653

Extrapolating from these numbers, we predict that five molecules should still be observable at Ct 31 with >95% confidence.

SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scitranslmed.abbj2266

Materials and Methods

Figs. S1 to S7

Tables S1 to S3

Data file S1

References (50, 51)

[View/request a protocol for this paper from Bio-protocol.](#)

REFERENCES AND NOTES

- R. Weissleder, H. Lee, J. Ko, M. J. Pittet, COVID-19 diagnostics in context. *Sci. Transl. Med.* **12**, eabc1931 (2020).
- R. Jaafar, S. Aherfi, N. Wurtz, C. Grimaldier, T. Van Hoang, P. Colson, D. Raoult, B. La Scola, Correlation between 3790 quantitative polymerase chain reaction–positives samples and positive cell cultures, including 1941 severe acute respiratory syndrome coronavirus 2 isolates. *Clin. Infect. Dis.* **72**, e921 (2021).
- C. Johnston, B. Healy, Interpretation of COVID-19 PCR testing—What surgeons need to know. *Br. J. Surg.* **107**, e367 (2020).
- G. B. Barra, T. H. Santa Rita, P. G. Mesquita, R. H. Jácomo, L. F. A. Nery, Overcoming supply shortage for SARS-CoV-2 detection by RT-qPCR. *Genes* **12**, 90 (2021).
- J. Bradley, "In scramble for coronavirus supplies, rich countries push poor aside," *New York Times*, 9 April 2020.
- L. Mutesa, P. Ndishimye, Y. Butera, J. Souopgui, A. Uwineza, R. Rutayisire, E. L. Nduricimpaye, E. Musoni, N. Rujeni, T. Nyatanyi, E. Ntagwabira, M. Semakula, C. Musanabaganwa, D. Nyamwasa, M. Ndashimye, E. Ujeneza, I. E. Mwikarago, C. M. Muvunyi, J. B. Mazarati, S. Nsanjimana, N. Turok, W. Ndifon, A pooled testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature* **589**, 276–280 (2021).
- R. Ben-Ami, A. Klochendler, M. Seidel, T. Sido, O. Gurel-Gurevich, M. Yassour, E. Meshorer, G. Benedek, I. Fogel, E. Oiknine-Djian, A. Gertler, Z. Rotstein, B. Lavi, Y. Dor, D. G. Wolf, M. Salton, Y. Drier; Hebrew University-Hadassah COVID-19 Diagnosis Team, Large-scale implementation of pooled RNA extraction and RT-PCR for SARS-CoV-2 detection. *Clin. Microbiol. Infect.* **26**, 1248–1253 (2020).
- J. S. Bloom, E. M. Jones, M. Gasperini, N. B. Lubock, L. Sathe, C. Munugala, A. S. Boeshaghi, O. F. Brandenberg, L. Guo, J. Boockock, S. W. Simpkins, I. Lin, N. LaPierre, D. Hong, Y. Zhang, G. Oland, B. J. Choe, S. Chandrasekaran, E. E. Hilt, M. J. Butte, R. Damoiseaux, A. R. Cooper, Y. Yin, L. Pachter, O. B. Garner, J. Flint, E. Eskin, C. Luo, S. Kosuri, L. Kruglyak, V. A. Arboleda, Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing. *medRxiv* (2020).
- R. Yelagandula, A. Bykov, A. Vogt, R. Heinen, E. Özkan, M. M. Strobl, J. C. Baar, K. Uzunova, B. Hajdusits, D. Kordic, E. Suljic, A. Kurtovic-Kozaric, S. Izetbegovic, J. Schaeffer, P. Hufnagl, A. Zoufaly, T. Seitz, VCDI, M. Födinger, F. Allerberger, A. Stark, L. Cochella, U. Elling, Multiplexed detection of SARS-CoV-2 and other respiratory infections in high throughput by SARSeq. *Nat. Commun.* **12**, 3132 (2021).
- K. U. Ludwig, R. M. Schmithausen, D. Li, M. L. Jacobs, R. Hollstein, K. Blumenstock, J. Liebing, M. Slabicki, A. Ben-Shmuel, O. Israeli, S. Weiss, T. S. Ebert, N. Paran, W. Rüdiger, G. Wilbring, D. Feldman, B. Lippke, N. Ishorst, L. M. Hochfeld, E. C. Beins, I. H. Kaltheuner, M. Schmitz, A. Wöhler, M. Döhla, E. Sib, M. Jentzsch, J. D. Borrajo, J. Strecker, J. Reinhardt, B. Cleary, M. Geyer, M. Hölzel, R. Macrae, M. M. Nöthen, P. Hoffmann, M. Exner, A. Regev, F. Zhang, J. L. Schmid-Burgk, LAMP-Seq enables sensitive, multiplexed COVID-19 diagnostics using molecular barcoding. *Nat. Biotechnol.* (2021).
- M.-M. Aynaud, J. J. Hernandez, S. Barutcu, U. Braunschweig, K. Chan, J. D. Pearson, D. Trcka, S. L. Prosser, J. Kim, M. Barrios-Rodiles, M. Jen, S. Song, J. Shen, C. Bruce, B. Hazlett, S. Poutanen, L. Attisano, R. Bremner, B. J. Blencowe, T. Mazzulli, H. Han, L. Pelletier, J. L. Wrana, A. Multiplexed, next generation sequencing platform for high-throughput detection of SARS-CoV-2. *Nat. Commun.* **12**, 1405 (2021).
- J. J. Credle, M. L. Robinson, J. Gunn, D. Monaco, B. Sie, A. Tchir, J. Hardick, X. Zheng, K. Shaw-Saliba, R. E. Rothman, S. H. Eshleman, A. Pekosz, K. Hansen, H. Mostafa, M. Steinegger, H. B. Larman, Highly multiplexed oligonucleotide probe-ligation testing enables efficient extraction-free SARS-CoV-2 detection and viral genotyping. *Mod. Pathol.* **34**, 1093–1103 (2021).
- A. S. Boeshaghi, N. B. Lubock, A. R. Cooper, S. W. Simpkins, J. S. Bloom, J. Gehring, L. Luebbert, S. Kosuri, L. Pachter, Reliable and accurate diagnostics from highly multiplexed sequencing assays. *Sci. Rep.* **10**, 21759 (2020).
- Q. Wu, C. Suo, T. Brown, T. Wang, S. A. Teichmann, A. R. Bassett, INSIGHT: A population-scale COVID-19 testing strategy combining point-of-care diagnosis with centralized high-throughput sequencing. *Sci. Adv.* **7**, (2021).
- D. Palmieri, J. Siddiqui, A. Gardner, R. Fishel, W. O. Miles, REMBRANDT: A high-throughput barcoded sequencing approach for COVID-19 screening. *bioRxiv* 2020.05.16.099747 [Preprint]. 17 May 2020. <https://doi.org/10.1101/2020.05.16.099747>.
- E. Yánguez, G. White, S. Kreutzer, L. Opitz, L. Poveda, T. Sykes, M. D. Moccia, C. Aquino, R. Schlapbach, HiDRA-seq: High-throughput SARS-CoV-2 detection by RNA barcoding and amplicon sequencing. *bioRxiv* 2020.06.02.130484 [Preprint]. 2 June 2020. <https://doi.org/10.1101/2020.06.02.130484>.
- D. Miller, M. A. Martin, N. Harel, O. Tirosh, T. Kustin, M. Meir, N. Sorek, S. Gefen-Halevi, S. Amit, O. Vorontsov, A. Shaag, D. Wolf, A. Peretz, Y. Shemer-Avni, D. Roif-Kaminsky, N. M. Kopelman, A. Huppert, K. Koelle, A. Stern, Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat. Commun.* **11**, 5518 (2020).
- M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in Europe and the US. *bioRxiv* 2020.05.21.109322 [Preprint]. 23 May 2020. <https://doi.org/10.1101/2020.05.21.109322>.
- P. Bost, A. Giladi, Y. Liu, Y. Bendjelal, G. Xu, E. David, R. Blecher-Gonen, M. Cohen, C. Medaglia, H. Li, A. Deczkowska, S. Zhang, B. Schwikowski, Z. Zhang, I. Amit, Host-viral infection maps reveal signatures of severe COVID-19 patients. *Cell* **181**, 1475–1488.e12 (2020).
- B. Udugama, P. Kadhiresan, H. N. Kozłowski, A. Malekjahani, M. Osborne, V. Y. C. Li, H. Chen, S. Mubareka, J. B. Gubbay, W. C. W. Chan, Diagnosing COVID-19: The disease and tools for detection. *ACS Nano* **14**, 3822–3835 (2020).
- A. Rahat, M. Adam, U. Shabi, M. Cohen, D. Kitsberg, M. Nissim, H. Turm, A. Klochendler, D. Joseph-Strauss, I. Sharkia, M. Lotem, G. Fialkoff, R. Sadeh, A. Chappleboim, Y. Dor, N. Friedman, D. Wolf, N. Habib, Sars-CoV-2 RNA purification with homemade SPRI beads for RT-qPCR test v1 (protocols.io.beswjefe), protocols.io, doi:10.17504/protocols.io.beswjefe.
- V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. W. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, D. G. Mulders, B. L. Haagmans, B. van der Veer, S. van den Brink, L. Wijsman, G. Goderski, J.-L. Romette, J. Ellis, M. Zambon, N. Peiris, H. Goossens, C. Reusken, M. P. G. Koopmans, C. Drosten, Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill.* **25**, 2000045 (2020).
- T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011).
- R. Farouni, H. Djambazian, L. E. Ferri, J. Ragoussis, H. S. Najafabadi, Model-based analysis of sample index hopping reveals its widespread artifacts in multiplexed single-cell RNA-sequencing. *Nat. Commun.* **11**, 2704 (2020).
- M. Costello, M. Fleharty, J. Abreu, Y. Farjoun, S. Ferreria, L. Holmes, B. Granger, L. Green, T. Howd, T. Mason, G. Vicente, M. Dasilva, W. Brodeur, T. DeSmet, S. Dodge, N. J. Lennon, S. Gabriel, Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332 (2018).
- Y. Onda, K. Takahagi, M. Shimizu, K. Inoue, K. Mochida, Multiplex PCR targeted amplicon sequencing (MTA-Seq): Simple, flexible, and versatile SNP genotyping by highly multiplexed PCR amplicon sequencing. *Front. Plant Sci.* **9**, 201 (2018).
- X. Lu, L. Wang, S. K. Sakthivel, B. Whitaker, J. Murray, S. Kamili, B. Lynch, L. Malapati, S. A. Burke, J. Harcourt, A. Tamin, N. J. Thornburg, J. M. Villanueva, S. Lindstrom, US CDC real-time reverse transcription PCR panel for detection of severe acute respiratory syndrome coronavirus 2. *Emerg. Infect. Dis.* **26**, 1654–1665 (2020).
- Y. Finkel, O. Mizrahi, A. Nachshon, S. Weingarten-Gabbay, D. Morgenstern, Y. Yahalom-Ronen, H. Tamir, H. Achdout, D. Stein, O. Israeli, A. Beth-Din, S. Melamed, S. Weiss, T. Israely, N. Paran, M. Schwartz, N. Stern-Ginossar, The coding capacity of SARS-CoV-2. *Nature* **589**, 125–130 (2021).
- C. B. F. Vogels, A. F. Brito, A. L. Wyllie, J. R. Fauver, I. M. Ott, C. C. Kalinich, M. E. Petrone, A. Casanovas-Massana, M. C. Muenker, A. J. Moore, J. Klein, P. Lu, A. Lu-Culligan, X. Jiang, D. J. Kim, E. Kudo, T. Mao, M. Moriyama, J. E. Oh, A. Park, J. Silva, E. Song, T. Takahashi, M. Taura, M. Tokuyama, A. Venkataraman, O.-E. Weizman, P. Wong, Y. Yang, N. R. Cheemarla, E. B. White, S. Lapidus, R. Earnest, B. Geng, P. Vijayakumar, C. Odio, J. Fournier, S. Bermejo, S. Farhadian, C. S. Dela Cruz, A. Iwasaki, A. I. Ko, M. L. Landry, E. F. Foxman, N. D. Grubaugh, Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT-qPCR primer-probe sets. *Nat. Microbiol.* **5**, 1299–1305 (2020).

30. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
31. FDA emergency use validation for BGI real-time fluorescent RT-PCR kit for detecting SARS-CoV-2 (2020); www.fda.gov/media/136472/download.
32. E. Krakoff, R. B. Gagne, S. VandeWoude, S. Carver, Variation in intra-individual lentiviral evolution rates: A systematic review of human, nonhuman primate, and felid species. *J. Virol.* **93**, e00538-19 (2019).
33. K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M. Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, R. Williams, G. Vernet, A. Justice, A. Green, S. M. Nicholls, M. A. Ansari, L. Abeler-Dörner, C. E. Moore, T. E. A. Peto, D. W. Eyre, R. Shaw, P. Simmonds, D. Buck, J. A. Todd; Oxford Virus Sequencing Analysis Group (OVSG), T. R. Connor, S. Ashraf, A. da Silva Filipe, J. Shepherd, E. C. Thomson; COVID-19 Genomics UK (COG-UK) Consortium, D. Bonsall, C. Fraser, T. Golubchik, SARS-CoV-2 within-host diversity and transmission. *Science* **372**, eabg0821 (2021).
34. J. R. Tyson, P. James, D. Stoddart, N. Sparks, A. Wickenhagen, G. Hall, J. H. Choi, H. Lapointe, K. Kamelian, A. D. Smith, N. Prystajecy, I. Goodfellow, S. J. Wilson, R. Harrigan, T. P. Snutch, N. J. Loman, J. Quick, Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv* 2020.09.04.283077 [Preprint]. 4 September 2020. <https://doi.org/10.1101/2020.09.04.283077>.
35. D. M. Gohl, J. Garbe, P. Grady, J. Daniel, R. H. B. Watson, B. Auch, A. Nelson, S. Yohe, K. B. Beckman, A rapid, cost-effective tailed amplicon method for sequencing SARS-CoV-2. *BMC Genomics* **21**, 863 (2020).
36. H. Tegally, E. Wilkinson, M. Giovanetti, A. Iranzadeh, V. Fonseca, J. Giandhari, D. Doolabh, S. Pillay, E. J. San, N. Msomi, K. Mlisana, A. von Gottberg, S. Walaza, M. Allam, A. Ismail, T. Mohale, A. J. Glass, S. Engelbrecht, G. Van Zyl, W. Preiser, F. Petruccione, A. Sigal, D. Hardie, G. Marais, N.-Y. Hsiao, S. Korsman, M.-A. Davies, L. Tyers, I. Mudau, D. York, C. Maslo, D. Goedhals, S. Abrahams, O. Laguda-Akingba, A. Alisoltani-Dehkordi, A. Godzik, C. K. Wibmer, B. T. Sewell, J. Lourenço, L. C. J. Alcantara, S. L. Kosakovsky Pond, S. Weaver, D. Martin, R. J. Lessells, J. N. Bhiman, C. Williamson, T. de Oliveira, Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
37. J. W. Tang, P. A. Tambyah, D. S. Hui, Emergence of a new SARS-CoV-2 variant in the UK. *J. Infect.* **82**, e27–e28 (2021).
38. Y. Wang, D. Wang, L. Zhang, W. Sun, Z. Zhang, W. Chen, A. Zhu, Y. Huang, F. Xiao, J. Yao, M. Gan, F. Li, L. Luo, X. Huang, Y. Zhang, S.-S. Wong, X. Cheng, J. Ji, Z. Ou, M. Xiao, M. Li, J. Li, P. Ren, Z. Deng, H. Zhong, X. Xu, T. Song, C. K. P. Mok, M. Peiris, N. Zhong, J. Zhao, Y. Li, J. Li, J. Zhao, Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. *Genome Med.* **13**, 30 (2021).
39. A. L. Valesano, K. E. Rumfeldt, D. E. Dimcheff, C. N. Blair, W. J. Fitzsimmons, J. G. Petrie, E. T. Martin, A. S. Lauring, Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLoS Pathog.* **17**, e1009499 (2021).
40. R. P. McNamara, C. Caro-Vegas, J. T. Landis, R. Moorad, L. J. Pluta, A. B. Eason, C. Thompson, A. Bailey, F. C. S. Villamor, P. T. Lange, J. P. Wong, T. Seltzer, J. Seltzer, Y. Zhou, W. Vahrson, A. Juarez, J. O. Meyo, T. Calabre, G. Broussard, R. Rivera-Soto, D. L. Chappell, R. S. Baric, B. Damania, M. B. Miller, D. P. Dittmer, High-density amplicon sequencing identifies community spread and ongoing evolution of SARS-CoV-2 in the southern United States. *Cell Rep.* **33**, 108352 (2020).
41. B. B. O. Munnink, D. F. Nieuwenhuijse, M. Stein, Á. O'Toole, M. Haverkate, M. Mollers, S. K. Kamga, C. Schapendonk, M. Pronk, P. Lexmond, A. van der Linden, T. Bestebroer, I. Chestakova, R. J. Overmars, S. van Nieuwkoop, R. Molenkamp, A. A. van der Eijk, C. GeurtsvanKessel, H. Vennema, A. Meijer, A. Rambaut, J. van Dissel, R. S. Sikkema, A. Timen, M. Koopmans; Dutch-Covid-19 response team, Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat. Med.* **26**, 1405–1410 (2020).
42. D. Butler, C. Mozsary, C. Meydan, J. Foox, J. Rosiene, A. Shaiber, D. Danko, E. Afshinnekoo, M. MacKay, F. J. Sedlazeck, N. A. Ivanov, M. Sierra, D. Pohle, M. Zietz, U. Gisladdottir, V. Ramlall, E. T. Scholle, E. J. Schenck, C. D. Westover, C. Hassan, K. Ryon, B. Young, C. Bhattacharya, D. L. Ng, A. C. Granados, Y. A. Santos, V. Servellita, S. Federman, P. Ruggiero, A. Fungtammasan, C.-S. Chin, N. M. Pearson, B. W. Langhorst, N. A. Tanner, Y. Kim, J. W. Reeves, T. D. Hether, S. E. Warren, M. Bailey, J. Gawrys, D. Meleshko, D. Xu, M. Couto-Rodriguez, D. Nagy-Szakal, J. Barrows, H. Wells, N. B. O'Hara, J. A. Rosenfeld, Y. Chen, P. A. D. Steel, A. J. Shemesh, J. Xiang, J. Thierry-Mieg, D. Thierry-Mieg, A. Iftner, D. Bezdán, E. Sanchez, T. R. Campion Jr., J. Siple, L. Cong, A. Craney, P. Velu, A. M. Melnick, S. Shapira, I. Hajirasouliha, A. Borczuk, T. Iftner, M. Salvatore, M. Loda, L. F. Westblade, M. Cushing, S. Wu, S. Levy, C. Chiu, R. E. Schwartz, N. Tatonetti, H. Rennett, M. Imielski, C. E. Mason, Shotgun transcriptome, spatial omics, and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification, and drug interactions. *Nat. Commun.* **12**, 1660 (2021).
43. T. P. Loka, S. H. Tausch, B. Y. Renard, Reliable variant calling during runtime of Illumina sequencing. *Sci. Rep.* **9**, 16502 (2019).
44. J. D. Buenostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
45. D. Joseph-Strauss, A. Rahat, I. Sharkia, A. Chappleboim, M. Adam, D. Kitsberg, G. Fialkoff, M. Lotem, O. Gershon, A. Kristina, E. Oiknine, A. Klochender, R. Sadeh, Y. Dor, D. Wolf, N. Habib, N. Friedman, SARS-CoV-2 detection with AphaSeq v2 (protocols.io.bjgukjww), protocols.io, doi:10.17504/protocols.io.bjgukjww.
46. E. E. Karrer, J. E. Lincoln, S. Hogenhout, A. B. Bennett, R. M. Bostock, B. Martineau, W. J. Lucas, D. G. Gilchrist, D. Alexander, In situ isolation of mRNA from individual plant cells: Creation of cell-specific cDNA libraries. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3814–3818 (1995).
47. A. Fungtammasan, M. Tomaszewicz, R. Campos-Sánchez, K. A. Eckert, M. DeGiorgio, K. D. Makova, Reverse transcription errors and RNA-DNA differences at short tandem repeats. *Mol. Biol. Evol.* **33**, 2744–2758 (2016).
48. N. S. Zuckerman, S. Fleishon, E. Bucris, D. Bar-Ilan, M. Linal, I. Bar-Or, V. Indenbaum, M. Weil, Y. Lustig, E. Mendelson, M. Mandelboim, O. Mor, N. Zuckerman; Israel National Consortium For Sars-CoV-Sequencing, A unique SARS-CoV-2 spike protein P681H variant detected in Israel. *Vaccines* **9**, 616 (2021).
49. Z. Liu, L. A. VanBargan, L.-M. Bloyet, P. W. Rothlauf, R. E. Chen, S. Stumpf, H. Zhao, J. M. Errico, E. S. Theel, M. J. Liebeskind, B. Alford, W. J. Buchser, A. H. Ellebedy, D. H. Fremont, M. S. Diamond, S. P. J. Whelan, Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* **29**, 477–488.e4 (2021).
50. J. Aslanzadeh, Preventing PCR amplification carryover contamination in a clinical laboratory. *Ann. Clin. Lab. Sci.* **34**, 389–396 (2004).
51. M. Johnson, PCR machines. *Mater. Methods* **3**, 193 (2013).

Acknowledgments: We thank M. Rabani for critical comments and support. We thank M. Bronstein, A. Nasereddin, I. Shiff, A. Turjeman, N. Barak, and M. Yassour for their help.

Funding: This work was supported, in part, by the Rothschild Foundation. A.C. is an Azrieli scholar and would like to thank the Azrieli Foundation for their support.

Author contributions: A.C., A.R., D.J.-S., R.S., N.H., and N.F. designed the assay. A.R., D.J.-S., and A.C. conducted and supervised all of the experiments with the following help: N.H., M.A., D.K., A.K., and A.-K.S. headed the SPRI-based branch of the experiments (Fig. 2A and fig. S1), and I.S., M.L., and O.G. conducted multiple calibration assays (figs. S2 and S3). E.O.-D. and D.W. provided samples and materials for the clinical validation of the protocol. A.C. performed the statistical and computational analysis with help from G.F. and N.F. A.C. wrote the manuscript with contributions from all coauthors. Y.D., N.H., and N.F. secured funding for this work. **Competing interests:** A patent application titled “Methods for nucleic acid detection” (PCT/IL2021/050400) that describes hybridization-based nucleic acid detection has been submitted by the Hebrew University of Jerusalem and by Hadassah Hebrew University Medical Center. R.S., I.S., and N.F. are founders of Senseera. **Data and materials availability:** All data and code to generate figures, summary data, and statistical tests in this study were uploaded to Zenodo and are available at 10.5281/zenodo.5069979.

Submitted 28 April 2021

Accepted 21 September 2021

Published First Release 30 September 2021

Published 3 November 2021

10.1126/scitranslmed.abj2266