analytical chemistry

pubs.acs.org/ac Article

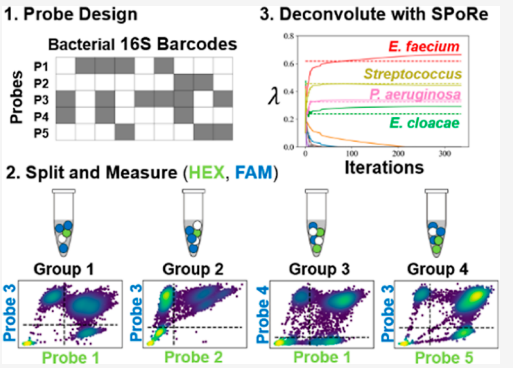
Expanded Multiplexing on Sensor-Constrained Microfluidic Partitioning Systems

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 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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

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

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

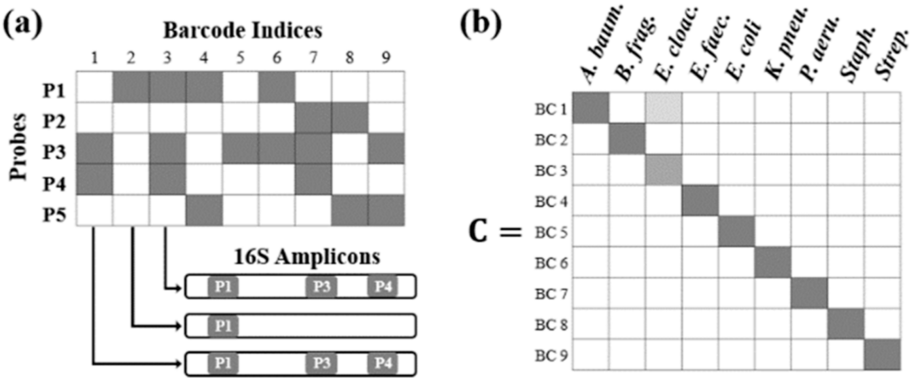


Figure 1. (a) Presence of nonspecific probe sequences in 16S gene copies defines their barcodes. (b) Accounting for barcode variability in 16S copies. The white values in **C** are zero with the darkest gray representing 1. Each column contains the proportions of the barcodes in each bacterial taxa's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

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

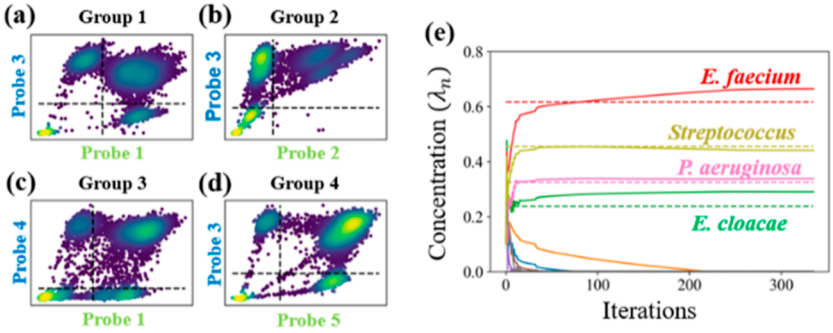


Figure 2. (a–d) Example of raw data from four groups of droplets, each from the same mixed bacterial sample. Green and blue axis labels indicate a 5′ HEX and FAM modification for the probes, respectively. Raw data are binarized by manual thresholding, overriding most of the effects due to PSC. (e) SPoRe 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, 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

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

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 $p(\mathrm{y}_d|\mathbf{x}) = \prod_m p(y_m|\mathbf{x})$ with $p(\mathrm{y}_m|\mathbf{x}) = 1$ if $\mathbf{x}$ has at least one copy of a gene that contains the corresponding probe, with $p(\mathrm{y}_m|\mathbf{x}) = 0$ 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

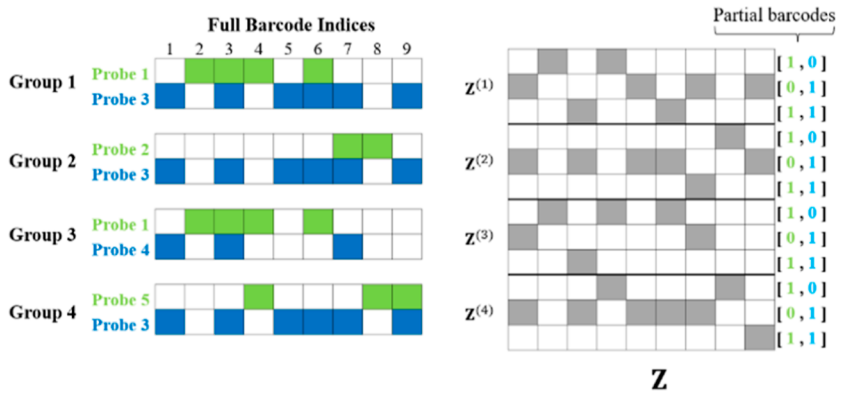


Figure 3. Formation of the linear system matrix **Z** that verifies the identifiability for our assay. Nonwhite squares are 1, and white squares are zero. Each group contributes three rows to **Z**, and rank (**Z**) must be N (Theorem 2.7, Supporting Information S2).

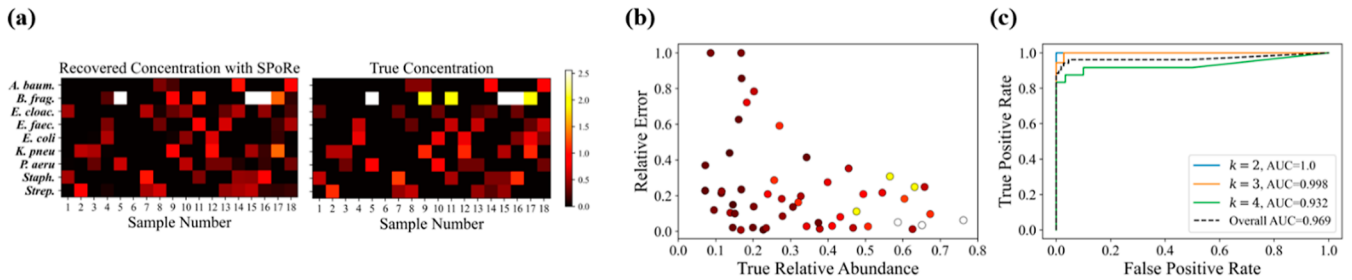


Figure 4. (a) Signal recovery results against the estimated ground truth. All colors are scaled against the maximum estimated ground truth concentration of $\lambda^* = 2.56$ for concentration 1 of *B. fragilis*. Sample 6 is a negative control with no bacterial gDNA added. (b) Relative error of the estimated 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 **x** to **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 (**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 **Z**, and if 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 **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 **Z**. It is not sufficient to simply capture each probe's information in at least one group. Also, for **Z** to be 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-

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

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

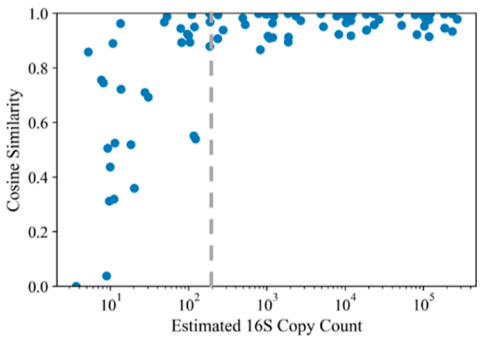


Figure 5. SPoRe's performance on random subsamples of experimental data. Each sample's set of prebinarized droplet measurements was subsampled by a factor of 10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}.

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 $\hat{\lambda}_{\mathrm{MLE}}$, 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

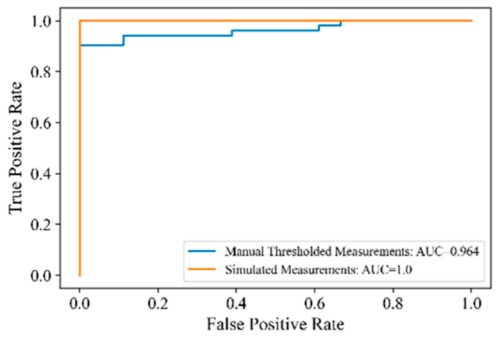


Figure 6. Flagging samples with out-of-panel bacteria. A χ^2 goodness of fit test is performed using the distribution of $\mathbf{y}$ expected given $\hat{\boldsymbol{\lambda}}$ and the observed distribution. The p value of the test is used as the metric for determining if a sample has 16S genes with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

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.

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

Supporting Information

Expanded Multiplexing on Sensor Constrained Microfluidic Partitioning Systems

Pavan K. Kota[1], Hoang-Anh Vu[1], Daniel LeJeune[2], Margaret Han[3], Saamiya Syed[4], Richard G. Baraniuk[2], and Rebekah A. Drezek*[1]

[1]Department of Bioengineering, Rice University
[2]Department of Electrical and Computer Engineering, Rice University
[3]Department of Biosciences, Rice University
[4]Department of Engineering Technology, University of Houston
*Corresponding author: drezek@rice.edu

S1 Probe Design

We used ThermoBLAST from DNA Software (Plymouth, MI) to align the 16S primers (27F and 1492R) against bacterial genomes and find amplicons. We passed these amplicons to a custom Matlab script to design probes. We chose a sequence length of 11 nucleotides with 5-8 GC nucleotides, without four consecutive G's or C's, and without a G on the 5' end to avoid self-quenching of the fluorophore. We used Smith-Waterman alignment in Matlab to pre-screen for probes that self-hybridize and to assess cross-hybridization of probes amongst the evolving candidate set. To assist in achieving near binary measurements, we considered perfect matches on all 16S copies to be "1" for a genome, and for imperfect homology, we filtered for sequences that had neither nine consecutive matches nor a single G-T mismatch. This latter filtering is a proxy for ensuring that probes have weak, negligible interactions against 16S sequences where they do not have perfect complementarity. The former filtering for positive hits was intended to avoid the issue of mixtures of barcodes for any particular bacteria for simplicity in our initial demonstration.

Given a set of filtered, candidate probes, we used a coordinate ascent strategy to iteratively optimize a set. We hypothesized that barcoding the full-length 16S gene with probes could achieve genus level resolution, as sequencing the full gene achieves a mix of genus and species resolution. As a result, we encouraged similarity of the three *Staphylococcus* species and the two *Streptococcous* species. Define $\mathcal{S}$ as a set of pairs of bacteria (b_i, b_j) within a genus that are similar. The complementary set $\mathcal{D}$ includes all other bacterial pairs. Let $\mathbf{k}_{p,i}$ represent the 11-mer barcode of bacteria i with probes indexed by p. Coordinate ascent sought to solve:

$$\arg\max_{p} \left[\sum_{(b_i,b_j)\in\mathcal{S}} -\|\mathbf{k}_{p,i} - \mathbf{k}_{p,j}\|_2^2 + \log\left(\sum_{(b_i,b_j)\in\mathcal{D}} \|\mathbf{k}_{p,i} - \mathbf{k}_{p,j}\|_2 \right) \right] + \theta \min_{(b_i,b_j)\in\mathcal{D}} \|\mathbf{k}_{p,i} - \mathbf{k}_{p,j}\|_2 \tag{S1}$$

The first term is taken from research in metric learning [1], and the second term (with a weight of $\theta = 10$) highly rewards some nonzero separation between all bacterial pairs that are intended to be discriminated between. We chose an initial random set of probes that passed our cross-hybridization check. We iteratively cycled through a shuffled order of the candidate set of probes, evaluating one probe at a time for replacement with any of the other probes that passed the initial filtering step. If replacing a probe improved the objective function, the probe set was updated and the search continued. The algorithm terminated when all probe sequences had been evaluated for replacement but not replaced. For the chosen set of sequences, we evaluated the alignment against 16S genes with imperfect homology (the zeroes in the barcodes). 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, positioning additional LNAs as necessary to reach a sufficient probe T_m.

S2 Theory: Identifiability with Common Types of Sensors

For estimating $\boldsymbol{\lambda}$, the property of *identifiability* means that there is a one-to-one correspondence between each realizable distribution of measurements and the Poisson rates $\boldsymbol{\lambda}$: if $p(\mathbf{y}|\boldsymbol{\lambda}) = p(\mathbf{y}|\boldsymbol{\lambda}')$ for all $\mathbf{y}$, then $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$. From an optimization perspective, identifiability implies that the $\boldsymbol{\lambda}^*$ is the unique global optimum

to the likelihood function if we have infinite measurements. Therefore, identifiability is a necessary condition for our method to work.

S2.1 Notation

We use bold face upper and lower case letters for matrices and vectors, respectively. Non-bold, lower case letters represent scalars. We denote the vector of all zeros as $\mathbf{0}$ with its dimensionality dependent on context. We use script letters ($\mathcal{A}$, $\mathcal{B}$, etc.) to denote sets. We denote $\mathbf{e}_j$ as the standard basis vector with $\mathbf{e}_j = 1$ and $\mathbf{e}_i = 0$ for all $i \neq j$. Let $\mathbf{a}$ and $\mathbf{b}$ be two arbitrary vectors of the same dimension, and let a_i and b_i denote their ith elements. We use $\text{supp}(\mathbf{a})$ to denote the support of vector $\mathbf{a}$ defined as the index set where $a_i > 0$ for $i \in \text{supp}(\mathbf{a})$. We use the notation $\mathbf{a} \succeq \mathbf{b}$ to imply that $a_i \geq b_i \ \forall i$, and we use $\mathbf{a} \succ \mathbf{b}$ to further imply the existence of at least one index i where $a_i > b_i$. A set in the subscript of a vector such as $\mathbf{x}_{\mathcal{A}}$ refers to the subvector of $\mathbf{x}$ indexed by the elements of $\mathcal{A}$. We make frequent use of the shorthand $\sum \mathbf{a}$ to denote the summation over elements of a vector $\mathbf{a}$.

S2.2 Definitions and Assumptions

We treat the dataset of measurements from all partitions in a sensor group as samples of a random variable $\mathbf{y}$. The signal (i.e., analyte quantities in a partition) $\mathbf{x}$ is N-dimensional with $\mathbf{x} \sim \text{Poisson}(\boldsymbol{\lambda}^*)$. Each signal is measured by M sensors to yield the observation vector $\mathbf{y}$ (e.g., M fluorescence measurements). We define the function $f : \mathbb{Z}_+^N \to \mathbb{R}^M$ that is composed of M scalar functions $f_m : \mathbb{Z}_+^N \to \mathbb{R}$. A particular measurement value y_m is determined by the sensor output $f_m(\mathbf{x})$ plus some additive, zero-mean random noise n_m.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_M \end{bmatrix} \tag{S2}$$

Note that the sensor functions f_m are group-dependent. For example, each group may have different probes.

We assume that our sensors are *monotonic* and that our analytes obey *responsiveness* and *fingerprint equivalence*. Given these properties, we prove sufficient conditions for identifiability. Without loss of generality, we will say that all M sensor functions are monotonic.

Definition S2.1 (Monotonic Sensors)**.** *A sensor function $f_m : \mathbb{Z}_+^N \to \mathbb{R}$ is monotonic increasing if $\mathbf{a} \succeq \mathbf{b} \Rightarrow f_m(\mathbf{a}) \geq f_m(\mathbf{b})$ and monotonic decreasing if $\mathbf{a} \succeq \mathbf{b} \Rightarrow f_m(\mathbf{a}) \leq f_m(\mathbf{b})$.*

Monotonic functions are very common and natural; for instance, many sensing modalities have a monotonic increasing sigmoidal response to their input. Any time a Lemma or Theorem relies on monotonicity, its proof will assume all M sensors are monotonic increasing without loss of generality. Next, we define the *responsiveness* property of analytes:

Definition S2.2 (Responsiveness)**.** *If $f(\mathbf{e}_i) \neq f(\mathbf{0})$, the analyte indexed by i is said to be responsive. If $f(\mathbf{e}_i) = f(\mathbf{0})$, then the analyte indexed by i is nonresponsive. If we let $\mathcal{B}$ define the set of indices for all such nonresponsive analytes ($i \in \mathcal{B}$), then for any two signals $\mathbf{x}$ and $\mathbf{x}'$, $f(\mathbf{x}) = f(\mathbf{x}')$ if $x_n = x'_n$ for all $n \notin \mathcal{B}$.*

In other words, a nonresponsive analyte does not influence the sensor output regardless of its quantity. An analyte is considered "responsive" if a single copy yields a different measurement than the null signal (e.g., an empty microfluidic partition).

We define a final intuitive condition on our system called *fingerprint equivalence*. The *fingerprint* of analyte n is the measurement yielded by an isolated copy of the analyte, or $f(\mathbf{e}_n)$. Among analytes with identical fingerprint responses within a sensor group, the total number of occurrences of these analytes dictates the output response. In other words, the sensors treat these analytes as interchangeable copies of each other.

Definition S2.3 (Fingerprint Equivalence)**.** *Let $\mathcal{X} \subseteq \{1, ..., N\}$ be an index set of analytes with identical fingerprints, i.e. $f(\mathbf{e}_i)$ is fixed for all $i \in \mathcal{X}$. A system has the fingerprint equivalence property if for any pair of vectors $\mathbf{x}$ and $\mathbf{x}'$ with $\text{supp}(\mathbf{x})$, $\text{supp}(\mathbf{x}') \subseteq \mathcal{X}$ and $\sum_n x_n = \sum_n x'_n$, we have $f(\mathbf{x}) = f(\mathbf{x}')$.*

Note that even if all analyte fingerprints are distinct, multiple signals can still map to the same measurement vector since we allow for cases of multi-analyte capture in the same partition. We define these signals as members of a *collision set*.

Definition S2.4 (Collision Sets)**.** *The collision set $\mathcal{C}_{\mathbf{x}}$ for signal $\mathbf{x}$ is the set of all signals $\mathbf{x}'$ that satisfy $f(\mathbf{x}') = f(\mathbf{x})$.*

We define $\mathcal{U}$ as the set of unique collision sets. In 2-channel ddPCR with binarized measurements, there are four collision sets in each sensor group ($\{0,1\}^2$). It will soon be clear that observations $\mathbf{y}$ are drawn from a mixture model. We can define each *mixture element* as follows:

Definition S2.5 (Mixture Element)**.** *The mixture element $\mathcal{E}_{\mathbf{x}}$ for signal $\mathbf{x}$ is the set of all signals $\mathbf{x}'$ that satisfy $p(\mathbf{y}|\mathbf{x}) \sim p(\mathbf{y}|\mathbf{x}')$.*

Note with any zero-mean noise, $p(\mathbf{y}|\mathbf{x}) \sim p(\mathbf{y}|\mathbf{x}') \Rightarrow f(\mathbf{x}) = f(\mathbf{x}')$ such that $\mathcal{E}_{\mathbf{x}} \subseteq \mathcal{C}_{\mathbf{x}}$. In some cases, such as additive white Gaussian noise, $\mathcal{E}_{\mathbf{x}} = \mathcal{C}_{\mathbf{x}}$. We define $\mathcal{V}$ as the set of unique mixture elements with arbitrary $\mathcal{E}_v \in \mathcal{V}$.

S2.3 Proof of Identifiability

With G different sensor groups indexed by g, we assume that the sensor group applied to a measurement $\mathbf{y}$ is known and deterministic. Each sensor group has a different function f that maps $\mathbf{x}$ to M-dimensional space (e.g., different probes in ddPCR). Identifiability means that $p(\mathbf{y}|\boldsymbol{\lambda}) = p(\mathbf{y}|\boldsymbol{\lambda}')\ \forall \mathbf{y}, g \Rightarrow \boldsymbol{\lambda} = \boldsymbol{\lambda}'$. Each $\boldsymbol{\lambda}$ must yield a unique set of G distributions of measurements.

We will use the notation $\mathcal{C}_u^g$ and $\mathcal{E}_v^g$ to specify the group g when necessary. For an arbitrary group, we can express $p(\mathbf{y}|\boldsymbol{\lambda})$ as:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\lambda}) &= \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}) P(\mathbf{x}|\boldsymbol{\lambda}) \\ &= \sum_{\mathcal{E}_v \in \mathcal{V}} p(\mathbf{y}|\mathbf{x} \in \mathcal{E}_v) P(\mathcal{E}_v|\boldsymbol{\lambda}), \end{aligned} \tag{S3}$$

$$P(\mathcal{E}_v|\boldsymbol{\lambda}) = \sum_{\mathbf{x} \in \mathcal{E}_v} P(\mathbf{x}|\boldsymbol{\lambda}). \tag{S4}$$

If a mixture distribution is identifiable, it means that identical distributions must come from the same set of weights on the mixture elements; in this context, $p(\mathbf{y}|\boldsymbol{\lambda}) \sim p(\mathbf{y}|\boldsymbol{\lambda}') \Rightarrow P(\mathcal{E}_v|\boldsymbol{\lambda}) = P(\mathcal{E}_v|\boldsymbol{\lambda}')\ \forall v$. Many finite mixtures (what we practically have in MMVP) and countably infinite mixtures with common noise distributions are identifiable [2-3], and we assume that the system noise characteristics lend to an identifiable mixture. However, we need to prove the identifiability of MMVP, or that equal mixture weights implies equal Poisson parameters: $P(\mathcal{E}_v^g|\boldsymbol{\lambda}) = P(\mathcal{E}_v^g|\boldsymbol{\lambda}')\ \forall v, g \Rightarrow \boldsymbol{\lambda} = \boldsymbol{\lambda}'$. Note that because $\mathcal{E}_{\mathbf{x}}^g \subseteq \mathcal{C}_{\mathbf{x}}^g$ and unique collision sets are disjoint, $P(\mathcal{E}_v^g|\boldsymbol{\lambda}) = P(\mathcal{E}_v^g|\boldsymbol{\lambda}')\ \forall v, g \Rightarrow P(\mathcal{C}_u^g|\boldsymbol{\lambda}) = P(\mathcal{C}_u^g|\boldsymbol{\lambda}')\ \forall u, g$.

We assume $P(\mathcal{C}_u^g|\boldsymbol{\lambda}) = P(\mathcal{C}_u^g|\boldsymbol{\lambda}')\ \forall u, g$ and prove the implication of $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$ given a set of monotonic sensors and with analytes exhibiting responsiveness and fingerprint equivalence in all G groups. We first focus on what can be concluded from a single, arbitrary sensor group (dropping the g superscript) with analytes potentially having nonunique fingerprints, and then we conclude with how multiple groups can be pooled to achieve identifiability. Again, note that our analysis will focus on monotonic increasing sensors without loss of generality.

Define $\mathcal{A} \subseteq \{1, ..., N\}$ such that analytes indexed by $a \in \mathcal{A}$ are all responsive such that there exists some m such that $f_m(\mathbf{e}_a) > f_m(\mathbf{0})$. Define the complementary set $\mathcal{B}$ with nonresponding analytes indexed by b.

Lemma S2.1. *If f_m is monotonic increasing for all $m \in \{1, ..., M\}$, and if only the analytes indexed by $a \in \mathcal{A} \subseteq \{1, ..., N\}$ are responsive, then $f(\mathbf{x}) = f(\mathbf{0})$ if and only if $\mathbf{x}_{\mathcal{A}} = \mathbf{0}$.*

Proof. Consider $\mathbf{z}$ such that $\mathbf{z}_{\mathcal{A}} = \mathbf{0}$. Note that $\text{supp}(\mathbf{z}) \subseteq \mathcal{B}$. Because analytes indexed by $b \in \mathcal{B}$ are nonresponding, $f(\mathbf{z}) = f(\mathbf{0})$ by definition. Next, we prove the forward condition, $f(\mathbf{x}) = f(\mathbf{0}) \Rightarrow \mathbf{x}_{\mathcal{A}} = \mathbf{0}$, by contradiction. Say $f(\mathbf{z}) = f(\mathbf{0})$ and let $\mathbf{z}$ satisfy $z_a \geq 1$ for some $a \in \mathcal{A}$. By definition of $\mathcal{A}$, $f(\mathbf{e}_a) > f(\mathbf{0})$, and $\mathbf{z} \succeq \mathbf{e}_a$. With monotonic functions, $f(\mathbf{z}) \succeq f(\mathbf{e}_a) \succ f(\mathbf{0})$ and we have arrived at a contradiction. □

The key concept to carry forward is that values of elements in $\mathbf{x}_\mathcal{B}$ are entirely arbitrary for the analysis of collision sets.

Lemma S2.2. *Let f_m be monotonic increasing for all $m \in \{1, ..., M\}$, and let only the analytes indexed by $a \in \mathcal{A} \subseteq \{1, ..., N\}$ be responsive. If $P(\mathcal{C}_\mathbf{0}|\boldsymbol{\lambda}) = P(\mathcal{C}_\mathbf{0}|\boldsymbol{\lambda}')$, then $\sum \boldsymbol{\lambda}_\mathcal{A} = \sum \boldsymbol{\lambda}'_\mathcal{A}$.*

Proof. By Lemma S2.1, $\mathcal{C}_\mathbf{0}$ contains all $\mathbf{x}$ with $\mathbf{x}_\mathcal{A} = \mathbf{0}$ with arbitrary values on $\mathbf{x}_\mathcal{B}$. Therefore, $P(\mathcal{C}_\mathbf{0}|\lambda) = P(\mathcal{C}_\mathbf{0}|\lambda')$ implies

$$P(\mathbf{x}_\mathcal{A} = \mathbf{0}|\boldsymbol{\lambda}) = P(\mathbf{x}_\mathcal{A} = \mathbf{0}|\boldsymbol{\lambda}') \tag{S5}$$

$$e^{-\sum \boldsymbol{\lambda}_\mathcal{A}} = e^{-\sum \boldsymbol{\lambda}'_\mathcal{A}} \tag{S6}$$

$$\sum \boldsymbol{\lambda}_\mathcal{A} = \sum \boldsymbol{\lambda}'_\mathcal{A}. \tag{S7}$$

□

Lemma S2.3. *Let f_m be monotonic increasing for all $m \in \{1, ..., M\}$, and let only the analytes indexed by $a \in \mathcal{A} \subseteq \{1, ..., N\}$ be responsive. For $a \in \mathcal{A}$, if for all $\mathbf{x} \in \mathcal{C}_{\mathbf{e}_a}$, x_a is the only nonzero value in $\mathbf{x}_\mathcal{A}$, then $\lambda_a = \lambda'_a$.*

Proof. We assume $P(\mathcal{C}_{\mathbf{e}_a}|\boldsymbol{\lambda}) = P(\mathcal{C}_{\mathbf{e}_a}|\boldsymbol{\lambda}')$. By definition of $\mathcal{A}$, $f(\mathbf{e}_a) \succ f(\mathbf{0})$. If x_a is the only nonzero value of $x_\mathcal{A}$, we have $\mathcal{C}_{\mathbf{e}_a} = \{c\mathbf{e}_a \colon c \in \mathcal{K} \subseteq \{1, 2, \ldots\}\}$. Then, $P(\mathcal{C}_{\mathbf{e}_a}|\boldsymbol{\lambda}) = P(\mathcal{C}_{\mathbf{e}_a}|\boldsymbol{\lambda}')$ implies

$$e^{-\sum \boldsymbol{\lambda}_\mathcal{A}} \sum_{c \in \mathcal{K}} \frac{\lambda_a^c}{c!} = e^{-\sum \boldsymbol{\lambda}'_\mathcal{A}} \sum_{c \in \mathcal{K}} \frac{{\lambda'_a}^c}{c!}. \tag{S8}$$

Using Lemma S2.2, $\sum_{c \in \mathcal{K}} \frac{\lambda_a^c}{c!} = \sum_{c \in \mathcal{K}} \frac{{\lambda'_a}^c}{c!}$, which implies $\lambda_a = \lambda'_a$ since the function on both sides is monotonic in λ_a. □

From here, we first derive results for the special case where all analytes indexed by $\mathcal{A}$ have unique single-copy fingerprints. Afterwards, we generalize to multiple groups, allowing for equal nonzero fingerprints within a group. The next Lemma guarantees at least one index a to which Lemma S2.3 can be applied.

Lemma S2.4. *Let f_m be monotonic increasing for all $m \in \{1, ..., M\}$, and let only the analytes indexed by $a \in \mathcal{A} \subseteq \{1, ..., N\}$ be responsive. If $f(\mathbf{e}_i) \neq f(\mathbf{e}_j)$ $\forall i, j \in \mathcal{A}$ with $i \neq j$, $\exists a \in \mathcal{A}$ such that all $\mathbf{x} \in \mathcal{C}_{\mathbf{e}_a}$ are nonzero in $\mathbf{x}_\mathcal{A}$ only on index a.*

Proof. First, with unique nonzero fingerprints in $\mathcal{A}$ and monotonic sensors, the fingerprint responses $f_m(\mathbf{e}_a)$ can be sorted. Starting arbitrarily with $m = 1$, we can select the minimal set $\mathcal{M} \subseteq \mathcal{A}$ that minimizes $f_1(\mathbf{e}_a)$ such that $\forall a \in \mathcal{M}, \forall j \in \mathcal{M}^c$, $f_1(\mathbf{e}_a) < f_1(\mathbf{e}_j)$. If $|\mathcal{M}| > 1$, then the process can be repeated with $m = 2$ (and so forth) on the subset $\mathcal{M}$ until there is one unique minimum and its corresponding index a.

For this $\mathbf{e}_a$, all $i \in \mathcal{A} \setminus \{a\}$ satisfy $f_m(\mathbf{e}_i) > f_m(\mathbf{e}_a)$ for at least one m. Therefore, signals in the collision set $\mathbf{x} \in \mathcal{C}_{\mathbf{e}_a}$ must satisfy $\mathbf{x}_i = 0$ for all $i \in \mathcal{A} \setminus \{a\}$. Signals with at least one $\mathbf{x}_i \geq 1$ would have at least one m where $f_m(\mathbf{x}) > f_m(\mathbf{e}_a)$, and therefore not be in the collision set by definition. This conclusion completes the proof by contradiction. □

Next, we show how this result chains to all analytes indexed in $\mathcal{A}$.

Lemma S2.5. *Let f_m be monotonic increasing for all $m \in \{1, ..., M\}$, and let only the analytes indexed by $a \in \mathcal{A} \subseteq \{1, ..., N\}$ be responsive. If $f(\mathbf{e}_i) \neq f(\mathbf{e}_j)$ $\forall i, j \in \mathcal{A}$ with $i \neq j$, $\lambda_a = \lambda'_a$ $\forall a \in \mathcal{A}$.*

Proof. Lemmas S2.3 and S2.4 guarantee at least one a that yields $\lambda_a = \lambda'_a$. Let us call this index a_1 and define the subset $\mathcal{S} \subseteq \mathcal{A}$, the subset of indices for which $\lambda_i = \lambda'_i$ $\forall i \in \mathcal{S}$. At this point, $\mathcal{S} = \{a_1\}$. Repeating the process in the proof of Lemma S2.4, we can find a new index a_2 that satisfies $f_m(\mathbf{e}_n) > f_m(\mathbf{e}_{a_2})$ $\forall n \in \mathcal{S}^c \setminus \{a_2\}$ for at least one m.

For the direct proof of identifiability, we assume $P(\mathcal{C}_{\mathbf{e}_{a_2}}|\boldsymbol{\lambda}) = P(\mathcal{C}_{\mathbf{e}_{a_2}}|\boldsymbol{\lambda}')$, or:

$$\sum_{\mathbf{x}\in\mathcal{C}_{\mathbf{e}_{a_2}}} P(\mathbf{x}|\boldsymbol{\lambda}) = \sum_{\mathbf{x}\in\mathcal{C}_{\mathbf{e}_{a_2}}} P(\mathbf{x}|\boldsymbol{\lambda}'). \tag{S9}$$

Among signals $\mathbf{x}$ in $\mathcal{C}_{\mathbf{e}_{a_2}}$, $x_n = 0$ for $n \in \mathcal{S}^c \setminus \{a_2\}$ because $f_m(\mathbf{e}_n) > f(\mathbf{e}_{a_2})$ for some m, and sensors are monotonic. These signals can also be partitioned into those with $x_{a_2} = 0$, and those with $x_{a_2} \geq 1$. If any of the former type exist, then $x_i > 0$ for some of the indices $i \in \mathcal{S}$. For instance, we could have $f(2\mathbf{e}_i) = f(\mathbf{e}_{a_2})$ with $x_i = 2$. These signals' terms in the summation follow the form $\left[\prod_{n\in\mathcal{S}} P(x_n|\lambda_n)\right] e^{-\sum_{i\in\mathcal{S}^c}\lambda_i}$. Note that $\lambda_i = \lambda'_i\ \forall i \in \mathcal{S}$ combined with Lemma S2.2 yields $\sum_{i\in\mathcal{S}^c}\lambda_i = \sum_{i\in\mathcal{S}^c}\lambda'_i$. Because $\lambda_i = \lambda'_i$ for $i \in \mathcal{S}$, the product component is equal on both sides as well. Therefore, these terms can be eliminated in Equation (S9). We will denote the set of remaining $\mathbf{x}$ in the summation as $\mathcal{C}'$.

In $\mathcal{C}'$, we have $x_{a_2} \in \mathcal{K} \subseteq \{1, 2, ...\}$. Therefore:

$$e^{\sum\boldsymbol{\lambda}_{\mathcal{A}}} \sum_{\mathbf{x}\in\mathcal{C}_{\mathbf{e}_{a_2}}} \prod_{n\in\mathcal{S}\cup\{a_2\}} \frac{\lambda_n^{x_n}}{x_n!} = e^{\sum\boldsymbol{\lambda}'_{\mathcal{A}}} \sum_{\mathbf{x}\in\mathcal{C}_{\mathbf{e}_{a_2}}} \prod_{n\in\mathcal{S}\cup\{a_2\}} \frac{{\lambda'_n}^{x_n}}{x_n!} \tag{S10}$$

$$\sum_{k\in\mathcal{K}} \frac{\lambda_{a_2}^k}{k!} h_k(\boldsymbol{\lambda}_{\mathcal{S}}) = \sum_{k\in\mathcal{K}} \frac{{\lambda'_{a_2}}^k}{k!} h_k(\boldsymbol{\lambda}'_{\mathcal{S}}), \tag{S11}$$

where h_k is the mapping $\boldsymbol{\lambda}_{\mathcal{S}} \mapsto \sum_{\mathbf{x}\in\mathcal{C}_{\mathbf{e}_{a_2}}:\, x_{a_2}=k} \prod_{n\in\mathcal{S}} \frac{\lambda_n^{x_n}}{x_n!}$. Because $\boldsymbol{\lambda}_{\mathcal{S}} = \boldsymbol{\lambda}'_{\mathcal{S}}$, we can replace both $h_k(\boldsymbol{\lambda}_{\mathcal{S}})$ and $h_k(\boldsymbol{\lambda}'_{\mathcal{S}})$ by constants H_k. Therefore,

$$\sum_{k\in\mathcal{K}} \frac{\lambda_{a_2}^k}{k!} H_k = \sum_{k\in\mathcal{K}} \frac{{\lambda'_{a_2}}^k}{k!} H_k. \tag{S12}$$

Because $H_k \geq 0$, both sides are monotonic in λ_{a_2} such that $\lambda_{a_2} = \lambda'_{a_2}$. Now, a_2 can be added to $\mathcal{S}$ and the process can be repeated until $\mathcal{S} = \mathcal{A}$ such that $\boldsymbol{\lambda}_{\mathcal{A}} = \boldsymbol{\lambda}'_{\mathcal{A}}$. □

Now, we will extend this result to the case of having equal fingerprints in the same sensor group—i.e., that $f(\mathbf{e}_i) = f(\mathbf{e}_j)$ for some pairs i, j.

Lemma S2.6. *Let f_m be monotonic increasing for all $m \in \{1, ..., M\}$, and let only the analytes indexed by $a \in \mathcal{A} \subseteq \{1, ..., N\}$ be responsive. Define the disjoint sets $\mathcal{A}_1, \mathcal{A}_2, ...\mathcal{A}_C$ indexed by c with $\cup_{c=1}^{C}\mathcal{A}_c = \mathcal{A}$ such that for all $i, j \in \mathcal{A}_c$, $f(\mathbf{e}_i) = f(\mathbf{e}_j)$. Then, $\sum\boldsymbol{\lambda}_{\mathcal{A}_c} = \sum\boldsymbol{\lambda}'_{\mathcal{A}_c}\ \forall c$.*

Proof. Note that the Poisson distribution has the property that if x_i are each independently drawn from Poisson(λ_i), then $\sum_{i\in\mathcal{A}_c} x_i \sim$ Poisson($\sum\boldsymbol{\lambda}_{\mathcal{A}_c}$). We can then simply define a dummy variables $\mathbf{x}_c^\dagger = \sum_{i\in\mathcal{A}_c} x_i$ and $\boldsymbol{\lambda}_c^\dagger$ such that $\mathbf{x}_c^\dagger \sim$ Poisson($\boldsymbol{\lambda}_c^\dagger$). This dummy variable represents an "analyte" that appears with a distribution governed by the total quantities of analytes with the same fingerprint. However, what matters for identifiability is the sensor functional values of signals, i.e. that $f(\mathbf{a}) = f(\mathbf{b})$ if for all c, $\sum\mathbf{a}_{\mathcal{A}_c} = \sum\mathbf{b}_{\mathcal{A}_c}$. Namely, $\mathbf{x}_c^\dagger \sim$ Poisson($\boldsymbol{\lambda}_c^\dagger$) by fundamental properties of the Poisson distribution, but it is only with the condition of fingerprint equivalence (Definition S2.3) that lets us apply all previous results that are based on collision sets, i.e., sets of signals with equal functional values. These yield $\boldsymbol{\lambda}_c^\dagger = {\boldsymbol{\lambda}'_c}^\dagger\ \forall c$, or $\sum\boldsymbol{\lambda}_{\mathcal{A}_c} = \sum\boldsymbol{\lambda}'_{\mathcal{A}_c}\ \forall c$. □

Theorem S2.7. *Let g index G different sensor groups that satisfy fingerprint equivalence and that contain monotonic, saturating sensors. For each group g, define the row vector $\mathbf{z}_c^g$ of zeros and ones with ones in the indices associated with $\mathcal{A}_c$. Define the N-column matrix $\mathbf{Z}_g$ whose C rows are comprised of $\mathbf{z}_c^g\ \forall c$. Define the N-column matrix $\mathbf{Z}$ as the vertical concatenation $\mathbf{Z}_g\ \forall g$. If $\text{rank}(\mathbf{Z}) = N$, then $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$.*

Proof. This theorem is a formal way of saying that Lemma S2.6 must yield N independent equations when applied to all groups where the sensing and system conditions hold. We can consider the system of equations yielded by Lemma S2.6 and represented by $\mathbf{Z}\boldsymbol{\lambda} = \mathbf{Z}\boldsymbol{\lambda}'$, or $\mathbf{Z}(\boldsymbol{\lambda} - \boldsymbol{\lambda}') = \mathbf{0}$. If $\text{rank}(\mathbf{Z}) = N$, then it follows that $\boldsymbol{\lambda} = \boldsymbol{\lambda}'$. Therefore, we have $P(\mathcal{C}_u^g|\boldsymbol{\lambda}) = P(\mathcal{C}_u^g|\boldsymbol{\lambda}')\ \forall(u, g) \Rightarrow \boldsymbol{\lambda} = \boldsymbol{\lambda}'$, concluding our proof of identifiability. □

S3 Special Cases of MLE with ddPCR

From eq 2 in the main text that describes the generalized gradient in MLE, we consider two commonly employed special cases. First, if samples are sufficiently dilute such that partitions are either empty ($\mathbf{x}_d = \mathbf{0}$) or have only one analyte, the goal is often to identify each nonzero signal independently with a classification process. In other words, assays must be designed such that $p(\mathbf{y}_d|\mathbf{x}) > 0$ *only* for $\mathbf{x}_d^*$ - the measurements are *unambiguous*. Setting the gradient equal to zero and simplifying leads to $\hat{\boldsymbol{\lambda}}_{MLE} = \frac{1}{D}\sum_{d=1}^{D}\mathbf{x}_d^*$. In practice, clusters of classes must have reliable decision boundaries and concentrations are estimated by totaling the classification results.

The second specialized case is common for ddPCR where for each PCR assay is specific for a target analyte and assigned to a particular channel. With M channels, $N = M$ and each measurement unambiguously determines the presence or absence of each target sequence. Precisely, $p(y_d|\mathbf{x})$ is one or zero, and considerations of each analytes' quantity x_n can be simplified to $x_n = 0$ (absent) or $x_n > 0$ (present). Each analyte n can be inferred independently such that $\lambda_n = -\log\frac{D_{0,n}}{D}$ where $D_{0,n}$ is the number of droplets that do *not* contain analyte n. This formula can be found by applying the above assumptions, setting eq 2 in the main text to zero, and simplifying.

S4 Exact Gradient Computation and $p(\mathbf{y}|\mathbf{x})$ Model for ddPCR

We first focus on the gradient resulting from a single probe group. In a single group, there are only four viable measurements with $\mathbf{y} \in \{0,1\}^2$. Let us define $\mathcal{Y} = \{0,1\}^2$, and $p_{\mathbf{y}}$ as the proportion of the D measurements that equal $\mathbf{y}$. We can then re-express the log-likelihood maximization as:

$$\widehat{\boldsymbol{\lambda}}_{MLE} = \arg\max_{\boldsymbol{\lambda}} \frac{1}{D}\sum_{d=1}^{D}\log\sum_{\mathbf{x}\in\mathbb{Z}_+^N} p(\mathbf{y}_d|\mathbf{x})P(\mathbf{x}|\boldsymbol{\lambda}) \tag{S13}$$

$$= \arg\max_{\boldsymbol{\lambda}} \sum_{\mathbf{y}\in\mathcal{Y}} p_{\mathbf{y}}\log\sum_{\mathbf{x}\in\mathbb{Z}_+^N} p(\mathbf{y}_d|\mathbf{x})P(\mathbf{x}|\boldsymbol{\lambda}) \tag{S14}$$

Here, we will use $\boldsymbol{\lambda} \equiv \boldsymbol{\lambda}^{(bact)}$ since we will be optimizing over the bacterial concentrations directly. In our case, *E. cloacae* is the only bacteria with a fractional abundance of a probe binding site - approximately 87.5% of its copies interact with probes 1, 3, and 4, and 12.5% interact with only probes 3 and 4 (Figure S3). Similarly to how we defined $\mathbf{C}$ in the Results and Discussion, we can define $\mathbf{C}^{(g)}$ for group g with each bacterium's fractional abundances of genes with a corresponding barcode. Figure S3 shows an example of $\mathbf{C}^{(1)}$, which can be generated with Figure 1 as a reference.

We define $p(\mathbf{y}|\mathbf{x}) = \prod_m p(y_m|\mathbf{x})$. For $p(y_m = 1|\mathbf{x})$, then $p(y_m|\mathbf{x}) = 1$ if the droplet has at least one copy of a gene that interacts with probe m and $p(y_m|\mathbf{x}) = 0$ otherwise. For $p(y_m = 0|\mathbf{x})$, this likelihood is 1 if none of the genes in the droplet interact with probe m and 0 otherwise.

However, with the analyte currently defined as a copy of the nth bacterium's 16S gene, we must be careful. For instance, with index 3 corresponding with *E. cloacae*, if $x_3 = 1$ in $\mathbf{x}$, $p(y_1|\mathbf{x})$ may not be 1 since one copy of *E. cloacae*'s 16S gene is not guaranteed to interact with probe 1. To resolve this, we will temporarily transform the problem to the space of gene barcodes for this group: $\boldsymbol{\lambda}^{(BC_1)} = \mathbf{C}^{(1)}\boldsymbol{\lambda}$. Note $\boldsymbol{\lambda}^{(BC_1)}$ is 4-dimensional. We can define $\mathbf{x}^{(BC_1)} = [x_{00}, x_{01}, x_{10}, x_{11}]^T$ as the vector representing the quantities of 16S genes from any source bacteria that interact with the probes in the pattern noted in the subscript, noting $\mathbf{x}^{(BC_1)} \sim \text{Poisson}(\boldsymbol{\lambda}^{(BC_1)})$. Lastly, let us define $\mathcal{X}_{\mathbf{y}}$ as that set where if $\mathbf{x}^{(BC_1)} \in \mathcal{X}_{\mathbf{y}}$, then $p(\mathbf{y}|\mathbf{x}^{(BC_1)}) = 1$. Now we can rewrite Equation (S14) as:

$$= \sum_{\mathbf{y}\in\mathcal{Y}} p_{\mathbf{y}}\log\sum_{\mathbf{x}^{(BC_1)}\in\mathcal{X}_{\mathbf{y}}} P(\mathbf{x}^{(BC_1)}|\boldsymbol{\lambda}^{(BC_1)}). \tag{S15}$$

The linearity of gradients allow us to treat this one $\mathbf{y}$ at a time, summing the contributions from each $\mathbf{y}$ at the end. In general, treat 00 as short for $[0,0]$, 01 for $[0,1]$, etc. Let $\nabla_{\boldsymbol{\lambda}}^{00}$ be the component of the gradient from $\mathbf{y} = [0,0]$, $\nabla_{\boldsymbol{\lambda}}^{01}$ from $\mathbf{y} = [0,1]$, etc. We will similarly define the mean log likelihood contributions as

ℓ^{00}, ℓ^{01}, etc. Similarly, define the rows of $\mathbf{C}^{(1)}$ as $\mathbf{C}^{(1)}_{00}, \mathbf{C}^{(1)}_{01}$, etc. By convention, $\boldsymbol{\lambda}$ and other vectors should be assumed to be column vectors, but the rows of $\mathbf{C}^{(1)}$ are row vectors. Thus we have:

$$\begin{aligned}
\ell^{00} &= p_{00} \log P(x_{01} = 0 \text{ and } x_{10} = 0 \text{ and } x_{11} = 0) && \text{(S16)}\\
&= p_{00} e^{-x_{01} - x_{10} - x_{11}} && \text{(S17)}\\
&= p_{00}(-\mathbf{C}^{(1)}_{01} - \mathbf{C}^{(1)}_{10} - \mathbf{C}^{(1)}_{11})\boldsymbol{\lambda} && \text{(S18)}
\end{aligned}$$

$$\nabla^{00}_{\boldsymbol{\lambda}} = p_{00}(-\mathbf{C}^{(1)}_{01} - \mathbf{C}^{(1)}_{10} - \mathbf{C}^{(1)}_{11})^T. \quad \text{(S19)}$$

In the first line, we define the conditions for $\mathbf{x}^{(BC_1)} \in \mathcal{X}_{00}$ and solve. Genes that interact with either probe cannot be in droplets that yield $\mathbf{y} = [0, 0]$. Next, for the $\mathbf{y} = [0, 1]$ response, at least one gene that interacts with the 2nd (FAM) probe must be present, and genes that interact with the HEX probe must be absent.

$$\begin{aligned}
\ell^{01} &= p_{01} \log P(x_{01} \geq 1 \text{ and } x_{10} = 0 \text{ and } x_{11} = 0) && \text{(S20)}\\
&= p_{01} \log(1 - e^{-x_{01}}) e^{-x_{10} - x_{11}} && \text{(S21)}\\
&= p_{01} \left(\log(1 - e^{-\mathbf{C}^{(1)}_{01}\boldsymbol{\lambda}}) - \mathbf{C}^{(1)}_{10}\boldsymbol{\lambda} - \mathbf{C}^{(1)}_{11}\boldsymbol{\lambda} \right) && \text{(S22)}
\end{aligned}$$

$$\nabla^{01}_{\boldsymbol{\lambda}} = p_{01} \left[\left(\frac{e^{-\mathbf{C}^{(1)}_{01}\boldsymbol{\lambda}}}{1 - e^{-\mathbf{C}^{(1)}_{01}\boldsymbol{\lambda}}} \right) \mathbf{C}^{(1)T}_{01} - \mathbf{C}^{(1)T}_{10} - \mathbf{C}^{(1)T}_{11} \right]. \quad \text{(S23)}$$

A virtually identical simplification for $\nabla^{10}_{\boldsymbol{\lambda}}$ is omitted here. Lastly, for $\mathbf{y} = [1, 1]$, we have:

$$\begin{aligned}
\ell^{11} &= p_{11} \log(P(x_{11} \geq 1) + P(x_{11} = 0 \text{ and } x_{01} > 0 \text{ and } x_{10} > 0) && \text{(S24)}\\
&= p_{11} \log \left((1 - e^{-x_{11}}) + e^{-x_{11}}(1 - e^{-x_{01}})(1 - e^{-x_{10}}) \right). && \text{(S25)}
\end{aligned}$$

The remaining algebraic steps are omitted, but the final result is

$$\nabla^{11}_{\boldsymbol{\lambda}} = p_{11} e^{-x_{11}} \frac{-\mathbf{C}^{(1)T}_{11} - e^{-x_{01}}(-\mathbf{C}^{(1)T}_{01} - \mathbf{C}^{(1)T}_{11}) - e^{-x_{10}}(-\mathbf{C}^{(1)T}_{10} - \mathbf{C}^{(1)T}_{11}) - e^{-x_{01} - x_{10}}(-\mathbf{C}^{(1)T}_{01} - \mathbf{C}^{(1)T}_{10} - \mathbf{C}^{(1)T}_{11})}{(1 - e^{-x_{11}}) + e^{-x_{11}}(1 - e^{-x_{01}})(1 - e^{-x_{10}})} \quad \text{(S26)}$$

where $\mathbf{C}^{(1)}_{ij}\boldsymbol{\lambda}$ can be substituted for any x_{ij}.

We can now say that for group 1, $\nabla^{(1)}_{\boldsymbol{\lambda}} = \nabla^{00}_{\boldsymbol{\lambda}} + \nabla^{01}_{\boldsymbol{\lambda}} + \nabla^{10}_{\boldsymbol{\lambda}} + \nabla^{11}_{\boldsymbol{\lambda}}$. The above process can be repeated for any group g. Therefore, the final gradient vector (arbitrarily scaled) is

$$\nabla_{\boldsymbol{\lambda}} = \sum_g p_g \nabla^{(g)}_{\boldsymbol{\lambda}} \quad \text{(S27)}$$

where p_g is the proportion of total droplets that come from group g.

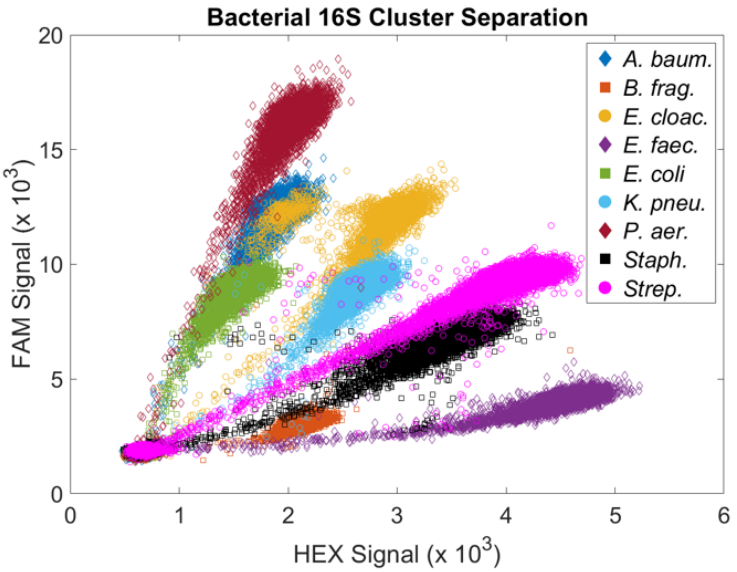


Figure S1: Separation of bacterial barcodes with amplitude multiplexing. Each cluster depicted is from a separate ddPCR reaction with one bacterial species in it. Data from the three *Staphylococcus* bacteria and the two *Streptococcus* bacteria were combined in this plot. Amplitude multiplexing is a technique to resolve more probes than the available number of color channels, but it is typically used with each unique probe participating in an orthogonal assay with its own primer pair. Here, we adjusted probe concentrations to "move" the cluster positions with a single pair of primers. Probes 1 and 5 were tagged with HEX, and Probes 2-4 were tagged with FAM. Probes 1, 2, and 4 at 125 nM, and with Probes 3 and 5 at 250 nM. Based on each 16S gene's barcode, droplets containing that gene will position in clusters whose channel intensities roughly correlate with the total probe concentration tagged with the corresponding fluorophore.

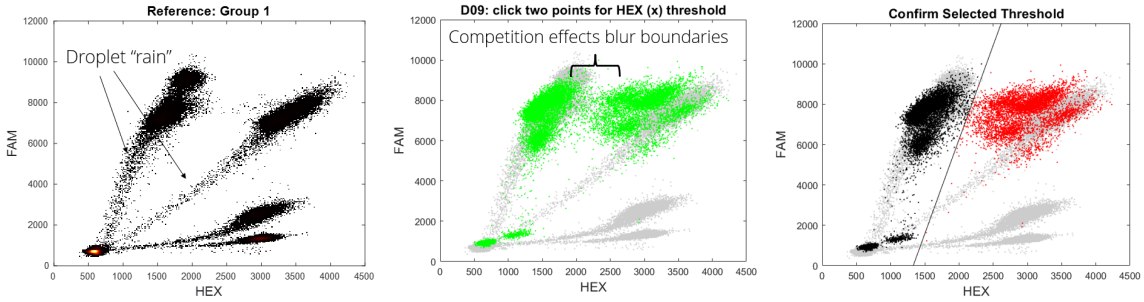
		K = 2 samples				**K = 3 samples**				**K = 4 samples**			
		1	2	3	4	5	6	7	8	9	10	11	12
Conc. #1	A	*A baum*	*B frag*	*E cloac*	*E faec*	*E coli*	*K pneu*	*P aeru*	*S aur*	*S epid*	*S sapr*	*S agal*	*S pneu*
Conc #2	B	*S pneu*	*S agal*	*S sapr*	*S epid*	*S aur*	*P aeru*	*K pneu*	*E coli*	*E faec*	*E cloac*	*B frag*	*A baum*
	C	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1	S1
	D	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2	S2
	E	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3	S3
	F	S4	S4	S4	S4	S4	S4	S4	S4	S4	S4	S4	S4
	G	S5	S5	S5	S5	S5	S5	S5	S5	S5	S5	S5	S5
	H	NTC	NTC	NTC	NTC	S6	S6	S6	S6	S6	S6	S6	S6
		G1	G2	G3	G4	G1	G2	G3	G4	G1	G2	G3	G4

Figure S2: Plate layout for ddPCR test samples. The first two rows served as references for ground truth concentration estimation of monomicrobial dilutions and manual thresholding of all wells. The colors of the wells in rows A and B correspond to the probe group applied. Random mixtures of bacteria were distribtued across the rest of the plate with each mixture being applied to four wells, each with a different subset of two probes defining the 16S barcodes.

$$\mathbf{C}^{(1)} =$$

	A. baum.	*B. frag.*	*E. cloac.*	*E. faec.*	*E. coli*	*K. pneu.*	*P. aeru.*	*Staph.*	*Strep.*
[0,0]	0	0	0	0	0	0	0	1	0
[0,1]	1	0	.125	0	1	0	1	0	1
[1,0]	0	1	0	1	0	0	0	0	0
[1,1]	0	0	.875	0	0	1	0	0	0

Figure S3: Example of partial barcode matrix for Group 1. *E. cloacae*'s amplicons appeared to always interact with Probes 3 and 4, but a small subcluster appeared to lack the HEX response to Probe 1 (Fig. S1a). We used our SPoRe algorithm to estimate the barcode abundances in reference wells A3 and B10 (Fig. S2) which both contained Probe 1. After manual thresholding, the binarized data and "analytes" with barcodes $[0,1], [1,0]$, and $[1,1]$ (ordered as [HEX, FAM]) were passed to SPoRe, and SPoRe estimated the abundances of the amplicons with these responses. In this case, the $[1,0]$ quantity was nearly zero, consistent with the expectation that *E. cloacae* always interacted with a FAM probe. The fraction of amplicons with the HEX probe was determined by $\lambda_{[1,1]}/(\lambda_{[1,1]} + \lambda_{[0,1]})$. In A3 and B10, these were estimated to be 0.832 and 0.828, respectively. Our sequence analysis found eight 16S copies in the *E. cloacae* genome, so it is possible that one amplicon had a sequencing error such that Probe 1 truly binds to 7/8 copies. Therefore, column 3 of matrix **C** had 0.875 of barcode [1,1] and 0.125 of barcode [0,1] for Groups 1 and 3.



(a) **(b)** **(c)**

Figure S4: Example process for manual thresholding with noted challenges. **(a)** All reference data (rows A and B in Figure S2) from the same probe group was pooled and displayed to serve as a visual reference. Droplet "rain" is evident in each cluster. Due to some mild "lean" and "lift" of the raw ddPCR clusters caused likely by partial probe interactions, we allowed any linear threshold for each fluorescence channel determined by two user-selected points. **(b)** Raw data from a polymicrobial sample was overlaid on the reference data with the same corresponding probe group. An example is shown with the raw data from D09 (Group 1, $k = 4$ sample number 2) overlaid with the Group 1 reference data. PSC effects, as expected, create subpopulations of droplet measurements between the binary clusters. We speculate that the small additional cluster near zero is due to droplets where probes partially interacted with amplicons due to imperfect sequence homology. **(c)** After the user selects two points to define a line for thresholding, the plot is updated to allow the user to visually confirm the results. Red points are assigned the value 1, and black points are assigned 0.

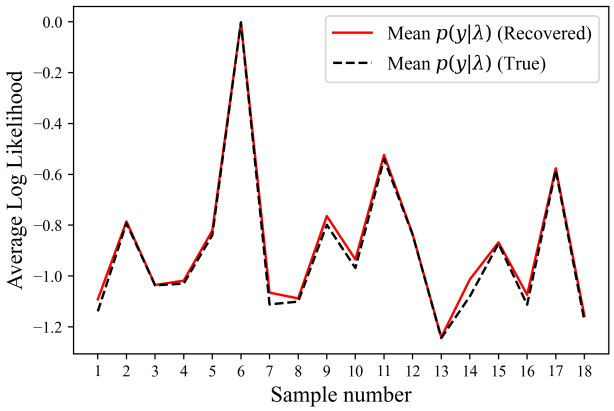


Figure S5: Likelihood comparison of SPoRe's solution against the estimated ground truth. SPoRe's solution exhibits higher average likelihood for the pre-binarized data that it is given.

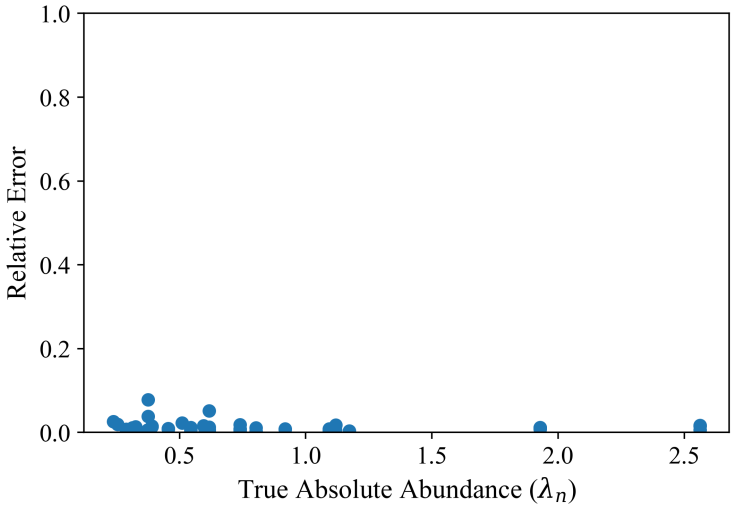


Figure S6: SPoRe's performance on simulations of experimental concentrations. Given the estimated ground truth concentrations ($\boldsymbol{\lambda}^*$), we simulated binary measurement data to pass to SPoRe. SPoRe returns virtually perfect results with mean cosine similarity of 0.9999. Here, we plot the relative error in estimates of the absolute abundance for each bacteria, represented by $|\hat{\lambda}_n - \lambda_n^*|/\lambda_n^*$ when running SPoRe on the simulated measurements.

References

[1] Xing, P. E.; Ng, A. Y.; Jordan, M. I.; Russell, S. Distance Metric Learning, with Application to Clustering with Side-Information. *Adv. Neural Inf. Process. Syst.* **2002**, 521-528

[2] Yakowitz, S. J.; Spragins, J. D. On the Identifiability of Finite Mixtures. *Ann. Math. Statist.* **1968**, *39* (1), 209-214

[3] Yang, L.; Xu, W. A New Sufficient Condition for Identifiability of Countably Infinite Mixtures. *Metrika*. **2013**, *77*, 377-387.