

Imperial College of Science, Technology and Medicine
Department of Electrical & Electronic Engineering
Centre for Bio-Inspired Technology

Intelligent Algorithms for DNA Detection, Quantification and Multiplexing

Ahmad Moniri

A thesis submitted for the degree of
Doctor of Philosophy (Ph.D.)

January 2021

Statement of Originality

I declare that the work presented in this thesis is the outcome of my research under the supervision of Dr Pantelis Georgiou and Dr Jesus Rodriguez-Manzano. Ideas and results from the work of others are appropriately acknowledged with the standard referencing practices employed in this discipline. The material in this thesis has not been submitted for any degree at any other academic or professional institution.

Ahmad Moniri

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Ahmad Moniri

To my family and friends... My signals in the noise.

Abstract

Nucleic acid detection, referred herein as ‘DNA detection’, is defined as the process of identifying the presence of a specific sequence of bases, and has a vital, far-reaching impact on healthcare applications such as pathogen detection and drug resistance screening. In particular, DNA amplification methods, such as those used in popular COVID-19 tests, present a rapid and affordable means of DNA detection. When performing DNA amplification, there are 3 main goals which are desired: detection, quantification, and multiplexing. In an ideal world, we would like: (i) detection to be rapid and robust; (ii) quantification to be accurate and precise; and (iii) the ability to multiplex several targets in the same chemical reaction, without increasing the complexity of the chemistry or instrument. The aim of this thesis is to develop intelligent algorithms which can be used to improve DNA detection, quantification and multiplexing.

In an effort to achieve the aforementioned goals, several detection platforms have been developed which can be broadly classed as: quantitative, digital, and electronic instruments, corresponding to the 3 Parts in this thesis. This work exhibits a common philosophy throughout: increasing the value of data through algorithms that operate in higher dimensions, compared with existing methods which significantly reduce the dimensionality of data.

In Part I, the use of multiple features from quantitative instruments are investigated by analysing each signal in multidimensions, compared to the traditional unidimensional perspective, where each signal is reduced to a single value. To this end, a new framework is proposed that formalises the process of quantification and leverages the benefits of multiple features using a novel concept called the multidimensional standard curve (MSC). This new method is shown to guarantee enhanced quantification and provides a natural method of outlier detection, solely through analysing existing data from a new perspective - incurring little to no cost. Subsequently, this new framework is extended to perform multiplexing, demonstrating the first data-driven method for simultaneous quantification and multiplexing of nucleic acids in a single fluorescent channel. In particular, this method showed an accuracy of 100.0% (using 11 clinical isolates) when multiplexing 4 of the most prominent carbapenem-resistant genes, which account for 97.1% of confirmed cases in the UK.

In Part II, the use of machine learning methods for multiplexing is explored by considering the vast volume of data from digital instruments. To this end, a supervised machine learning classifier is trained to automatically perform a new form of data-driven multiplexing, referred to as Amplification Curve Analysis (ACA). This demonstrates the first application of machine learning for multiplexing in digital instruments based on amplification curves, leading to breaking the barrier of multiplexing one target per fluorescent channel and final intensity

value. Moreover, a mathematical formula based on Multivariate Poisson statistics is derived to theoretically demonstrate the trade-off between multiplexing and quantification, and support evaluating quantification precision in practical settings. Following this work, a novel 3-step machine learning system called Amplification and Melting Curve Analysis (AMCA) is proposed, which combines the kinetic information from amplification curves with thermodynamic information from melting curves - leading to the largest digital multiplex (using cheap intercalating dyes) that has been reported in the literature. In particular, over 100,000 amplification events were analysed, where this method showed an accuracy of 99.3% (for positive events) when multiplexing 9 variants of mobilised colistin resistance, representing a 10.0% increase over existing methods.

In Part III, the use of intelligent algorithms for increasing the robustness of electronic devices based on sensor arrays in highly noisy environments is explored. In particular, recent devices based on Ion-Sensitive Field-Effect Transistor (ISFET) sensors present a new form of data which is correlated in both space and time (similar to a video), and can produce a vast volume of rich data. To this end, a new algorithm is proposed based on modelling sensor drift using adaptive signal processing and clustering sensors based on their behaviour using unsupervised machine learning. This robust algorithm is shown to enable a high precision application such as distinguishing bacterial from viral infections using host signature-based classification. In particular, this new method results in robust values from 100.0% of experiments (N=22), compared with 63.6% from an existing method.

In summary, as inspired by the field of Signal Processing and Machine Learning, this thesis develops new intelligent algorithms for a wide range of platforms, by analysing data in higher dimensions. Through exploring the data from instruments that represent *past, present and future* diagnostic systems, this thesis lays the foundation for maximising the value of amplification data, which will ultimately lead to better informed decisions across all healthcare applications.

Acknowledgements

First and foremost I would like to thank my supervisors Dr. Jesus Rodriguez-Manzano and Dr. Pantelis Georgiou for granting me the opportunity to work on a range of exciting projects under their great leadership and alongside a supportive team. Your guidance, support and leadership will stay with me as a model example of how I should treat my fellow colleagues in my future career. The cutting-edge research we conduct at the Centre for Bio-Inspired Technology (CBIT) is truly inspirational, impactful & intellectually stimulating, and I am very fortunate to say that I am very happy - something which is seldom voiced by PhD students from other groups.

This work would not have come to fruition without the entire CBIT team due to the highly interdisciplinary and collaborative nature of my thesis. In particular, I would like to thank Nicolas Moser for his constructive suggestions and pearls of wisdom; Kenny Malpartida-Cardenas and Ivana Pennisi for their patience and vital contribution to my PhD; Luca Miglietta for his fresh perspective and inspiring ambitions; Dan Terracina and Miguel Cacho-Soblechero for their uplifting humour and interesting collaboration; and John Daniels for our intellectually stimulating discussions about artificial intelligence.

Moreover, I am grateful to my previous colleagues under the supervision of Prof. Danilo Mandic for shaping me into the academic that I am today. In particular, I would like to thank Giuseppe Calvi, Bruno Scalzo Dees and Ilia Kisil for our discussions on multidimensional data structures and machine intelligence, and David Joseph for his uplifting humour.

Finally, I would like to express my deep and sincere gratitude to all my friends and family for their continuous love and support. A special thank you to my close friend Shamim Ahmed for proofreading this thesis. And last but not least, I am forever indebted to my father, Prof. Mansour Moniri, my mother, Mehrnaz, and my brothers, Ali and Erfan, for their endless encouragement, inspiration and mentoring.

Contents

Abstract	4
Acknowledgements	6
List of Figures	11
List of Tables	14
List of Abbreviations	17
List of Publications	19
Chapter 1: Introduction	23
1.1 Motivation	23
1.2 Research Aim & Objectives	26
1.3 Thesis Structure and Contributions	27
Chapter 2: DNA Amplification: Quantitative, Digital and Electronic	31
2.1 Chapter Overview	31
2.2 Polymerase Chain Reaction (PCR)	32
2.3 End-Point PCR	33
2.4 Quantitative PCR (qPCR)	34
2.5 Digital PCR (dPCR)	36
2.6 Loop Mediated Isothermal Amplification (LAMP)	38
2.7 Electronic LAMP (eLAMP): Combining LAMP with Semiconductor Technology	40
2.8 Chapter Summary and Reflection	42

Part I Multidimensional Standard Curves for Quantitative Instruments

Chapter 3: Framework for Enhanced Quantification and Outlier Detection 45

3.1 Chapter Overview	45
3.2 Introduction	46
3.3 Experimental Section	48
3.4 Results and Discussion	54
3.5 Conclusion	59

Chapter 4: Simultaneous Single-Channel Multiplexing and Quantification 63

4.1 Chapter Overview	63
4.2 Introduction	64
4.3 Experimental Section	66
4.4 Results and Discussion	69
4.5 Conclusion	73

Part II Supervised Machine Learning for Digital Instruments

Chapter 5: Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR 77

5.1 Chapter Overview	77
5.2 Introduction	78
5.3 Experimental Section	80
5.4 Results and Discussion	82
5.5 Conclusion	94

Chapter 6: High-level Multiplexing in Digital PCR with Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis 97

6.1 Chapter Overview	97
6.2 Introduction	98
6.3 Experimental Section	100
6.4 Results and Discussion	104
6.5 Conclusion	109

Part III Intelligent Algorithms for Electronic Instruments

Chapter 7: Robust Time-to-Positive from ISFET Arrays 113

7.1 Chapter Overview	113
7.2 Introduction	114
7.3 Experimental Section	117
7.4 Results and Discussion	122
7.5 Conclusion	126

Chapter 8: Applications and Analysis of ISFET Arrays 129

8.1 Chapter Overview	129
8.2 COVID-19	129
8.3 <i>Aspergillus Fumigatus</i>	130
8.4 Mobilised Colistin Resistance	131
8.5 Dengue	132

Chapter 9: Conclusion & Future Perspective 133

9.1 Contributions	133
9.2 Future Perspective	134

Appendix	137
A Supporting Information: Chapter 3	137
B Supporting Information: Chapter 4	151
C Supporting Information: Chapter 5	163
D Supporting Information: Chapter 6	165
E Supporting Information: Chapter 7	171
References	176

List of Figures

Chapter 1: Introduction	23
1.1 Desired Outcomes of DNA Amplification	24
1.2 Overview of Quantitative, Digital and Electronic Instruments	25
1.3 Thesis Organisation	27
Chapter 2: DNA Amplification: Quantitative, Digital and Electronic	31
2.1 Timeline of DNA Amplification	31
2.2 Polymerase Chain Reaction (PCR)	32
2.3 Principles of Gel Electrophoresis	33
2.4 Quantification using Real-time PCR	34
2.5 Principles of Melting Curve Analysis	35
2.6 Principles of Digital PCR	37
2.7 Mechanism of Loop Mediated Isothermal Amplification (LAMP)	38
2.8 Formation of Dumbell Structure in LAMP	39
2.9 Classes of LAMP Products	39
2.10 Depiction of the Structural Difference between a MOSFET and ISFET	40
2.11 LaceWing: An In-house Diagnostic Platform developed by Moser et al.	41
Chapter 3: Framework for Enhanced Quantification and Outlier Detection	45
3.1 Depiction of Quantification Metrics	46
3.2 Illustration of Proposed and Conventional Framework for Quantification	48

List of Figures	12
3.3 Evaluating Quantification Performance of MSC versus Single-Feature Methods .	55
3.4 Multidimensional Standard Curve with Intentional Outliers	56
3.5 Multidimensional Analysis for Detecting Non-Specific Outliers	57
3.6 The Effect of Changing Annealing Temperature on Outlier Detection	58
3.7 Improved Quantification Through Outlier Detection	58
Chapter 4: Simultaneous Single-Channel Multiplexing and Quantification	63
4.1 Experimental Workflow for Single-Channel Multiplexing using qPCR	66
4.2 Amplification and Melting Curves for CPE Genes using Multiplex Assay	69
4.3 Multidimensional Standard Curves for CPE Genes	70
4.4 Average Mahalanobis Distance Between MSC and Samples	71
4.5 Visualisation of Mahalanobis Space for Training and Test Points	72
4.6 Evaluating Quantification Performance of MSC versus Single-Feature Methods .	72
Chapter 5: Data-driven Multiplexing using Real-Time Digital PCR	77
5.1 Experimental Workflow for Single-Channel Multiplexing using digital PCR . . .	79
5.2 Visualisation of 3-plex Assay Data in qPCR	84
5.3 Visualisation of 3-plex Assay Data in Real-time dPCR	86
5.4 Multiplexing Based on Final Fluorescent Intensity	87
5.5 Visualising the Similarity Between Amplification Curves using TSNE	88
5.6 Performance of ACA in the Presence of Single and Multiple Targets	90
5.7 The Impact of Co-amplification Events on Quantification	92
Chapter 6: High-level Multiplexing by Coupling Amplification and Melting Curves	97

6.1	Depiction of Amplification and Melting Curve Analysis	99
6.2	Flowchart of AMCA Data Processing Workflow	103
6.3	Analysis of Amplification and Melting Curves using qPCR and dPCR Instruments	105
6.4	Distribution of FFI Values for 9 <i>mcr</i> Targets	106
6.5	Performance of All Methods for Multiplexing 9 <i>mcr</i> Targets	107
Chapter 7: Robust Time-to-Positive from ISFET Arrays		113
7.1	Sources of Noise in ISFET Sensors - a Data Perspective	115
7.2	Sample-to-Result Workflow using a Lab-on-a-Chip Device in Conjunction with the Proposed Algorithm	116
7.3	Proposed Method for Extracting Time-to-Positive Values From ISFET Arrays . .	118
7.4	Viral Versus Bacterial Classification - Performance of Microarray Analysis	122
7.5	RT-qLAMP Standard Curves for IFI44L and EMR1	123
7.6	Viral Versus Bacterial Classification - Performance of RT-qLAMP and RT-eLAMP	124
7.7	Translation of 2-Gene Signature From Microarray Analysis to RT-qLAMP to RT- eLAMP - Moving from Expensive Lab-based Discovery to Cheap Point-of-Care Diagnostics	125
9.1	Publication Trend in Quantitative, Digital and Electronic Nucleic Acid Amplifi- cation	135

List of Tables

Chapter 4: Simultaneous Single-Channel Multiplexing and Quantification	63
4.1 Confirmed Cases of CPE in the UK Laboratories (2003-2015)	64
4.2 Primers Used in qPCR Multiplexing Study	67
4.3 Bacterial Isolates Used in qPCR Multiplexing Study	68
4.4 Instance of Framework for qPCR Multiplexing Study	68
Chapter 5: Data-driven Multiplexing using Real-Time Digital PCR	77
5.1 Primers Specification	81
5.2 Final Fluorescent Intensity Classification Performance	88
5.3 ACA Classification Performance	90
Chapter 6: High-level Multiplexing by Coupling Amplification and Melting Curves	97
6.1 Primer Sequences and Meta Data of 9 <i>mcr</i> Targets	101
Chapter 7: Robust Time-to-Positive from ISFET Arrays	113
7.1 Primer Sequences for IFI44L and EMR1-ADGRE	120

List of Abbreviations

NA Nucleic Acid

DNA Deoxyribonucleic Acid

RNA Ribonucleic Acid

PCR Polymerase Chain Reaction

qPCR Quantitative PCR

dPCR Digital PCR

ePCR Electronic PCR

LAMP Loop Mediated Isothermal Amplification

qLAMP Quantitative LAMP

dLAMP Digital LAMP

eLAMP Electronic LAMP

RT- Reverse Transcriptase- (This is a prefix, e.g. RT-PCR or RT-LAMP)

ISFET Ion-Sensitive Field-Effect Transistor

MOSFET Metal-Oxide-Semiconductor Field-Effect Transistor

CMOS Complementary Metal-Oxide-Semiconductor

LoC Lab-on-a-Chip

MSC Multidimensional Standard Curve

RE Relative Error

CV Coefficient of Variation

LOOCV Leave-one-out Cross Validation

ACA Amplification Curve Analysis

MCA Melting Curve Analysis

FFI	Final Fluorescent Intensity
AMCA	Amplification and Melting Curve Analysis
ATP	Average Then Process
TTP	Time-to-Positive
<i>bla</i>	Beta-lactamase (e.g. <i>bla</i> _{OXA-48})
<i>mcr</i>	Mobilised Colistin Resistance (This is a prefix, e.g. <i>mcr-1</i>)
CPE	Carbapenem-producing Enterobacteriaceae

List of Publications

Accepted/Published Peer-Reviewed Journals

- [J1] A. Moniri*, J. Rodriguez-Manzano*, K. Malpartida-Cardenas, L. Yu, X. Didelot, A. Holmes and P. Georgiou, “Framework for DNA Quantification and Outlier Detection Using Multi-dimensional Standard Curves”, *ACS Analytical Chemistry*, vol. 91, no. 11, pp. 7426-7434, May 2019. *These authors contributed equally.
- [J2] J. Rodriguez-Manzano*, A. Moniri*, K. Malpartida-Cardenas, J. Dronavalli, F. Davies, A. Holmes and P. Georgiou, “Simultaneous Single-Channel Multiplexing and Quantification of Carbapenem-Resistant Genes using Multidimensional Standard Curves”, *ACS Analytical Chemistry*, vol. 91, no. 3, pp. 2013-2020, Jan 2019. *These authors contributed equally.
- [J3] A. Moniri*, L. Miglietta*, K. Malpartida-Cardenas, I. Pennisi, M. Cacho-Soblechero, N. Moser, A. Holmes, P. Georgiou and J. Rodriguez-Manzano*, “Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR”, *ACS Analytical Chemistry*, vol. 92, no. 19, pp. 13134–13143, Sept 2020. *These authors contributed equally.
- [J4] A. Moniri*, L. Miglietta*, A. Holmes, P. Georgiou and J. Rodriguez-Manzano, “High-level Multiplexing in Digital PCR With Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis”, *ACS Analytical Chemistry*, vol. 92, no. 20, pp. 14181-14188, Sept 2020. *These authors contributed equally.
- [J5] I. Pennisi, J. Rodriguez-Manzano, A. Moniri, M. Kaforou, J.A. Herberg, M. Levin and P. Georgiou, “Translation of a host blood RNA Signature distinguishing bacterial from viral infection into a platform suitable for development as a point-of-care test”, *JAMA Paediatrics*, Jan 2021.
- [J6] J. Rodriguez-Manzano, K. Malpartida-Cardenas, N. Moser, I. Pennisi, M. Cavuto, L. Miglietta, A. Moniri, R. Penn, G. Satta, P. Randell, F. Davies, F. Bolt, W. Barclay, A. Holmes and P. Georgiou, “A handheld point-of-care system for rapid detection of SARS-CoV-2 in under 20 minutes”, *ACS Central Sciences*, vol. 7, no. 2, pp. 307-317, Jan 2021
- [J7] L-S. Yu, J. Rodriguez-Manzano, N. Moser, A. Moniri, K. Malpartida-Cardenas, N. Miskourides, T. Sewell, T. Kochina, A. Brackin, J. Rhodes, A.H. Holmes, M.C. Fisher and P.

- Georgiou “Rapid Detection of Azole-Resistant *Aspergillus fumigatus* in Clinical and Environmental Isolates by Use of a Lab-on-a-Chip Diagnostic System”, *Journal of Clinical Microbiology*, vol. 58, no. 11, Nov 2020.
- [J8] J. Rodriguez-Manzano, N. Moser, K. Malpartida-Cardenas, A. Moniri, L. Fisarova, I. Pennisi, A. Boonyasiri, E. Jauneikaite, A. Abdolrasouli, F. Bolt, F. Davies, X. Didelot, A. Holmes and P. Georgiou, “Rapid Detection of Mobilized colistin Resistance using a nucleic Acid Based Lab-on-a-chip Diagnostic System”, *Scientific Reports*, vol. 10, no. 1, pp 1-9, May 2020.
- [J9] A. Moniri*, D. Terracina*, J. Rodriguez-Manzano, P.H. Strutton and P. Georgiou, “Real-Time Forecasting of sEMG Features for Trunk Muscle Fatigue using Machine Learning”, *IEEE Transactions on Biomedical Engineering (TBME)*, Jul 2020. *These authors contributed equally.
- [J10] D.P. Mandic, S. Kanna, Y. Xia, A. Moniri, A. Junyent-Ferre and A.G. Constantinides, “A Data Analytics Perspective of Power Grid Analysis - Part 1: The Clarke and Related Transforms”, *IEEE Signal Processing Magazine*, vol. 36, no. 2, pp. 110-116, Feb 2019.
- [J11] S. Kanna, A. Moniri, Y. Xia, A.G. Constantinides and D.P. Mandic, “A Data Analytics Perspective of Power Grid Analysis - Part 2: Teaching Old Power Systems New Tricks”, *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 110-117, May 2019.

Submitted & Prepared Journals

- [J12] M. Cacho-Soblechero, A. Moniri, M. Douthwaite, J. Rodriguez-Manzano, P. Georgiou, “Long-Term Ion Imaging with ISFET Arrays and Programmable Real-Time Drift Compensation”, In preparation for *IEEE Sensors*.
- [J13] I. Pennisi*, A. Moniri*, N. Miscourides, N. Moser, M. Kaforou, J.A. Herberg, M. Levin, J. Rodriguez-Manzano, and P. Georgiou, (Title TBD) “Translation of a host blood RNA Signature distinguishing bacterial from viral infection into a platform suitable for development as a point-of-care test”, In preparation for *Biosensors and Bioelectronics*. *These authors contributed equally.
- [J14] H. Kwong Li*, M. Kaforou*, J. Rodriguez-Manzano*, S. Channon, A. Moniri, D. Coote, R. Gupta, E. Mills, D. Arancon, J. Lin, R. Mehta, N. Obaray, J. Chiu, L. Miglietta, I. Pennisi, P. Georgiou, M. Levin, L. Shallcross, A. Mentzer, G. Cooke, M. Noursadeghi and S. Sriskandan, “Discovery and validation of a 3-gene transcriptional signature to distinguish COVID-19 and other viral infections from bacterial sepsis in adults; a case-control then

observational cohort study”, Submitted to *Lancet Infectious Diseases*. *These authors contributed equally.

- [J15] N. Moser, L-S. Yu, J. Rodriguez-Manzano, K. Malpartida-Cardenas, A. Au, C. Cicatiello, A. Moniri, W-H Wang, S-F. Wang, Y-H. Chen and P. Georgiou, “Dengue diagnostics in the cloud: Quantitative detection using a smartphone-connected handheld Lab-on-Chip platform”, In preparation for *Biosensors and Bioelectronics*.
- [J16] G.G. Calvi, A. Moniri, M. Mahfouz, Z. Yu, Q. Zhao and D.P. Mandic, “Compression and Interpretability of Deep Neural Networks via Tucker Tensor Layer: From First Principles to Back-Propagation”, Submitted to *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*.

Published Peer-Reviewed Conference Proceedings

- [C1] D. Terracina*, A. Moniri*, J. Rodriguez-Manzano, P.H. Strutton and P. Georgiou, “Real-Time Forecasting and Classification of Trunk Muscle Fatigue using Surface Electromyography”, *IEEE Conference on Biomedical Circuits and Systems (BioCAS)*, Oct 2019. *These authors contributed equally.
- [C2] A. Moniri, A.G. Constantinides and D.P. Mandic, “Smart DSP for a Smarter Power Grid: Teaching Power System Analysis through Signal Processing”, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7883-7887, May 2019.
- [C3] I. Kisil, A. Moniri and D.P. Mandic, “Tensor Ensemble Learning for Multidimensional Data”, *IEEE Global Conference on Signal and Information Processing*, pp. 1358-1362, Nov 2018.
- [C4] A. Moniri, A.G. Constantinides and D.P. Mandic, “Refreshing DSP Courses through Bio-presence in the Curriculum: A Successful Paradigm”, *IEEE Conference on Digital Signal Processing (DSP)*, Nov 2018.

Published & Filed Patents

- [P1] P. Georgiou, A. Moniri and J. Rodriguez-Manzano, “A method for analysis of real-time amplification data”, WO2019234247A1 (Assignee: Imperial Innovations Limited). Priority 08-06-2018 / Filed 07-06-2019 / Published 12-12-2019.
- [P2] J. Rodriguez-Manzano, A. Moniri, L. Miglietta and P. Georgiou, “Identifying a target nucleic acid”, GB2013035.7 (Assignee: Imperial Innovations Limited). Priority 20-08-2020 / Filed 20-08-2020.

- [P3] P. Georgiou, A. Moniri, N. Moser and J. Rodriguez-Manzano, “Devices and method for detecting an amplification event”, WO2019234451A1 (Assignee: Imperial Innovations Limited). Priority 08-06-2018 / Filed 07-06-2019 / Published 12-12-2019.
- [P4] I. Pennisi, J. Rodriguez-Manzano, N. Miscourides, A. Moniri, N. Moser, D. Habgood-Coote, M. Kaforou, J. A. Herberg, M. Levin and P. Georgiou, “A method for determining a diagnostic outcome”, (Assignee: Imperial Innovations Limited). Priority 24-08-2020 / Filed 24-08-2020.

Total: 24

Chapter 1

“DNA is like a computer program but far, far more advanced than any software we’ve ever created”

Bill Gates

Introduction

1.1 Motivation

Every year, technological innovations surprise us with extraordinary advancements in improving our quality of life and life-expectancy. From Google’s self-driving cars to Amazon’s recommender systems, it is natural to ask ourselves: What distinguishes these great successes from the rest? The answer is simple – data. More specifically, the ability to convert the data that surrounds us into meaningful information, and subsequently, informed decisions. In particular, there is a surge of research in applying *data science* to healthcare and biomedical-related fields, largely due to advances in automated digital sensors [1, 2]. This includes the rise of DNA sequencing instruments, super-resolution digital microscopy, mass spectrometry and magnetic resonance imagery [3, 4, 5]. As sensing technology becomes more advanced, the data also becomes more rich and challenging (in terms of 4 V’s of Big Data: volume, variety, velocity and veracity). One field which has seen major advancements in sensing technology is *nucleic acid detection*, herein referred to as ‘DNA detection’ - the subject of this thesis.

Nucleic acids, which can be classed as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), are carriers of genetic information present in all living organisms. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences. Hence, detecting a specific sequence of bases, also termed as *DNA detection*, is vital for many important healthcare applications such as diagnosing patients with infectious diseases or detecting resistances to antibiotics (commonly known as Antimicrobial Resistance).

Broadly speaking, there are 3 classes of DNA detection: DNA amplification, microarray analysis and DNA sequencing. Both microarray analysis and DNA sequencing have lots of literature surrounding data science (commonly called bioinformatics) since they output large volumes of data such that data analysis is a necessary part of the technology [6, 7]. For example,

in microarray analysis, over 50,000 genes can be screened in one go, and lots of work has been done with regards to selecting the most relevant genes for a given application [8]. Similarly, DNA sequencing outputs large volumes of data through determining the sequence of over 6 million base pairs [9]. The major issue with these two approaches is the high monetary cost and time consumption. Therefore, these two methods are typically used for discovery and research applications, as opposed to medical diagnostics. On the contrary, DNA amplification methods are relatively cheap and fast, yet because of the ‘simplicity’ of the data, less data scientists are attracted to this field. This serves as a motivation for this thesis to explore the hidden information in DNA amplification data by analysing data in higher dimensions, and extending this knowledge to state-of-the-art sensing instruments which *need* complex data analytics.

The basic idea of DNA detection based on ‘DNA amplification’, or the replication of DNA, involves designing a chemical reaction such that amplification occurs, if and only if, the *target* DNA sequence is present in the analysed sample. Therefore, measuring the quantity of DNA before and after (or during) the designed chemical reaction can determine whether the target DNA exists or not. In reality, DNA is measured through different proxies such as fluorescence or pH (or equivalently H^+ ions) depending on the sensing instrument. If DNA amplification is triggered (i.e. the target DNA exists) then the reaction is defined as *positive*, otherwise the reaction is described as *negative*.

When performing DNA amplification, there are 3 main goals, as depicted in Figure 1.1, which are desired: detection, quantification and multiplexing. In an ideal world, we would like: (i) Detection to be rapid and robust (in terms of reproducibility, sensitivity and specificity). This has major implications on clinical outcomes, such as the accuracy of diagnostic tests. (ii) Quantification of the amount of target DNA to be accurate and precise. Similarly, this is important for applications which administer different treatments based on the quantity of target. And (iii)

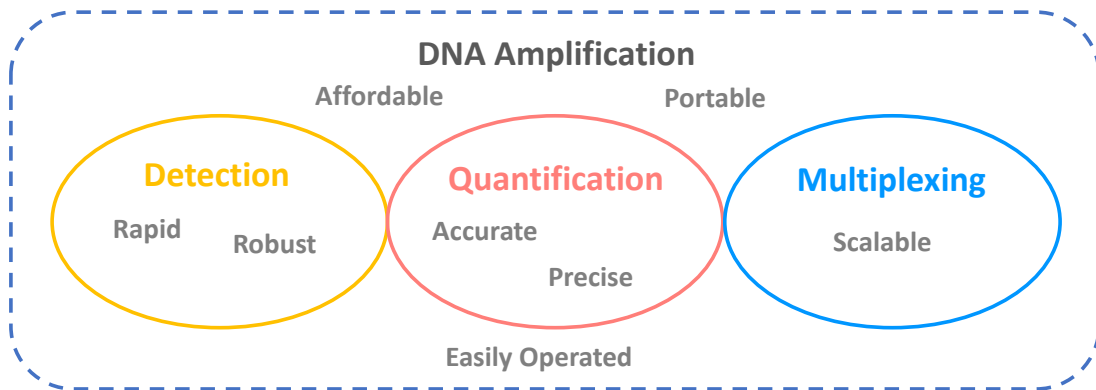


Figure 1.1: Desired Outcomes of DNA Amplification

Multiplexing, i.e simultaneously detecting several targets, to be scalable. The ability to multiplex without increasing the complexity of the chemistry or instrument has significant resource implications, reducing the cost of reagents, instruments and amount of sample required. Alongside the aforementioned points above, for practical and impactful DNA detection applications, especially at the point-of-care, sensing instruments need to be affordable and portable.

In an effort to achieve these goals, several amplification instruments have been developed over time. Figure 1.2 depicts 3 classes of instruments, namely the ‘gold standard’ quantitative polymerase chain reaction (qPCR), digital polymerase chain reaction (dPCR), and an electronic device named Lacewing (an in-house platform developed by our team at the Centre for Bio-Inspired Technology, Imperial College London). From a data perspective, quantitative instruments output a single time-series per chemical reaction (≈ 50 data points) which is typically reduced to a single value, denoted as C_q , to characterise the curve. Digital instruments typically output thousands of binary outputs which is characterised by the positive count, although with recent technology, the entire time-series for each independent output can be extracted (yielding $> 10^4$ data points). In contrast, the electronic platforms based on 2D arrays of sensors output a stream of images over time, similar to a video. Therefore the signal generated from electronic platforms is correlated in both space and time. The number of data points is dependant on the

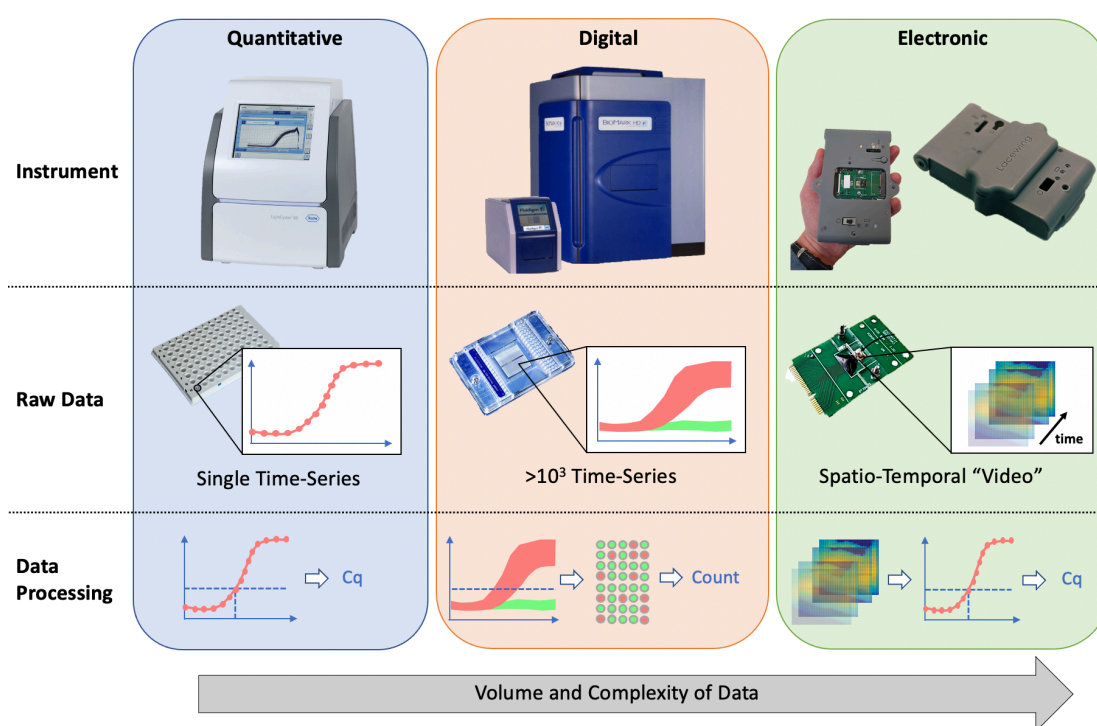


Figure 1.2: Overview of a quantitative instrument (Roche LightCycler[®] 96), digital instrument (Fluidigm BioMark HD) and state-of-the-art electronic platform (CBIT LaceWing)

sampling rate, where the number of data points can easily exceed 10^6 per reaction. Clearly, the complexity and volume of data, is increasing with more advanced sensing instruments. More interestingly, across all 3 strategies, the data is compressed into a low dimension, suggesting a potential loss of information. This motivates the use of signal processing and machine learning for enhancing the field of DNA detection by analysing data in higher dimensions, in order to squeeze as much value from the data as possible.

1.2 Research Aim & Objectives

The aim of this thesis is to investigate whether intelligent algorithms (signal processing and machine learning) can be used to enhance DNA detection, quantification and multiplexing through analysing data in higher dimensions. In doing so, general algorithms are developed to increase the value of existing data such that all applications of DNA amplification, such as detecting infectious diseases and antimicrobial resistance, can significantly benefit.

Since the aim of this work is broad and can be achieved through many different avenues, the following questions are the focus of this particular thesis:

- (i) *Quantitative*. Currently, the signal from qPCR is commonly reduced to a single feature. Can we enhance quantification of qPCR instruments by considering multiple features of the DNA amplification curve? Furthermore, does considering multiple features sufficiently increase the amount of information such that different targets can be distinguished, and therefore, multiplexed in a single reaction?
- (ii) *Digital*. Currently, analytics in dPCR is limited to considering whether each individual reaction is ‘positive’ or ‘negative’ based on simple thresholding methods. Can we use the entire amplification curve to extract more information, and consequently multiplex several targets? Moreover, with recent advances in sensing technology, does incorporating other sources of information (e.g. thermodynamic) increase multiplexing capabilities?
- (iii) *Electronic*. Currently, the output of semiconductor-based sensor arrays are averaged to yield a single time-series which is subsequently analysed. Can we use signal processing and machine learning techniques *before averaging* to make DNA detection using these devices more accurate and robust?

1.3 Thesis Structure and Contributions

This report is structured into 9 chapters with a common theme - increasing the value of nucleic acid amplification data. Chapter 1 gives a brief introduction into DNA detection and outlines the aims and objectives this thesis. This is followed by Chapter 2 which gives the interdisciplinary background necessary to understand the data-generating processes that are vital to grasp the technical content of this report. This is followed by the core technical contributions of this thesis which is separated into 3 parts, algorithms for: quantitative (Chapter 3 and 4), digital (Chapter 5 and 6), and electronic (Chapter 7 and 8) instruments. Given the broad contents of this thesis, each Part is self-contained (i.e. includes literature review and contribution to knowledge) such that the reader can skip to the part which interests them. The organisation of this thesis is depicted in Figure 1.3 and a short summary of each Chapter is as follows.

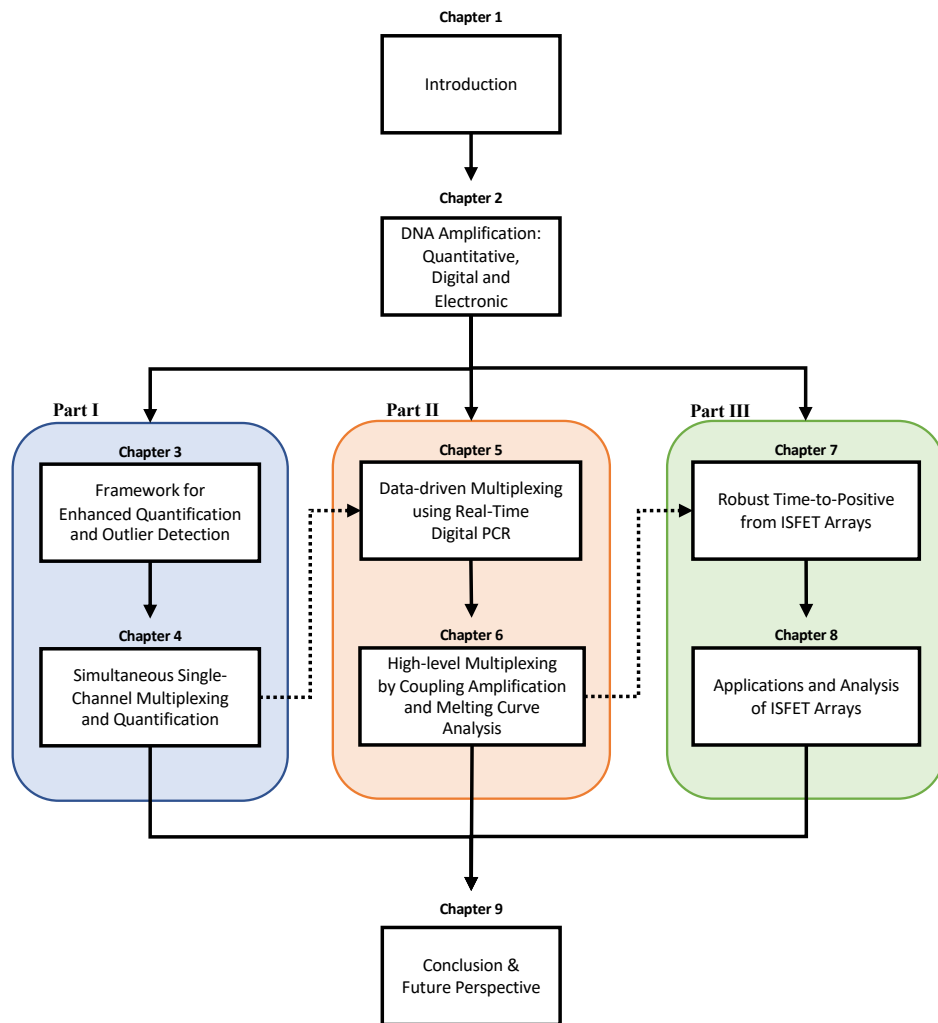


Figure 1.3: Thesis Organisation

Chapter 2: DNA Amplification: Quantitative, Digital and Electronic

This Chapter serves as an brief introduction to the field of DNA amplification, so as to familiarise readers from different backgrounds with the basics of this field. In particular, the history of DNA amplification is presented in chronological order, demonstrating the transition of research efforts from quantitative to digital to electronic instruments - corresponding to the 3 Parts of this thesis.

Part I: Multidimensional Standard Curves for Quantitative Instruments

Chapter 3: Framework for Enhanced Quantification and Outlier Detection

This Chapter investigates the benefits of extracting multiple features of the amplification curve for quantification and outlier detection, as opposed to characterising the entire amplification curve using a single value. In particular, a new framework is proposed to formalise the data processing workflow which involves constructing a novel concept called ‘multidimensional standard curves’. Moreover, a case study using synthetic DNA is implemented to show the effect of non-ideal conditions by varying reaction conditions.

Chapter 4: Simultaneous Single-Channel Multiplexing and Quantification

This Chapter extends the concepts of Chapter 3 to multiplexing, by constructing several multidimensional standard curves for various targets. It is shown that these objects do not overlap in multidimensional space, yielding a new method for simultaneous quantification and multiplexing in the same chemical reaction. In particular, this method is demonstrated on 4 of the most prominent antimicrobial resistant genes in the UK.

Part II: Supervised Machine Learning for Digital Instruments

Chapter 5: Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR

Following the idea that ‘more information exists in the amplification curve than previously anticipated’, this Chapter explores the use of machine learning algorithms to automatically learn how to distinguish various targets, as opposed to explicitly extracting features, by leveraging the vast quantity of data from real-time digital instruments. This demonstrates the first application of machine learning for multiplexing based on the kinetic information in amplification curves in a single digital panel and a single fluorescent channel.

Chapter 6: High-level Multiplexing in Digital PCR with Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis

This Chapter extends the aforementioned method by investigating the benefits of combining the kinetic information from amplification curves with thermodynamic information from melting curves, by proposing a new three step machine learning system. In terms of biological insight, this study indirectly tries to answer whether the non-mutual information contained in amplification and melting curves can have a significant impact on target identification. The proposed method, referred to as ‘amplification and melting curve analysis’, demonstrates the highest single-channel target identification in real-time digital PCR to-date - by multiplexing 9 clinically relevant antimicrobial resistant genes.

Part III: Intelligent Algorithms for Electronic Instruments

Chapter 7: Robust Time-to-Positive from ISFET Arrays

Although quantitative and digital instruments are highly related, electronic instruments present a new form of data that introduces its own benefits and challenges. To this end, this Chapter proposes an intelligent system to extract a time-to-positive from sensor arrays in highly noisy environments, by relaxing assumptions made by existing methods. In particular, this Chapter shows the first application of ISFET array based detection to RNA host transcriptomics, an area which requires high precision to be effective.

Chapter 8: Applications and Analysis of ISFET Arrays

This Chapter serves as a brief literature review of recent applications of electronic instruments that have arisen during the work of this thesis, to highlight the future direction of this field and the on-going research which will see a large benefit from incorporating more sophisticated data-driven methods.

Chapter 2

“Breakthrough innovation occurs when we bring down boundaries and encourage disciplines to learn from each other.”

Gyan Nagpal

DNA Amplification: Quantitative, Digital and Electronic

2.1 Chapter Overview

This Chapter aims to equip the reader with sufficient background knowledge on DNA amplification in order to understand the technical content in the thesis. More specifically, this chapter focuses on presenting the 3 methods which are used in the following chapters, namely, quantitative, digital and electronic. In order to keep this chapter concise, the relevant topics are presented in chronological order, as in Figure 2.1, to give the reader a sense of how the field of DNA amplification has developed and the future prospects it presents. Since the aim of this thesis is oriented around data processing and algorithms, this Chapter does not delve into the finer details of the multidisciplinary background, but rather serves as a brief summary of the area. For the interested reader, references are provided throughout the text for more detailed content.

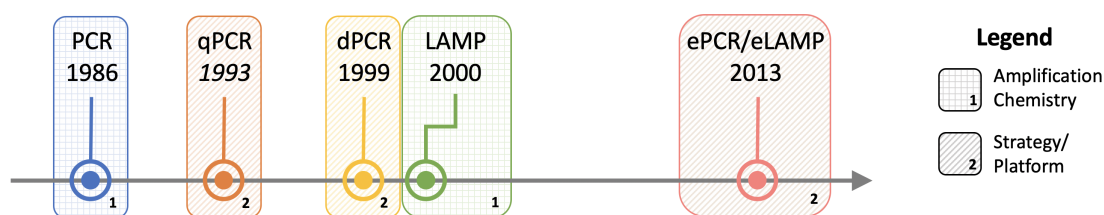


Figure 2.1: Timeline of DNA amplification techniques. PCR - Polymerase chain reaction, qPCR - quantitative PCR, dPCR - digital PCR, LAMP - loop-mediated isothermal amplification, ePCR - electronic PCR and eLAMP - electronic LAMP.

2.2 Polymerase Chain Reaction (PCR)

Polymerase chain reaction (PCR), invented in 1986 by Katy B. Mullis, is a widely used method in molecular biology for detecting DNA sequences [10]. Methods prior to this seminal work often consumed days of work, whereas PCR introduced the first DNA amplification method that was able to provide results in hours. It is a highly sensitive method that requires minimal template for the detection/amplification of specific sequences. Common applications of PCR include: amplifying DNA for subsequent experimental use, genetic testing and detecting pathogens.

For PCR, the amplification reaction consist of: the sample(s), DNA polymerase, magnesium, nucleotides and primers. Furthermore, the process of PCR requires thermal management, therefore an instrument capable of efficient heating/cooling is required. The PCR mechanism consists of three steps: 1) denaturation, where double-stranded DNA (dsDNA) is heated into two single DNA strands, 2) annealing, where primers align to the single DNA strands and 3) elongation, where the primers are extended by DNA polymerase, resulting in two copies of the original DNA strand. The denaturation, annealing, and elongation process is referred to as one cycle of amplification, where each step is typically optimised for the template and primer set used. This cycle is repeated approximately 20-50 times resulting in a high concentration of target DNA template, as depicted in Figure 2.2. As PCR is highly sensitive and is applicable to small volumes, it is very common to conduct technical replicates of each reaction in order to average out variations in reagents.

It is important to note the difference between DNA and RNA detection. Standard PCR is used for detecting DNA. In order to detect RNA, a reverse transcription step should precede PCR, to give rise to a method called reverse transcription PCR, or RT-PCR [11, 12]. Therefore, all methods in this thesis apply to both DNA *and* RNA.

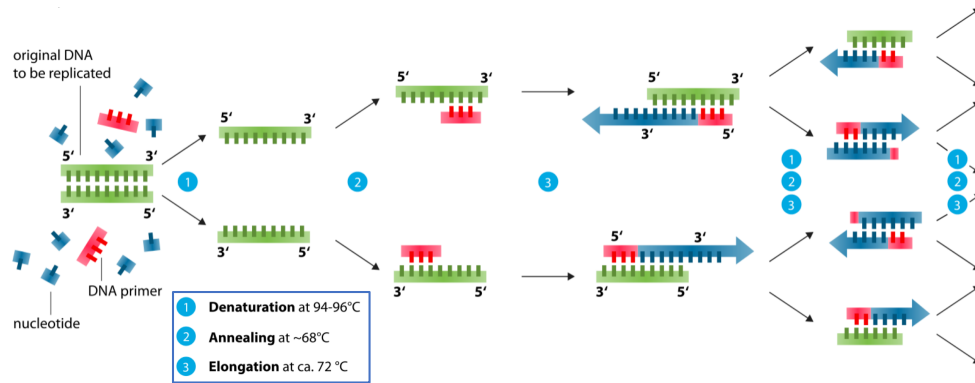


Figure 2.2: Polymerase Chain Reaction (PCR) [13]

2.3 End-Point PCR

The most basic form of PCR, referred to as end-point PCR, consists of taking a single measurement after conducting a certain number of PCR cycles. Following the reaction, a fluorescent dye is added to the solution which binds to any dsDNA. Although a high fluorescence may indicate a positive reaction, it is vital to confirm the amplification product to ensure a true positive. The most common method of determining specificity involves running a gel electrophoresis. This method essentially checks if the amplification product is the expected size (in bp). This process involves: (1) (optionally) cutting up the DNA from the amplification product using restriction enzymes, (2) loading the fragments into agarose gel, (3) and applying an electric field across the gel. Since DNA is inherently negatively charged, the fragments will move towards the positive electrode - with shorter fragments moving further than longer fragments (due to the molecular weight). If known DNA fragment sizes are also ran in parallel, commonly referred to as a *DNA ladder* and illustrated in 2.3 (A, B), then it is possible to identify if the amplification product has a high concentration of the desired amplicon size. Aside from the size of amplicon, further characterisation can be performed to detect a *specific* sequence of DNA using southern blotting (or northern blotting for RNA) [14]. It is important to note that end-point PCR is a simple tool for detection, however it is not able to absolutely quantify the amount of initial DNA copies. Some methods have been proposed, by running gel electrophoresis with varying known quantities of DNA templates, in order to use the fluorescent intensity as an indication for quantification. However, these methods are known to be inaccurate and are, therefore, seldom used and deemed to be ‘semi-quantitative’.

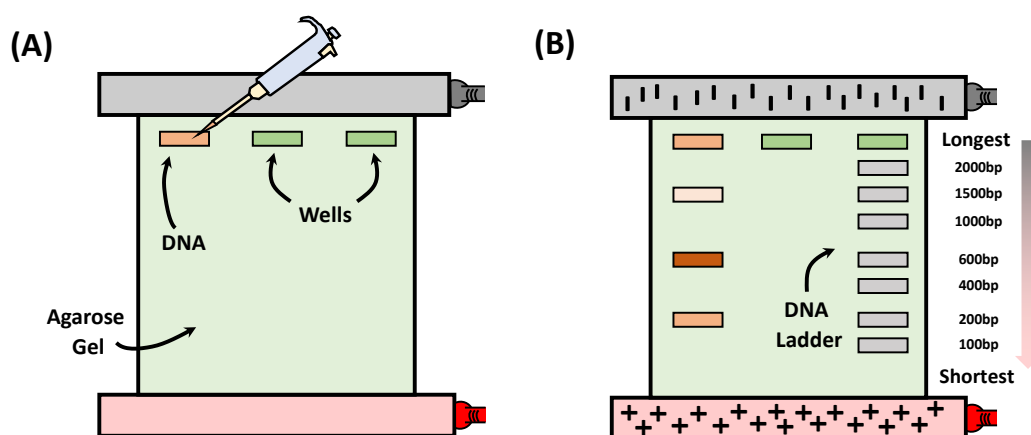


Figure 2.3: Principles of Gel Electrophoresis

2.4 Quantitative PCR (qPCR)

Quantitative PCR (qPCR), or RT-qPCR, is used to detect, characterise and quantify nucleic acids for numerous applications. As first described by Higuchi et al. (1993), the principle of qPCR is based on the idea that “the fewer cycles necessary to produce a detectable fluorescence, the greater the number of target sequences” [15]. Similar to end-point PCR, DNA is amplified by 3 repeating steps: denaturation, annealing and elongation. However, in qPCR, fluorescent labelling enables the collection of data in real-time. Hence, qPCR is commonly referred to as real-time PCR. The series of measurements over time (or cycles) is commonly known as the *amplification curve*, as visualised in Figure 2.4 (A). The amplification curve, like many natural processes, follows a sigmoidal shape and has 4 regions of interest: baseline, exponential, linear and plateau. More importantly, it is possible to quantify the amount of initial target DNA concentration using a standard curve, also referred to as a calibration curve. This is because the time-to-positive (TTP), defined as the cycle number where there is a detectable increase in fluorescence, is linearly proportional to the logarithm of the initial template concentration. Figure 2.4 (B) and (C) shows how to build a standard curve using a basic thresholding method, called the *C_t* method, in order to quantify unknown initial concentrations of target DNA [16, 17].

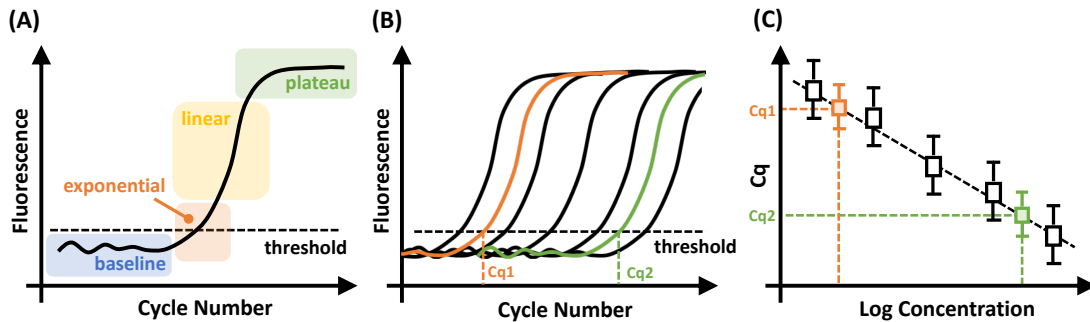


Figure 2.4: Quantification using Real-time PCR

Although qPCR can achieve quantification, it does not guarantee specific amplification of the target DNA. In other words, a positive reaction may be due to ‘undesired’ amplification. The method used to confirm the amplification product is related to the type of fluorescent labelling used: (1) dye-based and (2) probe-based.

1. In dye-based qPCR, fluorescent labelling allows the quantification of the amplified DNA molecules by employing the use of a dsDNA binding dye. The disadvantage of dye-based qPCR is that the dye will bind to any dsDNA and therefore only one target is typically examined at a time. Furthermore, it is possible that the signal is not specific to the target template, and is due

to multiple amplification products. Therefore, it is common to perform post-processing of the amplification product to confirm the specificity; through extra time-consuming methods such as agarose gel electrophoresis or melting curve analysis. When performing dye-based PCR, and since qPCR instruments have the capability for real-time measurements and thermal management, it is usually convenient to use melting curve analysis [18]. Contrary to gel electrophoresis, this method uses thermodynamic properties (e.g. GC content as well as amplicon length) as a proxy for specificity. As illustrated in Figure 2.5 (A) and (B), as temperature increases, double stranded DNA is dissociated into single-stranded DNA, causing a decrease in fluorescence. Plotting the negative derivative of fluorescence with respect to temperature, as shown in Figure 2.5 (C), will show a clear peak indicating a single amplification product. Multiple peaks may occur if the experiment contains more than 1 target, is contaminated, or due to unintended by-product from primers interacting with each other (called primer dimer).

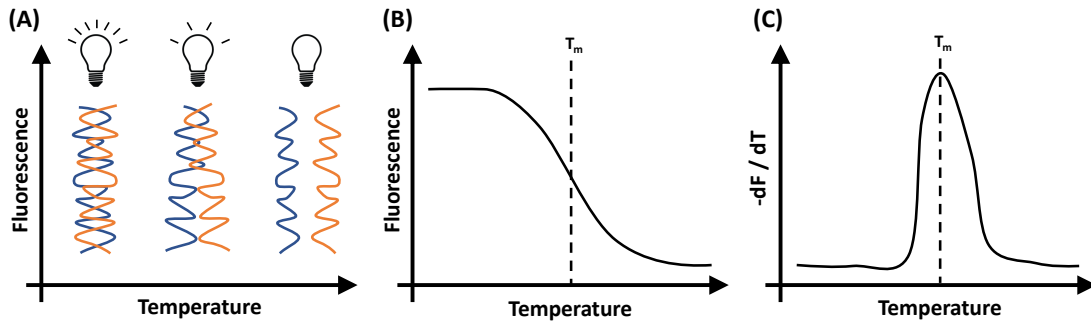


Figure 2.5: Principles of Melting Curve Analysis

2. In probe-based qPCR, fluorescent molecules are only released when specific amplification takes place. There are many types of probes available, however the most common is a hydrolysis probe (such as Taqman) which consists of a fluorophore and a quencher [19]. While the probe is intact, the fluorescence resonance energy prevents the emission of the fluorophore due to the quencher. During the PCR reaction, the probe is hydrolysed during primer extension and the fluorophore separates from the quencher, resulting in amplification-dependant increase in fluorescence. Thus, the signal from a probe-based qPCR reaction is proportional to the amount of the probe target sequence present in the sample. Since the amplification is specific, multiple fluorescent channels (i.e. colours) can also be used to multiplex several target templates.

Amplification Efficiency

In theory, the number of nucleic acids should double after every cycle. However, in practise, the number of molecules increases by a factor of $(1 + \eta)$, where $\eta \in [0, 1]$ [20, 21]. The quantity η , also referred to as *amplification efficiency*, is a key measure of how ideal the reaction is.

Theoretically, the number of molecules at cycle C is given by: $N_C = N_0(1 + \eta)^C$. Taking logarithms of both sides and rearranging yields,

$$\log(N_C) = \log\left(N_0(1 + \eta)^C\right) \quad (2.1)$$

$$= \log(N_0) + C \cdot \log(1 + \eta) \quad (2.2)$$

$$C = -\frac{1}{\log(1 + \eta)} \cdot \log(N_0) + \log\left(\frac{N_C}{1 + \eta}\right) \quad (2.3)$$

Assuming N_C and η are constants, the above equation is exactly the standard curve used for quantification. Therefore, η can be estimated from the slope of the standard curve. For example, an ideal amplification efficiency, $\eta = 1$, results in a slope equal to $-\frac{1}{\log(2)} \approx -3.3$. Although it is common to report a single value for η across a set of reactions [22, 23, 24], the efficiency changes between reactions and even cycle-to-cycle. This means that quantification can only be performed if the ‘test’ sample has an efficiency similar to the those used to construct the standard curve. Therefore, methods apart from the cycle-threshold method have been proposed to improve quantification under these circumstances, notably the C_{y0} method [25, 26].

2.5 Digital PCR (dPCR)

More recently, digital PCR was introduced as an alternative to qPCR, in order to quantify samples without the need for a standard curve [27, 28]. The first use of the term ‘digital PCR’ was used by Vogelstein et al. (1999), although the concept was initially explored in earlier years [29, 30, 31]. Digital PCR achieves quantification by partitioning the sample into many independent PCR reactions such that each ‘well’ only contains a single DNA or RNA molecule. Subsequently, the DNA/RNA is amplified using PCR and quantification can be simply performed by counting. The workflow of dPCR is depicted in Figure 2.6. The rise of dPCR in recent years can be attributed to the advances in efficient partitioning methods in the field of microfluidics. Note: dPCR is sometimes referred to conducting ‘single-molecule’ reactions (amplifying 1 molecule), as opposed to qPCR which conducts ‘bulk’ reactions (amplifying >1 molecule).

In practise, it is very difficult to isolate a single DNA molecule in every partition. It is typically assumed that each molecule has an equal chance of occupying any given partition. Therefore, the statistical foundations of quantification using dPCR is governed by the Binomial and Poisson distribution. This gives dPCR a theoretical advantage over qPCR since the accuracy of the quantification can also be approximated [32]. The average number of molecules per

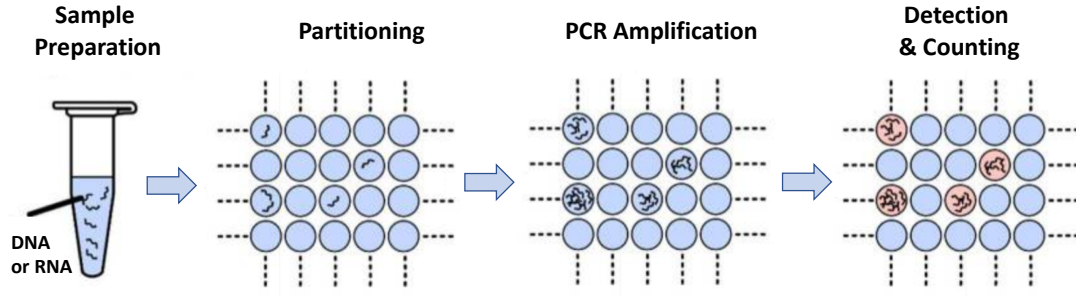


Figure 2.6: Principles of Digital PCR [28]

partition (or well) is denoted by λ and depends on the number of molecules in the sample, n , and the total number of wells, m , or equivalently, on the sample concentration, C , and the volume in each well, V_d . That is,

$$\lambda = \frac{n}{m} = C \cdot V_d \quad (2.4)$$

The binomial distribution is a discrete probability distribution that describes the likelihood of k successes in n trials, each with constant probability p . For example, when flipping a fair coin, the binomial distribution can estimate the probability of flipping exactly 4 heads out of 10 trials. In the context of dPCR, flipping the coin is analogous to “flipping” a molecule randomly into a panel of partitions, and a success is when the target falls into partition of interest. Assuming that the likelihood of a molecule falling within a partition is uniform, i.e. $\frac{1}{m}$, the probability of that a partition contains k molecules (i.e. successes) using the binomial distribution is given by:

$$p(k) = \binom{n}{k} \left(\frac{1}{m}\right)^k \left(1 - \frac{1}{m}\right)^{n-k}, \text{ where } \binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (2.5)$$

If m is large, then the probability of a successful trial is very low, and therefore the binomial distribution can be approximated using a Poisson distribution. That is,

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (2.6)$$

The mean and variance of the Poisson distribution are both equal to λ . Furthermore, the probability of an empty well, i.e. $p(0)$, is equal to $e^{-\lambda}$. Therefore, λ can be directly estimated from the proportion of empty partitions, E , using the equation: $\lambda = -\ln(E)$. Now, it is trivial to estimate the initial concentration or number of DNA molecules:

$$C = \frac{\lambda}{V_d} = -\frac{\ln(E)}{V_d} \quad (2.7)$$

$$n = m \cdot \lambda = -m \cdot \ln(E) \quad (2.8)$$

primer (BIP). (2) The target DNA strand is displaced by FIP. (3) A DNA polymerase is used to extend the primer and separate the DNA duplex. (4) Subsequently, the FOP anneals to an upstream target region. (5) As the strand is displaced, a new product is formed which contains a self-hybridizing loop. (6/7/8/9) A similar process occurs from the 3' side using the BIP/BOP to construct a short dumbbell structure.

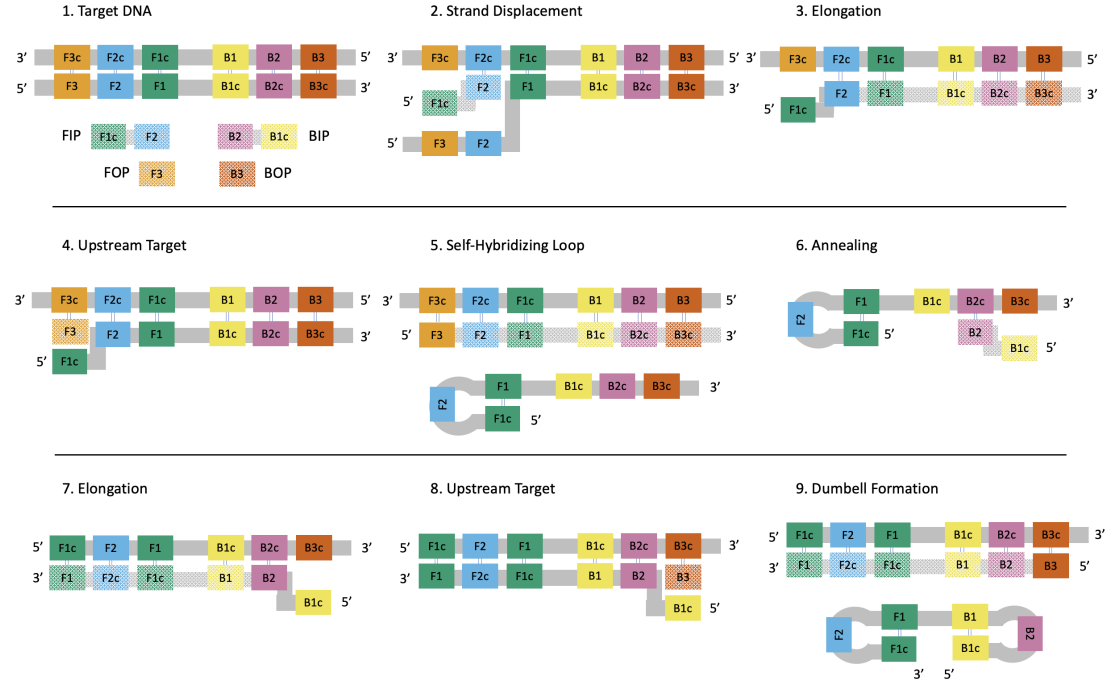


Figure 2.8: Formation of dumbbell structure in LAMP

Once the dumbbell is formed, there are multiple sites for initiation. Figure 2.9 shows the 4 classes of products which can be formed as a result of LAMP amplification. For more information on the kinetics of these products and how they are formed, please see the stoichiometric and pseudo kinetic model of LAMP recently proposed by Kaur et al. (2020) [40].

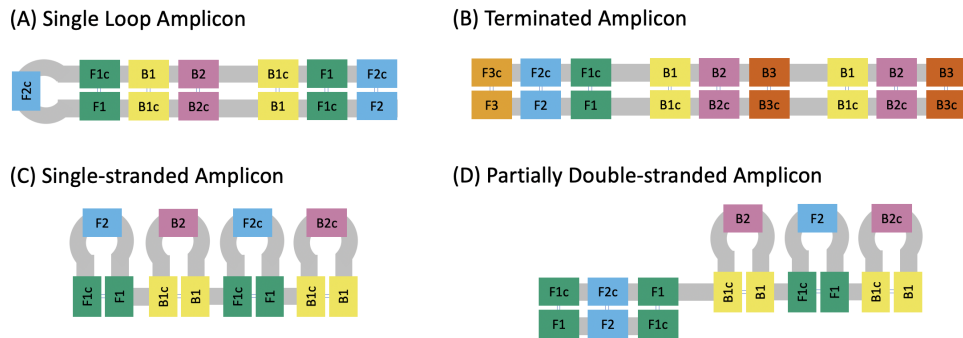


Figure 2.9: Classes of LAMP products

2.7 Electronic LAMP (eLAMP):

Combining LAMP with Semiconductor Technology

Nucleic acid amplification naturally releases byproducts such as protons (H^+) and pyrophosphate ions from the synthesis of the complementary chain by DNA polymerase. In PCR, insufficient amplification product is produced in order to detect these products - hence the need for fluorescent based sensing. However, LAMP produces > 50 fold more amplicons than PCR, enabling H^+ , or equivalently pH, as a proxy for amplification [35, 41]. Therefore, LAMP enables the use of pH-based sensing devices (as described in the following section), alleviating the need for bulky fluorescent instrumentation.

2.7.1 The Ion-Sensitive Field-Effect Transistor (ISFET)

The ion-sensitive field-effect transistor (ISFET) is a chemically sensitive transistor that was first introduced in 1970 by Bergveld [42]. It can be tailored to different chemicals by depositing an ion-selective membrane on the gate. Insulators, such as Aluminium Oxide (Al_2O_3), Silicon Nitride (Si_3N_4) and Silicon Dioxide (SiO_2), make the transistor sensitive to pH, and therefore, this makes the ISFET useful for DNA detection (using LAMP). Figure 2.10 shows a visual representation of the ISFET, compared with conventional metal-oxide-semiconductor field-effect transistors (MOSFET).

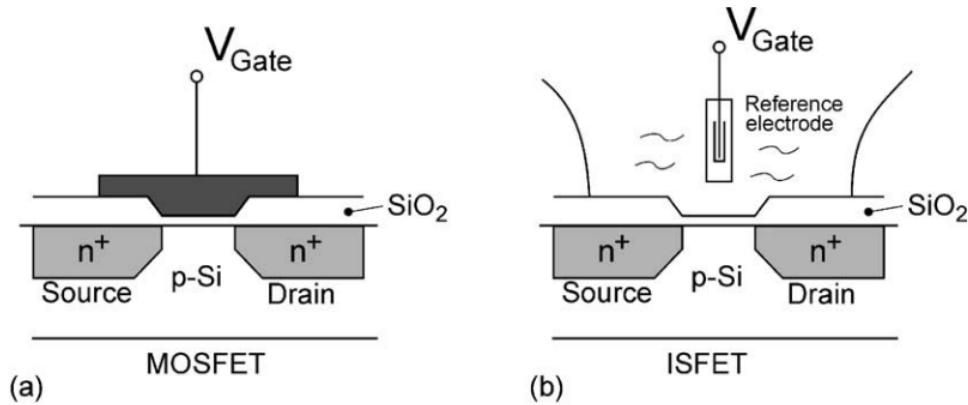


Figure 2.10: Depiction of the structural difference between a MOSFET and ISFET [43]

Given that the gate of the ISFET is essentially made of the chemical solution (and reference electrode), properties of the ISFET (e.g. threshold voltage) will be sensitive to pH fluctuations,

i.e. the number of protons released. Several circuits and modes of operation exist in the literature in order to generate signals proportional to pH [44].

2.7.2 ISFET Arrays for DNA Detection

Using a single ISFET sensor that is sensitive to pH limits the full potential of DNA detection, especially from a data perspective. Constructing an array of ISFETs allows for a chemical sensing platform where multiple realisations of the raw signal can be analysed as a function of time and space [44]. Given the output of the array is essentially an ‘image’ of pH values, a single ISFET ‘sensor’ can be interchangeably referred to as a ‘pixel’.

2.7.3 LaceWing: In-house Diagnostic Platform

Recently, a fully portable diagnostic platform using an ISFET array called ‘LaceWing’ has been proposed by Moser et al. (2018) [45]. This connected and affordable smartphone-based platform for DNA detection at the point-of-need offers a promising alternative to conventional laboratory-based instruments. The platform performs eLAMP on a disposable cartridge mounted onto a printed circuit board, with a smartphone application for data acquisition, visualisation and cloud connectivity, as illustrated in Figure 2.11. The ISFET array consists of 78×56 sensors, among which 3874 are chemical and 494 are temperature sensors. This technology has been demonstrated for several applications such as detecting pathogens and antimicrobial resistances, and will be the device used to generate data in Chapter 7 [46, 47].

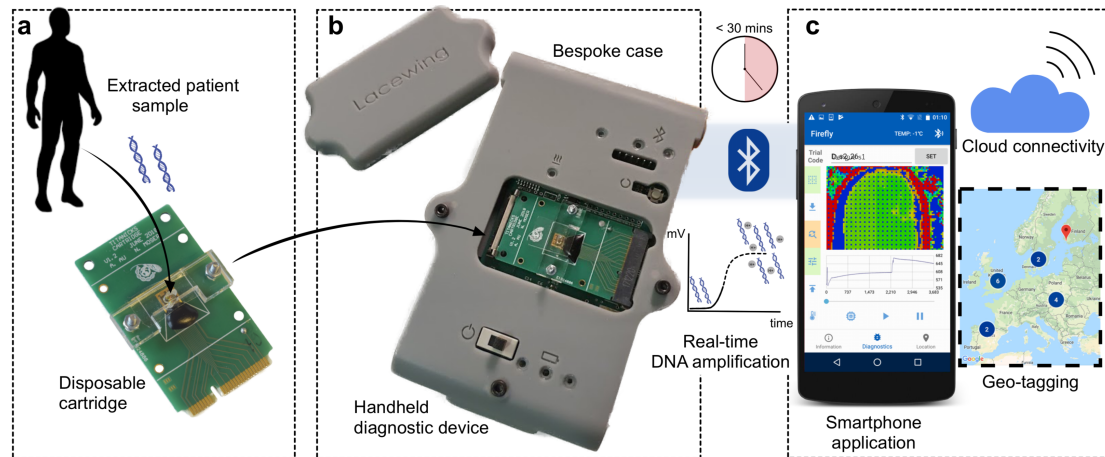


Figure 2.11: LaceWing: An In-house Diagnostic Platform developed by Moser et al. [45]

2.8 Chapter Summary and Reflection

In summary, this Chapter briefly reviewed the history of particular DNA amplification chemistries (PCR and LAMP) and relevant instrumentation strategies (end-point, quantitative, digital and electronic). To date, the focus of many researchers has been on the development of the chemistries and instrumentation. However, there are gaps in the literature regarding data-processing algorithms; especially to leverage the increasing trend in the volume and complexity of data. Therefore, this serves as a motivation for this thesis to delve deeper into this area in order to maximise the value of data across all 3 strategies: quantitative, digital and electronic.

Part I:

Multidimensional Standard Curves for Quantitative Instruments

“You can have data without information, but you cannot
have information without data.”

DANIEL KEYS MORAN

Chapter 3

Framework for Enhanced Quantification and Outlier Detection

3.1 Chapter Overview

Real-time PCR is a highly sensitive and powerful technology for the quantification of DNA and has become the method of choice in microbiology, bioengineering and molecular biology. Currently, the analysis of real-time PCR data is hampered by only considering a single feature of the amplification profile to generate a standard curve. The current “gold standard” is the cycle-threshold (C_t) method which is known to provide poor quantification under inconsistent reaction efficiencies. Multiple single-feature methods have been developed to overcome the limitations of the C_t method, however, there is an unexplored area of combining multiple features in order to benefit from their joint information. Here, a novel framework is proposed that combines existing standard curve methods into a *multidimensional standard curve*. This is achieved by considering multiple features together such that each amplification curve is viewed as a point in a multidimensional space. Contrary to only considering a single-feature, in the multidimensional space, data points do not fall exactly on the standard curve, which enables a similarity measure between amplification curves based on distances between data points. This framework is shown to expand the capabilities of standard curves in order to: optimise quantification performance, provide a measure of how suitable an amplification curve is for a standard, and thus automatically detect outliers and increase the reliability of quantification. This work aims to provide an affordable solution to enhance existing diagnostic settings through maximising the amount of information extracted from conventional instruments.

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

- A. Moniri*, J. Rodriguez-Manzano*, K. Malpartida-Cardenas, L. Yu, X. Didelot, A. Holmes and P. Georgiou, “Framework for DNA Quantification and Outlier Detection Using Multidimensional Standard Curves”, *ACS Analytical Chemistry*, vol. 91, no. 11, pp. 7426-7434, May 2019. *These authors contributed equally.

- P. Georgiou, A. Moniri and J. Rodriguez-Manzano, “A method for analysis of real-time amplification data”, WO2019234247A1 (Assignee: Imperial Innovations Limited). Priority 08-06-2018 / Filed 07-06-2019 / Published 12-12-2019.

3.2 Introduction

The real-time polymerase chain reaction (qPCR) has become a routine technique in microbiology, bioengineering and molecular biology for detecting and quantifying nucleic acids [15, 48, 49]. This is predominantly due to its large dynamic range (7-8 magnitudes), desirable sensitivity (5-10 molecules per reaction) and reproducible quantification results [50, 51, 52]. New methods to improve the analysis of qPCR data are invaluable to a number of application fields, including environmental monitoring and clinical diagnostics [53, 54, 55, 56].

The current “gold standard” for absolute quantification of DNA (or RNA if preceded by a reverse transcription step) using standard curves is the cycle-threshold (C_t) method [57, 16, 17]. The C_t value is a feature of the amplification curve defined as the cycle number in the exponential region from which there is a detectable increase in fluorescence. However, this method is known to provide inaccurate quantification under inconsistent reaction efficiencies [58].

Since the C_t method was proposed, several alternative methods have been developed to improve absolute quantification in terms of accuracy, precision and robustness. The focus of current research is based on the computation of single features, for example, C_{y_0} or $-\log_{10}(F_0)$, that are linearly related to the initial concentration, as in C_t [59, 60]. The C_{y_0} approach, proposed by Guescini et al. (2008), fits a sigmoid to the amplification curve and takes C_{y_0} as the intersection between the abscissa axis and the tangent of the inflection point [59]. On the other hand, F_0 , proposed by Rutledge et al. (2004), fits the sigmoid up to a “cut-off cycle” and takes F_0 as the fluorescence at cycle 0 [60]. Note that, $-\log_{10}(F_0)$ is used instead of F_0 since it is linearly related to initial DNA concentration. Figure 3.1 depicts the 3 aforementioned quantification metrics.

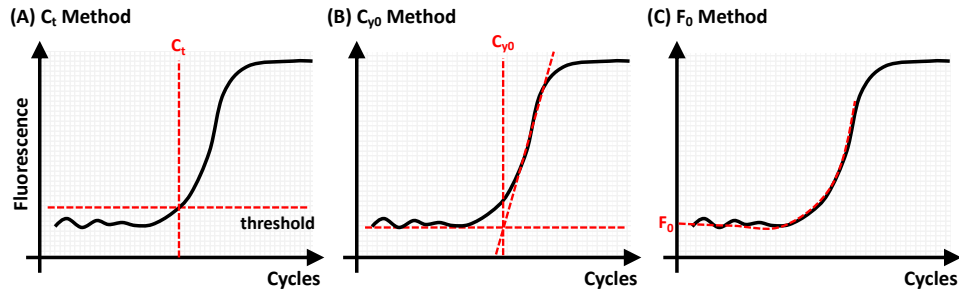


Figure 3.1: Depiction of quantification metrics in DNA amplification, namely, C_t , C_{y_0} and F_0 .

The three aforementioned features correspond to an underlying assumption. For example, the C_t approach assumes the PCR efficiency to be constant between reactions and cycles [57]. The Cy_0 approach allows for different efficiency between reactions but assumes a constant efficiency between cycles [59]. The third feature, $-\log_{10}(F_0)$, allows for different efficiency between reactions but additionally assumes that it decreases from cycle to cycle [60]. These single-feature methods provide a simple approach for absolute quantification, however, the degrees of freedom to implement more complex data analysis techniques is limited, and the use of multiple features together has been unexplored.

Inspired by the field of Machine Learning, this Chapter takes a multidimensional view, combining multiple features in order to take advantage of the information and principles behind all of the current standard curve methods developed. Here, a novel framework is presented that combines existing standard curve methods into a *multidimensional standard curve* (MSC). This is achieved by considering multiple features together such that each amplification curve is viewed as a point in a multidimensional space. Therefore, the standard curve in the multidimensional space should theoretically form a 1D line. Contrary to only considering a single-feature, in the multidimensional space, data points do not fall exactly on the standard curve and thus enables a similarity measure between amplification curves based on distances between data points.

This framework aims to expand the capabilities of standard curves in order to: optimise quantification performance, provide a measure of how suitable an amplification curve is for a standard, and thus automatically detect outliers and increase the reliability of quantification. Here, *outlier* refers to abnormal amplification data, due to non-specific target amplification or inconsistencies in amplification efficiency and reaction conditions (e.g. annealing temperature).

In order to show the effectiveness of this method, an MSC is constructed for phage lambda DNA and evaluated using a figure of merit combining accuracy, precision and overall predictive power. Following this, the framework is used for outlier detection using non-specific DNA targets where the notion of distance in the multidimensional space is explored to understand if it can be used as a similarity measure between amplification curves. Finally, reaction conditions, specifically annealing temperature, is varied as a proxy for amplification efficiency in order to investigate whether the MSC can be used to disregard specific outliers and enhance quantification.

3.3 Experimental Section

3.3.1 Proposed Framework

In order to understand the proposed framework, it is useful to have an overall picture of how standard curves are used for quantification. Here, two terms, namely *training* and *testing* are borrowed from Machine Learning to describe the construction of a standard curve and quantification of unknown samples respectively. Within the conventional single-feature approach, training is typically achieved through 4 stages: pre-processing, curve fitting, feature extraction and line fitting (linear regression). This is illustrated in Figure 3.2 (top branch and solid line). Testing is accomplished by extracting the same feature as in the training, and using the generated standard curve to quantify the concentration in unknown samples.

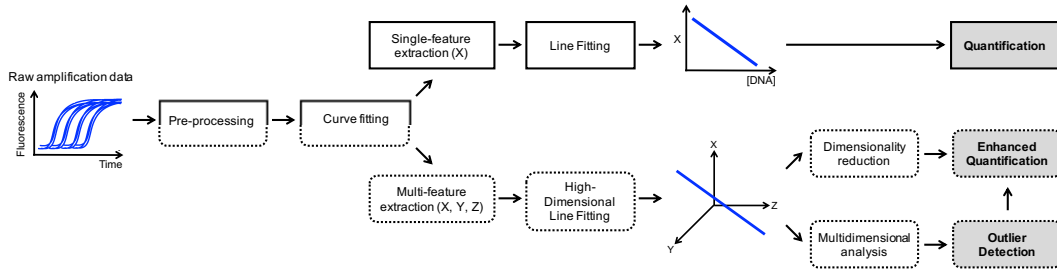


Figure 3.2: Block diagram showing the conventional method (top branch and solid line) compared to the proposed framework (bottom branch and dotted line) for target quantification. In both cases, raw amplification data for several known concentrations of the target are typically pre-processed and fitted with an appropriate curve. In the conventional case, a single feature such as the cycle threshold, C_t , is extracted from each curve. Subsequently, the extracted features are graphed as a function of concentration and a line is fit to the data in order to generate a standard curve and quantification of unknown samples. In the proposed framework, multiple features are extracted and thus a 1D line in high dimensional space (called the feature space) is fitted in order to construct a multidimensional standard curve. Through dimensionality reduction, enhanced quantification can be achieved and performing multidimensional analysis in the feature space allows for outlier detection. The quantification can be further assisted by disregarding outliers.

The proposed framework extends the conventional approach by increasing the dimensionality of the standard curve in order to explore and take advantage of using multiple features together. This new framework is presented in Figure 3.2 (bottom branch and dotted line). For training, there are 6 stages: pre-processing, curve fitting, multi-feature extraction, high dimensional line fitting, multidimensional analysis and dimensionality reduction. Within this framework, testing can be achieved through dimensionality reduction, and multidimensional analysis can be used to detect outliers and support quantification.

In contrast with conventional training, instead of extracting a single linear feature, multiple features are extracted from the processed amplification curves, for example, denoted using the

dummy labels X , Y and Z . Therefore, each amplification curve has been reduced to 3 values (e.g. X_1 , Y_1 and Z_1) and, consequently, can be viewed as a point in 3 dimensional space. *It is important to stress that any number of features could be used as long as they are linearly related to the initial target concentration.* Therefore, the training data should theoretically form a 1-D line in 3-D space. This line is approximated using high-dimensional line fitting and generates what is called the *multidimensional standard curve*. Although, the training data forms a line, it is important to understand that data points do not lie exactly on the line. Consequently, there is considerable room for exploring this multidimensional space, commonly referred to as the *feature space*.

For quantification purposes, the MSC needs to be mapped into a single dimension, denoted as M_0 , linearly related to the initial concentration of the target. In order to distinguish this curve from conventional standard curves, it is referred to here as the *quantification curve*. This can be achieved using dimensionality reduction techniques (DRT) [61]. Mathematically, this means that DRTs are multivariate functions of the form: $M_0 = \phi(X, Y, Z)$ where $\phi(\cdot) : \mathcal{R}^3 \rightarrow \mathcal{R}$. In fact, given that scaling features does not affect linearity, M_0 can be mathematically expressed as $M_0 = \phi(\alpha_1 X, \alpha_2 Y, \alpha_3 Z)$ where α_i for $i \in \{1, 2, 3\}$ are scalar constants. These weights provide a simple method for choosing the contribution of each individual feature in order to improve quantification. Furthermore, regardless of the weightings, all features will be considered for the multidimensional analysis.

3.3.2 Multidimensional Standard Curves

In order to demonstrate the benefits of this proposed framework, a specific *instance of framework* to construct the MSC is required. Therefore, for the purposes of this study, the following instance is used, although it is important to note that any step could be replaced with an appropriate alternative.

- 1) *Preprocessing*. The first step is to perform background subtraction. This is accomplished by subtracting the average of the fluorescent readings in the first five cycles from every amplification curve. More advanced methods could be considered to improve performance such as the *taking-difference linear regression method* [62].
- 2) *Curve Fitting*. In this study, the model used to fit the amplification curves is the 5-parameter sigmoid (Richards Curve) given by:

$$F(x) = F_b + \frac{F_{\max}}{(1 + e^{-(x-c)/b})^d} \quad (3.1)$$

where x is the cycle number, $F(x)$ is the fluorescence at cycle x , F_b is the background fluorescence, $F_{\max}$ is the maximum fluorescence, c is the fractional cycle of the inflection point, b is related to the slope of the curve and d allows for an asymmetric shape (Richard's coefficient). The optimisation algorithm used to fit the five parameters of this model to the data is the trust-region method and is based on the interior-reflective Newton method [63, 64]. Here, the trust-region method is chosen over the Levenberg-Marquardt algorithm since bounds for the 5 parameters can be chosen in order to encourage a unique and realistic solution [65, 66]. The lower and upper bounds for the 5 parameters, $[F_b, F_{\max}, c, b, d]$, are given as: $[-0.5, -0.5, 0, 0, 0.7]$ and $[0.5, 0.5, 50, 100, 10]$, respectively.

- 3) *Feature Extraction.* Three features were used to construct the multidimensional standard curve in this study: C_t , Cy_0 and $-\log_{10}(F_0)$. Therefore, each amplification curve can be represented as a point in 3-dimensional space, i.e. $\mathbf{p} = [C_t, Cy_0, -\log_{10}(F_0)]^T$ where $[\cdot]^T$ denotes the transpose operator. Note that by convention, vectors are denoted using bold lowercase letters.

The cycle-threshold, C_t , is computed by fitting the amplification curve with the 5-parameter sigmoid in equation 3.1, normalising the fitted function $F(x)$ with respect to $F_{\max}$, and then takes C_t as the time where $F(x)$ exceeds 0.2 (i.e. 20% of its maximum fluorescence). The Cy_0 approach also uses the 5-parameter sigmoidal curve-fitting and takes Cy_0 as the intersection between the abscissa axis and the tangent of the inflection point in the fitted $F(x)$. The third feature, $-\log_{10}(F_0)$ fits the sigmoid up to a “cut-off cycle” and takes F_0 as the fluorescence at cycle 0.

- 4) *Line Fitting.* Quantification is conventionally achieved by fitting a line to the Cq values as a function of concentration, as in Figure 2.4. On the contrary, in this work, a 1D line is fit to the training data in N-dimensional space to construct the MSC. In particular, the method of least squares is used, which is equivalent to using the first principal direction in principal component analysis (PCA) [67, 68]. If sufficient data exists, other methods such as random sample consensus (RANSAC) which are robust to outliers could be used [69].
- 5) *Similarity Measure.* There are two similarity measures explored in this study: Euclidean and Mahalanobis distance. The Euclidean distance between a point, $\mathbf{p}$, and the MSC can be calculated by orthogonally projecting the point onto the MSC and then using simple geometry to calculate the Euclidean distance, e , given by

$$P = \Phi(\mathbf{p}, \mathbf{q}_1, \mathbf{q}_2) = \frac{(\mathbf{p} - \mathbf{q}_1)^T (\mathbf{q}_2 - \mathbf{q}_1)}{(\mathbf{q}_2 - \mathbf{q}_1)^T (\mathbf{q}_2 - \mathbf{q}_1)} \quad (3.2)$$

$$e = |(\mathbf{p} - \mathbf{q}_1) - (\mathbf{q}_1 + P \cdot (\mathbf{q}_2 - \mathbf{q}_1))| \quad (3.3)$$

where Φ computes the projection of the point $\mathbf{p} \in \mathcal{R}^n$ onto the multidimensional standard curve, the points $\mathbf{q}_1, \mathbf{q}_2 \in \mathcal{R}^n$ are any two distinct points that lie on the standard curve, and $|\cdot|$ denotes the absolute value operator.

The Mahalanobis distance is defined as the distance between a point, $\mathbf{p}$, and a distribution, $\mathcal{D}$, in multidimensional space [70]. Similar to the Euclidean distance, the point is first projected onto the MSC and the following formula is applied to compute the Mahalanobis distance, d ,

$$d = \sqrt{(\mathbf{p} - P(\mathbf{q}_2 - \mathbf{q}_1))^T \boldsymbol{\Sigma}^{-1} (\mathbf{p} - P(\mathbf{q}_2 - \mathbf{q}_1))} \quad (3.4)$$

where $\mathbf{p}$, P , $\mathbf{q}_1$ and $\mathbf{q}_2$ are given in equation (3.2) and $\boldsymbol{\Sigma}$ is the co-variance matrix of the training data used to approximate the distribution $\mathcal{D}$.

It can be shown that if the data is approximately normally distributed then the Mahalanobis distance squared, d^2 , follows a χ^2 -distribution [71]. Therefore, a χ^2 -distribution table can be used to translate a specific p -value into a distance threshold. For instance, for a χ^2 -distribution with 2 degrees of freedom, a p -value of 0.001 corresponds to a Mahalanobis distance of 3.72.

- 6) *Feature Weights.* As mentioned previously, in order to maximise quantification performance, different weights, α , can be assigned to each feature. This can be accomplished by minimising an error measure on the training data using an optimisation algorithm. The specific error measure used in this study is described in Chapter 3.3.3. The optimisation algorithm is the Nelder-Mead simplex algorithm with weights initialised to unity, i.e. beginning with no assumption on the quantification performance of individual features [72, 73]. This is a standard algorithm and only 20 iterations are used to find the weights so that there is little computational overhead.
- 7) *Dimensionality Reduction.* In this study, every point of the MSC is mapped into an estimated concentration using principal component regression, i.e. $M_0 = P$ from equation (3.2). This is compared with projecting the MSC onto all three dimensions, i.e. C_t , C_{y_0} and $-\log_{10}(F_0)$.

3.3.3 Evaluating Standard Curves

In consistency with the current literature on evaluating standard curves, relative error (RE) and coefficient of variation (CV) are used to measure accuracy and precision respectively. The CV for each concentration is calculated after normalising the standard curves such that a fair

comparison across standard curves is achieved. The formula for RE is given by:

$$\text{RE}_i = 100 \times \left| \frac{\hat{x}_i}{x_i} - 1 \right| \quad (3.5)$$

where i is the index of a given training point, x_i is the true template concentration of the i^{th} training data, and $\hat{x}_i$ is the estimate of x_i using the standard curve. The CV for a given concentration is computed as:

$$\text{CV}_i = 100 \times \frac{\text{std}(\hat{\mathbf{x}}^i)}{\text{mean}(\hat{\mathbf{x}}^i)} \quad (3.6)$$

where i is the index of a given training point and $\hat{\mathbf{x}}^j$ is a vector of estimated concentrations for all training points with the same concentration as x_i . The sample standard deviation and sample mean are denoted by $\text{std}(\cdot)$ and $\text{mean}(\cdot)$, respectively. Alongside the RE and CV, the leave-one-out cross validation (LOOCV) error can be used as a measure for stability and overall predictive performance [67]. Stability refers to the predictive performance when points are removed from the training process. The LOOCV is given as:

$$\text{LOOCV}_i = 100 \times \left| \frac{\hat{z}_i}{x_i} - 1 \right| \quad (3.7)$$

where i is the index of a given training point, x_i is the true concentration of the i^{th} training data and $\hat{z}_i$ is the estimate of x_i using a standard curve generated without the i^{th} training point. In this study, the LOOCV is specified as a percentage in order to compare across different template concentrations, as shown in equation 3.7.

In order for the optimisation algorithm to compute α and simultaneously minimise the three aforementioned measures, it is convenient to introduce a figure of merit, Q , to capture all of the desired properties. For a given training point, the product between all three errors, Q_i , can be used to heuristically measure the quantification performance. Therefore, Q can be defined as the average over all Q_i , as shown in equation 3.8, and is the error measure that the optimisation algorithm will minimise.

$$Q = \frac{1}{N} \sum_{i=1}^N \text{RE}_i \times \text{CV}_i \times \text{LOOCV}_i \quad (3.8)$$

3.3.4 Statistical Analysis

The p -values used for assessing the significance between methods in absolute quantification were calculated using a paired, two-sided Wilcoxon signed rank test. Statistical significance was considered as * p -value < 0.05, ** p -value < 0.01, *** p -value < 0.001, **** p -value < 0.0001. Outliers using multidimensional standard curves were determined using a χ^2 -distribution with 2 degrees of freedom and statistical significance was assumed for a p -value < 0.001. The Henze-Zirkler test is used to determine multivariate normality with a p -value significance level of 0.05.

3.3.5 Fluorescence Datasets

For reproducibility purposes, the DNA targets and experimental conditions are described below. The reader may wish to skip directly to the Results and Discussion (section 3.4).

- 1) Standard curves were constructed using synthetic double-stranded DNA (gBlocks Fragments Genes) containing phage lambda DNA sequence (DNA concentration ranging from 10^2 to 10^8 copies per reaction). Reactions were performed at an annealing temperature of 62°C . See Table A.S1 for primer/sequence information.
- 2) Non-specific outlier detection experiments were performed using synthetic double-stranded DNA carrying *bla*_{OXA-48}, *bla*_{NDM} and *bla*_{KPC} genes, which in this work are referred to as outlier 1, 2 and 3 respectively. These are genes that are resistant to the carbapenem antibiotics. The details of these genes are not significant for this Chapter, although the reader may wish to review the following for more information: <https://www.cdc.gov/hai/organisms/cre/technical-info.html>. Reactions were performed at annealing temperature of 68°C . See Table A.S2-A.S4 for primer/sequence information.
- 3) Specific outlier detection experiments were performed using synthetic double-stranded DNA containing lambda DNA sequence at 10^5 copies per reaction. Reactions were performed at annealing temperatures ranging from 54.0 to 73.6°C . See Table A.S1 for primer and sequence information.

All oligonucleotides were synthesised by IDT (Integrated DNA Technologies, Germany) with no additional purification. The specific PCR primers for lambda phage were designed in-house using Primer3 (<http://biotools.umassmed.edu/bioapps/primer3-www.cgi>), whereas the primer pairs used for the outlier detection were taken from Monteiro et al. (2012) [74]. Real-time PCR amplifications were conducted using FastStart Essential DNA Green Master kit

(Roche Diagnostics, Germany) according to manufacturer's instructions. Each reaction consisted of 2.5 μL FastStart Essential DNA Green Master 2 $\times$ concentrated, 1 μL of PCR grade water, 0.5 μL of 10 $\times$ primer mixture at 5 μM and 1 μL of DNA at variable amounts, in a 5 μL final reaction volume. Thermocycling was performed using a LightCycler 96 (Roche) initiated by a 10 min incubation at 95 $^{\circ}\text{C}$, followed by 40 cycles: 95 $^{\circ}\text{C}$ for 20 sec; 62 $^{\circ}\text{C}$ (for lambda) or 68 $^{\circ}\text{C}$ (for non-specific outliers) for 45 sec; and 72 $^{\circ}\text{C}$ for 30 sec, with a single fluorescence reading taken at the end of each cycle. Each reaction combination was conducted in quintuplicates/octuplates. All the runs were completed with a melting curve analysis performed at 95 $^{\circ}\text{C}$ for 10 sec, 65 $^{\circ}\text{C}$ for 60 sec, and 97 $^{\circ}\text{C}$ for 1 sec (continuous reading from 65 to 97 $^{\circ}\text{C}$) to confirm the specificity of amplification and the lack of primer dimer. Appropriate positive and negative controls were included in each experiment.

3.4 Results and Discussion

In this study, a new framework is presented to construct multidimensional standard curves in order to: (i) optimise the quantification performance; (ii) detect outliers; and (iii) provide a heuristic measure for the similarity between an amplification curve and the MSC.

Optimising Quantification Performance. Synthetic phage lambda DNA was used to construct an MSC and evaluate its quantification performance relative to single feature methods. The resulting MSC, constructed using the features C_t , C_{y_0} and $-\log_{10}(F_0)$, is visualised in Figure 3.3 (a). The computed features and curve-fitting parameters for each amplification curve grouped by concentration, ranging from 10^2 to 10^8 copies per reaction, is presented in Table A.S5. For comparison, Figure 3.3 (b) shows the quantification curves for each single-feature method plus the multi-feature method M_0 which is obtained after dimensionality reduction through principal component regression.

The proposed framework enables the user to optimise quantification performance (through weighting each feature) in terms of a figure of merit, Q . The Q chosen in this work combines RE, CV and LOOCV. After 20 iterations of the optimisation algorithm, the weights α converged to $[-0.0741, 1.1185, 1.6574]$ corresponding to C_t , C_{y_0} and $-\log_{10}(F_0)$ respectively. It is important to stress that although the optimisation algorithm suggests different performances across the selected features, there is value in keeping all of them as it can assist outlier detection, as shown in the subsequent sections.

The average Q ($\pm$ standard deviation) for M_0 against the single-feature methods is visualised

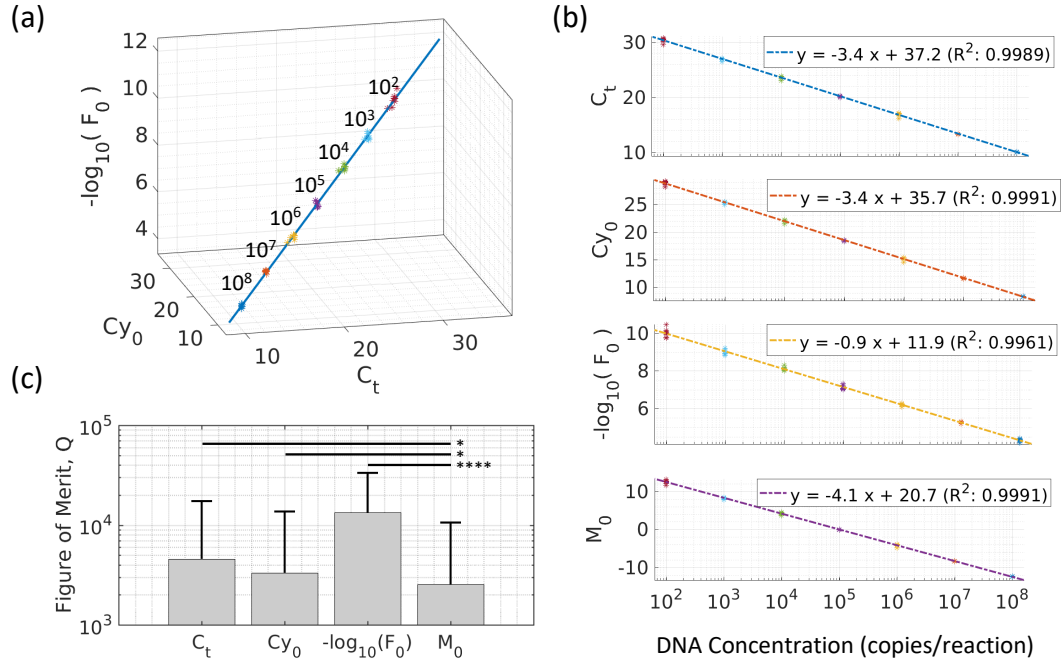


Figure 3.3: Evaluating quantification through the multidimensional standard curve and single-feature methods. (a) A multidimensional standard curve is constructed using C_t , Cy_0 and $-\log_{10}(F_0)$ for lambda DNA with concentration values ranging from 10^2 to 10^8 (top right to bottom left). (b) The constructed standard curves using single-feature methods along with M_0 . (c) The average figure of merit combining accuracy, precision and overall predictive power for all methods. A paired two-sided Wilcoxon signed rank test was performed between M_0 and the other methods (* p -value < 0.05 and **** p -value < 0.0001).

in Figure 3.3 (c), where for C_t , Cy_0 , $-\log_{10}(F_0)$ and M_0 it is 4587 ± 12799 , 3327 ± 10357 , 13384 ± 19966 and 2547 ± 8058 respectively. Therefore, in terms of average Q, M_0 enhances quantification by 17.44% ($p < 0.05$), 10.65% ($p < 0.05$) and 99.3% ($p < 0.0001$) compared to C_t , Cy_0 and $-\log_{10}(F_0)$ respectively. A summary and breakdown of each calculated error for all methods grouped by concentration are provided in Table A.S6-A.S9.

Outlier Detection. In this section, the concept of distance in the feature space is explored in order to demonstrate the capability of the framework for outlier detection. This can be caused by non-specific amplification or inconsistent amplification efficiencies and reaction conditions; both of which are investigated in this Chapter.

First, three non-specific DNA targets with respect to the phage lambda MSC, referred to as outlier 1, 2 and 3, were amplified. Subsequently, features were extracted from each amplification curve following the same procedure as the training data for the phage lambda MSC. Finally, the outliers were plotted in the feature space, as shown in Figure 3.4. It is visually clear that the outliers do not lie on the MSC and, therefore, this suggests that sufficient information is captured from the three features extracted from the amplification curves in order to distinguish

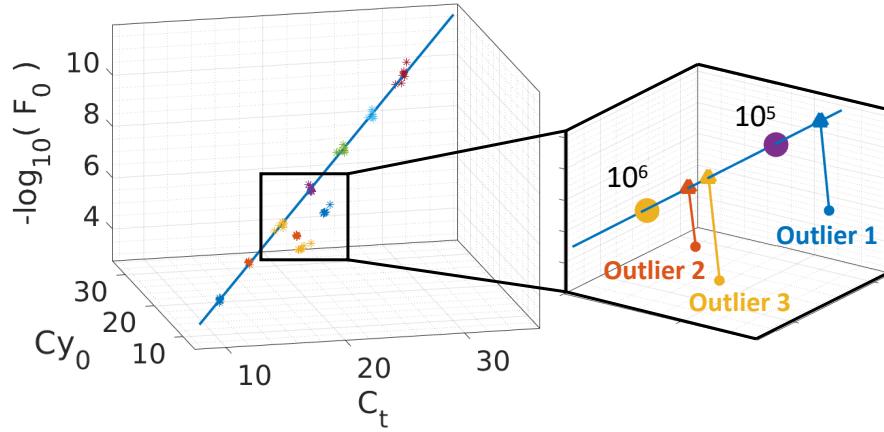


Figure 3.4: The multidimensional standard curve for lambda DNA along with three non-specific outliers. The right panel shows a zoomed region of the feature space with the mean of the replicates and the projection of the outliers onto the standard curve. The computed features/curve-fitting parameters of the three outliers and melting curve analysis are presented in Table A.S10 and Figure A.S1-A.S2 respectively.

the different targets from phage lambda DNA. Notice that without any secondary confirmation, e.g. from melting curves or agarose gels, the test data itself suggests it is not ‘similar’ to the training data. Furthermore, it is important to note that single feature methods are unable to distinguish these outliers (e.g. a given C_t value will always fall on a conventional standard curve).

In order to fully capture the position of the outliers in the feature space, it is convenient to remove the effect of concentration and view the feature space along the axis of the multidimensional standard curve. This is achieved by projecting all the data points in the feature space onto the plane perpendicular to the standard curve as illustrated in Figure 3.5 (a). The resulting projected points are shown in Figure 3.5 (b). It is clear that all three outliers can be clustered and clearly distinguished from the training data. Furthermore, the Euclidean distance, e , from the MSC to the mean of the outliers is given by $e_1 = 1.44$, $e_2 = 0.99$ and $e_3 = 1.66$. Given that the furthest training point from the MSC in terms of Euclidean distance is 0.36, the ratio between e_1 , e_2 , e_3 and the furthest training point is 4.00, 2.75 and 4.61, respectively. In other words, the mean of outlier 1 is 4 times further than the furthest training point. Therefore, this ratio can be used as a similarity measure and the three clusters could be classified as outliers using a threshold. However, this similarity measure has two limitations: (i) There is an assumption that distances in different directions are equally likely, which is intuitively untrue in the feature space because a change in one direction, e.g. C_t , does not impact the amplification curve as much as another, e.g. $-\log_{10}(F_0)$. (ii) There is no probabilistic measure that captures the distribution of the data and therefore the threshold for determining outliers must be chosen arbitrarily.

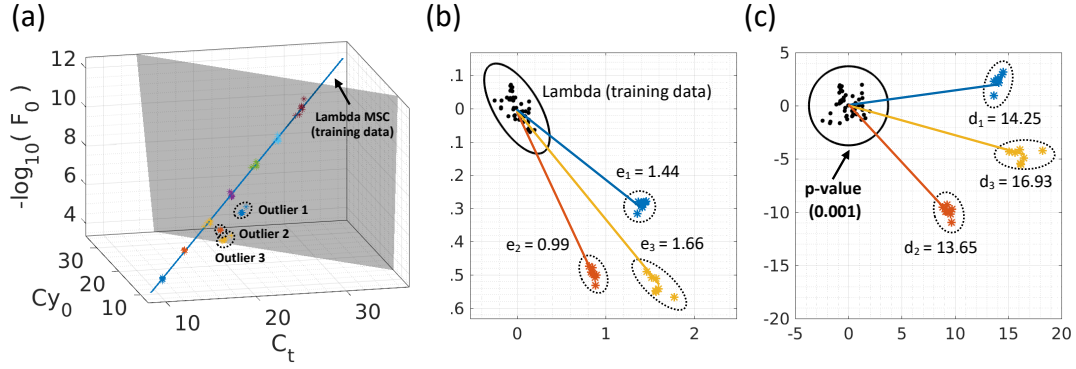


Figure 3.5: Multidimensional analysis using the feature space for detecting non-specific outliers. (a) MSC using C_t , C_{y0} and $-\log_{10}(F_0)$ for lambda DNA along with three non-specific outliers. An arbitrary hyperplane orthogonal to the MSC is shown in grey. (b) Euclidean space: the view of the feature space when all the data points have been projected onto the aforementioned hyperplane. The Euclidean distance between the mean of the training data and the outliers (e_1 , e_2 and e_3). (c) Mahalanobis space: a transformed space where the Euclidean distance is equivalent to the Mahalanobis distance, d , in the Euclidean space. The black circle corresponds to a p -value of 0.001 using a χ^2 -distribution with 2 degrees of freedom.

In order to tackle the two aforementioned limitations, the Mahalanobis distance, d , can be used. Clearly, by observing Figure 3.5 (b), the training data predominantly varies in a given direction. In order to visualise the Mahalanobis distance, the orthogonal view of the feature space (Figure 3.5 (b)) can be transformed into a new space (Figure 3.5 (c)) where the Euclidean distance is equivalent to the Mahalanobis distance in the original space. This is achieved by normalising the principal components of the training data.

The Mahalanobis distance from the multidimensional standard curve to the mean of the outliers is $d_1 = 14.25$, $d_2 = 13.65$ and $d_3 = 16.93$, respectively. In contrast with the Euclidean distances, it is observed that when considering the distribution of the data, the position of the outliers change. A useful property of d is that its squared value, d^2 , follows a χ^2 -distribution if the data is approximately normally distributed. The hypothesis that the data is normally distributed is confirmed using the Henze-Zirkler test with a significance level of 0.05. Therefore, the distance can be converted into a probability in order to determine if a data point is an outlier. Based on the χ^2 -distribution table with 2 degrees of freedom, any point further than 3.717 is 99.9% (p -value < 0.001) likely to be an outlier. Since all the outliers have a Mahalanobis distance significantly greater than 3.717, they are confidently classified as outliers.

Aside from non-specific DNA amplification, another cause of outliers, especially in resource-limited settings, is due to inconsistent reaction efficiencies between the training and test data. In the following study, the annealing temperature is used as a proxy for varying the efficiency of lambda DNA amplification. Figure 3.6 (a) shows the amplification curves for lambda DNA at 10^5 copies/reaction for temperatures ranging from $54.0 - 73.6^\circ C$. From observing the change in C_t it

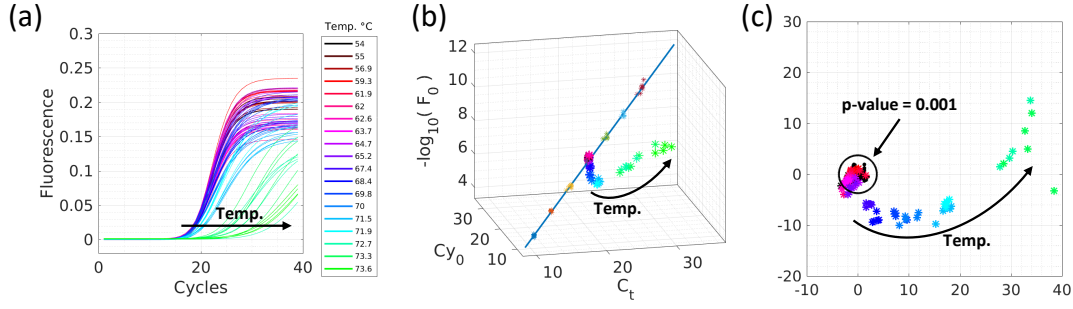


Figure 3.6: The effect of changing annealing temperature on detecting outliers using multidimensional standard curves. (a) Fluorescent amplification curves for lambda DNA (10^5 copies per reaction) at temperatures ranging from $54.0 - 73.6^\circ\text{C}$. (b) The MSC constructed at 62°C using the features C_t , C_{y0} and $-\log_{10}(F_0)$ along with data points obtained from the aforementioned fluorescent amplification curves. (c) The lambda standard and temperature variation data points in the Mahalanobis space. The black circle corresponds to a p -value of 0.001 using a χ^2 -distribution with 2 degrees of freedom.

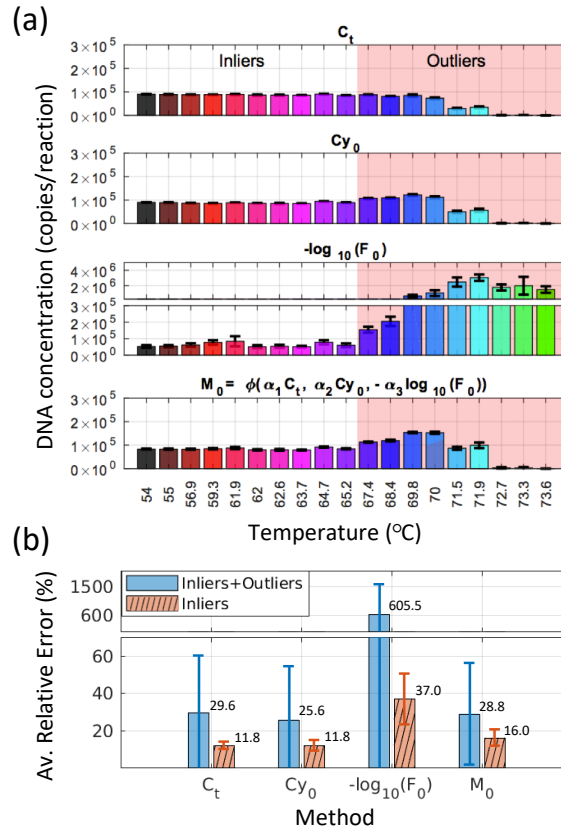


Figure 3.7: Merging quantification and outlier detection. (a) Average estimated quantification for lambda DNA at 10^5 copies per reaction for annealing temperatures ranging from $54.0 - 73.6^\circ\text{C}$ using C_t , C_{y0} , $-\log_{10}(F_0)$ and M_0 . The shaded region indicate outliers according to the MSC with p -value < 0.001 based on a χ^2 -distribution with 2 degrees of freedom. Details for the quantification and outlier detection are provided in Table A.S11 and A.S12. (b) Average relative error for estimated quantification of lambda DNA at 10^5 copies/reaction across all annealing temperatures for every method. The solid bars represent the average relative error for all data points (including outliers and inliers) whereas the dashed bars only considers inliers. A paired two-sided Wilcoxon signed rank test was performed between M_0 and the other methods with a confirmed significance (p -value < 0.0001).

can be observed that, even though the product is specific (see Figure A.S3 for melting analysis), the quantification performance can be drastically affected. Current standard curve approaches have no heuristic measure to indicate whether any of these curves will be quantified poorly. However, when the data is viewed in the feature space (Figure 3.6 (b)) and the Mahalanobis space (Figure 3.6 (c)), it can be observed that when the amplification shape diverges from the curves belonging to the MSC (especially for low efficiencies), the Mahalanobis distance between the test data and the MSC increases. Therefore, this raises the question: can d be used to disregard *specific* outliers and therefore support quantification?

Merging Quantification and Outlier Detection. In order to investigate the use of the MSC and Mahalanobis distance for supporting quantification, the effect of removing outliers on the estimated quantification can be explored. Figure 3.7 (a) shows the average quantification as a function of the temperature for C_t , Cy_0 , $-\log_{10}(F_0)$ & M_0 (bar plots) and visualises the temperatures at which the amplification curves are considered as outliers (shaded in red).

There are two key observations that can be made: (i) the quantification performance begins to deteriorate at temperatures above $65.2^\circ C$ for all methods; (ii) amplification curves at temperatures above $65.2^\circ C$ are considered as outliers based on a χ^2 -distribution with $p < 0.001$. These two observations coincide and, therefore, support the claim that the selected features extracted from the amplification curves contain sufficient information to disregard outliers and improve quantification performance. Figure 3.7 (b) shows the average relative error in the estimated quantification for all the considered methods when using all data points and also disregarding outliers. It can be observed that quantification is improved by 59.9%, 53.9%, 93.9% and 44.6% for C_t , Cy_0 , $-\log_{10}(F_0)$ and M_0 , respectively. Notice that the benefits of multidimensional analysis using *all* of the features extends to enhancing quantification performance of *any* method, including single-feature methods.

3.5 Conclusion

Absolute quantification of nucleic acids in real-time PCR using standard curves is exceedingly important in several fields of biomedicine, although research has saturated in recent years. This is partially due to the simplicity of standard curves and the movement of research towards digital PCR (dPCR) because of the advantages it holds over qPCR, such as removing the need for a standard curves. However, dPCR is currently not suitable for many applications given the cost and complexity of instruments [75, 76, 77]. This Chapter presents a framework that

shows the benefits of standard curves extend beyond absolute quantification when observed in a multidimensional environment. Consequently, this work opens the possibility for researchers from different fields to explore mathematical methods and applications that are enabled by this proposed framework.

The focus of current researchers is on the computation of a single value, referred to here as a *feature*, that is linearly related to the template concentration. Therefore, there has been a gap in the literature in taking advantage of multiple features together. The potential reason for a lack of research in this area is because of the non-trivial benefits of combining linear features. The only intuitive interpretation of using several features is in the reliability of quantification. For example, instead of trusting a single feature, e.g. C_t , other features such as C_{y_0} and $-\log_{10}(F_0)$ can be used to check if the quantification result is similar. This unidimensional way of thinking prevents several degrees of freedom and advantages that the proposed framework enables.

Three main capabilities are enabled by the framework proposed in this Chapter: (i) to optimise quantification performance based on a figure of merit; (ii) to detect outliers; and (iii) to measure how suitable an amplification curve is for the constructed MSC. The first capability provides a lower bound on the quantification performance of the framework to single best feature since this is a special case (e.g. $M_0 = C_t$ when $\alpha_1 = 1$, $\alpha_2 = 0$ and $\alpha_3 = 0$). The second and third capabilities are an application of the MSC that was enabled through exploring the information gain captured by the elements of the feature space (e.g. Mahalanobis distance); which are typically meaningless or not considered in the unidimensional approach. In fact, applications of the MSC have already been developed, as shown in Chapter 4 [47].

The multidimensional approach is not completely unfamiliar in absolute quantification. The shape based outlier detection (SOD) takes a multidimensional approach in order to define a similarity measure between amplification curves [78]. However, there are two fundamental differences with the work of this Chapter. The first is that SOD relies on using a specific model for amplification, namely the 5-parameter sigmoid, and is therefore not a general framework. The second difference is that the pattern between the features in SOD and initial target concentration is unknown, therefore the SOD cannot be naturally integrated into the quantification process and is typically used as an add-on [79]. In other words, the multidimensional approach is only considered for outlier detection and quantification is still considered as unidimensional.

The contribution of this Chapter can be accredited to the framework as a whole and the feature space which incorporates the multidimensional standard curve. Currently, the framework is limited to considering features that are linearly related to the initial target concentration. This limitation is in fact a design choice given that there is a lack of other types of features available in

the literature with non-linear relationships, and in order to reduce the complexity of the analysis. The second limitation is related to the feature space. The question arises as to whether sufficient information is captured between amplification curves in order to distinguish them in the feature space. For example, if two unrelated PCR reactions exhibit a perfectly symmetric sigmoidal amplification curve, their position in the feature space may potentially overlap. This limitation can be tackled from a molecular perspective by tuning the chemistry (e.g. amplicon length, primer location, GC content, etc) in order to sufficiently change amplification curves without compromising the performance of the reaction (e.g. speed, sensitivity, specificity, etc).

In terms of future directions, there are many research paths that can be explored. Both the theory of the framework and its applications can be investigated. The results presented in this Chapter raise a number of questions: Can the proposed framework be used for emerging isothermal amplification chemistries? Is there any benefit of using more than 3 features? How could the framework accommodate features with a non-linear relationship to the initial template concentrations? Can the MSC approach reduce the number of required technical replicates for quantifying unknown samples given the increased confidence provided by the distance measure?

In conclusion, this Chapter presents a framework, multidimensional standard curve and the feature space - which presents many opportunities for researchers to explore new techniques and ideas. I hope that by sharing these concepts, others will be able to adapt and enhance this work to meet their objectives and advance the field of nucleic acid research.

Chapter 4

Simultaneous Single-Channel Multiplexing and Quantification

4.1 Chapter Overview

Multiplexing and quantification of nucleic acids, both have, in their own right, significant and extensive use in biomedical related fields. Currently, the ability to detect several nucleic acid targets in a single-reaction scales linearly with the number of targets; an expensive and time-consuming feat. Here, the ideas in Chapter 3 are extended to propose a new methodology based on multidimensional standard curves that extends the use of real-time PCR data to multiplexing. By applying this novel methodology, simultaneous single-channel multiplexing and enhanced quantification of multiple targets using only real-time amplification data is achieved. This is obtained without the need of fluorescent probes, agarose gels, melting curves or sequencing analysis. Given the importance and demand for tackling challenges in antimicrobial resistance, the proposed method is applied to four of the most prominent carbapenem-resistant genes: *bla*_{OXA-48}, *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC}, which account for 97.1% of the UK's reported carbapenemase-producing Enterobacteriaceae.

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

- J. Rodriguez-Manzano*, A. Moniri*, K. Malpartida-Cardenas, J. Dronavalli, F. Davies, A. Holmes and P. Georgiou, “Simultaneous Single-Channel Multiplexing and Quantification of Carbapenem-Resistant Genes using Multidimensional Standard Curves”, *ACS Analytical Chemistry*, vol. 91, no. 3, pp. 2013-2020, Jan 2019. *These authors contributed equally.
- P. Georgiou, A. Moniri and J. Rodriguez-Manzano, “A method for analysis of real-time amplification data”, WO2019234247A1 (Assignee: Imperial Innovations Limited). Priority 08-06-2018 / Filed 07-06-2019 / Published 12-12-2019.

4.2 Introduction

This work demonstrates simultaneous multiplex qPCR and absolute quantification using standard curves, employing only a single fluorescence channel without post-PCR analysis; such that it can be used with conventional qPCR instruments. This is achieved by extending the quantification framework proposed by Chapter 3 to establish that multidimensional standard curves (MSCs) can also be used for target identification. The proposed method is validated for the detection of the β -lactamase genes *bla*_{OXA-48}, *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC} using bacterial isolates from clinical samples in a single-reaction; without fluorescent probes, agarose gels, melting curves or sequencing analysis.

The β -lactamases are enzymes produced by bacteria that provide multi-resistance to the β -lactam antibiotics (which includes penicillin, among others) [80]. In particular, a type of β -lactamase called carbapenemase is highly resistant to the class of carbapenem antibiotics - which is a treatment typically used as a last resort for many infections. More specifically, carbapenem-resistant Enterobacteriaceae (CRE), also known as carbapenemase-producing Enterobacteriaceae (CPE), are bacteria that produce carbapenamases [81]. Therefore, it is important to identify β -lactamase genes, so as to determine if the bacteria is resistant to certain carbapenem antibiotics. This is because misusing antibiotics leads to more antimicrobial resistance, making infections harder to treat and increasing the risk of disease spread, severe illness and death [82]. Table 4.1 summarises the breakdown of confirmed carbapenemase-producing Enterobacteriaceae (CPE) cases in the UK from 2003 to 2015. The chosen drug resistant genes in this study cover over 97.1% of the total reported cases [83].

Table 4.1: Laboratory confirmed cases of carbapenemase-producing Enterobacteriaceae from UK laboratories (2003-2015).

Carbapenemases	<i>bla</i> _{OXA-48}	<i>bla</i> _{NDM}	<i>bla</i> _{VIM}	<i>bla</i> _{KPC}	Others
Total cases	1325	1129	491	3260	189
Percentage (%)	20.73	17.67	7.68	51.01	2.96

Invasive infections with carbapenemase-producing strains are associated with high mortality rates (up to 40 - 50%) and represent a major public health concern worldwide [84, 85]. Rapid and accurate screening for carriage of CPE is essential for successful infection prevention and control strategies as well as bed management [86, 87]. However, routine laboratory detection of CPE based on carbapenem susceptibility is challenging [88]: (i) culture-based methods have limited sensitivity and long turnaround time; (ii) nucleic acid amplification techniques (NAATs) are often too expensive and require sophisticated equipment to be used as a screening tool in healthcare systems; and (iii) multiplexed NAATs have not been able to meet the demand for

high-level multiplexing using available technologies.

There is an unmet clinical need for new molecular tools that can be successfully adopted within existing healthcare settings. For example, the pros and cons of various CPE screening workflows in UK hospitals was investigated in Knight et al. (2018) [89]. The aforementioned study discusses the choice between fast & expensive PCR versus cheap & slow culture methods. In an ideal world, this choice would be driven solely by maximising patient care and outcomes, however this study exemplifies the reality behind choosing detection methods in the national healthcare system - minimising time and cost.

The proposed method allows *existing* technologies to benefit from the advantages of multiplex PCR assays whilst reducing the complexity of CPE screening; resulting in a time and cost-effective solution. This is enabled through changing the fundamental approach of current data analytic techniques for the quantification of nucleic acids from unidimensional to multidimensional. Figure 4.1 compares both (a) the conventional approach versus (b) the proposed method for single-channel multiplexing. In the first, conventional standard curves are constructed but multiple targets cannot be differentiated and quantified without post-PCR processing. In contrast, the proposed method constructs multidimensional standard curves, extracting more information from the amplification curves, allowing for simultaneous quantification and multiplexing in a single channel. This work uses CPE as a clinically relevant case study, however, the reader is encouraged to explore other targets and amplification chemistries in order to expand the capabilities of current state-of-the art technologies.

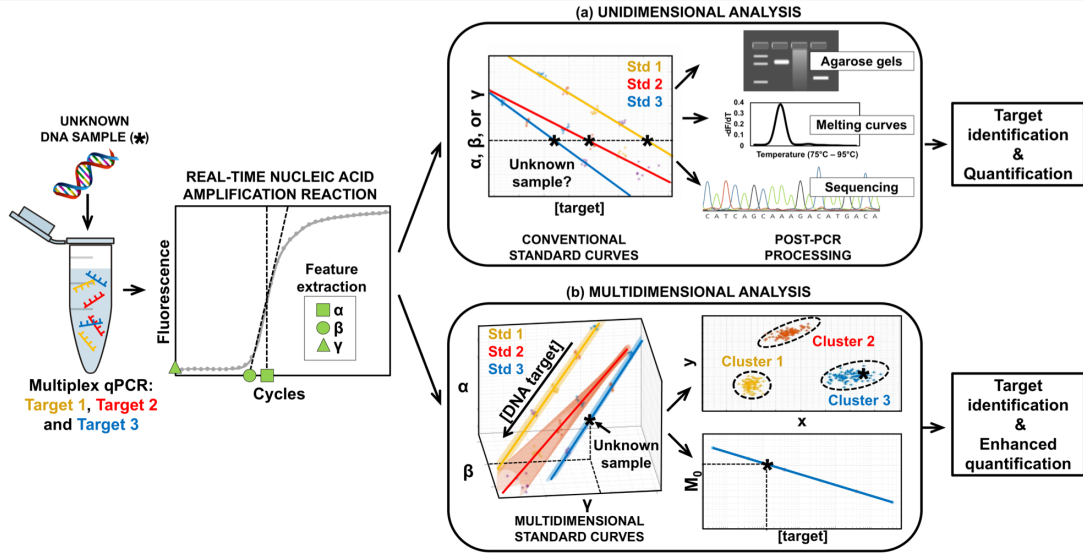


Figure 4.1: Illustration of experimental workflow for single-channel multiplex quantitative PCR using unidimensional and multidimensional approach. An unknown DNA sample is amplified by multiplex qPCR for targets 1, 2 and 3. Features denoted using dummy variables α , β and γ are extracted from the amplification curve. It is important to stress that any number of targets and features could have been chosen. (a) Unidimensional analysis. Three conventional standard curves are generated through serial dilution of the known targets using a single feature. Given it is not possible to identify the target based on these standard curves, post-PCR analysis is required for target identification and quantification. (b) Multidimensional analysis. Three multidimensional standard curves are constructed through serial dilution of the known targets using multiple features. The unknown samples can be confidently classified and an enhanced quantification can be achieved by combining all the features into a unified feature called M_0 .

4.3 Experimental Section

For reproducibility purposes, the DNA targets, experimental conditions and data processing workflow are described below. The reader may wish to skip directly to the Results and Discussion (section 4.4).

4.3.1 Primers and amplification reaction conditions

All oligonucleotides used in this study were synthesised by IDT (Integrated DNA Technologies, Germany) with no additional purification. Primers were previously reported by Monteiro et al. (see Table 4.2) [74]. Each amplification reaction was performed in 5 μ L of final volume with: 2.5 μ L FastStart Essential DNA Green Master 2 $\times$ concentrated (Roche Diagnostics, Germany), 1 μ L PCR Grade water, 0.5 μ L of 10 $\times$ multiplex PCR primer mixture containing the four primer sets (5 μ M of each primer) and 1 μ L of different concentrations of synthetic DNA or bacterial genomic DNA. PCR amplifications consisted of 10 min at 95 $^{\circ}$ C followed by 45 cycles at 95 $^{\circ}$ C

for 20 sec, 68 °C for 45 sec and 72 °C for 30 sec. In order to validate the proposed method, the results were compared against melting curve analyses. One melting cycle was performed at 95 °C for 10 sec, 65 °C for 60 sec and 97 °C for 1 sec (continuous reading from 65 °C to 97 °C). Each experimental condition was ran 5 to 8 times loading the reactions into Light Cycler 480 Multiwell Plates 96 (Roche Diagnostics, Germany) using a Light Cycler 96 Real-Time PCR System (Roche Diagnostics, Germany). Appropriate negative and positive controls were included in each experiment.

Table 4.2: Primers used for the multiplex qPCR assay. Primer names and sequences were previously reported by Monteiro et al (2012) [74]. Sequences are given in the 5' to 3' direction. Size denotes PCR amplification products.

Target	Primer	Sequence	Size (bp)
<i>bla</i> _{OXA-48}	OXA-48-F	TGTTTTTGGTGGCATCGAT	177
	OXA-48-R	GTAAMRATGCTTGGTTTCGC	
<i>bla</i> _{NDM}	NDM-F	TTGGCCTTGCTGTCCTTG	82
	NDM-R	ACACCAGTGACAATATCACCG	
<i>bla</i> _{VIM}	VIM-F	GTTTGGTTCGCATATCGCAAC	382
	VIM-R	AATGCGCAGCACCAGGATAG	
<i>bla</i> _{KPC}	KPC-F	TCGCTAAACTCGAACAGG	785
	KPC-R	TTACTGCCCCGTTGACGCCCAATCC	

4.3.2 Synthetic DNA and bacterial isolates

Four synthetic double-stranded DNA (gBlock Gene fragments) were purchased from IDT and re-suspended in TE buffer to 10 ng/μL stock solutions (stored at −20 °C until further use). The concentrations of all DNA stock solutions were determined using a Qubit 3.0 fluorimeter (Life Technologies). The synthetic templates contained the DNA sequence from: *bla*_{OXA-48}, *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC} genes required for the multiplex qPCR assay (Table B.S1). Standard curves were constructed utilising different DNA concentrations as follows: *bla*_{OXA-48} (10⁸-10⁴ copies/reaction), *bla*_{NDM} (10⁷-10¹ copies/reaction), *bla*_{VIM} (10⁸-10³ copies/reaction) and *bla*_{KPC} (10⁸-10³ copies/reaction). Pure bacterial cultures from clinical isolates were used in this study, as described in Table 4.3. One loop of colonies from each pure culture was suspended in 50 μL digestion buffer at pH 8.0 (Tris-HCl 10 mmol/L, EDTA 1 mmol/L and 5 U/μL lysozyme) and incubated at 37 °C for 30 min in a dry bath. Subsequently, 0.75 μL proteinase K at 20 μg/μL (Sigma) was added, and the solution was incubated at 56 °C for 30 min. Afterwards, the solution was boiled for 10 min to inactivate proteinase K, the samples were centrifuged at 10,000 ×g for 5 min and the supernatant was transferred in a new tube and stored at −80 °C before use. Sample 9 was generated by mixing sample 6 & 8 at equal proportions. Non-CPE producers *Klebsiella pneumoniae* and *Escherichia coli* were included as control strains.

Table 4.3: Bacterial isolates used in this study.

Sample ID	Bacterial Isolate	Carbapenemase genes
1	<i>Klebsiella pneumoniae</i>	<i>bla</i> _{OXA-48}
2	<i>Escherichia coli</i>	<i>bla</i> _{OXA-48}
3	<i>Citrobacter Freundii</i>	<i>bla</i> _{VIM}
4	<i>Escherichia coli</i>	<i>bla</i> _{NDM}
5	<i>Klebsiella pneumoniae</i>	<i>bla</i> _{OXA-48}
6	<i>Klebsiella pneumoniae</i>	<i>bla</i> _{NDM}
7	<i>Pseudomonas aeruginosa</i>	<i>bla</i> _{VIM}
8	<i>Klebsiella pneumoniae</i>	<i>bla</i> _{KPC}
9	<i>Klebsiella pneumoniae</i>	<i>bla</i> _{NDM} + <i>bla</i> _{KPC}
10	<i>Klebsiella pneumoniae</i>	non-producer
11	<i>Escherichia coli</i>	non-producer

4.3.3 Multidimensional standard curves

The framework shown in Chapter 3 provides a generalisation of the approach for quantification of nucleic acids using standard curves. The stages of data analysis are as follows: pre-processing, curve-fitting, multiple feature extraction, high-dimensional line fitting, similarity measure, feature weighting, and dimensionality reduction. The major difference with the conventional approach to quantification is that multiple features are extracted from amplification curves. The stages used in this work, also referred to as the instance of framework, are described below and summarised in Table 4.4.

Table 4.4: Instance of framework presented in Chapter 3 and Moniri et al. [90].

Data Analysis Stages	Method	References
Pre-processing	Baseline correction	-
Curve fitting	5-parameter sigmoid	[91]
Feature extraction	C_t , Cy_0 and $-\log_{10}(F_0)$	[57, 59, 60]
Line fitting	Method of least squares	[92]
Similarity measure	Mahalanobis distance: d	[70]
Feature weights	Minimise figure of merit: Q	[90]
Dimensionality reduction	Principal component regression: M_0	[92]

4.4 Results and Discussion

4.4.1 Target Discrimination using Multidimensional Analysis

Given that it is non-trivial that several targets can be multiplexed and differentiated using only fluorescent amplification data in a single channel, it is helpful to visualise an example. Figure 4.2 shows four amplification curves and their respective derived melting curves specific for *bla*_{OXA-48}, *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC} genes. The four curves have been chosen to have similar C_t (within 1.2 cycles). Using only this information, i.e. the conventional way of thinking, post-PCR processing such as melting curve analysis is needed to differentiate the targets. The same argument applies when solely observing Cy_0 or F_0 . This is an expected result given that these parameters are used for quantification and were not intended for target identification. However, can observing all three features together provide information for target discrimination?

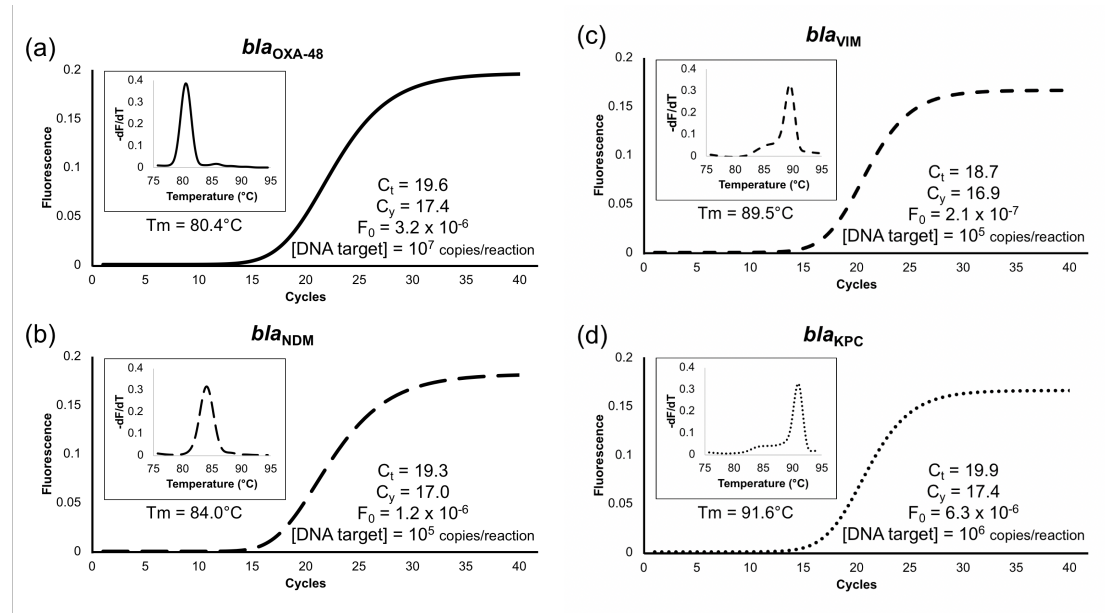


Figure 4.2: Experimentally obtained amplification and melting curves by single-channel multiplex quantitative PCR for: (a) *bla*_{OXA-48}, (b) *bla*_{NDM}, (c) *bla*_{VIM} and (d) *bla*_{KPC} genes. Extracted features (C_t , Cy_0 and F_0) are shown on the respective plots. Each plot has been generated by amplifying standard synthetic DNA containing the target of interest. Background subtraction has been performed on all amplification curves and all samples contain SYBR Green I dye.

In Chapter 3 it was shown that considering multiple features contains sufficient information gain in order to discriminate outliers from a specific target using a MSC. However, this raises the question: does the outlier lie on its own MSC? If so, can this property be taken advantage of in order to build several multidimensional standard curves in order to discriminate multiple specific targets?

To explore this new concept, MSCs are constructed using a single primer mix for the four target genes using C_t , C_{y0} and $-\log_{10}(F_0)$, as shown in Figure 4.3. It is visually observable that the 4 standards are sufficiently distant in multidimensional space, also termed the feature space, in order to distinguish them. That is, an unknown DNA sample can be potentially classified as one of the specific targets (or an outlier) solely using the extracted features from amplification curves in a single channel.

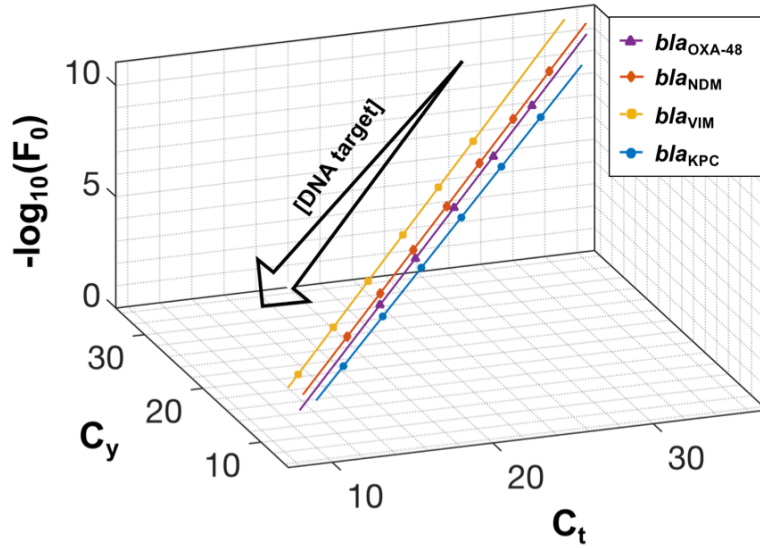


Figure 4.3: Multidimensional standard curves for detection of four carbapenemase genes: *bla*_{OXA-48} (purple line), *bla*_{NDM} (red line), *bla*_{VIM} (yellow line) and *bla*_{KPC} (blue line). They were constructed using C_t , C_{y0} and $-\log_{10}(F_0)$ features extracted from real-time amplification curves derived from amplifying 10-fold dilutions of synthetic DNA. From bottom left to top right, target concentrations range between: 10^8 - 10^4 copies/reaction for *bla*_{OXA-48}; 10^7 - 10^1 copies/reaction for *bla*_{NDM}; 10^8 - 10^3 copies/reaction for *bla*_{VIM}; and 10^8 - 10^3 copies/reaction for *bla*_{KPC}. Each concentration was repeated 5 to 8 times and the resulting average values are projected onto the standard curves. The computed features and curve-fitting parameters for each MSC is presented in Table B.S2.

In order to demonstrate the proposed method for multiplexing, 11 samples (bacterial isolates) given in Table 3 were tested against the multidimensional standards. The specificity of all results were validated using a melting curve analysis (Figure B.S1). As expected, samples 1 to 9 provided a positive outcome whereas samples 10 & 11 (control strains) showed no amplification. The similarity measure used to classify the unknown samples is the Mahalanobis distance using a p-value of 0.01 as the threshold to determine if the sample is an outlier. The results from testing can be succinctly captured within a bar chart shown in Figure 4.4. There are two main observations: (i) the mean of the test samples (bacterial isolates) which have a single resistance (samples 1-8) are correctly classified with a p-value < 0.01 . (ii) the target with multiple resistances (sample 9) is considered as an outlier for all of the targets. Although multiple resistances are not currently common for CPE in the UK (accounting for less than

1.3% of confirmed cases),[83] there is room to explore extending MSCs for other applications that require detecting multiple targets in a single reaction. However, this is outside the scope of this study.

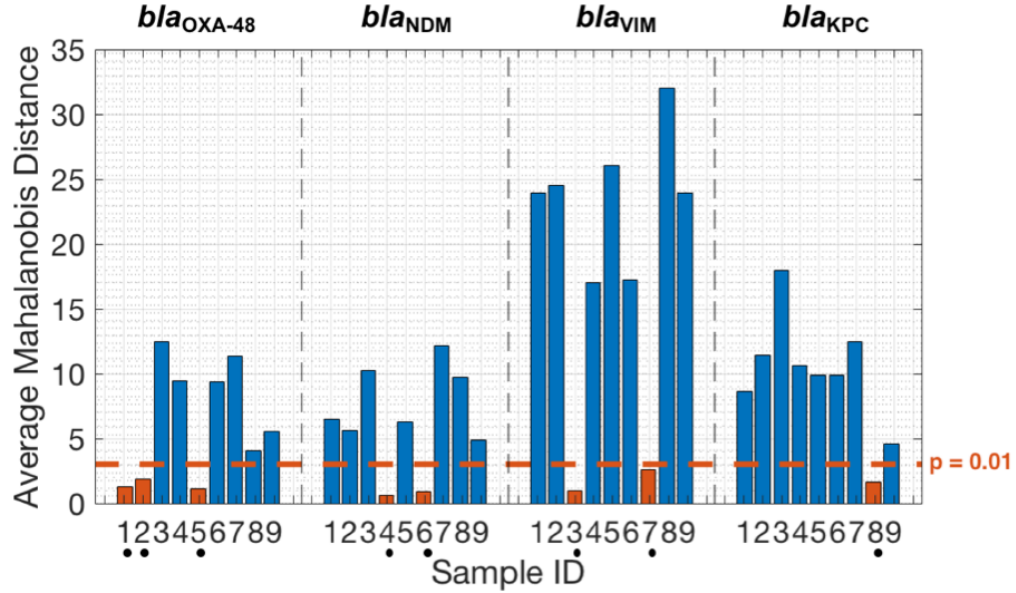


Figure 4.4: Average Mahalanobis distance between multidimensional standard curves and sample test points (bacterial isolates) used for target identification. Dots below sample ID indicate that the test sample is classified to the standard of interest with a p-value < 0.01.

In addition to observing the average Mahalanobis distance, it is important to visualise the data in order to confirm that the Mahalanobis distance is a suitable similarity measure. Figure 4.5 shows the Mahalanobis space for the four standards. This visualisation is constructed by projecting all data points onto an arbitrary hyperplane orthogonal to each multidimensional standard curve. When the training data points in the feature space are approximately normally distributed, then the distribution of the training data points in the Mahalanobis space is approximately circular - as seen in Figure 4.5. It can also be observed that the training points (synthetic DNA) from each standard curve are clustered together (i.e. not considered outliers) in its respective Mahalanobis space, however, they are considered outliers for the other MSCs. This corroborates the fact that there is sufficient information in the 3 chosen features to distinguish the 4 standard curves.

4.4.2 Enhanced Quantification

Given that multiplexing has been established for this case study, quantification can be trivially obtained using any conventional method such as the gold standard cycle threshold, C_t . However, as shown in Chapter 3, enhanced quantification can be achieved using a feature, M_0 , that

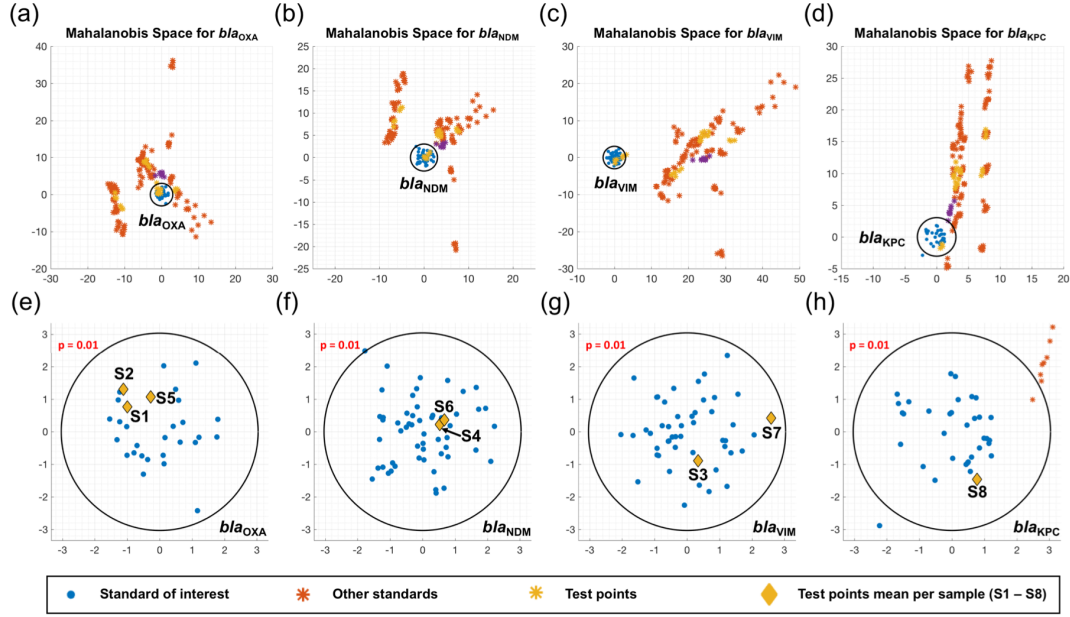


Figure 4.5: Multidimensional analysis using the feature space for identification of unknown samples. (a-d) All data points, including the replicates for each concentration for the four MSCs (training standard points) and nine unknown samples (test points), have been projected onto arbitrary hyperplanes orthogonal to each MSC. (e-h) The previous plots are magnified to visualise the location of the samples relative to each standard of interest. The blue dots represent the data points for each standard of interest (5 to 8 replicates per each concentration) and the black circle around them corresponds to a p-value of 0.01. Sample 1-8 are correctly classified with a p-value < 0.01 whereas sample 9 is considered an outlier for all standards. Please see Table B.S3 for details on the extracted features and sigmoidal curve fittings and Table B.S4 for the estimated quantification of samples using all methods.

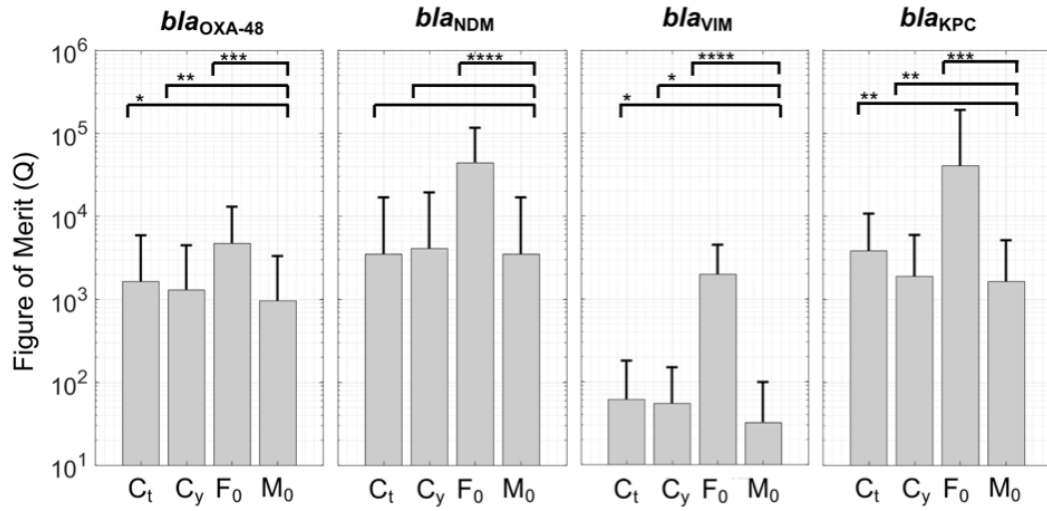


Figure 4.6: Figure of merit comparing conventional features with M_0 for absolute quantification. F_0 denotes $-\log_{10}(F_0)$. Bar plots represent the average figure of merit for each feature and the error bars indicate the standard deviation. Please see Figure B.S2 for details.

combines all of the features. This is enabled through weighting each feature by optimising an objective function and then applying a dimensionality reduction technique in order to create

a quantification curve for M_0 . The objective function in this study is a figure of merit, Q , that combines accuracy, precision, stability and overall predictive power as described in the experimental section. Figure 4.6 shows the average figure of merit (with standard deviation and p-values) for each target using the 3 chosen features (C_t , Cy_0 and $-\log_{10}(F_0)$) and M_0 . Please see Figure B.S2 for the quantification curves and details for the figure of merit. It can be observed from Figure 4.6 that quantification using M_0 performs as good, or better than any single feature, for any of the targets. This is expected given the nature of the multidimensional framework as M_0 is constructed using a linear combination of the other features in order to minimise the average figure of merit. It is important to stress that any figure of merit could be selected. For bla_{OXA-48} , bla_{VIM} and bla_{KPC} , the average figure of merit of M_0 was reduced by 26.5%, 41.1% and 12.9% compared with the best single feature - Cy_0 , with all p-values < 0.05 . Furthermore, for bla_{NDM} , the optimisation algorithm showed that M_0 converged to C_t and that the p-value between C_t and Cy_0 was not significant. Therefore, in this case, it is acceptable to use either C_t , Cy_0 or M_0 . However, M_0 provides an automated solution for quantification that is robust in the sense that it will always be the best performing method.

4.5 Conclusion

There has been a long-standing goal to meet the demand of methods for high-level multiplexing and enhanced quantification. Any advancement in this area would have a substantial positive impact on healthcare and patient outcomes. Alongside these challenges, there is a growing concern around antimicrobial resistance; the past 10 years has seen an explosion in molecular methods and instruments for rapid screening of drug resistant genes [93]. Here, a novel method is proposed that allows for simultaneous single channel multiplexing and enhanced quantification using existing technologies; without an increase in complexity or cost over using conventional single-plex qPCR reactions for detecting multiple targets. This method has been validated for the detection of four of the most prominent carbapenem-resistant genes: bla_{OXA-48} , bla_{NDM} , bla_{VIM} and bla_{KPC} .

Methods for multiplexing and quantification typically involve using fluorescent probes, melting curve analysis, agarose gels or sequencing - all of which are time-consuming or expensive processes. In the last few years there have been attempts to achieve simultaneous multiplexing and quantification in a single channel. For instance, single channel multiplexing has been achieved without melting curve analysis by altering cycling conditions and reading fluorescence at different temperatures [94]. This resulted in sufficient information gain in order to discriminate two targets. However, this method still uses a unidimensional approach to data analytics and increasing the number of targets is not trivial. By extracting multiple features from exist-

ing data, the proposed methodology represents an opportunity to evolve existing approaches in order to significantly increase the number of targets.

In order to implement this methodology, the following are required: (i) to build multiple multidimensional standard curves. This is generally a one-time procedure; however, MSCs may be affected by variations between experiments (such as changing reagent batches or instruments). Therefore, as with conventional standard curves, MSCs may have to be eventually re-calibrated; (ii) to design multiplex assays such that MSCs for each target are sufficiently distant in the feature space. This can be achieved by tuning reaction conditions or primer design. For example, by altering annealing temperature, amplicon length, introducing mismatches or primer mix concentration; and (iii) to perform data processing (such as multi-feature extraction) which is negligible given the power of computers today.

In addition, given the nature of the multidimensional framework, absolute quantification is enhanced through the use of M_0 by optimising a figure of merit combining accuracy, precision, stability and overall predictive power. This motivates the adoption of M_0 , as an alternative to conventional standard curves for absolute quantification, as it guarantees improved performance by combining the benefits of all the features it is derived from. This property results in M_0 offering a robust method of quantification in the sense that it provides the best quantification performance across targets. Furthermore, the capabilities of MSCs extend beyond quantification and allow for outlier detection and target identification.

Given the novelty of this work, there are many future directions and questions that can be addressed. In this Chapter the proposed method was applied to the rapid screening of the four most prominent carbapenemase genes in the UK. In future studies, it would be interesting to explore: other targets in order to develop new multiplexing panels associated with the most significant healthcare challenges; or more targets through incorporating additional MSCs into the feature space and/or using multiple fluorescent channels. It is also important to stress that the focus of this work was not on optimising the chemistry or data analytics for this specific set of targets. Thus, there is room to investigate whether the chemistry and the instance of framework can be optimised in order to maximise the separation of MSCs in the feature space for carbapenem-resistant genes.

In conclusion, this work has shown that is possible to simultaneously quantify and multiplex several targets in a single channel. This is achieved by changing the way amplification data obtained from existing technologies is analysed. I hope that by sharing these ideas, researchers and practitioners can implement and advance this work in order to provide novel and affordable tools that can be easily adopted by healthcare systems.

Part II:

Supervised Machine Learning for Digital Instruments

“Once we know something, we find it hard to imagine
what it was like not to know it.”

CHIP & DAN HEATH

Chapter 5

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

5.1 Chapter Overview

Information about the kinetics of PCR reactions are encoded in the amplification curve. In Chapter 4 it was shown that this information is sufficient for multiplexing in conventional qPCR instruments. However, in digital PCR (dPCR), this information is typically neglected by collapsing each amplification curve into a binary output (positive/negative). Here, it is demonstrated that the large volume of raw data obtained from real-time dPCR instruments can be exploited to perform data-driven multiplexing in a single fluorescent channel using machine learning methods, by virtue of the target-specific information in the amplification curve. This new approach, referred to as *amplification curve analysis* (ACA), is shown using an intercalating dye (EvaGreen), reducing the cost and complexity of the assay and enabling the use of melting curve analysis for validation. As a case study, 3 carbapenem-resistant genes are multiplexed to show the impact of this approach on global challenges such as antimicrobial resistance. Contrary to the ideas in Chapter 4, ACA does not explicitly extract any features to build an MSC, and implicitly learns the relevant information by the nature of supervised machine learning algorithms. For this case study, and in the presence of single targets, a classification accuracy of 99.1% ($N = 16188$) is reported, which represents a 19.7% increase compared to standard dPCR multiplexing based on the final fluorescent intensity. Considering all combinations of amplification events (including co-amplifications), the accuracy was shown to be 92.9% ($N = 10383$). To support the analysis, a formula is derived to estimate the occurrence of co-amplification in dPCR based on multivariate Poisson statistics, which suggests that reducing the digital occupancy in the case of multiple targets improves accuracy. 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, it is expected that combining this method with existing probe-based assays will increase multiplexing capabilities significantly, and that once emerging point-of-care technologies can reliably capture real-time data from isothermal chemistries, the ACA method will facilitate the implementation of dPCR outside of the lab.

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

- A. Moniri*, L. Miglietta*, K. Malpartida-Cardenas, I. Pennisi, M. Cacho-Soblechero, N. Moser, A. Holmes, P. Georgiou and J. Rodriguez-Manzano*, “Amplification Curve Analysis: Data-driven Multiplexing using Real-Time Digital PCR”, *ACS Analytical Chemistry*, vol. 92, no. 19, pp. 13134–13143, Sept 2020. *These authors contributed equally.
- J. Rodriguez-Manzano, A. Moniri, L. Miglietta and P. Georgiou, “Identifying a target nucleic acid”, GB2013035.7 (Assignee: Imperial Innovations Limited). Priority 20-08-2020 / Filed 20-08-2020.

5.2 Introduction

Digital PCR (dPCR) is a well-established method to detect and quantify nucleic acids [29, 95]. It is based on the amplification of single target DNA/RNA molecules in many separate reaction wells. This approach offers several advantages over conventional real-time PCR (qPCR), such as: (1) lack of references or standards; (2) high precision in quantification; (3) tolerance to inhibitors; and (4) the capability to analyse complex mixtures [96, 97, 75]. Therefore, dPCR has enabled scientific breakthroughs in cancer biomarker discovery, genetic alterations and infectious diseases, among others [98, 99, 100].

As the need for high throughput analysis of multiple targets continues to escalate, several approaches have been proposed to simultaneously detect and quantify multiple nucleic acids. Microfluidic solutions offer spatial multiplexing by subdividing the reaction into an increasing number of partitions [101]. Alternatively, multiplex dPCR assays can rely on the use of fluorescent probes (e.g. TaqMan) using multiple channels. The probe concentration can be optimised to increase the number of targets per channel by changing the final fluorescent intensity (FFI) [102]. However, probes are expensive and require time-consuming optimisation [103]. To achieve similar multiplexing capabilities, dye-based approaches (e.g. EvaGreen) have also been proposed, which alter primer concentration, changing the PCR efficiency and impacting the FFI [104]. Furthermore, instruments with melting capabilities, such as Fluidigm’s BioMark

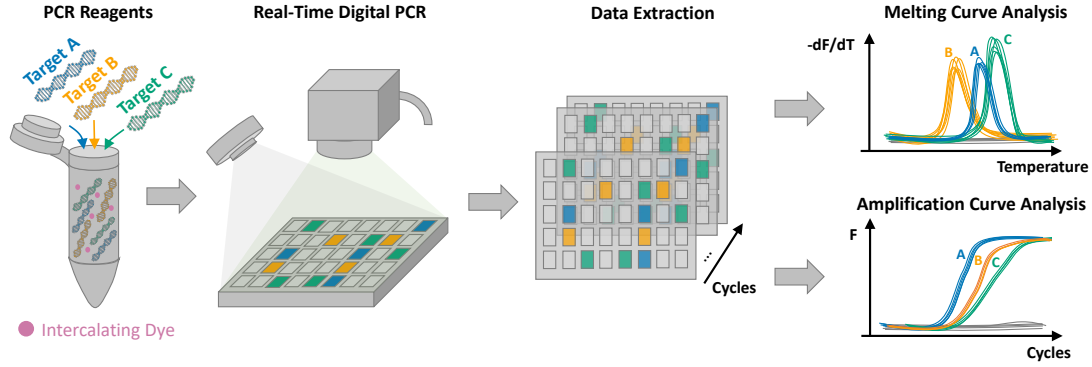


Figure 5.1: Experimental workflow. A multiplex PCR assay (with an intercalating dye) is developed for detecting targets A, B and C. Real-time digital PCR is used to perform single-molecule amplification to detect the targets. Melting curve analysis is used to validate the specificity of the amplification 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.

HD, can be used to perform melting curve analysis (MCA) when using intercalating dyes. The melting peak, T_m , is extracted to distinguish targets [105], although more recently, methods based on machine learning have also been proposed [106]. To date, there has been no report of multiplexing using the entire amplification curve in dPCR.

In Chapter 4, it was shown that in qPCR, sufficient information exists within the amplification curve which can enable it to distinguish several targets using multidimensional standard curves [47, 90]. 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, real-time single-molecule digital PCR using an intercalating dye and machine learning models 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). MCA is used as the “gold standard” to assess the performance of the proposed method, as illustrated in the experimental workflow depicted in Figure 5.1. This new method takes 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 in order to develop machine learning models, without explicitly extracting features of the amplification curve or compromising the assay performance (by modifying probe or primer concentration). Moreover, the amplification curves are normalised for the FFI, to show that this method can be combined with current approaches for dPCR multiplexing - breaking the barrier of 1 target for each level of FFI (in a given fluorescent channel). Finally, a theoretical derivation for the likelihood of multiple targets in a single well (i.e. co-amplification) is presented to understand the effect of this phenomenon on quantification and multiplexing.

As a clinically relevant application, the proposed method is applied to the global challenge of antimicrobial resistance [107]. In particular, the chosen targets are carbapenemases, which are β -lactamases (*bla*) that are resistant to the carbapenems, a class of highly effective antibiotic agents [108]. Therefore, a multiplex assay for the detection of 3 common carbapenem-resistant genes, namely *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC}, is also developed.

The vision for this work is three-fold: (1) maximise the capabilities of existing instruments and chemistries by extracting more information from existing data; (2) combine this approach with existing probe-based methods to increase multiplexing capabilities significantly; and (3) translate this methodology to isothermal chemistries and emerging point-of-care technologies to facilitate the implementation of dPCR outside of the lab [109].

5.3 Experimental Section

For reproducibility purposes, the DNA targets, experimental conditions and data analysis are described below. The reader may wish to skip directly to the Results and Discussion (section 5.4).

5.3.1 DNA Templates

Double-stranded synthetic DNA (gBlock Gene fragments) containing *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC} gene sequences (ranging from 801 to 917 bp). The sequences of these genes were downloaded from the GenBank web site (<http://www.ncbi.nlm.nih.gov/genbank/>) with accession numbers of NC_023908, NC_023274 and NC_014312 for *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC}, respectively. These genes belong to the class B metalloenzymes encoding *bla*_{NDM} and *bla*_{VIM}, plus the class A carbapenemases encoding *bla*_{KPC} type. They were purchased from Life Technologies and re-suspended in Tris-EDTA buffer to 10 ng/ μ L 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).

5.3.2 PCR Primer Design

Primers for the multiplex assay were designed to target the aforementioned referenced sequences. For each gene of interest, 1000 sequences were retrieved from NCBI blast (*in-silico*), to identify all the possible inclusive targets and exclude potential cross-reactivity sequences.

Alignments were performed using the MUSCLE algorithm [110], in Geneious Prime[®] 2020.1.2 (<https://www.geneious.com>) [111]. 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 [112]. The T_m of the amplification product of each primer set was determined by the Melting Curve Predictions Software (uMELT) package (<https://dna-utah.org/umelt/umelt.html>) [113]. All primers were synthesised by Life Technologies. Primer sequences are listed in Table 5.1.

Table 5.1: Primers Specification

Target	Primer Name	Sequence (5'-3')	Amplicon size (bp)
<i>bla</i> _{NDM}	NDM-F	CACACCAGTGACAATATCACCGTTG	85
	NDM-R	ACTTGGCCTTGCTGTCCTTGAT	
<i>bla</i> _{VIM}	VIM-F	CTTCGGTCCAGTAGAACTCT	258
	VIM-R	GTGTGCTTGAGCAAGTCT	
<i>bla</i> _{KPC}	KPC-F	TCGAACAGGACTTTGGCG	202
	KPC-R	GGAACCAGCGCATTTTTCG	

Primers have been developed in this study.

5.3.3 PCR Reaction Conditions

Real-time PCR. Each amplification reaction was performed in 10 μ L of final volume with 5 μ L of SsoFast EvaGreen Supermix with Low ROX (BioRad, UK), 3 μ L of PCR grade water, 1 μ L of 10 $\times$ multiplex PCR primer mixture containing the three primer sets (5 μ M of each primer), and 1 μ L of different concentrations of synthetic DNA (or controls). PCR amplifications consisted of 10 min at 95 $^{\circ}$ C, followed by 45 cycles at 95 $^{\circ}$ C for 20s, 65 $^{\circ}$ C for 45s, and 72 $^{\circ}$ C for 30s. In order to validate the proposed method, the results were compared against melting curve analysis. Melting curve analysis was performed with one cycle at 65 $^{\circ}$ C for 60s, and reading from 65 to 97 $^{\circ}$ C with an increment of 0.5 $^{\circ}$ C. The PCR machine used in this study was the Light Cycler 96 Real-Time PCR System (Roche Diagnostics).

Real-time Digital PCR. Each amplification reaction was performed in 4 μ L of final volume with 2 μ L of SsoFast EvaGreen Supermix with Low ROX (BioRad, UK), 0.4 μ L of 20 $\times$ GE Sample Loading Reagent (Fluidigm PN 85000746), 0.3 μ L of PCR grade water, 0.2 μ L of 20 $\times$ multiplex PCR primer mixture containing the three primer sets (10 μ M of each primer), and 1.2 μ L of different concentrations of synthetic DNA (or controls). PCR amplifications consisted of a hot start step for 10 min at 95 $^{\circ}$ C, followed by 45 cycles at 95 $^{\circ}$ C for 20s, 65 $^{\circ}$ C for 45s, and 72 $^{\circ}$ C for 30s. The results were validated using MCA which was performed with one cycle at 65 $^{\circ}$ C for 3s and reading from 65 to 97 $^{\circ}$ C with an increment of 0.5 $^{\circ}$ C. The integrated

fluidic circuit controller is used to prime and load qdPCR 37K[™] digital chips and Fluidigm's Biomark HD system to perform the dPCR experiments, following manufacturer's instructions. More specifically, each digital chip contains 48 inlets, where each inlet is connected to a panel consisting of 770 wells (0.85nL well volume) [114]. In this study, 3 digital chips were used, totalling 144 panels, including negative controls.

5.3.4 Data Analysis

Multiple in-house Python (v3.7) scripts were developed to extract and analyse the data using standard data science packages (e.g. NumPy, Pandas and Scikit-Learn). Complete details of the code can be found at www.github.com/am5113/pyACA or Appendix C. All graphics are made using the Matplotlib package and optimised for colour blindness [115]. The logistic regression and k-nearest neighbour models were implemented using the scikit-learn package with default parameters (for more information please see provided code and package documentation). The classification accuracy (i.e. proportion of correctly identified events), sensitivity (i.e. true positive rate) and specificity (i.e. true negative rate) values in Table 5.2 and 5.3 were computed for each binary classification sub-problem in the one-vs-one multi-class classification scheme.

5.4 Results and 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 target-specific information is visualised using unsupervised machine learning. This motivates 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.

5.4.1 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 are used to 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 optimisation to reliably distinguish the amplification products [102, 116].

First, a 3-plex assay for the detection of *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC} is developed. Figure 5.2 (A) shows the amplification curves and melting peaks for each target at concentrations ranging from 5×10^3 to 1×10^6 copies/reaction. Observe the melting peaks for *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC} can be distinguished from each other and are given as 84.7°C , 88.5°C and 89.7°C respectively. Moreover, Figure 5.2 (B) shows the corresponding standard curves illustrating the *Ct* 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) [18]. 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 5.2 (C) shows the amplification curves for the co-presence of targets and Figure 5.2 (D) 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 2 distinct peaks in the melting profile. However, the mixture containing *bla*_{VIM}+*bla*_{KPC} results in only a single peak. This is also observed in the mixture with all 3 targets as only 2 peaks are evident. This may suggest there are fewer amplification products. Through adding the pure melting profiles, the ‘expected’ melting curve for mixtures of products can be estimated, as in Figure 5.2 (E). Observe that for *bla*_{VIM} and *bla*_{KPC}, the expected melting curve only predicts a single peak. This demonstrates the uncertainty as to whether the single peak contains 1 or more products - representing one of the major challenges with using MCA for multiplexing in the presence of more than one target. Therefore, the user is forced to run post PCR analysis techniques such as gel electrophoresis or sequencing. Figure 5.2 (F) shows the gel electrophoresis image for the same reactions as above. It can be observed that each reaction contains the same number of bands as the expected number of targets at

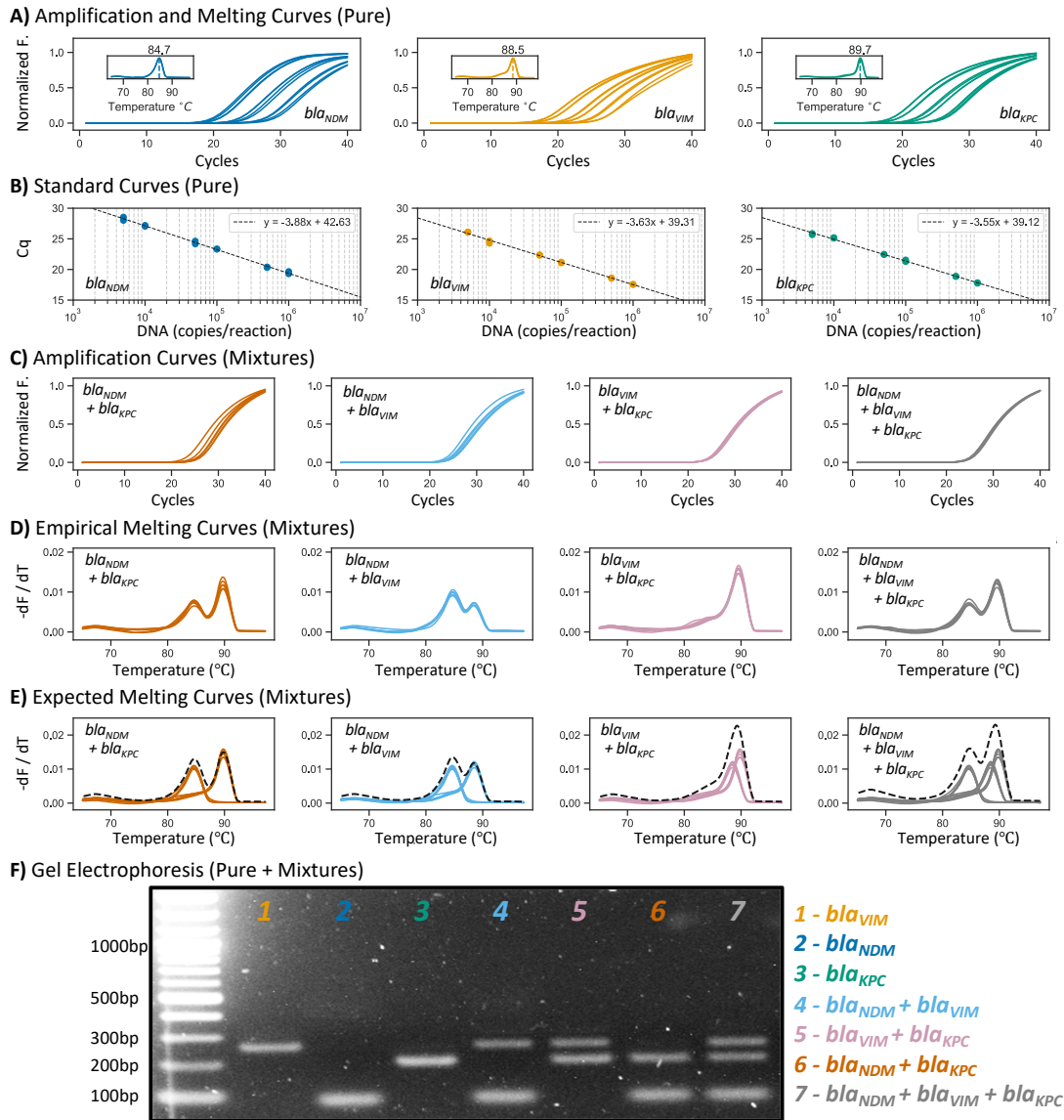


Figure 5.2: Real-time PCR Experiments showing the performance of a 3-plex assay in the presence of single and multiple targets. (A) Amplification curves for single targets (in a single reaction) with corresponding melting curves, where six different dilutions were used ranging from 5×10^3 to 1×10^6 copies/reaction. (B) Corresponding standard curves correlating the 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 correct amplicon length (see Table 5.1). Although gel electrophoresis can resolve 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 [117].

In Chapter 4, 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 [47, 90]. However, that 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.

5.4.2 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 3-plex assay from qPCR is translated into real-time dPCR. First, the multiplex assay in the presence of pure targets (in each digital panel) is investigated. Figure 5.3 (A) and (B) show the digital pattern and amplification curves for a serial dilution of the 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, in order to capture kinetic information in both domains. In total, there were 36960 amplification events with 16188 positive reactions: bla_{NDM} ($N = 4589$), bla_{VIM} ($N = 5682$) and bla_{KPC} ($N = 5917$). It is interesting to observe the Ct values as a function of the target concentration as seen in Figure 5.3 (C) since there is a clear separation between the single-molecule and the bulk regions. In the bulk region, the panels are saturated and therefore the target can be quantified using a standard curve (as in qPCR), whereas the low concentrations form a digital pattern that can be quantified using Binomial and Poisson statistics [96]. 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 [96, 97, 75]. Figure 5.3 (D) shows the distribution of the melting curve peaks (Tm) for each target. The maximum likelihood value of Tm for bla_{NDM} , bla_{VIM} and bla_{KPC} is 84.8°C , 88.6°C and 89.9°C , respectively. All Tm values are within 0.2°C of their respective qPCR quantities. The width of the distributions are

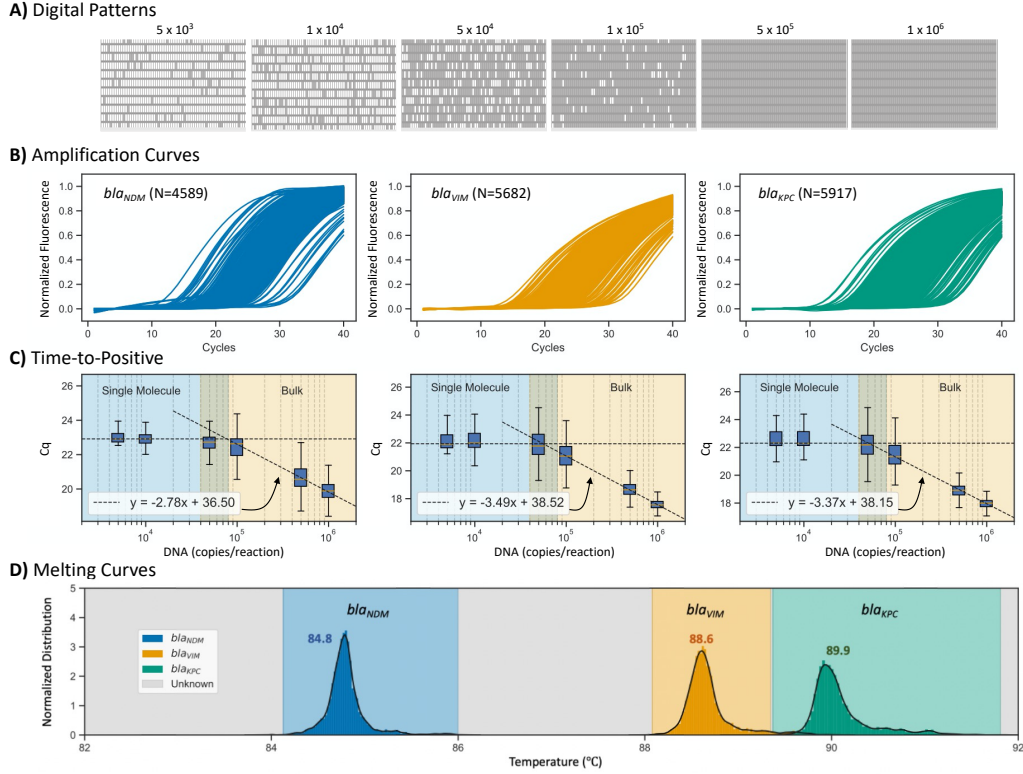


Figure 5.3: Real-time dPCR data. (A) Digital patterns for each panel at increasing concentrations. (B) Amplification curves for serial dilution of each target at concentrations ranging from 5×10^3 to 1×10^6 copies/reaction. (C) Standard curves correlating the C_t values with the concentration of each 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. All data presented in this figure were conducted in a single digital experiment.

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 which to distinguish the targets can be determined by considering the 1st and 99th percentile of T_m values. 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.

5.4.3 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 [116]. Figure 5.4 (A) shows the raw amplification curves with background subtraction. The associated FFI for each amplification event is shown in Figure 5.4 (B). It can be observed that there is an overlap between the distributions of FFI for the 3 targets. Based on these values, a machine learning model can be trained to learn the optimal boundaries to distinguish the targets. The dashed red

lines, T1 and T2, show the thresholds learned using a Logistic Regression classifier. Based on this classifier, the overall classification accuracy is computed as 79.42%, which is not adequate for many applications. This value is computed on the entire dataset (N=16188) based on a 10-fold cross-validation. More specifically, the model is trained on 14569 data points and tested on the remaining 1619. This process is repeated 10 times to provide the confusion matrix in Figure 5.4 (B) and details of the one-vs-one classifiers (which are combined to form the multi-class model) in Table 5.2. It can be observed that accurate multiplexing can be achieved for *bla*_{NDM} vs *bla*_{KPC} or *bla*_{NDM} vs *bla*_{VIM}, however the *bla*_{VIM} and *bla*_{KPC} are not separable, compromising the entire 3-plex. This demonstrates the need for optimising the reaction conditions (e.g. primer concentration) in order to obtain reasonable accuracy for 3 or more targets, due to the large variation of FFI values.

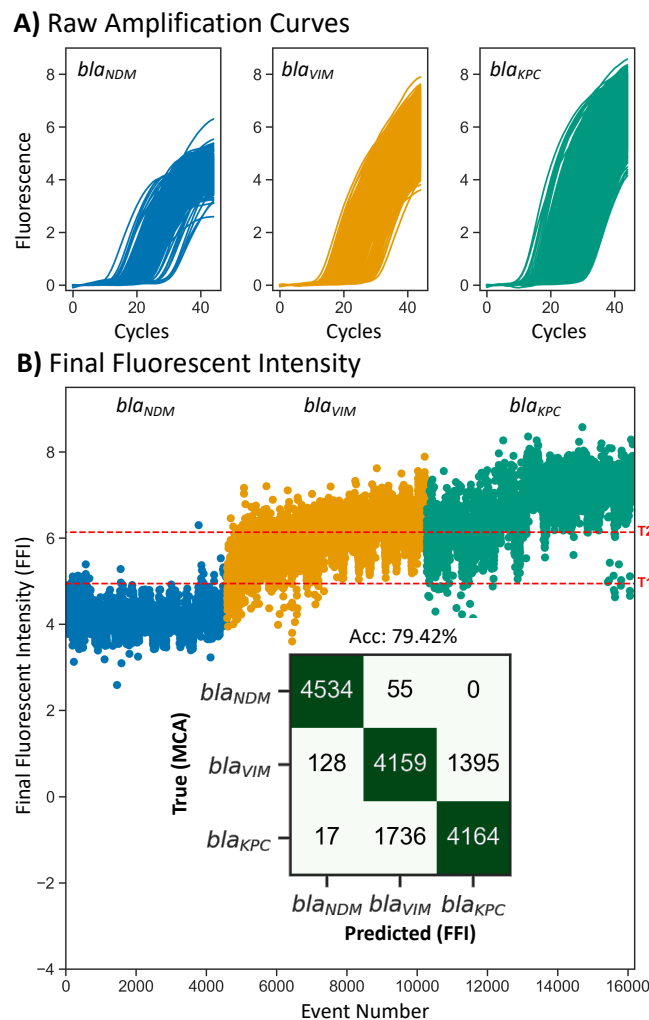


Figure 5.4: Multiplexing based on final fluorescent intensity. (A) Raw amplification curves with background subtraction. (B) Distribution of FFI values across the 3 targets. The red dashed lines (T1 and T2) indicate the thresholds generated from a Logistic Regression method for target classification. The predictions are shown in the overlay confusion matrix. Amplification events are ordered from low to high concentration.

Table 5.2: 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.

5.4.4 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, for a single digital chip, the melting data required 1.7Gb of extra memory (at just $0.5^{\circ}C$ resolution) in addition to the 1.1Gb required for the amplification data, translating to more time for acquisition & processing. Moreover, the MCA method cannot be extended to chemistries which are not compatible such as TaqMan assays or non-fluorescence based sensing. Therefore, the aim of this study is to explore a new method of multiplexing through the use of machine learning to extract specific kinetic information 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 visualisable 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

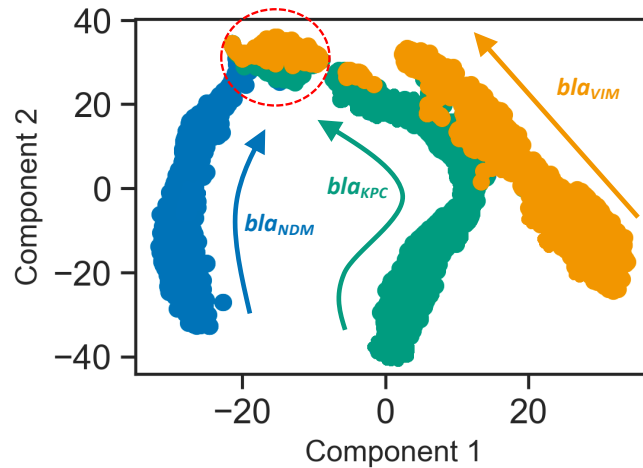


Figure 5.5: Visualising the similarity between amplification curves using the t-distributed stochastic neighbour embedding algorithm with 2 components and perplexity equal to 500. Direction of arrows indicate high to low concentration. Dashed red circle indicates curves that have not reached plateau.

neighbour embedding (t-SNE) algorithm, which has the ability to preserve the local structure [118]. It is important to understand that t-SNE is an unsupervised learning algorithm and therefore does not use the target labels. Figure 5.5 illustrates the t-SNE algorithm applied to the amplification curves with each target coloured for visualisation 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.

5.4.5 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 neighbours, support vector machines and deep neural networks [119, 92, 120]. The following section is demonstrated using the k-nearest neighbours algorithm which is a non-parametric method that is intuitive [121]. 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 5.3 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 5.6 (A) shows the prediction accuracy in a confusion matrix for the k-NN algorithm (for $k = 10$). The dark green squares indicate the single-target true positive classifications. The overall classification accuracy was 99.09% (CI: 99.08-99.09%). Moreover,

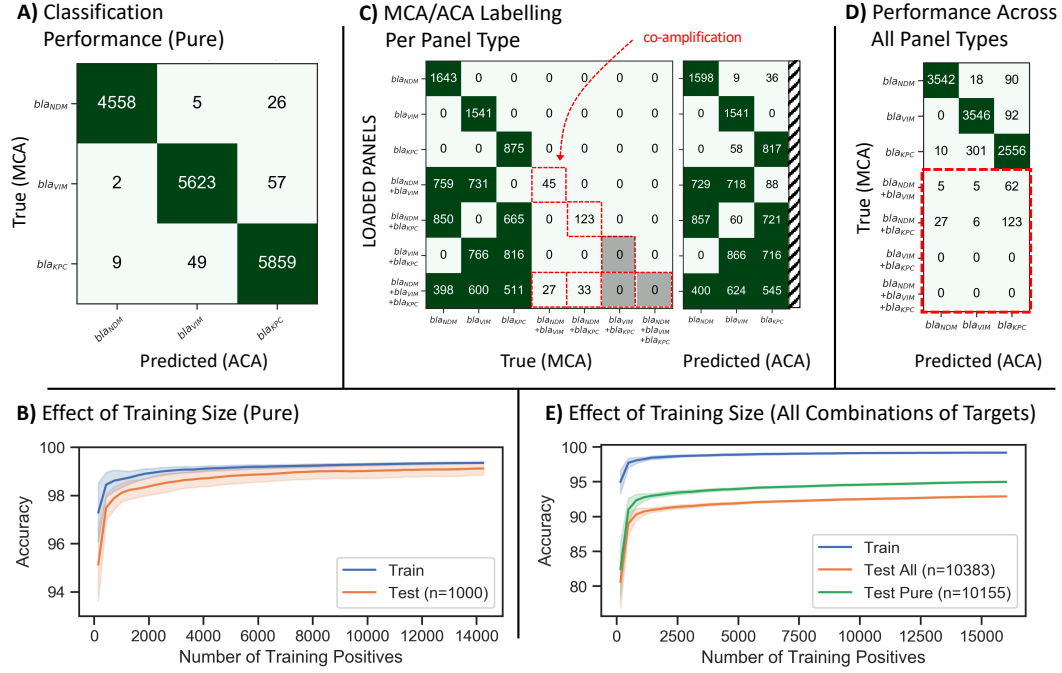


Figure 5.6: Performance of ACA in the presence of single and multiple targets. (A) Confusion matrix showing the predictions of ACA compared with MCA in the presence of single targets; (B) The effect of training size (with pure targets) on the train/test performance; (C) Matrices displaying the prediction of MCA and ACA per panel type; (D) Confusion matrix showing the predictions of ACA compared with MCA in the presence of multiple targets; (E) The effect of training size (with all combinations of targets) on the train/test performance. The data collected and analysed in this figure were conducted in 3 digital experiments on separate days. The confidence intervals were computed by bootstrapping 100 times using a stratified shuffle split and taking the standard deviation of the values.

the accuracy, sensitivity and specificity for the one-vs-one classifiers are given in Table 5.3. This result demonstrates that all 3 targets can be accurately distinguished from each other. Moreover, these results show the high concordance between MCA and ACA, suggesting that the amplification curve contains more information than commonly presumed. Compared to the FFI method, the overall classification accuracy was increased from 79.4% to 99.1%, representing a 19.7% increase in performance. Once again, this highlights the need to optimise the assay for the FFI method, which can be challenging in many applications.

Table 5.3: ACA Classification Performance

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

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

Volume of data required for accurate multiplexing. The volume of data required for training has significant practical implications in order to obtain a high test performance whilst

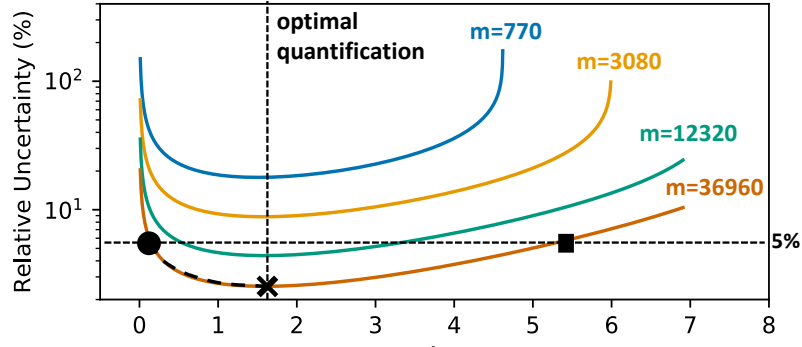
reducing the number of experiments. Figure 5.6 (B) shows the test accuracy for 1000 samples as a function of the number of training samples. This was computed through bootstrapping 100 times using a stratified shuffle split. As expected, as the number of training data increases, the out-of-sample performance increases. More interestingly, with as little as 100 training samples, the performance is at 95%, and increases to 98% before 1000 training samples.

Performance in the presence of multiple targets. Although theoretically with sufficient number of wells the challenges of multiple targets are mitigated, in reality the likelihood of co-amplification exists. Moreover, from a practical perspective, the training data is conducted in a different experiment to the test data, raising the possibility of inter-experiment variations. In this section, the data in Figure 5.3 is used as the training samples, and a different experiment which contains all possible combinations of the targets is used as test data.

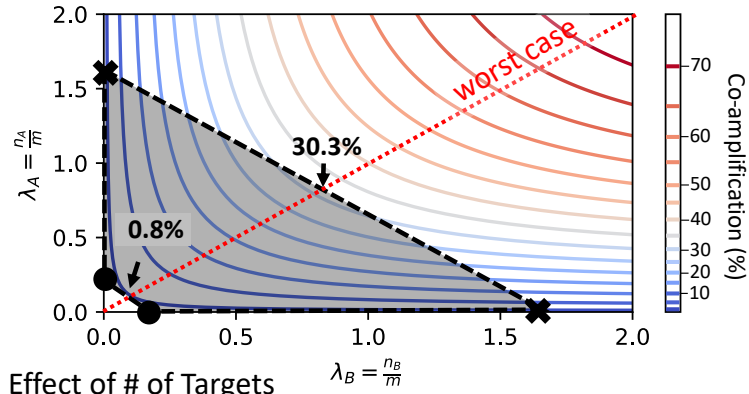
Figure 5.6 (C) illustrates the number of positives for each panel class, as determined by MCA and ACA. The dashed red boxes illustrate the co-amplification events. In total, 228 co-amplification events were observed. Moreover, the shaded boxes indicate the events where it is expected that co-amplification occurs, 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 5.6 (D). The overall classification accuracy (including the co-amplification events indicated by the dashed red line) is computed as 92.9%. However, only considering pure events yields an accuracy of 95.0%. Figure 5.6 (E) displays the accuracy for both, pure and all, amplification events as a function of the volume of training data. It can be observed the accuracy plateaus within 1000 training samples. The error due to the co-amplification events can be mitigated further by increasing the number of wells (or equivalently decreasing the digital occupancy).

Understanding the impact of co-amplification events. Quantification in dPCR is performed based on Binomial & Poisson statistics in order to estimate the number of molecules taking into account the probability of double, triple, etc. events [96]. 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 5.7 (A) shows the quantification precision as a function of the occupancy (based on the Wilson confidence interval). It can be observed that the optimal occupancy across all m is approximately 80% or $\lambda = 1.6$ (marked with a cross). However, an acceptable range of digital occupancy can be defined according the desired

A) Confidence Interval for Poisson Quantification



B) Probability of Co-amplification



C) Effect of # of Targets

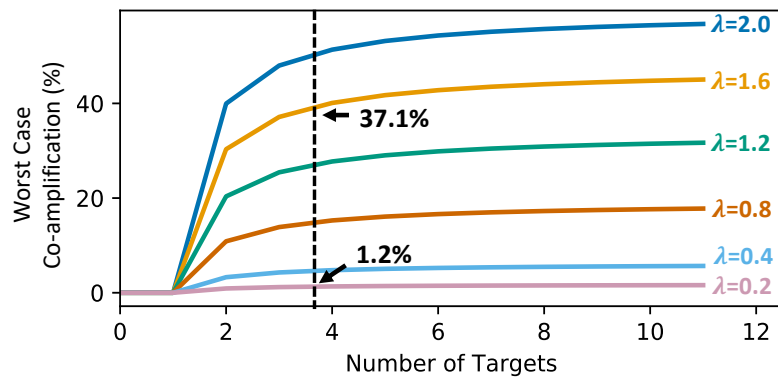


Figure 5.7: The impact of co-amplification events. (A) The relative uncertainty of Poisson quantification as a function of λ and the number of wells. (B) The probability of co-amplification in the presence of 2 targets. (C) The effect of the number of targets on the worst-case probability of co-amplification.

accuracy for a given application. For example, under the constraint of $m = 36960$ (number of wells in a Fluidigm[®] 37K chip), the uncertainty is below 5% between 16.7% occupancy ($\lambda = 0.2$, marked with a circle) and 99.3% occupancy ($\lambda = 5.0$, marked with a square).

Here, 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

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (5.1)$$

$$\lambda = \frac{n}{m} \quad (5.2)$$

Where n is the number of DNA molecules and m is the number of wells. Let $p(\mathbf{k})$ denote $p(k_1, \dots, k_K)$, 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

$$p(\mathbf{k}) = p(k_1) \dots p(k_K) \quad (5.3)$$

$$= \prod_{i=1}^K p(k_i) \quad (5.4)$$

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

$$P_0 = \prod_{i=1}^K p(k_i = 0) \quad (5.5)$$

$$P_1 = \sum_{j=1}^K \frac{p(k_j > 0)}{p(k_j = 0)} \prod_{q=1}^K p(k_q = 0) \quad (5.6)$$

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

$$P_C = 1 - \underbrace{\prod_{i=1}^K e^{-\lambda_i}}_{P_0} - \underbrace{\sum_{j=1}^K (e^{\lambda_j} - 1) \prod_{q=1}^K e^{-\lambda_q}}_{P_1} \quad (5.7)$$

Which can be simplified to

$$P_C = 1 - e^{-\lambda} \left(1 + \sum_{j=1}^K (e^{\lambda_j} - 1) \right) \quad (5.8)$$

$$\text{where } \lambda = \sum_{i=1}^K \lambda_i \quad (5.9)$$

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. $m \rightarrow \infty$, then $\lambda \rightarrow 0$, therefore $P_0 \rightarrow 1$ and $P_1 \rightarrow 0$, resulting in $P_C \rightarrow 0$. This demonstrates the error in ACA due to co-amplifications tends to zero as the number of wells increases. Figure 5.7 (B) 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 5.7 (C) shows the worst-case co-amplification proportion as a function of the number of targets. For 3 targets, P_C can be decreased from 37.1% down to 1.2% by reducing λ to 0.2. Moreover, the error starts to plateau (for all λ) above 6 targets. In fact, as the number of targets tends to infinity, P_C is equivalent to the probability of wells with more than 1 molecule independent of the number of targets. That is,

$$\lim_{K \rightarrow \infty} P_C = 1 - e^{-\lambda} - \lambda e^{-\lambda} \quad (5.10)$$

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

5.5 Conclusion

To date, no published study has reported dPCR (or droplet dPCR) multiplexing by considering the kinetic information encoded in the *entire* amplification curve. By leveraging the large volume of single-molecule data in real-time dPCR, a new data-driven method using supervised machine

learning is reported in this Chapter, referred to as amplification curve analysis, or ACA. This approach is validated on 3 drug-resistant genes: *bla*_{NDM}, *bla*_{VIM} and *bla*_{KPC}, by comparing to melting curve analysis as the “ground truth”. Although MCA is not ideal due to merging of peaks, it remains the only post PCR method to validate dPCR amplification products.

The results show that in the presence of single targets, the accuracy of ACA is 99.1% when training and testing within a digital experiment. This represents an improvement of 19.7% compared to the conventional method of multiplexing based on the final fluorescent intensity. Furthermore, when training and testing across experiments, 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, that suggests reducing the digital occupancy in the case of multiple targets in the same digital panel. This work raises several questions for future studies: (1) How robust is the ACA method to variations in reaction conditions? (2) Can the ACA be extended to perform higher-level multiplexing by considering the final fluorescent intensity? (3) Can this method be translated to probe-based or pH-based methods? (4) Can co-amplification events and non-specific products be classified using probabilistic models?

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

Chapter 6

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

6.1 Chapter Overview

Digital polymerase chain reaction (dPCR) is a mature technique that has enabled scientific breakthroughs in several fields. However, this technology is primarily used in research environments with high-level multiplexing representing a major challenge. In this Chapter, the concepts of amplification curve analysis (ACA) in Chapter 5 are extended to a new method, referred to as *amplification and melting curve analysis* (AMCA), which leverages the kinetic information in real-time amplification data and the thermodynamic melting profile using an affordable intercalating dye (EvaGreen). The method trains a system comprised of supervised machine learning models for accurate classification, by virtue of the large volume of data from dPCR platforms. As a case study, a new 9-plex assay is developed to detect 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 filing:

- [A. Moniri*](#), [L. Miglietta*](#), A. Holmes, P. Georgiou and J. Rodriguez-Manzano, “High-level Multiplexing in Digital PCR With Intercalating Dyes by Coupling Real-Time Kinetics and Melting Curve Analysis”, *ACS Analytical Chemistry*, vol. 92, no. 20, pp. 14181-14188,

Sept 2020. *These authors contributed equally.

- J. Rodriguez-Manzano, A. Moniri, L. Miglietta and P. Georgiou, “Identifying a target nucleic acid”, GB2013035.7 (Assignee: Imperial Innovations Limited). Priority 20-08-2020 / Filed 20-08-2020.

6.2 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 [15, 122, 49, 90]. 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 [123, 96, 97, 75]. Therefore, dPCR has enabled scientific breakthroughs in clinical microbiology, gene expression and precision cancer research, among others [124, 100, 125].

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 [126, 47]. Current approaches based on fluorescent probes are expensive and require lengthy optimisation which is challenging for high-throughput applications [127, 128]. 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 [128, 116, 129]. 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 [128]. Reported studies showed that specific target identification could be achieved through adjusting primer concentration to modulate the FFI value [116]. However, extensive optimisation 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 optimisation, a new method called amplification curve analysis (ACA) was recently proposed, to extract target-specific kinetic information from real-time amplification data using

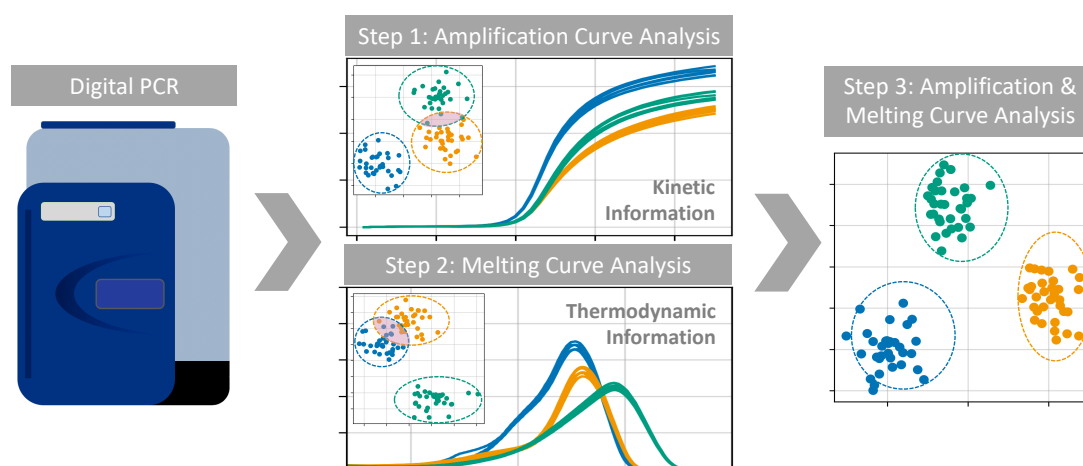


Figure 6.1: Concept of the proposed method. Amplification and melting curve data from real-time dPCR instrument (e.g. Fluidigm BioMark HD) is extracted. Subsequently, machine learning models are trained to classify multiple targets for both datasets individually. For high-level multiplexing, both methods provide insufficient accuracy (indicated by overlapping data distributions highlighted by the shaded regions). However, the proposed method, referred to as amplification and melting curve analysis, or AMCA, takes into account both kinetic and thermodynamic information in order to classify the targets accurately. Note: Three targets have been used to simplify the illustration of the concept.

supervised machine learning [130]. 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. Similar to ACA, high-level multiplexing with MCA also requires complex assay design to distinguish between close melting curve peaks.

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) [131]. Moreover, some studies have combined dPCR and melting curves, although they are restricted to end-point PCR which does not encode kinetic information [132, 106]. To date, there has been no report of enhancing multiplexing capabilities by combining amplification and melting curves.

In this Chapter, this concept is 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 6.1. The first step trains a model on the entire real-time amplification data and the second 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 [133]. In particular, colistin is a “last-line” antibiotic, reserved for the treatment of severe bacterial infections. The rise of mobilised colistin resistance (*mcr*) has been reported in over 40 countries across five different continents [134, 135, 108]. 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 [86, 136]. Incorrect diagnosis delays appropriate intervention, increases financial burdens for the healthcare system and complicates antimicrobial stewardship efforts [85]. Therefore, detecting variants of *mcr* is important to help treat and understand this emerging antimicrobial resistance. In this study, the first 9-plex PCR assay to detect *mcr*-1 to *mcr*-9 is developed.

The vision of this Chapter 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.

6.3 Experimental Section

For reproducibility purposes, the DNA targets, primers and experimental conditions are described below. The reader may wish to skip directly to the Data Analysis (section 6.3.3).

6.3.1 DNA Templates and Primer Design

Double-stranded synthetic DNA (gBlock Gene fragments) containing the entire coding sequences of *mcr*-1 to *mcr*-9 were used. The accession numbers from the NCBI GenBank web site (<http://www.ncbi.nlm.nih.gov/genbank/>) for each target are shown in Table 6.1. The gBlocks were purchased from Life Technologies (ThermoFisher Scientific) and re-suspended in Tris-EDTA buffer to 10 ng/μL stock solutions (stored at -80°C until further use). The concentrations of

all DNA stock solutions were determined using a Qubit 3.0 fluorimeter (Life Technologies).

To perform the (*in-silico*) design for the 9-plex, the first step was to conduct an NCBI blast (<https://blast.ncbi.nlm.nih.gov>) to ensure that each primer set binds to a conserved region. For each target, the blast was able to retrieve an average of 1000 sequences, which have been used to identify variation in the nucleotide sequence for all possible inclusive targets within the same gene and exclude potential cross-reactivity sequences (either within the *mcr* family or from a different species). Alignments were performed using the MUSCLE algorithm [110], in Geneious Prime® 2020.1.2 (<https://www.geneious.com>) [111]. Primer characteristics were analysed through the IDT OligoAnalyzer software using the J. SantaLucia thermodynamic table for melting temperature (T_m) evaluation [112]. 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 [137]. The T_m of the amplification product of each primer set was determined by the Melting Curve Predictions Software (uMELT) package (<https://dna-utah.org/umelt/umelt.html>) [113]. All primers were synthesised by Life Technologies (ThermoFisher Scientific). Primer sequences, amplicon length and GC content of the product are listed in Table 6.1.

Table 6.1: 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. (%)
<i>mcr</i> -1 (KP347127.1)	TGGCGTTTCAGCAGTCATTATGC	CAAATTGCGCTTTTGGCAGCTTA	516	50.0
<i>mcr</i> -2 (LT598652.1)	CTGTATCGGATAACTTAGGCTTT	ATACTGACTGCTAAATAGTCCAA	407	47.9
<i>mcr</i> -3 (KY924928.1)	AGACACCAATCCATTTACCAGTAA	GCGATTATCATCAAACCTCTTCT	136	47.1
<i>mcr</i> -4 (MF543359.1)	TTGCAGACGCCCATGGAATA	GCCGCATGAGCTAGTATCGT	207	45.4
<i>mcr</i> -5 (KY807921.1)	GGTTGAGCGGCTATGAAC	GAATGTTGACGTCACACTACGG	207	56.0
<i>mcr</i> -6 (MF176240.1)	GTCCGGTCAATCCCTATCTGT	ATCACGGGATTGACATAGCTAC	556	46.9
<i>mcr</i> -7 (MG267386.1)	TGCTCAAGCCCTTCTTTTCGT	TTGGCGACGACTTTGGCATC	466	56.2
<i>mcr</i> -8 (NG_061399.1)	CGAAACCGCCAGAGCACAGAATT	TCCCGGAATAACGTTGCAACAGTT	617	42.9
<i>mcr</i> -9 (NG_064792.1)	TATAAAGGCATTGCTTACCGTT	GGAAAGGCACTTTAGTCGTAAA	202	45.0

Primers have been developed in this study.

6.3.2 PCR Reaction Conditions

Real-time Digital PCR. Each amplification reaction was performed in 4 μ L of final volume with 2 μ L of 2 $\times$ SsoFast EvaGreen Supermix with Low ROX (BioRad, UK), 0.4 μ L of 20 $\times$ GE Sample Loading Reagent (Fluidigm PN 85000746), 0.4 μ L of 10 $\times$ multiplex PCR primer mixture containing the nine primer sets (5 μ M of each primer), and 1.2 μ L of different concentrations of synthetic DNA (or controls). PCR amplifications consisted of a hot start step for 10 min at 95 °C, followed by 45 cycles at 95 °C for 20s, 66 °C for 45s, and 72 °C for 30s. Melting curve analysis was performed with one cycle at 65 °C for 3s and reading from 65 to 97 °C with an increment of 0.5 °C. 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. More specifically, each digital chip contains 48 inlets, where each inlet is connected to a panel consisting of 770 wells (0.85nL well volume) [114]. In this study, 3 digital chips were used, totalling 144 panels (110880 wells), with experiments equally distributed across all *mcr* variants and negative controls. The number of positive reactions for each *mcr* variant is as follows: *mcr-1* (N=6767), *mcr-2* (N=6889), *mcr-3* (N=6159), *mcr-4* (N=6520), *mcr-5* (N=6424), *mcr-6* (N=6447), *mcr-7* (N=5919), *mcr-8* (N=6884) and *mcr-9* (N=6589).

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

6.3.3 Data Analysis

Multiplexing based on FFI. Final fluorescent intensity values were extracted from each amplification curve (as in [116]) 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, a high performance is not expected.

Amplification Curve Analysis, or ACA, consists of training a supervised machine learning model to distinguish targets based on the entire real-time amplification curve [130]. In this study, a deep neural network was chosen based on cross-validation. 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. The kernel size of 5 and 3 were chosen based on domain knowledge since the sampling period of the time-series is 1 minute and therefore the receptive field covers 7 minutes, which should capture the exponential, linear and plateau phases of the amplification curve. Moreover, the number of filters was chosen to be relatively small to keep the number of parameters low (both to keep training times low and to reduce chance of over-fitting). In total, the neural architecture consists of 3,305 parameters which is of comparable size to the

training set and can be trained in under 5 minutes on any modern computer. 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 global threshold.

Melting Curve Analysis, or MCA, consists of distinguishing the thermodynamic profile (i.e. $-\frac{dF}{dT}$) of the amplification product. In this study, and conventionally, this is achieved by distinguishing the melting peak, Tm , although methods have also been proposed to consider the entire curve [138, 139]. After peak detection, negative reactions can be confirmed by identifying curves with no peak. Subsequently, a supervised machine learning model can be trained to distinguish the Tm values. In this study, logistic regression was chosen as a classifier based on cross-validation.

The Proposed Method, termed 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 6.2. The output of ACA and MCA are probabilities for

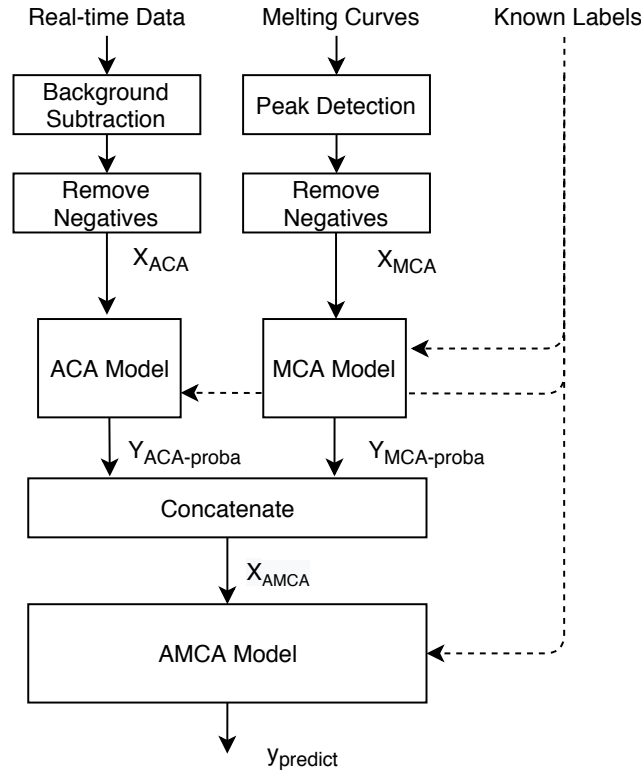


Figure 6.2: Flowchart to visualise the data processing workflow for the proposed method. Known labels (marked with a dashed line) are only required for training the models, as opposed to testing unknown samples. The input to the machine learning models are denoted as X_{ACA} , X_{MCA} and X_{AMCA} . The output probabilities of ACA and MCA are denoted as Y_{ACA} , Y_{MCA} . The final classification is given by $y_{predict}$. Vectors are given as lower case letters while matrices are upper case.

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.

6.3.4 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.

Data & Code availability - All data and code used in this study can be found at <https://github.com/am5113/pyAMCA>, with a brief description in Appendix D.

6.4 Results and Discussion

6.4.1 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 9-plex has been designed and validated using a conventional qPCR platform. Figure 6.3 (A)-(C) show the real-time amplification curves, melting peak distributions (extracted from melting curves) and standard curves for a serial dilution of each *mcr* target. Figure D.S1 and D.S2 show the raw melting curves before peak extraction and conventional standard curves, respectively. From Figure 6.3 (A), 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 6.3 (B), 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 6.3 (C) demonstrates that the multiplex assay is highly efficient (all > 95%) with a lower limit of detection (LoD) down to 10 copies per reaction for all targets (excluding *mcr*-9 which showed an LoD of 100 copies per reaction). All negative controls did not amplify before 45 cycles. The data

suggests that the co-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.

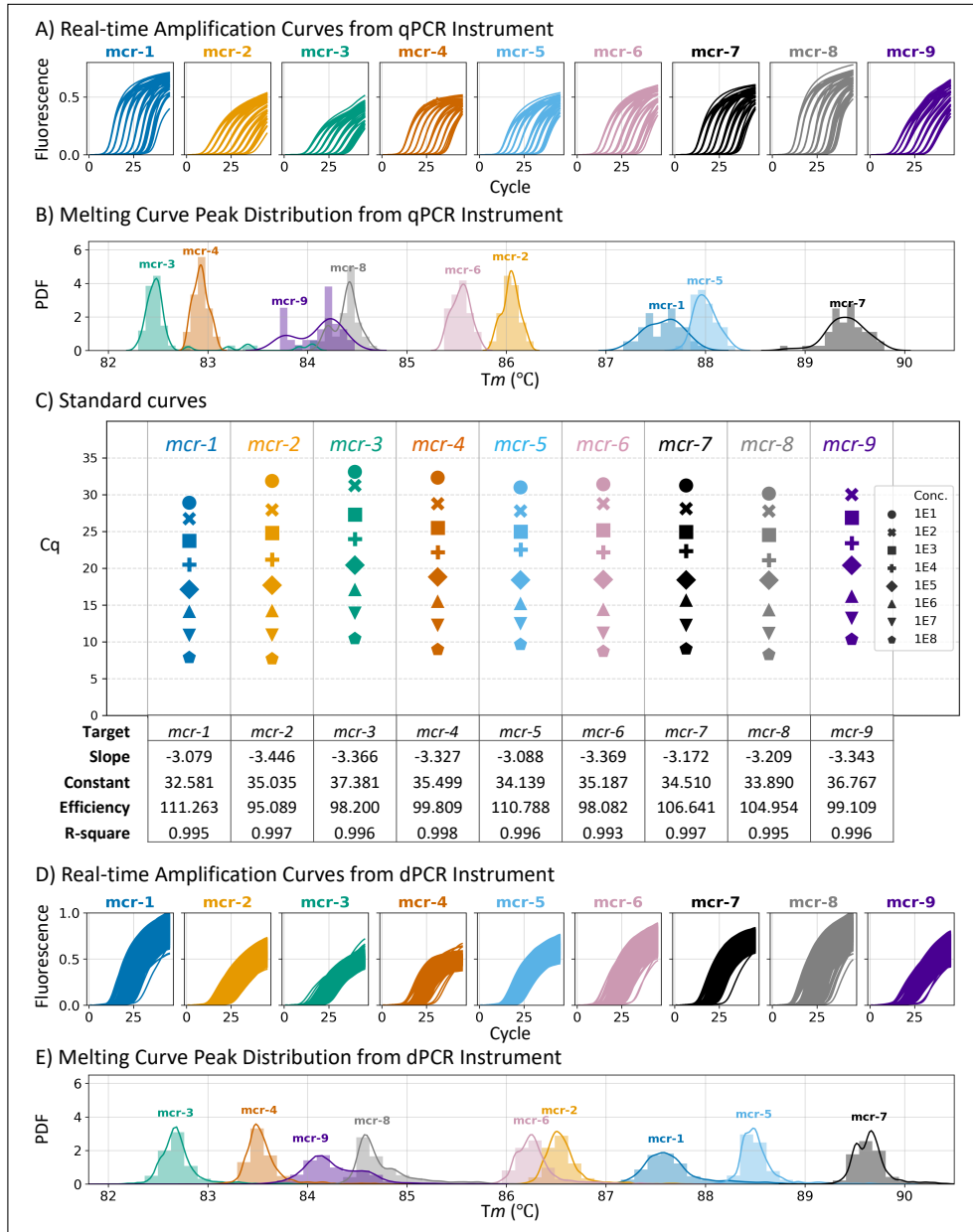


Figure 6.3: Analysis of real-time amplification and melting curves from qPCR and dPCR instruments. A) Real-time amplification curves from qPCR instrument. B) Melting curve peak distribution from qPCR instrument showing the probability density function (PDF) for each target. The mean $\pm$ std of *mcr-1* to *mcr-9* is $87.6 \pm 0.2^\circ\text{C}$, $86.0 \pm 0.1^\circ\text{C}$, $82.6 \pm 0.4^\circ\text{C}$, $82.9 \pm 0.1^\circ\text{C}$, $88.0 \pm 0.1^\circ\text{C}$, $85.5 \pm 0.1^\circ\text{C}$, $89.4 \pm 0.2^\circ\text{C}$, $84.4 \pm 0.1^\circ\text{C}$, $84.1 \pm 0.2^\circ\text{C}$, respectively. C) Visualisation and statistics of standard curves for a serial dilution of each target in qPCR using 9-plex assay. D) Real-time amplification curves from dPCR instrument. E) Melting curve peak distribution from dPCR instrument. The mean $\pm$ std of *mcr-1* to *mcr-9* is $87.7 \pm 0.3^\circ\text{C}$, $86.6 \pm 0.2^\circ\text{C}$, $82.7 \pm 0.2^\circ\text{C}$, $83.6 \pm 0.2^\circ\text{C}$, $88.5 \pm 0.2^\circ\text{C}$, $86.3 \pm 0.2^\circ\text{C}$, $89.7 \pm 0.2^\circ\text{C}$, $84.8 \pm 0.3^\circ\text{C}$, $84.3 \pm 0.3^\circ\text{C}$, respectively. Raw melting curves are shown in Figure D.S1

6.4.2 Classification accuracy of FFI, ACA and MCA 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 6.3 (D) and (E) 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 6.4 shows the FFI values for all amplification events grouped by target. It can be clearly observed that classification based on solely FFI values would be challenging. More specifically, a classification based on logistic regression yields an accuracy of $34.50 \pm 0.52\%$ (confusion matrix provided in Figure D.S3). As the results indicate, the FFI performance has low accuracy, although better than a random classifier (i.e. 11.1%), due to single-parameter usage, which contains little information specific to each target. Therefore, optimisation for primer concentration must be performed to achieve acceptable classification accuracy, as in McDermott *et al.* (2013), although this is neither trivial nor guaranteed for a 9-plex [116].

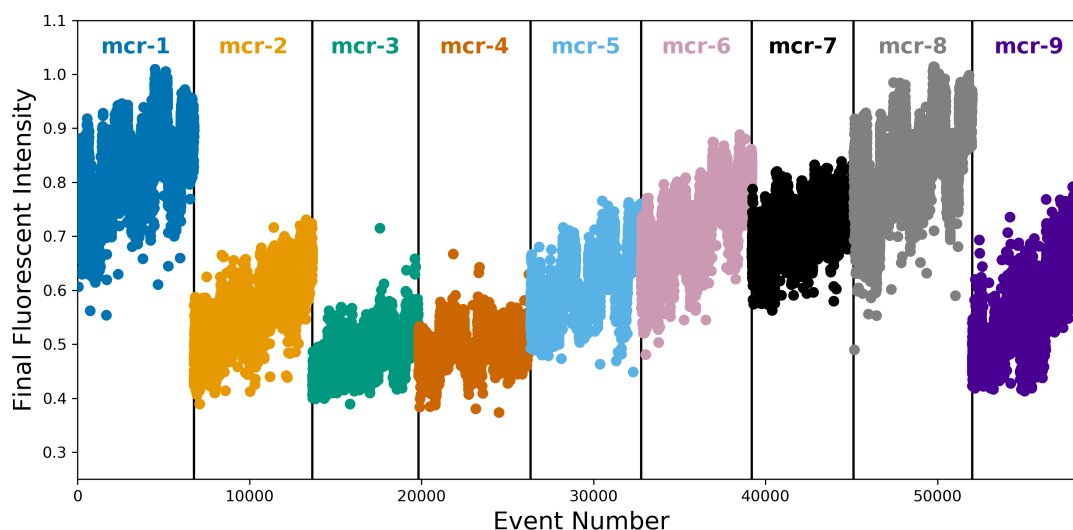


Figure 6.4: Distribution of FFI values for all targets and amplification events. For each target, amplification events are ordered from low to high concentration.

Figure 6.5 (A) and (B) show the confusion matrices, comparing the true and predicted targets for ACA and MCA, and the overall classification performance is $82.31 \pm 1.47\%$ and $89.34 \pm 0.33\%$, respectively. ACA, which was achieved by analysing the entire amplification curves using a neural network, boosts performance by 57.7% compared to the FFI method, by 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.

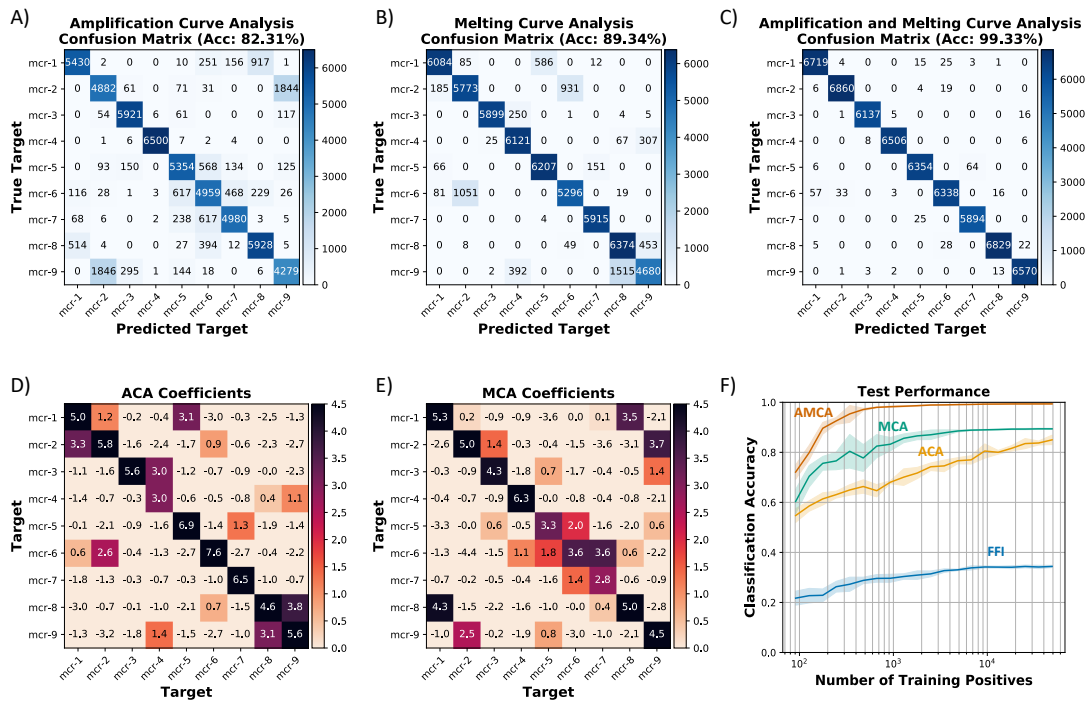


Figure 6.5: Performance of all methods for multiplexing the 9 *mcr* targets. A, B, C) Confusion matrices illustrating the predictions from ACA, MCA and AMCA (proposed method), respectively. Values indicate the number of amplification events with diagonal entries corresponding to correct predictions. 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.

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

Figure 6.5 (C) shows the confusion matrix comparing the predicted classification from the AMCA method to the true labels. It can be observed that the accuracy is $99.33 \pm 0.13\%$ and that no target is misclassified more than 1.7%, showing a significant improvement from ACA or MCA

individually (p -value $\ll 0.01$). Since the chosen supervised machine learning model for AMCA is linear, the coefficients can be investigated to understand how it weights the predictions from ACA and MCA. More specifically, the output of AMCA is defined by:

$$\mathbf{y} = \hat{\mathbf{W}}_{\text{ACA}} \mathbf{y}_{\text{ACA}} + \hat{\mathbf{W}}_{\text{MCA}} \mathbf{y}_{\text{MCA}} \quad (6.1)$$

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. Figure 6.5 (D) and (E) show the ACA and MCA coefficients in the form of a heatmap, respectively. It is interesting to observe that AMCA weights the prediction from ACA more heavily for targets which show poor classification in MCA, and vice-versa. For example, MCA misclassifies 1515 *mcr-9* reactions as *mcr-8*, therefore the AMCA positively weights the ACA prediction by 3.1 and negatively weights the MCA prediction by -2.1. Similarly, ACA misclassifies 1846 *mcr-9* reactions as *mcr-2* and the coefficients compensate for this phenomenon.

6.4.4 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 6.5 (F) shows the classification performance on 5000 out-of-sample data points (repeated 10 times) where n_{train} is between 1.0×10^2 and 5.3×10^4 for all models. It can be observed that all of the models perform better given more training data points. Since AMCA weights 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.

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

The same methodology (as in Figure 6.2) was applied to the qPCR data presented in Figure 6.3 (A) and (B). The classification accuracy for ACA, MCA and AMCA was shown to be $84.40 \pm 6.7\%$, $82.74 \pm 5.5\%$ and $95.98 \pm 3.4\%$, respectively. The confusion matrices for each method

and the model coefficients for AMCA are provided in Figure D.S4. These results suggest that the AMCA method works across real-time platforms, both quantitative and digital, although a further study to fully characterise the reliability in qPCR instruments (outside the scope of this study) is required.

6.5 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 instruments 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 optimisation.

On the other hand, the AMCA method requires training a supervised machine learning model which raises its own challenges. Firstly, since 3 models are required to be trained, especially if a neural network is used, this may take time and expertise in data science to perform. However, computational resources have negligible cost given the wide variety of open-source tools available for machine learning (such as *tensorflow* and *scikit-learn*). Secondly, it is important to ensure reproducibility of the experiment from a chemistry perspective in order for the training and testing data to be consistent. More specifically, if the instrument or laboratory approach show variability between experiments, then this needs to be accounted for from a data perspective (e.g. more data, pre-processing or data augmentation) or experimental procedures (i.e. consistent processes in the lab). However, since it was shown in this study that only 1000 amplification curves were required to achieve accurate multiplexing, it is possible to run training data within an experiment to avoid inter-experiment variations. For example, the Fluidigm qdPCR 37K digital chip contains 48 sample inlets (each connected to a panel of 770 wells), of which 9 panels can be used to generate the training data, one for each target. Assuming a digital occupancy of 80%, 9 panels translates to 5544 training data points, which based on Figure 6.5 (F), is expected to give an accuracy of 99.1%. From a practical point of view, this means that a single digital chip could accommodate screening 39 samples against 9 targets, whereas conventional spatial multiplexing (with single-plex assays) would only manage to screen 5 samples against the 9 targets.

As reported in a previous study, the ACA performance is degraded as a result of a phenomenon called ‘co-amplification’, which refers to the co-presence of multiple targets in a single chamber in dPCR instruments. This problem can be solved by keeping the occupancy of the digital panel (using Poisson statistics) within acceptable bounds in order to simultaneously reduce co-amplification and retain sufficient quantification precision. For example, for *mcr* genes, the vast majority of studies report the presence of a single *mcr* variant, and only few studies have reported the co-presence of 2 *mcr* variants in the same sample [140]. 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 [138, 139].

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

In conclusion, a new method for high-multiplexing in real-time digital PCR instruments with melting curve capabilities is proposed. This approach is based on training supervised machine learning algorithms to extract kinetic and thermodynamic information together, to enhance the classification accuracy in multiplexing. A classification accuracy of 99.3% was demonstrated for identifying 9 clinically relevant targets, namely mobilised colistin resistance, using a new multiplex assay based on an affordable intercalating dye. This represents the largest dPCR multiplex using intercalating dyes published in the literature to-date. Moreover, 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. 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. Extending this to conventional qPCR instruments and isothermal chemistries will be extremely beneficial for the wider scientific community.

Part III:

Intelligent Algorithms for Electronic Instruments

“We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision.”

ALAN TURING

Chapter 7

Robust Time-to-Positive from ISFET Arrays

7.1 Chapter Overview

Ion-sensitive field-effect transistor (ISFET) arrays, in conjunction with isothermal chemistries, are offering a promising alternative to conventional lab-based nucleic acid amplification instruments, owing it to their portable and affordable nature. Currently, the data analysis of ISFET arrays are restricted to established methods developed for qPCR instruments by averaging the output of every sensor to give a single time-series. This simple approach makes unrealistic assumptions, leading to insufficient performance for applications that require accurate quantification such as quantifying pathogen load or RNA host transcriptomics. In this Chapter, a novel data-driven algorithm for extracting time-to-positive values from ISFET arrays (fabricated in unmodified CMOS) is proposed, based on modelling sensor drift with adaptive signal processing and clustering sensors based on their behaviour with unsupervised learning methods. To validate this method, the global challenge of detecting bacterial infection is used as a case study, so as to take a step towards tackling antimicrobial resistance, using 22 clinical isolates (11 bacterial, 11 viral). First a 2-gene signature which has been reported to distinguish viral from bacterial infection using microarray analysis is validated. Subsequently, a novel reverse transcriptase quantitative loop-mediated isothermal amplification (RT-qLAMP) assay is developed for the 2-gene signature with high sensitivity, detecting down to 10 genomic copies per reaction within 25 minutes. Following this, the assay is translated onto an ISFET array device, and it is shown that averaging sensors leads to some of the reactions wrongly considered as negative, highlighting the need for a more robust algorithm. On the other hand, the proposed algorithm correctly outputs a time-to-positive for all the reactions, with a high correlation to RT-qLAMP (0.85, $R^2 = 0.98$, $p < 0.01$), and resulting in a classification accuracy of 100% (CI, 95 – 100%). By leveraging more advanced data processing methods for ISFET arrays, this work hopes to bridge the gap between translating assays from microarray analysis (expensive lab-based discovery method) to ISFET arrays (cheap point-of-care diagnostics).

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

- I. Pennisi, J. Rodriguez-Manzano, A. Moniri, M. Kaforou, J.A. Herberg, M. Levin and P. Georgiou, “Translation of a host blood RNA Signature distinguishing bacterial from viral infection into a platform suitable for development as a point-of-care test”, *JAMA Paediatrics*, Jan 2021.
- I. Pennisi, J. Rodriguez-Manzano, N. Miscourides, A. Moniri, N. Moser, D. Habgood-Coote, M. Kaforou, J. A. Herberg, M. Levin and P. Georgiou, “A method for determining a diagnostic outcome”, (Assignee: Imperial Innovations Limited). Priority 24-08-2020 / Filed 24-08-2020.

7.2 Introduction

Ion-sensitive field-effect transistor (ISFET) arrays, in conjunction with isothermal chemistries such as loop-mediated isothermal amplification (LAMP), are offering a promising alternative to conventional lab-based nucleic acid amplification instruments, owing it to their portable and affordable nature [141, 44]. These lab-on-a-chip platforms have made promising progress for point-of-need applications such as detecting malaria or mobilised colistin resistance, amongst others [46, 108, 142].

However, the ISFET is prone to several challenges which arise due to non-idealities of the sensor, stemming from the manufacturing process and experimental setup. These non-idealities are even more challenging when working with low-cost and accessible *unmodified* CMOS, i.e. sensors directly from fabrication, due to trapped charge and sensor degradation [143]. From a data perspective, these non-idealities can be broadly categorised by: (i) trapped charge, (ii) drift, (iii) temperature effects, and (iv) inherent noise. These effects are depicted in Figure 7.1.

- (i) *Trapped charge* induces an offset voltage in the sensor relative to the reference electrode, most likely due to charge transfer during chip production. This is particularly important if multiple sensors (such as an ISFET array) are used in conjunction with a single reference electrode, since each sensor will be operating in a different voltage range. The main consequence of this effect is that the reference voltage cannot be tuned for every sensor, and therefore, sensors may behave differently and some sensors may even fall out of the operating region and become inactive.
- (ii) *Drift* refers to the phenomena whereby the output of the sensor changes over time regardless of the input. This tends to be slow moving and linear in many sensing systems, such

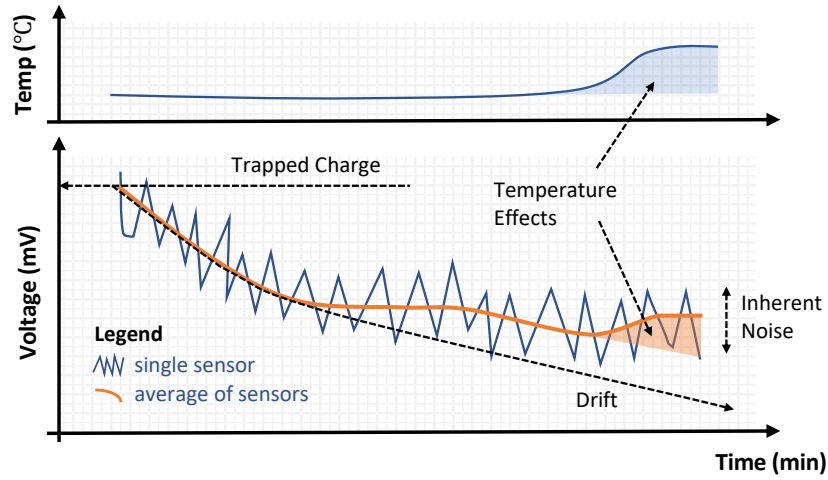


Figure 7.1: A data perspective of the sources of noise in ISFET sensors. These various non-idealities can be broadly classified as: (i) trapped charge - an offset in the signal; (ii) drift - a slow change in the output which occurs regardless of the input; (iii) temperature effects - changes in the output due to poor temperature regulation; and (iv) inherent noise - due to electronic and chemical noise.

as qPCR instruments, which can be compensated for. However, ISFETs are known to exhibit non-linear electronic *and* chemical drift. The electronic drift can be characterised as a function of the circuit and can be compensated for [39]. However, the chemical drift is a function of the pH itself, i.e. ions interacting with the insulator, rendering it very difficult to compensate for.

- (iii) *Temperature effects* can largely influence the output of ISFET sensors due to the underlying electron physics (and chemical effects). The effect of temperature on the drain-to-source current of the ISFET was measured in [144]. It has also been reported that the ISFET experiences 1pH change every 7°C around room temperature [145]. Clearly, performing thermocycling, as in PCR, poses a significant challenge for ISFETs, highlighting the advantage of isothermal chemistries.
- (iv) Other *Inherent noise* can be observed in ISFET sensors due to electronic and chemical noise. This includes flicker noise (i.e. a 1/f relationship in the frequency domain), where it has been shown that the flicker noise from chemical noise is one order of magnitude greater than its electronic counterpart [146].

Currently, the data analysis surrounding ISFET arrays consists of averaging the output of all sensors into a single time-series, and removing drift such that the output is similar to conventional lab-based instruments such as qPCR [46]. To this end, all the established methods for detecting amplification and computing a time-to-positive, such as the ‘gold standard’ cycle-threshold method or C_{y0} method can be used (see Chapter 3.2). However, sensor arrays in-

herently output a spatio-temporal signal and simply averaging sensors assumes that the sensors belong to the same data distribution. This is particularly important when fabricating ISFETs in unmodified CMOS technology due to significant trapped charge effects.

In this Chapter, a novel data-driven algorithm is proposed to combat the issues of drift and trapped charge. Since isothermal chemistries require simple thermal management, it can be assumed that the temperature is constant (and therefore temperature effects are negligible). The sample-to-result workflow is illustrated in Figure 7.2, where the focus of this Chapter is on the ‘Proposed Algorithm’. In particular, instead of globally compensating for the drift in order to find a ‘signal’, the data is modelled adaptively using ‘online learning’ methods, and the latent representation is used to represent ‘changes in behaviour’. Following this, the latent time-series for each sensor is clustered, so as to average sensors with the same behaviour. In order to validate the clustering, the clusters can be observed spatially to confirm the signal is not random. Although this algorithm does not directly model the spatio-temporal signal from ISFET arrays, it serves as a first step towards obtaining a robust signal from arrays which do not behave the same from experiment to experiment.

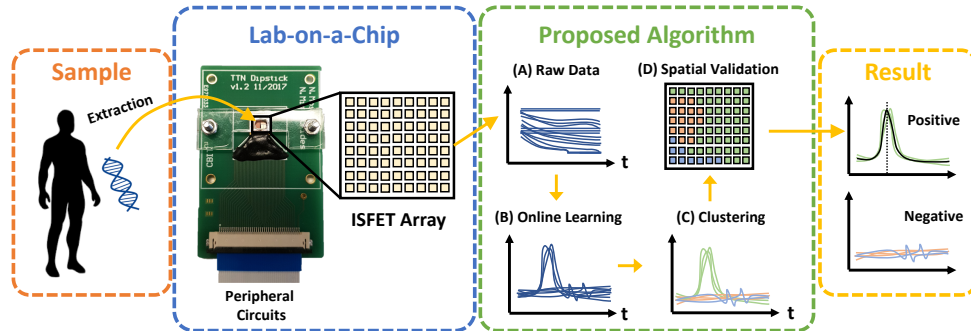


Figure 7.2: Depiction of the sample-to-result workflow using a lab-on-a-chip device with the proposed algorithm. First, an environmental or clinical sample is collected and the nucleic acid is extracted using an available kit. Subsequently, the sample is loaded onto the ISFET-based lab-on-a-chip platform with the necessary LAMP assay reagents. After heating to the required LAMP temperature, the amplification reaction begins. Each ISFET sensor yields a raw data time-series which is collectively input into the proposed algorithm. The algorithm consists of online learning, clustering and spatial validation, which is then used to determine a positive or negative result (and an associated time-to-positive if applicable).

As a case study, this method is applied to the detection of bacterial infection, so as to prevent the misuse of antibiotics. Recently, Herberg et al. (2016) reported 38 genes which are related to bacterial infection using microarray analysis [147, 148]. However, microarray analysis is an expensive lab-based method primarily used for discovery application, rather than cheap point-of-care settings. In an effort to translate the findings from microarray discovery to lab-on-a-chip (LoC) diagnostics, in this Chapter, 2 genes from the reported 38 genes are chosen to develop a specific reverse transcriptase LAMP (RT-LAMP) assay to detect them. This development is performed in a standard qPCR instrument, i.e. RT-qLAMP, which is then used on an ISFET-

based LoC platform yielding a method referred to as electronic RT-LAMP, or RT-eLAMP. To validate the approach, 22 clinical isolates (11 bacterial and 11 viral) are used to assess the classification performance using time-to-positive values extracted using the proposed algorithm.

The vision of this work is two-fold. The first vision is to show that the vast amount of rich data from ISFET sensor arrays enable a wide variety of algorithms, in order to provide more accurate results. And the second vision is to motivate a workflow from microarray analysis (expensive lab-based discovery) to lab-on-a-chip devices (cheap point-of-care diagnostics), in order to create a more efficient process for nucleic acid detection.

7.3 Experimental Section

7.3.1 Conventional Method

The conventional method of analysing amplification data from ISFET arrays typically revolves around averaging the signals from the sensors and subsequently processing the data (e.g. drift compensation). Therefore, herein, the conventional method is referred to as *average then process*, or ATP. The ATP method, as described in Malpartida et al (2019), is described below and the key assumptions which the proposed method aims to tackle are boxed in gray.

1. The signals from all ISFETs in contact with the chemical solution are averaged. The resulting signal is normalised by removing the DC component.

Assumption 1

The signal from all the sensors are drawn from the same distribution.

2. During the first few minutes of the chemical reaction, sensor drift is modelled using a stretched exponential function described by:

$$y[n] = e^{-t[n]^\beta} \quad (7.1)$$

Where $y[n]$ and $t[n]$ are the voltage and time values at sample n , and β is a scalar parameter. The drift is extrapolated across time, and subsequently subtracted from the signal to yield an amplification curve.

Assumption 2

The drift parameter β is stationary, or equivalently, β does not change with time.

3. The time-to-positive is extracted as described in Guescini et al (2008) via the $Cy0$ method.

Assumption 3

The signal after pre-processing can be modelled using a sigmoid.

7.3.2 Proposed Method

The proposed method, as depicted in Figure 7.3 (with Matlab code provided in Appendix E), follows 5 steps, namely: (i) pre-processing, (ii) adaptive signal processing, (iii) unsupervised clustering, (iv) spatial validation and (v) averaging & peak detection; where each step is described below. This algorithm will be compared against the ATP method in Chapter 7.4.3.

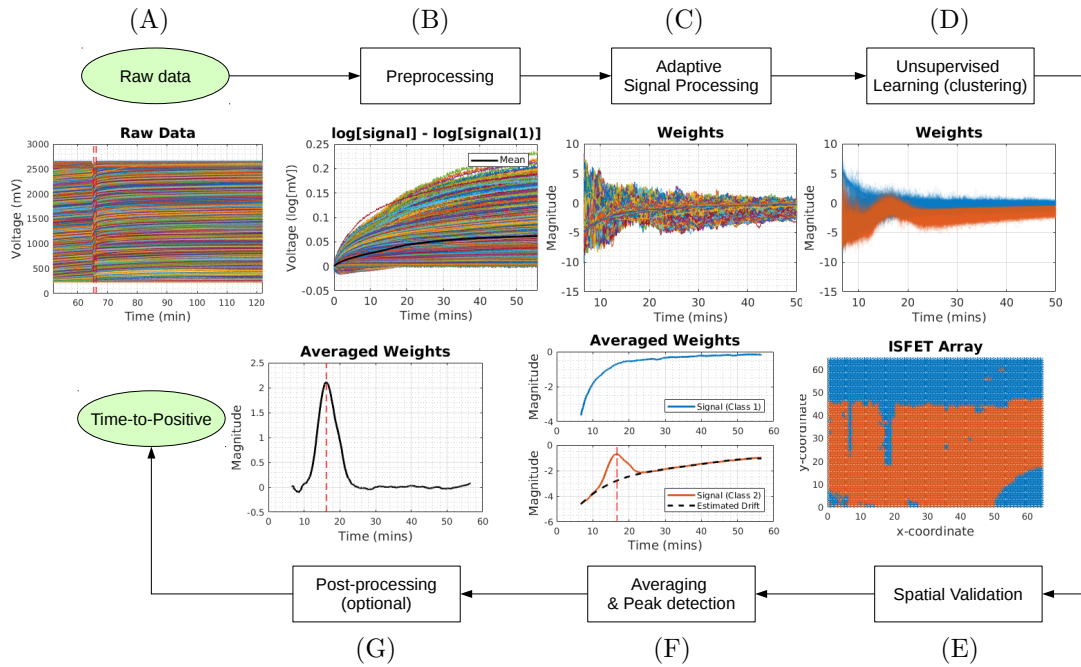


Figure 7.3: Proposed method for extracting time-to-positive values from ISFET arrays. (A) Raw data is collected from ISFET array. (B) Preprocessing is applied to reduce sensor-dependent artefacts. (C) Adaptive signal processing is used as a method of removing drift. (D) Unsupervised clustering groups similar behaving signals. (E) Spatial validation is performed to confirm non-random clustering. (F) Averaging and peak detection is used to determine positive/negative and a time-to-positive (if applicable). (G) Post-processing can be applied to improve time-to-positive extraction by removing background noise in the data.

Pre-processing. The background of each signal is removed by subtracting the average of the first 5 voltage measurements. An example of raw data and pre-processed signals are shown in Figure 7.3 (A) and (B). **Adaptive Signal Processing.** The drift is typically modelled via the stretched exponential given in equation 7.1. The parameter β is estimated via any iterative

gradient based optimisation algorithm, commonly of the form:

$$\beta_{i+1} = \beta_i - \alpha \frac{dy[n]}{d\beta} \quad (7.2)$$

Where i is the iteration number, α is a small constant (note: this can be adaptive), and $\frac{dy[n]}{d\beta}$ can be computed as:

$$\frac{dy[n]}{d\beta} = \log(t[n])(-e^{(-t[n]^\beta)})t[n]^\beta \quad (7.3)$$

The above analysis assumes that the drift parameter is constant through time. In order to relax this assumption, the proposed method takes an ‘adaptive signal processing’ approach whereby a linear model is used (i.e. $y[n] = \beta x[n]$), where the learned parameter is updated at each time step, so as to capture the behaviour over time. Therefore, each sensor will yield a β vector with the same length as the experiment, as depicted in Figure 7.3 (C). In this study, the normalised least mean square is used as an update rule [149]. **Unsupervised Clustering.** The β vectors are subsequently grouped into K clusters via K-Means clustering, whereby the value K is determined via the silhouette score [150]. Therefore, sensors are grouped by similar behaviour, as depicted in Figure 7.3 (D). **Spatial Validation.** In reality, we expect sensors which are spatially close, to have more similar behaviour. Therefore, to ensure the clustering is not random, the clusters can be visualised spatially, as in Figure 7.3 (E), to confirm this. This step can be achieved automatically via graph theoretic approaches for calculating the number of components [151]. **Averaging & Peak Detection.** The signals in each cluster are averaged to remove the effect of inherent noise, and a peak detection algorithm is used to extract a time-to-positive value, as shown in Figure 7.3 (E). Note: Since the behaviour of sensors are commonly drifting over time, an optional post-processing is to remove slow components in the signals, as in Figure 7.3 (F).

For reproducibility purposes, the RNA targets, primer design, reaction conditions and statistical analysis are described below. Details of the microarray analysis can be found in Herberg et al. (2016) [147] and the ISFET experimental setup can be found in Malpartida et al. (2019) [46]. The reader may wish to skip directly to the Results and Discussion (section 7.4).

7.3.3 RNA Targets and LAMP Primer Design

Based on Herberg et al. (2016), two genes were selected from the 38 highly correlated transcripts, namely, IFI44L (RefSeq ID: NM_006820) and EMR1-ADGRE (RefSeq ID: NM_001974.5 and

abbreviated herein to EMR1). Six different LAMP primers were designed to target eight specific region within the gene IFI44L and EMR1, as shown in Table 7.1. The primers specific for the two assays were selected according to the efficiency of the amplification reaction (limit-of-detection and reaction speed) using Primer Explorer V5 (Eiken Chemical Co. Ltd., Tokyo, Japan). The primers sequences were aligned to the reference genomic sequence using Geneious 10.0.5 software and manually optimised considering T_m , stability at the 3' end of each primers, GC content, and possible secondary structures such as primer dimer formation checked by using NUPACK37 software (<http://www.nupack.org/>). Primers were purchased from Integrated DNA Technologies and resuspended in TE to 100 μ M stock solutions. The stocks were stored at -20°C .

Table 7.1: Primer sequences for IFI44L and EMR1-ADGRE.

Target (accession number)	Primer	Sequence (5' $\rightarrow$ 3')
IFI44L (NM.006820)	F3	TGTGGTATAGCATATGTGGC
	B3	GAGAGGATGAGAATATCCTTCA
	FIP	CGGCTTTGAGAAATCATAGATCTAC-CTTGCTTACTAAAGTGGATGATTG
	BIP	ATGAATGTCCATAAAATGCTAGGCA-GTCCAGTTCCAAATCTGAAG
	LF	AGTTGTCTTGAAGAACCTCACTG
	LB	CCAATATTTTGATGGTTGGAAACTATG
EMR1 (NM.001974.5)	F3	CCACAGTCTTCTGGAGAGT
	B3	CCAGTGGTCTCTGTGGATTG
	FIP	CGTCCGAACAGCCGGAGTGATATGGAAAGCATGACACTGGC
	BIP	TGTGACGTTGGACTTGGTAGCCCCCTCAATTGTGGAACACCCG
	LF	TGCTGAGGGTTTCCAAAAAGAT
	LB	AAGGGGGATAAGATGAAGAT

Primers have been developed in this study.

7.3.4 LAMP Reaction Conditions

Each LAMP reaction mixture contained the following reagents: 0.5 μ l of 10 $\times$ isothermal buffer, 0.3 μ l of MgSO_4 (100mM stock), 0.7 μ l of dNTPs (10mM stock), 0.125 μ l of BSA (20mg/mL stock), 0.8 μ l of Betaine (5 M stock), 0.064 of AMV reverse transcriptase by Promega (25000U/mL stock), 0.125 μ l of SYTO 9 Green (20 μ M stock), 0.2 μ l of Bst 2.0 DNA Polymerase (8000 U/mL stock), 0.5 μ l of 10 $\times$ LAMP primer mixture (20 μ M BIP/FIP, 10 μ M LF/LB, and 2.5 μ M B3/F3) 1.68 μ l of extracted RNA or 1 μ l of synthetic RNA and enough nuclease-free water (Thermo Fisher Scientific) when necessary, to bring the volume to 5 μ l reactions. All reagents except the AMV RT were purchased from New England BioLabs. Reactions were performed at a single temperature of 63°C for 60 min and loaded into LightCycler 480 Multiwell Plates 96 (Roche Diagnostics) by using a LightCycler 96 Real-Time PCR System (Roche Diagnostics). One melting cycle was performed at 0.1°C/s from 65°C up to 97°C for validation of the specificity of the products. Each experimental condition was run in triplicates. For RT-eLAMP, the reagents were modified to obtain a low buffered solution with the aim to detect nucleotides by measuring

the solution's pH change. The reaction mixture contains the same reagents except the Tris buffer is removed to obtain a low buffered solution where pH change can be measured.

7.3.5 Statistical Analysis

To define a decision boundary between the bacterial and viral infections, a support vector classifier from the scikit-learn package in Python was used [152]. The underlying regression, i.e. linear combination of time-to-positive values were used as a score. To prevent over-fitting, 10-fold cross-validation was used to select hyper-parameters. The confidence interval for the classification accuracy was determined by 95% under the exact binomial distribution. The correlation between methods was determined using the Pearson correlation and a p-value using the F-statistic.

7.4 Results and Discussion

The following section is structured as follows. First, the 2-gene signature for distinguishing bacterial and viral infections are validated using microarray analysis. Subsequently, a new RT-qLAMP assay for detecting the same 2-gene is demonstrated. Following this, the ATP and proposed algorithm for extracting time-to-positive values from ISFET arrays are evaluated. Finally, the correlation between moving from microarray analysis to RT-qLAMP, and from RT-qLAMP to RT-eLAMP is studied to highlight the bottlenecks in the translation process.

7.4.1 Microarray Analysis

Figure 7.4 (A) shows the 2-gene signature (IFI44L and EMR1) for 11 definite bacterial (DB) and 11 definite viral (DV) isolates. It can be observed that the two groups are linearly separable. In particular, Figure 7.4 (A) shows the decision boundary for a trained support vector machine classifier with a black dashed line. The classification accuracy, as determined by cross-validation, is shown to be 100.0% (CI, 95.7 – 100.0%). Figure 7.4 (B) shows boxplots with scores which corresponding to the euclidean distance between each data point and the decision boundary. The log RNA expression can be found in Figure 7.4 (C). This analysis demonstrates that the 2-gene signature is in fact a successful method of distinguishing viral from bacterial infections in microarray analysis. However, microarray analysis is lab-based, expensive, and used for discovery applications, as opposed to diagnostics.

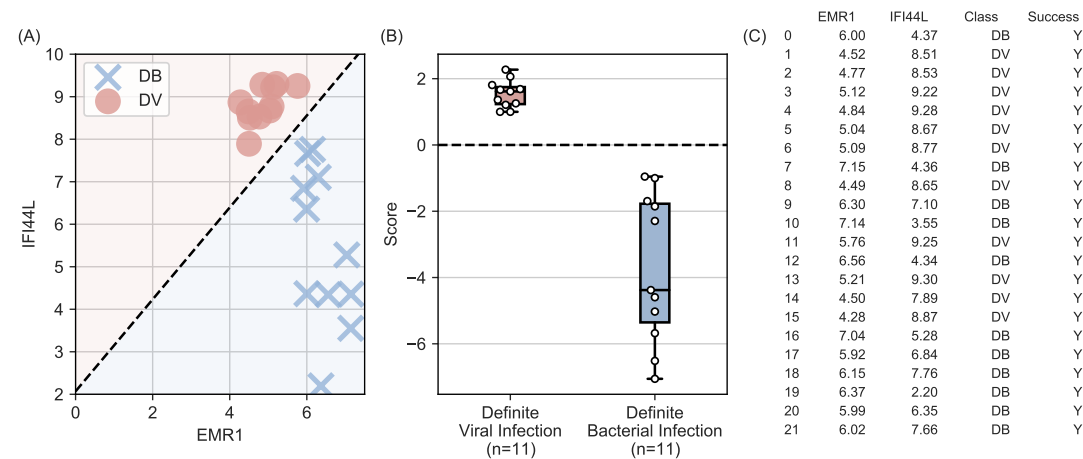


Figure 7.4: Viral versus bacterial classification - performance of microarray analysis. (A) Scatter plot of the log microarray expression values for the definite bacterial and definite viral isolates. (B) Box plot of the classification score, which corresponds to the distance to the decision boundary. (C) Table of log expression values for each isolate, along with their true class. Note: The ‘success’ column indicates whether the experiment gave a valid output for the two-gene signature. For example, if any one of the genes failed to give an output, the isolate would be deemed unsuccessful.

7.4.2 Quantitative LAMP

To this end, two new LAMP assays were developed to detect the 2-gene signature using a conventional qPCR platform. Figure 7.5 shows the standard curves for the new IFI44L and EMR1 assays with concentrations ranging from 10 to 10^6 copies per reaction, respectively. This demonstrates highly sensitive assays, with a lower limit of detection down to 10 copies. With regards to classification performance, Figure 7.6 (top panel) shows the RT-qLAMP assay also provides a 100.0% (CI, 95.7 – 100.0%) accuracy, illustrating the successful translation across platforms. Since microarray measures the RNA expression, it is expected that the time-to-positive value from LAMP is inversely proportion. As a sanity check, it can be observed that the position of DB and DV isolates are swapped in 2D space (i.e. for microarray, DB is in the lower right and DV is in the upper left, and vice-versa for RT-qLAMP).

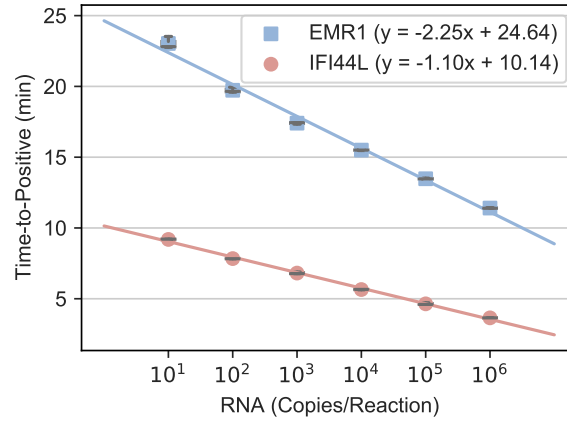
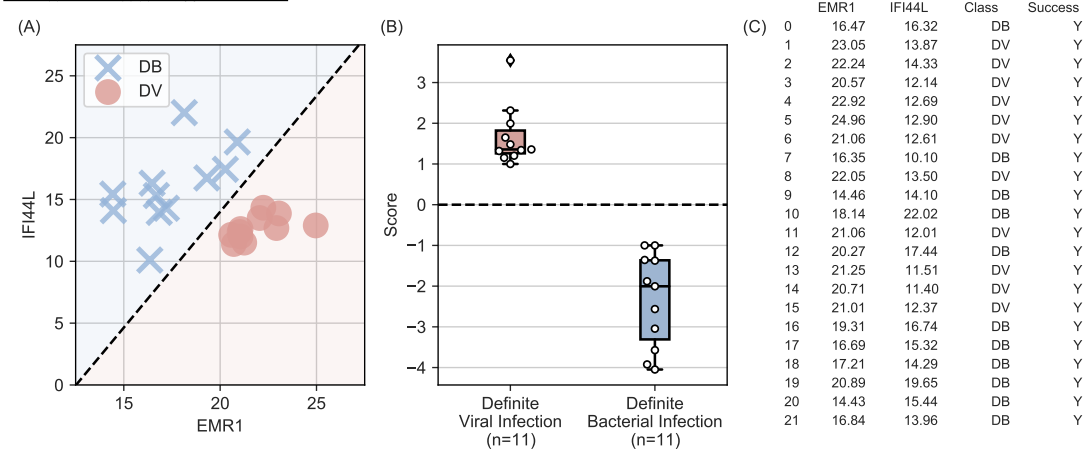


Figure 7.5: RT-qLAMP standard curves for IFI44L and EMR1.

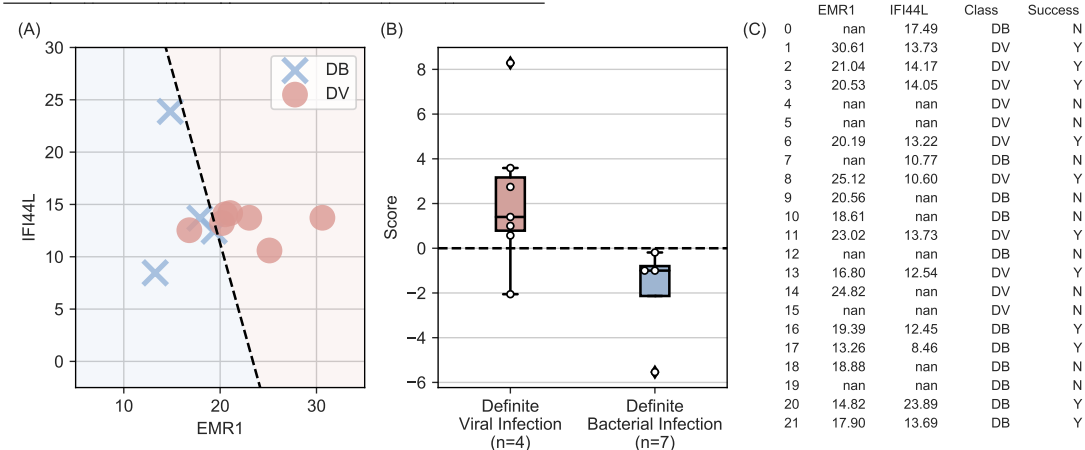
7.4.3 Electronic LAMP

Subsequently, the same assays/isolates were tested on the ISFET-based LoC platform using the ATP algorithm. The results are shown in Figure 7.6 (middle panel). It can be observed that 36.4% of the experiments were considered negative/failed, resulting in 50% of the isolates failing to provide a 2D signature. This highlights the challenge presented with ISFETs which have significant noise. Moreover, it can be observed that due to the lack of data, the decision boundary has a negative slope coefficient, which is incorrect. It is important to stress that this method is sufficient where most of the sensors have similar behaviour (e.g. low trapped charge and noise effects), as in Malpartida et al (2019). In contrast with the ATP method, the proposed method successfully called positive for all experiments. As shown in Figure 7.6 (bottom panel), the classification accuracy is 100.0% (CI, 95.7 – 100.0%) with a positive decision boundary slope, although the class separation is not as evident as microarray analysis or RT-qLAMP.

Top Panel - RT-qLAMP



Middle Panel - RT-eLAMP with ATP Method



Bottom Panel - RT-eLAMP with Proposed Method

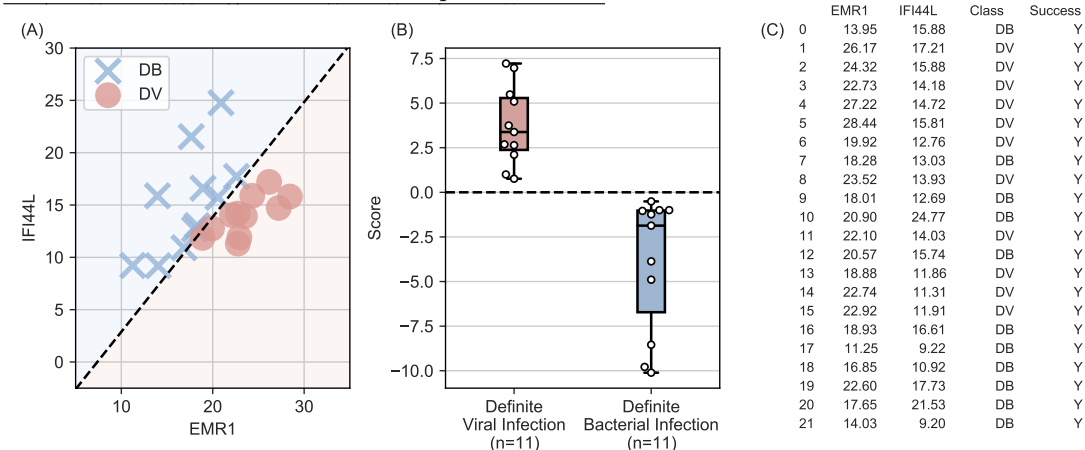


Figure 7.6: Viral Versus Bacterial Classification - Performance of RT-qLAMP (top panel), RT-eLAMP with ATP method (middle panel), and RT-eLAMP with proposed Method (bottom panel). (A) Scatter plot of the time-to-positive values for the definite bacterial and definite viral isolates. (B) Box plot of the classification score, which corresponds to the distance to the decision boundary. (C) Table of time-to-positive values for each isolate, along with their true class. Note: The 'success' column indicates whether the experiment gave a valid output for the two-gene signature. For example, if any one of the genes failed to give an output, the isolate would be deemed unsuccessful.

7.4.4 Evaluating the Translation Process

Although it has been shown that RT-qLAMP and RT-eLAMP can be used as a viable diagnostic solution for this specific application, microarray analysis remains a useful tool for identifying the signatures. Therefore, it is interesting to study the translation process from microarray analysis to RT-qLAMP to RT-eLAMP. This is particularly useful for identifying the bottlenecks in the process, so as to improve the class separation. Figure 7.7 shows a scatter plot between the RNA expression from microarray analysis and RT-qLAMP Cq values, and a scatter plot between RT-qLAMP and RT-eLAMP Cq values. The correlation coefficient between microarray analysis and RT-qLAMP is found to be -0.71 with $R^2 = 0.51$ and p -value $\ll 0.01$. However, the correlation between RT-qLAMP and RT-eLAMP is 0.85 with $R^2 = 0.98$ and p -value $\ll 0.01$. This demonstrates that the bottleneck in the translation process is between microarray analysis and qPCR, highlighting the need to focus research efforts on assay design.

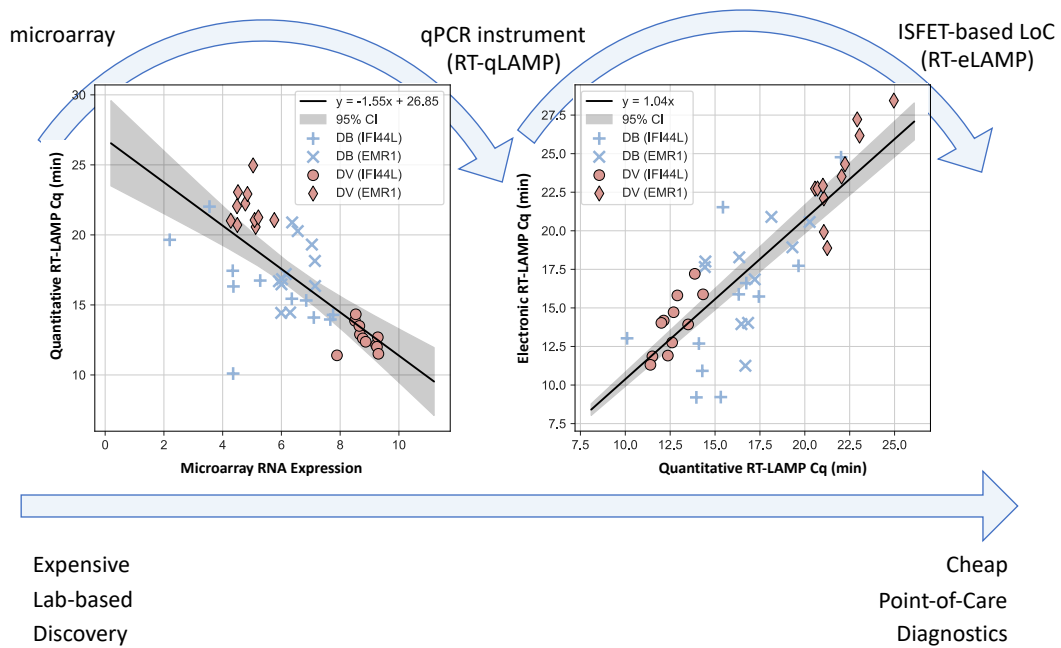


Figure 7.7: Translation of 2-Gene Signature from microarray analysis to RT-qLAMP to RT-eLAMP - moving from expensive lab-based discovery to cheap point-of-care diagnostics. (Left panel) Scatter plot of log microarray expression values against RT-qLAMP Cq values. As expected, as the expression increases, the time-to-positive decreases. (Right panel) Scatter plot of RT-qLAMP and RT-eLAMP (proposed method) Cq values.

7.5 Conclusion

As sensing systems become more advanced, the extracted data is becoming more rich and complex, presenting new challenges. In the case of diagnostic systems based on ISFET arrays, a major challenge is to combat the effects of various sources of noise. In this study, a new data-driven system was proposed to extract time-to-positives from nucleic acid amplification, which significantly improved the robustness compared with existing methods in the literature.

Although this study is a first step towards exploring new methods of extracting time-to-positive values from ISFET arrays, there are several factors which should be highlighted for future studies and improvements. Firstly, it is important to understand the cost of using the proposed algorithm from a computational and memory complexity stance. In terms of computational complexity, the added operations include: (i) fitting a model to each sensor, as opposed to the average of all sensors; (ii) unsupervised clustering such as the K-means algorithm; (iii) spatial validation; and (iv) peak detection. For (i), the operations consist of simple additions and multiplications, however the complexity scales with the number of sensors which may seem expensive for large arrays. However, the computation for each sensor is independent and therefore can be completely parallelised, yielding the same complexity as conventional drift fitting. The complexity of (ii)-(iv) largely depends on the chosen algorithms. However with modern computers, these are not significant. For example, the entire proposed algorithm (without visualisation) took only 0.96 ± 0.11 seconds on a MacBook Pro (16-inch, 2019)¹ running Matlab R2019B. In terms of space complexity, the additional memory usage is in the order of 10 Mb.

Aside from algorithmic factors, it is also important to evaluate the experimental design of the study. This includes the experimental conditions and sample set. In terms of the conditions, the experiments were performed similar to a real setting with the exception of using a thermocycler for thermal management. This was in fact a study design choice in order to remove noise from temperature effects since it can be assumed that this can be fixed using good thermal control, rather than increasing algorithmic complexity by accounting for temperature. However, a full study including this source of noise should be included in future studies. Moreover, this study used 22 isolates (11 bacterial and 11 viral), and although confidence intervals were used to describe the classification accuracy, a larger clinical study is required to ensure there was no unintended sampling bias.

¹The machine specs include 2.3 GHz 8-Core Intel Core i9 processor and 32 GB 2667 MHz DDR4 memory.

This study raises several interesting questions for future directions: What other applications may benefit significantly from this methodology? Is it possible to embed the proposed algorithm into the PoC device itself (rather than running on a computer or server)? And are there systematic or *in silico* approaches to design LAMP assays from microarray experiments to address bottlenecks in the translation process?

In conclusion, a new method for extracting time-to-positive values from nucleic acid amplification for ISFET arrays under high noise environments is proposed. This approach is primarily based on modelling the sensors adaptively to represent ‘changes in behaviour’, and clustering the time-series, so as to average sensors with the same behaviour. Using this new method was shown to successfully provide a high correlation between conventional lab-based RT-qLAMP and the ISFET-based LoC platform, yielding 100% accuracy when classifying viral and bacterial isolates with a 2-gene signature. Moreover, the entire workflow from microarray analysis (expensive lab-based discovery) to lab-on-a-chip devices (cheap point-of-care diagnostics) was studied, in order to highlight bottlenecks in the translation process. This work hopes to encourage researchers to apply this methodology to other applications, such that it benefits the wider scientific community.

Chapter 8

Applications and Analysis of IS-FET Arrays

8.1 Chapter Overview

The previous Chapter showed the application of ISFET-arrays for the detection of viral and bacterial infections. This presents a solution which is more portable, affordable and scalable than existing methods in healthcare systems. However, the biggest challenge for wide-scale adoption of this technology is to show the performance (relative to the gold standards) across a range of applications, so as to prove this is a viable solution. This Chapter serves as a short overview of some studies that have used ISFET arrays for nucleic acid detection, to identify applications and on-going research which will greatly benefit from algorithms such as that presented in Chapter 7. In particular, these studies show the first steps towards moving diagnostics (i.e. DNA detection, quantification and multiplexing) directly to the patient.

8.2 COVID-19

The following extract is taken from:

- J. Rodriguez-Manzano, K. Malpartida-Cardenas, N. Moser, I. Pennisi, M. Cavuto, L. Miglietta, A. Moniri, R. Penn, G. Satta, P. Randell, F. Davies, F. Bolt, W. Barclay, A. Holmes and P. Georgiou, “A handheld point-of-care system for rapid detection of SARS-CoV-2 in under 20 minutes”, *ACS Central Sciences*, vol. 7, no. 2, pp. 307-317, Jan 2021

The COVID-19 pandemic is a global health emergency characterised by the high rate of transmission and ongoing increase of cases globally. Rapid point-of-care diagnostics to detect the causative virus, SARS-CoV-2, are urgently needed to identify and isolate patients, contain its

spread and guide clinical management. In this work, the development of a rapid PoC diagnostic test (< 20 min) is reported based on RT-LAMP and semiconductor technology for the detection of SARS-CoV-2 from extracted RNA samples. The developed LAMP assay was tested on a real-time benchtop instrument (RT-qLAMP) showing a lower limit of detection of 10 RNA copies per reaction. It was validated against 183 clinical samples including 127 positive samples (screened by the CDC RT-qPCR assay). Results showed 90.55% sensitivity and 100% specificity when compared to RT-qPCR and average positive detection times of 15.45 ± 4.43 min. For validating the incorporation of the RT-LAMP assay onto the PoC platform (RT-eLAMP), a subset of samples was tested ($n=40$), showing average detection times of 12.89 ± 2.59 min for positive samples ($n=34$), demonstrating a comparable performance to a benchtop commercial instrument. Paired with a smartphone for results visualisation and geo-localization, this portable diagnostic platform with secure cloud connectivity will enable real-time case identification and epidemiological surveillance.

8.3 *Aspergillus Fumigatus*

The following extract is taken from:

- L-S. Yu, J. Rodriguez-Manzano, N. Moser, A. Moniri, K. Malpartida-Cardenas, N. Miskourides, T. Sewell, T. Kochina, A. Brackin, J. Rhodes, A.H. Holmes, M.C. Fisher and P. Georgiou “Rapid Detection of Azole-Resistant *Aspergillus fumigatus* in Clinical and Environmental Isolates by Use of a Lab-on-a-Chip Diagnostic System”, *Journal of Clinical Microbiology*, vol. 58, no. 11, Nov 2020.

Aspergillus fumigatus has widely evolved resistance to the most commonly used class of antifungal chemicals, the azoles. Current methods for identifying azole resistance are time-consuming and depend on specialized laboratories. There is an urgent need for rapid detection of these emerging pathogens at point-of-care to provide the appropriate treatment in the clinic and to improve management of environmental reservoirs to mitigate the spread of antifungal resistance. Our study demonstrates the rapid and portable detection of the two most relevant genetic markers linked to azole resistance, the mutations TR34 and TR46, found in the promoter region of the gene encoding the azole target, *cyp51A*. We developed a lab-on-a-chip platform consisting of: (1) tandem-repeat loop-mediated isothermal amplification, (2) state-of-the-art complementary metal-oxide-semiconductor microchip technology for nucleic-acid amplification detection and, (3) and a smartphone application for data acquisition, visualization and cloud connectivity. Specific and sensitive detection was validated with isolates from clinical and environmental

samples from 6 countries across 5 continents, showing a lower limit-of-detection of 10 genomic copies per reaction in less than 30 minutes. When fully integrated with a sample preparation module, this diagnostic system will enable the detection of this ubiquitous fungus at the point-of-care, and could help to improve clinical decision making, infection control and epidemiological surveillance.

8.4 Mobilised Colistin Resistance

The following extract is taken from:

- J. Rodriguez-Manzano, N. Moser, K. Malpartida-Cardenas, A. Moniri, L. Fisarova, I. Pennisi, A. Boonyasiri, E. Jauneikaite, A. Abdolrasouli, F. Bolt, F. Davies, X. Didelot, A. Holmes and P. Georgiou, “Rapid Detection of Mobilized colistin Resistance using a nucleic Acid Based Lab-on-a-chip Diagnostic System”, *Scientific Reports*, vol. 10, no. 1, pp 1-9, May 2020.

The increasing prevalence of antimicrobial resistance is a serious threat to global public health. One of the most concerning trends is the rapid spread of Carbapenemase-Producing Organisms (CPO), where colistin has become the last-resort antibiotic treatment. The emergence of colistin resistance, including the spread of mobilized colistin resistance (*mcr*) genes, raises the possibility of untreatable bacterial infections and motivates the development of improved diagnostics for the detection of colistin-resistant organisms. This work demonstrates a rapid response for detecting the most recently reported *mcr* gene, *mcr-9*, using a portable and affordable lab-on-a-chip (LoC) platform, offering a promising alternative to conventional laboratory-based instruments such as real-time PCR (qPCR). The platform combines semiconductor technology, for non-optical real-time DNA sensing, with a smartphone application for data acquisition, visualization and cloud connectivity. This technology is enabled by using loop-mediated isothermal amplification (LAMP) as the chemistry for targeted DNA detection, by virtue of its high sensitivity, specificity, yield, and manageable temperature requirements. Here, we have developed the first LAMP assay for *mcr-9* - showing high sensitivity (down to 100 genomic copies/reaction) and high specificity (no cross-reactivity with other *mcr* variants). This assay is demonstrated through supporting a hospital investigation where we analyzed nucleic acids extracted from 128 carbapenemase-producing bacteria isolated from clinical and screening samples and found that 41 carried *mcr-9* (validated using whole genome sequencing). Average positive detection times were 6.58 ± 0.42 min when performing the experiments on a conventional qPCR instrument ($n = 41$). For validating the translation of the LAMP assay onto a LoC platform, a subset of the samples

were tested ($n = 20$), showing average detection times of 6.83 ± 0.92 min for positive isolates ($n = 14$). All experiments detected mcr-9 in under 10 min, and both platforms showed no statistically significant difference ($p\text{-value} > 0.05$). When sample preparation and throughput capabilities are integrated within this LoC platform, the adoption of this technology for the rapid detection and surveillance of antimicrobial resistance genes will decrease the turnaround time for DNA detection and resistotyping, improving diagnostic capabilities, patient outcomes, and the management of infectious diseases.

8.5 Dengue

The following extract is taken from:

- N. Moser, L-S. Yu, J. Rodriguez-Manzano, K. Malpartida-Cardenas, A. Au, C. Cicatiello, A. Moniri, W-H Wang, S-F. Wang, Y-H. Chen and P. Georgiou, “Dengue diagnostics in the cloud: Quantitative detection using a smartphone-connected handheld Lab-on-Chip platform”, In preparation for *Biosensors and Bioelectronics*.

Dengue is one of the most prevalent infectious diseases in the world. Rapid and scalable diagnostics are key to patient management and epidemiological surveillance of the dengue virus (DENV), however current technologies do not match required clinical sensitivity and specificity or rely on large laboratory equipment. To address this global health issue, this work reports the development of the first smartphone-connected handheld molecular test for quantitative detection of dengue serotypes at the point of care (PoC). The custom hardware and software were designed for usability and portability. The extracted RNA from the plasma sample ($20\mu L$) is loaded into the reaction chamber of the lab-on-a-chip (LoC) cartridge and onto the Complementary Metal-Oxide-Semiconductor (CMOS) microchip integrating over 4,000 potentiometric sensors. The main device is paired with the smartphone application and performs a reverse-transcriptase electronic loop-mediated isothermal amplification (RT-eLAMP) assay to quantitatively detect the presence of a serotype of dengue. The method was shown to achieve similar performance to a gold-standard laboratory instrument (RT-qLAMP), with detection in under 15 minutes for DENV-1 and DENV-2 in our lab at Imperial College London, UK. The results are synchronised to a cloud server hosted by Amazon Web Services and geo-tagged on a map for real-time tracking of epidemics accessible from any electronic device. To validate the translation of the test in clinical settings, we demonstrated quantitative detection of 9 clinical samples in Kaohsiung Medical Hospital, Taiwan. When paired with sample extraction methods, the device may enable rapid dengue diagnostics in low-resource settings.

Chapter 9

Conclusion & Future Perspective

This thesis investigated the application of intelligent algorithms to the field of DNA detection, and showed that more value can be extracted from existing data by processing data in higher dimensions. In particular, 3 sources of data were explored: quantitative, digital, and electronic. Although each approach has a common thread, namely nucleic acid amplification curves, there are significant differences which warrant their own data processing strategies.

9.1 Contributions

Chapter 3 introduced a new framework for enhanced quantification and outlier detection. This was achieved by combining multiple features of the amplification curve to construct a multidimensional standard curve (MSC). Contrary to only considering a single-feature, data points do not fall exactly on the standard curve, which enables a similarity measure between amplification curves - leading to a natural method of outlier detection. The contributions of Chapter 3 led to the [J1] and [P1] (see List of Publications).

Chapter 4 extended the concept of MSCs to multiplexing. More specifically, it was shown that different targets can be designed to have non-overlapping MSCs, and therefore enabling the first data-driven method for simultaneous enhanced quantification and multiplexing in a single chemical reaction (and single fluorescent channel). The fact that the amplification curve is sufficient to differentiate targets suggests that there is target-specific kinetic information which is typically neglected. The contributions of Chapter 4 led to [J2] and [P1] (see List of Publications).

Chapter 5 demonstrated the first successful application of machine learning for multiplexing based on the entire amplification curve, by leveraging the vast volume of data from real-time digital instruments. This was achieved by training a supervised machine learning model to automatically benefit from target-specific kinetic information and classify targets, without explicitly extracting features. Since digital PCR inherently performs spatial multiplexing and relies on Poisson statistics to perform quantification (rather than standard curves), a formula was derived

to establish a trade-off between quantification and multiplexing. The contributions of Chapter 5 led to [J3] and [P2] (see List of Publications).

Chapter 6 demonstrated the largest dPCR multiplex using intercalating dyes that has been reported in the literature. This was achieved through combining the kinetic information from amplification curves and thermodynamic information from melting curves using a novel 3-step machine learning system. In particular, it was shown that although both amplification and melting curves are analysing the same product, they contain non-mutual information which is beneficial for target identification. The contributions of Chapter 6 led to [J4] and [P2] (see List of Publications).

Chapter 7 tackled the challenges of extracting time-to-positive (TTP) values from ISFET arrays in highly noisy environments. In particular, current methods make several assumptions regarding the data which are often violated. To this end, a new algorithm was proposed based on modelling sensor drift using adaptive signal processing and clustering sensors based on their behaviour with unsupervised learning methods. This algorithm was shown to have a significant impact on signature based classification, where it is vital that TTP values are accurate. The contributions of Chapter 7 led to [J5], [J13], and [P4] (see List of Publications).

Chapter 8 identified recent applications of ISFET arrays to highlight the future direction of this field and the on-going research which will see a large benefit from incorporating more sophisticated data-driven methods such as that in Chapter 7.

9.2 Future Perspective

There is evidently a trend in adopting nucleic acid amplification solutions to answer a wide variety of research questions. Figure 9.1 shows a graph of the estimated number of publications in this field (broken down by quantitative, digital and electronic instruments) as a function of time. It can be observed that the growth of publications is exponential (i.e. linear on log-scale) and there is no sign of slowing down. This reinforces the wide-scale impact of improving data processing algorithms since they are general methods that apply across all applications.

This thesis has shown that there is considerable room for research in this field from a data perspective. Therefore, it is interesting to question why the methods in this work were not previously developed. The literature regarding quantitative instruments is very mature and established (as evident from the $> 10^4$ publications every year). To this end, most research efforts in recent years that enhance the core technology have been incremental, and the great focus has

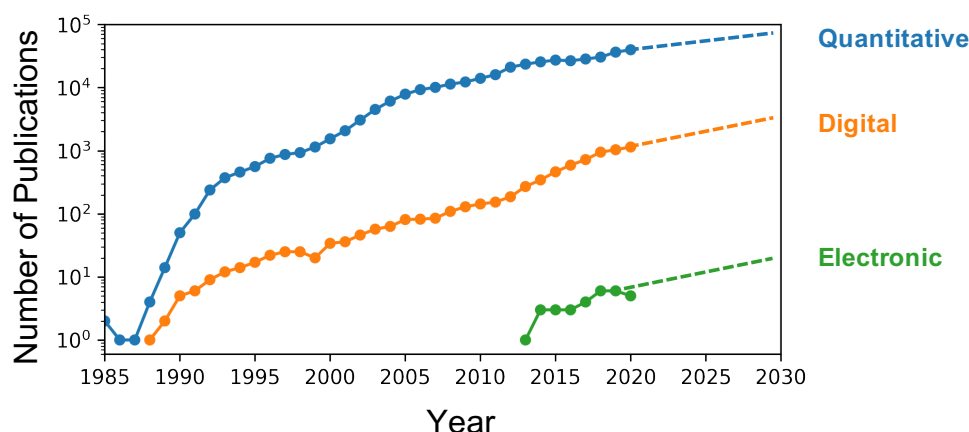


Figure 9.1: Publication trend in quantitative, digital and electronic nucleic acid amplification, as quantified using the Scopus database (on 28/01/2021) with the following search queries: Let variable %CHEM% = (“polymerase chain reaction” OR “isothermal amplification” OR “nucleic acid amplification” OR “DNA amplification”). For ‘Quantitative’: (“quantitative” OR “real-time”) AND %CHEM%. For “Digital”: (“digital” OR “single-molecule”) AND %CHEM%. For “Electronic”: (“ISFET” OR (“sensor” AND “FET”) OR (“biosensor” AND “FET”)) AND %CHEM%.

been on the applications (e.g. developing optimal PCR assays for COVID-19). Since quantitative PCR is the “gold standard”, qPCR data processing methods have directly been translated to other technologies. For example, digital instruments (with real-time capabilities) will perform the same data processes as qPCR, such as extracting Ct values. Or for electronic instruments, the noise is modelled/removed such that the data resembles a real-time amplification curve so that a Ct value is extracted. Although these methods provide promising results, there has not been a push to explore *new* data processing methods that are enabled by having a different source of data, or the advantages that these new methods may offer.

At its core, the central theme for this thesis is to analyse data in higher dimensions. Each technical Chapter highlighted interesting questions which open various future directions. Although the benefits of every method have been demonstrated in each Chapter, there is one key question which needs to be addressed and will likely drive future research: How do the fundamental biological factors influence the data? For example, to create amplification curves which have different shapes (and can therefore be multiplexed), one needs to understand how to change the shape (by changing factors such as primer design or experimental conditions).

Aside from the scientific challenge described above, the impact of this work and its future directions is also based on the adoption of these intelligent algorithms in real-world healthcare settings. Therefore, in terms of next steps, it is important to conduct a comprehensive cost analysis of developing and integrating software applications into diagnostic pathways within current healthcare systems.

It is clear that the contents of this thesis were the result of a highly interdisciplinary collaboration between multiple fields, namely: data science, engineering, and molecular biology. This may explain the reason for the limited amount of research efforts at this intersection - which this thesis hopes to serve as a starting point. There is clearly a growing trend for researchers from the biological sciences to learn more about data processing, and given the recent global pandemic, there has also been a surge of data scientists attracted to the field of DNA detection. Therefore, I hope that by sharing the ideas in this thesis, others will be able to adapt and enhance this work, so as to tackle the world's greatest healthcare challenges.

Appendix A

Supporting Information:

Chapter 3

This Appendix contains the following:

- Synthetic DNA sequences (Table A.S1 - A.S4)
- Extracted features and sigmoidal fittings for lambda DNA standard (Table A.S5)
- Breakdown for Figure of Merit of Ct , C_{y0} , $-\log_{10}(F_0)$ and M_0 (Table A.S6 - A.S9)
- Extracted features and sigmoidal fittings for non-specific outliers (Table A.S10)
- Extracted features and sigmoidal fittings for temperature variation experiment (Table A.S11)
- Estimated quantification for temperature variation experiment (Table A.S12)
- Melting curve analysis for lambda DNA standard experiment (Figure A.S1)
- Melting curve analysis for non-specific outlier detection experiment (Figure A.S2)
- Melting curve analysis for temperature variation experiment (Figure A.S3).

Table A.S1. Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing the lambda phage DNA target.

gBlock gene fragment for lambda	CAGGAACAGGGAATGCCGTTCTGCGAGGC GGTGGCAAGGGTAATGAGG TGCTTTATGACTCTGCCCGCTCATAA AATGGT ATGCCGAAAGGGATGCTGA AATTGAGAACGAAAAGCTGCCCGGGAGGTTGAAGAACTGCCGCAGGCCA GCCAGGCAGATCTCCAGCCAGGAACATTGAGTACGAACGCCATCGACTTACGCGTGCGCAGGCCGACGCACAGGA ACTGAAGAATGCCAG
---------------------------------	--

Forward lambda PCR primer in **bold-red** and reverse lambda primer in **bold-blue** (primers were designed in-house).

Table A.S2. Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing blaOXA-48 target.

Outlier 1 (5' - 3')	ATGCGTGATTAGCCTTATCGGCTG TGTTTTGGTGGCATCGAT TATCGGAATGCCTGCGGTAGCAAAGGAATGGCA AGAAAAACAAAGTTGGAATGCTCACTTTACTGAACATAAATCACAGGGCGTAGTTGTGCTCTGGAATGAGAATAAG CAGCAAGGATTTACCAATAATCTTAAACGGGCCAACCAAGCATTTTAC CCGCATCTACCTTTAAATCCCAATAGC TTGATCGCCCTCGATTGGCGTGTTAAGGATGAACACCAAGTCTTAAAGTGGGATGGACAGACGCGC
---------------------	--

OXA-48-F (TGTTTTGGTGGCATCGAT) **in red**;
OXA-48-R (GTAAMRATGCTTGGTTCGC) **in blue**;
amplicon size = 177 bp (**bold and underlined**);
primers from Monteiro et al 2012

Table A.S3. Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing blaNDM target.

Outlier 2 (5' - 3')	ATGGTTGCGGCGCAACACAGCCTGACTTTCGCCCAATGGCTGGGTGCAACAGCAACCGCGCCCACTTTGGCCCG CTCAAGGTATTTTACCCGGCCCCGCC ACACAGTGACAATATCACCGTTGGGATCGACGGCACCGACATCGCTTTTG GTGGCTGCTGATCAAGGACAGCAAGGCCAA GTGCTCGGCAATCTCGGTGATGCCGACACTGAGCACTACGCCGCG TCAGCGCGCGCTTGGTGCGGCGTTCCCAAGGCCAGCATGATCGTGATGAGCCATTCGCCCC
---------------------	--

NDM-F (TTGGCCTTGCTGTCTTG) **in red**;
NDM-R (ACACAGTGACAATATCACCG) **in blue**;
amplicon size = 82 bp (**bold and underlined**);
primers from Monteiro et al 2012

Table A.S4. Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing blaKPC target.

Outlier 3 (5' - 3')	AACCAT TCGCTAAACTCGAACAGG ACTTTGGCGGCTCCATCGGTGTGTACGCGATGGATACCGGCTCAGGCGCAACT GTAAGTTACCGCGCTGAGGAGCGCTTCCCACTGTGCAGCTATTCAAGGGCTTTCTTGCTGCCGCTGTGCTGGCTCGCA GCCAGCAGCAGGCCGGCTTGCTGGACACACCCATCCGTTACGGCAAAAATGCGCTGGTTCGGTGGTCAACCATCTCGG AAAAATATCTGACAACAGGCATGACGGTGGCGGAGCTGTCGCGGGCCGCGTGCAATACAGTGATAACGCCGCCGCC AATTTGTTGCTGAAGGAGTTGGGCGGCCGGCGGGCTGACGGCTTCATGCGCTCTATCGGCGATACCAGTTCCGT CTGGACCGCTGGGAGCTGGAGCTGAACTCCGCCATCCAGGCGATGCGCGCGATACCTCATCGCCGCGCGCGTGACG GAAAGCTTACAAAACTGACACTGGGCTCTGCACTGGCTGCGCCGACGCGGACAGTTTGTGATTGGCTAAAGGG AAACAGCAGCGGCAACACCGCATCCGCGGGCGGTGCCGGCAGACTGGGCAGTCGGAGACAAAACCGGAACCTGC GGAGTGTATGGCACGGCAAATGACTATGCCGTCGTCTGGCCACTGGGCGCGCACCTATTGTGTTGGCCGTCTACAC CGGGCGCTAAACAAGGATGACAAGCACAGCGAGGCCGTATCGCCGCTGCGGCTAGACTCGCGCTCGAGGGATTGG GCGTCAACGGGCGAGTAAG
---------------------	---

KPC-F (TCGCTAAACTCGAACAGG) **in red**;
KPC-R (TTACTGCCGTTGACGCCAATCC) **in blue**;
amplicon size = 785 bp (**bold and underlined**);
primers from Monteiro et al 2012

Table A.S5. Experimental values for lambda DNA standards.

Concentration (copies/reaction)	Sigmoidal Curve-fitting Parameters								
	Ct	Cy0	-log10(F0)	M0	Fb	Fmax	c	b	d
1.00E+02	29.60286	28.14662	9.76812	11.59576	0.00171	0.19864	31.20895	2.00134	1.26013
	30.36106	28.88565	10.01293	12.51620	0.00137	0.19362	31.89412	2.00278	1.28424
	30.49607	29.03997	10.08592	12.72201	0.00148	0.19703	32.06496	1.99031	1.27033
	30.18126	28.59652	9.74092	12.04612	0.00110	0.19011	31.17973	2.16333	1.54270
	30.61087	29.12458	9.93007	12.71221	0.00137	0.18160	32.38258	1.99474	1.19285
	30.52995	28.99942	10.05593	12.66146	0.00109	0.19989	31.77558	2.06643	1.41198
	30.80903	29.22164	10.46644	13.14969	-0.00004	0.19203	31.98968	1.99794	1.39437
1.00E+03	30.48359	28.99977	10.15566	12.72414	0.00105	0.20679	32.00059	1.97305	1.27604
	26.53766	24.98309	8.93031	7.78828	0.00113	0.19464	27.65128	2.11468	1.48081
	26.97111	25.38085	8.88703	8.17403	0.00141	0.19293	28.35725	2.08432	1.33897
	27.04113	25.43462	8.90557	8.24153	0.00131	0.19943	28.15692	2.13202	1.46509
	27.08741	25.45308	8.91666	8.26776	0.00132	0.20711	27.96269	2.20259	1.59319
	27.06787	25.39216	8.83712	8.15487	0.00100	0.18234	27.98259	2.20839	1.56360
	26.95429	25.32104	9.10802	8.25072	0.00095	0.19906	27.91180	2.12276	1.52155
1.00E+04	26.94251	25.33669	9.19890	8.32383	0.00052	0.19659	28.21981	2.02390	1.35499
	27.01954	25.37736	9.00308	8.24342	0.00082	0.20544	27.85523	2.14832	1.58779
	23.06449	21.45487	8.11253	3.61473	0.00082	0.17757	24.12603	2.05981	1.45392
	23.86092	22.23739	8.16096	4.45714	0.00126	0.20219	24.63902	2.19898	1.64095
	23.82437	22.16628	8.08453	4.33559	0.00102	0.20641	24.58259	2.20760	1.64317
	23.87834	22.20694	8.17261	4.43307	0.00094	0.20444	24.41637	2.24269	1.76205
	23.85718	22.20265	8.05278	4.35341	0.00102	0.20003	24.76700	2.20123	1.57268
1.00E+05	23.85051	22.14791	7.99930	4.26329	0.00104	0.20823	24.32139	2.30533	1.80903
	23.77082	22.13842	8.31389	4.45038	0.00017	0.20166	24.60059	2.09622	1.57798
	23.43357	21.87384	8.09965	4.04118	0.00084	0.20425	24.71569	2.05671	1.38458
	19.93326	18.28931	7.31713	-0.16854	-0.00006	0.18707	20.86677	2.03578	1.49672
	20.15876	18.53712	7.24532	0.04347	0.00037	0.18782	21.00119	2.09966	1.57509
	20.19548	18.50961	7.07905	-0.08905	0.00075	0.20186	20.68453	2.24574	1.78447
	20.18913	18.51850	7.06403	-0.08931	0.00091	0.20251	20.76798	2.23613	1.73927
1.00E+06	20.15086	18.48183	7.01717	-0.15675	0.00097	0.20562	20.79512	2.23583	1.70703
	20.27313	18.52587	7.02551	-0.10546	0.00041	0.20629	20.24835	2.36218	2.09772
	20.22384	18.50878	7.00447	-0.13653	0.00068	0.22824	20.47807	2.30008	1.92284
	20.08686	18.45829	7.18246	-0.07771	0.00049	0.19780	21.04207	2.08265	1.51051
	16.17976	14.51421	6.08100	-4.86072	0.00061	0.18155	16.77620	2.18851	1.71329
	17.10805	15.38710	6.23458	-3.85842	0.00036	0.19747	17.37334	2.25699	1.89780
	17.04889	15.30319	6.17615	-3.98202	0.00060	0.20297	17.04251	2.31918	2.06395
1.00E+07	17.06687	15.28346	6.09262	-4.05469	0.00068	0.21577	16.99645	2.36634	2.09972
	17.04314	15.32311	6.17842	-3.96003	0.00043	0.20119	17.40975	2.23856	1.83308
	17.00921	15.31046	6.28314	-3.90760	0.00043	0.19592	17.57628	2.18554	1.71424
	17.03086	15.30684	6.27657	-3.91537	0.00018	0.20555	17.30596	2.23191	1.88232
	16.65264	15.04565	6.21089	-4.22809	0.00047	0.19609	17.49631	2.11557	1.58828
	13.22484	11.56815	5.27775	-8.42135	0.00021	0.17499	13.59601	2.17619	1.83729
	13.35007	11.66265	5.30553	-8.30574	0.00019	0.17182	13.56278	2.23050	1.93622
1.00E+08	13.34625	11.65031	5.26970	-8.34097	0.00036	0.18093	13.33486	2.28242	2.08529
	13.35548	11.61685	5.17543	-8.43461	0.00050	0.19908	13.03536	2.37365	2.29371
	13.33545	11.64715	5.28802	-8.33280	0.00027	0.16484	13.51644	2.25462	1.96335
	13.35185	11.72146	5.31096	-8.24157	0.00017	0.18286	13.89558	2.14029	1.74200
	13.32309	11.68837	5.30690	-8.27843	0.00020	0.18031	13.83290	2.14862	1.76103
	13.28688	11.63747	5.28394	-8.34558	0.00017	0.18015	13.68564	2.17296	1.82563
	10.05275	8.29492	4.33749	-12.40689	-0.00017	0.17118	10.10286	2.24129	2.00471
1.00E+08	10.05378	8.26361	4.32252	-12.44862	-0.00009	0.16975	9.86191	2.30801	2.16317
	10.10387	8.30169	4.27329	-12.43992	0.00010	0.18066	9.33865	2.40134	2.59146
	10.08808	8.28186	4.27340	-12.46041	0.00001	0.18756	9.41329	2.39533	2.52000
	10.05353	8.32742	4.39515	-12.33715	-0.00016	0.16012	10.28119	2.17953	1.88943
	10.04862	8.31133	4.25989	-12.43861	-0.00012	0.20363	10.05777	2.23803	2.03842
	10.09425	8.37908	4.36445	-12.30285	-0.00013	0.18918	10.18256	2.19485	1.98332
	10.08292	8.36027	4.17914	-12.43852	-0.00027	0.22425	10.39602	2.18692	1.84830

Table A.S6. Breakdown for Figure of Merit of Ct

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
Quantification	1	9.86E+07	1.14E+07	1.54E+06	1.21E+05	1.44E+04	1.36E+03	1.70E+02
	2	9.85E+07	1.05E+07	8.21E+05	1.04E+05	8.39E+03	1.02E+03	1.02E+02
	3	9.52E+07	1.05E+07	8.54E+05	1.01E+05	8.60E+03	9.69E+02	9.29E+01
	4	9.62E+07	1.05E+07	8.44E+05	1.01E+05	8.29E+03	9.39E+02	1.15E+02
	5	9.85E+07	1.06E+07	8.58E+05	1.04E+05	8.41E+03	9.52E+02	8.60E+01
	6	9.88E+07	1.05E+07	8.78E+05	9.58E+04	8.45E+03	1.03E+03	9.08E+01
	7	9.58E+07	1.07E+07	8.65E+05	9.90E+04	8.92E+03	1.04E+03	7.51E+01
	8	9.66E+07	1.10E+07	1.12E+06	1.09E+05	1.12E+04	9.84E+02	9.37E+01

	CV	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
		1.51	3.12	25.54	7.31	22.67	13.24	28.57

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
RE	1	1.45	14.47	54.09	20.63	44.06	36.41	70.38
	2	1.52	5.15	17.93	3.51	16.09	1.65	1.85
	3	4.81	5.42	14.57	0.96	13.98	3.07	7.07
	4	3.79	4.76	15.61	1.40	17.08	6.07	15.06
	5	1.50	6.20	14.24	4.07	15.88	4.81	14.04
	6	1.17	5.02	12.24	4.22	15.50	2.82	9.18
	7	4.19	7.09	13.52	0.96	10.80	3.64	24.85
	8	3.45	9.75	11.79	8.69	12.14	1.64	6.28

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
LOOCV	1	1.54	15.06	55.68	21.04	45.23	37.94	75.74
	2	1.61	5.34	18.30	3.58	16.43	1.71	1.96
	3	5.10	5.63	14.88	0.98	14.28	3.18	7.49
	4	4.01	4.94	15.93	1.43	17.44	6.28	16.05
	5	1.59	6.43	14.53	4.14	16.21	4.99	14.85
	6	1.24	5.21	12.50	4.29	15.83	2.92	9.73
	7	4.44	7.36	13.80	0.98	11.04	3.78	26.20
	8	3.65	10.14	12.08	8.85	12.43	1.70	6.66

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
Q	1	3.369615427	680.1255825	76914.73423	3172.548417	45181.28958	18284.66405	152285.5769
	2	3.697706432	85.91882195	8380.724989	91.75630779	5994.743495	37.37121997	103.6361771
	3	37.0387709	95.22160658	5536.058531	6.890050818	4528.43351	129.1393161	1513.613468
	4	22.94081953	73.45389592	6350.202408	14.59457257	6751.192477	504.5186376	6904.466029
	5	3.616615164	124.5264551	5285.166586	123.272899	5836.543914	317.5217883	5953.015666
	6	2.205991195	81.67364176	3907.027556	132.3924554	5559.816658	109.0805965	2551.570218
	7	28.06870714	162.9962182	4766.226925	6.868853549	2702.049103	182.0832954	18605.78912
	8	19.0321599	308.559578	3636.013447	562.2663153	3422.092657	36.78207274	1194.270867

Av. Q	4587.427356
Std(Q)	12798.97639

Shaded region is not included in the mean and standard deviation calculation.

Table A.S7. Breakdown for Figure of Merit of Cy0

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
Quantification	1	9.89E+07	1.10E+07	1.51E+06	1.20E+05	1.43E+04	1.33E+03	1.59E+02
	2	1.01E+08	1.03E+07	8.41E+05	1.01E+05	8.42E+03	1.02E+03	9.66E+01
	3	9.84E+07	1.04E+07	8.90E+05	1.03E+05	8.83E+03	9.82E+02	8.71E+01
	4	9.98E+07	1.06E+07	9.02E+05	1.03E+05	8.60E+03	9.70E+02	1.17E+02
	5	9.67E+07	1.04E+07	8.78E+05	1.05E+05	8.62E+03	1.01E+03	8.22E+01
	6	9.78E+07	9.88E+06	8.86E+05	1.02E+05	8.94E+03	1.06E+03	8.95E+01
	7	9.34E+07	1.01E+07	8.88E+05	1.03E+05	9.00E+03	1.05E+03	7.71E+01
	8	9.46E+07	1.05E+07	1.06E+06	1.07E+05	1.08E+04	1.02E+03	8.94E+01

	CV	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
		2.61	3.09	22.80	5.67	20.53	10.92	26.76

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
RE	1	1.12	9.57	51.28	19.63	42.51	33.04	58.70
	2	0.99	2.83	15.86	1.28	15.78	1.83	3.42
	3	1.57	3.68	10.98	3.17	11.65	1.78	12.94
	4	0.24	6.04	9.79	2.55	14.03	2.99	17.29
	5	3.25	3.91	12.17	5.11	13.79	1.06	17.75
	6	2.20	1.16	11.42	2.05	10.55	6.01	10.53
	7	6.55	1.07	11.20	3.23	9.98	4.90	22.95
	8	5.37	4.58	5.84	6.79	7.54	2.07	10.55

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
LOOCV	1	1.18	9.95	52.77	20.02	43.63	34.41	63.06
	2	1.05	2.94	16.19	1.30	16.11	1.90	3.63
	3	1.66	3.82	11.22	3.23	11.90	1.85	13.69
	4	0.26	6.28	10.00	2.60	14.33	3.10	18.43
	5	3.45	4.05	12.42	5.21	14.08	1.10	18.76
	6	2.33	1.20	11.66	2.08	10.78	6.24	11.15
	7	6.94	1.11	11.44	3.29	10.20	5.09	24.21
	8	5.68	4.76	5.98	6.92	7.71	2.15	11.17

	Replicate	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
Q	1	3.446499155	293.8727203	61694.83744	2229.153151	38080.94797	12414.46875	99065.56028
	2	2.704335868	25.64423697	5856.179975	9.4490826	5217.224456	38.01580127	332.6606835
	3	6.773396564	43.44835471	2808.754941	58.02041459	2847.704792	35.87320758	4740.053865
	4	0.16475125	117.0191354	2234.028185	37.61831964	4129.709448	101.3143436	8528.344499
	5	29.24540678	48.93045044	3446.670889	150.9393028	3985.124014	12.75766633	8910.135307
	6	13.39485809	4.278144208	3035.066203	24.23629323	2336.676232	409.4133361	3143.759043
	7	118.5810533	3.654988773	2921.458507	60.18102025	2090.351446	272.132678	14865.29181
	8	79.49868599	67.2800651	795.926136	266.35222	1194.209994	48.59985426	3155.958584

Av. Q	3327.111107
Std(Q)	10357.03217

Shaded region is not included in the mean and standard deviation calculation.

Table A.S8. Breakdown for Figure of Merit of $-\log_{10}(F_0)$

	Replicate	Concentration (copies/reaction)						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
Estimated concentration	1	9.24E+07	9.38E+06	1.33E+06	6.57E+04	9.49E+03	1.30E+03	1.69E+02
	2	9.58E+07	8.77E+06	9.15E+05	7.82E+04	8.44E+03	1.44E+03	9.32E+01
	3	1.08E+08	9.56E+06	1.05E+06	1.17E+05	1.02E+04	1.38E+03	7.81E+01
	4	1.08E+08	1.20E+07	1.29E+06	1.22E+05	8.20E+03	1.34E+03	1.81E+02
	5	8.03E+07	9.15E+06	1.05E+06	1.36E+05	1.10E+04	1.63E+03	1.14E+02
	6	1.12E+08	8.65E+06	8.13E+05	1.34E+05	1.25E+04	8.43E+02	8.40E+01
	7	8.65E+07	8.74E+06	8.26E+05	1.41E+05	5.82E+03	6.75E+02	3.09E+01
	8	1.36E+08	9.24E+06	9.69E+05	9.12E+04	9.79E+03	1.09E+03	6.59E+01

	CV	Concentration						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
		17.12	11.62	18.87	25.79	21.25	26.44	49.95

	Replicate	Concentration (copies/reaction)						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
RE	1	7.65	6.22	32.90	34.29	5.09	29.83	69.15
	2	4.23	12.35	8.53	21.75	15.64	44.24	6.75
	3	7.96	4.37	5.44	17.25	1.60	37.88	21.92
	4	7.93	20.28	29.19	21.62	18.00	34.21	80.72
	5	19.74	8.54	4.86	36.30	9.76	62.86	14.07
	6	11.53	13.50	18.72	33.56	25.01	15.74	16.01
	7	13.51	12.64	17.41	40.58	41.85	32.45	69.06
	8	35.74	7.62	3.11	8.83	2.07	8.77	34.10

	Replicate	Concentration (copies/reaction)						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
LOOCV	1	8.10	6.44	33.78	34.79	5.20	31.05	74.40
	2	4.48	12.77	8.72	22.10	15.97	46.13	7.16
	3	8.47	4.53	5.56	17.59	1.64	39.47	23.14
	4	8.44	21.12	29.97	22.05	18.37	35.64	87.00
	5	20.79	8.83	4.97	37.07	9.99	65.67	14.99
	6	12.30	13.96	19.10	34.27	25.63	16.28	16.93
	7	14.27	13.07	17.77	41.45	42.59	33.47	71.48
	8	38.41	7.89	3.18	8.98	2.12	9.10	35.85

	Replicate	Concentration (copies/reaction)						
		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
Q	1	1060.418654	465.8733045	20973.87734	30776.12302	562.5879396	24487.22015	257007.1228
	2	323.8706979	1832.30566	1403.671426	12400.58407	5306.383246	53951.07103	2413.597403
	3	1154.52433	229.6986509	571.1291957	7827.451923	55.58405516	39524.89899	25335.17332
	4	1145.856701	4977.746725	16505.17504	12296.32648	7024.969033	32226.4529	350820.2001
	5	7024.242078	876.1035038	455.7804334	34709.15559	2071.538524	109116.6425	10533.75
	6	2426.695163	2189.612488	6748.799012	29662.98078	13620.99001	6773.502357	13541.05669
	7	3301.099969	1919.845431	5838.601616	43385.93276	37870.80065	28710.48399	246581.6781
	8	23496.23644	699.0459642	186.2176545	2043.665944	92.9962591	2109.96936	61074.02853

Av. Q	13383.6404
Std(Q)	19965.62983

Shaded region is not included in the mean and standard deviation calculation.

Table A.S9. Breakdown for Figure of Merit of M0

		Concentration						
Quantification	Replicate	1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
	1	9.79E+07	1.07E+07	1.49E+06	1.10E+05	1.34E+04	1.33E+03	1.60E+02
	2	1.00E+08	1.01E+07	8.51E+05	9.76E+04	8.42E+03	1.07E+03	9.61E+01
	3	9.97E+07	1.03E+07	9.12E+05	1.05E+05	9.01E+03	1.03E+03	8.57E+01
	4	1.01E+08	1.08E+07	9.49E+05	1.05E+05	8.54E+03	1.02E+03	1.25E+02
	5	9.42E+07	1.02E+07	9.01E+05	1.09E+05	8.92E+03	1.08E+03	8.62E+01
	6	9.97E+07	9.70E+06	8.75E+05	1.06E+05	9.38E+03	1.03E+03	8.87E+01
	7	9.24E+07	9.90E+06	8.79E+05	1.08E+05	8.46E+03	9.85E+02	6.76E+01
	8	9.97E+07	1.03E+07	1.05E+06	1.04E+05	1.06E+04	1.03E+03	8.56E+01

		Concentration						
CV		1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
		3.16	3.66	21.27	3.60	17.82	10.04	29.53

		Concentration						
RE	Replicate	1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
	1	2.08	7.18	48.51	9.81	34.47	32.59	60.20
	2	0.21	0.51	14.86	2.38	15.76	7.03	3.89
	3	0.27	2.50	8.81	5.07	9.88	3.10	14.26
	4	0.87	7.97	5.06	5.08	14.62	1.61	24.77
	5	5.80	2.04	9.92	9.09	10.76	8.18	13.80
	6	0.34	3.00	12.50	6.03	6.19	2.57	11.33
	7	7.58	1.00	12.13	7.87	15.44	1.51	32.38
	8	0.35	2.76	4.53	4.41	6.12	2.99	14.37

		Concentration						
LOOCV	Replicate	1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
	1	2.21	7.45	49.90	9.99	35.35	33.94	64.68
	2	0.23	0.53	15.17	2.43	16.09	7.30	4.12
	3	0.28	2.59	9.00	5.16	10.09	3.21	15.09
	4	0.93	8.28	5.17	5.18	14.94	1.67	26.44
	5	6.14	2.11	10.13	9.26	11.00	8.49	14.59
	6	0.36	3.11	12.77	6.14	6.32	2.67	12.00
	7	8.02	1.03	12.38	8.02	15.77	1.56	34.06
	8	0.37	2.87	4.64	4.49	6.27	3.10	15.19

		Concentration						
Q	Replicate	1.00E+08	1.00E+07	1.00E+06	1.00E+05	1.00E+04	1.00E+03	1.00E+02
	1	14.48355173	196.0455799	51492.29758	352.4987999	21721.63372	11102.42442	114977.4461
	2	0.151085871	0.996125511	4794.369153	20.80382594	4518.99283	515.2650383	473.4225522
	3	0.240514715	23.74490618	1688.21065	94.08412553	1776.523547	100.031615	6354.210365
	4	2.54494666	241.6626081	556.8389089	94.55462541	3892.730628	26.9552615	19337.51929
	5	112.3744416	15.77897784	2138.129748	302.7398044	2109.697609	697.2277884	5945.29875
	6	0.38928774	34.24071887	3395.910064	133.1499656	697.2143991	68.87569779	4015.152291
	7	191.6758201	3.782885288	3193.689604	226.9843175	4338.91238	23.57942485	32567.2971
	8	0.399221687	28.97160266	446.8837465	71.19091061	683.6123534	93.11477616	6444.590448

Av. Q	2546.596033
Std(Q)	8058.439539

Shaded region is not included in the mean and standard deviation calculation.

Table A.S10. Experimental values for non-specific outlier detection experiment.

Outlier ID	Ct	Cy0	-log10(F0)	M0	Sigmoidal Curve-fitting Parameters				
					Fb	Fmax	c	b	d
Outlier 1	22.18460	20.16701	6.24107	N/A	0.00108	0.16458	22.37300	2.98314	2.06180
	21.63717	19.66722	6.00429	N/A	0.00165	0.20330	21.78228	2.98460	2.09782
	21.49195	19.51880	6.04543	N/A	0.00127	0.17532	21.76057	2.94959	2.03027
	21.61322	19.64198	6.04332	N/A	0.00142	0.18405	21.85918	2.96545	2.04505
	21.55848	19.57242	6.02639	N/A	0.00113	0.19108	21.75289	2.94269	2.06273
	21.43270	19.45167	5.98519	N/A	0.00141	0.19163	21.45900	2.98925	2.15545
	21.44939	19.45573	5.98497	N/A	0.00132	0.18354	21.65420	2.97424	2.05931
	21.73830	19.77457	6.02388	N/A	0.00159	0.18908	22.14562	2.96286	1.97109
Outlier 2	18.44049	16.09981	5.61749	N/A	0.00098	0.19616	12.36939	3.70596	8.27322
	18.37323	16.03334	5.62648	N/A	0.00103	0.21221	12.06129	3.66681	8.86532
	18.38343	16.04607	5.64900	N/A	0.00101	0.20760	12.16554	3.66055	8.68182
	18.37301	16.01949	5.61605	N/A	0.00102	0.21130	12.00102	3.68541	8.92312
	18.43692	16.05071	5.62262	N/A	0.00082	0.20214	11.98671	3.73028	8.93332
	18.36191	16.05032	5.64676	N/A	0.00115	0.21558	12.02351	3.62638	9.07374
	18.34952	16.04050	5.68474	N/A	0.00099	0.21375	12.08867	3.59851	8.99036
	18.38125	16.04822	5.66437	N/A	0.00099	0.20719	12.08793	3.64746	8.88694
Outlier 3	19.93116	17.55704	5.13044	N/A	0.00124	0.20157	18.06961	3.73834	3.18304
	18.84150	16.52545	5.05112	N/A	0.00127	0.21137	16.20053	3.68401	3.79378
	18.89363	16.52140	5.05550	N/A	0.00116	0.20746	16.12094	3.72917	3.85576
	18.97990	16.62387	5.05235	N/A	0.00129	0.21675	16.25445	3.71713	3.82810
	19.15945	16.79429	5.13383	N/A	0.00101	0.19113	16.76119	3.71036	3.57039
	18.63558	16.31977	5.04156	N/A	0.00117	0.20856	15.72623	3.68477	4.02451
	18.53768	16.24235	5.07549	N/A	0.00099	0.20603	15.96533	3.58936	3.77196
	19.01092	16.68804	5.05829	N/A	0.00175	0.21230	16.77984	3.68891	3.45259

Table A.S11. Experimental values for temperature variation experiment.

Temperature (°C)	Ct	Cy0	-log10(F0)	M0	Sigmoidal Curve-fitting Parameters				
					Fb	Fmax	c	b	d
54.0	20.36472	18.71671	7.52739	0.40680	0.00019	0.16920	20.82675	2.14960	1.77959
	20.34576	18.70276	7.35635	0.28505	0.00062	0.20736	20.81465	2.18309	1.78861
	20.37727	18.70277	7.42406	0.32766	0.00029	0.18938	20.61377	2.21799	1.92214
	20.37547	18.71779	7.39350	0.32401	0.00044	0.18941	20.72816	2.20445	1.85649
	20.34697	18.73254	7.37018	0.32450	0.00078	0.19933	20.98318	2.14943	1.69672
55.0	20.35076	18.70896	7.41330	0.32720	0.00066	0.19168	20.74375	2.19071	1.83181
	20.34589	18.70457	7.38593	0.30547	0.00051	0.20034	20.88222	2.16314	1.74592
	20.39589	18.72499	7.38224	0.32448	0.00049	0.20133	20.56239	2.24494	1.97382
	20.39395	18.72994	7.33733	0.30143	0.00069	0.20441	20.67652	2.23723	1.90351
	20.42425	18.77627	7.46880	0.43186	0.00008	0.20008	20.88380	2.15034	1.78424
56.9	20.37027	18.75150	7.36937	0.34368	0.00045	0.19184	21.08252	2.12646	1.65080
	20.37564	18.74796	7.35993	0.33413	0.00040	0.19838	21.04501	2.13617	1.67370
	20.39991	18.77594	7.33519	0.34764	0.00067	0.20141	20.94867	2.17618	1.75209
	20.42563	18.77232	7.24463	0.28723	0.00078	0.21952	20.80642	2.23909	1.85573
	20.41883	18.78839	7.42493	0.41685	0.00030	0.18988	21.04506	2.13321	1.69389
59.3	20.37572	18.75679	7.19605	0.24051	0.00091	0.21538	21.08690	2.17507	1.66996
	20.38818	18.78148	7.31135	0.33836	0.00036	0.21527	21.24931	2.09322	1.57093
	20.39076	18.78063	7.25616	0.30289	0.00082	0.21567	21.08949	2.15789	1.67288
	20.40582	18.78285	7.17554	0.25470	0.00099	0.23383	21.07078	2.19077	1.69844
	20.40556	18.81165	7.34236	0.38905	0.00060	0.19931	21.21351	2.11276	1.60749
61.9	20.38439	18.74872	7.40130	0.36089	-0.00020	0.21077	20.97150	2.10500	1.70538
	20.33584	18.74770	7.31287	0.30417	0.00048	0.21610	21.27799	2.06409	1.52489
	20.37290	18.72687	7.14594	0.17816	0.00077	0.21721	20.89255	2.22958	1.78190
	20.43562	18.72142	7.03779	0.10500	0.00088	0.21644	20.39958	2.38595	2.12478
	20.36003	18.72843	7.28283	0.26554	0.00038	0.20818	20.99268	2.15006	1.69807
62.0	20.41051	18.76012	7.48086	0.42268	0.00014	0.14534	20.90459	2.17275	1.77418
	20.39197	18.76013	7.33928	0.33383	0.00056	0.17423	21.02490	2.16699	1.70140
	20.42831	18.78184	7.44075	0.42005	0.00032	0.15973	20.89417	2.18726	1.79439
	20.52006	18.79240	7.51415	0.47739	-0.00040	0.16194	20.55288	2.22705	2.03337
	20.38991	18.77583	7.35947	0.36270	0.00053	0.16913	21.01541	2.16170	1.71138
62.6	20.38280	18.73967	7.36062	0.32604	0.00078	0.16352	20.83232	2.23319	1.81676
	20.39852	18.77213	7.38646	0.37584	0.00047	0.16690	21.07822	2.14955	1.67338
	20.43431	18.78175	7.34838	0.36206	0.00042	0.17135	20.95350	2.19847	1.76730
	20.49831	18.82070	7.54599	0.52649	-0.00039	0.16673	20.85542	2.14898	1.83399
	20.46540	18.82146	7.40142	0.43649	0.00013	0.18356	21.01592	2.14776	1.73624
63.7	20.41626	18.77712	7.41314	0.39781	0.00047	0.16384	20.93468	2.19208	1.76969
	20.43215	18.79243	7.44092	0.43111	0.00035	0.16397	20.97034	2.17289	1.75241
	20.44733	18.78596	7.36589	0.37745	0.00046	0.16898	20.84960	2.22591	1.83667
	20.43170	18.79295	7.40845	0.41129	0.00045	0.16867	20.93550	2.18689	1.77562
	20.46906	18.81612	7.49067	0.48695	0.00020	0.16283	20.87581	2.18652	1.82512
64.7	20.36327	18.65122	7.34938	0.22752	0.00036	0.20699	20.18481	2.30557	2.19262
	20.31742	18.64796	7.22387	0.14525	0.00107	0.22005	20.55469	2.27274	1.93658
	20.37853	18.65491	7.25185	0.17024	0.00080	0.21916	20.15218	2.35872	2.22943
	20.37585	18.64610	7.16426	0.10621	0.00107	0.21438	20.11381	2.40256	2.25970
	20.34921	18.61443	7.25960	0.13315	0.00068	0.20605	20.05181	2.36940	2.27506
65.2	20.43822	18.71953	7.24305	0.23176	0.00105	0.18152	20.25138	2.40375	2.21454
	20.42829	18.71954	7.37928	0.31714	0.00048	0.17588	20.49530	2.29242	2.03097
	20.46003	18.73462	7.33735	0.30657	0.00071	0.17844	20.31844	2.35687	2.17127
	20.50101	18.76079	7.40170	0.37414	0.00035	0.17237	20.13990	2.36654	2.31995
	20.45599	18.70472	7.41535	0.32457	0.00052	0.16909	20.26672	2.34996	2.18669
67.4	20.45377	18.48566	6.92595	-0.20862	0.00127	0.20727	17.52534	2.90144	4.61491
	20.37989	18.45652	7.01548	-0.18293	0.00125	0.20657	18.26642	2.76729	3.73532
	20.39153	18.44212	6.95625	-0.23488	0.00119	0.21009	17.69905	2.84843	4.36246
	20.38347	18.42486	6.92194	-0.27427	0.00136	0.21405	17.28423	2.90740	4.85197
	20.39764	18.44359	7.02391	-0.19091	0.00094	0.20133	17.67030	2.83803	4.41447
68.4	20.53432	18.42980	6.89546	-0.28509	0.00043	0.16600	16.14991	3.10894	6.54409
	20.49266	18.44686	6.83167	-0.30764	0.00119	0.17262	16.70523	3.09419	5.59168
	20.53705	18.41536	6.81213	-0.35224	0.00072	0.17192	16.17232	3.16518	6.37963
	20.53589	18.43272	6.93089	-0.25985	0.00021	0.16609	16.23338	3.07248	6.46712
	20.51385	18.39090	6.79396	-0.38901	0.00063	0.17053	15.93598	3.17754	6.75246
69.8	20.41249	18.23849	6.40807	-0.78888	0.00109	0.20294	15.33253	3.36855	7.27589
	20.41747	18.26823	6.45102	-0.73121	0.00098	0.20575	15.22802	3.31863	7.63894
	20.41780	18.24772	6.44430	-0.75660	0.00104	0.20434	15.17342	3.35466	7.63455
	20.61922	18.32978	6.23095	-0.80469	0.00065	0.20668	14.60788	3.60700	8.50319
	20.42243	18.22635	6.41143	-0.79927	0.00084	0.19605	15.24092	3.38203	7.42636

Temperature (°C)	Sigmoidal Curve-fitting Parameters								
	Ct	Cy0	-log10(F0)	M0	Fb	Fmax	c	b	d
70.0	20.74806	18.44930	6.03475	-0.80365	0.00082	0.16904	15.16445	3.68778	7.46727
	20.60356	18.35253	6.26511	-0.75984	0.00088	0.16672	15.02772	3.60165	7.67607
	20.62427	18.34432	6.25985	-0.77153	0.00078	0.16474	15.05547	3.62830	7.57422
	20.64095	18.35823	6.40274	-0.66749	0.00008	0.15721	15.38570	3.48703	7.30104
71.5	20.76420	18.43413	6.06676	-0.79918	0.00049	0.16405	15.06840	3.68114	7.66155
	22.00085	19.55338	5.72672	0.14952	0.00110	0.15902	18.25436	3.96320	4.59345
	22.06136	19.64796	5.83418	0.31489	0.00066	0.15096	18.91699	3.87376	4.12782
	21.89426	19.47358	5.83790	0.13631	0.00063	0.15639	17.47154	3.94200	5.37595
71.9	21.95624	19.47948	6.01417	0.25321	-0.00028	0.14694	16.78302	3.85915	6.42304
	22.26285	19.82939	5.77029	0.46318	0.00048	0.15164	18.16812	3.95244	5.00574
	21.82854	19.45337	5.68766	0.02093	0.00126	0.19959	16.57795	4.00858	6.39325
	21.70507	19.32861	5.71016	-0.09441	0.00081	0.20207	17.06401	3.89975	5.71849
72.7	21.68929	19.30635	5.75117	-0.09177	0.00076	0.18518	16.97161	3.91519	5.79432
	21.68989	19.27806	5.72871	-0.13507	0.00061	0.19773	16.54539	3.95119	6.30403
	22.14798	19.75238	5.84279	0.42856	0.00075	0.16901	17.73599	3.95168	5.38159
	29.27482	27.23517	6.07609	8.33772	0.00139	0.13414	32.16381	3.44074	1.32696
73.3	27.30637	24.96968	5.89784	5.87651	0.00101	0.13731	27.47575	3.94222	2.20929
	26.29511	23.91018	5.88241	4.76767	0.00072	0.13070	25.93275	3.95333	2.41984
	27.14658	24.77912	5.92470	5.69576	0.00093	0.13004	26.75693	4.00640	2.44168
	37.14676	33.62933	6.07270	14.97709	0.00036	0.19567	36.62494	6.13472	2.48132
73.6	29.76530	27.62594	6.18037	8.80903	0.00138	0.15107	31.95165	3.60649	1.52664
	27.75944	25.50957	5.77067	6.35660	0.00145	0.16679	28.94448	3.76794	1.83080
	28.11433	25.71665	5.81668	6.60099	0.00131	0.16450	28.41290	3.99881	2.14824
	26.05641	23.67020	5.73693	4.42744	0.00113	0.17675	25.33975	3.94051	2.57152
73.6	32.59289	29.76913	6.51494	11.24574	0.00029	0.16556	32.74108	4.55253	2.21121
	33.94676	30.78618	5.96904	11.96032	0.00155	0.16958	33.74045	6.02777	2.43656
	32.61359	29.77271	5.99564	10.92393	0.00128	0.14406	33.53012	5.11220	2.05708
	35.49216	32.25844	5.97058	13.48930	0.00112	0.18466	36.00561	5.95166	2.22506
73.6	34.29546	31.07854	6.25448	12.44291	0.00102	0.15645	33.57587	5.94813	2.57798
	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table A.S12. Estimated quantification (copies/reaction) for temperature variation experiments

Temperature (°C)	Ct_quant	Cy_quant	-log10(F0)_quant	M0_quant	MahalDistSq	Outlier (p<0.01)	Outlier (p<0.001)
54.0	9.00E+04	8.98E+04	3.94E+04	7.98E+04	15.80888	1	1
	9.12E+04	9.06E+04	5.97E+04	8.54E+04	3.72041	0	0
	8.92E+04	9.06E+04	5.07E+04	8.34E+04	9.81176	1	0
	8.94E+04	8.97E+04	5.46E+04	8.35E+04	6.17608	0	0
	9.11E+04	8.88E+04	5.78E+04	8.35E+04	3.43553	0	0
55.0	9.09E+04	9.02E+04	5.20E+04	8.34E+04	6.61385	0	0
	9.12E+04	9.05E+04	5.56E+04	8.44E+04	5.09417	0	0
	8.81E+04	8.93E+04	5.61E+04	8.35E+04	6.29589	0	0
	8.82E+04	8.90E+04	6.26E+04	8.46E+04	3.44637	0	0
	8.64E+04	8.62E+04	4.54E+04	7.87E+04	9.51227	1	0
56.9	8.97E+04	8.77E+04	5.79E+04	8.26E+04	3.22295	0	0
	8.93E+04	8.79E+04	5.92E+04	8.31E+04	2.98724	0	0
	8.79E+04	8.63E+04	6.29E+04	8.25E+04	1.89064	0	0
	8.64E+04	8.65E+04	7.84E+04	8.53E+04	0.23503	0	0
	8.68E+04	8.55E+04	5.05E+04	7.93E+04	5.56744	0	0
59.3	8.93E+04	8.74E+04	8.82E+04	8.75E+04	0.97742	0	0
	8.86E+04	8.59E+04	6.66E+04	8.29E+04	1.59929	0	0
	8.84E+04	8.60E+04	7.62E+04	8.45E+04	1.04814	0	0
	8.75E+04	8.59E+04	9.27E+04	8.68E+04	1.19408	0	0
	8.75E+04	8.42E+04	6.18E+04	8.06E+04	2.38773	0	0
61.9	8.88E+04	8.79E+04	5.35E+04	8.18E+04	5.05873	0	0
	9.18E+04	8.79E+04	6.64E+04	8.45E+04	2.48206	0	0
	8.95E+04	8.92E+04	9.96E+04	9.06E+04	0.52527	0	0
	8.58E+04	8.95E+04	1.30E+05	9.43E+04	2.40121	0	0
	9.03E+04	8.91E+04	7.14E+04	8.63E+04	0.95897	0	0
62.0	8.73E+04	8.72E+04	4.41E+04	7.91E+04	10.94358	1	0
	8.84E+04	8.72E+04	6.23E+04	8.31E+04	2.18943	0	0
	8.62E+04	8.59E+04	4.86E+04	7.92E+04	7.36837	0	0
	8.10E+04	8.53E+04	4.07E+04	7.67E+04	23.23125	1	1
	8.85E+04	8.63E+04	5.93E+04	8.18E+04	2.64154	0	0
62.6	8.89E+04	8.84E+04	5.91E+04	8.34E+04	3.47163	0	0
	8.80E+04	8.65E+04	5.55E+04	8.12E+04	3.78358	0	0
	8.59E+04	8.59E+04	6.09E+04	8.18E+04	2.81780	0	0
	8.22E+04	8.37E+04	3.77E+04	7.47E+04	18.17028	1	1
	8.41E+04	8.37E+04	5.35E+04	7.85E+04	4.38302	0	0
63.7	8.69E+04	8.62E+04	5.20E+04	8.02E+04	5.41980	0	0
	8.60E+04	8.53E+04	4.86E+04	7.87E+04	6.82898	0	0
	8.51E+04	8.57E+04	5.84E+04	8.11E+04	3.93775	0	0
	8.60E+04	8.53E+04	5.26E+04	7.96E+04	4.92475	0	0
	8.39E+04	8.40E+04	4.31E+04	7.63E+04	10.74028	1	0
64.7	9.01E+04	9.38E+04	6.07E+04	8.81E+04	9.41522	1	0
	9.29E+04	9.40E+04	8.24E+04	9.23E+04	0.71597	0	0
	8.92E+04	9.36E+04	7.70E+04	9.10E+04	5.12620	0	0
	8.93E+04	9.41E+04	9.53E+04	9.43E+04	2.88542	0	0
	9.10E+04	9.62E+04	7.56E+04	9.29E+04	7.25503	0	0
65.2	8.56E+04	8.96E+04	7.87E+04	8.79E+04	3.65354	0	0
	8.62E+04	8.96E+04	5.65E+04	8.39E+04	9.87237	1	0
	8.44E+04	8.87E+04	6.26E+04	8.44E+04	8.83050	0	0
	8.21E+04	8.71E+04	5.35E+04	8.12E+04	15.21127	1	1
	8.46E+04	9.05E+04	5.17E+04	8.35E+04	19.65390	1	1
67.4	8.47E+04	1.05E+05	1.70E+05	1.12E+05	43.32151	1	1
	8.91E+04	1.07E+05	1.37E+05	1.11E+05	34.53592	1	1
	8.84E+04	1.08E+05	1.58E+05	1.14E+05	39.33326	1	1
	8.89E+04	1.09E+05	1.72E+05	1.16E+05	40.76519	1	1
	8.80E+04	1.08E+05	1.34E+05	1.11E+05	44.56879	1	1
68.4	8.02E+04	1.09E+05	1.83E+05	1.17E+05	95.13760	1	1
	8.25E+04	1.08E+05	2.14E+05	1.19E+05	66.43159	1	1
	8.01E+04	1.10E+05	2.24E+05	1.22E+05	97.58370	1	1
	8.01E+04	1.09E+05	1.68E+05	1.16E+05	97.36687	1	1
	8.13E+04	1.12E+05	2.35E+05	1.24E+05	97.38880	1	1
69.8	8.71E+04	1.24E+05	6.00E+05	1.55E+05	113.71959	1	1
	8.68E+04	1.21E+05	5.40E+05	1.50E+05	102.96621	1	1
	8.68E+04	1.23E+05	5.49E+05	1.52E+05	111.38571	1	1
	7.57E+04	1.16E+05	9.23E+05	1.56E+05	173.79202	1	1
	8.66E+04	1.25E+05	5.95E+05	1.56E+05	122.96730	1	1

Temperature (°C)	Ct_quant	Cy_quant	-log10(F0)_quant	M0_quant	MahalDistSq	Outlier (p<0.01)	Outlier (p<0.001)
70.0	6.94E+04	1.07E+05	1.49E+06	1.56E+05	196.83714	1	1
	7.65E+04	1.15E+05	8.49E+05	1.52E+05	154.38650	1	1
	7.55E+04	1.15E+05	8.60E+05	1.53E+05	168.08292	1	1
	7.46E+04	1.14E+05	6.08E+05	1.45E+05	166.31021	1	1
	6.86E+04	1.09E+05	1.38E+06	1.56E+05	207.39946	1	1
71.5	2.97E+04	5.12E+04	3.15E+06	9.20E+04	372.42053	1	1
	2.85E+04	4.80E+04	2.42E+06	8.40E+04	339.53794	1	1
	3.19E+04	5.40E+04	2.40E+06	9.27E+04	330.72542	1	1
	3.06E+04	5.38E+04	1.56E+06	8.69E+04	328.50617	1	1
	2.48E+04	4.25E+04	2.83E+06	7.73E+04	375.62239	1	1
71.9	3.33E+04	5.47E+04	3.46E+06	9.88E+04	346.99193	1	1
	3.62E+04	5.95E+04	3.28E+06	1.05E+05	332.46719	1	1
	3.66E+04	6.04E+04	2.96E+06	1.05E+05	323.48567	1	1
	3.66E+04	6.16E+04	3.13E+06	1.08E+05	338.15955	1	1
	2.68E+04	4.48E+04	2.37E+06	7.88E+04	337.40204	1	1
72.7	2.13E+02	2.93E+02	1.34E+06	9.77E+02	1353.71622	1	1
	8.10E+02	1.34E+03	2.07E+06	3.83E+03	940.99661	1	1
	1.61E+03	2.74E+03	2.15E+06	7.09E+03	777.50706	1	1
	9.02E+02	1.53E+03	1.94E+06	4.24E+03	890.94052	1	1
	1.02E+00	3.98E+00	1.36E+06	2.45E+01	3083.72685	1	1
73.3	1.53E+02	2.25E+02	1.04E+06	7.52E+02	1308.98610	1	1
	5.95E+02	9.34E+02	2.83E+06	2.94E+03	1143.44541	1	1
	4.68E+02	8.13E+02	2.53E+06	2.56E+03	1129.26146	1	1
	1.89E+03	3.22E+03	3.07E+06	8.56E+03	821.55818	1	1
	2.24E+01	5.33E+01	4.62E+05	1.95E+02	1490.49511	1	1
73.6	8.93E+00	2.69E+01	1.74E+06	1.31E+02	2257.84980	1	1
	2.21E+01	5.32E+01	1.64E+06	2.33E+02	1884.37851	1	1
	3.13E+00	1.00E+01	1.74E+06	5.60E+01	2647.27747	1	1
	7.05E+00	2.21E+01	8.71E+05	1.00E+02	2136.30201	1	1
	N/A	N/A	N/A	N/A	N/A	N/A	N/A

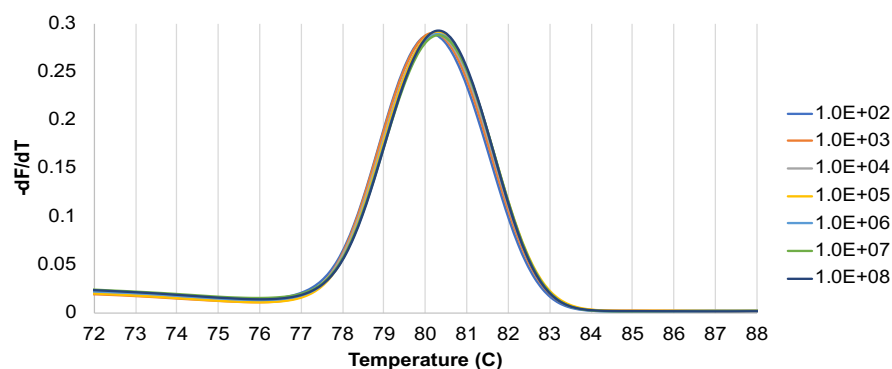


Figure A.S1. Melting curve analysis for lambda DNA standard experiment.

This figure shows average melting curves peaks for synthetic lambda DNA standard experiments using the 242bp double-stranded DNA molecule (gBlock gene fragment ordered from IDT) using in-house lambda primers. Ten-fold dilution from 10^8 to 10^2 copies/reaction were used in this experiment, 8-reactions per tested concentration. Average melting curve peak was 80.26°C (SD = 0.1°C) for all positive reactions and no secondary melting event was observed at other annealing temperatures.

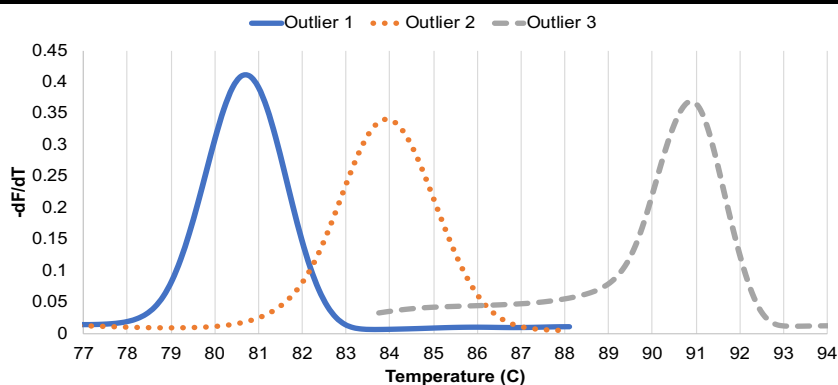


Figure A.S2. Melting curve analysis for non-specific outlier detection experiment.

This figure shows average melting curves peaks of 80.66°C (SD = 0.07°C) for Outlier 1, 83.97°C (SD = 0.10°C) for Outlier 2 and 90.76°C (SD = 0.10°C) for Outlier 3. Octuplicate reactions were performed per reaction condition. No secondary melting event was observed at other annealing temperatures. Specific primers sets were selected from Monteiro et al 2012.

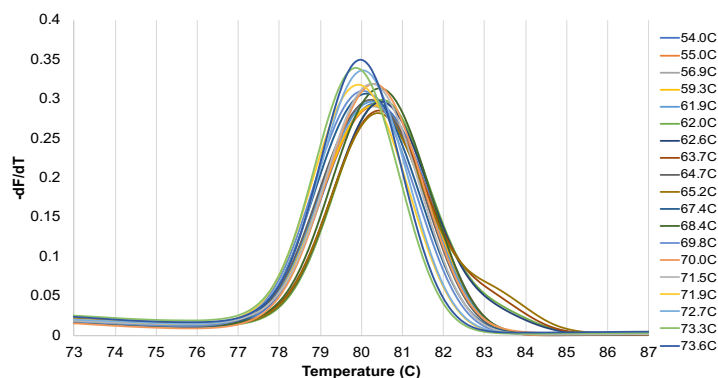


Figure A.S3. Melting curve analysis for temperature variation experiment.

This figure shows average melting curves peaks for temperature variation experiments ranging from 54°C to 73.6°C using phage lambda DNA and in-house primers. Observed average melting curve peaks for tested temperatures is 80.07°C (SD = 1.12 °C). Quintuplicate reactions were performed per tested temperature.

Appendix B

Supporting Information:

Chapter 4

This Appendix contains the following:

- Synthetic DNA sequences (Table B.S1)
- Extended numerical values of features and sigmoidal fittings of standards (Table B.S2)
- Extended numerical values of features and sigmoidal fittings of samples (Table B.S3)
- Estimated quantification for samples (Table B.S4)
- Melting curve analysis of standards and samples (Figure B.S1)
- Standard curves generated utilising different features (Figure B.S2)

Table B.S1. Synthetic DNA sequences

Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing bla_{OXA-48} target.

<i>bla</i> _{OXA-48} consensus (5' - 3')	ATGCGTGATTAGCCTTATCGGCTG <u>TGTTTTGGTGGCATCGATTATCGGAAT</u> <u>GCCTGCGGTAGCAAAGGAATGGCAAGAAAACAAAAGTTGGAATGCTCACTT</u> <u>TACTGAACATAAATCACAGGGCGTAGTTGTGCTCTGGAATGAGAATAAGCAG</u> <u>CAAGGATTTACCAATAATCTTAAACGG</u> <u>CGGAACCAAGCATT</u> <u>TTTAC</u> CCGCATC TACCTTTAAAATTCCCAATAGCTTGATCGCCCTCGATTGGGCGTGTTAAGG ATGAACACCAAGTCTTTAAGTGGGATGGACAGACGCGC
---	--

OXA-48-F (TGTTTTGGTGGCATCGAT) **in red**; OXA-48-R (GTAAMRATGCTTGGTTCGC) **in blue**; amplicon size = 177 bp (**bold and underlined**); primers from Monteiro et al 2012

blaOXA-48_consensus: 100 sequences were downloaded from NCBI GeneBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and were aligned by using MUSCLE algorithm (Edar, 2004) in Geneious V.10 (Kearse et al., 2012).

Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing bla_{NDM} target.

<i>bla</i> _{NDM} consensus (5' - 3')	ATGGTTGCGGCGCAACACAGCCTGACTTTCGCCGCCAATGGCTGGGTCGAA CCAGCAACCGCGCCCACTTTGCCCCGCTCAAGGTATTTACCCCGGCCCGG GCC <u>ACACCAGTGACAATATCACCG</u> <u>TTGGGATCGACGGCACCGACATCGCTTT</u> <u>TGGTGGCTGCCTGATC</u> <u>AAGGACAGCAAGGCCAA</u> GTGCTCGGCAATCTCGG TGATGCCGACACTGAGCACTACGCCGCGTCAGCGCGCGCGTTTGGTGCGGC GTTCCCCAAGGCCAGCATGATCGTGATGAGCCATTCCGCCCCC
--	---

NDM-F (TTGGCCTTGCTGTCCTTG) **in red**; NDM-R (ACACCAGTGACAATATCACCG) **in blue**; amplicon size = 82 bp (**bold and underlined**); primers from Monteiro et al 2012

blaNDM_consensus: 100 sequences were downloaded from NCBI GeneBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and were aligned by using MUSCLE algorithm (Edar, 2004) in Geneious V.10 (Kearse et al., 2012).

Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing bla_{VIM} target.

<i>bla</i> _{VIM} consensus (5' - 3')	GATGGT <u>GTTTGGTCGCATATCGCAAC</u> GCAGTCGTTGATGGCGCAGTCTACC <u>CGTCCAATGGTCTCATTGTCCGTGATGGTGATGAGTTGCTTTTGATTGATACA</u> <u>GCGTGGGGTGCGAAAACACAGCGGCACTTCTCGCGGAGATTGAGAAGCA</u> <u>AATTGGACTTCCTGTAACGCGTGCAGTCTCCACGCACTTTCATGACGACCGC</u> <u>GTCGGCGGCGTTGATGTCCTTCGGGCGGCTGGGGTGGCAACGTACGCATCA</u> <u>CCGTCGACACGCCGGCTAGCCGAGGTAGAGGGGAACGAGATTCCCACGCA</u> <u>CTCTCTAGAAGGACTCTCATCGAGCGGGGACGCAAGTGCCTTCGGTCCAGT</u> <u>AGAACTCTTCTATCCTGGTGCTGCGCATT</u> CGACCGACAACCTAGTTGTGTACG TCCCGTCTGCGAGTGTGCTCTATGGTGGTTGTGCGATTTATGAGTTGTCACG CACGTCTGCGGGGAACGTGGCCGATGCCGATCTGGCTGAATGGCCACCTC CATTGAGCGGATTCAACAACACTACCCGGAAGCACAGTTCGTATTCCGGGG CACGGCCTGCCGGGCGGTCTAGACTTGCTCAAGCACACAACGAATG
--	--

VIM-F (GTTTGGTCGCATATCGCAAC) **in red**; VIM-R (AATGCGCAGCACCAGGATAG) **in blue**; amplicon size = 382 bp (**bold and underlined**); primers from Monteiro et al 2012

blaVIM_consensus: 100 sequences were downloaded from NCBI GeneBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and were aligned by using MUSCLE algorithm (Edar, 2004) in Geneious V.10 (Kearse et al., 2012).

Nucleotide sequence for synthetic double-stranded DNA ordered from Integrated DNA Technologies containing bla_{KPC} target.

<i>bla</i>_{KPC} consensus (5' - 3')	AACCAT<u>TCGCTAAACTCGAACAGG</u>ACTTTGGCGGCTCCATCGGTGTGTACGC GATGGATACCGGCTCAGGCGCAACTGTAAGTTACCGCGCTGAGGAGCGCTT CCCACTGTGCAGCTCATTCAAGGGCTTTCTTGCTGCCGCTGTGCTGGCTCGC AGCCAGCAGCAGGCCGGCTTGCTGGACACACCCATCCGTTACGGCAAAAAT GCGCTGGTTCGGTGGTCACCCATCTCGGAAAAATATCTGACAACAGGCATG ACGGTGGCGGAGCTGTCCGCGGCCCGCCGTGCAATACAGTGATAACGCCGC CGCCAATTTGTTGCTGAAGGAGTTGGGCGGCCCGGGCTGACGGCCT TCATGCGCTCTATCGGCGATACCACGTTCCGTCTGGACCGCTGGGAGCTGG AGCTGAACTCCGCCATCCCAGGCGATGCGCGCGATACCTCATCGCCGCGCG CCGTGACGGAAAGCTTACAAAACTGACACTGGGCTCTGCACTGGCTGCGC CGCAGCGGCAGCAGTTTGTGATTGGCTAAAGGGAAACACGACCGGCAACC ACCGCATCCGCGCGGCGGTGCCGGCAGACTGGGCAGTCGGAGACAAAACC GGAACCTGCGGAGTGTATGGCACGGCAAATGACTATGCCGTCTGTCTGGCCC ACTGGGCGCGCACCTATTGTGTTGGCCGTCTACACCCGGGCGCCTAACAAG GATGACAAGCACAGCGAGGCCGTCATCGCCGCTGCGGCTAGACTCGCGCT CGAGGGATTGGGCGTCAACGGGCAGTAAG
---	--

KPC-F (TCGCTAAACTCGAACAGG) in red; KPC-R (TTACTGCCCGTTGACGCCCAATCC) in blue; amplicon size = 785 bp (**bold and underlined**); primers from Monteiro et al 2012

blaKPC_consensus: 100 sequences were downloaded from NCBI GeneBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and were aligned by using MUSCLE algorithm (Edar, 2004) in Geneious V.10 (Kearse et al., 2012).

Table B.S2. Extended numerical values of features and sigmoidal fittings of standards

Features and Sigmoidal Fit for bla _{OXA-48} Standard									
DNA (copies/reaction)	C _t	C _y	F ₀	M ₀	F _b	F _{max}	c	b	d
1.00E+04	31.22	29.39	4.74E-09	8.26E+00	3.35E-04	0.18	33.21	2.51	1.29
	31.16	29.31	5.01E-09	8.17E+00	2.52E-04	0.18	33.03	2.53	1.32
	31.09	29.27	4.51E-09	8.16E+00	1.93E-04	0.19	32.99	2.47	1.30
	31.35	29.58	5.13E-09	8.40E+00	3.73E-04	0.18	33.54	2.46	1.22
	31.00	29.11	5.32E-09	7.99E+00	6.47E-05	0.19	32.75	2.53	1.35
	31.13	29.24	4.80E-09	8.13E+00	1.18E-04	0.18	32.78	2.58	1.40
1.00E+05	27.38	25.45	3.89E-08	4.21E+00	2.19E-04	0.19	28.20	2.74	1.75
	27.30	25.32	4.42E-08	4.07E+00	4.37E-04	0.21	27.44	2.86	2.08
	27.30	25.37	4.13E-08	4.12E+00	1.90E-04	0.20	28.18	2.70	1.71
	27.21	25.17	4.28E-08	3.95E+00	1.18E-04	0.19	27.03	2.92	2.24
1.00E+06	23.59	21.54	3.15E-07	1.91E-01	4.32E-04	0.20	23.01	3.02	2.48
	23.39	21.31	2.74E-07	2.03E-02	1.81E-04	0.21	22.50	3.04	2.67
	23.39	21.34	2.57E-07	5.85E-02	3.18E-04	0.20	22.63	3.02	2.59
	23.38	21.31	2.98E-07	2.05E-04	2.86E-04	0.20	22.85	3.02	2.45
	23.19	21.09	3.12E-07	-2.04E-01	1.87E-04	0.19	22.21	3.06	2.72
	23.22	21.16	3.07E-07	-1.44E-01	1.63E-04	0.19	22.66	3.00	2.47
1.00E+07	19.53	17.35	3.02E-06	-4.12E+00	3.29E-04	0.18	17.55	3.31	3.41
	19.45	17.28	2.98E-06	-4.18E+00	2.91E-04	0.18	17.74	3.27	3.21
	19.56	17.39	3.21E-06	-4.10E+00	4.87E-04	0.20	17.66	3.31	3.35
	19.51	17.33	3.25E-06	-4.16E+00	3.23E-04	0.20	17.56	3.31	3.39
	19.41	17.23	2.81E-06	-4.21E+00	3.47E-04	0.19	17.60	3.29	3.28
	19.53	17.33	3.10E-06	-4.14E+00	3.11E-04	0.17	17.73	3.32	3.26
1.00E+08	15.76	13.34	4.69E-05	-8.37E+00	7.11E-04	0.20	12.80	3.80	3.99
	15.75	13.34	4.58E-05	-8.37E+00	7.18E-04	0.20	12.75	3.80	4.03
	16.51	14.08	3.52E-05	-7.63E+00	7.04E-04	0.18	13.30	3.90	4.17
	16.15	13.72	4.34E-05	-8.01E+00	7.32E-04	0.19	13.26	3.81	3.91
	16.08	13.60	3.76E-05	-8.07E+00	6.09E-04	0.18	13.10	3.86	3.94

Features and Sigmoidal Fit for bla _{NDM} Standard									
DNA (copies/reaction)	C _t	C _y	F ₀	M ₀	F _b	F _{max}	c	b	d
1.00E+01	32.95	31.21	2.75E-10	3.29E+01	2.75E-04	0.16	34.95	2.28	1.21
	33.39	31.85	2.04E-10	3.34E+01	2.09E-05	0.16	36.19	1.95	0.90
	32.11	30.34	4.71E-10	3.21E+01	1.13E-04	0.16	34.13	2.25	1.17
	32.61	30.95	3.72E-10	3.26E+01	2.03E-05	0.16	35.14	2.09	1.00
	32.63	31.00	3.86E-10	3.26E+01	3.95E-04	0.16	35.15	2.12	1.02
	33.11	31.41	2.71E-10	3.31E+01	2.78E-04	0.18	35.34	2.20	1.12
	33.10	31.40	3.24E-10	3.31E+01	2.71E-04	0.18	35.36	2.22	1.11
1.00E+02	29.29	27.33	6.18E-09	2.93E+01	5.54E-05	0.16	29.65	2.69	1.92
	29.36	27.39	4.34E-09	2.94E+01	-1.43E-04	0.16	29.51	2.68	2.03
	29.26	27.31	5.12E-09	2.93E+01	8.17E-05	0.16	29.65	2.65	1.90
	29.21	27.29	5.62E-09	2.92E+01	2.31E-04	0.16	29.71	2.63	1.85
	29.36	27.36	3.92E-09	2.94E+01	5.93E-05	0.15	28.80	2.81	2.45
	29.28	27.32	4.60E-09	2.93E+01	-2.14E-05	0.16	29.41	2.68	2.04
	29.17	27.23	6.51E-09	2.92E+01	2.44E-04	0.16	29.49	2.68	1.95
1.00E+03	26.03	23.87	3.89E-08	2.60E+01	-2.77E-05	0.18	23.99	3.11	3.48
	26.05	23.89	2.63E-08	2.60E+01	-8.04E-05	0.16	23.78	3.08	3.70
	25.86	23.74	3.09E-08	2.59E+01	2.52E-05	0.16	24.04	3.02	3.32
	25.84	23.74	3.41E-08	2.58E+01	7.87E-05	0.18	24.77	2.93	2.75
	25.78	23.68	3.14E-08	2.58E+01	-2.96E-05	0.17	24.48	2.96	2.92
	25.91	23.78	2.79E-08	2.59E+01	-1.17E-04	0.16	23.94	3.02	3.45
	25.91	23.78	3.19E-08	2.59E+01	6.18E-05	0.18	23.38	3.11	3.98
1.00E+04	26.04	23.88	4.64E-08	2.60E+01	5.96E-05	0.19	23.93	3.13	3.54
	22.68	20.43	1.63E-07	2.27E+01	6.89E-05	0.18	18.32	3.41	5.96
	22.56	20.34	1.30E-07	2.26E+01	9.24E-05	0.16	18.42	3.33	5.76
	22.52	20.30	1.27E-07	2.25E+01	-3.35E-05	0.17	18.27	3.31	5.94
	22.55	20.35	1.19E-07	2.26E+01	-1.68E-04	0.17	18.54	3.27	5.65
	22.56	20.34	1.27E-07	2.26E+01	-2.91E-05	0.17	18.50	3.30	5.67
	22.54	20.32	1.34E-07	2.25E+01	-1.19E-04	0.17	18.36	3.31	5.85
1.00E+05	22.62	20.40	1.59E-07	2.26E+01	5.12E-05	0.18	18.56	3.36	5.61
	22.66	20.44	1.96E-07	2.27E+01	1.91E-04	0.19	19.46	3.34	4.53
	19.34	16.98	1.20E-06	1.93E+01	-3.60E-05	0.18	13.17	3.62	8.73
	19.20	16.87	1.02E-06	1.92E+01	-1.06E-04	0.16	13.47	3.51	8.11
	19.22	16.88	1.06E-06	1.92E+01	-9.87E-06	0.18	13.03	3.55	9.00
	19.18	16.86	1.04E-06	1.92E+01	6.40E-05	0.17	13.41	3.50	8.21
	19.21	16.87	9.68E-07	1.92E+01	-1.30E-04	0.17	13.48	3.49	8.14
1.00E+06	19.24	16.91	1.02E-06	1.92E+01	-2.23E-04	0.18	13.47	3.50	8.22
	19.28	16.89	1.10E-06	1.93E+01	-9.76E-05	0.18	12.87	3.63	9.18
	15.94	13.44	8.51E-06	1.59E+01	-2.24E-04	0.18	9.07	3.75	9.69
	15.88	13.44	7.59E-06	1.59E+01	-1.99E-04	0.17	9.48	3.65	9.02
	15.87	13.44	5.39E-06	1.59E+01	-2.87E-04	0.18	9.40	3.62	9.29
	15.87	13.43	5.56E-06	1.59E+01	-2.56E-04	0.18	9.46	3.63	9.09
	15.87	13.43	5.53E-06	1.59E+01	-2.26E-04	0.17	9.48	3.63	9.06
1.00E+06	15.90	13.45	7.25E-06	1.59E+01	-2.66E-04	0.17	9.46	3.64	9.13
	15.90	13.46	7.79E-06	1.59E+01	-2.22E-04	0.18	9.53	3.63	8.99
	16.03	13.55	9.39E-06	1.60E+01	-1.89E-04	0.20	9.19	3.72	9.75

Features and Sigmoidal Fit for bla _{NDM} Standard									
DNA (copies/reaction)	C _t	C _y	F ₀	M ₀	F _b	F _{max}	c	b	d
1.00E+07	12.66	9.90	6.15E-05	1.27E+01	-7.75E-04	0.17	5.10	4.14	9.74
	12.63	9.95	5.18E-05	1.26E+01	-6.46E-04	0.16	5.40	3.98	9.57
	12.61	9.94	5.30E-05	1.26E+01	-6.25E-04	0.16	5.38	3.97	9.62
	12.61	9.97	5.09E-05	1.26E+01	-6.38E-04	0.17	5.46	3.93	9.60
	12.62	10.00	4.98E-05	1.26E+01	-5.72E-04	0.16	5.49	3.92	9.63
	12.64	10.02	4.68E-05	1.26E+01	-5.56E-04	0.16	5.57	3.92	9.51
	12.67	9.98	5.08E-05	1.27E+01	-6.37E-04	0.17	5.25	4.00	9.94
	12.85	10.06	8.64E-05	1.28E+01	-8.53E-04	0.20	5.22	4.16	9.78

Features and Sigmoidal Fit for bla _{VIM} Standard									
DNA (copies/reaction)	C _t	C _y	F ₀	M ₀	F _b	F _{max}	c	b	d
1.00E+03	25.56	24.02	2.59E-09	1.11E+01	2.73E-04	0.16	27.27	1.99	1.21
	25.48	23.97	2.58E-09	1.10E+01	5.19E-04	0.16	27.27	1.98	1.18
	25.67	24.11	3.16E-09	1.12E+01	5.33E-04	0.17	27.28	2.05	1.26
	25.56	24.06	3.28E-09	1.10E+01	7.70E-04	0.18	27.48	1.97	1.14
	25.45	23.93	3.09E-09	1.09E+01	6.20E-04	0.18	27.22	1.99	1.19
	25.58	24.01	2.73E-09	1.11E+01	4.44E-04	0.16	27.13	2.05	1.28
	25.43	23.90	2.29E-09	1.09E+01	2.25E-04	0.16	27.12	1.97	1.21
	25.43	23.92	3.21E-09	1.09E+01	6.45E-04	0.17	27.28	1.98	1.16
1.00E+04	22.09	20.43	1.91E-08	6.65E+00	5.27E-05	0.16	23.06	2.11	1.51
	22.09	20.46	2.22E-08	6.64E+00	4.94E-04	0.16	23.26	2.11	1.43
	22.20	20.50	3.00E-08	6.70E+00	5.18E-04	0.18	22.84	2.28	1.72
	22.18	20.51	2.31E-08	6.73E+00	4.89E-04	0.16	22.97	2.21	1.63
	22.14	20.45	2.73E-08	6.65E+00	5.79E-04	0.16	22.88	2.26	1.67
	22.22	20.52	2.66E-08	6.74E+00	4.04E-04	0.16	22.88	2.27	1.71
	22.22	20.51	2.47E-08	6.75E+00	3.66E-04	0.16	22.76	2.29	1.77
	22.10	20.47	2.27E-08	6.65E+00	4.89E-04	0.17	23.36	2.09	1.39
1.00E+05	18.70	16.91	2.20E-07	2.27E+00	4.58E-04	0.17	18.24	2.48	2.39
	18.60	16.90	1.90E-07	2.20E+00	2.84E-04	0.16	19.03	2.27	1.83
	18.69	16.88	2.41E-07	2.23E+00	4.78E-04	0.18	18.40	2.46	2.26
	18.58	16.89	2.15E-07	2.16E+00	5.40E-04	0.16	19.06	2.30	1.81
	18.59	16.90	1.92E-07	2.19E+00	3.13E-04	0.16	19.05	2.27	1.81
	18.62	16.91	2.04E-07	2.21E+00	3.34E-04	0.16	19.04	2.29	1.84
	18.71	16.94	2.10E-07	2.29E+00	3.61E-04	0.17	18.57	2.42	2.18
1.00E+06	15.19	13.43	1.66E-06	-2.15E+00	3.20E-04	0.16	14.81	2.39	2.33
	15.16	13.41	1.77E-06	-2.18E+00	3.78E-04	0.15	14.97	2.37	2.20
	15.17	13.42	1.79E-06	-2.18E+00	3.37E-04	0.16	14.99	2.36	2.19
	15.18	13.39	1.74E-06	-2.17E+00	3.56E-04	0.15	14.70	2.44	2.39
	15.17	13.43	1.87E-06	-2.17E+00	3.35E-04	0.16	15.10	2.35	2.13
	15.17	13.45	1.63E-06	-2.15E+00	2.82E-04	0.15	15.25	2.31	2.03
	15.17	13.48	1.43E-06	-2.12E+00	1.32E-04	0.14	15.45	2.24	1.90
	15.18	13.46	1.60E-06	-2.14E+00	2.02E-04	0.15	15.27	2.29	2.02
1.00E+07	11.75	9.97	1.49E-05	-6.51E+00	3.64E-04	0.15	10.66	2.49	2.89
	11.76	9.97	1.53E-05	-6.51E+00	3.33E-04	0.15	10.83	2.48	2.75
	11.72	9.81	1.63E-05	-6.60E+00	4.21E-04	0.14	9.32	2.74	4.11
	11.74	9.95	1.74E-05	-6.54E+00	4.51E-04	0.16	10.67	2.51	2.87
	11.73	9.99	1.51E-05	-6.52E+00	2.31E-04	0.16	11.48	2.35	2.24
	11.75	10.02	1.48E-05	-6.49E+00	2.39E-04	0.15	11.59	2.33	2.18
	11.74	10.05	1.36E-05	-6.48E+00	1.53E-04	0.15	11.83	2.27	2.03
	11.79	10.05	1.56E-05	-6.46E+00	1.70E-04	0.17	11.55	2.34	2.23
1.00E+08	8.31	6.39	1.73E-04	-1.10E+01	-4.59E-04	0.16	6.01	2.62	4.06
	8.29	6.39	1.82E-04	-1.10E+01	-6.91E-04	0.16	6.72	2.53	3.28
	8.28	6.33	1.88E-04	-1.10E+01	-7.01E-04	0.16	6.35	2.59	3.61
	8.29	6.38	1.72E-04	-1.10E+01	-5.88E-04	0.15	6.35	2.57	3.66
	8.28	6.43	1.54E-04	-1.10E+01	-6.33E-04	0.14	7.17	2.44	2.88
	8.27	6.40	1.69E-04	-1.10E+01	-7.57E-04	0.15	7.24	2.44	2.80
	8.29	6.45	1.62E-04	-1.09E+01	-5.27E-04	0.15	6.94	2.47	3.11
	8.31	6.46	1.68E-04	-1.09E+01	-4.82E-04	0.16	6.77	2.50	3.29

Features and Sigmoidal Fit for bla _{KPC} Standard									
DNA (copies/reaction)	C _t	C _y	F ₀	M ₀	F _b	F _{max}	c	b	d
1.00E+03	32.20	30.30	3.37E-08	1.38E+01	-9.83E-05	0.16	35.30	2.60	1.04
	31.62	29.82	2.97E-08	1.32E+01	-7.09E-06	0.15	34.89	2.47	0.97
	31.89	30.12	3.47E-08	1.36E+01	5.73E-05	0.16	35.41	2.39	0.90
	32.36	30.32	4.69E-08	1.38E+01	2.59E-07	0.18	34.98	2.92	1.24
	32.81	30.77	6.03E-08	1.43E+01	-9.23E-05	0.18	35.77	2.86	1.14
1.00E+04	27.80	25.70	2.05E-07	8.00E+00	1.55E-04	0.17	29.11	3.03	1.63
	27.71	25.64	1.91E-07	7.93E+00	-1.24E-04	0.18	29.08	2.96	1.60
	28.02	25.88	1.67E-07	8.25E+00	-1.44E-04	0.18	29.05	3.05	1.72
	28.06	25.97	1.85E-07	8.34E+00	-2.34E-05	0.18	29.38	2.98	1.61
	28.32	26.13	2.57E-07	8.52E+00	-8.28E-05	0.19	29.21	3.20	1.81
	28.55	26.50	2.09E-07	8.97E+00	-7.17E-05	0.18	30.41	2.95	1.45
1.00E+05	23.70	21.56	9.03E-07	2.78E+00	-5.86E-05	0.17	24.13	3.12	1.98
	23.70	21.43	1.09E-06	2.63E+00	8.85E-05	0.19	23.02	3.40	2.53
	23.71	21.48	1.03E-06	2.69E+00	3.03E-05	0.19	23.43	3.31	2.32
	23.78	21.50	9.76E-07	2.73E+00	-1.01E-04	0.18	23.21	3.37	2.46
	24.01	21.72	1.05E-06	2.99E+00	-4.32E-05	0.20	23.54	3.37	2.40
	23.99	21.69	9.24E-07	2.97E+00	-2.55E-05	0.19	22.80	3.43	2.81
1.00E+06	19.54	17.06	6.87E-06	-2.89E+00	2.05E-04	0.19	15.68	3.87	4.78
	19.51	17.06	7.13E-06	-2.89E+00	7.46E-05	0.20	16.33	3.78	4.19
	19.60	17.13	6.77E-06	-2.80E+00	2.23E-04	0.20	15.39	3.90	5.14
	19.73	17.22	6.67E-06	-2.68E+00	2.40E-04	0.19	15.59	3.92	5.01
	19.78	17.30	6.31E-06	-2.59E+00	2.24E-04	0.18	15.97	3.88	4.72
	19.85	17.34	6.79E-06	-2.54E+00	1.28E-04	0.19	16.11	3.92	4.61
1.00E+07	15.73	12.78	5.69E-05	-8.27E+00	3.55E-06	0.19	7.49	4.67	9.39
	15.69	12.75	5.62E-05	-8.31E+00	3.42E-05	0.19	7.45	4.67	9.42
	15.72	12.82	5.53E-05	-8.23E+00	8.81E-05	0.19	7.59	4.63	9.34
	15.80	12.87	5.39E-05	-8.16E+00	1.43E-05	0.19	7.63	4.64	9.34
	15.72	12.80	5.27E-05	-8.25E+00	6.03E-05	0.18	7.60	4.64	9.28
	16.02	12.98	5.69E-05	-8.01E+00	9.33E-05	0.20	7.27	4.82	9.83
1.00E+08	11.75	7.93	5.86E-04	-1.43E+01	-1.98E-03	0.20	1.39	5.61	9.93
	11.88	8.05	6.05E-04	-1.42E+01	-1.98E-03	0.21	1.67	5.64	9.61
	12.14	8.27	5.52E-04	-1.39E+01	-1.76E-03	0.20	1.95	5.74	9.33
	12.20	8.31	5.31E-04	-1.38E+01	-1.74E-03	0.19	1.94	5.78	9.36
	12.19	8.37	5.31E-04	-1.38E+01	-1.57E-03	0.19	1.82	5.73	9.68

Table B.S3. Extended numerical values of features and sigmoidal fittings of samples

Features and Sigmoidal Fit for Samples 1 - 9									
Sample ID	C _t	C _y	F ₀	M ₀	F _b	F _{max}	c	b	d
Sample_1	24.12	22.07	2.10E-07	4.77E+00	5.19E-04	0.16	23.69	3.02	2.39
	24.02	21.93	1.94E-07	4.61E+00	4.51E-04	0.15	23.10	3.08	2.67
	24.07	22.03	1.96E-07	4.72E+00	4.05E-04	0.15	23.82	2.99	2.29
Sample_2	26.40	24.43	5.34E-08	7.88E+00	2.94E-04	0.16	26.82	2.82	1.94
	26.27	24.31	5.22E-08	7.73E+00	1.42E-04	0.16	26.71	2.79	1.93
	26.19	24.14	5.21E-08	7.54E+00	1.74E-04	0.16	25.31	2.99	2.66
Sample_3	19.00	17.24	1.72E-07	-1.51E+00	1.75E-04	0.21	19.12	2.31	1.99
	18.82	17.09	1.92E-07	-1.72E+00	2.32E-04	0.20	19.06	2.28	1.92
	18.70	16.97	2.23E-07	-1.89E+00	2.69E-04	0.20	18.74	2.34	2.06
Sample_4	20.02	17.68	8.48E-07	-8.65E-01	1.27E-04	0.17	14.47	3.53	7.67
	19.96	17.61	8.18E-07	-9.45E-01	-8.96E-05	0.18	14.04	3.52	8.46
	19.83	17.47	1.12E-06	-1.14E+00	5.60E-05	0.18	13.73	3.61	8.61
Sample_5	26.19	24.19	7.20E-08	7.56E+00	1.34E-04	0.16	26.84	2.80	1.82
	25.94	23.89	7.80E-08	7.19E+00	2.15E-04	0.15	26.05	2.90	2.08
	25.85	23.80	8.61E-08	7.06E+00	2.39E-04	0.16	25.64	2.95	2.26
Sample_6	19.82	17.47	1.25E-06	-1.15E+00	1.57E-04	0.18	13.79	3.63	8.43
	19.33	16.98	1.11E-06	-1.77E+00	9.56E-05	0.16	13.42	3.62	8.20
	19.06	16.65	1.57E-06	-2.20E+00	1.16E-04	0.16	12.98	3.68	8.30
Sample_7	16.31	14.52	1.48E-06	-5.16E+00	6.53E-04	0.23	15.76	2.49	2.46
	16.29	14.51	1.34E-06	-5.17E+00	5.42E-04	0.23	15.70	2.46	2.49
	16.29	14.49	1.27E-06	-5.18E+00	4.65E-04	0.21	15.55	2.49	2.59
Sample_8	21.10	18.71	4.07E-06	3.44E-01	4.55E-04	0.19	19.38	3.67	3.10
	20.82	18.45	4.59E-06	-3.14E-03	5.04E-04	0.21	18.79	3.67	3.33
	21.09	18.68	4.71E-06	3.03E-01	4.91E-04	0.21	19.38	3.70	3.09
Sample_9	20.05	17.63	3.30E-06	-1.01E+00	1.78E-03	0.18	15.38	3.92	5.61
	20.03	17.62	2.42E-06	-9.99E-01	1.23E-03	0.18	14.80	3.84	6.48
	19.96	17.63	1.81E-06	-9.95E-01	1.06E-03	0.15	15.98	3.68	5.06
	20.08	17.66	2.74E-06	-9.61E-01	1.39E-03	0.21	14.11	3.92	7.52
	20.10	17.67	2.70E-06	-9.38E-01	1.39E-03	0.19	14.29	3.95	7.19
	20.06	17.65	2.66E-06	-9.73E-01	1.49E-03	0.18	14.91	3.89	6.28
	19.99	17.60	2.74E-06	-1.05E+00	1.59E-03	0.19	14.72	3.86	6.51
	20.38	17.99	1.96E-06	-5.17E-01	1.18E-03	0.21	14.69	3.88	7.19

Table B.S4. Estimated quantification for samples

Estimated copy number per reaction for samples 1 - 9 using extracted features and M0				
Sample ID	C _t	C _y	-log10(F0)	M ₀
Sample_1	9.05E+04	7.60E+04	1.09E+04	4.35E+04
	9.57E+04	8.16E+04	9.92E+03	4.64E+04
	9.27E+04	7.75E+04	1.00E+04	4.43E+04
Sample_2	2.48E+04	2.21E+04	2.08E+03	1.19E+04
	2.66E+04	2.35E+04	2.02E+03	1.27E+04
	2.79E+04	2.57E+04	2.02E+03	1.38E+04
Sample_3	1.65E+06	9.49E+05	8.59E+03	5.92E+05
	1.84E+06	1.02E+06	9.84E+03	6.46E+05
	1.96E+06	1.09E+06	1.17E+04	6.94E+05
Sample_4	9.28E+05	7.55E+05	5.94E+04	4.53E+05
	9.61E+05	7.81E+05	5.68E+04	4.69E+05
	1.03E+06	8.40E+05	8.30E+04	5.09E+05
Sample_5	2.78E+04	2.50E+04	2.99E+03	1.36E+04
	3.22E+04	2.92E+04	3.30E+03	1.59E+04
	3.39E+04	3.07E+04	3.71E+03	1.68E+04
Sample_6	1.04E+06	8.39E+05	9.47E+04	5.11E+05
	1.37E+06	1.09E+06	8.26E+04	6.62E+05
	1.60E+06	1.29E+06	1.25E+05	7.91E+05
Sample_7	7.61E+06	3.93E+06	1.16E+05	2.70E+06
	7.72E+06	3.96E+06	1.03E+05	2.72E+06
	7.71E+06	3.99E+06	9.66E+04	2.73E+06
Sample_8	5.03E+05	4.39E+05	3.96E+05	2.74E+05
	5.89E+05	5.03E+05	4.59E+05	3.17E+05
	5.05E+05	4.46E+05	4.74E+05	2.79E+05
Sample_9	9.14E+05	7.73E+05	3.08E+05	4.81E+05
	9.24E+05	7.76E+05	2.11E+05	4.79E+05
	9.59E+05	7.74E+05	1.49E+05	4.78E+05
	8.97E+05	7.63E+05	2.46E+05	4.72E+05
	8.85E+05	7.58E+05	2.42E+05	4.67E+05
	9.06E+05	7.67E+05	2.37E+05	4.74E+05
	9.43E+05	7.87E+05	2.46E+05	4.89E+05
	7.56E+05	6.41E+05	1.64E+05	3.92E+05

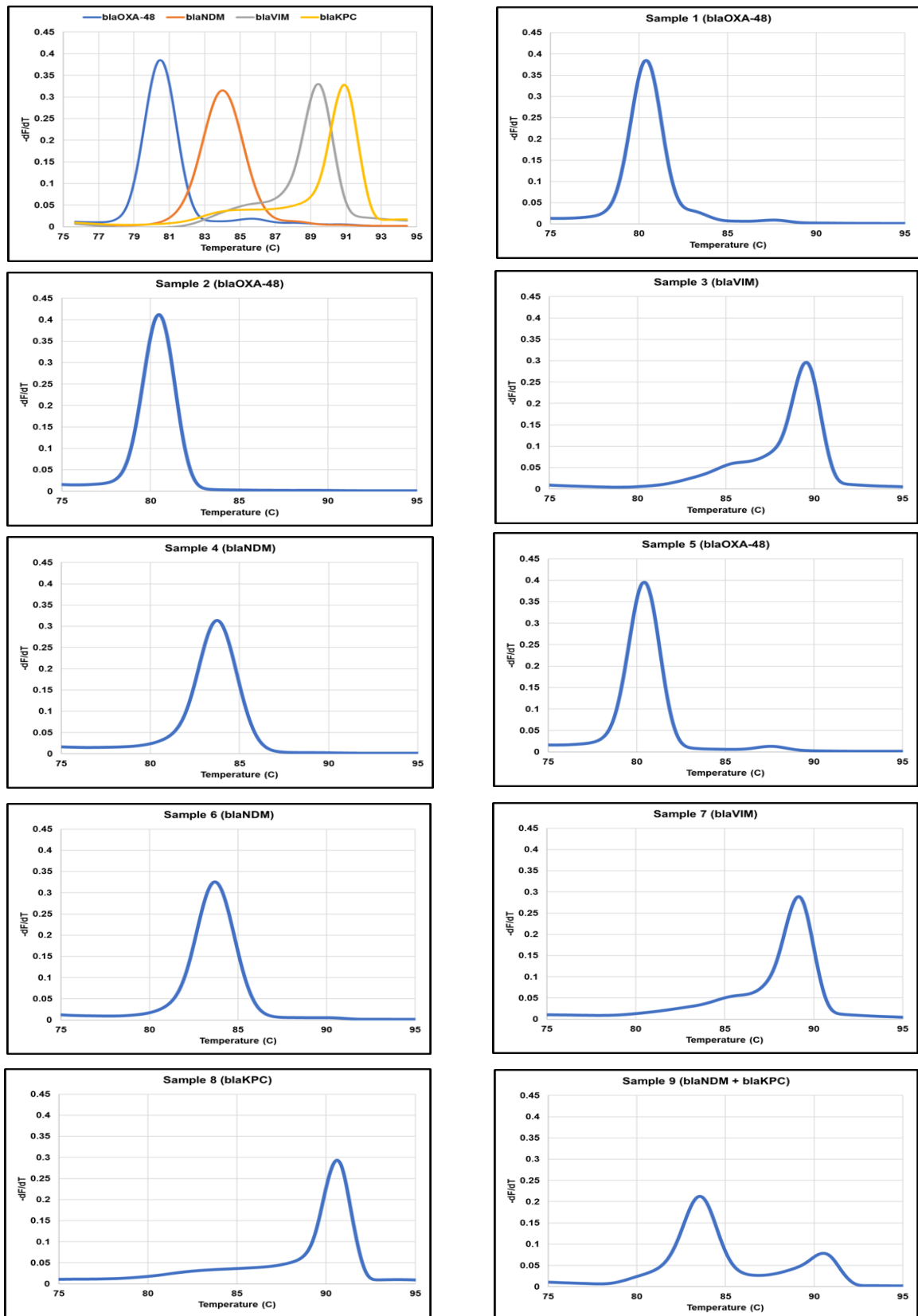
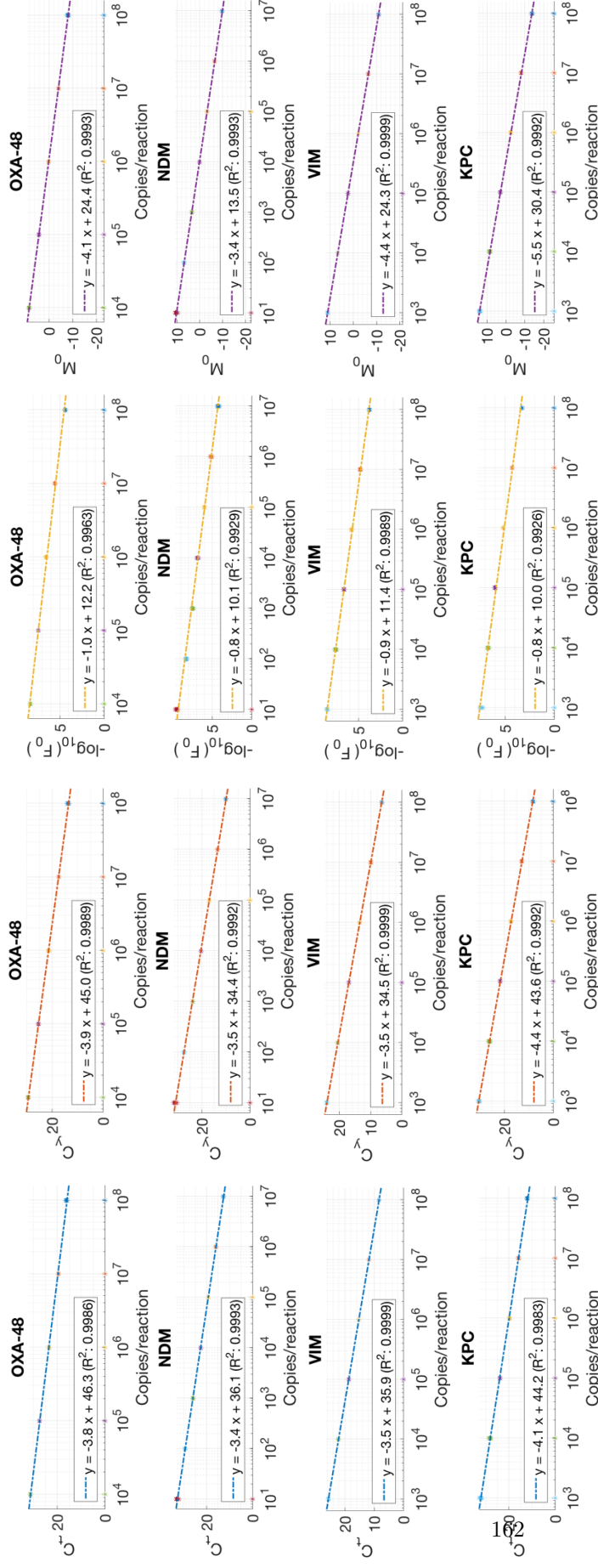


Figure B.S1. Melting curves analysis for standards and samples. Average melting curve peaks for standards (gblocks) and analysed samples (genomic DNA). Experiments were performed using multiplexed primer mix (Monteiro et al 2012).



DNA	Alphas for M0				P-values between M0 and Ct, Cy and -log10(F0)				Figure of Merit (Mean)				Figure of Merit (Stdev)			
	Ct	Cy	-log10(F0)	M0	Ct	Cy	-log10(F0)	M0	Ct	Cy	-log10(F0)	M0	Ct	Cy	-log10(F0)	M0
OXA-48	0.38	0.89	1.54		4.61E-02	6.63E-03	1.47E-04	1.64E+03	1.31E+03	4.75E+03	4.25E+02	9.59E+02	4.25E+03	3.16E+03	8.30E+03	2.35E+03
NDM	1.00	0.00	0.00		1.00E+00	7.14E-01	2.23E-09	3.51E+03	4.11E+03	4.46E+04	3.51E+03	3.51E+03	1.34E+04	1.52E+04	7.21E+04	1.34E+04
VIM	1.02	0.67	1.24		4.22E-02	3.01E-02	3.10E-09	6.14E+01	5.49E+01	1.99E+03	1.88E+03	3.23E+01	1.18E+02	9.40E+01	2.48E+03	6.62E+01
KPC	0.51	1.14	1.30		1.43E-03	1.35E-03	1.53E-04	3.85E+03	1.88E+03	4.07E+04	1.84E+03	1.64E+03	6.89E+03	4.05E+03	1.50E+05	3.49E+03

Figure B.S2.

(Top) Standard curves using C_t , C_y , F_0 or M_0 for each synthetic DNA containing bla_{OXA48} , bla_{NDM} , bla_{VIM} and bla_{KPC} genes .

(Bottom) Evaluation of standard curves. Alphas indicate the weights for each feature to generate M_0 for a given MSC; P-values between the Figure of Merit for M_0 and the rest of the methods; Mean and Standard Deviation (Stddev) for the Figure of Merit.

Appendix C

Supporting Information:

Chapter 5

The code and data required to reproduce the results in Chapter 5 are publicly available at www.github.com/am5113/pyACA. In particular, this link contains 3 Jupyter notebooks and 1 Python file, as follows:

- `01_qPCR_Data_Analysis.ipynb` - amplification curve analysis on qPCR data
- `02_dPCR_Data_Analysis.ipynb` - amplification curve analysis on dPCR data
- `03_Coamplification_Statistics.ipynb` - studying the effect of coamplification in dPCR in the presence of multiple targets
- `utilities.py` - functions used for prior notebooks

Appendix D

Supporting Information: Chapter 6

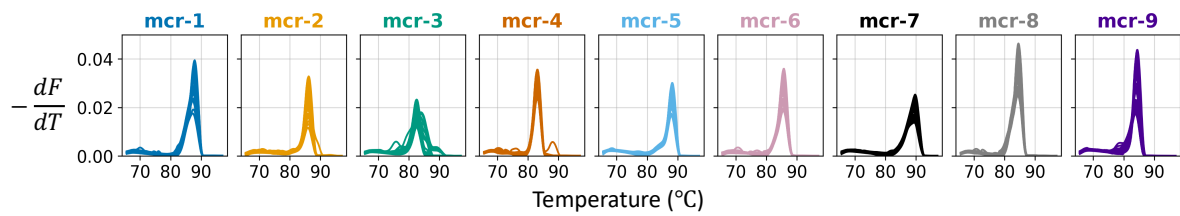
This Appendix contains the following:

- Raw melting curves from qPCR and dPCR (Figure D.S1)
- Standard curves for each mcr target using new 9-plex assay (Figure D.S2)
- Performance of high-level dPCR multiplexing with FFI (Figure D.S3)
- Performance of all methods in conventional qPCR (Figure D.S4)

The code and data required to reproduce the results in Chapter 6 are publicly available at www.github.com/am5113/pyAMCA. In particular, this link contains 2 Jupyter notebooks and 1 Python file, as follows:

- `Data_Analysis_qPCR.ipynb` - AMCA analysis on qPCR data
- `Data_Analysis_dPCR.ipynb` - AMCA analysis on dPCR data
- `utilities.py` - functions used for prior notebooks

A) Melting Curves from qPCR Instrument



B) Melting Curves from dPCR Instrument

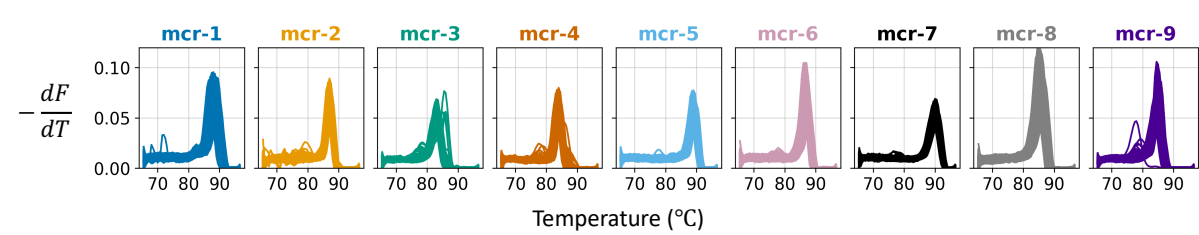


Figure D.S1. Raw melting curves from (A) qPCR and (B) dPCR instrument.

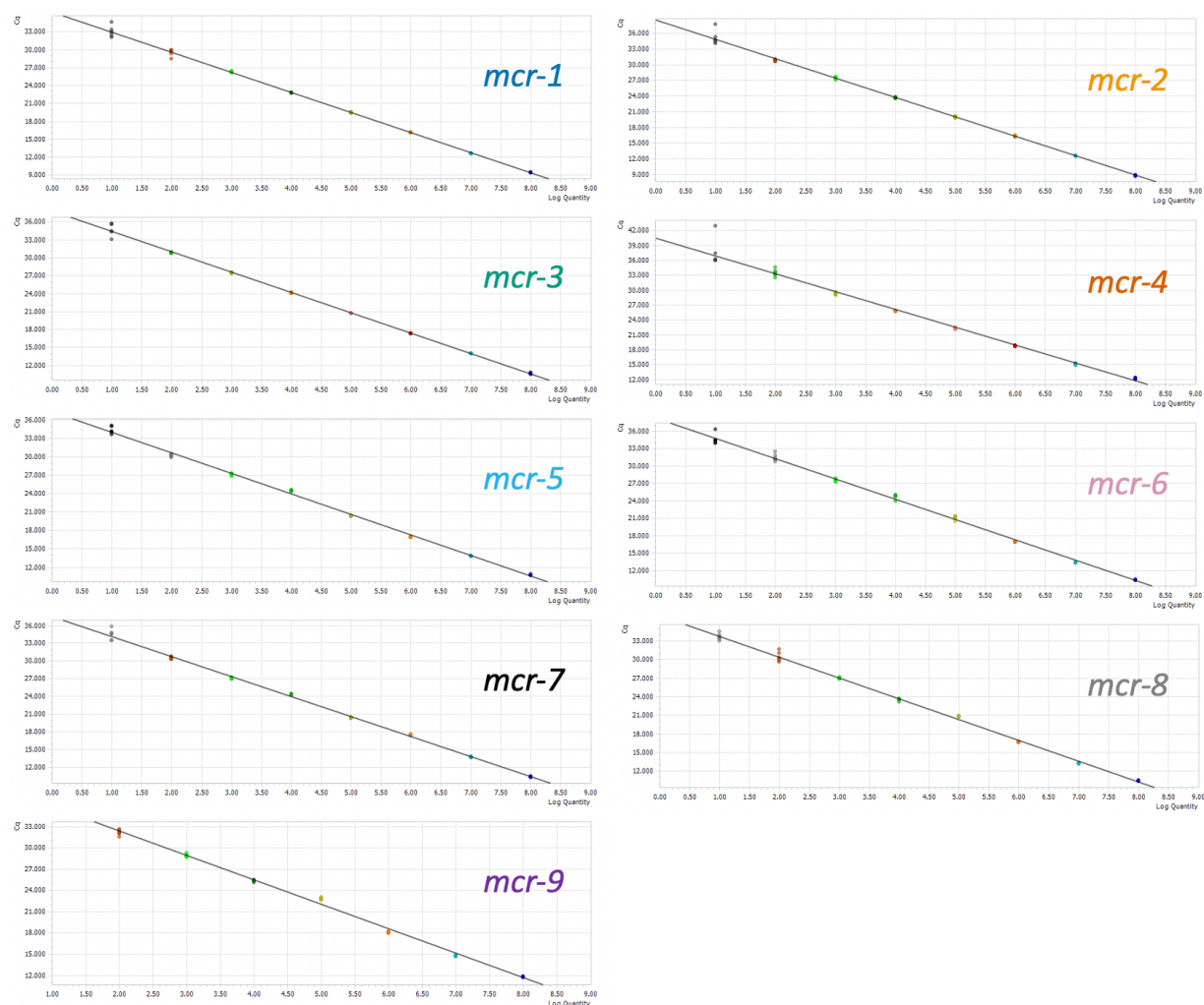


Figure D.S2. Standard curves for each *mcr* target using new 9-plex assay. (Top Panel) Plots were generated using Roche LightCycler software (version 1.1). (Bottom Panel) Table with relevant meta data for each standard curve.

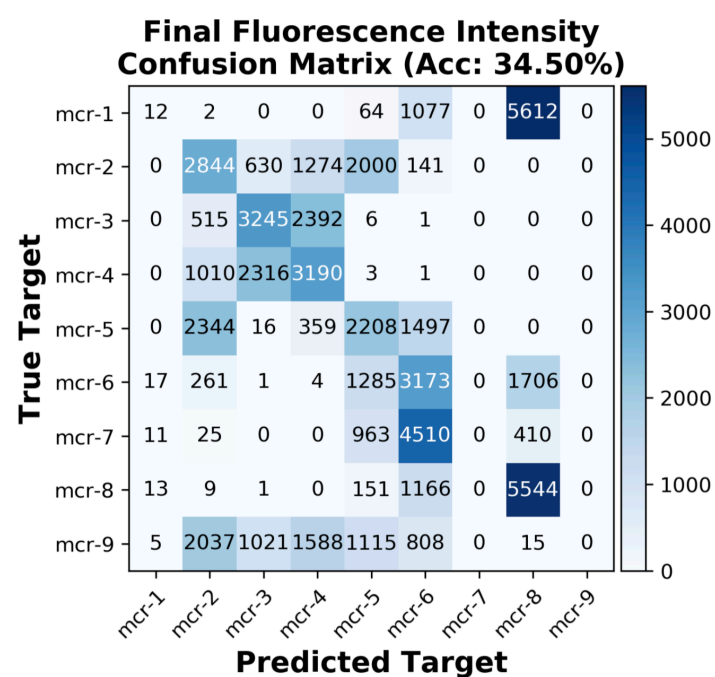


Figure D.S3. Multiplexing with FFI in dPCR (without optimization of primer concentration). Confusion matrix showing performance of logistic regression classifier on FFI values.

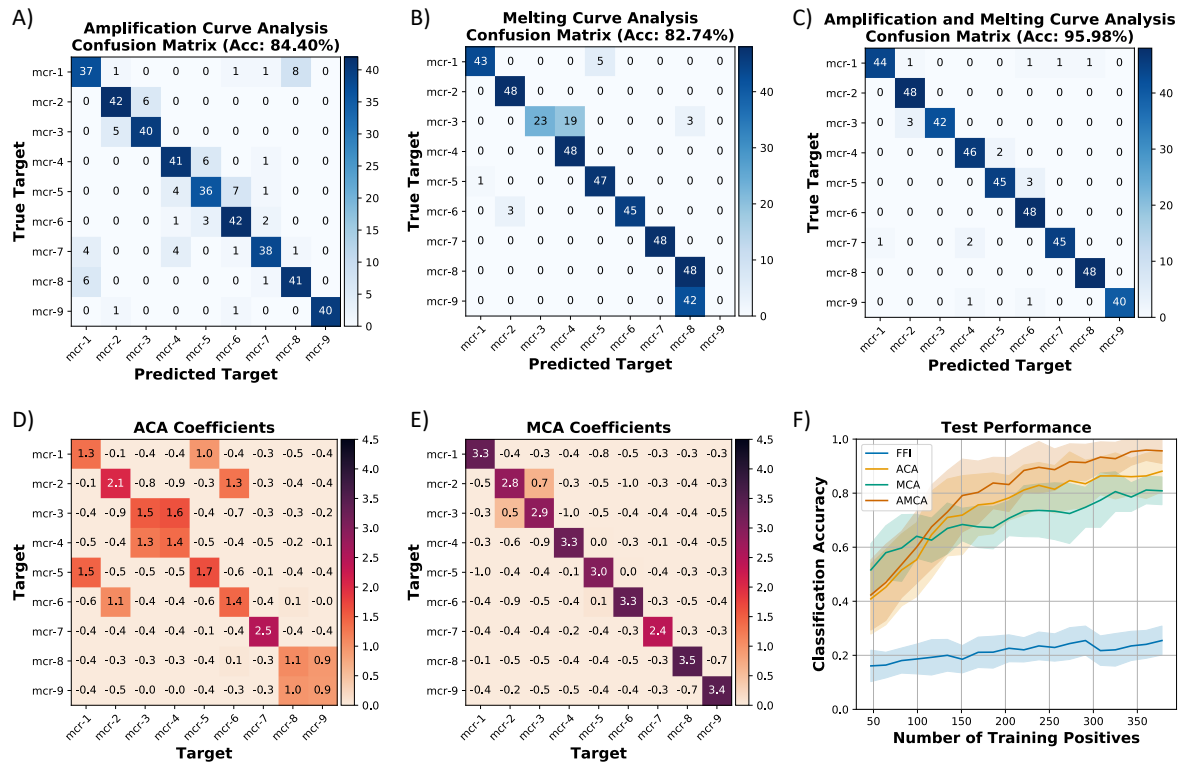


Figure D.S4. Performance of all methods for multiplexing the 9 *mcr* targets in conventional qPCR instrument. A, B, C) Confusion matrices illustrating the predictions from ACA, MCA and AMCA (proposed method), respectively. Values indicate the number of amplification events with diagonal entries corresponding 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 E

Supporting Information:

Chapter 7

This Appendix contains the following:

- Matlab code to implement the proposed algorithm (Listing E.1)

Listing E.1: Implementation of Proposed Algorithm

```

1  %% Clear workspace
2  clear all; close all;
3
4  % This function retrieves the raw data (in mV), time vector (in
   minutes), and dimensions of the array (MxN)
5  [raw_data, time, M, N] = getData();
6
7  % Plot Raw Data
8  f1 = figure;
9  plot(time, raw_data');
10 title('Raw Data', 'fontsize', 16);
11 xlabel('Time (min)', 'fontsize', 14);
12 ylabel('Voltage (mV)', 'fontsize', 14);
13
14 %% Pre-processing
15
16 % Remove baseline
17 raw_data_bs = raw_data - mean(raw_data(:, 1:5), 2);
18
19 % Plot preprocessed data
20 f2 = figure; hold on;
21 plot(time, raw_data_bs');
22 h = plot(time, mean(raw_data_bs), 'k', 'linewidth', 2);
23 legend(h, {'Mean'});
24 title('Preprocessed Data', 'fontsize', 16);
25 xlabel('Time (mins)', 'fontsize', 14);
26 ylabel('Voltage (mV)', 'fontsize', 14);
27
28 %% Adaptive Signal Processing
29 smoothing = 30;
30
31 output_weights = zeros(size(raw_data_bs));
32
33 lms = dsp.LMSFilter('Method', 'Normalized LMS',...
34                    'Length', 1,...

```



```

35         'StepSize', 0.1,...
36         'InitialConditions', 0,...
37         'WeightsOutput', 'All');
38
39 % For every sensor, apply the NLMS algorithm
40 for i = 1:size(raw_data_bs, 1)
41     [~, ~, wts] = lms(time', raw_data_bs(i, :));
42     output_weights(i, :) = smooth(diff(wts), smoothing);
43 end
44
45 % Plot the weight vectors
46 % (i.e. abstract representation of behaviour over time)
47 figure;
48 plot(time, output_weights);
49 title('Weights', 'fontsize', 16);
50 xlabel('Time (mins)', 'fontsize', 14);
51 ylabel('Magnitude', 'fontsize', 14);
52
53 %% Unsupervised Learning
54 % For illustration purposes, 2 clusters are manually chosen here.
55 % Although this should be done automatically (e.g. via silhouette
    score)
56 N_clusters = 2;
57
58 idx = kmeans(output_weights, N_clusters, 'Display', 'iter',...
59     'MaxIter', 10000,...
60     'Replicates', 3);
61
62 classes = unique(idx);
63
64 %% Spatial Validation
65 colors = [0      0.4470      0.7410;
66     0.8500      0.3250      0.0980;
67     0.9290      0.6940      0.1250;
68     0.4940      0.1840      0.5560;
69     0.4660      0.6740      0.1880;

```

```

70     0.3010     0.7450     0.9330;
71     0.6350     0.0780     0.1840];
72
73 X_2D = reshape(idx, [M, N]);
74 X_2D_classes = X_2D(:, :, 1);
75
76 % Plot spatial representation of the clusters
77 figure; hold on;
78 for i = 1:M
79     for j = 1:N
80         plot(i, j, '*', 'color', colors(X_2D_classes(i,j),:));
81     end
82 end
83 xlim([0, M+1]); ylim([0, N+1]);
84
85 title('ISFET Array', 'fontsize', 16);
86 xlabel('x-coordinate', 'fontsize', 14);
87 ylabel('y-coordinate', 'fontsize', 14);
88
89 % Plot time-domain signals in each cluster
90 figure; hold on
91 for i = 1:length(classes)
92     plot(time, output_weights(idx==i, :), 'color', [colors(i, :),
93         0.05]);
94 end
95 title('Weights', 'fontsize', 16);
96 xlabel('Time (mins)', 'fontsize', 14);
97 ylabel('Magnitude', 'fontsize', 14);
98
99 %% Averaging and Post-processing
100 for i = 1:N_clusters
101     figure; hold on;
102     y = mean(output_weights(idx==i, :), 2);
103
104     % Algorithm to find location of peak
105     peak = findpeak(time, y);

```

```
105     plot(time, y, 'color', colors(i,:), 'linewidth', 2)
106     % Algorithm to identify low-frequency background
107     background = subtract_background(time, y);
108     plot(time, background, 'k--', 'linewidth', 2);
109
110     title('Averaged Weights', 'fontsize', 16);
111     xlabel('Time (mins)', 'fontsize', 14);
112     ylabel('Magnitude', 'fontsize', 14);
113     legend({'Signal', 'Estimated Drift'},...
114           'location', 'southeast')
115
116     figure;
117     plot(time, y - background, 'k', 'linewidth', 2)
118     title('Averaged Weights', 'fontsize', 16);
119     xlabel('Time (mins)', 'fontsize', 14);
120     ylabel('Magnitude', 'fontsize', 14);
121 end
```


References

- [1] R. F. Colclough G, Dorling G and S. A. Ghafur S, “Harnessing Data Science and AI in Healthcare: From policy to practice.,” World Innovation Summit for Health, 2018.
- [2] M. C. Schatz, “Biological data sciences in genome research,” Genome Research, vol. 25, pp. 1417–1422, oct 2015.
- [3] J. A. Reuter, D. V. Spacek, and M. P. Snyder, “High-Throughput Sequencing Technologies,” Molecular Cell, vol. 58, pp. 586–597, may 2015.
- [4] C. Toumazou, L. M. Shepherd, S. C. Reed, G. I. Chen, A. Patel, D. M. Garner, C. J. A. Wang, C. P. Ou, K. Amin-Desai, P. Athanasiou, H. Bai, I. M. Brizido, B. Caldwell, D. Coomber-Alford, P. Georgiou, K. S. Jordan, J. C. Joyce, M. La Mura, D. Morley, S. Sathyavruthan, S. Temelso, R. E. Thomas, and L. Zhang, “Simultaneous DNA amplification and detection using a pH-sensing semiconductor system,” Nature Methods, vol. 10, pp. 641–646, jul 2013.
- [5] J. Vangindertael, R. Camacho, W. Sempels, H. Mizuno, P. Dedecker, and K. P. Janssen, “An introduction to optical super-resolution microscopy for the adventurous biologist,” Methods and Applications in Fluorescence, vol. 6, mar 2018.
- [6] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, “Microarray data analysis: From disarray to consolidation and consensus,” Nature Reviews Genetics, vol. 7, no. 1, pp. 55–65, 2006.
- [7] D. Somnath and D. Nettleton, Statistical Analysis of Next Generation Sequencing Data. Springer International Publishing, 2014.
- [8] M. J. Heller, “Dna microarray technology: devices, systems, and applications,” Annual review of biomedical engineering, vol. 4, no. 1, pp. 129–153, 2002.
- [9] J. Shendure and H. Ji, “Next-generation dna sequencing,” Nature biotechnology, vol. 26, no. 10, pp. 1135–1145, 2008.
- [10] K. Mullis, F. Faloona, S. Scharf, R. K. Saiki, G. T. Horn, and H. Erlich, “Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction,” in Cold Spring Harbor symposia on quantitative biology, vol. 51, pp. 263–273, Cold Spring Harbor Laboratory Press, 1986.

- [11] D. A. Rappolee, D. Mark, M. J. Banda, and Z. Werb, “Wound Macrophages Express TGF- α and Other Growth Factors in Vivo: Analysis by mRNA Phenotyping,” Science, vol. 241, no. 4866, pp. 708–712, 1988.
- [12] D. A. Rappolee, C. A. Brenner, R. Schultz, D. Mark, and Z. Werb, “Developmental expression of PDGF, TGF- α , and TGF- β genes in preimplantation mouse embryos,” Science, vol. 241, no. 4874, pp. 1823–1825, 1988.
- [13] Enzoklop, “Schematic drawing of the PCR cycle.” https://commons.wikimedia.org/wiki/File:Polymerase_chain_reaction.svg, accessed 2016-01-27.
- [14] E. Southern, “Southern blotting,” Nature protocols, vol. 1, no. 2, p. 518, 2006.
- [15] R. Higuchi, C. Fockler, G. Dollinger, and R. Watson, “Kinetic pcr analysis: real-time monitoring of dna amplification reactions,” Nature Biotechnology, vol. 11, no. 9, p. 1026, 1993.
- [16] C. Wittwer, K. Ririe, and R. Rasmussen, Fluorescence monitoring of rapid cycle PCR for quantification. Springer, 1998.
- [17] W. M. Freeman, S. J. Walker, and K. E. Vrana, “Quantitative rt-pcr: pitfalls and potential,” Biotechniques, vol. 26, no. 1, pp. 112–125, 1999.
- [18] K. M. Ririe, R. P. Rasmussen, and C. T. Wittwer, “Product differentiation by analysis of dna melting curves during the polymerase chain reaction,” Analytical biochemistry, vol. 245, no. 2, pp. 154–160, 1997.
- [19] ThermoFisher Scientific, “How taqman assays work.” <https://www.thermofisher.com/uk/en/home/life-science/pcr/real-time-pcr/real-time-pcr-learning-center/real-time-pcr-basics/how-taqman-assays-work.html>, 2020.
- [20] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim, “Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia,” Science, vol. 230, no. 4732, pp. 1350–1354, 1985.
- [21] R. G. Rutledge, “Mathematics of quantitative kinetic PCR and the application of standard curves,” Nucleic Acids Research, vol. 31, pp. 93e–93, aug 2003.
- [22] C. S. Booth, E. Pienaar, J. R. Termaat, S. E. Whitney, T. M. Louw, and H. J. Viljoen, “Efficiency of the polymerase chain reaction,” Chemical Engineering Science, vol. 65, pp. 4996–5006, sep 2010.

- [23] P. Keohavong and W. G. Thilly, “Fidelity of DNA polymerases in DNA amplification,” Proceedings of the National Academy of Sciences of the United States of America, vol. 86, no. 23, pp. 9253–9257, 1989.
- [24] H. Li, U. B. Gyllensten, X. Cui, R. K. Saiki, H. A. Erlich, and N. Arnheim, “Amplification and analysis of DNA sequences in single human sperm and diploid cells,” Nature, vol. 335, no. 6189, pp. 414–417, 1988.
- [25] M. Guescini, D. Sisti, M. B. Rocchi, L. Stocchi, and V. Stocchi, “A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition,” BMC Bioinformatics, vol. 9, jul 2008.
- [26] M. Guescini, D. Sisti, M. B. Rocchi, R. Panebianco, P. Tibollo, and V. Stocchi, “Accurate and Precise DNA Quantification in the Presence of Different Amplification Efficiencies Using an Improved Cy0 Method,” PLoS ONE, vol. 8, jul 2013.
- [27] A. A. Morley, “Digital PCR: A brief history,” Biomolecular Detection and Quantification, vol. 1, no. 1, pp. 1–2, 2014.
- [28] P. L. Quan, M. Sauzade, and E. Brouzes, “DPCR: A technology review,” Sensors (Switzerland), vol. 18, apr 2018.
- [29] B. Vogelstein and K. W. Kinzler, “Digital pcr,” Proceedings of the National Academy of Sciences, vol. 96, no. 16, pp. 9236–9241, 1999.
- [30] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich, “Primer-directed enzymatic amplification of dna with a thermostable dna polymerase,” Science, vol. 239, no. 4839, pp. 487–491, 1988.
- [31] P. Sykes, S. Neoh, M. Brisco, E. Hughes, J. Condon, and A. Morley, “Quantitation of targets for pcr by use of limiting dilution,” Biotechniques, vol. 13, no. 3, pp. 444–449, 1992.
- [32] A. S. Basu, “Digital Assays Part I: Partitioning Statistics and Digital PCR,” SLAS TECHNOLOGY: Translating Life Sciences Innovation, vol. 22, pp. 369–386, aug 2017.
- [33] J. E. Kreutz, T. Munson, T. Huynh, F. Shen, W. Du, and R. F. Ismagilov, “Theoretical design and analysis of multivolume digital assays with wide dynamic range validated experimentally with microfluidic digital PCR,” Analytical Chemistry, vol. 83, pp. 8158–8168, nov 2011.

- [34] F. Shen, B. Sun, J. E. Kreutz, E. K. Davydova, W. Du, P. L. Reddy, L. J. Joseph, and R. F. Ismagilov, "Multiplexed quantification of nucleic acids with large dynamic range using multivolume digital RT-PCR on a rotational SlipChip tested with HIV and hepatitis C viral load," Journal of the American Chemical Society, vol. 133, pp. 17705–17712, nov 2011.
- [35] T. Goda, M. Tabata, and Y. Miyahara, "Electrical and electrochemical monitoring of nucleic acid amplification," Frontiers in bioengineering and biotechnology, vol. 3, p. 29, 2015.
- [36] L. Yan, J. Zhou, Y. Zheng, A. S. Gamson, B. T. Roembke, S. Nakayama, and H. O. Sintim, "Isothermal amplified detection of dna and rna," Molecular BioSystems, vol. 10, no. 5, pp. 970–1003, 2014.
- [37] L. M. Zanolli and G. Spoto, "Isothermal amplification methods for the detection of nucleic acids in microfluidic devices," Biosensors, vol. 3, no. 1, pp. 18–43, 2013.
- [38] T. Notomi, "Loop-mediated isothermal amplification of DNA," Nucleic Acids Research, vol. 28, pp. 63e–63, jun 2000.
- [39] N. Moser, "Integrated Auto-Calibration and In-Pixel Quantisation Methodologies for CMOS ISFET Arrays in Point-of-Care Diagnostics." PhD thesis. Imperial College London, Dept of Electrical and Electronic Engineering, 2018.
- [40] N. Kaur, N. Thota, and B. J. Toley, "A stoichiometric and pseudo kinetic model of loop mediated isothermal amplification," Computational and structural biotechnology journal, vol. 18, pp. 2336–2346, 2020.
- [41] M. Safavieh, M. K. Kanakasabapathy, F. Tarlan, M. U. Ahmed, M. Zourob, W. Asghar, and H. Shafiee, "Emerging loop-mediated isothermal amplification-based microchip and microdevice technologies for nucleic acid detection," ACS biomaterials science & engineering, vol. 2, no. 3, pp. 278–294, 2016.
- [42] P. Bergveld, "Short Communications: Development of an Ion-Sensitive Solid-State Device for Neurophysiological Measurements," IEEE Transactions on Biomedical Engineering, vol. BME-17, no. 1, pp. 70–71, 1970.
- [43] P. Bergveld, "Thirty years of ISFETOLOGY: What happened in the past 30 years and what may happen in the next 30 years," Sensors and Actuators, B: Chemical, vol. 88, pp. 1–20, jan 2003.

- [44] N. Moser, T. S. Lande, C. Toumazou, and P. Georgiou, “Isfets in cmos and emergent trends in instrumentation: A review,” IEEE Sensors Journal, vol. 16, no. 17, pp. 6496–6514, 2016.
- [45] N. Moser, J. Rodriguez-Manzano, T. S. Lande, and P. 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, vol. 12, pp. 390–401, apr 2018.
- [46] K. Malpartida-Cardenas, N. Miscourides, J. Rodriguez-Manzano, L.-S. Yu, N. Moser, J. Baum, and P. Georgiou, “Quantitative and rapid plasmodium falciparum malaria diagnosis and artemisinin-resistance detection using a cmos lab-on-chip platform,” Biosensors and Bioelectronics, vol. 145, p. 111678, 2019.
- [47] J. Rodriguez-Manzano, A. Moniri, K. Malpartida-Cardenas, J. Dronavalli, F. Davies, A. Holmes, and P. Georgiou, “Simultaneous single-channel multiplexing and quantification of carbapenem-resistant genes using multidimensional standard curves,” Analytical chemistry, vol. 91, no. 3, pp. 2013–2020, 2019.
- [48] C. A. Heid, J. Stevens, K. J. Livak, and P. M. Williams, “Real time quantitative pcr.,” Genome research, vol. 6, no. 10, pp. 986–994, 1996.
- [49] T. R. Gingeras, R. Higuchi, L. J. Kricka, Y. D. Lo, and C. T. Wittwer, “Fifty years of molecular (dna/rna) diagnostics,” Clinical chemistry, vol. 51, no. 3, pp. 661–671, 2005.
- [50] I. M. Mackay, K. E. Arden, and A. Nitsche, “Real-time pcr in virology,” Nucleic acids research, vol. 30, no. 6, pp. 1292–1305, 2002.
- [51] S. A. Bustin, “Absolute quantification of mrna using real-time reverse transcription polymerase chain reaction assays,” Journal of molecular endocrinology, vol. 25, no. 2, pp. 169–193, 2000.
- [52] T. Nolan, R. E. Hands, and S. A. Bustin, “Quantification of mrna using real-time rt-pcr,” Nature protocols, vol. 1, no. 3, p. 1559, 2006.
- [53] R. Girones, M. A. Ferrús, J. L. Alonso, J. Rodriguez-Manzano, B. Calgua, A. de Abreu Corrêa, A. Hundesa, A. Carratala, and S. Bofill-Mas, “Molecular detection of pathogens in water—the pros and cons of molecular techniques,” Water research, vol. 44, no. 15, pp. 4325–4339, 2010.
- [54] A. M. Caliendo, T. C. Quinn, F. C. Tenover, D. Alland, A. J. Blaschke, R. A. Bonomo, K. C. Carroll, D. N. Gilbert, M. J. Ferraro, L. R. Hirschhorn, W. P. Joseph, T. Karchmer, A. T. MacIntyre, L. B. Reller, A. F. Jackson, C. C. Ginocchio, K. E. Hanson, and L. May,

- “Better tests, better care: Improved diagnostics for infectious diseases,” Clinical Infectious Diseases, vol. 57, pp. S139–S170, 12 2013.
- [55] A. C. Ghani, D. H. Burgess, A. Reynolds, and C. Rousseau, “Expanding the role of diagnostic and prognostic tools for infectious diseases in resource-poor settings,” Nature, vol. 528, no. 7580, pp. S50–S52, 2015.
- [56] M. Misyura, M. A. Sukhai, V. Kulasingam, T. Zhang, S. Kamel-Reid, and T. L. Stockley, “Improving validation methods for molecular diagnostics: application of bland-altman, deming and simple linear regression analyses in assay comparison and evaluation for next-generation sequencing,” Journal of clinical pathology, vol. 71, no. 2, pp. 117–124, 2018.
- [57] C. T. Wittwer, M. G. Herrmann, A. A. Moss, and R. P. Rasmussen, “Continuous fluorescence monitoring of rapid cycle dna amplification,” Biotechniques, vol. 22, no. 1, pp. 130–139, 1997.
- [58] L. Raeymaekers, “A commentary on the practical applications of competitive pcr.,” Genome Research, vol. 5, no. 1, pp. 91–94, 1995.
- [59] M. Guescini, D. Sisti, M. B. Rocchi, L. Stocchi, and V. Stocchi, “A new real-time pcr method to overcome significant quantitative inaccuracy due to slight amplification inhibition,” BMC bioinformatics, vol. 9, no. 1, p. 326, 2008.
- [60] R. Rutledge, “Sigmoidal curve-fitting redefines quantitative real-time pcr with the prospective of developing automated high-throughput applications,” Nucleic acids research, vol. 32, no. 22, pp. e178–e178, 2004.
- [61] L. Van Der Maaten, E. Postma, and J. Van den Herik, “Dimensionality reduction: a comparative,” J Mach Learn Res, vol. 10, pp. 66–71, 2009.
- [62] X. Rao, D. Lai, and X. Huang, “A new method for quantitative real-time polymerase chain reaction data analysis,” Journal of Computational Biology, vol. 20, no. 9, pp. 703–711, 2013.
- [63] T. F. Coleman and Y. Li, “An interior trust region approach for nonlinear minimization subject to bounds,” SIAM Journal on optimization, vol. 6, no. 2, pp. 418–445, 1996.
- [64] Y. Coleman T Li, “On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds,” Mathematical Programming, vol. 67, p. 189224, 1994.
- [65] K. Levenberg, “A method for the solution of certain non-linear problems in least squares,” Quarterly of applied mathematics, vol. 2, no. 2, pp. 164–168, 1944.

- [66] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," Journal of the society for Industrial and Applied Mathematics, vol. 11, no. 2, pp. 431–441, 1963.
- [67] J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning, vol. 1. Springer series in statistics New York, NY, USA, 2001.
- [68] D. P. Mandic, S. Kanna, Y. Xia, A. Moniri, A. Junyent-Ferre, and A. G. Constantinides, "A data analytics perspective of power grid analysis-part 1: The clarke and related transforms [lecture notes]," IEEE Signal Processing Magazine, vol. 36, no. 2, pp. 110–116, 2019.
- [69] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 24, no. 6, pp. 381–395, 1981.
- [70] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," Chemometrics and intelligent laboratory systems, vol. 50, no. 1, pp. 1–18, 2000.
- [71] D. Coomans, I. Broeckaert, M. Derde, A. Tassin, D. Massart, and S. Wold, "Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles," Computers and biomedical research, vol. 17, no. 1, pp. 1–14, 1984.
- [72] J. A. Nelder and R. Mead, "A simplex method for function minimization," The computer journal, vol. 7, no. 4, pp. 308–313, 1965.
- [73] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder–mead simplex method in low dimensions," SIAM Journal on optimization, vol. 9, no. 1, pp. 112–147, 1998.
- [74] J. Monteiro, R. H. Widen, A. C. Pignatari, C. Kubasek, and S. Silbert, "Rapid detection of carbapenemase genes by multiplex real-time pcr," Journal of Antimicrobial Chemotherapy, vol. 67, no. 4, pp. 906–909, 2012.
- [75] D. Witters, B. Sun, S. Begolo, J. Rodriguez-Manzano, W. Robles, and R. F. Ismagilov, "Digital biology and chemistry," Lab on a Chip, vol. 14, no. 17, pp. 3225–3232, 2014.
- [76] B. Sun, J. Rodriguez-Manzano, D. A. Selck, E. Khorosheva, M. A. Karymov, and R. 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, vol. 53, no. 31, pp. 8088–8092, 2014.

- [77] J. Rodriguez-Manzano, M. A. Karymov, S. Begolo, D. A. Selck, D. V. Zhukov, E. Jue, and R. F. Ismagilov, "Reading out single-molecule digital rna and dna isothermal amplification in nanoliter volumes with unmodified camera phones," *ACS nano*, vol. 10, no. 3, pp. 3102–3113, 2016.
- [78] D. Sisti, M. Guescini, M. B. Rocchi, P. Tibollo, M. D'Atri, and V. Stocchi, "Shape based kinetic outlier detection in real-time pcr," *BMC bioinformatics*, vol. 11, no. 1, p. 186, 2010.
- [79] M. Guescini, D. Sisti, M. B. Rocchi, R. Panebianco, P. Tibollo, and V. Stocchi, "Accurate and precise dna quantification in the presence of different amplification efficiencies using an improved cy0 method," *PloS one*, vol. 8, no. 7, p. e68481, 2013.
- [80] C. L. Tooke, P. Hinchliffe, E. C. Bragginton, C. K. Colenso, V. H. Hirvonen, Y. Takebayashi, and J. Spencer, " β -lactamases and β -lactamase inhibitors in the 21st century," *Journal of molecular biology*, vol. 431, no. 18, pp. 3472–3500, 2019.
- [81] P. Nordmann, T. Naas, and L. Poirel, "Global spread of carbapenemase-producing enterobacteriaceae," *Emerging infectious diseases*, vol. 17, no. 10, p. 1791, 2011.
- [82] C. L. Ventola, "The antibiotic resistance crisis: part 1: causes and threats," *Pharmacy and therapeutics*, vol. 40, no. 4, p. 277, 2015.
- [83] P. H. England, "Carbapenemase-producing Enterobacteriaceae: laboratory confirmed cases, 2003 to 2015." <https://www.gov.uk/government/publications/carbapenemase-producing-enterobacteriaceae-laboratory-confirmed-cases/carbapenemase-producing-enterobacteriaceae-laboratory-confirmed-cases-2003-to-2013>, accessed 2016-01-27.
- [84] D. van Duin and Y. Doi, "The global epidemiology of carbapenemase-producing enterobacteriaceae," *Virulence*, vol. 8, no. 4, pp. 460–469, 2017.
- [85] J. Otter, P. Burgess, F. Davies, S. Mookerjee, J. Singleton, M. Gilchrist, D. Parsons, E. Brannigan, J. Robotham, and A. Holmes, "Counting the cost of an outbreak of carbapenemase-producing enterobacteriaceae: an economic evaluation from a hospital perspective," *Clinical Microbiology and Infection*, vol. 23, no. 3, pp. 188–196, 2017.
- [86] J. A. Otter, M. Doumith, F. Davies, S. Mookerjee, E. Dyakova, M. Gilchrist, E. T. Brannigan, K. Bamford, T. Galletly, H. Donaldson, *et al.*, "Emergence and clonal spread of colistin resistance due to multiple mutational mechanisms in carbapenemase-producing *klebsiella pneumoniae* in london," *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

- [87] Y. Carmeli, M. Akova, G. Cornaglia, G. Daikos, J. Garau, S. Harbarth, G. Rossolini, M. Souli, and H. Giamarellou, "Controlling the spread of carbapenemase-producing gram-negatives: therapeutic approach and infection control," Clinical microbiology and infection, vol. 16, no. 2, pp. 102–111, 2010.
- [88] R. Viau, K. M. Frank, M. R. Jacobs, B. Wilson, K. Kaye, C. J. Donskey, F. Perez, A. Endimiani, and R. A. Bonomo, "Intestinal carriage of carbapenemase-producing organisms: current status of surveillance methods," Clinical microbiology reviews, vol. 29, no. 1, pp. 1–27, 2016.
- [89] G. M. Knight, E. Dyakova, S. Mookerjee, F. Davies, E. T. Brannigan, J. A. Otter, and A. H. Holmes, "Fast and expensive (pcr) or cheap and slow (culture)? a mathematical modelling study to explore screening for carbapenem resistance in uk hospitals," BMC medicine, vol. 16, no. 1, pp. 1–11, 2018.
- [90] A. Moniri, J. Rodriguez-Manzano, K. Malpartida-Cardenas, L.-S. Yu, X. Didelot, A. Holmes, and P. Georgiou, "Framework for dna quantification and outlier detection using multidimensional standard curves," Analytical chemistry, vol. 91, no. 11, pp. 7426–7434, 2019.
- [91] A.-N. Spiess, C. Feig, and C. Ritz, "Highly accurate sigmoidal fitting of real-time pcr data by introducing a parameter for asymmetry," BMC bioinformatics, vol. 9, no. 1, p. 221, 2008.
- [92] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.
- [93] J. Osei Sekyere, U. Govinden, and S. Essack, "Review of established and innovative detection methods for carbapenemase-producing gram-negative bacteria," Journal of applied microbiology, vol. 119, no. 5, pp. 1219–1233, 2015.
- [94] Y.-J. Lee, D. Kim, K. Lee, and J.-Y. Chun, "Single-channel multiplexing without melting curve analysis in real-time pcr," Scientific reports, vol. 4, no. 1, pp. 1–6, 2014.
- [95] A. A. Morley, "Digital pcr: a brief history," Biomolecular detection and quantification, vol. 1, no. 1, pp. 1–2, 2014.
- [96] P.-L. Quan, M. Sauzade, and E. Brouzes, "dpcr: a technology review," Sensors, vol. 18, no. 4, p. 1271, 2018.

- [97] K. R. Sreejith, C. H. Ooi, J. Jin, D. V. Dao, and N.-T. Nguyen, "Digital polymerase chain reaction technology—recent advances and future perspectives," *Lab on a chip*, vol. 18, no. 24, pp. 3717–3732, 2018.
- [98] G. Perkins, H. Lu, F. Garlan, and V. Taly, "Droplet-based digital pcr: application in cancer research," in *Advances in clinical chemistry*, vol. 79, pp. 43–91, Elsevier, 2017.
- [99] S. Olmedillas-López, M. García-Arranz, and D. García-Olmo, "Current and emerging applications of droplet digital PCR in oncology," *Molecular diagnosis & therapy*, vol. 21, no. 5, pp. 493–510, 2017.
- [100] Y. Tong, S. Shen, H. Jiang, and Z. Chen, "Application of digital pcr in detecting human diseases associated gene mutation," *Cellular Physiology and Biochemistry*, vol. 43, no. 4, pp. 1718–1730, 2017.
- [101] F. Shen, B. Sun, J. E. Kreutz, E. K. Davydova, W. Du, P. L. Reddy, L. J. Joseph, and R. F. Ismagilov, "Multiplexed quantification of nucleic acids with large dynamic range using multivolume digital rt-pcr on a rotational slipchip tested with hiv and hepatitis c viral load," *Journal of the American Chemical Society*, vol. 133, no. 44, pp. 17705–17712, 2011.
- [102] A. Rajagopal, D. Yurk, C. Shin, K. Menge, L. Jacky, S. Fraser, T. A. Tombrello, and G. J. Tsongalis, "Significant expansion of real-time pcr multiplexing with traditional chemistries using amplitude modulation," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.
- [103] M. Tajadini, M. Panjehpour, and S. H. Javanmard, "Comparison of sybr green and taqman methods in quantitative real-time polymerase chain reaction analysis of four adenosine receptor subtypes," *Advanced biomedical research*, vol. 3, 2014.
- [104] J. Ruijter, C. Ramakers, W. Hoogaars, Y. Karlen, O. Bakker, M. Van den Hoff, and A. Moorman, "Amplification efficiency: linking baseline and bias in the analysis of quantitative pcr data," *Nucleic acids research*, vol. 37, no. 6, pp. e45–e45, 2009.
- [105] J. Tanaka, T. Nakagawa, A. Shiratori, Y. Shimazaki, C. Uematsu, M. Kamahori, T. Yokoi, K. Harada, and Y. Kohara, "Kras genotyping by digital pcr combined with melting curve analysis," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.
- [106] P. Athamanolap, K. Hsieh, C. M. O'Keefe, Y. Zhang, S. Yang, and T.-H. Wang, "Nanoarray digital polymerase chain reaction with high-resolution melt for enabling broad bacteria identification and pheno-molecular antimicrobial susceptibility test," *Analytical chemistry*, vol. 91, no. 20, pp. 12784–12792, 2019.

- [107] D. Hammoudi Halat and C. Ayoub Moubareck, “The current burden of carbapenemases: Review of significant properties and dissemination among gram-negative bacteria,” *Antibiotics*, vol. 9, no. 4, p. 186, 2020.
- [108] J. Rodriguez-Manzano, N. Moser, K. Malpartida-Cardenas, A. Moniri, L. Fisarova, I. Penlisi, A. Boonyasiri, E. Jauneikaite, A. Abdolrasouli, J. A. Otter, *et al.*, “Rapid detection of mobilized colistin resistance using a nucleic acid based lab-on-a-chip diagnostic system,” *Scientific Reports*, vol. 10, no. 1, pp. 1–9, 2020.
- [109] T. Gou, J. Hu, W. Wu, X. Ding, S. Zhou, W. Fang, and Y. Mu, “Smartphone-based mobile digital pcr device for dna quantitative analysis with high accuracy,” *Biosensors and Bioelectronics*, vol. 120, pp. 144–152, 2018.
- [110] R. C. Edgar, “Muscle: multiple sequence alignment with high accuracy and high throughput,” *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.
- [111] M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, *et al.*, “Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data,” *Bioinformatics*, vol. 28, no. 12, pp. 1647–1649, 2012.
- [112] J. SantaLucia Jr and D. Hicks, “The thermodynamics of dna structural motifs,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 33, pp. 415–440, 2004.
- [113] Z. Dwight, R. Palais, and C. T. Wittwer, “umelt: prediction of high-resolution melting curves and dynamic melting profiles of pcr products in a rich web application,” *Bioinformatics*, vol. 27, no. 7, pp. 1019–1020, 2011.
- [114] 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%90906896/qdpcr-37k-dpcr-qr-100%E2%80%90906896/fluidigm%3Afile>, 2014.
- [115] B. Wong, “Points of view: Color blindness,” *Nature Methods*, vol. 8, no. 6, pp. 441–441, 2011.
- [116] G. P. McDermott, D. Do, C. M. Litterst, D. Maar, C. M. Hindson, E. R. Steenblock, T. C. Legler, Y. Jouvenot, S. H. Marrs, A. Bemis, *et al.*, “Multiplexed target detection using dna-binding dye chemistry in droplet digital pcr,” *Analytical chemistry*, vol. 85, no. 23, pp. 11619–11627, 2013.

- [117] B. C. Millar, J. Xu, and J. 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, vol. 40, no. 5, pp. 1575–1580, 2002.
- [118] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [119] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, Learning from data, vol. 4. AML-Book New York, NY, USA:, 2012.
- [120] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.
- [121] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE transactions on information theory, vol. 13, no. 1, pp. 21–27, 1967.
- [122] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative pcr and the 2- $\delta\delta$ ct method," methods, vol. 25, no. 4, pp. 402–408, 2001.
- [123] A. S. Whale, J. F. Huggett, S. Cowen, V. Speirs, J. Shaw, S. Ellison, C. A. Foy, and D. J. Scott, "Comparison of microfluidic digital pcr and conventional quantitative pcr for measuring copy number variation," Nucleic acids research, vol. 40, no. 11, pp. e82–e82, 2012.
- [124] J. Kuypers and K. R. Jerome, "Applications of digital pcr for clinical microbiology," Journal of clinical microbiology, vol. 55, no. 6, pp. 1621–1628, 2017.
- [125] M. E. Dueck, R. Lin, A. Zayac, S. Gallagher, A. K. Chao, L. Jiang, S. S. Datwani, P. Hung, and E. Stieglitz, "Precision cancer monitoring using a novel, fully integrated, microfluidic array partitioning digital pcr platform," Scientific reports, vol. 9, no. 1, pp. 1–9, 2019.
- [126] G. Zangenberg, R. Saiki, and R. Reynolds, "Multiplex pcr: optimization guidelines," in PCR Applications, pp. 73–94, Elsevier, 1999.
- [127] C. T. Wittwer, M. G. Herrmann, C. N. Gundry, and K. S. Elenitoba-Johnson, "Real-time multiplex pcr assays," Methods, vol. 25, no. 4, pp. 430–442, 2001.
- [128] A. S. Whale, J. F. Huggett, and S. Tzonev, "Fundamentals of multiplexing with digital pcr," Biomolecular Detection and Quantification, vol. 10, pp. 15–23, 2016.
- [129] S. I. Fraley, J. Hardick, B. Jo Masek, P. Athamanolap, R. E. Rothman, C. A. Gaydos, K. C. Carroll, T. Wakefield, T.-H. Wang, and S. Yang, "Universal digital high-resolution

- melt: a novel approach to broad-based profiling of heterogeneous biological samples,” Nucleic acids research, vol. 41, no. 18, pp. e175–e175, 2013.
- [130] A. Moniri, L. Miglietta, K. Malpartida-Cardenas, I. Pennisi, M. Cacho-Soblechero, N. Moser, A. Holmes, P. Georgiou, and J. Rodriguez-Manzano, “Amplification curve analysis: Data-driven multiplexing using real-time digital pcr,” Submitted to Analytical Chemistry, 2020.
- [131] J. C. Rolando, E. Jue, J. T. Barlow, and R. 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, vol. 48, no. 7, pp. e42–e42, 2020.
- [132] T. Nakagawa, J. Tanaka, K. Harada, A. Shiratori, Y. Shimazaki, T. Yokoi, C. Uematsu, and Y. Kohara, “10-plex digital polymerase chain reaction with four-color melting curve analysis for simultaneous kras and braf genotyping,” Analytical Chemistry, vol. 92, no. 17, pp. 11705–11713, 2020.
- [133] W. H. Organization et al., Antimicrobial resistance: global report on surveillance. World Health Organization, 2014.
- [134] T. Lima, S. Domingues, and G. J. Da Silva, “Plasmid-mediated colistin resistance in salmonella enterica: a review,” Microorganisms, vol. 7, no. 2, p. 55, 2019.
- [135] L. M. Carroll, A. Gaballa, C. Guldemann, G. Sullivan, L. O. Henderson, and M. Wiedmann, “Identification of novel mobilized colistin resistance gene mcr-9 in a multidrug-resistant, colistin-susceptible salmonella enterica serotype typhimurium isolate,” mBio, vol. 10, no. 3, pp. e00853–19, 2019.
- [136] P. Manohar, T. Shanthini, R. Ayyanar, B. Bozdogan, A. Wilson, A. J. Tamhankar, R. Nachimuthu, and B. S. Lopes, “The distribution of carbapenem-and colistin-resistance in gram-negative bacteria from the tamil nadu region in india,” Journal of medical microbiology, vol. 66, no. 7, pp. 874–883, 2017.
- [137] 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.
- [138] P. Athamanolap, V. Parekh, S. I. Fraley, V. Agarwal, D. J. Shin, M. A. Jacobs, T.-H. Wang, and S. Yang, “Trainable high resolution melt curve machine learning classifier for large-scale reliable genotyping of sequence variants,” PloS one, vol. 9, no. 10, p. e109094, 2014.

- [139] D. O. Velez, H. Mack, J. Jupe, S. Hawker, N. Kulkarni, B. Hedayatnia, Y. Zhang, S. Lawrence, and S. I. Fraley, “Massively parallel digital high resolution melt for rapid and absolutely quantitative sequence profiling,” *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.
- [140] V. García, I. García-Meniño, A. Mora, S. C. Flament-Simon, D. Díaz-Jiménez, J. E. Blanco, M. P. Alonso, and J. 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*, vol. 52, no. 1, pp. 104–108, 2018.
- [141] N. Miscourides, L.-S. Yu, J. Rodriguez-Manzano, and P. Georgiou, “A 12.8 k current-mode velocity-saturation isfet array for on-chip real-time dna detection,” *IEEE transactions on biomedical circuits and systems*, no. 99, pp. 1–13, 2018.
- [142] J. Rodriguez-Manzano, N. Miscourides, K. Malpartida-Cardenas, I. Pennisi, N. Moser, A. Holmes, and P. Georgiou, “Rapid detection of *klebsiella pneumoniae* using an auto-calibrated isfet-array lab-on-chip platform,” in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4, IEEE, 2019.
- [143] N. Miscourides, “Linear ISFET Arrays and Optimisation Methods for DNA detection.” PhD thesis. Imperial College London, Dept of Electrical and Electronic Engineering, 2018.
- [144] Y. L. Chin, J. C. Chou, T. P. Sun, W. Y. Chung, and S. K. Hsiung, “A novel pH sensitive ISFET with on chip temperature sensing using CMOS standard process,” *Sensors and Actuators, B: Chemical*, vol. 76, pp. 582–593, jun 2001.
- [145] M. Daniel, M. Janicki, and A. Napieralski, “Simulation of Ion Sensitive Transistors Using a SPICE Compatible Model,” in *Proceedings of IEEE Sensors*, vol. 2, pp. 543–548, 2003.
- [146] Y. Liu, P. Georgiou, T. Prodromakis, T. G. Constandinou, and C. Toumazou, “An extended CMOS ISFET model incorporating the physical design geometry and the effects on performance and offset variation,” *IEEE Transactions on Electron Devices*, vol. 58, pp. 4414–4422, dec 2011.
- [147] J. A. Herberg, M. Kaforou, V. J. Wright, H. Shailes, H. Eleftherohorinou, C. J. Hoggart, M. Cebey-López, M. J. Carter, V. A. Janes, S. Gormley, *et al.*, “Diagnostic test accuracy of a 2-transcript host rna signature for discriminating bacterial vs viral infection in febrile children,” *Jama*, vol. 316, no. 8, pp. 835–845, 2016.

-
- [148] M. Kaforou, J. A. Herberg, V. J. Wright, L. J. Coin, and M. Levin, “Diagnosis of bacterial infection using a 2-transcript host rna signature in febrile infants 60 days or younger,” Jama, vol. 317, no. 15, pp. 1577–1578, 2017.
- [149] S. S. Haykin, Adaptive filter theory. Pearson Education India, 2005.
- [150] C. Yuan and H. Yang, “Research on k-value selection method of k-means clustering algorithm,” J—Multidisciplinary Scientific Journal, vol. 2, no. 2, pp. 226–235, 2019.
- [151] J. Hopcroft and R. Tarjan, “Algorithm 447: efficient algorithms for graph manipulation,” Communications of the ACM, vol. 16, no. 6, pp. 372–378, 1973.
- [152] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.