Imperial College London

Department of Electrical and Electronic Engineering
Centre for Bio-Inspired Technology
&
Department of Infectious Diseases
Centre for Antimicrobial Resistance Optimisation

The Art of PCR Assay Development: Data-Driven Multiplexing

Luca Miglietta

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy (Ph.D.) in Data Science and Molecular Medicine
of Imperial College London and the Diploma of Imperial College (DIC)
February 2022

To my father...

for illuminating my path in dark places
when all other lights had gone out

I am forever grateful.

Abstract

The present thesis describes the discovery and application of a novel methodology, named Data-Driven Multiplexing, which uses artificial intelligence and conventional molecular instruments to develop rapid, scalable and cost-effective clinical diagnostic tests.

Detection of genetic material from living organisms is a biologically engineered process where organic molecules interact with each other and with chemical components to generate a meaningful signal of the presence, quantity or quality of target nucleic acids. Nucleic acid detection, such as DNA or RNA detection, identifies a specific organism based on its genetic material. In particular, DNA amplification approaches, such as for antimicrobial resistance (AMR) or COVID-19 detection, are crucial for diagnosing and managing various infectious diseases. One of the most widely used methods is Polymerase Chain Reaction (PCR), which can detect the presence of nucleic acids rapidly and accurately. The unique interaction of the genetic material and synthetic short DNA sequences called primers enable this harmonious biological process. This thesis aims to bioinformatically modulate the interaction between primers and genetic material, enhancing the diagnostic capabilities of conventional PCR instruments by applying artificial intelligence processing to the resulting signals.

To achieve the goal mentioned above, experiments and data from several conventional platforms, such as real-time and digital PCR, are used in this thesis, along with state-of-the-art and innovative algorithms for classification problems and final application in real-world clinical scenarios. This work exhibits a powerful technology to optimise the use of the data, conveying the following message: the better use of the data in clinical diagnostics enables higher throughput of conventional instruments without the need for hardware modification, maintaining the standard practice workflows.

In Part I, a novel method to analyse amplification data is proposed. Using a state-of-the-art digital PCR instrument and multiplex PCR assays, we demonstrate the simultaneous detection of up to nine different nucleic acids in a single-well and single-channel format. This novel concept called Amplification Curve Analysis (ACA) leverages kinetic information encoded in the amplification curve to classify the biological nature of the target of interest. This method is applied to the novel design of PCR assays for multiple detections of AMR genes and further validated with clinical samples collected at Charing Cross Hospital, London, UK. The ACA

showed a high classification accuracy of 99.28% among 253 clinical isolates when multiplexing. Similar performance is also demonstrated with isothermal amplification chemistries using synthetic DNA, showing a 99.9% of classification accuracy for detecting respiratory-related infectious pathogens.

In Part II, two intelligent mathematical algorithms are proposed to solve two significant challenges when developing a Data-driven multiplex PCR assay. Chapter 7 illustrates the use of filtering algorithms to remove the presence of outliers in the amplification data. This demonstrates that the information contained in the kinetics of the reaction itself provides a novel way to remove non-specific and not efficient reactions. By extracting meaningful features and adding custom selection parameters to the amplification data, we increase the machine learning classifier performance of the ACA by 20% when outliers are removed. In Chapter 8, a patented algorithm called Smart-Plexer is presented. This allows the hybrid development of multiplex PCR assays by computing the optimal single primer set combination in a multiplex assay. The algorithm's effectiveness stands in using experimental laboratory data as input, avoiding heavy computation and unreliable predictions of the sigmoidal shape of PCR curves. The output of the Smart-Plexer is an optimal assay for the simultaneous detection of seven coronavirus-related pathogens in a single well, scoring an accuracy of 98.8% in identifying the seven targets correctly among 14 clinical samples. Moreover, Chapter 9 focuses on applying novel multiplex assays in point-of-care devices and developing a new strategy for improving clinical diagnostics.

In summary, inspired by the emerging requirement for more accurate, cost-effective and higher throughput diagnostics, this thesis shows that coupling artificial intelligence with assay design pipelines is crucial to address current diagnostic challenges. This requires crossing different fields, such as bioinformatics, molecular biology and data science, to develop an optimal solution and hence to maximise the value of clinical tests for nucleic acid detection, leading to more precise patient treatment and easier management of infectious control.

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr Jesus Rodriguez-Manzano, for shaping me into the researcher I am today, guiding every step of this thesis, and Prof. Pantelis Georgiou, for his support towards completing my PhD. I was granted the opportunity to work on a range of exciting projects under their outstanding leadership and alongside sensational collaborators such as Dr Myrsini Kaforou, Prof. Shiranee Sriskandan and Prof. Alison Holmes. The cutting-edge research we conduct at the Center for Antimicrobial Resistance Optimisation (CAMO) and the Centre for Bio-Inspired Technology (CBIT) is truly intellectually stimulating.

This work would not have come to fruition without the support of both teams due to my thesis's highly interdisciplinary nature. In particular, I would like to thank Dr Ahmad Moniri, for taking me through the first step in the data-driven research; Dr Ivana Pennisi and Matthew Cavuto for being invaluable friends and pioneers of novel ideas; Ke Xu for his uplifting humour and his empowering knowledge that made possible our last two excellent studies; Louis Kreitmann, Giselle D'Sousa and Dr Heather Jackson for their innovation and support.

Moreover, I am grateful to my colleagues at DNA nudge. To the father of the nudgers, Professor Christofer Toumazou, and my mentor Dr Rashmita Sahoo for their unconditional support throughout my studies and for letting me be part of the nudge's exciting journey. I thank the science team: Dr Sara De Mateo-Lopez, for being the best person that I have ever worked with; Dr Nicola Casali for directing my career progression, Chris Icely and Stavros Pournias, for being my dearest fantastic friends. Thanks to Dr Felice Leung, Dr Mohammadreza Sohbati, Dr Linglan Zhang, Dr Cathal McElgunn, Neeti Viswanathan and John Tassone.

Finally, I would like to express my deep and sincere gratitude to my future wife, Dr Linhongjia Xiong for her endless love and for being the only support during the most stressful and difficult periods of my life. Thanks to my non-blood brother Alex and the De Pascalis Family for always leaving the door open to me, and for the amazing cuisine of Mother Daniela. A warm thanks goes to all my relatives, especially to my cousins, her Majesty Livia and the Baron of Vasto, Mario De Marinis. To my beautiful god-mother/aunt Cristina, my god-father Giovanni and my aunt Lucia and uncle Pietro. And last but not least, I am forever indebted to my father, Leonardo Miglietta, my mother, Teresa, my sister, Valentina, my brother-in-law Federico, and my loving nephew, Lorenzo, for their endless encouragement, support and love.

Declaration of Originality

This thesis is a presentation of my own original work. All sources and materials which do not belong to my research work have been properly referenced in this thesis.

Luca Miglietta

Copyright Declaration

Luca Miglietta

Contents

List of Tables

List of Figures

List of Abbreviations

- **AC:** Amplification Curve
- **ACA:** Amplification Curve Analysis
- **AI:** Artificial Intelligence
- **AMCA:** Amplification and Melting Curve Analysis
- **AMR:** Antimicrobial Resistance
- **CNN:** Convolutional Neural Networks
- **CPOs:** Carbapenemase-Producing Organisms
- **DBSCAN:** Density-based Spatial Clustering of Applications with Noise
- **dPCR:** real-time Digital PCR
- **dsDNA:** Double-strand DNA
- **FFI:** Final Fluorescence Intensity
- **KNN:** the K-Nearest Neighbour
- **LOF:** Outlier Factor
- **MC** Melting Curve
- **MCA:** Melting Curve Analysis
- **ML:** Machine Learning
- **MSA:** Multiple Sequence Alignment
- **MSC:** Multi-dimensional Standard Curves
- **NAAT:** Nucleic Acid Amplification Test
- **NCBI:** National Center for Biotechnology Information

- **OC-SVM:** One-Class Support Vector Machine
- **PCA:** principal component analysis
- **PCR:** Polymerase Chain Reaction
- **PoC:** Point-of-Care
- **qdPCR:** real-time digital PCR
- **qPCR:** real-time Quantitative PCR
- **RF:** Random Forest
- **ROX:** 6-Carboxyl-X-Rhodamine
- **RTI:** Respiratory Tract Infection
- **ssDNA:** Single-strand DNA
- **SVM:** Support Vector Machines
- **t-SNE:** t-distributed Stochastic Neighbor Embedding
- **WMP:** Wrong Melting Percentage

List of Publications

Accepted/Published Peer-Reviewed Journals

[J1] Moniri A*, Miglietta L*, Malpartida-Cardenas K, Pennisi I, Moser N, Holmes A, Georgiou P, Rodriguez-Manzano J. "Amplification Curve Analysis: Data-Driven Multiplexing Using Real-Time Digital PCR." *ACS Analytical Chemistry*, 2020 Oct 6:92(19):13134-13143.

[J2] Moniri A*, Miglietta L*, Holmes A, Georgiou P, Rodriguez-Manzano J. "High-Level Multiplexing in Digital PCR with Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis." *ACS Analytical Chemistry*, 2020 Oct 20:92(20):14181-14188.

[J3] Rodriguez-Manzano J, Malpartida-Cardenas K, Moser N, Pennisi I, Cavuto M, Miglietta L, Moniri A, Penn R, Satta G, Randell P, Davies F, Bolt F, Barclay W, Holmes A, Georgiou P. "Handheld Point-of-Care System for Rapid Detection of SARS-CoV-2 Extracted RNA in under 20 min." *ACS Central Science*, 2021 Feb 24:7(2):307-317.

[J4] Li HK, Kaforou M, Rodriguez-Manzano J, Channon-Wells S, Monir A, Habgood-Coote D, Gupta RK, Mills EA, Lin J, Chiu YH, Pennisi I, Miglietta L, et al. "Discovery and validation of a 3-gene signature to distinguish COVID-19 and other viral infections in emergency infectious disease presentations; a case-control then observational cohort study." *The Lancet Microbe*, vol. 2, no. 11 (2021): e594-e603.

[J5] Miglietta L, Moniri A, Pennisi I, Malpartida-Cardenas K, Abbas H, Hill-Cawthorne K, Bolt F, Jauneikaite E, Davies F, Holmes A, Georgiou P. "Coupling machine learning and high throughput multiplex digital PCR enables accurate detection of carbapenem-resistant genes in clinical isolates". *Frontiers in molecular biosciences*, 2021:8:775299.

[J6] Malpartida-Cardenas K*, Miglietta L*, Peng T, Moniri A, Holmes A, Georgiou P, Rodriguez Manzano J. "Single-channel digital LAMP multiplexing using amplification curve analysis." *Sensors & Diagnostics*, 2022 May 19:1(3):465-8.

[J7] Cavallo FR, Mirza KB, de Mateo S, Miglietta L, Rodriguez-Manzano J, Nikolic K, Toumazou C. A point-of-care device for fully automated, fast and sensitive protein quantification via qPCR. *Biosensors*, 2022 Jul 19:12(7):537.

[J8] Pennisi I, Moniri A, Miscourides N, Miglietta L, Moser N, Habgood-Coote D, Herberg JA, Levin M, Kaforou M, Rodriguez-Manzano J, Georgiou P. "Discrimination of bacterial and viral infection using host-RNA signatures integrated in a lab-on-chip platform". *Biosensors and Bioelectronics*, 2022 Aug 24:114633.

[J9] Moser N, Yu LS, Rodriguez Manzano J, Malpartida-Cardenas K, Au A, Arkell P, Cicatiello C, Moniri A, Miglietta L, Wang WH, Wang S, Holmes AH, Chen YH, Georgiou P. "Quantitative detection of dengue serotypes using a smartphone-connected handheld Lab-on-Chip platform". *Frontiers in Bioengineering and Biotechnology*, 2022 Sep 15:892853.

[J10] Miglietta L*, Xu K*, Chhaya PM, Kreitmann L, Hill-Cawthorne K, Bolt F, Holmes AH, Georgiou P, Rodriguez-Manzano J. "Adaptive Filtering Framework to Remove Nonspecific and Low-Efficiency Reactions in Multiplex Digital PCR Based on Sigmoidal Trends". *ACS Analytical Chemistry*, 2022 Oct 3;94(41):14159-68.

[J11] Kreitmann L, Miglietta L, Xu K, Malpartida-Cardenas K, D'Souza G, Kaforou M, Brengel-Pesce K, Drazek L, Holmes A, Rodriguez-Manzano J, "Next-generation molecular diagnostics: leveraging digital technologies to enhance multiplexing in real-time PCR." *TrAC Trends in Analytical Chemistry*, 2023 Mar 1:vol.160.

Submitted & Prepared Journals

[S1] Boonyasiri A, Myall A, Wan Y, Bolt F, Ledda A, Mookerjee S, Weiße AY, Turton JF, Abbas H, Prakapaite R, Sabnis A. Alireza Abdolrasouli, Malpartida-Cardenas K, Miglietta L, Donaldson H, Gilchrist M, Hopkins KL, Ellington MJ, Otter JA, Larrouy-Maumus G, Edwards AM, Rodriguez-Manzano J, Didelot X, Barahona M, Holmes AH, Jauneikaite E, Davies F. "Integrated patient network and genomic plasmid analysis reveal

a regional, multi-species outbreak of carbapenemase-producing Enterobacterales carrying both blaIMP and mcr-9 genes". *Under review* - `https://www.medrxiv.org/content/10.1101/2021.10.28.21265436v1`.

[S2] Miglietta L, Chen Y, Luo Z, Xu K, Ding N, Peng T, Moniri A, Kreitmann L, Cacho-Soblechero M, Holmes A, Georgiou P, Rodriguez-Manzano J. "Smart-Plexer: a breakthrough workflow for hybrid development of multiplex PCR assays". *Under review* - `https://doi.org/10.21203/rs.3.rs-1765213/v1`.

[S3] Mao Y*, Xu K*, Miglietta L, Kreitmann L, Moser N, Georgiou P, Holmes A, Rodriguez-Manzano J. "Deep Domain Adaptation Enhances Amplification Curve Analysis for Single-Channel Multiplexing in Real-Time PCR". *Under review* - `https://www.techrxiv.org/articles/preprint/Deep_Domain_Adaptation_Enhances_Amplification_Curve_Analysis_for_Single-Channel_Multiplexing_in_Real-Time_PCR/21334701`.

[S4] Jackson H*, Miglietta L*, Habgood-Coote D, ..., Rodriguez-Manzano J, Kaforou M, Levin M. "Diagnosis of multi-system inflammatory syndrome in children by a whole-blood transcriptional signature". *Under review.*

Published & Filed Patents

[P1] Rodriguez-Manzano J, Moniri A, Miglietta L and Georgiou P. "Identifying a target nucleic acid", WO2022038279A1, Assignee: Imperial Innovations Limited, 2020.

[P2] Rodriguez-Manzano J, Moniri A, Miglietta L and Georgiou P. "Method of assay design", GB2108339.9, Assignee: Imperial Innovations Limited, 2021.

[P3] Rodriguez-Manzano J, Jackson H, Miglietta L, Habgood-Coote D, Kaforou M. "A Method to optimise transcriptomic signatures", GB2211707.1, Assignee: Imperial Innovations Limited, 2022.

[P4] Levin M, Kaforou M, Rodriguez-Manzano J, Jackson H, Miglietta L. Diagnosis of multi-system inflammatory syndrome in children by a whole-blood transcriptional". Signature Assignee: Imperial Innovations Limited, 2023.

* *First joint authorship.*

Chapter 1

Introduction

1.1 Motivation

The life sciences have a long history of dealing with a large amount of data, and current advances in instrument throughput have increased the capability of analysing and storing data [1]. There is a gap between the number of produced data and their analytical use and interpretation [2]. Better usage of the data can lead us to rapid, cost-effective and precise solutions for several research fields such as genomics, agriculture, environmental protection, cancer and clinical diagnostics [3]. One example which affected us most during the past two years is the recent coronavirus pandemic, where massive efforts have been made in order to control this disease [4]. What could have been done better to reduce the spreading of such outbreaks? The answer is straightforward - screen pathogenic diseases in a faster, cost-effective and scalable manner, knowing the source of infection and developing approaches that can allow rapid diagnostics. One solution has been provided by the detection of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) through molecular tests [5, 6].

Nucleic acids contain genetic information in living organisms, like byte-compiled code for a virtual machine. The language of DNA is digital but not binary; in fact, unlike the binary encoding with zeros and ones, the DNA has four different nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T) [7]. Like in the alphabet, the order, or the sequence, of the

chemical bases, defines the genetic code and the necessary information to ensure the optimal functioning of living organisms. The genetic code also contains the fingerprint of the biological classification of organisms and through DNA detection humans are capable of identifying living species at any level. This feature is crucial for many healthcare applications such as identifying pathogenic agents (e.g. coronaviruses), but also in recognising a person's identity or revealing the genetics of human diseases such as cancer.

From Marshall Nirenberg, who first sequenced DNA bases [8], to Frederick Sanger, who sequenced the full human genome for the first time [9], DNA detection has been the focus of many research works and applications for over 80 years. Various techniques have been developed to determine the sequence of a nucleic acid chain, but the gold-standard method widely used for DNA detection remains the Polymerase Chain Reaction or PCR. The success of the PCR technique derives from its simplicity in development and application, its time and cost-effectiveness and lastly its robustness [10]. Moreover, unlike sequencing approaches that require heavy computational power to process large amounts of genetic information correctly, PCR can be seen as a simple binary signal, where the occurrence of a sigmoidal amplification trend indicates the presence of a desired target. The aim of this thesis is to challenge the idea that PCR signals can only be interpreted as a binary outcome, showing the potential in the full use of the hidden information in the kinetics and thermodynamics behaviour of the DNA amplification event and data.

Performing PCR requires five core 'ingredients': (i) the DNA template to be copied; (ii) primers, short stretches of DNA that initiate the PCR reaction, designed to bind to either side of the section of DNA to copy; (iii) DNA nucleotide bases (A, C, G and T) to construct the new strand of DNA; (iv) the Polymerase enzyme to add in the new DNA bases; (v) various buffers to ensure the right conditions for the reaction. All these ingredients undergo a process of heating and cooling called thermal cycling performed by a machine in three main steps: (i) Denaturation, when the double-stranded template DNA is heated to separate it into two single strands; (ii) Annealing, when the temperature is lowered to enable the DNA primers to attach to the template DNA. (iii) Extending, when the temperature is raised, and the new strand of DNA is made by the Polymerase enzyme. These steps are repeated a certain number of times

to ensure that the number of DNA molecules is doubled on each cycle till a biological signal is observed [11].

Although PCR is an accurate, reliable, well-established and routinely used technology, the need for higher throughput of a single PCR reaction has always been the main focus for scaling the detection capabilities of this technique. Therefore, the concept of multiplex PCR was first described in 1988 by Jeffrey S. Chamberlain [12], allowing simultaneous detection of two or more regions of interest in the target genome. As shown in Figure 1.1, producing multiplex PCR systems is as simple as combining three (or more) reactions in a single tube to produce three different outputs based on the presence of the target in the tube. This is extremely beneficial to reduce the cost of the reaction, the usage of clinical samples and the throughput provided by a single PCR.

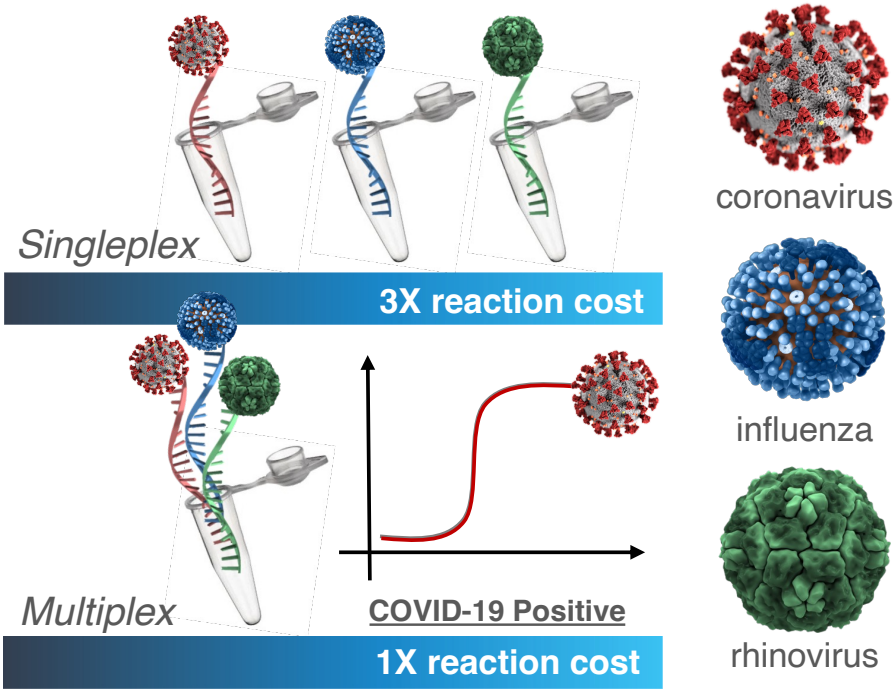


Figure 1.1: Singleplex and Multiplex PCR. In the top part of the figure, the concept of singleplex is depicted, where a pathogen (in this case, viruses) is detected. The tube has all the reagents for detecting a single specific virus; therefore, when the chemical signal is detected from that particular reaction, it is possible to identify which of the three viruses is amplified by looking at the label on the tube. This concept is also called spatial multiplexing. The bottom part of the figure shows the concept of multiplexing in a single well reaction, where any of the three pathogens can be detected simultaneously, through a biochemical tag (such as a fluorophore with different colours) or advanced data processing.

Moreover, other more complex strategies have been successfully established for multiplexing, leveraging the capabilities of more sophisticated and higher throughput PCR machines [13]. From PCR instruments with multiple fluorescence channels (up to six in a single machine) to the more recent digital PCR instruments for single-molecule amplification, the throughput of a simple PCR reaction has been enhanced drastically. Moreover, during the COVID-19 pandemic, a rise in Point-of-Care instrumentation has been seen in the effort to perform PCR in a portable manner. All these developments resulted in an enormous quantity of data to analyse and interpret the molecular test outcome accurately [14].

Previous work by Dr. Jesus Rodriguez-Manzano and Dr. Ahmad Moniri was conducted in the effort of leveraging the value of the data from multiplex PCR to enhance absolute quantification using a multidimensional standard curve (MSC) [15, 16, 17]. These studies highlighted the complexity and volume of data produced in the PCR world, which are largely increasing with more advanced instruments. With the rise of Artificial Intelligence in medicine, PCR data can be analysed in more depth, moving towards more data-driven approaches [3]. Coupling Machine Learning with multiplex PCR will benefit the field of DNA detection by analysing data with more features and working in higher dimensionality to get the most value from PCR data.

1.2 Introduction to Data-driven Multiplexing

This thesis describes a novel technique named data-driven multiplexing. This method benefits from state-of-the-art machine learning algorithms to diagnose the presence of multiple nucleic acids (such as bacteria, viruses, fungi, genetic variants and more) in a single chemical reaction, using gold-standard techniques such as real-time PCR (qPCR). Furthermore, this approach can also be used to design and optimise the multiplex assay, reducing the number of experiments and laboratory costs when developing molecular diagnostic tests. Using this technique, it is possible to reduce: (i) the amount of sample needed for screening of multiple genetic locations or pathogens, (ii) the time for multiple targets screening and (iii) the overall cost, drastically

reducing the cost of molecular diagnostics.

Data-driven multiplexing uses artificial intelligence algorithms and tailored chemistries to extract more information from real-time amplification data from conventional PCR instruments in a single-well and single-channel multiplex PCR reaction. The technology does not require novel instrumentation or hardware modifications, but standard assay development pipelines (performed in a molecular biology laboratory) and software development to perform post-test analysis and final output.

1.3 Hypothesis, Objectives & Research Questions

Hypothesis: the information encoded in the amplification and melting curve of a PCR reaction represents the relationship between a target genetic material and the primers used to detect it. Modulating this interaction can be used to develop molecular tests for clinical diagnostics.

Objective: using novel data-driven analysis methods, this thesis aims to investigate the possibility of improving the precision and throughput of diagnosis outcomes without hardware modifications and additional costs at current molecular diagnostic platforms.

Research Questions: as Figure 1.2 shows, the strategy focuses on improving bioinformatics for assay design with tailored chemistries and enhancing the value of the molecular test data through advanced data analytic approaches. To achieve this, the following research questions are the focus of this work:

(i) Is it possible to recognise the nature of a target nucleic acid by the kinetic and thermodynamic information contained in its resulting PCR amplification reaction? Furthermore, can this information differentiate among multiple targets in a single-well and single-channel multiplex PCR reaction? (investigated in Chapter 3 and 4)

(ii) Can data-driven multiplexing be translated to different chemistries, such as isothermal, and across different platforms so it can be applied in the clinic and/or integrated into point-of-care platforms? (investigated in Chapter 5, 6 and 9)

(iii) Is it possible to identify and filter out non-specific and not efficient reactions by looking merely at amplification curves? In doing so, can amplification data be confidently used from singleplex assays to reduce the laboratory testing of multiplex PCR assays and develop high-level data-driven multiplexing assays in a time- and cost-effective manner? (investigated in Chapter 7 and 8)

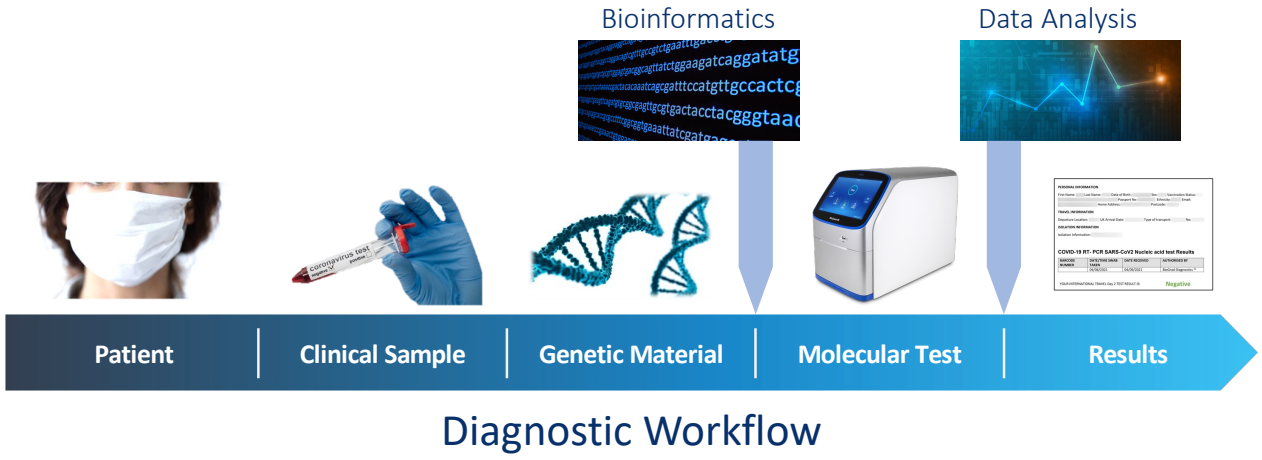


Figure 1.2: Example of a Molecular Diagnostic Pipeline. The horizontal blue arrow indicates the conventional diagnosis workflow from patient to result, where the patient sample is collected from different sources (e.g., nasopharyngeal swabs). Subsequently, nucleic acids are extracted, and the most appropriate genetic test is developed. The first aim of this thesis is to develop novel bioinformatics pipelines to increase the throughput of standard molecular tests (first vertical arrow). The test is performed with the PCR instrument. The second aim is to develop better data analytic approaches using machine learning algorithms to ensure reliable and accurate results (second vertical arrow).

1.4 Thesis Structure and Contribution

As described in Figure 1.3, this thesis is separated into ten chapters, taking the readers through the journey of data-driven multiplexing. This Chapter 1 introduces PCR and the importance of simultaneous detection of nucleic acids, plus the outline of hypothesis, objectives and research questions. Chapter 2 gives technical insights into understanding multiplex PCR and the algorithms used for the data-driven multiplexing approach.

The central technological aspect of this thesis, Part I, explains the discovery of the method (Chapters 3-4), its use with isothermal chemistries (Chapter 5) and its clinical application

(Chapter 6). Part II focuses on the optimisation of data-driven multiplexing (Chapter 7), the *How-To* (Chapter 8) and other applications of intelligent assay design strategies (Chapter 9). The structure of this thesis is depicted in 1.3 and a summary of each Chapter is as follows.

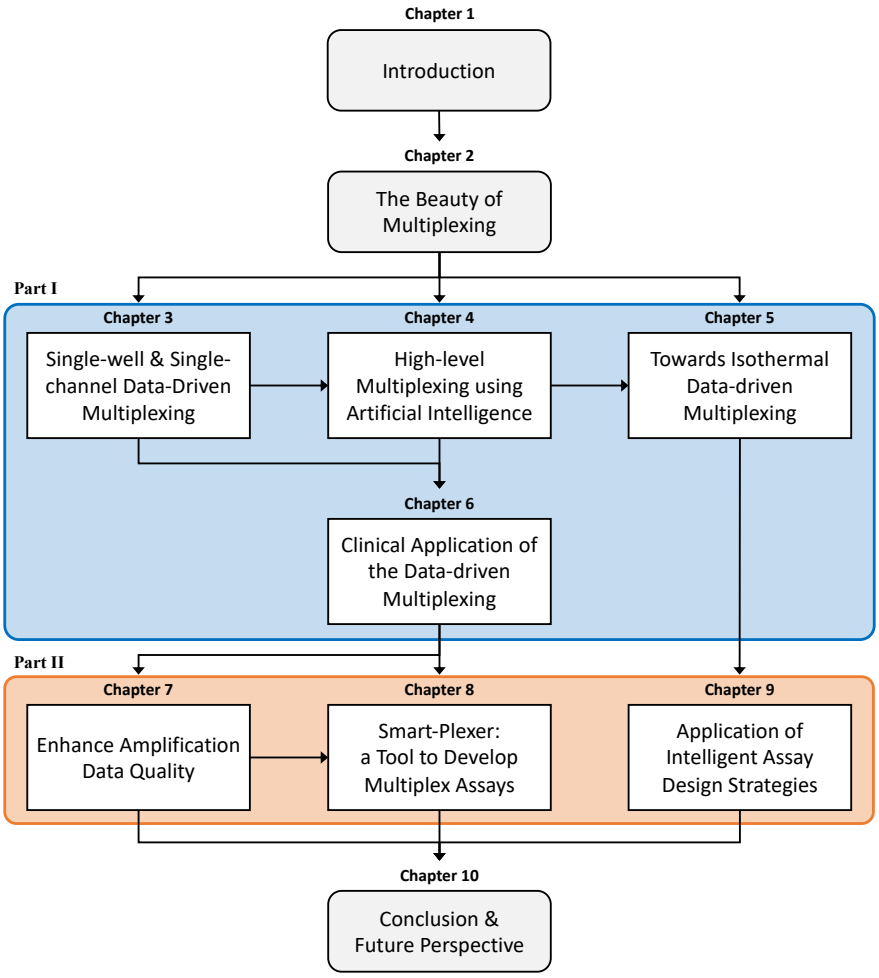


Figure 1.3: Thesis Organisation

1.4.1 Thesis Chapters

Chapter 2: The Beauty of Multiplexing

This Chapter reviews the history of multiplex PCR and its application in real-time PCR and digital PCR instruments. An assessment of several multiplex PCR methodologies is presented, explaining previous and more novel data-driven approaches from both chemical and hardware perspectives.

Part I: The Journey of Data-driven Multiplexing

Chapter 3: Single-well and Single-channel Data-Driven Multiplexing

This Chapter introduces a novel data-driven approach called Amplification Curve Analysis (ACA) to perform single-well and single-channel multiplex PCR reactions. The multi-target classification was achieved by leveraging the kinetic information of the amplification curve. Here the first application of machine learning on the entire amplification curve data is used to distinguish three different targets simultaneously.

Chapter 4: High-level Multiplexing using Artificial Intelligence

This Chapter extends the previous ACA method by exploring the use of melting curves to perform a higher level of classification using nine different targets in a single-well and single-channel multiplex PCR. In biological terms, the melting curves are related to the thermodynamics of the amplification reaction; therefore, using both kinetic and thermodynamic information it is possible to increase the complexity of the multiplex PCR assay, consequently detecting more targets in a single test. This method is called Amplification and Melting Curve Analysis (AMCA).

Chapter 5: Towards Isothermal Data-driven Multiplexing

In this Chapter, data-driven multiplexing is also applied to another kind of amplification chemistry, such as isothermal. Loop-mediated isothermal amplification (LAMP) is used to classify five different targets in a single-well and single-channel multiplex LAMP. This opens the

future perspective of the approach for its use on machines which does not require thermocycling, such as Point-of-Care devices.

Chapter 6: Clinical application of the Data-driven Multiplexing

This Chapter illustrates the clinical application of the AMCA method tackles the burden of antimicrobial resistance screening in hospitals. A total of 253 clinical isolates from patients' samples were screened in time and cost-effective manner coupling a novel 5-plex assay and the data-driven multiplexing approach. In particular, this Chapter shows the success of the first-ever application of the methods to clinical diagnostics.

Part II: Intelligent Algorithms to Optimise Data-driven Multiplexing

Chapter 7: Enhance Amplification Data Quality

The novel approaches depicted in the previous chapters were optimised by removing outliers from amplification events during multiplex PCR reactions. To validate the efficacy of the approach, comparisons with melting curve data are conducted, leading to the finding that thermodynamic information is also contained in the sigmoidal shape of the amplification curves. The Chapter outcome is a universal approach for removing non-specific and low-efficiency events, which makes data-driven multiplexing more accurate when only amplification curves are generated.

Chapter 8: Smart-Plexer: A Tool to Develop Multiplexing Assays

To fully express the potential of data-driven multiplexing and spread over the vast scientific community, a development pipeline for optimal multiplex assays is needed. This Chapter explores the use of an intelligent algorithm called Smart-Plexer, that can generate optimal primer set combinations for ACA approaches. The automation of this process is achieved by a hybrid assay development, coupling laboratory testing and mathematical computation of suitable multiplex assays for single-well and single-channel PCR reactions.

Chapter 9: Application of Intelligent Assay Design Strategies

This Chapter serves as a brief literature review of recent applications of novel assay design

strategies for the application of PCR- and LAMP-based assays in different fields. Here are two examples: (i) the application of LAMP assays for the detection of COVID-19 in Point-of-Care devices; (ii) a novel approach to design PCR reaction for the translation of RNA signature, from RNA sequencing, to a fast and cost-effective diagnostic test. This Chapter highlights the future direction of this field and the ongoing research with insight into the benefits of incorporating more sophisticated data-driven methods.

Chapter 2

The Beauty of Multiplexing

2.1 Chapter Overview

This Chapter provides to the reader an overview of the techniques used in the thesis. Here, several fields are involved, and relevant concepts are explained. More specifically, the first section is focused on the basics of Polymerase Chain Reaction (PCR) and the kinetics and thermodynamics trends of amplification and melting curves, respectively. The following section explains the computational part of PCR amplification, where bioinformatics tools are the focus. PCR amplification is only possible when short sequences of single-stranded DNA called Primers create a specific and efficient bond with the DNA/RNA target (or template). Knowing the target sequence and an optimal area to design primers is one of the primary needs for successful PCR assays. The following section is the core of the introduction: the fundamentals of multiplex PCR. Moreover, a short review of current and novel techniques to perform multiplex assays is presented for PCR and isothermal chemistries. The last section focuses on the basics of digital PCR and its use for several applications, particularly for clinical diagnostics and infectious diseases, which are also the case studies of this thesis.

To emphasise the interdisciplinarity of this work, Figure 2.1 illustrate the connections between each field enclosed in the thesis.

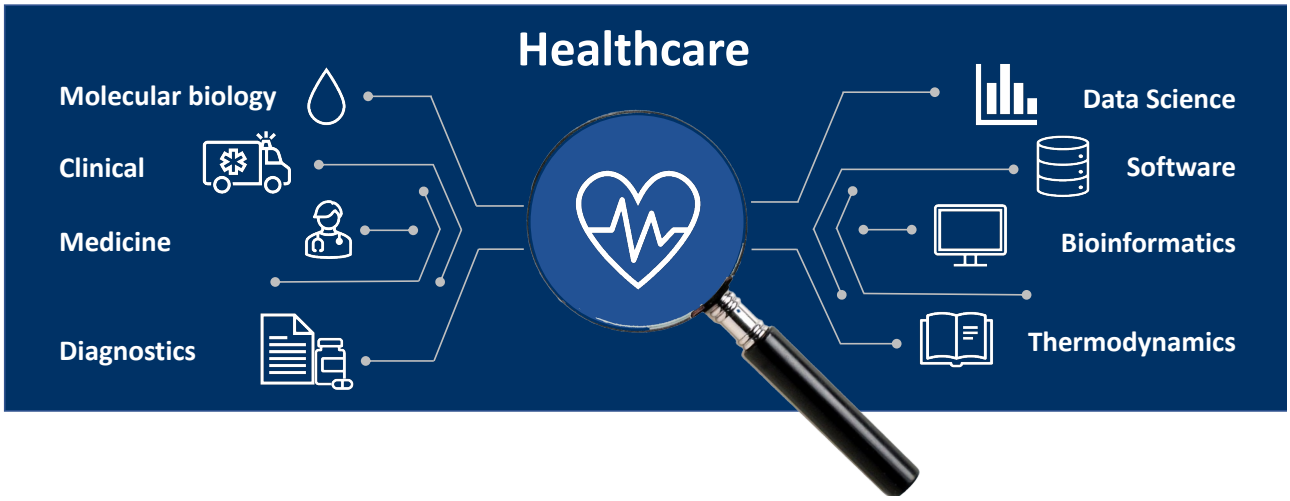


Figure 2.1: Thesis topics network within healthcare sector

2.2 Quantitative PCR

Nucleic Acid Amplification Tests (NAATs), such as Polymerase Chain Reaction (PCR), are the fundamental procedures in life sciences research, bioengineering, and diagnostics. Introduced in 1986 by Katy B. Mullis, PCR is an in-vitro technique capable of amplifying DNA or RNA, generating millions of copies of a specific fragment from a minimum amount of starting material [11]. As the name suggests, the driving force of PCR is the enzyme (Polymerase) that is capable of chaining nucleotides and generating new identical molecules [18]. In molecular biology, this process is coupled with a sequence of temperature cycles commonly repeated 20 to 50 times. The cycling is needed to repetitively denature the DNA duplex at high temperature (typically 95°C), hybridise two DNA oligonucleotides flanking the target sequence (primers) at a temperature between 55°C to 65°C [19], and allow the Polymerase to copy the DNA template using the primers as starting input. As Figure 2.2a shows, each cycle doubles the quantity of target DNA molecules exponentially, and after n cycles, 2^n copies can theoretically be created. Once PCR reagents run out and accumulated PCR products self-anneal, the amplification process saturates and hits a plateau, prohibiting any further amplification [20].

Real-time PCR uses a fluorescent readout to detect the amount of PCR product after each round of amplification [22]. A typical real-time PCR amplification plot is a sigmoidal-shaped

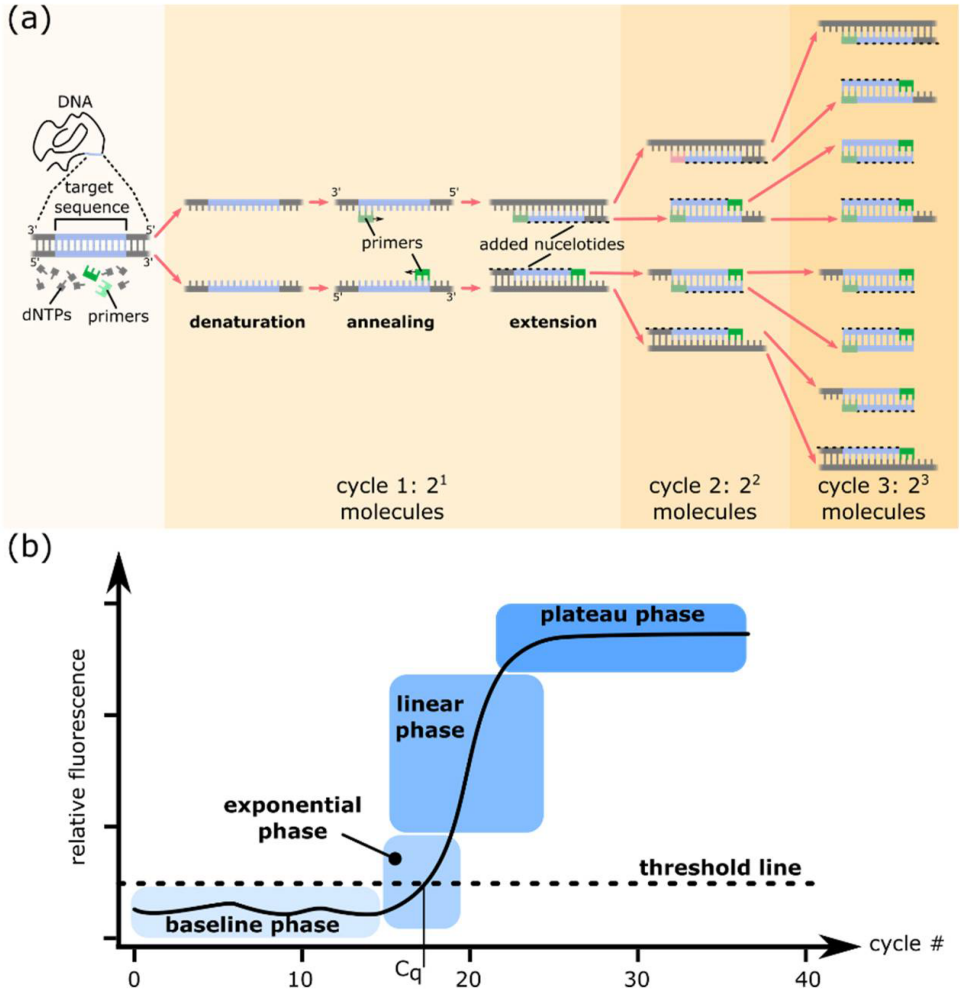


Figure 2.2: Principles of the polymerase chain reaction (PCR). (a) Each PCR cycle includes three steps: (1) Denaturation of double-stranded DNA by heat; (2) Annealing of primers to their complementary target DNA sequences; (3) Extension of primers by a thermostable DNA polymerase. A typical PCR reaction is cycled 20–40 times. Each cycle can theoretically result in a doubling of the number of molecules of the target sequence; (b) Different phases of a real-time PCR amplification plot on a linear scale [21]

curve (on a linear scale) with a baseline phase, an exponential phase, and a linear phase that approaches a plateau Figure 2.2b. The exponential phase of amplification is the most efficient, and if the amplification efficiency is 100%, the amount of PCR products doubles with each cycle. The relative quantification of a target to a calibrator is possible using real-time PCR. When calibrated using a standard curve (made by the use of data from the exponential phase), the procedure is quantitative or qPCR. This approach assumes that the sample and standard amplification efficiencies are equal. Differences in PCR efficiencies can significantly affect the quantification accuracy [23].

Two main ways to add fluorescent labels into PCR are intercalating fluorescent dye and fluorescent probe. Intercalating dyes (e.g., SYBR Green I) can bind to double-strand DNA (dsDNA) non-specifically but are unable to bind single-strand DNA (ssDNA). Fluorescence is emitted when the dye binds to dsDNA, and as the PCR cycling proceeds and dsDNA products accumulate, fluorescent intensity proportionally increases [24]. On the other hand, fluorescent probes are sequence-specific oligonucleotides with a fluorophore at the 5'-end and a quencher, which can inhibit fluorophore emission at the 3'-end. During the DNA amplification, when the Polymerase encounters the probe bound to the target, the exonuclease activity of Taq-polymerase cuts the probe sequence releasing the fluorophore and the quencher in the solution. Fluorescence can now be emitted and accumulated in each amplification cycle when the quencher's inhibition is removed [25, 26].

2.2.1 The Amplification Curve from qPCR

The kinetic information of the PCR reaction is encoded in the resulting amplification curves, and quality of a qPCR result [27]. As PCR is characterised by a series of cycles doubling the starting DNA/RNA material, the exponential behaviour results in a sigmoidal signal, with the horizontal axis being the PCR cycle number and the vertical axis the fluorescence intensity. The most notable feature of the sigmoidal trend in PCR is the Cycle threshold (C_t), representing the intersection between an amplification curve and a threshold line [28]. The threshold line can be arbitrarily set, but typically, it is placed between 10-20% of the final fluorescence intensity (FFI) value as shown in Figure 2.3b. Many factors could potentially influence an amplification curve's C_t, such as the target DNA's initial concentration, the reaction's efficiency and the presence of inhibitory agents [23, 29].

It is essential to mention that many efforts have been put into understanding the kinetics of the amplification curves, and researchers have explored mainly how to extract the most crucial feature from the sigmoidal trend. However, little research has explored the full use of such features [20, 30, 31], and therefore the primary aim of the thesis is to move away from the old concept of PCR as a binary signal and explore in-depth the full use of amplification data.

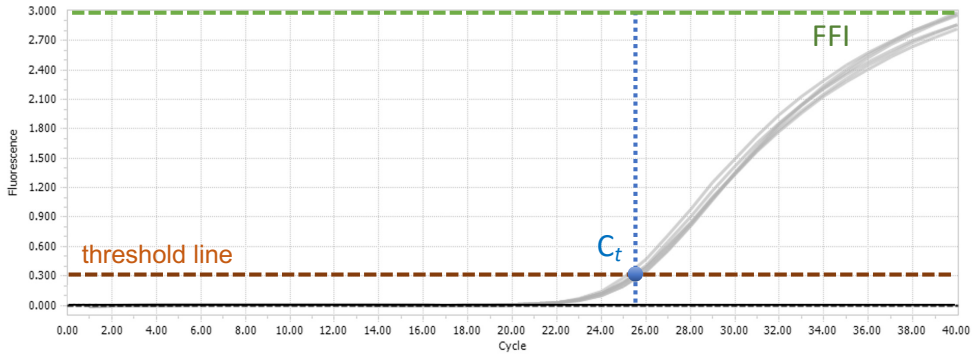


Figure 2.3: The Amplification Curve. Illustration of several replicates of a specific target Amplification Curve in real-time PCR. The orange segmented line indicates the threshold line set at 10% of the FFI value (segmented green line). The C_t is indicated with a blue dot.

2.2.2 The Melting Curve from qPCR using Intercalating Dyes

Melting Curve Analysis involves the assessment of the disassociation characteristics of the DNA strand during heating. It can only be performed with real-time PCR detection technologies using intercalating fluorescent dyes. During the denaturation stage in PCR, the various DNA templates denature at different temperatures. The relation between fluorescence and temperature during the melting step is represented by a bell-shaped graph. As shown in Figure 2.4, the melting temperature (T_m) is the temperature at which half the DNA strands are in the denatured (single-stranded) state [32]. Melting curve analysis is used to identify the targets and to indicate if a reaction is specific to a particular target [33]. The DNA hydrogen bonds rules this behaviour; the A-T pair has two hydrogen bonds, while the C-G pair has three hydrogen bonds; therefore, the temperature required to break C-G bounds is higher, resulting in a later temperature peak [7]. It is important to note that other factors can contribute to modifying the melting curves and their relative peaks, such as secondary structures or nucleotide variations in the amplified regions [34].

In traditional PCR, intercalating dyes bind to any dsDNA; thus, the amplification curve cannot clearly determine which targets are being amplified. Melting curves are often used to validate the identification of the amplified targets. The ability to distinguish between targets by their specific melting peak allows melting curve analysis to be used as a gold standard validation for multi-target detection. However, melting curve analysis is not always accessible,

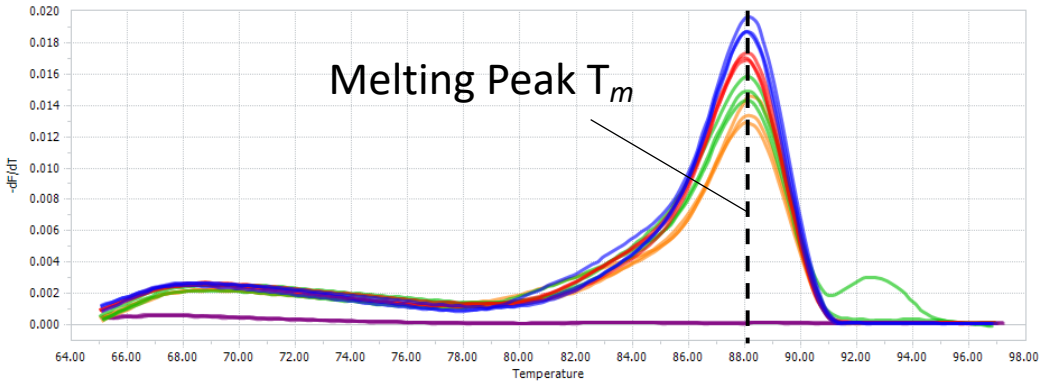


Figure 2.4: The Melting Curve. Illustration of several replicates of a specific target Melting Curve in real-time PCR using intercalating dyes. The black segmented line indicates the Melting Peak of the curve.

especially in Point-of-Care devices, or amplification chemistries such as TaqMan. This further motivates this work as one of the aims of the thesis is the use of amplification curves to identify multiple targets in a single channel without using melting curve analysis.

2.3 Bioinformatics in PCR Assay Development

A 2008 study from A.L. Robertson *et al.*, underlies the lack of bioinformatics knowledge in undergraduate students negatively impact the experiment performance of PCR due to inappropriate primer design and evaluation of annealing temperature of PCR cycling [35]. Understanding primer quality, directionality, and specificity through bioinformatics tools are crucial to PCR assay design, defining its success or failure. Database analysis is vital for accurate DNA detection during the assay design process. Defining boundaries between sequences that have to be detected (inclusivity) and sequences that are not relevant for the PCR test (exclusivity) are significant challenges in the development of primer sets [36]. To give a concrete example, human genomes are 99.9% similar to each others. Even though they are entirely different organisms, human genetic makeup scores over 60% similarity with the genome of a banana [37]. Designing an assay to identify human genomes using PCR primers, attaching random positions of the human genome, has a 60% chance of detecting the genome of a banana as well. This fun fact highlights the importance of exclusivity and the need for bioinformatics pipelines capable

of analysing several genomes from different species to ensure that the PCR primers specifically bind to the desired target (in this case, human genomes). Moreover, PCR is also used to detect single-base variation among organisms of the same species, called Single Nucleotide Polymorphism (SNP). During the COVID-19 pandemic, massive efforts in researching viral variants of the Severe Acute Syndrome Coronavirus 2 (SARS-CoV-2) have been made to tackle the boost in virulence and lethality capability of the SARS-CoV-2 variants. This behaviour was possible by changing a single nucleotide in the entire genomic sequence coding for the spike protein on the virus surface (1 out of 29,903 nucleotides). Detecting such variation is crucial for the surveillance and infection control of COVID-19 [38]. Identifying SNPs across SARS-CoV-2 genomes is possible by comparing millions of sequences and using tailored designed PCR primers on a set of inclusive genomes (specific for the desired variant) is likely to detect circulating variants in communities [39].

2.3.1 Nucleotide Sequence Databases

Biological data, such as DNA or RNA nucleotide sequences, are stored in databases and available to the public. One of the fastest growing repositories of known nucleotide sequences is GeneBank (Genetic Sequence Databank) from the National Center for Biotechnology Information (NCBI) [40], along with other important bioinformatics databases such as the EMBL (European Molecular Biology Laboratory) [41], GISAID (Global Initiative on Sharing Avian Influenza Data) [42], GOLD (Genomes Online Database at the University of Illinois) [43], dbSNP (Database of Single Nucleotide Polymorphisms) [44], CARD (Comprehensive Antibiotic Resistance Database), PATRIC (Pathosystems Resource Integration Center) [45] and many more. The most common nucleotide sequences file format used in these databases is the FASTA format, which is text-based for representing either nucleotide sequence in single letters. The first line in a FASTA file starts with a ">" (greater-than) symbol where the name of the sequence is stored, followed by the genomic sequences in the next line [46].

Usually, databases contain several thousands of sequences of the same organism, and a species can be an object of several studies from different groups around the globe. For example,

the simplest way to compare two sequences is to calculate the number of matching symbols after alignment. The value that measures the degree of sequence similarity is the alignment score of two sequences. In bioinformatics, the Basic Local Alignment Search Tool (BLAST) is the most common algorithm and program used for comparing biological sequence information and detecting similarities between them [47]. As shown in Figure 2.5, to ensure that the sequence coverage is high across genomes or genes of the same species worldwide, it is essential to perform a multiple sequence alignment (MSA) of the retrieved sequences from the BLAST search [48]. Web tools such as Clustal Omega [49] or software like Geneious Prime (Biomatters Ltd) compute rapid and efficient MSA. In this thesis, MSA are heuristic-based aligners, meaning that a local alignment search is used to operate faster than optimal or exact methods (which, in the case of large sequences, can require a lot of computational power). Identifying the most conserved region across genomes or genes of a target species is the foundation for the following step of assay design: Primer Design.

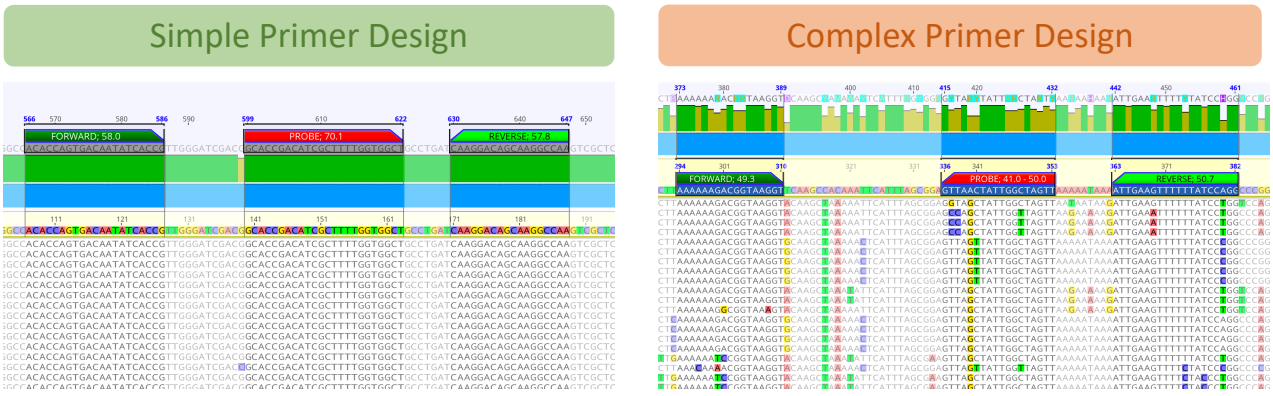


Figure 2.5: Primer Design Coverage. On the left, the GENEious visualisation of a multiple sequence alignment (MSA) with high similarity among different entries of the target DNA (i.e. a bacterial or human genome). The Assay is shown on the top of the graph, with the arrows indicating the direction of the primers. Here it can be observed that the higher similarity makes it simple to design a primer keeping high coverage. On the right, the GENEious visualisation of MSA with lower similarity among different entries. Inclusivity of the PCR (TaqMan) assay has to be ensured by designing primers in more conserved regions, avoiding as many mismatches as possible (coloured DNA bases indicate mismatches).

2.3.2 Primer Design

Primer Design is the critical step in developing PCR assays as the correct strategy in choosing primer sets can minimise troubleshooting by avoiding lengthy and costly laboratory testing and ensure that the PCR reaction works efficiently in a sensitive and specific manner. Primers are short sequences of ssDNA that bind specifically to the template DNA, seeding the Polymerase to initiate the amplification event. The terms primers or primer set or assay, or singleplex, are all indicating the composition of the PCR assay. Two primers are required during the amplification reaction, one called forward and the other reverse, as they bind to both leader and lagging DNA strands, respectively.

A primer is a short synthetic oligonucleotide which is used in PCR and other techniques such as sequencing. These primers are designed to have a sequence which is the reverse complement of a region of template or target DNA to which the primer has to anneal. When designing primers for PCR, bioinformatics analysis is necessary to make predictions about the performance of the primers, for parameter such as:

- Primer and sequence target GC content (%)
- Primer and sequence target Length (nt)
- maximum and minimum primer T_m
- maximum and minimum primer 3' clamp
- maximum and minimum primer hairpin T_m
- maximum and minimum primer cross-dimer T_m

Several programs (i.e. primer3 [50]) will perform these calculations on any primer sequence or pair. The Success of primer design is fundamental to develop efficient and highly sensitive assay. Another important aspect is to ensure that the designed primers do not present cross-reactivity with undesired organisms. This can be evaluated by a primer BLAST search [51]. When cross-reactivity is a concern, it is import to fine-tune the primer design by targeting only

conserved regions for the inclusive targets or target regions where the exclusive species exhibit a low similarity score compared to the binding region of the primer (hence high presence of mismatches between undesired targets and primer binding sites).

2.4 Fundamentals of Multiplex PCR

In the previous section, biological process and the bioinformatics are described to design PCR assays. The next complexity level in the hierarchy of qPCR is the design of Multiplex PCR. Multiplexing expands the PCR capabilities, allowing multi-target detection simultaneously [52, 53]. As shown in Figure 1.1, multiplex PCR provides a practical solution for nucleic acid detection in a single reaction, reducing the time, cost, and amount of the sample required at the expense of technical complexity. In many clinical applications, it is important to detect several DNA targets simultaneously and in one reaction, reducing the sample consumption, the time and the cost of the reaction. Furthermore, because dozens of different pathogens can be responsible for similar clinical manifestations, their concomitant detection in limited amounts of patient samples can be both an important diagnostic endpoint and a technical challenge [54].

Multiplexing can be achieved through several strategies, such as spatial approaches, probe-based methods or melting curve analysis. Spatial techniques, i.e. leveraging microfluidics systems, segregate PCR reactions in many compartments, allowing for parallel amplification and identification of different targets location-based. Probe-based approaches rely on fluorophores with varying emission wavelengths [55]. Additional optical enables target identification through a specific colour-sequence mapping. As long as the emission wavelengths of the different probes do not overlap, the number of target genes for concurrent detection is theoretically unlimited. However, both approaches present several limitations. On the one hand, spatial multiplexing requires multiple reactions of the same sample, consuming a high quantity of reagents and samples. On the other hand, multiplexing with probe-based approaches is expensive as uncommonly used fluorophore (e.g. Cy5.5) can double the price of the assay compared to a standard fluorophore (e.g. FAM) [52, 56]. Instead, melting curve analysis represents a simplified optical

detection system using a single intercalating dye. The melting step is performed after the PCR thermocycling. If multiple targets are present, the melting peak T_ms are the differentiation factor as each represents a target as shown in Figure 2.6. However, because T_m ranges are limited, and close melting curves may be challenging to discern, primer design becomes more complex, plus nucleotide sequence becomes a limiting factor for primer location (highlighted grey area in Figure 2.6) [57, 58]. This highlights the need to develop novel techniques to achieve rapid and cost-effective multiplexing solutions, which is the main aim of this thesis.

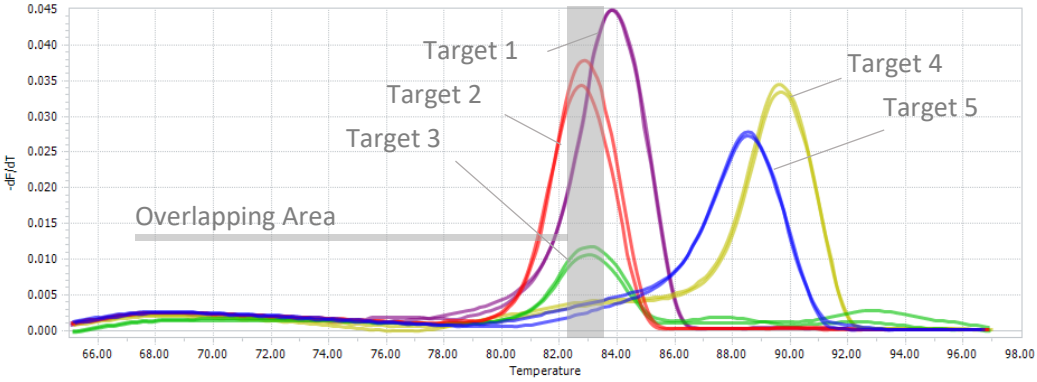


Figure 2.6: Multiplexing Melting Curves. Illustration of five targets' specific Melting Curve in real-time PCR using intercalating dyes. The highlighted area represents the temperature where melting curves overlap.

2.5 Machine Learning for Multiple Nucleic Acids Detection

Machine learning (ML), which relies on the confluence of statistics and computer science, as well as the basis of data science and artificial intelligence (AI), is currently one of the fastest growing fields widely used in research, innovative technology, and nearly every facet of human society [59]. This thesis is an example of applying novel machine learning and signal processing approaches for biochemical data to improve their throughput, analysis and interpretation. A regular job in machine learning is researching and developing algorithms (models) that can learn and predict data [60]. ML algorithms require a training or learning dataset, which is utilised to fit the model. The testing or validation dataset is then predicted using the fitted

model. By using a separate testing or validation dataset to evaluate the model's performance while modifying its parameters, we can obtain an unbiased evaluation of the model that is not biased towards the training data, and estimate its generalization error to ensure that it can perform well on new, unseen data [60]. In general, parameters of a model are modified based on the training dataset. The validation dataset is then used to evaluate the performance of the model during the training process, and to determine when to stop training in order to avoid overfitting. The testing dataset is used to assess the final performance of the model after training is completed and the parameters are fixed.

It is important to note that the testing dataset should not be used to modify the parameters of the model, as this would introduce bias and invalidate the assessment of the model's performance on new, unseen data. Instead, the testing dataset should be used solely for evaluation purposes, and the parameters of the model should be fixed based on the results from the validation dataset.

Machine learning algorithms are divided into two categories: supervised and unsupervised learning. The task of learning a function that maps an input to an output based on example input-output pairs is known as supervised learning. This necessitates "labelled" datasets, in which the desired result is ascribed to each input data point. Support vector machines (SVM), the k-nearest neighbour (KNN) technique, linear/logistic regression, neural networks, and other supervised machine learning algorithms will be investigated throughout this study [61]. While supervised learning relies on labeled data to train a model to predict outcomes, unsupervised learning involves finding patterns and structures in data without prior knowledge of the outcome or labels. This allows for the emergence of self-organization and the capturing of patterns through neuronal predilections or probability densities, enabling a deeper understanding of the underlying data distribution. Although the thesis focuses mainly on supervised learning algorithms, standard unsupervised methods such as K-Means clustering and principal component analysis (PCA) will also be considered [62].

The use of AI ("software" or data-driven solutions) to extract information from amplification reactions has been relatively unexplored, particularly about the sigmoidal trend of PCR

reactions. The majority of the scientific community still relies on rudimentary data processing methods, as the PCR signal is mainly used for binary identification of the positive or negative presence of a single target nucleic acid. As a result, valuable information – present in most molecular platforms – that could be used to enhance PCR performance is discarded, compromising time, overall cost and patient outcomes. However, much progress has been made recently at the intersection of ML and molecular biology to leverage information in amplification and melting curves for accurately classifying multiple DNA targets in a single reaction. AI using amplification and melting curves or isothermal chemistries leads to improved molecular diagnostic assays without the need to change hardware or reaction chemistry. Therefore, the following sub-sections describe those advances with a particular focus on data-driven solutions to increase the multiplexing capabilities of diagnostic instruments.

2.5.1 Coupling Melting Curve Multiplexing with AI

The analytical process to extract information from melting curves influences has seen gradual improvement over the years which includes several steps: (i) background fluorescence subtraction and normalisation; (ii) curve overlay, a "temperature shifting" of curves that allows correction of minor temperature errors between samples and experiments; (iii) variant clustering, using hierarchical clustering algorithms; (iv) computation of dissimilarity plots, fluorescence subtraction in each variant cluster from the average fluorescence of a reference cluster; (v) computation of negative first derivative plots of normalised melting data using Savitzky–Golay polynomial estimation [63, 64].

In 2011 Dwight *et al* propose for the first time an algorithm to *in-silico* predict melting curves: uMELT [65]. The uMELT algorithm is a method for predicting the melting temperature (Tm) and melting curve shape of DNA or RNA sequences. It is a widely used and accurate method for predicting the effects of single nucleotide polymorphisms (SNPs), mutations, and other sequence variations on the Tm and melting curve of nucleic acids. The uMELT algorithm is based on the thermodynamics of DNA or RNA melting, where the temperature at which the double-stranded DNA or RNA molecule becomes single-stranded is called the melting temper-

ature (Tm). The Tm is affected by various factors such as sequence composition, length, GC content, and salt concentration. The uMELT algorithm takes into account all these factors and predicts the Tm and melting curve shape of a given nucleic acid sequence. The algorithm uses nearest-neighbor thermodynamic parameters to calculate the Tm and melting curve shape.

Classifying several melting curves to recognise specific nucleotide sequences in multi-target test benefits from using ML algorithms. Athamanolap *et al.* generated melting curves related to a fragment of the capsule polysaccharide synthesis (cps) gene locus of 92 serotypes of *Streptococcus pneumonia* in-silico. They trained an ensemble of linear kernel SVM algorithm, resulting in an average classification accuracy of 99.9% [65]. In-vitro validation of the algorithm was performed using sequence variants of a cancer-related gene, scoring 100% accuracy with three training data points per variant. In the following work, the same team generated an experimental library of melting curves of long amplicons (> 1000 bp) related to the 16S gene of 37 microorganisms. Training a nested SVM classifier, the group obtained high accuracy with bacterial isolates but a limited classification performance on clinical samples [66]. Lastly, Convolutional Neural Networks (CNN) were used to classify high resolution melting (HRMC) data converted into images through recurrence plots [67].

2.5.2 Coupling Amplification Curve Multiplexing with AI

As discussed before, the most commonly considered feature of a sigmoidal signal of the amplification reaction present is the C_t, the value at which the fluorescence of PCR products reaches a specific threshold, indicating a positive presence of the target nucleic acid. Another widely used feature of the amplification curve is Final Fluorescence Intensity (FFI). In 2019 Rajagopal *et al.* engineered the PCR endpoint signal intensity (or FFI) by changing the probe concentration to perform multiplex PCR with a single fluorescent [68]. The method's success resulted in the High Definition PCR (HDPCR) breakthrough and is now commercialised by the molecular diagnostics company ChromaCode (Carlsbad, USA). Similarly, Lee *et al.* have devised a technique called MuDT (Multiple Detection Temperatures), which enables the detection of two targets simultaneously in a single fluorescent channel by using only the amplification signal

[69]. MuDT relies on the Tagging Oligonucleotide Cleavage and Extension (TOCE) technique, where indirect temperature-dependent signals are generated at each cycle through two oligonucleotide probes, namely the Pitcher and the Catcher. By designing Extender and Catcher sequences with specific T_m, this fluorescence signal can be measured during a temperature-specific detection phase at the end of each annealing step, enabling real-time duplex detection and quantification (using the Ct value) as well as resolution of co-amplification events.

The previously described technique has great potential in the scalability of multiplexing. However, developing such tailored chemistries is not always achievable in a timely and cost-effective manner. In addition, using a single data point, such as FFI, reduces the value of the entire amplification curve signal. More features can be extracted from the sigmoidal trend, such as the background fluorescence (or the fluorescence at the start of the reaction) or the intersection of a line tangent to the curve at the first derivative with the baseline-subtracted signal level [15]. In 2019 Rodriguez-Manzano *et al.* simultaneously used those parameters to develop the Multi-dimensional Standard Curves (MSCs) in a 4plex single-channel PCR assay, leveraging multiple physical features of the reaction in a shared analytics framework and identifying non-specific amplification events (outlier removal) improving DNA quantification quality [16]. This work provided an affordable solution to maximise the amount of information extracted using a ML algorithm coupled with conventional PCR instruments, requiring minimum assay optimisation and no hardware modification. Moving in this direction, the work in this thesis aims to optimise such techniques further to extend the level of multiplexing, providing an effective way to develop and perform Data-driven Multiplexing.

2.6 Beyond qPCR: Isothermal Amplification

Apart from gold-standard PCR methods, DNA detection can also be performed without a thermocycler using isothermal chemistries as Loop-mediated isothermal amplification (LAMP), introduced in 2000 by Notomi *et al.* [70]. In contrast with PCR, LAMP relies on forming a dumbbell structure using a strand displacing DNA polymerase (e.g. *Bacillus stearothermophilus*

DNA polymerase or Bst), which has multiple sites for initiation of synthesis, resulting in exponential amplification. The reaction mix contains target DNA and four sets of primers defined as the forward outer primer (F3), backward outer primer (B3), forward inner primer (FIP, with F1c and F2 fragments) and backward inner primer (BIP, with B1c and B2 fragments) [70]. As Figure 2.7 shows, the reaction starts with strand invasion from the F2 part of the forward primer FIP, complementary to the F2c region, and the Bst initiates the synthesis. It is noticeable that the 5'-end of the FIP remains overhanging as it is complementary to the F1c region (it will be explained later in the loop structure). Now the forward outer primer F3 bind to the F3c region with the only purpose of dissociating the newly formed ssDNA created. This process is repeated using the BIP primer, and after strand dissociation, carried by the B3 primer, the first LAMP product (or amplicon) is formed. The reverse complementary F1 directly hybridise with the F1c sequence comprising a loop structure. The same happens with the B1 and B1c, creating the dumbbell structure. From now on, all primers can bind to generate more products. To enhance the efficiency of product formation and amplification speed, loop primers can be designed for the LAMP assay [71]. Finally, the structure that the LAMP reaction formed can be detected through fluorescence emission using intercalating dyes when the target DNA is present.

This nucleic acid amplification method offers a rapid, accurate, and easy-to-use diagnosis of infectious diseases, especially in limited-resource settings. For this reason, LAMP has become extremely popular as it requires simpler instrumentation [72]. One of the thesis objectives is to demonstrate the application of Data-driven methodology for multiplexing using isothermal methods such as LAMP.

2.7 Fundamentals of Digital PCR

Using data-driven approaches, especially when ML algorithms are involved, can represent a challenge if more than available data is needed to train ML models. Conventional qPCR instruments usually have low throughput as only a few reactions can be performed, and obtaining

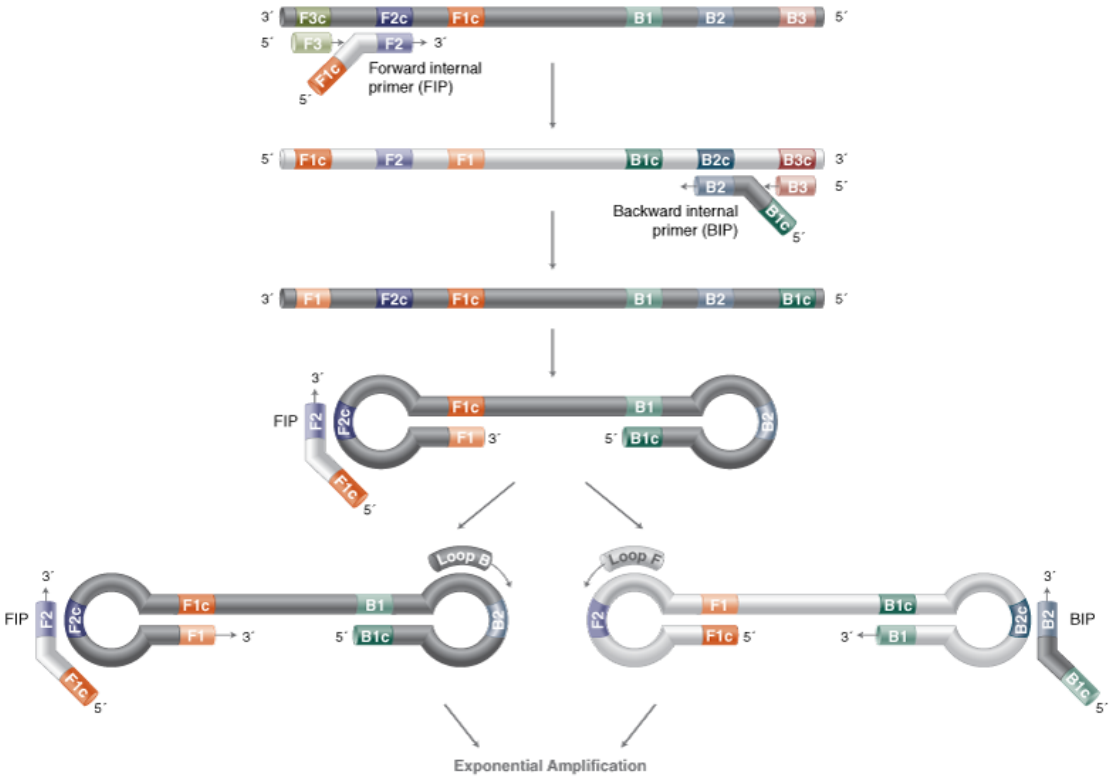


Figure 2.7: The Loop-Mediated Isothermal Amplification (LAMP) Reaction. LAMP uses 4-6 primers recognizing 6-8 distinct regions of target DNA. A strand-displacing DNA polymerase initiates synthesis and 2 of the primers form loop structures to facilitate subsequent rounds of amplification.

more data points is time and resource-consuming. Digital PCR allows the sample and the PCR reaction to be divided into numerous distinct PCR sub-reactions, with each partition containing either a few or no target sequences Figure 2.8. Moreover, dPCR enables absolute quantification of target nucleic acids by counting the single-molecule positive reactions, overcoming the limitations of qPCR [73, 74]. Amplification-positive wells are utilised to quantify the target sequence concentration using Poisson's statistics with a statistically determined precision [75, 76]. Sample partitioning, it turns out, effectively concentrates the target sequences within the isolated microreactors. Because of the concentration effect, template competition is reduced, allowing rare mutations to be detected in the background of wild-type sequences. It may also qualify for a more robust tolerance to inhibitors found in food [21].

It is important to note that the partitions can be created using a number of different mechanisms, such as emulsified microdroplets suspended in oil (droplet digital PCR, ddPCR), microwells, or microfluidic valving [77]. Amplification of target sequences can be detected by

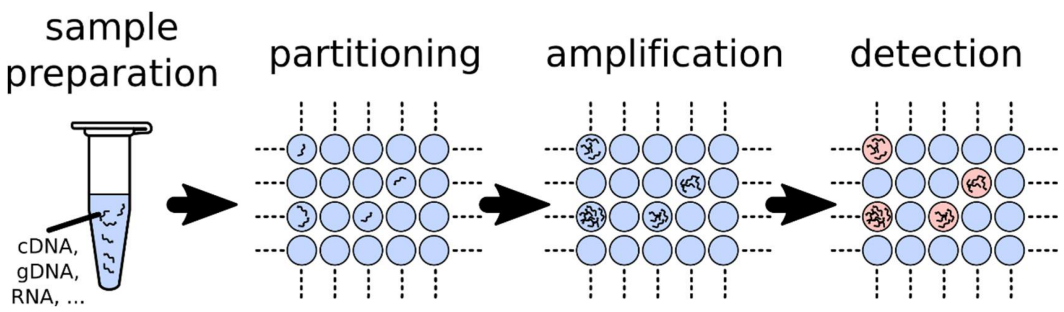


Figure 2.8: Principles of digital PCR. The sample is divided into many independent partitions such that each contains either a few or no target sequences. The distribution of target sequences in the partitions can be approximated with a Poisson's distribution. Each partition acts as an individual PCR microreactor, and partitions containing amplified target sequences are detected by fluorescence. The ratio of positive partitions (presence of fluorescence) over the total number allows for determining the concentration of the target in the sample [21].

endpoint fluorescence, and some machines also provide real-time fluorescence data. Unlike qPCR, dPCR back-calculates the target concentration using the number of positive partitions over the total available. This means that, unlike qPCR, dPCR does not need a calibration curve for sample quantification. The underlying accuracy and performance characteristics of dPCR quantification are formally defined by binomial statistics [78]. Typically, the confidence interval is derived using functions that may be calculated immediately. These forecasts are based on assumptions that have direct implications for the estimates. To calculate the probability p of a partition containing at least one target sequence, the case of the random distribution of m molecules into n partitions has to be considered. This situation is equivalent to a binomial behaviour where the outcome of each drawing can be present or absent (m times):

- The chance of a target sequence being present in a partition is $\frac{1}{n}$ because it results from random or independent events.
- The probability p is the complementary chance of the partition to be empty after the m target sequences are distributed.
- A partition has m chances to receive one target sequence.
- The possibility for a partition to be empty is then $1 - \frac{1}{n}$ after one draw and $\left(1 - \frac{1}{n}\right)^m$ after m attempts.

- The probability is then equal to $p = 1 - \left(1 - \frac{1}{n}\right)^m$
- When n is large (then $\frac{1}{n}$ is very small), the probability of p can be approximated to $p = 1 - e^{-\lambda}$ where $\lambda = \frac{m}{n}$

This formula defines the probability function of a Poisson distribution with the parameter lambda. When the average number of occurrences (λ) is known, the Poisson distribution explains the probability distribution of independent events. The fraction of partitions containing a given number of target sequences is predicted by the Poisson distribution. On the other hand, knowing the distribution allows you to calculate the average number of target sequences in the sample. Nonetheless, the ratio of positive partitions k (including some target sequences) to total partitions n is sufficient to forecast the target sequence's starting concentration in the sample with:

$$\lambda = \ln\left(1 - \frac{k}{n}\right) \tag{2.1}$$

The confidence interval in estimating the target concentration depends on the number of empty partitions. It is typically assessed using functions that can be directly calculated, for example, using Wald or Wilson methods [79, 80]. Those estimations suggest that there is a value of λ for which the initial template concentration can be estimated with the highest confidence. In cases of 10,000 or more partitions, the maximal confidence is obtained for a λ value of about 1.6, which corresponds to a proportion of 20% of empty partitions Figure 2.9. As noted previously, the precision is poor for low values of λ, reaching an optimal for a λ of 1.6 before slowly declining with increasing values of λ, which corresponds to a saturation of the partitions. The accuracy of the estimation of λ rises with the number of partitions, and the optimal precision (at $\lambda = 1.6$) scales as the inverse square root of the number of partitions.

All these assumptions drive us to explore the potential of dPCR as a research method for "digital" biology and chemistry where detecting single biological entities such as molecules are made possible. Spatial compartmentalisation, which entails splitting a solution or suspension of entities into various subunits, plays a crucial function in this context. Low ratios of biological

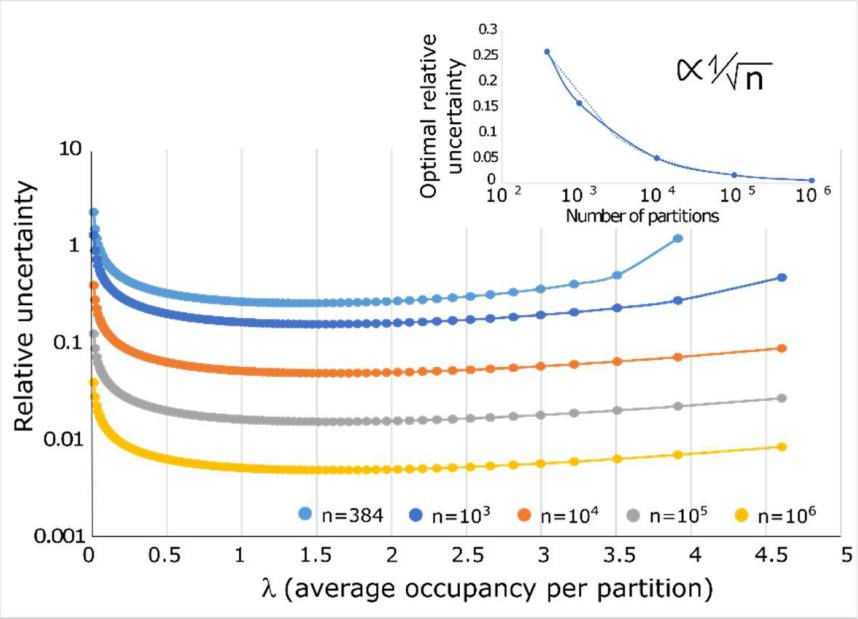


Figure 2.9: Quantification accuracy of dPCR. The precision of dPCR is non-uniform and depends on the average occupancy of the target sequence per partition. The precision of dPCR also increases with an increasing number of partitions (distinct colours). The inset shows that the evolution of the relative uncertainty (taken at $\lambda \approx 1.6$) decays as an inverted square root of the number of partitions [21].

entities to reaction compartments allow single entities to be captured per compartment [10]. Poisson statistics are often helpful in this regard. Microfluidics is essential for constructing and manipulating small fluidic chambers that hold a single biological sample. Microfluidic large-scale integration (LSI) is a technology broadly used for studying single cells and molecules, and it is currently used in digital PCR instruments [81, 82]. Using multilayer soft-lithography, it uses pneumatic valves that are monolithically generated in the silicone elastomer polydimethylsiloxane (PDMS). The compartmentalisation of a sample is achieved by an array of binary valve patterns that, when closed, can partition a network of microfluidic channels into several sections [21, 83].

For dPCR, it is possible to achieve both singleplex and multiplex assaying, as each target-containing well proceeds with the amplification process only with their specific primers, and no reaction will occur in wells without targeted DNA molecules if the single molecule occupancy is maintained. This allows high and low-abundance targets to be evaluated in a single experiment without being concerned about highly concentrated targets "swamping out" the lower ones

[73, 83].

2.8 Chapter Summary and Reflection

In summary, this Chapter reviewed the basics of DNA detection with a significant focus on PCR. The knowledge and understanding of the biological process and bioinformatics tools to design optimal PCR assay is the arsenal needed to proceed to more complex assay designs such as multiplex PCR and isothermal (LAMP). In section 2.5, Machine Learning concepts have also been introduced, which are the cardines of the data-driven multiplexing method exposed in this thesis. The literature review in paragraphs 2.5.1 and 2.5.2 highlights the gaps regarding PCR data usage to leverage the information encoded in the amplification and melting curve. Integration machine learning algorithm it is crucial to fill this gap and optimise the use of data to enhance diagnostic capabilities of state-of-the-art PCR or Point-of-Care instruments.

Part I:

The Data-driven Multiplexing Discovery

"Focus on the journey not the destination. Joy is found not in finishing an activity but in doing it."

GREG ANDERSON

Chapter 3

Single-well & Single-channel Data-Driven Multiplexing

3.1 Chapter Overview

Polymerase Chain Reaction (PCR) has been used as the gold standard to identify the presence or absence of a specific target nucleic acid. However, this binary usage of the PCR (positive/negative) neglects the true potential of this powerful technique by discarding the kinetic information contained in this sigmoidal signal. This Chapter demonstrates that the large volume of raw data obtained from real-time digital PCR (qdPCR) 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), by using an intercalating dye (EvaGreen), reduces the cost and complexity of the assay, and enables the use of melting curve analysis for validation. As a case study, three carbapenem-resistant genes are multiplexed in a single reaction, targeting global challenges such as antimicrobial resistance. In the presence of single targets, a classification accuracy of 99.1% ($N = 16,188$) is reported, representing 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 92.9%. To support the analysis, a formula to estimate co-amplification occurrence in dPCR based on multivariate Poisson statistics is derived, suggesting that reducing the dPCR occupancy improves the

digital count when multiple targets in the same digital panel are present. The ACA approach takes a step towards maximising 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, combining this method with existing probe-based assays will increase multiplexing capabilities significantly and facilitate the implementation of amplification chemistries outside the lab.

The concepts in this Chapter resulted in the following journal article and patent application:

- Moniri A*, Miglietta L*, Malpartida-Cardenas K, Pennisi I, Moser N, Holmes A, Georgiou P, Rodriguez-Manzano J. "Amplification Curve Analysis: Data-Driven Multiplexing Using Real-Time Digital PCR." *ACS Analytical Chemistry*, 2020 Oct 6;92(19):13134-13143. *First joint authorship.
- Rodriguez-Manzano J, Moniri A, Miglietta L and Georgiou P. "Identifying a target nucleic acid", WO2022038279A1, Assignee: Imperial Innovations Limited, 2020.

3.2 Introduction

Digital PCR (dPCR) is a well-established method to detect and quantify nucleic acid [73, 84]. 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: (i) lack of references or standards; (ii) high precision in quantification; (iii) tolerance to inhibitors; and (iv) the capability to analyse complex mixtures [21, 85, 86]. Therefore, dPCR has enabled scientific breakthroughs in cancer biomarker discovery, genetic alterations and infectious diseases, among others [87, 88, 89].

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. Currently, most multiplex dPCR assays rely on the use of fluorescent probes (e.g. TaqMan), such that the probe concentration can be optimised to distinguish between the targets using the final fluorescent intensity (FFI) [68]. However, probes are expensive and require time-consuming optimization [90]. In an effort to achieve similar multiplexing capabilities, dye-based approaches (e.g. EvaGreen) have also been proposed which alter primer concentration

in order to change the PCR efficiency and impact the FFI [91]. The aforementioned methods require extensive optimization to achieve accurate multiplexing without compromising assay performance. Current methods rely on the FFI only and there has been no report of alternative methods such as analysing the entire amplification curve in real-time 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 [16, 15]. 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, machine learning models and multiplex qdPCR outputs are combined , to prove that sufficient kinetic information exists in the amplification curve to perform data-driven multiplexing - referred to as amplification curve analysis (ACA). Melting curve analysis (MCA) were used as the 'gold standard' method to assess the performance of the proposed approach, as illustrated in the experimental workflow depicted in Figure 3.1. Taking advantage of the large volume of raw data extracted from real-time dPCR ($> 10^4$ reactions per experiment) and the high likelihood of single-molecule events, a machine learning model is developed without explicitly extracting features of the amplification curve or compromising the assay performance (by modifying probe or primer concentration). Moreover, normalisation of the FFI is performed, showing that this method can be combined with current approaches for dPCR multiplexing - breaking the barrier of one target for each level of FFI (in a given fluorescent channel). Finally, a theoretical derivation for the likelihood of multiple targets in a single well (i.e. co-amplification) is provided to understand the effect of this phenomenon on quantification and multiplexing.

As a clinically relevant application, this methodology is applied to the global challenge of antimicrobial resistance [92]. In particular, the carbapenemases are the focus. Carbapenemases are β-lactamases (*bla*) that are resistant to the carbapenems, a class of highly effective antibiotic agents [93]. Therefore, a multiplex assay is developed for the detection of three common carbapenem-resistant genes, namely bla_{NDM}, bla_{VIM} and bla_{KPC}.

The vision for this work is three-fold: (1) maximise the capabilities of existing instruments and chemistries by extracting more information from the data that already exists; (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

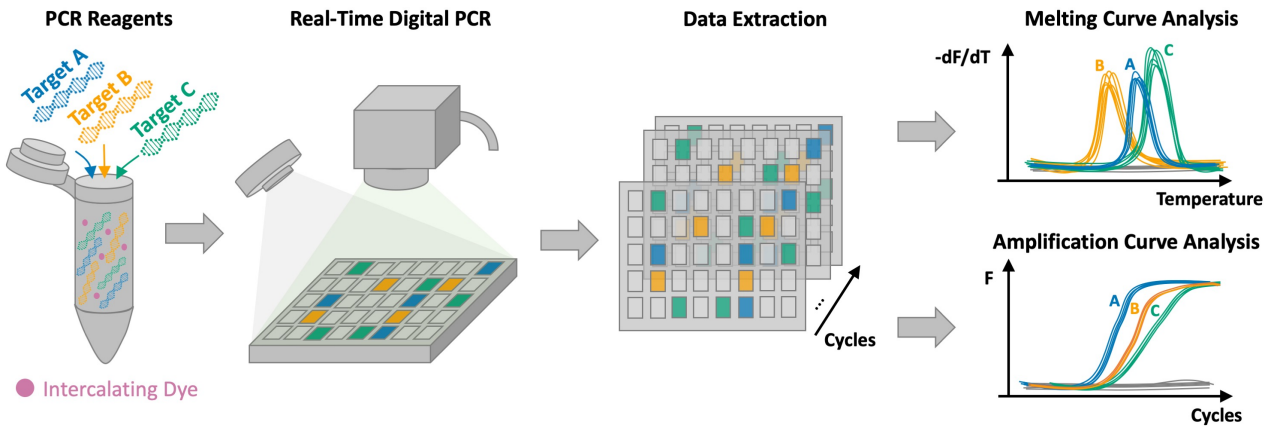


Figure 3.1: Experimental workflow. A multiplex PCR assay (with an intercalating dye) is developed for detecting targets A, B and C. Real-time digital PCR is used to perform single-molecule amplification to detect the targets. Melting curve analysis is used to validate the specificity of the amplification 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.

technologies to facilitate the implementation of dPCR outside of the lab.

3.3 Experimental Section

3.3.1 DNA Templates

Double-stranded synthetic DNA (gBlock^{TM} Gene fragments) containing bla_{NDM}, bla_{VIM} and bla_{KPC} gene sequences (ranging from 801 to 917 bp) is used. The sequences of these genes were downloaded from the NCBI 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 (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).

Table 3.1: 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 [97].

3.3.2 PCR Primer Design

Primers for the multiplex assay were designed to target the aforementioned referenced sequences. For each gene of interest, 1,000 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 [94], in Geneious Prime® 2020.1.2 [95]. Primer characteristics were analysed through the IDT OligoAnalyzer software using the J.SantaLucia thermodynamic table for melting temperature (T_m) evaluation, hairpin, self-dimer and cross-primer formation [96]. The T_m of the amplification product of each primer set was determined by the Melting Curve Predictions Software (uMELT) package [65]. All primers were synthesised by Life Technologies (ThermoFisher Scientific). Primer sequences are listed in Table 3.1.

3.3.3 PCR Reaction Conditions

Real-time PCR. Each amplification reaction was performed in 10 µL of final volume with 5 µL of SsoFast EvaGreen Supermix with Low ROX (6-Carboxyl-X-Rhodamine) (BioRad, UK), 3 µL of PCR grade water, 1 µL of 10× multiplex PCR primer mixture containing the three primer sets (5 µM of each primer), and 1 µL of different concentrations of synthetic DNA. PCR amplifications consisted of 10 min at 95 °C, followed by 45 cycles at 95 °C for 20s, 65 °C for 45s, and 72 °C for 30s. In order to validate the proposed method, the results were compared against melting curve analysis. One melting cycle was performed at 95 °C for 10s, 65 °C for 60s, and 97 °C for 1 s (continuous reading from 65 to 97 °C). Each experimental condition was run in triplicates, loading the reactions into a 96-well plate using a Light Cycler 96 Real-Time PCR

System (Roche Diagnostics, Germany). Moreover, negative and positive controls were included in each experiment.

Real-time Digital PCR (qdPCR). Each amplification reaction was performed in 4 µL of final volume with 2 µL of SsoFast EvaGreen 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 (0.25 µM of each primer), and 1.2 µL of different concentrations of synthetic DNA. 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 melting curve analysis. One melting cycle was performed at 65 °C for 3s and continuous reading from 65 to 97 °C with an increment of 0.5 °C every 3s. The reactions are loaded into Juno or FC1™ cycler or Biomark HD/Biomark (Fluidigm Corporation, South San Francisco, California, United States) using the qdPCR 37K™ integrated fluidic circuit (IFC) provided by the same company. Moreover, negative and positive controls were included in each.

3.3.4 Data Analysis

Multiple in-house Python (v3.7) scripts were developed to extract and analyse the data using standard data science packages including: 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 optimised for colour blindness[98].

3.4 Results & Discussion

In this Chapter, 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 visualised 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.

3.4.1 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 analyse 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 [68, 99].

First, a 3plex assay for the detection of bla_{NDM}, bla_{VIM} and bla_{KPC} is developed. Figure 3.2A shows the amplification curves and melting peaks for each target at concentrations ranging from 5×10^3 to 1×10^6 copies/reaction. Observe that the melting peaks for bla_{NDM}, bla_{VIM} and bla_{KPC} can be distinguished from each other and are given as 84.7°C, 88.5°C and 89.7°C respectively. Moreover, Figure 3.2B shows the corresponding standard curves illustrating the C_t 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. T_m, 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) [100]. 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 3.2C shows the amplification curves for the co-presence of targets and Figure 3.2D shows the corresponding melting curves. It can be observed that the bla_{NDM}+bla_{KPC} and bla_{NDM}+bla_{VIM} peaks are sufficiently different in order to identify two distinct peaks in the melting profile. However,

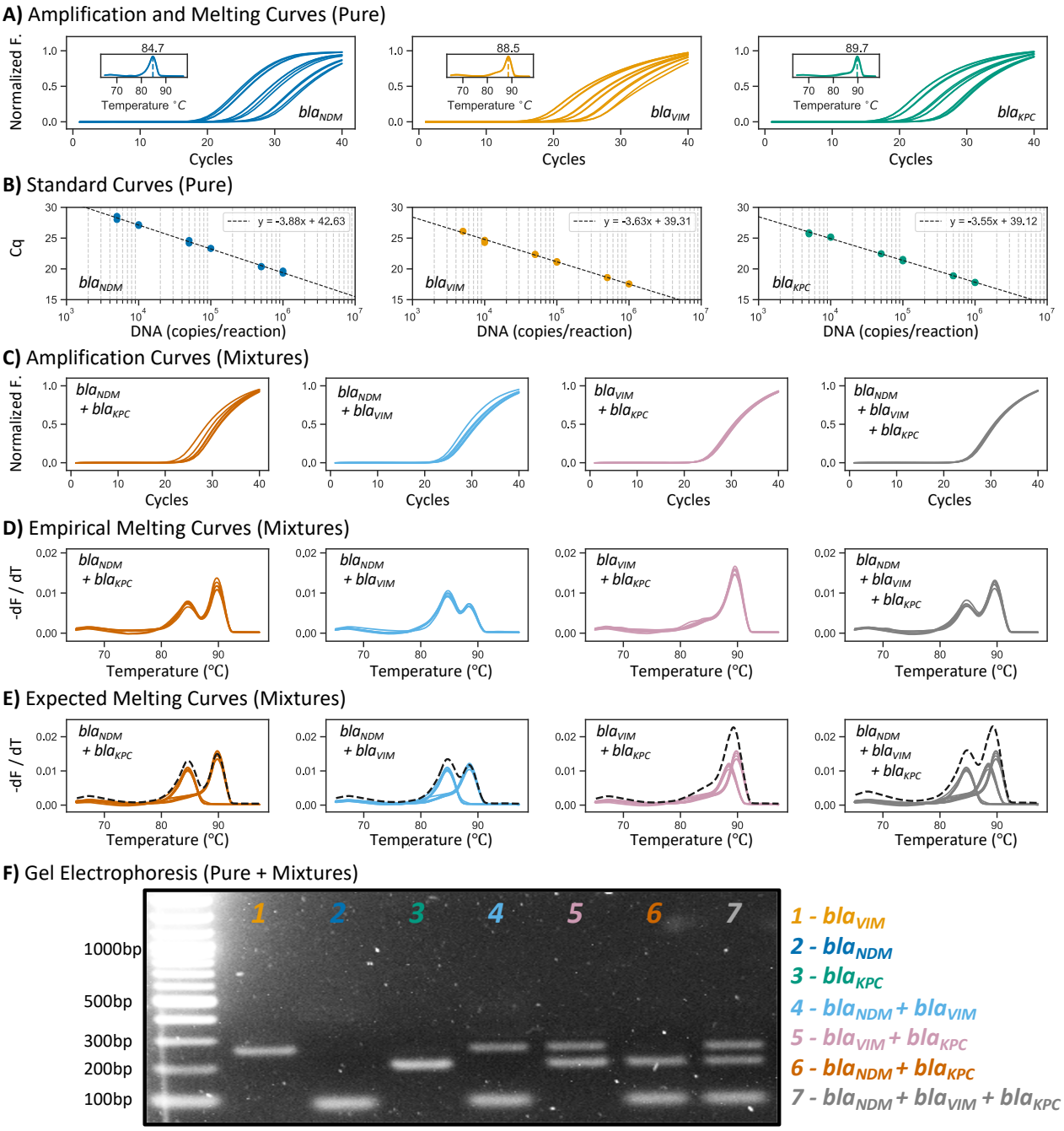


Figure 3.2: Real-time PCR Experiments showing the performance of a 3plex assay in the presence of single and multiple targets. (A) Amplification curves for single targets (in a single reaction) with corresponding melting curves, where six different dilutions were used ranging from 5×10^3 to 1×10^6 copies/reaction. (B) Corresponding standard curves correlating the C_t 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.

the mixture containing bla_{VIM}+bla_{KPC} results in only a single peak. This is also observed in the mixture with all three targets as only two peaks are evident. This may suggest there are fewer amplification products. Through adding the pure melting profiles, the 'expected' melting curve for mixtures of products is estimated, as in Figure 3.2E. 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, it must be run post PCR analysis techniques such as gel electrophoresis or sequencing. Figure 3.2F 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 [101].

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 [16, 15]. 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.

3.4.2 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, the translation from the 3plex assay in qPCR to qdPCR is performed. First, the multiplex assay in the presence of pure targets in each digital panel is investigated. Figure 3.3A and 3.3B show the digital pattern and amplification curves for a serial dilution of the

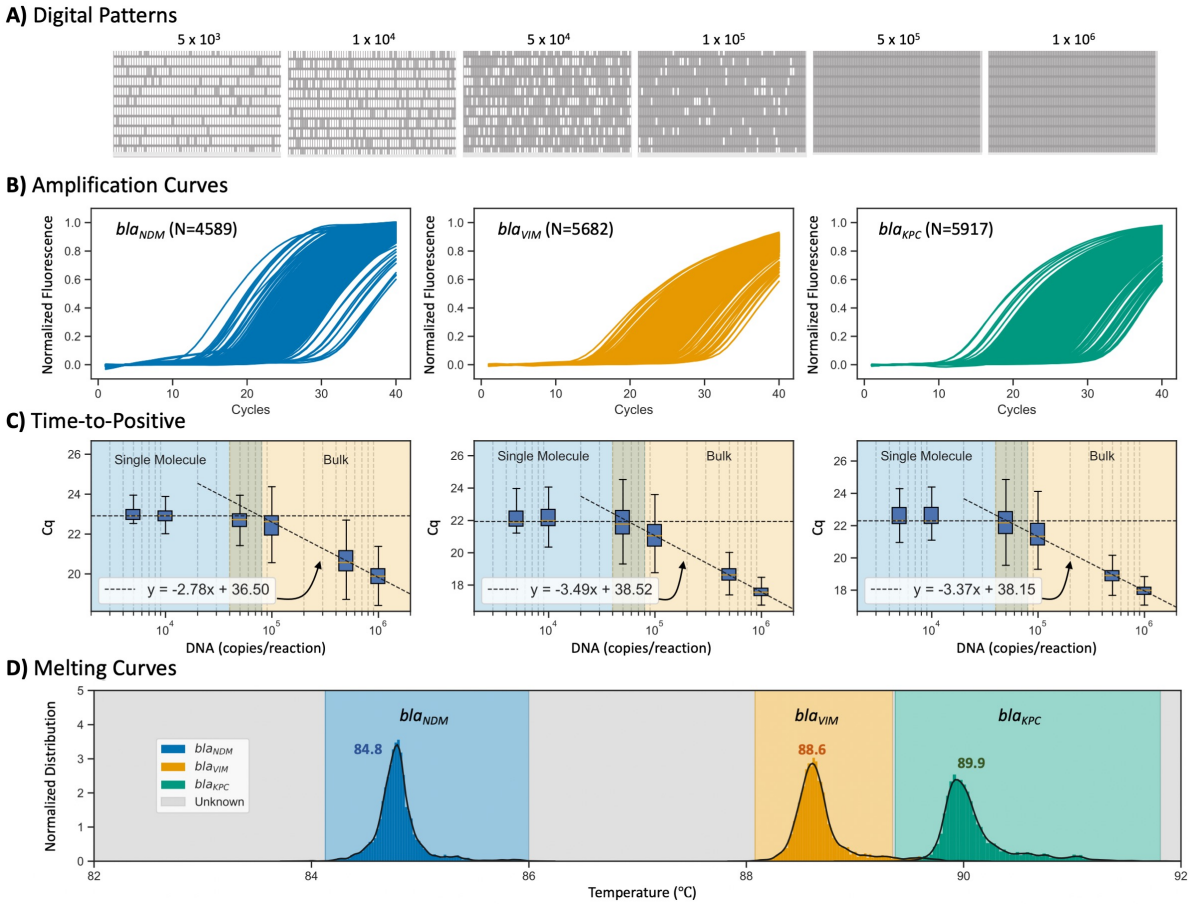


Figure 3.3: Real-time dPCR data. (A) Digital patterns for each panel at increasing concentrations. (B) Amplification curves for serial dilution of each target at concentrations ranging from 5×10^3 to 1×10^6 copies/reaction. (C) Standard curves correlating the C_t values with the concentration of each 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) Normalised distribution of the melting curve peaks, i.e. T_m, for each target.

targets. Concentrations ranging from 5×10^3 to 1×10^6 copies/reaction were chosen such that amplification events in both single-molecule and bulk regions are observed, capturing the 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 C_t values as a function of the target concentration as seen in Figure 3.3C 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 that can be quantified using Binomial and Poisson statistics [21]. 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[21, 85, 86]. Figure 3.3D shows the distribution of the melting curve peaks (T_m) for each target. The maximum likelihood value of T_m for bla_{NDM}, bla_{VIM} and bla_{KPC} is 84.8°C, 88.6°C and 89.9°C, respectively. All T_m 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, the bounds for distinguishing the targets are determined by considering the 1^{st} and 99^{th} 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.

3.4.3 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 [99]. Figure 3.4A shows the raw amplification curves with background subtraction. The associated FFI for each amplification event is shown in Figure 3.4B. 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 (based on 10-fold cross-validation) is computed as 79.42%, which is not adequate for many applications. In particular, the confusion matrix demonstrating the predictions is given in Figure 3.4B and details of the one-vs-one classifiers which are combined to form the multi-class model are given in Table 3.2. 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 which compromises the entire 3plex. This demonstrates the challenge of scaling up the FFI method to three or higher targets due to the large variation of FFI values.

Table 3.2: 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

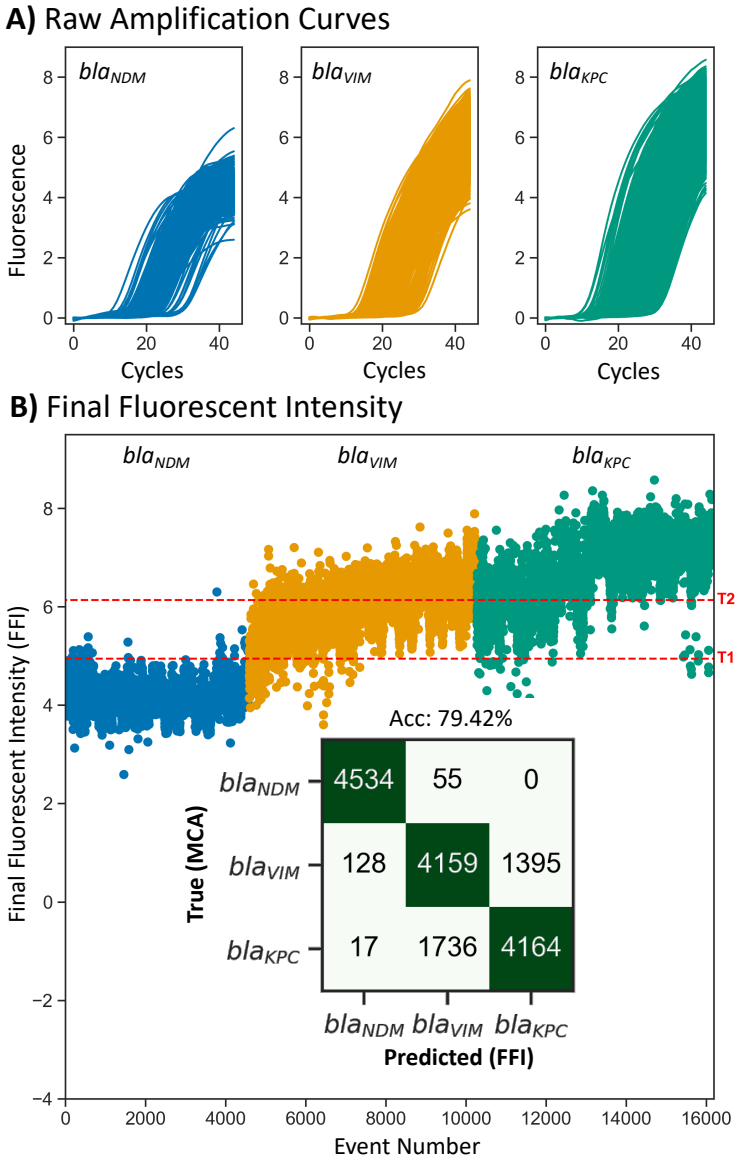


Figure 3.4: Multiplexing based on final fluorescent intensity. (A) Raw amplification curves with background subtraction. (B) Distribution of FFI values across the 3 targets. The red dashed lines (T1 and T2) indicate the thresholds generated from a Logistic Regression method for target classification. The predictions are shown in the overlay confusion matrix.

3.4.4 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, this

process required 1.7Gb extra memory (at just 0.5°C resolution), more time for acquisition & processing, and cannot be extended to chemistries which are not compatible with MCA such as TaqMan assays or pH-based sensing. Therefore, in this study, a new method of multiplexing through the use of machine learning is explored, leveraging specific kinetic information extracted directly from the amplifications curve. First, unsupervised machine learning is used to visualise 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 [102]. It is important to understand that t-SNE is an unsupervised learning algorithm and therefore does not use the target labels. Figure 3.5 illustrates the t-SNE algorithm (perplexity=500) applied to the amplification curves with each target coloured 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, it is demonstrated that even after normalising 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.

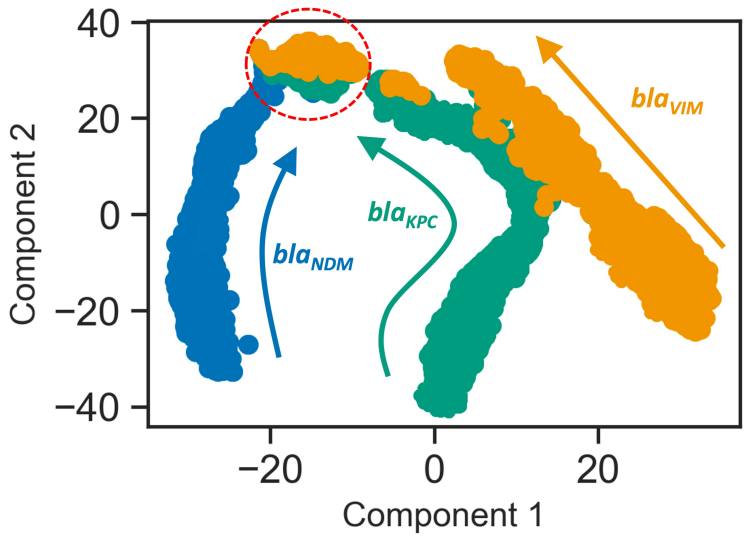


Figure 3.5: Visualising the similarity between amplification curves using the t-distributed stochastic neighbor embedding algorithm with 2 components. Direction of arrows indicate high to low concentration. Dashed red circle indicates curves that have not reached plateau.

3.4.5 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 (KNN), support vector machines and deep neural networks [103, 104, 105]. The following section is demonstrated using the k-nearest neighbors algorithm which is a non-parametric method that is intuitive [106]. 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 3.3 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 3.6A shows the prediction accuracy in a confusion matrix for the KNN algorithm (for number of neighbors $k = 10$). The dark green squares indicate the single-target true positive classifications. The overall classification accuracy was 99.1% (CI: 99.08-99.09%). Moreover, the accuracy, sensitivity and specificity for the one-vs-one classifiers are given in Table 3.3. 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.

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

Table 3.3: 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

reducing the number of experiments. Figure 3.6B 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 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, previous data is used as the training samples and run a different experiment which contains all possible combinations of the targets.

Figure 3.6C 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 co-amplifications is expected 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 3.6D. 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 accuracy of 95.0%. Figure 3.6E 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).

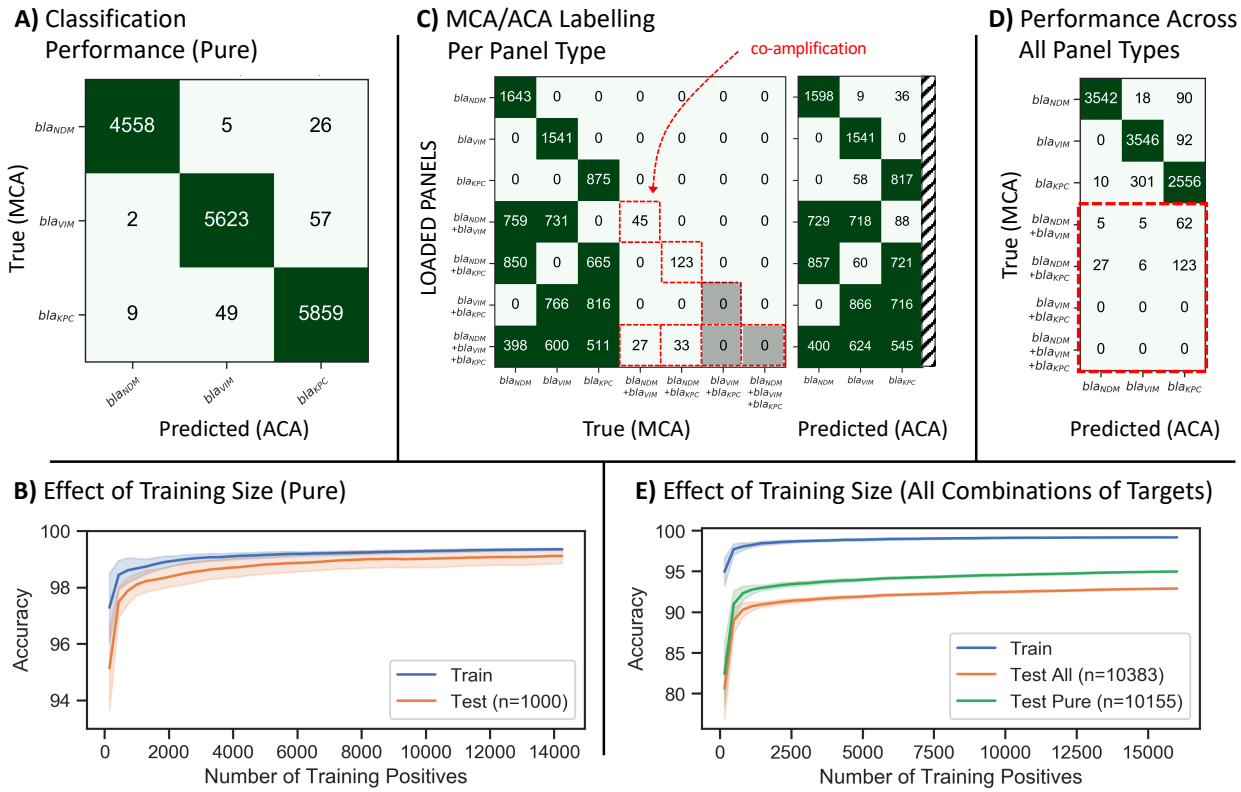


Figure 3.6: Performance of ACA in the presence of single and multiple targets. (A) Confusion matrix showing the predictions of ACA compared with MCA in the presence of single targets; (B) The effect of training size (with pure targets) on the train/test performance; (C) Matrices displaying the prediction 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.

3.4.6 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[21]. 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 3.7A 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$^{\text{TM}}$ chip), the uncertainty is below 5% between 16.7% occupancy ($\lambda = 0.2$, marked with a circle) and 99.3% occupancy ($\lambda = 5.0$,

marked with a square).

Here, the Poisson statistics is extended 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:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.1}$$

$$\lambda = \frac{n}{m} \tag{3.2}$$

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

$$p(\mathbf{k}) = p(k_1) \ldots p(k_K) \tag{3.3}$$

$$= \prod_{i=1}^{K} p(k_i) \tag{3.4}$$

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

$$P_0 = \prod_{i=1}^{K} p(k_i = 0) \tag{3.5}$$

$$P_1 = \sum_{j=1}^{K} \frac{p(k_j > 0)}{p(k_j = 0)} \prod_{q=1}^{K} p(k_q = 0) \tag{3.6}$$

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

$$P_C = 1 - \underbrace{\prod_{i=1}^{K} e^{-\lambda_i}}_{P_0} - \underbrace{\sum_{j=1}^{K} (e^{\lambda_j} - 1) \prod_{q=1}^{K} e^{-\lambda_q}}_{P_1} \tag{3.7}$$

which can be simplified to

$$P_C = 1 - e^{-\lambda}\big(1 + \sum_{j=1}^{K}(e^{\lambda_j} - 1)\big) \tag{3.8}$$

$$\text{where} \quad \lambda = \sum_{i=1}^{K} \lambda_i \tag{3.9}$$

Using this formula, the theoretical error introduced by co-amplifications can be estimated. 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 3.7B shows the proportion of co-amplification events for two targets (A and B) as a function of λ_A and λ_B. 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 3.7C shows the worst-case co-amplification proportion as a function of the number of targets. For three targets, the P_C is reduced from 37.1% down to 1.2% by decreasing 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,

$$\lim_{K \to \infty} P_C = 1 - e^{-\lambda} - \lambda e^{-\lambda} \tag{3.10}$$

Contrary to single target Poisson quantification, to maximise ACA multiplexing performance, λ should be decreased without compromising quantification significantly.

3.5 Conclusion

No previous 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, a new data-driven method using supervised machine learning, referred to as amplification curve analysis or ACA, is reported. The validation

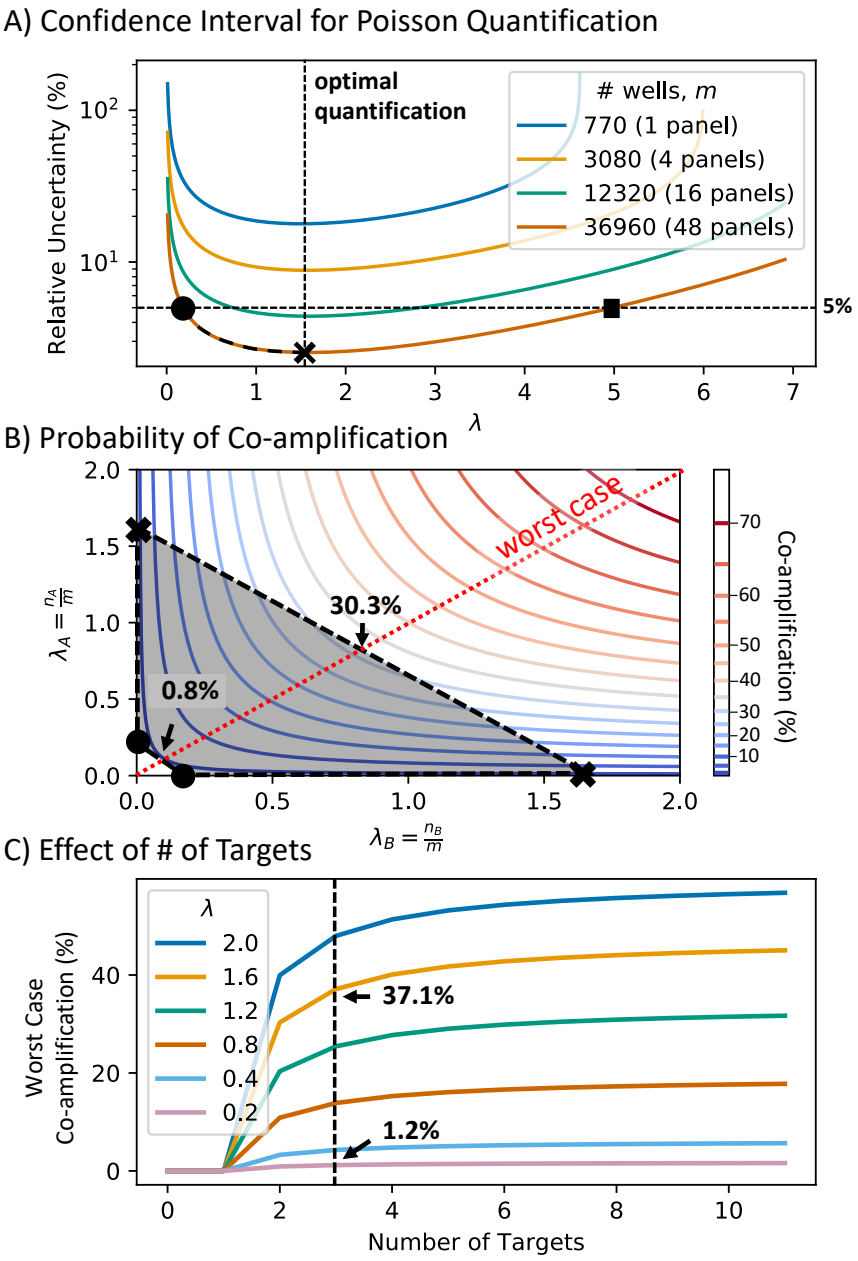


Figure 3.7: The impact of co-amplification events. (A) The relative uncertainty of Poisson quantification as a function of λ and the number of wells. (B) The probability of co-amplification in the presence of 2 targets. (C) The effect of the number of targets on the worst-case probability of co-amplification.

of this approach is performed through detection of three 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, an accuracy of 95.0% is observed. However, this promising performance was reduced to 92.9% due to the presence of co-amplification in a single well. To support the analysis, a formula to estimate the occurrence of co-amplification is derived, suggesting reducing the digital occupancy in the case of multiple targets in the same digital panel.

CHAPTER LESSON

This Chapter showed a novel artificial intelligence method to multiplex in single-well and single-channel, suggesting that the entire amplification curve contains more information than commonly presumed. The method is cross-validated with melting curve analysis (MCA) showing high concordance with the Amplification Curve Analysis (ACA). The implications of this method motivate further research in maximising the value of nucleic acid amplification data, by uniquely merging molecular biology and data science.

TAKEAWAY QUESTION

"Can the level of multiplexing (or the number of targets multiplexed in a single-well reaction) be further increased using these data-driven approaches?"

Chapter 4

High-level Multiplexing using Artificial Intelligence

4.1 Chapter Overview

The previous Chapter explored how to better use the data from PCR signals to perform data-driven multiplexing in single-well and single-channel reactions. This diagnostic solution shows excellent potential for multiple target detection rapidly and cost-effectively. However, when it is required to identify a higher number of nucleic acids in a single reaction, seeking new features of the amplification event is needed. Here, an expansion of the previous Amplification Curve Analysis (ACA) method is proposed, referred to as Amplification and Melting Curve Analysis (AMCA), which besides leveraging the kinetic information in real-time amplification data, also accounts for the thermodynamic melting profile. 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, a new 9plex assay is developed to detect nine mobilised colistin resistant (*mcr*) genes as clinically relevant targets for antimicrobial resistance. Over 100,000 amplification events have been analysed, 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.

The concepts in this Chapter resulted in the following journal article and patent application:

- Moniri A*, Miglietta L*, Holmes A, Georgiou P, Rodriguez-Manzano J. "High-Level Multiplexing in Digital PCR with Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis." *ACS Analytical Chemistry*, 2020 Oct 20;92(20):14181-14188. *First joint authorship.

- Rodriguez-Manzano J, Moniri A, Miglietta L and Georgiou P. "Identifying a target nucleic acid", WO2022038279A1, Assignee: Imperial Innovations Limited, 2020.

4.2 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 [107, 108, 109, 15, 110, 111, 112]. 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 analyse complex mixtures [21, 85, 86, 113]. Therefore, dPCR has enabled scientific breakthroughs in clinical microbiology, gene expression and precision cancer research, among others [114, 89, 115].

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 [16, 116]. Current approaches based on fluorescent probes are expensive and require lengthy optimization which is challenging for high-throughput applications [55, 117]. Intercalating 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 analyses such as gel electrophoresis, melting curve analysis or sequencing methods.

Current multiplex dPCR methods that are dependent on intercalating dyes are either limited to analysing real-time amplification data or performing melting curve analysis, since gel electrophoresis or sequencing is not possible [99, 118]. 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 [117]. Reported studies showed

that specific target identification could be achieved through adjusting primer concentration to modulate the FFI value [99]. 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 [97]. 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 [100]. Similar to ACA, high-level multiplexing with MCA also requires complex assay design to distinguish between close melting curve peaks [97].

Although the ACA and MCA methods are analysing 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) [119]. Moreover, some studies have combined dPCR and melting curves, although they are restricted to end-point PCR which does not encode kinetic information [120, 121]. To date, there has been no report of enhancing multiplexing capabilities by combining amplification and melting curves.

In this Chapter, this concept was explored 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 4.1. 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 [122]. In particular, colistin is a "last-line" antibiotic, reserved for the treatment

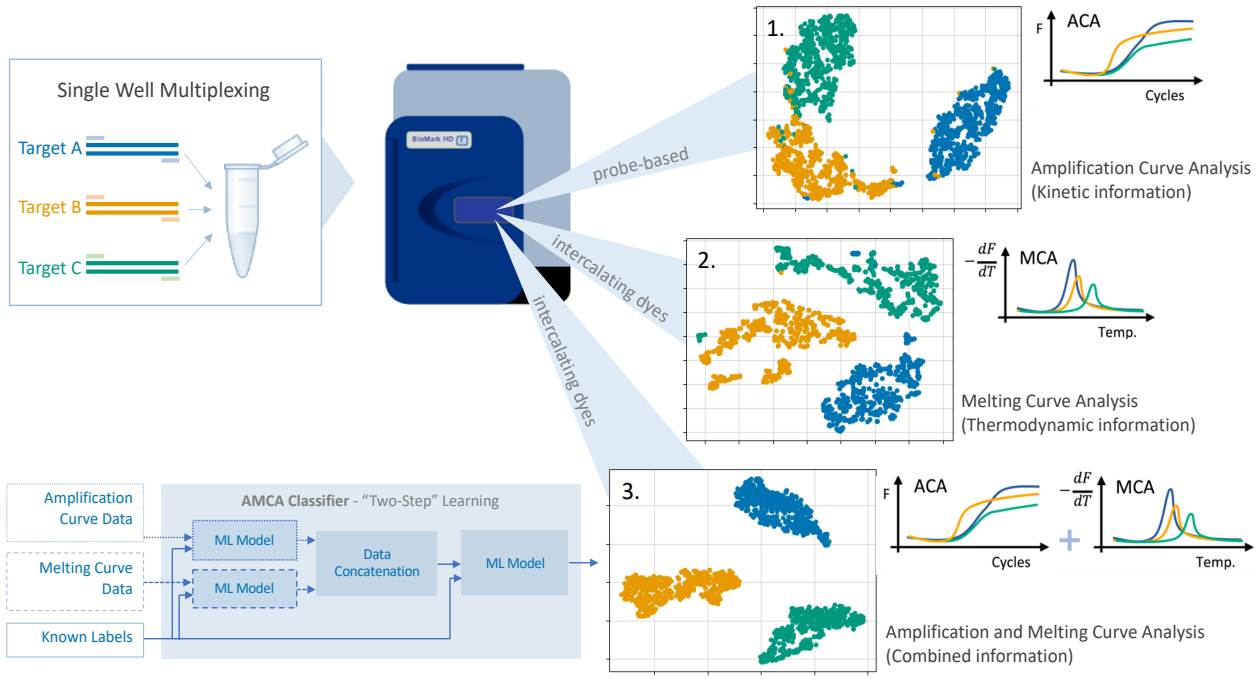


Figure 4.1: Concept of the proposed method. Amplification and melting curve data from real-time dPCR instrument (e.g. Fluidigm BioMark HD) is extracted. Subsequently, machine learning models are trained to classify multiple targets for both datasets individually. For high-level multiplexing, combining both methods can provide higher accuracy. Therefore, 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.

of severe bacterial infections. The rise of mobilised colistin resistance (*mcr*) has been reported in over 40 countries across five different continents [123, 124, 125]. Colistin resistant genes are often co-localised 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 [126, 127]. Incorrect diagnosis delays appropriate intervention, increases financial burdens for the healthcare system and complicates antimicrobial stewardship efforts [128]. Therefore, detecting variants of *mcr* is important to help treat and understand this emerging antimicrobial resistance. In this study, the first 9plex PCR assay to detect *mcr*-1 to *mcr*-9 in a single-well and single-channel was developed.

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.

4.3 Experimental Section

4.3.1 DNA Templates

Double-stranded synthetic DNA (gBlock^{TM} 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 4.1. The gBlocks^{TM} 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).

4.3.2 Multiplex Primer Design

To perform the (*in-silico*) design for the 9plex, 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 1,000 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 [94], in Geneious Prime® 2020.1.2 [95]. Primer characteristics were analysed through the IDT OligoAnalyzer software using the J. SantaLucia thermodynamic table for melting temperature (T_m) evaluation [96]. 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 [129]. The T_m of the amplification product of each primer set was determined by the Melting Curve Predictions Software (uMELT) package [65]. All primers were synthesised by Life Technologies (ThermoFisher Scientific). Primer sequences, amplicon length and GC content of the product are listed in Table 4.1.

4.3.3 PCR Reaction Condition

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

Table 4.1: 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 [130].

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 minutes at 95 °C, followed by 45 cycles at 95 °C for 20 seconds, 66 °C for 45 seconds, and 72 °C for 30 seconds. Melting curve analysis was performed with one cycle at 65 °C for 3 seconds and reading from 65 to 97 °C with an increment of 0.5 °C. The integrated fluidic circuit controller is used to prime and load qdPCR 37K$^{\text{TM}}$ digital chips. The Fluidigm's Biomark HD system performs the dPCR experiments. Each digital chip contains 48 inlets, where each inlet is connected to a panel consisting of 770 wells (0.85nL well volume) [131]. In this study, three digital chips were used, totalling 144 panels (110,880 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 = 6,767), *mcr*-2 (N = 6,889), *mcr*-3 (N = 6,159), *mcr*-4 (N = 6,520), *mcr*-5 (N = 6,424), *mcr*-6 (N = 6,447), *mcr*-7 (N = 5,919), *mcr*-8 (N = 6,884) and *mcr*-9 (N = 6,589).

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 20 seconds, 66 °C for 45 seconds, and 72 °C for 30 seconds. Melting curve analysis was performed with one cycle at 65 °C for 60 seconds, 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).

4.3.4 Data Analysis

Multiplexing based on FFI. Final fluorescent intensity values were extracted from each amplification curve (as in [99]) and used to train a logistic regression classifier to distinguish targets. It is important to stress that the primer mix concentration was not optimised to improve classification, therefore higher performance is expected if optimisation is conducted.

Amplification Curve Analysis (ACA). ACA consists of training a supervised machine learning model to distinguish targets based on the entire real-time amplification curve [97]. 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 (MCA). 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, T_m, although methods have also been proposed to consider the entire curve [66, 132]. 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 T_m values. In this study, logistic regression was chosen as a classifier based on cross-validation.

The Proposed Method. The amplification and melting curve analysis, or AMCA, trains a supervised machine learning model to combine the predictions of ACA and MCA. This process is visualised in Figure 4.2. 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.

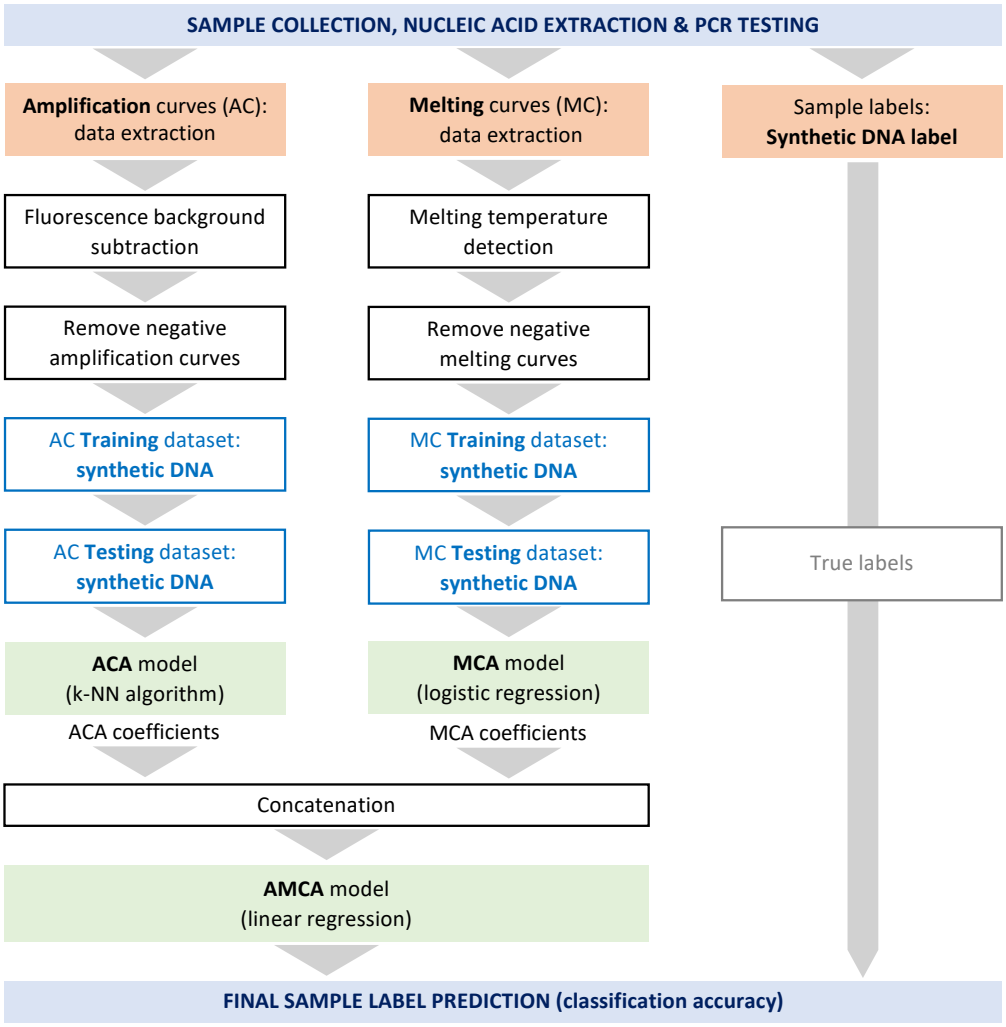


Figure 4.2: Flowchart to visualise the data processing workflow for the proposed method. True labels (marked as Synthetic DNA label from manufacture) are only required for training the models, as opposed to testing unknown samples. The input to the machine learning models are denoted as AC Training and MC Training. The output coefficients of ACA and MCA are concatenated and used for the final AMCA model sample prediction.

4.3.5 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. All data and code in this study can be found at

`https://github.com/am5113/pyAMCA`.

4.4 Results & Discussion

4.4.1 A new multiplex assay for mobilised 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 9plex has been designed and validated using a conventional qPCR platform. Figure 4.3A - 4.3C show the real-time amplification curves, melting peak distributions (extracted from melting curves) and standard curves for a serial dilution of each *mcr* target. Appendix Figure A.1 and A.2 show the raw melting curves before peak extraction and conventional standard curves, respectively. From Figure 4.3A, it can be observed that the final fluorescence and shape can vary between targets, although the precise overlap cannot be visualised. On the other hand, as in Figure 4.3B, the melting peak distributions have distinct mean T_m values, although some targets (e.g. *mcr*-1 and *mcr*-5) have overlapping distributions, compromising MCA multiplexing classification. Figure 4.3C 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 presence of *mcr* variants, by virtue of the overlapping T_m 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.

4.4.2 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 analysed, 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 4.3D and

4.3E show the amplification and T_m 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 4.4A and 4.4B show the confusion matrices, comparing the true and predicted targets for ACA and MCA, and the overall classification performance is 82.31 ± 1.47% and 89.34 ± 0.33%, respectively. Furthermore, a naive classification based on FFI gives an overall accuracy of 24.59 ± 0.52% (confusion matrix and FFI distributions are provided in Appendix Figure A.3). 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 9plex [99]. On the other hand, analysing the entire amplification curves (without normalising for FFI) using a neural network boosts performance by 57.7%, extracting relevant kinetic information from each event. The third method, MCA, analysed 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.

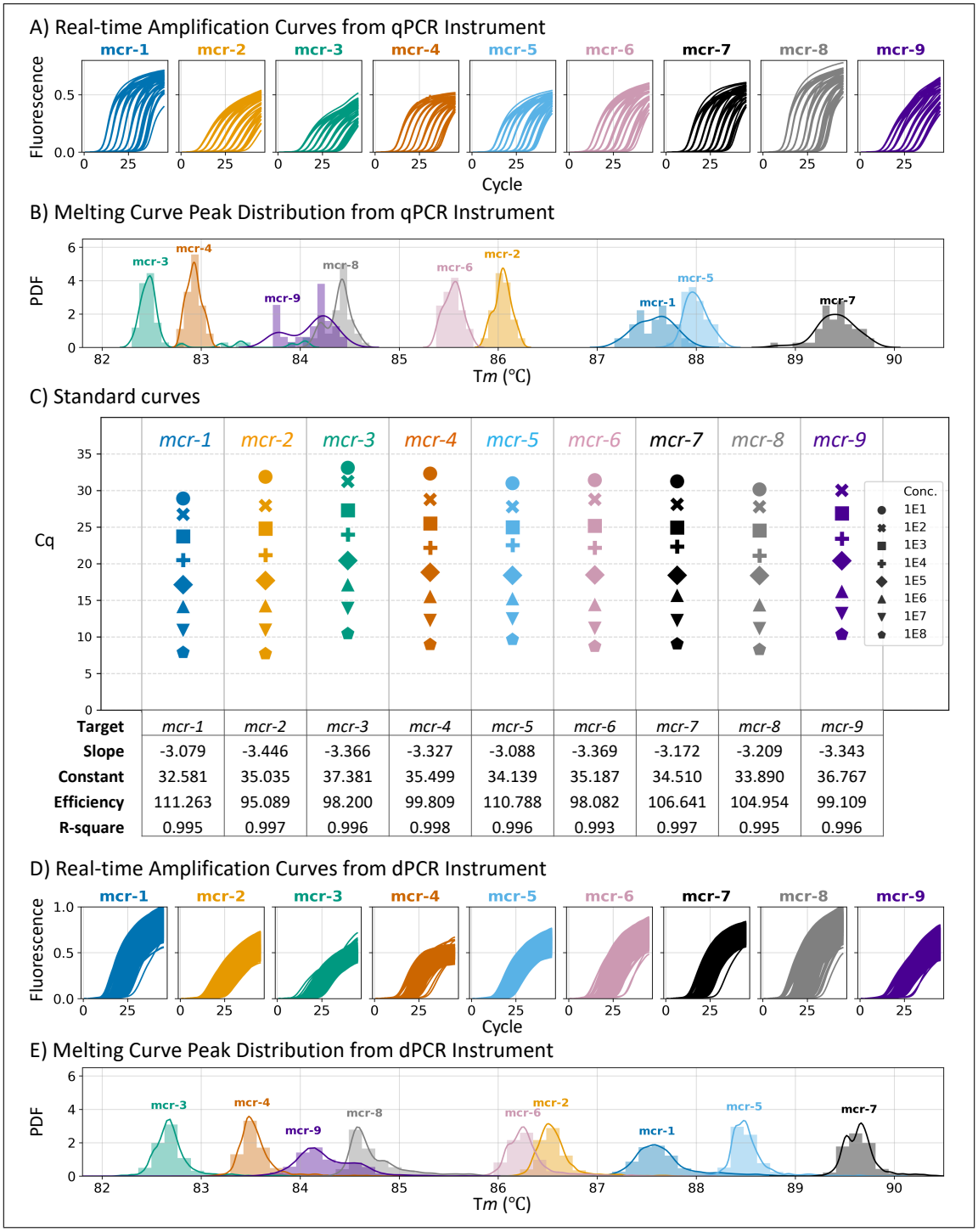


Figure 4.3: Analysis of real-time amplification and melting curves from qPCR and dPCR instruments. A) Real-time amplification curves from qPCR instrument. B) Melting curve peak distribution from qPCR instrument showing the probability density function (PDF) for each target. The mean ± std of *mcr*-1 to *mcr*-9 is 87.6 ± 0.2°C, 86.0 ± 0.1°C, 82.6 ± 0.4°C, 82.9 ± 0.1°C, 88.0 ± 0.1°C, 85.5 ± 0.1°C, 89.4 ± 0.2°C, 84.4 ± 0.1°C, 84.1 ± 0.2°C, respectively. C) Visualization and statistics of standard curves for a serial dilution of each target in qPCR using 9plex 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 ± 0.3°C, 86.6 ± 0.2°C, 82.7 ± 0.2°C, 83.6 ± 0.2°C, 88.5 ± 0.2°C, 86.3 ± 0.2°C, 89.7 ± 0.2°C, 84.8 ± 0.3°C, 84.3 ± 0.3°C, respectively. Raw melting curves are shown in Appendix Figure A.1.

4.4.3 AMCA method increases classification accuracy compared to ACA or MCA individually

Figure 4.4C shows the confusion matrix comparing the predicted classification from the AMCA method to the true labels. It can be observed that the accuracy is $99.33 \pm 0.13\%$ 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:

$$\mathbf{y} = \hat{\mathbf{W}}_{\text{ACA}}\,\mathbf{y}_{\text{ACA}} + \hat{\mathbf{W}}_{\text{MCA}}\,\mathbf{y}_{\text{MCA}} \tag{4.1}$$

Where $\mathbf{y}_{\text{ACA}} \in \mathbb{R}^9$ and $\mathbf{y}_{\text{MCA}} \in \mathbb{R}^9$ are the probability vectors outputted from the ACA and MCA models, $\hat{\mathbf{W}}_{\text{ACA}} \in \mathbb{R}^{9\times 9}$ and $\hat{\mathbf{W}}_{\text{MCA}} \in \mathbb{R}^{9\times 9}$ are the model coefficients, respectively. This method is one of the simplest forms of "stacking" [133], which is a special case of ensembling, where after training and getting the coefficients of both ACA and MCA models their a linear regression is applied to further enhance the classification performance. Here, predictions are made by selecting the maximum entry for the y vectors (containing arbitrary non-negative real numbers) and selecting the corresponding *mcr* label. Figure 4.4D and 4.4E 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 $1,515$ *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 $1,846$ *mcr*-9 reactions as *mcr*-2 and the coefficients compensate for this phenomenon.

4.4.4 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 4.4F shows the classification performance on 5000 out-of-sample data points (repeated 10 times) where n_{train}

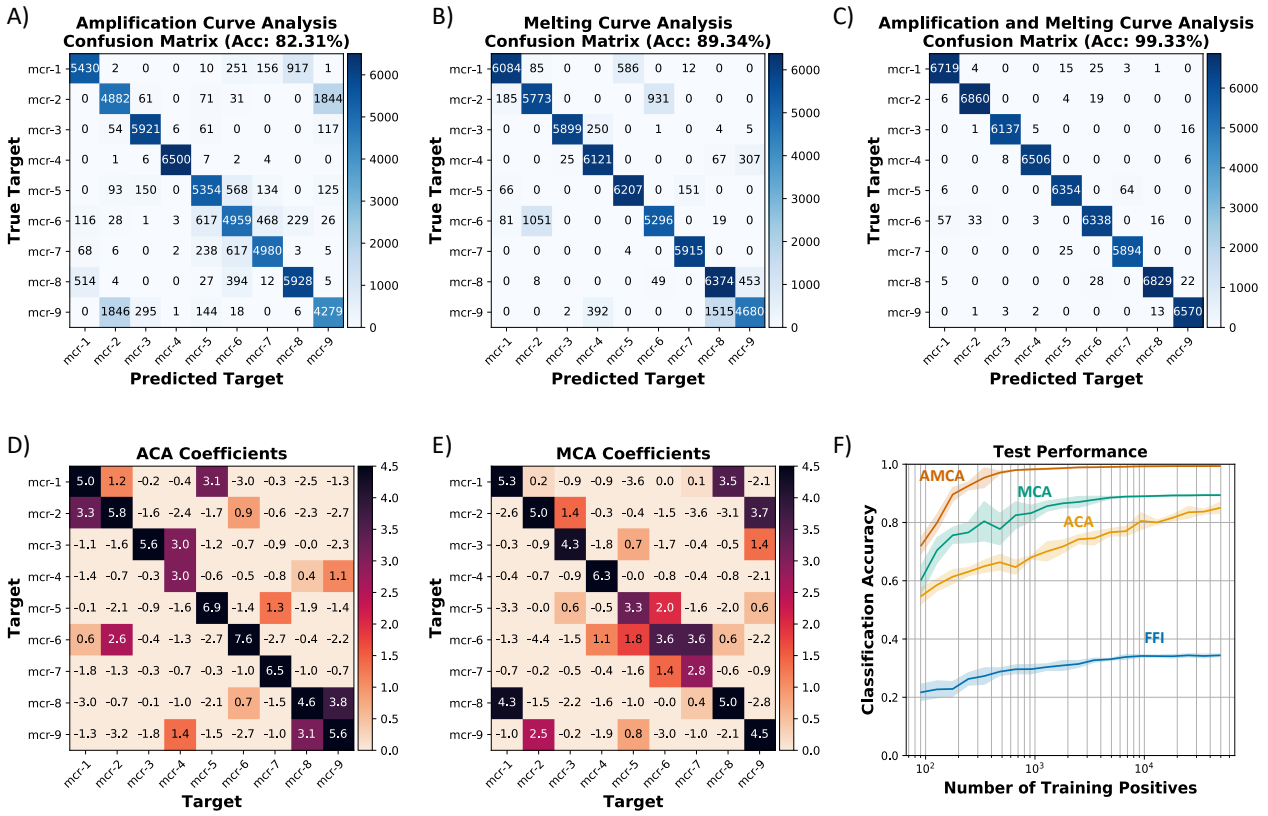


Figure 4.4: Performance of all methods for multiplexing the 9 *mcr* targets. A, B, C) Confusion matrices illustrating the predictions from ACA, MCA and AMCA (proposed method), respectively. Values indicate the number of amplification events with diagonal entries corresponding to correct predictions. D, E) Coefficients of the AMCA model weighting the predictions from the ACA and MCA methods, respectively. Darker colours 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.

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.

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

The same methodology (as in Figure 4.2) was applied to the qPCR data presented in Figure 4.3A and 4.3B. The classification accuracy for ACA, MCA and AMCA was shown to be 84.40±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 Appendix Figure A.4. These results suggest that the AMCA method works across real-time platforms, both quantitative and digital.

4.5 Conclusion

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^{TM} 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 4.4F, 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 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 presence of two *mcr* variants in the same sample [134]. Therefore, as in Moniri *et al.* (2020), considering the presence of 2 targets and under the constraint of $36,960$ chambers (Fluidigm 37K^{TM} chip), the quantification uncertainty is below 5% between 16.7% and 99.3% digital occupancy [97]. 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 [66, 132].

This study showed a the application of AMCA 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. An accuracy of 99.3% is reported to identify the nine colistin resistance genes, using affordable intercalating dye. Observing that the AMCA classification accuracy is better than solely analysing 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 characterise the contribution of ACA and MCA across all targets to optimise the final classification of each amplification event.

CHAPTER LESSON

This Chapter showed the application of data-driven multiplexing high-multiplexing in real-time PCR instruments with melting curve capabilities. This approach extracts kinetic and thermodynamic information, to enhance the classification accuracy in single-well and single-channel multiplex assays using machine learning algorithms.

TAKEAWAY QUESTION

"Can data-driven multiplexing be translated to other amplification chemistries such as isothermal-based?"

Chapter 5

Towards Isothermal Data-driven Multiplexing

5.1 Chapter Overview

The previous Chapter explored how to expand the data-driven multiplexing capabilities by leveraging Kinetic and thermodynamic information encoded in the amplification event to triple the number of detectable targets. The method is further developed here, and data-driven approaches are translated to isothermal chemistries, in particular to Loop-mediated isothermal amplification (LAMP). LAMP assays are currently limited to one target per reaction in the absence of melting curve analysis, molecular probes or restriction enzyme digestion. Here, multiplexing of five targets in a single fluorescent channel is demonstrated using digital LAMP and the machine learning-based method Amplification Curve Analysis, resulting in a classification accuracy of 91.33% on $54,186$ positive amplification events.

The concepts in this Chapter resulted in the following journal article:

- Malpartida-Cardenas K*, Miglietta L*, Peng T, Moniri A, Holmes A, Georgiou P, Rodriguez Manzano J. "Single-channel digital LAMP multiplexing using amplification curve analysis." *Sensors & Diagnostics*, 2022 May 19;1(3):465-8. *First joint authorship.

5.2 Introduction

Nucleic acid amplification tests for diagnosis and epidemiological surveillance of infectious disease are essential in the fight against outbreaks such as the ongoing COVID-19 pandemic. In addition to the gold standard polymerase chain reaction (PCR), loop-mediated isothermal amplification (LAMP) has become a popular alternative due to its high sensitivity, specificity, and rapidness. Although numerous LAMP assays have been developed in the last two decades, they have commonly been restricted to detect one target per reaction, limiting the throughput of technologies that rely on LAMP. Several methods have been employed to increase the number of targets in a single LAMP reaction, including: (i) fluorescence-based detection at different excitation wavelengths through the incorporation of a specific quencher-fluorophore pair per each target [135, 136, 137, 138], (ii) DNA restriction enzyme digestion followed by gel electrophoresis [139, 140], and (iii) melting curve analysis [141, 142]. However, probe-based approaches are still limited by the number of fluorescence channels present in the PCR platform and the increased cost of reagents, whereas post-PCR analysis requires more complex instrumentation, longer protocols and exposes the reaction to a greater risk of DNA contamination [143]. As demostrated in the previous Chapter, kinetic information embedded in an amplification curve can be used to distinguish nucleic acid targets [97, 144, 130]. This novel approach, named as data-driven multiplexing, utilise mathematical algorithms to extract target specific features from real-time amplification data which can be used as classifiers' input. In particular, this work explores the use of the Amplification Curve Analysis (ACA) classifier, which consists of a supervised machine learning model (i.e., k-nearest neighbours) using the entire real-time curve from each amplification event. This study demonstrates for the first time the applicability of ACA in digital LAMP (dLAMP) for multiplexing five LAMP assays (5plex-LAMP) in a single reaction with a non-specific intercalating dye (EvaGreen), therefore using a single-fluorescent channel in digital PCR. As a case study, this work focuses on the detection of five respiratory pathogens which present similar flu-like symptoms [145]: human influenza A virus (IAV), human influenza B virus (IBV), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), human adenovirus (hAdV) and Klebsiella Pneumoniae (KP).

5.3 Experimental Section

5.3.1 LAMP primer sequences

Primer sequences for each of the targets are summarised in 5.1. A LAMP assay was designed for the detection of the "M" gene of the influenza A virus. Genomic sequences were retrieved from NCBI (`http://www.ncbi.nlm.nih.gov/genbank/`) and sequence alignment was performed using the MUSCLE algorithm [94]. A conserved region of interest was selected, and the sequence was uploaded into Primer Explorer v5 software for the generation of several sets of LAMP assays. Further manual optimisation and design of loops primers were performed using GENEious Prime 2020.1.2 (`https://www.geneious.com`). Primer sequences specific to each of the targets were analysed with IDT OligoAnalyzer software (`https://eu.idtdna.com/pages/tools/oligoanalyzer`) using the J. SantaLucia thermodynamic table for melting temperature (T_m) evaluation, hairpin, self-dimer, and cross-primer formation. Primers were purchased from IDT and rehydrated in TE (pH 8.00) at 500 µM. A 50X primer mix was prepared for each target and subsequently, the 10X 5plex-LAMP was obtained by mixing each specific primer mix at equitable volumes.

5.3.2 Multiplex real-time LAMP

Real-time LAMP reactions consisted of 6 µL final reaction volume including: 0.60 µL of 10x custom isothermal buffer, 0.30 µL of Mg SO4 (100 mM stock), 0.34 µL of dNTPs (25 mM stock), 0.36 µL of BSA (20 mg/mL), 0.48 µL of Betaine (5 M stock), 0.60 µL of 10X 5plex LAMP primer mix, 0.15 µL of NaOH (0.2 M stock), 0.03 µL of Bst 2.0 DNA polymerase (120 kU/µL stock), 0.30 µL of EvaGreen (20X stock), 1.8 µL of the target oligonucleotide and enough nuclease free water to have a final volume of 6 µL. Amplification reaction was performed at 63°C for 35 cycles of 60 seconds duration reading at the end of each cycle. Melting curve analysis was performed after the amplification reaction and consisted of 1 cycle at 95°C for 10 seconds, 65°C for 60 seconds, and gradual temperature change from 65°C to 97°C with a step of 2.2°C/s reading every 0.2°C. The LAMP protocol was based and adapted from Rodriguez-Manzano *et al.* [146].

Table 5.1: Primer sequences of the LAMP assays used for the 5plex-LAMP.

Assay	Primer	Sequence 5'→ 3'	Gene	LOD[a]	Author
LAMP-FA1	F3	GGCTATGGAGCAAATGGCTG			
LAMP-FA1	B3	CACTTGAACCGTTGCATCTG			
LAMP-FA1	LF	CTGACTAGCAACCTCCATGG	M	180 copies/reaction	This study [147]
LAMP-FA1	LB	GCTGGTCTGAAAAATGATCTTCTTG			
LAMP-FA1	FIP	CGCTTGCACCATTTGCCTAGCGATCGAGTGAGCAAGCAGC			
LAMP-FA1	BIP	TGGGACTCATCCTAGCTCCAGTCACCCCCATTCGTTTCTGA			
LAMPcov	F3	ACCAATAGCAGTCCAGATGA			
LAMPcov	B3	CACGATTGCAGCATTGTTAGC			
LAMPcov	LF	GGACTGAGATCTTTCATTTTACCGT	N	10 copies/reaction	[146]
LAMPcov	LB	ACTGAGGGAGCCTTGAATACA			
LAMPcov	FIP	TCTGGCCCAGTTCCTAGGTAGTCCAGACAAATTCGTGGTGG			
LAMPcov	BIP	GGACTTCCCTATGGTGCTAACAAACGGGTGCCAATGTGATCT			
LAMP-FB1	F3	AGGGACATGAACAACAAAGA			
LAMP-FB1	B3	CAAGTTTAGCAACAAGCCT			
LAMP-FB1	LF	TCAAACGGAACTTCCCTTCTTTC	NS1	1 copy/reaction	[141]
LAMP-FB1	LB	GGATACAAGTCCTTATCAACTCTGC			
LAMP-FB1	FIP	TCAGGGACAATACATTACGCATATCGATAAAGGAGGAAGTAAACACTCA			
LAMP-FB1	BIP	TAAACGGAACATTCCTCAAACACCACTCTGGTCATAGGCATTC			
LAMP-HAdV	F3	GTGCGACAGGACCATGTG			
LAMP-HAdV	B3	GGTAGACGGCCTCGATGA			
LAMP-HAdV	LF	GGCCCCCCATGGACATGAA	HEXON	180 copies/reaction	[148]
LAMP-HAdV	LB	CCACCCTGCTTTATCTTCTTTTCG			
LAMP-HAdV	FIP	AGCATGTTCTGTCCCAGGTCGGCATTCCCTTCTCCAGCAA			
LAMP-HAdV	BIP	GAGGTGGATCCCATGGATGAGCACTCTGACCACGTCGAARAC			
LAMP-KPn	F3	GGATATCTGACCAGTCGG			
LAMP-KPn	B3	GGGTTTTGCGTAATGATCTG			
LAMP-KPn	LB	GAAGACTGTTTCGTGCATGATGA	RCSA	10 copies/reaction	[149]
LAMP-KPn	FIP	CGACGTACAGTGTTTCTGCAGTTTTAAAAAACAGGAAATCGTTGAGG			
LAMP-KPn	BIP	CGGCGGTGGTGTTTCTGAATTTTGCGAATAATGCCATTACTTTC			

[a] Limit of Detection (LOD)

5.3.3 Multiplex real-time digital LAMP

Real-time digital LAMP reactions consisted of 6 µL final reaction volume including: 0.024 µL of ROX (50 µM stock), 0.60 µL of 20x GE Sample Loading Reagent (Fluidigm), 0.60 µL of 10x custom isothermal buffer, 0.30 µL of $MgSO_4$ (100 mM stock), 0.34 µL of dNTPs (25 mM stock), 0.36 µL of BSA (20 mg/mL), 0.48 µL of Betaine (5 M stock), 0.60 µL of 10X 5plex LAMP primer mix, 0.15 µL of NaOH (0.2 M stock), 0.03 µL of Bst 2.0 DNA polymerase (120,000 U/µL stock), 0.30 µL of EvaGreen (20X stock), 1.8 µL of the target oligonucleotide and enough nuclease free water to have a final volume of 6 µL. The qdPCR 37K^{TM} integrated fluidic circuit (IFC) was used to perform the dLAMP experiments. Firstly, the 48.48 control lines fluid were injected into each accumulator of the qdPCR 37K^{TM} IFC and primed in the IFC Controller MX. Secondly, reactions and 1X GE were loaded into the qdPCR 37K^{TM} IFC following manufacturer's instructions and the qdPCR 37K^{TM} IFC was loaded IFC Controller MX. Finally, the qdPCR 37K^{TM} IFC was placed into the Fluidigm's Biomark HD system and the amplification reaction was performed at 63°C for 35 cycles of 1 min duration reading at the end of each cycle. Melting curve analysis was performed after the amplification reaction and consisted of 1 cycle at 95°C for 10 s, 65°C for 60 s, and gradual temperature change from 65°C to 97°C with a step of 2.2 °C/s reading every 0.2 °C. The qdPCR 37K^{TM} IFC contains 48 inlets

which correspond to 48 panels. Each of the panels contains 770 wells with a volume of 0.85 nL.

5.3.4 Evaluation of the 5plex-LAMP assay

Performance of the 5plex-LAMP was evaluated by using 10-fold serial dilutions of synthetic DNA of each of the targets at concentrations ranging from 11.8×10^8 to 1.8×10^2 copies per reaction. A total of 8 replicates were performed per each concentration and target. Specificity of the 5plex-LAMP assay was evaluated *in-silico* by testing the primers with the sequences of the target pathogens and experimentally by cross-testing each LAMP assay with all the other targets including non-template controls (NTC). Synthetic oligonucleotides (gBlock^{TM} Gene Fragment) for each of the targets were purchased from IDT and resuspended at 5 ng/µL.

5.3.5 Machine learning methods for the detection of amplification events: ACA, MCA and FFI

Multiple standard packages and in-house scripts in Python (v3.7) were developed to analyse the data: (i) FFI values were extracted from each amplification curve, considering only the last values in the cycle time series. The FFI model consisted in a logistic regression classifier to distinguish different targets (please note that these assays are not optimised for an improved FFI classification). (ii) A k-Nearest neighbor model was used to implement the ACA model using scikit-learn package with default parameters (for more information please see provided code and package documentation). The ACA classification accuracy (i.e., proportion of correctly identified events), sensitivity (i.e., true positive rate), and specificity (i.e., true negative rate) values in Tables 1 were computed for each binary classification subproblem in the one-vs-one multiclass classification scheme. (iii) The MCA classifier distinguished the melting peak temperature or peak Tm, using a supervised machine learning classifier. Here a logistic regression was used. Performance of the models was 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, shuffled stratified split was performed five times, with $5,000$ test samples. All data and code used in this study can be found at `https://github.com/LMigliet/pyiACA`.

5.4 Results & Discussion

Publicly available assays were used to demonstrate the applicability of the ACA method for multiplexing in dLAMP without lengthy assay optimisation. All primer LAMP sequences used in this study are detailed in 5.1. Please note that the LAMP assay for IAV (targeting M gene) was designed in-house. Performance of the 5plex-LAMP assay was evaluated with a fluorescence-based real-time instrument (LightCycler96 system, Roche) using a 10-fold serial dilution of synthetic DNA, at concentrations ranging from 1.8×10^7 to 1.8×10^2 copies per reaction. All assays amplified their specific target down to 180 copies per reaction. Melting curve analysis was used to confirm the target-specific amplification; obtained melting temperature peak values (T_m) for IAV, IBV, SARS-CoV-2, hAdV and KP were 88.5°C, 83.5°C, 86.5°C, 89.5°C and 88°C, respectively. Self-dimer or cross-primer formation was not observed in the non-template control (NTC) during the 35 cycles (1 min/cycle) run.

The 5plex-LAMP was then tested in a digital real-time instrument, dLAMP. In total, $110,880$ amplification events were generated including 54,186 positive amplification reactions. Time-to-positive distribution obtained with the 5plex-LAMP assay are provided in Appendix Figure B.2. Between 6,000 to 14,000 positive amplification events were obtained per target, and an adequate number of NTC reactions (N = 6,930) were included to verify the absence of contamination, formation of any detectable secondary structure or primer dimerisation.

The obtained data was first evaluated by unsupervised machine learning using the Uniform Manifold Approximation and Projection (UMAP) method to visualise how distinguishable the amplification curves were per target [150]. Classification and clustering considered all available real-time data (in this case, 40 data-point per amplification reaction). After dimensionality reduction into a 3D space (Figure 5.1A), it can be observed that amplification curves obtained per each target formed distinguishable clusters.

As shown in Figure 5.1A, supervised machine learning was employed to classify the amplification curves demonstrating the capability of the ACA method for single-channel multiplexing in dLAMP. The selected classification algorithm was k-nearest neighbor (KNN, with parameter k = 10) [144, 151]. The overall classification accuracy of the ACA method was 91.33% ± 0.33% (mean ± std), represented by the confusion matrix shown in Figure 5.1B. In addition, the accuracy, sensitivity, and specificity for the one-vs-one classifiers is shown in Table 5.2, which

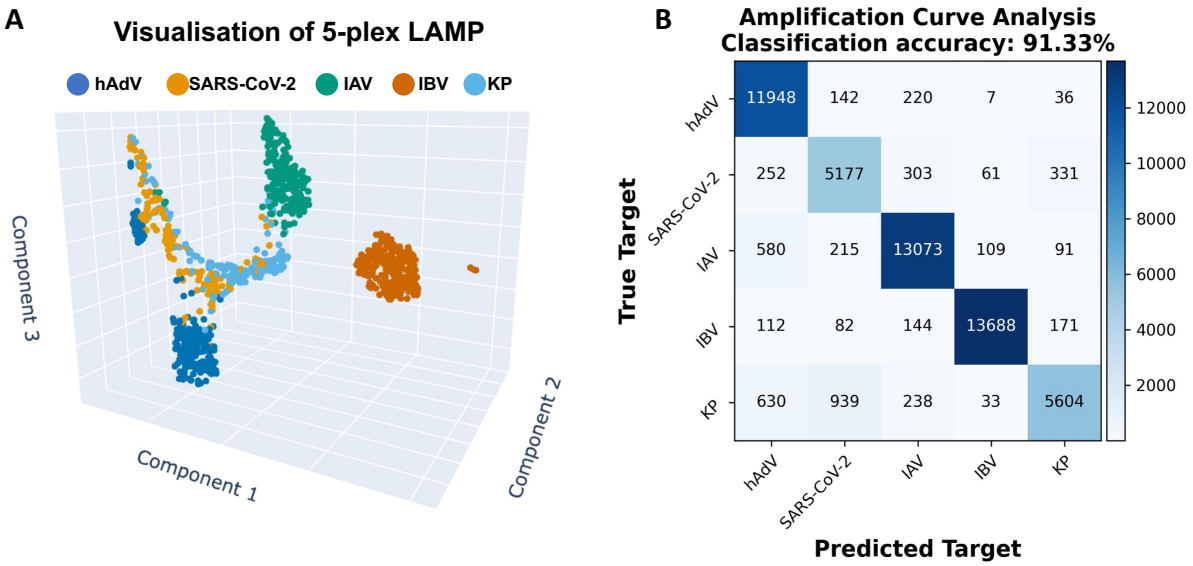


Figure 5.1: Performance of the multiplex LAMP assay using the ACA machine-learning based method in real-time digital LAMP. (A) Visualisation of the similarity of real-time LAMP amplification curves using the Uniform Manifold Approximation and Projection algorithm. (B) Confusion matrix showing prediction performance of ACA for each of the selected targets in the 5plex-LAMP: human influenza A virus (IAV), human influenza B virus (IBV), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), human adenovirus (hAdV) and Klebsiella Pneumoniae (KP).

demonstrates that the 5 targets can be distinguished with a classification accuracy ranging from 91.10% to 99.15%.

Table 5.2: ACA classification performance by one-vs-one classifiers.

Targets	Accuracy	Sensitivity	Specificity
hAdV vs SARS-CoV-2	97.40%	98.74%	94.69%
hAdV vs IAV	97.22%	98.32%	96.25%
hAdV vs IBV	99.15%	99.88%	98.51%
hAdV vs KP	97.55%	99.42%	94.45%
SARS-CoV-2 vs IAV	97.03%	94.02%	98.34%
SARS-CoV-2 vs IBV	98.64%	98.64%	98.63%
SARS-CoV-2 vs KP	91.10%	93.08%	89.48%
IAV vs IBV	98.96%	99.30%	98.63%
IAV vs KP	97.94%	99.03%	95.86%
IBV vs KP	98.25%	97.93%	98.86%

Furthermore, these results are compared with two alternative machine learning-based methods commonly used for the identification of multiple targets in single-well PCR multiplex assays; Final Fluorescence Intensity (FFI) and Melting Curve Analysis (MCA). The obtained classification accuracy of the MCA method was 94.55% ± 0.33% (melting curves distribution and confusion matrix are shown in Figure 5.2A-B), which represents a 3.41% ± 0.33% improve-

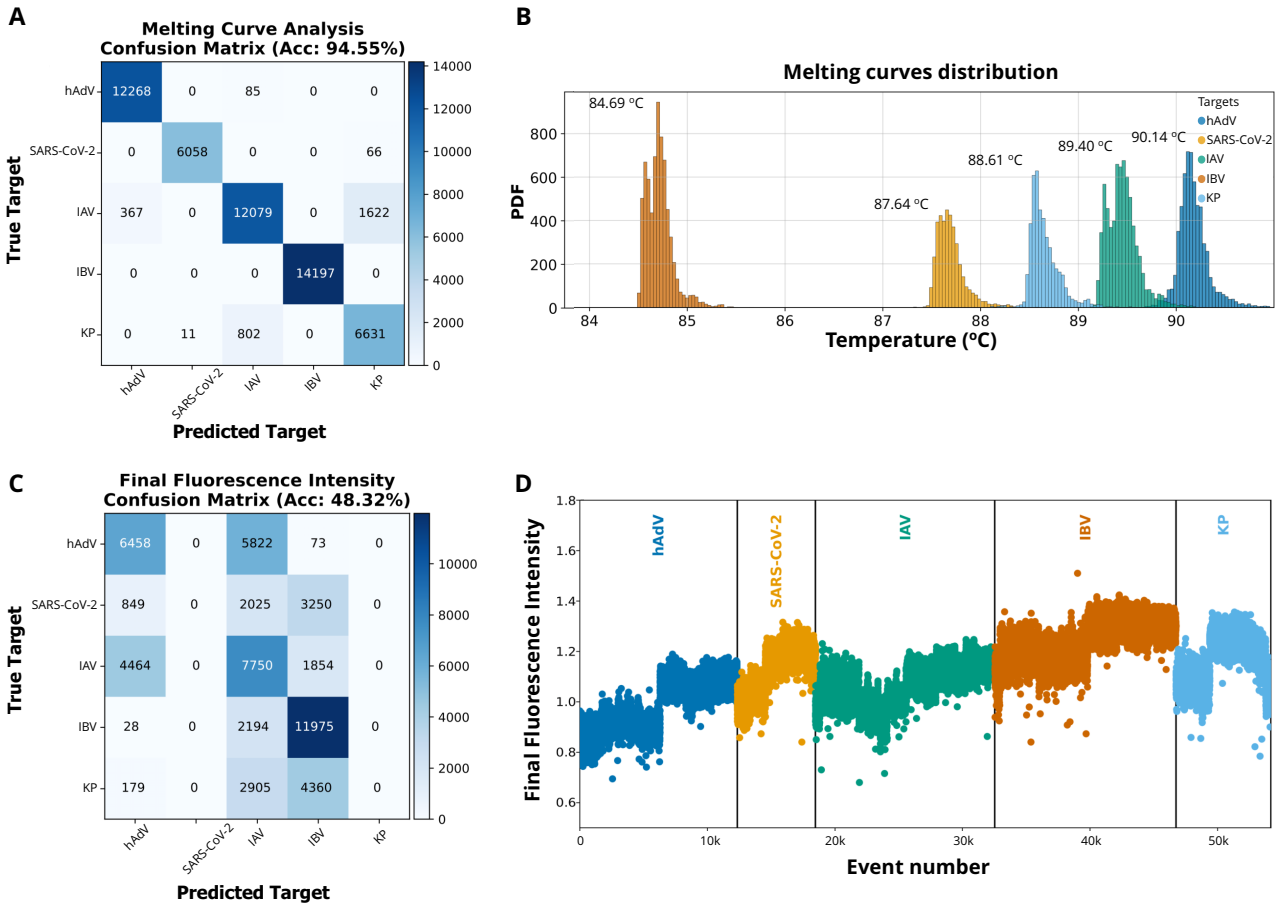


Figure 5.2: Performance of Melting Curve Analysis (MCA) and Final Fluorescence Intensity (FFI) machine-learning based methods in real-time digital LAMP. (A) Confusion matrix showing the prediction performance of MCA for each of the targets in the respiratory panel. (B) Melting curve distributions for each target showing the median temperature of the distribution. (C) Confusion matrix showing the prediction performance of FFI for each of the targets in the respiratory panel. (D) Distribution of FFI across the five targets.

ment compared to the ACA. The results obtained with the FFI method reported a classification accuracy of 48.32% ± 0.56% (Figure 5.2C), showing a 43.01% ± 0.56% decreased classification accuracy compared to ACA method. The FFI values were similar across different assays, and consequently the LAMP mechanism are not suitable for FFI classification-based (Figure 5.2D).

It is important to note that the 5plex-LAMP has not been optimised for any of the used methods, neither for ACA, MCA nor FFI analysis, therefore obtained results could have been improved. Furthermore, this is the first time FFI has been applied for target identification in LAMP. The combination of ACA and MCA methods, named Amplification and Melting Curve Analysis (AMCA) has been previously reported by Moniri *et al.* [130] and Miglietta *et al.* [144] as an approach that combines coefficients from both classifiers improving overall accuracy (as shown in Figure 5.3). As depicted in Appendix Figure B.1, all methods except FFI achieved

a classification accuracy superior to 90% requiring 103 training data points. Although MCA and AMCA have shown superior performance compared to the ACA, the limitations that MCA impose in terms of accurate thermal control restrict its future use in combination with LAMP, particularly for point-of-care applications.

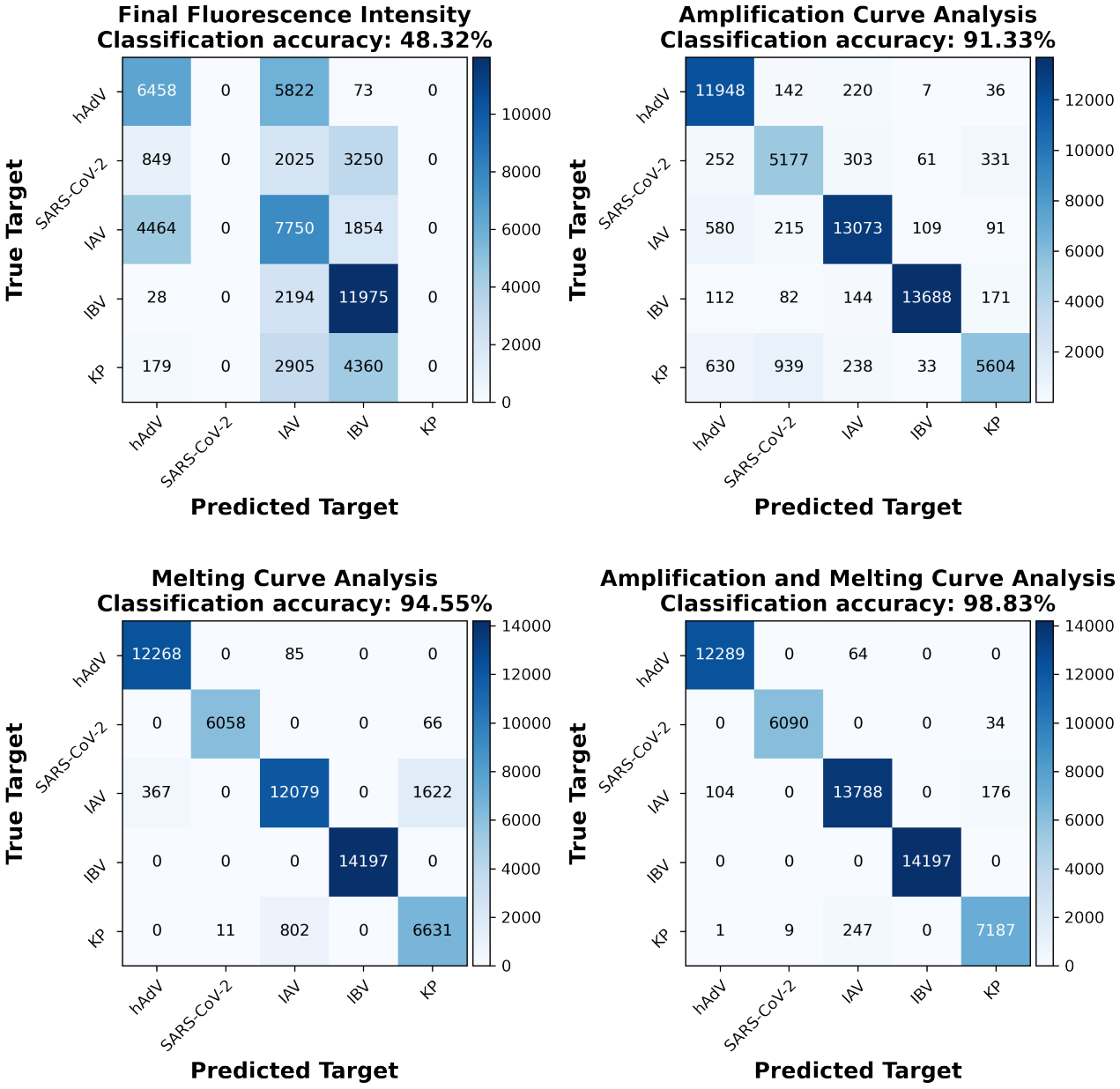


Figure 5.3: Confusion matrices showing the prediction performance of the four methods evaluated: FFI, ACA, MCA and AMCA.

5.5 Conclusion

The achieved throughput and turnaround time (< 35 min) in a single well reaction leverages target identification accuracy of several pathogens. This proof-of-concept study demonstrates that the ACA method can be used to multiplex LAMP assays using only the amplification curves. No further primer design optimisation, modifications in the reaction, incorporation of molecular probes or accurate thermal cycling are needed. Furthermore, it is observed that 5plex LAMP assays did not generate non-specific products (e.g., primer dimerisation). Although there may be a limitation in the maximum number of assays that can be multiplexed in a single well, the 5plex-LAMP used here has proven to be equal or higher than the currently used methods for multiplexing in LAMP which rely on molecular probes [135], melting curve analysis or restriction enzyme digestion [140].

Notwithstanding the achieved results, limitations to this study include the fact that real-time digital instruments are not commonly available, and therefore, the performance of the evaluated methods for target classification using data from a conventional real-time instrument should be further assessed. This will also require verifying if the trained data is transferable across instruments such that the proposed methodologies could be implemented in conventional real-time instruments, and ultimately in affordable devices for point-of-care diagnostics. Lastly, the conducted experiments for the demonstration of data-driven multiplexing with LAMP only considered the presence of synthetic pure DNA targets. Co-infections are likely to occur, especially in the field of infectious diseases where it is common to find patients presenting with more than one disease. The use of dLAMP with single molecule resolution will increase the accuracy in determining the presence of co-infections. This could also be further investigated in future work, as well as the validation of the proposed method with clinical samples to determine its robustness and performance for multiplexing.

CHAPTER LESSON

This Chapter demonstrated that multiplexing five LAMP assays in a single well reaction using a single fluorescent channel can be achieved with the Amplification Curve Analysis (ACA) in a highly accuracy manner without the need of downstream experiments. The Chapter vision is to apply the proposed method for multiplexing any desired isothermal assay at standard laboratory settings enhancing the current testing capabilities, and at the point-of-care once integrated in portable devices that acquire real-time data.

TAKEAWAY QUESTION

"Can data-driven multiplexing be validated in real-world clinical diagnostics scenarios, for example in hospitals?"

Chapter 6

Clinical Application of the Data-driven Multiplexing

6.1 Chapter Overview

Previous chapters have demonstrated the ability to combine machine learning algorithms with real-time PCR instruments to increase classification accuracy of multiplex PCR assays when using synthetic DNA templates. The next study aims to determine if this novel methodology could be applied to improve identification of the five antimicrobial resistance genes in clinical isolates, which would represent a leap forward in the use of PCR-based data-driven diagnostics for clinical applications. A total of 253 clinical isolates (including 221 positive samples) were collected and a novel 5plex PCR assay for detection of bla_{IMP}, bla_{KPC}, bla_{NDM}, $bla_{\text{OXA-48}}$ and bla_{VIM} was developed. Combining the described ML method "Amplification and Melting Curve Analysis" (AMCA) with the abovementioned multiplex assay, the performance of the AMCA method is assessed for the detection of these five genes. The AMCA classifier demonstrated excellent predictive performance with 99.6% (CI 97.8-99.9%) accuracy (only one misclassified sample out of the 253, with a total of 160,041 positive amplification events), which represents a 7.9% increase (p-value < 0.05) compared to conventional melting curve analysis. This work demonstrates the use of the AMCA method to increase the throughput and performance of state-of-the-art molecular diagnostic platforms, without hardware modifications and additional costs, thus potentially providing substantial clinical utility on screening patients for CPO car-

riage.

The concepts in this Chapter resulted in the following journal article:

- Miglietta L, Moniri A, Pennisi I, Malpartida-Cardenas K, Abbas H, Hill-Cawthorne K, Bolt F, Jauneikaite E, Davies F, Holmes A, Georgiou P. "Coupling machine learning and high throughput multiplex digital PCR enables accurate detection of carbapenem-resistant genes in clinical isolates". *Frontiers in molecular biosciences*, 2021;8:775299.

6.2 Introduction

This Chapter demonstrates that machine learning (ML) approaches coupled with high throughput real-time digital PCR (dPCR) can be used to increase detection accuracy of multiplex PCR assays when screening clinical isolates for the presence of carbapenemase-producing organisms (CPOs). A recently reported ML method called Amplification and Melting Curve Analysis (AMCA), which leverages the target-specific information encoded in each amplification event (via real-time data), was used to identify the nature of nucleic acid molecules [97]. The AMCA approach is based on training supervised machine learning algorithms to extract kinetic and thermodynamic information from PCR amplification and melting curves to enhance the classification accuracy in multiplexing. Validation of this methodology using clinical isolates has never been reported before; therefore, this work represents a step forward towards the implementation of this method into clinical microbiology laboratories. Nucleic acid amplification tests (NAATs) that incorporate the AMCA classifier for multiple target detection will greatly improve their specificity, sensitivity and turn-around time to result, reducing overall resource consumptions and improving diagnostic performance.

Antimicrobial resistance (AMR) is a serious global threat and poses a challenge for modern medicine, compromising effective infectious disease management [152, 153]. One of the most concerning forms of AMR is the rapid spread of CPOs; bacteria producing enzymes that inactivate the potent antibiotics, carbapenems. Whilst overall UK incidence is low, there are centres nationally facing increasing rates and outbreaks, including Imperial College Healthcare NHS Trust (ICHNT), and it is endemic in many other regions worldwide [126, 125]. CPO infections are associated with higher morbidity and mortality than susceptible strains, in part because

their resistance can lead to ineffective empirical therapy and suboptimal treatment [154, 155]. Therapeutic options are severely restricted, and in many cases clinical management relies on "last line" antibiotics that are less effective and have more side effects [156].

Patients infected with CPOs present significant challenges for diagnostics and infection control. There is an urgent need for accurate and timely diagnosis to improve patient outcomes and prevent the spread of AMR. Carbapenemase resistance genes are often co-localised on highly transmissible plasmids and are readily shared between bacterial species, providing the ideal conditions for multidrug resistant organisms [157]. Incorrect diagnosis delays appropriate intervention, increases financial burdens for the healthcare system, and complicates antimicrobial stewardship efforts [158]. A local ICHNT economic analysis estimated the cost of a large hospital outbreak ($\approx$ 100 infections) of carbapenemase producing *Klebsiella pneumoniae* to be £1M. Some of the increased expenditure was associated with increased screening, bed closures, medication and patient bed-days; better diagnostics could reduce these costs [128, 126].

Diagnosis of CPOs is often too complicated and time-consuming, as it is normally based upon multiple tests which employ a wide range of instruments and diagnostic tests. Phenotypic methods typically target carbapenemase production and provide no information on the underlying resistance mechanism [159]. These tests represent a low-cost (£2-15 per sample) and robust methodology; however, they rely on pure culture which increases turnaround times (12-24h) [160]. A variety of molecular methods, including amplification (PCR-based), microarray and sequencing assays have been developed and are frequently used in microbiology laboratories [161, 162]. Microarray and sequencing are time consuming (>12-48h), expensive (>£50K platforms and >£80 per sample), and require bioinformatic expertise. Conversely, NAATs are commonly cheaper (£15-30 per sample) and faster (1-2h), whereas instrument price significantly ranges between tens to hundreds of thousands of pounds for conventional and digital PCR platforms, respectively [21, 163]. Furthermore, the application of sophisticated data processing for its optimisation (as done with microarray and sequencing methods) has been largely unexplored [164, 165]. As a result of all aforementioned limitations, implementation of microarrays, sequencing and molecular methods for CPO diagnosis into routine practice is often limited.

Recently, our group has demonstrated that the large volume of data obtained from real-time digital PCR (dPCR) instruments can be exploited to perform data-driven multiplexing

in a single fluorescent channel, reporting a 99.33 ± 0.13% classification accuracy when using synthetic DNA in a 9-plex format [130]. This result represented an increase of 10% over using melting curve analysis, indicative of the potential benefits of this methodology for diagnostic and screening applications. The ML method used (AMCA) leverages kinetic and thermodynamic information encoded in the amplification and melting curves to perform target identification in multiplexed environments [16, 15]. For the first time, the analytical performance of AMCA method was compared to Xpert Carba-R Cepheid and Resist-3 O.K.N assays on clinical isolates for detection of the most common types of serine-beta-lactamases (bla_{KPC} and $bla_{\text{OXA-48}}$) and metallo-beta-lactamases (bla_{IMP}, bla_{VIM} and bla_{NDM}) [166, 167]. Results were compared against another ML based classifier 'Melting Curve Analysis' (MCA), which uses the thermodynamic information contained in PCR melting curves for identification of multiple targets in a single well reaction [66, 97]. A 5plex PCR assay was developed in-silico and validated with synthetic DNA templates. The performance of the AMCA method, using this 5plex, was further assessed with 253 clinical isolates provided by the microbiology department at Charing Cross Hospital, ICHNT. All samples were analysed in real-time dPCR, using an intercalating dye (EvaGreen) in a single-fluorescent channel. This work demonstrates that the AMCA method can be integrated with conventional clinical diagnostic workflows in combination with real-time dPCR platforms, as it does not require any hardware modification. Increasing multiplexing capabilities enables improved workflow efficiency while reducing per sample cost, and it is beneficial to a number of application fields beyond clinical diagnostics, such as veterinary and environmental fields, where multiple targets need to be analysed simultaneously (e.g., SNP genotyping, forensic studies and gene deletion analysis). Figure 6.1 illustrates the concept of data-driven multiplexing, where tailored PCR-based amplification chemistries combined with advance data analytics can be seamlessly integrated into existing diagnostics pipelines which utilise real-time platforms.

6.3 Experimental Section

6.3.1 Synthetic DNA

Double-stranded synthetic DNA (gBlock$^{\text{TM}}$ Gene Fragments) containing the entire coding sequences of bla_{IMP}, bla_{KPC}, bla_{NDM}, $bla_{\text{OXA-48}}$ and bla_{VIM} genes was used for quantitative real-time PCR (qPCR) experiments when determining the limit-of-detection of the 5plex PCR assay,

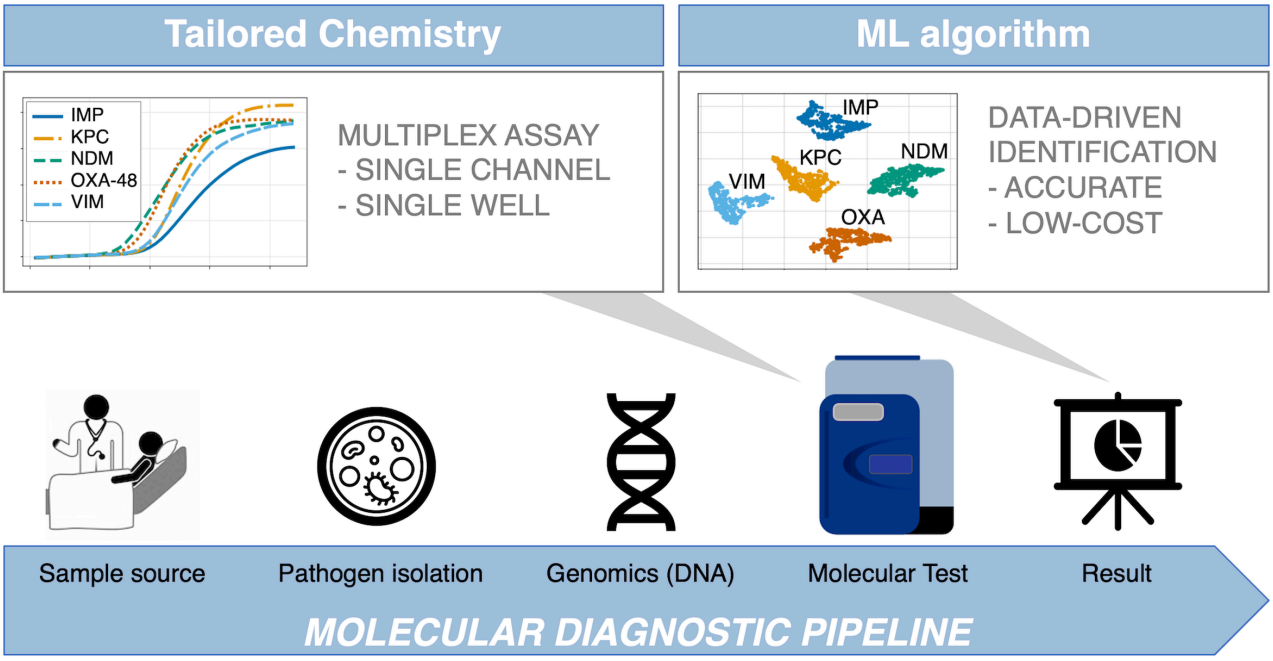


Figure 6.1: Integration of data-driven approaches to standard diagnostic workflows. The blue arrow indicates the conventional diagnosis pipeline from patient to result, where patient sample is collected from different sources (e.g., eye swab, nasopharyngeal swab, throat swab, urine, or rectal swab). Subsequently, samples are cultured, and nucleic acids are extracted in a microbiology lab. Following this, the most suitable genetic test is developed in-silico, comprising of specialised assays capable of multi target detection in a single reaction (first grey arrow). The test is performed in the dPCR instrument, outputting large amounts of data, which are analysed by a machine learning supported algorithm to ensure reliable and accurate results (second grey arrow). This is where the AMCA methodology is applied.

and in dPCR experiments for generating the digital bulk standards and training the mathematical models. The gene fragments (ranging from 900 to 1000 bp) were purchased from Integrated DNA Technologies Ltd (IDT) and resuspended in Tris-EDTA buffer to 10 ng/µL stock solutions (stored at −80°C until further use). The DNA stock concentration for all targets was estimated by dPCR using the Fluidigm's Biomark HD system. The following NCBI accession numbers are used as reference for the gBlocks^{TM} synthesis: NG_049172 (bla_{IMP}), NC_016846 (bla_{KPC}), NC_023908 (bla_{NDM}), NG_049762 ($bla_{\text{OXA-48}}$) and NG_050336 (bla_{VIM}).

6.3.2 Clinical isolates: Bacterial Strains and Culture Condition

A total of 253 non-duplicated Enterobacteriaceae isolates were collected between 2012-2020 from clinical or screening samples routinely processed by Microbiology Department at Charing Cross Hospital, ICHNT (Ethics protocol 06/Q0406/20). Species identification was performed using

MALDI-TOF MS and carbapenemase mechanisms were determined using the Xpert Carba-R (Cepheid) or Resist-3 O.K.N assay (Corisbio). The isolates were subcultured on appropriate growth media and incubated at 37°C overnight, and the genomic DNA was extracted using GenElute Bacterial Genomic DNA kit (Sigma-Aldrich) following the manufacturer's instructions.

6.3.3 Primer Design

The genes used in this study belong to (i) class A carbapenemase encoding for bla_{KPC} type, (ii) class D oxacillinases encoding $bla_{\text{OXA-48}}$ and (iii) class B metalloenzymes encoding bla_{NDM}, bla_{IMP} and bla_{VIM}. The sequences of these genes were downloaded from the NCBI GenBank databse [168]. Based on the comprehensive analyses and alignments of each carbapenemase type using the MUSCLE algorithm, primers were specifically designed to amplify all alleles of each carbapenemase gene family described above [94]. Design and in-silico analysis were conducted using GENEious Prime 2020.1.2 [95]. Primer characteristics were analysed through IDT OligoAnalyzer software (`https://eu.idtdna.com/pages/tools/oligoanalyzer`) using the J. SantaLucia thermodynamic table for melting temperature (T_m) evaluation, hairpin, self-dimer, and cross-primer formation [96]. The T_m of the amplification product of each gene was determined by Melting Curve Predictions Software (uMELT) package [65]. To confirm the specificity of the real-time digital PCR assays, the primers were first evaluated in a singleplex PCR environment to ensure that they correctly amplified their respective loci and that the amplicons showed the predicted T_m and after that in multiplex format. All primers were synthesised by IDT (Coralville, IA, USA). Primer sequences and amplicon information are listed in Table 6.1.

Table 6.1: The 5plex PCR assay primer sets.

CPE Target	**Forward primer sequence (5'→ 3')**	**Reverse primer sequence (5'→ 3')**	**Amplicon size (bp)**	**Amplicon T_m (°C)**
bla_{IMP}	CAGCAGAGYCTTTGCCAGATT	GCCACGYTCCACAAACCAA	203	86.5
bla_{KPC}	GGCTCAGGCGCAACTGTAA	GCCCAACTCCTTCAGCAACAA	273	95.5
bla_{NDM}	CGCGTGCTGKTGGTCGATA	GGCGAAAGTCAGGCTGTGTTG	240	96.0
$bla_{\text{OXA-48}}$	CGATTTGGGCGTGGTTAAGGAT	GTCGAGCCARAAACTGTCTAC	235	88.5
bla_{VIM}	CGAGGYAGAGGGGARCGAGATT	CTSTGCTTCCGGGTAGTGTT	275	94.0

Primers have been developed in this study [144].

6.3.4 Multiplex real-time digital PCR

Each amplification mix for dPCR experiments contained the following: 2 µM of SsoFast EvaGreen Supermix with Low ROX (BioRad, UK), 0.4 µL of 20X GE Sample Loading Reagent (Fluidigm PN 85000746), 0.2 µL of PCR grade water, 0.2 µL of 20X multiplex PCR primer mixture containing the five primer sets (10 µM of each primer), and 1.2 µL of different concentrations of synthetic DNA, samples or controls to bring the final volume to 4 µL. PCR cycling condition consisted of a hot start step for 10 minutes at 95°C, followed by 45 cycles at 95°C for 20 seconds, 67°C for 45 seconds, and 72°C for 30 seconds. Melting curve analysis was performed with one cycle at 65°C for 3 s and reading from 65 to 97°C with an increment of 0.5°C. The integrated fluidic circuit controller was used to prime and load qdPCR 37K digital chips and Fluidigm's Biomark HD system to perform the dPCR experiments, following manufacturer's instructions. Each digital chip contains 48 inlets, where each inlet is connected to a microfluidic panel consisting of 770 partitions or wells (0.85 nL well volume). In this study, a total of seven qdPCR 37K digital chips were used, totalling 336 panels and 189,206 positive amplification reactions (29,165 from training and 160,041 from testing experiments).

6.3.5 Limit of detection for the 5plex PCR assay

Analytical sensitivity was evaluated with 10-fold dilutions of gBlocks^{TM} containing the sequence for the five carbapenemase genes, ranging from 10^1 to 10^6 DNA copies per reaction. Each experimental condition was run in triplicate. 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 five primer sets (5 µM of each primer), and 1 µL of different concentrations of synthetic DNA, clinical sample or controls. The reaction consisted of 10 minutes at 95°C, followed by 45 cycles at 95°C for 20 seconds, 67°C for 45 seconds, and 72°C for 30 seconds. Melting curve analysis was performed with one cycle at 65°C for 60 seconds and reading from 65°C 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).

6.3.6 Quantification of clinical isolates

Clinical isolates were quantified by real-time dPCR following the methodology proposed by Moniri *et al.* [97]. Thus, using Poisson statistics when the microfluidic panel occupancy was $\leq 85\%$ (a maximum of 665 positive amplification events for a given panel) and quantification cycle (C_t) interpolation from digital bulk standards when panel occupancy was $> 85\%$. Digital bulk standards were generated by serial dilutions of the gBlocks^{TM} Gene Fragments containing the sequence for the "big 5" carbapenemase genes ranging from 10^1 to 10^5 DNA copies per panel. The C_t values are calculated by the Fluidigm Digital PCR Analysis software 2.1.1.

6.3.7 Machine learning-based methods

The proposed method, AMCA, trains a supervised machine learning model in which the best fit linear line and the optimal value of intercept and coefficient are calculated to minimise error when combining the predictions of amplification curve analysis (ACA) and MCA [97, 130]. In this study, the ACA consists of applying a k-nearest neighbors (KNN) model (with parameter k=10) to the entire real-time curve from each amplification event, whereas the MCA method consists of applying a logistic regression model to T_m values extracted from each melting curve. Both ACA and MCA output 5 probabilities associated with each target in the 5plex. Therefore, as showed in the flowchart in the Figure 4.2, these probabilities are concatenated into 10 values which are the input to the AMCA method. It is important to note that this classifier is tuned with its own cross-validation step to avoid over fitting. The classifier threshold for positive samples has been set at 5% of panel occupancy.

6.3.8 Statistical Analysis

(i) Sample size: A sufficient number of samples was determined to provide statistically significant results via the binomial proportion confidence interval method [169]. Under the assumption that the test has a sensitivity and specificity of 95% with a 5% margin of error, the number of samples were determined as 72 (which is significantly smaller than 221 used in this study). (ii) AMCA cross-validation performance: Prior to evaluating the in-sample performance of the model, by using the 221 clinical isolates, the out-of-sample classification accuracy was estimated

by 10-fold cross-validation on the training data (using stratified splits). (iii) AMCA accuracy: The two-sided t-test with unknown variances was used to determine statistical significance for comparing the classification accuracy of AMCA against MCA. 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.

6.4 Results & Discussion

6.4.1 Primer characterisation for optimal multiplex PCR assay performance

***in-silico* analysis**. To test the inclusivity and exclusivity of the 5plex PCR assay, primers were subjected to a general NCBI BlastN search against more than 500 sequences per target. Inclusivity results showed over 99% identity coverage for each target (inclusivity alignments are provided in Appendix Figure C.1 - C.5. For exclusivity analysis, BlastN hits with an identity score lower than 80% were regarded as negative [47]. No cross-reactivity was observed with other sequences deposited in the database.

Experimental results in qPCR. The 5plex PCR assay has been validated using a conventional qPCR platform with synthetic DNA templates at concentrations ranging from 10^1 to 10^6 DNA copies/reaction. Appendix Figure C.6 shows the real-time amplification, melting and standard curves obtained from analytical sensitivity experiments. The amplification and melting curves have distinct shape and T_m value distribution for each target, respectively, which is beneficial for AMCA classification. Observed T_m values for bla_{IMP}, bla_{KPC}, bla_{NDM}, $bla_{\text{OXA-48}}$ and bla_{VIM} are 81.4°C, 89.5°C, 90.2°C, 83.8°C and 87.9°C, respectively. Moreover, each primer set (in a multiplex environment) shows an excellent Limit-of-Detection (LOD) of 10 DNA copies/reaction. Corresponding standard curves, illustrating the C_t value as a function of the target concentration, yield an assay efficiency of 87.3%, 103.5%, 105.7%, 98.7%, 88.1%, respectively. PCR products were absent in all the negative controls.

Experimental results in real-time dPCR. The 5plex PCR assay was further validated in the dPCR platform with synthetic DNA templates at concentrations ranging from 10^1 to 10^5

DNA copies per panel, which were chosen such that amplification events in both-single and bulk regions can be observed to capture kinetic information in both domains. Figure 6.2A shows end-point photographs (cycle 45) of panels at increasing amount of DNA. A total of 29,165 positive amplification reactions were performed. As shown in Figure 6.2B, a digital bulk standard curve for each target was build using the real-time dPCR instrument. As this microfluidic platform is capable of real-time data collection, quantification cycle values were used to generate the standard curves by plotting the C_t (or C_q) values against $log[quantity]$ of a ten-fold (10X) serial dilution of each DNA target. It can be observed that there is a clear separation between the single-molecule (10^1 to 10^2 copies/panel) and the bulk regions (10^4 to 10^5 copies/panel) based on C_t value ranges, where 103 copies/panel acts as a transition region across all the targets. In the none-saturated panels a digital pattern (number of ONs and OFFs) is observed at the end of the reaction and the amount input molecules is calculated using binomial and Poisson statistics, whereas in the saturated panels the amount input molecules is quantified using the digital bulk standard curve (as in qPCR) [21]. Digital bulk standard curves yield an assay efficiency of 118.1%, 98.7%, 86.2%, 100.8% and 90.2% efficiency for bla_{IMP}, bla_{KPC}, bla_{NDM}, $bla_{\text{OXA-48}}$ and bla_{VIM} assays, respectively. Appendix Figure C.6 reports the standard curve parameters for each assay, digital count and panel occupancy. Figure 6.3A and 6.3B, respectively, show the amplification and melting curves for the five carbapenem-resistant genes and the average characteristic sigmoidal shape for each target (black solid line) in real-time dPCR. Figure 6.3C represents the distribution of melting temperature, where the T_m range for each target is computed as: bla_{IMP} (81.3°C, 83.2°C), bla_{KPC} (89.0°C, 91.5°C), bla_{NDM} (90.0°C, 92.7°C), $bla_{\text{OXA-48}}$ (83.7°C, 86.6°C) and bla_{VIM} (87.7°C, 90.8°C). After peak detection, negative reactions can be confirmed by identifying curves with no peak.

6.4.2 Clinical isolates

As depicted in Appendix Figure C.10, the 253 pure bacterial strains were identified from MALDI-TOF MS as *Acinetobacter spp.* (N = 2), *Citrobacter spp.* (n = 16), *Enterobacter spp.* (N = 37), *Escherichia spp.* (N = 57), *Klebsiella sp.* (N = 133), *Proteus sp.* (N = 1), *Pseudomonas sp.* (N = 5) and *Serratia sp.* (N = 2). Carbapenemase genes were determined as a single enzyme in 220 strains (bla_{IMP} = 45; bla_{KPC} = 9; bla_{NDM} = 74; $bla_{\text{OXA-48}}$ = 84; bla_{VIM} = 8) and as a combination in one isolate (bla_{NDM} and $bla_{\text{OXA-48}}$). Thirty-two isolates

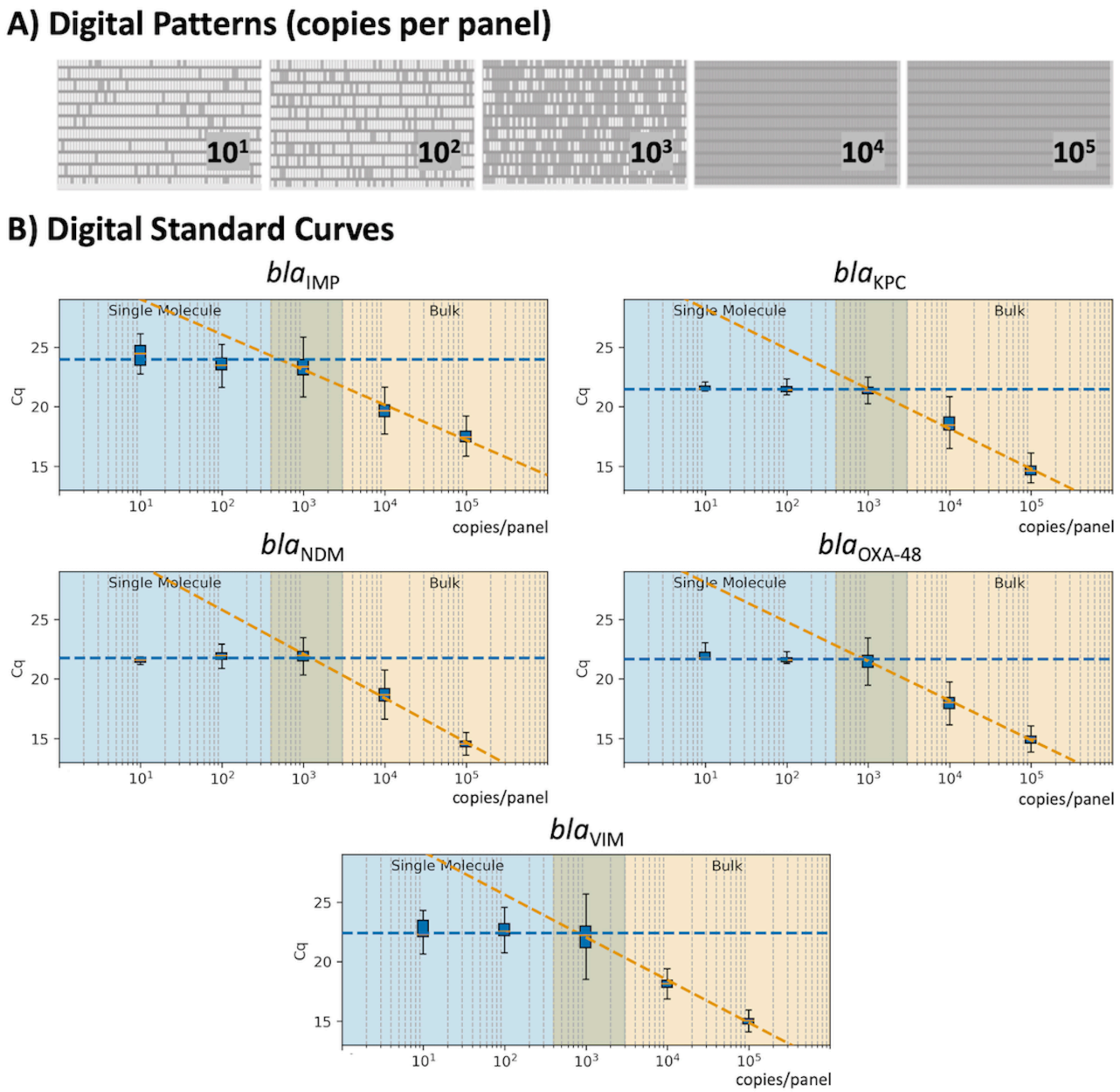


Figure 6.2: Standard Curve in real-time digital PCR. (A) Digital patterns for each microfluidic panel at increasing concentrations (770 reaction chambers per panel; 0.85 nL volume per chamber). (B) Standard curves correlating the C_q values with the concentration of each target; shaded blue area indicates the single-molecule region; shaded orange shows the bulk region; and the middle area displays the theoretical transition between the single-molecule and bulk.

were confirmed as negative for the five carbapenemase genes. A more detailed description of each isolate, including bacterial species, date of sampling, specimen type, antibiotic resistance mechanisms and concentration (copies/µL of extracted DNA) can be found in Appendix Table C.1 - C.5.

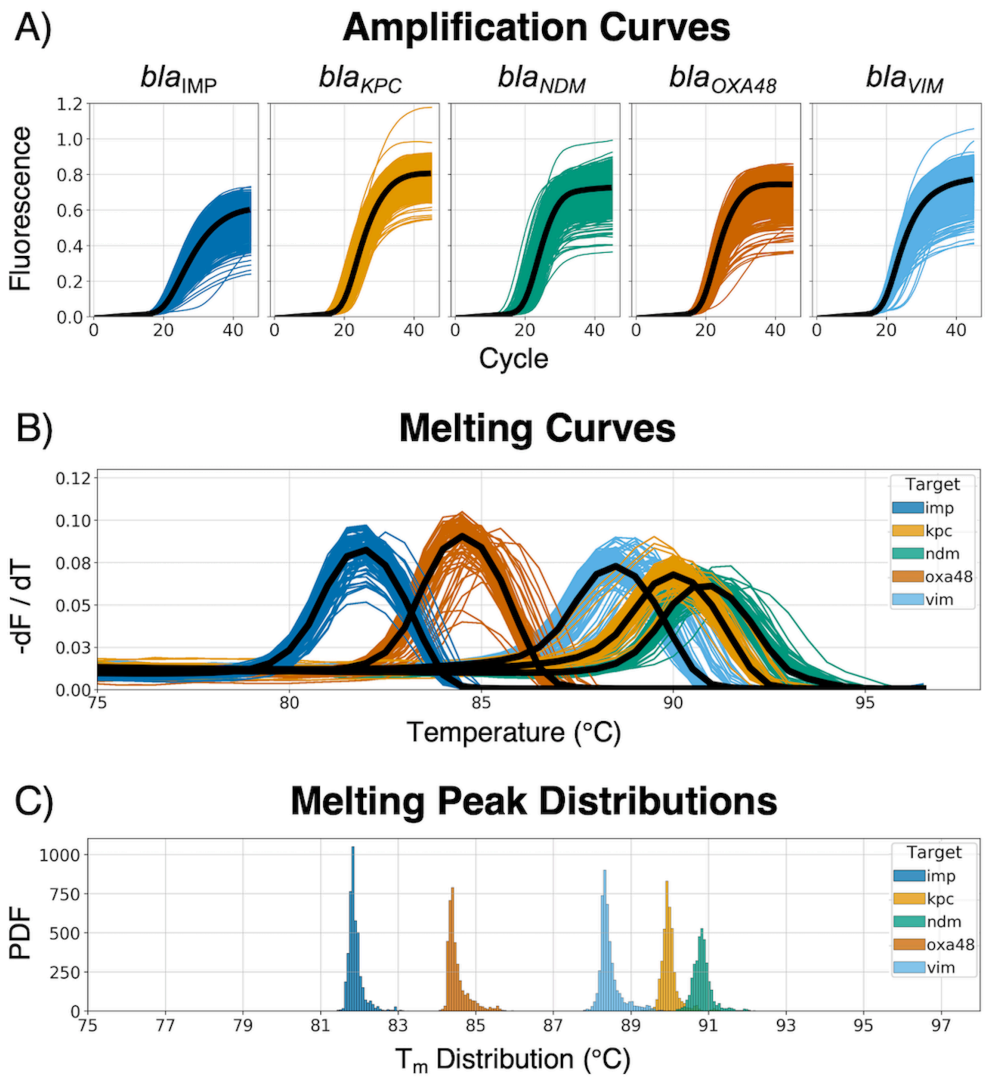


Figure 6.3: Real-time amplification and melting curves obtained from the dPCR instrument. (A) Raw amplification curves at different concentrations from synthetic DNA templates; the black line represents the average trend of the kinetic information based on each specific target-primer interaction. (B) Melting curves across the five different CPO; the black line represents the average trend of the thermodynamic information based on each specific target-primer interaction. (C) Melting peak (T_m) distribution from the dPCR instrument, showing the probability density function (PDF) for each target.

6.4.3 The AMCA model: training and cross-validation

Our study aims to validate the performance of the AMCA method for detection of carbapenem-resistant genes in clinical isolates compared with the MCA approach. To train both models, a total of 99,860 amplification events were generated using synthetic DNA templates, of which 29,165 were positive: bla_{IMP} (N = 4,941), bla_{KPC} (N = 5,940), bla_{NDM} (N = 5,870), $bla_{\text{OXA-48}}$ (N = 4,333) and bla_{VIM} (N = 8,081). Observed overall classification performance of training dataset for the MCA and AMCA methods was 94.9% ± 21.99% and 99.2% ± 8.86%, respec-

tively. Appendix Figure C.7 shows the confusion matrices comparing the true and predicted targets for both methods. It can be observed that the bla_{NDM} and bla_{KPC} targets are misclassified by the MCA methods, whereas the AMCA considerably improves the prediction of both targets: from 804 to 52 amplification events for bla_{NDM}, and from 511 to 46 for bla_{KPC}. No other target was misclassified more than 1.26% for either method.

6.4.4 The AMCA model: validation on clinical isolates

A total of 253 clinical isolates, including 221 positives, and 224,840 amplification events (of which 160,041 positives) were used for the clinical validation. Compared to results obtained with the Xpert Carba-R Cepheid and Resist-3 O.K.N assays, the overall observed accuracy for MCA was 91.7% (CI 87.59% to 94.79%) and 99.6% (CI 97.82% to 99.99%) for AMCA, which represent a 7.9% increase (p-value < 0.01) (see Appendix Figure C.8). A total of 21 clinical isolates were misclassified for the MCA method and considered false positives (FP) as shown in Table 6.2, whereas the AMCA reduced the number of misclassified samples to 1 only (see Table 6.3). All the false positive samples were identified as double infection with CPO because of the overlapping distribution in the T_m, as shown in Figure 6.3. Performance improvement in the AMCA method is due to the addition of real-time amplification data, contrary to the MCA approach that only takes into account the melting curve distribution. Further details on AMCA coefficient contributions (i.e., ACA and MCA weights) are shown in Appendix Figure C.9. Moreover, 32 bacterial isolates not carrying the five carbapenemase genes were used to evaluate the assay specificity. The 5plex PCR assay showed negative results in the absence of the specific target.

Table 6.2: Classification of clinical isolates when using the ML-based MCA method

Target	**N**	**TP**	**TN**[a]	**FP**	**FN**	**SEN**	**SPE**	**Accuracy** (CI)
bla_{IMP}	45	45	32	0	0	100.0%	100.0%	100.0% (95.32 to 100.00%)
bla_{KPC}	9	8	32	1[b]	0	100.0%	96.97%	97.56% (87.14 to 99.94%)
bla_{NDM}	74	54	32	20[c]	0	100.0%	61.54%	81.13% (72.38 to 88.08%)
$bla_{\text{OXA-48}}$	84	84	32	0	0	100.0%	100.0%	100.0% (96.87 to 100.00%)
bla_{VIM}	8	8	32	0	0	100.0%	100.0%	100.0% (91.19 to 100.00%)
$bla_{\text{OXA-48}}$ & bla_{NDM}	1	1	32	0	0	100.0%	100.0%	100.0% (97.24 to 100.00)
Total	221	200	32	21	0	100.0%	60.38%	91.70% (87.59 to 94.79%)

Abbreviations:
N: number of samples; TP: True Positive TN: True Negative; FP: False Positive;
FN: False Negative; SEN: Sensitivity; SPE: Specificity; CI: Confidence Interval.

[a] A total 32 negatives samples are considered across all the groups for sensitivity, specificity and accuracy calculation
[b] This isolate was misclassified as bla_{NDM} and bla_{KPC} double infection
[c] These isolates were misclassified as bla_{NDM} and bla_{KPC} double infections

Table 6.3: Classification of clinical isolates based on ML-based AMCA method

Target	**N**	**TP**	**TN**[a]	**FP**	**FN**	**SEN**	**SPE**	**Accuracy** (CI)
bla_{IMP}	45	45	32	0	0	100.0%	100.0%	100.0% (95.32 to 100.00%)
bla_{KPC}	9	9	32	0	0	100.0%	100.0%	100.0% (91.40 to 100.00%)
bla_{NDM}	74	73	32	1[b]	0	100.0%	96.97%	99.06% (94.86% to 99.98%)
$bla_{\text{OXA-48}}$	84	84	32	0	0	100.0%	100.0%	100.0% (96.87 to 100.00%)
bla_{VIM}	8	8	32	0	0	100.0%	100.0%	100.0% (91.19 to 100.00%)
$bla_{\text{OXA-48}}$ & bla_{NDM}	1	1	32	0	0	100.0%	100.0%	100.0% (97.24 to 100.00)
Total	221	220	32	1	0	100.0%	96.97%	99.60% (97.82 to 99.99%)

Abbreviations:
N: number of samples; TP: True Positive TN: True Negative; FP: False Positive;
FN: False Negative; SEN: Sensitivity; SPE: Specificity; CI: Confidence Interval.

[a] A total 32 negatives samples are considered across all the groups for sensitivity, specificity and accuracy calculation
[b] This isolate was misclassified as bla_{NDM} and bla_{KPC} double infection

6.5 Conclusion

In the last decade, novel pandemic outbreaks and the continued threats of emerging multi-drug resistant microorganisms have significantly increased the demand for molecular tests, in particular PCR-based methods [170, 171]. To respond to this need, the AMCA technology has been designed to increase the throughput of real-time molecular platforms. Seamlessly integrated with conventional diagnostic workflows, this machine learning based approach can enhance multiplexing capabilities of traditional qPCR and state-of-the art dPCR instruments, increasing the number of nucleic acid targets that can be identified in a single fluorescent channel without hardware modifications. Individual primer sets produce amplification products at a sequence-specific amplification rate and efficiency, which generate unique amplification and melting curves for different target concentrations. Such curves can be captured as time-series data by real-time instruments, feed into machine learning models and used to identify multidimensional patterns (or signatures) specific to each primer set. Therefore, enabling the identification of multiple DNA targets per fluorescent channel using only real-time data (i.e., data-driven multiplexing). In this study, a clinical validation on diagnostic accuracy of the AMCA methodology was assessed considering the "big 5" carbapenem-resistant genes (bla_{IMP}, bla_{KPC}, bla_{NDM}, $bla_{\text{OXA-48}}$ and bla_{VIM}) in multiplex PCR. A 5plex PCR assay was developed and characterised in both real-time qPCR and dPCR instruments, and the AMCA performance investigated through the identification of 253 clinical isolates from patients' samples. The MCA was used as a reference method to compare results.

A 99.2% accuracy is achieved for identifying the five carbapenem-resistant genes in the clinical isolates. The AMCA method was shown to enhance the classification performance by 7.9% compared to MCA. The AMCA takes advantage of the volume of raw data extracted from amplification and melting curves, whereas the MCA only considers melting curves. It is interesting to observe that the overlapping melting curve distribution in Figure 6.3B (e.g. bla_{NDM} and bla_{KPC}) represents a misclassification of 1303 reactions (509 bla_{KPC} as bla_{NDM}, and 804 bla_{NDM} as bla_{KPC}) and 21 clinical isolates (20 bla_{NDM} and 1 bla_{KPC} as co-infections) when using the MCA, but it only represents a misclassification of 99 reactions and 1 clinical isolate for the AMCA method. As described in previous publications from Moniri *et al.* [97], these results support the hypothesis that the underlying biological factors driving these methods for target identification are fundamentally different. As observed in Appendix Figure C.9,

machine learning methods can be used to exploit the distinctive information contained on the amplification and melting curves by weighting the predictions from the ACA and MCA to optimally combine them and maximise the AMCA performance.

Although dPCR is not likely to replace all qPCR assays in the clinical laboratory due to associated instrument costs and greater complexity, it has several specific advantages over qPCR. The vast number of partitions reduce the likelihood of co-amplification and inhibitors in a single reaction, facilitating accurate detection of multiple analytes; and the large amount of data enables the use of advance machine learning algorithms to detect subtle kinetic and thermodynamic differences encoded in the real-time amplification data. On the other hand, real-time dPCR platforms enable the use of digital bulk standards and offer a valuable solution for absolute quantification of clinical isolates (equivalently to conventional qPCR standards) even when the panels are saturated, expanding the dynamic range of quantification of the microfluidic chips and eliminating the need of testing the samples at multiple dilutions to ensure that at least one of them falls within the conventional dPCR range (i.e. panels at occupancy $< 85\%$). As shown in Figure 6.2, it is possible to create a standard curve in real-time dPCR by extracting C_q values as a function of the target concentration because there is a clear separation between the single-molecule and the bulk regions. It is expected that coupling real-time dPCR instruments with data-driven multiplexing will expand the use of these platforms in clinical microbiology laboratories.

The results presented in this Chapter represent a step forward in the use of PCR-based data-driven diagnostics for clinical applications. However, there are several aspects that need to be further investigated. Firstly, the performance of the AMCA method is evaluated in clinical isolates using pure bacterial cultures, therefore a follow-up study needs to be conducted to evaluate the performance of the method directly from clinical samples (work currently on going). Secondly, it is important to identify co-presence of infections for patient treatment, however it was encountered in only one sample with a double infection; a larger study will be required to test the effectiveness of the AMCA in double pathogen identification. Depending on the sample concentration, this might not limit multiplexing capabilities in dPCR, but it could represent a challenge when qPCR instruments are used.

CHAPTER LESSON

The development of the data-driven multiplexing made possible to use the AMCA approach as a diagnostics solution for the accurate detection of AMR genes in clinical isolates in a rapid and cost-effective manner. This Part 1 final Chapter highlights the importance of integrating artificial intelligence for diagnosis and how effectively it increases result reliability of state-of-the-art PCR instruments. So far, the thesis has described the evolution of data-driven multiplexing and its final use in clinical isolates. The next Part will focus on the optimisation, bioinformatics implementation and further application of the ACA.

TAKEAWAY QUESTION

"In the previous studies, the incorrectly classified samples from the Amplification Curve Analysis (ACA) were corrected by the Melting Curve Analysis (MCA). Can ACA be improved to guarantee higher accuracy in classification when melting curve capabilities are absent?"

Part II:

Optimising the use of

Data-driven Methodologies

"Champions keep playing until they get it right"

BILLIE JEAN KING

Chapter 7

Enhance Amplification Data Quality

7.1 Chapter Overview

The previous Chapter has shown the potential of data-driven multiplexing in clinical settings, unlocking the use of artificial intelligence for innovative scientific breakthroughs, particularly in the field of molecular diagnostics for infectious diseases. This data-driven approach enhances the level of multiplexing in single fluorescent channel PCR by extracting target-specific kinetic and thermodynamic information contained in amplification curves. However, accurate target classification can be compromised by the presence of undesired amplification events and non-ideal reaction conditions. Therefore, this Chapter proposes a novel framework to identify and filter out non-specific and low efficient reactions from real-time digital Polymerase Chain Reaction (qdPCR) data using outlier detection algorithms purely based on sigmoidal trends of amplification curves. As a proof-of-concept, this framework is implemented to improve the classification performance of the Amplification Curve Analysis (ACA) using the data presented in Chapter 6. Furthermore, a novel strategy, named Adaptive Mapping Filter (AMF), is developed to adjust the percentage of outliers removed according to the number of positive counts in qdPCR. From an overall total of 152,000 amplification events, 116,222 positive amplification reactions were evaluated before and after filtering by comparing against melting peak distribution, proving that abnormal amplification curves (outliers) are linked to shifted melting distribution or decreased PCR efficiency. The ACA was applied to assess classification performance before and after AMF, showing an improved sensitivity of 1.2% when using inliers compared to

a decrement of 19.6% when using outliers (p-value < 0.0001), removing 53.5% of all wrong melting curves based only on the amplification shape. This Chapter explores the correlation between the kinetics of amplification curves and the thermodynamics of melting curves, and it demonstrates that filtering out non-specific or low efficient reactions can significantly improve the classification accuracy for cutting-edge multiplexing methodologies.

The concepts in this Chapter resulted in the following journal article:

- Miglietta L*, Xu K*, Chhaya PM, Kreitmann L, Hill-Cawthorne K, Bolt F, Holmes AH, Georgiou P, Rodriguez-Manzano J. "Adaptive Filtering Framework to Remove Nonspecific and Low-Efficiency Reactions in Multiplex Digital PCR Based on Sigmoidal Trends". *ACS Analytical Chemistry*, 2022 Oct 1. *First joint authorship.

7.2 Introduction

This Chapter demonstrates that undesired amplification reactions from real-time digital PCR (qdPCR) can be detected and filtered out by only evaluating the sigmoidal shape of an amplification curve. This study proposes a novel methodology that can be used with multiplex PCR assays without the need of post-amplification analysis, increasing results accuracy and reliability [99, 117].

During the last decade, gold standard PCR technologies along with other nucleic acid amplification chemistries have resulted in key procedures for molecular diagnostic in both academic and clinical environments [107, 112, 109, 172, 173]. However, limitations such as sample availability, trained personnel, and overall laboratory costs can represent obstacles to the scalability and adoption of PCR-based approaches [174]. To overcome these barriers, multiplexing has been used to unlock the potential of conventional instruments, increasing the number of targets that can be detected in a single reaction [16, 175, 176]. Since the adoption of multiplexing techniques, researchers and industries have successfully applied them to different areas such as molecular diagnostics, RNA signature polymorphism, and quantitative analysis [18, 177]. Moreover, in an effort to increase overall multiplex PCR capabilities, several studies have recently been published on the use of Machine Learning (ML) to identify the biological nature of an amplification event, improving throughput, clinical and analytical reliability, and sample

classification accuracy [15, 178]. As described by Athamanolap *et al.* in 2014, ML methods were applied to High-Resolution Melt Curve to increase both the tolerance of melting temperature (T_m) deviation among targets and reliability of classification for genetic variants (such as polymorphic genetic loci) [66]. In Jacky *et al.* 2021, ML techniques were used to enable high-level multiplexing using TaqMan probes by leveraging on single-feature classification (i.e. final fluorescence intensity or FFI) and PCR platforms with multiple fluorescent channels [53]. While data driven methods have mostly been employed to improve the accuracy of target identification, with the aim to increase multiplexing capability, some groups have also explored such techniques for outlier removal, both in digital and bulk PCR. For instance, Yao *et al.* [179] developed a process-based classification model to identify false positive curves in dPCR (leading to a 64% improvement compared with classical techniques), and Burdukiewicza *et al.* [180] developed an algorithm to automatically detect hook effect-like curvatures, allowing for streamlined quality control in qPCR.

In 2020 Moniri *et al.* proposed a new approach called Amplification Curve Analysis (ACA) for single channel multiplexing without explicitly extracting features [97]. The ACA method comprises a supervised ML classifier to analyse kinetic information encoded in the entire amplification curve, by looking into sigmoidal shapes across different targets [181]. Furthermore, using ACA along with Melting Curve Analysis (MCA), a new method called Amplification and Melting Curve Analysis (AMCA) was developed, enabling higher-level multiplexing in a single channel [130]. While the melting curve is determined by thermodynamic properties of the amplicon, mainly related to its nucleotide sequence, the features of the amplification curve are also influenced by the concentration of templates and amplicon, as well as PCR efficiency (and its cycle-to-cycle variation), thus also providing information on the kinetics of the amplification reaction. The AMCA couples both ACA and MCA coefficients from the classifier to improve classification accuracy. This has been demonstrated through the detection of nine mobilised colistin resistance genes and clinical isolates containing five common carbapenemase resistance genes [144]. Moreover, multiplex PCR (coupled with innovative approaches such as ACA or AMCA) is bringing about a change of paradigm in molecular diagnostics by enabling faster, more accurate and higher throughput detection of several biomarkers in one reaction. Its applications are wide-ranging, including precision medicine in cancer, genetic testing, and syndromic testing in clinical microbiology and infectious diseases, where it enables precise multi-target identification of multiple pathogens and antimicrobial resistance genes.

A barrier to wider adoption of the aforementioned techniques is that they may be limited by instrumentation specifications such as thermal profile performance, available optical channels/filters, and software setup. For example, MCA methodologies are particularly limited in point-of-care devices, as many do not have melting curve capabilities. Furthermore, in assays based on probe-based chemistries (such as TaqMan), where intercalating dyes are not present, the melting curve cannot be generated. In these circumstances, the ACA method still stands as a valid option for multiplexing and therefore it has been the methodology of choice for the work proposed in this Chapter. However, across all these ML-based multiplexing strategies, the ACA approach can be negatively affected by the presence of abnormal amplification products, due to primer dimerization, amplification of undesired targets, the miscalibration of the instrument, and intra-molecule secondary structures. These abnormal behaviours tend to alter the kinetic information of the sigmoidal curves, causing low efficiency or delaying the amplification reaction [31, 182]. As represented in Figure 7.1, when considering shapes of amplification curves from a multiplex assay, similarities among different targets can reduce the accuracy of the ACA classifier, as the presence of non-specific or low efficient reactions results in blurred boundaries among clusters. To overcome this problem, an intelligent algorithm was developed to filter out outliers from multiplex amplification events. Furthermore, to validate the correctness of outlier removal, amplification curves (inliers and outliers) are compared with labelled melting curves ("correct" and "wrong").

This work demonstrated that non-specific and low efficient PCR reactions affect the shape of the amplification curve and therefore, they can be filtered out considering only the sigmoidal trend. Furthermore, an outlier removal algorithm called Adaptive Mapping Filter (AMF) was developed, which in combination with the ACA approach was used to improve the multi-target classification accuracy. This represents a step forward to incorporate ACA in clinical applications and ensure that by filtering in correct amplification curves, higher diagnosic reliability is delivered to the patient. These concepts were explored using data obtained from qdPCR experiments reported by Miglietta et al, 2021 [144]. As a case study, three of the "The big 5" carbapenemase genes (bla_{NDM}, bla_{IMP}, and $bla_{\text{OXA-48}}$) were considered in this study.

The vision of this Chapter is to significantly improve the quality of data from qdPCR instruments and enhance the sensitivity and accuracy of ML-based multiplexing methods relying only on amplification curves. Moreover, extending this framework to other amplification

chemistries and real-time platforms will improve multiplexing capabilities of existing diagnostic workflows and platforms.

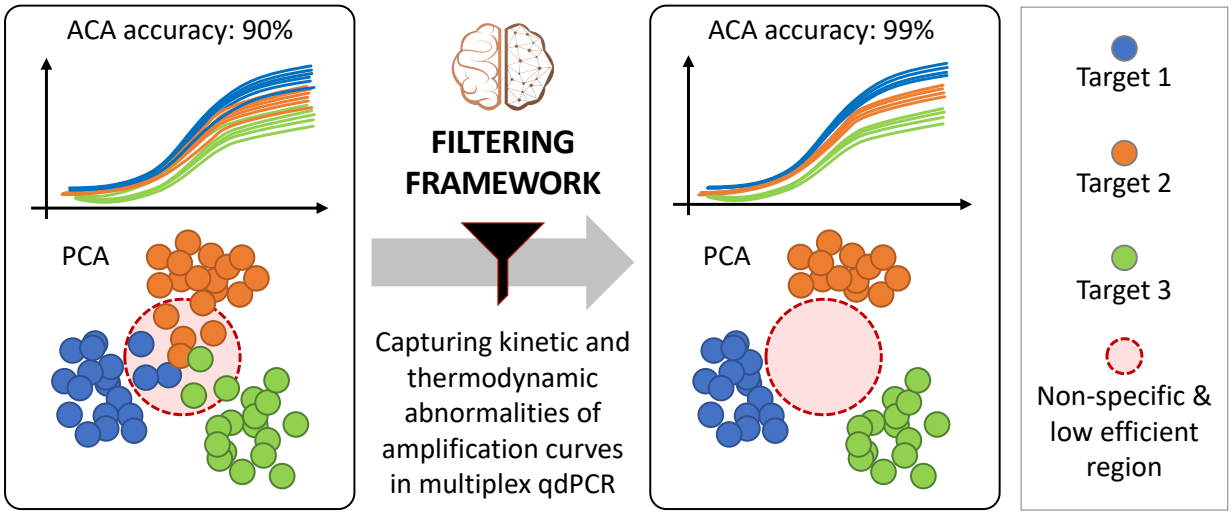


Figure 7.1: Filtering Amplification Curve Concept. Left: raw amplification curves and their corresponding ACA clusters (represented by principal component analysis or PCA) include non-specific and low efficient reactions (confined in the red-circled region). The presence of outliers blurs the boundaries of the different clusters, negatively impacting ACA classification accuracy. By applying the proposed filtering framework, kinetic and thermodynamic abnormalities from amplification events can be captured. Right: Outliers are removed from the original data, resulting in more separated clusters and clearer boundaries. Therefore, ACA classification accuracy is improved.

7.3 Experimental Section

In this section, a new framework for outlier removal in qdPCR is proposed. As depicted in Figure 7.2, this framework took raw amplification curve data as input, and applied baseline and flat/late curve removal in the processing step. Then each processed curve was fitted by a sigmoid function and the fitted parameters, as well as a newly developed feature referred as S_{end}, were used as input for a filtering algorithm which identified outliers automatically. Finally, the framework output the amplification curves after filtering, marked as inliers.

7.3.1 Data input

As a case study, data from Miglietta *et al.* 2021 was used in this work [144]. Data from synthetic DNA (gBlocks^{TM} gene fragments, IDT) containing bla_{NDM} (N $= 18,480$), bla_{IMP} (N

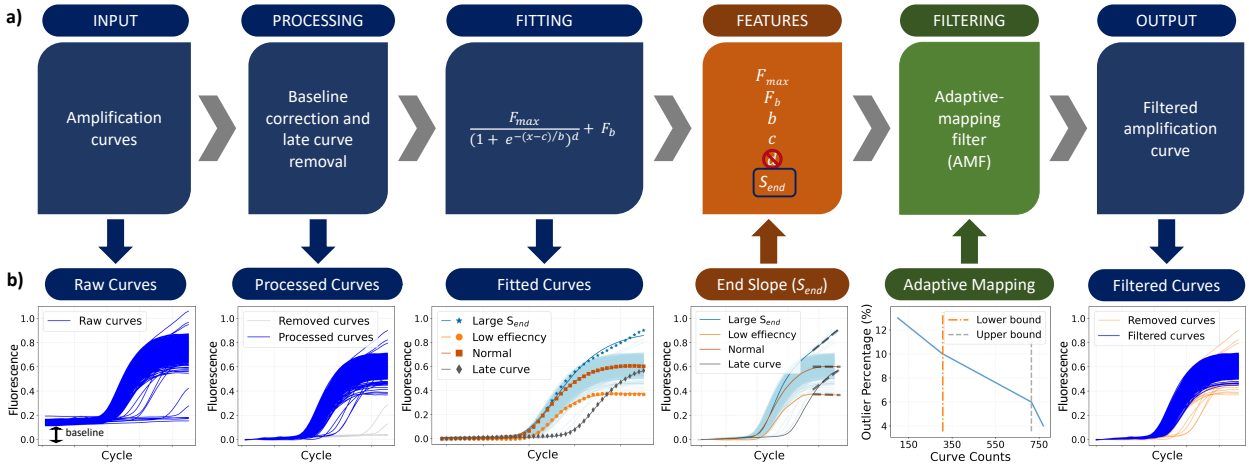


Figure 7.2: Amplification Curve Filter framework. a) Framework steps: raw data input, processing, curve fitting, feature extraction, Adaptive Mapping Filtering (AMF) and filtered curve output. b) Input or output of each step. From left to right, the input of the framework were raw amplification curves, some of which are flat or late curves. By applying the processing step, the baselines were removed, and flat/late curves were discarded. Following this, the processed curves were fitted using a five-parameter sigmoid function, after which each curve was condensed into five features. A new feature S_{end} plus four of the parameters were used to form a set, which is the input of the filtering step. The d parameter was discarded from the feature set for filtering as it is unsuitable for the used algorithms. The AMF was further optimised with a monotonic decreasing map between positive curve numbers within a panel and the outlier percentage. The outputs of the framework are the curves after filtering (inliers).

$= 17,710$), and $bla_{\text{OXA-48}}$ (N $= 17,710$) gene sequences were used as the training dataset. From the original study, a total of 198 clinical isolates labelled with these three targets were used as the testing samples in order to maintain a balanced dataset and due to their high prevalence and clinical significance in UK hospitals. Each sample contained 770 raw curves for a total of 152,460 curves across all the samples, within which 116,222 were positive after the processing step. It is expected that data from clinical isolates are much noisier and thus contain more outliers than those from gBlocks$^{\text{TM}}$.

7.3.2 Data processing

The first step of the framework is processing the raw curves using a baseline correction and a flat/late curve removal to exclude the negative curves of the unprocessed data from the qdPCR output. The baseline of real-time PCR reaction during the initial cycles presents little change in fluorescent signal. The low-level signal of the baseline equates with the back-ground or noise of the reaction. Therefore, the baseline of each raw curve was processed by averaging

the fluorescent value of the first five cycles and subtracting it from the time series. Following this, flat/late curves were removed by applying an upper and lower fluorescence threshold at the 40^{th} cycle, as suggested by the manufacturer [183].

7.3.3 Fitting and feature extraction

Following the processing step, a curve fitting step was introduced to represent the processed amplification curves with sigmoid parameters, which were later down-selected and used as input features for outlier removal and classification algorithms. A 5-parameter sigmoid model, [31] which is shown below, was used to fit the amplification curves:

$$F(t) = F_b + \frac{F_{max}}{\left(1 + e^{-\frac{(t-c)}{b}}\right)^d} \tag{7.1}$$

where t is the PCR cycle number, $F(t)$ is the fluorescence at the t^{th} cycle, F_b is the background fluorescence, F_{max} is the maximum fluorescence, b relates to the slope of the curve, c is the fractional cycle of the inflection point, and d is the asymmetric parameter. To solve the nonlinear least-square-optimization problem for the curve fitting, the Trust Region Reflective (TRF) algorithm with specific bounds was used [184]. Here, the upper and lower bounds were set to $[10, 0.3, 10, 50, 100]$ and $[0, -0.1, -10, -50, -10]$, respectively, as for the search of the 5-parameter set $p = [F_{max}, F_b, b, c, d]$. The initial parameter set p_0 was optimised through pivot fitting on 5% of the training data. After fitting, each amplification curve was given as five parameters, which are condensed representations of curve information. The fitting quality was assessed using Mean Squared Error (MSE) and reported in Figure 7.3. All parameters except for d were considered as input features for outlier removal algorithms because parameter values of outliers may have significant differences from those of normal curves. The d parameter shows a bimodal distribution with two distant peaks, which is unsuitable for the outlier removal step because many of the outlier algorithms require a unimodal distribution of features. Therefore, the d parameter was discarded from the feature set for filtering.

In addition, a new feature called the end slope (S_{end}) was introduced aiming to provide further information about the amplification curve shape. This was calculated by taking the

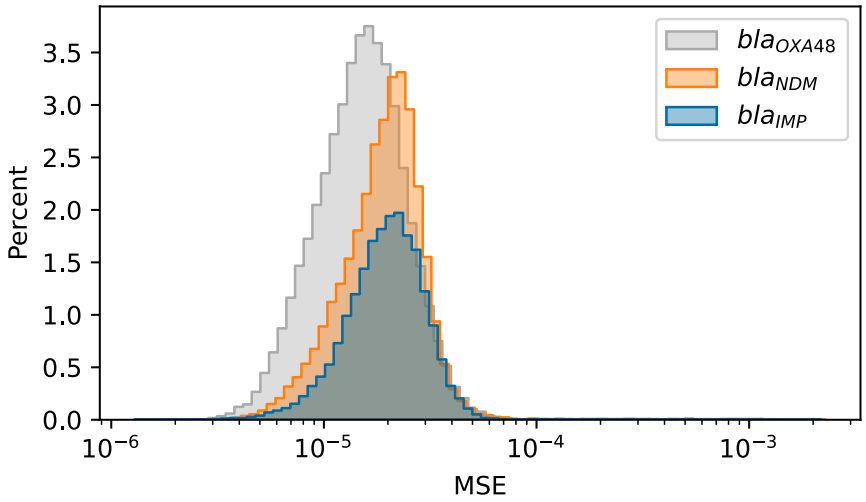


Figure 7.3: Mean Squared Error distributions. Mean Squared Error (MSE) is used to monitor the fitting performance and the quality/correctness of fitted parameters. As shown in the figure, for the majority of curves a MSE smaller than 10-4 can be observed, indicating good fitting considering the fluorescence values ranged between 0 and 1.

average of the first derivatives at the last five cycles of the amplification curve:

$$S_{end} = \frac{1}{5}\left[D\left(N-4\right) \quad D\left(N-3\right) \quad \cdots \quad D\left(N\right)\right] {e_5}^T \tag{7.2}$$

where:

$$D(x) = \left.\frac{d\ F(t)}{dt}\right|_{t=x} \tag{7.3}$$

$$e_5 = [1 \quad 1 \quad 1 \quad 1 \quad 1] \tag{7.4}$$

and N is the total cycle number.

Using the S_{end} feature, the information in the tail of amplification curves was extracted, which contributes to distinguishing inliers and outliers. For example, as illustrated in the "Fitting Curves" step of Figure 7.2b, curves that do not reach the plateau may have larger end slopes. These curves cannot be precisely represented by the fitted parameters since the fitting equation is not capable to capture this non-plateaued trend. Therefore, S_{end} would benefit the result of outlier removal by providing additional information to the feature set. Including S_{end} and discarding d, the final feature set for outlier removal algorithms is $x_f = [F_{max},\ F_b,\ b,\ c,\ S_{end}]$.

7.3.4 Outlier removal algorithms

In this research, seven outlier removal algorithms were considered, which can be split into the following categories according to their principal ideas of filtering: proximity-based, linear, outlier ensembles, and angle-based algorithms. (i) Proximity-based outlier detection algorithms rely on using a distance metric (e.g. Euclidean or Manhattan) to identify outliers. Two proximity-based algorithms, which are Local Outlier Factor (LOF) and Density-based Spatial Clustering of Applications with Noise (DBSCAN), were applied [185, 186]. The LOF algorithm considers the k-nearest neighbors (KNN) to every point in the dataset and computes a local outlier factor for each of them. DBSCAN classifies the points into the core, border, and noise of clusters based on the number of points (min points) within the radius (epsilon) of the considered point. (ii) The linear outlier detection methods used were One-Class Support Vector Machine (OC-SVM) and Elliptical Envelope [187, 188]. OC-SVM applies the concept of finding a hyperplane that separates the inlier points from the origin, such that the hyperplane is closest to the inlier points as possible. The Elliptical Envelope aims to fit the smallest ellipse possible to the core cluster of data points, with any point outside being considered outliers. (iii) Outlier ensemble-based detection methods considered were Isolation Forest and feature bagging [189, 190]. Isolation Forest uses random forests to recursively randomly partition data, after which datapoints with fewer partitions to isolate are marked as outliers. Feature bagging considers multiple outlier algorithms and randomly selects a group of features. From those features, the resulting outlier scores from each algorithm are merged to find the strongest outliers. (iv) Angle-based Outlier Detection considers the angles made by a point with all other pairs of points in the dataset [191]. For each point, the variance is calculated from all the angles obtained, where for a potential outlier the variance is small, since the point is distant from the main cluster of data.

7.3.5 Adaptive mapping filter (AMF)

Most of the outlier detectors explained in the previous section require a hyperparameter called "contamination ratio" or "outlier percentage", which represents the percentage of outliers to be removed from the original data. To adaptively set up this hyperparameter, a mapping strategy that maps the number of positive qdPCR reactions per panel (processed curves) to the contamination ratio was developed and used in the outlier removal algorithm.

In digital PCR, as the number of positive curves increases, the probability of having more than one molecule in a single well increases, resulting in a shift of reaction state from digital to bulk. Moreover, as the reaction goes toward the bulk region, a higher number of positive curves will be present in a panel, which can result in a lower probability of observing a non-specific or low efficient reaction (outlier) in a well [97, 192]. Let us suppose that for each well the probability of observing an outlier is $p(M_i)$, where M_i is the number of processed curves for the i^th sample. Since $p(M_i)$ are independent and identical distributed ($i.i.d.$) for all the wells, the total number of outliers X_i observed in the i^{th} sample follows the distribution of $X_i \sim B(M_i,\ \ p(M_i))$. Therefore, the expected percentage of outliers in the i^{th} sample should be:

$$\text{outlier percentage} = \frac{E\left(X_i\right)}{M_i} = \frac{M_i\ p(M_i)}{M_i} = p(M_i) \tag{7.5}$$

which means that the expected outlier percentage is a monotonic decreasing function to the number of positive curves. As illustrated in the filtering step of Figure 7.2B, a piecewise linear function with empirical turning points was applied.

Coupling the adaptive mapping with an outlier removal algorithm, a novel method called Adaptive Mapping Filter (AMF) was developed, which takes as input the feature set and outputs the inliers.

7.3.6 Melting Labeling

An algorithm was developed to automatically label the melting curves as specific (referred to as "correct") or non-specific (referred as "wrong") ones. By using this methodology, the percentage of wrong melting curves within all the curves of a sample (Wrong Melting Percentage or WMP) was calculated, and this WMP further served as a metric for performance evaluation.

To apply melting labeling, the reference melting peak for each target needs to be determined. For a target $tg \in [bla_{\text{NDM}}, bla_{\text{IMP}}, bla_{\text{OXA-48}}]$, a reference melting peak temperature T_m^{tg} was given by calculating the median value of all the melting peak temperatures of the gBlock$^{\text{TM}}$ curves with target tg. After that, the steps below were followed to label every single melting curve of the clinical dataset:

1. Find the global maximum melting peak's temperature T_m^g of the current melting curve.

2. If $T_m^g \notin [T_m^{tg} - \frac{W}{2}, \quad T_m^{tg} + \frac{W}{2}]$, where W is the tolerance width of the T_m^g distribution, the current curve is labeled directly as a wrong melting curve. Here, considering the instrument resolution for melting curve analysis the W is equal to ± 0.5°C.

3. Otherwise, find the local maximum melting peaks' temperatures on the left and right sides of T_m^g on the current curve, mark them as T_m^l and T_m^r respectively. Note that either T_m^l or T_m^r may not exist. If neither exists, the current curve will be labeled as a correct melting curve.

4. If at least one of T_m^l and T_m^r exists, a set of this (these) local melting peak(s) will be constructed. For each element T_m^e in this set, check whether

$$H_e \in [\ H_{mean} - \ 4H_{std,} \quad H_{mean} + \ 4H_{std}\ ] \tag{7.6}$$

where H_e is the height of the current melting curve at temperature T_m^e, H_{mean} and H_{std} are the mean and standard deviation of $\left[\ H_{1,T_m^e} \quad H_{2,T_m^e} \ \ldots \ H_{M,T_m^e}\right]$, in which H_{n,T_m^e} means the height of the nth melting curve of the sample at temperature T_m^e, and M is the total curve number in the sample. If at least one of the above tests fails, the current curve will be labeled as a wrong melting curve. Otherwise, it will be marked as a correct one.

With the above steps, it is ensured that both curves with large deviations of T_m^g from reference melting peaks and curves with large non-specific local melting peaks can be labeled as wrong. In this way, all the curves had been marked as either "*correct*" or "*wrong*", and further used to calculate the Wrong Melting Percentage (WMP):

$$\text{WMP} \ = \ \frac{N_{wrong}}{N_{total}} \ \times \ 100 \tag{7.7}$$

where N_{wrong} is the number of wrong melting curves within the sample, and N_{total} is the total number of curves in the sample.

It is worth mentioning that the proposed algorithm of automatic melting labeling is not a part of the filtering framework. The labeling was used to calculate the WMP which functioned as a metric for filtering evaluation, where a lower WMP indicates better filtering performance.

7.3.7 Data visualization

Visualization is a vital step for understanding the distribution of a given dataset. In this article, Principal Component Analysis (PCA) with two components was used to visualise the feature sets of the curves before and after applying the outlier removal algorithm into scatter plots. Visual inspection was performed to illustrate how separated the clusters of different targets were. Following this, several metrics for measuring density and degree of separation among those clusters were used to quantitively evaluate how well they were divided.

Specifically, after the PCA of the feature set $x_f = [F_{max},\ F_b,\ b,\ c,\ S_{end}]$ from the amplification curves of each target, the Silhouette Coefficient for each feature set was calculated [193]. The mean value of these coefficients, known as the mean Silhouette Score, was then used to indicate how well the curves of the same targets are clustered. A Higher Silhouette Score implies denser and better-separated clusters observed. Two additional metrics, the Calinski-Harabasz score and the Davies-Bouldin score, were also implemented for clustering evaluation, where a higher Calinski-Harabasz score or a lower Davies-Bouldin score relates to larger inter-cluster distances among targets [194, 195].

7.3.8 Classification of amplification curves – data-driven multiplexing

The ACA method uses kinetic information encoded in the amplification curve to classify different nucleic acid molecules from a PCR test. As shown in Table 7.1, the performance of the ACA was assessed using different curve representations, and the five fitted parameters were used in this study. To illustrate the influence of the AMF on the ACA, a random forest classifier with 100 trees was applied to the feature set $x_c = [F_b,\ F_{max},\ b,\ c,\ d]$, which differs from the x_f used for outlier removal algorithms. Here, parameter d was reintroduced because more curve-related information is needed, provided that the proposed classifier is relatively less sensitive to the feature distributions. S_{end} was discarded for classification because, after outlier removal, abnormal curves with large end slopes were not present in the data set. For the remaining curves, S_{end} were extremely close to zero, thus it was not necessary for S_{end} to be included again. All the other features were normalised with the mean and the variance of the training data before being input into the classifier.

In this research, after applying data processing and feature extraction on both training and testing set, the extracted features of the training set were used to train a Random Forest classifier. This trained classifier was then evaluated on the testing set with or without Adaptive Mapping Filtering (the progress of AMF is totally unsupervised so it can be applied on testing dataset without the true labels). For the testing set, both the inliers and the outliers marked by the aforementioned AMF algorithm were tested. As a comparison, two randomly down-selected datasets with the same numbers of curves as the inliers and the outliers were also constructed and tested.

Table 7.1: Performance comparison between the original ACA method and the proposed method, before applying AMF.

Target	**Precision (%)**		**Sensitivity (%)**		**F1-score**	
	KNN*	RF^x	KNN*	RF^x	KNN*	RF^x
bla_{NDM}	65.5	80.0	83.7	94.4	0.73	0.87
bla_{IMP}	67.8	75.3	96.0	97.8	0.80	0.85
$bla_{\text{OXA-48}}$	84.4	94.8	52.8	70.6	0.65	0.81
Accuracy (%)					**71.7**	**83.9**

* 45 cycles + KNN algorithm
^{x} 5 fitting parameters + Random Forest algorithm

7.3.9 Statistical Analysis

Two-sided Wilcoxon signed-rank tests were used to determine the statistical significance of the changes of WMP and melting peak distributions (distributions of melting peak temperature, T_m, and height, H_m) before and after outlier removal. Two-sided Mann-Whitney U rank tests were used to compare the distributions of C_t, FFI, and maximum slopes between inlier and outlier amplification curves. Those three metrics were chosen for their relationship with the amplification curve efficiency. Many studies suggest that sigmoidal modeling of the entire amplification curve can be used to define the rate of the PCR efficiency. Therefore, low-efficiency PCR reactions are related to low fluorescent values and low maximum slope [30, 196].

Moreover, the significance of the comparison between inliers and outliers in clustering Silhouette coefficients was determined by a two-sided Wilcoxon signed-rank test. This test was also used in the evaluation of the classification performance. A p-value of 0.001 with Bonferroni correction was used as the threshold for statistical significance.

7.4 Results & Discussion

In this study, a new framework is presented to detect outliers from amplification reaction in qdPCR. The outlier identification relies on the AMF, which is comprised of an outlier detection algorithm and a mapping strategy to adapt the contamination ratio hyperparameter to the positive amplification reaction counts (or positive wells) of the qdPCR chip.

7.4.1 Evaluation of outlier detection algorithms

As shown in Figure 7.4a, the detection performance of seven outlier removal algorithms on filtering amplification curves against outlier percentages were evaluated using three metrics: (i) Wrong Melting Percentage (WMP), (ii) Melting Curve T_m variance, (iii) Melting Curve H_m variance. The changing values of metrics for different algorithms with fixed outlier percentages from 0.1% to 40% are shown in Figure 7.4a. After the filtering is applied, the WMP shows a significant reduction from 1.1% (from the unfiltered dataset) to a maximum of 0.9% after filtering across all the algorithms. The graph depicts that outlier percentage and WMP are inversely proportional, but the trend can vary among methods. Proximity-based outlier detectors perform worse overall compared to the rest so they are unable to achieve a dramatic decrease in WMP, even with very large contamination ratios. On the other hand, ensemble-based detectors such as Feature Bagging and Isolation Forest have better performance with the lowest WMP among all the outlier percentages. As shown in the center and right end graphs, the variances of T_m and H_m have a decreasing trend that can be observed as the outlier percentage increases, indicating that both of their distributions are narrowed down. In the T_m variance plot, it is noticed that DBSCAN achieves better performance at lower outlier percentages, but this trend reaches a plateau as the outlier percentage further increases. Once again, ensemble-based methods have similar behavior for the T_m variance as for the WMP. For instance, Isolation Forest outperforms all other detectors after the outlier percentage reaches 12%. Moreover, Isolation Forest and elliptic envelope show the best performance for H_m variance up to 26% contamination ratio.

In this analysis, WMP was used to show the change of wrong melting proportion after applying outlier detection algorithms, indicating the direct effect of the filtering on removing wrong melting curves. It is important to consider that wrong meltings are not related to wrong

target sequences, as the true nature of the amplicons resulting from the PCR reaction can only be established by sequencing, which is impractical in digital PCR. The WMP is used to evaluate the shift of melting peak or the presence of multiple low-intensity peaks which result from non-specific or low-efficiency amplification reactions. This can largely affect the ACA classification depending on the presence of the abovementioned phenomena, therefore filtering such events can result in improved target identification. Moreover, a smaller T_m variance indicates a narrower T_m distribution, which in combination with the WMP methods shows that curves with large deviations from the reference T_m^{tg} are removed by the filtering algorithm. In molecular biology, those curves may be generated after non-specific events such as undesired target interaction or primer dimerization [197]. In addition, melting curves presenting low $-df/dt$ (or H_m) are associated with low efficient amplification reactions. Therefore, narrowed distribution of H_m indicates that low efficient curves, which are present at the tail of the distributions, are removed. All the algorithms provide better performance compared to the original benchmark calculated on the unfiltered data. However, it is noticed that Isolation Forest is always among one of the best methods for all the metrics and does not show any defects, which is common for other algorithms. In the following sections, the Isolation Forest algorithm is used to further demonstrate the proposed framework.

7.4.2 Filtering performance analysis of the AMF

In the following step, AMF was applied to the unfiltered data, and the distributions of inner-sample WMP, T_m and H_m variances are illustrated in Figure 7.2b. Across these three metrics, significant shifts of distributions to smaller values are shown after filtering, supported by all the p-values < 0.0001. This indicates that the proposed AMF can significantly remove both non-specific and low efficiency curves only by looking at amplification curves. This proves the hypothesis that amplification curves contain not only kinetic but also thermodynamic information as numbers of outliers correspond to wrong melting curves.

An example of the AMF visual performance on a clinical isolate containing the carbapenemase gene $bla_{\text{OXA-48}}$ is illustrated in Figure 3c. Columns represent both amplification and melting curves of: (i) correct melting and predicated inliers (N = 731, 94.9%), (ii) wrong melting and predicted outliers (N = 19, 2.5%), (iii) correct melting and predicted outliers (N = 12, 1.6%), (iv) wrong melting and predicted inliers (N = 8, 1%). The first column shows

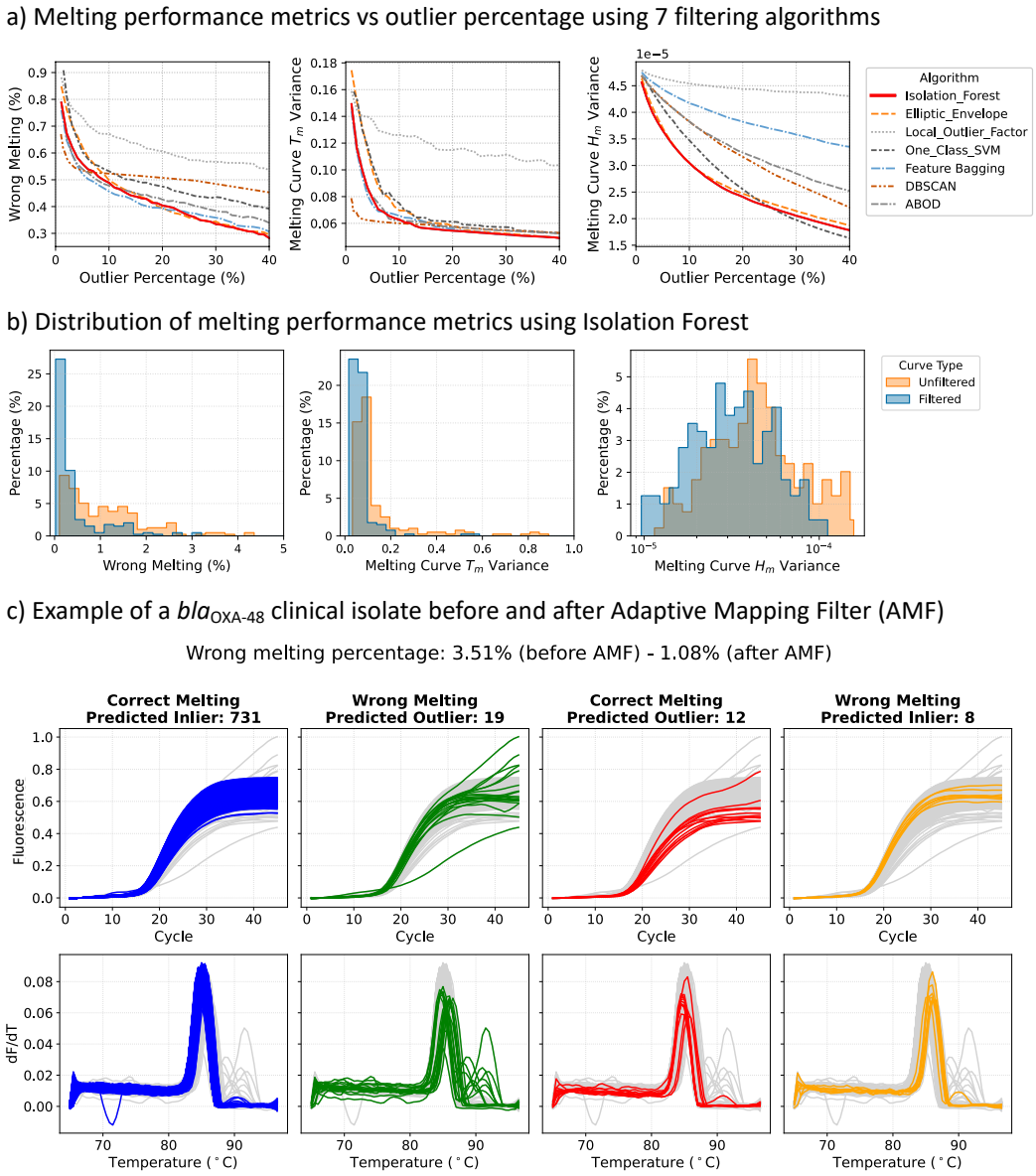


Figure 7.4: Melting curve analysis on filtering results. a) Melting performance shown with Wrong Melting Percentage (WMP), T_m and H_m variances versus fixed outlier percentage. As the outlier percentage increases, all the metrics show decreasing trends which tend to plateau after a certain percentage. As illustrated by the firm red line, Isolation Forest performs the best overall for the three metrics. b) The distribution of melting performance metrics shows that, after filtering, the WMP becomes significantly smaller, and T_m and H_m have a narrower distribution. c) An example of $bla_{\text{OXA-48}}$ clinical isolate. Each column shows the amplification curve and corresponding melting curve of the correct melting and predicted inliers (N = 731), wrong melting and predicted outliers (N = 19), correct melting and predicted outliers (N = 12), wrong melting and predicted inliers (N = 8).

the correctly identified inliers representing specific products of PCR tests. In the second column, non-specific reactions are correctly identified and labeled as outliers, which emphasises the effectiveness of the filtering. It can be noticed that a small number of specific curves were

predicted as outliers, as shown in the third column of Figure 7.2c. This phenomenon does not deny the efficacy of the filter, as these "incorrectly" removed curves have: (i) significantly larger C_t values, (ii) significantly smaller FFI, (iii) and smaller values of maximum slope compared to the inliers. Across the entire clinical isolate dataset (N = 116, 222), compared to melting curve analysis, 115,535 were correctly predicted inliers and 791 were correctly predicted outliers. Furthermore, 5,861 were wrongly classified as outliers whereas 687 were wrongly classified as inliers. Further statistical analyses on the entire dataset also endorse these significant differences between inliers and outliers for C_t, FFI and maximum slope values, as illustrated in Table 7.2. This indicates that AMF removes certain curves because they are of low amplification efficiencies even though they have "correct" melting peaks. A few curves labeled as "wrong" melting may be predicted as inliers, as shown in the fourth column of Figure 7.2c. This can be explained by the relatively low temperature resolution of the equipment which results in mislabeled wrong melting curves due to the large quantization noise of T_m^g during temperature measurement. In fact, by visually inspecting the last column of Figure 7.2c, it can be seen that amplification curves are of very similar shapes to correctly predicted inliers (shown in the first column of Figure 7.2c). The WMP of the illustrated sample has dropped from 3.51% to 1.08%. Overall, in this demonstrated dataset 1.2% of wrong meltings were reported before filtering, and after applying AMF, the WMP was reduced by half to 0.59%.

Table 7.2: Comparison of C_t, FFI and maximum slope between predicted inliers and outliers with correct melting peaks.

Target	C_t ($mean \pm std$)		FFI ($mean \pm std$)		Max Slope ($mean \pm std$)	
	Inliers	Outliers	Inliers	Outliers	Inliers	Outliers
bla_{NDM}	21.45 ± 3.28	26.40 ± 6.11	0.67 ± 0.06	0.60 ± 0.12	0.07 ± 0.01	0.06 ± 0.01
bla_{IMP}	30.33 ± 2.23	31.25 ± 3.43	0.44 ± 0.06	0.41 ± 0.07	0.0276 ± 0.003	0.0271 ± 0.01
$bla_{\text{OXA-48}}$	18.82 ± 3.08	21.03 ± 4.34	0.65 ± 0.08	0.51 ± 0.16	0.05 ± 0.01	0.04 ± 0.02

For all the targets, inliers have significantly smaller C_t and larger FFI and max slope.
All p-values < 0.0001.

7.4.3 Feature set visualization

To visualise the effect of the AMF, PCA-based feature visualization before and after filtering is depicted in Figure 7.5. On the left of the figure, the unfiltered data shows larger overlapping within clusters of different targets and a higher number of outliers compared to the data after filtering. The segmented squares are used to emphasise the differences in cluster overlapping

before and after the AMF, where clearer boundaries between bla_{IMP}, and both $bla_{\text{OXA-48}}$ and bla_{NDM} can be seen. These differences highlight that: (i) outliers can be effectively removed by the AMF, and (ii) removing outliers enhance the separation and reduce the overlap among different target clusters, which will ease the classification of the ACA method. To numerically evaluate the degree of separation across target clusters, the mean Silhouette score of all the datapoints was calculated before and after filtering, showing an increment from 0.378 to 0.399 (p-value < 0.0001). In addition, the Calinski-Harabasz score increased from $101,002.729$ to $130,134.802$, and the Davies-Bouldin score dropped from 0.886 to 0.839. All those results indicate that AMF makes target clusters denser and better separated.

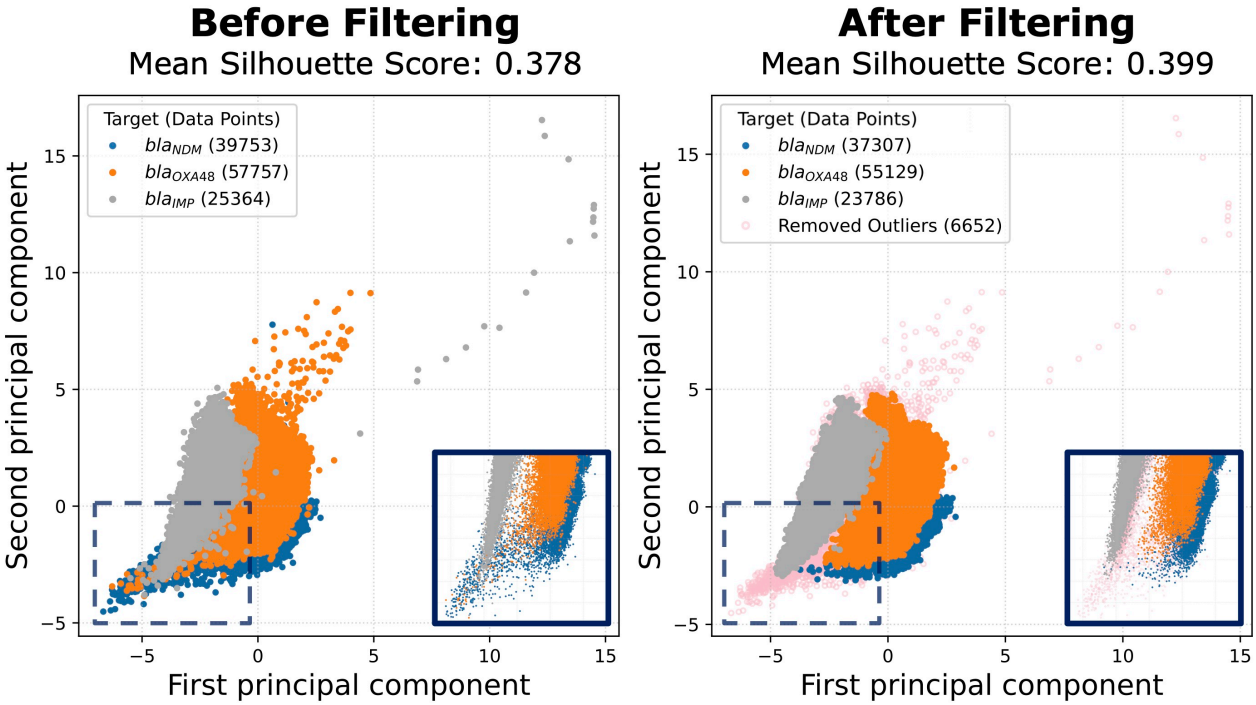


Figure 7.5: Data visualised using 2-D Principle Component Analysis before and after filtering. The processed data plot shows that most outliers have been removed from the original unfiltered data, and the clusters are more separated with clearer boundaries and fewer overlaps. The segmented squares on the bottom side of both figures show the areas where cluster overlapping is more evident, thus they are zoomed. The mean Silhouette Score rises from 0.378 to 0.399 after filtering.

7.4.4 ACA classification

After demonstrating that removing outliers improves the overall distance among clusters, further exploration on its impact on the ACA classification was explored for both inlier and outliers against randomly down-selected datasets with the same numbers of curves. In Figure 7.6a, the confusion matrix shows that the sensitivity for the inliers is 88.96%, which is an increase of 1.13% compared to the randomly down-selected ones in Figure 7.6b. For all the targets, a significant sensitivity improvement can be observed of 1.06%, 0.95% and 1.39% for bla_{IMP}, bla_{NDM}, and $bla_{\text{OXA-48}}$, respectively. Moreover, the overall classification accuracy was 84.94% for inliers and 83.76% for randomly down-selected curves, showing a 1.18% improvement (p-value < 0.0001), which is in line with the overall WMP before filtering (WMP = 1.2%). Applying the filter will help increase the overall performance and specificity of the dataset. This supports the hypothesis that melting information or thermodynamics are contained in the amplification curve.

To show that the removed outliers are less informative for target recognition and harmful for the overall classification, Figure 7.6c and 7.6d show the confusion matrices of the classification using both removed outliers and a randomly down-selected dataset with the same size. As expected, the performance for outliers is significantly worse than the randomly down-selected ones, with only 68.2% and 54.78% sensitivity and accuracy respectively for outliers (p-values < 0.0001). This dramatic sensitivity decrement of 19.57% strongly suggests that outliers have less useful information for the classification of the selected targets.

Furthermore, the statistical analysis on the two randomly down-selected datasets shows no significant differences of in-sample accuracy with p-value = 0.448, which is in line with the central limit theorem as they originate from the same distribution. This is a further proof that the efficacy of the proposed framework is not related to the size of the data.

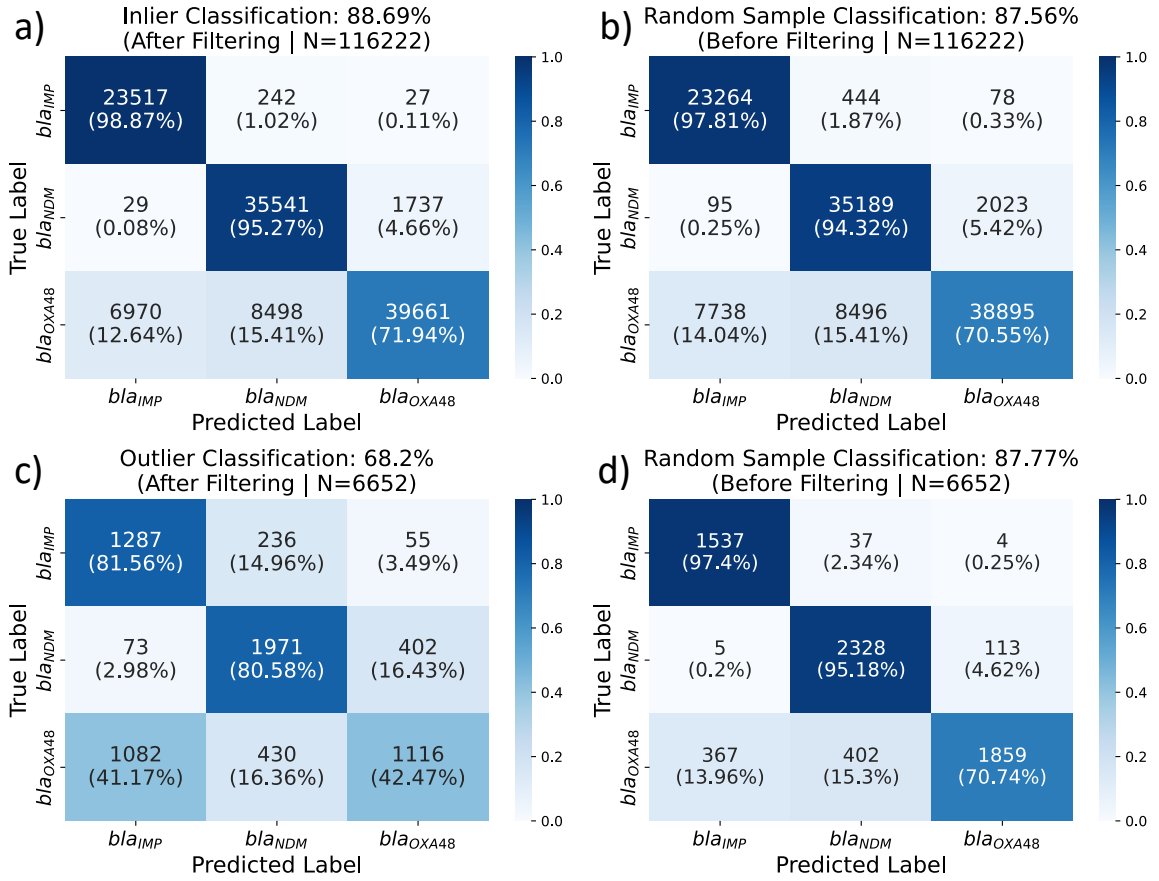


Figure 7.6: Confusion matrices for inlier and outlier classification. The four confusion matrices are shown for: a) inliers, b) randomly down-selected data with the same curve numbers as inliers, c) outliers, and d) randomly down-selected data with the same curve numbers as outliers. The title of each matrix reports the sensitivity of the model. Moreover, each square of the matrix has the number of predicted curves for the corresponding true label and the respective sensitivity of the square.

7.5 Conclusion

This chapter presents a novel framework to adaptively remove abnormal curves from PCR amplification reactions. The method takes the raw input from a qdPCR run and processes it in three steps: background subtraction, late curve removal, and sigmoidal fitting. Moreover, a new feature called end slope (S_{end}) is developed in this study which, along with sigmoidal parameters, is used in the Adaptive Mapping Filter (AMF). The AMF is capable of removing non-specific and low efficient amplification curves, which are labeled as outliers. Melting curves of the outliers, previously removed, were compared with melting curves of inliers using both Wrong Melting Percentage (WMP) and melting peak distributions. Results show that non-specific and low efficient curves can be removed from amplification reaction by purely considering the sigmoidal trend. Further validation of the framework performance was conducted by assessing

the classification accuracy and sensitivity of the ACA classifier on both inliers and outliers. This reinforces the hypothesis that removing abnormalities of amplification reaction in real-time PCR instruments would benefit data-driven multiplexing by removing undesired information.

This research uses data from qdPCR presented in the previous Chapter to demonstrate the effectiveness of the proposed framework, but its generality has not been tested in other settings. Future work will focus on evaluating this methodology on real-time data originating from various qPCR instruments, from different chemistries (such as isothermal amplification), and from point-of-care devices. Digital PCR allows to generate amplification curves at low concentrations of samples, enabling the use of the developed framework. However, future work will focus on the application of this novel method in bulk reactions. Moreover, in the event of secondary amplification, the curve may show a second increasing phase with a large FFI and different shape from sigmoid. However, as shown in Figure 7.2b fitting step, the approximate shape of the distorted curve can still be depicted by the 5-parameter model, with still relatively small fitting error. After fitting, certain parameter values of the secondary amplification events will be different and distant from normal reactions and these events can be identified easily by the outlier detector. Regarding the presence of multiple targets in a single well, a normal sigmoidal trend is expected, therefore the fitting error (MSE) will be low without affecting the AMF progress. However, the ACA classification of such event may be challenging. The previous Chapter demonstrated that presence of double targets can be resolved by using AMCA approach (with the help of melting curves), and other solutions such as FFI modulation by changing probe concentration in TaqMan assay may also help tackle this issue. Finally, upcoming work will focus on introducing advanced machine learning techniques to enhance the classification efficacy of the ACA classifier, and then on making this approach more reliable for use in clinical diagnostics.

CHAPTER LESSON

This Chapter reveals the interconnection between the kinetics of the amplification curve and the thermodynamics of the melting curves. For the first time, a framework is introduced which is capable of removing abnormalities in kinetic and thermodynamic information by purely screening amplification curves, improving data-driven methods performance and quantification accuracy in qdPCR.

TAKEAWAY QUESTION

"If the Adaptive Mapping Filter (AMF) notably improves the ACA performance (relying only on the amplification curve), can data-driven multiplexing be applied for multiple target detection using TaqMan probe assays (where melting curve capabilities are absent)?"

Chapter 8

Smart-Plexer: a Tool to Develop Multiplex Assays

8.1 Chapter Overview

Developing multiplex PCR assays requires an extensive amount of experimental testing, the number of which exponentially increases by the number of multiplexed targets. Dedicated efforts must be devoted to the design of optimal multiplex assays for specific and sensitive identification of multiple analytes in a single well reaction. Inspired by data-driven approaches, this Chapter describes the process of reinventing the way of designing and developing multiplex assays by proposing a hybrid, easy-to-use workflow, named Smart-Plexer, which couples empirical testing of singleplex assays and computer simulation of multiplexing. The Smart-Plexer leverages kinetic inter-target distances among amplification curves to generate optimal multiplex PCR primer sets for accurate multi-pathogen identification. The optimal single-channel assays, together with a novel data-driven approach, Amplification Curve Analysis (ACA), were demonstrated to be capable of classifying the presence of desired targets in a single test for seven common respiratory infection pathogens.

The concepts in this Chapter resulted in the following submitted article and patent:

- Miglietta L, Chen Y, Luo Z, Xu K, Ding N, Peng T, Moniri A, Kreitmann L, Cacho-Soblechero M, Holmes A, Georgiou P, Rodriguez-Manzano J. "Smart-Plexer: a break-

through workflow for hybrid development of multiplex PCR assays". *Under review* - https://doi.org/10.21203/rs.3.rs-1765213/v1.

- Rodriguez-Manzano J, Moniri A, Miglietta L and Georgiou P. "Method of assay design", GB2108339.9, Assignee: Imperial Innovations Limited, 2021

8.2 Introduction

Quantitative Polymerase Chain Reaction (qPCR) allows to continuously monitor the kinetic signature of a specific amplification event due to the mutual interaction of oligonucleotides and their specific template [107, 198, 199]. The extraordinary ease and reliability of this golden standard method for Nucleic Acid Amplification Tests (NAATs) have improved routine diagnostics in several fields and, more recently, played a crucial role during the COVID-19 pandemic, one of the ten deadliest infectious diseases in history [200, 201, 202]. This epidemic has further highlighted the need for more cost-effective and provisional diagnoses, and for enhancing the diagnostic capabilities of conventional instruments along with point-of-care devices [146, 203, 204, 205]. As the pandemic comes to an end, the focus on developing NAATs for the simultaneous detection of multiple respiratory pathogens alongside COVID-19 has drastically increased [206, 207, 208]. There is an emerging demand for rapid, affordable, and reliable molecular tests for multiple identifications of infectious disease [204].

Current screening strategies of multiple pathogens are reported to be expensive, sample consuming and, in some cases, inaccurate [209, 210]. As a result, multiplex PCR is emerging as an inexpensive alternative for multi-target identification [141, 211]. Many efforts have been made in developing novel methods to increase the number of targets detected by multiplex assays and to enhance the accurate identification of multiple infectious sources in a single test [212, 213]. Advances in multi-pathogen detection include the use of High-Resolution Melting Analysis (HRMA), fluorescent probe-based method, or restriction enzyme digestion [214, 215, 216]. Recently, the emergence of machine learning approaches in clinical diagnostics has highlighted the potential of data-driven multiplexing, which, compared to conventional methods, unbars limitations in terms of throughput, costs, time and reliability [3]. A few methods have been proposed using either melting curve analysis (intercalating dye-based chemistries) or the final fluorescence intensity (probe-based assays) as features for machine

learning algorithms [53, 67]. Moreover, using cutting-edge signal processing and tailored amplification chemistries, state-of-the-art identification performance has been achieved by leveraging the kinetic information encoded in the entire amplification curve from multiplex PCR assays. A novel learning-based methodology called Amplification Curve Analysis (ACA) has been recently reported as a digital tool to expand multiplex capabilities of real-time PCR-based diagnostic platforms, increasing the number of detectable targets per fluorescent channel in a single reaction without hardware modification [97, 144, 130].

However, the development of multiplex PCR assays is still restrained as extensive experimental testing is required to assess the assay's analytical performance, such as cross-reactivity, specificity, and sensitivity [212, 217]. One of the biggest challenges in multiplexing is the complexity of assay design, which dramatically increases with the number of targets, making the development costly, lengthy and resource consuming in the wet laboratory [141, 218]. For N_t multiplexed targets, if N_{Ps} candidate primer sets are designed for each of them (which is trivial progress for well-designed singleplex assays), the total number of possible multiplex assay combinations is $N_c = {N_{Ps}}^{N_t}$ (e.g. $N_c = 16,384$ when $N_{Ps} = 4$ and $N_t = 7$). The N_c increases exponentially with N_t, making it impractical to find the optimal combination by wet-lab experiments in high-level multiplexing. Therefore, an *in-silico* simulation method is required for fast screening and for narrowing down selections of multiplex assays.

This problem is addressed with the Smart-Plexer, a mathematical algorithm capable of simulating thousands of possible multiplex assay combinations based on singleplex real-time digital PCR (qdPCR) data. The use of this new methodology is addressed by developing a TaqMan-based multiplex assay, in a single fluorescent channel, for the specific and sensitive detection of seven common respiratory tract infection (RTI) pathogens. This work is two-fold: First, the Smart-Plexer is validated by comparing the performance of all possible simulated and empirical combinations in 3plex, showing a strong correlation between *in-silico* and lab-tested multiplexes; second, the proposed pipeline is assessed in high-level multiplex (7plex) by evaluating the ACA classification performance on synthetic DNA and clinical samples. Out of 4,608 simulated combinations, an optimal multiplex assay could be developed using this novel framework to detect seven common respiratory pathogens accurately in qdPCR.

8.3 Experimental Section

8.3.1 Synthetic Double-stranded DNA Templates & Clinical Samples

Synthetic DNA. Double-stranded synthetic DNA was used in this study to develop and assess the performance of all singleplex assays. In particular, the entire coding sequence used are the hexon protein gene (HEX gene) for human adenovirus (HAdV), and the nucleocapsid protein gene (N gene) of human coronavirus OC43 (HCoV-OC43), HKU1 (HCoV-HKU1), 229E (HCoV-229E), NL63 (HCoV-NL63), Middle East respiratory syndrome-related coronavirus (MERS-CoV) and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The following NCBI accession numbers were used as references for the gBlock synthesis: NC_001405, NC_006213, NC_006577, NC_002645, NC_005831, NC_019843 and NC_045512, respectively. The synthetic constructs were used for qPCR experiments when determining the limit-of-quantification of each PCR assay, and in qdPCR experiments for generating the dataset used in the simulation of the multiplexes and their empirical testing. The gene fragments (ranging from 1,134 to 1,558 bp) were purchased from Integrated DNA Technologies Ltd. (IDT) 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).

Commercial Clinical Sample. Whole pathogen control panels were purchased from Randox Laboratories Ltd, including MERS-CoV (catalog no. QAV154181), CoV-OC43, NL63 (catalog no. QAV164189), and SARS-CoV-2 (catalog no. SCV2QC). Samples were extracted using the QIAamp Viral RNA Mini Kits (catalog no. 52906). Viral nucleic acid was extracted using the manufacturer-recommended protocol. Viral RNA was reverse transcribed to cDNA using Fluidigm reverse transcription master mix (catalog no. SKU 100-6299). Viral cDNA was further pre-amplified using Fluidigm Preamp master mix (catalog no. PN 100-5744). Reverse transcription and pre-amplification were conducted according to the Fluidigm manufacturer's protocol (Fluidigm document number: 101-7571 A2 and 100-5876 C2).

8.3.2 PCR Assay Design

The sequences of each gene were downloaded from the NCBI GenBank website [168]. Based on the comprehensive analyses and alignments of each type using the MUSCLE algorithm [94], primers were specifically designed to amplify all sequence variations within each gene belonging to their specific target (inclusivity) and to exclude closely related but not inclusive sequences (exclusivity). Design and *in-silico* analysis were conducted using GENEious Prime 2022.0.1 [95]. Primer characteristics were analysed through IDT OligoAnalyzer software using the J. SantaLucia thermodynamic table for melting temperature (T_m) evaluation, hairpin, self-dimer, and cross-primer formation [96]. To confirm the specificity of the real-time digital PCR assays, the primers were first evaluated in a singleplex PCR environment to address their specificity and sensitivity for both singleplex and multiplex assays. All primers were synthesised by IDT (Coralville, IA, United States). Primer sequences for both 3plex and 7plex are provided in Appendix Table D.1 and D.3, plus assay details in Appendix Table D.2 and D.4, respectively.

8.3.3 Real-time Digital PCR and Limit-of-Quantification (LoQ)

For real-time amplification experiments, the BioMark HD (Fluidigm) and the QIAquant 96 5plex (catalog no. 9003011) were used. The master mix used was the PrimeTime Gene Expression Master Mix from Integrated DNA Technologies (IDT, catalog no. 1055772) supplemented with ROX passive reference dye and pre-mixed following manufacture guidelines. The qdPCR was performed with Fluidigm qdPCR 37k^{TM} integrated fluidic circuits (IFC) (catalog no. SKU100-6152) and was supplemented with Fluidigm 20X GE loading buffer (PN 85000746). The priming and loading steps of the IFC were followed as the supplier's protocol (Fluidigm document number: 100-6896 Rev 03). Each amplification mix for the qdPCR experiment contained 3 µL 2X IDT PrimeTime Gene Expression Master Mix (with passive ROX), 0.6 µL 20X GE, 0.6 µL 10X Primer mixture, 1.8 µL DNA templates from synthetic DNA, pre-amplified cDNA, or controls, and to bring the final volume to 6 µL. A total of 4.5 µL of reaction mix was transferred to each inlet (or panel) of a Fluidigm 37k^{TM} IFC for the thermal cycling step. Thermal-cycle conditions consisted of a hot start step for 3 minutes at 95°C, followed by 45 cycles at 95°C for 15 seconds and 60°C for 45 seconds. Real-time data of the amplification events were exported as a text file for each bulk by Fluidigm Digital PCR Analysis software

(version 4.1.2).

Experiments in qPCR are used to evaluate the Limit of quantification (LoQ) of the selected 7plex assay. Standard curves were generated with synthetic DNA ranging from 10^7 to 10^1, apart from SARS-CoV-2 whose concentration was from 10^5 to 10^1 because of limitations due to pandemic suppliers and contamination in the manufactures. PCR data were extracted and processed according to the data processing step. Standard curve plots and statistical values are reported in Appendix Figure D.2. The Absence of amplification signals was detected in Negative Template Control (NTC).

8.3.4 Data Processing

The processing of raw amplification curves is comprised of three parts. Firstly, to ensure all curves start from approximately zero fluorescence value and to normalise the starting cycles of the curve across the entire time series, the background information was removed, which can be expressed as:

$$Fl_{br}\left(t\right) = Fl\left(t\right) - avg_{back} \tag{8.1}$$

where $Fl_{br}\left(t\right)$ represents a curve with the background removed and $Fl\left(t\right)$ is the raw fluorescence values for each cycle $t = 1,\ 2,\ \cdots,\ T$. Here T indicates the total number of cycles for each amplification curve (45 in this case), and avg_{back} is the average background value. In order to avoid instrumental noise commonly found at the beginning of the PCR reaction, the avg_{back} value was estimated as the average value of the first several cycles' fluorescence, excluding the initial ones. In this case, five cycles were considered for the flat phase and the first three cycles were skipped. Secondly, late amplification filtering was applied to select curves that reached the plateau phase. The basic idea is to estimate the cycle threshold value (Est_{Ct}) for each curve, which can be represented as:

$$Est_{C_t} = min\ ts.t.\ \ \frac{Fl_{br}\left(t\right) - F_{min}}{F_{max} - F_{min}} \geq F_{th} \tag{8.2}$$

where $t \in \{1,\ 2,\ \cdots,\ T\}$, and F_{max} and F_{min} represent maximum and minimum fluorescence values of the entire reaction respectively for each curve. F_{th} is the fluorescence threshold and curves whose Est_{Ct} are above the cycle threshold ($C_t = 30$ as suggested by the manufacturer) were removed. Lastly, a filter was applied to remove non-sigmoidal curves with excessive noisy

signals. The sigmoidal trend of a noisy curve may contain certain notches. Based on this feature, the first derivative of each curve was estimated:

$$Fl_{br}{}'(t) = Fl_{br}(t) - Fl_{br}(t-1), \qquad t = 2, \cdots, T \tag{8.3}$$

The number of zero-crossing points in $Fl_{br}{}'(t)$ is related to the number of notches in the curve. Therefore, noisy curves should have significantly more zero-crossing points in their first derivatives compared with smooth sigmoidal curves. The curves that satisfied the following condition were regarded as noisy and removed:

$$\sum_{t} \frac{-sgn\left[Fl_{br}{}'(t)\right] + 1}{2} > N_{zc} \tag{8.4}$$

where $sgn[\bullet]$ is the sign function and N_{zc} is the given threshold value ($N_{zc} = 9$ in this research).

8.3.5 Five-parametric Sigmoidal Fitting

Since amplification curves contain several information (such as background, plateau phase, and slope), the most representative features are represented using various sigmoidal equations. The chosen model in this study for curve fitting is the five-parametric sigmoid function, whose equation is given below:

$$(t, \boldsymbol{p}) = \frac{a}{(1 + \ exp^{-c(t-d)})^{e}} + b \tag{8.5}$$

$$\boldsymbol{p} = [a, \ b, \ c, \ d, \ e]^{T} \tag{8.6}$$

where t is the amplification cycle, $\boldsymbol{p}$ is the parameter vector, $f(t, \boldsymbol{p})$ is the fluorescence at cycle t. The mathematical function of these parameters and their corresponding representations in amplification curves are listed below:

- Parameter $\boldsymbol{a}$: it represents the amplitude of the function in the y-axis and it affects the maximum fluorescence that the amplification curves can reach.
- Parameter $\boldsymbol{b}$: it represents the vertical shift of the function along the y-axis and it affects the maximum fluorescence together with parameter a.

- Parameter $\boldsymbol{c}$: it represents the maximum slope of the sigmoid function and it's related to the efficiency of PCR reactions.
- Parameter $\boldsymbol{d}$: it represents the horizontal shift of the function x-axis and the fractional cycle of the inflection point. It's also related to C_t values of the PCR reaction.
- Parameter $\boldsymbol{e}$: it's the Richard's coefficient and it is related to the asymmetry of the sigmoidal trend.

To reduce optimisation iterations and unsuccessful fitting, a pivot fitting is applied on a subset of data ($\boldsymbol{D_s}$) to evaluate the optimal initial parameters $\boldsymbol{p_0^{opt}}$ for the equation before searching on the entire dataset ($\boldsymbol{D}$). First, a non-linear Least Square function $LS(\boldsymbol{p})$ is defined, showing the equation below:

$$LS(\boldsymbol{p}) = \sum_{t=1}^{T} \left(f(t, \boldsymbol{p}) - Fl_{br}(t)\right)^2 \tag{8.7}$$

To apply the pivot fitting, it is first initialised $\boldsymbol{p_0} = [0,\ 0,\ 0,\ 0,\ 0]^T$. Then, for the i^{th} curve Fl_{br}^i within the dataset $\boldsymbol{D_s}$, the following optimisation problem was solved to find the fitted parameter vector:

$$\boldsymbol{p_i} = \underset{\boldsymbol{B_{low}} < \boldsymbol{p} < \boldsymbol{B_{up}}}{\operatorname{argmin}} LS(\boldsymbol{p}) \tag{8.8}$$

where the lower bound $\boldsymbol{B_{low}}$ and the upper bound $\boldsymbol{B_{up}}$ for all the parameters are -100 and 100, respectively. After all the curves were fitted, the mean vector of all the $\boldsymbol{p_i}$ was used as the optimal $\boldsymbol{p_0^{opt}}$.

With the outcome from the pivot fitting, all curves in $\boldsymbol{D}$ are fitted starting from $\boldsymbol{p_0^{opt}}$. In addition, to get better fitting performance, the maximum number of fitting iterations (maxfev) was increased to a sufficiently large value ($1,000,000$ in this case). The same $\boldsymbol{B_{low}}$ and $\boldsymbol{B_{up}}$ were used for the pivot fitting.

8.3.6 Calculating Average Distance Score (ADS) and Minimum Distance Score (MDS) for Multiplex Assays

There are four curve representations for calculating ADS and MDS, which are: raw curves (45-D), normalised curves (45-D), fitted parameters (5-D) and c parameter (1-D). Two steps were

taken before the score calculation: (i) Extract the median feature vectors of each target for 45-D, 5-D and 1-D feature arrays. The median value was taken on each dimension, and the median feature vector with the same dimension was generated. It is assumed that the distribution of each target is Gaussian. However, outliers can affect the distribution unexpectedly. Therefore, the median value is a more robust representative compared to the average value, and $\boldsymbol{N}_t$ median vectors corresponding to $\boldsymbol{N}_t$ targets were constructed. (ii) Calculate Euclidean distance between each pair of targets, where given $\boldsymbol{N}_t$ targets, the total number of distances $\boldsymbol{N}_d$ is:

$$\boldsymbol{N}_d = \binom{\boldsymbol{N}_t}{2} = \frac{N_t\,(N_t - 1)}{2} \tag{8.9}$$

The vector of distances for each pair of targets is defined as:

$$\boldsymbol{S_D} = [d_{ij} \mid \ for\ each\ i = 2,\ \ldots, \boldsymbol{N}_t,\ f\ j = 1, 2, ..i - 1] \tag{8.10}$$

where d_{ij} represents the Euclidean distance between extracted median vectors of target i and target j. With the constructed distance set, the ADS and MDS were calculated as the average and the minimum value of all elements in $\boldsymbol{S_D}$, respectively:

$$\text{ADS} \ = \ \text{mean}(\boldsymbol{S}_D) \tag{8.11}$$

$$\text{MDS} \ = \ \min(\boldsymbol{S}_D) \tag{8.12}$$

8.3.7 The Smart-Plexer Ranking System

The inputs of the ranking system are simulated ADS and MDS. To increase the likelihood of choosing an optimal assay for data-driven multiplexing approaches, assays with the highest ADS and MDS (S_{BEST}) are selected from the entire combination set (S_{ALL}). Provided the number of the best combinations to be selected as N_{BEST} and the number of total combinations as N_c, the following steps were applied:

The proposed Algorithm 1 is used to pick the best simulated multiplexes based on the developed metrics ADS and MDS, and these assays are further tested empirically to select the optimal one for the diagnostic use. Moreover, to verify the correlation of the Smart-Plexer ranking with the ACA performance, the algorithm was used to select the bottom multiplexes

Algorithm 1

1: **Initialize:**
   N_{BEST} as required, $S_{BEST} \leftarrow \emptyset$
2: **for** $n_e = N_{BEST},\ N_{BEST}+1, \ldots,\ N_c$ **do**
3: $S_{BEST}^{MDS} \triangleq \{x | x$ are the top n_e combinations in S_{ALL} with largest $MDS\}$
4: $S_{BEST}^{ADS} \triangleq \{x | x$ are the top n_e combinations in S_{ALL} with largest $ADS\}$
5: $S_{BEST} \leftarrow (S_{BEST}^{MDS} \cap S_{BEST}^{ADS}) \cup S_{BEST}$
6: **if** $|S_{BEST}| \geq N_{BEST}$ **then**
7: **return** S_{BEST}
8: **end if**
9: **end for**

with the lowest ADS and MDS, by modifying step 3 and 4, so that the smallest instead of the largest ADS and MDS are applied.

8.3.8 The Smart-Plexer: Workflow Recap

The complete workflow of utilising the Smart-Plexer in a real laboratory setting is illustrated in Figure 8.1 and depicted as follows: given a number of target genes to be identified, several candidate primer sets are first *in-silico* designed and tested in singleplex format for each target, resulting in real-time PCR amplification curves for all the assays. The obtained data are further processed using the background, late curve, and noisy curve removal techniques mentioned in the *Data Processing* section. The processed curves are then fitted with the sigmoidal function from which the c parameters are extracted. For each potential combination of primer sets, inter-target distances of c parameters from singleplex curves are calculated and function as simulated alternatives for empirical multiplex curve distances. In this way, the best candidates for multiplex assays can be selected by choosing the combinations with the most distant target clusters (represented by c) in the simulation. This progress is achieved by calculating the c parameter-based ADS and MDS of each combination and finding the best ones using the ranking system mentioned above. The best candidate assays shortlisted from simulated multiplexes further go through wet-lab tests on synthetic DNA templates, and the ACA-based target identification is applied to the empirical multiplex data. The final winner assay with the highest ACA classification performance on synthetic DNA is labelled as the optimal assay, which is the final output of the entire Smart-Plexer workflow. The ACA model was a KNN classifier with 10 neighbours (k = 10).

3plex validation. Synthetic DNA of Adenovirus (HAdV), Human coronavirus HKU1 (HCoV-HKU1) and Middle East respiratory syndrome-related coronavirus (MERS-CoV) targets were selected for a 3plex validation, and all the data were generated in real-time digital PCR (qdPCR). Three primer sets were designed as candidates for each target, resulting in 27 potential combinations of multiplex assays in total. Because of the relatively small number of candidate assays, it is possible to perform wet-lab experiments for all combinations and analyse the relationship between simulated and empirical multiplex curve distances. Simulated ADS and MDS were calculated on different levels of curve representations (raw curves, FFI-normalised curves, and fitted parameters), and their correlations with the same metrics derived from empirical multiplex data were analysed. Furthermore, the ADS and MDS of c parameters, which are more concise indicators for inter-target curve distances, were generated and compared between simulated and empirical multiplexes. ACA performance against simulated ADS and MDS was depicted, and the t-SNE of the selected assays' results were illustrated.

7plex validation. Following the 3plex validation, seven targets were used to further validate the Smart-Plexer performance, where each target had at least two different assays, resulting in a total of 24 singleplexes and $4,608$ candidate combinations. Unlike for 3plex, the mass number of combinations makes it impossible to empirically test all the assays in multiplex settings. Instead, representative groups of assays were chosen for the laboratory validation. Following the aforementioned Smart-Plexer workflow, after calculating simulated ADS and MDS on c parameters, six highest ranked (BEST) and six lowest ranked (BOT) combinations were picked out using the Ranking System. In addition, six middle-distant combinations (MID) were selected with the step below:

Algorithm 2

1: **Initialize:**
   N_{MID} as required, $S_{MID} \leftarrow \emptyset$,
   MDS_{max} and ADS_{max} the maximum MDS and ADS among all combinations,
   $ADS_{bias} = MDS_{bias} \leftarrow 0.001$
2: $R_{MDS} \triangleq \left(\frac{MDS_{max}}{2} - MDS_{bias},\ \frac{MDS_{max}}{2} + MDS_{bias}\right)$
3: $R_{ADS} \triangleq \left(\frac{ADS_{max}}{2} - ADS_{bias},\ \frac{ADS_{max}}{2} + ADS_{bias}\right)$
4: $S_{MID}^{tmp} \triangleq \{x | MDS_x \in R_{MDS} and ADS_x \in R_{ADS}, \forall x \in S_{ALL}\}$
5: $S_{MID} \leftarrow$ apply **Algorithm 1** $on S_{MID}^{tmp}$ with N_{MID}
6: **return** S_{MID}

TOP-ADS and TOP-MDS (N = 6) assays were selected empirically with large ADS but small MDS, and large MDS but small ADS, respectively. Similarly to the 3plex validation, the

relationship between simulated and empirical scores of the selected assays was explored by correlations of simulated and empirical metrics and comparisons of c parameter distributions. ACA was also applied to different groups of combinations. The complete pipeline of the 7plex validation is illustrated in Appendix Figure D.3.

8.4 Results & Discussion

This Chapter describes the Smart-Plexer, a framework that uses singleplex PCR reactions as a 'card deck' to generate a 'winning combination' of the multiplex assay. After deciding the number of targets intended to multiplex, the Smart-Plexer takes as input a dataset generated from real-time PCR reactions with a single primer set (or singleplex assay) and a single target. Given the desired number of targets to multiplex in a single channel PCR, sigmoidal curves generated from all the singleplex/target interactions can be combined to simulate curves from a multiplex assay Figure 8.1. These simulations of assay combinations are then empirically tested in wet-lab multiplex tests for each target to evaluate changes in the curve shape of the amplification reaction during the transition from singleplex to multiplex environment (empirical multiplex). Moreover, to identify multiple targets with empirical multiplexes, this framework was coupled and evaluated with the ACA methodology.

As the ACA is a classifier recognising clusters from different amplification shapes (which, in this case, represent different targets), it is crucial to maintain differences among sigmoidal trends *in-silico*. Therefore, those differences across targets can be computed using the Smart-Plexer method through distance measurements (such as Euclidian distance). This novel framework is capable of distance calculation from either the entire amplification curve or its sigmoidal features. The average of computed distances among all the targets is used to rank each combination of singleplex (or simulated multiplex) from high to low inter-curve similarity values. Moreover, the ranking system takes the minimum distance between the two closest targets to ensure that simulated multiplex with high average values is not dependent on the high difference of only a group of curves. When two amplification curves have high similarity, hence a small distance value, the ACA classifier will not work efficiently to identify either target. Therefore, the rank of the combination depends on both average and minimum distance scores. A set of singleplex assays from the top ranks were selected as simulated multiplex for the empirical

validation in the laboratory, and the ACA performance was assessed.

To compute distances between amplification curves, the Smart-Plexer requires a filtering process where the amplification data generated undergo the following steps: (i) subtraction of curve background to remove the fluorescence signal noise at the starting cycles, (ii) removal of late amplification curves to exclude non-plateau reactions, (iii) removal of noisy curves to exclude non-sigmoidal shapes as result of operator error or instrumentation faults [219]. The following step comprised of a fitting equation using the 5-parameter model proposed by Spiess et al [31].

8.4.1 Selection of representative amplification curve

The ACA method uses the entire amplification curve as a time series where fluorescence values change as the number of cycles increases. Firstly, the entire raw amplification curve generated from the real-time PCR reaction is used as the input of the Smart-Plexer. Secondly, the framework is evaluated using curves normalised with the final fluorescence intensity (FFI) as input to assess performance changes by removing the absolute fluorescence information. To further investigate changes related to different curve representations and different levels of data abstractions (feature dimensions) provided to the Smart-Plexer, sigmoidal parameters generated from a fitting model are also used as input to assess the influence on this framework.

To evaluate the best fitting model, primary efforts have been focused on the selection of an appropriate equation. Several methods have been proposed to efficiently model the real-time PCR sigmoid, such as four, five, and six-parametric functions [220, 221, 31]. As a case study, the amplification curve data previously reported by Moniri *et al.*, 2020 were retrived [15]. Using raw curves as input, after sigmoidal fitting, the Mean Square Error (MSE) between the raw and the fitted curves for the entire dataset was calculated. The lowest MSE is achieved with the five-parametric model (MSE = 0.0036). The rising MSE in six-parameter sigmoid fitting is caused by unsuccessful optimisation resulting from a larger searching dimension. Based on the lowest MSE value, it is determined to utilise the five-parameter sigmoid function to extract features, and the equation is given below:

$$f(t) = \frac{a}{(1 + \ exp^{-c(t-d)})^{e}} + b \tag{8.13}$$

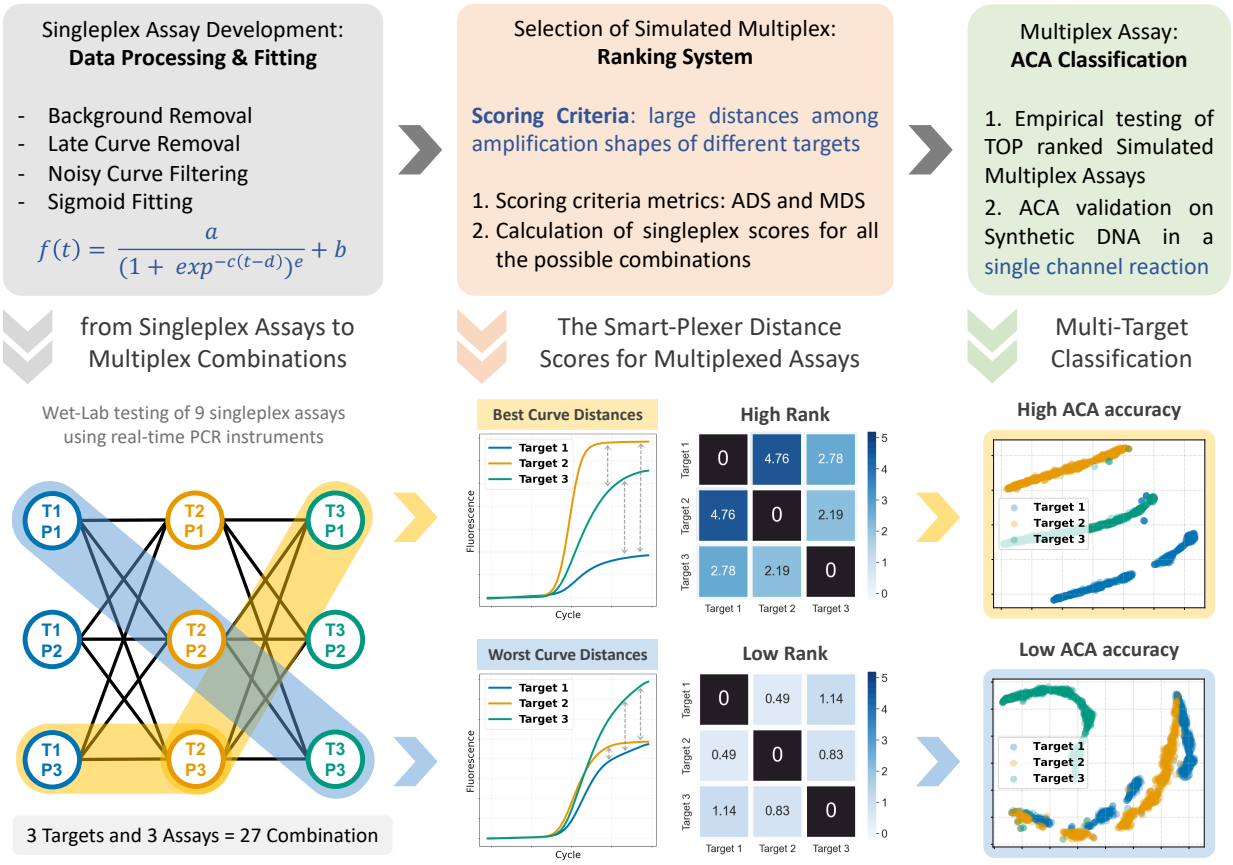


Figure 8.1: Smart-Plexer workflow. **a.** Given a dataset of singleplex real-time PCR reactions (real-time amplification curves), a processing step is applied (**a.i-a.iii**). The processed curves are fitted following the equation depicted in step **a.iv**. An example is given in **b.**, where each curve resulting from singleplex reactions is used in a simulation of multiplex assays. Three targets are considered, and each of them has three unique singleplex assays (a total of 27 simulated combinations). **c.** The simulated multiplex scores are calculated from the Smart-Plexer according to the Scoring Criteria. **d.** Distances within curves from different targets are calculated based on mathematical algorithms (such as Euclidean), and as shown in the confusion matrices, resulting values are used to rank multiplex assays from high (high distances within targets) to low (low distances within targets). **e.** High-rank multiplex assays are chosen for empirical testing, and the ACA method is used to evaluate the classification performance on target identification of each selected multiplex. **f.** Cluster visualisation with 2-D t-SNE represents the difference in inter-target distances between a High-Rank and a Low-Rank multiplex, resulting in high and low ACA classification accuracy, respectively.

where t is the amplification time (or PCR cycle), $f(t)$ is the fluorescence at time t, a is the maximum fluorescence, b is the baseline of the sigmoid, c is related to the slope of the curve, d is the fractional cycle of the inflection point, and e allows for an asymmetric shape (Richard's coefficient).

The three different curve representations (raw curves, FFI normalised curves and fitted parameters) were further used to evaluate the transferability from singleplex to multiplex reactions in the Smart-Plexer.

8.4.2 Average Distance Score (ADS) and Minimum Distance Score (MDS) based on curve distances to rank multiplex assays.

Since it is hypothesised that distances between amplification curves should be maintained during the transition from singleplex to multiplex environments, two distance metrics to measure transferability from simulated to empirical multiplexes are developed in this study.

It is possible to calculate distances between two distinct curves by considering them as two data points in the multidimensional space and quantify their distances using various metrics (i.e., Euclidian, Cosine and Manhattan). In a single channel multiplex assay, the number of primer sets present in the reaction equals the number of targets (N_t), therefore the number of distances (N_d) among curves of different targets is represented by the following formula:

$$N_d = \binom{\boldsymbol{N}_t}{2} = \frac{N_t\,(N_t - 1)}{2} \tag{8.14}$$

The average of all the distances is used to assign a score to the multiplex assay called Average Distance Score (ADS). The ADS provides information on the overall distances across targets, and the higher its values are, the more distant the curves are, and better ACA performance is expected (as distances are related to data point clusters). A high ADS does not guarantee a large distance between every two targets of the multiplex. To overcome this limitation, a second metric, called Minimum Distance Score (MDS), is used to account the distance value of the two closest curves (minimum value of the given N_d distances).

The ADS and MDS narrow down the selection of empirical testing for the highest performing multiplexes using a ranking system. Moreover, they are used to validate that inter-curve distance information is maintained during the transition from simulated to empirical multiplexes, and they can be used to develop assays *in-silico* more suitable for ACA, skipping costly and timely laboratory testing.

8.4.3 Smart-Plexer validation using a 3plex assay

To assess the performance of the Smart-Plexer for both *in-silico* multiplex development and ACA classification accuracy, three primer sets were designed to detect the three selected targets

using synthetic DNA and laboratory tests were conducted in real-time digital PCR (qdPCR). The considered viruses are Adenovirus (HAdV), Human coronavirus HKU1 (HCoV-HKU1) and Middle East respiratory syndrome-related coronavirus (MERS-CoV). As shown in Figure 8.1, the number of combinations to test using N_t targets ($N_t = 3$) and N_{Ps} assays for each target ($N_{Ps} = 3$) is 27 ($N_c = {N_{Ps}}^{N_t} = 27$ combinations, listed Appendix Table D.5. Three targets were chosen to validate the Smart-Plexer because a complete comparison of all the 27 simulated and empirical multiplex assays can be experimentally conducted as the number of wet-lab experiments is achievable ($N_c \times N_t = 81$ tests).

The wet-lab testing of each primer set (or singleplex assay) was conducted, and the resulting raw data were combined in a total of 27 simulated multiplexes as explained before. Similarly, experiments were carried out on combinations of primer sets (or empirical multiplex assays) in a single channel reaction. A group of amplification curves, which can be considered as data points in multidimensional spaces, were generated from a unique interaction between each assay and its specific target. The median of these data points was calculated to represent each group of curves. Furthermore, distances among all the curve medians were used to generate the ADS and MDS of all the possible combinations Figure 8.2a-b visually represent the correlation between the *in-silico* and wet-lab tested assays using ADS and MDS in simulated and empirical multiplexes. Pearson coefficients were reported for both ADS as 0.301, 0.972 and 0.607, and MDS as 0.092, 0.761 and 0.686, for raw curve, normalised curve and fitted parameters, respectively (visual representations of each curve type/parameters are depicted in Figure 8.2c, and ADS and MDS for all the curve types/combinations are reported in Appendix Table D.7).

It can be observed that normalised curve correlations scored higher than the rest in both ADS and MDS, showing that simulated and empirical multiplex are correlated if FFI is discarded. It is also important to note that the use of all the five curve parameters worsens the correlation as the bimodal distribution of parameter e negatively influences the correlation, as discussed by Miglietta *et al.*, 2022 [219]. Moreover, the correlation from singleplex to multiplex might be affected by the fact that the d parameter is related to the cycle threshold (C_t) of the amplification curve. Target concentration can be influenced by instrumentation, operator, and experimental errors; therefore, variabilities of C_t can easily mislead the correlation of the five parameters using d. Moreover, the scope of conducting this correlation is to compare purely

sigmoidal shapes, and concentrations of the nucleic acid targets should not affect the distance values of two curves. In addition, the use of parameter a and b is redundant as: (i) a is related to the FFI, and as shown in the middle plot of Figure 8.2a-b, FFI is not relevant to the distance correlation and (ii) all curves present in this dataset were processed with a background removal (baseline correction) and all b parameters were levelled to almost zero.

These discoveries on the correlation between simulated and empirical multiplex distances inspired us to seek a more representative feature which would maintain the information of distances during the translation from a singleplex to a multiplex environment. As mentioned before, the parameter a, b, d and e can negatively influence the correlation for both ADS and MDS; therefore, the c parameter is the focus of this study.

8.4.4 The key parameter for curve distance correlation in multiplex assays: the "slope"

The previous section reported all the correlation coefficients for ADS and MDS between simulated and empirical multiplexes, in concomitance with different curve representations: raw curves, normalised curves, and fitting parameters. Both ADS and MDS showed the maximum correlation values when considering normalised curves. Those results, along with the discussion on the fitted parameters in the previous section, indicate that reducing the information contained in the amplification curve is beneficial. This section explores how the c parameter preserves distance information from singleplex to multiple environments of each primer set/target reaction.

In the 3plex validation, each singleplex assay was tested against its specific target (N=9), resulting in 27 different combinations of simulated multiplexes. Moreover, the c parameters were fitted and extracted from 27 empirically tested multiplex assays (81 tests). Appendix Figure D.1 shows the correlation between simulated and empirical ADS and MDS calculated from c parameters with correlation coefficients of 0.973 and 0.774, respectively. To further evaluate whether c parameter distributions were maintained in the translation to empirical multiplexes, their three distributions (where three is equal to the number of multiplexed targets) from the singleplex reaction were compared with their corresponding distributions in empirical multiplex reactions. As illustrated in Figure 8.3a-c, distributions of three different multiplex assays are

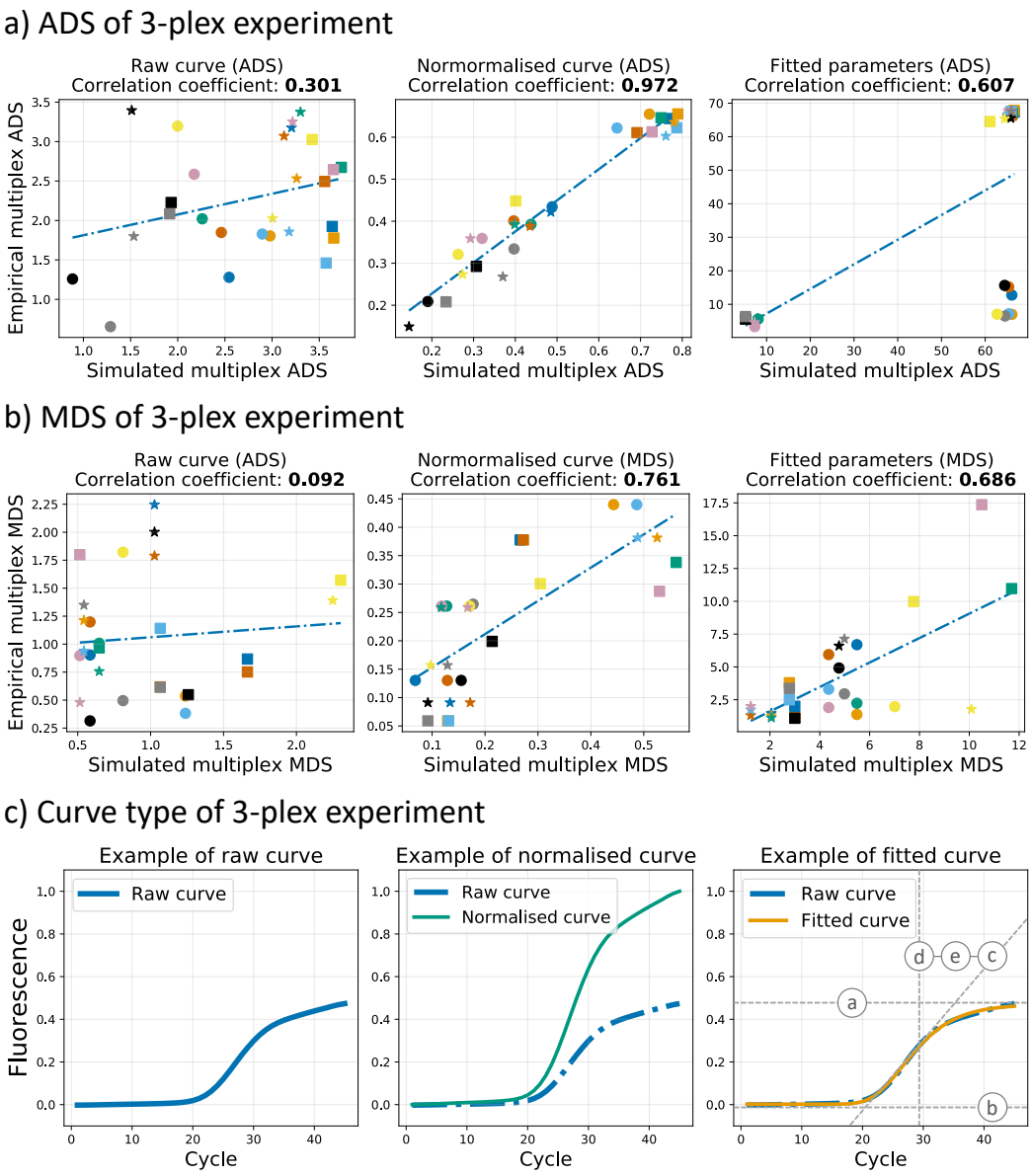


Figure 8.2: Representative features investigation based on the 3plex assay. a) The correlations of the Average distance score (ADS) between simulated and empirical multiplexes for the three types of curves/parameters (Raw curve, normalised curve and fitted parameters) are presented (from left to right in the same order). For each plot, each point with unique colour and shape corresponds to combination 1 to 27. The blue dashed lines are computed using linear regression. The Pearson coefficients for all three plots are calculated. b) Similarly, the correlations of Minimum distance score (MDS) are depicted for the three curve representations. c) Illustration of the three types of curve representations. Examples of raw amplification curve (after data processing), normalised curve (computed based on the FFI) and fitted curve/parameters are presented from left to right. The fitted curve is computed with a 5-parameter Sigmoid function using raw curves. As a result of this, both fitted parameters (a, b, c, d, e) and fitted curve (predicted fluorescence values corresponding to each cycle from the 5-parameter Sigmoid model with fitted parameters) are obtained.

visualised with their relative mean values represented by the dashed/dotted lines. The figures show the capabilities of the c parameter to maintain distance information going from simulation to empirical test. It can be observed that in most cases, the location of the parameter distribu-

tion for each target is maintained. In other situations, the distribution may be shifted from the singleplex events; however, the relative distance relationship of c values is kept. Figure 8.3a illustrates the c parameter distribution of a low-rank ADS/MDS multiplex, showing overlaps for all the three singleplex assays in both simulated and empirical multiplexes. As distances among amplification curve shapes can significantly affect the ACA classifier, reduced performance is expected for multi-target identification. Another distribution trend among multiplex assays is represented in Figure 8.3b, where the selected Primer Mix (PM3.01) has a high simulated ADS value (0.117) but low MDS (0.003). Moreover, the ADS value for distributions in Figure 8.3c equals 0.138, which differs only 0.21 from the combination PM3.01. However, PM3.12 has an MDS value of 0.075, representing an increase of 0.072 compared to PM3.01. This highlights the importance of considering minimum distances between c parameter distributions of the two closest targets: a small MDS value indicates a less separable group of target clusters, resulting in low ACA accuracies for multi-pathogen identification in a single fluorescent channel reaction. To numerically report how distributions are related in the translation from simulated to empirical multiplexes, the Rooted Mean Squared Error (RMSE) was calculated as follows:

$$\text{RMSE} = \sqrt{\frac{(\boldsymbol{D_s} - \boldsymbol{Dm})^T \ (\boldsymbol{D_s} - \boldsymbol{Dm})}{N_d}} \tag{8.15}$$

where $\boldsymbol{D_s}$ and $\boldsymbol{D_m}$ are vectors for distances among targets in singleplex and multiplex, respectively. RMSE values of all the 3plex combinations range from 0.003 to 0.050, which are negligible considering the range of the c parameters. The ADS, MDS and RMSE values for all the 3plex combinations are reported in Appendix Table D.6. These results emphasise that distances between simulated and empirical multiplex share high similarity across different ranks, ensuring that the scoring system (based on ADS and MDS) is not affected whether in singleplex or multiplex environments.

Accuracy of all the possible combinations in 3plex assays

One of the aims of the Smart-Plexer is to improve the classification of multiplex assays, in this case, related to the ACA method. As demonstrated in the previous section, distances among amplification curves of empirical multiplex assays are similar to those generated in simulated multiplexes. Therefore, leveraging ADS and MDS, simulated multiplexes can be used to rank

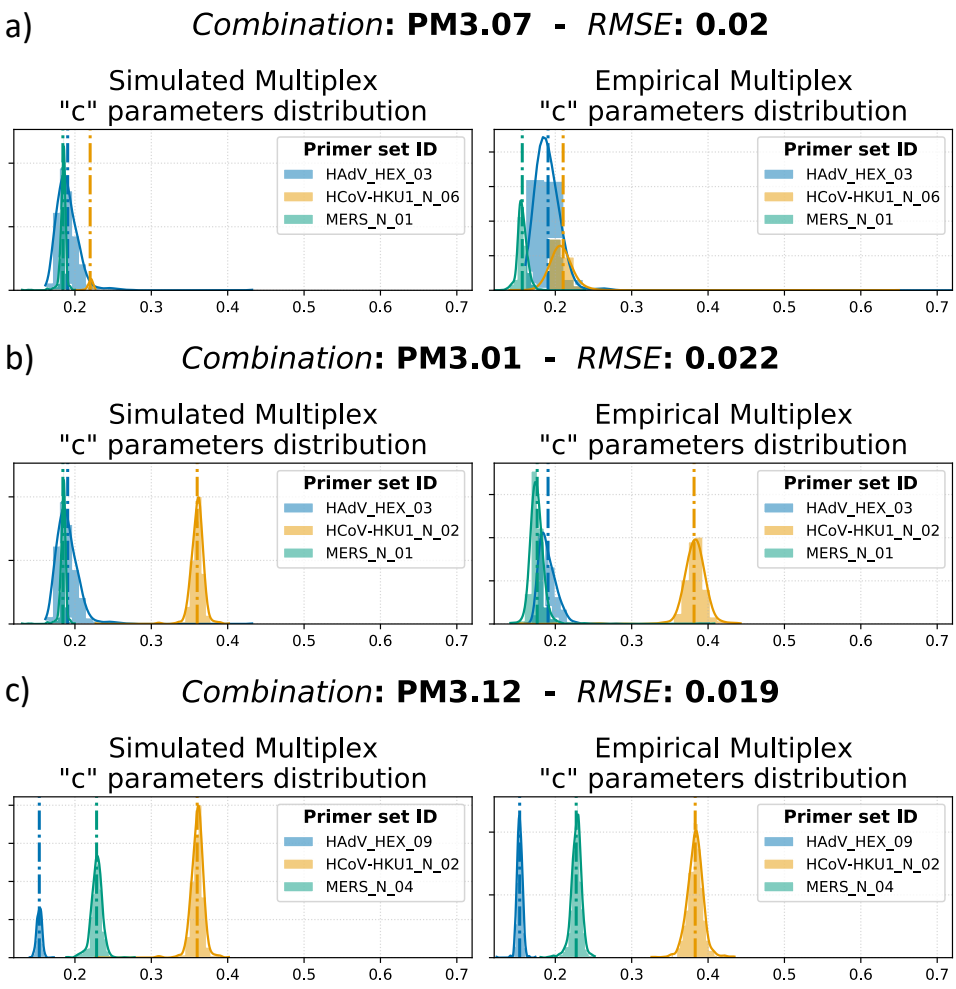


Figure 8.3: Relative c parameter distributions of three different multiplex assays. a) Primer Mix 3.07 (PM3.07) illustrates the c parameter distribution of a low-rank ADS/MDS multiplex; b) PM3.01 as an example of high ADS but low MDS multiplexes. c) Multiplex assay with high ADS and MDS with clearly separated distributions. For each subplot, the left graph shows the distributions of c parameters for the Simulated Multiplex. The right plot represents the corresponding distributions according to the empirical multiplex data. The vertical dashed lines correspond to the mean of the distribution computed for different targets. To quantitatively verify that the distances are maintained in transition from simulated to empirical multiplexes, the RMSE of distances is calculated and displayed on the graph title.

each combination and find the optimal assays with the largest inter-target distances for the ACA classifier. To further demonstrate that the ADS and MDS are crucial to improving multi-target identification in single well PCR reactions, the classification performance of the ACA method was assessed by using 10-fold cross-validation and the k-Nearest Neighbors (KNN) algorithm. Figure 8.4a shows a 3-D graph where both ADS and MDS of the c parameters are correlated to the ACA accuracy. Accuracy percentages ranged from 98.63% to 100% for each multiplex. The rainbow plane, which is fitted with linear regression on all the visualised data points, represents the gradient of the classification accuracy, showing an upward trend as

ADS and MDS increase, which is consistent with the hypothesis that the ACA classification performs better with larger inter-target distances. Moreover, the plane on the left of Figure 8.4a has a grey highlight zone called Vacuumed Area, where data points cannot fall inside as it is mathematically impossible to have an average distance value smaller than the minimum distance. It is also defined another area called Forbidden Area, as visualised in the rotated 3-D plot on the right of Figure 8.4a, where it is expected that no point will be founded, provided high values for ADS and MDS.

Both 3-D plots have circled points labelled as the top combination (TOP), bottom combination with lowest ADS (BOT ADS), bottom combination with lowest MDS (BOT MDS), and outlier combination (OUTLIER), with ACA classification accuracies of 99.9%, 99.89%, 98.06%, 99.01%, 99.82% and 99.87%, respectively. Although the overall classification performance for all the 27 combinations shows a high average of 99.51% $\pm$ 0.41%, an increase of 1.84% is observed for the top ADS/MDS data point compared to the bottom one. Furthermore, as depicted in Figure 8.4b-e, by applying 2-D t-distributed stochastic neighbor embedding (t-SNE) visualisation on curves generated by the top and bottom-ranked primer combinations, more condensed target clusters and better separated inter-target boundaries can be seen for top-ranked assays [222]. This results in more distinguishable curve shapes and larger curve distances among targets, which benefits the ACA classification. Numerical analysis of the visualised clusters was assessed using the Mean Silhouette Scores (MSS). As reported by Kaufman *et al.* 2009, Silhouette scores between 0.51-0.70 are considered more effective in cluster separation than values below 0.50 [223, 193]. The reported MMS scores show significantly larger inter-cluster distances for the top combinations, with values higher than 0.61 as opposed to the bottom ones of less than 0.27 (in Appendix Table D.6, ADS, MDS, MSS and ACA accuracies for each combination of the 3plex experiment are reported). This finding proves that the ADS and MDS metrics are valid indicators for predicting optimal primer set combinations for the ACA classifier. Relying on the Smart-Plexer for selecting multiplex assays from singleplexes, the likelihood of accurate multi-target identification in a single fluorescent channel reaction is significantly increased using the ACA methodology.

As mentioned above, Figure 8.4a highlights the presence of outlier combinations where small ADS/MDS with high ACA accuracy are reported (instead, low accuracy for the ACA classifier is expected). However, the existence of such data points does not deny the effectiveness

of the proposed method. It is important to emphasise that the overall ACA accuracy for 3plex is inherently high because of the low levels of multiplexing. Classifying three different curve shapes does not represent a major challenge for this Machine Learning method, and targets with minor curve-shape differences can be easily separated in the feature space. Considering this, along with the prevalent randomness that exists in the ACA method for 3plex, accuracies higher and lower than expected may occur in the given dataset. In fact, in the area with low ADS/MDS, a large standard deviation for accuracies among data points which fall beneath and above the fitted plane are observed. Regardless of the accidentally high accuracies and low ADS/MDS caused by randomness, Figure 8.4f-g evidence that these outlier combinations will face more challenges when used for multi-target identification in larger scale multiplexes (or high-level multiplexing). In the outliers, the mapped target clusters are largely overlapped with unclear boundaries and small MSS even in 3plex assays. Therefore, the next section demonstrates that the higher the level of multiplexing is, the more difficult the target separations are in the feature space when using these outliers.

Although low ADS/MDS combinations may occasionally show good performances, the proposed method ensures that all predicted optimal multiplex assays with high ADS/MDS show high accuracies in ACA and never the opposite. As illustrated in the 3-D plots of Figure 8.4a, the forbidden area (the red triangular prism) has no data point falling in, which highlights the effectiveness of the ADS/MDS ranking system. This is a first ever demonstration that multiplex assays tailored to the ACA method can be *in-silico* developed starting from singleplex PCR reactions. This not only increases the likelihood of accurate multi-pathogen identification, but also allows for a higher level of multiplexing in a single fluorescent channel. To demonstrate the capabilities of the Smart-Plexer in developing optimal high-level multiplex assays for data-driven approaches, in the following section, its performance with seven different targets is assessed.

8.4.5 Smart-Plexer for development of 7plex assays

The focus of the previous section was on using a small number of targets to demonstrate that the developed ADS and MDS used to correlate distances between curves in both simulated and empirical multiplex assays were maintained. Moreover, accuracies among all the different combinations were evaluated using the ACA methodology, where high ADS/MDS multiplex

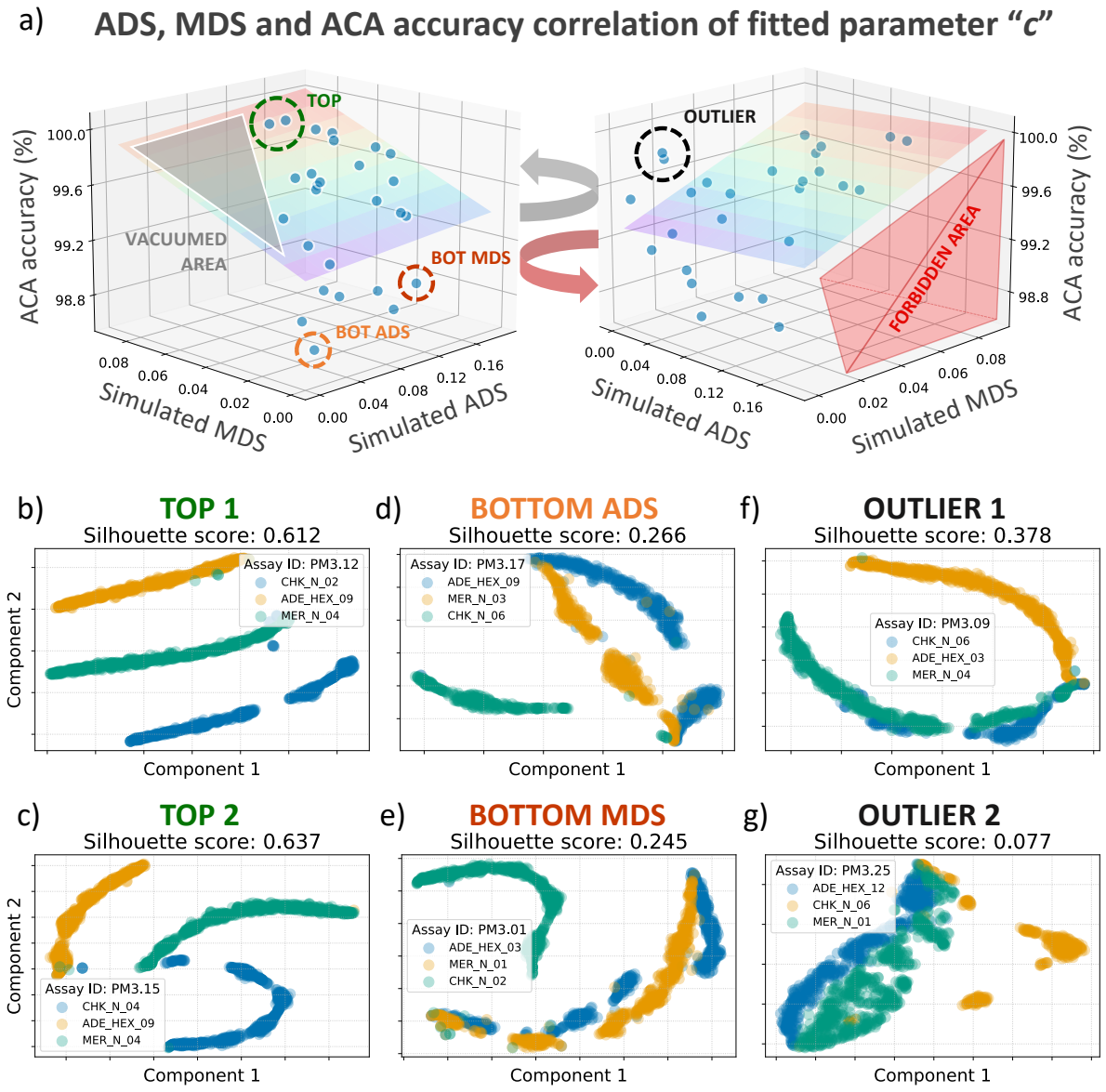


Figure 8.4: The influence of ADS/MDS on the ACA performance for all possible 3plex combinations. a) 3-D plot of ACA classification accuracy for each combination versus simulated ADS and simulated MDS computed based on the c parameter. The rainbow plane is calculated using linear regression. In the left 3-D figure, the grey highlighted area is called Vacuumed Area, where simulated MDS is larger than simulated ADS (combinations in this area are mathematically impossible to be found). The right 3-D figure is a rotation of the left one, where a red is highlighted named Forbidden Area. In this region, high ADS/MDS combinations possess low ACA accuracies; however, no combinations were found. b-g) For the combination circled (TOP, BOT MDS, BOT ADS and OUTLIER) in a), 2-D t-SNE was applied on raw curves. In addition, for quantitative verification, the Mean Silhouette Scores (MSS) of target clusters were reported in the subplot title.

assays show the highest likelihood of correct multi-target classification. These previous results indicate that the Smart-Plexer is a promising technique for optimal selection of primer set

combinations in data-driven multiplexing.

Next, the Smart-Plexer was challenged to develop an optimal 7plex assay, which through the ACA method, is able to accurately identify the following Respiratory Tract Infection (RTI) pathogens in a single fluorescent channel using qdPCR: Human adenovirus (HAdV), Human coronavirus OC43 (HCoV-OC43), Human coronavirus HKU1 (HCoV-HKU1), Human coronavirus 229E (HCoV-229E), Human coronavirus NL63 (HCoV-NL63), Middle East respiratory syndrome-related coronavirus (MERS-CoV), and Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). At least two different assays for each target were developed, for a total of 24 singleplexes across the seven pathogens, as shown in Appendix Table D.8. Each primer set was tested using synthetic DNA of its correspondent pathogenic target. Following the previous 3plex experimental workflow, the resulting raw curves were processed, fitted, and passed to the Smart-Plexer to calculate all possible 7plex combinations (N = 4608) and compute their ADS/MDS. Based on c parameter distances from fitted simulated multiplexes, Figure 8.5a shows how the ADS and MDS can be visualised in a two-dimensional space. By considering the mean and standard deviation of the two scores, boundaries to the ADS/MDS distribution for all the combinations are set up, dividing the space into four separate regions, with the purpose of showing how empirical multiplexes would perform for the ACA method depending on their ADS/MDS. The black horizontal segmented line in Figure 8.5a divides high and low MDS, and the vertical one separates the two ADS regions, resulting in four distinct areas. By testing different multiplexes from each of these regions, a further demonstration that the chance of developing a reliable multiplex can vary based on the selected regions or selection criteria is conducted. Therefore, multiplex assays from different areas are chosen and categorised into five classes, which were empirically tested with synthetic DNA in qdPCR: BOT (N = 6), MID (N = 6), BEST (N = 6), TOP-ADS and TOP-MDS (N = 6) values (detailed selection criteria are reported in the methodology section).

After the empirical testing, the distances of the c parameters of each selected multiplex were compared to the simulated one, resulting in a correlation coefficient of 0.99, as shown in the middle graph of Figure 8.5b. Moreover, empirical multiplex amplification events were visualised using 3-D t-SNE, and distances across target clusters were calculated with the MSS. As shown in the left plot of Figure 8.5b, clusters of the selected BOT combination have an MSS of 0.12, whereas for the BEST one the score is 0.67. It can be observed that there is a

clear difference in clustering between the two selected multiplex assays, where the BEST one shows clear separation among different targets (in line with the 3plex results), and is expected to converge in better ACA classification. The opposite scenario is shown in the BOT combination.

It was validated that in higher level multiplexing, distance distributions of the c parameters were still maintained from simulated to empirical testing; therefore, the RMSE of the chosen tested combinations was computed. Figure 8.5c-d illustrate side-by-side c parameter distributions for each target in both simulated (left) and empirical (right) multiplexes, showing a small RMSE for both BOT and BEST assays (0.012 and 0.031), and confirming the distance-maintaining hypothesis validated in the 3plex experiments. Moreover, the ACA accuracy using training and testing datasets obtained in different experimental settings (different days, operators, and reagents) is tested to ensure the reproducibility of the methodology. As expected, the performance of the BEST combination was significantly higher than the BOT one, with a 39.42% increase in accuracy. Furthermore, in Appendix Table D.9, the ADS, MDS and accuracy values for the 24 selected multiplex assays are reported. In Appendix Figure D.2 is visualised the standard curve for each target using the BEST 7plex assay to evaluate primer sensitivity and specificity. The chosen multiplex reached a limit of quantification equal to 102 for all the respiratory pathogens using synthetic DNA in real-time PCR.

As described before, ACA performances were evaluated using training and testing datasets from different experimental settings with the same sample size. All the selected 24 multiplexes were empirically tested, and their multi-target identification performances were assessed. In Figure 8.5e, accuracies and standard deviations of each group of multiplexes were reported and visualised as box plots. The best-combination group scored an average ($\pm$ standard deviation) classification performance of 95% ($\pm$ 0.04%) using a KNN classifier, which is the highest average and the lowest standard deviation among all the groups. There is a decreasing trend in the average accuracy, and an increasing trend in the standard deviation as the ADS/MDS values become smaller. Previously, the 3plex validation showed the presence of outliers in low ADS/MDS rank with high ACA classification accuracy, which is also observed in these 7plex tests. However, the standard deviation indicates that the Smart-Plexer does provide a robust and solid solution (even at high-level multiplexing) to significantly increase the likelihood of choosing an optimal multiplex for data-driven multiplexing (i.e. ACA methodology).

8.4.6 Clinical validation results

The final step was to validate that the Smart-Plexer is capable of easing the laboratory workload in developing multiplex assays. After testing six potential best combinations based on ADS/MDS, the one with the highest ACA classification accuracy on synthetic DNA (PM7.2151) was selected. To clinically validate the selected 7plex for multi-pathogen identification, inactivated clinical samples were purchased from Randox Laboratories (UK) and extracted using a gold standard kit (QIAGEN mini amp). The extracted samples were used as the testing dataset (7,638 positive amplification reactions), while curves resulting from synthetic DNA amplification reactions (5,207 positive amplification reactions) were the training. The classifier used was a KNN with the number of neighbours equal to 10. As shown in Table 8.1, a total of 14 positive samples were classified in qdPCR using the ACA methodology. The predicated label of a sample is given by selecting the most predictable label within all the in-sample curves. The confidence level was given as the percentage of the amplification curves with the most predicted label. Using the Smart-Plexer selected candidate assay, all the pathogens were correctly identified with high confidence level (median = 95.46%).

It is important to note that this study faced a seven-class classification problem, where the accuracy of a "random guess" (or a random classifier as convention) equals 14.3% under a balanced dataset. All the confidence levels were much higher than the random guess accuracy, indicating solid and robust predictions with the selected optimal multiplex assay. Although the number of clinical samples was limited by the number of pathogens provided by the manufacturer, the proposed framework, in combination with the ACA methodology, achieved a highly accurate identification of multiple pathogens by using an optimal multiplex assay in a single fluorescent channel reaction. The Smart-Plexer can leverage the capability of the data-driven multiplexing to an easy-to-develop, robust, and cost-effective molecular diagnostic solution.

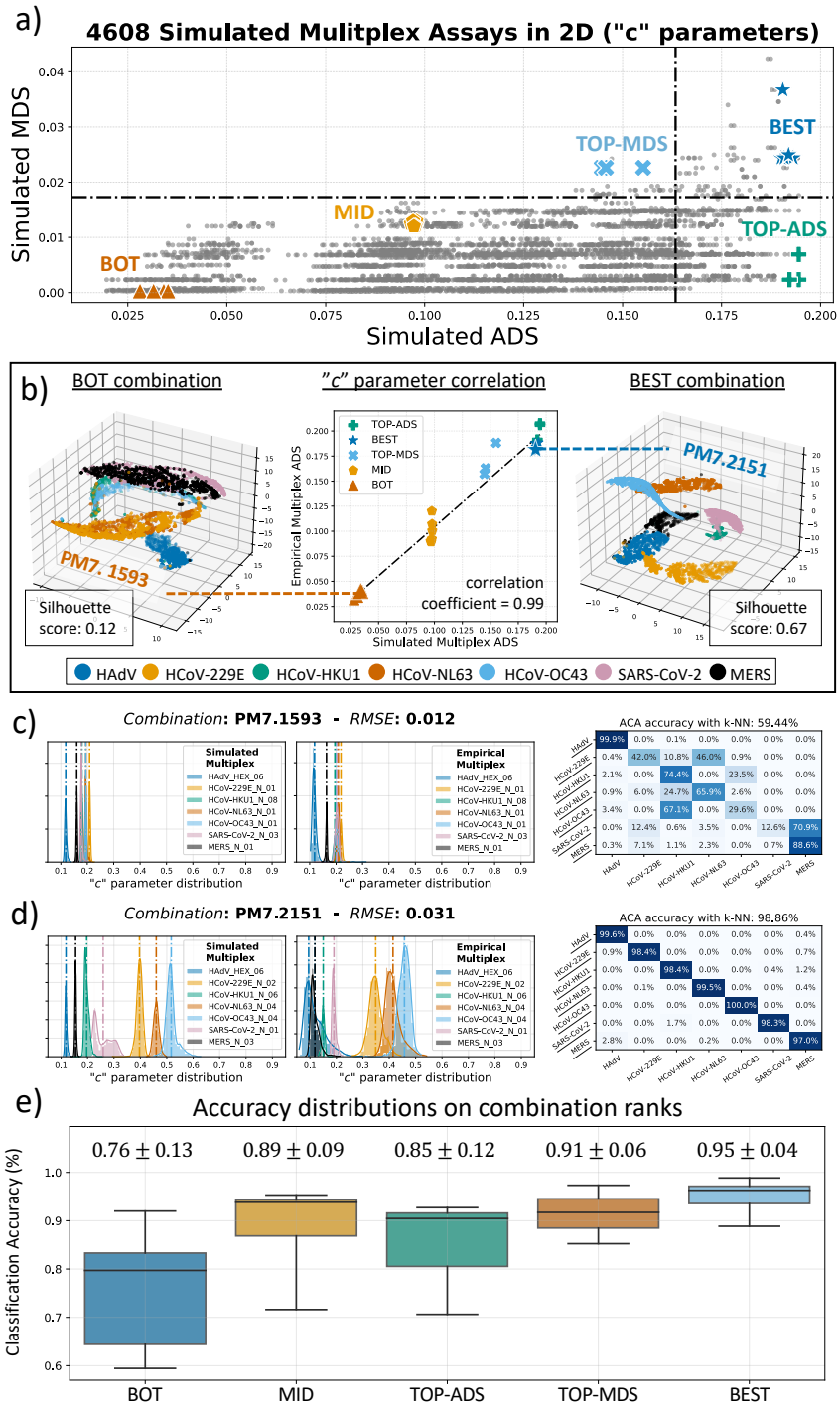


Figure 8.5: Validation of Smart-Plexer based on 7plex assays. a) 2-D ranking results for all 4608 combinations (ADS vs MDS). The selected combinations were used for multiplex empirical testing. b) The 2-D plot in the middle depicts the relationship between empirical and simulated scores based on c parameters. Enlarged data points are the 3-D plots of the BOT (PM7.1593) and BEST (PM7.2151) combinations. c-d) Simulated and Empirical c distribution plots of the selected combinations (RMSE values in subplot titles). The vertical dashed lines correspond to the mean of the distribution. On the right, the confusion matrixes of ACA performance for both with KNN accuracy. e) The box plot of ACA classification accuracy for each selected group (mean and std above).

Table 8.1: Clinical validation results.

Sample index	Panel ID (Randox, UK)	Expected Pathogen (True Label)	ACA Classified Pathogen (Predicted Label)	AC count	Confidence Level (%)	Outcome
1	QAV164189	HAdV	HAdV	14	100.0	detected
2	QAV164189	HCoV-NL63	HCoV-NL63	770	100.0	detected
3	QAV164189	HCoV-NL63	HCoV-NL63	545	96.15	detected
4	QAV164189	HCoV-OC43	HCoV-OC43	94	78.72	detected
5	SCV2QC	SARS-COV-2	SARS-COV-2	769	69.96	detected
6	SCV2QC	SARS-COV-2	SARS-COV-2	631	94.77	detected
7	SCV2QC	SARS-COV-2	SARS-COV-2	766	100.0	detected
8	SCV2QC	SARS-COV-2	SARS-COV-2	756	99.34	detected
9	SCV2QC	SARS-COV-2	SARS-COV-2	748	99.20	detected
10	QAV154181	MERS	MERS	287	60.98	detected
11	QAV154181	MERS	MERS	770	96.49	detected
12	QAV154181	MERS	MERS	770	79.09	detected
13	QAV154181	MERS	MERS	698	91.69	detected
14	QAV154181	MERS	MERS	20	70.00	detected

8.5 Conclusion

This Chapter describes the Smart-Plexer, an innovative framework which combines wet-lab experiments and computational algorithms to generate optimal multiplex assays for data-driven approaches using real-time PCR data. The method leverages mathematical metrics to construct an advanced ranking system to increase the throughput of conventional molecular tests by optimising their chemical peculiarities. To reveal the potential of this powerful approach, a recently reported machine learning method, named Amplification Curve Analysis (ACA), is used to identifying multiple nucleic acid targets in a single fluorescent channel with conventional PCR instruments. As the ACA leverages kinetic information encoded in the amplification curve, multiple targets can be classified based on the unique interaction with their assigned primer sets. However, constructing different amplification curve shapes for each multiplexed target is one of the major challenges for the ACA approach. The Smart-Plexer solves this problem by providing an easy-to-use framework for multiplex assay development, enabling high-level and highly accurate data-driven multiplexing.

This study shows the progression of the Smart-Plexer starting from a simple three-target classification problem. From the wet-lab testing of three singleplex assays for each of the three targets, a total of 27 combinations (in this case 3plex assays) can be generated *in-silico* (simulated multiplex) and ranked based on the mathematical curve-shape distances. Using

synthetic DNA in qdPCR and a single fluorescent channel, the assays were empirically tested (empirical multiplex), and the ACA classification accuracies were evaluated for all the possible combinations. The distance scores computed from the Smart-Plexer for multiplex assay ranking are linearly correlated between simulated and empirical multiplexes. Moreover, it is showed a further correlation between high-rank multiplexes and a high probability of increasing the ACA accuracies, confirming that the metrics used in this novel framework are theoretically connected to the distance measurement of the machine learning classifier.

As the complexity of developing multiplex assays exponentially increases with the number of targets, the Smart-Plexer was further challenged by designing a 7plex assay to identify common respiratory tract infection (RTI) pathogens. Consistent with the 3plex validation, the correlation between simulated and empirical multiplex is also maintained in 7plex. Regarding the ACA classification, it is logical that higher similarities among curves exist in a scenario with a higher number of targets, making it harder to develop multiplex assays. Nevertheless, the Smart-Plexer brilliantly generated an optimal multiplex assay, which correctly identified pathogens presented in 14 commercial clinical samples. It was further demonstrated that, since ACA is a clustering method, it requires a large minimum distance between the two closest clusters and a large average distance among all clusters in the multiplex. Therefore, the Smart-Plexer ranking system enables the development of optimal multiplex assays for data-driven multiplexing.

Apart from the scalability of multiplexing that the Smart-Plexer can provide to the ACA method, it is demonstrated for the first time that machine learning approaches can be applied to probe-based multiplexes, in this case, TaqMan. Probe-based assays, together with the use of intercalating dyes and isothermal chemistries, are expanding the boundaries of data-driven multiplexing and opening new windows for its application in commercial, research and clinical fields. The Smart-Plexer eases the development of any novel multiplex panel or molecular assays, enabling the use of the ACA as an emerging diagnostic tool. Through this hybrid method, it is possible to select the highest rank combination *in-silico* with wet-lab tested singleplexes, avoiding performing expensive and time-consuming multiplex assay development phases.

While this novel framework is validated with high-level multiplexing (7plex), it is essential to highlight that distances between amplification curves can be a limiting factor in single flu-

orescent channel multiplexing. This affects the Smart-Plexer since the inter-target differences of fitting parameters considered for the distance measurement become smaller as the target number increases. In this work, linear distance measurements are used, but more advanced metrics (e.g. Minkowski, Chebyshev or Cosine) can be adopted to improve the ranking performance. Moreover, when a higher level of multiplexing is required, the use of probe-based chemistries such as TaqMan comes handy. By leveraging the optical capability of real-time PCR instruments, a multiplex assay using multiple-channel detection can double or triple the number of targets in a single reaction. All these strategies aim to improve the ACA classification through a more innovative development from the chemistry perspective, while from the machine learning view, the current classifiers rely on state-of-the-art algorithms which shine for their robustness but are limited for tailoring to specific datasets. Previous studies demonstrated that more advanced classifiers such as convolutional neural networks (CNN) could extend the ACA capability to classify targets for higher-level multiplex assays. However, as a novel technique, data-driven multiplexing requires more optimisation and development of algorithms.

The Smart-Plexer represents a solution for developing multiplex assays by utilising both empirical testing and *in-silico* computation. The hybrid nature of this framework still requires wet-lab experiments; therefore, certain limitations exist in terms of staff training and time requirements. However, future work will focus on the full automation of developing such assays. Novel methodologies to predict amplification curve behaviours will be developed. One example is the brand-new algorithm for designing multiplex PCR primers using Dimer Likelihood Estimation by Xie *et al.* 2021 [212]. Another future aspect of this research is to further increase the level of multiplexing by using more fluorescent channels and by increasing inter-target curve shape differences. The development of a 21-plex using probe-based chemistries for three different fluorescent channels is an ongoing work (see Supplementary Figure D.4). Moreover, studies on the modulation of the amplification curve are conducted by changing the concentration levels of the fluorescent probe, increasing inter-target distances of amplification curves, easing the ACA classification with better clustering performance. All the above-mentioned future works will inspire the use of the ACA method for a broad range of applications and significantly increase its flexibility and scalability.

CHAPTER LESSON

In this Chapter, for the first time, a complete pipeline for developing optimal data-driven multiplex assays was presented, opening the usage of the Amplification Curve Analysis (ACA) method to the broad scientific community. The development of novel molecular tests is finally revealed, enabling easy-to-develop, easy-to-use, rapid and cost-effective data-driven molecular diagnostics solutions.

TAKEAWAY QUESTION

"Can intelligent assay design and development be utilised for point-of-care instruments or in other fields (outside of infectious diseases)?"

Chapter 9

Application of Intelligent Assay Design Strategies

9.1 Chapter Overview

The previous Chapter showed the development of Data-driven approaches, their application to different chemistries and instruments, and the optimisation process from improving development pipelines to increasing the accuracy of the methods for detecting viral and bacterial infections. This presents a more rapid, affordable and scalable solution than existing healthcare systems methods without the changing standard diagnostics pipelines. Moreover, several fields can benefit from adopting tailored chemistries and novel data analytics algorithms for wide-scale applications such as Point-of-Care (PoC) or RNA signature translation. This first part of this Chapter serves as a short overview of some studies that have used Ion-Sensitive Field-Effect Transistor (ISFET) arrays for nucleic acid detection to identify applications and ongoing research which will significantly benefit from algorithms such as that presented in Chapters 4 and 7. In particular, these studies show the first steps towards moving diagnostics (i.e. DNA detection, quantification and multiplexing) directly to the patient. The second part overviews a novel technique that enables the development of nucleic acids-based assays for optimal translation of transcriptomic diagnostic and prognostic signatures, primarily (but not limited) from high-throughput sequencing data to a PCR-based platform.

9.2 COVID-19 detection with Point-of-Care Devices

A worldwide health emergency, the COVID-19 pandemic is characterised by a rapid transmission rate and a steady rise in cases worldwide. To identify and isolate patients, stop the transmission of the virus, and direct clinical management, quick point-of-care diagnostics to identify the causal virus, SARS-CoV-2, are urgently required. The creation of a quick Point-of-Care (PoC) diagnostic test ($<$ 20 min) based on RT-LAMP and semiconductor technology is disclosed in this work to detect SARS-CoV-2 from extracted RNA samples. The following extract is taken from:

- Rodriguez-Manzano J, Malpartida-Cardenas K, Moser N, Pennisi I, Cavuto M, Miglietta L, Moniri A, Penn R, Satta G, Randell P, Davies F, Bolt F, Barclay W, Holmes A, Georgiou P. "Handheld Point-of-Care System for Rapid Detection of SARS-CoV-2 Extracted RNA in under 20 min." *ACS Central Science*, 2021 Feb 24;7(2):307-317.

9.2.1 ISFET array

In 1970, Bergveld introduced the Ion-Sensitive Field-Effect Transistor (ISFET), a field sensitive transistor used to measure ion concentration in solution [224]. ISFET has a gate electrode separated from the channel by a barrier sensitive to hydrogen (H^+). As nucleic acid amplification naturally releases protons (H^+), ISFETs are ideal for DNA detection because insulators like silicon dioxide (SiO_2), silicon nitride (Si_3N_4), and aluminium oxide (Al_2O_3) are ideal candiates for monitoring the concentration of released proton. Chemistry such as LAMP produces 50X more amplicons than PCR, so the production of H^+ is much larger. As a consequence, LAMP is perfect chemistry to couple with Point-of-Care ISFET-based devices [225].

ISFETs produced using unaltered complementary metal-oxide semiconductor (CMOS) technology are used in an embedded lab-on-chip (LoC) device that we recently reported [125, 226] for label-free electrochemical biosensing applications [227]. This device is compatible with isothermal tests, has integrated heat management, and can detect nucleic acids by keeping track of pH variations during nucleic acid amplification. As Figure 9.1 shows, the platform shows adaptability to a variety of targets. When used with a sample preparation module, it is compatible with real-time RT-LAMP (RT-eLAMP) and various sample types.

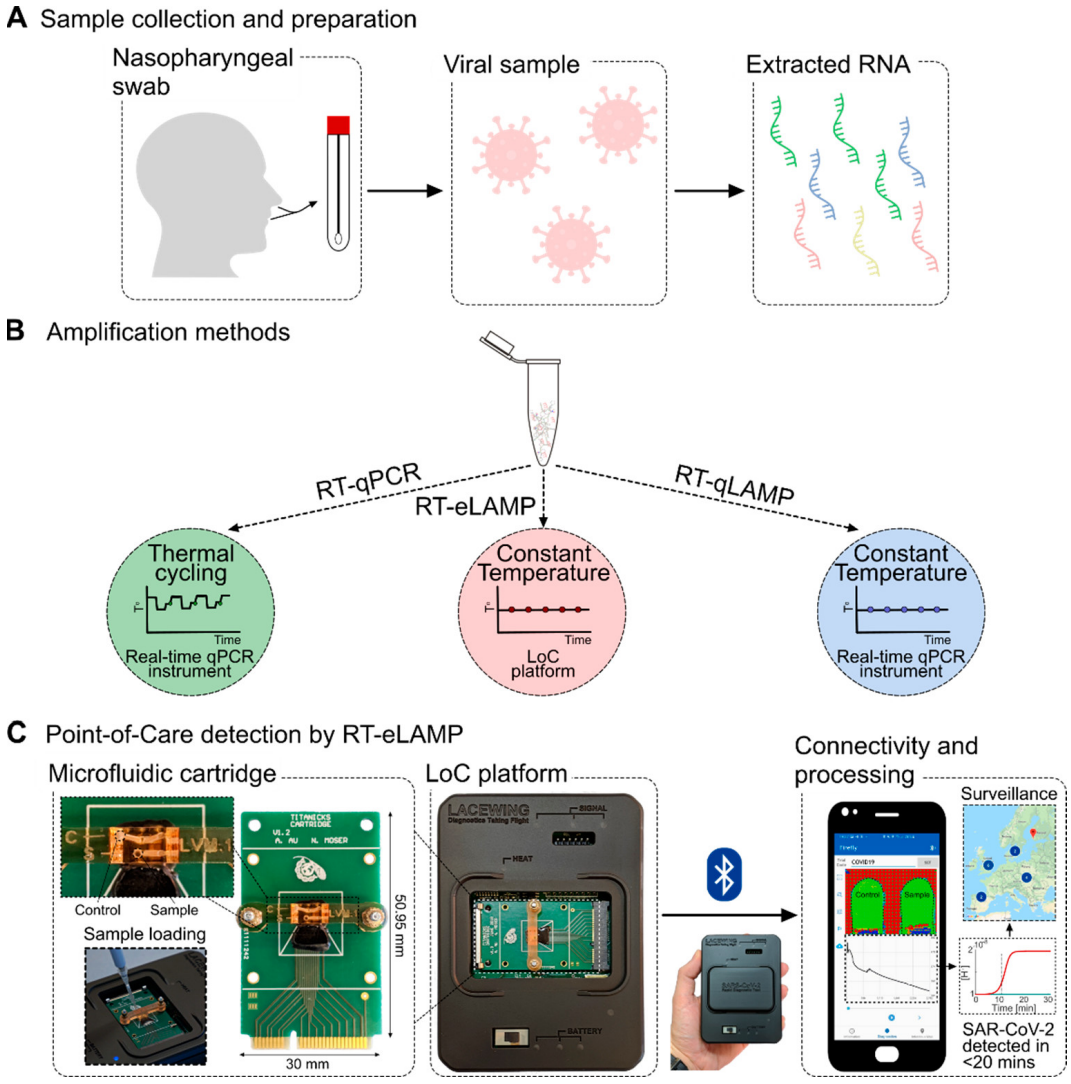


Figure 9.1: PoC diagnostic workflow. (A) Sample collection and preparation illustrating nasopharyngeal swab and RNA extraction. (B) Nucleic acid amplification methods for SARS-CoV-2 RNA detection were used in this study (RT-qPCR, RT-qLAMP, and RT-eLAMP) [146]. Thermal profiles are illustrated for comparison of the assays. (C) Point-of-care diagnostics by RT-eLAMP showing the proposed handheld LoC platform including the microfluidic cartridge with control and sample inlets, and the smartphone-enabled application for geolocalization and real-time visualization of results.

9.2.2 LAMP Assay Optimisation using Tailored Assay Design

In this work, we designed and optimized an RT-LAMP assay targeting the nucleocapsid (N) gene of SARS-CoV-2 based on collated sequences from available databases [168]. To validate the assay, we used a real-time benchtop instrument (RT-qLAMP). We have designed and optimized an RT-LAMP assay targeting the N gene of SARS-CoV-2, named LAMPcov. The N gene was selected as the optimal target since it is conserved across available sequences and more resilient to emergent mutations.

Extensive database analysis was conducted in NCBI and GISAID EpiCov databases. We developed an algorithm using python, and local blast query to analyse a total of 8,921 sequences across different countries such as China, USA, and United Kingdom [229, 42]. After inclusivity and exclusivity analysis, we were able to detect the most conserved region with the highest coverage and using an optimised primer design tool based on primer3 we designed our covid assay [50]. Primer sequences and the location in the gene can be found in Figure 9.2. Moreover,

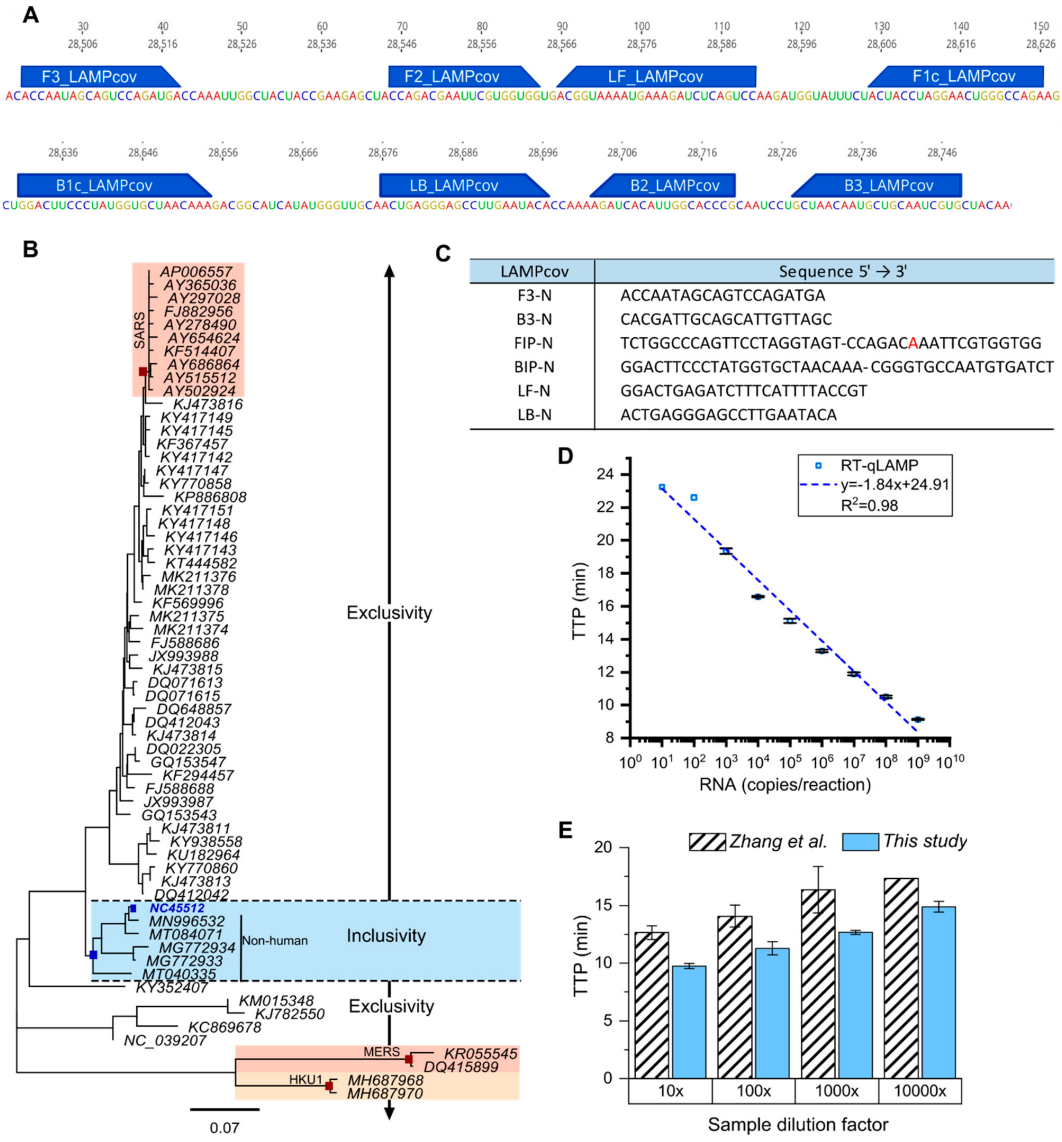


Figure 9.2: Phylogenetic analysis and LAMPcov assay design. (A) Reference sequence NC_45512 SARS-CoV-2 showing priming regions. (B) Phylogenetic tree showing the specificity of the amplicon for SARS-CoV-2 detection. Clades shadowed in blue include the reference sequence NC_45512. Clades highlighted in light red include HKU1, SARS, and MERS, all distant from the inclusivity clade. (C) Sequences of primers of the LAMPcov assay. One mismatch was introduced in F2 to avoid hairpin formation of the primer (in red). (D) A standard curve with RT-qLAMP using a control RNA at concentrations ranging between 10^1 and 10^9 copies per reaction. (E) Comparison between our assay (blue bars) and the published assay by Zhang et al [228]. (striped bars). Concentrations (dilution factor) of a clinical sample are plotted against TTP (minutes).

we compared our assay performance with available COVID-19 LAMP assays and summarized the results in Appendix Figure E.1.

9.2.3 Case Study

The previously described real-time benchtop device (RT-qLAMP) was used to test the newly constructed COVID-19 LAMP assay, and it revealed a lower limit of detection of 10 RNA copies per reaction, using viral particles. To clinically validate the use of the newly developed LoC device and the LAMP assay, a total of 183 clinical samples, including 127 positive ones, were used to validate this test and compared with gold-standard CDC COVID-19 RT-PCR tests [230].

When compared to RT-qPCR, the results demonstrated 90% sensitivity and 100% specificity, with average positive detection times of 15.45 ± 4.43 min. A subset of samples was examined (N = 40) to validate the integration of the RT-LAMP assay onto the PoC platform (RT-eLAMP), and the results showed average detection times of 12.89 ± 2.59 min for positive samples (N = 34). This performance was comparable to that of a tabletop commercial instrument. This portable diagnostic tool with secure cloud connectivity will allow real-time case identification and epidemiological surveillance when paired with a smartphone for result visualisation and geo-localization.

9.3 From Sequencing Data to PCR-based Diagnostics

This Section focuses on the need for novel frameworks for biomarker discovery and tailored assay design for PCR-based platforms based on user-provided or public RNA-Sequencing data. A case study is presented to showcase the impact on the translation of RNA signatures when primer design constraints are not considered. With the growing popularity of RNA signature-based tests, it is expected that this research will build a bridge between high-throughput experiments and molecular validation in conventional and Point-of-Care PCR-based instruments. This method has been patented, and the following is taken from a paper in preparation for submission:

- Jackson H*, Miglietta L*, Habgood-Coote D, ..., Rodriguez-Manzano J, Kaforou M, Levin M. "Diagnosis of multi-system inflammatory syndrome in children by a whole-blood transcriptional signature". *In preparation for submission*, 2022. *First joint authorship.
- Rodriguez-Manzano J, Jackson H, Miglietta L, Habgood-Coote D, Kaforou M. "A Method to optimise transcriptomic signatures", GB2211707.1, Assignee: Imperial Innovations Limited, 2022.
- Levin M, Kaforou M, Rodriguez-Manzano J, Jackson H, Miglietta L. Diagnosis of multi-system inflammatory syndrome in children by a whole-blood transcriptional". Signature Assignee: Imperial Innovations Limited, 2023.

9.3.1 RNA Diagnostics Signatures

The diagnostic signature is defined as a small number of host molecules (in this case RNA transcripts) that when combined, can distinguish between groups of interest. The discovery of diganostic signature occurs using high-dimensional 'omic data obtained from individuals in disease groups of interest (e.g. bacterial and/or viral infections), including, but not limited to, RNA-sequencing (RNA-seq) gene counts, exon counts, or microarray data [231]. Infection gives rise to certain changes in gene expression, leading to disease-specific patterns of RNA transcripts [232]. Discovering an RNA signature involves identifying a small number of transcripts that can distinguish between disease groups of interest, such as bacterial infections vs. viral infections

[233]. The initial step to define these patterns is to sequence the hosts' transcriptomes using a high-throughput method such as RNA-seq and develop a bioinformatics pipeline to identify the key combination of transcripts that characterise a certain disease. First, a filtering method such as differential expression analysis (e.g., DESeq2, EdgeR) is used to reduce the number of genes considered and retrieve only genes that contribute to disease-specific features, for example using statistical significance measures such as p-values and $log2$ fold-change values in RNA-seq counts. Using the filtered features, a feature selection algorithm is applied to the data to identify a signature composed of a small number of features that can distinguish between the disease groups of interest.

9.3.2 Novel Bioinformatics Pipeline to Translate RNA Signatures to PCR-based Tests

A new paradigm of diagnostic testing is urgently needed to guide the clinical care of patients where disease or pathogen identification is insufficient to guide treatment and prognosis, or in cases where traditional diagnostic methods fail to identify the disease-causing organism. High-throughput host transcriptomics, such as through RNA-seq, offers an alternative to traditional diagnostic processes. Despite the extensive benefits of RNA-seq or other high-throughput methods, they cannot be directly used for clinical diagnostics due to high costs and lengthy laboratory and analytical stages. On the other hand, state-of-the-art PCR assays are a much more viable alternative to host transcriptomics and can feasibly be integrated into existing clinical practices. Companies are tying to fill the gap barrier, in using host gene expression for molecular diagnostics, one example ius the Cephid Xpert MTB/RIF test. However, there are still limitations and barriers in moving from the highly accurate RNA-seq analysis to a sensitive and specific host response RT-qPCR-based test. A key reason for this gap is the signatures' compromised performance when transferred to more straightforward detection and quantification platforms.

PCR assay design constraints must be considered when designing RNA signature-based diagnostics tests. When targeting a gene using bioinformatic approaches, the amplicon has several features to consider, such as GC content, sequence length and secondary structure formations. equally for the primer design process, GC content, T_m, 3'-clamp, hairpin and cross-

priming have to be considered. We developed a new method that integrates these constraints when translating RNA signatures into PCR-based diagnostic tests.

9.3.3 Case Study

Our method for identifying optimal RNA targets based on RT-qPCR design constraints offers multiple unique contributions to the field of host diagnostics. For example, it is a workflow that does not depend on commercial primers, thus avoiding their high costs and design constraints. Our method incorporates a bespoke primer design, which provides more control over the targets of interest. For example, during the exploration genes related to Multisystem inflammatory syndrome (or MIS-C) [234] several exons associated with LEPROT gene [235] showed a strong correlation. To develop diagnostic tests targeting LEPROT exons, evaluation of a commercial primer pair and an *in-house* optimising primer (based on RT-qPCR) constraints were tested. The location of the primer binding sites and the exon were different:

- location 1:65,900,457-65,900,598 (BIORAD LEPROT gene, human [236])
- location 1:65,425,301-65,425,378 *in-house* design, reference transcript: ENSE00003644138

Using our *in-house* primer design approach, we tested several primer sets for genes included in the MIS-C signature. For each gene, we found the optimal primer that best translated the RNA signature to the molecular method approach. The approach is shown in Figure 9.3 for one of the genes of interest, where several assays have been designed with the purpose of generating a primer pair that best describes the differential expression of genes related to MIS-Cand not in children with Kawasaki disease, bacterial infections, or viral infections [234]. The figure shows four different assays, which we named VIP_01, VIP_02, VIP_03 and VIP_04. After testing them at the same condition with 48 clinical samples of patients with and without MIS-C diagnosis, we plot the C_t of each of them in a box plot format, from which we can observe that:

- VIP_01 had good C_t, but the difference with MIS-C or not MIS-C samples could not be appreciated because of the C_t distributions.
- VIP_02 assay worked for the majority of samples, but 7 of them didn't show signal

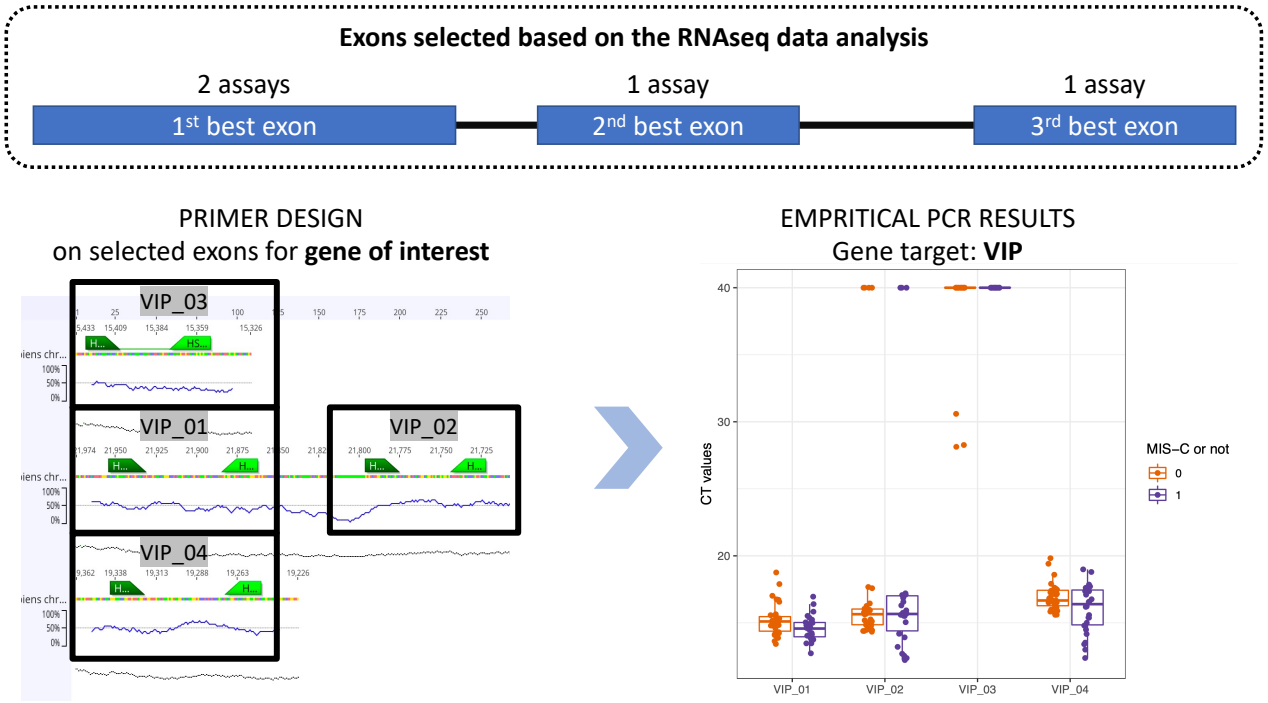


Figure 9.3: Assay design strategy and performance in qPCR instrument

- VIP_03 failed for all but three samples.
- VIP_04 had good C_t and translated the signature as the fold-change within MIS-C or not MIS-C samples could be appreciated.

This experiment highlighted that not all the exons could be considered for the molecular method in PCR as primer design constraints and the sequence of the targeted nucleic acid has a massive impact on the performance of each assay. Moreover, laboratory testing confirms that using standard primer design parameters does not necessarily result in the best-performing assay being translated across platforms from high-throughput to low-throughput methods. Firstly, this confirms that the RNA-seq and primer design constraints must be combined to increase the success rate of the assay. Furthermore, this shows that primer design parameters must be adjusted according to the target to obtain the best discrimination of the groups of interest. Another factor that can be missed when designing PCR assays is the presence of the targeted region as it can be subject to splicing events, or gene expression can be modified by the presence of other unknown or undiagnosed conditions.

This study represents a shift in the entire process of discovery of gene signatures. These pipelines are built with the final validation platform in mind, which is not done in transcrip-

tomics. Developing discovery pipelines based on the shortcomings, requirements and capabilities of the intended validation platforms is a novel concept in the field of transcriptomics. Furthermore, a unique aspect of this work is that it considers the different regions of genes, determining the most optimal area for cross-platform translation. Appendix Figures E.2 and E.3 are provided to further clarify the methodology workflow.

CHAPTER LESSON

In this Chapter, we have delineated the application of innovative assay design methods and data-driven techniques. Specifically, the utilization of Point-of-Care instruments and chemistries to enhance their throughput by employing advanced data analytics methods is an exceedingly attractive future prospect for diagnostic purposes in low and middle-income countries. Furthermore, the potential of these innovations can shed light on novel molecular tests, such as RNA signature translation, which are currently unattainable. The scope of data-driven methods for molecular diagnostics is extensive and untapped, but its potential is immense and ready to be explored.

Chapter 10

Conclusion & Future Perspective

This thesis explored the use of novel Data-driven approaches in the field of DNA detection and demonstrated the value that could be obtained from the available data from the sigmoidal trend of a PCR-based amplification reaction. Particularly, PCR and LAMP (as an isothermal alternative) were used to investigate the application of Data-driven Multiplexing in both quantitative and digital molecular instruments. In this work, we explored the use of Artificial Intelligence algorithms (such as Machine Learning) to enhance the throughput of conventional instruments, and we proposed a new perspective for the hybrid development of high-level multiplex assays (up to nine targets) using laboratory tests, bioinformatics tools and mathematical modelling. With this thesis, I would like to pass on the following message:

"*Sigmoidal curves from amplification reactions are not Binary, and the better use of the data is essential to develop novel higher throughput molecular tests.*" - Luca Miglietta

10.1 Contribution

Chapter 3 revealed that machine learning algorithms coupled with the large amount of data from real-time digital instruments can be used for PCR multiplexing based on the entire amplification curve. This was accomplished by explicitly training a supervised machine learning model to classify targets using target-specific kinetic information automatically. A formula was developed to provide a trade-off between quantification and multiplexing because digital PCR executes spatial multiplexing by default and quantification by Poisson statistics (rather than

standard curves) by default. Chapter 3's contributions resulted in [J1] and [P1] (see List of Publications): Paper conceptualisation, writing original draft, final review, data collection, algorithm development, assay design and experimental testing.

In Chapter 4, a high-level digital PCR multiplex using intercalating dyes was demonstrated. This was accomplished using a revolutionary three-step machine learning approach that combined the kinetic information from amplification curves and the thermodynamic information from melting curves. It was demonstrated that despite analysing the same nucleic acid product, both amplification and melting curves contain non-mutual valuable information for target identification. Chapter 4's contributions resulted in [J2] and [P1] (see List of Publications): paper conceptualisation, writing original draft, final review, data collection, data processing, algorithm development, assay design and experimental testing.

Chapter 5 extended the use of Data-driven multiplexing to isothermal chemistries, particularly LAMP. This was the first demonstration of applying an AI-based method to identify 5 respiratory pathogens with a 5plex assay in a single reaction using only sigmoidal shape information from LAMP. The contributions of Chapter 5 led to the [J6] (see List of Publications): paper conceptualisation, final review, data collection, data processing, algorithm development, assay design and experimental testing.

Chapter 6 demonstrated the first successful application of data-driven multiplexing for clinical diagnostics. A multiplex assay for the detection of the "big 5" carbapenem resistance genes was developed in this study. Coupling a Machine Learning classifier and the novel 5plex assay, using the information encoded in amplification and melting curves, the classification of 253 clinical isolates was achieved with over 99 % accuracy. Moreover, the demonstration of the digital standard curve was introduced in this Chapter for the first time. The contributions of Chapter 6 led to the [J5] (see List of Publications): paper conceptualisation, writing original draft, final review, data collection, data processing, algorithm development, assay design and experimental testing.

Chapter 7 introduced a new framework for enhanced data quality from digital PCR instruments and outlier detection. This was achieved by fitting the amplification curve and combining the resulting feature with novel ones. Removing outliers based on the sigmoidal trend allows more precise detection in instruments and chemistries without melting capabilities and enables better quantification in digital applications. The contributions of Chapter 7 led to the [J10]

and [S3] (see List of Publications). Contribution: paper conceptualisation, writing original draft, final review, data collection, data processing, algorithm development, assay design and experimental testing.

Chapter 8 demonstrated the first-ever hybrid pipelines to develop Data-driven Multiplex assays. Through better data use, the Smart-Plexer utilises laboratory assays (in singleplex) to *in-silico* compute optimal simulated multiplex assays. This drastically reduces the resources needed in developing high-level multiplex for data-driven approaches. We demonstrate the concept in a 7plex assay for Respiratory tract infectious pathogens (COVID-19 related). The contributions of Chapter 8 led to the [S2] and [P2] (see List of Publications): paper conceptualisation, writing original draft, final review, data collection, data processing, algorithm development, assay design and experimental testing.

Chapter 9 identified recent applications of intelligent assay design pipelines to highlight the future direction of this field. The ongoing research on Point-of-Care and RNA signature diagnostics tests will benefit from incorporating more sophisticated data-driven methods like that in Chapters 3 and 7. The contributions of Chapter 9 led to the [J3], [J4], [J8], [J9] and [S5] (see List of Publications): paper conceptualisation, final review, data collection, data processing, assay design and experimental testing.

10.2 Remarks, Impact and Future Perspective

This thesis directly targets the need for rapid diagnostic tests. As the COVID-19 pandemics highlighted, shortages in test kits paralysed the health system in many countries. Data-driven methods for multiple target detection aim to ease the burden of molecular diagnostic tests in hospitals (such as NHS) and reduces diagnostics cost for low and middle-income countries' applications. Furthermore, the apporach represent an answer in new infectious outbreaks and optimise diagnosis outcomes for patient treatment.

The social impact of new diagnostic solutions allows better infection prevention and reduces morbidity associated with delayed or inappropriate treatment. Accurate diagnosis informs real-time bed planning and escalation of care pathways improving patient and healthcare worker safety. Furthermore, improved tools with faster turnaround time prevent transmission. The

proposed research can be applied to other field requiring nucleic acid detection and, when combined with POC instruments, data-driven approaches enable better diagnostics outside of the lab and in low- and middle-income countries.

There is also an economic impact related to this work. More diagnostic companies are moving towards integrating artificial intelligence and advanced data processing to enhance the throughput of their chemistries or instruments (such as ChromaCode, Diacarta and Diagnostic.ai). This is especially the case for multiplexing. Several digital PCR providers are developing software that, by integrating Machine Learning algorithms in their data processing pipelines, can perform multiple detections in a single-molecule reaction [237]. This is highly crucial when we analyse the cost of individual PCR reactions. As Figure 10.2 shows, the price of PCR can vary based on chemistry and level of multiplexing. It can be seen that probe-based assays are 54% more expensive than intercalating dye chemistries for single target detection, reaching even 157% increase with three target detection [238, 239]. This highlights how PCR cost can be reduced in intercalating dye multiplexing. We broadly illustrate how the AMCA method can achieve five target detection with a single intercalating dye in a single well reaction [144].

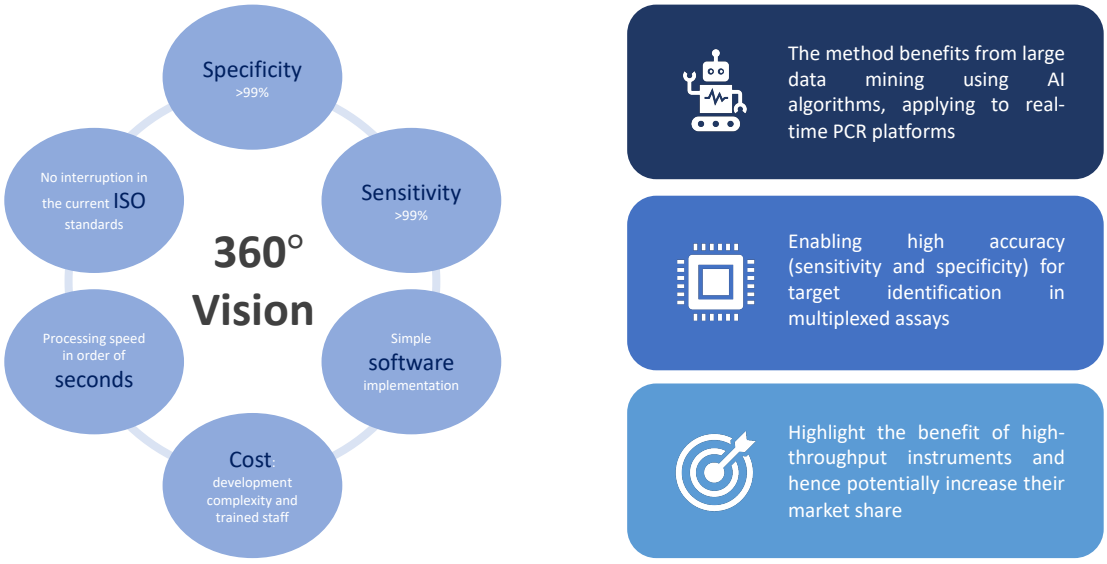


Figure 10.1: Vision of Data-driven Multiplexing

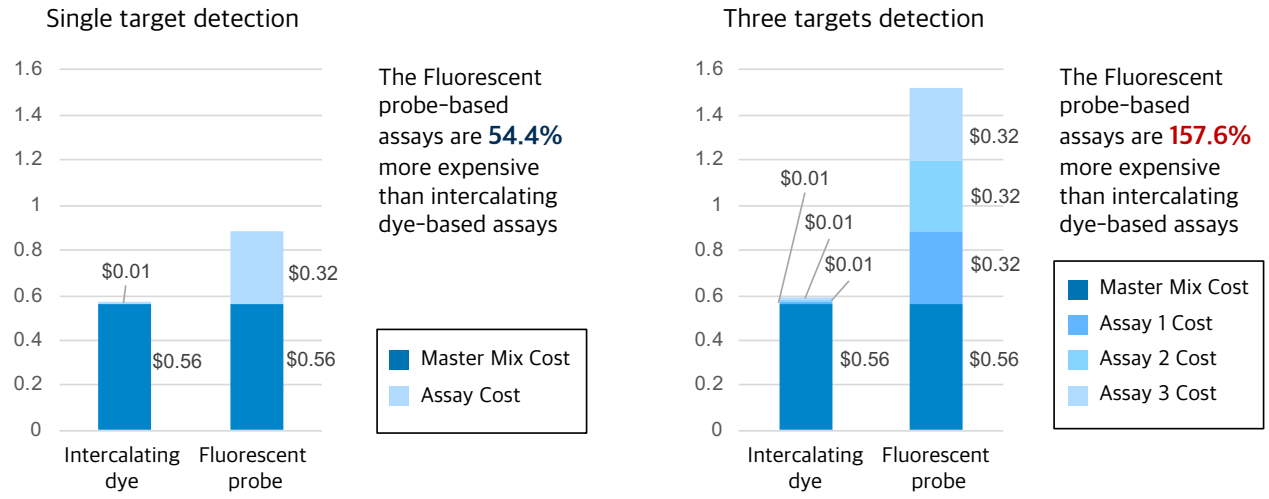


Figure 10.2: The cost of PCR per reaction. On the right, the cost of PCR per reaction when detection of single target is performed (intercalating dye-based vs fluorescent probe-based assay). On the left, the cost of PCR per reaction when detection of three targets is performed (intercalating dye-based vs fluorescent probe-based assay).

10.3 Future Work

1. Optimisation of data-driven multiplexing classification by using deep-learning and transfer learning methods (the first study has been already conducted [240]).

2. the expansion of the multiplex level by using multiple colour with TaqMan probes (see the 21-plex RTI panel in three fluorescent channels using the ACA method, Supplementary Figure D.4).

3. the integration of the developed data-driven multiplexing (i.e. ACA) into a Point-of-Care platforms.

4. mathematical modelling of the amplification curve kinetics and thermodynamics to development of data-driven multiplex assays fully *in-silico*.

The contents of this thesis are the outcome of a highly multidisciplinary collaboration amongst several disciplines, specifically data science, engineering, and molecular biology. This may help to explain why there hasn't been much research done on this topic; this thesis intends to fill that gap. There is a growing trend among biologists to learn more about data processing, and given the worldwide focus on COVID-19 pandemic, the field of DNA detection has seen a surge in interest among data scientists. I therefore expect that by disseminating the concepts in

this thesis, others will be able to modify and improve this work in order to address the biggest healthcare concerns in the world.

Appendices

Appendix A

Supporting Information: Chapter 3

This Appendix contains the following:

- Raw melting curves from qPCR and dPCR (Figure A.1)
- Standard curves for each mcr target using new 9-plex assay (Figure A.2)
- Performance of high-level dPCR multiplexing with FFI (Figure A.3)
- Performance of all methods in conventional qPCR (Figure A.4)

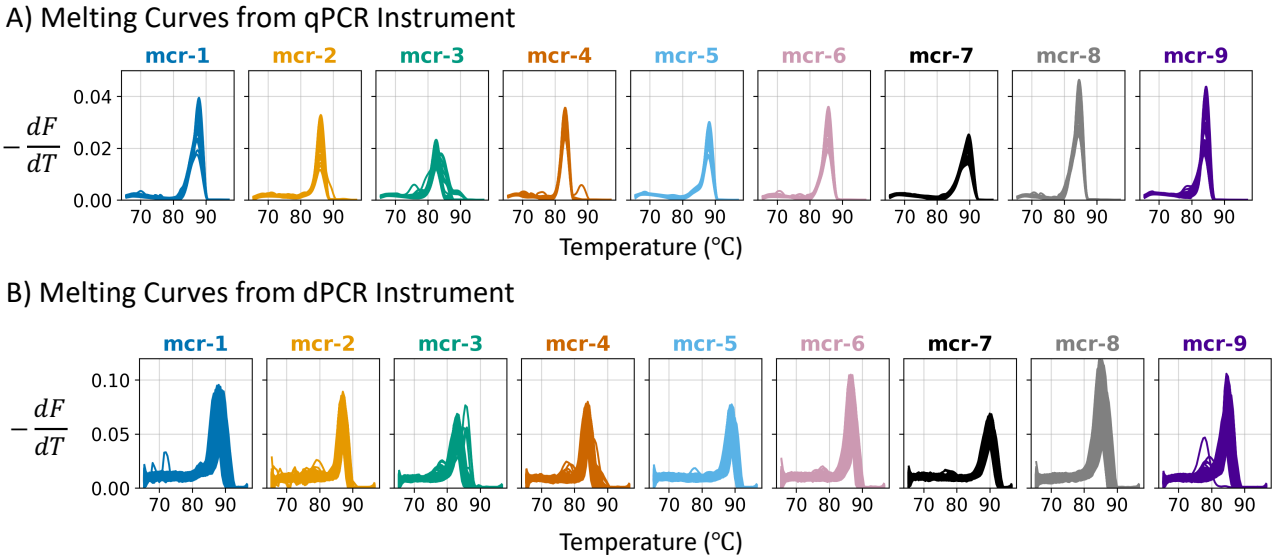
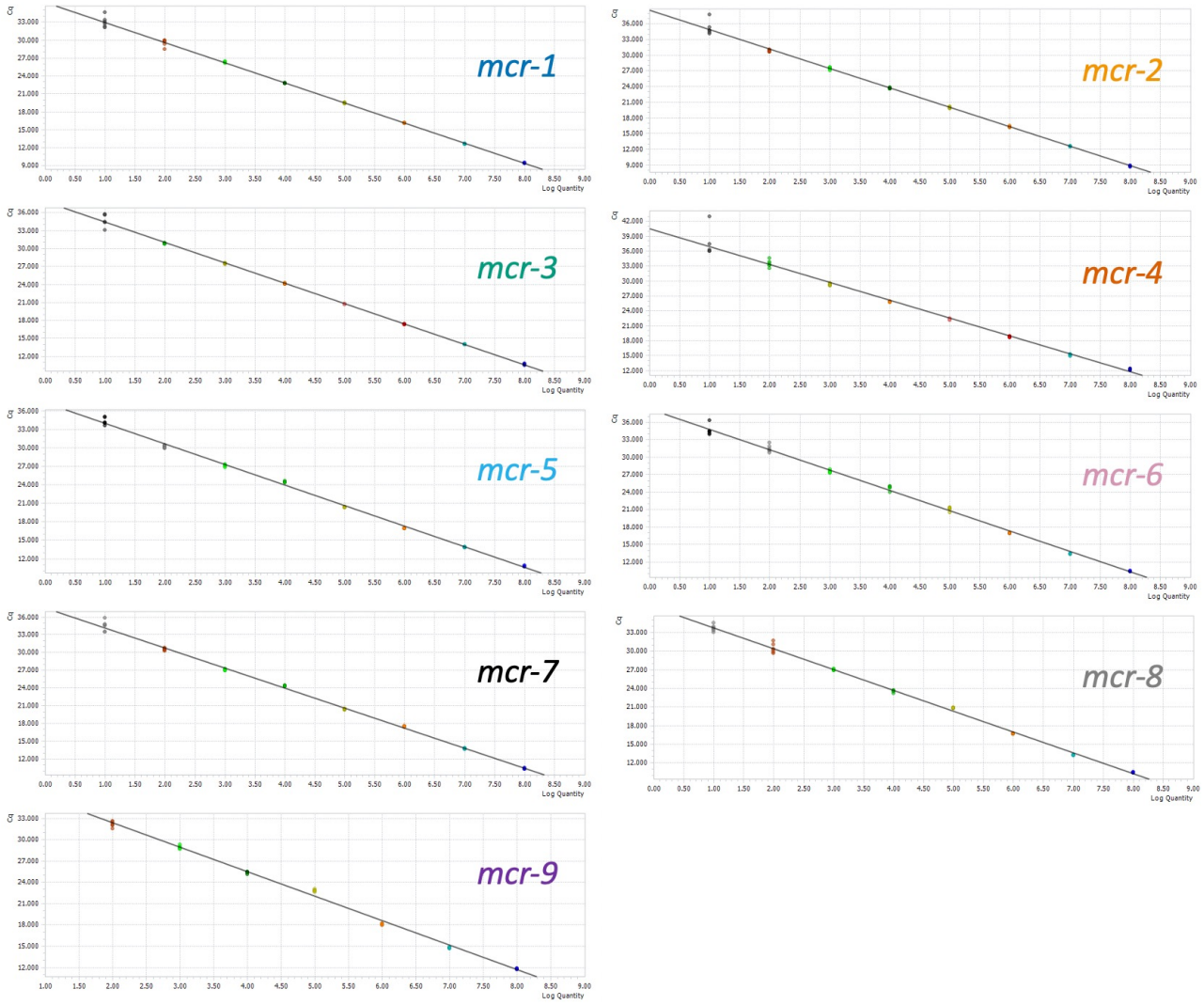


Figure A.1: Raw melting curves from (A) qPCR and (B) dPCR instrument.



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 A.2: Standard curves for each *mcr* target using new 9plex assay. (Top Panel) Plots were generated using Roche LightCycler software (version 1.1). (Bottom Panel) Table with relevant meta data for each standard curve.

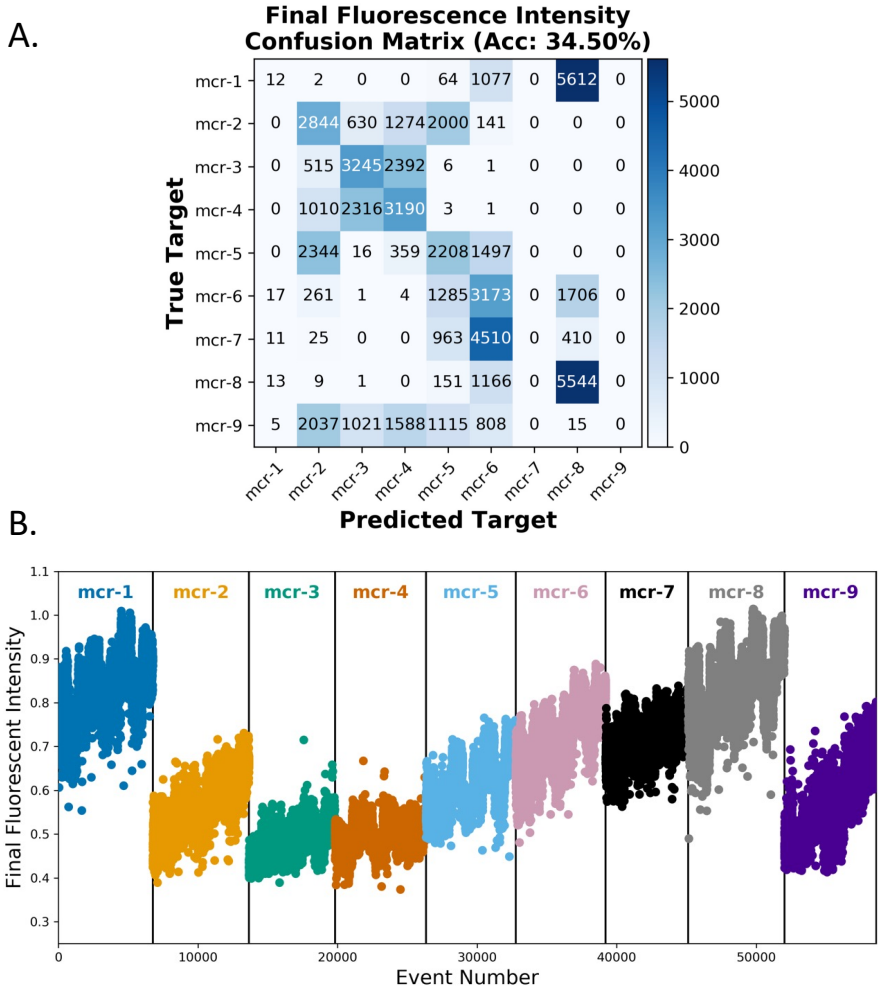


Figure A.3: Multiplexing with FFI in dPCR (without optimization of primer concentration). (Top panel) Confusion matrix showing performance of logistic regression classifier on FFI values. (Bottom panel) Visualisation of the distribution of FFI values for all targets and amplification events. For each target, amplification events are ordered from low to high concentration.

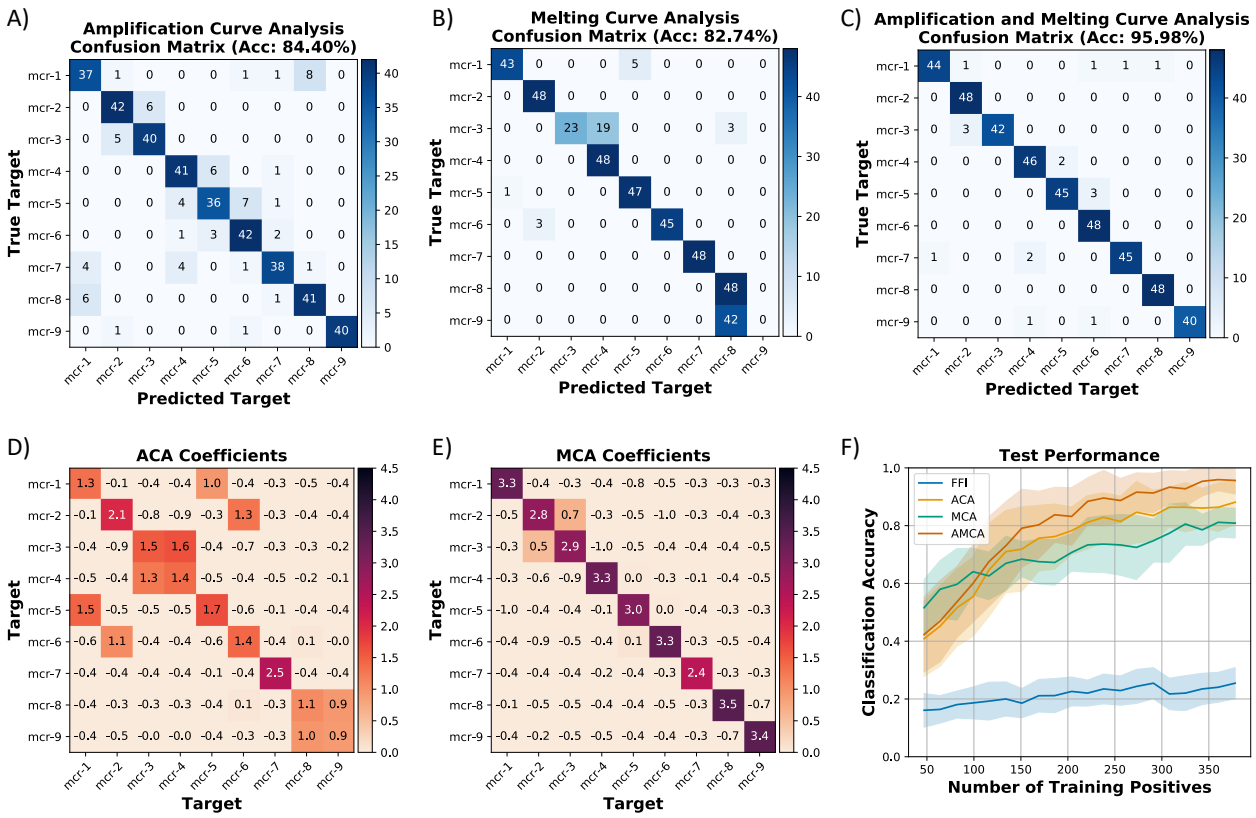


Figure A.4: Performance of all methods for multiplexing the 9 *mcr* targets in conventional qPCR instrument. A, B, C) Confusion matrices illustrating the predictions from ACA, MCA and AMCA (proposed method), respectively. Values indicate the number of amplification events with diagonal entries corresponding to correct predictions. D, E) Coefficients of the AMCA model weighting the predictions from the ACA and MCA methods, respectively. Darker colours 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.

Appendix B

Supporting Information: Chapter 5

This Appendix contains the following:

- Effect of training data size on the classification accuracy (Figure B.1)
- Distribution of Time-To-Positive in 5plex LAMP (Figure B.2)

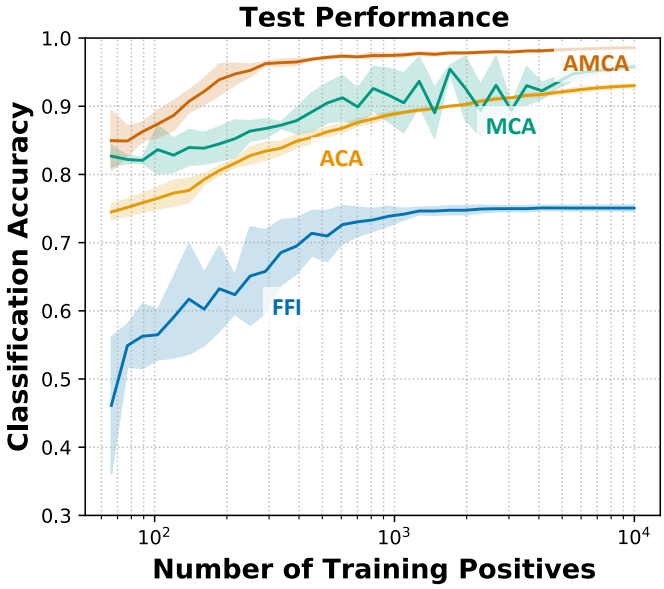


Figure B.1: Effect of training data size on the classification accuracy using 5,000 out-of-sample data points (10 iterations).

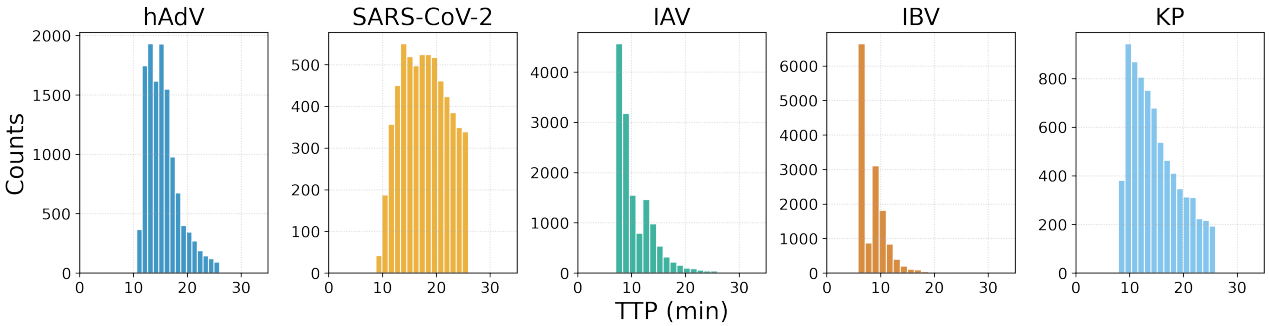


Figure B.2: Distribution of Time-To-Positive in 5plex LAMP. Histogram showing the distribution of Time-To-Positive (TTP) values of the 5plex-LAMP in dLAMP using a single fluorescence channel.

Appendix C

Supporting Information: Chapter 6

This Appendix contains the following:

Figure List

- Inclusivity alignment of $bla_{\text{OXA-48}}$ (Figure C.1)
- Inclusivity alignment of bla_{IMP} (Figure C.2)
- Inclusivity alignment of bla_{NDM} (Figure C.3)
- Inclusivity alignment of bla_{KPC} (Figure C.4)
- Inclusivity alignment of bla_{VIM} (Figure C.5)
- Analysis of real-time amplification and melting curves from qPCR instruments (Figure C.6)
- Performance of the MCA and AMCA in the training dataset using synthetic DNA templates (Figure C.7)
- Performance of MCA and AMCA methods in clinical isolates (Figure C.8)

Table List

- Clinical Enterobacteriaceae isolates used in this study (Figure C.10)
- Bacterial isolates used in this study [144]. (Table C.1 - C.5)

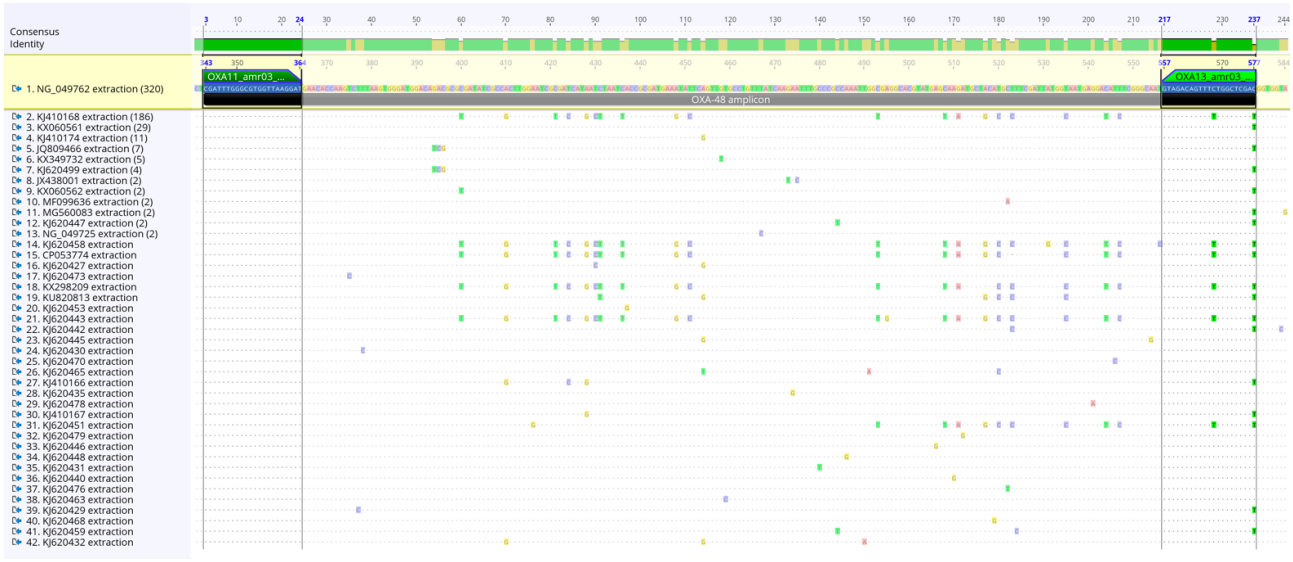


Figure C.1: Inclusivity alignment of $bla_{\text{OXA-48}}$. Sequences retrieved from nr/nt NCBI database (N = 603) with a coverage of 100% for each primer binding region. The alignment shows only unique sequences that differ from the reference NG_049762 in the amplification region. The sequences are ordered from the largest number of unique sequences to the lowest as shown in bracket (after the NCBI accession number) on the left side of the alignment.

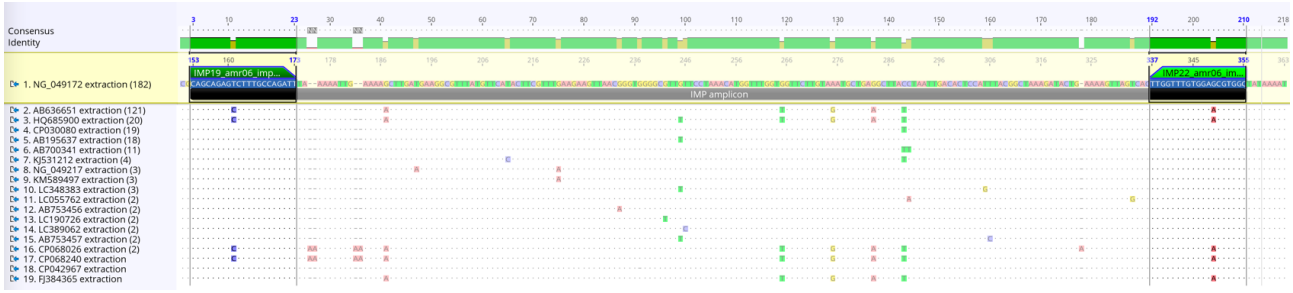


Figure C.2: Inclusivity alignment of bla_{IMP} (including $bla_{\text{IMP-1}}$ and $bla_{\text{IMP-4}}$ groups). Sequences retrieved from nr/nt NCBI database (N = 400) with a coverage of 100% for each primer binding region. The alignment shows only unique sequences that differ from the reference NG_049172 in the amplification region. The sequences are ordered from the largest number of unique sequences to the lowest as shown in bracket (after the NCBI accession number) on the left side of the alignment.

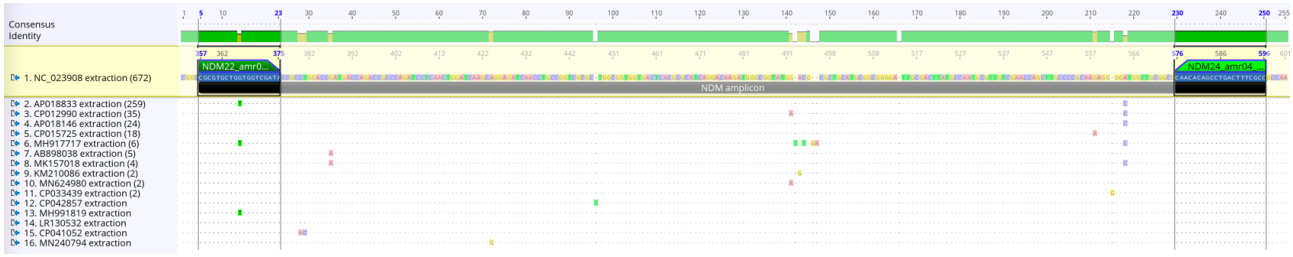


Figure C.3: Inclusivity alignment of bla_{NDM}. Sequences retrieved from nr/nt NCBI database (n = 1,035) with a coverage of 100% for each primer binding region. The alignment shows only unique sequences that differ from the reference NC_023908 in the amplification region. The sequences are ordered from the largest number of unique sequences to the lowest as shown in bracket (after the NCBI accession number) on the left side of the alignment.

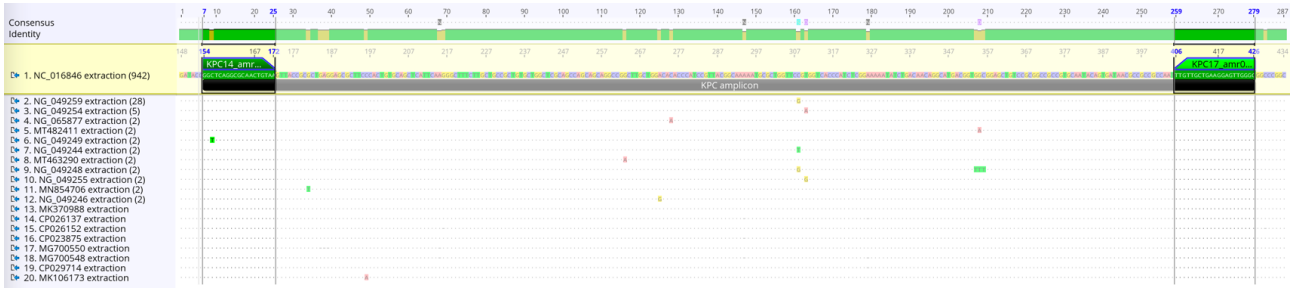


Figure C.4: Inclusivity alignment of bla_{KPC}. Sequences retrieved from nr/nt NCBI database (n = 1,001) with a coverage of 99.9% for each primer binding region. The alignment shows only unique sequences that differ from the reference NC_016846 in the amplification region. The sequences are ordered from the largest number of unique sequences to the lowest as shown in bracket (after the NCBI accession number) on the left side of the alignment.

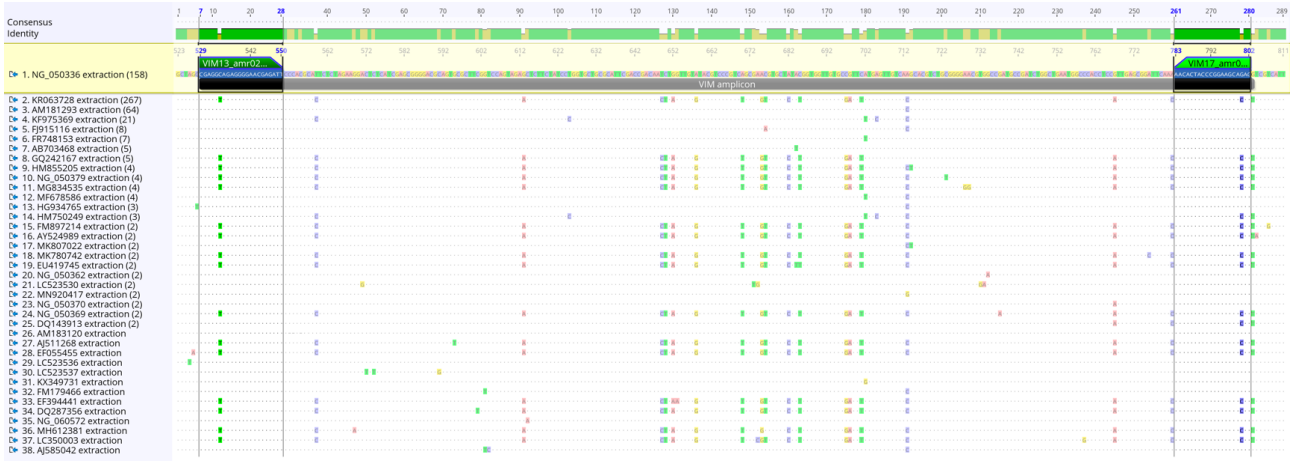
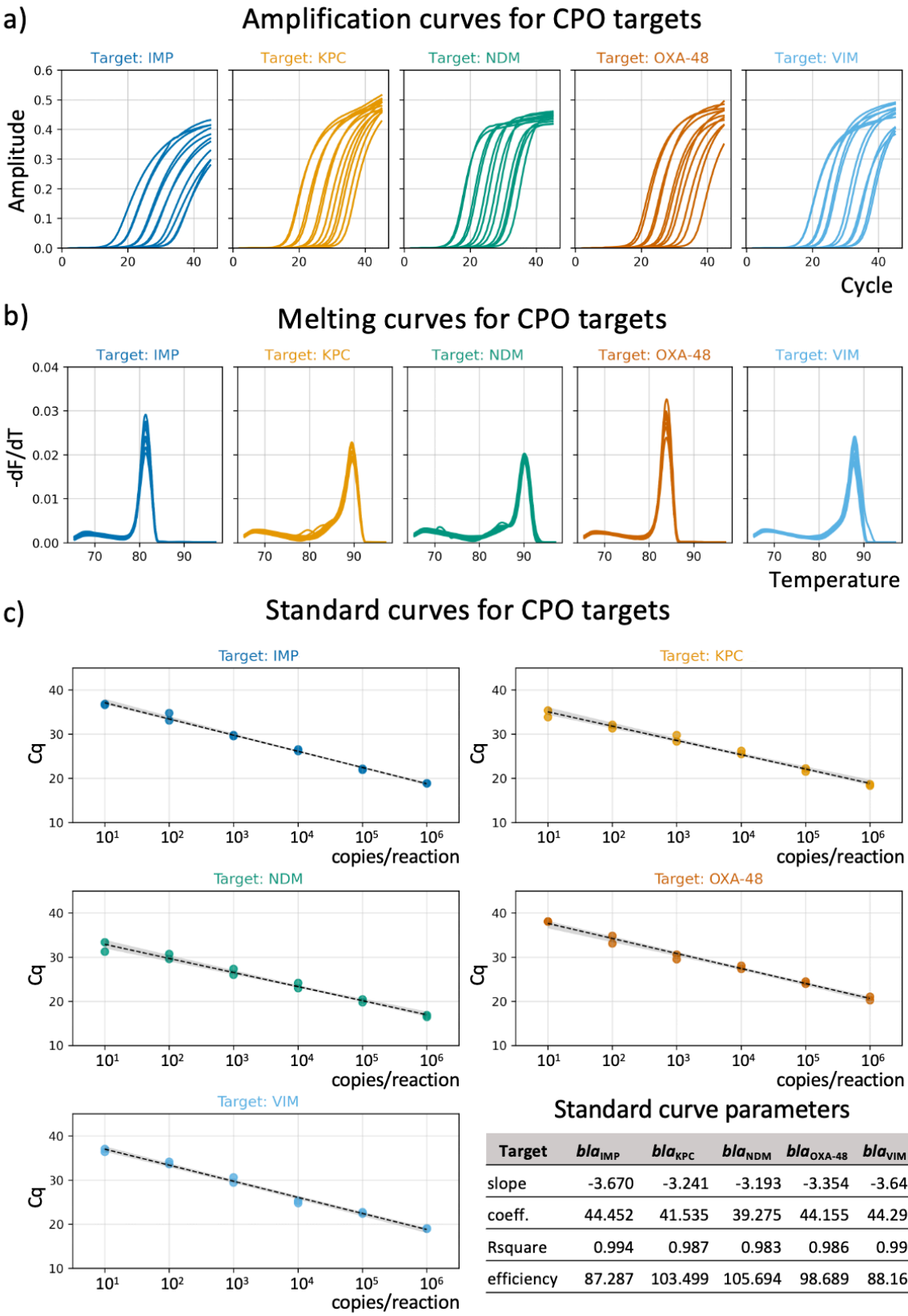


Figure C.5: Inclusivity alignment of bla_{VIM}. Sequences retrieved from nr/nt NCBI database (n = 593) with a coverage of 99.9% for each primer binding region. The alignment shows only unique sequences that differ from the reference NG_050336 in the amplification region. The sequences are ordered from the largest number of unique sequences to the lowest as shown in bracket (after the NCBI accession number) on the left side of the alignment.



Standard curve parameters

Target	bla_{IMP}	bla_{KPC}	bla_{NDM}	bla_{OXA-48}	bla_{VIM}
slope	-3.670	-3.241	-3.193	-3.354	-3.643
coeff.	44.452	41.535	39.275	44.155	44.291
Rsquare	0.994	0.987	0.983	0.986	0.990
efficiency	87.287	103.499	105.694	98.689	88.161

Figure C.6: Analysis of real-time amplification and melting curves from qPCR instruments. (a) Raw real-time amplification curves. (b) Raw melting curve for each target. (c) Standard curves for each target using our new 5plex PCR assay. (Bottom Panel) Table with relevant meta data for each standard curve.

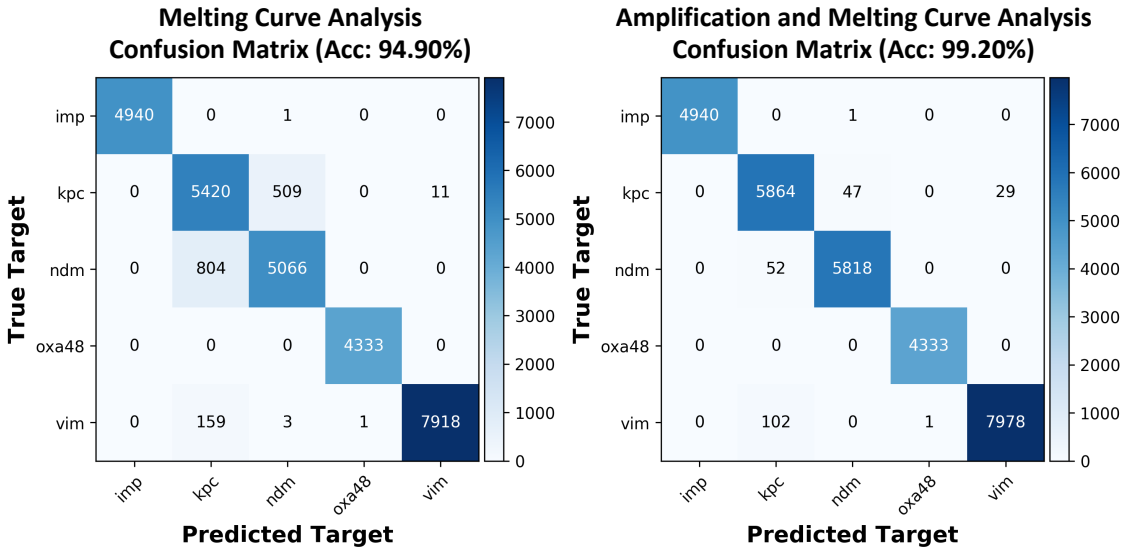


Figure C.7: Performance of the MCA and AMCA methods for multiplexing the five carbapenemase gene targets in the training dataset using synthetic DNA templates. (left) Confusion matrix illustrating the predictions from Melting Curve Analysis (MCA) algorithm. (right) Confusion matrix illustrating the predictions from Amplification and Melting Curve Analysis (AMCA) algorithm. Values in the matrices indicate the number of positive amplification events (N = 29,165) with diagonal entries corresponding to correct predictions.

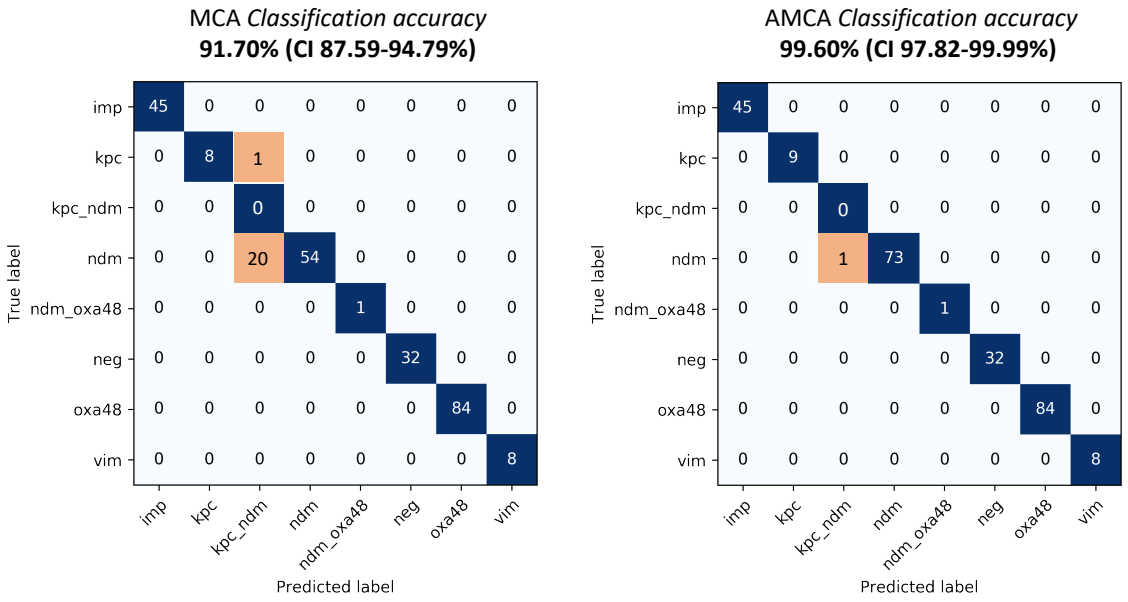


Figure C.8: Performance of MCA and AMCA methods for multiplexing the five carbapenemase gene targets in clinical isolates. (left) Confusion matrix illustrating the predictions from Melting Curve Analysis (MCA) algorithm. (right) Confusion matrix illustrating the predictions from Amplification and Melting Curve Analysis (AMCA) algorithm. Values indicate the number of clinical isolates (N=253) with diagonal entries corresponding to correct predictions.

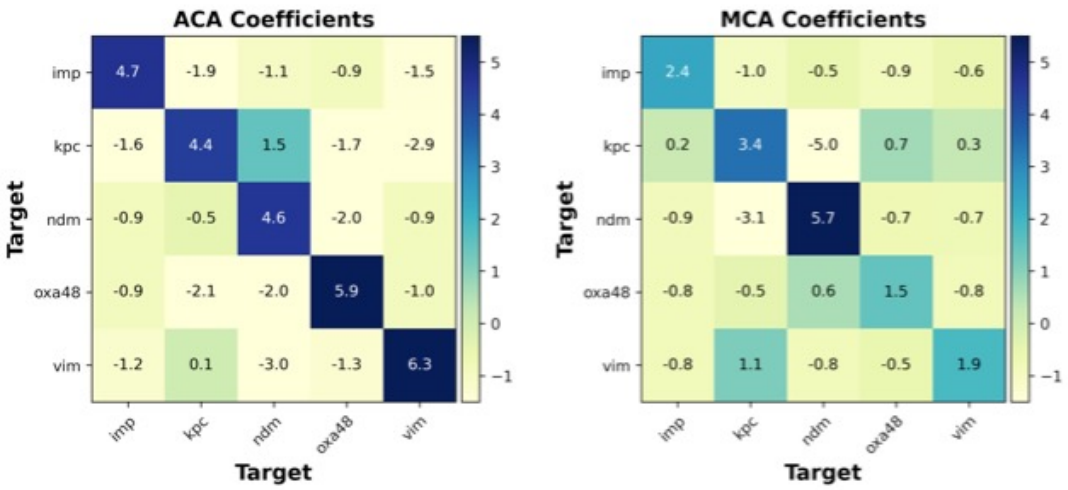


Figure C.9: The coefficients of the AMCA model. The values in the confusion matrices, ranging from -5 to 6.3, indicate the predictions weights from the ACA and MCA methods, respectively. Darker colours indicate more positive weighting. For example, as it can be observed, the AMCA weighs the prediction from ACA more heavily for bla_{IMP} target (4.7 in the ACA model compared to the 2.4 of the MCA model), instead for bla_{NDM} the situation is the opposite (5.7 in the MCA model compared to the 4.6 of the ACA model).

Species (MALDI-TOF MS)	Carbapenemase gene	Number of isolates
Citrobacter spp.	bla_{IMP}	1
	bla_{KPC}	2
	bla_{NDM}	1
	bla_{OXA-48}	10
	bla_{VIM}	1
Enterobacter spp.	bla_{IMP}	20
	bla_{NDM}	7
	bla_{OXA-48}	2
	bla_{VIM}	2
Escherichia spp.	bla_{IMP}	7
	bla_{NDM}	14
	bla_{NDM} and bla_{OXA-48}	1
	bla_{OXA-48}	26
Klebsiella pneumoniae	bla_{IMP}	15
	bla_{KPC}	6
	bla_{NDM}	51
	bla_{OXA-48}	45
	bla_{VIM}	3
Proteus mirabilis	bla_{NDM}	1
Pseudomonas aeruginosa	bla_{IMP}	2
	bla_{VIM}	2
Serratia marcescens	bla_{KPC}	1
	bla_{OXA-48}	1
Multiple species*	negative	32

Figure C.10: Clinical Enterobacteriaceae isolates.
*CPO-negative species: *Acinetobacter baumannii, Citrobacter freundii, Enterobacter spp., Escherichia coli, Klebsiella pneumoniae, and Pseudomonas aeruginosa.*

Table C.1: Bacterial isolates from clinical samples (part 1)

Sample ID	Specimen	Source	Collection	CPE	AMCA	Conc. cp/uL
CPO001	*Acinetobacter baumannii*	Bronchoalveolar	26/02/2015	oxa23	neg	0.00E+00
CPO002	*Klebsiella pneumoniae*	Urine	15/04/2013	neg	neg	0.00E+00
CPO003	*Klebsiella pneumoniae*	Rectal swab	29/05/2018	neg	neg	0.00E+00
CPO004	*Enterobacter cloacae*	Right leg tissue	13/07/2018	neg	neg	0.00E+00
CPO005	*Escherichia coli*	Sputum	10/10/2015	neg	neg	0.00E+00
CPO006	*Escherichia coli*	Urine	04/01/2016	neg	neg	0.00E+00
CPO007	*Klebsiella pneumoniae*	Urine	29/02/2016	neg	neg	0.00E+00
CPO008	*Citrobacter freundii*	Urine	09/03/2016	neg	neg	0.00E+00
CPO009	*Klebsiella pneumoniae*	Urine	10/07/2016	neg	neg	0.00E+00
CPO010	*Enterobacter cloacae*	Bronchoalveolar	15/08/2016	neg	neg	0.00E+00
CPO011	*Klebsiella pneumoniae*	Perineum	05/10/2016	neg	neg	0.00E+00
CPO012	*Klebsiella pneumoniae*	Right leg tissue	23/10/2016	neg	neg	0.00E+00
CPO013	*Klebsiella pneumoniae*	Rectal swab	26/12/2016	neg	neg	0.00E+00
CPO014	*Escherichia coli*	Urine	18/06/2017	neg	neg	0.00E+00
CPO015	*Klebsiella pneumoniae*	Urine	18/06/2017	neg	neg	0.00E+00
CPO016	*Klebsiella pneumoniae*	Urine	05/08/2017	neg	neg	0.00E+00
CPO017	*Enterobacter cloacae*	Sputum	18/08/2017	neg	neg	0.00E+00
CPO018	*Escherichia coli*	Urine	18/08/2017	neg	neg	0.00E+00
CPO019	*Klebsiella pneumoniae*	Urine	27/10/2017	neg	neg	0.00E+00
CPO020	*Klebsiella pneumoniae*	Urine	05/01/2018	neg	neg	0.00E+00
CPO021	*Enterobacter cloacae*	Wound swab	14/11/2017	neg	neg	0.00E+00
CPO022	*Escherichia coli*	Rectal swab	22/01/2018	neg	neg	0.00E+00
CPO023	*Escherichia coli*	Urine	22/01/2018	neg	neg	0.00E+00
CPO024	*Klebsiella pneumoniae*	Urine	26/01/2018	neg	neg	0.00E+00
CPO025	*Enterobacter spp*	Rectal swab	16/05/2019	neg	neg	0.00E+00
CPO026	*Escherichia coli*	Rectal swab	16/05/2019	neg	neg	0.00E+00
CPO027	*Enterobacter cloacae*	Rectal swab	12/05/2018	neg	neg	0.00E+00
CPO028	*Escherichia coli*	Rectal swab	22/04/2018	neg	neg	0.00E+00
CPO029	*Klebsiella pneumoniae*	Rectal swab	07/01/2019	neg	neg	0.00E+00
CPO030	*Pseudomonas aeruginosa*	Rectal swab	12/01/2019	neg	neg	0.00E+00
CPO031	*Escherichia coli*	Rectal swab	11/02/2019	neg	neg	0.00E+00
CPO032	*Acinetobacter baumannii*	Rectal swab	13/03/2019	neg	neg	0.00E+00
CPO033	*Pseudomonas aeruginosa*	Throat swab	31/03/2015	imp	imp	2.42E+03
CPO034	*Escherichia coli*	Rectal swab	12/05/2018	imp	imp	1.64E+04
CPO035	*Klebsiella pneumoniae*	Rectal swab	18/05/2018	imp	imp	2.06E+03
CPO036	*Enterobacter cloacae*	Rectal swab	23/06/2018	imp	imp	9.88E+01
CPO037	*Klebsiella pneumoniae*	Rectal swab	03/07/2018	imp	imp	8.00E+02
CPO038	*Escherichia coli*	Rectal swab	13/01/2019	imp	imp	3.24E+04
CPO039	*Klebsiella pneumoniae*	Rectal swab	14/01/2019	imp	imp	3.28E+04
CPO040	*Escherichia coli*	Rectal swab	30/01/2019	imp	imp	1.36E+04
CPO041	*Enterobacter cloacae*	Rectal swab	27/07/2019	imp	imp	2.11E+02
CPO042	*Klebsiella pneumoniae*	Rectal swab	24/07/2019	imp	imp	9.11E+03
CPO043	*Klebsiella pneumoniae*	Rectal swab	29/07/2019	imp	imp	3.69E+03
CPO044	*Enterobacter cloacae*	Rectal swab	26/08/2019	imp	imp	6.58E+03
CPO045	*Enterobacter cloacae*	Rectal swab	18/08/2019	imp	imp	1.67E+04
CPO046	*Klebsiella pneumoniae*	Rectal swab	06/05/2019	imp	imp	2.59E+04
CPO047	*Enterobacter cloacae*	Rectal swab	09/05/2019	imp	imp	1.46E+02
CPO048	*Enterobacter cloacae*	Rectal swab	11/05/2019	imp	imp	1.75E+04
CPO049	*Enterobacter spp*	Rectal swab	13/06/2019	imp	imp	3.27E+04
CPO050	*Escherichia coli*	Rectal swab	20/06/2019	imp	imp	9.15E+03
CPO051	*Enterobacter cloacae*	Rectal swab	09/09/2017	imp	imp	2.32E+04

Table C.2: Bacterial isolates from clinical samples (part 2)

Sample ID	Specimen	Source	Collection	CPE	AMCA	Conc. cp/uL
CPO052	*Klebsiella pneumoniae*	Rectal swab	03/10/2017	imp	imp	4.27E+01
CPO053	*Enterobacter cloacae*	Rectal swab	05/10/2017	imp	imp	2.18E+04
CPO054	*Enterobacter spp*	Rectal swab	11/10/2017	imp	imp	1.60E+04
CPO055	*Citrobacter freundii*	Rectal swab	28/10/2017	imp	imp	5.55E+02
CPO056	*Klebsiella pneumoniae*	Rectal swab	22/01/2018	imp	imp	1.20E+03
CPO057	*Klebsiella pneumoniae*	Rectal swab	28/01/2018	imp	imp	9.89E+02
CPO058	*Enterobacter cloacae*	Urine	06/02/2018	imp	imp	7.53E+01
CPO059	*Enterobacter spp*	Rectal swab	21/02/2018	imp	imp	7.92E+02
CPO060	*Klebsiella pneumoniae*	Rectal swab	10/09/2018	imp	imp	1.89E+02
CPO061	*Enterobacter cloacae*	Rectal swab	07/09/2018	imp	imp	1.83E+02
CPO062	*Escherichia hermannii*	Rectal swab	20/11/2018	imp	imp	1.47E+02
CPO063	*Escherichia coli*	Rectal swab	21/02/2018	imp	imp	1.05E+03
CPO064	*Enterobacter cloacae*	Rectal swab	28/01/2018	imp	imp	1.20E+02
CPO065	*Klebsiella pneumoniae*	Rectal swab	02/04/2018	imp	imp	3.48E+04
CPO066	*Klebsiella pneumoniae*	Rectal swab	22/04/2018	imp	imp	3.45E+04
CPO067	*Enterobacter cloacae*	Rectal swab	18/08/2018	imp	imp	1.05E+02
CPO068	*Klebsiella pneumoniae*	Rectal swab	07/01/2019	imp	imp	1.72E+02
CPO069	*Pseudomonas aeruginosa*	Rectal swab	12/01/2019	imp	imp	4.95E+03
CPO070	*Enterobacter cloacae*	Rectal swab	11/02/2019	imp	imp	2.41E+02
CPO071	*Klebsiella pneumoniae*	Rectal swab	13/03/2019	imp	imp	4.20E+02
CPO072	*Escherichia hermannii*	Rectal swab	27/05/2019	imp	imp	3.01E+02
CPO073	*Enterobacter spp*	Rectal swab	08/05/2019	imp	imp	1.32E+02
CPO074	*Klebsiella pneumoniae*	Swab	11/05/2019	imp	imp	3.45E+01
CPO075	*Enterobacter cloacae*	Rectal swab	16/06/2016	imp	imp	5.43E+01
CPO076	*Enterobacter cloacae*	Rectal swab	16/05/2019	imp	imp	2.41E+04
CPO077	*Enterobacter cloacae*	Rectal swab	12/05/2018	imp	imp	1.51E+04
CPO078	*Klebsiella pneumoniae*	Wound swab	08/10/2012	kpc	kpc	5.83E+03
CPO079	*Citrobacter spp*	Rectal Swab	01/10/2017	kpc	kpc	9.73E+03
CPO080	*Klebsiella pneumoniae*	Urine	22/03/2014	kpc	kpc	1.22E+04
CPO081	*Klebsiella pneumoniae*	Rectal Swab	10/09/2017	kpc	kpc	3.48E+03
CPO082	*Citrobacter spp*	Rectal Swab	15/10/2017	kpc	kpc	3.45E+03
CPO083	*Klebsiella pneumoniae*	Urine	08/04/2015	kpc	kpc	7.61E+03
CPO084	*Serratia marcescens*	Rectal Swab	10/10/2017	kpc	kpc	8.66E+03
CPO085	*Klebsiella pneumoniae*	Rectal Swab	11/09/2017	kpc	kpc	3.98E+03
CPO086	*Klebsiella pneumoniae*	Rectal Swab	11/09/2017	kpc	kpc	1.82E+04
CPO087	*Escherichia coli*	Rectal Swab	01/08/2016	ndm	ndm	4.23E+03
CPO088	*Klebsiella pneumoniae*	Urine	25/12/2015	ndm	ndm	5.39E+03
CPO089	*Klebsiella pneumoniae*	Rectal Swab	18/12/2015	ndm	ndm	6.47E+03
CPO090	*Escherichia coli*	Bone (Tibia)	14/01/2015	ndm	ndm	3.69E+03
CPO091	*Klebsiella pneumoniae*	Throat swab	02/08/2016	ndm	ndm	3.66E+03
CPO092	*Klebsiella pneumoniae*	Rectal Swab	18/04/2015	ndm	ndm	1.19E+04
CPO093	*Klebsiella pneumoniae*	Urine	23/04/2015	ndm	ndm	4.26E+03
CPO094	*Klebsiella pneumoniae*	Rectal Swab	25/04/2015	ndm	ndm	2.22E+03
CPO095	*Proteus mirabilis*	Urine	07/02/2014	ndm	ndm	1.40E+03
CPO096	*Klebsiella pneumoniae*	Rectal Swab	04/12/2016	ndm	ndm	2.63E+03
CPO097	*Klebsiella pneumoniae*	Mouth Swab	29/01/2015	ndm	ndm	2.48E+03
CPO098	*Klebsiella pneumoniae*	Rectal Swab	20/04/2015	ndm	ndm	2.39E+03
CPO099	*Klebsiella pneumoniae*	Perinrum swab	10/05/2015	ndm	ndm	1.00E+04
CPO100	*Escherichia coli*	Vaginal swab	05/03/2015	ndm	ndm	1.40E+04
CPO101	*Escherichia coli*	Perinrum swab	09/04/2015	ndm	ndm	9.82E+03

Table C.3: Bacterial isolates from clinical samples (part 3)

Sample ID	Specimen	Source	Collection	CPE	AMCA	Conc. cp/uL
CPO102	*Klebsiella pneumoniae*	Perinrum swab	09/12/2014	ndm	ndm	1.56E+04
CPO103	*Klebsiella pneumoniae*	Rectal Swab	20/04/2015	ndm	ndm	1.66E+04
CPO104	*Klebsiella pneumoniae*	Rectal Swab	15/04/2015	ndm	ndm	1.06E+04
CPO105	*Escherichia coli*	Faeces	23/11/2015	ndm	ndm	9.28E+03
CPO106	*Escherichia coli*	Rectal Swab	26/12/2015	ndm	ndm	7.51E+03
CPO107	*Klebsiella pneumoniae*	Perinrum swab	07/05/2015	ndm	ndm	6.29E+03
CPO108	*Klebsiella pneumoniae*	Rectal Swab	20/04/2015	ndm	ndm	1.85E+04
CPO109	*Klebsiella pneumoniae*	Sputum	23/03/2018	ndm	ndm	9.74E+03
CPO110	*Klebsiella pneumoniae*	Rectal Swab	29/04/2015	ndm	ndm	1.27E+04
CPO111	*Klebsiella pneumoniae*	Perinrum swab	07/05/2015	ndm	ndm	1.05E+04
CPO112	*Klebsiella pneumoniae*	Catheter Urine	02/07/2014	ndm	ndm	1.21E+04
CPO113	*Klebsiella pneumoniae*	Perinrum swab	09/12/2014	ndm	ndm	1.50E+04
CPO114	*Klebsiella pneumoniae*	Urine	08/03/2015	ndm	ndm	8.01E+03
CPO115	*Klebsiella pneumoniae*	Urine	10/04/2015	ndm	ndm	2.06E+04
CPO116	*Klebsiella pneumoniae*	Perinrum swab	19/04/2015	ndm	ndm	1.46E+04
CPO117	*Klebsiella pneumoniae*	Rectal Swab	18/04/2015	ndm	ndm	2.83E+04
CPO118	*Klebsiella pneumoniae*	Rectal Swab	17/05/2015	ndm	ndm	9.15E+03
CPO119	*Klebsiella pneumoniae*	Urine	23/04/2015	ndm	ndm	2.22E+04
CPO120	*Klebsiella pneumoniae*	Rectal Swab	25/04/2015	ndm	ndm	1.41E+04
CPO121	*Klebsiella pneumoniae*	Perinrum swab	10/05/2015	ndm	ndm	8.60E+03
CPO122	*Klebsiella pneumoniae*	Rectal Swab	03/10/2015	ndm	ndm	2.08E+04
CPO123	*Klebsiella pneumoniae*	Rectal Swab	16/08/2015	ndm	ndm	1.96E+04
CPO124	*Klebsiella pneumoniae*	Wound swab	19/04/2015	ndm	ndm	6.44E+03
CPO125	*Klebsiella pneumoniae*	Rectal Swab	16/09/2015	ndm	ndm	1.68E+04
CPO126	*Klebsiella pneumoniae*	Rectal Swab	07/06/2015	ndm	ndm	1.64E+04
CPO127	*Klebsiella pneumoniae*	Rectal Swab	07/07/2015	ndm	ndm	2.29E+04
CPO128	*Klebsiella pneumoniae*	Rectal Swab	03/06/2015	ndm	ndm	6.54E+03
CPO129	*Klebsiella pneumoniae*	Rectal Swab	20/05/2015	ndm	ndm	1.18E+04
CPO130	*Klebsiella pneumoniae*	Rectal Swab	10/08/2015	ndm	ndm	7.64E+03
CPO131	*Klebsiella pneumoniae*	Perinrum swab	12/07/2015	ndm	ndm	6.86E+03
CPO132	*Klebsiella pneumoniae*	Rectal Swab	21/07/2015	ndm	ndm	2.40E+04
CPO133	*Escherichia coli*	Rectal swab	29/06/2018	ndm	ndm	1.02E+04
CPO134	*Klebsiella pneumoniae*	Blood culture	23/09/2016	ndm	ndm	3.14E+04
CPO135	*Enterobacter cloacae*	Rectal swab	24/06/2018	ndm	ndm_kpc	5.75E+04
CPO136	*Escherichia coli*	Urine	12/03/2019	ndm	ndm	1.83E+04
CPO137	*Escherichia coli*	Rectal swab	28/01/2019	ndm	ndm	1.29E+04
CPO138	*Klebsiella pneumoniae*	Rectal swab	23/02/2019	ndm	ndm	2.41E+04
CPO139	*Klebsiella pneumoniae*	Rectal swab	25/03/2019	ndm	ndm	5.85E+04
CPO140	*Enterobacter cloacae*	Rectal swab	03/04/2019	ndm	ndm	1.24E+04
CPO141	*Enterobacter cloacae*	Rectal swab	21/01/2019	ndm	ndm	8.38E+03
CPO142	*Citrobacter freundii*	Abdomen	08/04/2019	ndm	ndm	1.34E+04
CPO143	*Klebsiella pneumoniae*	Urine	23/04/2015	ndm	ndm	1.11E+04
CPO144	*Klebsiella pneumoniae*	Leg tissue	29/07/2015	ndm	ndm	7.75E+03
CPO145	*Klebsiella pneumoniae*	Abdomen	22/09/2016	ndm	ndm	1.55E+04
CPO146	*Escherichia coli*	Urine	04/05/2017	ndm	ndm	1.73E+04
CPO147	*Escherichia coli*	Catheter urine	06/06/2019	ndm	ndm	6.45E+03
CPO148	*Escherichia coli*	Rectal swab	12/01/2019	ndm	ndm	1.13E+05
CPO149	*Escherichia coli*	Rectal swab	27/05/2019	ndm	ndm	2.44E+05
CPO150	*Enterobacter cloacae*	Rectal swab	16/06/2016	ndm	ndm	2.72E+06
CPO151	*Klebsiella pneumoniae*	Rectal swab	16/05/2019	ndm	ndm	7.59E+04

Table C.4: Bacterial isolates from clinical samples (part 4)

Sample ID	Specimen	Source	Collection	CPE	AMCA	Conc. cp/uL
CPO152	*Enterobacter cloacae*	Rectal swab	12/05/2018	ndm	ndm	2.48E+06
CPO153	*Enterobacter cloacae*	Rectal swab	22/04/2018	ndm	ndm	1.80E+06
CPO154	*Escherichia coli*	Rectal swab	18/08/2018	ndm	ndm	3.59E+05
CPO155	*Klebsiella pneumoniae*	Rectal swab	08/05/2019	ndm	ndm	1.87E+06
CPO156	*Enterobacter cloacae*	Rectal swab	12/05/2018	ndm	ndm	1.11E+05
CPO157	*Klebsiella pneumoniae*	Rectal swab	07/01/2019	ndm	ndm	1.54E+06
CPO158	*Klebsiella pneumoniae*	Rectal swab	12/01/2019	ndm	ndm	2.27E+06
CPO159	*Klebsiella pneumoniae*	Rectal swab	13/03/2019	ndm	ndm	3.15E+04
CPO160	*Klebsiella pneumoniae*	Rectal swab	27/05/2019	ndm	ndm	9.98E+04
CPO161	*Escherichia coli*	Rectal Swab	01/12/2014	ndm_oxa48	ndm_oxa48	1.70E+04
CPO162	*Escherichia coli*	Wound swab	14/03/2014	oxa48	oxa48	2.45E+04
CPO163	*Escherichia coli*	Rectal Swab	20/10/2017	oxa48	oxa48	3.76E+03
CPO164	*Citrobacter freundii*	Rectal Swab	06/01/2016	oxa48	oxa48	5.56E+03
CPO165	*Escherichia coli*	Urine	04/04/2015	oxa48	oxa48	2.48E+03
CPO166	*Klebsiella pneumoniae*	Rectal Swab	07/01/2016	oxa48	oxa48	1.50E+04
CPO167	*Escherichia coli*	Wound swab	25/11/2012	oxa48	oxa48	2.22E+03
CPO168	*Escherichia coli*	Blood culture	21/07/2013	oxa48	oxa48	1.17E+04
CPO169	*Serratia marcescens*	Bone (Tibia)	28/01/2015	oxa48	oxa48	6.25E+03
CPO170	*Citrobacter freundii*	Rectal Swab	13/12/2015	oxa48	oxa48	3.23E+03
CPO171	*Klebsiella pneumoniae*	Urine	02/07/2014	oxa48	oxa48	1.46E+03
CPO172	*Klebsiella pneumoniae*	Abdomen Fluid	22/01/2015	oxa48	oxa48	4.55E+03
CPO173	*Escherichia coli*	Urine	04/04/2015	oxa48	oxa48	2.25E+04
CPO174	*Escherichia coli*	Rectal swab	28/06/2018	oxa48	oxa48	1.16E+04
CPO175	*Klebsiella pneumoniae*	Urine	08/05/2018	oxa48	oxa48	3.36E+04
CPO176	*Klebsiella pneumoniae*	Blood culture	13/06/2018	oxa48	oxa48	2.77E+04
CPO177	*Klebsiella pneumoniae*	Blood culture	15/10/2018	oxa48	oxa48	3.40E+04
CPO178	*Klebsiella pneumoniae*	Blood culture	21/02/2015	oxa48	oxa48	8.70E+03
CPO179	*Citrobacter freundii*	Rectal swab	14/01/2018	oxa48	oxa48	4.21E+04
CPO180	*Citrobacter freundii*	Urine	16/01/2018	oxa48	oxa48	3.12E+04
CPO181	*Klebsiella pneumoniae*	Rectal swab	04/03/2019	oxa48	oxa48	6.24E+04
CPO182	*Klebsiella pneumoniae*	Urine	30/08/2018	oxa48	oxa48	2.14E+04
CPO183	*Klebsiella pneumoniae*	Urine	23/04/2019	oxa48	oxa48	8.64E+04
CPO184	*Klebsiella pneumoniae*	Urine	01/06/2015	oxa48	oxa48	2.82E+05
CPO185	*Escherichia coli*	Wound swab	08/07/2015	oxa48	oxa48	1.58E+04
CPO186	*Klebsiella pneumoniae*	Urine	04/01/2016	oxa48	oxa48	9.71E+04
CPO187	*Citrobacter amalonaticus*	Rectal swab	18/01/2016	oxa48	oxa48	3.38E+04
CPO188	*Klebsiella pneumoniae*	Urine	08/02/2016	oxa48	oxa48	8.94E+05
CPO189	*Klebsiella pneumoniae*	Urine	08/02/2016	oxa48	oxa48	4.47E+05
CPO190	*Klebsiella pneumoniae*	Wound swab	01/07/2016	oxa48	oxa48	5.80E+05
CPO191	*Klebsiella pneumoniae*	Pleural fluid	10/07/2016	oxa48	oxa48	7.23E+05
CPO192	*Escherichia coli*	Rectal swab	15/08/2016	oxa48	oxa48	4.37E+04
CPO193	*Escherichia coli*	Urine	26/08/2016	oxa48	oxa48	2.14E+04
CPO194	*Klebsiella pneumoniae*	Urine	22/10/2016	oxa48	oxa48	1.35E+05
CPO195	*Escherichia coli*	Urine	27/11/2016	oxa48	oxa48	1.11E+04
CPO196	*Klebsiella pneumoniae*	Rectal swab	18/03/2017	oxa48	oxa48	5.28E+05
CPO197	*Escherichia coli*	Wound swab	18/04/2017	oxa48	oxa48	9.71E+03
CPO198	*Citrobacter freundii*	Urine	05/05/2017	oxa48	oxa48	5.71E+03
CPO199	*Klebsiella pneumoniae*	Wound swab	06/08/2017	oxa48	oxa48	1.29E+04
CPO200	*Klebsiella pneumoniae*	Abdomen	17/01/2018	oxa48	oxa48	1.73E+03
CPO201	*Klebsiella pneumoniae*	Rectal swab	22/01/2018	oxa48	oxa48	1.78E+03

Table C.5: Bacterial isolates from clinical samples (part 5)

Sample ID	Specimen	Source	Collection	CPE	AMCA	Conc. cp/uL
CPO202	*Klebsiella pneumoniae*	Urine	27/01/2018	oxa48	oxa48	1.26E+05
CPO203	*Escherichia coli*	Rectal swab	22/04/2018	oxa48	oxa48	1.02E+06
CPO204	*Citrobacter freundii*	Rectal swab	18/08/2018	oxa48	oxa48	5.77E+05
CPO205	*Escherichia coli*	Rectal swab	07/01/2019	oxa48	oxa48	2.69E+05
CPO206	*Escherichia coli*	Rectal swab	12/01/2019	oxa48	oxa48	1.78E+04
CPO207	*Citrobacter freundii*	Rectal swab	11/02/2019	oxa48	oxa48	2.92E+06
CPO208	*Escherichia coli*	Rectal swab	13/03/2019	oxa48	oxa48	1.57E+04
CPO209	*Klebsiella pneumoniae*	Rectal swab	27/05/2019	oxa48	oxa48	2.45E+06
CPO210	*Klebsiella pneumoniae*	Rectal swab	08/05/2019	oxa48	oxa48	2.03E+06
CPO211	*Klebsiella pneumoniae*	Rectal swab	11/05/2019	oxa48	oxa48	2.41E+05
CPO212	*Klebsiella pneumoniae*	Rectal swab	16/06/2016	oxa48	oxa48	3.61E+05
CPO213	*Escherichia coli*	Rectal swab	16/05/2019	oxa48	oxa48	3.29E+04
CPO214	*Klebsiella pneumoniae*	Rectal swab	12/05/2018	oxa48	oxa48	1.21E+06
CPO215	*Klebsiella pneumoniae*	Rectal swab	22/04/2018	oxa48	oxa48	1.57E+06
CPO216	*Klebsiella pneumoniae*	Rectal swab	18/08/2018	oxa48	oxa48	4.18E+05
CPO217	*Klebsiella pneumoniae*	Rectal swab	07/01/2019	oxa48	oxa48	2.43E+06
CPO218	*Escherichia coli*	Rectal swab	12/01/2019	oxa48	oxa48	1.80E+05
CPO219	*Escherichia coli*	Rectal swab	11/02/2019	oxa48	oxa48	2.78E+03
CPO220	*Klebsiella pneumoniae*	Rectal swab	13/03/2019	oxa48	oxa48	3.75E+05
CPO221	*Escherichia coli*	Rectal swab	27/05/2019	oxa48	oxa48	5.40E+06
CPO222	*Klebsiella pneumoniae*	Rectal swab	08/05/2019	oxa48	oxa48	9.42E+03
CPO223	*Klebsiella pneumoniae*	Rectal swab	11/05/2019	oxa48	oxa48	3.71E+05
CPO224	*Citrobacter freundii*	Rectal swab	16/06/2016	oxa48	oxa48	7.02E+05
CPO225	*Escherichia coli*	Rectal swab	12/05/2018	oxa48	oxa48	8.26E+05
CPO226	*Klebsiella pneumoniae*	Rectal swab	22/04/2018	oxa48	oxa48	7.85E+05
CPO227	*Escherichia coli*	Rectal swab	18/08/2018	oxa48	oxa48	4.46E+05
CPO228	*Citrobacter freundii*	Rectal swab	07/01/2019	oxa48	oxa48	5.04E+05
CPO229	*Klebsiella pneumoniae*	Rectal swab	11/02/2019	oxa48	oxa48	2.57E+05
CPO230	*Klebsiella pneumoniae*	Rectal swab	13/03/2019	oxa48	oxa48	4.65E+05
CPO231	*Klebsiella pneumoniae*	Rectal swab	08/05/2019	oxa48	oxa48	4.16E+05
CPO232	*Klebsiella pneumoniae*	Rectal swab	11/05/2019	oxa48	oxa48	1.03E+05
CPO233	*Klebsiella pneumoniae*	Rectal swab	07/01/2019	oxa48	oxa48	2.01E+06
CPO234	*Klebsiella pneumoniae*	Rectal swab	12/01/2019	oxa48	oxa48	6.13E+04
CPO235	*Escherichia coli*	Rectal swab	11/02/2019	oxa48	oxa48	2.54E+05
CPO236	*Klebsiella pneumoniae*	Rectal swab	13/03/2019	oxa48	oxa48	2.26E+05
CPO237	*Escherichia coli*	Rectal swab	27/05/2019	oxa48	oxa48	2.22E+05
CPO238	*Escherichia coli*	Rectal swab	11/05/2019	oxa48	oxa48	8.02E+05
CPO239	*Klebsiella pneumoniae*	Rectal swab	16/06/2016	oxa48	oxa48	2.90E+05
CPO240	*Klebsiella pneumoniae*	Rectal swab	16/05/2019	oxa48	oxa48	2.93E+05
CPO241	*Escherichia coli*	Rectal swab	22/04/2018	oxa48	oxa48	5.33E+05
CPO242	*Enterobacter cloacae*	Rectal swab	18/08/2018	oxa48	oxa48	8.86E+04
CPO243	*Klebsiella pneumoniae*	Rectal swab	16/06/2016	oxa48	oxa48	4.46E+06
CPO244	*Klebsiella pneumoniae*	Rectal swab	18/08/2018	oxa48	oxa48	2.11E+06
CPO245	*Enterobacter spp*	Rectal swab	27/05/2019	oxa48	oxa48	4.32E+04
CPO246	*Pseudomonas aeruginosa*	Wound swab	25/03/2015	vim	vim	2.28E+03
CPO247	*Citrobacter freundii*	Rectal Swab	02/04/2016	vim	vim	1.33E+04
CPO248	*Enterobacter cloacae*	Bone (Tibia)	14/01/2015	vim	vim	5.66E+03
CPO249	*Pseudomonas aeruginosa*	Sputum	01/11/2013	vim	vim	5.19E+03
CPO250	*Enterobacter cloacae*	Bone (Tibia)	14/01/2015	vim	vim	2.95E+04
CPO251	*Klebsiella pneumoniae*	Rectal swab	11/02/2019	vim	vim	1.49E+05
CPO252	*Klebsiella pneumoniae*	Rectal swab	08/05/2019	vim	vim	1.90E+05
CPO253	*Klebsiella pneumoniae*	Rectal swab	11/05/2019	vim	vim	1.39E+05

Appendix D

Supporting Information: Chapter 8

This Appendix contains the following:

Figure List

- Correlation of c ADS and MDS for 3plex (Figure D.1)
- Standard curves for all targets in the BEST selected 7plex (Figure D.2)
- Overall development of Smart-Plexer (Figure D.3)
- The 21-plex for RTI detection using three fluorescent channels (Figure D.4)

Table List

- Primer table for 3plex (Table D.1)
- Assay table for 3plex (Table D.2)
- Primer table for 7plex (Table D.3)
- Assay table for 7plex (Table D.4)
- Assay Combination table for 3plex (Table D.5)
- The c parameter stats for 3plex (Table D.6)

- ADS and MDS scores for the three curve representations in 3plex (Table D.7)
- Assay Combination table for 7plex (Table D.8)
- The c parameter stats for 7plex (Table D.9)

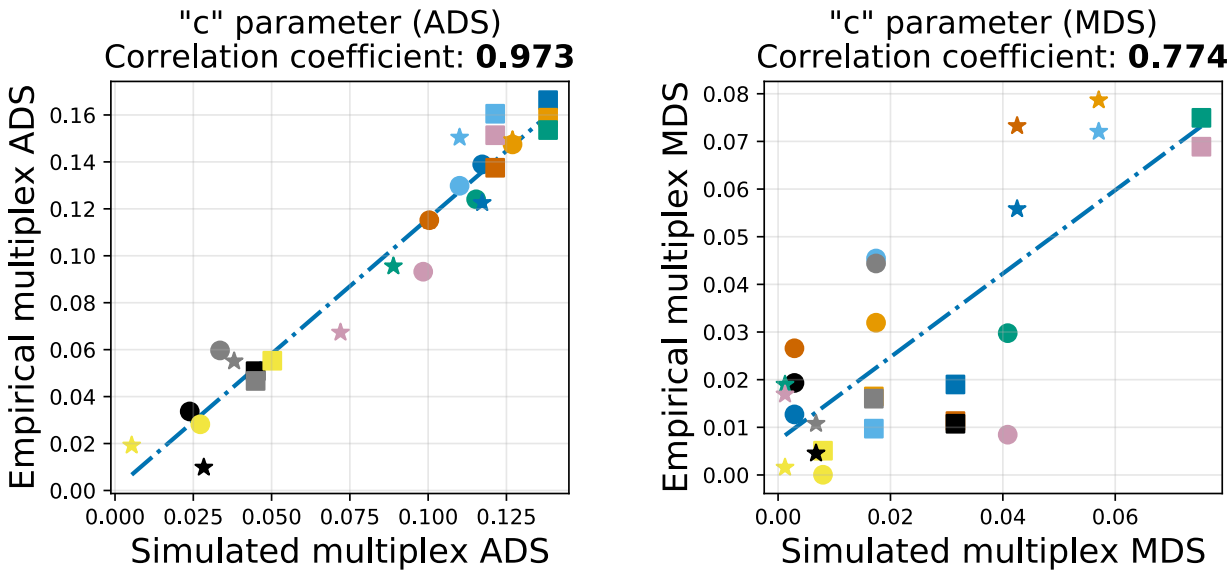


Figure D.1: Correlation of c ADS and MDS for 3plex.

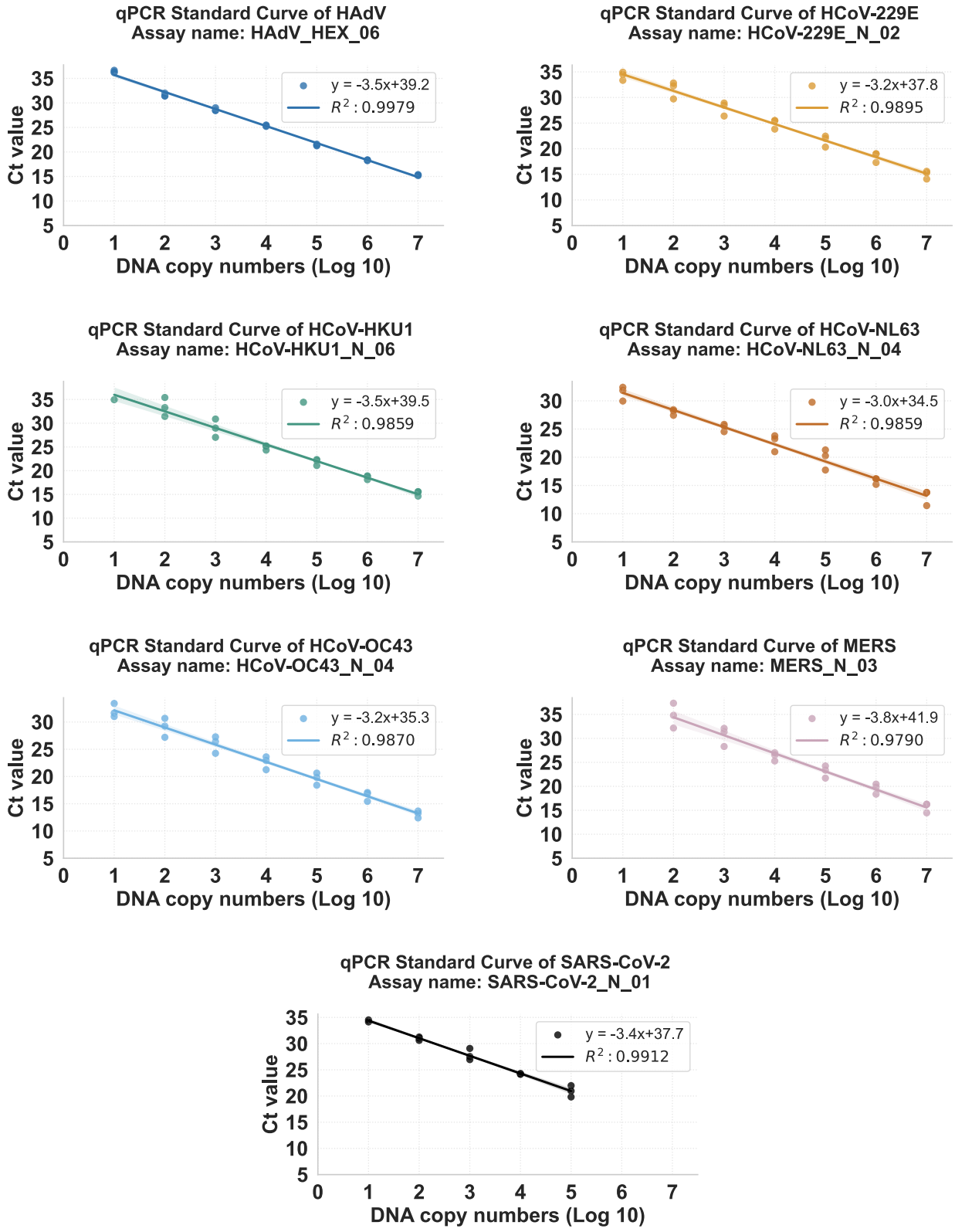


Figure D.2: Standard curves for all targets in the BEST selected 7plex.

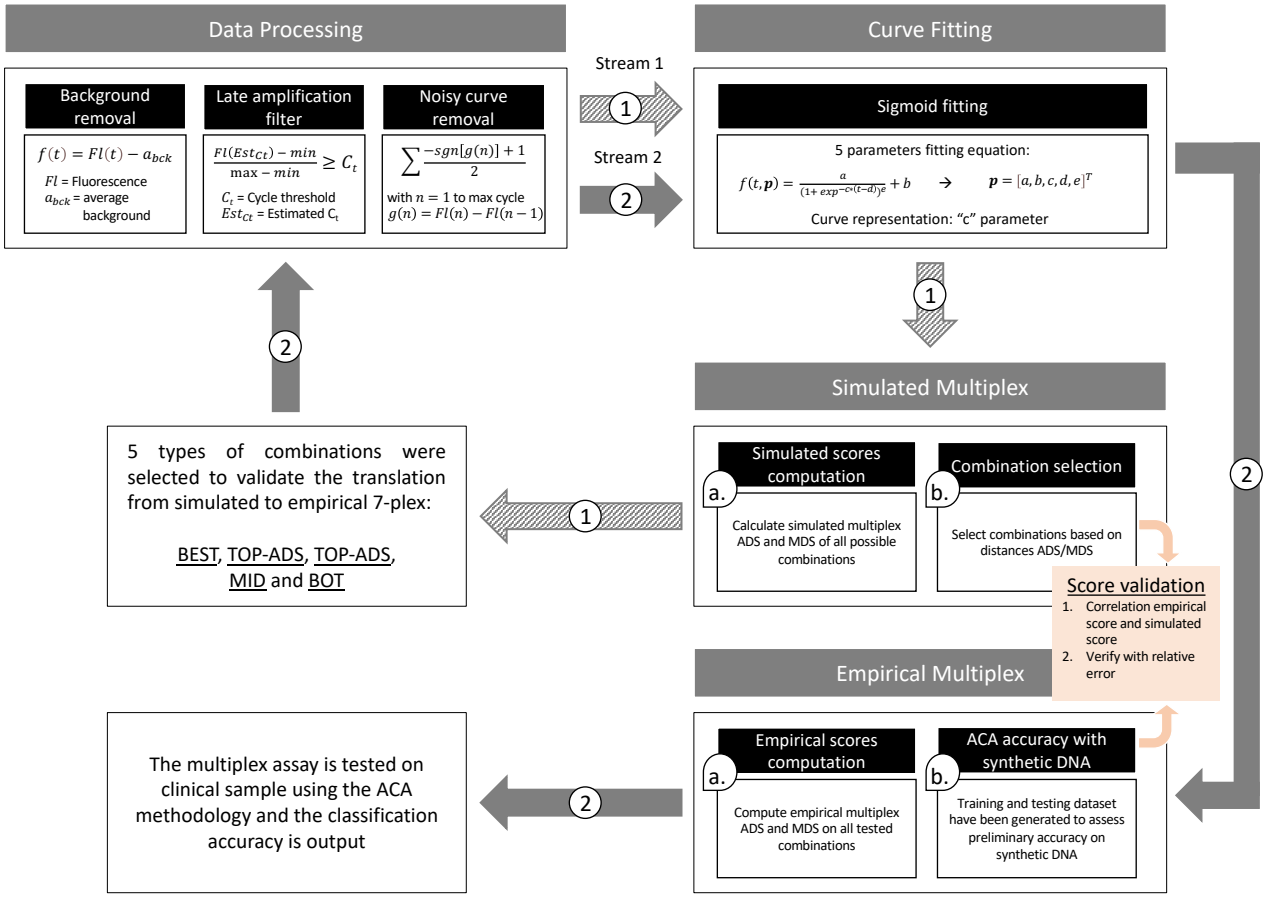


Figure D.3: Overall development of Smart-Plexer. Stream 1. Pipeline for combo selection based on simulated multiplex assays. Before combination selection, operations including 3-fold data manipulation (Background removal, Late amplification filter and Noisy curve removal), data processing (Sigmoid fitting and Curve FFI normalization) and simulated score computation (The types of data are raw curve, normalized curve, fitted parameters and "c" parameter) are conducted. The principle of selection is then based on MDS-ADS ranking system. Combinations from 5 groups (BEST, TOP-ADS, TOP-MDS, MID, BOT) are chosen for validation progress. Stream 2. Pipeline for result validation based on empirical multiplex assays. With empirical experiment, same pre-operations (3-fold data manipulation and data processing) are taken. Then, the empirical scores are computed, and the distributions of classification accuracy are evaluated versus scores. The last validation step is based on clinical samples with best assay combination developed so far.

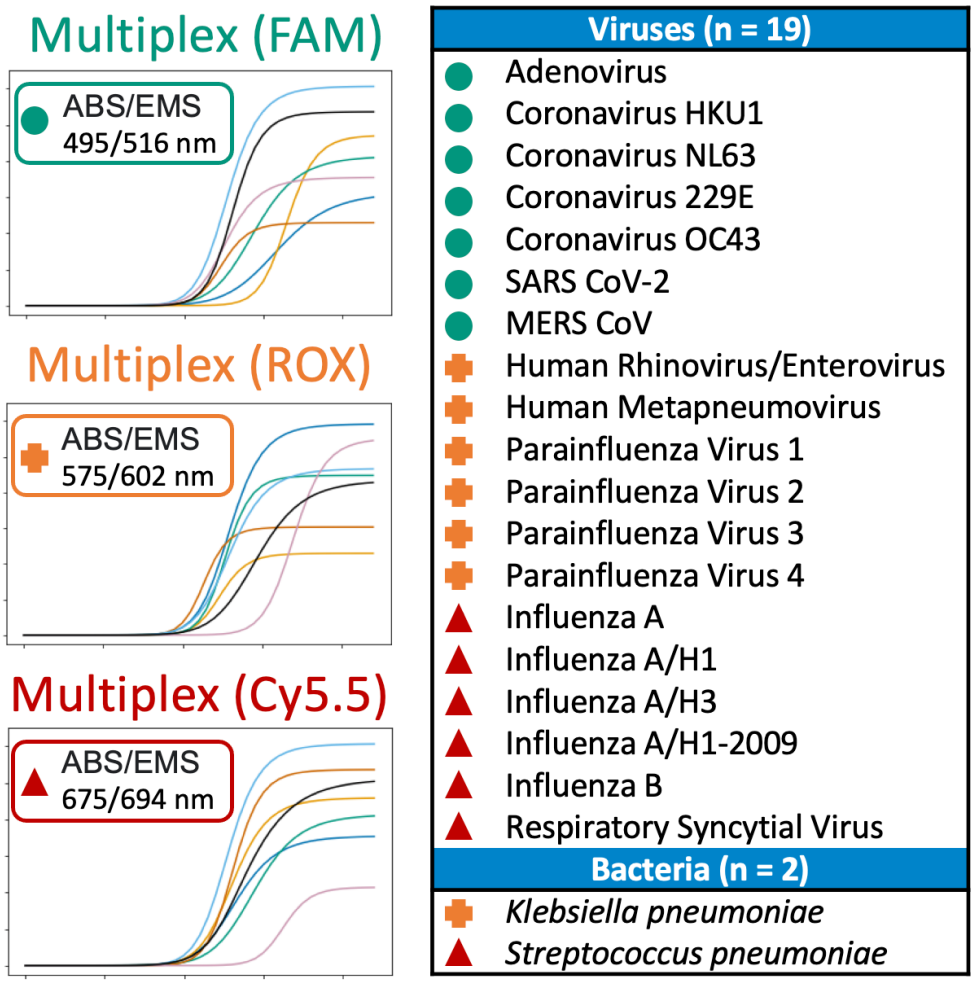


Figure D.4: The 21-plex for RTI detection using three fluorescent channels (data-driven multiplexing). Multiplexed panel of 21 respiratory pathogens, coupling the ACA method and TaqMan probe chemistries, using three different fluorescent channels in qPCR.

Table D.1: Primer table for 3plex

Oligo name	Target	Oligo type	Oligo modification	Oligo sequence
HAdV_01	HEX	forward		CCCTTCGATGATGCCGCA
HAdV_02	HEX	forward		CGCAGTGGTCTTACATGCACATCTC
HAdV_03	HEX	probe	56-FAM / ZEN / 3IABkFQ	CCTCGGAGTACCTRAGCCCCGG
HAdV_04	HEX	probe	56-FAM / ZEN / 3IABkFQ	CCGCGCCACCGAGACGTACTTCAG
HAdV_05	HEX	reverse		CAGGCTGAAGTACGTCTCGGT
HAdV_07	HEX	reverse		CGCAGCGTCAAACGCTG
HCoV-HKU1_02	N	forward		TCAAGAAGCTATCCCTACTAGGT
HCoV-HKU1_03	N	probe	56-FAM / ZEN / 3IABkFQ	CGCCTGGTACGATTTTGCCTCAAGGCT
HCoV-HKU1_05	N	reverse		AGACCTTCCTGAGCCTTCAACA
HCoV-HKU1_06	N	reverse		CTATTAGAAGCAGACCTTCCTGA
HCoV-HKU1_08	N	reverse		GCGATCTCATCAGCCATATCAGGT
MERS-CoV_01	N	forward		ACGCGGAACCCTAACAATGATT
MERS-CoV_02	N	probe	56-FAM / ZEN / 3IABkFQ	TGCCTCCAGTCCCCTCAATGTGGA
MERS-CoV_03	N	reverse		GCTAGAGGCTCTTGAAGATGATTGA
MERS-CoV_04	N	forward		CCACAAGCGCACTTCCACCAA
MERS-CoV_05	N	probe	56-FAM / ZEN / 3IABkFQ	TTCCCTGGAGGTCTCCTGGTCCGC
MERS-CoV_06	N	reverse		GTGGGTCCTCAGTGCCGAGT

Table D.2: Assay table for 3plex

Assay_ID	Forward_ID	Probe_ID	Reverse_ID
HAdV_HEX_03	HAdV_01	HAdV_04	HAdV_05
HAdV_HEX_09	HAdV_01	HAdV_03	HAdV_07
HAdV_HEX_12	HAdV_02	HAdV_04	HAdV_07
HCoV-HKU1_N_02	HCoV-HKU1_02	HCoV-HKU1_03	HCoV-HKU1_05
HCoV-HKU1_N_04	HCoV-HKU1_02	HCoV-HKU1_03	HCoV-HKU1_06
HCoV-HKU1_N_06	HCoV-HKU1_02	HCoV-HKU1_03	HCoV-HKU1_08
MERS-CoV_N_01	MERS-CoV_01	MERS-CoV_02	MERS-CoV_03
MERS-CoV_N_03	MERS-CoV_01	MERS-CoV_05	MERS-CoV_06
MERS-CoV_N_04	MERS-CoV_04	MERS-CoV_05	MERS-CoV_06

Table D.3: Primer table for 7plex

Oligo name	Target	Oligo type	Oligo modification	Oligo sequence
HAdV_01	HEX	forward		CCCTTCGATGATGCCGCA
HAdV_02	HEX	forward		CGCAGTGGTCTTACATGCACATCTC
HAdV_03	HEX	probe	56-FAM / ZEN / 3IABkFQ	CCTCGGAGTACCTRAGCCCCGG
HAdV_04	HEX	probe	56-FAM / ZEN / 3IABkFQ	CCGCGCCACCGAGACGTACTTCAG
HAdV_06	HEX	reverse		GCCACCGTGGGGTTTCTAAACTTG
HAdV_07	HEX	reverse		CGCAGCGTCAAACGCTG
HCoV-229E_01	N	forward		CAGTCAAATGGGCTGATGCA
HCoV-229E_02	N	probe	56-FAM / ZEN / 3IABkFQ	ACCCTGACGACCACGTTGTGGTTCA
HCoV-229E_03	N	reverse		TTGTTCACTATCAACAAGCAAAGG
HCoV-229E_04	N	forward		GAAATGCAAAAGCCACGGTGGAA
HCoV-229E_05	N	probe	56-FAM / ZEN / 3IABkFQ	AGTTGTGGTCAAGGTCTCTGGGGCC
HCoV-229E_06	N	reverse		AGCTCAGCAAATTGTGGATAGCC
HCoV-HKU1_02	N	forward		TCAAGAAGCTATCCCTACTAGGT
HCoV-HKU1_03	N	probe	56-FAM / ZEN / 3IABkFQ	CGCCTGGTACGATTTTGCCTCAAGGCT
HCoV-HKU1_05	N	reverse		AGACCTTCCTGAGCCTTCAACA
HCoV-HKU1_06	N	reverse		CTATTAGAAGCAGACCTTCCTGA
HCoV-HKU1_07	N	probe	56-FAM / ZEN / 3IABkFQ	ACG[+T]TCTC[+A]ATCA[+C]GTGG[+A]CCC
HCoV-HKU1_08	N	reverse		GCGATCTCATCAGCCATATCAGGT
HCoV-NL63_01	N	forward		TGGTTAGTTCTGATAAGGCACC
HCoV-NL63_02	N	probe	56-FAM / ZEN / 3IABkFQ	TGGAATGTTCAAGAGCGTTGGCGTATGCG
HCoV-NL63_03	N	reverse		GGAGGCAAATCAACACGTTG
HCoV-NL63_04	N	forward		GGTGCTAAAACTGTTAATACCAGT
HCoV-NL63_05	N	probe	56-FAM / ZEN / 3IABkFQ	AGGTTTCTGATTACGTTTGCGATTACCA
HCoV-NL63_06	N	reverse		GCAATAGAGAACTTTGGTTCCA
HCoV-OC43_01	N	forward		CTTGGTTCTCTGGAATTACTCA
HCoV-OC43_02	N	probe	56-FAM / ZEN / 3IABkFQ	AGAAGGACAAGGTGTGCCTATTGCACCA
HCoV-OC43_03	N	reverse		GTTCCCAGATAGTAAAAATACCAT
HCoV-OC43_04	N	forward		GGTGGAGAAATGTTAAAACTTGGAACT
HCoV-OC43_06	N	probe	56-FAM / ZEN / 3IABkFQ	TCCCCATTCTTGCAGAACTCGCACCCA
HCoV-OC43_07	N	reverse		CCAAAGAAAAACGCACCAGCTG
SARS-CoV-2_01	N	forward		ATAATGGACCCCAAAATCAGCGA
SARS-CoV-2_02	N	probe	56-FAM / ZEN / 3IABkFQ	CACCCCGCATTACGTTTGGTGGACC
SARS-CoV-2_03	N	reverse		TCTGGTTACTGCCAGTTGAATCTG
SARS-CoV-2_04	N	forward		CTGATTACAAACATTGGCCGCA
SARS-CoV-2_05	N	probe	56-FAM / ZEN / 3IABkFQ	TGCACAATTTGCCCCCAGCGCTTCAG
SARS-CoV-2_06	N	reverse		ATGCGCGACATTCCGAAGAA
SARS-CoV-2_12	N	forward		GACCCCAAAATCAGCGAAAT
SARS-CoV-2_13	N	probe	56-FAM / TAMRA	ACCCCGCATTACGTTTGGTGGACC
SARS-CoV-2_14	N	reverse		TCTGGTTACTGCCAGTTGAATCTG
MERS-CoV_01	N	forward		ACGCGGAACCCTAACAATGATT
MERS-CoV_02	N	probe	56-FAM / ZEN / 3IABkFQ	TGCCTCCAGTCCCCTCAATGTGGA
MERS-CoV_03	N	reverse		GCTAGAGGCTCTTGAAGATGATTGA
MERS-CoV_04	N	forward		CCACAAGCGCACTTCCACCAA
MERS-CoV_05	N	probe	56-FAM / ZEN / 3IABkFQ	TTCCCTGGAGGTCTCCTGGTCCGC
MERS-CoV_06	N	reverse		GTGGGTCCTCAGTGCCGAGT

Table D.4: Assay table for 7plex

Assay_ID	**Forward_ID**	**Probe_ID**	**Reverse_ID**
HAdV_HEX_06	HAdV_02	HAdV_03	HAdV_06
HAdV_HEX_09	HAdV_01	HAdV_03	HAdV_07
HAdV_HEX_12	HAdV_02	HAdV_04	HAdV_07
HCoV-229E_N_01	HCoV-229E_01	HCoV-229E_02	HCoV-229E_03
HCoV-229E_N_02	HCoV-229E_04	HCoV-229E_05	HCoV-229E_06
HCoV-HKU1_N_02	HCoV-HKU1_02	HCoV-HKU1_03	HCoV-HKU1_05
HCoV-HKU1_N_04	HCoV-HKU1_02	HCoV-HKU1_03	HCoV-HKU1_06
HCoV-HKU1_N_06	HCoV-HKU1_02	HCoV-HKU1_03	HCoV-HKU1_08
HCoV-HKU1_N_08	HCoV-HKU1_02	HCoV-HKU1_07	HCoV-HKU1_08
HCoV-NL63_N_01	HCoV-NL63_01	HCoV-NL63_02	HCoV-NL63_03
HCoV-NL63_N_02	HCoV-NL63_01	HCoV-NL63_02	HCoV-NL63_06
HCoV-NL63_N_04	HCoV-NL63_04	HCoV-NL63_05	HCoV-NL63_06
HCoV-OC43_N_01	HCoV-OC43_01	HCoV-OC43_02	HCoV-OC43_03
HCoV-OC43_N_02	HCoV-OC43_01	HCoV-OC43_02	HCoV-OC43_07
HCoV-OC43_N_04	HCoV-OC43_04	HCoV-OC43_06	HCoV-OC43_07
SARS-CoV-2_N_01	SARS-CoV-2_01	SARS-CoV-2_02	SARS-CoV-2_03
SARS-CoV-2_N_02	SARS-CoV-2_04	SARS-CoV-2_05	SARS-CoV-2_06
SARS-CoV-2_N_03	SARS-CoV-2_12	SARS-CoV-2_13	SARS-CoV-2_14
MERS-CoV_N_01	MERS-CoVS-CoV_01	MERS-CoVS-CoV_02	MERS-CoVS-CoV_03
MERS-CoV_N_02	MERS-CoVS-CoV_01	MERS-CoVS-CoV_02	MERS-CoVS-CoV_06
MERS-CoV_N_03	MERS-CoVS-CoV_01	MERS-CoVS-CoV_05	MERS-CoVS-CoV_06
MERS-CoV_N_04	MERS-CoVS-CoV_04	MERS-CoVS-CoV_05	MERS-CoVS-CoV_06

Table D.5: Assay Combination table for 3plex

Multiplex assay name	**HAdV singleplex**	**HCoV-HKU1 singleplex**	**MERS-CoV singleplex**
PM3.01	HAdV_HEX_03	HCoV-HKU1_N_02	MERS-CoV_N_01
PM3.02	HAdV_HEX_03	HCoV-HKU1_N_02	MERS-CoV_N_03
PM3.03	HAdV_HEX_03	HCoV-HKU1_N_02	MERS-CoV_N_04
PM3.04	HAdV_HEX_03	HCoV-HKU1_N_04	MERS-CoV_N_01
PM3.05	HAdV_HEX_03	HCoV-HKU1_N_04	MERS-CoV_N_03
PM3.06	HAdV_HEX_03	HCoV-HKU1_N_04	MERS-CoV_N_04
PM3.07	HAdV_HEX_03	HCoV-HKU1_N_06	MERS-CoV_N_01
PM3.08	HAdV_HEX_03	HCoV-HKU1_N_06	MERS-CoV_N_03
PM3.09	HAdV_HEX_03	HCoV-HKU1_N_06	MERS-CoV_N_04
PM3.10	HAdV_HEX_09	HCoV-HKU1_N_02	MERS-CoV_N_01
PM3.11	HAdV_HEX_09	HCoV-HKU1_N_02	MERS-CoV_N_03
PM3.12	HAdV_HEX_09	HCoV-HKU1_N_02	MERS-CoV_N_04
PM3.13	HAdV_HEX_09	HCoV-HKU1_N_04	MERS-CoV_N_01
PM3.14	HAdV_HEX_09	HCoV-HKU1_N_04	MERS-CoV_N_03
PM3.15	HAdV_HEX_09	HCoV-HKU1_N_04	MERS-CoV_N_04
PM3.16	HAdV_HEX_09	HCoV-HKU1_N_06	MERS-CoV_N_01
PM3.17	HAdV_HEX_09	HCoV-HKU1_N_06	MERS-CoV_N_03
PM3.18	HAdV_HEX_09	HCoV-HKU1_N_06	MERS-CoV_N_04
PM3.19	HAdV_HEX_12	HCoV-HKU1_N_02	MERS-CoV_N_01
PM3.20	HAdV_HEX_12	HCoV-HKU1_N_02	MERS-CoV_N_03
PM3.21	HAdV_HEX_12	HCoV-HKU1_N_02	MERS-CoV_N_04
PM3.22	HAdV_HEX_12	HCoV-HKU1_N_04	MERS-CoV_N_01
PM3.23	HAdV_HEX_12	HCoV-HKU1_N_04	MERS-CoV_N_03
PM3.24	HAdV_HEX_12	HCoV-HKU1_N_04	MERS-CoV_N_04
PM3.25	HAdV_HEX_12	HCoV-HKU1_N_06	MERS-CoV_N_01
PM3.26	HAdV_HEX_12	HCoV-HKU1_N_06	MERS-CoV_N_03
PM3.27	HAdV_HEX_12	HCoV-HKU1_N_06	MERS-CoV_N_04

Table D.6: The c parameter stats for 3plex

Combo	sADS	eADS	sMDS	eMDS	RMSE	MSS	ACA*
PM3.01	0.117	0.139	0.003	0.013	0.022	0.218	98.97%
PM3.02	0.127	0.147	0.017	0.032	0.018	0.365	98.86%
PM3.03	0.115	0.124	0.041	0.030	0.018	0.264	99.90%
PM3.04	0.100	0.115	0.003	0.027	0.018	0.291	99.55%
PM3.05	0.110	0.130	0.017	0.045	0.022	0.399	98.92%
PM3.06	0.098	0.093	0.041	0.008	0.024	0.306	99.66%
PM3.07	0.024	0.034	0.003	0.019	0.020	0.249	99.28%
PM3.08	0.034	0.060	0.017	0.044	0.028	0.342	98.86%
PM3.09	0.027	0.028	0.008	0.001	0.009	0.241	99.82%
PM3.10	0.138	0.166	0.032	0.019	0.041	0.310	99.75%
PM3.11	0.138	0.159	0.017	0.016	0.027	0.210	99.35%
PM3.12	0.138	0.154	0.075	0.075	0.019	0.376	99.90%
PM3.13	0.121	0.138	0.032	0.011	0.031	0.274	99.75%
PM3.14	0.121	0.161	0.017	0.010	0.050	0.122	99.88%
PM3.15	0.121	0.151	0.075	0.069	0.040	0.485	99.90%
PM3.16	0.045	0.051	0.032	0.011	0.021	0.365	99.50%
PM3.17	0.045	0.047	0.017	0.016	0.003	0.211	98.63%
PM3.18	0.050	0.055	0.008	0.005	0.008	0.282	99.04%
PM3.19	0.117	0.123	0.043	0.056	0.009	0.391	99.95%
PM3.20	0.127	0.149	0.057	0.079	0.025	0.396	99.64%
PM3.21	0.089	0.096	0.001	0.019	0.014	0.152	99.75%
PM3.22	0.100	0.115	0.043	0.073	0.022	0.452	100.0%
PM3.23	0.110	0.150	0.057	0.072	0.044	0.414	99.52%
PM3.24	0.072	0.067	0.001	0.017	0.017	0.220	99.74%
PM3.25	0.028	0.010	0.007	0.005	0.027	0.093	99.88%
PM3.26	0.038	0.055	0.007	0.011	0.020	0.253	99.29%
PM3.27	0.005	0.019	0.001	0.002	0.019	0.147	99.51%

Combo: Combination or mulitplex assay name
sADS: Simulated ADS
eADS: Empirical ADS
sMDS: Simulated MDS
eMDS: Empirical MDS
RMSE: Rooted Mean Squared Error
MSS: Mean Silhouette Score
ACA: ACA accuracy

Table D.7: ADS and MDS scores for the three curve representations in 3plex

		Raw curve		FFI Normalised		5 fitted parameters	
Combination	**Multiplex type**	**ADS**	**MDS**	**ADS**	**MDS**	**ADS**	**MDS**
PM3.01	Simulated	2.544	0.585	0.434	0.130	66.089	5.489
PM3.01	Empirical	1.279	0.903	0.488	0.068	12.764	6.693
PM3.02	Simulated	2.978	1.239	0.655	0.440	66.115	5.489
PM3.02	Empirical	1.804	0.537	0.721	0.443	6.999	1.378
PM3.03	Simulated	2.260	0.647	0.393	0.261	8.067	5.489
PM3.03	Empirical	2.021	1.008	0.438	0.128	5.599	2.232
PM3.04	Simulated	2.462	0.585	0.401	0.130	65.369	4.357
PM3.04	Empirical	1.848	1.198	0.396	0.129	15.182	5.940
PM3.05	Simulated	2.896	1.239	0.622	0.440	65.395	4.357
PM3.05	Empirical	1.828	0.382	0.644	0.487	7.165	3.297
PM3.06	Simulated	2.174	0.514	0.359	0.261	7.290	4.357
PM3.06	Empirical	2.586	0.898	0.320	0.118	3.336	1.905
PM3.07	Simulated	0.884	0.585	0.209	0.130	64.450	4.770
PM3.07	Empirical	1.258	0.314	0.190	0.155	15.637	4.916
PM3.08	Simulated	1.286	0.810	0.334	0.265	64.535	4.989
PM3.08	Empirical	0.654	0.495	0.397	0.178	6.503	2.948
PM3.09	Simulated	1.999	0.810	0.321	0.261	62.730	7.012
PM3.09	Empirical	3.197	1.822	0.263	0.171	7.041	1.976
PM3.10	Simulated	3.638	1.664	0.644	0.378	66.659	2.998
PM3.10	Empirical	1.925	0.867	0.769	0.266	67.390	2.001
PM3.11	Simulated	3.654	1.066	0.656	0.059	66.601	2.779
PM3.11	Empirical	1.777	0.619	0.789	0.129	67.853	3.792
PM3.12	Simulated	3.734	0.647	0.646	0.338	66.257	11.699
PM3.12	Empirical	2.674	0.966	0.749	0.561	66.970	10.969
PM3.13	Simulated	3.558	1.664	0.611	0.378	65.969	2.998
PM3.13	Empirical	2.494	0.751	0.692	0.272	66.407	1.132
PM3.14	Simulated	3.574	1.066	0.622	0.059	65.910	2.779
PM3.14	Empirical	1.461	1.141	0.787	0.131	67.556	2.493
PM3.15	Simulated	3.650	0.514	0.613	0.287	65.509	10.501
PM3.15	Empirical	2.645	1.799	0.728	0.530	67.005	17.366
PM3.16	Simulated	1.930	1.258	0.292	0.199	5.179	2.998
PM3.16	Empirical	2.228	0.548	0.307	0.214	5.590	1.088
PM3.17	Simulated	1.914	1.066	0.208	0.059	5.178	2.779
PM3.17	Empirical	2.084	0.615	0.234	0.092	6.334	3.364
PM3.18	Simulated	3.425	2.304	0.448	0.301	61.079	7.767
PM3.18	Empirical	3.026	1.571	0.401	0.304	64.531	9.985
PM3.19	Simulated	3.207	1.027	0.422	0.091	65.674	2.055
PM3.19	Empirical	3.176	2.246	0.485	0.134	65.996	1.461
PM3.20	Simulated	3.261	0.542	0.636	0.382	65.703	2.055
PM3.20	Empirical	2.532	1.213	0.781	0.525	66.492	1.275
PM3.21	Simulated	3.301	0.647	0.392	0.259	7.947	2.055
PM3.21	Empirical	3.375	0.757	0.399	0.117	5.731	1.125
PM3.22	Simulated	3.127	1.027	0.388	0.091	65.057	1.230
PM3.22	Empirical	3.071	1.789	0.436	0.172	66.130	1.302
PM3.23	Simulated	3.181	0.542	0.603	0.382	65.085	1.230
PM3.23	Empirical	1.857	0.940	0.761	0.488	66.533	1.742
PM3.24	Simulated	3.217	0.514	0.359	0.259	7.272	1.230
PM3.24	Empirical	3.252	0.479	0.292	0.168	4.574	1.999
PM3.25	Simulated	1.511	1.027	0.149	0.091	65.860	4.770
PM3.25	Empirical	3.396	2.002	0.145	0.092	65.538	6.607
PM3.26	Simulated	1.534	0.542	0.268	0.157	65.947	4.989
PM3.26	Empirical	1.799	1.349	0.370	0.129	67.646	7.135
PM3.27	Simulated	3.004	2.249	0.273	0.157	64.435	10.085
PM3.27	Empirical	2.027	1.391	0.274	0.097	65.362	1.782

Table D.8: Assay Combination table for 7plex

Assay Name	HAdV Singleplex	Coronavirus 229E Singleplex	Coronavirus HKU1 Singleplex	Coronavirus NL63 Singleplex	Coronavirus OC43 Singleplex	SARS-CoV-2 Singleplex	MERS Singleplex
PM7.1176	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_02	HCoV-NL63_N_01	HCoV-OC43_N_02	SARS-CoV-2_N_03	MERS_N_04
PM7.1191	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_02	HCoV-NL63_N_02	HCoV-OC43_N_01	SARS-CoV-2_N_01	MERS_N_03
PM7.1286	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_02	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_01	MERS_N_02
PM7.1294	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_02	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_03	MERS_N_02
PM7.1318	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_04	HCoV-NL63_N_01	HCoV-OC43_N_02	SARS-CoV-2_N_03	MERS_N_02
PM7.1319	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_04	HCoV-NL63_N_01	HCoV-OC43_N_02	SARS-CoV-2_N_03	MERS_N_03
PM7.1339	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_04	HCoV-NL63_N_02	HCoV-OC43_N_01	SARS-CoV-2_N_02	MERS_N_03
PM7.1430	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_04	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_01	MERS_N_02
PM7.1449	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_06	HCoV-NL63_N_01	HCoV-OC43_N_01	SARS-CoV-2_N_03	MERS_N_01
PM7.1451	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_06	HCoV-NL63_N_01	HCoV-OC43_N_01	SARS-CoV-2_N_03	MERS_N_03
PM7.1593	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_08	HCoV-NL63_N_01	HCoV-OC43_N_01	SARS-CoV-2_N_03	MERS_N_01
PM7.1595	HAdV_HEX_06	HCoV-229E_N_01	HCoV-HKU1_N_08	HCoV-NL63_N_01	HCoV-OC43_N_01	SARS-CoV-2_N_03	MERS_N_03
PM7.2014	HAdV_HEX_06	HCoV-229E_N_02	HCoV-HKU1_N_04	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_03	MERS_N_02
PM7.2151	HAdV_HEX_06	HCoV-229E_N_02	HCoV-HKU1_N_06	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_01	MERS_N_03
PM7.2155	HAdV_HEX_06	HCoV-229E_N_02	HCoV-HKU1_N_06	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_02	MERS_N_03
PM7.2203	HAdV_HEX_06	HCoV-229E_N_02	HCoV-HKU1_N_08	HCoV-NL63_N_02	HCoV-OC43_N_01	SARS-CoV-2_N_02	MERS_N_03
PM7.2295	HAdV_HEX_06	HCoV-229E_N_02	HCoV-HKU1_N_08	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_01	MERS_N_03
PM7.2302	HAdV_HEX_06	HCoV-229E_N_02	HCoV-HKU1_N_08	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_03	MERS_N_02
PM7.2303	HAdV_HEX_06	HCoV-229E_N_02	HCoV-HKU1_N_08	HCoV-NL63_N_04	HCoV-OC43_N_04	SARS-CoV-2_N_03	MERS_N_03
PM7.2601	HAdV_HEX_09	HCoV-229E_N_01	HCoV-HKU1_N_06	HCoV-NL63_N_01	HCoV-OC43_N_01	SARS-CoV-2_N_03	MERS_N_01
PM7.2602	HAdV_HEX_09	HCoV-229E_N_01	HCoV-HKU1_N_06	HCoV-NL63_N_01	HCoV-OC43_N_01	SARS-CoV-2_N_03	MERS_N_02
PM7.4382	HAdV_HEX_12	HCoV-229E_N_02	HCoV-HKU1_N_06	HCoV-NL63_N_02	HCoV-OC43_N_04	SARS-CoV-2_N_01	MERS_N_02
PM7.4441	HAdV_HEX_12	HCoV-229E_N_02	HCoV-HKU1_N_06	HCoV-NL63_N_04	HCoV-OC43_N_02	SARS-CoV-2_N_01	MERS_N_01
PM7.4443	HAdV_HEX_12	HCoV-229E_N_02	HCoV-HKU1_N_06	HCoV-NL63_N_04	HCoV-OC43_N_02	SARS-CoV-2_N_01	MERS_N_03

Table D.9: The c parameter stats for 7plex (tested)

Combo	sADS	eADS	sMDS	eMDS	RMSE	MSS	ACA*
PM7.1176	0.098	0.101	0.012	0.002	0.011	0.382	59.44%
PM7.1191	0.098	0.107	0.013	0.018	0.009	0.441	76.54%
PM7.1286	0.190	0.188	0.024	0.003	0.026	0.345	82.88%
PM7.1294	0.193	0.189	0.024	0.006	0.019	0.379	91.99%
PM7.1318	0.098	0.092	0.012	0.002	0.029	0.381	60.38%
PM7.1319	0.097	0.090	0.012	0.002	0.031	0.521	83.45%
PM7.1339	0.097	0.089	0.013	0.004	0.018	0.426	94.01%
PM7.1430	0.191	0.184	0.024	0.010	0.026	0.330	71.62%
PM7.1449	0.034	0.039	0.000	0.004	0.012	0.591	93.64%
PM7.1451	0.035	0.041	0.000	0.001	0.014	0.557	95.32%
PM7.1593	0.034	0.041	0.000	0.001	0.012	0.555	92.92%
PM7.1595	0.035	0.041	0.000	0.003	0.010	0.393	94.41%
PM7.2014	0.195	0.206	0.007	0.007	0.035	0.576	90.46%
PM7.2151	0.190	0.182	0.037	0.012	0.031	0.456	97.10%
PM7.2155	0.190	0.182	0.037	0.016	0.023	0.484	70.63%
PM7.2203	0.098	0.120	0.012	0.008	0.033	0.520	92.71%
PM7.2295	0.192	0.180	0.025	0.009	0.036	0.574	95.51%
PM7.2302	0.195	0.208	0.002	0.005	0.024	0.561	88.85%
PM7.2303	0.192	0.191	0.002	0.006	0.041	0.538	84.60%
PM7.2601	0.028	0.031	0.000	0.001	0.006	0.626	98.86%
PM7.2602	0.032	0.034	0.000	0.001	0.010	0.527	97.34%
PM7.4382	0.155	0.188	0.023	0.033	0.033	0.691	97.16%
PM7.4441	0.145	0.157	0.023	0.018	0.026	0.612	91.72%
PM7.4443	0.146	0.163	0.023	0.022	0.031	0.424	85.25%

Combo: Combination or mulitplex assay name
sADS: Simulated ADS
eADS: Empirical ADS
sMDS: Simulated MDS
eMDS: Empirical MDS
RMSE: Rooted Mean Squared Error
MSS: Mean Silhouette Score
ACA: ACA accuracy

Appendix E

Supporting Information: Chapter 9

This Appendix contains the following:

- Summary of reported assays for nucleic-acid amplification of SARS CoV-2 (Figure E.1)
- Classification accuracy or AUC of two genes and four primer sets (single RNA signature) (Figure E.2)
- RNA signature translation to development of tailored molecular tests based on amplification chemistries (Figure E.3)

Author	Assay Type	Detection	Annealing Temp. (°C)	Gene Target	Limit of detection (LOD)	Time	Volume (uL)	Ref.
CDC	TaqMan	NA	55	N	NA	<45 cycles	20 uL total 5 uL sample	*12*
Lamb et al.	RT-qLAMP	colorimetric	63	ORF1ab (pp1ab)	0.2 fg/reaction (387 cp/reaction)	<30 min	25 µL total	*14*
Mohamed et al.	qLAMP	colorimetric	63	RdRP	70 cp/reaction (Optigene mix)	<50 min	10 µL total 1 µL template	*15*
Zhang et al.	RT-qLAMP	colorimetric	65	ORF1a	120 cp/reaction	<30 min	20 uL total 3 uL sample	*13*
Zhang et al.	RT-qLAMP	colorimetric	65	N	120 cp/reaction	<30 min	20 µL total 3 µL sample	*13*
Yang et al.	RT-LAMP	turbidimetry	63	ORF1ab	20-fold	<60 min	25 µL total 2 µL sample	*18*
Yang et al.	RT-LAMP	turbidimetry	63	N	160-fold	<60 min	25 µL total 2 µL sample	*18*
Yang et al.	RT-LAMP	turbidimetry	63	E	40-fold	<60 min	25 µL total 2 µL sample	*18*
Yu et al.	RT-LAMP	colorimetric	65	ORF1ab	10 cp/reaction	<40 min	20 µL total 1 µL sample	*16*
Park et al.	RT-LAMP	colorimetric	69	Spike (S), Orf8, N*	100 cp/reaction	<30 min	15 µL total	*19*
Jiang et al.	RT-LAMP	colorimetric	63	N	500 cp/mL	<30 min	25 µL total 2 µL sample	*20*
Zhu et al.	RT-LAMP-NBS	colorimetric (NPs, LF)	63	ORF1ab, N	12 cp/ reaction	<1 hr (sample to result)	25 µL total 5 µL sample	*21*
Ding et al.	AIOD-CRISPR (RPA-based)	colorimetric (LED)	37	N	1.3 cp of plasmid	<40 min	25 µL total 1 µL sample	*22*
Lu et al.	RT-LAMP	colorimetric	63	ORF1ab (RdRP)	300 cp/reaction	<40 min	25 µL total	*23*
This study	RT-qLAMP	electrochemical	63	N	10 cp/reaction	<20 min	5 µL total 2 µL sample	*This study*

Figure E.1: Summary of reported assays for nucleic-acid amplification of SARS CoV-2.

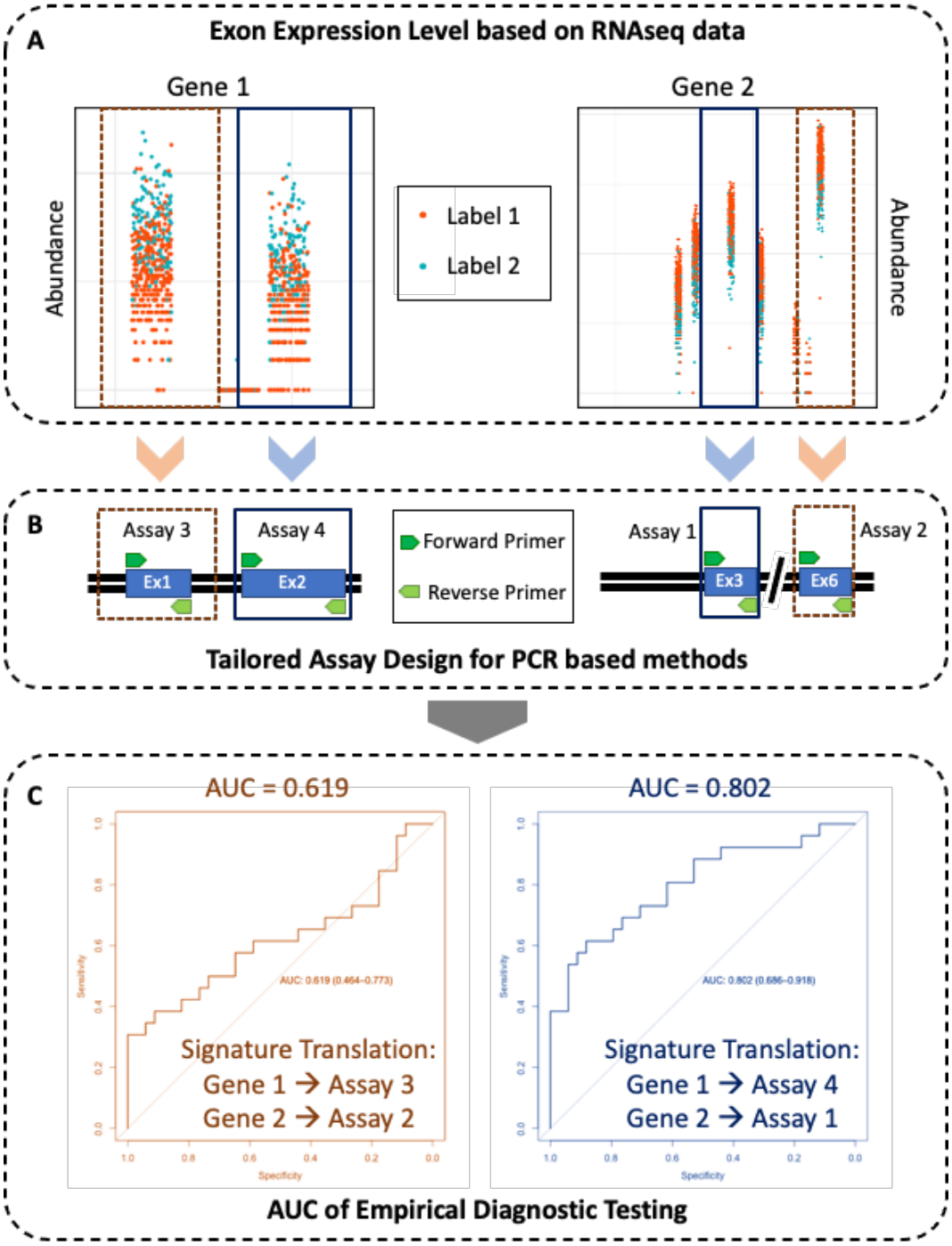


Figure E.2: Classification accuracy or AUC of two genes and four primer sets (single RNA signature).

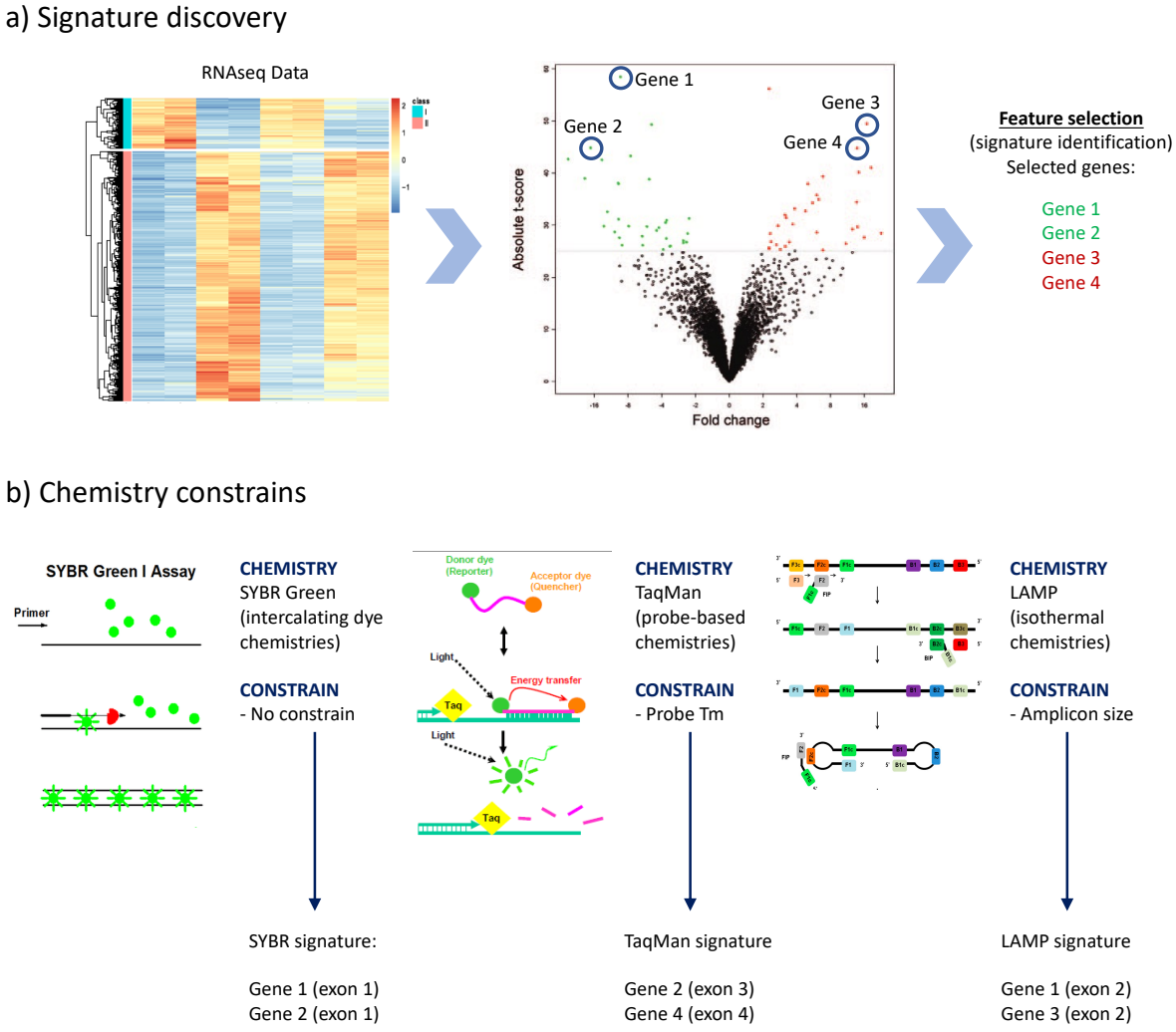


Figure E.3: RNA signature translation to development of tailored molecular tests based on amplification chemistries.

Bibliography

[1] Sangeeta Dhami, Deidre Thompson, Maha El Akoum, David W Bates, Roberto Bertollini, and Aziz Sheikh. Data-enabled responses to pandemics: policy lessons from COVID-19. *Nature Medicine*, pages 1–4, 2022.

[2] Michael C Schatz. Biological data sciences in genome research. *Genome research*, 25(10):1417–1422, 2015.

[3] Timothy Miles Rawson, Nathan Peiffer-Smadja, and Alison Holmes. Artificial Intelligence in Infectious Diseases. *Artificial Intelligence in Medicine*, pages 1–14, 2020.

[4] Yi-Chi Wu, Ching-Sung Chen, and Yu-Jiun Chan. The outbreak of COVID-19: An overview. *Journal of the Chinese medical association*, 83(3):217, 2020.

[5] Dana Trevas, Angela M Caliendo, Kimberly Hanson, Jaclyn Levy, and Christine C Ginocchio. Diagnostic tests can stem the threat of antimicrobial resistance: infectious disease professionals can help. *Clinical Infectious Diseases*, 72(11):e893–e900, 2021.

[6] Jocelyne Piret and Guy Boivin. Pandemics throughout history. *Frontiers in microbiology*, 11:631736, 2021.

[7] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[8] Charles Yanofsky. Establishing the triplet nature of the genetic code. *Cell*, 128(5):815–818, 2007.

[9] John Walker. Frederick Sanger (1918–2013). *Nature*, 505(7481):27–27, 2014.

[10] Hanliang Zhu, Haoqing Zhang, Ying Xu, Soňa Laššáková, Marie Korabečná, and Pavel Neužil. PCR past, present and future. *BioTechniques*, 69(4):317–325, 2020.

[11] Kary Mullis, Fred Faloona, Stephen Scharf, Randall Saiki, Glenn Horn, and Henry Erlich. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. In *Cold Spring Harbor symposia on quantitative biology*, volume 51, pages 263–273. Cold Spring Harbor Laboratory Press, 1986.

[12] Jeffrey S Chamberlain, Richard A Gibbs, Joel E Rainer, Phi Nga Nguyen, and C Thomas. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic acids research*, 16(23):11141–11156, 1988.

[13] Dasheng Lee, Pei-Jer Chen, and Gwo-Bin Lee. The evolution of real-time PCR machines to real-time PCR chips. *Biosensors and Bioelectronics*, 25(7):1820–1824, 2010.

[14] Qi Song, Xindi Sun, Ziyi Dai, Yibo Gao, Xiuqing Gong, Bingpu Zhou, Jinbo Wu, and Weijia Wen. Point-of-care testing detection methods for COVID-19. *Lab on a Chip*, 21(9):1634–1660, 2021.

[15] Ahmad Moniri, Jesus Rodriguez-Manzano, Kenny Malpartida-Cardenas, Ling-Shan Yu, Xavier Didelot, Alison Holmes, and Pantelis Georgiou. Framework for DNA quantification and outlier detection using multidimensional standard curves. *Analytical chemistry*, 91(11):7426–7434, 2019.

[16] Jesus Rodriguez-Manzano, Ahmad Moniri, Kenny Malpartida-Cardenas, Jyothsna Dronavalli, Frances Davies, Alison Holmes, and Pantelis Georgiou. Simultaneous single-channel multiplexing and quantification of carbapenem-resistant genes using multidimensional standard curves. *Analytical chemistry*, 91(3):2013–2020, 2019.

[17] Ahmad Moniri. Intelligent algorithms for DNA detection, quantification and multiplexing. 2021.

[18] Manit Arya, Iqbal S Shergill, Magali Williamson, Lyndon Gommersall, Neehar Arya, and Hitendra RH Patel. Basic principles of real-time quantitative PCR. *Expert review of molecular diagnostics*, 5(2):209–219, 2005.

[19] Phouthone Keohavong and William G Thilly. Fidelity of DNA polymerases in DNA amplification. *Proceedings of the National Academy of Sciences*, 86(23):9253–9257, 1989.

[20] RG Rutledge and C Cote. Mathematics of quantitative kinetic PCR and the application of standard curves. *Nucleic acids research*, 31(16):e93–e93, 2003.

[21] Phenix-Lan Quan, Martin Sauzade, and Eric Brouzes. dPCR: a technology review. *Sensors*, 18(4):1271, 2018.

[22] M Tevfik Dorak. *Real-time PCR*. Taylor & Francis, 2007.

[23] David Svec, Ales Tichopad, Vendula Novosadova, Michael W Pfaffl, and Mikael Kubista. How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments. *Biomolecular detection and quantification*, 3:9–16, 2015.

[24] Heather D VanGuilder, Kent E Vrana, and Willard M Freeman. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*, 44(5):619–626, 2008.

[25] Edith Frahm and Ursula Obst. Application of the fluorogenic probe technique (TaqMan PCR) to the detection of Enterococcus spp. and Escherichia coli in water samples. *Journal of microbiological methods*, 52(1):123–131, 2003.

[26] Louis Kreitmann, Luca Miglietta, Ke Xu, Kenny Malpartida-Cardenas, Giselle D'Souza, Myrsini Kaforou, Karen Brengel-Pesce, Laurent Drazek, Alison Holmes, and Jesus Rodriguez-Manzano. Next-generation molecular diagnostics: Leveraging digital technologies to enhance multiplexing in real-time PCR. *TrAC Trends in Analytical Chemistry*, page 116963, 2023.

[27] E van Pelt-Verkuil, WB van Leeuwen, and R te Witt. Molecular Diagnostics: Part 1: Technical Backgrounds and Quality Aspects, 2019.

[28] Leah M Dignan, Rachelle Turiello, Tiffany R Layne, Killian C O'Connell, Jeff Hickey, Jeff Chapman, Melinda D Poulter, and James P Landers. An ultrafast SARS-CoV-2 virus enrichment and extraction method compatible with multiple modalities for RNA detection. *Analytica Chimica Acta*, 1180:338846, 2021.

[29] C Schrader, A Schielke, L Ellerbroek, and R Johne. PCR inhibitors–occurrence, properties and removal. *Journal of applied microbiology*, 113(5):1014–1026, 2012.

[30] Robert G Rutledge and Don Stewart. A kinetic-based sigmoidal model for the polymerase chain reaction and its application to high-capacity absolute quantitative real-time PCR. *BMC biotechnology*, 8(1):1–28, 2008.

[31] Andrej-Nikolai Spiess, Caroline Feig, and Christian Ritz. Highly accurate sigmoidal fitting of real-time PCR data by introducing a parameter for asymmetry. *BMC bioinformatics*, 9(1):1–12, 2008.

[32] Kenneth J Breslauer, Ronald Frank, Helmut Blöcker, and Luis A Marky. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences*, 83(11):3746–3750, 1986.

[33] Robert J Pryor and Carl T Wittwer. Real-time polymerase chain reaction and melting curve analysis. In *Clinical applications of PCR*, pages 19–32. Springer, 2006.

[34] Milo M Lin, Lars Meinhold, Dmitry Shorokhov, and Ahmed H Zewail. Unfolding and melting of DNA (RNA) hairpins: the concept of structure-specific 2D dynamic landscapes. *Physical Chemistry Chemical Physics*, 10(29):4227–4239, 2008.

[35] Amber L Robertson and Allison R Phillips. Integrating PCR theory and bioinformatics into a research-oriented primer design exercise. *CBE—Life Sciences Education*, 7(1):89–95, 2008.

[36] Heike Margot, Roger Stephan, Salvatore Guarino, Balamurugan Jagadeesan, David Chilton, Emer O'Mahony, and Carol Iversen. Inclusivity, exclusivity and limit of detection of commercially available real-time PCR assays for the detection of Salmonella. *International journal of food microbiology*, 165(3):221–226, 2013.

[37] Jonathan Marks. What is molecular anthropology? What can it be? *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 11(4):131–135, 2002.

[38] Gareth Iacobucci. Covid-19: New UK variant may be linked to increased death rate, early data indicate. *bmj*, 372(230):n230, 2021.

[39] Chiara Ippoliti, Flavio De Maio, Giulia Santarelli, Simona Marchetti, Antonietta Vella, Rosaria Santangelo, Maurizio Sanguinetti, and Brunella Posteraro. Rapid detection of the omicron (B. 1.1. 529) SARS-CoV-2 variant using a COVID-19 diagnostic PCR assay. *Microbiology Spectrum*, 10(4):e00990–22, 2022.

[40] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, Barbara A Rapp, and David L Wheeler. GenBank. *Nucleic acids research*, 28(1):15–18, 2000.

[41] Wendy Baker, Alexandra van den Broek, Evelyn Camon, Pascal Hingamp, Peter Sterk, Guenter Stoesser, and Mary Ann Tuli. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 28(1):19–23, 2000.

[42] Yuelong Shu and John McCauley. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[43] Supratim Mukherjee, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A Chen, Nikos C Kyrpides, and TBK Reddy. Genomes OnLine Database (GOLD) v. 8: overview and updates. *Nucleic acids research*, 49(D1):D723–D733, 2021.

[44] Elizabeth M Smigielski, Karl Sirotkin, Minghong Ward, and Stephen T Sherry. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research*, 28(1):352–355, 2000.

[45] James J Davis, Alice R Wattam, Ramy K Aziz, Thomas Brettin, Ralph Butler, Rory M Butler, Philippe Chlenski, Neal Conrad, Allan Dickerman, Emily M Dietrich, et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic acids research*, 48(D1):D606–D612, 2020.

[46] William R Pearson and David J Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448, 1988.

[47] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[48] Francois Jeanmougin, Julie D Thompson, Manolo Gouy, Desmond G Higgins, and Toby J Gibson. Multiple sequence alignment with Clustal X. *Trends in biochemical sciences*, 23(10):403–405, 1998.

[49] David James Russell. *Multiple sequence alignment methods*. Springer, 2014.

[50] Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C Faircloth, Maido Remm, and Steven G Rozen. Primer3—new capabilities and interfaces. *Nucleic acids research*, 40(15):e115–e115, 2012.

[51] Jian Ye, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, and Thomas L Madden. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics*, 13:1–11, 2012.

[52] Haoqing Zhang, Zhiqiang Yan, Xinlu Wang, Martina Ganova, Honglong Chang, Sona Lassakova, Marie Korabecna, and Pavel Neuzil. Determination of advantages and limitations of qPCR duplexing in a single fluorescent channel. *ACS omega*, 6(34):22292–22300, 2021.

[53] Lucien Jacky, Dominic Yurk, John Alvarado, Paul Belitz, Kristin Fathe, Chris MacDonald, Scott Fraser, and Aditya Rajagopal. Robust multichannel encoding for highly multiplexed quantitative PCR. *Analytical Chemistry*, 93(9):4208–4216, 2021.

[54] J Gray and LJ Coupland. The increasing application of multiplex nucleic acid detection tests to the diagnosis of syndromic infections. *Epidemiology & Infection*, 142(1):1–11, 2014.

[55] Carl T Wittwer, Mark G Herrmann, Cameron N Gundry, and Kojo SJ Elenitoba-Johnson. Real-time multiplex PCR assays. *Methods*, 25(4):430–442, 2001.

[56] Sean Taylor, Michael Wakem, Greg Dijkman, Marwan Alsarraj, and Marie Nguyen. A practical approach to RT-qPCR—publishing data that conform to the MIQE guidelines. *Methods*, 50(4):S1–S5, 2010.

[57] Isa S Abubakar, Saidu B Abubakar, Abdulrazaq G Habib, Abdulsalam Nasidi, Nandul Durfa, Peter O Yusuf, Solomon Larnyang, John Garnvwa, Elijah Sokomba, Lateef Salako, et al. Randomised controlled double-blind non-inferiority trial of two antivenoms for saw-scaled or carpet viper (Echis ocellatus) envenoming in Nigeria. *PLoS Neglected Tropical Diseases*, 4(7):e767, 2010.

[58] Chase E Guion, Theresa J Ochoa, Christopher M Walker, Francesca Barletta, and Thomas G Cleary. Detection of diarrheagenic Escherichia coli by use of melting-curve analysis and real-time multiplex PCR. *Journal of clinical microbiology*, 46(5):1752–1757, 2008.

[59] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[60] P Russel Norvig and S Artificial Intelligence. A modern approach. *Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems*, 90:33–48, 2002.

[61] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[62] Geoffrey Hinton and Terrence J Sejnowski. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[63] Jesse Montgomery, Carl T Wittwer, Robert Palais, and Luming Zhou. Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis. *Nature protocols*, 2(1):59–66, 2007.

[64] Valin Reja, Alister Kwok, Glenn Stone, Linsong Yang, Andreas Missel, Christoph Menzel, and Brant Bassam. ScreenClust: Advanced statistical software for supervised and unsupervised high resolution melting (HRM) analysis. *Methods*, 50(4):S10–S14, 2010.

[65] Zachary Dwight, Robert Palais, and Carl T Wittwer. uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application. *Bioinformatics*, 27(7):1019–1020, 2011.

[66] Pornpat Athamanolap, Vishwa Parekh, Stephanie I Fraley, Vatsal Agarwal, Dong J Shin, Michael A Jacobs, Tza-Huei Wang, and Samuel Yang. Trainable high resolution melt curve machine learning classifier for large-scale reliable genotyping of sequence variants. *PloS one*, 9(10):e109094, 2014.

[67] Fatma Ozge Ozkok and Mete Celik. Convolutional neural network analysis of recurrence plots for high resolution melting classification. *Computer Methods and Programs in Biomedicine*, 207:106139, 2021.

[68] Aditya Rajagopal, Dominic Yurk, Claudia Shin, Karen Menge, Lucien Jacky, Scott Fraser, Thomas A Tombrello, and Gregory J Tsongalis. Significant expansion of real-time PCR multiplexing with traditional chemistries using amplitude modulation. *Scientific reports*, 9(1):1–8, 2019.

[69] Young-Jo Lee, Daeyoung Kim, Kihoon Lee, and Jong-Yoon Chun. Single-channel multiplexing without melting curve analysis in real-time PCR. *Scientific reports*, 4(1):1–6, 2014.

[70] Tsugunori Notomi, Hiroto Okayama, Harumi Masubuchi, Toshihiro Yonekawa, Keiko Watanabe, Nobuyuki Amino, and Tetsu Hase. Loop-mediated isothermal amplification of DNA. *Nucleic acids research*, 28(12):e63–e63, 2000.

[71] Kentaro Nagamine, Tesu Hase, and TJMCP Notomi. Accelerated reaction by loop-mediated isothermal amplification using loop primers. *Molecular and cellular probes*, 16(3):223–229, 2002.

[72] Nicolas Moser, Tor Sverre Lande, Christofer Toumazou, and Pantelis Georgiou. ISFETs in CMOS and emergent trends in instrumentation: A review. *IEEE Sensors Journal*, 16(17):6496–6514, 2016.

[73] Bert Vogelstein and Kenneth W Kinzler. Digital pcr. *Proceedings of the National Academy of Sciences*, 96(16):9236–9241, 1999.

[74] Olga Kalinina, Irina Lebedeva, James Brown, and Jonathan Silver. Nanoliter scale PCR with TaqMan detection. *Nucleic acids research*, 25(10):1999–2004, 1997.

[75] Simant Dube, Jian Qin, and Ramesh Ramakrishnan. Mathematical analysis of copy number variation in a DNA sample using digital PCR on a nanofluidic device. *PloS one*, 3(8):e2876, 2008.

[76] Alexandra S Whale, Simon Cowen, Carole A Foy, and Jim F Huggett. Methods for applying accurate digital PCR analysis on low copy DNA samples. *PloS one*, 8(3):e58177, 2013.

[77] Li Ling Tan, Nitin Loganathan, Sushama Agarwalla, Chun Yang, Weiyong Yuan, Jasmine Zeng, Ruige Wu, Wei Wang, and Suhanya Duraiswamy. Current commercial dPCR platforms: Technology and market review. *Critical Reviews in Biotechnology*, pages 1–32, 2022.

[78] Jim F Huggett. The digital MIQE guidelines update: minimum information for publication of quantitative digital PCR experiments for 2020. *Clinical Chemistry*, 66(8):1012–1029, 2020.

[79] Sean Wallis. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, 20(3):178–208, 2013.

[80] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[81] Elizabeth A Ottesen, Jong Wook Hong, Stephen R Quake, and Jared R Leadbetter. Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *science*, 314(5804):1464–1467, 2006.

[82] Feng Shen, Wenbin Du, Jason E Kreutz, Alice Fok, and Rustem F Ismagilov. Digital PCR on a SlipChip. *Lab on a Chip*, 10(20):2666–2672, 2010.

[83] Amar S Basu. Digital assays part I: partitioning statistics and digital PCR. *SLAS technology*, 22(4):369–386, 2017.

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

[85] Kamalalayam Rajan Sreejith, Chin Hong Ooi, Jing Jin, Dzung Viet Dao, and Nam-Trung Nguyen. Digital polymerase chain reaction technology–recent advances and future perspectives. *Lab on a chip*, 18(24):3717–3732, 2018.

[86] Daan Witters, Bing Sun, Stefano Begolo, Jesus Rodriguez-Manzano, Whitney Robles, and Rustem F Ismagilov. Digital biology and chemistry. *Lab on a Chip*, 14(17):3225–3232, 2014.

[87] G Perkins, H Lu, F Garlan, and V Taly. Droplet-based digital PCR: application in cancer research. In *Advances in clinical chemistry*, volume 79, pages 43–91. Elsevier, 2017.

[88] Susana Olmedillas-López, Mariano García-Arranz, and Damián García-Olmo. Current and emerging applications of droplet digital PCR in oncology. *Molecular diagnosis & therapy*, 21(5):493–510, 2017.

[89] Yu Tong, Shizhen Shen, Hui Jiang, and Zhi Chen. Application of digital PCR in detecting human diseases associated gene mutation. *Cellular Physiology and Biochemistry*, 43(4):1718–1730, 2017.

[90] Mohamadhasan Tajadini, Mojtaba Panjehpour, and Shaghayegh Haghjooy Javanmard. Comparison of SYBR Green and TaqMan methods in quantitative real-time polymerase chain reaction analysis of four adenosine receptor subtypes. *Advanced biomedical research*, 3, 2014.

[91] JM Ruijter, C Ramakers, WMH Hoogaars, Y Karlen, O Bakker, MJB Van den Hoff, and AFM Moorman. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic acids research*, 37(6):e45–e45, 2009.

[92] Dalal Hammoudi Halat and Carole Ayoub Moubareck. The Current Burden of Carbapenemases: Review of Significant Properties and Dissemination among Gram-Negative Bacteria. *Antibiotics*, 9(4):186, 2020.

[93] Jesus Rodriguez-Manzano, Nicolas Moser, Kenny Malpartida-Cardenas, Ahmad Moniri, Lenka Fisarova, Ivana Pennisi, Adhiratha Boonyasiri, Elita Jauneikaite, Alireza Abdolrasouli, Jonathan A Otter, et al. Rapid Detection of Mobilized Colistin Resistance using a Nucleic Acid Based Lab-on-a-Chip Diagnostic System. *Scientific Reports*, 10(1):1–9, 2020.

[94] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[95] Matthew Kearse, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, Alex Cooper, Sidney Markowitz, Chris Duran, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012.

[96] John SantaLucia Jr and Donald Hicks. The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, 33:415–440, 2004.

[97] Ahmad Moniri, Luca Miglietta, Kenny Malpartida-Cardenas, Ivana Pennisi, Miguel Cacho-Soblechero, Nicolas Moser, Alison Holmes, Pantelis Georgiou, and Jesus Rodriguez-Manzano. Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR. *Submitted to Analytical Chemistry*, 2020.

[98] Bang Wong. Points of view: Color blindness. *Nature Methods*, 8(6):441–441, 2011.

[99] Geoffrey P McDermott, Duc Do, Claudia M Litterst, Dianna Maar, Christopher M Hindson, Erin R Steenblock, Tina C Legler, Yann Jouvenot, Samuel H Marrs, Adam Bemis, et al. Multiplexed target detection using DNA-binding dye chemistry in droplet digital PCR. *Analytical chemistry*, 85(23):11619–11627, 2013.

[100] Kirk M Ririe, Randy P Rasmussen, and Carl T Wittwer. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Analytical biochemistry*, 245(2):154–160, 1997.

[101] B Cherie Millar, Jiru Xu, and John E Moore. Risk assessment models and contamination management: implications for broad-range ribosomal DNA PCR as a diagnostic tool in medical bacteriology. *Journal of Clinical Microbiology*, 40(5):1575–1580, 2002.

[102] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[103] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, NY, USA:, 2012.

[104] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[105] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[106] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[107] Russell Higuchi, Carita Fockler, Gavin Dollinger, and Robert Watson. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/technology*, 11(9):1026–1030, 1993.

[108] Kenneth J Livak and Thomas D Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the 2- $\Delta\Delta$CT method. *methods*, 25(4):402–408, 2001.

[109] Thomas R Gingeras, Russell Higuchi, Larry J Kricka, YM Dennis Lo, and Carl T Wittwer. Fifty years of molecular (DNA/RNA) diagnostics. *Clinical chemistry*, 51(3):661–671, 2005.

[110] Amal Bouzid, Ibtihel Smeti, Amine Chakroun, Salma Loukil, Abdullah Ahmed Gibriel, Mhamed Grati, Abdelmonem Ghorbel, and Saber Masmoudi. CDH23 Methylation status and presbycusis risk in elderly women. *Frontiers in aging neuroscience*, 10:241, 2018.

[111] Shohda A El-Maraghy, Ola Adel, Naglaa Zayed, Ayman Yosry, Saeed M El-Nahaas, and Abdullah A Gibriel. 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. *Journal of Advanced Research*, 22:57–66, 2020.

[112] Stephen A Bustin, Vladimir Benes, Jeremy A Garson, Jan Hellemans, Jim Huggett, Mikael Kubista, Reinhold Mueller, Tania Nolan, Michael W Pfaffl, Gregory L Shipley, et al. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments, 2009.

[113] Alexandra S Whale, Jim F Huggett, Simon Cowen, Valerie Speirs, Jacqui Shaw, Stephen Ellison, Carole A Foy, and Daniel J Scott. Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. *Nucleic acids research*, 40(11):e82–e82, 2012.

[114] Jane Kuypers and Keith R Jerome. Applications of digital PCR for clinical microbiology. *Journal of clinical microbiology*, 55(6):1621–1628, 2017.

[115] Megan E Dueck, Robert Lin, Andrew Zayac, Steve Gallagher, Alexander K Chao, Lingxia Jiang, Sammy S Datwani, Paul Hung, and Elliot Stieglitz. Precision cancer monitoring using a novel, fully integrated, microfluidic array partitioning digital PCR platform. *Scientific reports*, 9(1):1–9, 2019.

[116] G Zangenberg, RK Saiki, and R Reynolds. Multiplex PCR: optimization guidelines. In *PCR Applications*, pages 73–94. Elsevier, 1999.

[117] Alexandra S Whale, Jim F Huggett, and Svilen Tzonev. Fundamentals of multiplexing with digital PCR. *Biomolecular Detection and Quantification*, 10:15–23, 2016.

[118] Stephanie I Fraley, Justin Hardick, Billie Jo Masek, Pornpat Athamanolap, Richard E Rothman, Charlotte A Gaydos, Karen C Carroll, Teresa Wakefield, Tza-Huei Wang, and Samuel Yang. Universal digital high-resolution melt: a novel approach to broad-based profiling of heterogeneous biological samples. *Nucleic acids research*, 41(18):e175–e175, 2013.

[119] Justin C Rolando, Erik Jue, Jacob T Barlow, and Rustem F Ismagilov. Real-time kinetics and high-resolution melt curves in single-molecule digital LAMP to differentiate and study specific and non-specific amplification. *Nucleic acids research*, 48(7):e42–e42, 2020.

[120] Tatsuo Nakagawa, Junko Tanaka, Kunio Harada, Akiko Shiratori, Yuzuru Shimazaki, Takahide Yokoi, Chihiro Uematsu, and Yoshinobu Kohara. 10-plex digital polymerase chain reaction with four-color melting curve analysis for simultaneous KRAS and BRAF genotyping. *Analytical Chemistry*, 92(17):11705–11713, 2020.

[121] Pornpat Athamanolap, Kuangwen Hsieh, Christine M O'Keefe, Ye Zhang, Samuel Yang, and Tza-Huei Wang. Nanoarray digital polymerase chain reaction with high-resolution melt for enabling broad bacteria identification and pheno–molecular antimicrobial susceptibility test. *Analytical chemistry*, 91(20):12784–12792, 2019.

[122] World Health Organization et al. *Antimicrobial resistance: global report on surveillance*. World Health Organization, 2014.

[123] Tiago Lima, Sara Domingues, and Gabriela Jorge Da Silva. Plasmid-mediated colistin resistance in Salmonella enterica: a review. *Microorganisms*, 7(2):55, 2019.

[124] Laura M Carroll, Ahmed Gaballa, Claudia Guldimann, Genevieve Sullivan, Lory O Henderson, and Martin Wiedmann. Identification of Novel Mobilized Colistin Resistance Gene mcr-9 in a Multidrug-Resistant, Colistin-Susceptible Salmonella enterica Serotype Typhimurium Isolate. *mBio*, 10(3):e00853–19, 2019.

[125] Jesus Rodriguez-Manzano, Nicolas Moser, Kenny Malpartida-Cardenas, Ahmad Moniri, Lenka Fisarova, Ivana Pennisi, Adhiratha Boonyasiri, Elita Jauneikaite, Alireza Abdolrasouli, Jonathan A Otter, et al. Rapid Detection of Mobilized colistin Resistance using a nucleic Acid Based Lab-on-a-chip Diagnostic System. *Scientific Reports*, 10(1):1–9, 2020.

[126] Jonathan A Otter, Michel Doumith, Frances Davies, Siddharth Mookerjee, Eleonora Dyakova, Mark Gilchrist, Eimear T Brannigan, Kathleen Bamford, Tracey Galletly, Hugo Donaldson, et al. Emergence and clonal spread of colistin resistance due to multiple mutational mechanisms in carbapenemase-producing Klebsiella pneumoniae in London. *Scientific reports*, 7(1):1–8, 2017.

[127] Prasanth Manohar, Thamaraiselvan Shanthini, Ramankannan Ayyanar, Bulent Bozdogan, Aruni Wilson, Ashok J Tamhankar, Ramesh Nachimuthu, and Bruno S Lopes. The distribution of carbapenem-and colistin-resistance in Gram-negative bacteria from the Tamil Nadu region in India. *Journal of medical microbiology*, 66(7):874–883, 2017.

[128] JA Otter, P Burgess, F Davies, S Mookerjee, J Singleton, M Gilchrist, D Parsons, ET Brannigan, J Robotham, and AH Holmes. Counting the cost of an outbreak of carbapenemase-producing Enterobacteriaceae: an economic evaluation from a hospital perspective. *Clinical Microbiology and Infection*, 23(3):188–196, 2017.

[129] 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.

[130] Ahmad Moniri, Luca Miglietta, Alison Holmes, Pantelis Georgiou, and Jesus Rodriguez-Manzano. High-Level Multiplexing in Digital PCR with Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis. *Analytical Chemistry*, 92(20):14181–14188, 2020.

[131] 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.

[132] Daniel Ortiz Velez, Hannah Mack, Julietta Jupe, Sinead Hawker, Ninad Kulkarni, Behnam Hedayatnia, Yang Zhang, Shelley Lawrence, and Stephanie I Fraley. Massively parallel digital high resolution melt for rapid and absolutely quantitative sequence profiling. *Scientific reports*, 7(1):1–14, 2017.

[133] Nithum Thain, Christopher Le, Aldo Crossa, Shama Desai Ahuja, Jeanne Sullivan Meissner, Barun Mathema, Barry Kreiswirth, Natalia Kurepina, Ted Cohen, and Leonid Chindelevitch. Towards better prediction of Mycobacterium tuberculosis lineages from MIRU-VNTR data. *Infection, Genetics and Evolution*, 72:59–66, 2019.

[134] Vanesa García, Isidro García-Meniño, Azucena Mora, Saskia C Flament-Simon, Dafne Díaz-Jiménez, Jesús E Blanco, María Pilar Alonso, and Jorge Blanco. 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). *International journal of antimicrobial agents*, 52(1):104–108, 2018.

[135] Jin Hwa Kim, Minhee Kang, Eunkyoung Park, Doo Ryeon Chung, Jiyeon Kim, and Eung Soo Hwang. A simple and multiplex loop-mediated isothermal amplification (LAMP) assay for rapid detection of SARS-CoV. *BioChip Journal*, 13(4):341–351, 2019.

[136] Jeeyong Kim, Borae G Park, Da Hye Lim, Woong Sik Jang, Jeonghun Nam, Do-CiC Mihn, and Chae Seung Lim. Development and evaluation of a multiplex loop-mediated isothermal amplification (LAMP) assay for differentiation of Mycobacterium tuberculosis and non-tuberculosis mycobacterium in clinical samples. *Plos one*, 16(1):e0244753, 2021.

[137] Ikuyo Takayama, Mina Nakauchi, Hitoshi Takahashi, Kunihiro Oba, Shohei Semba, Atsushi Kaida, Hideyuki Kubo, Shinji Saito, Shiho Nagata, Takato Odagiri, et al. Development of real-time fluorescent reverse transcription loop-mediated isothermal amplification assay with quenching primer for influenza virus and respiratory syncytial virus. *Journal of virological methods*, 267:53–58, 2019.

[138] Nathan A Tanner, Yinhua Zhang, and Thomas C Evans Jr. Simultaneous multiple target detection in real-time loop-mediated isothermal amplification. *Biotechniques*, 53(2):81–89, 2012.

[139] Rungong Yang, Honghong Zhang, Xiaoxia Li, Ling Ye, Meiliang Gong, Jinghui Yang, Jihong Yu, and Jie Bai. A multiplex loop-mediated isothermal amplification assay for rapid screening of Acinetobacter baumannii and D carbapenemase OXA-23 gene. *Bioscience reports*, 38(5), 2018.

[140] Lan-Lan Zhong, Qian Zhou, Cui-Yan Tan, Adam P Roberts, Mohamed Abd El-Gawad El-Sayed Ahmed, Guanping Chen, Min Dai, Fan Yang, Yong Xia, Kang Liao, et al. Multiplex

loop-mediated isothermal amplification (multi-LAMP) assay for rapid detection of mcr-1 to mcr-5 in colistin-resistant bacteria. *Infection and drug resistance*, pages 1877–1887, 2019.

[141] James Mahony, Sylvia Chong, David Bulir, Alexandra Ruyter, Ken Mwawasi, and Daniel Waltho. Multiplex loop-mediated isothermal amplification (M-LAMP) assay for the detection of influenza A/H1, A/H3 and influenza B can provide a specimen-to-result diagnosis in 40 min with single genome copy sensitivity. *Journal of Clinical Virology*, 58(1):127–131, 2013.

[142] Ningwei Liu, Dayang Zou, Derong Dong, Zhan Yang, Da Ao, Wei Liu, and Liuyu Huang. Development of a multiplex loop-mediated isothermal amplification method for the simultaneous detection of Salmonella spp. and Vibrio parahaemolyticus. *Scientific reports*, 7(1):1–7, 2017.

[143] Zablon Kithinji Njiru, Andrew Stanislaw John Mikosza, Tanya Armstrong, John Charles Enyaru, Joseph Mathu Ndung'u, and Andrew Richard Christopher Thompson. Loop-mediated isothermal amplification (LAMP) method for rapid detection of Trypanosoma brucei rhodesiense. *PLoS neglected tropical diseases*, 2(2):e147, 2008.

[144] Luca Miglietta, Ahmad Moniri, Ivana Pennisi, Kenny Malpartida-Cardenas, Hala Abbas, Kerri Hill-Cawthorne, Frances Bolt, Elita Jauneikaite, Frances Davies, Alison Holmes, et al. Coupling machine learning and high throughput multiplex digital PCR enables accurate detection of carbapenem-resistant genes in clinical isolates. *Frontiers in molecular biosciences*, 8, 2021.

[145] Xiaojuan Zhu, Yiyue Ge, Tao Wu, Kangchen Zhao, Yin Chen, Bin Wu, Fengcai Zhu, Baoli Zhu, and Lunbiao Cui. Co-infection with respiratory pathogens among COVID-2019 cases. *Virus research*, 285:198005, 2020.

[146] Jesus Rodriguez-Manzano, Kenny Malpartida-Cardenas, Nicolas Moser, Ivana Pennisi, Matthew Cavuto, Luca Miglietta, Ahmad Moniri, Rebecca Penn, Giovanni Satta, Paul Randell, et al. Handheld point-of-care system for rapid detection of SARS-CoV-2 extracted RNA in under 20 min. *ACS central science*, 7(2):307–317, 2021.

[147] Kenny Malpartida-Cardenas, Luca Miglietta, Tianyi Peng, Ahmad Moniri, Alison Holmes, Pantelis Georgiou, and Jesus Rodriguez-Manzano. Single-channel digital LAMP

multiplexing using amplification curve analysis. *Sensors & Diagnostics*, 1(3):465–468, 2022.

[148] Fan Li, LQ Zhao, Jie Deng, RN Zhu, Yu Sun, LY Liu, YY Li, and Yuan Qian. Detecting human adenoviruses in respiratory samples collected from children with acute respiratory infections by loop-mediated isothermal amplification. *Zhonghua er ke za zhi= Chinese journal of pediatrics*, 51(1):52–57, 2013.

[149] Derong Dong, Wei Liu, Huan Li, Yufei Wang, Xinran Li, Dayang Zou, Zhan Yang, Simo Huang, Dongsheng Zhou, Liuyu Huang, et al. Survey and rapid detection of Klebsiella pneumoniae in clinical samples targeting the rcsA gene in Beijing, China. *Frontiers in microbiology*, 6:519, 2015.

[150] L McInnes, J Healy, and J Melville. arXiv e-prints. *arxiv*, 2018.

[151] Zhongheng Zhang. Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11), 2016.

[152] Karen Bush and Jed F Fisher. Epidemiological expansion, structural studies, and clinical challenges of new β-lactamases from gram-negative bacteria. *Annual review of microbiology*, 65:455–478, 2011.

[153] LS Tzouvelekis, A Markogiannakis, M Psichogiou, PT Tassios, and GL Daikos. Carbapenemases in Klebsiella pneumoniae and other Enterobacteriaceae: an evolving crisis of global dimensions. *Clinical microbiology reviews*, 25(4):682–707, 2012.

[154] Elizabeth A Neuner, Jun-Yen Yeh, Gerri S Hall, Jennifer Sekeres, Andrea Endimiani, Robert A Bonomo, Nabin K Shrestha, Thomas G Fraser, and David van Duin. Treatment and outcomes in carbapenem-resistant Klebsiella pneumoniae bloodstream infections. *Diagnostic microbiology and infectious disease*, 69(4):357–362, 2011.

[155] Mitchell J Schwaber and Yehuda Carmeli. An ongoing national intervention to contain the spread of carbapenem-resistant Enterobacteriaceae. *Clinical Infectious Diseases*, 58(5):697–703, 2014.

[156] Dina Bleumin, Matan J Cohen, Olivier Moranne, Vincent LM Esnault, Shmuel Benenson, Ora Paltiel, Keren Tzukert, Irit Mor-Yosef Levi, Iddo Z Ben-Dov, Ronen Levi,

et al. Carbapenem-resistant Klebsiella pneumoniae is associated with poor outcome in hemodialysis patients. *Journal of Infection*, 65(4):318–325, 2012.

[157] Anna Johnning, Nahid Karami, Erika Tång Hallbäck, Vilhelm Müller, Lena Nyberg, Mariana Buongermino Pereira, Callum Stewart, Tobias Ambjörnsson, Fredrik Westerlund, Ingegerd Adlerberth, et al. The resistomes of six carbapenem-resistant pathogens–a critical genotype–phenotype analysis. *Microbial genomics*, 4(11), 2018.

[158] Esmita Charani, Martin McKee, Raheelah Ahmad, Manica Balasegaram, Candice Bonaconsa, Gemma Buckland Merrett, Reinhard Busse, Vanessa Carter, Enrique Castro-Sanchez, Bryony D Franklin, et al. Optimising antimicrobial use in humans–review of current evidence and an interdisciplinary consensus on key priorities for research. *The Lancet Regional Health-Europe*, 7:100161, 2021.

[159] Francis S Codjoe and Eric S Donkor. Carbapenem resistance: a review. *Medical Sciences*, 6(1):1, 2017.

[160] Eoin Moloney, Kai Wai Lee, Dawn Craig, A Joy Allen, Sara Graziadio, Michael Power, and Carolyn Steeds. A PCR-based diagnostic testing strategy to identify carbapenemase-producing Enterobacteriaceae carriers upon admission to UK hospitals: early economic modelling to assess costs and consequences. *Diagnostic and prognostic research*, 3(1):1–9, 2019.

[161] Yasufumi Matsumura and Johann D Pitout. Recent advances in the laboratory detection of carbapenemase-producing Enterobacteriaceae. *Expert review of molecular diagnostics*, 16(7):783–794, 2016.

[162] Daniel Hussien Reta, Tesfaye Sisay Tessema, Addis Simachew Ashenef, Adey Feleke Desta, Wajana Lako Labisso, Solomon Tebeje Gizaw, Solomon Mequanente Abay, Daniel Seifu Melka, and Fisseha Alemu Reta. Molecular and immunological diagnostic techniques of medical viruses. *International Journal of Microbiology*, 2020, 2020.

[163] Jim F Huggett, Simon Cowen, and Carole A Foy. Considerations for digital PCR as an accurate molecular diagnostic tool. *Clinical chemistry*, 61(1):79–88, 2015.

[164] Gary S Collins and Karel GM Moons. Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181):1577–1579, 2019.

[165] Martina Beinhauerova, Vladimir Babak, Barbara Bertasi, Maria Beatrice Boniotti, and Petr Kralik. Utilization of digital PCR in quantity verification of plasmid standards used in quantitative PCR. *Frontiers in molecular biosciences*, 7:155, 2020.

[166] Florian P Maurer, Claudio Castelberg, Chantal Quiblier, Guido V Bloemberg, and Michael Hombach. Evaluation of carbapenemase screening and confirmation tests with Enterobacteriaceae and development of a practical diagnostic algorithm. *Journal of Clinical Microbiology*, 53(1):95–104, 2015.

[167] YJ Lim, HY Park, JY Lee, SH Kwak, MN Kim, H Sung, S-H Kim, and SH Choi. Clearance of carbapenemase-producing Enterobacteriaceae (CPE) carriage: a comparative study of NDM-1 and KPC CPE. *Clinical Microbiology and Infection*, 24(10):1104–e5, 2018.

[168] David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1):D13–D21, 2007.

[169] Nathaniel D Mercaldo, Kit F Lau, and Xiao H Zhou. Confidence intervals for predictive values with an emphasis to case–control studies. *Statistics in medicine*, 26(10):2170–2183, 2007.

[170] Toshihiro Nishizawa and Hidekazu Suzuki. Mechanisms of Helicobacter pylori antibiotic resistance and molecular testing. *Frontiers in molecular biosciences*, 1:19, 2014.

[171] Antti Vasala, Vesa P Hytönen, and Olli H Laitinen. Modern tools for rapid diagnostics of antimicrobial resistance. *Frontiers in Cellular and Infection Microbiology*, 10:308, 2020.

[172] Mark A Valasek and Joyce J Repa. The power of real-time PCR. *Advances in physiology education*, 29(3):151–159, 2005.

[173] SA Bustin, V Benes, T Nolan, and MW Pfaffl. Quantitative real-time RT-PCR–a perspective. *Journal of molecular endocrinology*, 34(3):597–601, 2005.

[174] Bernhard Kaltenboeck and Chengming Wang. Advances in real-time PCR: Application to clinical laboratory diagnostics. *Advances in clinical chemistry*, 40:219, 2005.

[175] Elfath M Elnifro, Ahmed M Ashshi, Robert J Cooper, and Paul E Klapper. Multiplex PCR: optimization and application in diagnostic virology. *Clinical microbiology reviews*, 13(4):559–570, 2000.

[176] P Markoulatos, N Siafakas, and M Moncany. Multiplex polymerase chain reaction: a practical approach. *Journal of clinical laboratory analysis*, 16(1):47–51, 2002.

[177] Abdullah AY Gibriel. Options available for labelling nucleic acid samples in DNA microarray-based detection methods. *Briefings in functional genomics*, 11(4):311–318, 2012.

[178] Jonas A Otoo and Travis S Schlappi. REASSURED Multiplex Diagnostics: A Critical Review and Forecast. *Biosensors*, 12(2):124, 2022.

[179] Jia Yao, Yuanyuan Luo, Zhiqi Zhang, Jinze Li, Chuanyu Li, Chao Li, Zhen Guo, Lirong Wang, Wei Zhang, Heming Zhao, et al. The development of real-time digital PCR technology using an improved data classification method. *Biosensors and Bioelectronics*, 199:113873, 2022.

[180] Michał Burdukiewicz, Andrej-Nikolai Spiess, Konstantin A Blagodatskikh, Werner Lehmann, Peter Schierack, and Stefan Rödiger. Algorithms for automated detection of hook effect-bearing amplification curves. *Biomolecular detection and quantification*, 16:1–4, 2018.

[181] Luca Miglietta, Yuwen Chen, Zhi Luo, Ke Xu, Ning Ding, Tianyi Peng, Ahmad Moniri, Louis Kreitmann, Miguel Cacho-Soblechero, Alison Holmes, et al. Smart-Plexer: a breakthrough workflow for hybrid development of multiplex PCR assays. *ResearchSquare*, 2022.

[182] Amelia L Markey, Stephan Mohr, and Philip JR Day. High-throughput droplet PCR. *Methods*, 50(4):277–281, 2010.

[183] Maja Sidstedt, Erica L Romsos, Ronny Hedell, Ricky Ansell, Carolyn R Steffen, Peter M Vallone, Peter Rådström, and Johannes Hedman. Accurate digital polymerase chain reaction quantification of challenging samples applying inhibitor-tolerant DNA polymerases. *Analytical chemistry*, 89(3):1642–1649, 2017.

[184] Thomas F Coleman and Yuying Li. On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Mathematical programming*, 67(1):189–224, 1994.

[185] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[186] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[187] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[188] Peter J Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

[189] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[190] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 157–166, 2005.

[191] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–452, 2008.

[192] Bing Sun, Jesus Rodriguez-Manzano, David A Selck, Eugenia Khorosheva, Mikhail A Karymov, and Rustem F Ismagilov. 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*, 53(31):8088–8092, 2014.

[193] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[194] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1(2):224–227, 1979.

[195] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[196] RG Rutledge. Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. *Nucleic acids research*, 32(22):e178–e178, 2004.

[197] Adrián Ruiz-Villalba, Elizabeth van Pelt-Verkuil, Quinn D Gunst, Jan M Ruijter, and Maurice JB van den Hoff. Amplification of nonspecific products in quantitative polymerase chain reactions (qPCR). *Biomolecular detection and quantification*, 14:7–18, 2017.

[198] Christian A Heid, Junko Stevens, Kenneth J Livak, and P Mickey Williams. Real time quantitative PCR. *Genome research*, 6(10):986–994, 1996.

[199] DV Rebrikov and D Yu Trofimov. Real-time PCR: a review of approaches to data analysis. *Applied biochemistry and microbiology*, 42(5):455–463, 2006.

[200] Thirumalaisamy P Velavan and Christian G Meyer. COVID-19: a PCR-defined pandemic. *International Journal of Infectious Diseases*, 103:278–279, 2021.

[201] Roman Wölfel, Victor M Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A Müller, Daniela Niemeyer, Terry C Jones, Patrick Vollmar, Camilla Rothe, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature*, 581(7809):465–469, 2020.

[202] Chianru Tan, Dongdong Fan, Nan Wang, Fang Wang, Bo Wang, Lingxiang Zhu, and Yong Guo. Applications of digital PCR in COVID-19 pandemic. *View*, 2(2):20200082, 2021.

[203] Dami A Collier, Sonny M Assennato, Ben Warne, Nyarie Sithole, Katherine Sharrocks, Allyson Ritchie, Pooja Ravji, Matthew Routledge, Dominic Sparkes, Jordan Skittrall, et al. Point of care nucleic acid testing for SARS-CoV-2 in hospitalized patients: a clinical validation trial and implementation study. *Cell Reports Medicine*, 1(5):100062, 2020.

[204] Olivier Vandenberg, Delphine Martiny, Olivier Rochas, Alex van Belkum, and Zisis Kozlakidis. Considerations for diagnostic COVID-19 tests. *Nature Reviews Microbiology*, 19(3):171–183, 2021.

[205] Anaïs Scohy, Ahalieyah Anantharajah, Monique Bodéus, Benoît Kabamba-Mukadi, Alexia Verroken, and Hector Rodriguez-Villalobos. Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis. *Journal of Clinical Virology*, 129:104455, 2020.

[206] K Sreenath, Priyam Batra, EV Vinayaraj, Ridhima Bhatia, KVP SaiKiran, Vishwajeet Singh, Sheetal Singh, Nishant Verma, Urvashi B Singh, Anant Mohan, et al. Coinfections with other respiratory pathogens among patients with COVID-19. *Microbiology spectrum*, 9(1):e00163–21, 2021.

[207] Karam Khaddour, Anna Sikora, Nayha Tahir, Daniel Nepomuceno, and Tian Huang. Case report: the importance of novel coronavirus disease (COVID-19) and coinfection with other respiratory pathogens in the current pandemic. *The American Journal of Tropical Medicine and Hygiene*, 102(6):1208, 2020.

[208] Hanliang Zhu, Haoqing Zhang, Sheng Ni, Marie Korabečná, Levent Yobas, and Pavel Neuzil. The vision of point-of-care PCR tests for the COVID-19 pandemic and beyond. *TrAC Trends in Analytical Chemistry*, 130:115984, 2020.

[209] James B Mahony, Gord Blackhouse, Jesse Babwah, Marek Smieja, Sonya Buracond, Sylvia Chong, William Ciccotelli, Tim O'Shea, Daifallah Alnakhli, May Griffiths-Turner, et al. Cost analysis of multiplex PCR testing for diagnosing respiratory virus infections. *Journal of clinical microbiology*, 47(9):2812–2817, 2009.

[210] Joan Barenfanger, Cheryl Drake, Nidia Leon, Tina Mueller, and Tammy Troutt. Clinical and financial benefits of rapid detection of respiratory viruses: an outcomes study. *Journal of clinical microbiology*, 38(8):2824–2828, 2000.

[211] Mary C Edwards and Richard A Gibbs. Multiplex PCR: advantages, development, and applications. *Genome Research*, 3(4):S65–S75, 1994.

[212] Nina G Xie, Michael X Wang, Ping Song, Shiqi Mao, Yifan Wang, Yuxia Yang, Junfeng Luo, Shengxiang Ren, and David Yu Zhang. Designing highly multiplex PCR primer sets

with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE). *Nature communications*, 13(1):1–10, 2022.

[213] Linda Strömqvist Meuzelaar, Owen Lancaster, J Paul Pasche, Guido Kopal, and Anthony J Brookes. MegaPlex PCR: a strategy for multiplex amplification. *Nature methods*, 4(10):835–837, 2007.

[214] Jared S Farrar and CT Wittwer. High-resolution melting curve analysis for molecular diagnostics. In *Molecular diagnostics*, pages 79–102. Elsevier, 2017.

[215] Qi Zhang, Feng Yang, Jie Gao, Weimin Zhang, and Xingang Xu. Development of multiplex TaqMan qPCR for simultaneous detection and differentiation of eight common swine viral and bacterial pathogens. *Brazilian Journal of Microbiology*, 53(1):359–368, 2022.

[216] Jongho Lee, Jichan Jang, Bongjoon Kim, Jeongho Kim, Gajin Jeong, and Hongui Han. Identification of Lactobacillus sakei and Lactobacillus curvatus by multiplex PCR-based restriction enzyme analysis. *Journal of Microbiological Methods*, 59(1):1–6, 2004.

[217] John Rachlin, Chunming Ding, Charles Cantor, and Simon Kasif. Computational tradeoffs in multiplex PCR assay design for SNP genotyping. *BMC genomics*, 6(1):1–11, 2005.

[218] Yuki Ozaki, Shingo Suzuki, Koichi Kashiwase, Atsuko Shigenari, Yuko Okudaira, Sayaka Ito, Anri Masuya, Fumihiro Azuma, Toshio Yabe, Satoko Morishima, et al. Cost-efficient multiplex PCR for routine genotyping of up to nine classical HLA loci in a single analytical run of multiple samples by next generation sequencing. *BMC genomics*, 16(1):1–12, 2015.

[219] Luca Miglietta, Ke Xu, Priya Chhaya, Louis Kreitmann, Kerri Hill-Cawthorne, Frances Bolt, Alison Holmes, Pantelis Georgiou, and Jesus Rodriguez-Manzano. Adaptive Filtering Framework to Remove Nonspecific and Low-Efficiency Reactions in Multiplex Digital PCR Based on Sigmoidal Trends. *Analytical Chemistry*, 94(41):14159–14168, 2022.

[220] Weibo Liu, Zidong Wang, Yuan Yuan, Nianyin Zeng, Kate Hone, and Xiaohui Liu. A novel sigmoid-function-based adaptive weighted particle swarm optimizer. *IEEE transactions on cybernetics*, 51(2):1085–1093, 2019.

[221] Joanna Ukalska and Szymon Jastrzebowski. Sigmoid growth curves, a new approach to study the dynamics of the epicotyl emergence of oak. *Folia Forestalia Polonica*, 61:30–41, 2019.

[222] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[223] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[224] Piet Bergveld. Development of an ion-sensitive solid-state device for neurophysiological measurements. *IEEE Transactions on biomedical engineering*, 1(1):70–71, 1970.

[225] Christofer Toumazou, Leila M Shepherd, Samuel C Reed, Ginny I Chen, Alpesh Patel, David M Garner, Chan-Ju A Wang, Chung-Pei Ou, Krishna Amin-Desai, Panteleimon Athanasiou, et al. Simultaneous DNA amplification and detection using a pH-sensing semiconductor system. *Nature methods*, 10(7):641–646, 2013.

[226] Ling-Shan Yu, Jesus Rodriguez-Manzano, Kenny Malpartida-Cardenas, Thomas Sewell, Oliver Bader, Darius Armstrong-James, Matthew C Fisher, and Pantelis Georgiou. Rapid and sensitive detection of azole-resistant Aspergillus fumigatus by tandem repeat loop-mediated isothermal amplification. *The Journal of Molecular Diagnostics*, 21(2):286–295, 2019.

[227] Nicolas Moser, Jesus Rodriguez-Manzano, Tor Sverre Lande, and Pantelis Georgiou. 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*, 12(2):390–401, 2018.

[228] Yinhua Zhang, Nelson Odiwuor, Jin Xiong, Luo Sun, Raphael Ohuru Nyaruaba, Hongping Wei, and Nathan A Tanner. Rapid molecular detection of SARS-CoV-2 (COVID-19) virus RNA using colorimetric LAMP. *MedRxiv*, 2020.

[229] David W Mount. Using the basic local alignment search tool (BLAST). *Cold Spring Harbor Protocols*, 2007(7):pdb–top17, 2007.

[230] CDC. Real-time rt-pcr for diagnosing covid-19. `https://www.fda.gov/media/134922/download`. Accessed: 2020.

[231] D Habgood-Coote, V Wright, Jethro Herberg, Jane Burns, Ulrich von Both, Enitan D Carrol, A Cunnington, Ronald de Groot, Marieke Emonts, Taco W Kuijpers, et al. MIN-

IMAL OPTIMISED HOST BLOOD GENE EXPRESSION SIGNATURE OF BACTERIAL INFECTION. In *38th Annual Meeting of the European Society for Paediatric Infectious Diseases (ESPID)*, 2020.

[232] Ivana Pennisi, Ahmad Moniri, Nicholas Miscourides, Luca Miglietta, Nicolas Moser, Dominic Habgood-Coote, Jethro A Herberg, Michael Levin, Myrsini Kaforou, Jesus Rodriguez-Manzano, et al. Discrimination of bacterial and viral infection using host-RNA signatures integrated in a lab-on-chip platform. *Biosensors and Bioelectronics*, 216:114633, 2022.

[233] Ho Kwong Li, Myrsini Kaforou, Jesus Rodriguez-Manzano, Samuel Channon-Wells, Ahmad Moniri, Dominic Habgood-Coote, Rishi K Gupta, Ewurabena A Mills, Dominique Arancon, Jessica Lin, et al. Discovery and validation of a three-gene signature to distinguish COVID-19 and other viral infections in emergency infectious disease presentations: a case-control and observational cohort study. *The Lancet Microbe*, 2(11):e594–e603, 2021.

[234] Daniela Guimarães, Rita Pissarra, Ana Reis-Melo, and Hercília Guimarães. Multisystem inflammatory syndrome in children (MISC): a systematic review. *International Journal of Clinical Practice*, 75(11):e14450, 2021.

[235] Emel Rothzerg, Xuan D Ho, Jiake Xu, David Wood, Aare Märtson, Katre Maasalu, and Sulev Koks. Alternative splicing of leptin receptor overlapping transcript in osteosarcoma. *Experimental Biology and Medicine*, 245(16):1437–1443, 2020.

[236] Primepcr sybr green assay: Leprot, human. `https://commerce.bio-rad.com/en-uk/prime-pcr-assays/assay/qhsaced0037872-primepcr-sybr-green-assay-leprot-human`. Accessed: 2022-12-07.

[237] David Dobnik, Dejan Štebih, Andrej Blejec, Dany Morisset, and Jana Žel. Multiplex quantification of four DNA targets in one reaction with Bio-Rad droplet digital PCR system for GMO detection. *Scientific reports*, 6(1):1–9, 2016.

[238] Robert R Kitchen, Mikael Kubista, and Ales Tichopad. Statistical aspects of quantitative real-time PCR experiment design. *Methods*, 50(4):231–236, 2010.

[239] Stephen A Bustin. Why the need for qPCR publication guidelines?—The case for MIQE. *Methods*, 50(4):217–226, 2010.

[240] Ye Mao, Ke Xu, Luca Miglietta, Louis Kreitmann, Nicolas Moser, Pantelis Georgiou, Alison Holmes, and Jesus Rodriguez-Manzano. Deep Domain Adaptation Enhances Amplification Curve Analysis for Single-Channel Multiplexing in Real-Time PCR. 2022.