High-level Multiplexing in Digital PCR With Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis

Ahmad Moniri,[†,¶] Luca Miglietta,[†,¶] Alison Holmes,[‡]
Pantelis Georgiou,[†] and Jesus Rodriguez-Manzano[*,†,‡]

†*Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering, Imperial College London, UK*
‡*NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, Department of Infectious Disease, Imperial College London, UK*
¶*These authors contributed equally.*

E-mail: j.rodriguez-manzano@imperial.ac.uk
Phone: +44 (0)20 7594 0843

Abstract

Digital polymerase chain reaction (dPCR) is a mature technique that has enabled scientific breakthroughs in several fields. However, this technology is primarily used in research environments with high-level multiplexing representing a major challenge. Here, we propose a novel method for multiplexing, referred to as *amplification and melting curve analysis* (AMCA), which leverages the kinetic information in real-time amplification data and the thermodynamic melting profile using an affordable intercalating dye (EvaGreen). The method trains a system comprised of supervised machine learning models for accurate classification, by virtue of the large volume of data from dPCR platforms. As a case study, we develop a new 9-plex assay to detect mobilized colistin resistant (*mcr*) genes as clinically relevant targets for antimicrobial resistance. Over 100,000 amplification events have been analyzed, and for the positive reactions, the AMCA approach reports a classification accuracy of $99.33 \pm 0.13\%$, an increase of 10.0% over using melting curve analysis. This work provides an affordable method of high-level multiplexing without fluorescent probes, extending the benefits of dPCR in research and clinical settings.

Introduction

Detecting and quantifying nucleic acids are important tasks in several fields, where the real-time polymerase chain reaction (qPCR) remains the most common technique.[1–7] More recently, the use of digital PCR (dPCR) has been flourishing due to the several advantages over conventional qPCR, such as: (i) lack of references or standards; (ii) high precision in quantification; (iii) tolerance to inhibitors; and (iv) the capability to analyze complex mixtures.[8–11] Therefore, dPCR has enabled scientific breakthroughs in clinical microbiology, gene expression and precision cancer research, among others.[12–14]

Multiplex assays provide a practical solution for nucleic acid detection in a single reaction, reducing the time, cost and amount of sample required, at the expense of technical complexity.[15,16] Current approaches based on fluorescent probes are expensive and require lengthy optimization which is challenging for high-throughput applications.[17,18] Intercalating

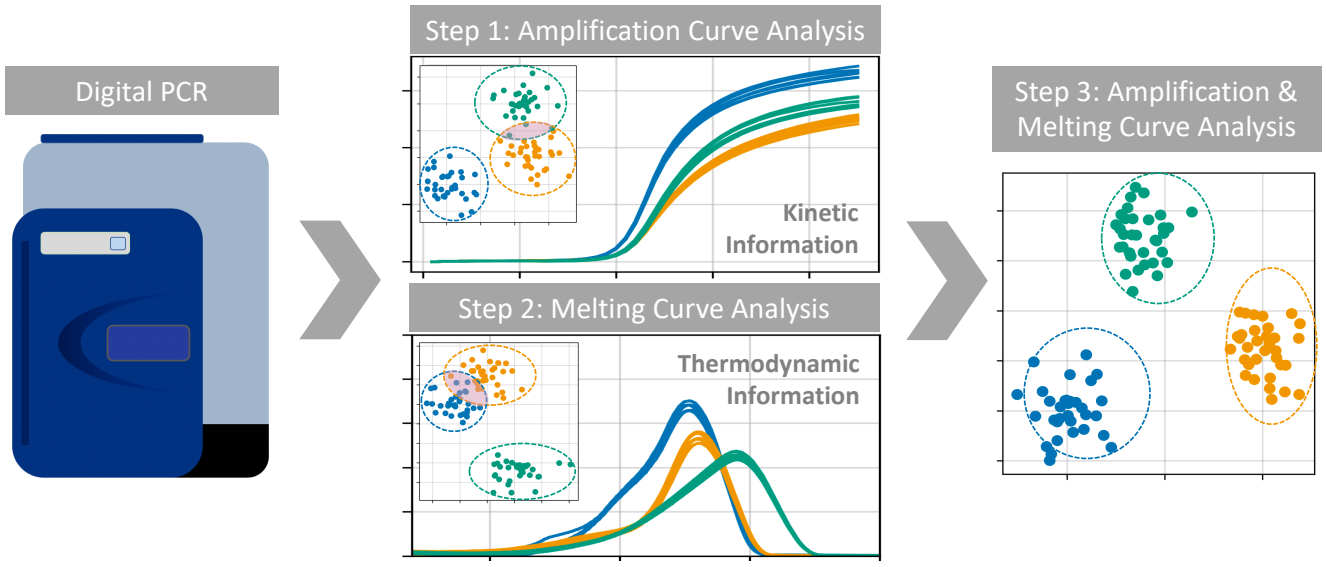


Figure 1: Concept of the proposed method. Amplification and melting curve data from real-time dPCR instrument (e.g. Fluidigm BioMark HD) is extracted. Subsequently, machine learning models are trained to classify multiple targets for both datasets individually. For high-level multiplexing, both methods provide insufficient accuracy (indicated by overlapping data distributions highlighted by the shaded regions). However, the proposed method, referred to as amplification and melting curve analysis, or AMCA, takes into account both kinetic and thermodynamic information in order to classify the targets accurately. Note: Three targets have been used to simplify the illustration of the concept.

dyes provide a suitable and alternative chemistry which is affordable and does not require in-silico design. However, since intercalating dyes bind to any double-stranded DNA, the prospect of non-specific amplification are typically addressed with further post-PCR analyzes such as gel electrophoresis, melting curve analysis or sequencing methods.

Current multiplex dPCR methods that are dependent on intercalating dyes are either limited to analyzing real-time amplification data or performing melting curve analysis, since gel electrophoresis or sequencing is not possible.[19,20] Since most commercially available platforms (such as Fluidigm EP1, Bio-Rad QX200 and Stilla Naica systems) do not have real-time data acquisition, the most common approach for multiplexing uses the final fluorescent intensity (FFI) of the amplification curve to distinguish between targets.[18] Reported studies showed that specific target identification could be achieved through adjusting primer concentration to modulate the FFI value.[19] However, extensive optimization is required and the number of targets is limited due to the variation of FFI values. In an effort to reduce the need for lengthy optimization, a new method called amplification curve analysis (ACA) was recently proposed, to extract target-specific kinetic information from real-time amplification data using supervised machine learning.[21] However, for the ACA approach, there is currently no systematic method of shaping the amplification curve and this presents a challenge for high-level multiplexing. Alternatively, some dPCR instruments offer the capability of melting curve analysis (MCA), providing a post-PCR method to identify specific targets with established literature and tools to assist assay design.[22] Similar to ACA, high-level multiplexing with MCA also requires complex assay design to distinguish between close melting curve peaks.[21]

Although the ACA and MCA methods are analyzing the same amplification product, they take advantage of different information to distinguish between targets. The amplification curve encodes target-specific kinetic information (i.e. complex reaction efficiency from cycle-to-cycle) while the melting curve is the result of thermodynamic properties of the amplicon

(e.g. GC content and length). Recently, it was shown that kinetic and thermodynamic parameters can be combined to detect non-specific amplification product in real-time digital loop-mediated isothermal amplification (LAMP).[23] However, there has been no report of enhancing multiplexing capabilities by combining the amplification and melting curves.

In this paper, we explore this concept using a commercially available dPCR platform (Fluidigm's BioMark HD) with an intercalating dye (EvaGreen) to demonstrate that non-mutual information from amplification and melting curves can improve multiplexing accuracy. The proposed method, referred to as amplification and melting curve analysis (AMCA), leverages the large volume of data from real-time dPCR and trains a "three-step" machine learning system, as depicted in Figure 1. The first step trains a model on the entire real-time amplification data and the seconds step trains a model using melting curve information. The final step combines the resulting outputs into a final classification for each amplification event.

As a case study, this work applies the AMCA method to the global challenge of antimicrobial resistance.[24] In particular, colistin is a "last-line" antibiotic, reserved for the treatment of severe bacterial infections. The rise of mobilized colistin resistance (*mcr*) has been reported in over 40 countries across five different continents.[25–27] Colistin resistant genes are often co-localized on highly transmissible plasmids with carbapenemase genes and are readily shared between bacterial species, providing the ideal conditions for multi-drug resistant organisms, and raising the possibility of untreatable infections.[28,29] Incorrect diagnosis delays appropriate intervention, increases financial burdens for the healthcare system and complicates antimicrobial stewardship efforts.[30] Therefore, detecting variants of *mcr* is important to help treat and understand this emerging antimicrobial resistance. In this study, we develop the first 9-plex PCR assay to detect *mcr*-1 to *mcr*-9.

Our vision is that by sharing this new method, researchers and practitioners can use affordable multiplex assays, compatible with dPCR platforms, for their clinically relevant applications. Moreover, extending this methodology to conventional qPCR instruments will be beneficial for the wider scientific community.

EXPERIMENTAL SECTION

DNA Templates

Double-stranded synthetic DNA (gBlock Gene fragments) containing the entire coding sequences of *mcr*-1 to *mcr*-9 were used. The accession numbers from the NCBI GenBank web site for each target are shown in Table 1. The gBlocks were purchased from Life Technologies (ThermoFisher Scientific) and re-suspended in Tris-EDTA buffer to 10 ng/µL stock solutions (stored at −80 °C until further use). The concentrations of all DNA stock solutions were determined using a Qubit 3.0 fluorimeter (Life Technologies).

Multiplex Primer Design

To perform the (*in-silico*) design for the 9-plex, the first step was to conduct an NCBI blast (`https://blast.ncbi.nlm.nih.gov`) to ensure that each primer set binds to a conserved region. For each target, the blast was able to retrieve an average of 1000 sequences, which have been used to identify variation in the nucleotide sequence for all possible inclusive targets within the same gene and exclude potential cross-reactivity sequences (either within the *mcr* family or from a different species). Alignments were performed using the MUSCLE algorithm,[31] in Geneious Prime® 2020.1.2.[32] Primer characteristics were analyzed through the IDT OligoAnalyzer software using the J. SantaLucia thermodynamic table for melting temperature (Tm) evaluation.[33] Moreover, to avoid secondary structure formation such as hairpin and primer-dimer (including self-dimer and cross-primer), the Multiple Primer Analyzer (ThermoFisher Scientific) was used.[34] The Tm of the amplification product of each primer set was determined by the Melting Curve Predictions Software (uMELT) package.[35] All primers were synthesized by Life Technologies (ThermoFisher Scientific). Primer sequences,

Table 1: Primer sequences and relevant meta data regarding the amplicon for all nine *mcr* targets.

Target (accession number)	**Forward primer** (5'→ 3')	**Reverse primer** (5' → 3')	**Amplicon length** (bp)	**Amplicon GC cont.** (%)
mcr-1 (KP347127.1)	TGGCGTTCAGCAGTCATTATGC	CAAATTGCGCTTTTGGCAGCTTA	516	50.0
mcr-2 (LT598652.1)	CTGTATCGGATAACTTAGGCTTT	ATACTGACTGCTAAATAGTCCAA	407	47.9
mcr-3 (KY924928.1)	AGACACCAATCCATTTACCAGTAA	GCGATTATCATCAAACTCCTTTCT	136	47.1
mcr-4 (MF543359.1)	TTGCAGACGCCCATGGAATA	GCCGCATGAGCTAGTATCGT	207	45.4
mcr-5 (KY807921.1)	GGTTGAGCGGCTATGAAC	GAATGTTGACGTCACTACGG	207	56.0
mcr-6 (MF176240.1)	GTCCGGTCAATCCCTATCTGT	ATCACGGGATTGACATAGCTAC	556	46.9
mcr-7 (MG267386.1)	TGCTCAAGCCCTTCTTTTCGT	TTGGCGACGACTTTGGCATC	466	56.2
mcr-8 (NG_061399.1)	CGAAACCGCCAGAGCACAGAATT	TCCCGGAATAACGTTGCAACAGTT	617	42.9
mcr-9 (NG_064792.1)	TATAAAGGCATTGCTTACCGTT	GGAAAGGCACTTTAGTCGTAAA	202	45.0

All primers have been fully developed in-house and published for the first time in this study.

amplicon length and GC content of the product are listed in Table 1.

PCR Reaction Conditions

Real-time Digital PCR.

Each amplification reaction was performed in 4 µL of final volume with 2 µL of 2× SsoFast EvaGreen Supermix with Low ROX (BioRad, UK), 0.4 µL of 20× GE Sample Loading Reagent (Fluidigm PN 85000746), 0.4 µL of 10× multiplex PCR primer mixture containing the nine primer sets (5 µM of each primer), and 1.2 µL of different concentrations of synthetic DNA (or controls). PCR amplifications consisted of a hot start step for 10 min at 95 °C, followed by 45 cycles at 95 °C for 20s, 66 °C for 45s, and 72 °C for 30s. Melting curve analysis was performed with one cycle at 65 °C for 3s and reading from 65 to 97 °C with an increment of 0.5 °C. We used the integrated fluidic circuit controller to prime and load qdPCR 37K™ digital chips and Fluidigm's Biomark HD system to perform the dPCR experiments, following manufacturer's instructions. More specifically, each digital chip contains 48 inlets, where each inlet is connected a panel consisting of 770 wells (0.85nL well volume).[36] In this study, we used 3 digital chips, totalling 144 panels (110880 wells), with experiments equally distributed across all *mcr* variants and negative controls. The number of positive reactions for each *mcr* variant is as follows: *mcr-1* (N=6767), *mcr-2* (N=6889), *mcr-3* (N=6159), *mcr-4* (N=6520), *mcr-5* (N=6424), *mcr-6* (N=6447), *mcr-7* (N=5919), *mcr-8* (N=6884) and *mcr-9* (N=6589).

Real-time PCR.

Each amplification reaction was performed in 10 µL of final volume with 5 µL of 2× SsoFast EvaGreen Supermix with Low ROX (BioRad, UK), 3 µL of PCR grade water, 1 µL of 10× multiplex PCR primer mixture containing the nine primer sets (5 µM of each primer), and 1 µL of different concentrations of synthetic DNA (or controls). The reaction consisted of 10 min at 95 °C, followed by 45 cycles at 95 °C for 20s, 66 °C for 45s, and 72 °C for 30s. Melting curve analysis was performed with one cycle at 65 °C for 60s, and reading from 65 to 97 °C with an increment of 0.2 °C. The PCR machine used in this study was the Light Cycler 96 Real-Time PCR System (Roche Diagnostics, Germany).

Data Analysis

Multiplexing based on FFI.

Final fluorescent intensity values were extracted from each amplification curve (as in [19]) and used to train a logistic regression classifier to distinguish targets. It is important to stress that the primer mix concentration was not optimized to improve classification, therefore we do not expect high performance.

Amplification Curve Analysis,

or ACA, consists of training a supervised machine learning model to distinguish targets based on the entire real-time amplification curve.[21] In this study, a deep neural network was chosen based on cross-validation score. In particular, the neural architecture consists of

two convolutional layers in order to extract temporal dynamics of the curve whilst keeping training times low (compared to recurrent architectures such as long short-term memory or gated recurrent unit networks). The first layer consists of 16 filters (kernel size of 5) and the second layer has 8 filters (kernel size of 3), where both layers have a rectified linear unit activation function. Prior to training the model, amplification curves were pre-processed using background subtraction (removing the mean of the first 5 fluorescent measurements) and subsequently calling positive/negative curves based on an arbitrary threshold.

Melting Curve Analysis,

or MCA, consists of distinguishing the thermodynamic profile (i.e. $-\frac{dF}{dT}$) of the amplification product. In this study, and conventionally, this is achieved by distinguishing the melting peak, Tm, although methods have also been proposed to consider the entire curve.[37,38] After peak detection, negative reactions can be confirmed by identifying curves with no peak. Subsequently, a supervised machine learning model can be trained to distinguish the Tm values. In this study, logistic regression was chosen as a classifier based on cross-validation.

The Proposed Method,

amplification and melting curve analysis, or AMCA, trains a supervised machine learning model to combine the predictions of ACA and MCA. This process is visualized in Figure 2. The output of ACA and MCA are probabilities for the amplification event belonging to each target of interest. In the training process, these probabilities are concatenated and used to train a model. In this study, a logistic regression classifier was chosen. It is important to note that this classifier is tuned with its own cross-validation step in order to avoid over-fitting.

Statistical Analysis

Performance of the models were evaluated based on out-of-sample classification accuracy, as determined by 10-fold cross-validation (using stratified splits). In order to assess the performance as a function of the volume of training data, a shuffled stratified split was performed 5 times, with 5000 test samples. The two-sided t-test with unknown variances was used to determine statistical significance for comparing the classification accuracy of different models. Prior to this test, a Lilliefors test was used to determine normality of the distributions and the Bartlett test for equal/unequal variances. A p-value of 0.05 was used as a threshold for statistical significance for all tests.

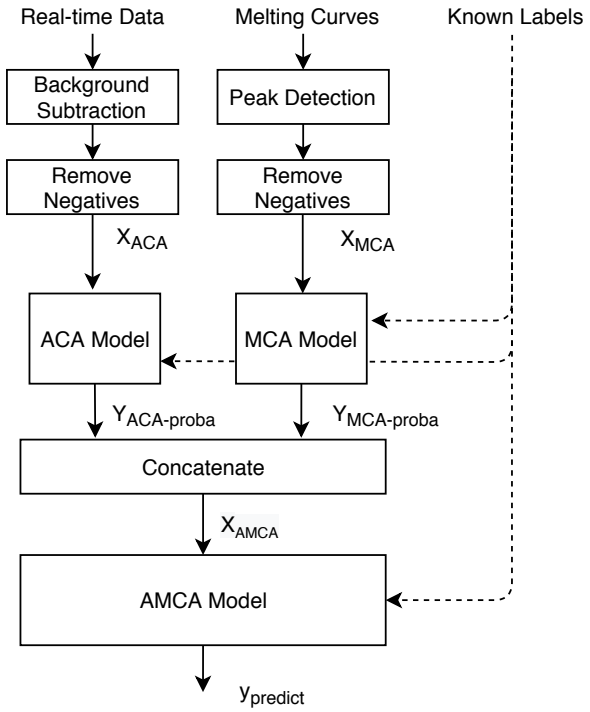


Figure 2: Flowchart to visualize the data processing workflow for the proposed method. Known labels (marked with a dashed line) are only required for training the models, as opposed to testing unknown samples. The input to the machine learning models are denoted as X_{ACA}, X_{MCA} and X_{AMCA}. The output probabilities of ACA and MCA are denoted as Y_{ACA}, Y_{MCA}. The final classification is given by $\mathrm{y}_{\mathrm{predict}}$. Vectors are given as lower case letters while matrices are upper case.

Data & Code availability

All data and code used in this study can be found at `https://github.com/am5113/pyAMCA`.

RESULTS

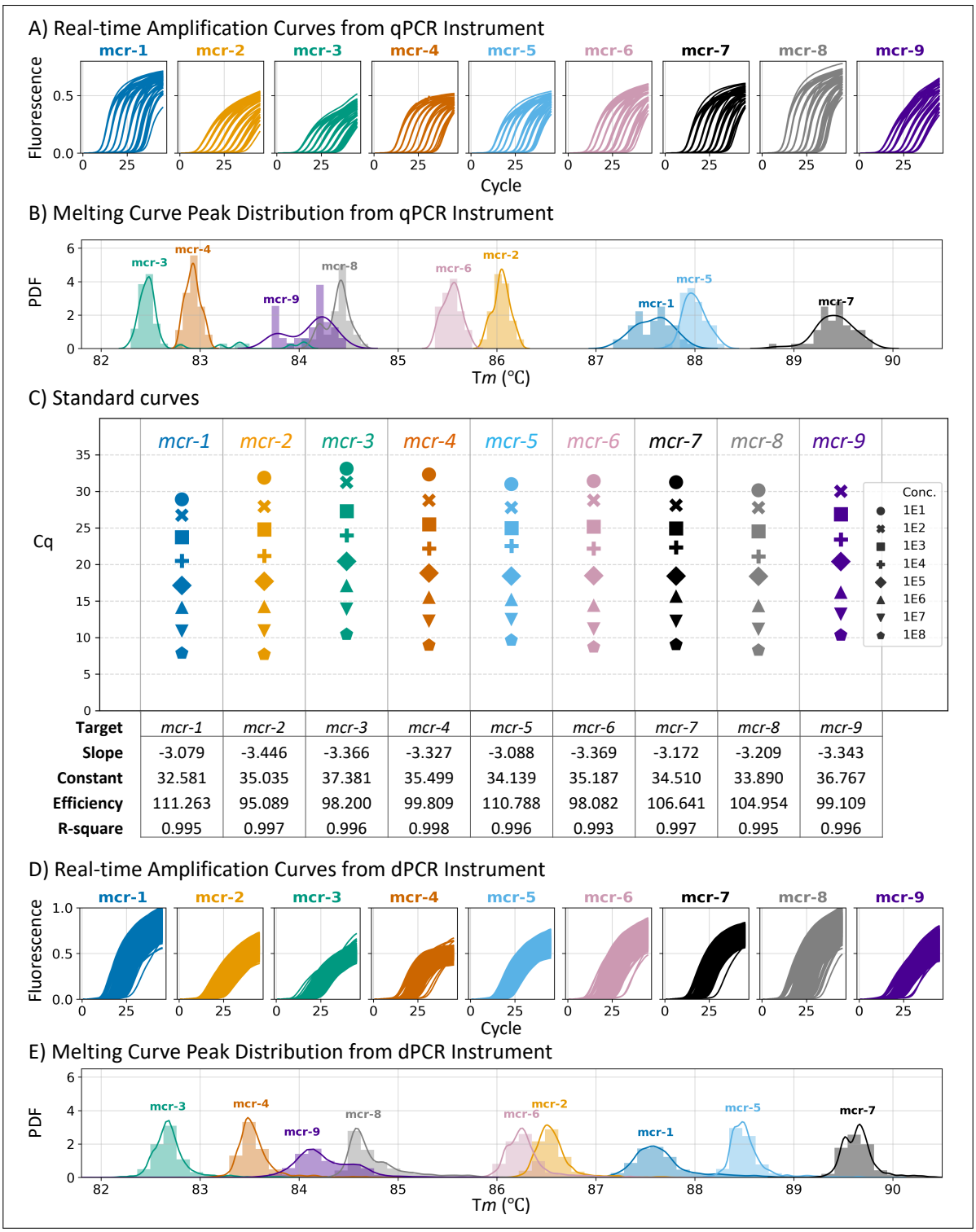
A new multiplex assay for mobilized colistin resistance which is highly sensitive and efficient

To date, there has been no report of multiplexing *mcr*-1 to *mcr*-9. Here, a new 9-plex has been designed and validated using a conventional qPCR platform. Figure 3 (A)-(C) show the real-time amplification curves, melting peak distributions (extracted from melting curves) and standard curves for a serial dilution of each *mcr* target. Figure S1 and S2 show the raw melting curves before peak extraction and conventional standard curves, respectively. From Figure 3 (A), it can be observed that the final fluorescence and shape can vary between targets, although the precise overlap cannot be visualized. On the other hand, as in Figure 3 (B), the melting peak distributions have distinct mean Tm values, although some targets (e.g. *mcr*-1 and *mcr*-5) have overlapping distributions, compromising MCA multiplexing classification. Figure 3 (C) demonstrates that the multiplex assay is highly efficient (all $> 95\%$) with a lower limit of detection (LoD) down to 10 copies per reaction for all targets (excluding *mcr*-9 which showed an LoD of 100 copies per reaction). All negative controls did not amplify before 45 cycles. The data suggests that the copresence of *mcr* variants, by virtue of the overlapping Tm distributions, raise the possibility of a single melting peak with multiple amplification products - leading to unavoidable misclassification using MCA. This motivates the use of digital PCR due to physical (single-molecule) partitioning.

Classification accuracy of FFI, ACA and MCA in dPCR is limited

To assess the performance of previously reported methods for dPCR multiplexing, 110,880 amplification reactions were analyzed, of which 58,598 are considered positive. To train the ACA model to be invariant to template concentration, experiments included concentrations ranging from single-molecule (digital pattern) to bulk reactions (saturated panels). Figure 3 (D) and (E) show the amplification and Tm distributions resulting from the dPCR platform, respectively. It is interesting to observe that the amplification curves and melting peak distributions resemble the qPCR data (within 0.8 °C), highlighting the consistency and reproducibility of the PCR chemistry and multiplex assay across platforms. The discrepancy between the distributions from qPCR to dPCR can be explained by the change in instrument resolution (from 0.2 °C to 0.5 °C) and the volume of data. The reason for selecting a lower resolution in dPCR, was such that a manageable volume of data was extracted via the Fluidigm digital PCR analysis software.

Figure 4 (A) and (B) show the confusion matrices, comparing the true and predicted targets for ACA and MCA, and the overall classification performance is $82.31 \pm 1.47\%$ and $89.34 \pm 0.33\%$, respectively. Furthermore, a naive classification based on FFI gives an overall accuracy of $24.59 \pm 0.52\%$ (confusion matrix and FFI distributions are provided in Figure S3). As the results indicate, the FFI performance has low accuracy, although better than a random classifier (i.e. 11.1%), due to single-parameter usage, which contains little information specific to each target. Therefore, optimization for primer concentration must be performed to achieve acceptable classification accuracy, as in McDermott *et al.* (2013), although this is neither trivial nor guaranteed for a 9-plex.[19] On the other hand, analyzing the entire amplification curves (without normalizing for FFI) using a neural network boosts performance by 57.7%, extracting relevant kinetic information from each event. The third method, MCA, analyzed thermodynamic information encoded in the melting profiles, showing a further increase of 7.0% in classification accuracy. It is interesting to observe that there is no obvious misclassification of any target which is common in both ACA and MCA, suggesting that the two methods extract non-mutual information.



Target	*mcr-1*	*mcr-2*	*mcr-3*	*mcr-4*	*mcr-5*	*mcr-6*	*mcr-7*	*mcr-8*	*mcr-9*
Slope	-3.079	-3.446	-3.366	-3.327	-3.088	-3.369	-3.172	-3.209	-3.343
Constant	32.581	35.035	37.381	35.499	34.139	35.187	34.510	33.890	36.767
Efficiency	111.263	95.089	98.200	99.809	110.788	98.082	106.641	104.954	99.109
R-square	0.995	0.997	0.996	0.998	0.996	0.993	0.997	0.995	0.996

Figure 3: Analysis of real-time amplification and melting curves from qPCR and dPCR instruments. A) Real-time amplification curves from qPCR instrument. B) Melting curve peak distribution from qPCR instrument showing the probability density function (PDF) for each target. The mean ± std of *mcr*-1 to *mcr*-9 is $87.6 \pm 0.2°C$, $86.0 \pm 0.1°C$, $82.6 \pm 0.4°C$, $82.9 \pm 0.1°C$, $88.0 \pm 0.1°C$, $85.5 \pm 0.1°C$, $89.4 \pm 0.2°C$, $84.4 \pm 0.1°C$, $84.1 \pm 0.2°C$, respectively. C) Visualization and statistics of standard curves for a serial dilution of each target in qPCR using 9-plex assay. D) Real-time amplification curves from dPCR instrument. E) Melting curve peak distribution from dPCR instrument. The mean ± std of *mcr*-1 to *mcr*-9 is $87.7 \pm 0.3°C$, $86.6 \pm 0.2°C$, $82.7 \pm 0.2°C$, $83.6 \pm 0.2°C$, $88.5 \pm 0.2°C$, $86.3 \pm 0.2°C$, $89.7 \pm 0.2°C$, $84.8 \pm 0.3°C$, $84.3 \pm 0.3°C$, respectively. Raw melting curves are shown in Figure S1.

AMCA method increases classification accuracy compared to ACA or MCA individually

Figure 4 (C) shows the confusion matrix comparing the predicted classification from the AMCA method to the true labels. It can be observed that the accuracy is 99.33 ± 0.13% and that no target is misclassified more than 1.7%, showing a significant improvement from ACA or MCA individually (p-value « 0.01). Since the chosen supervised machine learning model for AMCA is linear, the coefficients can be investigated to understand how it weighs the predictions from ACA and MCA. More specifically, the output of AMCA is defined by:

$$\mathbf{y} = \hat{\mathbf{W}}_{\mathrm{ACA}}\,\mathbf{y}_{\mathrm{ACA}} + \hat{\mathbf{W}}_{\mathrm{MCA}}\,\mathbf{y}_{\mathrm{MCA}} \qquad (1)$$

Where $\mathbf{y}_{\mathrm{ACA}} \in \mathbb{R}^9$ and $\mathbf{y}_{\mathrm{MCA}} \in \mathbb{R}^9$ are the probability vectors outputted from the ACA and MCA models, $\hat{\mathbf{W}}_{\mathrm{ACA}} \in \mathbb{R}^{9\times9}$ and $\hat{\mathbf{W}}_{\mathrm{MCA}} \in \mathbb{R}^{9\times9}$ are the model coefficients, respectively. Figure 4 (D) and (E) show the ACA and MCA coefficients in the form of a heatmap, respectively. It is interesting to observe that AMCA weighs the prediction from ACA more heavily for targets which show poor classification in MCA, and vice-versa. For example, MCA misclassifies 1515 *mcr*-9 reactions as *mcr*-8, therefore the AMCA positively weighs the ACA prediction by 3.1 and negatively weights the MCA prediction by -2.1. Similarly, ACA misclassifies 1846 *mcr*-9 reactions as *mcr*-2 and the coefficients compensate for this phenomenon.

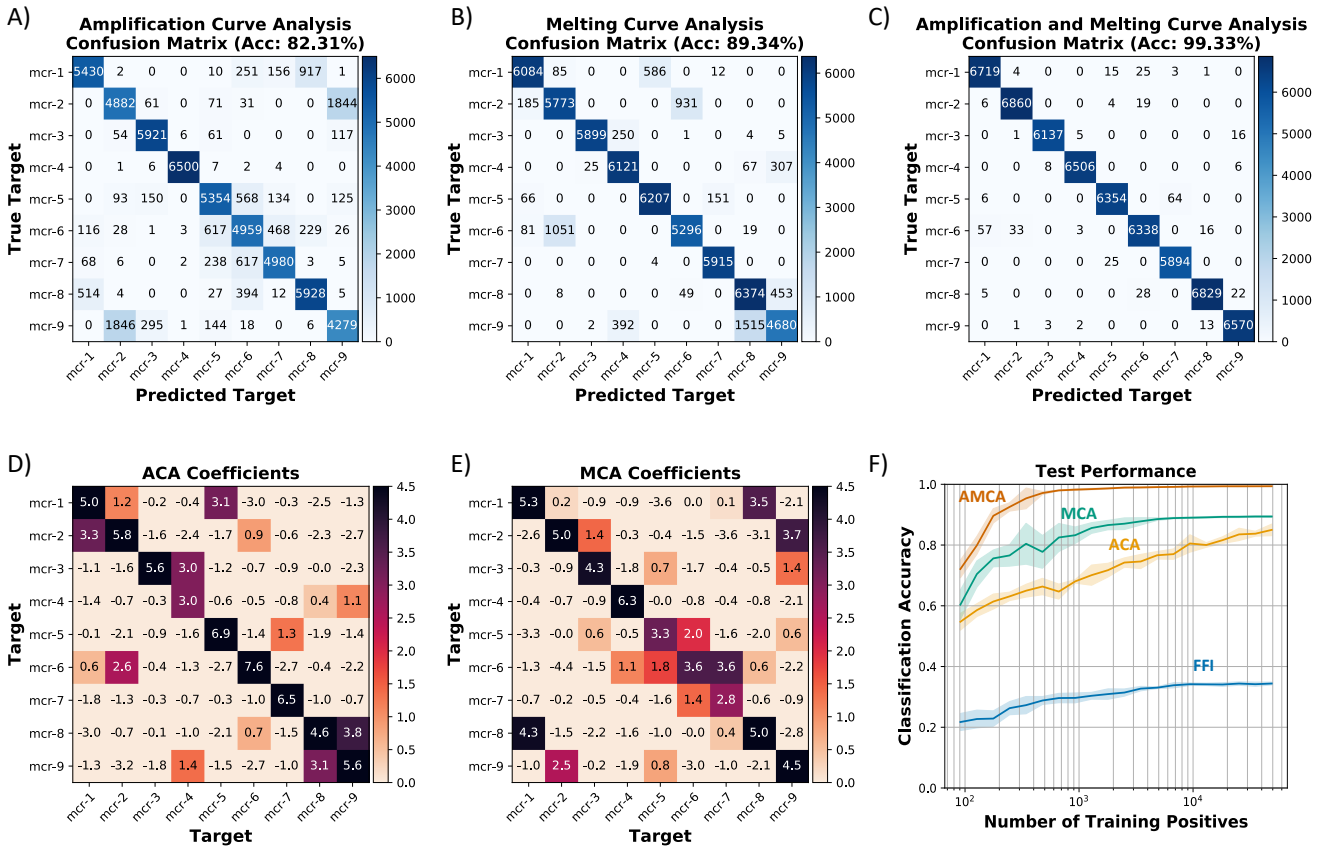


Figure 4: Performance of all methods for multiplexing the 9 *mcr* targets. A, B, C) Confusion matrices illustrating the predictions from ACA, MCA and AMCA (proposed method), respectively. Values indicate the number of amplification events with diagonal entries corresponding to correct predictions. D, E) Coefficients of the AMCA model weighting the predictions from the ACA and MCA methods, respectively. Darker colors indicate more positive weighting. F) The effect of the number of training data points on the overall classification accuracy for all methods. The shaded regions correspond to ± 1 standard deviation.

AMCA method reaches high accuracy with only 1000 training data points

From a practical perspective, it is important to understand the volume of training data required for the AMCA model, denoted by n_{train}, for accurate classification. Figure 4 (F) shows the classification performance on 5000 out-of-sample data points (repeated 10 times) where n_{train} is between 1.0×10^2 and 5.3×10^4 for all models. It can be observed that all of the models perform better given more training data points. Since AMCA weighs ACA and MCA, it is unlikely to perform worse than either of its constituents with sufficient data. In fact, the AMCA model consistently outperforms the other models for all training data sizes and repeats. Through observing the enhanced multiplexing accuracy, it can be concluded that the target-specific kinetic information (provided by ACA) and thermodynamic information (provided by MCA) is non-mutual.

AMCA method shows promising classification accuracy in conventional real-time PCR platform

The same methodology (as in Figure 2) was applied to the qPCR data presented in Figure 3 (A) and (B). The classification accuracy for ACA, MCA and AMCA was shown to be $84.40 \pm 6.7\%$, $82.74 \pm 5.5\%$ and $95.98 \pm 3.4\%$, respectively. The confusion matrices for each method and the model coefficients for AMCA are provided in Figure S4. These results suggest that the AMCA method works across real-time platforms, both quantitative and digital, although a further study to fully characterize the reliability in qPCR instruments (outside the scope of this manuscript) is required.

DISCUSSION

The AMCA method was shown to enhance the capability of high-level multiplexing in real-time digital PCR platforms, increasing the classification accuracy by combining kinetic information (through ACA) and thermodynamic information (through MCA). Currently, most instrument that have melting curve capabilities also integrate a real-time system for extracting amplification curves, which allows this method to be widely applicable to many labs. Furthermore, this method shows that even a non-ideal multiplex based on ACA or MCA may in fact contain sufficient information when combined together to perform accurate multiplexing, reducing the need for further time and resource consuming optimization .

On the other hand, the AMCA method requires training a supervised machine learning model which raises its own challenges. Firstly, since 3 models are required to be trained, especially if a neural network is used, this may take time and expertise in data science to perform. However, computational resources have negligible cost given the wide variety of open-source tools available for machine learning (such as *tensorflow* and *scikit-learn*). Secondly, it is important to ensure reproducibility of the experiment from a chemistry perspective in order for the training and testing data to be consistent. More specifically, if the instrument or laboratory approach show variability between experiments, then this needs to be accounted for from a data perspective (e.g. more data, pre-processing or data augmentation) or experimental procedures (i.e. consistent processes in the lab). However, since it was shown in this study that only 1000 amplification curves were required to achieve accurate multiplexing, it is possible to run training data within an experiment to avoid inter-experiment variations. For example, the Fluidigm qdPCR 37K digital chip contains 48 sample inlets (each connected to a panel of 770 wells), of which 9 panels can be used to generate the training data, one for each target. Assuming a digital occupancy of 80%, 9 panels translates to 5544 training data points, which based on Figure 4 (F), is expected to give an accuracy of 99.1%. From a practical point of view, this means that a single digital chip could accommodate screening 39 samples against 9 targets, whereas conventional spatial multiplexing (with single-plex assays) would only manage to screen 5 samples against the 9 targets.

As reported in a previous study, the ACA performance is degraded as a result of a phenomenon called 'co-amplification', which refers to the co-presence of multiple targets in a single chamber in dPCR instruments. This problem can be solved by keeping the occupancy of the digital panel (using Poisson statistics) within acceptable bounds in order to simultaneously reduce co-amplification and retain sufficient quantification precision. For example, for *mcr* genes, the vast majority of studies report the presence of a single *mcr* variant, and only few studies have reported the co-presence of 2 *mcr* variants in the same sample.[39] Therefore, as in Moniri *et al.* (2020), considering the co-presence of 2 targets and under the constraint of 36960 chambers (Fluidigm® 37K chip), the quantification uncertainty is below 5% between 16.7% and 99.3% digital occupancy.[21] Currently, there is no method of identifying co-amplification events in qPCR platforms using only the real-time amplification profile. However, melting curves can be used to circumvent this issue, although MCA is also limited when two melting peaks are close, e.g. within 1.0 °C. Recent studies show that using the entire melting profile using machine learning methods can be beneficial for classification purposes.[37,38]

This study raises several interesting questions for future directions: (i) Is it possible to identify co-amplification events in a single well using the entire melting curve profile? (ii) What is the highest number of targets AMCA can accurately and reliably multiplex? (iii) Does this method translate to other amplification chemistries including isothermal methods?

CONCLUSION

In conclusion, we propose a new method for high-multiplexing in real-time digital PCR instruments with melting curve capabilities. This approach is based on training supervised machine learning algorithms to extract kinetic and thermodynamic information together, to enhance the classification accuracy in multiplexing. We successfully show a 99.3% accuracy for identifying 9 clinically relevant targets, namely mobilized colistin resistance, using a new multiplex assay based on an affordable intercalating dye. Observing that the AMCA classification accuracy is better than solely analyzing amplification or melting curves demonstrates that the underlying biological factors driving these methods for target identification are fundamentally different. This biological insight is seen in the parameters of the machine learning model, which characterize the contribution of ACA and MCA across all targets to optimize the final classification of each amplification event. The implications of this method motivate further research in maximizing the value of nucleic acid amplification data, by uniquely merging molecular biology and data science. Extending this to conventional qPCR instruments and isothermal chemistries will be extremely beneficial for the wider scientific community.

ACKNOWLEDGMENTS

This work was supported by: NIHR Imperial Biomedical Research Centre [P80763]; the Imperial College's Centre for Antimicrobial Optimization (CAMO) and EPSRC DTP (EP/N509486/1 to A.M.). Please note that authors (J.R-M and A.H.) are affiliated with the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Imperial College London in partnership with Public Health England (PHE) in collaboration with, Imperial Healthcare Partners, University of Cambridge and University of Warwick. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or Public Health England. Professor Alison Holmes is a National Institute for Health Research (NIHR) Senior Investigator.

SUPPORTING INFORMATION

Raw melting curves from qPCR and dPCR instruments, standard curves for 9 *mcr* targets, multiplexing with final fluorescent intensity, and performance of method in qPCR.

Conflict of interest statement.

None declared.

References

(1) Higuchi, R.; Fockler, C.; Dollinger, G.; Watson, R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Nat. Biotechnol.* **1993**, *11*, 1026–1030.

(2) Livak, K. J.; Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2- $\Delta\Delta$CT method. *Methods* **2001**, *25*, 402–408.

(3) Gingeras, T. R.; Higuchi, R.; Kricka, L. J.; Lo, Y. D.; Wittwer, C. T. Fifty years of molecular (DNA/RNA) diagnostics. *Clin. Chem.* **2005**, *51*, 661–671.

(4) Moniri, A.; Rodriguez-Manzano, J.; Malpartida-Cardenas, K.; Yu, L.-S.; Didelot, X.; Holmes, A.; Georgiou, P. Framework for dna quantification and outlier detection using multidimensional standard curves. *Anal. Chem.* **2019**, *91*, 7426–7434.

(5) Bouzid, A.; Smeti, I.; Chakroun, A.; Loukil, S.; Gibriel, A. A.; Grati, M.; Ghorbel, A.; Masmoudi, S. CDH23 Methylation status and presbycusis risk in elderly women. *Front. Aging Neurosci.* **2018**, *10*, 241.

(6) El-Maraghy, S. A.; Adel, O.; Zayed, N.; Yosry, A.; El-Nahaas, S. M.; Gibriel, A. A. Circulatory miRNA-484, 524, 615 and 628 expression profiling in HCV mediated HCC among Egyptian patients; implications for diagnosis and staging of hepatic cirrhosis and fibrosis. *J. Adv. Res.* **2020**, *22*, 57–66.

(7) Bustin, S. A.; Benes, V.; Garson, J. A.; Hellemans, J.; Huggett, J.; Kubista, M.; Mueller, R.; Nolan, T.; Pfaffl, M. W.; Shipley, G. L., et al. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. 2009.

(8) Whale, A. S.; Huggett, J. F.; Cowen, S.; Speirs, V.; Shaw, J.; Ellison, S.; Foy, C. A.; Scott, D. J. Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. *Nucleic Acids Res.* **2012**, *40*, e82–e82.

(9) Quan, P.-L.; Sauzade, M.; Brouzes, E. dPCR: a technology review. *Sensors* **2018**, *18*, 1271.

(10) Sreejith, K. R.; Ooi, C. H.; Jin, J.; Dao, D. V.; Nguyen, N.-T. Digital polymerase chain reaction technology–recent advances and future perspectives. *Lab Chip* **2018**, *18*, 3717–3732.

(11) Witters, D.; Sun, B.; Begolo, S.; Rodriguez-Manzano, J.; Robles, W.; Ismagilov, R. F. Digital biology and chemistry. *Lab Chip* **2014**, *14*, 3225–3232.

(12) Kuypers, J.; Jerome, K. R. Applications of digital PCR for clinical microbiology. *J. Clin. Microbiol.* **2017**, *55*, 1621–1628.

(13) Tong, Y.; Shen, S.; Jiang, H.; Chen, Z. Application of digital PCR in detecting human diseases associated gene mutation. *Cell. Physiol. Biochem.* **2017**, *43*, 1718–1730.

(14) Dueck, M. E.; Lin, R.; Zayac, A.; Gallagher, S.; Chao, A. K.; Jiang, L.; Datwani, S. S.; Hung, P.; Stieglitz, E. Precision cancer monitoring using a novel, fully integrated, microfluidic array partitioning

digital PCR platform. *Sci. Rep.* **2019**, *9*, 1–9.

(15) Zangenberg, G.; Saiki, R.; Reynolds, R. *PCR Applications*; Elsevier, 1999; pp 73–94.

(16) Rodriguez-Manzano, J.; Moniri, A.; Malpartida-Cardenas, K.; Dronavalli, J.; Davies, F.; Holmes, A.; Georgiou, P. Simultaneous single-channel multiplexing and quantification of carbapenem-resistant genes using multidimensional standard curves. *Anal. Chem.* **2019**, *91*, 2013–2020.

(17) Wittwer, C. T.; Herrmann, M. G.; Gundry, C. N.; Elenitoba-Johnson, K. S. Real-time multiplex PCR assays. *Methods* **2001**, *25*, 430–442.

(18) Whale, A. S.; Huggett, J. F.; Tzonev, S. Fundamentals of multiplexing with digital PCR. *Biomol. Detect. Quantif.* **2016**, *10*, 15–23.

(19) McDermott, G. P.; Do, D.; Litterst, C. M.; Maar, D.; Hindson, C. M.; Steenblock, E. R.; Legler, T. C.; Jouvenot, Y.; Marrs, S. H.; Bemis, A., et al. Multiplexed target detection using DNA-binding dye chemistry in droplet digital PCR. *Anal. Chem.* **2013**, *85*, 11619–11627.

(20) Fraley, S. I.; Hardick, J.; Jo Masek, B.; Athamanolap, P.; Rothman, R. E.; Gaydos, C. A.; Carroll, K. C.; Wakefield, T.; Wang, T.-H.; Yang, S. Universal digital high-resolution melt: a novel approach to broad-based profiling of heterogeneous biological samples. *Nucleic Acids Res.* **2013**, *41*, e175–e175.

(21) Moniri, A.; Miglietta, L.; Malpartida-Cardenas, K.; Pennisi, I.; Cacho-Soblechero, M.; Moser, N.; Holmes, A.; Georgiou, P.; Rodriguez-Manzano, J. Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR. *Anal. Chem.* **2020**,

(22) Ririe, K. M.; Rasmussen, R. P.; Wittwer, C. T. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Anal. Biochem.* **1997**, *245*, 154–160.

(23) Rolando, J. C.; Jue, E.; Barlow, J. T.; Ismagilov, R. F. Real-time kinetics and high-resolution melt curves in single-molecule digital LAMP to differentiate and study specific and non-specific amplification. *Nucleic Acids Res.* **2020**, *48*, e42–e42.

(24) Organization, W. H., et al. *Antimicrobial resistance: global report on surveillance*; World Health Organization, 2014.

(25) Lima, T.; Domingues, S.; Da Silva, G. J. Plasmid-mediated colistin resistance in Salmonella enterica: a review. *Microorganisms* **2019**, *7*, 55.

(26) Carroll, L. M.; Gaballa, A.; Guldimann, C.; Sullivan, G.; Henderson, L. O.; Wiedmann, M. Identification of Novel Mobilized Colistin Resistance Gene mcr-9 in a Multidrug-Resistant, Colistin-Susceptible Salmonella enterica Serotype Typhimurium Isolate. *mBio* **2019**, *10*, e00853–19.

(27) Rodriguez-Manzano, J.; Moser, N.; Malpartida-Cardenas, K.; Moniri, A.; Fisarova, L.; Pennisi, I.; Boonyasiri, A.; Jauneikaite, E.; Abdolrasouli, A.; Otter, J. A., et al. Rapid Detection of Mobilized colistin Resistance using a nucleic Acid Based Lab-on-a-chip Diagnostic System. *Sci. Rep.* **2020**, *10*, 1–9.

(28) Otter, J. A.; Doumith, M.; Davies, F.; Mookerjee, S.; Dyakova, E.; Gilchrist, M.; Brannigan, E. T.; Bamford, K.; Galletly, T.; Donaldson, H., et al. Emergence and clonal spread of colistin resistance due to multiple mutational mechanisms in carbapenemase-producing Klebsiella pneumoniae in London. *Sci. Rep.* **2017**, *7*, 1–8.

(29) Manohar, P.; Shanthini, T.; Ayyanar, R.; Bozdogan, B.; Wilson, A.; Tamhankar, A. J.; Nachimuthu, R.; Lopes, B. S. The distribution of carbapenem-and colistin-resistance in Gram-negative bacteria from the Tamil Nadu region in India. *J. Med. Microbiol.* **2017**, *66*, 874–883.

(30) Otter, J.; Burgess, P.; Davies, F.; Mookerjee, S.; Singleton, J.; Gilchrist, M.; Parsons, D.; Brannigan, E.; Robotham, J.; Holmes, A. Counting the cost of an outbreak of carbapenemase-producing Enterobacteriaceae: an economic evaluation from a hospital perspective. *Clin. Microbiol. Infect.* **2017**, *23*, 188–196.

(31) Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.

(32) Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C., et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649.

(33) SantaLucia Jr, J.; Hicks, D. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 415–440.

(34) ThermoFisher Scientific, Multiple Primer Analyzer. `https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html`, 2020.

(35) Dwight, Z.; Palais, R.; Wittwer, C. T. uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics* **2011**, *27*, 1019–1020.

(36) Fluidigm, Digital PCR with the qdPCR 37K IFC Using Gene-Specific Assays. `https://www.fluidigm.com/binaries/content/documents/fluidigm/resources/qdpcr-37k-dpcr-qr-100%E2%80%906896/qdpcr-37k-dpcr-qr-100%E2%80%906896/fluidigm%3Afile`, 2014.

(37) Athamanolap, P.; Parekh, V.; Fraley, S. I.; Agarwal, V.; Shin, D. J.; Jacobs, M. A.; Wang, T.-H.; Yang, S. Trainable high resolution melt curve machine learning classifier for large-scale reliable genotyping of sequence variants. *PLoS One* **2014**, *9*, e109094.

(38) Velez, D. O.; Mack, H.; Jupe, J.; Hawker, S.; Kulkarni, N.; Hedayatnia, B.; Zhang, Y.; Lawrence, S.; Fraley, S. I. Massively parallel digital high resolution melt for rapid and absolutely quantitative sequence profiling. *Sci. Rep.* **2017**, *7*, 1–14.

(39) García, V.; García-Meniño, I.; Mora, A.; Flament-Simon, S. C.; Díaz-Jiménez, D.; Blanco, J. E.; Alonso, M. P.; Blanco, J. Co-occurrence of mcr-1, mcr-4 and mcr-5 genes in multidrug-resistant ST10 Enterotoxigenic and Shiga toxin-producing Escherichia coli in Spain (2006-2017). *Int. J. Antimicrob. Agents* **2018**, *52*, 104–108.

Graphical TOC Entry