Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR

SCHOLARONE™
Manuscripts

Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR

Ahmad Moniri,[†,§] Luca Miglietta,[†,§] Kenny Malpartida-Cardenas,[†] Ivana Pennisi,[†,‡]
Miguel Cacho-Soblechero,[†] Nicolas Moser,[†] Alison Holmes,[¶]
Pantelis Georgiou,[†] and Jesus Rodriguez-Manzano*[,†,¶,§]

[†]*Centre for Bio-Inspired Technology, Department of Electrical and Electronic Engineering,
Imperial College London, UK*
[‡]*Section of Paediatric Infectious Disease, Department of Infectious Disease, 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

Information about the kinetics of PCR reactions are encoded in the amplification curve. However, in digital PCR (dPCR), this information is typically neglected by collapsing each amplification curve into a binary output (positive/negative). Here, we demonstrate that the large volume of raw data obtained from real-time dPCR instruments can be exploited to perform data-driven multiplexing in a single fluorescent channel using machine learning methods, by virtue of the information in the amplification curve. This new approach, referred to as *amplification curve analysis* (ACA), was shown using an intercalating dye (EvaGreen), reducing the cost and complexity of the assay and enabling the use of melting curve analysis for validation. As a case study, we multiplexed 3 carbapenem-resistant genes to show the impact of this approach on global challenges such as antimicrobial resistance. In the presence of single targets, we report a classification accuracy of 99.1% ($N = 16188$) which represents a 19.7% increase compared to multiplexing based on the final fluorescent intensity. Considering all combinations of amplification events (including co-amplifications), the accuracy was shown to be 92.9% ($N = 10383$). To support the analysis, we derived a formula to estimate the occurrence of co-amplification in dPCR based on multivariate Poisson statistics, and suggest reducing the digital occupancy in the case of multiple targets in the same digital panel. The ACA approach takes a step towards maximizing the capabilities of existing real-time dPCR instruments and chemistries, by extracting more information from data to enable data-driven multiplexing with high accuracy. Furthermore, we expect that combining this method with existing probe-based assays will increase multiplexing capabilities significantly. We envision that once emerging point-of-care technologies can reliably capture real-time data from isothermal chemistries, the ACA method will facilitate the implementation of dPCR outside of the lab.

Introduction

Digital PCR (dPCR) is a well-established method to detect and quantify nucleic acids.[1,2] It is based on the amplification of single target DNA/RNA molecules in many separate reaction wells. This approach offers several advantages over conventional real-time PCR (qPCR), such as: (1) lack of references or standards; (2) high precision in quantification; (3) tolerance to inhibitors; and (4) the capability to analyze complex mixtures.[3–5] Therefore, dPCR has enabled scientific breakthroughs in cancer biomarker discovery, genetic alterations and infectious diseases, among others.[6–8]

As the need for high throughput analysis of multiple targets continues to escalate, several approaches have been proposed to simultaneously detect and quantify multiple nucleic acids. Microfluidic solutions offer spatial multiplexing by subdividing the reaction into an increasing number of partitions.[9] Alternatively, multiplex dPCR assays can rely on the use of fluorescent probes (e.g. TaqMan) using multiple channels. The probe concentration can be optimized to increase the number of targets per channel by changing the final fluorescent intensity (FFI).[10] However, probes are expensive and require time-consuming optimization.[11] To achieve similar multiplexing capabilities, dye-based approaches (e.g. EvaGreen) have also been proposed, which alter primer concentration, changing the PCR efficiency and impacting the FFI.[12] Furthermore, instruments with melting capabilities, such as Fluidigm's BioMark HD, can be used to perform melting curve analysis (MCA) when using intercalating dyes. The melting peak, Tm, is extracted to distinguish targets,[13] although more recently methods based on machine learning have also been proposed.[14] To the best of our knowledge, there has been no report of multiplexing using the entire amplification curve in dPCR.

Recently, in qPCR it was shown that sufficient information exists within the amplification curve so as to distinguish several targets using multidimensional standard curves.[15,16] However, since the volume of data from qPCR is limited ($< 10^2$ reactions per experiment), explicit features of the amplification curve were extracted to perform reliable multiplexing in a single-channel. In this study, we combined real-time single-molecule digital PCR using an intercalating dye and machine learning models, to prove that sufficient kinetic information exists in the amplification curve to perform data-driven multiplexing - referred to as amplification curve analysis (ACA). We use MCA as the "gold standard" to assess the performance of the proposed method, as illustrated in the experimental workflow depicted in Figure 1. We take advantage of the volume of raw data extracted from real-time dPCR ($> 10^4$ reactions per experiment) and the high likelihood of single-molecule events in order to develop machine learning models, without explicitly extracting features of the amplification curve or compromising the assay performance (by modifying probe or primer concentration). Moreover, we normalize for the FFI, to show that this method can be combined with current approaches for dPCR multiplexing - breaking the barrier of 1 target for each level of FFI (in a given fluorescent channel). We also provide a theoretical derivation for the likelihood of multiple targets in a single well (i.e. co-amplification) in order to understand the effect of this phenomenon on quantification and multiplexing.

As a clinically relevant application, we apply this methodology to the global challenge of antimicrobial resistance.[17] In particular, we focused on carbapenemases, which are β-lactamases (bla) that are resistant to the carbapenems, a class of highly effective antibiotic agents.[18] Therefore, we developed a multiplex assay for the detection of 3 common carbapenem-resistant genes, namely bla_{NDM}, bla_{VIM} and bla_{KPC}.

Our vision for this work is three-fold: (1) maximize the capabilities of existing instruments and chemistries by extracting more information from existing data; (2) combine this approach with existing probe-based methods to increase multiplexing capabilities significantly; and (3) translate this methodology to isothermal chemistries and emerging point-of-care technologies to facilitate the implementation of dPCR outside of the lab.[19]
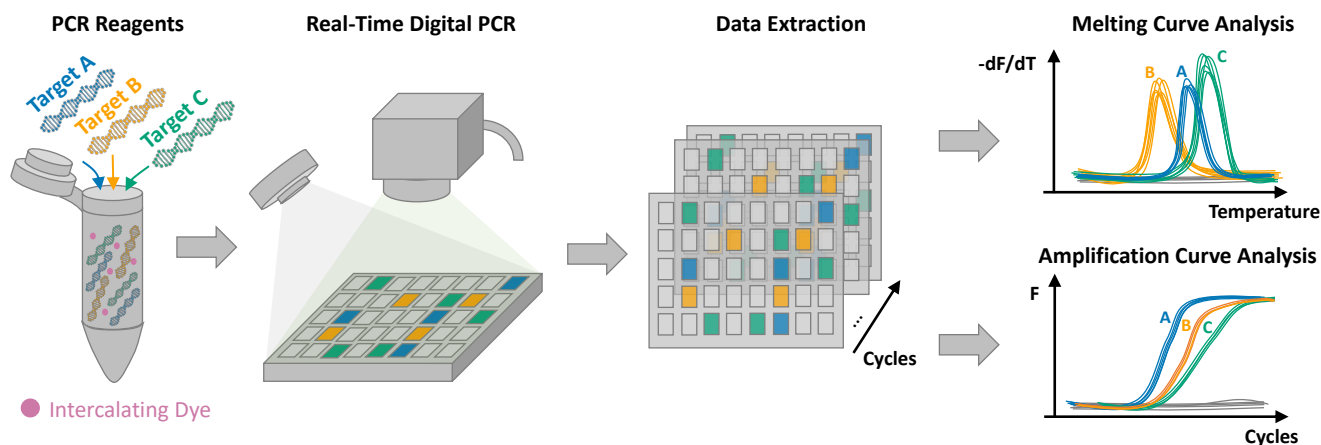
Figure 1: Experimental workflow. A multiplex PCR assay (with an intercalating dye) is developed for detecting targets A, B and C. Real-time digital PCR is used to perform single-molecule amplification to detect the targets. Melting curve analysis is used to validate the specificity of the amplification product. The output of real-time dPCR is a sequence of images, from which the time-series of the amplification and melting curves can be extracted. Subsequently, supervised machine learning using the amplification curves, referred to as *amplification curve analysis*, can be used to distinguish the targets, and melting curve analysis can be used to evaluate the performance.

Experimental Section

DNA Templates

We used double-stranded synthetic DNA (gBlock Gene fragments) containing bla_{NDM}, bla_{VIM} and bla_{KPC} gene sequences (ranging from 801 to 917 bp). The sequences of these genes were downloaded from the GenBank web site (http://www.ncbi.nlm.nih.gov/genbank/) with accession numbers of NC_023908, NC_023274 and NC_014312 for bla_{NDM}, bla_{VIM} and bla_{KPC}, respectively. These genes belong to the class B metalloenzymes encoding bla_{NDM} and bla_{VIM}, plus the class A carbapenemases encoding bla_{KPC} type. They were purchased from Life Technologies and resuspended in Tris-EDTA buffer to $10\,\text{ng/µL}$ stock solutions (stored at $-80\,°\text{C}$ until further use). The concentrations of all DNA stock solutions were determined using a Qubit 3.0 fluorimeter (Life Technologies).

PCR Primer Design

Primers for the multiplex assay were designed to target the aforementioned referenced sequences. For each gene of interest, 1000 sequences were retrieved from NCBI blast (*in-silico*), to identify all the possible inclusive targets and exclude potential cross-reactivity sequences. Alignments were performed using the MUSCLE algorithm,[20] in Geneious Prime® 2020.1.2 (https://www.geneious.com).[21] Primer characteristics were analyzed through the IDT OligoAnalyzer software using the J.SantaLucia thermodynamic table for melting temperature (Tm) evaluation, hairpin, self-dimer and cross-primer formation.[22] The Tm of the amplification product of each primer set was determined by the Melting Curve Predictions Software (uMELT) package (https://dna-utah.org/umelt/umelt.html).[23] All primers were synthesized by Life Technologies. Primer sequences are listed in Table 1.

PCR Reaction Conditions

Real-time PCR.

Each amplification reaction was performed in $10\,\text{µL}$ of final volume with $5\,\text{µL}$ of SsoFast EvaGreen Supermix with Low ROX (BioRad, UK), $3\,\text{µL}$ of PCR grade water, $1\,\text{µL}$ of $10\times$ multiplex PCR primer mixture containing the three primer sets ($5\,\text{µM}$ of each primer), and $1\,\text{µL}$ of different concentrations of synthetic DNA (or controls). PCR amplifications consisted of 10 min at $95\,°\text{C}$, followed by 45 cycles at $95\,°\text{C}$ for 20s, $65\,°\text{C}$ for 45s, and $72\,°\text{C}$ for 30s. In order to

Table 1: Primer Specification

Target	Primer Name	Sequence (5'–3')	Amplicon size (bp)
bla_{NDM}	NDM-F	CACACCAGTGACAATATCACCGTTG	85
	NDM-R	ACTTGGCCTTGCTGTCCTTGAT	
bla_{VIM}	VIM-F	CTTCGGTCCAGTAGAACTCT	258
	VIM-R	GTGTGCTTGAGCAAGTCT	
bla_{KPC}	KPC-F	TCGAACAGGACTTTGGCG	202
	KPC-R	GGAACCAGCGCATTTTTGC	

Primers have been developed in this study.

validate the proposed method, the results were compared against melting curve analysis. Melting curve analysis was performed with one cycle at 65 °C for 60s, and reading from 65 to 97 °C with an increment of 0.5 °C. The PCR machine used in this study was the Light Cycler 96 Real-Time PCR System (Roche Diagnostics).

Real-time Digital PCR.

Each amplification reaction was performed in 4 µL of final volume with 2 µL of SsoFast Eva-Green Supermix with Low ROX (BioRad, UK), 0.4 µL of 20× GE Sample Loading Reagent (Fluidigm PN 85000746), 0.3 µL of PCR grade water, 0.2 µL of 20× multiplex PCR primer mixture containing the three primer sets (10 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, 65 °C for 45s, and 72 °C for 30s. The results were validated using MCA which 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 (IFC) 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).[24] In this study, we used 3 digital chips, totalling 144 panels, including negative controls.

Data Analysis

Multiple in-house Python (v3.7) scripts were developed to extract and analyze the data using standard data science packages (e.g. NumPy, Pandas and Scikit-Learn). Complete details of the code can be found at `www.github.com/am5113/pyACA`. All graphics are made using the Matplotlib package and optimized for color blindness.[25] The logistic regression and k-nearest neighbour models were implemented using the scikit-learn package with default parameters (for more information please see provided code and package documentation). The classification accuracy (i.e. proportion of correctly identified events), sensitivity (i.e. true positive rate) and specificity (i.e. true negative rate) values in Table 1 and 2 were computed for each binary classification sub-problem in the one-vs-one multi-class classification scheme.

Results and Discussion

In this paper, it is shown, for the first time, that data-driven multiplexing can be achieved by ACA at the single-molecule level using intercalating dyes, by only considering the amplification curve. The following section is structured as follows. First, the challenges of qPCR multiplexing in the presence of multiple targets are illustrated, which motivate the use of dPCR. Second, the limitation of dPCR multiplexing based on final fluorescent intensity is demonstrated, highlighting the need to extract more information from the amplification curve for high-level multiplexing. Subsequently, this kinetic information is visualized in the *entire*

amplification curve using unsupervised machine learning. This enables the use of supervised machine learning to perform data-driven multiplexing - called amplification curve analysis. Therefore, the performance of ACA in the presence of single and multiple targets is assessed, and the impact of co-amplification in dPCR using multivariate Poisson statistics is explored.

Challenges of qPCR multiplexing in the presence of multiple targets in a single reaction

Performing multiplexing in a single fluorescent channel using intercalating dyes presents a major challenge since the measured fluorescence is proportional to *all* double-stranded DNA produced in the reaction. To this end, several methods analyze the amplification product through approaches such as melting curve analysis and gel electrophoresis in order to distinguish the targets from each other (and from non-specific products). In general, the presence of multiple targets in the same reaction is either neglected because it is a rare event or it is solved through lengthy and expensive optimization to reliably distinguish the amplification products.[10,26]

First, we developed a 3-plex assay for the detection of bla_{NDM}, bla_{VIM} and bla_{KPC}. Figure 2 (A) shows the amplification curves and melting peaks for each target at concentrations ranging from 5×10^3 to 1×10^6 copies/reaction. Observe the melting peaks for bla_{NDM}, bla_{VIM} and bla_{KPC} can be distinguished from each other and are given as $84.7^\circ C$, $88.5^\circ C$ and $89.7^\circ C$ respectively. Moreover, Figure 2 (B) shows the corresponding standard curves illustrating the Cq value as a function of the target concentration, yielding an assay efficiency of 80.5%, 88.6% and 92.2% for targets bla_{NDM}, bla_{VIM} and bla_{KPC}, respectively.

Typically, a single value, i.e. Tm, is used to identify the specificity of the melting peak. However, information is also contained in the width of the melting peaks (due to GC content and amplicon length).[27] In the co-presence of multiple targets in a single reaction, the width is important since it defines the ability to resolve two peaks. For example, Figure 2 (C) shows the amplification curves for the co-presence of targets and Figure 2 (D) shows the corresponding melting curves. It can be observed that the $bla_{\mathrm{NDM}}+bla_{\mathrm{KPC}}$ and $bla_{\mathrm{NDM}}+bla_{\mathrm{VIM}}$ peaks are sufficiently different in order to identify 2 distinct peaks in the melting profile. However, the mixture containing $bla_{\mathrm{VIM}}+bla_{\mathrm{KPC}}$ results in only a single peak. This is also observed in the mixture with all 3 targets as only 2 peaks are evident. This may suggest there are fewer amplification products. Through adding the pure melting profiles, we can estimate the 'expected' melting curve for mixtures of products, as in Figure 2 (E). Observe that for bla_{VIM} and bla_{KPC}, the expected melting curve only predicts a single peak. This demonstrates the uncertainty as to whether the single peak contains 1 or more products - representing one of the major challenges with using MCA for multiplexing in the presence of more than one target. Therefore, we are forced to run post PCR analysis techniques such as gel electrophoresis or sequencing. Figure 2 (F) shows the gel electrophoresis image for the same reactions as above. It can be observed that each reaction contains the same number of bands as the expected number of targets at the correct amplicon length (see Experimental Section). Although gel electrophoresis can resolve the multiple products, it is time-consuming, increases the risk of contamination and is impractical for many applications due to the protocol and components of the gel.[28]

Recently, it was shown that kinetic information in the amplification curve can be used to multiplex without the need for melting curve analysis or gel electrophoresis using multidimensional standard curves.[15,16] However, this work did not explore the presence of co-amplification and explicit features of the amplification curve were extracted due to the limited amount of data in qPCR.
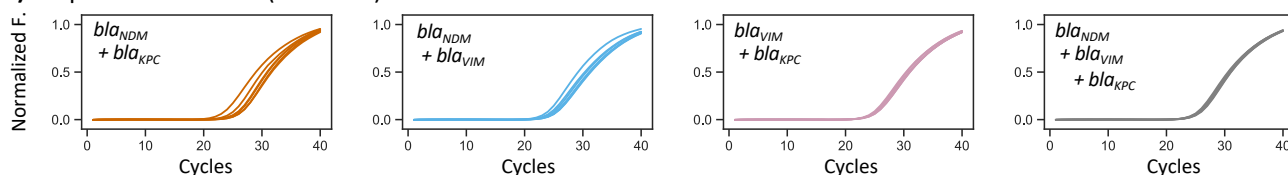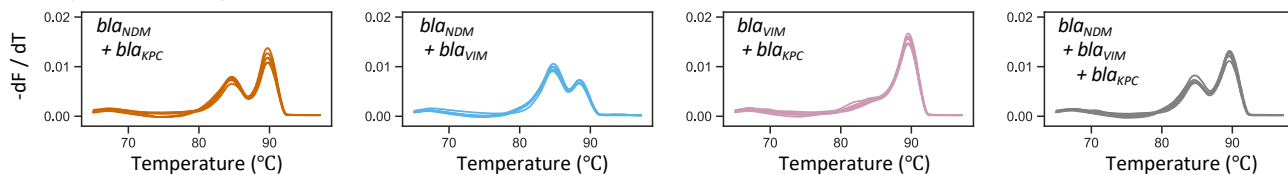
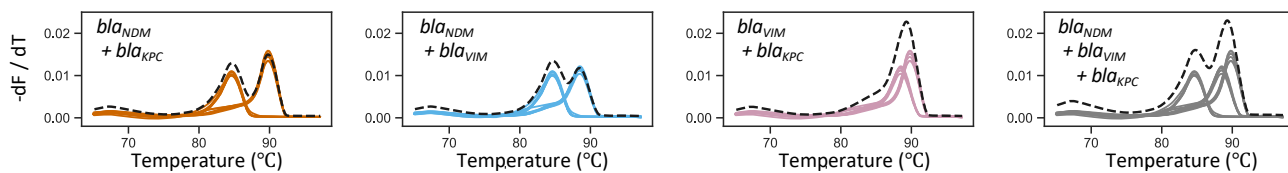**A)** Amplification and Melting Curves (Pure)
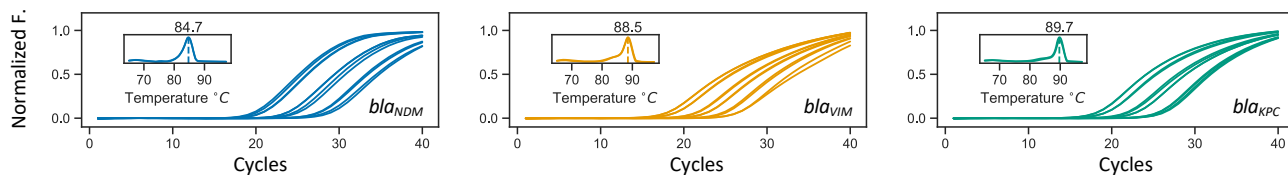
B) Standard Curves (Pure)
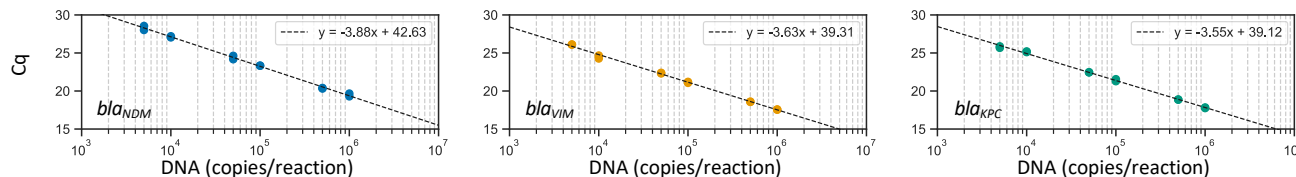
C) Amplification Curves (Mixtures)

D) Empirical Melting Curves (Mixtures)

E) Expected Melting Curves (Mixtures)

F) Gel Electrophoresis (Pure + Mixtures)

1 - bla_{VIM}
2 - bla_{NDM}
3 - bla_{KPC}
4 - bla_{NDM} + bla_{VIM}
5 - bla_{VIM} + bla_{KPC}
6 - bla_{NDM} + bla_{KPC}
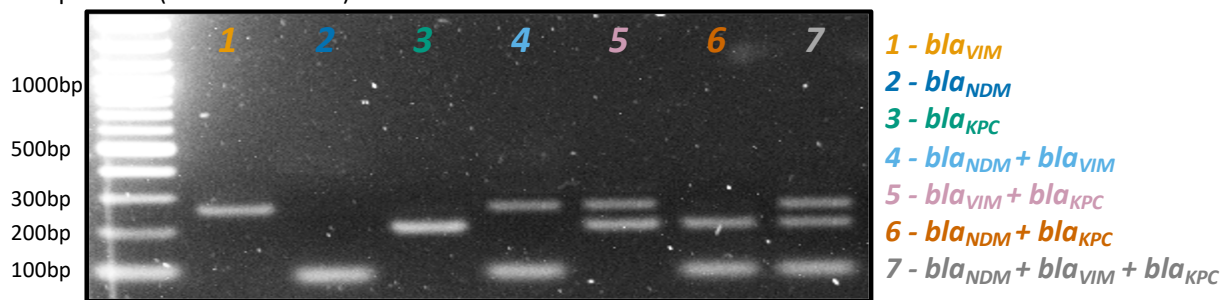7 - bla_{NDM} + bla_{VIM} + bla_{KPC}

Figure 2: Real-time PCR Experiments showing the performance of a 3-plex assay in the presence of single and multiple targets. (A) Amplification curves for single targets (in a single reaction) with corresponding melting curves, where six different dilutions were used ranging from 5×10^3 to 1×10^6 copies/reaction. (B) Corresponding standard curves correlating the Cq values with the concentration of each target. (C) Amplification curves for the co-presence of targets and (D) respective empirical melting curves. (E) Prediction of melting curves for co-presence of targets: solid lines indicate single target meltings; dashed line is an estimation of the expected melting curve for mixture of products. (F) Gel electrophoresis image of each reaction type.

Real-time dPCR Multiplexing based on Melting Curve Analysis

The aforementioned limitations motivate the use of real-time dPCR as a method of multiplexing for two main reasons: (1) the vast number of partitions reduce the likelihood of co-amplification in a single reaction significantly; and (2) the large volume of data enables the use of advanced machine learning algorithms to detect subtle kinetic differences encoded in the amplification curves.

Here, we translate the 3-plex assay from qPCR to real-time dPCR. First, we investigate the multiplex assay in the presence of pure targets in each digital panel. Figure 3 (A) and (B) show the digital pattern and ampli-

fication curves for a serial dilution of the targets. Concentrations ranging from 5×10^3 to 1×10^6 copies/reaction were chosen such that we observe amplification events in both single-molecule and bulk regions in order to capture kinetic information in both domains. In total, there were 36960 amplification events with 16188 positive reactions: bla_{NDM} ($N = 4589$), bla_{VIM} ($N = 5682$) and bla_{KPC} ($N = 5917$). It is interesting to observe the Cq values as a function of the target concentration as seen in Figure 3 (C) since there is a clear separation between the single-molecule and the bulk regions. In the bulk region, the panels are saturated and therefore the target can be quantified using a standard curve (as in qPCR), whereas the low concentrations form a digital pattern
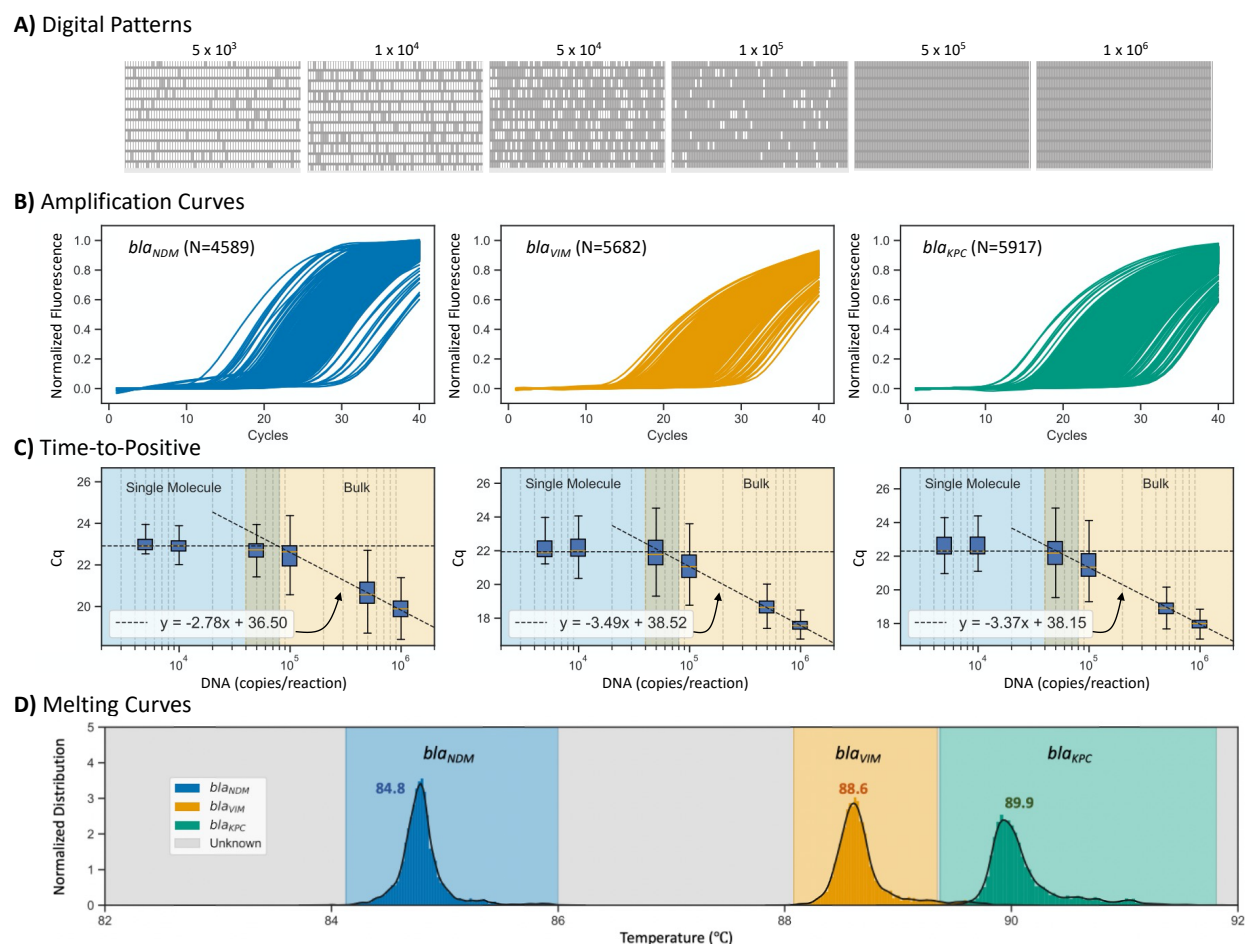

Figure 3: Real-time dPCR data. (A) Digital patterns for each panel at increasing concentrations. (B) Amplification curves for serial dilution of each target at concentrations ranging from 5×10^3 to 1×10^6 copies/reaction. (C) Standard curves correlating the Cq values with the concentration of each target; shaded blue area indicates single-molecule region; shaded orange shows the bulk region; and the middle area displays the transition between single-molecule and bulk. (D) Normalized distribution of the melting curve peaks, i.e. Tm, for each target. All data presented in this figure were conducted in a single digital experiment.

that can be quantified using Binomial and Poisson statistics.[3] Moreover, it is observed that the assay efficiency in digital PCR is 129.0%, 93.4% and 98.2% for target bla_{NDM}, bla_{VIM} and bla_{KPC} respectively. This is a 48.5%, 4.8% and 6.0% increase compared to qPCR, which is expected due to several factors such as: less inhibition and high local concentration.[3–5] Figure 3 (D) shows the distribution of the melting curve peaks (Tm) for each target. The maximum likelihood value of Tm for bla_{NDM}, bla_{VIM} and bla_{KPC} is $84.8°C$, $88.6°C$ and $89.9°C$, respectively. All Tm values are within $0.2°C$ of their respective qPCR quantities. The width of the distributions are related to the resolution of the measurements. To obtain a manageable volume of data from the dPCR platform, a resolution of $0.5°C$ was used for the melting curve analysis. Based on this, we can determine the bounds for which to distinguish the targets by considering the 1st and 99th percentile. The lower and upper bounds for bla_{NDM}, bla_{VIM} and bla_{KPC} were computed as $(84.1°C, 86.0°C)$, $(88.1°C, 89.3°C)$ and $(89.4°C, 91.8°C)$ respectively.

Real-time dPCR Multiplexing using Final Fluorescent Intensity

In the literature, the current method of multiplexing with intercalating dyes in dPCR is based on differentiating the final fluorescent intensity (FFI) between the targets.[26] Figure 4 (A) shows the raw amplification curves with background subtraction. The associated FFI for each amplification event is shown in Figure 4 (B). It can be observed that there is an overlap between the distributions of FFI for the 3 targets. Based on these values, a machine learning model can be trained to learn the optimal boundaries to distinguish the targets. The dashed red lines, T1 and T2, show the thresholds learned using a Logistic Regression classifier. Based on this classifier, the overall classification accuracy is computed as 79.42%, which is not adequate for many applications. This value is computer on the entire dataset (N=16188) based on 10-fold cross-validation. More specifically, we train the model on 14569 data points and test on the remaining 1619.

This process is repeated 10 times to provide the confusion matrix in Figure 4 (B) and details of the one-vs-one classifiers (which are combined to form the multi-class model) in Table 2. It can be observed that accurate multiplexing can be achieved for bla_{NDM} vs bla_{KPC} or bla_{NDM} vs bla_{VIM}, however the bla_{VIM} and bla_{KPC} are not separable, compromising the entire 3-plex. This demonstrates the need for optimizing the reaction conditions (e.g. primer concentration) in order to obtain reasonable accuracy for 3 or more targets, due to the large variation of FFI values.

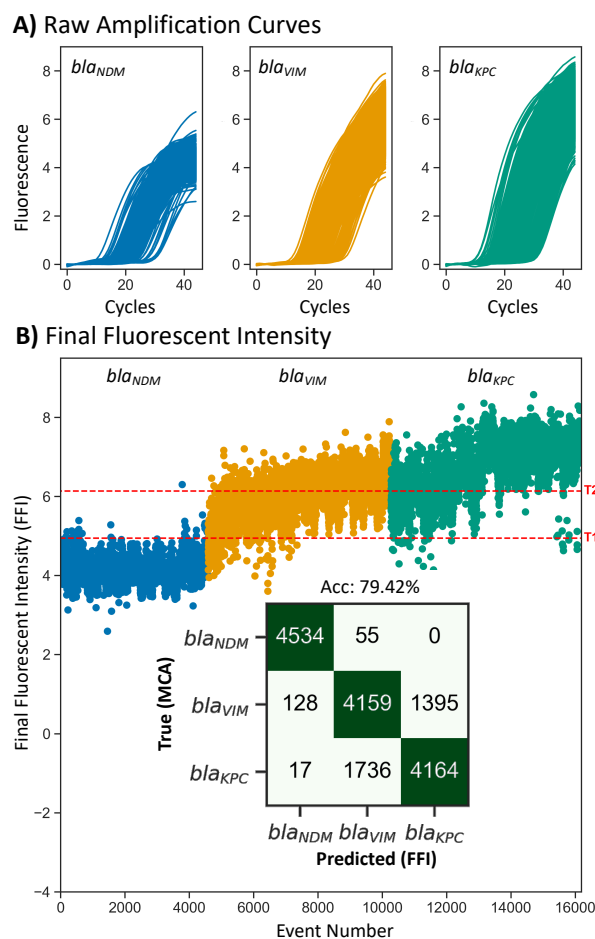Figure 4: Multiplexing based on final fluorescent intensity. (A) Raw amplification curves with background subtraction. (B) Distribution of FFI values across the 3 targets. The red dashed lines (T1 and T2) indicate the thresholds generated from a Logistic Regression method for target classification. The predictions are shown in the overlay confusion matrix. Amplification events are ordered from low to high concentration.

Table 2: Final Fluorescent Intensity Classification Performance

Classifier	Acc.	Sens.	Spec.
bla_{NDM} vs bla_{VIM}	98.2%	97.3%	99.0%
bla_{NDM} vs bla_{KPC}	99.5%	99.3%	99.6%
bla_{VIM} vs bla_{KPC}	72.9%	71.0%	74.9%

Acc = Accuracy, Sens = Sensitivity, Spec = Specificity.

Information in the Amplification Curve

The findings above suggest that more information than the FFI is needed. The MCA clearly encoded this information since it is able to distinguish the 3 targets in dPCR. However, for a single digital chip, the melting data required 1.7Gb extra memory (at just $0.5°C$ resolution) in addition to the 1.1Gb required for the amplification data, translating to more time for acquisition & processing. Moreover, the MCA method cannot be extended to chemistries which are not compatible such as TaqMan assays or non-fluorescence based sensing. Therefore, in this study, we aim to explore a new method of multiplexing through the use of machine learning to extract specific kinetic information directly from the amplifications curve. First, we use unsupervised machine learning in order to visualize the complex interaction from cycle to cycle, by embedding the high dimensional amplification curves (i.e. 40 cycles) into a visualizable low dimensional space (e.g. 2 or 3). That is, amplification curves which are more similar are mapped to points which are close in lower-dimensional space. This can be achieved using the t-distributed stochastic neighbor embedding (t-SNE) algorithm, which has the ability to preserve local structure.[29] It is important to understand that t-SNE is an unsupervised learning algorithm and therefore does not use the target labels. Figure 5 illustrates the t-SNE algorithm (perplexity=500) applied to the amplification curves with each target colored for visualization purposes. It can be observed that the different targets fall in a different region of this embedding and can therefore be distinguished

automatically using statistical machine learning. Therefore, we demonstrate that even after normalizing for fluorescent intensity, the kinetic information which is encoded in the amplification curve can provide sufficient information to perform data-driven multiplexing. Moreover, it is interesting to observe that the region indicated within the dashed red circle shows amplification curves which do not fully plateau, and therefore are similar across the 3 targets. This suggests that the *entire* curve is necessary to extract sufficient kinetic information.
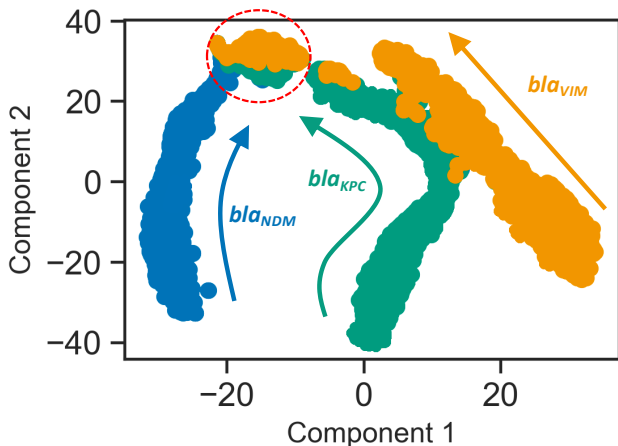

Figure 5: Visualizing the similarity between amplification curves using the t-distributed stochastic neighbor embedding algorithm with 2 components. Direction of arrows indicate high to low concentration. Dashed red circle indicates curves that have not reached plateau.

Amplification Curve Analysis: Data-Driven Multiplexing using Supervised Machine Learning

After establishing that information exists in the amplification curve using unsupervised methods, supervised learning methods can be used to exploit this information to perform multiplexing. Several machine learning algorithms exist for classification tasks such as k-nearest neighbors, support vector machines and deep neural networks.[30–32] The following section is demonstrated using the k-nearest neighbors algorithm which is a non-parametric method that is intuitive.[33] In order to assess the performance of this new form of data-driven multiplexing, referred to in this report as *amplification curve*

analysis, four questions were answered:

1. What is the performance of ACA in the presence of single targets?

2. How much data is required to perform accurate ACA multiplexing?

3. What is the performance of ACA in the presence of multiple targets?

4. What is the impact of co-amplification events on ACA?

Performance in the presence of a single target

Melting curve analysis can be used as the "gold standard" to evaluate the performance of ACA. The data presented in Figure 3 can be used to estimate the out-of-sample (or test) accuracy in the presence of a single target using 10-fold cross-validation. Figure 6 (A) shows the prediction accuracy in a confusion matrix for the k-NN algorithm (for $k = 10$). The dark green squares indicate the single-target true positive classifications. The overall classification accuracy was 99.09% (CI: 99.08-99.09%). Moreover, the accuracy, sensitivity and specificity for the one-vs-one classifiers are given in Table 3. This result demonstrates that all 3 targets can be accurately distinguished from each other. Moreover, these results show the high concordance between MCA and ACA, suggesting that the amplification curve contains more information than commonly presumed. Compared to the FFI method, the overall classification accuracy was increased from 79.4% to 99.1%, representing a 19.7% increase in performance. Once again, this highlights the need to optimize the assay for the FFI method, which can be challenging in many applications.

Table 3: ACA Classification Performance

Classifier	Acc.	Sens.	Spec.
bla_{NDM} vs bla_{VIM}	99.8%	99.9%	99.7%
bla_{NDM} vs bla_{KPC}	99.7%	99.9%	99.5%
bla_{VIM} vs bla_{KPC}	99.1%	99.1%	99.0%

Acc = Accuracy, Sens = Sensitivity, Spec = Specificity.

Volume of data required for accurate multiplexing

The volume of data required for training has significant practical implications in order to obtain high test performance whilst reducing the number of experiments. Figure 6 (B) shows the test accuracy for 1000 samples as a function of the number of training samples. This was computed through bootstrapping 100 times using a stratified shuffle split. As expected, as the number of training data increases, the out-of-sample performance increases. More interestingly, with as little as 100 training samples, the performance is at 95%, and increases to 98% before 1000 training samples.

Performance in the presence of multiple targets

Although theoretically with sufficient number of wells the challenges of multiple targets are mitigated, in reality the likelihood of co-amplification exists. Moreover, from a practical perspective, the training data is conducted in a different experiment to the test data, raising the possibility of inter-experiment variations. In this section, we use the previous data as the training samples and run a different experiment which contains all possible combinations of the targets.

Figure 6 (C) illustrates the number of positives for each panel class, as determined by MCA and ACA. The dashed red boxes illustrate the co-amplification events. In total, 228 co-amplification events were observed. Moreover, the shaded boxes indicate the events where we expect co-amplifications to occur, but MCA is not able to detect due to the merging melting peaks discussed previously. Furthermore, it can be observed that compared to MCA, some of the panels show misclassified reactions using ACA. More specifically, observing each amplification event at the single-molecule level independently, the overall predictions of ACA are described in the confusion matrix illustrated in Figure 6 (D). The overall classification accuracy (including the co-amplification events indicated by the dashed red line) is computed as 92.9%. However, only considering pure events yields an
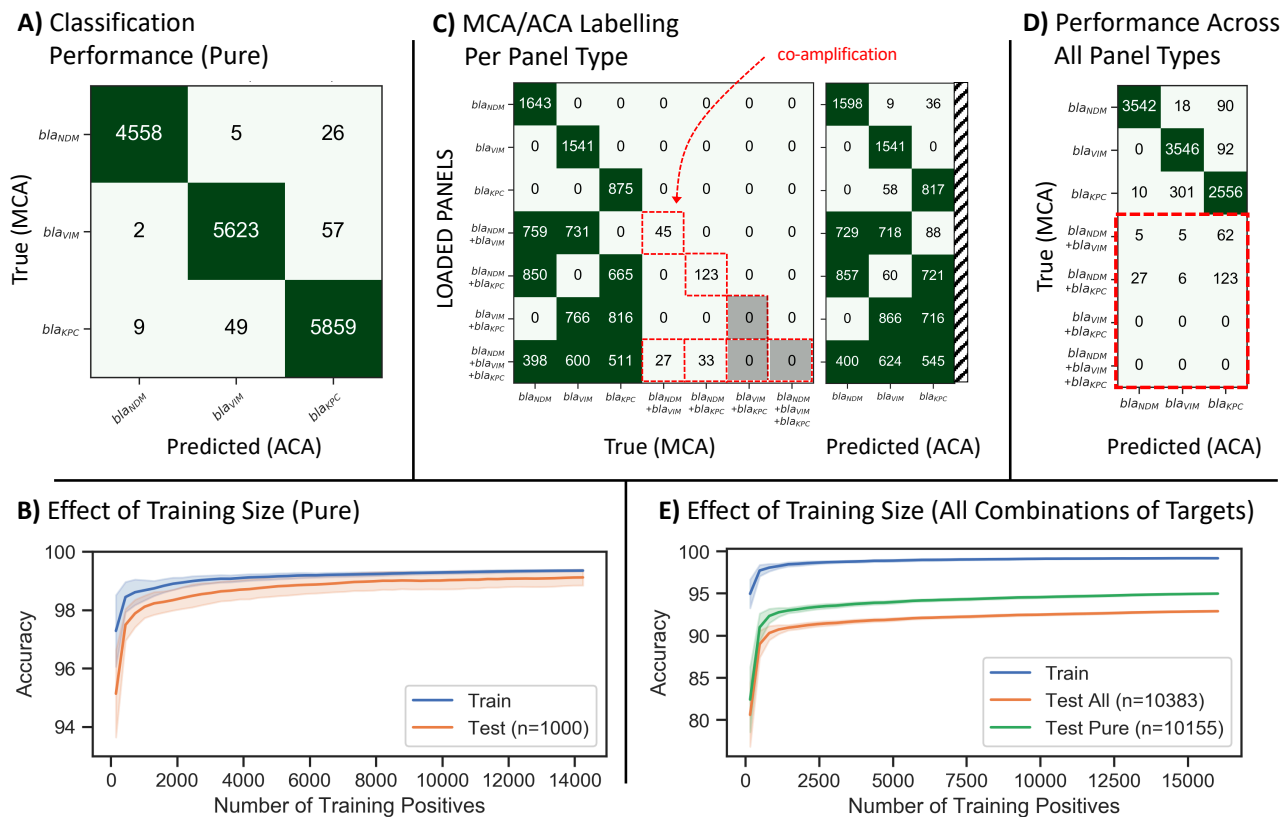
Figure 6: Performance of ACA in the presence of single and multiple targets. (A) Confusion matrix showing the predictions of ACA compared with MCA in the presence of single targets; (B) The effect of training size (with pure targets) on the train/test performance; (C) Matrices displaying the prediction of MCA and ACA per panel type; (D) Confusion matrix showing the predictions of ACA compared with MCA in the presence of multiple targets; (E) The effect of training size (with all combinations of targets) on the train/test performance. The data collected and analysed in this figure were conducted in 3 digital experiments on separate days. The confidence intervals were computed by bootstrapping 100 times using a stratified shuffle split and taking the standard deviation of the values. The analysis in A and B consisted of panels with varying digital occupancies, ranging from empty to saturated panels. All co-amplification experiments in C, D and E, showed digital patterns (with λ's ranging from 0.2 to 1.6).

accuracy of 95.0%. Figure 6 (E) displays the accuracy for both, pure and all, amplification events as a function of the volume of training data. It can be observed the accuracy plateaus within 1000 training samples. The error due to the co-amplification events can be mitigated further by increasing the number of wells (or equivalently decreasing the digital occupancy).

Understanding the impact of co-amplification events

Quantification in dPCR is performed based on Binomial & Poisson statistics in order to estimate the number of molecules taking into account the probability of double, triple, etc. events.[3] This analysis assumes that the DNA

molecules are independently and uniformly distributed across the digital array. The advantage of dPCR is that the accuracy of the quantification can be estimated using the confidence interval in the Poisson parameter estimation. Figure 7 (A) shows the quantification precision as a function of the occupancy (based on the Wilson confidence interval). It can be observed that the optimal occupancy across all m is approximately 80% or $\lambda = 1.6$ (marked with a cross). However, an acceptable range of digital occupancy can be defined according the desired accuracy for a given application. For example, under the constraint of $m = 36960$ (number of wells in a Fluidigm® 37K chip), the uncertainty is below 5% between 16.7% occupancy ($\lambda = 0.2$, marked with a circle) and 99.3% oc-

11

cupancy ($\lambda = 5.0$, marked with a square).

Here, we extend the Poisson statistics to derive a formula to estimate the theoretical number of wells with more than one target, i.e. wells that represent a challenge for ACA. The probability that k molecules fall within a well can be described by the Poisson distribution given by

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{1}$$

$$\lambda = \frac{n}{m} \tag{2}$$

Where n is the number of DNA molecules and m is the number of wells. Let $p(\mathbf{k})$ denote $p(k_1, \ldots, k_K)$, the joint probability distribution of having k_i molecules from target i in each well (where K is the number of targets). Under the independence assumption, the joint distribution can be given as

$$p(\mathbf{k}) = p(k_1) \ldots p(k_K) \tag{3}$$

$$= \prod_{i=1}^{K} p(k_i) \tag{4}$$

The proportion of co-amplification, denoted by P_C, is defined as having more than 1 target in a well. Or equivalently, it is defined as $1 - P_0 - P_1$ where P_0 is the probability of having no targets and P_1 is the probability of having a single or multiple molecules of the same target. Therefore, using equation (3)-(4), P_0 and P_1 are given as
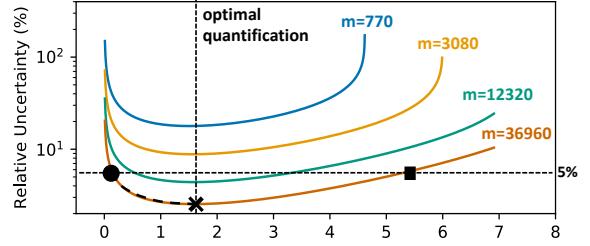
$$P_0 = \prod_{i=1}^{K} p(k_i = 0) \tag{5}$$

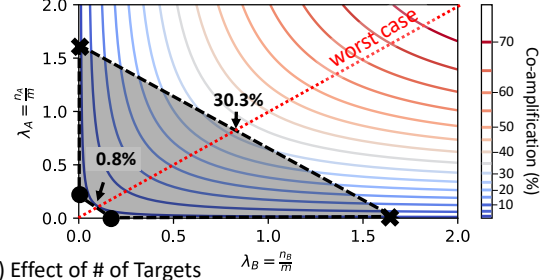$$P_1 = \sum_{j=1}^{K} \frac{p(k_j > 0)}{p(k_j = 0)} \prod_{q=1}^{K} p(k_q = 0) \tag{6}$$

Substituting equation (1) into the above and using the identity $p(k > 0) = 1 - p(k = 0)$ yields

$$P_C = 1 - \underbrace{\prod_{i=1}^{K} e^{-\lambda_i}}_{P_0} - \underbrace{\sum_{j=1}^{K} (e^{\lambda_j} - 1) \prod_{q=1}^{K} e^{-\lambda_q}}_{P_1} \tag{7}$$
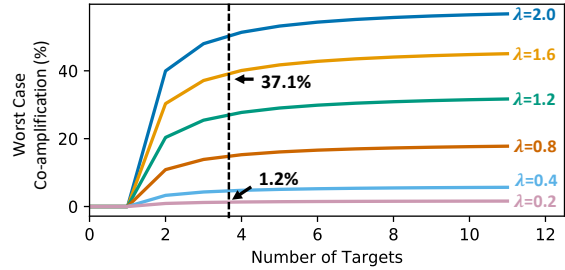

Figure 7: The impact of co-amplification events. (A) The relative uncertainty of Poisson quantification as a function of λ and the number of wells. (B) The probability of co-amplification in the presence of 2 targets. (C) The effect of the number of targets on the worst-case probability of co-amplification.

Which can be simplified to

$$P_C = 1 - e^{-\lambda} \left(1 + \sum_{j=1}^{K} (e^{\lambda_j} - 1)\right) \tag{8}$$

$$\text{where} \quad \lambda = \sum_{i=1}^{K} \lambda_i \tag{9}$$

Using this formula, we can estimate the theoretical error introduced by co-amplifications. In the ideal scenario, as the number of wells tends to infinity, i.e. $m \to \infty$, then $\lambda \to 0$, therefore $P_0 \to 1$ and $P_1 \to 0$, resulting in $P_C \to 0$. This demonstrates the error in ACA due to co-amplifications tend to zero as the number of wells increases. Figure 7 (B) shows the proportion of co-amplification events for two targets (A and B) as a function of λ_A and λ_B.

12

It can be observed that an increase in the total $\lambda = \lambda_A + \lambda_B$, causes an increase in the likelihood of co-amplification events. Moreover, the worst-case scenario is experienced when $\lambda_A = \lambda_B$ as marked with a dashed red line. The shaded region indicates the range of λ's between 0.4 and 1.6. Therefore, the intersection of the shaded region and the worst-case scenario shows 30.3% for $\lambda = 1.6$ and 0.8% for $\lambda = 0.2$. Figure 7 (C) shows the worst-case co-amplification proportion as a function of the number of targets. For 3 targets, we can decrease P_C from 37.1% down to 1.2% by reducing λ to 0.2. Moreover, the error starts to plateau (for all λ) above 6 targets. In fact, as the number of targets tends to infinity, P_C is equivalent to the probability of wells with more than 1 molecule independently of the number of targets. That is,

$$\lim_{K \to \infty} P_C = 1 - e^{-\lambda} - \lambda e^{-\lambda} \qquad (10)$$

Contrary to single target Poisson quantification, to maximize ACA multiplexing performance, λ should be decreased without compromising quantification significantly. However, in a realistic scenario, the sample concentration is not known beforehand. Therefore, this analysis is intended to be used after an experiment, such that the user can estimate how much error may be present in the results due to co-amplification - similar to the quantification precision (i.e. confidence interval) when performing Poisson statistics.

Conclusion

To the best of our knowledge, no published study has reported dPCR (or droplet dPCR) multiplexing by considering the kinetic information encoded in the *entire* amplification curve. By leveraging the large volume of single-molecule data in real-time dPCR, we report a new data-driven method using supervised machine learning, referred to as amplification curve analysis, or ACA. We validated this approach for 3 drug-resistant genes: bla_{NDM}, bla_{VIM} and bla_{KPC}, by comparing to melting curve analysis as the "ground truth". Although

MCA is not ideal due to merging of peaks, it remains the only post PCR method to validate dPCR amplification products.

The results show that in the presence of single targets, the accuracy of ACA is 99.1% when training and testing within a digital experiment. This represents an improvement of 19.7% compared to the conventional method of multiplexing based on the final fluorescent intensity. Furthermore, when training and testing across experiments, we observed an accuracy of 95.0%. However, this promising performance was reduced to 92.9% due to the presence of co-amplification in a single well. To support the analysis, we derived a formula to estimate the occurrence of co-amplification and suggest reducing the digital occupancy in the case of multiple targets in the same digital panel. This work raises several questions to for future studies: (1) How robust is the ACA method to variations in reaction conditions? (2) Can we extend ACA to perform higher-level multiplexing by considering the final fluorescent intensity? (3) Can we translate this method to probe-based or pH-based methods? (4) Can we classify co-amplification and non-specific products?

This paper showed the high concordance between MCA and ACA, suggesting that the amplification curve contains more information than commonly presumed. We expect that combining ACA with existing probe-based methods will increase multiplexing capabilities significantly. Moreover, translating this methodology to isothermal chemistries and point-of-care technologies, such as ISFET arrays and SlipChip, will facilitate the implementation of real-time dPCR outside of the lab.[34–40]

Acknowledgments

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.

References

(1) Vogelstein, B.; Kinzler, K. W. Digital pcr. *Proceedings of the National Academy of Sciences* **1999**, *96*, 9236–9241.

(2) Morley, A. A. Digital PCR: a brief history. *Biomolecular detection and quantification* **2014**, *1*, 1–2.

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

(4) 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 on a chip* **2018**, *18*, 3717–3732.

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

(6) Perkins, G.; Lu, H.; Garlan, F.; Taly, V. *Advances in clinical chemistry*; Elsevier, 2017; Vol. 79; pp 43–91.

(7) Current and emerging applications of droplet digital PCR in oncology, author=Olmedillas-López, Susana and García-Arranz, Mariano and García-Olmo, Damián. *Molecular diagnosis & therapy* **2017**, *21*, 493–510.

(8) Tong, Y.; Shen, S.; Jiang, H.; Chen, Z. Application of digital PCR in detecting human diseases associated gene mutation. *Cellular Physiology and Biochemistry* **2017**, *43*, 1718–1730.

(9) Shen, F.; Sun, B.; Kreutz, J. E.; Davydova, E. K.; Du, W.; Reddy, P. L.; Joseph, L. J.; Ismagilov, R. F. Multiplexed quantification of nucleic acids with large dynamic range using multivolume digital RT-PCR on a rotational SlipChip tested with HIV and hepatitis C viral load. *Journal of the American Chemical Society* **2011**, *133*, 17705–17712.

(10) Rajagopal, A.; Yurk, D.; Shin, C.; Menge, K.; Jacky, L.; Fraser, S.; Tombrello, T. A.; Tsongalis, G. J. Significant expansion of real-time PCR multiplexing with traditional chemistries using amplitude modulation. *Scientific reports* **2019**, *9*, 1–8.

(11) Tajadini, M.; Panjehpour, M.; Javanmard, S. H. Comparison of SYBR Green and TaqMan methods in quantitative real-time polymerase chain reaction analysis of four adenosine receptor subtypes. *Advanced biomedical research* **2014**, *3*.

(12) Ruijter, J.; Ramakers, C.; Hoogaars, W.; Karlen, Y.; Bakker, O.; Van den Hoff, M.; Moorman, A. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic acids research* **2009**, *37*, e45–e45.

(13) Tanaka, J.; Nakagawa, T.; Shiratori, A.; Shimazaki, Y.; Uematsu, C.; Kamahori, M.; Yokoi, T.; Harada, K.; Kohara, Y. KRAS genotyping by digital PCR combined with melting curve analysis. *Scientific reports* **2019**, *9*, 1–8.

(14) Athamanolap, P.; Hsieh, K.; O'Keefe, C. M.; Zhang, Y.; Yang, S.;

Wang, T.-H. Nanoarray Digital Polymerase Chain Reaction with High-Resolution Melt for Enabling Broad Bacteria Identification and Pheno–Molecular Antimicrobial Susceptibility Test. *Analytical chemistry* **2019**, *91*, 12784–12792.

(15) 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. *Analytical chemistry* **2019**, *91*, 2013–2020.

(16) 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. *Analytical chemistry* **2019**, *91*, 7426–7434.

(17) Hammoudi Halat, D.; Ayoub Moubareck, C. The Current Burden of Carbapenemases: Review of Significant Properties and Dissemination among Gram-Negative Bacteria. *Antibiotics* **2020**, *9*, 186.

(18) 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. *Scientific Reports* **2020**, *10*, 1–9.

(19) Gou, T.; Hu, J.; Wu, W.; Ding, X.; Zhou, S.; Fang, W.; Mu, Y. Smartphone-based mobile digital PCR device for DNA quantitative analysis with high accuracy. *Biosensors and Bioelectronics* **2018**, *120*, 144–152.

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

(21) 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.

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

(23) 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.

(24) 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.

(25) Wong, B. Points of view: Color blindness. *Nature Methods* **2011**, *8*, 441–441.

(26) 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. *Analytical chemistry* **2013**, *85*, 11619–11627.

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

(28) Millar, B. C.; Xu, J.; Moore, J. E. Risk assessment models and contamination management: implications for broad-range ribosomal DNA PCR as a diagnostic tool in medical bacteriology. *Journal of Clinical Microbiology* **2002**, *40*, 1575–1580.

(29) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.

(30) Abu-Mostafa, Y. S.; Magdon-Ismail, M.; Lin, H.-T. *Learning from data*; AMLBook New York, NY, USA:, 2012; Vol. 4.

(31) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*; Springer Science & Business Media, 2009.

(32) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.

(33) Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory* **1967**, *13*, 21–27.

(34) Sun, B.; Rodriguez-Manzano, J.; Selck, D. A.; Khorosheva, E.; Karymov, M. A.; Ismagilov, R. F. Measuring Fate and Rate of Single-Molecule Competition of Amplification and Restriction Digestion, and Its Use for Rapid Genotyping Tested with Hepatitis C Viral RNA. *Angewandte Chemie International Edition* **2014**, *53*, 8088–8092.

(35) Rodriguez-Manzano, J.; Karymov, M. A.; Begolo, S.; Selck, D. A.; Zhukov, D. V.; Jue, E.; Ismagilov, R. F. Reading out single-molecule digital RNA and DNA isothermal amplification in nanoliter volumes with unmodified camera phones. *ACS nano* **2016**, *10*, 3102–3113.

(36) Moser, N.; Rodriguez-Manzano, J.; Lande, T. S.; Georgiou, P. A scalable ISFET sensing and memory array with sensor auto-calibration for on-chip real-time DNA detection. *IEEE transactions on biomedical circuits and systems* **2018**, *12*, 390–401.

(37) Miscourides, N.; Yu, L.-S.; Rodriguez-Manzano, J.; Georgiou, P. A 12.8 k current-mode velocity-saturation ISFET array for on-chip real-time DNA detection. *IEEE transactions on biomedical circuits and systems* **2018**, *12*, 1202–1214.

(38) Malpartida-Cardenas, K.; Miscourides, N.; Rodriguez-Manzano, J.; Yu, L.-S.; Moser, N.; Baum, J.; Georgiou, P. Quantitative and rapid Plasmodium falciparum malaria diagnosis and artemisinin-resistance detection using a CMOS Lab-on-Chip platform. *Biosensors and Bioelectronics* **2019**, *145*, 111678.

(39) Cacho-Soblechero, M.; Malpartida-Cardenas, K.; Cicatiello, C.; Rodriguez-Manzano, J.; Georgiou, P. A Dual-Sensing Thermo-Chemical ISFET Array for DNA-Based Diagnostics. *IEEE Transactions on Biomedical Circuits and Systems* **2020**,

(40) Kalofonou, M.; Malpartida-Cardenas, K.; Alexandrou, G.; Rodriguez-Manzano, J.; Yu, L.-S.; Miscourides, N.; Allsopp, R.; Gleason, K. L.; Goddard, K.; Fernandez-Garcia, D., et al. A novel hotspot specific isothermal amplification method for detection of the common PIK3CA p. H1047R breast cancer mutation. *Scientific Reports* **2020**, *10*, 1–10.

Graphical TOC Entry

Amplification Curve Analysis

Digital PCR   Real-Time Data   Machine Learning