Rapid Tracing of Resistance Plasmids in a Nosocomial Outbreak Using Optical DNA Mapping

Vilhelm Müller,[1] Nahid Karami,[2] Lena K. Nyberg,[1] Christoffer Pichler,[3] Paola C Torche,[3] Saair Quaderi,[1,3] Joachim Fritzsche,[4] Tobias Ambjörnsson,[3] Christina Åhrén,[2] and Fredrik Westerlund[1]

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, 412 96 Gothenburg, Sweden
[2]Department of Infectious Diseases, Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, 413 46 Gothenburg, Sweden
[3]Department of Astronomy and Theoretical Physics, Lund University, Sölvegatan 14A, 223 62 Lund, Sweden
[4]Department of Applied Physics, Chalmers University of Technology, Kemivägen 9, 412 96 Gothenburg, Sweden

[18 pages, 4 figures, 4 tables]

S.M Supplementary Methods

S.M.1 Problem formulation

In the main text we want to evaluate the similarity of plasmid barcodes in order to monitor the spread of plasmids. This is done by answering the following two questions:

Q1. "Are the two barcodes identical (at some level of significance)?"

Q2. "Is the shorter barcode a part of the longer barcode (at some level of significance)?".

We start with asking Q1. If the answer to this question is negative, we proceed to answer Q2. In order to define the theoretical problem at hand, namely provide means for answering Q1 and Q2, consider two circular barcodes (barcode 1 and barcode 2). In the paper both barcodes are consensus barcodes. Before turning to the general problem of dealing with noisy barcodes, let us assume that the barcodes are noiseless. In such a scenario, despite the absence of noise, two barcodes originating from identical DNA sequences will not be identical, since the intact DNA of interest herein is in its circular form. Hence, if barcode 1 and 2 originate from the same DNA sequence (Q1) barcode 2 must be "slided" across barcode 1 (to the optimal shift position, $\hat{d}$) before they can be compared. Also, as the orientation of one barcode with respect to the other is not known, one must compare the two barcodes for both flip directions, and find the optimal flip, $\hat{f}$. Further, as we describe in the next section, in order to answer Q2, we must circularly shift barcode 2, and find an optimal shift, $\hat{\Delta}$.

In the presence of experimental noise, the problem becomes more challenging. Even if the two barcodes originate from the same underlying DNA sequence they will, due to the noise, differ slightly. It is the purpose here to provide means for answering Q1 and Q2 above for two such noisy barcodes. To that end, we first, in Sec. S.M.2, introduce the Pearson correlation coefficient. This quantity, evaluated at optimal parameters ($\hat{d}$, $\hat{f}$ and $\hat{\Delta}$), is denoted $\hat{C}$ ("best Pearson correlation coefficient") and quantifies the similarity of two barcodes. However, the average best Pearson correlation coefficient for "non-match" barcodes decreases with the length of the barcodes. Hence, one cannot directly use $\hat{C}$ to quantify whether two barcodes are identical. In Sec. S.M.3, we therefore introduce a quantity, which we refer to as p-value. The p-value utilizes randomized barcodes as reference and turns a particular $\hat{C}$ into a quantity insensitive to barcode length. By applying a universal threshold to the p-value we can deem two barcodes are the same, or the same + insert (at some level of significance).

S.M.2 Quantifying the similarity of two barcodes

In this section we introduce the Pearson correlation coefficient which quantifies the similarity of two barcodes. We find that answering Q1 (Q2), introduced in Sec. S.M.1, requires us to introduce two (three) optimal slide/flip/(shift) parameters.

S.M.2.1 Q1. Detecting barcodes originating from identical DNA sequences

In order to introduce means for addressing Q1 (Sec. S.M.1), consider two simple, noiseless barcodes consisting of three pixels each, see Figure S1 (Top). The intensity levels of the barcodes are A, B, C. In our example, barcode 1 is described by an intensity vector $(B_1(1), B_1(2), B_1(3)) = (A, B, C)$ and barcode 2 has intensity levels $(B_2(1), B_2(2), B_2(3)) = (C, A, B)$. Are the two barcodes from an identical DNA sequence? To address this question, we slide barcode 2 across barcode 1 (remember the circular nature of the plasmids) with a shift d, see Figure S1 (Top). For $d = 0$ we compare barcode 2 to the original barcode 1, for $d = 1$ we compare barcode 2 to

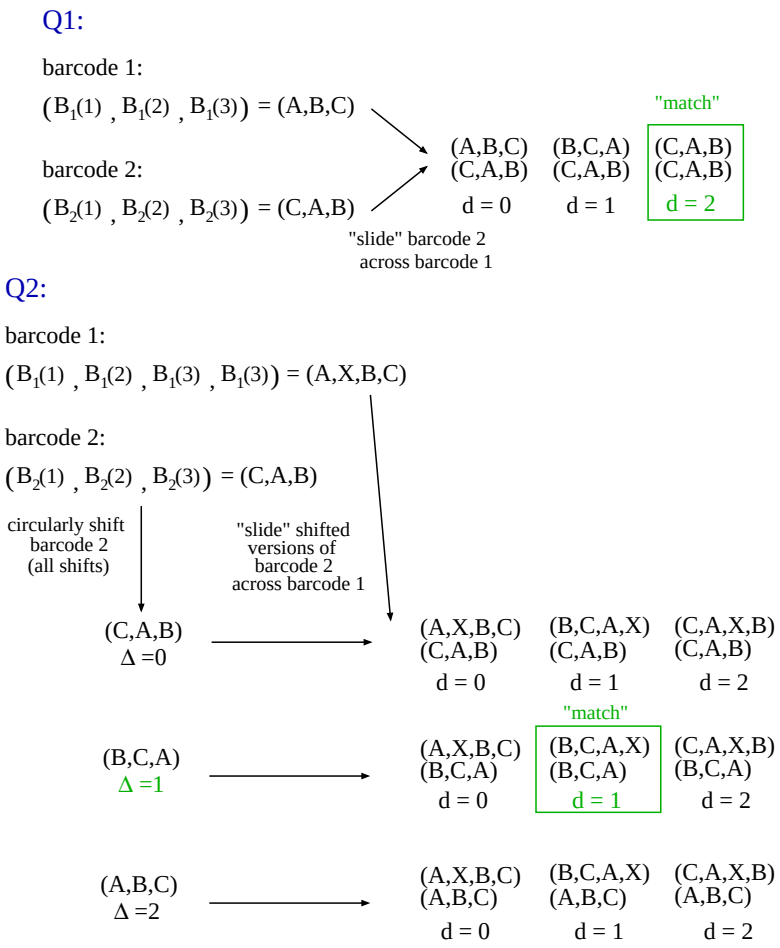

Figure S1: **Finding optimal slide/flip/shift parameters when comparing two barcodes.** (Top) in order to check the similarity of two circular barcodes (1 and 2), barcode 2 is slided across barcode 1 (here, ignoring the unknown relative orientation of the two barcodes). The optimal position is denoted by $\hat{d}$ (here $\hat{d} = 2$). For noisy barcodes the same procedure is applied, but then the optimal position and optimal flip ($\hat{f}$) are determined by maximizing the Pearson correlation coefficient, $C(d, \Delta = 0, f)$ in Eq. (S.1). (Bottom) In order to identify whether barcode 1 has an insert (X), but is otherwise identical (or, highly similar, for noisy barcodes) to barcode 2, the steps under (Top) are performed for all circularly shifted versions of barcode 2. For noisy barcodes the same procedure is applied, but then the optimal position, optimal flip and optimal circular shift ($\hat{\Delta}$) are determined by maximization of the Pearson correlation coefficient, $C(d, \Delta, f)$ in Eq. (S.1).

(B, C, A) and for $d = 2$ we compare barcode 2 to (C, A, B). Thus, for a shift of two pixels, $d = 2$ (and no flip), we get perfect agreement between the two barcodes. This optimal shift is denoted by $\hat{d}$ (here, $\hat{d} = 2$).

Above, we considered noiseless barcodes. Real, experimental, barcodes are however subject to noise. How can we determine $\hat{d}$ and the optimal flip, $\hat{f}$ in such a scenario? In subsection S.M.2.3, we define the Pearson correlation coefficient, $C(d, f)$ [see Eq. (S.1)] between two barcodes at given shift positions, d and flip direction, f. The Pearson correlation coeffcient takes values between -1 and 1. A Pearson correlation coefficient of 1 means that the barcodes are identical, or perfectly correlated, and -1 is then the opposite, perfectly anti-correlated. Two uncorrelated barcodes will have $C = 0$ on average. The optimal parameters $\hat{d}$ and $\hat{f}$ are, for noisy barcodes, simply obtained the by maximizing $C(d, f)$ with respect to d and f. The Pearson correlation coefficient at these best parameters is denoted by $\hat{C}$.

S.M.2.2 Q2. Detecting barcodes originating from identical DNA sequences but where one barcode has an insert

Let us now consider the more challenging question Q2 from Sec. S.M.1. To that end, consider again noiseless barcodes, and assume that barcode 1 has an insert, $(B_1(1), B_1(2), B_1(3), B_1(4)) = (A, X, B, C)$, where X is the inserted region (we, again, ignore the flip), see Figure S1. Barcode 2 is $(B_2(1), B_2(2), B_2(3)) = (C, A, B)$ as before. Now, if we were to proceed as in the previous subsection we would compare barcode 2, i.e. (C, A, B), to either (A, X, B), (X, B, C), (B, C, A) or (C, A, X), and, in neither case will there be a match. Thus, we conclude that the two barcodes are not the same, which is indeed true since barcode 1 has an extra region inserted.

To be able to find if there might be an inserted region, both barcodes have to be shifted, i.e. we now have two parameters d (measuring how much the shorter barcode is slided across the longer one as before) and a parameter Δ which measures how much the shorter barcode is shifted before sliding. We now circularly shift barcode 2 using all possible shifts Δ ($\Delta = 0, ..., N_2 - 1$, where N_2 is the number of pixels in barcode 2). Thus, we now have three versions of barcode 2: $\Delta = 0$ for which barcode 2 is (C, A, B), $\Delta = 1$ for which barcode 2 is shifted to (B, C, A), and $\Delta = 2$, where we have a shifted barcode (A, B, C). For each Δ we go through all possible shifts d, in exactly the same way as in the previous subsection. In the present example we then find that for $d = \hat{d} = 2$ and $\Delta = \hat{\Delta} = 1$ (optimal shift) we compare two identical sequences [the first three pixels in (B, C, A, X) are then compared to (B, C, A)]. Thus, we conclude that barcode 1 is equal to barcode 2 but with an inserted region.

What if the barcodes are "noisy"? In this case, just like in the previous subsection, we determine optimal sliding positions, $\hat{d}$, flip directions, $\hat{f}$, and shift $\hat{\Delta}$ simply by maximizing $C(d, \Delta, f)$ with respect to d, Δ and f. The cross correlation value at these best parameters is, as before, denoted by $\hat{C}$.

S.M.2.3 Pearson correlation coefficient for comparing two barcodes

In this subsection we formally define the Pearson correlation coefficient, which was used in the previous discussions.

Consider two noisy circular consensus barcodes, $B_1(x)$ and $B_2(x)$.[1] Barcode 1 has length N_1 (long barcode), i.e. $x = 1, ..., N_1$ and barcode 2 has length N_2 (short barcode). Without loss of generality we assume that $N_1 \geq N_2$. The sample estimator for the Pearson correlation

[1] A consensus barcode is an average over several barcodes, circularly shifted to optimal positions, with the ends "masked" over a distance equal to three times the standard deviation of the point spread function.

coefficient for comparing the two barcodes is:

$$C(d,\Delta,f)=\frac{1}{N_2-1}\,\frac{\sum_{x=1}^{N_2}[B_1(x+d)-\mu_1(d)][B_2(x+\Delta)-\mu_2]}{\sigma_1(d)\sigma_2}, \tag{S.1}$$

where $d=0,...,N_1-1$ is the "sliding" parameter of the short barcode along the longer one and $\Delta=0,...,N_2-1$ measures the circular shift of the short barcode. The parameter f labels orientation of barcode 2: $f=0$ (original orientation) or $f=1$ (flipped orientation). We leave this parameter implicit in all expressions below. Due to the circular symmetry we have $B_1(x+N_1)=B_1(x)$ and $B_2(x+N_2)=B_2(x)$. The mean barcode intensity, μ_2 and the associated standard deviation, σ_2, for the short barcode are:

$$\mu_2=\frac{1}{N_2}\sum_{x=1}^{N_2}B_2(x), \tag{S.2}$$

and

$$[\sigma_2]^2=\frac{1}{N_2-1}\sum_{x=1}^{N_2}[B_2(x)-\mu_2]^2. \tag{S.3}$$

We also introduce the *local* mean value, $\mu_1(d)$ and *local* standard deviation $\sigma_1(d)$ for the long barcode according to:

$$\mu_1(d)=\frac{1}{N_2}\sum_{x=1}^{N_2}B_1(x+d), \tag{S.4}$$

and

$$[\sigma_1(d)]^2=\frac{1}{N_2-1}\sum_{x=1}^{N_2}[B_1(x+d)-\mu_1(d)]^2. \tag{S.5}$$

Eqs. (S.1)-(S.5) define the Pearson correlation coefficient used throughout this study. When addressing Q1 (Sec. S.M.1) we do not need to shift barcode 2 and hence set $\Delta=0$ above.

Let us finally address computational costs. In a "brute force" approach, the number of operations required for evaluating $C(d,\Delta,f)$ for all d, Δ and f is proportional to $2N_1N_2^2$. However, because of the convolution type structure of the Pearson correlation coefficient, Fast Fourier Transforms (FFT) can be used to bring down computational costs to $N_1N_2\log(N_2)$ scaling.[2]

S.M.3 Turning Pearson correlation coefficients into p-values using a probabilistic framework

In this section we introduce a method for turning the Pearson correlation coefficients introduced in the previous section into a probabilistic framework, by defining a p-value. Our p-value is defined in the usual way, *i.e.*, it is the probability that a Pearson correlation coefficient is larger than or equal to the measure value, given some "zero model". In [1] we based our zero model on *random sequence* barcodes. Here, we introduce a different approach for generating our zero model, phase randomization [2, 3]. Our new method has three advantages compared to the

[2]In order to show explicitly how FFT can be used to evaluate Eq. (S.1) it is convenient to first rescale the shorter barcode to have mean 0 and standard deviation = 1, i.e., define a rescaled barcode, $b_2(x)$ according to: $b_2(x)=[B_2(x)-\mu_2]/\sigma_2$. This rescaling requires on the order of $2N_2$ operations, i.e. is computationally cheap. Eq. (S.1) now becomes (using the fact that $\sum_{x=1}^{N_2}b_2(x+\Delta)=0$): $C(d,\Delta,f)=A\sum_{x=1}^{N_2}[B_1(x+d)b_2(x+\Delta)]$ with $A=1/\sigma_1(d)$. For a given value of d, this expression has the form of a convolution, and can hence be evaluated using FFT. The computational cost is proportional to $N_2\log(N_2)$. Since d takes on N_1 possible values, the total computational cost can be brought to scale as $N_1N_2\log(N_2)$.

previous approach: (i) phase randomization works also in the case when we do not know the underlying DNA sequence, i.e. can be applied directly to experimental barcodes from DNA with unknown sequence, (ii) it is computationally fast (see discussion at the end of subsection S.M.3.1), (iii) the phase randomized barcodes have a degree of statistical similarity to the input barcode (same autocorrelation function), and are therefore very "realistic" looking, see Figure S2.

In subsection S.M.3.1 we introduce our method for generating "randomized' barcodes and in S.M.3.2 we use these randomized barcodes to calculate p-values.

S.M.3.1 Generating random barcodes using Phase randomization

Let us now introduce our method for generating the randomized barcodes using phase randomization [2, 3]. The phase randomization procedure takes one "realistic" barcodes as input, and produces several randomized output barcodes. These output barcodes have identical autocorrelation function as the input barcode. As input barcode we herein use an average (defined below) over a set of all available plasmid theory barcodes from the RefSeq database.[3]

Our algorithm for producing "realistic looking" zero model barcodes is the following:

1. Provide a set of J "realistic" barcodes, $\{B_j(x)\}$, where $j = 1, ..., J$. If experimental barcodes obtained under different experimental conditions are used, then stretch/compress all barcodes to have the same kbp/pixel value, S_{input}, using linear interpolation. The length (in pixels) of barcode j is now N_j, and the pixels for barcode j are labeled by $x = 0, ..., N_j - 1$. The barcode set consists of experimental barcodes or theory barcodes. In this study, we use the $J = 3127$ theory plasmid barcodes from the RefSeq database, which all were rescaled to have mean zero and standard deviation = 1, i.e.,

$$\begin{aligned} \bar{B}_j = \frac{1}{N_j} \sum_{x=0}^{N_j-1} B_j(x) &= 0, \\ \sigma_j^2 = \frac{1}{N_j - 1} \sum_{x=0}^{N_j-1} [B_j(x) - \bar{B}_j]^2 &= 1. \end{aligned} \tag{S.6}$$

   Denote by $N_{\max} = \max\{N_j\}$ the length of the longest barcode in the set.

2. Calculate the Fourier amplitudes (absolute value of the discrete Fourier transforms, DFT) for the set of barcodes from 1. above, i.e.

$$\tilde{B}_j(f_n) = |\text{FFT}\{B_j(x)\}|, \tag{S.7}$$

   where $|...|$ corresponds to taking the absolute value. In practice, the discrete Fourier transform above is evaluated using the fast Fourier Transform (FFT) algorithm. Above, we have frequencies: $f_n = n/N$, with $n = -N/2, ..., 0, ..., N/2 - 1$ if N is even, and $n = -(N-1)/2, ..., 0, ..., (N-1)/2$ if N is odd. In Fourier space, Eqs. (S.6) become

$$\tilde{B}(f_0) = 0, \tag{S.8}$$

$$\frac{1}{N_j(N_j - 1)} \sum_n [\tilde{B}_j(f_n)]^2 = 1, \tag{S.9}$$

   where Eq. (S.8) follows from the DFT definition, and Eq. (S.9) is the Parseval's theorem for DFTs [4].

[3] Plasmid DNA sequences were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/refseq/, June 2015). Based on these sequences, theory barcodes were calculated using the transfer matrix method, see Ref. [1] for details.

3. Interpolate all $\tilde{B}_j(f_n)$ barcodes to the length $N_{\max}$. When performing the interpolation we use that (since $B(x)$ is real valued) $\tilde{B}(f_n)$ is symmetric, $\tilde{B}(-f_n) = \tilde{B}(f_n)$. We leave the Fourier amplitude for f_0 "untouched". We then interpolate the positive frequencies (f_n for $n > 0$) and "fold" the interpolated result to negative frequencies (for N odd). For N even, we keep the Fourier amplitudes for f_0 and f_n with $n = -N/2$ "untouched". After interpolation, we normalize the Fourier amplitudes so that the last equation in Eqs. (S.8) and (S.9) is satisfied. This normalization makes sure that, in real space, the interpolated barcodes have mean 0 and standard deviation 1. We now have J Fourier amplitudes, $\{B_j^{(\text{interp})}(f)\}$, each consisting of $N_{\max}$ frequencies.

4. Average the squared Fourier amplitudes:

$$[\bar{B}(f_n)]^2 = (1/J)\sum_{j=1}^{J}[\tilde{B}_j^{(\text{interp})}(f_n)]^2 \tag{S.10}$$

for all n. The quantity $\bar{B}(f_n)$ serves as our input barcode to be used for phase randomization. Note that the averaging method above makes sure that the average Fourier amplitudes, $\bar{B}(f_n)$, satisfy the relations in Eqs. (S.8) and (S.9). Hence, the corresponding real space barcode will have mean 0 and standard deviation equal to 1.

5. The average Fourier amplitudes, $\bar{B}(f_n)$, allows us to generate randomized barcodes of arbitrary length, N, and with kbp/pixel value, S_{input}. To that end, we interpolate $\bar{B}(f_n)$ to have N frequencies. As above, we interpolate the positive frequencies ($f_n > 0$), and then "fold" the interpolated result to negative frequencies. We also, as above, make sure that Eqs. (S.8) and (S.9) are satisfied for the interpolated barcodes. Then a zero model barcode is obtained by multiplying $\bar{B}(f_n)$ with "symmetrized" random phase factors[4], and the inverse Fourier transform is applied [2, 3]

$$\bar{B}^{(\text{PR})}(x) = \text{IFFT}\{\bar{B}^{(\text{PR})}(f)\}. \tag{S.13}$$

This procedure yields a real valued randomized barcode, $\bar{B}^{(\text{PR})}(x)$, of (arbitrary) length N.

By repeating the last two steps n times, we generate n randomized barcodes, which are used to calculate the p-value, see subsection S.M.3.2.

Figure S2A displays our averaged Fourier amplitudes, $\bar{B}(f)$, based on the plasmid barcode database. Figure S2 B-C shows two barcodes generated using the phase randomization procedure. Note that they look clearly distinguishable, but still contain general features which visually make them resemble each other.

The procedure above only works if S_{input} is the same as the kbp/pixel value for the experimental barcode, S_{exp}. If, $S_{\text{input}} \neq S_{\text{exp}}$ some modification to step 5 has to be done. If the target length, for the fully prepared random barcode, is N, then $\bar{B}(f_n)$, should be interpolated

[4]In practice, this step is performed by drawing $K = (N-1)/2$ (for N odd), or $K = (N-2)/2$ (for N even) uniformly distributed random numbers, $R_k \in [0, 1]$, $k = 0, ..., K$. Then, for positive frequencies we calculate:

$$\bar{B}^{(\text{PR})}(f_k) = \bar{B}(f_k)\exp(2\pi i R_k). \tag{S.11}$$

For negative frequencies we calculate

$$\bar{B}^{(\text{PR})}(f_{-k}) = \bar{B}(f_{-k})\exp(-2\pi i R_k). \tag{S.12}$$

For $f_0 = 0$ we set the phase factor equal to 1. Also, for the case where N is even, we set the phase factor for the frequency $f_{-N/2}$ to 1 (this choice makes sure that the phase randomized barcodes are real).

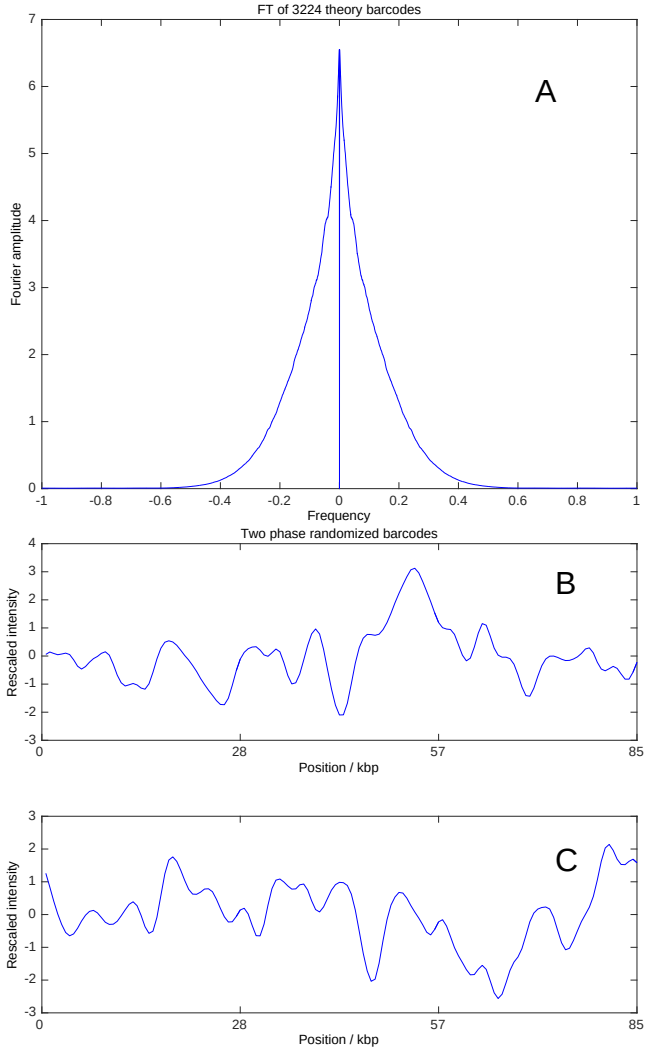

Figure S2: **Phase randomization of plasmid barcodes generates randomized, "realistic looking", DNA barcodes.** (A) Averaged Fourier amplitudes $\bar{B}(f_n)$, based on theory barcodes from the RefSeq plasmid database. (B and C) Two zero model barcodes as obtained using phase randomization based on $\bar{B}(f_n)$. We use all 3127 plasmid barcodes for the average Fourier amplitudes and the randomized barcodes.

to length $N_{\text{raw}} = N(S_{\text{exp}}/S_{\text{input}})$. The new barcode $\bar{B}^{(\text{PR})}(x)$, with length N_{raw}, can now be interpolated to length N in order to change S_{input} with the factor N_{raw}/N. This results in $S_{\text{input,new}} = S_{\text{input}}(N_{\text{raw}}/N) = S_{\text{input}}(S_{\text{exp}}/S_{\text{input}}) = S_{\text{exp}}$ which is the correct kbp/pixel value. After interpolation $\bar{B}^{(\text{PR})}(x)$, has the correct length, N, as well as the correct $S_{\text{input,new}}$ value.

In [1] a different way of calculating random barcodes was introduced: generate random DNA sequences and then make theoretical barcodes out of those. However, the computational cost is much higher for this method compared to the phase randomization method. Using a standard desktop computer of today, generating 1000 barcodes using the present method takes roughly 20 seconds. On the other hand, generating 1000 barcodes from random sequences takes roughly 2700 seconds. Thus, using phase randomization speeds things up with at least a factor 100. Because of the speed advantage, as well as the fact that it represents barcodes better than random sequence barcodes (same average autocorrelation function as of the input barcodes), phase randomization is herein used to generate random zero model barcodes.

S.M.3.2 p-value, definition

The p-value, used in the main text, is defined as

$$p - \text{value} = \int_{\hat{C}}^{\infty} \phi(\hat{C}')d\hat{C}'. \tag{S.14}$$

The quantity $\phi(\hat{C})$ is the probability density for the best cross correlation values obtained when matching a particular barcode to a set of phase randomized barcodes. From the definition above follows that the p-value has the following properties: (1) $0 \leq p-\text{value} \leq 1$, (2) $\langle p-\text{value}\rangle = 1/2$ under the null hypothesis.[5] A p-value smaller than a specifically chosen (small) threshold, p_{thresh}, indicates that there is a significant resemblance (compared to the zero model) between the two barcodes being compared.

Depending on which question, (Q1) or (Q2) in Sec. S.M.1, we are interested in our matching procedures are slightly different:

Q1. When addressing Q1 from Sec. S.M.1, the two barcodes being compared are stretched to the same length, N_{exp} (two barcodes, which are not identical after being stretched to the same length, cannot originate from the same DNA sequence). The same procedure is applied to the random barcodes, i.e. the target length N (as previously discussed) is the same as the length of the two experimental barcodes after stretching. When comparing one of the experimental barcodes to a random barcode, there will be $2N$ number of Pearson correlation coefficients. If this is done for, say, n random barcodes, there will be $2nN$ Pearson correlation coefficients instead.

Q2. If, on the other hand, it is suspected that one of the two barcodes in a pair has an insert, then the barcodes are instead stretched to the same kbp/pixel-value. The stretching is done to ensure that the pixels from one barcode contain the same amount of information as the pixels from the other. A concern here is that the kbp/pixel-values might have some experimental uncertainty associated with them, and thus each barcode is allowed

[5] What is the meaning of $\langle p-\text{value}\rangle = 1/2$ under the null hypothesis? This means that if we match a "new" zero model barcode to the set of all "previous" zero model barcodes, then the value of this match will on average give p-value $= 1/2$. This result follows immediately because p-values, under the null hypothesis, are uniformly distributed on $[0, 1]$. Mathematically the latter result is straightforward to derive: the PDF of a p-value$=p$ is formally given by $\rho(p) = \int_{-\infty}^{\infty} \delta[p - \int_{\hat{C}}^{\infty} \phi(\hat{C}')d\hat{C}']\phi(\hat{C})d\hat{C}$, where $\delta[z]$ is the Dirac delta-function. Making the change of variables, $t = \int_{\hat{C}}^{\infty} \phi(\hat{C}')d\hat{C}'$ we find $\rho(p) = \int_0^1 \delta(p-t)dt$, i.e. $\rho(s) = 1$ if $0 \leq p \leq 1$ and zero otherwise. Thus, indeed, p-values are uniformly distributed on $[0, 1]$.

to be stretched, within some uncertainty interval, when doing the comparison. Since the barcodes are not the same length, and both have to be circularly permuted, each comparison generates $2N_1N_2$ correlation coefficients, where N_1 and N_2 are the lengths of the first and the second barcode, respectively. If this is done using n random barcodes, there will be $2nN_1N_2$ Pearson correlation coefficients generated. Adding stretching for uncertainty in kbp/pixel value as a parameter (ignoring the change in length N_1), there will be another factor, a (number of stretches) to consider, and the total number of Pearson correlation coefficients becomes $2anN_1N_2$. We use $a = 11$ and allow changes around the experimental value, $S_{\text{exp}} = 0.5$ kbp/pixel, by 5 percent.

What form will $\phi(\hat{C})$ have? To answer this question, assume that we have K best Pearson correlation coefficients, $\hat{C}_k$ for $k = 1, ..., K$. For large enough number of best Pearson correlation coefficients, these $\hat{C}$s are expected to be distributed according to the Gumbel PDF (probability density function for the best) [5]

$$\phi(\hat{C}) = \frac{1}{\beta} \exp[-(y + e^{-y})], \tag{S.15}$$

where $y = y(\hat{C}) = (\hat{C} - \kappa)/\beta$, with parameters κ and β. To be able to fit these two parameters, we use method of moments (moment matching). For large K, we first estimate β using

$$\beta = \frac{\sqrt{6}}{\pi}\sigma, \tag{S.16}$$

where $\sigma^2 = [1/(K-1)] \sum_{k=1}^{K} (\hat{C}_k - \mu)^2$, i.e., σ^2 is the sample variance of the $\hat{C}_k$s with mean $\mu = (1/K) \sum_{k=1}^{K} \hat{C}_k$. The parameter κ is subsequently determined by

$$\kappa = \mu - \beta\gamma, \tag{S.17}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. Representative histograms alongside moment matching fitted $\phi(\hat{C})$ for two plasmids are displayed in Figure S3.

Plugging Eq. (S.15) in to Eq. (S.14), and using the known cumulative distribution for the Gumbel PDF, we obtain the explicit expression

$$p - \text{value} = 1 - \exp[-e^{-(\hat{C}-\kappa)/\beta}], \tag{S.18}$$

which is used throughout this study.

S.M.3.3 Symmetrizing p-values

When turning a Pearson correlation coefficient, obtained by comparing barcode 1 (length N_1) and barcode 2 (length N_2), into an p-value, certain care is required, in particular if the two barcodes are of different lengths and/or the barcodes are very dissimilar. As seen in Tables S1 and S2 we get slightly different results if barcode 1 is matched to phase randomized barcodes of length N_2, than if barcode 2 is matched to phase randomized barcodes of length N_1. This asymmetry effect is, in general, minor but, nevertheless, needs to be addressed.

In order to tackle the asymmetry issue, let us denote by p_1 the p-value obtained by matching to phase randomized barcodes of length N_1, and by p_2 the p-value obtained by matching to phase randomized barcodes of length N_2. A standard method for combining p-values is Fisher's combined probability test [6]. One introduces

$$q_i = -\log(p_i), \qquad i = 1, ..., N, \tag{S.19}$$

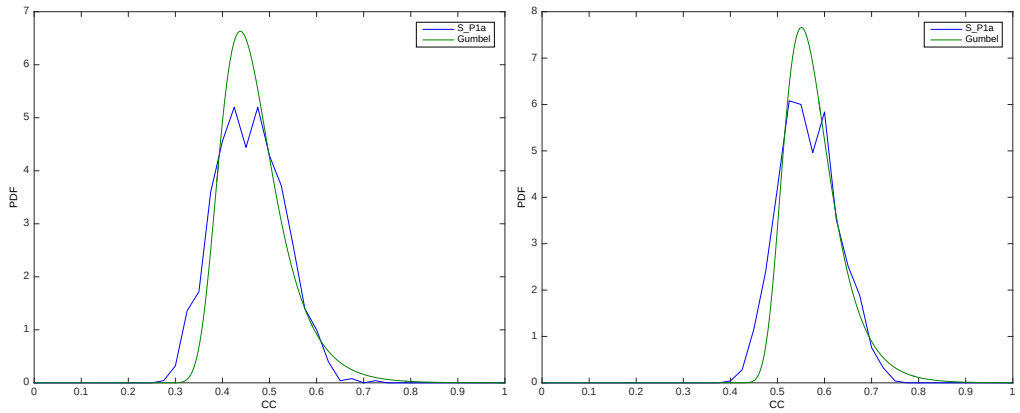

Figure S3: **Fitting of $\hat{C}$ to a Gumbel probability density.** Illustration of the normalized histogram $\phi(\hat{C})$ and fitted Gumbel probability densities using moment matching, see Eq. (S.15). The histogram uses data obtained by matching consensus barcodes for plasmid from the shared plasmid in isolate P1a (see main text) to 1000 phase randomized barcodes, see subsection S.M.3.1. These fits provides estimates for the Gumbel parameters, β and κ, which are subsequently used to calculate p-values using Eq. (S.18). (Left) Plasmid barcodes where treating using the Q1 methodology, see Sec. S.M.3.2. The common length was 81 kbps. (Right) Barcodes treated using the Q2 methodology. The plasmid's the target length was 74 kbp (length of plasmid in isolate P2a) and we used $a = 11$ stretching factors within 5 percent of this value. Notice that we, due to the stretching approach for Q2, generate more Pearson correlation coefficients (more "attempts") for Q2 than for Q1. This results in the average $\hat{C}$ being higher for Q2 than for Q1.

where, here, $N = 2$. Since the p_i are uniformly distributed (see footnote in Sec. S.M.3.2), the q_i are exponentially distributed, i.e. the associated PDF is $P(q) = \exp(-q)$. In Fisher's method one then defines the sum

$$\kappa = 2\sum_{i=1}^{N} q_i \tag{S.20}$$

and denotes by $\varrho(\kappa)$ the PDF for κ. The combined p-value is then calculated as

$$p_{\text{tot}} = \int_{\hat{\kappa}}^{\infty} \varrho(\kappa) d\kappa, \tag{S.21}$$

where $\hat{\kappa}$ is the observed value for κ as defined in Eq. (S.20). As noted in the footnote in Sec. S.M.3.2, p-values should be uniformly distributed on $[0, 1]$ under the null hypothesis. By definition, p_{tot} as defined above satisfies this condition.

What is $\varrho(\kappa)$ appearing in Eq. (S.21)? If the p_i are *independent*, then κ has a χ^2 distribution with $2N$ degrees of freedom [6]. However, in our case p_1 and p_2 are highly correlated. For instance, if p_1 is small, then so is p_2. If the two p-values were perfectly correlated the normalized histogram of the average, $\kappa/(2N)$, will be identical to the normalized histogram of q_1 values. Thus, in this perfectly correlated scenario, we have $P(\kappa) = [1/(2N)] \exp[-\kappa/(2N)]$. Using this as an approximation to the "true" PDF for κ, Eq. (S.21) becomes

$$p_{\text{tot}} = \exp[-\hat{\kappa}/(2N)] = \exp[-(1/N)\sum_{i=1}^{N} \hat{q}_i] = \exp[(1/N)\sum_{i=1}^{N} \log[\hat{p}_i]], \tag{S.22}$$

where $\hat{q}_i$ and $\hat{p}_i$ are observed values. Simplifying the expression above we obtain:

$$p_{\text{tot}} = [\prod_{i=1}^{N} \hat{p}_i]^{1/N}. \tag{S.23}$$

For our case, namely $N = 2$, the method for computing the symmetrized p-value is then simply to take the geometric mean,

$$p_{\text{tot}} = \sqrt{p_1 p_2}, \tag{S.24}$$

of the two constituting p-values.

S.T Supplementary Tables

From the outbreak, the shared plasmid detected in all nine isolates were compared with the larger plasmid found in all ST131 samples, using the p-value methodology (see Supplementary Methods). Furthermore, consensus barcodes (of three previously sequenced plasmids found in the NCBI RefSeq Database), R100 from Shigella flexneri 2b (measured size 90.0 ± 5.6 kbp)[7], RP1 from Pseudomonas aeruginosa (measured size 58.5 ± 4.0 kbp)[8], and pUUH239.2 (pUUH) from Klebsiella pneumoniae (measured size 221.3 ± 10.3 kbp)[9] were included as controls.

In Tables S1 and S2, we show p-values for all plasmid pairs for Q1 and Q2 analysis respectively. Tables S3 and S4 show the associated symmetrized p-values introduced in Sec. S.M.3.3.

	P1a S	P1b S	P2a S	P2b S	P2c S	P3a S	P3b S	P4a S	P4b S	R100	RP1	pUUH	P1a L	P1b L	P2a L	P2c L	P4a L
P1a S	0,00	0,03	0,01	0,23	4,38	0,03	3,44	2,14	3,89	36,21	-	-	-	-	-	-	-
P1b S	0,13	0,00	0,11	0,12	37,06	0,27	41,00	38,84	33,92	20,90	-	-	-	-	-	-	-
P2a S	0,01	0,02	0,00	0,04	6,21	0,01	32,03	23,03	11,40	34,53	-	-	-	-	-	-	-
P2b S	0,63	0,11	0,20	0,00	43,47	0,57	75,67	63,09	34,05	39,69	-	-	-	-	-	-	-
P2c S	9,83	29,60	11,26	34,06	0,00	6,48	0,06	0,18	0,72	99,99	-	-	-	-	-	-	-
P3a S	0,04	0,09	0,02	0,22	4,00	0,00	23,80	34,83	7,07	86,49	-	-	-	-	-	-	-
P3b S	2,53	17,17	22,39	43,29	0,01	16,59	0,00	0,01	0,03	95,23	-	-	-	-	-	-	-
P4a S	1,27	13,64	12,95	27,00	0,02	19,43	0,01	0,00	0,04	99,28	-	-	-	-	-	-	-
P4b S	0,57	3,89	2,60	5,03	0,01	0,84	0,00	0,00	0,00	77,82	-	-	-	-	-	-	-
R100	69,61	20,04	58,40	44,99	100,0	97,99	99,97	100,0	99,91	0,00	-	-	-	-	-	-	-
RP1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
pUUH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
P1a L	-	-	-	-	-	-	-	-	-	-	-	-	0,00	0,02	0,01	0,35	0,06
P1b L	-	-	-	-	-	-	-	-	-	-	-	-	0,01	0,00	0,00	0,03	0,00
P2a L	-	-	-	-	-	-	-	-	-	-	-	-	0,00	0,00	0,00	0,02	0,00
P2c L	-	-	-	-	-	-	-	-	-	-	-	-	0,04	0,01	0,00	0,00	0,01
P4a L	-	-	-	-	-	-	-	-	-	-	-	-	0,02	0,01	0,00	0,04	0,00

Table S1. **p-values for all plasmid pairs for Q1.** p-values for all pairs of barcodes differing less than 20% in length (for larger difference they can not be regarded as the same), calculated using the method in Sec. S.M.3. S = shared plasmid which was detected in all isolated. R100, RP1 and pUUH = consensus of sequenced plasmids from the RefSeq database. L = large plasmids found in ST131 isolates. All the p-values are in %.

	P1a S	P1b S	P2a S	P2b S	P2c S	P3a S	P3b S	P4a S	P4b S	R100	RP1	pUUH	P1a L	P1b L	P2a L	P2c L	P4a L
P1a S	0,00	0,08	0,03	0,67	0,23	0,09	0,09	0,06	0,23	7,08	44,23	69,76	38,35	29,93	53,31	69,63	70,81
P1b S	0,42	0,00	0,30	0,38	0,23	0,78	0,21	0,45	3,07	52,71	97,76	32,16	34,63	23,04	13,92	44,39	56,38
P2a S	0,02	0,04	0,00	0,13	0,24	0,03	0,06	0,05	0,14	30,39	56,51	57,47	34,89	23,61	43,40	69,68	53,97
P2b S	2,04	0,23	0,60	0,00	0,43	1,29	0,17	0,43	1,54	81,91	99,88	13,43	57,26	75,43	45,12	100,0	71,37
P2c S	0,73	0,32	0,89	0,67	0,00	0,43	0,14	0,37	4,24	13,67	42,66	79,51	27,56	19,80	25,29	10,91	63,25
P3a S	0,09	0,20	0,05	0,67	0,23	0,00	0,20	0,10	0,16	69,84	99,73	33,84	51,95	90,07	66,70	57,52	49,62
P3b S	0,16	0,14	0,15	0,11	0,03	0,18	0,00	0,02	0,04	1,66	11,07	50,79	86,14	59,14	61,93	70,65	91,08
P4a S	0,10	0,34	0,16	0,27	0,12	0,10	0,02	0,00	0,08	15,60	31,30	33,01	83,71	97,24	56,40	86,05	88,86
P4b S	0,06	0,60	0,08	0,42	0,52	0,04	0,01	0,02	0,00	35,35	22,56	92,85	25,93	98,59	94,01	99,63	31,69
R100	22,57	72,50	61,10	95,42	23,98	95,66	8,36	61,26	97,50	0,00	34,73	10,73	49,03	11,36	11,32	1,10	7,64
RP1	40,26	78,25	46,60	95,05	29,43	99,92	11,26	38,02	40,47	13,46	0,00	14,27	73,89	54,65	20,99	26,71	31,58
pUUH	99,09	68,49	94,76	45,35	98,56	72,88	88,59	83,71	100,0	22,11	54,29	0,00	13,19	97,76	42,72	91,11	55,36
P1a L	93,58	64,36	79,99	98,09	69,80	98,08	100,0	100,0	89,18	77,34	100,0	22,52	0,00	0,04	0,02	1,08	0,15
P1b L	57,79	28,77	38,41	94,67	31,82	99,56	92,99	100,0	100,0	16,83	98,61	96,87	0,02	0,00	0,00	0,09	0,01
P2a L	96,94	22,01	72,99	77,32	43,83	93,99	96,78	99,29	100,0	13,92	78,97	42,88	0,01	0,00	0,00	0,05	0,01
P2c L	81,76	32,77	63,19	99,85	12,00	58,78	76,19	97,72	100,0	0,93	37,23	61,69	0,14	0,03	0,01	0,00	0,04
P4a L	99,35	80,13	84,80	94,61	82,68	74,65	99,95	100,0	89,14	11,63	95,21	51,06	0,07	0,01	0,01	0,17	0,00

Table S2. ***p*-values for all plasmid pairs for Q2.** p-values for all pairs of barcodes, calculated using the method in Sec. S.M.3. S = shared plasmid which was detected in all isolated. R100, RP1 and pUUH = consensus of sequenced plasmids from the RefSeq database. L = large plasmids found in ST131 isolates. All the p-values are in %.

	P1a S	P1b S	P2a S	P2b S	P2c S	P3a S	P3b S	P4a S	P4b S	R100	RP1	pUUH	P1a L	P1b L	P2a L	P2c L	P4a L
P1a S	-																
P1b S	0,07	-															
P2a S	0,01	0,04	-														
P2b S	0,38	0,11	0,09	-													
P2c S	6,56	33,12	8,37	38,48	-												
P3a S	0,03	0,15	0,02	0,36	5,09	-											
P3b S	2,95	26,53	26,78	57,23	0,02	19,87	-										
P4a S	1,65	23,02	17,27	41,28	0,07	26,01	0,01	-									
P4b S	1,48	11,49	5,45	13,09	0,09	2,43	0,00	0,01	-								
R100	50,21	20,47	44,91	42,26	100,0	92,06	97,57	99,64	88,18	-							
RP1	-	-	-	-	-	-	-	-	-	-	-						
pUUH	-	-	-	-	-	-	-	-	-	-	-	-					
P1a L	-	-	-	-	-	-	-	-	-	-	-	-	-				
P1b L	-	-	-	-	-	-	-	-	-	-	-	-	0,03	-			
P2a L	-	-	-	-	-	-	-	-	-	-	-	-	0,01	0,00	-		
P2c L	-	-	-	-	-	-	-	-	-	-	-	-	0,69	0,06	0,03	-	
P4a L	-	-	-	-	-	-	-	-	-	-	-	-	0,14	0,01	0,01	0,09	-

Table S3. **Symmetrized *p*-values for all plasmid pairs for Q1.** Symmetrized p-values, p_{tot}, see Eq. (S.24), for all pairs of barcodes differing less than 20% in length (for larger difference they can not be regarded as the same). S = shared plasmid which was detected in all isolated. R100, RP1 and pUUH = consensus of sequenced plasmids from the RefSeq database. L = large plasmids found in ST131 isolates. All the p-values are in %.

	P1a S	P1b S	P2a S	P2b S	P2c S	P3a S	P3b S	P4a S	P4b S	R100	RP1	pUUH	P1a L	P1b L	P2a L	P2c L	P4a L
P1a S	-																
P1b S	0,19	-															
P2a S	0,03	0,11	-														
P2b S	1,17	0,30	0,28	-													
P2c S	0,41	0,27	0,46	0,53	-												
P3a S	0,09	0,40	0,03	0,93	0,31	-											
P3b S	0,12	0,17	0,10	0,14	0,07	0,19	-										
P4a S	0,07	0,39	0,09	0,34	0,21	0,10	0,02	-									
P4b S	0,12	1,36	0,10	0,81	1,49	0,08	0,02	0,04	-								
R100	12,64	61,81	43,09	88,41	18,11	81,74	3,73	30,91	58,71	-							
RP1	42,20	87,46	51,32	97,44	35,43	99,82	11,17	34,50	30,21	21,62	-						
pUUH	83,14	46,93	73,79	24,68	88,52	49,66	67,08	52,57	96,36	15,40	27,84	-					
P1a L	59,91	47,21	52,83	74,94	43,86	71,38	92,81	91,49	48,08	61,58	85,96	17,23	-				
P1b L	41,59	25,75	30,11	84,50	25,10	94,70	74,16	98,61	99,29	13,83	73,41	97,31	0,03	-			
P2a L	71,89	17,50	56,28	59,07	33,29	79,18	77,42	74,84	96,96	12,55	40,71	42,80	0,01	0,00	-		
P2c L	75,45	38,14	66,35	99,92	11,44	58,15	73,37	91,70	99,81	1,01	31,54	74,97	0,39	0,05	0,02	-	
P4a L	83,88	67,21	67,65	82,17	72,32	60,86	95,41	94,26	53,15	9,43	54,83	53,16	0,10	0,01	0,01	0,08	-

Table S4. **Symmetrized p-values for all plasmid pairs for Q2.** Symmetrized p-values, p_{tot}, see Eq. (S.24). S = shared plasmid which was detected in all isolated. R100, RP1 and pUUH = consensus of sequenced plasmids from the RefSeq database. L = large plasmids found in ST131 isolates. All the p-values are in %.

As noted in Supplementary Methods, p-values should be uniformly distributed on $[0, 1]$ for "non-match" plasmid barcodes. A consequence of this is that that the expected p-value should equal $1/2$ and the standard deviation should be $\sqrt{1/12} \approx 0.2887$ under the null hypothesis. To validate our method, we calculated the mean and standard deviation from Table S4 of symmetrized p-values for Q2 for all non-match cases. We found: mean = 0.5801 and standard deviation = 0.2810, thus validating the p-value method for identifying "match"/"non-match" plasmid pairs.

S.F Supplementary Figures

In Figure S4, we turn the p-value (Supplementary Method) into a categorization tool. If the p-value is smaller than some threshold p_{thresh}, then the two plasmids are deemed significantly similar (Q1 in Supplementary Methods visualized in green). For two plasmids not the same, i.e. for p-value$\geq p_{\text{thresh}}$, a plasmid can still be deemed "the same + an insert" (Q2 in Supplementary Methods) if below the threshold for Q2 (yellow). We here use $p_{\text{thresh}} = 0.01 = 1\%$.

As seen in Figure S4, neither of the three controls resulted in p-values below s_{thresh} for either Q1 or Q2 when compared to the consensus barcodes of the plasmids found in the resistance outbreak. Furthermore, none of the shared plasmids found in all nine isolates showed high enough similarity in order to be regarded the same (+ insert Q2) as the larger plasmid found in all ST131 isolates.

However, when comparing the shared plasmids, all of them (except the combination of P1b and P4b) showed p-values lower than s_{thresh} for Q1 or/and Q2. As discussed in the main text, the isolates in which the shared plasmid visually displayed an inserted DNA region (P2c, P3b, P4a and P4b), were also separated from the remaining five isolates by using Q1 and Q2. It should also be noted that all of the large plasmids found in the ST131 isolates rendered p-values below p_{thresh}, indicating that it is the same plasmid that is found both over time (P1a vs P1b and P2a vs P2c) and in different patients (P1, P2 and P4) suggesting that the analysis is correct.

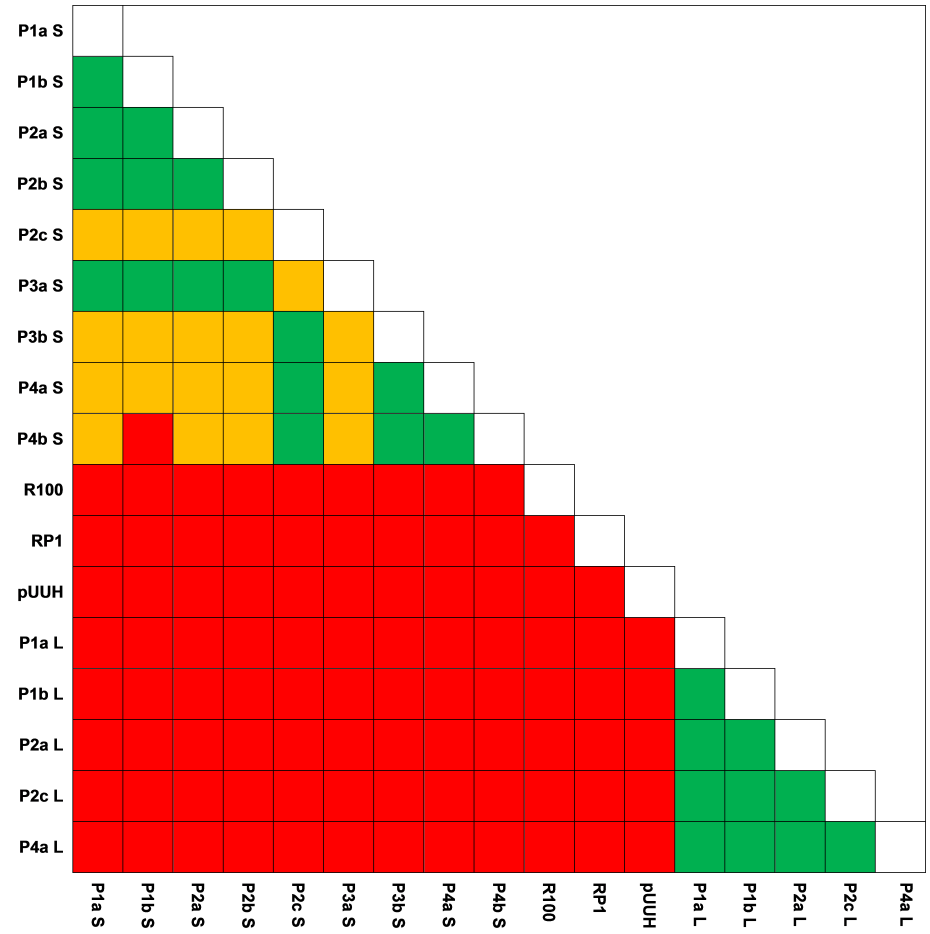

Figure S4: **Plasmid pair categorization.** Imposing a p-value threshold on the symmetrized p-values in Tables S3 and S4 allow us to categorize a particular pair of barcodes into one of three categories: green = plasmids are deemed significantly similar, yellow = the pair are significantly similar with an insert, red = not identical. S = shared plasmid which was detected in all isolated. R100, RP1 and pUUH = consensus of sequenced plasmids from the RefSeq database. L = large plasmids found in ST131 isolates. We used $p_{\text{thresh}} = 0.01 = 1\%$.

References

[1] Nilsson, A. N., Emilsson, G., Nyberg, L.K., Noble, C., Svensson Stadler, L., Fritzsche, J., Moore, E.R.B., Tegenfeldt, J.O., Ambjörnsson, T., & Westerlund, F. Competitive binding-based optical DNA mapping for fast identification of bacteria-multi-ligand transfer matrix theory and experimental applications on Escherichia coli. Nucleic acids research **42**, e118 (2014).

[2] Schreiber, T. Constrained randomization of time series data. Physical Review Letters **80**, 2105-2108 (1998).

[3] Schreiber, T., & Schmitz, A. Surrogate time series. Physica D: Nonlinear Phenomena, **142**, 346-382 (2000).

[4] Easton Jr, R.L. Fourier methods in imaging. John Wiley & Sons (2010).

[5] Karlin, S., and Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proceedings of the National Academy of Sciences **87**, 2264-2268 (1990).

[6] Whitlock, M.C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. Journal of evolutionary biology, **18**, 1368-1373 (2005).

[7] Silver, L., Chandler, M., la Tour, de, E. B. and Caro, L. Origin and direction of replication of the drug resistance plasmid R100. 1 and of a resistance transfer factor derivative in synchronized cultures. Journal of Bacteriology **131**, 929-942 (1977).

[8] Pansegrau, W. et al. Complete Nucleotide Sequence of Birmingham IncPα Plasmids: Compilation and Comparative Analysis. Journal of Molecular Biology **239**, 623-663 (1994).

[9] Sandegren, L., Linkevicius, M., Lytsy, B., Melhus, A. and Andersson, D. I. Transfer of an Escherichia coli ST131 multiresistance cassette has created a Klebsiella pneumoniae-specific plasmid associated with a major nosocomial outbreak. Journal of Antimicrobial Chemotherapy **67**, 74 - 83 (2011).