Scalable search of massively pooled nucleic acid samples enabled by a molecular database query language

Joseph D. Berleant[1,†], James L. Banal[1,4,†], Dhriti K. Rao[3], and Mark Bathe[1,2*]
[1]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02139 USA
[3]University of Cambridge, Cambridge UK
[4] Present address: Cache DNA, Inc. 733 Industrial Rd., San Carlos, CA 94070 USA

[†]These authors contributed equally to this work.

*Correspondence should be addressed to mark.bathe@mit.edu

ABSTRACT

The surge in nucleic acid analytics requires scalable storage and retrieval systems akin to electronic databases used to organize digital data. Such a system could transform disease diagnosis, ecological preservation, and molecular surveillance of biothreats. Current storage systems use individual containers for nucleic acid samples, requiring single-sample retrieval that falls short compared with digital databases that allow complex and combinatorial data retrieval on aggregated data. Here, we leverage protective microcapsules with combinatorial DNA labeling that enables arbitrary retrieval on pooled biosamples analogous to Structured Query Languages. Ninety-six encapsulated pooled mock SARS-CoV-2 genomic samples barcoded with patient metadata are used to demonstrate queries with simultaneous matches to sample collection date ranges, locations, and patient health statuses, illustrating how such flexible queries can be used to yield immunological or epidemiological insights. The approach applies to any biosample database labeled with orthogonal barcodes, enabling complex post-hoc analysis, for example, to study global biothreat epidemiology.

INTRODUCTION

The archival large-scale storage of DNA and RNA is essential to numerous biotechnological areas including population-scale disease tracking(*1*), precision medicine(*2–4*), forensics(*5*), and worldwide ecological record-keeping(*6*, *7*). For example, datasets of nucleic acids collected from

large patient cohorts can link DNA and RNA biomarkers to disease variants and onset likelihood, and subsequently determine personalized treatment plans via *in silico* statistical analysis and *in vitro* molecular sensors(*4*, *8–12*). In general, raw nucleic acid samples provide the most complete representation of genomic information, including epigenetic marks. While low-cost molecular diagnostics such as polymerase chain reaction (PCR) provide the means to identify nucleic acids, these tests only survey a small fragment of the entire genome. Alternatively, sequencing is an important tool for systematically collecting nucleic acids samples from the environment at the point-of-collection or after short-term storage(*13–15*). However, the high cost of sequencing means that the number of samples one can fully characterize is likely to remain only a small fraction of the number of samples one can collect. The existence of low-probability yet catastrophic events, such as the COVID-19 pandemic, heightens the importance of low-cost, large-scale archiving of physical samples of nucleic acids to create a biosample database without immediate characterization to allow for retrospective analyses like understanding the emergence of novel pathogens or pathogen variants. However, building a scalable and automated infrastructure to store nucleic acids in native form to support multiple sequencing modalities, yet also allow for direct access to nucleic acids, presents significant logistical, technological, and economic challenges.

Biosamples, stored in barcoded tubes within freezers or liquid nitrogen tanks, face scalability issues as biosample collections reach into the millions. For these vast biosample databases of nucleic acid samples, sophisticated systems are required for biosample search and retrieval to be efficient. Automated systems notwithstanding, the effective utility of these biosample databases is hampered by their low storage density and the inadequate parallelization of sample access. In contrast, molecular biology approaches enable the pooling of millions to billions of unique DNA biosamples per tube, effectively creating a highly dense biosample database. Retrieval of specific samples can be achieved using biochemical approaches, such as PCR(*16*), magnetic pulldown(*17*), or fluorescence-activated sorting (FAS) (*18*, *19*) of molecular labels, such as primers, affinity tags, or fluorescent dyes. Unlike conventional manual search of individual tubes, biochemical approaches expedite retrieval through execution of millions to billions of concurrent molecular search and retrieval operations in solution. Yet even with these biochemical approaches, previously demonstrated sample retrieval using molecular labels representing categorical metadata are elementary when compared with the query abilities of digital databases.

Within a biosample database that employs biochemical querying, the categorization of each biosample is facilitated by the assignment of biochemical identifiers. These identifiers act as an

organization system that facilitates the retrieval process, whether for individual biosamples or subsets of related biosamples. Previous implementations within these databases have employed molecular labels, such as primers and DNA barcodes. These molecular labels provide a method of identification that capitalizes on the innate specificity of DNA hybridization, for example in polymerase chain reaction(*16*), magnetic pulldown(*20*), and FAS(*18*, *19*), thus ensuring the precision of retrieval. To aid the retrieval of sets of related biosamples, molecular labels can encode information about the encapsulated nucleic acid samples, such as Boolean metadata. For example, we previously demonstrated Boolean retrieval using a pool of twenty distinct samples storing images encoded in plasmid DNA and a biosample database containing SARS-CoV-2 RNA, human-derived total RNA, methicillin-resistant *S. aureus* genome, and bovine whole genome. In these examples, encapsulation of plasmid DNA, genomic DNA, and RNA in silica microparticles allowed for the long-term protection and room temperature storage as well as physical separation between the nucleic acid biosamples and DNA barcodes that represent metadata about the underlying encapsulated nucleic acid, such as cats, dogs, and buildings(*19*) or Animalia and Eukarya. This physical separation is integral to the microcapsule architecture, enabling the precise identification and selection of specific nucleic acids or groups thereof from the biosample database through the use of fluorescently-labeled DNA probes. These probes selectively bind to the DNA barcodes corresponding to the nucleic acid samples of interest, facilitating their retrieval via FAS. Despite the precision of these biochemical methodologies in biosample identification and retrieval, they have a limited scope of query expressiveness, especially when compared to the broader querying capabilities afforded by basic digital databases.

While DNA barcodes are crucial for efficiently locating and differentiating between biosamples in a pool, this molecular labeling approach faces a scalability challenge. The number of available DNA barcodes is limited to only about 240,000 25-nt unique sequences(*21*) carefully designed and validated to be orthogonal from one another and to maintain consistent melting temperatures for reliability in biochemical reactions. This limited number of barcodes is insufficient for the vast array of biosamples needing identification, creating a bottleneck for various research applications that require unique labeling for each sample. Furthermore, the current DNA barcoding approach struggles with handling numerical metadata, for example the collection dates of samples. While a DNA barcode can represent a single date(*18*), efficiently querying for samples collected within a specific date range, or before or after a certain date, is not feasible with this system. Each date query would require individually checking multiple barcodes, a process that is both time-

consuming and impractical. In contrast, digital databases, exemplified by Project Gutenberg or Google Books, can effortlessly manage queries for items within specific date ranges, exemplifying a flexibility and efficiency that biosample databases currently lack.

Here, we develop a scalable biosample database that enables queries approaching the expressiveness of modern digital database query languages, permitting metadata queries for numerical ranges, such as ranges of dates or ages; categories, such as cities or countries; and Boolean classifications (i.e., either true or false). We apply the approach to a database of synthetic DNA barcodes and RNA from SARS-CoV-2 encapsulated in silica microcapsules, with each microcapsule labeled with single-stranded DNA barcodes, representing metadata features, attached to the microcapsule surface(*19*, *22*). We choose an example application of SARS-CoV-2 sequencing to demonstrate the expressiveness of queries of our barcoding approach, for which we simulate a scenario in which airplane passengers entering a major international airport are comprehensively swabbed for SARS-CoV-2 to record the ongoing state of the pandemic (**Fig. 1A**). Although immediate sequencing of all samples for the presence of SARS-CoV-2 would be impractical, samples may be stored archivally in microcapsules labeled with patient information such as age, health status, and flight information. For demonstration purposes, we synthesized a database of 96 mock SARS-CoV-2 samples, each labeled with up to 14 unique barcodes, which is almost a five-fold increase over previous molecular data storage implementations in the number of barcodes(*17*, *19*, *22–24*). Using this test database, we considered retrospective analyses of several interesting epidemiological and immunological questions via queries of increasing complexity, including queries for a particular health status, for three distinct age ranges, and for samples simultaneously matching date range, location, and health status criteria. These examples illustrate a query language with many of the features of digital database query languages like SQL, permitting arbitrary logical expressions composed of numerical range, categorical, and Boolean metadata criteria. Silica encapsulation also enables long-term sample preservation at room temperature, obviating the need for refrigeration and cold-chain logistics. Thus, we demonstrate a general-purpose query language for scalable molecular databases that approaches the expressiveness of modern digital databases (**Fig. 1B**). This framework offers broad applicability to molecular diagnostics, pathogen surveillance, and ecological preservation to address the bottleneck in nucleic acid sample storage and retrieval capabilities.

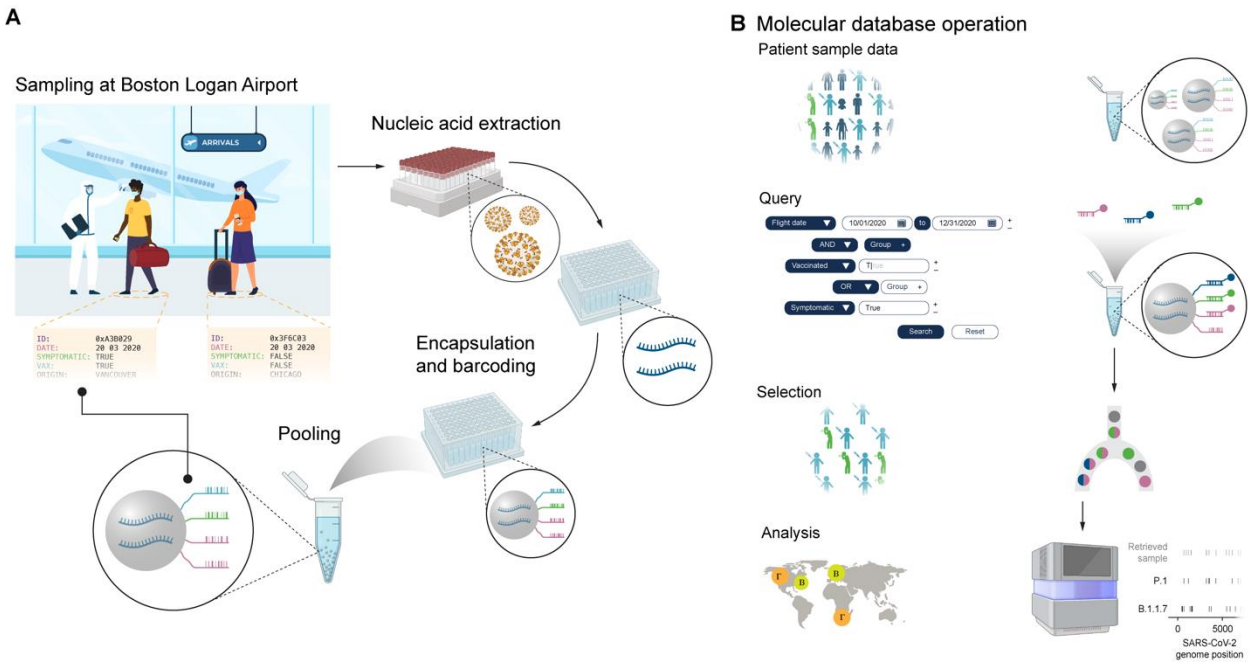

Fig. 1. Application of a molecular database to simulated SARS-CoV-2 tracking. (**A**) Simulated scenario of sample collection at Boston Logan airport with subsequent pooling and nucleic acid extraction, encapsulation, and barcoding of samples using our proposed molecular filesystem. (**B**) Workflow for querying and analyzing samples within a molecular database, shown side-by-side with generic database operations.

Database construction

Ninety-six mock "patient samples" were separately encapsulated in 5 µm silica microcapsules (**Fig. 1A**). These microcapsules simulated samples drawn from passengers flying into Logan International Airport in Boston, Massachusetts, and subsequently stored archivally for future diagnostic or epidemiological testing, if needed. Each sample consisted of at most one variant of the SARS-CoV-2 genome using either synthetic P.1 and B.1.1.7 variants and a unique internal 85-nt barcode whose purpose was to aid validation of microcapsule retrieval by making possible downstream identification of de-encapsulated microcapsules in a pool. Variants were quasi-randomly assigned to each sample, with approximately a 15% chance of some variant being present. Encapsulation followed the procedure previously used (see Methods)(*19*, *22*). In contrast to our previously encapsulated nucleic acids(*19*, *22*), the DNA encapsulated here consisted of linear DNA of differing lengths: short internal barcodes of 85-nt that represented unique identification barcodes for each "patient" and 5-kb fragments of long genomic synthetic SARS-CoV-2 RNA. However, because the encapsulation process relied only on charge interactions between the negatively charged phosphate backbone of nucleic acids, we expected successful incorporation of all nucleic acids irrespective of length (**fig. s1**), as previously demonstrated(*22*).

We validated nucleic acid encapsulation and de-encapsulation with approximately 1 copy of SARS-CoV-2 genome and ~10^8 copies of the internal barcode per microcapsule.

Database labeling and querying

The ideal SARS-CoV-2 genomic database would label each sample with metadata features, such as its unique identifier, patient health status, sample acquisition date, and flight origin. For our example use case, several features were chosen to describe a variety of relevant metadata, including patient age and month and year of the arrival flight, which are examples of numerical metadata; vaccination status, which is an example of Boolean metadata; and flight number and city of origin, which are examples of categorical metadata. For each microcapsule, each feature value was encoded into a set of barcodes to be displayed on its exterior (**Fig. 1A**), with a total of either 13 or 14 unique 25-nt ssDNA barcodes per microcapsule (see Methods). Differing encoding strategies were employed based on the type of metadata (numerical, Boolean, or categorical), which enabled type-specific queries of each feature, such as queries for numerical ranges. More detail about each encoding and retrieval strategy is given in Methods.

For numerical metadata, we encoded each numerical value as a sequence of digits in a mixed-radix representation, corresponding to a sequence of barcodes with one barcode per digit. Queries for an exact numerical match are performed by probing for every barcode in its representation, while ranges of various sizes are possible by omitting one or more of the less significant digits (**Fig. 2**). Categorical metadata were encoded using distinct combinations of *k* barcodes for each possible value, so that the value of *k* determined the number of barcodes required to identify a sample's specific feature value. This approach scales to accommodate an extensive number of feature values. Boolean metadata, on the other hand, were straightforward, with the presence or absence of a single barcode indicating the feature's state.

When performing a query, the search string is first translated into a DNA barcode list (**Fig. 2**). This list not only facilitates the query but also informs the selection of unique dyes needed for FAS. For NOT logic, the barcode that denotes the unwanted condition is tagged with a fluorescent dye. During sorting, microcapsules that do not show this fluorescence are selected, thus excluding the condition indicated by the dye. OR logic employs a single dye for all relevant barcodes, selecting samples with the matching dye. For AND logic, each query barcode is tagged with a distinct dye, and only samples displaying all unique dyes are selected. This intricate logic necessitates a

careful selection of dyes and strategic grouping of search parameters, ensuring precise sample retrieval.

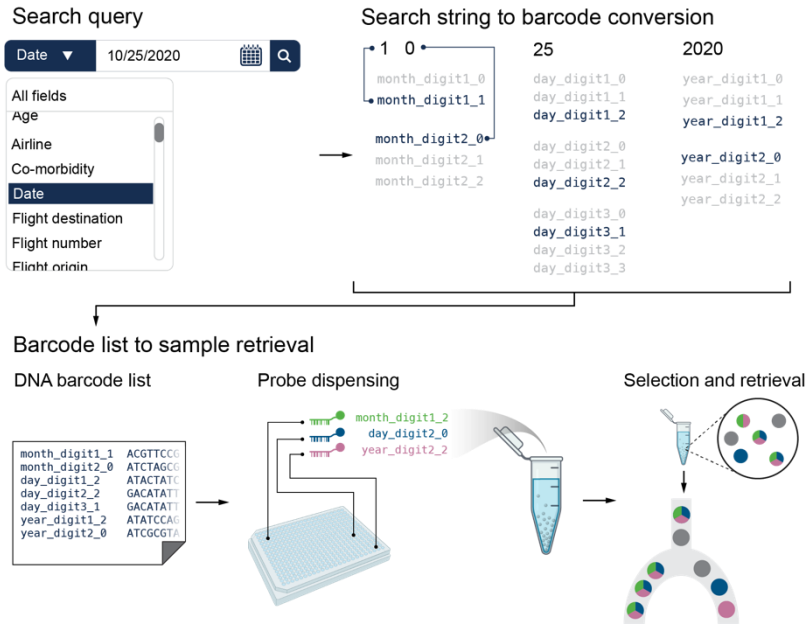

Fig. 2. Conversion of search query to operations on a molecular database. Numerical, Boolean, and categorical metadata are converted into a corresponding logical expression comprising AND, OR, NOT operations on several barcodes. This logical expression then guides the dye-labeling strategy, search grouping, and dispensing of dye-labeled DNA barcodes for sample selection. Selected samples are then retrieved using optical sorting. In this schematic, a specific date is represented by a total of seven barcodes (two for the month, three for the day, and two for the last two digits of the year). Here, month and year are represented in base 10 while the day is represented in a combination of bases 3 and 4. Ranges of contiguous dates are specified by omitting one or more barcodes.

Immunological case study

To demonstrate the application of a large-scale molecular database with advanced search query capabilities, we performed several database queries on the synthetic database of 96 SARS-CoV-2 samples hypothetically collected from airplane passengers entering Boston, MA. We designed search queries to demonstrate the breadth of the queries enabled by the sample labeling approach, as well as to show how an actual database of this type could be used to answer valuable retrospective epidemiological and immunological questions.

We began with the immunological question of whether specific SARS-CoV-2 variants were present in asymptomatic passengers. To answer this question, we used a query for when the

Boolean feature “symptomatic” was false. For each sample, the barcode “bc_symptomatic” was present when this feature was true, and absent when this feature was false. Thus, our query “NOT symptomatic” should retrieve exactly those samples not displaying this barcode.

The presence of the barcode “bc_symptomatic” on each microcapsule was determined using a fluorescent probe combined with fluorescence-associated sorting (FAS). However, in comparison to previous work using FAS for barcode detection on microcapsules(*19*, *22*), the number of distinct barcodes attached per microcapsule was increased from 3 to up to 14, leading to a proportional decrease in the copy number of each barcode per microcapsule. To compensate for this copy number reduction, we used fluorescence amplifiers routinely used to detect low copy numbers of RNA in cells in flow cytometry(*25*) (**Fig. 3A**). The pool of microcapsules was mixed with this amplifier probe and subsequently passed through a FAS instrument, which generates a stream of droplets. Fluorescence from each droplet was measured, and selected populations were defined using fluorescence intensity to distinguish microcapsules with and without the fluorescent probe attached. The low-intensity population was separated from the rest of the pool for subsequent de-encapsulation and sequencing.

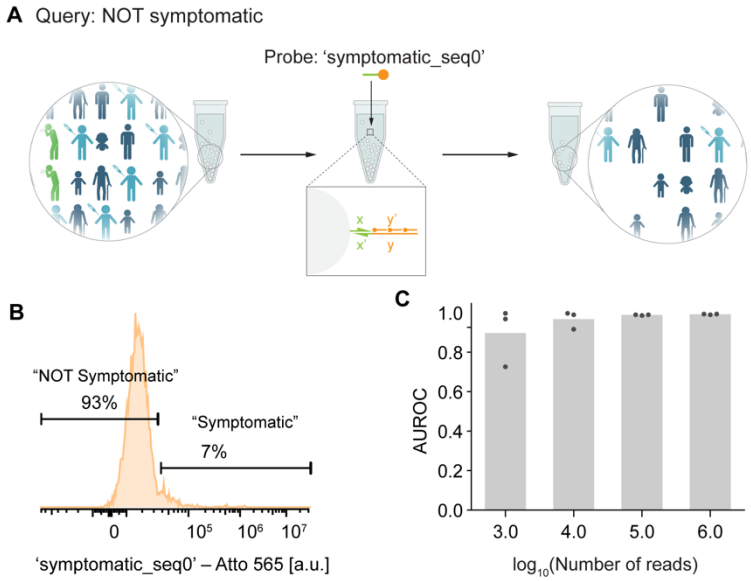

Fig. 3. Database querying results for “NOT symptomatic”. (**A**) For this query, the probe included a region complementary to barcode “bc_symptomatic” followed by a repeating sequence that allowed 3-fold fluorescence amplification. (**B**) Histogram of FAS results showing 93% of microcapsules with low fluorescence indicating absence of the “bc_symptomatic” barcode. c) The area under the ROC curve (AUC), which measures quality of a binary classifier, increases as more sequencing reads are processed. This is expected behavior as the variance in the number of reads for each sample is reduced with more reads.

To validate correct retrieval, sequencing reads of the 85-nt internal barcodes were matched to the known sequences to identify the samples in the sequencing data. The number of reads of each sample showed clear separation between the retrieved and non-retrieved microcapsules (**Fig. 3B**). To quantify the retrieval performance, receiver operator characteristic (ROC) analysis of the number of reads per sample was performed and the area under the ROC curve (AUC) computed. The AUC value is interpreted as a quality score between 0.0 and 1.0, with 1.0 indicating ideal behavior (all correct samples were present at higher levels than any incorrect samples) and 0.0 indicating maximally incorrect behavior (all incorrect samples present at higher levels than correct samples) (**Fig. 3C**). For this query, the mean AUC over 3 replicates was computed as increasing numbers of sequencing reads were processed. The AUC was high (AUC=0.99) when the number of sequencing reads surpassed 10^5, indicating enrichment of the desired populations.

Next, we demonstrated how range queries on patient age could be used to explore if certain age groups were more susceptible to different SARS-CoV-2 variants(*26*). We considered three age range queries of different size: “age = 0”, “15 ≤ age ≤ 19”, and “50 ≤ age ≤ 74”. For the narrowest query “age = 0”, we selected samples labeled with barcodes bc_age_x25_seq0, bc_age_x5_seq0, and bc_age_x1_seq0 using the same type of multi-stranded branched probes as previously described, labeled with fluorophores Atto 565, Alexa Fluor 647, and Alexa Fluor 750, respectively. For the moderate-range query “15 ≤ age ≤ 19”, we selected samples labeled with barcodes bc_age_x25_seq0 and bc_age_x5_seq3, using probes labeled with fluorophores Atto 565 and Alexa 647. For the query of broad range “50 ≤ age ≤ 74”, we selected samples labeled with the barcode bc_age_x25_seq2 using a probe labeled with fluorophore Atto 565. In each case, sequencing the internal 85-nt barcodes showed that correct samples were retrieved, as indicated by AUC values of 1.0, 1.0, and 0.9, respectively, for at least 10^5 sequencing reads (**Fig. 4**).

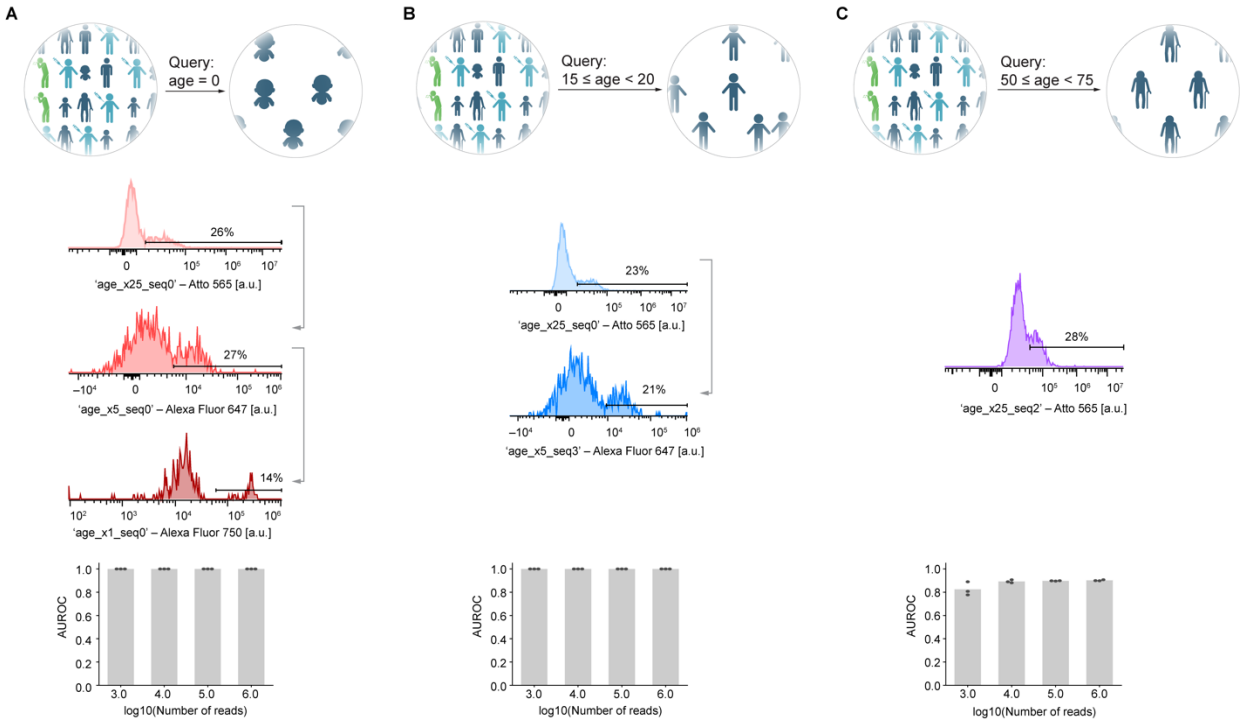

Fig. 4. Age range queries, demonstrating range queries of different size. (**A**) "age = 0", (**B**) "15 ≤ age ≤ 19", and (**C**) "50 ≤ age ≤ 74".

Epidemiological case study

We then sought to demonstrate how a composite query involving multiple metadata types can be used to understand the transmission of SARS-CoV-2 infection from specific areas, date ranges, and flight city(*27*). For the epidemiological case study, we illustrated two aspects of our database: first, efficient representation and querying of categorical features; and second, the composition of several smaller queries into arbitrarily complex logical expressions. This was performed via retrieval of all samples for passengers flying from Chicago between July and September 2020, who were either symptomatic or unvaccinated. This was equivalent to the query "(symptomatic OR NOT vaccinated) AND flight_city = Chicago AND 6 ≤ arrival_month ≤ 8 AND arrival_year = 2020", which combines queries on two numerical features, one categorical feature, and two Boolean features. Such a query necessitated examining eight barcodes: bc_vaccinated, bc_symptomatic, bc_city_seq0, bc_city_seq3, bc_city_seq4, bc_flight_month_x3_seq2, bc_flight_year_x10_seq2, and bc_flight_year_seq0, making it the most complex query tested on any molecular database to the best of our knowledge, both semantically and in terms of the number of barcodes tested. To further exhibit the flexibility of our approach to a variety of

fluorescent channels and probe design methodologies, we selected a new set of dyes for this query, which modified the bandwidth of our dye markers (**Fig. 5A**). Specifically, we transitioned from Atto 565, Alexa Fluor 647, and Alexa Fluor 750 to Atto 488, Atto 565, and Alexa Fluor 647. The reduced brightness of Atto 488 relative to the other dyes necessitated the use of branched probe designs that amplify fluorescence signals, similar to branched probed designs used to improve the relative brightness of low-copy targets in cell imaging(*28*, *29*). This strategy allowed the amplification of the net fluorescent signal by increasing the number of dye markers per barcode.

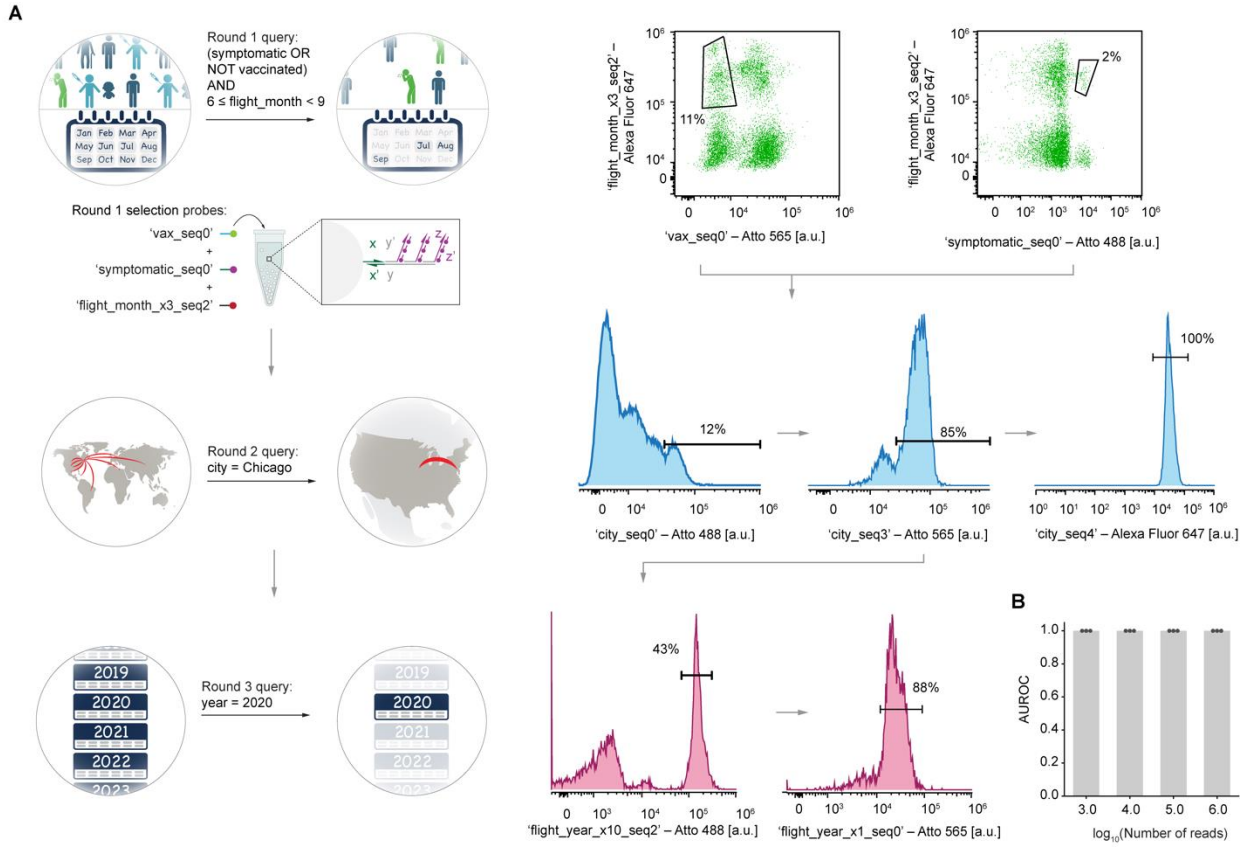

Fig. 5. "(symptomatic OR NOT vaccinated) AND flight_city = Chicago AND 6 ≤ arrival_month ≤ 8 AND arrival_year = 2020". (**A**) For this query, branched probes providing 9-fold fluorescence amplification were used. Scatterplots (top-right) from the first round of FAS illustrate the cluster separation observed. Histograms for the second and third rounds of FAS are shown (middle, bottom). (**B**) The AUROC score remains at 1.0 down to 1000 reads, indicating all correct samples had higher sequencing counts than all incorrect samples.

This query was performed over three FAS passes, using probes with fluorophores Atto 488, Atto 565, and Alexa Fluor 647 for bc_vaccinated, bc_symptomatic, and bc_flight_month_x3_seq2,

respectively, for the first pass. The sorted populations were stripped of their fluorescent probes using a denaturation buffer. For the second round of selection, fluorophores Atto 488, Atto 565, and Alexa Fluor 647 were used for selecting bc_city_seq0, bc_city_seq3, bc_city_seq4, respectively. Again, the previous probes were removed prior to subjecting the sorted population to the next round of selection. For the final FAS pass, Atto 488 and Atto 565 were used for bc_flight_year_x10_seq2 and bc_flight_year_seq0, respectively. Sequencing of the 85-nt internal barcodes after all three passes indicated correct retrieval of all samples (AUC = 1.0) (**Fig. 5B**). This demonstration illustrates two important features of our approach: first, the ability to implement a single molecular database query that describes criteria spanning many features, such as a numerical feature falling in a specified range and multiple Boolean features having the desired truth values; and second, the ability to split complex queries over several FAS passes without loss of retrieval fidelity.

SARS-CoV-2 sequencing

Identifying the dominant SARS-CoV-2 variant is essential for assessing the virulence of emerging strains, forecasting outbreaks, and expediting vaccine development(*30*). We aimed to identify the predominant SARS-CoV-2 variant in our queries, given that our samples were encapsulated with either the Alpha or Gamma variants. After sorting 100,000 to 700,000 microcapsules for each selection, we selectively sequenced samples that yielded a positive result from a specifically designed end-point tiling amplicon PCR for SARS-CoV-2(*31*) (**fig. s3**). Computational demultiplexing using Freyja(*32*) revealed that all samples showed mostly Alpha variants followed by Gamma variants and other variants that were not assigned by Freyja (**Fig. 6**). In all cases, the ratio of expected Alpha to Gamma abundance: 8:3 for selections from **Fig. 3**, 2:1 for selections from **Fig. 4C**, and 1:0 for selections from **Fig. 5**, closely matched the measured abundances in **Fig. 6**, further providing support to the high retrieval precision observed using an orthogonal sequencing approach.

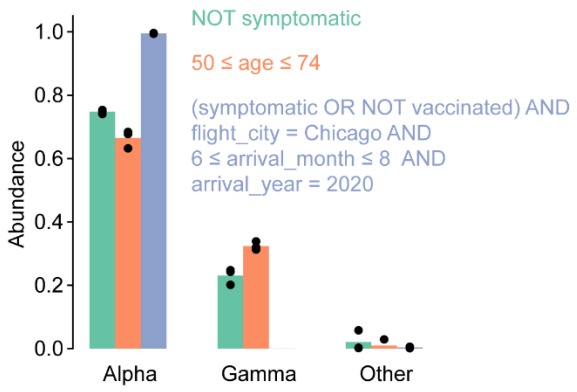

Fig. 6. Sequencing results SARS-CoV-2 samples. For the three database queries "NOT symptomatic", "50 ≤ age ≤ 74", and "(symptomatic OR NOT vaccinated) AND flight_city = Chicago AND 6 ≤ arrival_month ≤ 8 AND arrival_year = 2020", the variants in each sample were quantified. These variants should correspond to the synthetic Alpha or Gamma that had been encapsulated in each of the 96 mock patient samples.

DISCUSSION AND CONCLUSION

The large-scale, global collection of genomic DNA and RNA across health and security sectors ranging from oncology and reproductive health to pathogen surveillance calls for scalable and efficient storage and retrieval capabilities. In this work, we demonstrated an intuitive yet powerful sample labeling strategy that significantly expands querying capabilities within a pooled molecular database, rendering it analogous to common digital file databases such as public datasets hosted by Google BigQuery, Microsoft Azure, and Amazon Web Services(*33–35*). The molecular database query language can accommodate arbitrary logical combinations of ranged queries, categorical queries, and truth queries on features that have been encoded into barcodes displayed on each nucleic acid specimen, encapsulated within a microcapsule for long-term stability. In our demonstration of the search query language, we showed how the molecular database could be used to answer several examples of retrospective epidemiological and immunological questions by analyzing sequencing results from cohorts retrieved from a database of simulated SARS-CoV-2 samples.

Our demonstration used a database of 96 samples with up to 14 barcodes per microcapsule, which allowed all 96 samples to be consolidated into a single tube resulting in a hundredfold reduction in storage space while maintaining a broad range of sophisticated search capabilities, including combinations of numerical range, categorical, and Boolean queries. This scalability and efficiency in sample management have direct implications for enhancing biosurveillance strategies as proposed by the Nucleic Acid Observatory(*36*). By enabling the pooling of thousands

or millions of samples into fewer tubes, our approach offers a significant footprint reduction over methods that require a separate vessel for each sample, thereby streamlining the process of monitoring and responding to pathogenic threats. Consider the traditional storage of 1 million nucleic acid samples. Storing these samples in –80°C freezers, each accommodating 40,000 samples, would require 25 freezers. In contrast, the microcapsule architecture, capable of retrieval from a pool of 10,000 distinct nucleic acids in just one tube at room temperature as we have previously shown(*19*), would only require 100 tubes for storage. An entire tube can be queried in around 15 minutes, assuming a FAS rate of 1000 microcapsules per second and a redundancy of 100 microcapsules per sample. This streamlined approach not only offers a potential solution to the challenges outlined by the Nucleic Acid Observatory in deploying biosurveillance approaches but also underscores the potential for our technology to facilitate rapid and efficient global health responses.

The approach demonstrated in this manuscript represents a significant advance in expanding our ability to store, organize, and access nucleic acids, implementing the ability to perform the search functions that are essential for modern digital databases. Work remains to further refine the retrieval processes necessary to make such a database useful; however, this demonstration shows the feasibility and utility of building a large-scale nucleic acid database with expressive search query abilities. Should the integration of automation and microfluidic technologies, along with the advent of alternative encapsulation chemistries beyond silica, streamline the encapsulation and barcoding processes, we foresee the inclusion of additional biomarkers such as proteins and metabolites in our storage and query system, forming a general-purpose biosample database. Moreover, with the ongoing transformation of immunoassays(*37*) and spatial tissue data(*38*) into DNA molecules, we anticipate that the encapsulation and barcoding approach that we have demonstrated here can be used to store and query a comprehensive range of genomic, transcriptomic, and proteomic data. Finally, the prospect of encoding digital data, such as health records, into DNA, envisages a future where vast biological information could be efficiently stored, marking a significant leap forward in the compact and versatile storage of biological information in the palm of the hand.

METHODS

General materials. All DNA oligonucleotides (oligos)—including internal barcodes, splint adapters, 5'-amino-modified DNA barcodes, dye-labeled probes, preamplifier probes, amplifier sequences, master forward and reverse primers (**table s1.1**), random hexamers, and 20-mer oligodeoxythymidine—were synthesized and processed by Integrated DNA Technologies (IDT). Specifically, 5'-amino-modified barcodes in Echo 384 polypropylene microplates and internal barcodes in 96-deep well plates were purchased as desalted and delivered in nuclease-free water at 1000 μM and in 1× TE at 500 μM, respectively. Dye-labeled probes, adapter sequences, and branch sequences were received in 1× TE at concentrations of 100 μM, 33 μM, and 10 μM, respectively. The master primers were desalted and resuspended at 100 μM with nuclease-free water, while random hexamers and 20-mer oligodeoxythymidine were purified using ion-exchange high-performance liquid chromatography and resuspended at 50 μM with nuclease-free water. SARS-CoV-2 RNA controls were sourced from Twist Bioscience with catalog numbers 103909 and 104044. All oligos were stored at –20 °C. Silica particles of 5 μm diameter with hydroxy-terminated surfaces (catalog number: DNG-B017) were obtained from Creative Diagnostics, and *N*-[3-(Trimethoxysilyl)propyl]-*N*,*N*,*N*-trimethylammonium chloride (TMAPS; 50% methanol; catalog number: H66414) was acquired from Alfa Aesar. Chemicals such as tetraethyl orthosilicate (TEOS), N-(2-aminoethyl)-3-aminopropyltrimethoxysilane (AEAPTS), N-methyl-2-pyrrolidone (NMP), isopropanol, and ethanol were sourced from Millipore Sigma, bearing catalog numbers: 131903, 8.19172, 270458, 278475, and 459836, respectively. DBCO-PEG5-tetra ester (1260-10) and azidoacetic acid N-hydroxysuccinimide ester (1070-100) were from Click Chemistry Tools. Additionally, carbonate buffer (500 mM, pH 9.0; catalog number: J63899.AK) and saline sodium citrate (SSC, 20×; catalog number: 15557044) were purchased from Thermo Fisher, while dextran sulfate (50% in water; catalog number: S4030) came from Millipore Sigma, 10% Tween 20 from VWR (catalog number: 97063-980), formamide from VWR (catalog number: 97062-006), and 5M sodium chloride from VWR (catalog number: 97062-858). The preparation of ammonium-functionalized silica particles followed previously published procedures using TMAPS(*19*, *22*).

Internal barcode generation. A subset of primers, 700 in total, from the validated 240,000 primer library(*21*) was checked for alignment against SARS-CoV-2 using BLAST(*39*). The alignment XML output file from BLAST was then parsed to create a list of primers orthogonal to the SARS-CoV-2 genome. The resulting primer list was further filtered by selecting primers with 60–65°C melting temperature using the melting temperature module in Biopython.

Internal barcodes were generated by first picking a master primer pair from the filtered primer list, forward: 5'–GGCTATGAGACTGTTCGCTAATCAC–3' and reverse: 5'–CCCTTTGTGGGCACAGTTTAGTCTC–3', which flanked a unique barcode taken from the primer list. Five nucleotide randomer spacers (N) were also added between the master primers and the unique barcode to increase the sequence diversity of the internal barcodes for downstream sequencing. Together, the 85-nucleotide internal barcode sequence structure is GGCTATGAGACTGTTCGCTAATCACNNNNNUUUUUUUUUUUUUUUUUUUUUUUUUUNNNNNCCCTTTGTGGGCACAGTTTAGTCTC, where U is the unique barcode sequence.

96-well encapsulation. For each well in a Nunc 96 deep-well polypropylene plates (Thermo Fisher; catalog number: 278752), a total of 2 mg of ammonium-functionalized silica particles and 800 μl of 0.1% (v/v) Tween 20 in nuclease-free water were added. A total of 5 nanomoles of each internal barcode were added to their respective wells. Ten million copies of SARS-CoV-2 standards were added to each randomly selected well. A volume of 10 μl of TMAPS and 5 μl of TEOS were then added to each well. The plate was then covered with a chemically resistant silicone mat (Axygen; catalog number: AM-2ML-RD-S) and shaken for four days at 1500 rpm using a BioShake iQ thermal mixer (Bulldog Bio; catalog number: 1808-0506).

Encoding of features with DNA barcodes. A full list of features is given in (**table s2.3**). For each feature, a set of barcodes was allocated from which a subset was drawn to encode each feature value. The sets of barcodes for each feature were disjoint from each other. The encoding strategy for feature values differed based on the type of the metadata (numerical, Boolean, or categorical), which enabled type-specific queries of each feature, such as queries for feature values matching particular numerical ranges.

Numerical features. Numerical features were encoded using a mixed-radix number system (i.e. a sequence of digits with the base allowed to vary between positions) (**fig. s6a**). A position with base *n* was allocated *n* distinct barcode sequences, one for each of the *n* possible digit values at that position. Thus, a feature value represented by *k* digits was encoded on each microcapsule with a collection of *k* distinct barcodes. Ranges of varying size could be specified by allowing some number of the least significant digits to vary (i.e. placing "wildcards" at these digits) (**fig. s6b**). Experimentally, any numerical range specified in this manner can be retrieved by omitting the corresponding complementary probes during sorting. The base at each position was chosen to provide a good compromise between compression (number of barcodes required on each microcapsule) and the variety of the range sizes that could be represented.

Boolean features. Each Boolean feature was encoded using a single barcode assigned to that feature, similar to barcoding previously used to indicate image content in a database of images(*19*). When the value of a Boolean feature was TRUE for a sample, the barcode was displayed on the microcapsule; a value of FALSE was indicated by the absence of that barcode.

Categorical features. For each categorical feature we used a combinatorial number system(*40*) to associate each possible feature value with a distinct *k*-combination drawn from a chosen set of *n* barcodes, where *k* is the number of distinct barcodes used to represent this feature on each microcapsule. The number of possible feature values that may be represented is $\binom{n}{k}$, which grows rapidly with both *n* and *k*. The combinatorial number system provides a method to associate each possible *k*-combination with a unique integer value between 0 and $\binom{n}{k} - 1$. Each feature value was assigned a unique numerical index in that range, from which the corresponding *k*-combination of barcodes was determined. The values of *k* and *n* were chosen to provide a reasonable compromise between feature width (the number of barcodes required on a microcapsule to represent its feature value) and the number of barcodes that needed to be allocated for this feature.

Barcoding of individual samples and pooling. After encapsulation, the plate was centrifuged for 1 minute at 1000 relative centrifugal force (rcf). The supernatant was removed then backfilled with 1000 μl of 0.1% (v/v) Tween 20. A volume of 10 μl of AEAPTS was then added. The plate was then covered with a chemically resistant silicone mat and shaken for 1 day at 1500 rpm using a BioShake iQ thermal mixer.

To wash the encapsulated microparticles after amino modification, the plate was centrifuged for 1 minute at 1000 rcf. The supernatant was removed and then backfilled with 1000 μl of NMP. The washing step was repeated thrice and finally resuspended with 500 μl of NMP. A mass of 1 mg of azidoacetic acid *N*-hydroxysuccinimide ester was added to each well. The plates were then re-sealed with a chemically resistant silicone mat and shaken for 3 hours at 1500 rpm using a BioShake iQ thermal mixer. After azide modification, the previous wash steps were repeated, and the microparticles were finally resuspended in 500 μl of NMP. A mass of 0.5 mg of DBCO-PEG5-tetrafluorophenyl ester was added to each well, re-sealed with a chemically resistant silicone mat, and shaken for 5 hours at 1500 rpm using a BioShake iQ thermal mixer. The plate was washed with NMP thrice and then resuspended with 200 μl of NMP.

Barcode combinations were dispensed in a 96-well plate using an Echo 550 liquid handler and then transferred to the 96-deep-well plate containing the encapsulated microparticles. To each well of the 96 deep-well plates was added 800 μl of 50 mM carbonate buffer. The plate was re-sealed with a chemically resistant silicone mat and then shaken for 12 hours at 1500 rpm using a BioShake iQ thermal mixer. The barcoded microparticles were centrifuged at 1000 rcf then the supernatant was removed. The resulting pellet was washed with 20 mM Tris, 1 mM EDTA, and 0.1% (v/v) Tween 20 through repeated centrifugation, removal of the supernatant, and redispersing of the supernatant with 20 mM Tris, 1 mM EDTA, and 0.1% Tween 20 for three times. The microparticles were finally redispersed in 1000 μl of 20 mM Tris, 1 mM EDTA, and 0.1% (v/v) Tween 20. A volume of 500 μl from each well was taken and pooled together to create the sample library. No barcode degradation was observed over 4 months of room-temperature storage.

Selection of microparticles. An aliquot of 500 μl of the sample library was placed in a 1.5-ml tube and then centrifuged at 1000 rcf to sediment the microparticles. The supernatant was removed then the microparticles were re-dispersed with 200 μl of hybridization buffer (10× SSC, 10% (v/v) dextran sulfate, 10% (v/v) formamide, and 0.05% (v/v) Tween 20). Separately, equivolume of barcodes probes, adapters, and fluorescent probes, and 1× SSC were pre-hybridized using the following method: 98°C for 10 seconds, 40°C for 5 minutes, 20°C for 2 minutes. A volume of 20 μl of prehybridized probe solutions were then added to the microparticle suspension. The resulting mixture was shaken at 1500 rpm for 15 minutes at 35°C using a BioShake iQ thermal mixer then centrifuged at 1000 rcf to sediment the microparticles. The supernatant was removed, and the microparticles were re-dispersed 1000 μl with the sorting buffer, composed of 1× SSC and 0.05% (v/v) Tween 20, to wash the microparticles. The microparticle sedimentation and washing steps were repeated thrice. The microparticles were finally resuspended in 1000 μl of sorting buffer. The fluorescently labeled microparticles were sorted using a Sony SH800 sorter equipped with a 100-μm sorting chip.

In the multi-pass selection process, existing fluorescent probes from previously sorted populations were meticulously removed before initiating subsequent rounds. Populations from an earlier selection were initially centrifuged in a 1.5-ml tube at 1000 rcf for 20 seconds, after which the sheath buffer was gently discarded. Next, 1000 μl of a denaturation buffer, comprised of 0.2 M NaOH in 90% formamide and pre-heated to 70°C, was added. This mixture was swiftly vortexed for 5 seconds and then incubated at 70°C in a BioShake iQ thermal mixer for 15 minutes. Following this incubation period, another 20-second centrifugation at 1000 rcf was performed then

denaturation buffer was discarded. Then, 1000 µl of a denaturation wash buffer, composed of 0.05% Tween 20 and pre-heated to 70°C, was added. After a brief 5-second vortex and a 20-second centrifugation at 1000 rcf, the denaturation wash buffer supernatant was carefully siphoned off. Finally, 200 µl of hybridization buffer was added, preparing the sample for the next selection phase.

Sorted microparticles were de-encapsulated using 10 µl of electronics-grade 5:1 buffered oxide etch (VWR, catalog number: JT5192-3) and then diluted to 50 µl with nuclease-free water. The released samples were immediately used for Illumina sequencing library preparation.

Internal barcode validation using short-read sequencing. A volume of 1 µl of 50 µM of combined master forward (5'–TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNGGCTATGAGACTGTTCGCTAAT*C*A*C –3') and reverse (5'–GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAGACTAAACTGTGCCCACAAA*G*G*G –3') primers with three consecutive phosphorothioates from the 3'-end, 44 µl of nuclease-free water, and 50 µl of repliQa HiFi ToughMix (Quantabio, catalog number: 95200-500) were added to 5 µl of released sample. The samples were amplified for 20 cycles using the manufacturer's protocol then purified using 1× of AMPure XP beads (Beckman Coulter, catalog number: A63881). Samples were eluted from the magnetics beads using 22 µl of 20 mM Tris with 0.05% (v/v) Tween 20. Concentration of the amplified sample was measured using Qubit fluorescence assay (Thermo Fisher, catalog number: Q33231). Ten nanograms from the first PCR amplification were then taken to cycle-limited indexing PCR. Indexing PCR includes 10 µl of an indexing primer set from IDT® for Illumina® DNA/RNA UD Indexes (Illumina, catalog number: 20027213) or Nextera™ DNA CD Indexes (Illumina, catalog number: 20018708) and repliQa HiFi ToughMix as PCR master mix. Indexed samples were cleaned using 1× of AMPure XP beads and then quantified using quantitative PCR (qPCR). Final pooled libraries were then sequenced using an Illumina Nexsteq 2000 (800 pM loading concentration) using P1 flow cell with 150×2 reads, 50–80% human genome spike-in for nucleotide diversity, and 2% PhiX internal standard.

Synthetic SARS-CoV-2 sequencing. A portion of the released samples were then processed for sequencing SARS-CoV-2 samples using NEBNext® ARTIC SARS-CoV-2 FS Library Prep Kit for Illumina (New England Biolabs, catalog number: E7658) using the VarSkip primers with several modifications. Complementary DNA synthesis was performed using SuperScript™ IV

(catalog number: 18091200), 12 μl of released sample, and 1 μl of 50 μM 20-mer oligodeoxythymidine and 50 μM random hexamers for primers. Reverse transcription reactions were incubated at 50 °C for 1 hour. Finally, amplicons were amplified for 40 cycles.

Resulting libraries following the NEBNext® ARTIC SARS-CoV-2 FS Library Prep protocol were quantified using qPCR and then sequenced on a Nextseq 2000 (750 pM loading concentration) using a P1 flow cell with 150×2 reads and 1–10% PhiX internal standard.

Sequencing reads were first aligned with the SARS-CoV-2 Wuhan sequence (NC_045512.2) using minimap2(*41*) (2.24-r1122). Human-readable sequence alignment maps were converted to binary alignment maps using samtools(*42*) (v1.13). Resulting binary alignment map files were then used to demix the SARS-CoV-2 variants for each samples using Freyja(*32*) (v1.4.5).

ACKNOWLEDGEMENTS

M.B. and J.D.B. were supported by the Office of Naval Research (N00014-21-1-4013), the Army Research Office (ICB Subaward KK1954), and the National Science Foundation (CBET-1729397, OAC-1940231, and CCF-1956054). Additional funding to J.D.B. was provided through a National Science Foundation Graduate Research Fellowship (Grant No. 1122374). J.L.B. acknowledges support in part by the National Science Foundation SBIR Phase I 2136447, UCSF Parnassus Flow CoLab RRID:SCR_018206, DRC Center Grant NIH P30 DK063720, UCSF Center for Advanced Technology at Mission Bay, and Illumina. This research was also supported by a core center grant from the National Institute of Environmental Health Sciences, National Institutes of Health (P30-ES002109). We thank G. Paradis, M. Jennings, and M. Griffin of the Flow Cytometry Core at the Koch Institute at the Massachusetts Institute of Technology (MIT) for flow sorting assistance. We thank G. Tikhorimov for providing access to a Beckman Coulter Labcyte Echo 550.

COMPETING INTERESTS

The Massachusetts Institute of Technology has filed a patent related to this work on behalf of the authors. J.L.B. and M.B. are co-founders and equity shareholders of Cache DNA. J.L.B. is an

employee of Cache DNA and an independent contractor of OpenAI. D.K.R. was an intern at Cache DNA for the period of this work.

REFERENCES

1. F. Kreier, The myriad ways sewage surveillance is helping fight COVID around the world. *Nature*, doi: 10.1038/d41586-021-01234-1 (2021).

2. F. S. Collins, H. Varmus, A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).

3. A. J. Vargas, C. C. Harris, Biomarker development in the precision medicine era: lung cancer as a case study. *Nat. Rev. Cancer* **16**, 525–537 (2016).

4. S. Tarazona, A. Arzalluz-Luque, A. Conesa, Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat. Comput. Sci.* **1**, 395–402 (2021).

5. S. B. Lee, K. C. Clabaugh, B. Silva, K. O. Odigie, M. D. Coble, O. Loreille, M. Scheible, R. M. Fourney, J. Stevens, G. R. Carmody, T. J. Parsons, A. Pozder, A. J. Eisenberg, B. Budowle, T. Ahmad, R. W. Miller, C. A. Crouse, Assessing a novel room temperature DNA storage medium for forensic biological samples. *Forensic Sci. Int. Genet.* **6**, 31–40 (2012).

6. O. A. Ryder, A. McLaren, S. Brenner, Y.-P. Zhang, K. Benirschke, DNA Banks for Endangered Animal Species. *Science* **288**, 275–277 (2000).

7. P. Brandies, E. Peel, C. J. Hogg, K. Belov, The Value of Reference Genomes in the Conservation of Threatened Species. *Genes* **10**, 846 (2019).

8. C. Kieffer, A. J. Genot, Y. Rondelez, G. Gines, Molecular Computation for Molecular Classification. *Adv. Biol.* **7**, 2200203 (2023).

9. D. Y. Zhang, G. Seelig, Dynamic DNA nanotechnology using strand-displacement reactions. *Nat. Chem.* **3**, 103–113 (2011).

10. R. Lopez, R. Wang, G. Seelig, A molecular multi-gene classifier for disease diagnostics. *Nat. Chem.* **10**, 746–754 (2018).

11. C. Zhang, Y. Zhao, X. Xu, R. Xu, H. Li, X. Teng, Y. Du, Y. Miao, H. Lin, D. Han, Cancer diagnosis with DNA molecular computation. *Nat. Nanotechnol.* **15**, 709–715 (2020).

12. F. Yin, H. Zhao, S. Lu, J. Shen, M. Li, X. Mao, F. Li, J. Shi, J. Li, B. Dong, W. Xue, X. Zuo, X. Yang, C. Fan, DNA-framework-based multidimensional molecular classifiers for cancer diagnosis. *Nat. Nanotechnol.* **18**, 677–686 (2023).

13. R. A. Bull, T. N. Adikari, J. M. Ferguson, J. M. Hammond, I. Stevanovski, A. G. Beukers, Z. Naing, M. Yeang, A. Verich, H. Gamaarachchi, K. W. Kim, F. Luciani, S. Stelzer-Braid, J.-S. Eden, W. D. Rawlinson, S. J. van Hal, I. W. Deveson, Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat. Commun.* **11**, 6272 (2020).

14. T. D. Minogue, J. W. Koehler, C. P. Stefan, T. A. Conrad, Next-Generation Sequencing for Biodefense: Biothreat Detection, Forensics, and the Clinic. *Clin. Chem.* **65**, 383–392 (2019).

15. L. Whitmore, M. McCauley, J. A. Farrell, M. R. Stammnitz, S. A. Koda, N. Mashkour, V. Summers, T. Osborne, J. Whilde, D. J. Duffy, Inadvertent human genomic bycatch and intentional capture raise beneficial applications and ethical concerns with environmental DNA. *Nat. Ecol. Evol.* **7**, 873–888 (2023).

16. L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, K. Strauss, Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).

17. K. J. Tomek, K. Volkel, A. Simpson, A. G. Hass, E. W. Indermaur, J. M. Tuck, A. J. Keung, Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synth. Biol.* **8**, 1241–1248 (2019).

18. J. L. Banal, M. Bathe, Scalable Nucleic Acid Storage and Retrieval Using Barcoded Microcapsules. *ACS Appl. Mater. Interfaces* **13**, 49729–49736 (2021).

19. J. L. Banal, T. R. Shepherd, J. Berleant, H. Huang, M. Reyes, C. M. Ackerman, P. C. Blainey, M. Bathe, Random access DNA memory using Boolean search in an archival file storage system. *Nat. Mater.* **20**, 1272–1280 (2021).

20. K. J. Tomek, K. Volkel, A. Simpson, A. G. Hass, E. W. Indermaur, J. M. Tuck, A. J. Keung, Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synth. Biol.* **8**, 1241–1248 (2019).

21. Q. Xu, M. R. Schlabach, G. J. Hannon, S. J. Elledge, Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci.* **106**, 2289–2294 (2009).

22. J. L. Banal, M. Bathe, Scalable Nucleic Acid Storage and Retrieval Using Barcoded Microcapsules. *ACS Appl. Mater. Interfaces* **13**, 49729–49736 (2021).

23. S. Kashiwamura, M. Yamamoto, A. Kameda, T. Shiba, A. Ohuchi, “Hierarchical DNA Memory Based on Nested PCR” in *DNA Computing*, M. Hagiya, A. Ohuchi, Eds. (Springer, Berlin, Heidelberg, 2003)*Lecture Notes in Computer Science*, pp. 112–123.

24. M. Yamamoto, S. Kashiwamura, A. Ohuchi, M. Furukawa, Large-scale DNA memory based on the nested PCR. *Nat. Comput.* **7**, 335–346 (2008).

25. F. Porichis, M. G. Hart, M. Griesbeck, H. L. Everett, M. Hassan, A. E. Baxter, M. Lindqvist, S. M. Miller, D. Z. Soghoian, D. G. Kavanagh, S. Reynolds, B. Norris, S. K. Mordecai, Q. Nguyen, C. Lai, D. E. Kaufmann, High-throughput detection of miRNAs and gene-specific mRNA at the single-cell level by flow cytometry. *Nat. Commun.* **5**, 5641 (2014).

26. E. Goldstein, M. Lipsitch, M. Cevik, On the Effect of Age on the Transmission of SARS-CoV-2 in Households, Schools, and the Community. *J. Infect. Dis.* **223**, 362–369 (2021).

27. J. R. Fauver, M. E. Petrone, E. B. Hodcroft, K. Shioda, H. Y. Ehrlich, A. G. Watts, C. B. F. Vogels, A. F. Brito, T. Alpert, A. Muyombwe, J. Razeq, R. Downing, N. R. Cheemarla, A. L.

Wyllie, C. C. Kalinich, I. M. Ott, J. Quick, N. J. Loman, K. M. Neugebauer, A. L. Greninger, K. R. Jerome, P. Roychoudhury, H. Xie, L. Shrestha, M.-L. Huang, V. E. Pitzer, A. Iwasaki, S. B. Omer, K. Khan, I. I. Bogoch, R. A. Martinello, E. F. Foxman, M. L. Landry, R. A. Neher, A. I. Ko, N. D. Grubaugh, Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990-996.e5 (2020).

28. J. Y. Kishi, S. W. Lapan, B. J. Beliveau, E. R. West, A. Zhu, H. M. Sasaki, S. K. Saka, Y. Wang, C. L. Cepko, P. Yin, SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nat. Methods* **16**, 533–544 (2019).

29. A. N. Player, L.-P. Shen, D. Kenny, V. P. Antao, J. A. Kolberg, Single-copy Gene Detection Using Branched DNA (bDNA) In Situ Hybridization. *J. Histochem. Cytochem.* **49**, 603–611 (2001).

30. K. Tao, P. L. Tzou, J. Nouhin, R. K. Gupta, T. de Oliveira, S. L. Kosakovsky Pond, D. Fera, R. W. Shafer, The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.* **22**, 757–773 (2021).

31. Y. Bei, K. Pinet, K. B. Vrtis, J. G. Borgaro, L. Sun, M. Campbell, L. Apone, B. W. Langhorst, N. M. Nichols, Overcoming variant mutation-related impacts on viral sequencing and detection methodologies. *Front. Med.* **9**, 989913 (2022).

32. S. Karthikeyan, J. I. Levy, P. De Hoff, G. Humphrey, A. Birmingham, K. Jepsen, S. Farmer, H. M. Tubb, T. Valles, C. E. Tribelhorn, R. Tsai, S. Aigner, S. Sathe, N. Moshiri, B. Henson, A. M. Mark, A. Hakim, N. A. Baer, T. Barber, P. Belda-Ferre, M. Chacón, W. Cheung, E. S. Cresini, E. R. Eisner, A. L. Lastrella, E. S. Lawrence, C. A. Marotz, T. T. Ngo, T. Ostrander, A. Plascencia, R. A. Salido, P. Seaver, E. W. Smoot, D. McDonald, R. M. Neuhard, A. L. Scioscia, A. M. Satterlund, E. H. Simmons, D. B. Abelman, D. Brenner, J. C. Bruner, A. Buckley, M. Ellison, J. Gattas, S. L. Gonias, M. Hale, F. Hawkins, L. Ikeda, H. Jhaveri, T. Johnson, V. Kellen, B. Kremer, G. Matthews, R. W. McLawhon, P. Ouillet, D. Park, A. Pradenas, S. Reed, L. Riggs, A. Sanders, B. Sollenberger, A. Song, B. White, T. Winbush, C. M. Aceves, C. Anderson, K. Gangavarapu, E. Hufbauer, E. Kurzban, J. Lee, N. L. Matteson, E. Parker, S. A. Perkins, K. S. Ramesh, R. Robles-Sikisaka, M. A. Schwab, E. Spencer, S. Wohl, L. Nicholson, I. H. McHardy, D. P. Dimmock, C. A. Hobbs, O. Bakhtar, A. Harding, A. Mendoza, A. Bolze, D. Becker, E. T. Cirulli, M. Isaksson, K. M. Schiabor Barrett, N. L. Washington, J. D. Malone, A. M. Schafer, N. Gurfield, S. Stous, R. Fielding-Miller, R. S. Garfein, T. Gaines, C. Anderson, N. K. Martin, R. Schooley, B. Austin, D. R. MacCannell, S. F. Kingsmore, W. Lee, S. Shah, E. McDonald, A. T. Yu, M. Zeller, K. M. Fisch, C. Longhurst, P. Maysent, D. Pride, P. K. Khosla, L. C. Laurent, G. W. Yeo, K. G. Andersen, R. Knight, Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* **609**, 101–108 (2022).

33. BigQuery public datasets, *Google Cloud*. https://cloud.google.com/bigquery/public-data.

34. Open Datasets Documentation - Tutorials, API reference - Azure - Azure Open Datasets. https://learn.microsoft.com/en-us/azure/open-datasets/.

35. Open Data on AWS. https://aws.amazon.com/opendata/.

36. The Nucleic Acid Observatory Consortium, A Global Nucleic Acid Observatory for Biodefense and Planetary Health. [Preprint] (2021). arXiv:2108.02678.

37. G. H. Eldjarn, E. Ferkingstad, S. H. Lund, H. Helgason, O. Th. Magnusson, K. Gunnarsdottir, T. A. Olafsdottir, B. V. Halldorsson, P. I. Olason, F. Zink, S. A. Gudjonsson, G. Sveinbjornsson, M. I. Magnusson, A. Helgason, A. Oddsson, G. H. Halldorsson, M. K. Magnusson, S. Saevarsdottir, T. Eiriksdottir, G. Masson, H. Stefansson, I. Jonsdottir, H. Holm, T. Rafnar, P. Melsted, J. Saemundsdottir, G. L. Norddahl, G. Thorleifsson, M. O. Ulfarsson, D. F. Gudbjartsson, U. Thorsteinsdottir, P. Sulem, K. Stefansson, Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).

38. T. Zhao, Z. D. Chiang, J. W. Morriss, L. M. LaFave, E. M. Murray, I. Del Priore, K. Meli, C. A. Lareau, N. M. Nadaf, J. Li, A. S. Earl, E. Z. Macosko, T. Jacks, J. D. Buenrostro, F. Chen, Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).

39. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

40. D. E. Knuth, *The Art of Computer Programming, Volume 4, Fascicle 2: Generating All Tuples and Permutations* (Addison-Wesley, 2005).

41. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

42. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

43. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

44. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

45. N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L. Tan, L. M. Paul, D. E. Brackney, S. Grewal, N. Gurfield, K. K. A. Van Rompay, S. Isern, S. F. Michael, L. L. Coffey, N. J. Loman, K. G. Andersen, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).

46. I. Aksamentov, C. Roemer, E. Hodcroft, R. Neher, Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.* **6**, 3773 (2021).

47. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).