Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array

Alison M. Devault[1], Crystal Jaing[2], Shea Gardner[2], Teresita M. Porter[3], Jacob M. Enk[1,3], James Thissen[2], Jonathan Allen[2], Monica Borucki[2], Sharon N. DeWitte[4], Anna N. Dhody[5], Kevin McLoughin[2], and Hendrik N. Poinar[1,3,6 *]

1. McMaster Ancient DNA Centre, Department of Anthropology, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L9, Canada
2. Lawrence Livermore National Laboratory, Livermore, CA 94551, USA
3. Department of Biology, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L8, Canada
4. Departments of Anthropology and Biological Sciences, University of South Carolina, Columbia, SC, USA
5. The College of Physicians of Philadelphia, Mütter Museum, 19 S 22nd St, Philadelphia, PA 19103, USA
6. Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, 1280 Main St W, Hamilton, Ontario L8S4L8, Canada

Supplementary Information for

Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array

TABLE OF CONTENTS

I. Supplementary Methods 3
A. Sample preparation 3
B. Shotgun HTS sequencing 3
C. Pathogen HTS assemblies 3
D. HTS BLAST & MEGAN metagenomic analysis 3
E. LLMDA analysis 4
i. LLMDA v5 design 4
ii. LLMDA analyses 4
II. Supplementary Results 6
A. Tables 6
1. Table S1 – LLMDA and HTS analysis, full results ..... *EXCEL*
2. Table S2 – HTS BLASTN/MEGAN metagenomic profiles, full results ..... *EXCEL*
3. Table S3 – LLMDA probe data supporting *V. cholerae* and *Y. pestis* hits..... *EXCEL*
B. Figures 7
1. Figure S1 – LLMDA detected probe distributions for *V. cholerae* and *Y. pestis* pathogens 7
2. Figure S2 – Flowchart of workflow 9
3. Figure S3 – Average LLMDA probe GC% vs. average LLMDA probe log intensity, by family 10
4. Figure S4 – Number of HTS reads vs. HTS %GC, by family 11
5. Figure S5 – Examples of LLMDA detected probe distributions..... 12
III. Supplementary References 13

I. Supplementary Methods

A. Sample preparation

Sample 3090.13 is a preserved intestinal specimen from a victim of the 1849 Philadelphia cholera epidemic, sealed in a glass jar with alcohol, and stored in the collections of the Mütter Museum (Philadelphia, PA, USA).[1] This specimen was sub-sampled, extracted, and libraries suitable for sequencing on the Illumina platform were prepared as described in reference 1. Specimen 8291 is a tooth from a victim of the Black Death buried at the East Smithfield cemetery in London in 1348-1349.[2] This specimen was sampled and extracted using the same methods as described in reference 2. Libraries suitable for sequencing on the Illumina platform were prepared just as for 3090.13 (above), as described in reference 1.

B. Shotgun HTS sequencing

Prior to sequencing, additional indexing amplification was performed in 8 reactions each sample (5 µl 0.1x diluted template DNA in 50 µl total reaction volume) of indexed library, using 400nM each indexing primer, and 11 cycles for 3090.13 and 20 cycles for 8291. The purified libraries were pooled in equimolar ratio on one lane of Illumina HiSeq 1000. Sequencing was performed by the Farncombe Family Digestive Health Research Institute (McMaster University). 100bp paired-end read chemistry was used, with one indexing read. The lane yielded 141,039,627 reads each direction from 3090.13 and 122,830,910 reads each direction from 8291.

C. Pathogen HTS assemblies

Raw R1 reads from each sample were trimmed to remove residual adaptor sequence using cutadapt (v.1)[3] with the parameters: error rate (0.16), minimum overlap (1). Reads <28bp were removed from a 24,000,000 subset of each sample, leaving 12,946,441 for 3090.13 and 12,076,222 for 8291. To calculate HTS pathogen percentages, remaining reads were aligned using *bowtie* v.0.12.7[4] with default settings to the O395 strain *V. cholerae* reference genome (NC_009456, NC_009457) for sample 3090.13 and to the CO92 strain *Y. pestis* reference genome and 3 plasmids pCD1, pPCP1, and pMT1 (NC_003143, NC_003131, NC_003132, NC_003134) for sample 8291. For 3090.13, 6,938 aligned (0.054% of reads ≥28bp, 0.029% of total reads), and for 8291, 18,931 aligned (0.157% of reads ≥28bp, 0.079% of total reads).

D. HTS BLAST & MEGAN metagenomic analysis

Raw reads from each sample were trimmed using cutadapt (v.1.1) with the parameters -b (13bp adaptor sequence), –e (errors allowed) 0, – m (minimum length, bp) 20, -q (Phred scaled quality cutoff) 20, and -O (overlap, bp) 13, leaving 118,859,751 reads from 3090.13 and 89,321,997reads from 8291 for further processing. Reads were subjected to local BLASTN-megablast analysis[5] (v.2.2.26+) using a local copy of the refseq_genomic database[6] (downloaded October 16, 2012), using the parameters: -task megablast, -word_size 28, -evalue 1e-10, -num_descriptions 100, -num_alignments 100. BLAST reports were parsed using MEGAN4 (v.4.70.4) using the default lowest common ancestor (LCA) parameters.[7] Full results of this analysis can be found in Table S2.

E. LLMDA Analysis

i. LLMDA v5 design

All completely sequenced genomes or elements (chromosomes, mitochondria, plasmids) as of December 20, 2011 were obtained from public sources (NCBI, J. Craig Venter Institute, etc.). These included assembled draft and finished sequences for viruses, bacteria, archaea, fungi, and the subset of protozoa known to be human pathogens or their near neighbors. These were grouped by kingdom and family. LLMDAv5 was designed using substantially the same approach as previous versions,[8] namely, finding family-specific regions in the available complete sequences, and selecting probes within those regions such that all targets are represented by both conserved and discriminating probes. The LLMDAv5 135K design has approximately 135,000 unique target probes. Conserved probes were selected favoring the most within-family conserved, thermodynamically optimal probes, so that all targets were represented by at least 15 conserved probes. Discriminating probes were selected favoring the least conserved probes for each sequence, with at least 2 per genome or sequence element. On the 135K design, only probes from families containing at least one species known to infect vertebrates were included for the viruses, bacteria, and fungi. All archaea families were included since there were few enough probes to include them all, as well as all the pathogenic protozoa previously selected for probe design. Vertebrate infecting bacterial, viral, and fungal families were selected based on literature (PubMed) and web searches to determine whether any members of a family have been found to infect vertebrates or were involved in clinical infections, and all members of a family were included even if only some of them were vertebrate-infecting. The array also included several thousand negative control probes with random sequences designed to match the length and GC% distribution of the target probes. The following numbers of species were represented: 3,521 microbial species total, including 1,856 viral species, 1,398 bacterial species, 125 archaeal species, 94 protozoan species, and 48 fungal species.

ii. LLMDA analyses

LLMDA arrays were analyzed using the CLiMax (Composite Likelihood Maximization) algorithm, described in detail previously[4], followed by some additional processing steps. We measured probe

intensities on each array using NimbleScan software (Roche NimbleGen) and reduced them to vectors of binary probe detection indicators, by comparing each target probe intensity to the 95th percentile of the negative control probe intensities. The CLiMax software processes this indicator data using a greedy iterative procedure to predict a series of targets likely to be present in the sample. In the first iteration, a target is selected by computing, for each genome in a reference target database, the log-odds of the observed probe detection data if that genome were present in the sample; the target with the highest log-odds score becomes the first element of the series. In each subsequent iteration, a conditional log-odds score is computed for each remaining target, representing the likelihood of the data if the target were added to the series, relative to the likelihood given the previously predicted targets. The target with the largest conditional log-odds score is then appended to the series. Iterations continue until there are no additional targets with positive conditional log-odds scores, meaning that no further improvement in the likelihood can be obtained by predicting additional targets.

After the initial CLiMax analysis, we filtered the list of genomes predicted to be present by rejecting those for which the array detected only a small subset of the genome regions covered by probes. In our past experience, targets with this pattern of detected probes are likely to be false positives, resulting from cross-hybridization to a similar region in another genome. **Figure S5** shows examples of targets that were accepted and rejected under this filtering strategy. We aligned probes matching each selected target sequence to genome positions using BLAST. We used Gaussian kernel density estimates to approximate the positional distribution functions for all probes matching the target (with predicted detection probabilities greater than 0.85), and for the subset of these probes with intensities above the 95th percentile of negative controls, taking care to use the same bandwidth for both estimates. To quantify the difference between these two distributions, we computed the Kullback-Leibler divergence (D_{KL}) between the two density estimates. If $f_{pred}(x)$ and $f_{det}(x)$ are, respectively, the estimated density functions for the probes predicted to bind the target and the probes actually detected, evaluated at discrete positions x, then the K-L divergence is computed as $D_{KL}(f_{pred} \,||\, f_{det}) = \sum_x f_{pred}(x) \log (f_{pred}(x) / f_{det}(x))$. Targets with $D_{KL} > 4\times10^{-4}$ were removed from the predicted set; this threshold was chosen by analysis of samples of known composition, to provide a reasonable compromise between sensitivity and specificity. The numbers of target sequences predicted to be present in sample 8291 were 398 total and 204 after filtering; for sample 3090.13 the target counts were 430 total and 217 after filtering.

Finally, to enable comparison of the LLMDA results with the family-level results produced by BLAST and MEGAN analysis of HTS data, we grouped the filtered targets by family, and summed log-odds scores over targets to produce an aggregate score for each family.

Author Contribution Statement

HNP, MB, CJ conceived of the research. AMD, CJ, JT, JME designed and performed experiments. AMD, CJ, SG, TMP, JME, JT, JA, KM and HNP analyzed data. AND and SND provided samples. All authors contributed to the manuscript preparation.

Competing Financial Interests

The author(s) declare no competing financial interests.

II. Supplementary Results

Table S1. LLMDA and HTS analysis, full results

SEE EXCEL FILE

Table S2. HTS BLASTN/MEGAN metagenomic profiles, full results

SEE EXCEL FILE

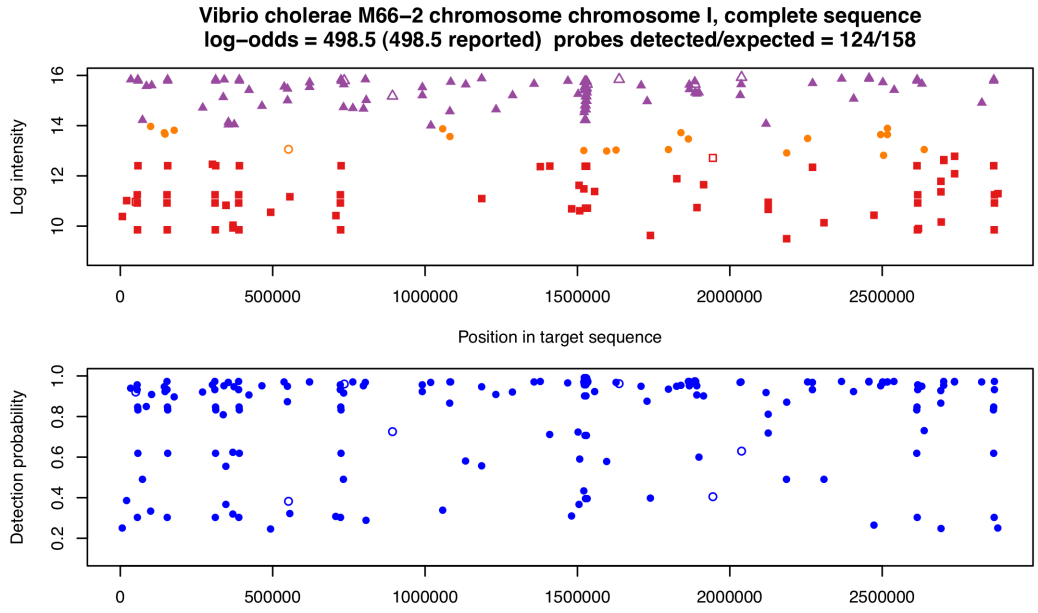
Table S3. LLMDA probe data supporting *V. cholerae* and *Y. pestis* hits

SEE EXCEL FILE

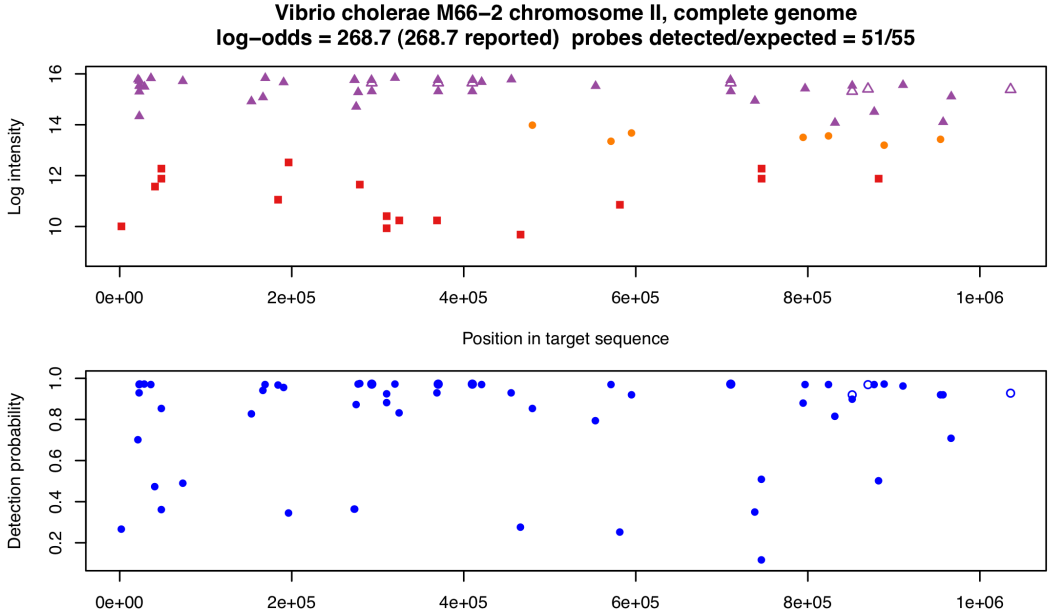
Figure S1. LLMDA detected probe distributions for *V. cholerae* and *Y. pestis* pathogens

Log intensities vs. genome position (upper graphs) and probe detection probabilities (based on similarity to target sequence) vs. position (lower graphs) for probes targeting (A) *V. cholerae* M66-2 chromosome I on array hybridized to sample 3090.13, (B) *V. cholerae* M66-2 chromosome II in sample 3090.13, and (C) *Y. pestis* CO92 plasmid pPCP1 in sample 8291. Purple triangles indicate that intensity was above the 99th percentile of the negative controls; orange circles indicated intensities between the 99th and the 95th percentiles; red squares indicate intensities below the 95th percentile. Open symbols represent probes that were excluded from the score computation, because they light up non-specifically even when there is no sample present in the hybridization mixture.

A



B



C

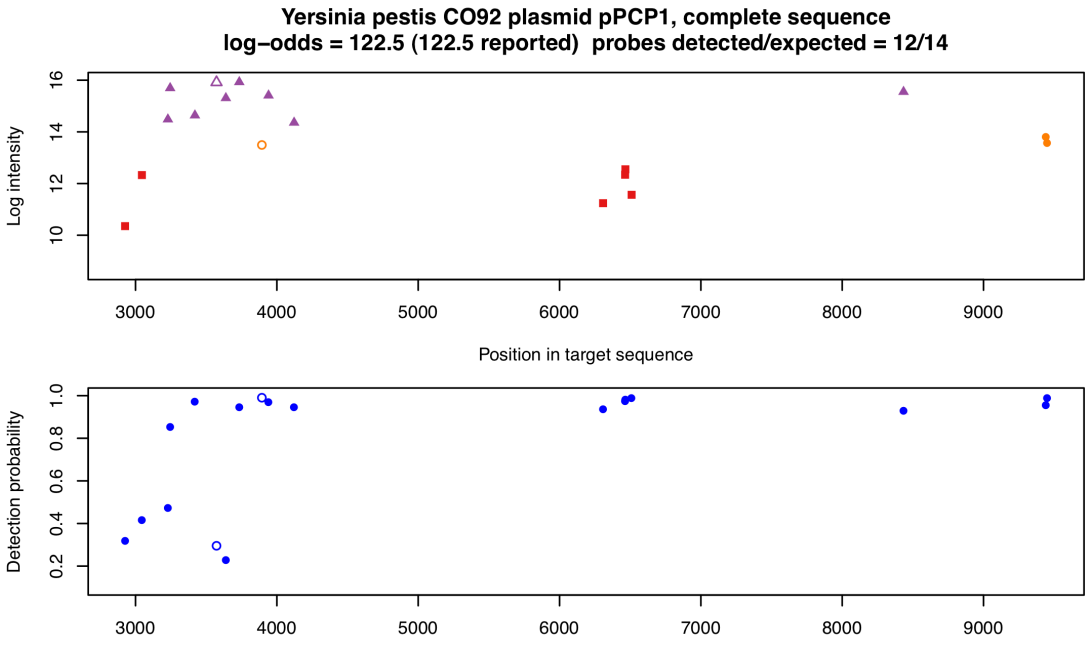

Yersinia pestis CO92 plasmid pPCP1, complete sequence
log−odds = 122.5 (122.5 reported) probes detected/expected = 12/14
Log intensity
10
12
14
16
3000
4000
5000
6000
7000
8000
9000
Position in target sequence
Detection probability
0.2
0.4
0.6
0.8
1.0
3000
4000
5000
6000
7000
8000
9000

Figure S2. Flowchart of workflow

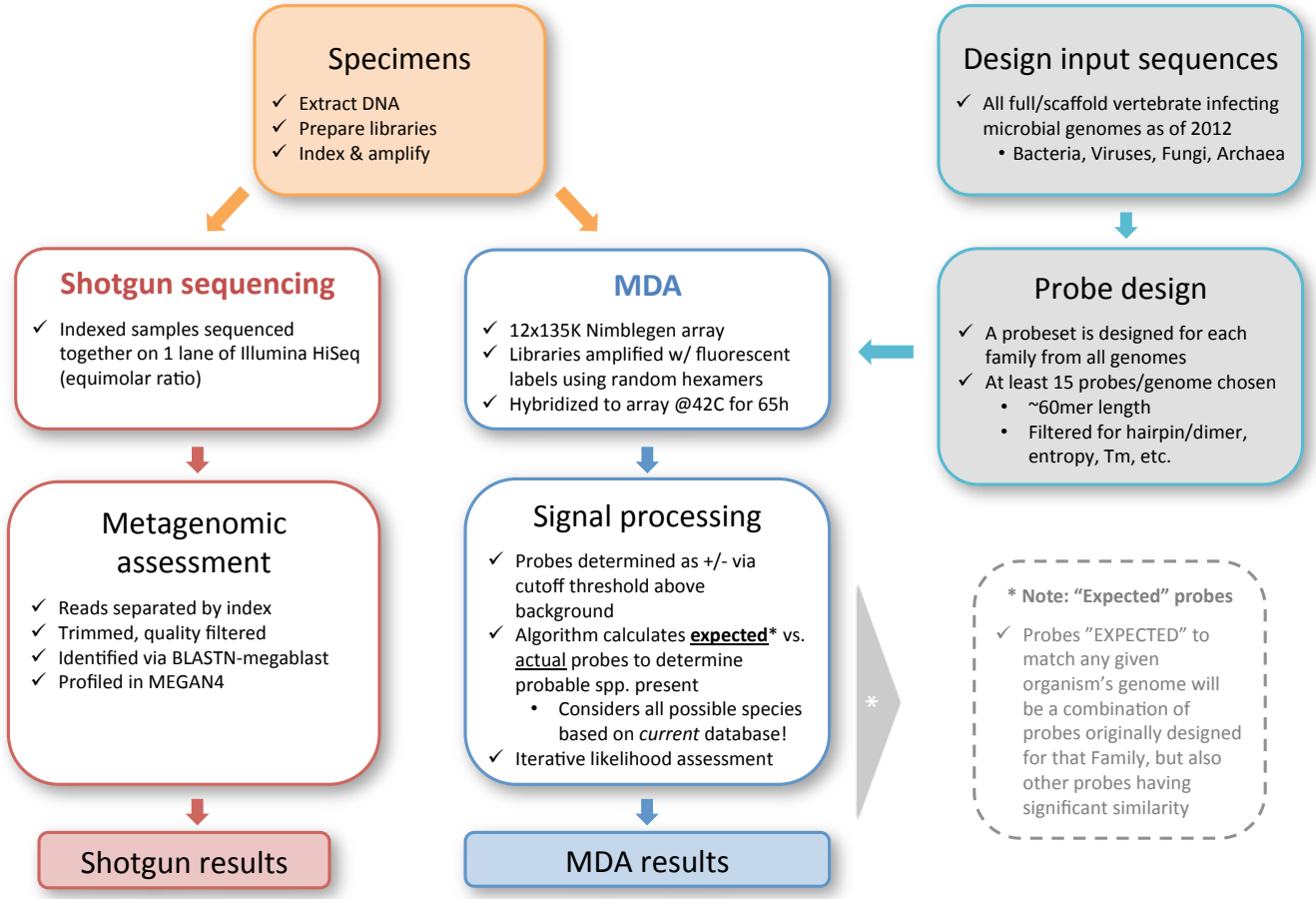


Figure S3. Average LLMDA probe GC% vs average LLMDA probe log intensity, by family.
Analysis is restricted to families used for LLMDA probe design. Families not detected with HTS are represented with blue triangles. Families detected with both methods are represented with red circles.

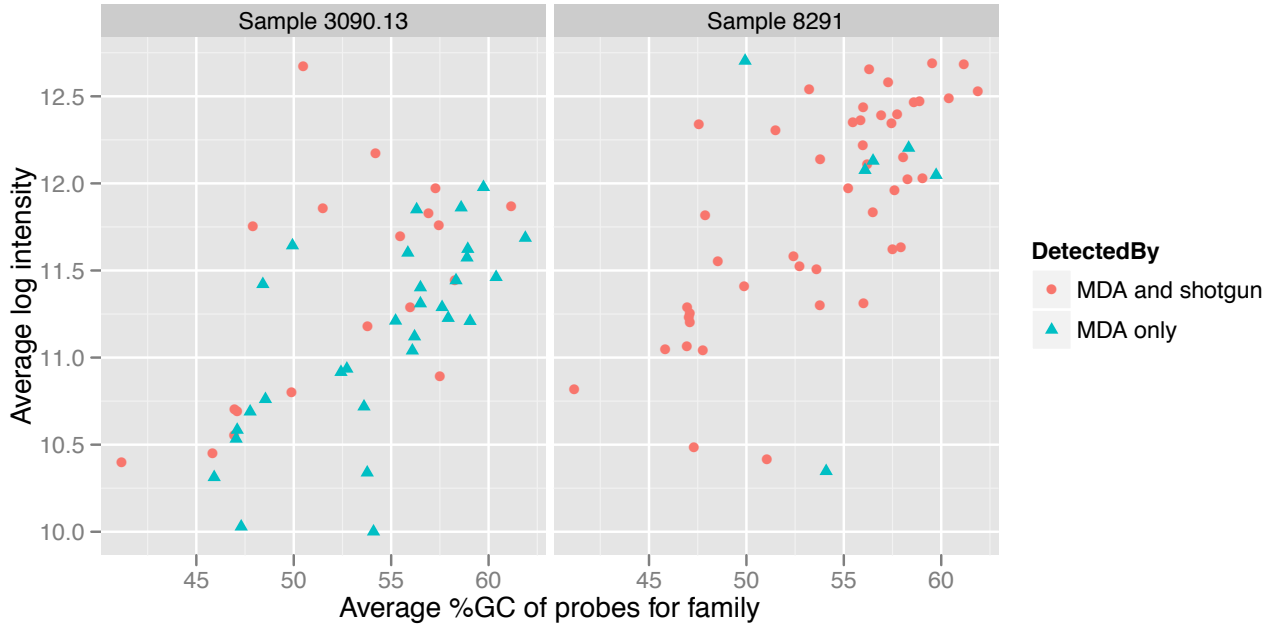
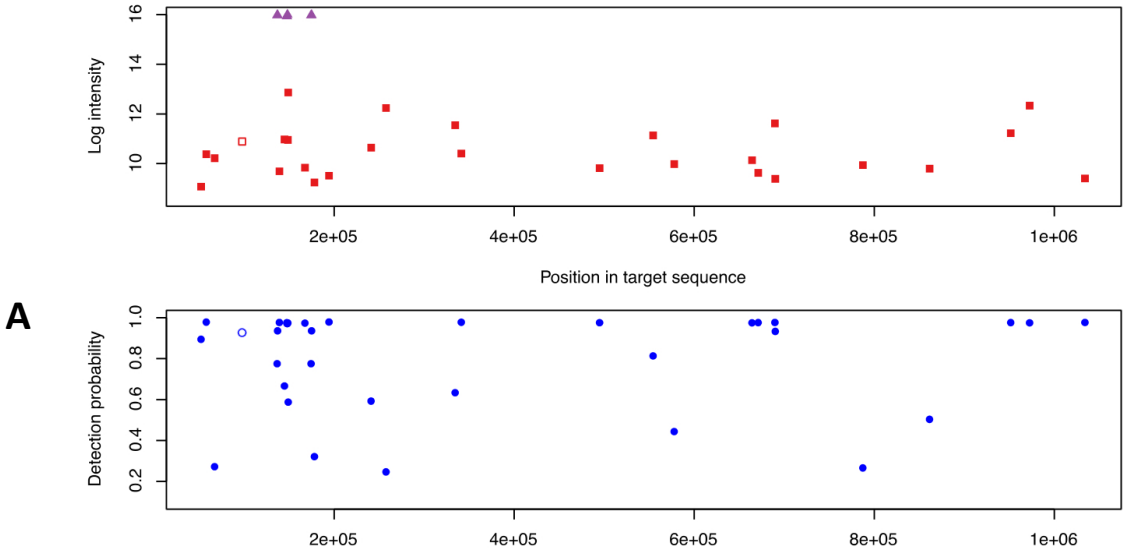


Figure S4. Number of HTS reads vs. HTS GC%, by family

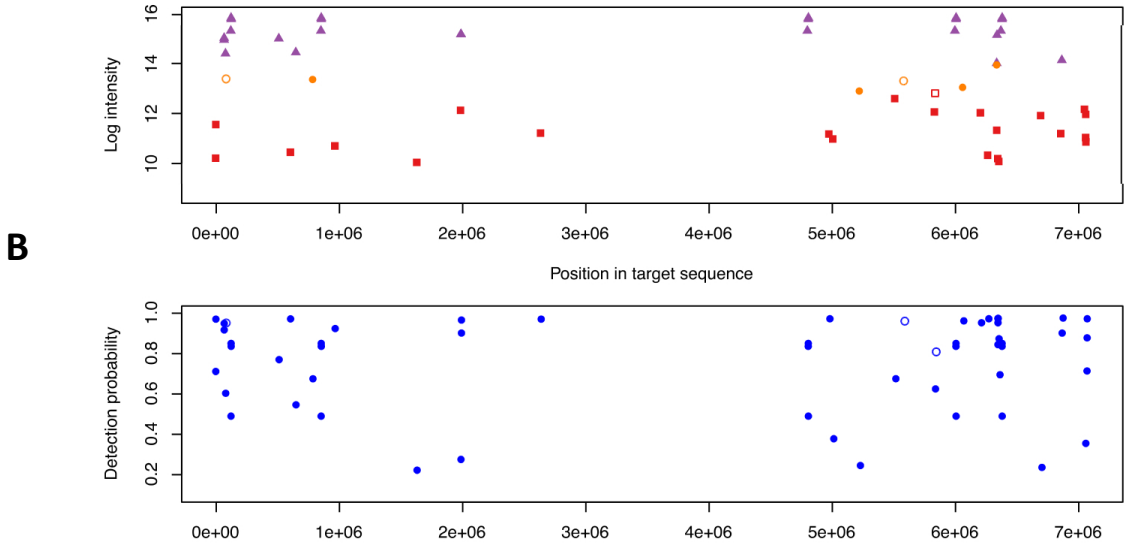
For each bacterial family detected by HTS with probes present on the MDA v5, plots of the total number of HTS reads assigned to that family versus GC% of the HTS reads. Data only shown for those families with HTS representation (3090.13 = 24; 8291 = 81). Blue triangles = families not detected via LLMDA. Red circles = families detected by LLMDA.

Figure S5. Examples of LLMDA detected probe distributions

A: Log intensities vs. genome position for probes targeting *Chlamydia muridarum* on array hybridized to sample 8291, and probe detection probabilities (based on similarity to target sequence) vs. position. Purple triangles indicate that intensity was above the 99th percentile of the negative controls; orange circles indicated intensities between the 99th and the 95th percentiles; red squares indicate intensities below the 95th percentile. Open symbols represent probes that were excluded from the score computation, because they light up non-specifically even when there is no sample present in the hybridization mixture. This target was removed from the predicted set because the only high-intensity probes came from a narrow region of the genome.



B: Log intensities and detection probabilities vs. genome position for probes targeting *Pseudomonas fluorescens* Pf-5 on array hybridized to sample 3090.13. This target was included in the predicted set, since high-intensities are found from most regions of the genome that are covered by high-probability probes.



III. Supplementary References

1 Devault, A. M. *et al.* Second-Pandemic Strain of *Vibrio cholerae* from the Philadelphia Cholera Outbreak of 1849. *N. Engl. J. Med.* **370**, 334-340, doi:doi:10.1056/NEJMoa1308663 (2014).

2 Bos, K. I. *et al.* A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478**, 506-510, doi:10.1038/nature10549 (2011).

3 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17** (2011).

4 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

5 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403-410, doi:10.1006/jmbi.1990.9999 (1990).

6 Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130-D135, doi:10.1093/nar/gkr1079 (2012).

7 Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552-1560, doi:10.1101/gr.120618.111 (2011).

8 Gardner, S., Jaing, C., McLoughlin, K. & Slezak, T. A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics* **11**, 668, doi:10.1186/1471-2164-11-668 (2010).