

Universal microbial diagnostics using random DNA probes

Amirali Aghazadeh,^{1*} Adam Y. Lin,^{1*} Mona A. Sheikh,^{1*} Allen L. Chen,¹ Lisa M. Atkins,² Coreen L. Johnson,² Joseph F. Petrosino,² Rebekah A. Drezek,¹ Richard G. Baraniuk^{1†}

2016 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC). 10.1126/sciadv.1600025

Early identification of pathogens is essential for limiting development of therapy-resistant pathogens and mitigating infectious disease outbreaks. Most bacterial detection schemes use target-specific probes to differentiate pathogen species, creating time and cost inefficiencies in identifying newly discovered organisms. We present a novel universal microbial diagnostics (UMD) platform to screen for microbial organisms in an infectious sample, using a small number of random DNA probes that are agnostic to the target DNA sequences. Our platform leverages the theory of sparse signal recovery (compressive sensing) to identify the composition of a microbial sample that potentially contains novel or mutant species. We validated the UMD platform *in vitro* using five random probes to recover 11 pathogenic bacteria. We further demonstrated *in silico* that UMD can be generalized to screen for common human pathogens in different taxonomy levels. UMD's unorthodox sensing approach opens the door to more efficient and universal molecular diagnostics.

INTRODUCTION

The accurate, efficient, and rapid identification of microbial organisms, such as bacteria and viruses, is of mounting importance in the fields of health care, environmental monitoring, defense, and beyond (1–3). Sepsis from bacterial infection is currently the 11th leading cause of death in the United States, and the mortality rate of bloodstream infections is high (14 to 50%) (4–6).

Conventional strategies for microbial detection are based on microbe-specific genomic or proteomic markers and protocols. Polymerase chain reaction (PCR)-based approaches rely on the binding of specific capture probes with unique genomic identifiers, such as the 16S ribosomal DNA (rDNA) subunit in bacteria. Although these methods show promise as highly specific tools for microbial identification (7, 8), they have limitations in clinical, industrial, and defense settings (9, 10). In the case of an epidemic, the detection of a newly mutated species using current PCR methods would require entirely new capture probes to be manufactured, introducing additional costs and delays. For bacterial detection, blood cultures typically require 48 to 72 hours to produce reliable results (11–14). During this waiting period, administration of broad-spectrum antibiotics breeds further threats of bacterial resistance and missed coverage (15). DNA microarrays also require many target-specific probes to detect multiple pathogens and lie dormant against unknown organisms. Whole-genome sequencing (WGS), currently the most complete and accurate technique, is not yet conducive to point-of-care diagnostics; it requires millions of expensive sequencing reads to assemble or align with genomic identifiers. It follows that there is a critical need for a new means of microbial detection: a universal (that is, works for bacteria outside of the target library), inexpensive (that is, requires minimal resources for acquisition, such as DNA probes and sequencing reads, etc.), and rapid sensing platform capable of identifying known and novel species with high phylogenetic power.

Here, we report on the design and validation of a new microbial diagnostic platform that satisfies the above desiderata. In common with

microarrays and PCR-based techniques, our Universal Microbial Diagnostics (UMD) platform exposes a microbial sample (which may contain more than one genus/species) to a collection of DNA probes. However, in sharp contrast to conventional methods, the probes are randomly generated (and hence target-agnostic) permutations of nucleotides that freely hybridize to different spots and to different extents on different bacterial genomes. By measuring the degree to which the sample hybridizes with the collection of random probes, we set up a statistical inverse problem to detect the presence and estimate the concentrations of the various bacteria in the sample. Using signal recovery techniques from the recently developed theory of compressive sensing (16, 17), we show below that it is possible to stably solve this inverse problem even when the number of probes is significantly smaller than the size of the library of possible bacteria of interest. Thanks to the random structure of the probes and the variabilities that bacterial organisms exhibit in their genomes, UMD is universal, inexpensive, rapid, and phylogenetically informative (random probes bind to arbitrary spots on the genome). Moreover, because of the universal nature of its probe design, UMD can classify not only known organisms but also novel mutants with their closest known relatives.

Universal Microbial Diagnostics

In UMD (Fig. 1A), the genomic DNA of an infectious sample is extracted and exposed to a small number M of DNA probes, which hybridize to the genomic DNA at various locations; this hybridization is experimentally quantified, producing a probe-binding (or hybridization) vector y whose entries correspond to the hybridization binding level of each probe with the microbial sample.

A priori, the hybridization binding level of each probe to a reference database of N bacterial genomes, is obtained and stored in an $M \times N$ hybridization affinity matrix (Fig. 1B). The hybridization affinity matrix can be either measured experimentally *in vitro* or predicted computationally *in silico*. Here, to speed up the probe design and prove the concept of UMD, we predict the affinity matrix using a thermodynamic model *in silico*. To compute the entry ϕ_{ij} in the matrix Φ (the hybridization binding level of probe i to genome j), we first perform a rapid

¹Rice University, Houston, TX 77005, USA. ²Baylor College of Medicine, Houston, TX 77030, USA. *These authors contributed equally to this work. †Corresponding author. Email: richb@rice.edu

thermodynamic alignment of the sequence of the probe to the sequence of the genome using the alignment model described in the study by SantaLucia and Hicks (18). Next, we extract sequence fragments from the genome sequence, which contain a significant hybridization affinity with the probe sequence. The fragment-probe mixture is then fed into a thermodynamics-based hybridization model (18). This model predicts all possible stable probe-bacteria fragment bindings along with their resulting concentrations for a given set of experimental conditions (Fig. 2, B and C). The overall hybridization affinity ϕ_{ij} is computed by summing the concentrations of all predicted and stable probe-fragment bindings for a unit concentration of bacterial genome.

Because of an excess concentration of probes as compared to sample DNA, the probe-binding vector \mathbf{y} can be closely approximated as a linear combination of the predicted hybridization affinities of the species in the reference genome database (the columns of the matrix) weighted by their concentrations \mathbf{x} ; that is, $\mathbf{y} = \Phi\mathbf{x} + \mathbf{n}$, where the vector \mathbf{n} accounts for noise and modeling errors (see Materials and Methods and Fig. 1C).

The two key capabilities of the UMD platform are to (i) detect the presence and (ii) estimate the concentrations \mathbf{x} of a potentially large number N of reference microbial genomes in an infectious sample given only a small number M of probe-binding measurements \mathbf{y} . Simply

inverting the matrix is impossible in this case because it has many more columns than rows. Fortunately, it is reasonable to assume that only a small number K of microbial genomes will be present in a given sample, in which case the concentration vector \mathbf{x} is sparse, with K nonzero and $N - K$ zero (or close to zero) entries; when $K < M$, one can hope to invert to estimate the K nonzero concentrations. More rigorously (see the Supplementary Materials for details), when the columns of Φ are sufficiently incoherent (close to orthogonal) and when $M = cK\log(N/K)$, where c is a small constant, we can apply the theory of compressive sensing (16, 17) to recover the concentrations \mathbf{x} from the measurements \mathbf{y} via a sparse optimization of the form

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \text{ subject to } \|\mathbf{y} - \Phi\mathbf{x}\|_2 < \sigma$$

Here, $\|\mathbf{x}\|_0$ counts the number of nonzero values in the vector \mathbf{x} , and σ bounds the energy of the noise vector \mathbf{n} . Because $M = cK\log(N/K)$ scales logarithmically with N , the UMD platform has the potential to identify and estimate the concentrations of a large number N of potential microbial genomes using only a small number M of measurement probes. To ensure that the columns of Φ are incoherent, we use DNA probes whose sequences are generated via a random permutation of

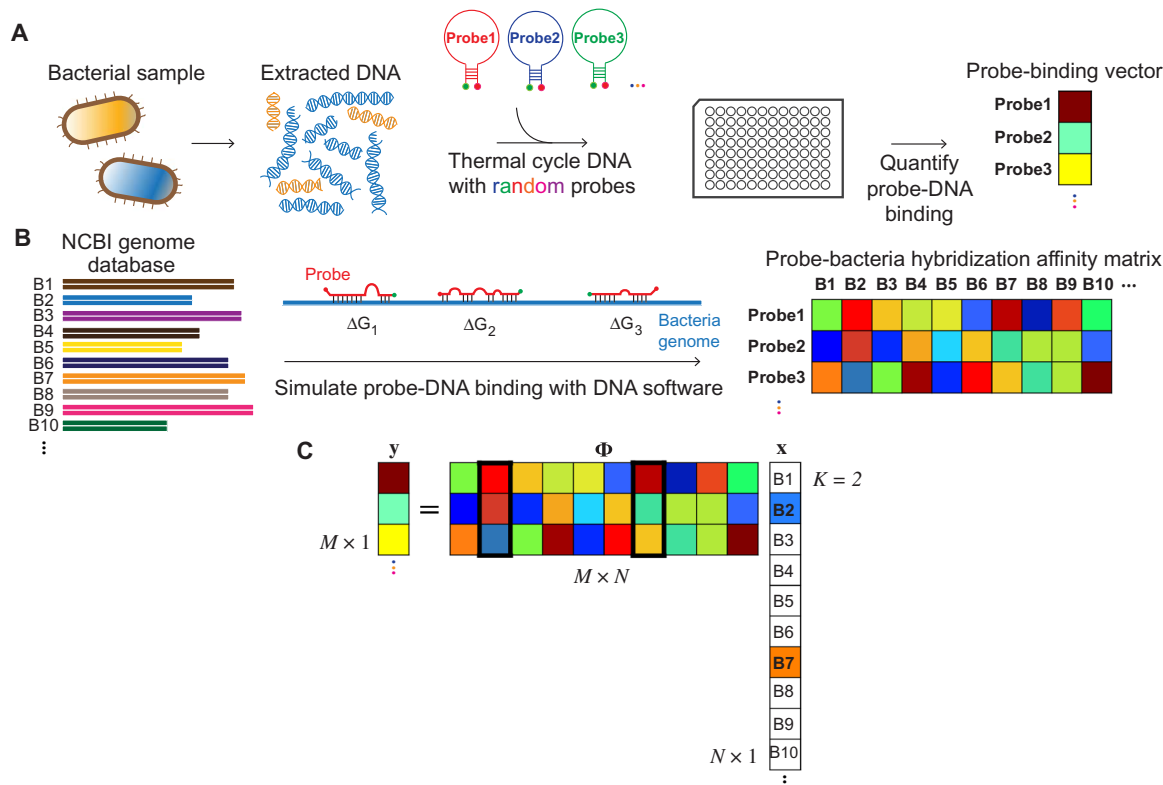


Fig. 1. Schematic of UMD platform. (A) Genomic DNA is extracted from a bacterial sample and thermal-cycled with M random DNA probes. The genome-probe binding is quantified, producing a probe-binding vector \mathbf{y} ; in this study, the random probes are in the form of MBs, and the DNA-probe binding is quantified by the ratio of open/hybridized to closed/nonhybridized MBs. (B) The hybridization binding level of each probe to a potentially large reference database of N bacterial genomes (B1, B2, ..., BN) is predicted using a thermodynamic model and stored in an $M \times N$ hybridization affinity matrix Φ . NCBI, National Center for Biotechnology Information. (C) Assuming that K bacterial species comprises the sample, the probe-binding vector \mathbf{y} is a sparse linear combination of the corresponding K columns of the matrix Φ weighted by the bacterial concentrations \mathbf{x} , that is, $\mathbf{y} = \Phi\mathbf{x} + \mathbf{n}$, where the vector \mathbf{n} accounts for noise and modeling errors. When K is small enough and M is large enough, Φ can be effectively inverted using techniques from compressive sensing, yielding the estimate for the microbial makeup of the sample \mathbf{x} ; in this illustration, the $K = 2$ bacteria-labeled B2 and B7 are present in the sample.

nucleotides (Fig. 2A). We demonstrate that a small fixed set of randomly selected probes induces sufficiently incoherent hybridization patterns across the columns of Φ and enables us to screen for a group of pathogenic organisms in vitro (Fig. 3). In addition, we showcase the average performance of several sets of randomly selected probes in universal pathogenic detection in silico (Fig. 4).

This universal sensing strategy can take on any physical embodiment (for example, quantitative real-time PCR, DNA microarray, or WGS) for the detection of any DNA sequence (bacterial, viral, or fungal). To test and validate the concept, we recovered pathogenic bacteria using random probes in the form of mismatch-tolerant sloppy molecular beacons (MBs) (19). In a conventional MB (20) for bacterial detection, the loop sequence is designed to target specific regions (for example, 16S rDNA) within a single bacterium (7) or multiple bacteria (19).

In the MB probes for UMD, the loop sequence is selected as a random sequence (Fig. 2A, fig. S1, and Materials and Methods) of length 38 nucleotides (nt), and the 4-nt-long stem sequence is consistent across all probes, although other choices might be used [different design trade-offs are discussed in the study by Sheikh (21)]. The unusually long loop and short stem enable our random probes to form hybrids with several base pair mismatches across the entire bacterial genome and compensate for the lower signal intensity in the absence of DNA amplification methods, such as PCR (Fig. 2C).

RESULTS

Experimental proof of concept

To prove the UMD concept, we mixed five UMD MBs [as shown in Fig. 2A and characterized in fig. S2 with guanine-cytosine (GC) contents 50, 56.5, 60.8, 50, and 52.7%, identical melting temperature

of 40°C, and concentration of 1 μ M] in separate tubes (to prevent cross hybridization of probes) with genomic DNA from each of nine human infectious bacterial strains grouped into three categories: (i) exact sequence known (*Escherichia coli*, *Francisella tularensis*, *Staphylococcus aureus*, *Campylobacter jejuni*, and *Proteus mirabilis*), (ii) exact sequence unknown (*Cupriavidus metallidurans* and *Micrococcus luteus*), and (iii) clinical isolates, whose exact sequence is unknown (*Bacteroides fragilis* and *Enterobacter aerogenes*) (the identification of *Pseudomonas aeruginosa* and *Bifidobacterium dentium* strains was tested using four random probes; see figs. S3 to S5 for detection results). For bacteria in groups II and III, the DNA sequences in the database might not exactly match the sequences present in the bacterial samples.

For each MB-bacterial species pair, equal volumes of probe and bacterial DNA were combined and subjected to a thermal cycling process of denaturing (95°C) and binding/cooling to 4°C overnight. To quantify probe-DNA binding, the MB probes' Cy3 and Cy5 fluorescence intensities were measured with a fluorometer, and the fluorescence resonance energy transfer (FRET) ratios (a decrease represents MB opening due to DNA binding) were calculated by computing Cy5 intensity over total fluorescence intensity (Cy3 + Cy5). The FRET ratios from binding of the nine bacteria to the MBs are depicted in Fig. 3A.

To estimate the bacterial concentrations in physical units, we translated the FRET ratio of each bacterium-MB pair into the concentration of opened MBs or hybridization affinity, represented in units of molarity. For this, we experimentally obtained and fitted FRET ratios for each of the five MBs as a function of the concentration of their exact probe complements, using an optimization method described in the study by Jeričević and Kušter (22) (see table S1 and Materials and Methods for the fit curve parameters and fit method, respectively, and fig. S2.). The coefficient of determination (R^2) values for the fits ranged from 0.97 to 0.99, suggesting a satisfactory fit. On the basis of the fit equations, the

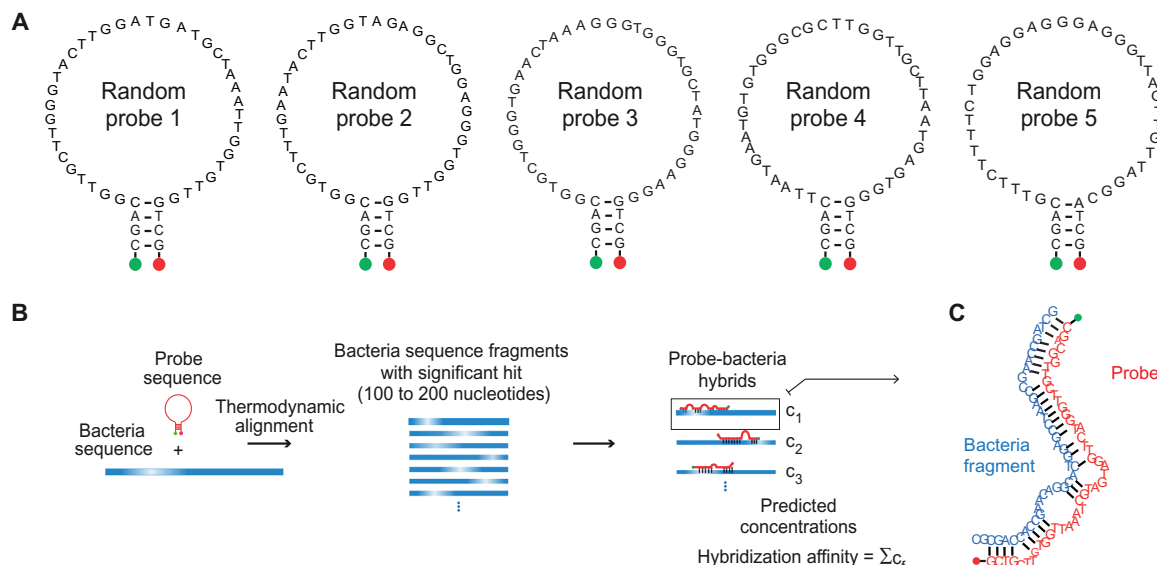


Fig. 2. Random probe design and hybridization affinity computation process. (A) DNA sequence structure of five test random DNA probes. (B) Both strands of the bacterial genome (blue lines) are first thermodynamically aligned with the probe sequence. The sequence of the bacteria is segmented into fragments of roughly equal length (~100 to 200 nt), each containing a significant hybridization affinity with the probe. Then, all of the bacterial fragments and probe sequences along with the experimental conditions are fed into the DNA software (18) to predict all stable probe-bacteria complexes and concentrations. These concentrations, in aggregate, determine the concentration of opened MBs, which is defined as the hybridization affinity of the probes to the bacterial genome. (C) Example of a predicted probe-bacteria fragment binding with many base pair mismatches.

hybridization affinities corresponding to the FRET ratios for all bacteria were calculated. We refer to these measurements as the measured hybridization affinity vectors and show them in Fig. 3B. Our challenge is to decode the experimentally measured affinities of the bacterial species samples reacting to UMD MBs using compressive sensing recovery techniques. With the predicted hybridization affinities of $N = 9$ bacteria to $M = 5$ random probes stored in the computationally obtained $\Phi_{5 \times 9}$, we used a variant of the orthogonal matching pursuit (OMP) algorithm (23) and successfully identified the species present in each of the samples (Fig. 3). UMD estimated the relative bacterial concentrations with an average error of 11.5% (fig. S6).

To provide the physician or scientist with a metric quantifying how close the measured hybridization affinity vector is to that of each bacteria in the database and thus how confident the OMP detection results are, UMD can output the inner products between the normalized measured hybridization affinity vectors from the nine experiments and the columns of the centered and normalized matrix $\Phi_{5 \times 9}$ (Fig. 3C). This metric measures the similarity of a pathogenic sample to bacteria in the UMD database. Using this metric, we characterized the performance of UMD in identifying the nine pathogens in terms of false positives and false negatives. We constructed the receiver operating characteristic (ROC) curve (Fig. 3D), where each point on the curve corresponds to a certain universal detection threshold in the range $[-1, 1]$ for all nine independent bacterial experiments. Inner product values above/below the detection threshold were considered as a positive/negative outcome, respectively. The area under the corresponding ROC

curve (AUC = 0.91) suggests successful screening performance. In addition, Fig. 3E shows the consistency of the measured and simulated hybridization affinities of nine bacteria to five random probes (different probes are shown in different colors). The normalized root mean square error (NRMSE = 12%) suggests that our thermodynamic modeling of bacteria-probe hybridization is accurate.

Next, we assessed the performance of UMD in detecting our nine test species from a list of common pathogens using our five DNA probes. We expanded the reference genome database to contain 40 genera (that is, $M = 5 \ll N = 40$), including bacterial pathogens listed by the Centers for Disease Control and Prevention (CDC) as the most common notifiable human diseases (24). With the most common pathogens' genomes in the database, the detection performance remained above AUC = 0.84, suggesting a high recovery rate with only five random probes.

Extension by simulation

Thus far, we have presented an experimental proof of concept that validates UMD's ability to detect 11 test species among a list of pathogens using a fixed set of five randomly selected test probes. We next numerically demonstrate that, if a sufficient number of probes is used, then any group of randomly selected probes will detect the presence of one ($K = 1$) or a mixture of several ($K = 2, 3, \dots$) pathogenic organisms in a sample out of a database of 40 pathogenic organisms. We introduced additive white Gaussian noise to the simulated hybridization affinity vectors to capture the variance in the hybridization affinities among

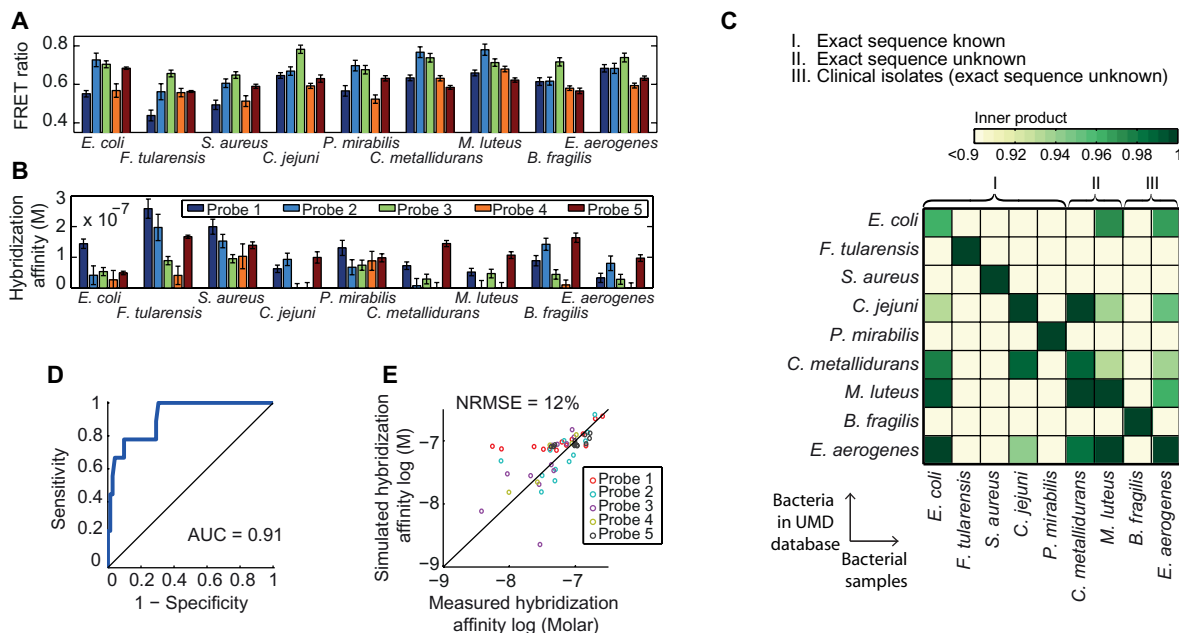


Fig. 3. Binding patterns of five random probes correctly identify the bacteria present in nine diverse bacterial samples. (A) Experimentally measured FRET ratios to quantify hybridization between bacterial DNA and probes 1 to 5. (B) Hybridization affinity between DNA samples and probes converted from FRET ratio through the probe characteristic curve fit equations (table S1 and fig. S2). (C) Heat map of normalized inner products between the experimentally obtained hybridization affinity and predicted hybridization affinities (by thermodynamic model) for nine DNA samples as a measure of the similarity of the probe measurements to the bacteria in the data set. DNA samples are clustered into three groups: (i) exact sequence known, (ii) exact sequence unknown, and (iii) clinical isolates (whose exact sequence is unknown). UMD correctly recovers the diagonally highlighted bacterium (with inner product >0.9). (D) The average ROC curve of UMD in detecting nine bacteria, assuming the independence of the different experiments. Each point on the curve corresponds to a threshold value between $[-1, 1]$. UMD achieves high values of the AUC (AUC > 0.9). (E) Correlation of measured and simulated hybridization affinities and the NRMSE of the prediction (straight line corresponds to maximum correlation). All experiments were performed in triplicate, and the results shown here average over the trials with the error bars representing SEM.

the independent test trials in Fig. 3. The noise levels were extrapolated from the above 11 test bacteria experiments, with the noise variance set to $\sigma_0 = 2.4 \times 10^{-8}$ M. To control for differences in the genome size of each organism, we normalized numerical simulations to unit weight of bacterial DNA.

In Fig. 4A, we first demonstrate the detection performance of UMD in identifying a single bacterium ($K = 1$) among the pathogen database at different noise levels. As the ROC curves suggest, UMD's detection performance improves when the noise variance decreases. With only a fivefold decrease in the noise variance, UMD almost perfectly identifies all 40 bacteria in the database ($AUC > 0.95$), using only five randomly selected probes.

The ability of UMD to universally detect target species can be improved by increasing the number of random probes. Figure 4B demonstrates that UMD almost perfectly identifies all 40 bacteria in the CDC database ($AUC = 0.95$) with any $M = 15$ randomly selected MBs when the noise variance is similar to that measured experimentally (Fig. 3).

UMD has the unique advantage that it can recover more than a single ($K > 1$) organism in an infectious sample. To evaluate the minimum number of probes M required for this task, we used the basis pursuit denoising algorithm (as described in the Supplementary Materials) to identify the composition of a sample containing $K = \{2,3\}$ equiconcentration bacterial species (Fig. 4C). We found that any set of $M = 15$ randomly selected probes will recover all $\binom{40}{K}$ possible

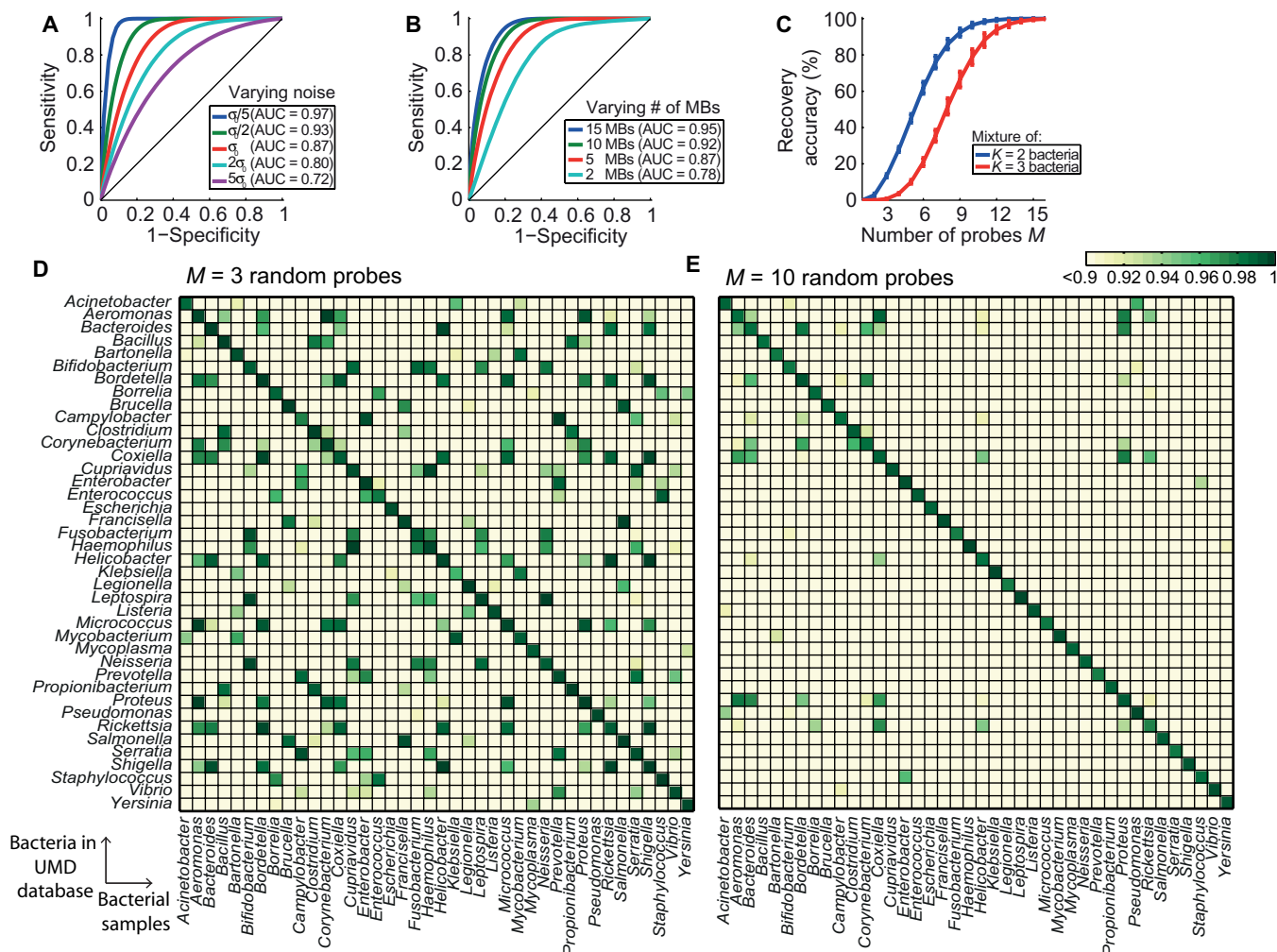


Fig. 4. Performance of UMD platform in genus-level recovery of 40 species listed as the most common human infectious genera by CDC with different number of random probes M and noise variance σ . (A) The ROC curve in detecting single bacterium ($K = 1$) with different noise levels. $\sigma_0 = 2.4 \times 10^{-8}$ M denotes the variance of the additive white Gaussian noise used in the simulation. This value is obtained from the experiments in Fig. 3 by calculating the propagated variance of measured FRET ratios. UMD performs more accurately with lower noise variance. The detection is almost perfect ($AUC > 0.95$) under noise variance $\sigma = \sigma_0/5$. (B) The average ROC curve in detecting single bacterium using different number of random probes M and fixed noise variance $\sigma = \sigma_0$. The detection performance universally improves over all the 40 species by increasing the number of random probes. With 15 random probes, UMD achieves almost perfect detection performance ($AUC > 0.95$). (C) The percentage of simulated trials, where K bacteria present in the samples were recovered correctly with zero false positives, among all possible $\binom{40}{K}$ bacteria mixtures (blue and red curves corresponding to $K = 2$ and $K = 3$ bacteria, respectively). Simulations were repeated 1000 times with randomly selected MBs, and error bars represent SD. (D and E) Confusion matrices illustrating the detection result of UMD using $M = 3$ and $M = 10$ probes selected by the GPS algorithm.

mixtures of $K = \{2,3\}$ pathogenic species in the CDC database. The error bars show SD over 1000 test trials with different sets of random MBs. This result confirms that the incoherence requirement for compressive sensing is empirically satisfied for the pathogenic strains in the CDC database and thus that UMD is capable of screening for pathogenic bacteria at the genus level.

We next evaluated UMD's performance for species-level bacterial detection. We focused on differentiating among 24 species of *Staphylococcus* genus and 23 species of *Vibrio* genus in silico. We identified the composition of samples containing *Staphylococcus* species using 11 random probes (fig. S7A) and the composition of samples containing *Vibrio* species using 18 random probes (fig. S7B) with high sensitivity and specificity (AUC > 0.95). This underscores UMD's potential to differentiate pathogens at high taxonomic resolution.

Using the UMD platform, one can trade off between universality (detecting species outside of the library) and cost efficiency (number of probes). That is, it is possible to select a set of probes that achieves better detection performance in terms of specificity and sensitivity than the average performance of random probe sets at the cost of universality. For example, in Fig. 4C, some of the probe sets achieved 100% accuracy using three fewer probes than the number required for universal recovery. To capitalize on this phenomenon, we developed a greedy probe selection (GPS) algorithm that rapidly selects these optimized probes given a very large database of genome-probe hybridization affinities (see Materials and Methods for more details on GPS).

Figure 4 (D and E) illustrates the UMD confusion matrices in detecting pathogenic bacteria using $M = 3$ and $M = 10$ probes selected using GPS. The false-positive rate drops for all the bacteria in the database as the number of probes increases from $M = 3$ (AUC > 0.95) to $M = 10$ (AUC > 0.99). Although the performance detection is high (AUC > 0.99), the confusion matrix shows few cases where the inner product values for possible species (for example, *Coxiella*, *Aeromonas*, and *Proteus* when the actual sample contains *Coxiella*) are only slightly separated. Figure S8 shows that greater separation between inner product values for candidate bacterial species can be achieved by using a larger number of UMD probes. This can increase the robustness of the UMD system (ensure low false-positive rate) for noisier environments.

Although mainly intended to rapidly screen for pathogens at higher taxonomy levels, UMD can also provide strain-level information to the physician using additional GPS-selected probes. In fig. S9, we demonstrate that GPS selects UMD probes that differentiate among nine strains of *E. coli* (eight pathogenic and one nonpathogenic) with high detection accuracy (AUC > 0.95) in silico.

The theory behind the UMD can be extended to identify more complex samples using a relatively small number of random probes. We verified that the UMD platform can recover complex microbial samples containing up to 100 active species out of a large dictionary of $N = 1500$ bacterial genera. We first computed the hybridization affinity of a set of random probes to $N = 1500$ representative species of all sequenced bacterial genera in the NCBI website. Then, we used the resulting hybridization affinity matrix to identify the composition of samples containing K unique species with equal concentrations (the most difficult case where the sample contains multiple species all at significant concentration). For each value of K , fig. S10 shows the minimum number of random probes M required to identify the composition of 1000 simulated complex samples containing K species randomly selected from $N = 1500$ genera. Figure S10 illustrates that the number of required probes M ($R^2 = 0.98$) closely follows the compressive

sensing theory $M = cK \log(N/K)$ with constant $c = 2.94$. That is, UMD requires a number of probes that grows logarithmically with the number of target bacteria N and sublinearly with the number of active bacteria K in the sample.

DISCUSSION

UMD probes are universal in the sense that a fixed set of probes captures the salient information required to distinguish between members of a large and growing database of species (25). This gives UMD a potentially important future proof property: a fixed set of measurement probes can be used to detect and estimate the concentration of newly sequenced species not yet present in the library. To detect a new organism, the software merely has to be adjusted to take into account how the new organism will react to the existing probe set; however, new capture probes are not required. Moreover, because the number of probes grows only logarithmically in the size of the library, the UMD platform naturally contends with the data deluge (26) from new microbial species being discovered and sequenced every day.

Several other pathogen detection schemes are currently under investigation (19, 27–29). To the best of our knowledge, UMD is the only technique that enables a unified representation of bacterial organisms in a low-dimensional geometric space. The theory of compressive sensing provides rigorous recovery guarantees and suggests algorithms to leverage this geometry to both efficiently detect bacteria and estimate their concentrations. Our successful implementation of UMD confirms that a small number of random DNA probes satisfy the incoherency requirements of compressive sensing theory and can be used for universal microbial sensing.

The UMD platform has the potential to rapidly direct physicians to use appropriate antibiotics or treatment and thus minimize the risk of antibiotic resistance. It can also be used in biodefense applications to classify multiple novel and mutant agents. With further optimization of the probe design and detection schemes, we expect that UMD will be able to sense an even wider range of organisms (for example, viruses and fungi) and various biomolecules of interest (genes and proteins).

Finally, the theory behind the UMD platform can be applied to DNA sensing in several other incarnations, including reads from a sequencer; for example, we may be able to quickly identify a bacterium from a subset of reads, rather than requiring full alignment or assembly. Application of such signal acquisition principles to biological sensing systems will shape the future of microbial diagnostics.

MATERIALS AND METHODS

Random DNA probe construction and preparation

To implement the UMD platform, we obtained DNA oligonucleotides for the random DNA probes and their exact complements from Integrated DNA Technologies. The sequences are provided below. MgCl₂, KCl, and sterile nuclease-free water for making the MB buffer were purchased from Fisher Scientific. Tris-HCl solution (1 M; pH 8.3) and tris-EDTA buffer [TE buffer; 10 mM tris-HCl, 0.1 mM EDTA (pH 8)] were obtained from Teknova. To prevent nuclease contamination, all work surfaces and materials were routinely cleaned with RNase-OFF decontamination solution.

Random probe 1, 5′-5Cy5/CGACGGTTGCTTGGGTACTTG-GATGATGCTAAATTGGTGTGGTTCG/3Cy3Sp/-3′; random probe 2, 5′-5Cy5/CGACGGTGCTTTGAATACTTGGTAGAGGCTG-GAGGGTGGTTGGTTCG/3Cy3Sp/-3′; random probe 3, 5′-5Cy5/CGACGGTGCTGGGTGAACTAAAGGGTGGGTGCTATGG-GAAGGGTCG/3Cy3Sp/-3′; random probe 4, 5′/5Cy5/CGACTTAAT-GAATGTGTGGGCGCTTGGTTGCTTAATGAGTGGGTTCG/3Cy3Sp/-3′; and random probe 5, 5′-5Cy5/CGACGTTTCTTTCTG-GAGGAGGGAGGGTTAGTTGTTAGGCAGTCG/3Cy3Sp/-3′.

Random probe complement 1, 5′-CGACCAACACCAATTTAG-CATCATCCAAGTACCCAAGCAACCGTCG-3′; random probe complement 2, 5′-CGACCAACCACCCTCCAGCCTCTACCAAG-TATTCAAAGCACCGTCG-3′; random probe complement 3, 5′-CGACCTTCCCATAGCACCCACCTTTAGTTCACCCAG-CACCGTCG-3′; random probe complement 4, 5′-CGACCCACT-CATTAAGCAACCAAGCGCCACACATTCATTAAGTCG-3′; and random probe complement 5, 5′-CGACTGCCTAACAAC-TAACCTCCCTCCTCCAGAAAAGAAACGTCG-3′.

Random DNA probe design

In the design of MBs (20) for random DNA probes, the length and GC content (ratio of G + C to other nucleotides) of the probe loop and stem sequences were considered to strike a balance between two factors: fluorescence signal level and probe stability. Signal intensity was especially important in our detection scheme, because no DNA amplification method (such as PCR) was used. Similar to sloppy MBs (19), we selected the random probe loop sequence to be longer than traditional MBs. In addition, we made the stem sequence 1 nt shorter to introduce additional sloppiness (that is, hybridization in presence of more base pair mismatches).

Our challenge was to find probes that maintain the MB's signature hairpin structure over a wide range of temperatures (4° to 50°C) after introducing additional sloppiness. To produce random MBs, we followed the following procedure: we first generated 1 million random sequences of length 46 nt with fixed stem sequences on both ends (Fig. 2A). Then, we used a package in the DNA software (Visual OMP DE) to generate all the possible stable and secondary structures of the sequences in the experimental thermodynamic conditions. We parsed the output of the DNA software and filtered out the probes with undesired secondary structures or melting temperatures. Figure S1 shows the gain in hybridization affinity obtained using the random MBs in comparison with traditional MBs and sloppy MBs in binding to the *E. coli* genome. By no means is this the only method to generate random probes for a UMD platform; any method that produces probes with a stable hairpin structure and uniform melting temperature while providing the required signal intensity can be used.

Generation of random probe characteristic curves

The experimentally measured FRET, defined as the ratio of Cy5 intensity over total fluorescence intensity (Cy3 + Cy5), is a function of the concentration of open random probes in the solution, that is, the probe-target hybridization affinity. To discern the hybridization affinity between a probe and target in units of molarity rather than as a FRET ratio, a characteristic curve was constructed for each probe. These curves presented the FRET ratio as a function of the concentration of open probes in molarity.

To obtain the characteristic curves, random probes were diluted to 1 μM in 1× MB buffer [4 mM MgCl₂, 50 mM KCl, 10 mM tris-HCl (pH 8), in sterile ribonuclease (RNase)-free water]. DNA oligonucleotides

perfectly complementary to the random probes were diluted using 1× TE buffer to 10⁻⁵, 10⁻⁶, 8 × 10⁻⁷, 6 × 10⁻⁷, 4 × 10⁻⁷, 2 × 10⁻⁷, 10⁻⁷, 8 × 10⁻⁸, 6 × 10⁻⁸, 4 × 10⁻⁸, 2 × 10⁻⁸, 10⁻⁸, 10⁻⁹, 10⁻¹⁰, or 10⁻¹¹ M concentration. Twenty-five microliters of 1 μM random DNA probes (diluted in MB buffer) was added to 25 μl of perfect complement DNA of various concentrations or to the TE buffer-only control.

The DNA mixture was briefly centrifuged with a mini centrifuge (VWR) to collect all DNA to the bottom of the tube. Then, the DNA was hybridized using a MyCycler Thermal Cycler (Bio-Rad) under the following conditions: 95°C for 5 min, 50°C for 2 min, 30°C for 1 min, 20°C for 1 min, and 4°C for 2 min. Forty-five microliters of each thermal cycled mixture was added to 155 μl of 1× MB buffer in a black flat-bottom 96-well plate (Corning) and kept at 4°C overnight. A nonlinear optimization algorithm (22) was used to fit the parameters *a*, *b*, *n*, and FRET₀ to the characteristic curve $FRET(c) = FRET_0 + a/[1 + b(10^{-6} - c)^{-n}]$ (fig. S2). The *R*² and RMSE for the curve fits are reported in table S1.

Bacterial DNA extraction

Overnight cultures of *S. aureus* USA300 and *E. coli* MG1655 were used to inoculate fresh cultures grown in 50 to 100 ml of 2xTY or lysogeny broth, respectively. *F. tularensis* Live Vaccine Strain was obtained from DynPort Vaccine Company LLC (derived from NDBR101 Lot 4) and grown in modified Mueller-Hinton cation-adjusted (MHII) broth (Becton Dickinson) supplemented with sterile 0.1% glucose, sterile 0.025% ferric pyrophosphate, and 2% reconstituted IsoVitalEx (Becton Dickinson). Cultures were pelleted and washed three times with sterile phosphate-buffered saline (PBS). To release chromosomal DNA, cells were resuspended in TE buffer and mixed with 10% SDS and proteinase K at 65°C overnight. DNA was isolated using phenol/chloroform and precipitated via ethanol precipitation [protocol adopted from the study by Sambrook *et al.* (30)]. DNA pellets were resuspended in 50 μl of TE buffer and stored at -20°C.

Bacterial strains *C. jejuni*, *P. mirabilis*, *C. metallidurans*, *M. luteus*, *B. dentium*, *E. aerogenes*, *B. fragilis*, and *P. aeruginosa* were grown overnight in 30 ml of Brain Heart Infusion medium (Becton Dickinson) at 37°C. Bacterial cells were pelleted, washed two times with sterile 1× PBS, and resuspended in TE buffer. Proteinase K (1 mg/ml; Sigma-Aldrich) and 0.5% SDS were added to the bacterial cells, which were then incubated overnight at 55°C on an orbital shaker. The samples were then mixed with phenol/chloroform (Invitrogen) and centrifuged; supernatants were transferred to a fresh tube. This aqueous phase was then mixed with an equal volume of chloroform and centrifuged (and repeated). Finally, 1/10 volume of 2 M sodium chloride and an equal volume of isopropanol were added to precipitate the DNA. This mixture was incubated at -20°C for 30 min and centrifuged. The pellets were rinsed with 70% ethanol, air-dried, and resuspended in TE buffer.

Random probe and bacterial DNA hybridization

Bacterial DNA was diluted to approximately 500 ng/μl using TE buffer and kept at -20°C until use. The random MB probes were diluted to 1 μM MB buffer before use. Twenty-five microliters of 1 μM random probes (diluted in MB buffer) was added to 25 μl of TE buffer control or *E. coli*, *F. tularensis*, *S. aureus*, *C. jejuni*, *P. mirabilis*, *C. metallidurans*, *M. luteus*, *B. dentium*, *E. aerogenes*, *B. fragilis*, or *P. aeruginosa* DNA. The DNA mixture was briefly centrifuged to collect DNA and then hybridized using MyCycler Thermal Cycler (Bio-Rad) under the following conditions: 95°C for 5 min, 50°C for 2 min, 30°C for 1 min, 20°C for 1 min, and 4°C for 2 min. Forty-five microliters of each thermal cycled mixture was added

to 155 μl of $1\times$ MB buffer in a black flat-bottom 96-well plate (Corning) and kept at 4°C overnight.

Measuring FRET ratio through fluorescence as indicator of random probe-bacteria hybridization

The FRET ratio for the genomic DNA samples following hybridization with each of the random probes was determined by reading the Cy3 and Cy5 fluorescence using a FluoroLog-3 spectrofluorometer (Jobin Yvon Horiba) coupled with a MicroMax 384 MicroWell Plate Reader and water-cooled photomultiplier tube detector. Samples were excited at 545 nm, and single-point fluorescence measurements were taken at 562- and 677-nm emission (optimal wavelengths determined through excitation-emission matrix analysis) to measure the Cy3 and Cy5 fluorescence, respectively. The FRET ratio was calculated as $\text{Cy5}/(\text{Cy5} + \text{Cy3})$.

Determining DNA hybridization affinity via SantaLucia thermodynamic models

A comprehensive thermodynamic model by SantaLucia and Hicks (18) was used to predict the hybridization of probes to bacterial genomes. The SantaLucia model incorporates thermodynamic parameters for mismatches between two DNA sequences. We used two software packages: ThermoBlast DE, which performs fast alignment of sequences against large genome databases to discover thermodynamically stable hybridizations, and Visual OMP DE, which simulates hybridization experiments with detailed solution conditions and generates results for melting temperature, Gibbs free energy, and the percentage-based concentration of each resultant species after experiment. The secondary structure of each monomer, homodimer, and heterodimer species formed from the constituent probes, and target fragments could also be visualized.

To calculate the hybridization affinity of a genome to a probe, we first used the ThermoBlast package and thermodynamically aligned the sequence of the random probe to both complement strands of the target genome. We extracted all of the sequence fragments of the genome (100 to 200 nt) that aligned with the probe sequence with a predicted melting temperature within approximately 35°C of the melting temperature of the sequence genome. We then used Visual OMP DE to simulate the hybridization between the probe and the target genome using the target fragments (Fig. 2). Every simulation contained information on the probe sequence, the target fragment sequences, and conditions for the experiment, including probe concentration (1 μM), unit target concentration (500 ng/ μl for all bacteria), assay temperature (4°C), hybridization buffer composition (4 mM Mg^{++} , 50mM Na^+ , 0 M glycerol, 0 M dimethyl sulfoxide, 0 M formamide, 0 M tetramethylammonium chloride, 0 M betaine), and pH (8). This procedure was repeated for each probe-target genome pair. We used the percentage of probe-target heterodimer structures formed, that is, the percentage of probes that were bound to target fragments, as an estimate for the hybridization affinity of the probe to each target (Fig. 2B).

Linearity assumption considerations in UMD

In the UMD platform, the probe concentration (1×10^{-6} M) was in far excess of the target concentrations ($\sim 1 \times 10^{-10}$ M); therefore, we were able to linearly combine the hybridization affinity signatures that we measured for individual targets using the hybridization model. Because of the flooding of excessive number of probes, each target fragment had

its choice of binding/not binding to the probes, and thus, we could safely sum together multiple target interactions of the same probe, assuming them to be independent.

ROC analysis

ROC analysis was performed by plotting an ROC curve showing the sensitivity and (1-specificity) for 1000 threshold values ranging from -1 to 1 . For each threshold value, the following procedure was performed on the data matrix of normalized inner products between the experimentally obtained hybridization affinity and predicted hybridization affinities (by thermodynamic model) for the nine independent bacterial DNA samples (Fig. 3C): each entry in the inner product data matrix was compared with the threshold value to determine the number of true positives, false positives, true negatives, and false negatives. True positives were identified when values in the diagonal entries of the inner product data matrix were greater than the threshold value because diagonal entries represented the correct classification of the bacterial sample with its corresponding genus in the database. False positives were identified as off-diagonal values that were greater than the threshold value. True negatives were identified as off-diagonal values that were less than or equal to the threshold value. False negatives were identified as diagonal values that were less than or equal to the threshold value. For each threshold value, sensitivity was defined as $[\# \text{ true positives}/(\# \text{ true positives} + \# \text{ false negatives})]$, and specificity was defined as $[\# \text{ true negatives}/(\# \text{ true negatives} + \# \text{ false positives})]$.

Greedy probe selection

Given a set of P random probes, finding the set of M probes with the best detection performance in terms of sensitivity and specificity is an extremely challenging problem. A brute-force search would require one to search among all $\binom{M}{P}$ possible combinations of M probes to find the optimal probe set. This combinatorial search algorithm grows quadratically with P and thus becomes computationally intractable when the number of probes grows. We thus developed a rapid probe selection method that we dub GPS. With a small sacrifice in sensitivity, GPS finds the best performance probe in a few seconds: exponentially faster than the naïve search method. The algorithm in each iteration finds the probe that maximizes a detection performance criterion (here, the maximum pairwise correlation of bacteria) and adds it to the list of probes picked from the previous iterations. GPS stops when the maximum desired number of probes is reached.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/2/9/e1600025/DC1>

Mathematical formulation of the comprehensive sensing (CS) detection and estimation algorithms

Complete list of bacterial strains used in UMD simulations

fig. S1. Comparison of the sloppiness of random probes to other MBs.

fig. S2. Random probes' characteristic curves.

fig. S3. Experimentally measured FRET ratios to quantify hybridization between 11 bacterial DNA samples and probes 1 to 5.

fig. S4. Hybridization affinity between 11 bacterial DNA samples and probes 1 to 5.

fig. S5. Detection performance of 11 bacterial samples using five random probes.

fig. S6. Comparison of the predicted concentrations of bacterial DNA with the experimentally measured values.

fig. S7. Performance of UMD in species-level recovery of 24 strains of *Staphylococcus* and 23 strains of *Vibrio*.

fig. S8. Performance of UMD in identifying pathogens in genus level using 15 GPS probes.

fig. S9. Performance of UMD in identifying eight pathogenic and one nonpathogenic *E. coli* strains using GPS probes.

fig. S10. Performance of UMD in identifying the composition of several complex samples.

table S1. The fitted parameter to the probes' characteristic curves.

References (31–35)

REFERENCES AND NOTES

- M. Klompas, D. S. Yokoe, R. A. Weinstein, Automated surveillance of health care–associated infections. *Clin. Infect. Dis.* **48**, 1268–1275 (2009).
- V. N. Pinto, Bioterrorism: Health sector alertness. *J. Nat. Sci. Biol. Med.* **4**, 24–28 (2013).
- B. Van Dorst, J. Mehta, K. Bekaert, E. Rouah-Martin, W. De Coen, P. Dubruel, R. Blust, J. Robbens, Recent advances in recognition elements of food and environmental biosensors: A review. *Biosens. Bioelectron.* **26**, 1178–1194 (2010).
- G. S. Martin, D. M. Mannino, S. Eaton, M. Moss, The epidemiology of sepsis in the United States from 1979 through 2000. *N. Engl. J. Med.* **348**, 1546–1554 (2003).
- D. L. Hoyert, J. Xu, Deaths: Preliminary data for 2011. *Natl. Vital Stat. Rep.* **61**, 1–51 (2012).
- J. M. Mylotte, A. Tayara, Blood cultures: Clinical aspects and controversies. *Eur. J. Clin. Microbiol. Infect. Dis.* **19**, 157–163 (2000).
- P. M. Dark, P. Dean, G. Warhurst, Bench-to bedside review: The promise of rapid infection diagnosis during sepsis using polymerase chain reaction–based pathogen detection. *Crit. Care* **13**, 217 (2009).
- A. Pechorsky, Y. Nitzan, T. Lazarovitch, Identification of pathogenic bacteria in blood cultures: Comparison between conventional and PCR methods. *J. Microbiol. Methods* **78**, 325–330 (2009).
- S. Sontakke, M. B. Cadenas, R. G. Maggi, P. P. V. P. Diniz, E. B. Breitschwerdt, Use of broad range 16S rDNA PCR in clinical microbiology. *J. Microbiol. Methods* **76**, 217–225 (2009).
- C. D. Sibley, G. Peirano, D. L. Church, Molecular methods for pathogen and microbial community detection and characterization: Current and potential application in diagnostic microbiology. *Infect. Genet. Evol.* **12**, 505–521 (2012).
- S. Riedel, K. C. Carroll, Blood cultures: Key elements for best practices and future directions. *J. Infect. Chemother.* **16**, 301–316 (2010).
- M. Paolucci, M. P. Landini, V. Sambri, Conventional and molecular techniques for the early diagnosis of bacteraemia. *Int. J. Antimicrob. Agents* **36**, S6–S16 (2010).
- M. Bauer, K. Reinhart, Molecular diagnostics of sepsis—Where are we today? *Int. J. Med. Microbiol.* **300**, 411–413 (2010).
- R. P. H. Peters, M. A. van Agtmael, S. A. Danner, P. H. M. Savelkoul, C. M. J. E. Vandembroucke-Grauls, New developments in the diagnosis of bloodstream infections. *Lancet Infect. Dis.* **4**, 751–760 (2004).
- Centers for Disease Control and Prevention Antibiotic Resistance threats in the United States, available at www.cdc.gov/drugresistance/ [accessed November 2013].
- D. L. Donoho, Compressed sensing. *IEEE T. Inform. Theory* **52**, 1289–1306 (2006).
- R. G. Baraniuk, Compressive sensing. *IEEE Signal Proc. Mag.* **24**, 118–121 (2007).
- J. SantaLucia Jr., D. Hicks, The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 415–440 (2004).
- S. Chakravorty, B. Aladegebami, M. Burday, M. Levi, S. A. E. Marras, D. Shah, H. H. El-Hajj, F. R. Kramer, D. Alland, Rapid universal identification of bacterial pathogens from clinical cultures by using a novel sloppy molecular beacon melting temperature signature technique. *J. Clin. Microbiol.* **48**, 258–267 (2010).
- S. Tyagi, F. R. Kramer, Molecular beacons: Probes that fluoresce upon hybridization. *Nat. Biotechnol.* **14**, 303–308 (1996).
- M. A. Sheikh, thesis, Rice University (2010).
- Ž. Jeričević, Ž. Kušter, Non-linear optimization of parameters in Michaelis-Menten kinetics. *Croat. Chem. Acta* **78**, 519–523 (2005).
- A. M. Bruckstein, M. Elad, M. Zibulevsky, On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE T. Inform. Theory* **54**, 4813–4820 (2008).
- Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report, available at www.cdc.gov/mmwr/ [accessed November 2013].
- M. A. Davenport, P. T. Boufounos, M. B. Wakin, R. G. Baraniuk, Signal processing with compressive measurements. *IEEE J. Sel. Top. Signa.* **4**, 445–460 (2010).
- R. G. Baraniuk, More is less: Signal processing and the data deluge. *Science* **331**, 717–719 (2011).
- W. Dai, M. A. Sheikh, O. Milenkovic, R. G. Baraniuk, Compressive sensing DNA microarrays. *EURASIP J. Bioinform. Syst. Biol.* 162824 (2009).
- M. Mohtashemi, D. K. Walburger, M. W. Peterson, F. N. Sutton, H. B. Skaer, J. C. Diggans, Open-target sparse sensing of biological agents using DNA microarray. *BMC Bioinformatics* **12**, 314 (2011).
- H. J. Chung, C. M. Castro, H. Im, H. Lee, R. Weissleder, A magneto-DNA nanoparticle system for rapid detection and phenotyping of bacteria. *Nat. Nanotechnol.* **8**, 369–375 (2013).
- J. Sambrook, E. F. Fritsch, T. Maniatis, *Molecular Cloning* (Cold Spring Harbor Lab. Press, ed. 3, 2001).
- D. L. Donoho, J. Tanner, Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9446–9451 (2005).
- D. L. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition. *IEEE T. Inform. Theory* **47**, 2845–2862 (2001).
- J. A. Tropp, Greed is good: Algorithmic results for sparse approximation. *IEEE T. Inform. Theory* **50**, 2231–2242 (2004).
- T. T. Cai, L. Wang, Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE T. Inform. Theory* **57**, 4680–4688 (2011).
- J. A. Tropp, A. C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit. *IEEE T. Inform. Theory* **53**, 4655–4666 (2007).

Acknowledgments: We thank J. Moake for allowing use of his laboratory's thermal cycler. **Funding:** A.A., M.A.S., and R.G.B. were supported by the NSF (CCF-0728867). A.Y.L. was supported by the Medical Scientist Training Program at Baylor College of Medicine, the Edward and Josephine Hudson Scholarship, and the Ruth L. Kirschstein National Research Service Awards for Individual Predoctoral MD/PhD Fellows (5F30CA165686) of the NIH and National Cancer Institute. A.L.C. was supported by the NSF through a Graduate Research Fellowship (0940902) and the Keck Center of the Gulf Coast Consortia through the Nanobiology Interdisciplinary Graduate Training Program (NIH grant T32EB009379). J.F.P. and L.M.A. were supported by the NIH Biodefense and Emerging and Infectious Disease Regional Center of Excellence Program (U54 AI057156). **Author contributions:** A.A., M.A.S., and R.G.B. designed the random probes, performed the DNA software simulations, conducted the compressive sensing modeling, analyzed and interpreted the simulation and experimental results, and drafted and revised the manuscript. A.Y.L., A.L.C., and R.A.D. designed and conducted the experiments, contributed to compressive sensing modeling design, analyzed and interpreted the simulation and experimental results, and drafted and revised the manuscript. L.M.A., C.L.J., and J.F.P. participated in the design of bacterial experiments and interpretation of results, prepared and characterized the bacterial samples, and assisted in the drafting and revision of the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 7 January 2016

Accepted 19 August 2016

Published 28 September 2016

10.1126/sciadv.1600025

Citation: A. Aghazadeh, A. Y. Lin, M. A. Sheikh, A. L. Chen, L. M. Atkins, C. L. Johnson, J. F. Petrosino, R. A. Drezek, R. G. Baraniuk, Universal microbial diagnostics using random DNA probes. *Sci. Adv.* **2**, e1600025 (2016).

Universal microbial diagnostics using random DNA probes

Amirali Aghazadeh, Adam Y. Lin, Mona A. Sheikh, Allen L. Chen, Lisa M. Atkins, Coreen L. Johnson, Joseph F. Petrosino, Rebekah A. Drezek and Richard G. Baraniuk

Sci Adv 2 (9), e1600025.
DOI: 10.1126/sciadv.1600025

ARTICLE TOOLS	http://advances.sciencemag.org/content/2/9/e1600025
SUPPLEMENTARY MATERIALS	http://advances.sciencemag.org/content/suppl/2016/09/26/2.9.e1600025.DC1
REFERENCES	This article cites 30 articles, 3 of which you can access for free http://advances.sciencemag.org/content/2/9/e1600025#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the [Terms of Service](#)

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Advances* is a registered trademark of AAAS.