analytical chemistry

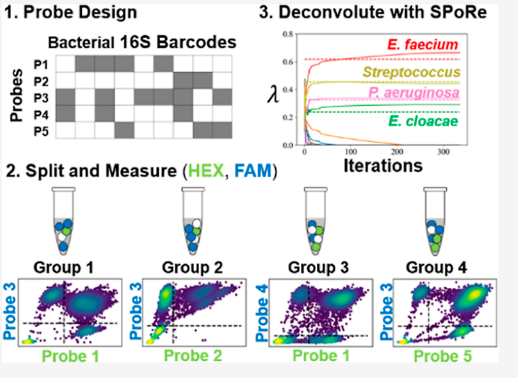
pubs.acs.org/ac Article

Expanded Multiplexing on Sensor-Constrained Microfluidic Partitioning Systems

Pavan K. Kota, Hoang-Anh Vu, Daniel LeJeune, Margaret Han, Saamiya Syed, Richard G. Baraniuk, and Rebekah A. Drezek*

Cite This: *Anal. Chem.* 2023, 95, 17458−17466

ABSTRACT: Microfluidics can split samples into thousands or millions of partitions, such as droplets or nanowells. Partitions capture analytes according to a Poisson distribution, and in diagnostics, the analyte concentration is commonly inferred with a closed-form solution via maximum likelihood estimation (MLE). Here, we present a new scalable approach to multiplexing analytes. We generalize MLE with microfluidic partitioning and extend our previously developed Sparse Poisson Recovery (SPoRe) inference algorithm. We also present the first in vitro demonstration of SPoRe with droplet digital PCR (ddPCR) toward infection diagnostics. Digital PCR is intrinsically highly sensitive, and SPoRe helps expand its multiplexing capacity by circumventing its channel limitations. We broadly amplify bacteria with 16S ddPCR and assign barcodes to nine pathogen genera by using five nonspecific probes. Given our two-channel ddPCR system, we measured two probes at a time in multiple groups of droplets. Although individual droplets are ambiguous in their bacterial contents, we recover the concentrations of bacteria in the sample from the pooled data. We achieve stable quantification down to approximately 200 total copies of the 16S gene per sample, enabling a suite of clinical applications given a robust upstream microbial DNA extraction procedure. We develop a new theory that generalizes the application of this framework to many realistic sensing modalities, and we prove scaling rules for system design to achieve further expanded multiplexing. The core principles demonstrated here could impact many biosensing applications with microfluidic partitioning.



The advent of microfluidics in biosensing has led to portable, cost-effective, and automated assays on chips manufactured with the same platforms that spurred the computing revolution.[1,2] However, the core methods of biosensing have largely rested on the paradigm of designing a specific sensor for each analyte. For situations with many target analytes to consider, this one-to-one principle scales poorly: many sensors must be embedded on a single device, samples must be concentrated enough such that a representative subsample can be applied to each sensor, and cross-reactivity of sensors and analytes scales combinatorially.[3,4] Our motivating application is in bacterial and fungal infection diagnostics where one or a few out of hundreds of plausible pathogens may be responsible for a patient's condition, but samples may exhibit very low microbial concentrations.[5,6] For instance, a milliliter of blood can have as low as one colony-forming unit or on the order of 10^2 to 10^3 equivalent genomic copies of microbial DNA.[7]

Scalable coverage of many analytes is viable with nonspecific sensing modalities that each generate measurements from multiple analytes. Such approaches need a postprocessing method for inferring the presence or quantities of individual analytes. For nucleic acid diagnostics, DNA sequencing is often the method-of-choice. Metagenomic sequencing analyzes the contents of virtually any sample with raw sequence reads analyzed and interpreted with bioinformatics, but this approach has limited sensitivity in the presence of high background such as host DNA in blood.[8] Amplicon sequencing is an alternative for microbial diagnostics in both microbiome analysis and infections.[9,10] These approaches conduct PCR on rRNA genes (e.g., 16S for bacteria, 18S or 28S for fungal, among others) that are flanked by conserved regions for priming and exhibit internal sequence differences for taxonomic discrimination.[9,11] While sequencing has made strides in clinical practice, its expense required expertise, and complex workflows have hindered its routine use.[12,13]

Another category of nonspecific sensing involves "fingerprinting" where a general sensing modality assigns unique signatures to analytes such that an unknown sample can be read and matched against a database. Spectroscopic methods

Received: March 17, 2023
Revised: November 2, 2023
Accepted: November 2, 2023
Published: November 16, 2023

https://doi.org/10.1021/acs.analchem.3c01176
Anal. Chem. 2023, 95, 17458−17466

are common in this class, profiling a wide range of proteins, metabolites, and cells. In clinical infections, mass spectrometry has been applied to rRNA amplicons[14] although its use to identify clinical isolates from positive culture is gaining much more traction.[15,16] However, these approaches often struggle to analyze mixtures of analytes.[17] For mass spectrometry, this limitation manifests as a need to analyze clinical isolates from polymicrobial samples one at a time.

Microfluidic partitioning technologies offer an avenue for the high-throughput, sensitive, and quantitative characterization of heterogeneous samples via a fingerprinting approach.[17] These systems split an initial sample into thousands or millions of partitions such as droplets or nanowells.[18] If the analytes are at a limiting dilution, most partitions will be empty, with an occasional analyte isolated in its own partition. Formally, analytes are captured according to a Poisson distribution with the Poisson rate parameter λ_n such that $P(x_n) = {\lambda_n}^{x_n} e^{-\lambda_n}/x_n!$ represents the probability of capturing nonnegative integer (x_n) copies of analyte n in a given partition.[19,20] Among N total analytes indexed by n that are distributed independently among the partitions, single-analyte capture is very likely if $\sum_n \lambda_n < 0.1$ with most partitions remaining empty. This approach with dilute samples guides much of the research in single-cell and single-molecule analysis.

Given the probabilistic isolation of individual analytes, nonempty partitions can be classified one at a time against a database. Researchers have demonstrated classification with high-resolution melt curve analysis of individual 16S gene amplicons captured in droplets.[21] Also, surface-enhanced Raman spectroscopy (SERS) of isolated bacterial cells[22] can assign unique spectra to species, and digital SERS with microfluidic capture has been proposed.[17] However, such approaches rest on the assumption that partitions must be individually classified. From a data perspective, these systems are dependent on reliable decision boundaries between N analyte classes which makes them highly sensitive to measurement noise.[23] Acquiring enough information from each partition for reliable classification limits the throughput of acquiring partition measurements, and therefore, the volume of sample that can be analyzed.[24] Moreover, in the diagnostics sample, concentrations are rarely known a priori. Multianalyte capture in the same partition in concentrated samples can cause errors in classification approaches that assume single-analyte capture.

Our group recently built on ideas from compressed sensing (CS) to address these challenges. CS seeks to infer sparse signals efficiently: faster or with fewer sensors.[25,26] In biosensing, samples are sparse when among many possible analytes only a handful are present in any given sample. For instance, a patient could be infected with any of hundreds of pathogens, but only one or a few are responsible for the current infection.[27] In this application, CS is analogous to quantifying analyte fingerprints from mixed measurements.[28] We recently developed new theory and a new Sparse Poisson Recovery (SPoRe) algorithm that couple principles of CS with microfluidic partitioning.[29] SPoRe performs maximum likelihood estimation (MLE) via gradient ascent over a generalized likelihood function. While we were initially motivated by microfluidics' high sensitivity via single-molecule analysis, we also found fundamental advantages from a signal processing perspective. Most notably, leveraging the Poisson-distributed capture of analytes enables improved rates of multiplexing (fewer sensors and more analytes), tolerates multianalyte capture in the same partition, withstands very high measurement noise, and can enable partial fingerprints to be captured separately in sensor-constrained systems.

This latter concept of asynchronous fingerprinting enables high-throughput, sensor-constrained microfluidics systems to achieve both sensitive detection and efficient multiplexing of analytes. The key insight is that individual partitions can be entirely ambiguous in their analyte content, but the distribution of all partition measurements can be used to solve for the analyte concentrations. In this work, we first extend our statistical theory to cover a broad class of realistic sensors. Next, we present the first in vitro demonstration of our framework toward bacterial infection diagnostics, quantifying 12 bacterial species at the genus level with only one 16S primer pair and five orthogonal DNA probes in two-channel droplet digital PCR (ddPCR). The probes assign "barcodes" to the 16S genes. While there are other probe-based methods for higher order multiplexing in channel-constrained ddPCR,[30] our oligo-efficient approach with nonspecific probes ameliorates cross-reactivity issues that otherwise scale combinatorially. Although our approach is not mutually exclusive to these techniques, their combination is beyond the scope of this work. For our bacterial panel, we selected species based on high prevalence and cause for concern due to growing drug resistance.[31−33] We characterize the performance of our assay with 18 samples each with a mixture of 2−4 bacteria, demonstrating accurate polymicrobial quantification down to approximately 200 total copies of the 16S gene. Finally, we show how our probabilistic framework enables the flagging of samples with 16S barcodes outside the designed panel. Our goal is that the promising practical results of our demonstration motivate further theoretical research, a refinement of our particular assay toward scalable infection diagnostics, and broader applications of our new framework to multiplexed biosensing.

■ EXPERIMENTAL SECTION

Bacterial Panel. We ordered bacterial species' genomic DNA (gDNA) from the American Type Culture Collection (ATCC, Manassas, VA). The species' names and their ATCC identifiers are as follows: *Acinetobacter baumannii* (BAA-1605), *Bacteroides fragilis* (25285), *Enterobacter cloacae* (13047), *Enterococcus faecium* (BAA-23200, *Escherichia coli* (11775), *Klebsiella pneumoniae* (13883), *Pseudomonas aeruginosa* (BAA-1744), *Staphylococcus aureus* (12600), *Staphylococcus epidermidis* (14990), *Staphylococcus saprophyticus* (15305), *Streptococcus agalactiae* (13813), and *Streptococcus pneumoniae* (33400). Particular strains were selected based on their availability at the time of purchase and only if ATCC provided whole genome sequence information for the isolate. DNA was resuspended and aliquoted according to ATCC's instructions at approximately 10^6 genome copies per microliter. DNA aliquots were stored at −4 °C until use.

Probe Design. All oligonucleotides were acquired from Integrated DNA Technologies (Coralville, IA) with HPLC purification and are given in Table 1. All probes had a 3′ Iowa Black quencher. HEX and FAM 5′ modifications are indicated in each experiment's context. We used ThermoBLAST from DNA Software (Plymouth, MI) to align the 16S primers (27F and 1492R from a previous study[11]) against bacterial genomes and find amplicons. Hydrolysis probes for barcoding must hit multiple bacterial taxa, and shorter probes are naturally less specific. We spiked probes with locked nucleic acids (LNAs) to

Table 1. Oligonucleotides Used in This Study

oligo name	sequence
primer 27F	AGAGTTTGATCMTGGCTCAG
primer 1492R	TACGGYTACCTTGTTAYGACTT
probe 1	TA+A+C+GGC+T+C+AC
probe 2	CTT+T+CGC+C+C+AT
probe 3	A+TT+C+C+GA+CT+TC
probe 4	A+C+C+AA+T+C+CATC
probe 5	A+A+G+CA+C+TCCGC

achieve a sufficient melting temperature (T_m). To avoid combinatorially increasing our probe search space, we deferred LNA positioning until after sequence selection.

Full details of our sequence selection process are provided in Supporting Information S1 (Supp. S1). We chose a length of 11 nucleotides for flexibility in LNA positioning and a sufficient T_m. We used heuristics based on the GC content and alignment to filter the 4^{11} possible 11-mers to avoid heterodimers and weak mismatches. As much as possible, we positioned LNAs at mismatch sites to improve the thermodynamic discrimination against these sequences. We evaluated all T_m's in IDT's OligoAnalyzer. Each 16S gene elicits a binary barcode response to the set of five candidate probes based on the presence or absence of the probe sequences in the gene (Figure 1a). We used coordinate ascent optimization to select a final probe set that separated the bacterial barcodes by genus. Particularly, we grouped together the three species of *Staphylococcus* and two species of *Streptococcus*.

Droplet Digital PCR. We used the Bio-Rad Qx 200 (Bio-Rad Laboratories, Hercules, CA, USA) which has two fluorescence channels (FAM and HEX) for multiplexed PCR with hydrolysis probes. Primers were at 900 nM, and for polymicrobial samples, all probes were at 125 nM. We used Bio-Rad's ddPCR Multiplex Supermix and prepared master mixes, generated droplets, and read droplets according to the manufacturer's instructions. For PCR cycling, extension times were set to 7 min because of the long amplicon (approximately 1500 base pairs) that is atypical in ddPCR, partly following guidance from a previous study[34] and internal data (not shown). PCR cycling was as follows: 95 °C for 10 min (initial denaturation and hot-start deactivation), 60 cycles of 94 °C for 30 s (denaturation), 60 °C for 7 min (annealing and extension), and 98 °C for 10 min. Ramp rates during cycling were set to 2 °C/s. Samples were refrigerated at 4 °C for 30 min prior to droplet readout.

Barcode Validation. We prepared ddPCR reactions with individual microbial gDNA and a no template control (NTC). We used amplitude multiplexing[30] to measure five probes in a single well with the two-channel system by adjusting individual probe concentrations (Figure S1).

Preparation of Polymicrobial Samples. We prepared monomicrobial dilutions of gDNA in Milli-Q purified water. One dilution was prepared for each bacterium, approximately targeting a concentration λ_n between 0.2 and 2 ("Concentration 1"). We diluted each of these by 1/2 to yield "Concentration 2." We used a custom script to assign random combinations of these bacterial dilutions to samples, generating five samples with $k = 2$ unique bacteria and six samples of $k = 3$ and $k = 4$. We reserved one sample as an NTC with water alone. The probability of drawing each bacterium was adaptively weighted to encourage an approximately even representation of taxa across the samples. Each sample was split across four wells of a ddPCR plate (Figure S2).

RESULTS AND DISCUSSION

Overview of the Approach. In ddPCR, samples are split into thousands of droplets to stochastically capture nucleic acids. End point PCR measurements form binary clusters that indicate the presence or absence of target sequences.[30,35] In this study, we use nonspecific probes that "barcode" 16S genes based on their binary pattern of response in ddPCR, and we statistically infer bacterial concentrations from partial barcode measurements. We present theoretical results and characterize the performance in an in vitro demonstration.

We must first account for the intragenomic sequence variability of copies of the 16S gene. Although we attempted to design probes such that each genus had a unique, consistent barcode for all copies, *E. cloacae* appeared to exhibit a small proportion of variant barcodes, a fraction which we computed experimentally (Figures S1, S3). Such variation is likely inevitable, especially in larger-scale systems, but can be accounted for. We store each pathogen's fractional barcode distribution across its copies in a column of a matrix **C** (Figure 1b). Note the ordering of barcodes is arbitrary in constructing **C** and that because of only slight variation between 16S copies within a genome,[11] **C** is nearly the identity matrix in practice. In the rows of **C**, we also ignore the barcodes that are not

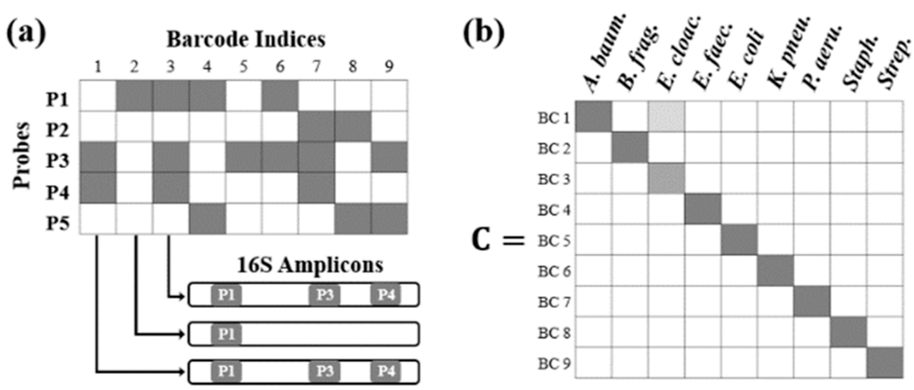


Figure 1. (a) Presence of nonspecific probe sequences in 16S gene copies defines their barcodes. (b) Accounting for barcode variability in 16S copies. The white values in **C** are zero with the darkest gray representing 1. Each column contains the proportions of the barcodes in each bacterial taxa's 16S gene copies.

https://doi.org/10.1021/acs.analchem.3c01176
Anal. Chem. 2023, 95, 17458−17466

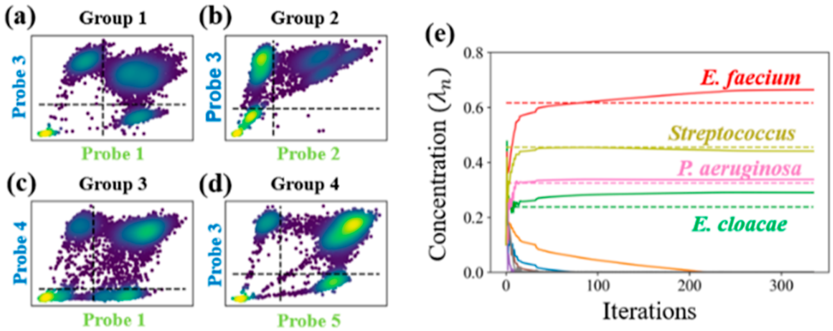


Figure 2. (a−d) Example of raw data from four groups of droplets, each from the same mixed bacterial sample. Green and blue axis labels indicate a 5′ HEX and FAM modification for the probes, respectively. Raw data are binarized by manual thresholding, overriding most of the effects due to PSC. (e) SPoRe algorithm optimizes over all groups simultaneously, accurately reflecting the estimated ground truth (dashed lines).

elicited by the combination of the probes and the bacterial panel. Our optimization estimates the nine-dimensional Poisson parameter vector $\boldsymbol{\lambda}$ that represents nine analyte concentrations. With nine unique barcodes and bacterial genera, the term "analytes" could refer to either. If the analytes are the barcodes, then $\lambda_n^{(\mathrm{BC})}$ is the concentration of the total 16S genes from any source bacteria that exhibits the nth barcode. If the analytes are the bacteria, then $\lambda_n^{(\mathrm{bact})}$ is the concentration of the nth bacterium's 16S genes, regardless of the particular barcodes of individual genes. Because $\boldsymbol{\lambda}^{(\mathrm{BC})} = \mathbf{C}\boldsymbol{\lambda}^{(\mathrm{bact})}$, our results currently depend on a $\mathbf{C}$ matrix of rank N to readily convert between the barcode concentrations $\boldsymbol{\lambda}^{(\mathrm{BC})}$ and the bacterial concentrations $\boldsymbol{\lambda}^{(\mathrm{bact})}$. We make use of both definitions of "analyte", carefully clarify which we are using at any time, and often drop the superscript.

Ideally, we could estimate $\boldsymbol{\lambda}^{(\mathrm{BC})}$ by simply capturing individual 16S genes in droplets with all five probes. However, if multiple genes appear (with distinct sequences) in the same droplet, an effect known as partition-specific competition (PSC) occurs, and fluorescent intensities can decrease.[31] PSC makes it difficult to differentiate many barcodes in a single reaction. Second, unique clusters for every barcode cannot scale beyond this study if the eventual goal is to quantify dozens to hundreds of microbes.

Instead, we generate four sensor groups of droplets each with a different subset of two probes (Figure 2). We call this concept asynchronous fingerprinting and describe the allocation of probes to each group in our theory (Supporting Information S2). Despite PSC effects, raw droplets can still be reasonably thresholded above zero in each channel.[36] Although the 16S barcode content in each droplet is made entirely ambiguous, we infer bacterial concentrations in the sample from the pooled, binarized data from the four groups of droplets. SPoRe essentially finds the solution that best explains the distribution of droplet measurements across the four groups (Figure 2e).

Generalized MLE with Microfluidic Partitioning. The standard for quantification in digital microfluidics data is based on MLE.[37] We generalize MLE in our new framework and apply it to ddPCR. Respectively, the general terms used in this section analyte, partition, and measurement vector correspond with the physical concepts of a barcode or bacterium, a droplet, and the two prebinarized measurements acquired from each droplet. Supporting Information S2 contains detailed clarification of our mathematical notation.

Let N and D define the number of unique analytes in the assay and the number of partitions, respectively. We let $\mathbf{x}_d$ be an N-dimensional nonnegative integer vector representing the quantities of each analyte in partition d. We say that $\boldsymbol{\lambda}$ is k-sparse if k elements are nonzero. With microfluidic partitioning, $\mathbf{x}_d$ is distributed as Poisson($\boldsymbol{\lambda}$) where $\boldsymbol{\lambda}$ is the N-dimensional parameter vector that characterizes the rate of capture of each of the N analytes. Let $\mathbf{y}_d$ represent the measurement vector acquired from partition d (e.g., in our assay, $\mathrm{y}_d \in \{0,1\}^2$). Note that while $\mathbf{y}_d$ is observed directly, $\boldsymbol{\lambda}$ must be inferred, and $\mathbf{x}_d$ is latent. We use an asterisk ($\boldsymbol{\lambda}^*$, $\mathbf{x}_d^*$) to denote true values and a hat ($\hat{\boldsymbol{\lambda}}$, $\hat{\mathbf{x}}_d$) to denote estimates. In MLE, an estimate of $\hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}$ maximizes the likelihood of the observed measurements

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\mathrm{MLE}} &= \arg\max_{\boldsymbol{\lambda}} \prod_{d=1}^{D} p(\mathbf{y}_d|\boldsymbol{\lambda}) \\ &= \arg\max_{\boldsymbol{\lambda}} \prod_{d=1}^{D} \sum_{\mathbf{x}\in\mathbb{Z}_+^N} p(\mathbf{y}_d|\mathbf{x})P(\mathbf{x}|\boldsymbol{\lambda})\end{aligned} \tag{1}$$

Denoting the likelihood function from the right-hand side of eq 1 as l, the gradient is

$$\nabla_{\boldsymbol{\lambda}} l = \frac{1}{D}\sum_{d=1}^{D} \frac{\sum_{\mathbf{x}\in\mathbb{Z}_+^N} p(\mathbf{y}_d|\mathbf{x})P(\mathbf{x}|\boldsymbol{\lambda})\boldsymbol{x}}{\boldsymbol{\lambda}\sum_{\mathbf{x}\in\mathbb{Z}_+^N} p(\mathbf{y}_d|\mathbf{x})P(\mathbf{x}|\boldsymbol{\lambda})} - 1 \tag{2}$$

Although fairly obtuse, this equation leads to two commonly used equations in specialized implementations of MLE that use digital fingerprinting or orthogonal assays (Supporting Information S3). In our ddPCR assay, droplets may contain multiple gene copies, including some with zero probe response. Despite this ambiguity, SPoRe uses gradient descent to solve eq 1.

We made two modifications to the original SPoRe implementation. First, SPoRe is modular for the appropriate sensing model, $p(\mathbf{y}_d|\mathbf{x})$, for the application. We used a simple model for our ddPCR assay. For $\mathrm{y}_d \in \{0,1\}^2$, with $M = 2$ for the two fluorescence channels, we say $p(\mathbf{y}_d|\mathbf{x}) = \prod_m p(y_m|\mathbf{x})$ with $p(\mathbf{y}_m|\mathbf{x}) = 1$ if $\mathbf{x}$ has at least one copy of a gene that contains the corresponding probe, with $p(\mathbf{y}_m|\mathbf{x}) = 0$ otherwise. In our implementation, we define the analyte as the bacterial content and account for the fractional barcode content in the gradient computations. Second, our earlier work used Monte Carlo approximations of the gradient (eq 2) on batches of observed measurements. Here, with the finite measurement space of ddPCR, these gradients can be computed quickly and exactly over all measurements (Supporting Information S4). This enables a backtracking line search to speed up convergence of

https://doi.org/10.1021/acs.analchem.3c01176
Anal. Chem. 2023, 95, 17458−17466

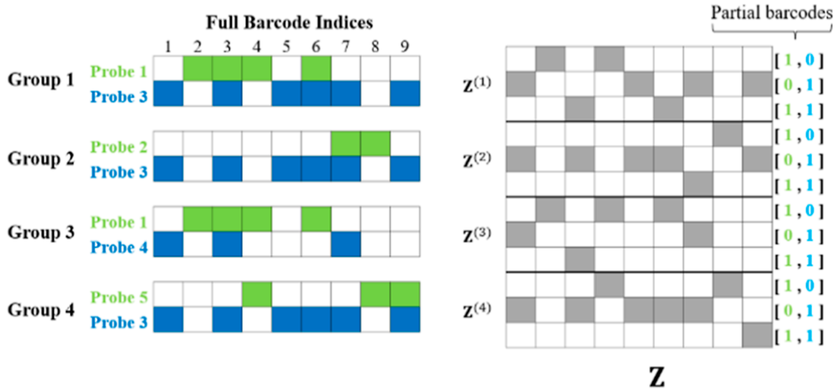


Figure 3. Formation of the linear system matrix **Z** that verifies the identifiability for our assay. Nonwhite squares are 1, and white squares are zero. Each group contributes three rows to **Z**, and rank (**Z**) must be N (Theorem 2.7, Supporting Information S2).

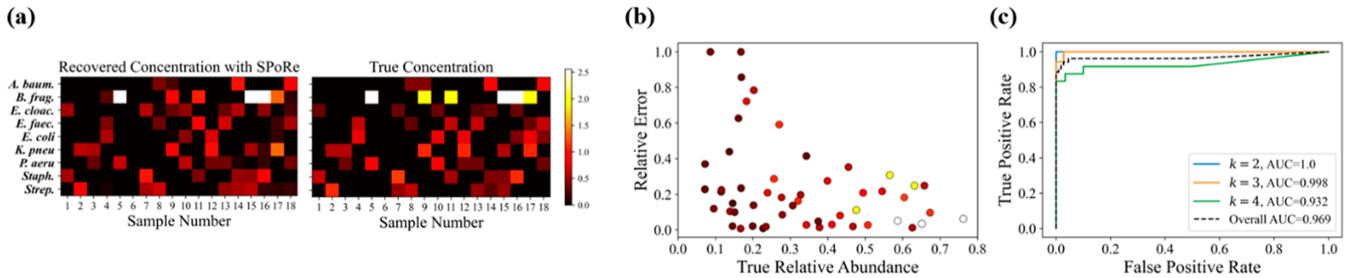


Figure 4. (a) Signal recovery results against the estimated ground truth. All colors are scaled against the maximum estimated ground truth concentration of $\boldsymbol{\lambda}^* = 2.56$ for concentration 1 of *B. fragilis*. Sample 6 is a negative control with no bacterial gDNA added. (b) Relative error of the estimated relative abundance versus true relative abundance. Data points use the same colormap in (a) to indicate the estimated absolute abundance of the bacterium. (c) Receiver operating characteristic curves and their area under the curves on aggregate data within different sparsity levels and across all samples.

gradient descent to estimate $\hat{\lambda}_{\mathrm{MLE}}{}^{(\mathrm{bact})}$. The exact gradient, while cumbersome to derive for the 2-channel ddPCR system, could be similarly calculated for any number of channels and is computationally cheap. This implementation and our raw data are available at https://github.com/pavankkota/SPoRe.

Identifiability of the System. Gradient descent will always converge to some solution, but we need to develop some assurance that it is the correct solution. Proving the identifiability of a model ensures that there is a unique global optimum for the likelihood function given infinite data. Identifiability states that if $p(\mathbf{y}|\boldsymbol{\lambda}) = p(\mathbf{y}|\boldsymbol{\lambda}')$ $\forall \mathbf{y}$ in each sensor group, then $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$.

We formally define terms and prove sufficiency conditions for identifiability (Supporting Information S2). Briefly, we find that $\boldsymbol{\lambda}$ can be inferred uniquely if the sensing functions that map $\mathbf{x}$ to $\mathbf{y}$ are monotonic. Monotonic functions do not change direction in the output; in biosensing, most outputs increase with increases in the input such that the sensors are monotonic increasing. We also assert that if one copy of an analyte does not yield a nonzero measurement, then the analyte is considered nonresponsive such that its content in a partition has no influence on the measurement. Lastly, we also impose a system-wide condition called fingerprint equivalence, which (informally) states that analytes with the same single-molecule fingerprints in a given sensing group behave interchangeably.

To align with these conditions in our proof, we define the analytes as the barcodes. Under reasonable PSC effects (e.g., not an overwhelming diversity of 16S genes in any given droplet), the addition of new barcodes to a droplet cannot reduce the binarized measurements (monotonicity). The binary data are determined by the presence of a 16S gene with a complementary probe sequence; without such a binding site, the gene is nonresponsive. Lastly, gene copies with the same combination of probe-binding sites are interchangeable (fingerprint equivalence). Our theorem (Theorem S2.7, Supporting Information S2) proves the sufficiency of these conditions for $\boldsymbol{\lambda}^{(\mathrm{BC})} = \boldsymbol{\lambda}'^{(\mathrm{BC})}$, and with rank $(\mathbf{C}) = N$, $\boldsymbol{\lambda}^{(\mathrm{BC})} = \boldsymbol{\lambda}'^{(\mathrm{BC})} \Rightarrow \boldsymbol{\lambda}^{(\mathrm{bact})} = \boldsymbol{\lambda}'^{(\mathrm{bact})}$.

Figure 3 illustrates a key process in our theorem for this particular assay. Our theorem defines a matrix $\mathbf{Z}^{(g)}$ whose rows indicate the positions of analytes with equal, nonzero fingerprint responses in the sensor group g. Stacking these matrices for each g yields a matrix $\mathbf{Z}$, and if $\mathrm{rank}(Z) = N$, then the system is identifiable. With two binary measurements per group, there are $2^2 - 1 = 3$ nonzero barcode measurements. For instance, note that in Group 1, original barcode indices 2 and 4 share a [1,0] response, yielding the first row of $\mathbf{Z}^{(1)}$. Each group contributes three rows to matrix $\mathbf{Z}$.

This result implies that we cannot arbitrarily assign probes to each group, and interestingly, using probes in multiple groups can be beneficial by adding rows to $\mathbf{Z}$. It is not sufficient to simply capture each probe's information in at least one group. Also, for $\mathbf{Z}$ to be $\mathrm{rank}(N)$, we can derive a simple rule of thumb for binary ddPCR with M channels: $G(2^M - 1) \geq N$ is necessary for the conditions of our theorem. Although we had access to a two-channel Bio-Rad Qx200, this result also indicates promise for applying our framework to digital PCR systems with more than two channels: up to $N = 15G$ analytes on the 4-channel QuantStudio Absolute Q (ThermoFisher Scientific, Waltham, MA), or up to $63G$ analytes on new 6-

channel systems from Bio-Rad and Roche (Basel, Switzerland). Note that we do not intend to rank these instruments as other factors such as the volume that can be processed, the number of partitions that can be generated, automation of workflows, etc. require application-specific consideration.

The implication of identifiability must be approached with caution: given the infinite data ($D \to \infty$) in each group, the true $\boldsymbol{\lambda}^*$ is the global optimum of the likelihood function. Our results fall short of a recovery guarantee for finite D partitions. In our earlier work,[29] we derived the insight that less sparse $\boldsymbol{\lambda}^*$ (more analytes with nonzero quantities) necessitate more partitions for stable recovery. Nonetheless, in contrast to typical applications of CS, there is no explicit maximum for the sparsity level as any $\boldsymbol{\lambda}$ is identifiable under our result.

Demonstration of Polymicrobial Quantification. We tested SPoRe's ability to quantify bacterial loads in mixed samples of purified gDNA. We used reference wells with individual bacterial dilutions to estimate the ground truth concentrations and assist with manual thresholding (Figure S4) to binarize the data. We passed this prebinarized data to SPoRe.

Figure 4a illustrates the quantitative results. For a general performance evaluation on polymicrobial samples, we used the cosine similarity metric to capture concordance with the true relative abundances of bacteria in the sample. We found an average cosine similarity of 0.97, indicating our ability to very reliably capture the dominant bacteria in a sample while making some errors on the relatively less abundant bacteria. These errors are further characterized in Figure 4b, which indicates that relative error in the estimated relative abundances decreases for higher relative abundance bacteria. Of course, $\hat{\boldsymbol{\lambda}}_{\mathrm{MLE}}$ estimates absolute abundance. Sweeping a global threshold on $\hat{\lambda}_n$ to make a binary call on bacterial presence yielded a receiver operating characteristic (ROC) curve with an AUC of 0.969 (Figure 4c) and a downward trend in AUC when increasing k. Less sparse samples are subject to higher estimation variance which may explain this effect.[29] A fixed threshold of $\lambda_n = 0.15$ achieves an overall sensitivity of 96% and specificity of 95%.

We investigated the source of error in the signal recovery. Differing in concentration estimates for the bacteria truly in the sample is to be expected due to pipetting volume variability and sampling variability. However, in some samples, SPoRe missed a bacterium of low abundance (a false negative) while it included a bacterium that is absent in the sample (a false positive). First, we confirmed that in all samples, $p(\mathbf{y}|\hat{\boldsymbol{\lambda}}) > p(\mathbf{y}|\boldsymbol{\lambda}^*)$ on average; the recovered solutions better explained the data given to SPoRe than the estimated ground truth (Figure S5). Thus, the local optima are unlikely to be the issue.

Next, we hypothesized that mistakes in thresholding propagated to SPoRe. Informally, given warped data, SPoRe returns a warped solution that could appear to have higher mean likelihood than the estimated ground truth. SPoRe's sensing model, $p(\mathbf{y}|\mathbf{x})$, assumes that binary measurements perfectly reflect the presence or absence of an amplicon with the corresponding probe sequence. However, in Figure S4, we illustrate and discuss how challenges with droplet rain, "lean" and "lift", and PSC cause some data points to fall ambiguously between clusters. While these effects are common and have some popular tools to help disambiguate droplets,[38,39] we decided to use manual clustering as these tools are generally not designed for the conditions of our assay. We designed a simulated experiment to evaluate SPoRe's performance in the idealized absence of cluster ambiguity. Given $\boldsymbol{\lambda}^*$ and the droplet counts in each group, we simulated the underlying droplet gene content ($\mathbf{X}$ with $\mathbf{x}_d \sim \mathrm{Poisson}(\boldsymbol{\lambda}^*)$ and the resulting binary measurements using our $p(\mathbf{y}|\mathbf{x})$ model. On this simulated data, SPoRe returned virtually perfect solutions with a mean cosine similarity of 0.9999 (Figure S6). This finding highlights the possibility that future research could focus on closing the gap between the modeled $p(\mathbf{y}|\mathbf{x})$ and experimental reality, perhaps via conditions that result in clearer cluster boundaries or probabilistic models for $p(\mathbf{y}|\mathbf{x})$ that account for assay-specific noise.

Characterization of Limit of Quantification. In infection diagnostics, pathogen loads can vary by several orders of magnitude. Tolerating multigene capture reduces the risk that high concentration samples flood a system and allows design for microfluidics systems with fewer partitions (e.g., smaller form factors with nanowells instead of droplets). We designed samples such that total concentrations ($\sum_n \lambda_n^*$) would be between 1 and 5 to illustrate this ability. However, demonstrating this capability on the Bio-Rad Qx200 means that our samples have 16S concentrations that are unrealistically high for most clinical presentations. We characterized the limit of quantification in terms of 16S copy counts per sample for partitioning systems that may still result in multianalyte capture (e.g., via spatial constraints that limit D) by randomly subsampling our experimental data. For each sample, we subsampled 10, 1, 0.1, and 0.01% of the droplet data and passed it to SPoRe. We estimate the 16S copy count in this data as the product of the number of subsampled droplets and the total estimated ground truth concentration $D_S(\sum_n \lambda_n^*)$.

Figure 5 shows how SPoRe maintains a strong recovery down to approximately 200 copies of the 16S gene. Depending

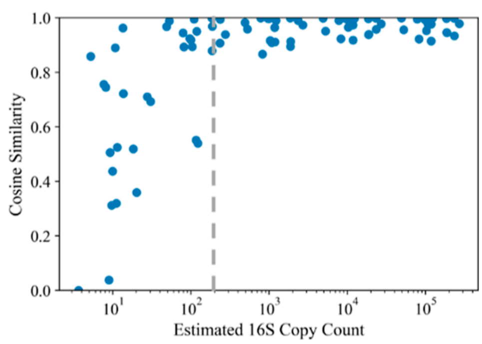


Figure 5. SPoRe's performance on random subsamples of experimental data. Each sample's set of prebinarized droplet measurements was subsampled by a factor of 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}.

on the quality of a future upstream microbial DNA isolation procedure, this limit could be potent for many applications in infection diagnostics. For instance, for a bacterial genome with five 16S copies, an initial sample volume of 5 mL, a DNA isolation procedure with 20% yield, and the ability to pass the entire elution volume across multiple groups in the digital PCR assay, our result would translate to a limit of quantification of 40 genome copies/mL.

Of course, a final system may have the flexibility to generate many partitions, driving lower magnitudes of $\boldsymbol{\lambda}$ which empirically help recovery.[29] Intuitively, signal inference can only gain information by capturing measurements from individual molecules rather than their combined effects.

https://doi.org/10.1021/acs.analchem.3c01176
Anal. Chem. 2023, 95, 17458−17466

Moreover, in ddPCR, single-molecule capture would avoid PSC altogether, such that thresholding may be more reliable.

Flagging Samples with Unknown Barcodes. Given a set of droplet measurements, the MLE will always report some solution even if the sample contains a bacterium with a 16S barcode distribution outside the panel given to SPoRe. However, this probabilistic approach allows us to assess the recovered solution and detect such anomalies. Given a recovered $\hat{\lambda}_{MLE}$, we can characterize the expected distribution of the discrete measurements and perform a χ^2 goodness of fit test between the expected and observed distributions. A poor match between these distributions would indicate a faulty solution that could be due to an out-of-panel bacterium.

We used the p value of the χ^2 goodness of fit test as a metric to detect faulty solutions. For each tested polymicrobial sample, we simulated the effect of having an "unknown" bacterium in it by removing each of the correct, present bacteria (one at a time) from SPoRe's database before running the algorithm. We repeated this process for both the manually thresholded and the simulated data. In both cases, the p value of the test is a highly reliable metric for flagging samples with out-of-panel bacteria, as indicated by the ROC curves (Figure 6). With simulated data, the separation is perfect with an area under the curve (AUC) of 1.0. Indeed, the minimum p value for SPoRe on simulated data with the full database of microbes was 0.783, and all cases in which SPoRe was deliberately not given one of the present bacteria in the sample returned $p = 0$. Given a reliable measurement model that corresponds with real-world data, the significant presence of a microbial barcode outside the provided database could be detected with a threshold on the p value. With manual thresholding, note that small mistakes in binarizing the data may make the observed distribution of measurements improbable for any $\boldsymbol{\lambda}$. As a result, samples that contain only bacteria in the panel may nonetheless return results that are flagged as faulty. We see this effect in the diminished (but still strong) performance with an AUC of 0.964.

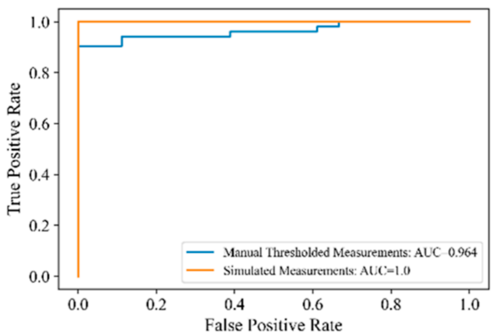


Figure 6. Flagging samples with out-of-panel bacteria. A χ^2 goodness of fit test is performed using the distribution of $\mathbf{y}$ expected given $\hat{\boldsymbol{\lambda}}$ and the observed distribution. The p value of the test is used as the metric for determining if a sample has 16S genes with barcodes that are unaccounted for in SPoRe.

Reporting "unknown bacteria" is likely far more useful to a clinician than reporting a "negative" result that would be returned from panels designed by specific sensors. This ability mirrors that of mass spectrometry and other fingerprinting systems, but the statistical underpinning could lead to theoretically grounded approaches with more research. Based on limitations in hard thresholding, our current assay would more likely only be able to report "faulty solution" since "unknown bacteria" is a more specific call with a different clinical decision pathway.

■ CONCLUSIONS

We present a new scalable framework for infection diagnostics that leverages the sparsity of samples and the Poisson distribution of microfluidic capture. We showed how analyte concentrations in a sample can be inferred from a population of partition measurements, despite ambiguity in the content of any individual partition. Tolerating this ambiguity enables the use of nonspecific probes for efficient multiplexing and asynchronous fingerprinting to circumvent channel limitations in common microfluidic systems. In an in vitro demonstration, we achieved clinically relevant limits of quantification of nine pathogen taxa with only five DNA probes and two primers.

Our ddPCR assay has a few areas for improvement. First, we designed probes to assign unique barcodes to the bacterial taxa in our chosen panel. Future bioinformatics tools could account for bacteria outside the panel that could plausibly appear in a sample to ensure that the designed barcodes are specific to the intended microbes. Second, our PCR cycling time was over 8 h driven by a long extension time to efficiently amplify the full 16S gene. Future iterations of our approach could employ custom master mixes with faster polymerases, restrict the amplicon to a shorter 16S segment, or replace hard thresholding with probabilistic noise models at faster cycling conditions. Third, many clinical infections may be caused by bacteria or fungi. Multiplexing primers to include eukaryotic marker genes along with the 16S primers for bacteria could enable the broadening of the panel. Lastly, automatic thresholding or postprocessing would be necessary for practical routine use. Internal controls and unsupervised clustering could help flexibly account for variable PSC effects.

While there is room for improvement in the ddPCR approach, our theory and algorithm open additional routes to improve this microbial assay or expand it to other applications. Our conditions on identifiability cover many realistic sensing modalities that could enhance performance. For instance, expanding to nonbinary measurements would enable fewer sensors to assign unique fingerprints to analytes at a higher rate. Moreover, our identifiability conditions are sufficient but not necessary, and our SPoRe algorithm is modular for any user-defined sensing function. We encourage users to proceed with simulations, even if their sensing model is outside the scope of our currently developed theory. Combining conventional sensors with new techniques in microfluidics and signal processing will offer a suite of new interdisciplinary approaches to scalable, multiplexed biosensing.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c01176.

Detailed description of algorithmic and theoretical considerations, including the probe sequence selection algorithm, formal proof of the identifiability result, comparisons to conventional MLE with digital microfluidics, modifications to the gradient computations, detailed description of amplitude multiplexing for barcode validation, experimental setup, thresholding ddPCR data, estimating intragenomic barcode varia-

https://doi.org/10.1021/acs.analchem.3c01176
Anal. Chem. 2023, 95, 17458−17466

bility, likelihood analysis of SPoRe's solutions, and performance on simulated versions of experimental samples (PDF)

AUTHOR INFORMATION

Corresponding Author

Rebekah A. Drezek − *Department of Bioengineering, Rice University, Houston, Texas 77005, United States; Department of Engineering Technology, University of Houston, Houston, Texas 77204, United States;* Email: drezek@rice.edu

Authors

Pavan K. Kota − *Department of Bioengineering, Rice University, Houston, Texas 77005, United States;* Present Address: P.K.K. is at Anvil Diagnostics Inc; orcid.org/0000-0001-7008-8405

Hoang-Anh Vu − *Department of Bioengineering, Rice University, Houston, Texas 77005, United States;* Present Address: H.-A.V. is at the Department of Bioengineering at U.C. Berkeley.

Daniel LeJeune − *Department of Electrical and Computer Engineering, Rice University, Houston, Texas 77005, United States*

Margaret Han − *Department of Biosciences, Rice University, Houston, Texas 77005, United States*

Saamiya Syed − *Department of Bioengineering, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0002-6635-9919

Richard G. Baraniuk − *Department of Electrical and Computer Engineering, Rice University, Houston, Texas 77005, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.3c01176

Author Contributions

All authors contributed to the preparation of the manuscript and have given approval to the final version.

Notes

The authors declare the following competing financial interest(s): Rice University has filed a patent application related to CS with microfluidic partitioning on which R.A.D., R.G.B., P.K.K., D.L., and H.-A.V. are co-inventors. Since the completion of this research, P.K.K. has joined Anvil Diagnostics Inc. which seeks to commercialize relevant technologies.

ACKNOWLEDGMENTS

This work was supported by NSF grant CBET 2017712 and the Rice Institute of Biosciences and Bioengineering. P.K.K. was supported by the NLM Training Program (T15LM007093). D.L. and R.G.B. were supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014-18-12571, N00014-20-1-2534, and MURI N00014-20-1-2787; AFOSR grant FA9550-22-1-0060; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047.

REFERENCES

(1) Luka, G.; Ahmadi, A.; Najjaran, H.; Alocilja, E.; Derosa, M.; Wolthers, K.; Malki, A.; Aziz, H.; Althani, A.; Hoorfar, M. *Sensors* **2015**, *15* (12), 30011−30031.

(2) Bashir, R. *Adv. Drug Delivery Rev.* **2004**, *56* (11), 1565−1586.

(3) Palka-Santini, M.; Cleven, B. E.; Eichinger, L.; Krönke, M.; Krut, O. *BMC Microbiol.* **2009**, *9*, 1−14.

(4) Macbeath, G. *Nat. Genet.* **2002**, *32* (S4), 526−532.

(5) Sinha, M.; Jupe, J.; Mack, H.; Coleman, T. P.; Lawrence, S. M.; Fraley, S. I. *Clin. Microbiol. Rev.* **2018**, *31* (2), 1−26.

(6) Opota, O.; Jaton, K.; Greub, G. *Clin. Microbiol. Infect.* **2015**, *21* (4), 323−331.

(7) Bacconi, A.; Richmond, G. S.; Baroldi, M. A.; Laffler, T. G.; Blyn, L. B.; Carolan, H. E.; Frinder, M. R.; Toleno, D. M.; Metzgar, D.; Gutierrez, J. R.; Massire, C.; Rounds, M.; Kennel, N. J.; Rothman, R. E.; Peterson, S.; Carroll, K. C.; Wakefield, T.; Ecker, D. J.; Sampath, R. *J. Clin. Microbiol.* **2014**, *52* (9), 3164−3174.

(8) Moragues-Solanas, L.; Scotti, R.; O'Grady, J. *Expert Rev. Mol. Diagn.* **2021**, *21* (4), 371−380.

(9) Schlaberg, R. *Clin. Chem.* **2020**, *66* (1), 68−76.

(10) Wellinghausen, N.; Kochem, A. J.; Disqué, C.; Mühl, H.; Gebert, S.; Winter, J.; Matten, J.; Sakka, S. G. *J. Clin. Microbiol.* **2009**, *47* (9), 2759−2765.

(11) Johnson, J. S.; Spakowicz, D. J.; Hong, B. Y.; Petersen, L. M.; Demkowicz, P.; Chen, L.; Leopold, S. R.; Hanson, B. M.; Agresta, H. O.; Gerstein, M.; Sodergren, E.; Weinstock, G. M. *Nat. Commun.* **2019**, *10* (1), 5029−5111.

(12) Schlaberg, R.; Chiu, C. Y.; Miller, S.; Procop, G. W.; Weinstock, G. *Arch. Pathol. Lab Med.* **2017**, *141* (6), 776−786.

(13) Gardy, J. L.; Loman, N. J. *Nat. Rev. Genet.* **2018**, *19* (1), 9−20.

(14) Ecker, D. J.; Sampath, R.; Massire, C.; Blyn, L. B.; Hall, T. A.; Eshoo, M. W.; Hofstadler, S. A. *Nat. Rev. Microbiol.* **2008**, *6* (7), 553−558.

(15) Singhal, N.; Kumar, M.; Kanaujia, P. K.; Virdi, J. S. *Front. Microbiol.* **2015**, *6* (AUG), 1−16.

(16) Özenci, V.; Patel, R.; Ullberg, M.; Strålin, K. *Clin. Infect. Dis.* **2018**, *66* (3), 452−455.

(17) Zhang, Y.; Hu, A.; Andini, N.; Yang, S. *Biotechnol. Adv.* **2019**, *37* (3), 476−490.

(18) Guo, M. T.; Rotem, A.; Heyman, J. A.; Weitz, D. A. *Lab Chip* **2012**, *12* (12), 2146−2155.

(19) Basu, A. S. *SLAS Technol.* **2017**, *22* (4), 369−386.

(20) Moon, S. J.; Ceyhan, E.; Gurkan, U. A.; Demirci, U. *PLoS One* **2011**, *6* (7), No. e21580.

(21) Velez, D. O.; Mack, H.; Jupe, J.; Hawker, S.; Kulkarni, N.; Hedayatnia, B.; Zhang, Y.; Lawrence, S.; Fraley, S. I. *Sci. Rep.* **2017**, *7* (1), 42326−42414.

(22) Dina, N. E.; Zhou, H.; Colniţă, A.; Leopold, N.; Szoke-Nagy, T.; Coman, C.; Haisch, C. *Analyst* **2017**, *142* (10), 1782−1789.

(23) Langouche, L.; Aralar, A.; Sinha, M.; Lawrence, S. M.; Fraley, S. I.; Coleman, T. P. *Bioinformatics* **2021**, *36* (22−23), 5337−5343.

(24) Zhu, Y.; Fang, Q. *Anal. Chim. Acta* **2013**, *787*, 24−35.

(25) Baraniuk, R. *IEEE Signal Process. Mag.* **2007**, *24* (4), 118−121.

(26) Donoho, D. L. *IEEE Trans. Inf. Theor.* **2006**, *52* (4), 1289−1306.

(27) Peters, B. M.; Jabra-rizk, M. A.; O'May, G. A.; Costerton, J. W.; Shirtliff, M. E. *Clin. Microbiol. Rev.* **2012**, *25* (1), 193−213.

(28) Aghazadeh, A.; Lin, A. Y.; Sheikh, M. A.; Chen, A. L.; Atkins, L. M.; Johnson, C. L.; Petrosino, J. F.; Drezek, R. A.; Baraniuk, R. G. *Sci. Adv.* **2016**, *2* (9), No. e1600025.

(29) Kota, P. K.; LeJeune, D.; Drezek, R. A.; Baraniuk, R. G. *IEEE Trans. Signal Process.* **2022**, *70*, 2388−2401.

(30) Whale, A. S.; Huggett, J. F.; Tzonev, S. *Biomol. Detect. Quantif.* **2016**, *10*, 15−23.

(31) Flores-Mireles, A. L.; Walker, J. N.; Caparon, M.; Hultgren, S. J. *Nat. Rev. Microbiol.* **2015**, *13* (5), 269−284.

(32) Mayr, F. B.; Yende, S.; Angus, D. C. *Virulence* **2014**, *5* (1), 4−11.

(33) Mulani, M. S.; Kamble, E. E.; Kumkar, S. N.; Tawre, M. S.; Pardesi, K. R. *Front. Microbiol.* **2019**, *10* (April), 539.

(34) Lasham, A.; Tsai, P.; Fitzgerald, S. J.; Mehta, S. Y.; Knowlton, N. S.; Braithwaite, A. W.; Print, C. G. *Cancers* **2020**, *12* (3), 769.

(35) Quan, P. L.; Sauzade, M.; Brouzes, E. *Sensors* **2018**, *18* (4), 1271.
(36) Whale, A. S.; De Spiegelaere, W.; Trypsteen, W.; Nour, A. A.; Bae, Y.-K.; Benes, V.; Burke, D.; Cleveland, M.; Corbisier, P.; Devonshire, A. S.; Dong, L.; Drandi, D.; Foy, C. A.; Garson, J. A.; He, H.-J.; Hellemans, J.; Kubista, M.; Lievens, A.; Makrigiorgos, M. G.; Milavec, M.; Mueller, R. D.; Nolan, T.; O'Sullivan, D. M.; Pfaffl, M. W.; Rödiger, S.; Romsos, E. L.; Shipley, G. L.; Taly, V.; Untergasser, A.; Wittwer, C. T.; Bustin, S. A.; Vandesompele, J.; Huggett, J. F. *Clin. Chem.* **2020**, *66* (8), 1012−1029.
(37) Majumdar, N.; Wessel, T.; Marks, J. *PLoS One* **2015**, *10* (3), 0118833.
(38) Brink, B. G.; Meskas, J.; Brinkman, R. R. *Bioinformatics* **2018**, *34* (15), 2687−2689.
(39) Jones, M.; Williams, J.; Gärtner, K.; Phillips, R.; Hurst, J.; Frater, J. *J. Virol. Methods* **2014**, *202*, 46−53.

https://doi.org/10.1021/acs.analchem.3c01176
Anal. Chem. 2023, 95, 17458−17466