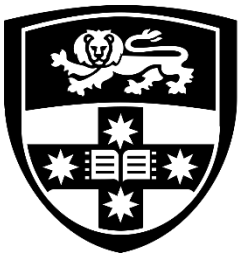


THE UNIVERSITY OF
SYDNEY

Discovery and Development of Novel Precision Medicine Tools for Breast Cancer

Twingle Daniel

BSc, The University of Sydney 2018

BSc Honours Class I, The University of Sydney 2019

A thesis submitted to fulfil requirements for the degree of Doctor of Philosophy at The University of Sydney in 2024

Faculty of Medicine and Health

Abstract

Breast cancer is a heterogeneous disease that can be classified into clinically different subtypes based on histopathological and molecular features. Estrogen receptor positive (ER+) breast cancer is classified into two intrinsic molecular subtypes: luminal A and luminal B. Outcomes are overall poorer and proliferation rates are higher in the luminal B subtype compared to luminal A. Commercial multigene tests exploit this molecular difference to accurately predict outcome in ER+ cases. However, in Australia, their use is limited by high cost. For triple negative breast cancers (TNBC) lacking ER, progesterone receptor or amplification of human epidermal growth factor receptor 2 (HER2), outcomes are overall poorer and there are currently no robust predictive tools.

This thesis describes the development and validation of the PROSPER proliferation-based gene signature assay, as an affordable test to accurately predict recurrence in ER+ cases. In addition, novel approaches were used to discover gene expression signatures that stratified TNBC into clinically distinct groups.

The PROSPER signature was integrated into a multiplexed reverse transcription quantitative PCR (RT-qPCR) assay, which produced robust results in formalin-fixed paraffin-embedded (FFPE) breast cancer cohorts. PROSPER results significantly correlated with the EndoPredict commercial test and proliferation marker Ki-67, demonstrating its utility for ER+ breast cancers.

TNBC can be classified into biologically different subtypes. A sequential subtyping approach was used to stratify TNBC cases, and subtype-specific prognostic gene signatures were identified using elastic net regularisation. While this approach demonstrated significant value in predicting disease-free outcome in the test dataset, predictive values were lower in the validation cohort suggesting further work is required.

This study makes an important contribution towards equitable delivery of precision tests for breast cancer patients.

Statement of originality

This is to certify that, to the best of my knowledge; the content of this thesis is my own work. This thesis has not been submitted for any degree or other purpose.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Twingle Daniel		27 June 2024
Name	**Signature**	**Date**

Supervisor statement

In my opinion, the thesis:

- is sufficiently well presented to be examined; and
- does not exceed the prescribed word limit

Dinny Graham		26 June 2024
Supervisor Name	**Supervisor Signature**	**Date**

Contributions by others

The processing, H&E staining, and sectioning of FFPE samples in Chapter 4 were performed by Barb Guild in the Translational Breast Cancer Genomics Research Group at WIMR. Bioinformatic analyses in Chapter 5 were assisted by Tram Doan from the Translational Breast Cancer Genomics Research Group.

Financial support

My PhD candidature was supported by a University of Sydney Postgraduate Award scholarship from The University of Sydney. Funding for research included in this thesis was provided by a Senior Research Award from Tour de Cure and Rapid Applied Research Translation Award from the Medical Research Future Fund.

Acknowledgement

I would like to express my deepest appreciation and gratitude towards my supervisor, Dr Dinny Graham. Dinny's guidance, support and mentorship over the past 6 years was invaluable. I am extremely privileged to be able to learn from her. Her expertise has allowed me to complete this research. Without her tireless efforts, patience and constant encouragement, this work would not have been possible. I would also like to thank Dr Nirmala Pathmanathan, for her clinical expertise and insights that have been invaluable in shaping this research.

A special thank you to the Breast group, specifically Barb, who has been a constant source of support throughout a lot of lab work and especially in guiding me through numerous experiments. I am also deeply grateful to Tram. Her bioinformatics knowledge is limitless, and I cannot thank her enough for her patience and knowledge which immensely helped me in completing my research.

Another special thank you to the SpeeDx team, especially Nicole and Ally. Their expertise in molecular diagnostics, overall guidance and emotional support has been invaluable.

Additionally, I would like to extend my sincere gratitude to my friends and family. They have been constant pillars of support, providing emotional encouragement throughout this challenging journey. At the same time, they provided me with distraction from the research at critical times when I needed them.

Finally, to my partner in crime, Tony. His unwavering support and belief in me have pushed me to finish this research. Thank you for being the listening ear and thank you for being there for me.

Table of contents

Abstract 2
Statement of originality 4
Supervisor statement 5
Contributions by others 6
Financial support 6
Acknowledgement 7
Table of contents 8
List of figures 12
List of tables 15
List of abbreviations and acronyms 16
1 Chapter 1 19
1.1 Introduction 19
1.2 Breast Cancer 19
1.2.1 Traditional breast cancer classification and treatment 20
1.2.2 Tumour size, nodal status, and distant metastasis staging 25
1.2.3 Intrinsic molecular subtype 28
1.2.4 Current clinical management and limitations 32
1.2.5 Commercial multi-gene tests for more precise risk stratification 33
1.2.6 The challenge of classifying TNBC 37
1.3 Prognostic tests guiding treatment of ER+ breast cancer 40
1.3.1 Current commercial multi-gene tests are expensive 40
1.3.2 The prognostic power of multigene tests reflects proliferative activity 41
1.3.3 PROSPER signature with the NanoString nCounter platform. 47
1.3.4 qPCR as an affordable technique for routine clinical use 49
1.3.5 Normalisation of gene expression in RT-qPCR 56
1.4 Lack of prognostic tests guiding treatment of TNBC 58
1.5 Aims and hypotheses 59
1.5.1 Aims 59
1.5.2 Hypotheses 60
2 Chapter 2 61

2.1 Material 61
2.1.1 FFPE and flash frozen cell line samples 65
2.1.2 Breast cancer patient samples 65
2.2 Methods 69
2.2.1 Tissue culture 69
2.2.2 H&E staining of FFPE samples 71
2.2.3 Processing of FFPE samples for nucleic acid extraction 71
2.2.4 Standard one-step RT- qPCR conditions 72
2.2.5 Standard nucleic acid and oligonucleotide quantity and quality procedures 74
3 Chapter 3 75
3.1 Introduction 75
3.2 Aims and Hypotheses 78
3.2.1 Aims 78
3.2.2 Hypotheses 78
3.3 Approach 79
3.4 PROSPER qPCR assay design 82
3.4.1 RNA-based PROSPER qPCR assay 83
3.4.2 TNA-based qPCR assay 84
3.4.3 Oligonucleotide design for measurement of the PROSPER signature 85
3.5 Results 89
3.5.1 Performance of each qPCR assay in a single plex 89
3.5.2 Performance of assay in 2-plex and 3-plex format 92
3.5.3 Performance of assay in 4-plex format 93
3.5.4 Performance of assay in 5-plex format 95
3.5.5 Transcription of DNA and RNA in the PROSPER gene signature 98
3.5.6 Validation of the 5-plex assay in cell line TNA samples 101
3.6 Discussion 105
3.6.1 Summary of Results 105
3.6.2 Computational tools for optimising primer design for multiplex qPCR assays . 106
3.6.3 Experimental validation of *in silico* designed primers 107
3.6.4 Detection of nucleic acids: TNA-based and RNA-based assays 109
3.6.5 Assay validation with synthetic progestin treatment 110
3.6.6 Limitations and future directions 111

3.6.7 Conclusion 112
4 Chapter 4 114
4.1 Introduction 114
4.2 Aims and hypotheses 117
4.2.1 Aims 117
4.2.2 Hypotheses 117
4.3 Approach 118
4.4 Results 120
4.4.1 Selection of appropriate nucleic acid extraction kits 120
4.4.2 Evaluation of initial sample preparation for TNA yield 127
4.4.3 Evaluation of incubation time on TNA yield 130
4.4.4 Enhancement of the multiplexed PCR assay 131
4.4.5 Evaluation of the optimised multiplexed assay in stored RNA 135
4.4.6 Redesign and optimisation of the PROSPER multiplexed PCR assay 137
4.4.7 Comparative analysis of PROSPER index 144
4.5 Discussion 150
4.5.1 Summary of results 150
4.5.2 Integrating the multiplexed qPCR assay for laboratory diagnostics 155
4.5.3 Optimising the extraction protocol and qPCR multiplexed assay 156
4.5.4 Implementation of PROSPER signature as multiplexed qPCR assay 159
4.5.5 Study strengths, limitations and future directions 162
4.5.6 Conclusions 162
5 Chapter 5 164
5.1 Introduction 164
5.2 Aims and hypotheses 167
5.2.1 Aims 167
5.2.2 Hypothesis 167
5.3 Approach 168
5.4 Results 170
5.4.1 Data search and selection 170
5.4.2 Assembly of the TNBC cohort 171
5.4.3 Data preparation and preprocessing 174
5.4.4 TNBC subtype classification and dataset segmentation 174

5.4.5 Feature selection and prognostic signature identification 175
5.4.6 Prognostic utility of the models 185
5.4.7 Independent validation of the prognostic gene signatures 188
5.5 Discussion 191
5.5.1 Summary of results 191
5.5.2 Preliminary exploration of subtype-specific prognostic signatures in TNBC 193
5.5.3 The impact of TNBC subtype-specific prognostication on treatment 195
5.5.4 Validation of the TNBC subtype-specific prognostic gene signature 197
5.5.5 Study strengths, limitations and future directions 199
5.5.6 Conclusion 201
6 Chapter 6 203
6.1 Final Remarks 203
7 Chapter 7 214

List of figures

Figure 1.1: Normal Breast structure and breast cancer subtypes. 31
Figure 1.2: Proliferation reflects the power of multigene tests. 42
Figure 1.3: Flowchart of the discovery and validation of PROSPER using the METABRIC cohort. 44
Figure 1.4: Performance of the PROSPER signature in the validation cohort of the ER+ METABRIC dataset. 45
Figure 1.5: A Prospective study that compares the PROSPER index and EndoPredict score (EPClin) in a cohort of 130 patients. 46
Figure 1.6: Association between PROSPER index and survival in a retrospective study with 142 samples with extensive follow-up data. 47
Figure 1.7: NanoString Assay Chemistry. 48
Figure 1.8: A qPCR workflow. 51
Figure 1.9: PCR amplification phases. 52
Figure 1.10: Principle of detection using SYBR Green and TaqMan Probe. 54
Figure 1.11: Schematic representation of *PlexZyme* assay components and signal detection. 56
Figure 3.1: Workflow of synthetic Progestin, ORG2058, treatment of T-47D cell lines. 80
Figure 3.2: Graphical summary of Chapter 3. 81
Figure 3.3: Linear correlation between the 5 gene and 4 gene PROSPER index. 83
Figure 3.4: Variance of PROSPER signature housekeeping genes. 84
Figure 3.5: *PlexZyme* assay design. 86
Figure 3.6: Multiplex PCR Primer and *PlexZyme* Design workflow. 87
Figure 3.7: Single plex amplification plots of the PROSPER targets. 90
Figure 3.8: Cq values and ΔCq differences of single plex assays. 91
Figure 3.9: Compatibility of the 3-plex assay on TNA samples from flash-frozen T-47D cell lines. 93
Figure 3.10: Compatibility of the 4-plex assay on TNA samples from flash-frozen T-47D cell lines. 94
Figure 3.11: Compatibility of 5-plex assay on TNA samples from flash-frozen T-47D cells. 95

Figure 3.12: Comparison of Cq values of RNA-based assay. 96
Figure 3.13: Comparison of Cq values of TNA-based assay. 98
Figure 3.14: Comparison of ΔCq with and without reverse transcriptase reactions. 101
Figure 3.15: Performance of the 5-plex assay across various sample types. 102
Figure 3.16: Validation of PROSPER multiplexed qPCR assay. 104
Figure 4.1: Workflow of the two commercial FFPE extraction kits. 119
Figure 4.2: TNA-based qPCR assay performance on TNA extracted using the two kits in the FFPE patient sample from Cohort 1. 123
Figure 4.3: TNA-based assay performance on TNA extracted from the Quick Kit in FFPE samples from Cohort 1. 124
Figure 4.4: Comparison of TNA extracted from commercial TNA extraction kits in a fresh FFPE patient sample. 126
Figure 4.5: H&E staining and annotation of Cohort 1 FFPE patient samples, along with the TNA-based assay performance. 129
Figure 4.6: Cq values from the optimised TNA extraction kit on FFPE patient samples in Cohort 1 130
Figure 4.7: TNA-based assay performance on TNA extracted using different incubation times from FFPE patient samples in Cohort 1. 131
Figure 4.8: Gene 4 assay optimisations on TNA extracted from fresh T-47D cell pellet. 133
Figure 4.9: Assessment of the optimised TNA extraction and multiplexed PCR assay performance on FFPE samples from Cohort 1. 134
Figure 4.10: Assessment of the RNA-based assay performance on stored residual RNA from Cohort 1. 136
Figure 4.11: Transcription of TNA using the smaller amplicon assay. 138
Figure 4.12: Assessment of the smaller amplicon assay multiplexed PCR assay performance on TNA extracted from FFPE patient samples from Cohort 1. 139
Figure 4.13: Performance of smaller amplicon TNA-based assay on Cohort 1 samples. 140
Figure 4.14: Assessment of the smaller amplicon multiplexed PCR assay performance on RNA extracted from FFPE blocks from Cohort 1. 141
Figure 4.15: Assessment of the smaller amplicon multiplexed PCR assay performance on residual RNA from archival FFPE samples in Cohort 2 143

Figure 4.16: Comparison of PROSPER index measurements from Cohort 1 obtained from the multiplexed RNA-based PCR across different assays and clinical parameters. 145
Figure 4.17: Comparison of PROSPER index measurements obtained from the multiplexed TNA-based PCR across different assays and clinical parameters. 147
Figure 4.18: Comparison of PROSPER index measurements obtained from different assays tested on a subset of samples from Cohort 1. 149
Figure 5.1: Graphical overview of chapter five. 169
Figure 5.2: Gene set sizes across different alpha values obtained from elastic net regularisation modelling. 184
Figure 5.3 Gene set sizes and their associated Cox regression p-values obtained from elastic net regularisation modelling. 185
Figure 5.4: Performance of subtype specific gene signatures in discriminating patient samples according to risk of recurrence. 187
Figure 5.5: Performance of the non-subtype specific gene signature in discriminating patient samples according to risk of recurrence..... 188
Figure 5.6: Performance of subtype-specific gene signatures in discriminating patient samples according to the risk of recurrence in the METABRIC TNBC validation cohort. 189
Figure 5.7: Performance of non-subtype specific gene signatures in discriminating patient samples according to the risk of recurrence in the METABRIC TNBC cohort. 190

List of tables

Table 1.1 Elston-Ellis modified Nottingham grading system 23
Table 1.2 TNM classification..... 26
Table 1.3 Breast cancer stages based on the TNM classification..... 27
Table 1.4 The associated turnaround times and costs of commercially available multi-gene tests 41
Table 2.1 Cell line 61
Table 2.2 Reagents 61
Table 2.3 Nucleic acid extraction kits 63
Table 2.4 Equipment 63
Table 2.5 Software 64
Table 2.6 Cohort characteristics 67
Table 3.1 Fold difference in single plex assay 91
Table 3.2 ΔCq values from 5-plex PCR assay for the detection of RNA transcript levels in the PROSPER gene signature with RNA HK Gene 1 100
Table 3.3 ΔCq values from 5-plex PCR assay for the detection of RNA and DNA transcription in the PROSPER gene signature with NED Gene 1 100
Table 4.1 Specifications of the two TNA extraction kits..... 121
Table 4.2 Cq values were obtained from the amplification of RNA utilising the smaller amplicon multiplexed assay on Cohort 1 samples..... 141
Table 4.3 Summary of the comparison of the PROSPER index against markers. 153
Table 5.1 Overview of triple-negative breast cancer microarray studies. 172
Table 5.2 Clinical characteristics of the TNBC patients..... 173
Table 5.3 Classification of samples from the TNBC cohort into TNBC subtypes..... 175
Table 5.4 Significantly associated ($p < 0.04$) gene from the TNBC subtypes obtained from univariate Cox regression. 177
Table 5.5 Genes significantly associated ($p < 0.01$) with DFS from the non-subtype specific TNBC cohort obtained from univariate Cox regression. 181

List of abbreviations and acronyms

ATCC	American type culture collection
BL	Basal like TNBC subtype
BL1	Basal like 1 TNBC subtype
BL2	Basal like 2 TNBC subtype
BLBC	Basal-like breast cancer
BLIA	Basal like immune activated TNBC subtype
BLIS	Basal like immunosuppressed TNBC subtype
cDNA	Complementary DNA
Cq	Quantification cycle
DFS	Disease-free survival
DNA	Deoxyribonucleic acid
DSS	Disease-specific survival
EP score	EndoPredict score
EPclin score	EndoPredict clinical score
ER	Estrogen receptor
ER+	Estrogen receptor positive
FFPE	Formalin-fixed paraffin-embedded
GEO	Gene expression omnibus
H&E	Haematoxylin and eosin
HER2	Human epidermal growth factor receptor 2
HER2+	Human epidermal growth factor receptor 2 positive

HK genes	Housekeeping genes
HREC	Human Research Ethics Committee
IHC	Immunohistochemistry
IM	Immunomodulatory TNBC subtype
LAR	Luminal androgen receptor TNBC subtype
M	Mesenchymal tnbc subtype
MasterPure Kit	MasterPure complete DNA and RNA purification kit
MES	Mesenchymal TNBC subtype
METABRIC	Molecular taxonomy of breast cancer international consortium
mRNA	Messenger RNA
MSL	Mesenchymal stem-like tnbc subtype
NED	Non-transcribed DNA
NED Gene 1	Non-transcribed DNA gene 1
NOS	Not otherwise specified
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PR	Progesterone receptor
PR+	Progesterone receptor positive
PROSPER	PROliferation Signature for Prognosticating ER+ breast cancer
qPCR	Quantitative polymerase chain reaction
Quick Kit	Quick-DNA/RNA FFPE kit
RFU	Relative fluorescence units
RMA	Robust multi-array average

RNA	Ribonucleic acid
RNA HK Gene 1	RNA housekeeping gene 1
RS	Recurrence score
RT	Reverse transcriptase
RT-qPCR	Reverse transcription-quantitative polymerase chain reaction
Thrombin	Thromborel S
TNA	Total nucleic acid
TNBC	Triple-negative breast cancer
TNM	Tumour node metastasis
TrypLE	TrypLE express enzyme (1X), no phenol red
WIMR	Westmead Institute of Medical Research
WSLHD	Western Sydney Local Health District
α	alpha

Chapter 1

Introduction

1.1 Introduction

Carcinogenesis occurs when cells undergo genetic mutations that disrupt normal cellular processes, leading to uncontrolled cell growth and the formation of cancer. The key mechanisms driving its progression include sustained proliferation, evasion of growth suppressors, resistance to cell death, enhanced angiogenesis, replicative immortality, and activated invasion and metastasis (1, 2). Cancer is a complex and multifaceted disease that is fundamentally caused by mutations arising from genetic predispositions, endogenous DNA replication errors and damage arising from environmental exposures. The increasing number of deaths related to cancer makes it one of the leading causes of mortality worldwide (3).

1.2 Breast Cancer

Breast cancer is a global problem. It is the most commonly diagnosed cancer in women, with more than 2.3 million new cases and it is also the leading cause of cancer death, with an estimated 685,000 deaths reported in 2020 (4, 5). Although survival rates have improved significantly over time due to recent advances in early detection and treatment, the incidence of breast cancer continues to rise worldwide. Current projections indicate that by 2030, the global annual number of new breast cancer cases diagnosed will reach 2.7 million, with an annual death toll estimated to be 0.87 million (6). Even though the exact cause of breast cancer is unknown, multiple factors

have been linked to the development and progression of breast cancer including age, gender, genetic predisposition such as germline mutations of breast cancer-disposing genes including BRCA1/2, PALB2, ATM, and TP53, early menarche, late menopause, nulliparity, obesity, high breast tissue density, long-term or high-dose ovarian hormone replacement therapy, alcohol consumption, and exposure to ionizing radiation (4, 7, 8).

1.2.1 Traditional breast cancer classification and treatment

The epithelial component of the breast consists of glandular tissue comprising breast lobes and ducts. Between these, are fibrous connective tissue and adipose tissue that fill in the gaps between these structures **(Figure 1.1 A**) (9). The breast of an adult woman consists of 15–20 lobes, each lobe containing 20–40 lobules. These lobules are connected to small milk ducts, which eventually join to form larger collecting ducts that open at the nipple. The terminal duct lobular unit is the functional unit of the breast responsible for milk synthesis and is composed of both the lobules and the associated ducts (**Figure 1.1 C**) (9). Terminal duct lobular units are epithelial structures comprised of luminal epithelial cells, which are responsible for milk production, surrounded by contractile myoepithelial cells that facilitate milk ejection; they are also the site of origin for most breast cancers. Extensive studies have revealed that both the luminal and myoepithelial cells of the breast arise from a single quiescent stem cell type, which gives rise to the breast epithelium via a complex hierarchy of progenitor and mature cell stages (10). When stimulated to produce breast structures, at puberty and during the first trimester of pregnancy, these stem cells undergo asymmetric division to produce bipotent progenitors, which in turn produce committed luminal and myoepithelial progenitor and mature cell populations. It is this somewhat unique adult replicative potential that is believed to render the epithelium of the breast to be vulnerable to

malignancy. Moreover, it is hypothesized that the heterogeneity of the various epithelial cell populations results in the development of a spectrum of distinct breast cancer subtypes that differ in their biological functions and gene expression. This influences the responsiveness to specific treatments for different breast cancer subtypes (9).

Breast cancer diagnosis and prognostication involves a multi-dimensional approach incorporating histopathological assessment, evaluation of clinical characteristics, and in some cases advanced molecular analysis. Non-invasive procedures such as mammography and MRI can predict the presence of malignant tumours (11, 12). However, a definitive diagnosis of breast cancer requires histopathological examination of the tumour tissue. Tumour samples are assessed at initial screening diagnosis via examination of fine needle aspirates or, more commonly, core needle biopsy specimens. The post-surgical excision specimen is also formalin-fixed and paraffin-embedded (FFPE) for extensive pathological and histological review, which includes assessment of surgical margins, tumour grading and measurement of immunohistochemical markers, including those for receptors and proliferative rate.

These results are then reviewed to determine the stage – *in situ* or invasive – and the histological subtype of the primary breast cancer, such as ductal or lobular carcinoma (**Figure 1.1 D-G**). *In situ* carcinomas are breast cancers that are confined to the ducts or lobules by an intact myoepithelial layer and have not invaded surrounding tissues, while invasive carcinoma refers to cancers that have spread into the surrounding stromal, and sometimes lympho-vascular space (13). Ductal carcinomas represent around 75% of breast cancers, with around 15% being lobular carcinomas. A small number of so-called ‘special types’ are also diagnosed, including medullary, mucinous, metaplastic, papillary and apocrine breast cancers, which together make up around 5-10% of breast cancers.

Treatment of cancer is typically guided by the stratification of cases according to their anatomical stage, and by estimates of their proliferative and invasive potential, generally defined by a grading system. Several grading scales exist, and these vary between cancer types. In breast cancer, the assessment of histological grade involves evaluating tumour biology using the Elston-Ellis modified Nottingham grading system **(Table 1.1**) (14, 15). This classification assesses the degree to which the tumour has lost the tissue architecture of the normal breast epithelium, the extent of nuclear pleomorphism and mitotic count. Each feature is scored with a three-tier system, and the final grade (Grade 1, 2 or 3) is determined by adding the individual scores (15). The histologic grade is an independent predictor of breast cancer prognosis, being a predictor of both breast cancer-specific survival and disease-free survival, with higher-grade tumours associated with poorer outcomes (16). However, the grading system has poor reproducibility between observers and facilities due to its subjective elements (16, 17). Reproducibility is highest at the ends of the spectrum of grade where the malignant potential of the tumour is less likely to be debatable, with the highest interobserver agreement observed for tubule formation compared with other components of the grading system (18). In contrast, reproducibility is poor for intermediate grade cases where some elements of the scoring system are equivocal. Additionally, while the grade components are given equal weight for the score, they have different predictive values, with tubule formation being least predictive and nuclear pleomorphism and mitotic index having higher predictive value (19).

Table 1.1 Elston-Ellis modified Nottingham grading system

Feature	Subcategory	Score
Tumour tubule and gland formation	> 75% of tumour	1
	10 - 75% of tumour	2
	< 10% of tumour	3
Nuclear pleomorphism	Minimal variation in nuclear size and shape	1
	Moderate variation in nuclear size and shape	2
	Marked variation in nuclear size and shape	3
Mitotic count (per 10 high power fields) *dependent on microscope field area, e.g. for field diameter of 0.59 mm	0-9	1
	10-19	2
	> 20	3
Total score	**Grade 1 (well differentiated)**	**3 - 5**
	Grade 2 (moderately differentiated)	**6 or 7**
	Grade 3 (poorly differentiated)	**8 or 9**

In addition to characterising the tissue appearance of breast cancers in the form of histological grade, breast tumours are stratified into three clinical subtypes based on the expression of the ovarian hormone receptors, estrogen receptor (ER) and progesterone receptor (PR), and amplification of the ERBB2 gene encoding the human epidermal growth factor receptor 2 (HER2): ER+, HER2+ and triple-negative breast cancer (TNBC). These clinical subtypes have distinct biological characteristics and different responses to treatment (4).

The expression of hormone receptors is routinely tested by immunohistochemistry on FFPE tissue samples. Any nuclear staining in >1% of invasive tumour cells is considered positive for ER and PR expression. HER2 status is assessed using immunohistochemistry and fluorescence *in situ* or chromogenic *in situ* hybridization. Approximately, 70% of breast cancers are ER+, while HER2+ tumours comprise around 15-20% of cases, and can be ER+ or ER- (4). The remaining cases are termed triple negative due to their lack of any of these three markers.

Proliferative rate is a key predictor of the risk of relapse and response to chemotherapeutic agents and is measured in several ways, including estimation of mitotic rate and detection of Ki-67 protein by immunohistochemistry (IHC) (20). However, Ki-67 is only informative for ER+, HER2- breast cancers. Due to their almost universally high proliferative rate, HER2+ breast cancers and most TNBCs are not distinguishable by proliferation markers and virtually all require chemotherapy. Thus Ki-67 does not contribute to therapeutic decision-making in these subtypes. While there are active efforts to better define recommended protocols for the detection and reporting of Ki-67 in breast cancer (21), there is currently a lack of standardisation for determining Ki-67 levels (22). The commonly used threshold is 15% stained nuclei in invasive tumour cells; less than 15% indicates low proliferation, while over 30% indicates high proliferation (20).

TNBC is known to have an aggressive clinical course associated with younger age, higher grade at diagnosis, and overall poor prognosis and accounts for a disproportionate number of breast cancer-related deaths (4, 23). Due to the lack of targetable receptors, treatment options are difficult for TNBC. However, gene expression studies have identified subtypes of TNBC, which may have prognostic and therapeutic implications although their clinical utility remains to be proven (**1.2.6.1 Molecular subtyping of TNBC**).

1.2.2 Tumour size, nodal status, and distant metastasis staging

The degree of dissemination of breast cancer at the time of diagnosis is a strong determinant of outcome. Consequently, breast cancer staging is a critical prognostic factor, which is considered at treatment planning (24). Early-stage breast cancer has 10-year survival rates of over 90%. In contrast, metastatic breast cancer has 5-year survival rates of only around 25% and a median overall survival of approximately 2 years (25). This difference in the outcome is observed regardless of the histological type, grade, or molecular features of the disease. However, due to its commonly more aggressive character, TNBC is more frequently diagnosed at a more advanced stage and therefore is over-represented in the poor outcome group. The most commonly used staging system for breast cancer is the Tumour Node Metastasis (TNM) staging system which classifies cancers based on the size of the tumour, lymph node involvement, and distant organ spread (**Table 1.2**) (26). In addition to these factors, specific clinical presentations such as inflammatory carcinoma, skin ulceration and chest wall infiltration are also recognised for their distinct behavioural patterns within breast cancer, impacting clinical management (24). All these factors are combined to stratify the disease into one of five stages: 0, I II III IV (**Table 1.3**).

Table 1.2 TNM classification(27)

Tumour Size	
T0	No primary tumour
Tis	Tumour only in breast ducts or lobules
T1	<3 cm
T2	>2 - <4 cm
T3	>5 cm
T4	Tumour of any size with extension to chest wall/skin (ulceration or skin nodules)
Lymph Node Status	
N0	No lymph node metastases
N1	Metastases in 1-3 axillary lymph nodes
N2	Metastases in 4-9 axillary lymph nodes
N3	Metastases in infra or supraclavicular lymph nodes or >9 axillary lymph nodes
Distant Metastasis	
Mx	Metastasis can't be measured or found
M0	No distant metastases
M1	Distant metastases found

Tis = Tumour *in situ*

Table 1.3 Breast cancer stages based on the TNM classification(27)

Overall Breast Cancer Stage	T category	N category	M category
Stage 0	Tis	N0	M0
Stage I	T1	N0	M0
Stage IIA	T0	N1	M0
	T1	N1	M0
	T2	N0	M0
Stage IIB	T2	N1	M0
	T3	N0	M0
Stage IIIA	T0	N2	M0
	T1	N2	M0
	T2	N2	M0
	T3	N1	M0
	T3	N2	M0
Stage IIIB	T4	Any N	M0
Stage IIIC	Any T	N3	M0
Stage IV	Any T	Any N	M1

Tis = Tumour *in situ*

1.2.3 Intrinsic molecular subtype

The advent of high throughput genomic profiling approaches dramatically expanded the understanding of the inherent characteristics of individual breast tumour tissues and revealed significant correlates to clinical behaviour that could be exploited to guide treatment. Several classification systems have been proposed for breast cancer subtypes, but the best recognised of these is the intrinsic subtypes discovered by Sorlie, Perou and colleagues, which have been adopted into clinical practice (28, 29). The intrinsic molecular subtype of breast cancer refers to the categorisation of breast tumours based on their gene expression profiles, which provides insight into their underlying biology and potential therapeutic strategies (4). Four distinct intrinsic molecular subtypes of breast cancer were identified: luminal A, luminal B, HER2-enriched and basal-like, each with unique clinical characteristics and treatment responses (**Figure 1.1H**) (28, 29). These molecular subtypes also demonstrate the relevance of immunophenotypic classification by hormone receptors and HER2 status where there are strong correlations. ER expression stratifies breast cancers into 2 distinct clusters: ER+ and ER−, which are reflected in the intrinsic molecular subtypes. Luminal A and B subtypes are enriched with ER+ cancers, whereas HER2-overexpressing and basal-like are ER-. Apart from the expression of ER and ER-related genes, different subtypes show differential expression of genes related to proliferation, HER2 expression, and cancer stem cell characteristics. These different intrinsic subtypes reveal differences in incidence, clinical and pathological features, and largely overlap with the established clinical and histopathologic classifications (4).

Luminal A tumours account for approximately 50% of cases and are characterized by higher expression of hormone receptors (ER or PR or both) (30). They express a high luminal gene signature that includes ESR1, GATA3, XBP1 and FOCA1, are generally HER2 negative and

express low levels of Ki-67 protein, indicating a lower proliferation rate. They are typically low-grade tumours and have better relapse-free survival and overall survival post-treatment, compared to other breast cancer subtypes (4). They respond better to endocrine therapy such as tamoxifen or aromatase inhibitors and generally do not benefit from cytotoxic chemotherapy (24).

Luminal B tumours also express hormone receptors but have higher proliferation rates and their HER2 status is variable; they can be either HER2+ or HER2-. They account for 20-30% of cases and are higher grade associated with a higher risk of relapse. A study of the key distinguishing features of luminal A and B tumours revealed PR expression and proliferation as fundamental classifiers, with luminal A tumours generally expressing high PR and low proliferation markers, and conversely luminal B tumours being low for PR and more highly proliferative (31). Clinically, these tumours often require more aggressive treatment strategies, including chemotherapy in addition to hormone therapy.

HER2-enriched breast cancer accounts for 15-20% of breast cancer cases. They are characterized by overexpression of HER2, often with low expression of both ER and PR. These tumours highly express proliferation-related genes and are associated with a poorer prognosis compared to luminal subtypes. Importantly, while approximately 70% of HER2+, ER− tumours are HER2-enriched, small subsets of HER2+ tumours also exist within the basal-like and luminal B subtypes (4). The HER2-enriched subtype is characterised by higher mitotic activity, increased invasiveness, and cellular motility, leading to earlier and more frequent relapses of the disease after primary treatment (4).

Basal-like breast cancer accounts for approximately 15% of breast cancer cases. This subtype of breast cancer is highly proliferative and is characterised by high expression of basal markers such

as cytokeratins 5/6 and EGFR and low expression of luminal A signature genes (4). These tumours are typically hormone receptor and HER2 negative, and are associated with a poor prognosis due to their aggressive behaviour and higher likelihood of metastasis (4). While approximately 80% of basal-like tumours are triple negative, the terms are not interchangeable. Interestingly, although TNBC patients have a higher likelihood for early relapse, compared with luminal tumours, an estimated 40% of patients with basal-like disease display high sensitivity to chemotherapy and high rates of pathological complete response, indicating the complete eradication of tumour cells (32). For the TNBC cases that respond well to systemic treatments in the first two years after diagnosis, the long-term outlook becomes as good or better than the other subtypes.

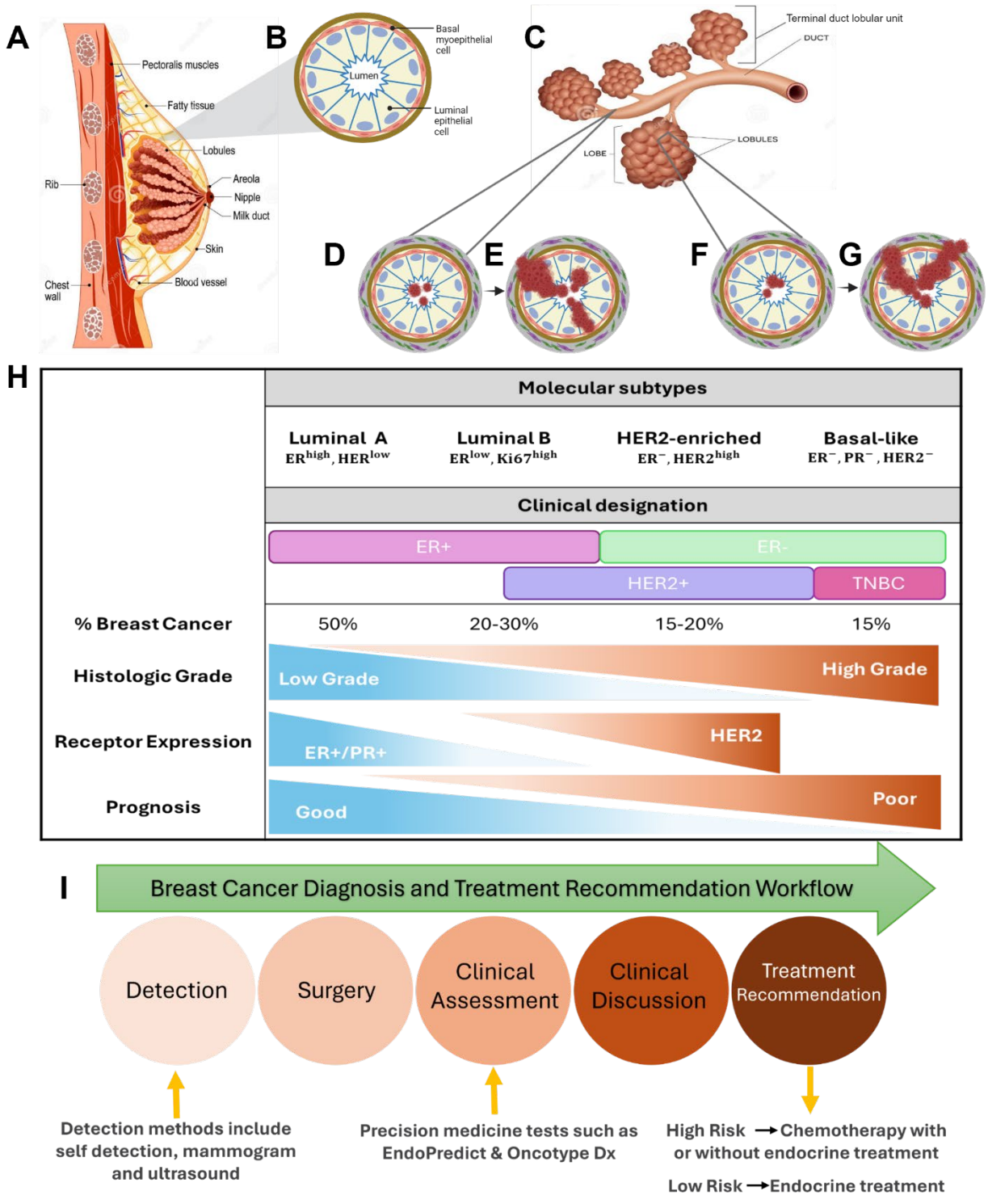


Figure 1.1: Normal Breast structure and breast cancer subtypes. (A)Schematic representation of the human breast, showing the lobular and ductal system. (B) Cross-section view of the normal mammary duct showing the basal myoepithelial cells and the luminal epithelial cells. (C) Detailed magnified view of the milk duct illustrating the lobe, lobules, duct and the terminal ductal lobular unit, the functional unit of the breast where most tumours arise. (D) Representative images of ductal carcinoma in situ and (E) invasive ductal carcinoma. (F) Representative images of lobular carcinoma in situ and (G) invasive lobular

carcinoma. (H) Molecular breast cancer subtype summary. (I) Flowchart of breast cancer diagnosis to treatment recommendation. Created with BioRender.com

1.2.4 Current clinical management and limitations

Current clinical management of breast cancer involves, firstly the detection of the tumour in the breast and the diagnosis of malignancy and the extent of disease by mammography, ultrasound, core biopsy examinations, and wider imaging if dissemination is suspected (33, 34). Once the disease is diagnosed, the tumour is then surgically removed via breast-conserving surgery or mastectomy, which involves the complete removal of the breast epithelium (33, 34). Additionally, some or all of the axillary lymph nodes are removed, to allow assessment of tumour spread and also to limit local recurrence. The tumour and lymph node tissues are transferred to the clinical pathology department for assessment of disease features, including tumour size, grade, lymph node involvement, receptor levels and proliferative markers such as Ki-67.

This clinical assessment is reported to the multi-disciplinary clinical oncology team that recommends appropriate treatments based on these results. Very often, radiation and adjuvant therapy are recommended after surgery to prevent recurrence and to eliminate breast tumour cells that might have spread (34). Examples of adjuvant therapy are cytotoxic chemotherapy, endocrine therapy, and other targeted therapy (34, 35). While, for some cases, the potential for the disease to recur is clearly low or high, in a subset of 30-50% of cases the treatment recommendation is difficult because the risk of recurrence is equivocal based on traditional markers. In those cases, the Ki-67 positivity, which predicts proliferation level is important (36). Unfortunately, this marker is subject to considerable inter-observer and between facility variation (36). So, a more accurate test that returns an objective numerical output is valuable in these cases.

Although adjuvant therapy improves overall survival, lengthens the disease-free interval of patients with early breast cancer and prevents metastatic recurrence, the patient's quality of life is severely affected. Side effects include nausea, vomiting, cognitive dysfunction and fatigue (34, 37). In fact, only 40% of breast cancer patients administered adjuvant treatment, relapse and die of metastatic breast cancer (35). Thus, indicating many patients are over-treated and suffer unnecessary side effects. Only 15% of patients treated with tamoxifen after surgery will have distant recurrence, indicating that 85% of patients will be overtreated if chemotherapy is recommended (34, 38). Moreover, in cases where the proliferative rate of the tumour is low, chemotherapy provides little or no benefit since it targets dividing cells. This has been clearly demonstrated in early neoadjuvant chemotherapy trials that demonstrated negligible tumour shrinkage in low proliferation tumours (39). One solution to overcome overtreatment is through identifying patients with high and low risk of metastasis using prognostic and predictive biomarkers.

1.2.5 Commercial multi-gene tests for more precise risk stratification

The traditional clinicopathological characteristics currently used in standard practice can predict prognosis and likelihood of response to systemic therapies in cases where the likely disease course is clear (4, 40). However, for luminal breast cancer, a spectrum of outcomes and clinical responses can be expected, and there is uncertainty about the best treatment approach as clinicians aim to prevent both over-treatment and under-treatment (41) . Therefore, a molecular analysis of cancer has greater power to highlight its genetic heterogeneity and enhance clinical management by accurately predicting prognosis and treatment response in this broad group representing three-quarters of breast cancers (29, 42).

The power of genome-wide breast tumour gene expression profiles to discriminate cases into subgroups with clearly different biology and clinical outcomes was first identified in academic research. However, it was quickly recognised that this information could be commercially exploited to produce precision medicine products that could guide clinical decision-making much more accurately than traditional pathology approaches. In the case of breast cancer, the greatest discriminatory power from gene expression profiles is derived for separating luminal A and B breast cancers. As a result, companies worldwide have utilized molecular profiling findings in breast cancer to develop multi-gene tests that can assist in refining therapy selection for breast cancer patients with ER+, HER2- luminal early breast cancer. These tests yield a numerical value that is correlated with a defined per cent risk of recurrence in five years if treated with endocrine therapy alone, and identify patients who are at elevated risk of recurrence and likely to benefit from treatment escalation to adjuvant chemotherapy (43). Thus, these multi-gene tests can also help identify which ER+ breast cancer patients can safely avoid potentially damaging chemotherapy. Consequently, there are various personalized multi-gene tests available globally aimed at guiding the treatment of this type of breast cancer (4). The three most commonly applied tests are Oncotype DX, EndoPredict and Prosigna.

1.2.5.1 Oncotype DX

Oncotype DX, a 21-gene signature (16 outcome-associated and 5 housekeeping genes), was developed to predict the likelihood of distant metastases in node-negative, endocrine therapy-treated breast cancer patients. The assay uses a reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) in tumour RNA derived from FFPE tissue (38). The test provides a recurrence score (RS) which predicts the likelihood of breast cancer recurrence within 10 years of the initial diagnosis and reports on chemotherapy benefits. The RS classifies patients into three risk groups:

low risk (0-15), intermediate risk (16-25), and high risk of recurrence (> 25). Extensive studies have demonstrated benefits from adjuvant chemotherapy in patients with a high recurrence score and benefits from adjuvant endocrine-only therapy in patients with a low recurrence score (38, 44-46). The TAILORx nine-year follow-up phase III clinical study involving 10,273 women with ER+, HER2– breast cancer showed that patients with an intermediate risk score who were randomly assigned to either endocrine therapy alone or endocrine therapy plus chemotherapy had similar rates of disease-free survival (83.3% for chemo-endocrine vs 84.3% for endocrine therapy alone), suggesting that they may be able to avoid chemotherapy (47-49). The only exception to this was the minor subset of patients in this group who were 50 years of age or younger, where a small benefit from the addition of chemotherapy was observed (46). Although Oncotype DX was initially designed and validated only for the assessment of recurrence risk in node-negative disease, the subsequent RxPONDER trial of 5083 women (33.2% pre-menopausal and 66.8% post-menopausal) with one to three positive lymph nodes, also demonstrated no difference in disease-free survival at five years for post-menopausal women treated with endocrine-only (91.9%) or chemo-endocrine therapy (91.3%) in the RS ≤ 25 group, but a modest benefit for pre-menopausal women (89.0% disease-free survival with endocrine-only and 93.9% with chemo-endocrine therapy) demonstrating the utility of this assay for all women with ER+ breast cancer (50).

1.2.5.2 EndoPredict

EndoPredict is based on a twelve-gene molecular signature panel, of which eight are cancer-related genes and four are housekeeping genes. The assay can be used to predict the risk of recurrence in patients with early-stage, ER+, HER2− breast cancer with up to 3 positive lymph nodes. The assay calculates the EndoPredict score (EP score) which is a weighted linear algorithm that combines the normalised expression levels of the eight test genes assessed by quantitative RT-qPCR in FFPE-

derived total RNA. The combination of the EP score with two clinical risk factors – nodal status and tumour size – is used to calculate the EPclin score (51). The score predicts the risk of recurrence of breast cancer over 10 years, with an EPclin cut-off at 3.3 which is equivalent to a 10% risk of recurrence if treated for five years with endocrine therapy alone (52). The inclusion of other clinical information such as the tumour grade, ER status, and age at diagnosis increases the accuracy of the score, providing better risk prediction and thus allowing for a more informed decision about chemotherapy choice (53, 54). EndoPredict was retrospectively validated in 1702 cases of ER+/HER2- breast cancer, treated only with endocrine agents, from the ABCSG-6 (tamoxifen arm) and ABCSG-8 clinical trials (55).

1.2.5.3 Prosigna

Prosigna a 50-gene molecular test based on the PAM50 gene signature that classifies breast cancers into their intrinsic subtypes (56, 57). The test was developed to estimate the risk of distant recurrence within 10 years of diagnosis of ER+ breast cancer with up to 3 positive lymph nodes in postmenopausal women (58, 59). The assay uses the NanoString nCounter platform on frozen or FFPE tissue-derived total RNA (60, 61). The assay can predict the risk of distant recurrence after 5 years of endocrine therapy and the benefit of extending endocrine therapy beyond 5 years (62). Similar to Oncotype DX, the assay computes Prosigna risk of recurrence scores that divide patients into low-, intermediate-, or high-risk groups. Recently, the assay was reported to be prognostic for long-term breast cancer survival, irrespective of menopausal status. Furthermore, compared with Oncotype DX, fewer patients fall into the intermediate group in this test, therefore offering a better risk stratification (63). The American Society of Clinical Oncology guidelines recommend chemotherapy for patients with a high Prosigna risk of recurrence, and no chemotherapy is recommended for patients with a low Prosigna risk of recurrence, as they are predicted not to

benefit from it (40). The OPTIMA (Optimal Personalized Treatment of Early Breast Cancer Using Multi-Parameter Analysis) de-escalation, partially blinded, clinical trial is currently underway to prospectively validate Prosigna for tailoring adjuvant treatment of luminal breast cancers (64).

1.2.6 The challenge of classifying TNBC

TNBC is one of the most aggressive subtypes and often occurs in younger patients. It is linked to poor prognosis and limited treatment options due to the absence or low expression of ER, PR, and HER2 (**Figure 1.1**). TNBC may be considered a single disease; however, clinical, histological, and molecular profiling show its intrinsic heterogeneity (4, 65). This variability leads to differences in treatment response and overall survival among TNBC patients. While chemotherapy is typically recommended for TNBC patients, their responses vary greatly. Approximately 30-40% of patients with TNBC treated with chemotherapy achieve a pathological complete response. However, the rest do not respond well to chemotherapy treatments and have higher chances of disease recurrence with poorer outcomes (66, 67). These non-responders endure multiple cycles of chemotherapy without significant benefits, emphasizing the need for predictive biomarkers that can prevent unnecessary toxicity and costs by better stratifying cases according to clinical outcomes. Therefore, effective classification systems highlighting targetable differences are urgently required to improve the treatment outcomes for TNBC patients by better understanding the tumour heterogeneity.

1.2.6.1 Molecular subtyping of TNBC

Gene expression profiling has emerged as an important tool for stratifying TNBC into biologically different subgroups. However, defining subtypes that are clinically different in terms of response

to treatment and disease recurrence has been challenging. Lehmann et al. identified six different TNBC gene expression profile subtypes (TNBCtype-6 classification) from 587 TNBC cases identified in 21 gene expression data sets, performing k-means clustering on the most differentially expressed genes (68), The subtypes were named according to their expression patterns; the basal-like (BL) subtypes, which were divided into BL1 and BL2, immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), and luminal androgen receptor (LAR) (11). These subtypes were characterized based on gene ontologies and differential gene expression that cluster with genes implicated in specific pathways. The gene ontologies for the BL1 subtype are enriched in DNA damage response (ATR/BRCA), cell cycle regulation and proliferation pathways (MYC, NRAS, AURKA). The BL2 subtype, which is enriched with genes involved in growth factor signalling and metabolic pathway activity, shares a highly proliferative phenotype. The IM subtype displays gene ontology enrichment involving immune cell processes, while mesenchymal-like TNBC subtypes, M and MSL observe gene ontologies related to cell motility and cell differentiation pathways. Unique to the MSL subtype are genes enriched in angiogenesis and genes associated with quiescent stem cells (ABCA8, PROCR and ENG). The LAR subtype is heavily enriched in hormonally regulated pathways with elevated mRNA and protein levels of androgen receptor (AR), and 82% of genes in the LAR signature are also contained within the luminal A or luminal B intrinsic subtype signatures. Five years later, the group refined the classification after observing that the IM and MSL subtypes mainly reflected the influence of immune infiltration and high stromal fibroblast content, respectively, rather than the gene signatures arising from the tumour cell component. Thus, the TNBC subtypes were refined as BL1, BL2, M, and LAR (TNBCtype-4 classification) (69, 70).

Complementing the Lehmann subtyping is the work of Burstein and colleagues. They identified four TNBC subtypes based on copy number variation analysis and gene expression profiling techniques: LAR, MES (mesenchymal), BLIS (basal-like immunosuppressed) and BLIA (basal-like immune activated) (71). These subtypes were identified using a reduced 80-gene RNA-based classifier and were shown to differ in survival and to exhibit distinctive patterns of copy number variations, gene expression and pathway enrichment. Among these four subtypes, the prognostic analysis showed that disease-free survival (DFS) was in the order BLIA > M > LAR > BLIS ($p = 0.019$) and disease-specific survival (DSS) was BLIA > M > LAR > BLIS ($p = 0.07$) (71, 72).

A study by Jézéquel et al. identified three distinct subtypes (C1, C2 and C3) through transcriptomic profiling using a fuzzy clustering strategy for prognosis (73). The C1 cluster is correlated with the molecular apocrine phenotype and has a favourable prognosis. Meanwhile, both C2 and C3 are enriched in basal-like genes but exhibit contrasting immune responses: the C2 cluster is associated with an immunosuppressive response, while the C3 cluster is linked to an adaptive immune response (74).

1.3 Prognostic tests guiding treatment of ER+ breast cancer

1.3.1 Current commercial multi-gene tests are expensive

In clinical practice, the key question is the identification of patients who are at elevated risk of recurrence, and discrimination between patients who will or will not benefit from therapies. By using molecular assays, more patients can be spared adjuvant chemotherapy, and by defining objective numerical outputs, with clear cut-offs, these assays can overcome some of the individual variations in pathology assessment, as well as subjectivity that comes with trying to interpret all the histopathological factors together. However, despite the advantages of utilising these multi-gene assays to predict the risk of relapse over traditional prognostic tools, the uptake of these assays in Australia has been limited by their high associated cost (**Table 1.4**). To date, the Oncotype DX test is the most widely used commercial multi-gene test in Australia, likely due to its earlier establishment, subsidisation by some private insurers and greater uptake internationally. However, due to the high cost of commercial tests and limited public or private subsidy, only a minority of breast tumours are assessed using multi-gene tests in Australia. In addition to their high cost, not all tests are locally available, often delaying treatment. Specifically, for Oncotype DX, the FFPE tissue blocks must be submitted to a centralised testing facility in the United States for processing, resulting in significant delays. Therefore, discovery and development of an affordable alternative multigene test to predict the risk of recurrence in ER+ breast cancer will be a breakthrough in Australia.

Table 1.4 Comparison between PROSPER and commercially available multi-gene tests

	Oncotype DX	EndoPredict	Prosigna	PROSPER
Submitted material				
Methodology	RT-qPCR[a]	RT-qPCR	NanoString nCounter	RT-qPCR using PlexZymes
Facility accessibilty	Genomic Health USA	Australian Clinical Labs Victoria	Selected private pathology labs	Any local public or private pathology
Turnaround Time	2-3 weeks	4-5 business days	10 business days	2-3 days
Subsidised by Medicare	No	Partial	No	Not applicable[b]
Cost/test (AU$)	5,000	2980	3,300	< 200

[a] RT-qPCR: reverse transcription and quantitative PCR
[b] PROSPER is not yet approved for clinical application

1.3.2 The prognostic power of multigene tests reflects proliferative activity

There is objective evidence that the predictive power of multigene tests derives from their ability to estimate proliferative rate. In fact, the prognostic power of these multi gene tests is lost if genes that are correlated with proliferation are excluded. Venet and colleagues revealed that the more a gene signature was correlated with proliferation-associated genes, the stronger its association with patient outcome (**Figure 1.2**) (75). The group developed a 131-gene panel, which they termed meta-PCNA, comprising the top 1% of transcripts most positively correlated with the proliferation marker, proliferating cell nuclear antigen (PCNA), across a compendium of 36 normal tissues, and demonstrated that the meta-PCNA signature performed equally well or better than many of the published gene signatures in predicting breast cancer outcome (75). Moreover, when meta-PCNA was used to adjust for the influence of proliferation genes in all other signatures, most breast cancer

predictive signatures lost their prognostic power suggesting that the proliferation genes were the major drivers predicting outcome in these tests (75).

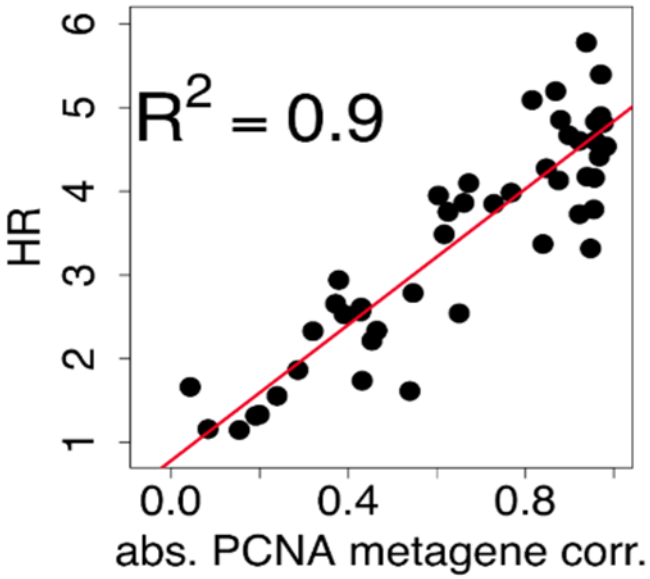


Figure 1.2: Proliferation reflects the power of multigene tests. Each point corresponds to a signature (75). The x-axis denotes correlation with the meta-PCNA proliferation signature and the y-axis denotes the hazard ratio of the individual signature for prediction of breast cancer recurrence.

The number of genes contained in meta-PCNA was arbitrarily chosen. While proliferation genes were demonstrated to drive the power of predictive signatures, the redundancy of measuring multiple targets evaluating a single function would suggest that the level of significance achieved would rise the most steeply at low numbers of geneset members then reach a plateau. Because there is an immediate need for a more cost-effective assay for accurately predicting the risk of recurrence in Australia, the PROliferation Signature for Prognosticating ER+ breast cancer (PROSPER) test was developed in our lab by determining the minimum subset of genes derived from meta-PCNA that accurately classified ER+, HER2– breast cancers into good and poor outcome subgroups. Performed as an RT-qPCR assay, with optimal automation, it is hypothesized that the reagent and labour costs for an in-house gene signature-based prognostic test are comparable to the current immunohistochemical test for the proliferation marker Ki-67, which is

delivered in Australia for under $200. An objective of developing a new gene signature-based predictive test is that it would replace the Ki-67 test with a more accurate molecular test for a comparable price. The PROSPER gene signature was composed of five gene targets along with 5 housekeeping genes.

The individual prognostic power of multiple combinations of proliferation-associated genes was investigated *in silico,* in the ER+, HER2- breast cancer cases from the publicly available Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) gene expression dataset of over 2000 tumour samples, to identify the minimum number of predictive genes that could be grouped into a signature to predict risk of recurrence with a high level of statistical power. Given the large size and statistical power of the source cohort, the dataset was divided into equally sized discovery and validation cohorts, and the PROSPER signature was then designed and tested, by estimating individual Cox regression p values for correlations between expression of each gene with disease-specific survival, then performing elastic net regularisation to find the optimal minimal gene list that predicted outcome without loss of statistical significance (**Figure 1.3**). After correlation-based weighting, the gene expression results were combined with key clinical data to derive a PROSPER index, which was positively correlated with proliferative rate and disease recurrence. In fact, in the validation dataset, PROSPER performed as well or better than all commercial gene signature-based tests assessed, when cases were stratified into predicted low and high-risk groups, as for EndoPredict, and after tertiary stratification as for Oncotype DX (**Figure 1.4**).

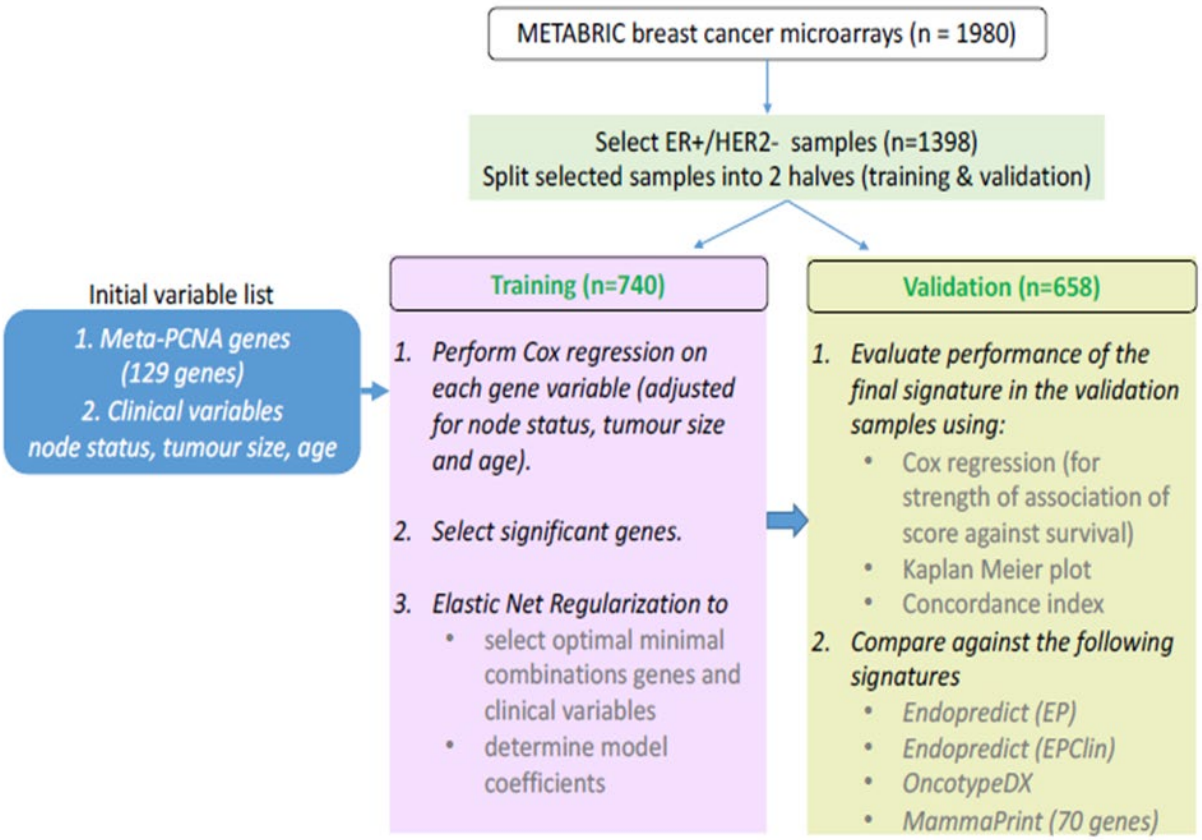


Figure 1.3: Flowchart of the discovery and validation of PROSPER using the METABRIC cohort. Gene expression data of tumours from 1398 patients with ER+ breast cancers were randomly split into two groups (training = 740, validation = 658). Cox regression was performed on the training set to select genes that were significantly associated with outcome. Elastic net regularisation was utilised to select the optimal minimal gene and clinical variable combination as well as to determine model coefficients. The performance of the PROSPER signature was then evaluated in the validation cohort and compared against other multigene signatures, including those signatures underpinning the leading commercial tests.

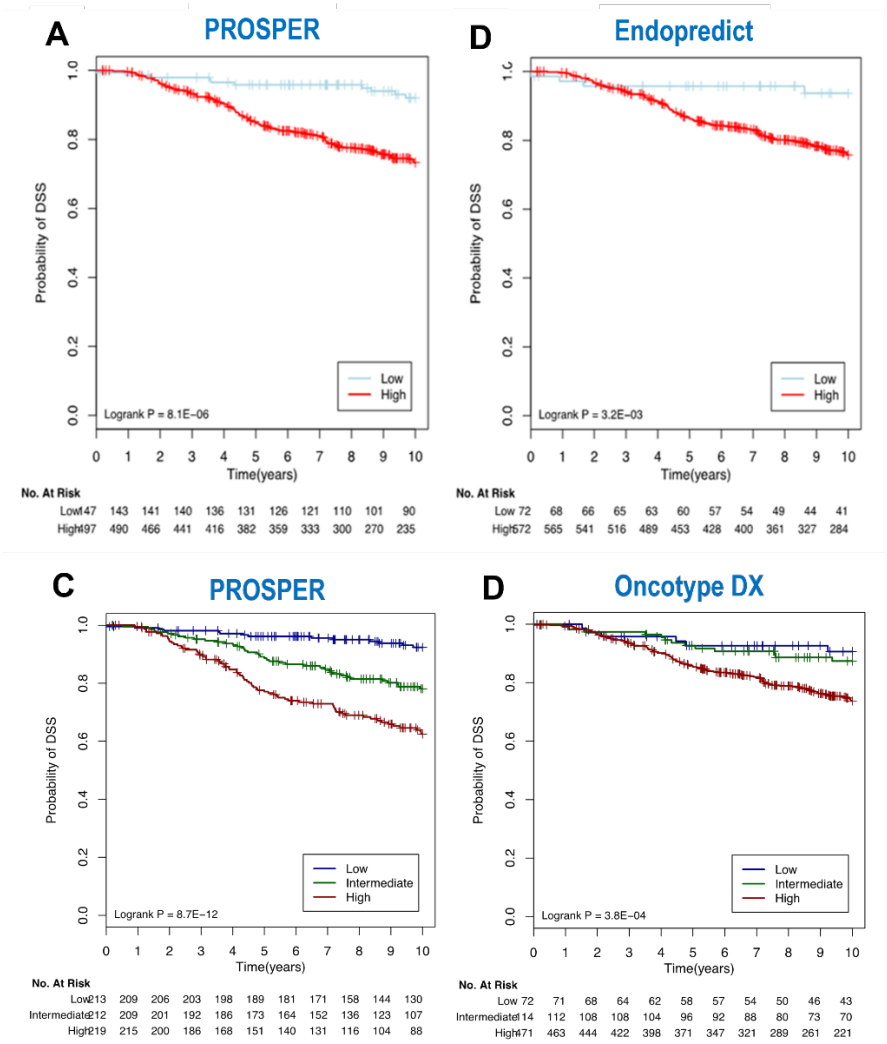


Figure 1.4: Performance of the PROSPER signature in the validation cohort of the ER+ METABRIC dataset. Kaplan-Meier survival plot of METABRIC validation cohort samples stratified into low and high (A) PROSPER index with binary stratification. (B) EndoPredict EPclin score, (C) PROSPER index with tertiary stratification, (D)

The PROSPER test was then validated in a cohort of patients, treated at Westmead Breast Cancer Institute, Westmead Hospital, who had been recruited to receive EndoPredict testing as a part of a study assessing the influence of molecular testing on adjuvant treatment recommendation (76). After EndoPredict testing, the PROSPER gene signature was measured in the same residual RNA using the NanoString nCounter platform. The PROSPER index was correlated with EPclin (**Figure 1.5A**) and was statistically significantly different between groups stratified into EPclin low and

high risk (**Figure 1.5B**), suggesting that it could be a substitute for the EndoPredict test. The correlation between the PROSPER index and EndoPredict EPclin values (R^2=0.61, **Figure 1.5**A), was comparable to the correlation between PROSPER and Ki-67 hotspot estimate (R^2=0.60, **Figure 1.5**C) and better than the correlation between EndoPredict EPclin and Ki-67 hotspot (R^2=0.55, **Figure 1.5**D).

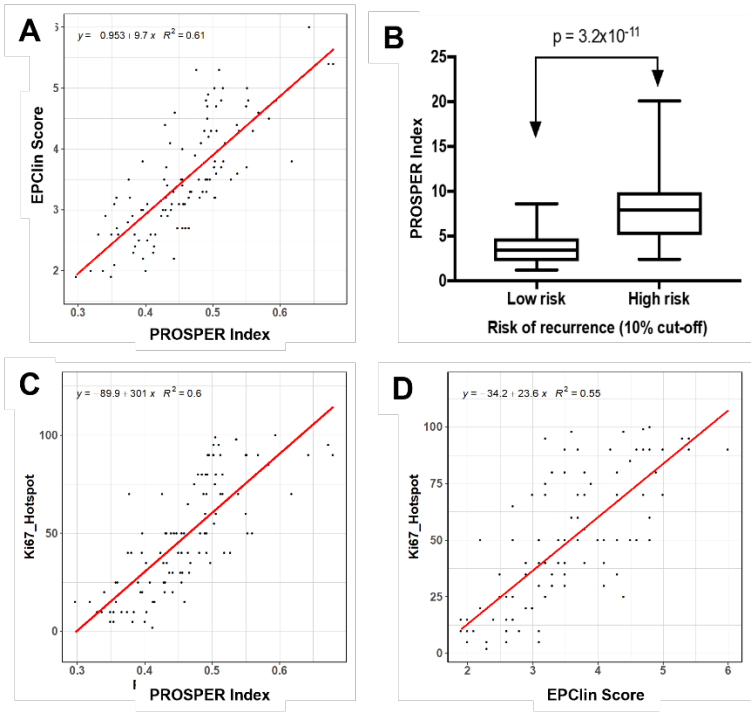


Figure 1.5: A Prospective study that compares the PROSPER index and EndoPredict score (EPclin) in a cohort of 130 patients. A) Correlation between PROSPER index and EPclin score. B) PROSPER index values in breast cancer cases stratified into low and high EPclin scores, based on the assay cut-off. C) Correlation between PROSPER index and Ki67 hotspot estimate. D) Correlation between EndoPredict EPclin and Ki67.

Since PROSPER needs to be translated to clinical application, the next step involved determining the correlation between PROSPER and outcome in a cohort of breast cancer patients with long-term follow-up. We measured PROSPER in a retrospective cohort of 142 breast cancers with

known long-term outcomes. The PROSPER index accurately stratified cases into low and high risk of recurrence in both disease-specific survival and overall survival (**Figure 1.6).**

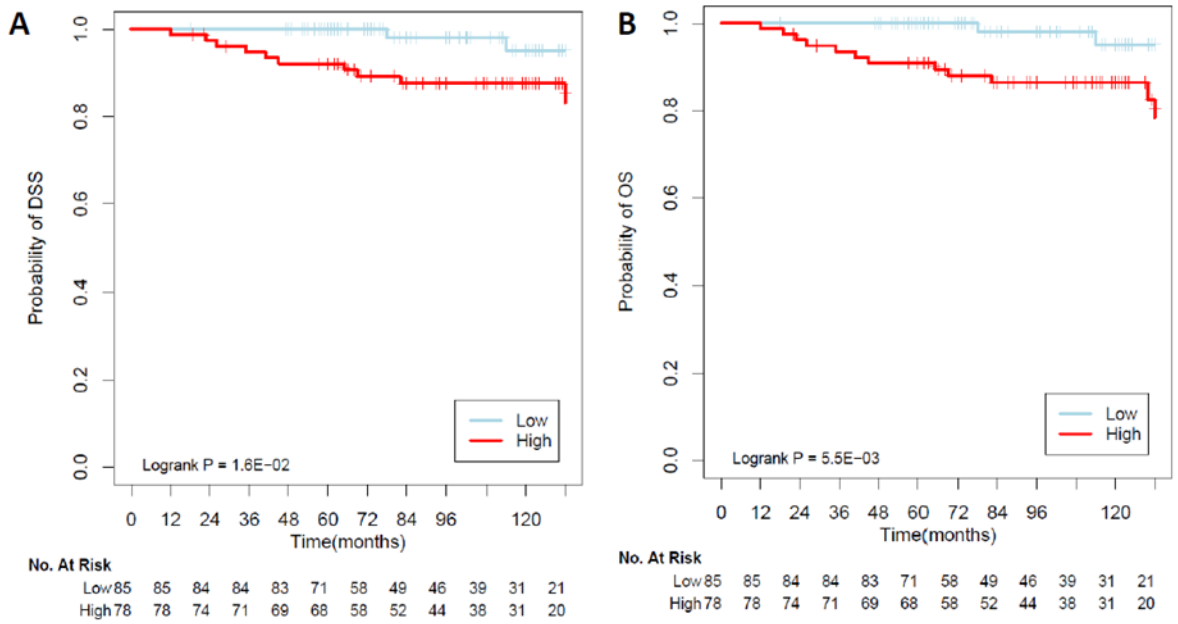


Figure 1.6: Association between PROSPER index and survival in a retrospective study with 142 samples with extensive follow-up data. Kaplan – Meier plots reveal statistically significant differences in (A) disease-specific survival and (B) overall survival between cases stratified into low or high PROSPER index.

The demonstration that PROSPER performed as well or better than existing commercial tests for predicting risk of recurrence in ER+ breast cancer retrospective data, suggests that the signature could be used as an alternative, more affordable assay, with equivalent sensitivity and accuracy to existing tests. The inclusion of target genes across a broad dynamic range of expression levels accounts for the high sensitivity and accuracy of the assay.

1.3.3 PROSPER signature with the NanoString nCounter platform.

For all validation studies performed to date, the PROSPER signature was measured using the NanoString nCounter platform. The NanoString assay is based on direct transcript counting, by using multiplexed, colour-coded probe pairs (**Figure 1.7**) (77). Total RNA was extracted from FFPE samples and mixed with a set of nucleotide probes targeting the genes from the PROSPER signature.

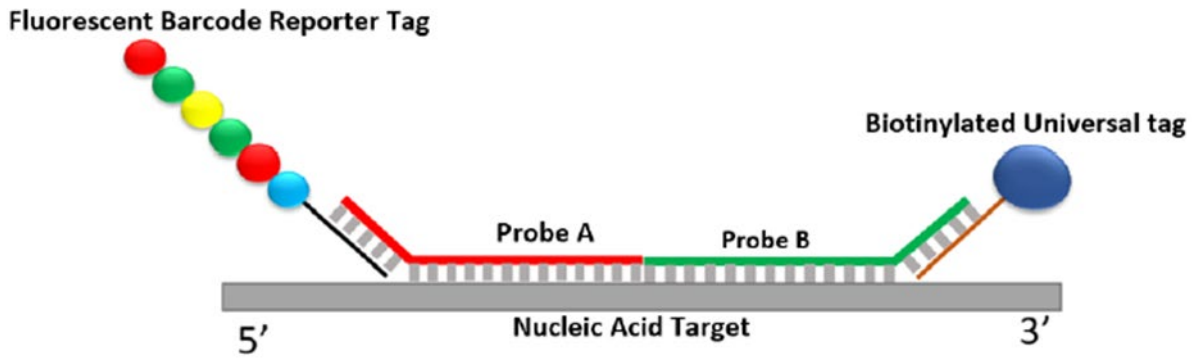


Figure 1.7: NanoString Assay Chemistry. The uniquely barcoded fluorescently labelled reporter tag hybridises to a pair of sequence-specific oligonucleotide probes along with a universal tag. The hybridised universal tag anchors each target to the surface of a streptavidin-coated slide to immobilise the target regions. Finally, the immobilised complex is imaged using the NanoString nCounter, allowing for single molecule counting and data analysis.

A strength of the NanoString platform is that it allows for the accurate quantitation of gene expression levels in highly degraded archival tissues (78, 79). The assay directly quantitates gene expression through a hybridization-based method and does not require reverse transcription of mRNA and subsequent cDNA amplification. This leads to the absence of amplification bias and control samples are not required, since absolute transcript abundance is determined for each single sample and normalized against the expression of housekeeping genes in that same RNA (80, 81). Additionally, the NanoString nCounter platform allows for highly specific multiplexing and molecular counting of up to 800 different targets in a single sample with a high signal-to-noise ratio. This is partly due to the two different probes required for this quantification, which decrease the odds of non-specific detection of gene expression (79). In fact, Prosigna, a commercial multigene test, employs the NanoString nCounter platform to evaluate patients' risk of breast cancer recurrence (82). Even though, the platform was selected because of its ease of use within the research setting and high signal-to-noise performance, the Prosigna assay is expensive, costing over $3000 per sample, and a major component of that cost derives from the specialized NanoString reagents and consumables. Therefore, in the clinical setting, the NanoString based test

is not accessible to all patients. Furthermore, because NanoString detects by reading the molecular barcodes attached to each probe, the results generated must be analysed using the NanoString platform's algorithm, which is not revealed to the user and restricted to the application through the NanoString platform, to protect their intellectual property. As a result, it is not possible to bioinformatically interrogate the rigour of the internal analysis (79). While the NanoString platform is suitable in the context of a private diagnostic laboratory, the high cost of the equipment limits the test from being implemented in smaller local facilities. These factors limit the availability of rapid and affordable molecular prognostic testing for breast cancer treatment.

1.3.4 qPCR as an affordable technique for routine clinical use

The two most frequently used breast cancer commercial tests in Australia, EndoPredict and Oncotype DX utilise an RT-qPCR assay to evaluate the expression of target genes (38, 51). The precedent created by Oncotype DX and EndoPredict demonstrates that optimizing the PROSPER test into an RT-qPCR assay is feasible. RT-qPCR is not only cost-efficient but has high sensitivity due to the exponential amplification of the template nucleic acid samples (83, 84). The assay can also be very specific when using gene-specific primers in the synthesis of cDNA and DNA amplification step. Moreover, the full workflow of sample to results, using RT-qPCR, can be completed in one to two working days (84).

Within the field of molecular analysis, the advent of quantitative polymerase chain reaction (qPCR) has been transformative. Since its development in the 1990s, qPCR swiftly emerged as the preferred analytical technique for detecting and quantifying low levels of DNA in a wide range of samples from many sources, such as tissues, blood, and cultured cells originating from humans, animals, plants, and bacteria (85). Its inherent versatility expanded to the quantification of RNA by incorporating a reverse-transcription step prior to qPCR. Renowned for its sensitivity,

specificity, and reproducibility, qPCR is extensively utilised in many studies for the discovery of biomarkers for applications in clinical studies such as evaluating the status of certain cancers or for monitoring disease progression or treatment responses (83).

The technique combines polymerase chain reaction (PCR) and fluorescent reporter chemistry (**Figure 1.8**) enabling real-time nucleic acid amplification and detection (84). The quantification of gene (or transcript) numbers is determined during the exponential phase of the PCR amplification when the numbers of amplicons detected are directly proportional to the initial numbers of target sequences present within the sample (86). Due to the closed nature of the reaction system, the exponential phase is short, resulting in the depletion of reactants, enzyme activity and other factors, while the products accumulate over time (86-88). Thus, a qPCR reaction can be divided into four reaction phases known as Baseline, Exponential, Linear and Plateau phases (**Figure 1.9**). The target specificity of any qPCR assay is determined by the design of the primers and in some cases an internal probe, allowing quantification of gene expression (86). Furthermore, the incorporation of fluorescence-based detection offers greater sensitivity and enables discrimination of transcript numbers across a wider dynamic range (86).

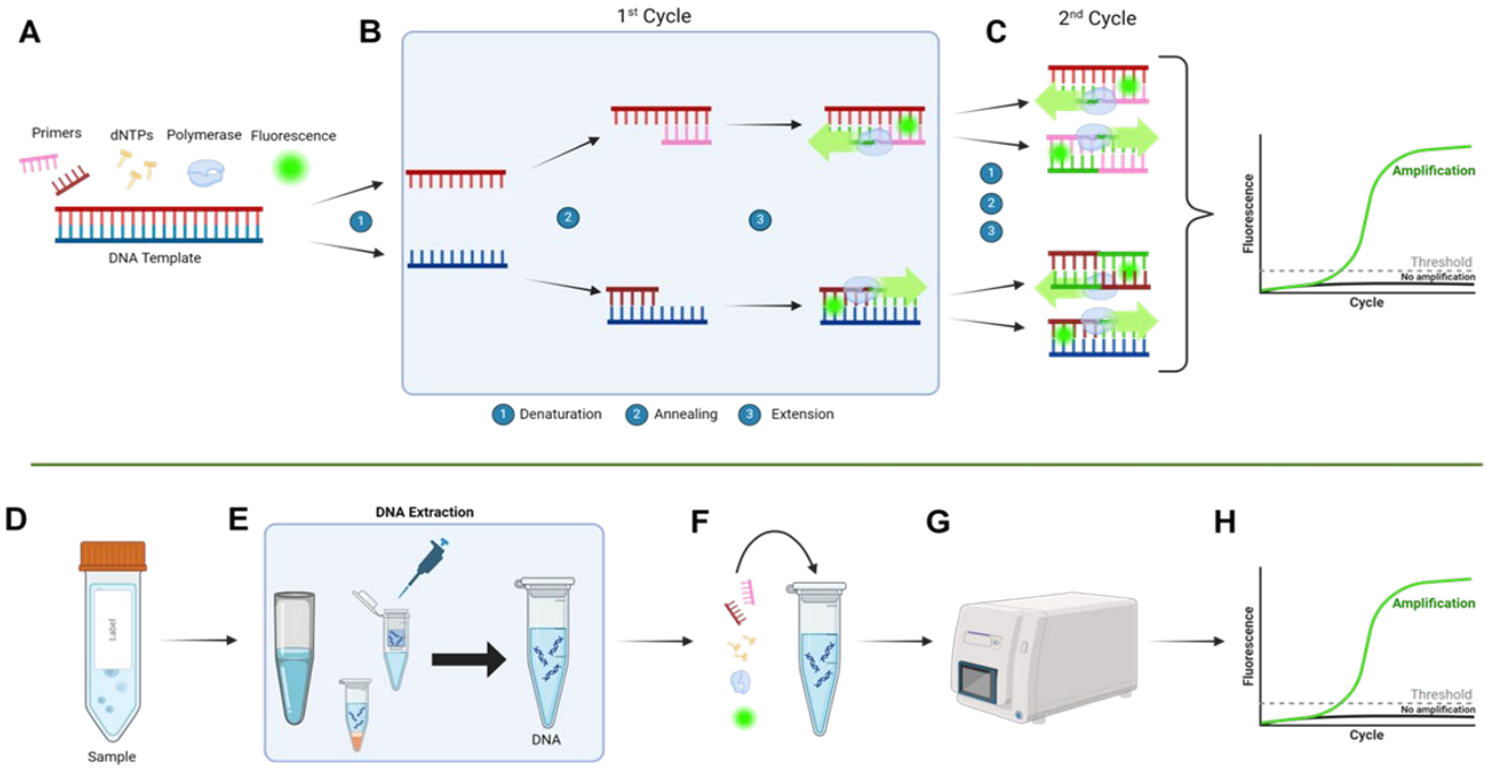


Figure 1.8: A qPCR workflow. (A) qPCR reactions contain oligonucleotide primers, dNTPs, polymerase and the fluorescent tag. (B) During the 1st cycle, the double-stranded DNA (dsDNA) is denatured by high temperature, followed by the annealing of the target specific primers to the single strand DNA (ssDNA). Polymerase will then synthesize the new strand of complementary DNA using the dNTPs for the extension process. The fluorescent tag will then bind to the dsDNA and fluoresce. Once cycle 1 is completed, (C) the DNA is again heated to denature, and the old and new strands are served as templated for DNA synthesis. The process of denaturation, annealing of the primer to the template, and DNA synthesis (extension) is called a cycle. (D) TNA is isolated from a sample using (E) a nucleic acid extraction protocol. RNA is reverse transcribed to cDNA (F) template is mixed with primers, dNTPs and polymerase in a buffer. (G) The mix is then loaded onto a qPCR system which (H) quantifies the level of fluorescent signal which is correlated to the level of PCR product in each cycle. Figure created with BioRender.com.

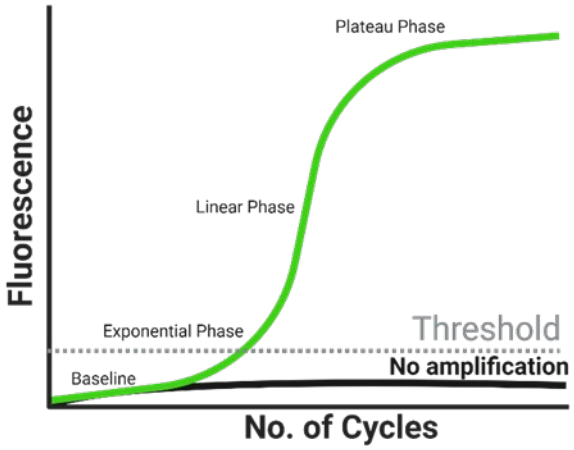


Figure 1.9: PCR amplification phases. During the baseline phase, the produced fluorescence is below the detection level. Then in the exponential phase, the fluorescent signal increases significantly above the background, the cycle at which occurs is called the cycle threshold (Ct). The PCR is in its optimal amplification stage with doubling PCR products in every cycle during the linear phase. is observed. At the exponential phase, quantification is possible. Figure created with BioRender.com.

The two most used qPCR methods for quantitative gene expression analysis are generic non-sequence-specific double-stranded DNA binding dyes such as SYBR Green, and sequence-specific probes such as TaqMan probes and ***PlexZyme*** (**Figure 1.10**)(89). Because SYBR green binds to all double-stranded DNA, it is essential to use primer pairs that are highly specific to their target sequence to avoid the generation of nonspecific products that will contribute to the fluorescent signal, resulting in an overestimation of the target (90). Extensive optimisations of primers for SYBR green qPCR assays are required to ensure that only the targeted product is formed. Primer pairs that exhibit self-complementarity should be avoided to prevent primer–dimer formation (91, 92).

The TaqMan probe method utilizes a fluorescently labelled probe that hybridizes to an additional conserved region that lies within the target amplicon sequence (89). The proximity of the probe of the quencher molecule to the fluorophore prevents it from fluorescing due to fluorescent resonance

energy transfer. During the annealing step of each cycle of the PCR, primers and the intact probe bind to their target sequences. During subsequent template extension, the exonuclease activity of the Taq polymerase enzyme cleaves the fluorophore from the TaqMan probe and a fluorescent signal is detected as the fluorophore is no longer near the quencher (**Figure 1.10**) (89).

While SYBR Green offers cost-effectiveness and ease of adoption its drawback lies in non-specific binding to any double-stranded DNA (93). In contrast, TaqMan probes, being sequence-specific and dual fluorophore-labelled DNA oligonucleotides, exhibit increased sensitivity and reproducibility, particularly advantageous for measuring lowly expressed genes. TaqMan probes can be labelled with different fluorophores, facilitating the development of multiplexed qPCR protocols whereby different targets can be coamplified and quantified within a single reaction. Therefore, multiplexing allows for the measurement of the expression levels of several targets in one reaction. This minimises assay costs and maximises the amount of information obtained from one sample. Additionally, multiplexing provides the option to include more than one reference control and the inclusion of positive and negative controls within the same reaction as the target (93). Although target-binding probes used in TaqMan improve specificity, new probes must be designed and optimised for new targets and extensive optimisation is required when multiplexing targe-specific probes together. In fact, very few researchers attempt to multiplex more than two targets together using TaqMan probes, and unfortunately, the amount of optimisation required to set up multiplexed assays reduces downstream cost benefits (94).

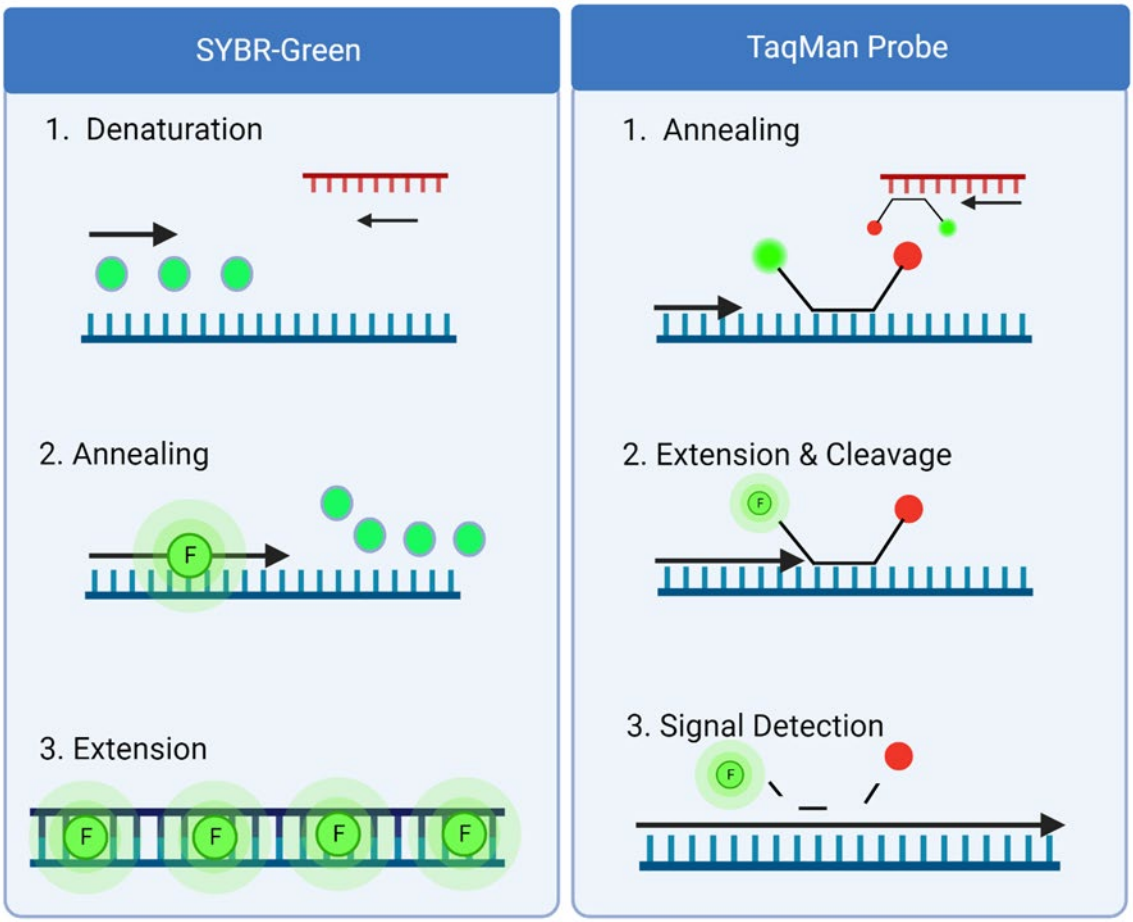


Figure 1.10: Principle of detection using SYBR Green and TaqMan Probe. SYBR Green, a DNA-binding dye, fluoresces once bound to dsDNA. During the extension phase, more SYBR binds to the PCR product, resulting in an increased fluorescence. TaqMan probes are hydrolysis probes that are complementary to the target sample. During the annealing phase, the primers and probe will anneal to the target sample. The probe is labelled with a fluorophore (green dot) and a quencher (red dot). During the extension phase, the probe is cleaved by the polymerase, thereby releasing the fluorophore from the quencher, producing fluorescence. At each amplification step the PCR product and fluorescence generated are directly proportional. Figure created with BioRender.com.

In contrast, other probe-based technologies enable a robust and inexpensive way to fully utilise the multiplex capacity of qPCR methods, such as ***PlexZyme*** (originally called MNAzyme) (95, 96). They have been widely used for the diagnosis of many infectious diseases and cancers. (97-99).

Plex*Zymes* are catalytic enzymes coupled with fluorescent probes (**Figure 1.11**). With ***Plex****Zymes* being multiple turnover enzymes, the target detection sensitivity is enhanced by multiple probes being cleaved during each qPCR cycle. Similarly, specificity is remarkably high due to the need for four distinct target-binding events to occur in a precise arrangement for signal generation: the hybridization of two PCR primers and two partzymes (97). Whereas TaqMan relies on three determinants of specificity involving two PCR primers and one target-specific probe. The heightened specificity of ***Plex****Zymes* is especially valuable in clinical contexts for accurate detection of gene expression.

Superior multiplexing capacity is possible with ***Plex****Zyme* technology, since each enzyme specifically recognises and binds to the target sequence, and then separately binds and modifies a specific fluorescent probe. The partzymes are single-stranded DNA oligonucleotides and therefore cheap to manufacture, whilst the fluorescent probe is universal, offering specific target interrogation and a generic readout (95), eliminating probe-related development costs and reducing assay development time.

The probes have previously been designed and tested to ensure no cross-reactivity with other probes that may be present in the same reaction, facilitating the detection of new assays without cross-reactivity or the need for new gene-specific probes. This stands in contrast to TaqMan, which require gene-specific probes designed for each new assay, leading to time-consuming and costly assay development (96).

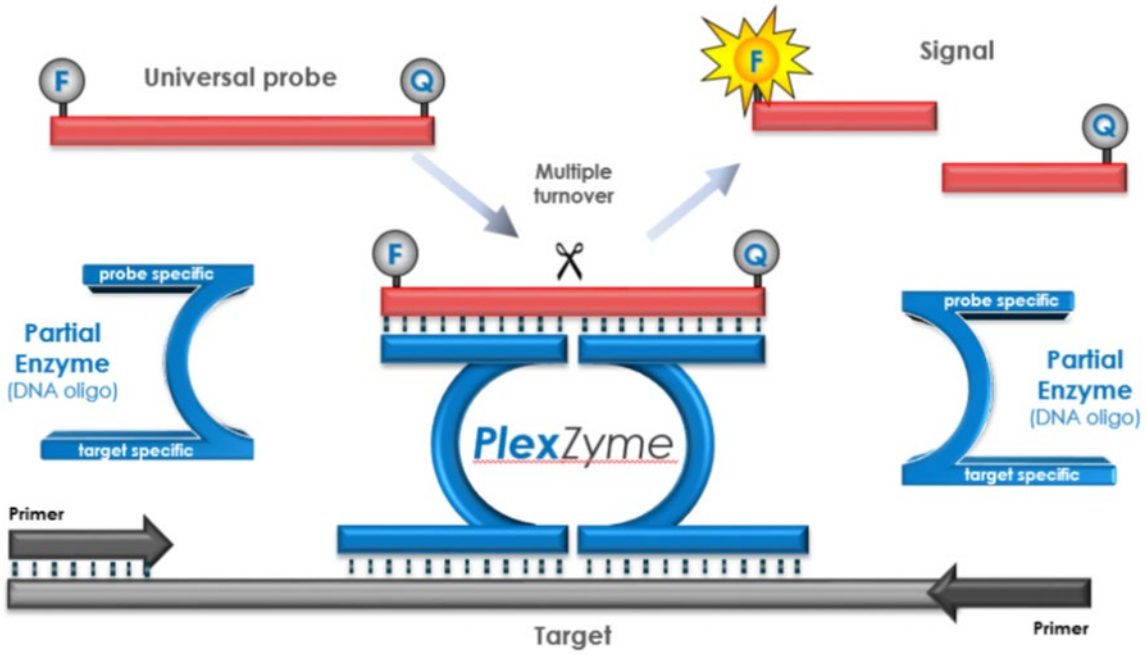


Figure 1.11: Schematic representation of *PlexZyme* assay components and signal detection. Image obtained from SpeeDx Pty Ltd. *PlexZyme* is composed of partzymes A and B which specifically bind to the target. Partzymes contain a target-specific arm, a partial catalytic core, and a probe-specific arm. Once the catalytic core is activated, forming a ***Plex****Zyme*, with the presence of a target, the fluorescently labelled probe is cleaved generating a signal that is monitored in real-time.

1.3.5 Normalisation of gene expression in RT-qPCR

RT-qPCR is a crucial molecular technique that combines reverse transcription and PCR principles to amplify and quantify specific RNA molecules present in samples. The use of RNA as a target in RT-qPCR provides insight into gene expression levels, enabling researchers and clinicians to understand cellular processes, disease mechanisms, and treatment responses at the molecular level. Both EndoPredict and Oncotype Dx use RT-qPCR to prognosticate breast cancer outcomes. While RT-qPCR offers great promise for diagnostic usage in public pathology, it is not without limits. It is well known that RNA can be unstable in clinical samples and requires more careful handling than DNA alone. However, other factors may also influence the accuracy of results based on the molecular methods used for analysis. Particularly, extraction methods need to be specific to extract

RNA only with no contaminating DNA, the normalisation method relies on having a stable molecule as a control however housekeeping genes used for this purpose are known to also change in their expression therefore judicious selection of housekeeping gene or housekeeping gene panel (100, 101).

To overcome these limitations, SpeeDx researchers have developed a method, termed ***In****Signia*®. It is a gene normalisation method, that relies on the co-amplification of both DNA and RNA within a single sample, extracted as total nucleic acid (TNA). This amplification of TNA from the sample allows for detecting the gene and the transcript expressed from that gene. The ***In****Signia* assay also detects a non-transcribed DNA sequence (NED). Gene and transcripts and NED have independent readouts and comparing both, through analysis of Cq values, allows the determination of an internal ratio of the number of RNA transcripts per DNA copy. This provides a normalisation method for mRNA expression or transcriptional activity based on an invariable DNA sequence, making the analysis of gene expression more reliable, and removing the need to select a normalising gene based on disease state or tissue. Furthermore, the method also circumvents the need to control for DNA contamination in an RNA sample, since both nucleic acids form part of the method.

1.4 Lack of prognostic tests guiding treatment of TNBC

In recent years, the development and application of omics techniques have greatly enhanced breast cancer research. This has yielded profound changes in the way that breast cancer is clinically assessed and treated. However, the benefit of this revolution in prognostic assessment has been seen only for ER+ breast cancers. While TNBC is highly heterogeneous, methods to stratify cases into different outcome groups have been elusive. What has been achieved, is an appreciation of the heterogeneity of the biology of TNBC, which is defined by the molecular subtypes observed by Lehmann and Burstein (68, 69, 71). The key driver of clinical molecular subtyping in ER+ breast cancer is proliferation. So, given that TNBC is characterised by an almost universally high proliferative rate, it is unsurprising that the efforts to apply a similar strategy to the development of precision molecular tools in this subset of breast cancers have been unsuccessful. A possible way forward, however, is to recognise that TNBC is a collection of biologically different tumour groups, which each could be assessed separately with different subtype-specific molecular signatures, that exploit their unique biology and disease characteristics. In summary, the way forward in prognosticating TNBC may be to further subdivide it before attempting to predict the outcome.

1.5 Aims and hypotheses

1.5.1 Aims

The first goal within this thesis was to optimise the performance of the PROSPER signature assay using RT-qPCR and ***Plex****Zyme* technology. **Chapter 3** explored the optimisation process for the implementation of the PROSPER gene signature into the ***Plex****Zyme* technology. Both ***In****Signia* (TNA-based) and RNA-based multiplexed assays were developed and validated on a variety of samples. The successful multiplexed assay was further optimised in **Chapter 4** to determine the performance and the clinical utility of the assay for the assessment of the PROSPER index in breast cancer samples. This chapter explored the feasibility of extracting nucleic acids from archival FFPE samples and further optimised the multiplexed assay in measuring gene expression in FFPE-derived material. Finally, the PROSPER index obtained from the multiplexed RT-qPCR assay was compared and validated against established markers such as the EPclin score, the NanoString-generated PROSPER assay results and Ki-67 results.

TNBC is recognised as one of the most aggressive breast cancer subtypes. However, due to the lack of receptors, the treatment options for TNBC are limited, making prognosis and personalized treatment strategies crucial. There is a growing interest in identifying gene signatures that can accurately predict survival outcomes in TNBC patients. Despite these molecular advances, the lack of specific prognostic markers for TNBC continues to hinder the effective management of this aggressive cancer subtype. The heterogeneity of TNBC necessitates the development of subtype-specific gene signatures that can accurately predict prognosis and guide personalized treatment decisions. Our lab developed a novel approach to subtype TNBC (manuscript under review). **Chapter 5** discusses the discovery and validation of a new prognostic subtype-specific gene

signature for TNBC using *in silico* methods using our lab's subtyping method. This study highlighted the significance of subtype-specific prognostic gene signatures in TNBC and provides a foundation for further research in the classification of TNBC.

1.5.2 Hypotheses

- The PROSPER assay can be translated to a multiplex RT-qPCR using the ***Plex****Zyme* technology as an affordable alternative to the current commercial multi-gene assay in predicting the risk of recurrence in ER+ breast cancer patients.

- The PROSPER test can be used to guide treatment recommendations for ER+ breast cancer patients.

- A more sophisticated exploration of the subtypes of TNBC could aid in the development of superior molecular prognostic tools for this challenging breast cancer type.

Chapter 2

Material and Methods

2.1 Materials

Table 2.1 Cell line

Cell Line	Source	Product number	Supplier	Culture medium
T-47D	T-47D are epithelial cells isolated from a pleural effusion obtained from a 54-year-old female patient with an infiltrating ductal carcinoma of the breast (102)	HTB-133	American Type Culture Collection (ATCC) Manassas, VA, USA	RPMI1640 medium supplemented with 10% fetal calf serum, and 0.25 U/ml insulin

Table 2.2 Reagents

Name	Product number	Supplier	Location
70% Ethanol	ETHABS70W/V5LT	POCD HealthCare	North Rocks, NSW, Australia
100% Absolute Ethanol	ETHABS2.5P	POCD HealthCare	North Rocks, NSW, Australia
Citrolene	Citrolene2.5LT	POCD HealthCare	North Rocks, NSW, Australia
Dulbecco's Phosphate-buffered saline (PBS)	17-512F	Lonza	Belgium

Eosin Alcoholic 1%	FNNII017	Thermofisher Scientific	Waltham, USA
Fetal calf serum	12003C	Sigma-Aldrich	St. Louis, USA
Formaldehyde solution 37%	818708	Sigma Aldrich	St. Louis, USA
Human Plasma	Acquired from human blood sample	-	-
Insulin	-	Novo Nordisk A/S	Bagsværd, Denmark
Mayer Haematoxylin with RIPE Solution	MH-1L	Amber Scientific	Midvale, Western Australia
Nuclease free water	11-05-01-04	Thermofisher Scientific	Waltham, USA
Reverse transcriptase (from SentiFAST Probe No-ROX One-Step Kit)	BIO-76005	Meridian Bioscience	Ohio, USA
RiboSafe RNase Inhibitor (from SentiFAST Probe No-ROX One-Step Kit)	BIO-76005	Meridian Bioscience	Ohio, USA
SensiFAST Probe No-ROX (SentiFAST)	BIO-76005	Meridian Bioscience	Ohio, USA
Thromborel S (thrombin)	-	Siemens Healthineers	Erlangen, Germany
TrypLE Express Enzyme (1X), no phenol red (TrypLE)	12604013	Thermofisher Scientific	Waltham, USA
Ultramount no.4 mounting medium	FNN11065C	Thermofisher Scientific	Waltham, USA

Table 2.3 Nucleic acid extraction kits

Name	Catalogue #	Sample Type	Nucleic Acids extracted	Supplier	Location
Complete MasterPure Purification Kit	MC85200	FFPE tissue	RNA, DNA and co-purification of RNA and DNA (total nucleic acid, TNA)	Astral Scientific	Taren Point, NSW, Australia
ISOLATE II RNA Mini Kit	BIO-52072	Cultured Cells	RNA	Meridian Bioscience	Ohio, USA
Roche High Pure RNA Paraffin kit	03270289001	FFPE tissue	RNA	Sigma Aldrich	St. Louis, USA
Quick DNA/RNA FFPE Kit	R1009	FFPE tissue	TNA	Integrated Sciences	Chatswood, NSW, Australia

Table 2.4 Equipment

Name	Supplier	Location
4200 TapeStation	Agilent Technologies	California, USA
CFX96TM Real-time PCR detection system	Bio-Rad	California, USA
Heraeus Biofuge fresco Kendro	Laboratory Products	Waltham, USA
Heraeus MultiFuge centrifuge x3R	Thermofisher Scientific	Waltham, USA
Microscope BH-2	Olympus	Shinjuku, Japan
Mini hybridisation oven	Thermofisher Scientific	Waltham, USA
Nanophotometer	Implen	Munich, Germany

Nanozoomer Slide Scanner	Hamamatsu	Japan
Rotary microtome cut 4060	Micro Tec	Walldorf, Germany
Shake 'n' Stack Hybridization Oven	Thermofisher Scientific	Waltham, USA
Heraeus Pico 17 microcentrifuge	Thermofisher Scientific	Waltham, USA
Trinean Xpose spectrophotometer	Integrated Sciences	Chatswood, NSW, Australia

Table 2.5 Software

Name	Company	Location
CFX Maestro	Bio-Rad	California, USA
FastPCR	PrimerDigital	Helsinki, Finland
MAD	SpeeDx	Eveleigh, NSW, Australia
Microsoft Excel	Office 365	Redmond, America
NDP view 2	Hamamatsu	Shizuoka, Japan
OligoAnalyzer	IDT DNA	Iowa, USA
Prism 10	GraphPad	California, USA
Primer-BLAST	National Institutes of Health	Maryland, USA
RStudio	RStudio, Inc.	Boston, USA

2.1.1 FFPE and flash frozen cell line samples

FFPE and flash frozen cell lines were utilised in this project. Mainly the T-47D cell line was used for RT-qPCR optimisations for the PROSPER assay.

2.1.2 Breast cancer patient samples

The PROSPER assay was validated in two separate cohorts comprising archival FFPE breast tumour tissues and associated patient information. Cohort 1 was recruited between 2017-2018 for a prospective study of the application of a molecular test (76) and was annotated with detailed histopathological data and results from EndoPredict prognostic testing (54), as well as the results of routine follow-up surveillance. Cohort 2 comprised breast cancer cases treated between 1987 and 1994 and was annotated with extensive clinical and survival information. Both cohorts underwent standard of care treatment which included a comprehensive histological assessment. Haematoxylin and eosin (H&E)-stained sections of the FFPE patient samples were used to assess the tumour content to assess appropriate sample input for nucleic acid extractions for this project. The use of both cohorts in this project was reviewed and approved by the Western Sydney Local Health District (WSLHD) Human Research Ethics Committee (HREC) (Approval reference 2019/ETH09574), the Westmead Institute of Medical Research (WIMR) and Westmead Hospital Research Governance Offices. Both cohorts were supplied by the Westmead Breast Cancer Institute.

2.1.2.1 Cohort 1

Cohort 1 was assembled by consecutive prospective recruitment of patients diagnosed with ER+, HER2- early-stage breast cancer, with tumour size < 5 cm and 0-3 involved nodes, at the Westmead

Breast Cancer Institute, between 2017-2018 (76). Information on patient recruitment and characteristics has been described previously (76). The original purpose of the recruitment was to determine extent of treatment recommendation change after incorporating the results of the commercial multi-gene test, EndoPredict into the multidisciplinary team discussion. Consequently, all FFPE tumour tissues had been tested using the EndoPredict commercial assay and the total RNA prepared from FFPE sections for the EndoPredict test was retained for future use and made available for this study (54). The study cohort consisted of 123 breast cancer cases, who had 5 years follow-up data. The original study received ethical approval from the WSLHD HREC (HREC/17/WMEAD/211). Study participants provided consent for the use of their samples and data in future unspecified research. Cohort 1 characteristics are summarised in **Table 2.6**.

2.1.2.2 Cohort 2

Cohort 2 consisted of patients diagnosed with clinical stage I or II invasive breast cancers, without lymph node involvement, who were treated with breast-conserving surgery and radiotherapy at Westmead Hospital, Sydney between 1987-1994. Information on patient recruitment, inclusion criteria and their characteristics has been described previously (103, 104). Aside from surgical and radiation treatment, these patients did not receive any other systemic treatment. The study cohort included a total of 164 breast cancer cases. The original study received ethical approval from the Sydney West Local Health District HREC. Informed consent for the inclusion of the cohort in this study was waived due to the due to the age of the tissue samples. The samples were anonymised, minimising any risk to participants. Cohort 2 characteristics are also summarised in **Table 2.6**.

Table 2.6 Cohort characteristics

Tumour characteristic	Cohort 1 (n = 123); n (%)	Cohort 2 (n=164); n (%)
Age at diagnosis (years)		
< 50	24 (20)	84 (51)
≥ 50	99 (80)	80 (49)
Tumour size (mm)		
0 – 5	0 (0)	6 (4)
6 – 10	8 (7)	23 (14)
11 – 20	70 (57)	98 (60)
> 20	45 (37)	36 (22)
Not reported	-	1 (1)
Histological grade		
1	27 (22)	39 (24)
2	58 (47)	68 (41)
3	38 (31)	57 (35)
Estrogen receptor intensity		
1	5 (4)	40 (24)
2	35 (28)	78 (48)
3	82 (67)	24 (15)
3+	1 (1)	-
Not reported	-	22 (13)
Progesterone receptor intensity		
0	3 (2)	-
1	15 (12)	42 (26)
2	45 (37)	73 (45)
3	57 (46)	27 (16)
3+	1 (1)	-
Not reported	2 (2)	22 (13)
EPclin score[a]		
1.9 – 3.3 (Low Risk)	61 (50)	-
3.4 – 6 (High Risk)	62 (50)	-
Ki-67[b]		
< 14 %	33 (27)	77 (47)
14–30%	42 (34)	40 (24)
> 30%	48 (39)	39 (24)
Not reported	-	8 (5)

Follow-up data		
Data available	116 (94)	164 (100)
Lost to follow-up	7 (6)	-
Events		
Recurrences	7 (6)	-
Breast cancer specific deaths	2 (2)	16 (10)
No events	112 (91)	143 (87)
Non-breast cancer specific death	-	5 (3)

[a]EPclin score: EndoPredict risk of recurrence score incorporating clinical factors

[b] Ki-67: Marker for proliferation

2.2 Methods

2.2.1 Tissue culture

All tissue culture work was performed using sterile materials and aseptic techniques, in a Class II biological safety cabinet. Cells were grown in a humidified incubator at 37°C with 5% CO_2. Cells were routinely tested for mycoplasma contamination.

2.2.1.1 Cell revival

Frozen cells were thawed in warm water and then transferred to a 75 cm^2 flask containing 25 mL of media. They were incubated at 37°C for 5 h to promote cell revival and adhesion to the flask. After this period, any non-adhered cells were collected by transferring the medium from the flask to a 50ml centrifuge tube and centrifuging at 200 x g for 5 min. Fresh medium (20 mL) was then added to the flask. After centrifugation, the supernatant was discarded, and the cell pellet was resuspended in 5 mL of fresh medium and returned to the flask for further incubation. The cells were passaged 2 days later.

2.2.1.2 Cell passaging

For cell passaging, the medium was discarded from the flask containing the cells. The cell monolayer was briefly rinsed with TrypLE, which was then removed and discarded. Fresh TrypLE (2 mL) was added to the monolayers, and the cells were incubated at 37°C for 3-5 min to dislodge them from the flask. To neutralise the TrypLE, fresh medium (8 mL) was added to the flask, followed by a rinse with 5 mL of medium. This suspension was then transferred to a 50 mL centrifuge tube. The cell suspension was centrifuged at 200 x g for 5 min, the supernatant was discarded, and the cells were resuspended in 10 mL of fresh media. A 0.5 mL sample of this

suspension was used to count cells using a haemocytometer. Based on this count, cells were distributed into new 75 cm² flasks containing medium to achieve a density that would reach 70-80% confluency in 7 days based on their doubling time (33 to 36 h). The cells were then either harvested as flash frozen pellets or made into FFPE blocks (**2.2.1.3**).

For frozen cell pellets, each harvested cell pellet was washed twice with cold Dulbecco's PBS, and centrifuged between each wash at 400 x g for 5 min removing the supernatant each time. The pellets were flash frozen by placing them on dry ice for 2 min. Frozen cells were then stored in a -80°C freezer.

2.2.1.3 Thrombin clotting of cell pellets for FFPE blocks

Cells were harvested and collected into a pellet by centrifugation, as described for routine cell line passaging. The pellet was washed twice with cold Dulbecco's PBS, centrifuging between each wash at 400 x g for 5 min and removing the supernatant. The cells were then fixed by resuspending the pellet in freshly made formaldehyde (3.7%; 1 mL) and incubated for 15 min at room temperature. Following fixation, cells were centrifuged for 5 min, the fixative was removed, and the cells were washed twice in cold Dulbecco's PBS centrifuging each time to re-pellet the cells.

The cells were then resuspended in human plasma (70 μL) and thrombin (70 μL) for the formation of the thrombin clot, which was incubated for 5 min at 37°C. Formaldehyde (3.7%; 0.5 mL) was added to the thrombin clot for incubation at room temperature on a rocking platform for 1 h. The clot was then placed in a CellSafe Biopsy Capsule (CellPath, Newtown, UK) The embedding cassette was processed through an automated processor program for 4 hours, and the processed thrombin clot was subsequently paraffin-embedded. The FFPE blocks were stored at room temperature.

2.2.2 H&E staining of FFPE samples

FFPE sections (5 μm) were sliced using a microtome and floated in a waterbath for collection onto positively-charged microscope slides. After drying for 1 h at 60°C, the sections were dewaxed by submersion in citrolene twice with 5 min incubations. The slides were then rehydrated by sequentially incubating them in 1) 100% ethanol twice, then 2) 95% ethanol once, and 3) finally 70% ethanol once, each for 5 min. After rehydration, the slides were washed in tap-water and subsequently rinsed in deionized water.

The nuclei were stained by placing the slides in haematoxylin solution for 5 min, followed by rinsing the slides in running tap water. The slides were then quickly dipped in 0.3% acid alcohol before submerging the slides in running hot tap water for 10 min to differentiate and stain the nuclei with blue colour. Next, the slides were submerged in 1% alcoholic-eosin for 2 min to stain the cytoplasm pink. This was followed by rinsing the slides in tap water. The slides were dipped 10 times in 70% ethanol and in 100% ethanol to dehydrate the sections. The slides were then incubated in citrolene three times for 5 min each. Finally, coverslips were mounted onto the slides using Ultramount mounting medium. The slides were stored at room temperature.

2.2.3 Processing of FFPE samples for nucleic acid extraction

Before nucleic acid extractions, an appropriate amount of FFPE material was sectioned at 10μm into curls using a microtome according to the manufacturer's instructions from the relevant nucleic acid extraction kit (**Table 2.3**) and then manually deparaffinised. The amount of tissue processed was dependent on the specific reagent kit used: up to 50mm^2 for MasterPure, 100mm^2 for Roche High Pure, or 40mm^2 for Zymo RNA/DNA Quick Kit. For manual deparaffinisation, the FFPE

curls were first washed three times in citrolene (800 µL), with centrifugation and removal of the supernatant between each wash. All centrifugation for the deparaffinisation steps were performed at 21,100 x g for 2 min. Next, 100% ethanol (800 µL) was added to the sample, followed by a 5 min incubation at room temperature and centrifugation to remove the supernatant. This step was repeated with an additional ethanol rinse (100% ethanol; 1 mL). The tissue pellets were then placed into an incubator at 55°C for 10 min to remove excess ethanol from the samples. After this, samples were taken immediately for nucleic acid extraction, which was performed according to the instructions for each of the specific kits.

2.2.4 Standard one-step RT- qPCR conditions

RT-qPCR reactions were performed using a consistent RNA or TNA input volume, which contained either undiluted or 1 in 10 diluted products from the nucleic acid extraction kits used. Although RNA quantity and quality assessment were performed, input amounts were not normalised to a uniform nucleic acid mass for three main reasons:

1. An objective of the assay was to maximize template input, and therefore sensitivity, while avoiding inhibition by possible contaminants. Thus, a consistent sample eluate input volume was used.
2. Due to the degraded nature of the input material, it was not possible to ensure consistent input of RNA or DNA template by normalising the RNA input quantity used.
3. It was not possible to quantify the RNA component present in TNA extracts.

The RT-qPCR reactions were prepared using the following components and conditions:

- Total Reaction Volume: 20 µL
- Sample Volume: 5 µL (sample or water for no template control)

- Reaction Volume (excluding sample): 15 µL

Components:

- Forward Primer: 40 nM
- Reverse Primer: 200 nM
- Partzyme A: 200 nM
- Partzyme B: 200 nM
- Substrate: 200 nM
- SensiFAST Mix: 1x
- RiboSafe RNase Inhibitor: 0.2 U/µL
- Reverse Transcriptase: 1U
- Nuclease-Free Water: To adjust the final volume to 15 µL

The following steps were undertaken to set up the RT-qPCR reactions:

1. **Preparation of Reaction Mixture**: The forward primer, reverse primer, partzyme A, partzyme B, substrate, SensiFAST mix, RiboSafe RNase Inhibitor, and nuclease-free water were combined to achieve a total volume of 15 µL per reaction.
2. **Final Reaction Volume Adjustment**: The reaction mixture was adjusted with nuclease-free water to ensure a total reaction volume of 15 µL per well.
3. **Sample Addition**: Each reaction well received 5 µL of either the sample (containing RNA template) or nuclease-free water (for no template control), bringing the final reaction volume to 20 µL.

Thermal Cycling parameters:

All reactions were performed on a BioRad CFX96 thermocycler. The thermal cycling parameters were as follows:

1) Initial incubation for reverse transcription: 55°C for 10 min
2) Polymerase activation: 95°C for 2 min
3) Touchdown PCR: 10 cycles of 95°C for 5 s and 61°C for 30 s (decreasing by 0.5°C per cycle)
4) Amplification: 40 cycles of 95°C for 5 s and 52°C for 50 s.

2.2.5 Standard nucleic acid and oligonucleotide quantity and quality procedures

2.2.5.1 RNA quantification and quality assessment

RNA quality and concentration were determined by measurement of absorbance at 260 and 280 nm using a nanophotometer after correction for background readings measured using the elution buffer (provided by the respective kits utilised for RNA extraction (**Table 2.3**) used to reconstitute the specific sample. Quality assessment of RNA samples was performed using an RNA ScreenTape (Agilent Technologies, California, USA) with a 4200 Tape Station.

2.2.5.2 Oligonucleotide resuspension and quantification

All oligonucleotides were synthesised and produced by Integrated DNA Technologies (Iowa, USA). All oligonucleotides were acquired lyophilized and subsequently resuspended in nuclease free water prior to use. The concentration of oligonucleotides was determined by measuring their absorbance at 260nm using the Trinean Xpose spectrophotometer.

Chapter 3

Development of a novel breast cancer precision test: From target gene primer selection to a multiplex test

3.1 Introduction

Approximately 70-80% of breast cancers are ER+ and can be classified into two intrinsic molecular subtypes: luminal A and luminal B (105, 106). Luminal B breast cancers generally have poorer outcomes compared to luminal A breast cancer due to their higher proliferation rates. While luminal A tumours respond better to hormone therapy (107), luminal B tumours may require additional chemotherapy (4). Therefore, accurately distinguishing between these subtypes is crucial for recommending appropriate treatment options, highlighting the importance of this distinction in ER+ breast cancer patients' outcomes.

In clinical practice, traditional pathological features such as tumour grade, size and receptor and proliferation marker staining are typically used to predict the prognosis and likelihood of response to systemic therapies (4, 108). However, in certain cases of luminal breast cancers, limited technical reproducibility, subjective interpretation, and qualitative readouts make it difficult to recommend treatments based on these traditional features when distinguishing between these patients, particularly in the cases where intermediate and equivocal results are returned (109-111).

Since the advent of molecular profiling approaches, companies worldwide have leveraged this technique to develop commercial multi-gene tests that can accurately predict recurrence in luminal breast cancers. Examples include Oncotype Dx (38), Prosigna (112) and EndoPredict (113). These

tests aid in eliminating the subjectivity in interpreting pathology assessment by setting a clear cut-off point for classifying patients into low and high-risk categories. Moreover, the use of a panel of genes instead of a single marker introduces redundancy in the assessment, which improves both the sensitivity and accuracy of the test (43). Since the introduction of Oncotype Dx, there has been a reduction in the recommendation of adjuvant chemotherapy to luminal breast cancer patients.

However, despite these advantages, the adoption of these multigene tests in Australia has been limited due to their associated high cost. Oncotype Dx, the most widely used commercial multi-gene test, is not subsidised by Medicare. Furthermore, tumour tissue blocks must be sent to their centralised testing facility in the United States, causing delays in treatment decisions. While EndoPredict testing is partially subsidised by Medicare, its uptake remains low due to the high cost to the patient. Given these challenges, there is an urgent need for a more cost-effective, standardised multi-gene assay for routine clinical use.

To address this need, our lab identified a gene signature called PROSPER that could be used to predict the risk of recurrence in ER+ breast cancers. The markers were identified and validated in the public METABRIC breast cancer dataset (114) and revealed that the PROSPER gene signature performed equally well or better than the predictive signatures underpinning current commercial tests when cases were stratified into predicted low and high-risk groups. When the PROSPER signature was assessed using the NanoString nCounter platform on Cohort 1 samples (**Table 2.6**), the PROSPER index was significantly associated with clinical grade, proliferation marker Ki-67 and the results of the EndoPredict commercial multi-gene test. Additionally, when PROSPER was assessed on Cohort 2 (**Table 2.6**), the PROSPER index was significantly associated with Ki-67, and disease-specific survival and overall survival.

In this chapter, through collaboration with a Sydney-based biotech company, SpeeDx Pty Ltd, the PROSPER signature was optimised into a cost-effective, PCR-based multiplexed assay for use in standard PCR thermocyclers. For real-time detection, the SpeeDx propriety ***Plex****Zyme* technology was employed, due to its superior multiplexing capability (115). The SpeeDx ***In****Signia* (TNA-based assay) technology was also applied for the normalisation of gene expression relative to a stable molecule of DNA, contrasting with traditional normalisation methods that use variable housekeeping genes to correct the output from assessing RNA samples. This chapter outlines the optimisation process, from design to validation, highlighting the potential for this affordable assay to guide treatment decisions in ER+ breast cancer.

3.2 Aims and Hypotheses

3.2.1 Aims

1) To design and optimise real-time RT-qPCR assays targeting each gene in the PROSPER gene signature and normalising genes, using ***Plex****Zyme* technology.

2) To develop and assess the performance of the PROSPER gene real-time assays into a multiplexed reaction.

3) To validate the performance, specificity, and robustness of the PROSPER multiplexed assay across various sample types.

3.2.2 Hypotheses

It was hypothesised that the PROSPER signature can be successfully incorporated in an RNA and TNA-based multiplexed RT-qPCR assay utilising SpeeDx proprietary ***Plex****Zyme* technology. Additionally, when the multiplexed assay is assessed on a synthetic progestin, ORG2058 treated T-47D cell line, the control group will exhibit a higher PROSPER index compared to the drug-treated group.

3.3 Approach

In this chapter, the PROSPER signature was optimised into TNA-based and RNA-based RT-qPCR multiplexed assays. Four of the PROSPER signature genes were included in the multiplexed assay along with either an RNA housekeeping gene (HK Gene 1) or a non-transcribed control DNA region (NED Gene 1) as the normalisation genes for both RNA and TNA-based assays respectively. Flash-frozen T-47D cells were used as the primary samples for the development and optimisation of the multiplexed assays. To comprehensively evaluate the assay performance, low (40 transcript copies) and high (1000 transcript copies) nucleic acid targets per sample were analysed. These were calibrated by reference to a standard curve of known target amplicon input numbers. The extraction method outlined in **Chapter 2** was employed to extract nucleic acids from primary samples for the optimisation of both RT-qPCR (hereafter termed qPCR).

The initial step in assay development involved designing oligonucleotides (primer pairs and ***Plex****Zymes*) for the PROSPER targets and the normalisation genes. This was accomplished using various computational tools, including Primer-BLAST (116), FastPCR (117), OligoAnalyzer Tool (IDT, Iowa, USA) and MAD (SpeeDx, Eveleigh, NSW, Australia). Following this, experimental validation of the *in silico* designed oligonucleotides was performed. The PCR reaction was refined from a single-plex format, where each PCR well tests a single gene of interest, to a multiplexed format, where all gene targets are detected in a single PCR well. The broader applicability and robustness of the assay were assessed by testing the assay's performance on various sample types. Finally, the multiplexed assay was validated on paired samples of T-47D cells treated with the growth inhibitory synthetic progestin (ORG2058) or vehicle control (**Figure 3.1**). A graphical summary of this chapter is depicted in **Figure 3.2.**

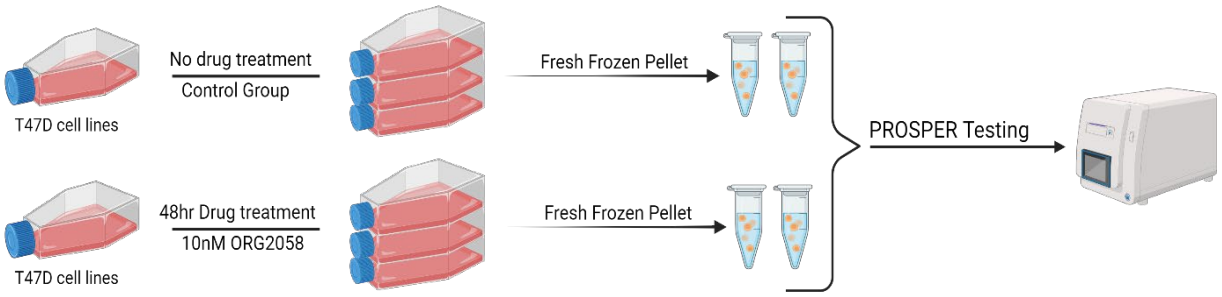


Figure 3.1: Workflow of synthetic Progestin, ORG2058, treatment of T-47D cell lines. T-47D cells were seeded in a T75 flask. At 50% confluency, cells were treated with 10nM ORG2058 or vehicle (ethanol), as indicated. After 48h, the cells were harvested, and flash frozen in cell pellets. RNA and TNA were extracted for PROSPER multiplexed qPCR testing.

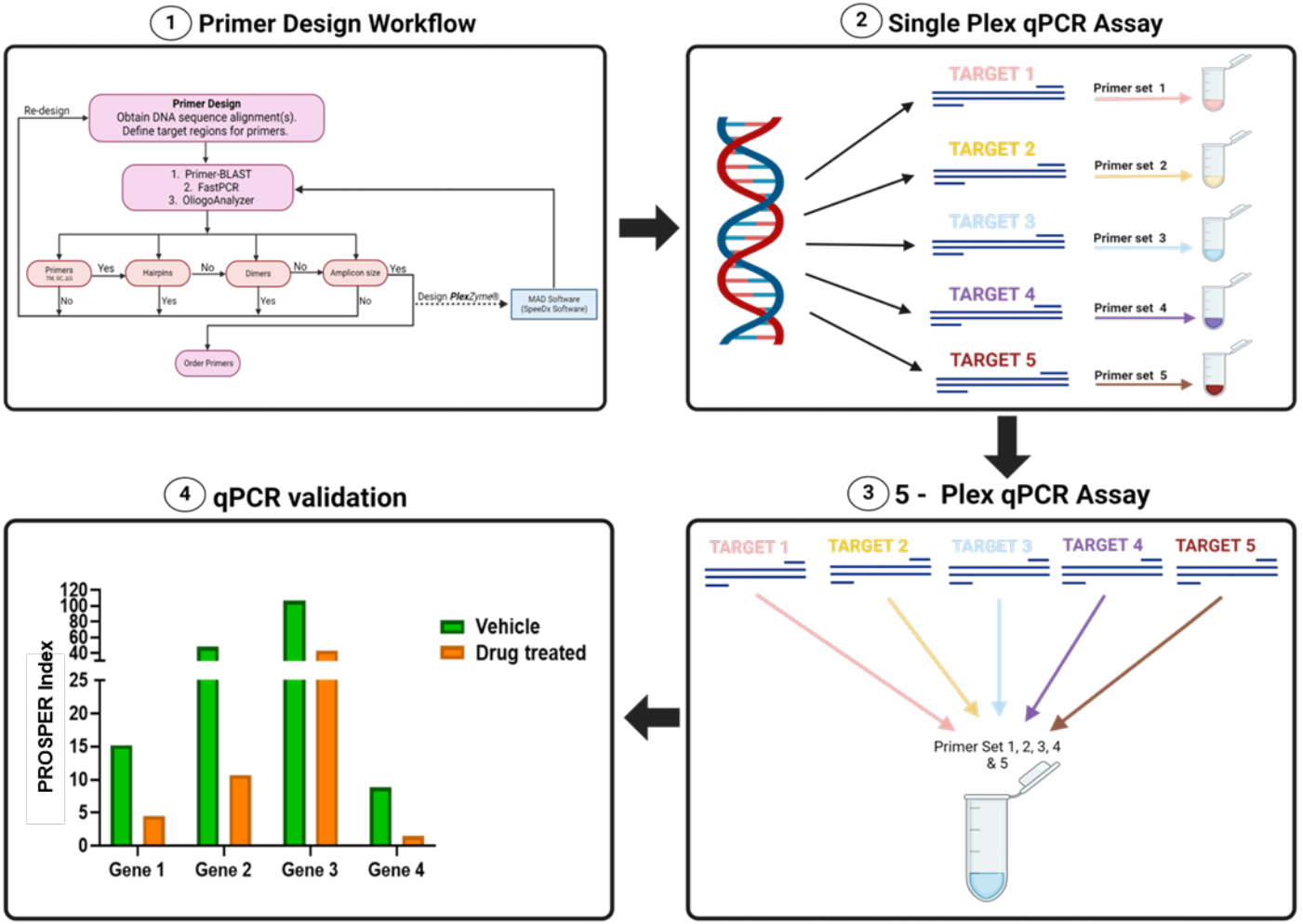


Figure 3.2: Graphical summary of Chapter 3. Firstly, a structured primer design approach was established, which allowed for a systematic workflow in designing oligonucleotides for each target. This approach is shown and discussed in detail in a later section. Secondly, the *in silico* designed oligonucleotides were experimentally validated and assessed for the performance of the RNA and TNA-based assays from single plex to a multiplex format. Steps 3 and 4 were the assessment of the 5-plex assays in various samples and the validation of the 5-plex assays in determining different proliferation levels in drug-treated T-47D cell line. Figure created with BioRender.com

3.4 PROSPER qPCR assay design

The original PROSPER gene signature was composed of five gene targets along with a panel of housekeeping genes. To protect intellectual property, the target genes are designated as Gene 1 through Gene 5 in this thesis. To facilitate the implementation of the PROSPER signature into a qPCR format, the assay was designed to be compatible with the widely used BioRad CFX Real-Time Quantitative PCR, thermocycling instrument. The platform is equipped to simultaneously detect five fluorescent signals across a detection range of 510–730 nm, including HEX, FAM, CY5, Quasar and TX4. *In silico* analysis was conducted to determine which target gene could be removed without affecting the accuracy of the assay so the 5th channel could be allocated to a normalising gene target. Linear regression analysis of the combined NanoString gene expression data from Cohorts 1 and 2 was used to compare the five-gene PROSPER index with each possible permutation of the PROSPER index after the removal of one of the five genes. This identified the optimal 4-gene PROSPER signature, which showed a high correlation with the 5-gene signature, with a linear correlation score of $r^2 = 1$ and slope of best fit of 0.996 (**Figure 3.3**), suggesting that the removal of the least influential gene from the PROSPER signature had limited impact. PROSPER genes 1- 4 were selected to be developed as a multiplexed qPCR assay.

The forward and reverse primer pairs for these targets were designed to target both the DNA and RNA components of the target gene sequence. Since the PROSPER signature was being optimised as an RNA-based and TNA-based ***In****Signia* ***Plex****Zyme* multiplexed assay, to streamline the experimental workflow, the ***Plex****Zyme* assay was optimised in both TNA and RNA-only samples.

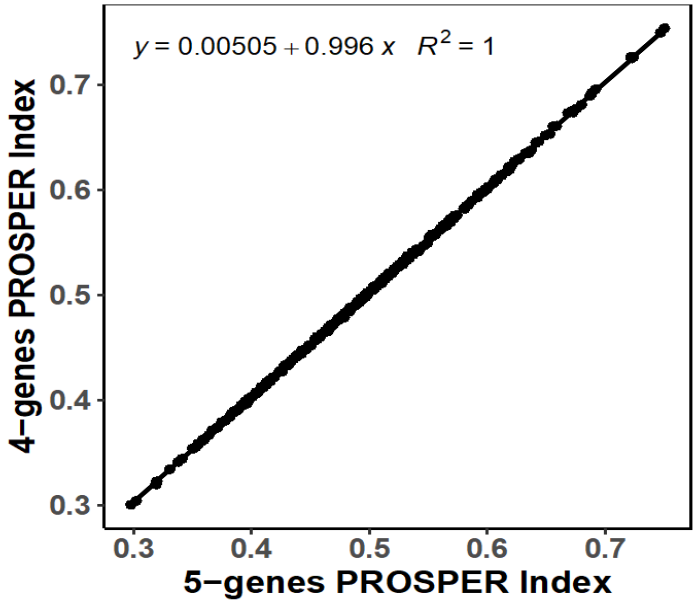
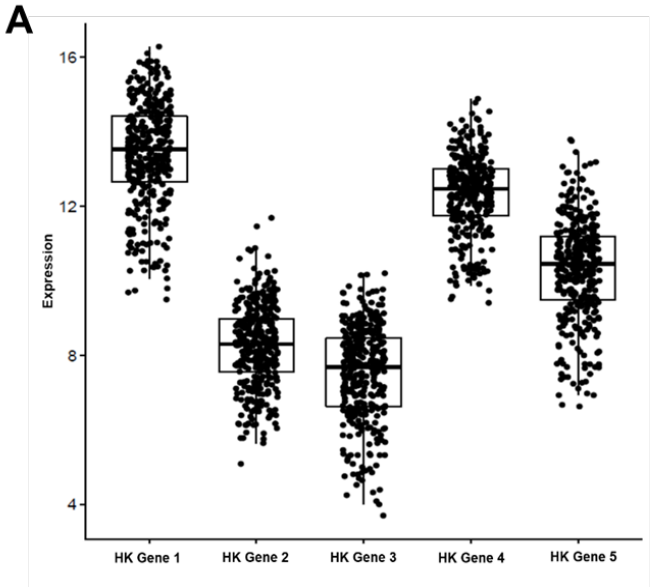


Figure 3.3: Linear correlation between the 5 gene and 4 gene PROSPER index. A scatter plot of the 5-gene and 4-gene PROSPER index was computed using Cohort 1 and Cohort 2 samples.

3.4.1 RNA-based PROSPER qPCR assay

To determine a robust control for the RNA-based qPCR assay, *in silico* analysis was performed to identify the most stable RNA housekeeping gene. Although the original PROSPER gene signature consisted of five housekeeping (HK) genes, due to the limited channels (n= 5) available to multiplex in the qPCR assay, only one channel was allocated for the most stable housekeeping gene from the PROSPER signature. The log2-transformed raw transcript counts from Cohort 1 and 2 of all RNA housekeeping genes within the PROSPER signature were assessed using the NanoString Platform. Among these genes, HK Gene 1 exhibited the highest expression followed by HK Gene 4 and HK Gene 5 (**Figure 3.4A**). Notably, HK Gene 1 had the highest variance value of 1.14, followed by HK Gene 2 (1.04). In contrast, HK Gene 4 demonstrated the lowest variance,

measured at 0.78 (**Figure 3.4B**). Therefore, HK Gene 4 (referred to hereafter as RNA HK Gene 1) was selected for the RNA-based qPCR assay development and optimisation.



B

Housekeeping gene	Variance
HK Gene 1	1.14
HK Gene 2	1.04
HK Gene 3	1.03
HK Gene 4	0.78
HK Gene 5	0.86

Figure 3.4: Variance of PROSPER signature housekeeping genes. (A) Box plots illustrating the expression levels of the five housekeeping genes in the PROSPER signature measured on the NanoString platform in Cohort 1 and Cohort 2 samples. (B) The table illustrates the variance in the expression of the housekeeping genes in both cohorts.

3.4.2 TNA-based qPCR assay

For the DNA control for the TNA-based qPCR assay, the NED control was designed by the SpeeDx team. The essential criteria for the selection of a NED region were that the DNA sequence should be a stable copy number, and never transcribed, thus only found within a human cell in DNA form. Sequences found were then shortlisted by including those GC content between 40 – 60%. These then had primer pairs designed and were tested experimentally comparing Cq values in reactions with and without reverse transcription. Regions with elevated delta Cq values were excluded and regions with no change in Cq had a ***PlexZyme*** assay designed for target detection in quantitative

PCR. The best region was selected based on assay efficiency and non-homology to other regions of interest. Additionally, it is known that cancer can cause genomic disruption that could potentially impact the reliability of using NED as a PCR control. Therefore, detailed genomic analyses on ER+ breast cancer was performed to screen for regions with reliable detection of DNA controls. Consequently, NED (NED Gene 1) was selected as a DNA control for TNA-based qPCR assay development and optimisation.

3.4.3 Oligonucleotide design for measurement of the PROSPER signature

Four ***Plex****Zyme* assays were designed to target the four selected gene transcripts from the PROSPER signature, and two ***Plex****Zyme* assays were designed for RNA HK Gene 1 and NED Gene 1 for the RNA and TNA-based qPCR assays, respectively.

3.4.3.1 Oligonucleotide characteristics

The primer pairs were designed to create an amplicon shorter than 150bp and consisted of DNA oligonucleotides, specific to the gene sequence, that were approximately 20 bases in length. The ***Plex****Zyme* designs for each target included two 40 base partzymes (partzyme A and partzyme B) (**Figure 3.5A**). They were synthesised to specifically bind to the target gene sequence and a fluorescently labelled probe. The target-specific and the probe-specific arms were designed to have a melting temperature between 59ºC and 62ºC. To minimize excessive core hybridization of partzymes to the amplicon, designs avoided more than three self-complementary bases in the partzyme core binding area (**Figure 3.5B**). The probe comprised 20 DNA bases with two central RNA bases (guanine and uracil), labelled with a fluorophore at the 5' end and a quencher at the 3' end.

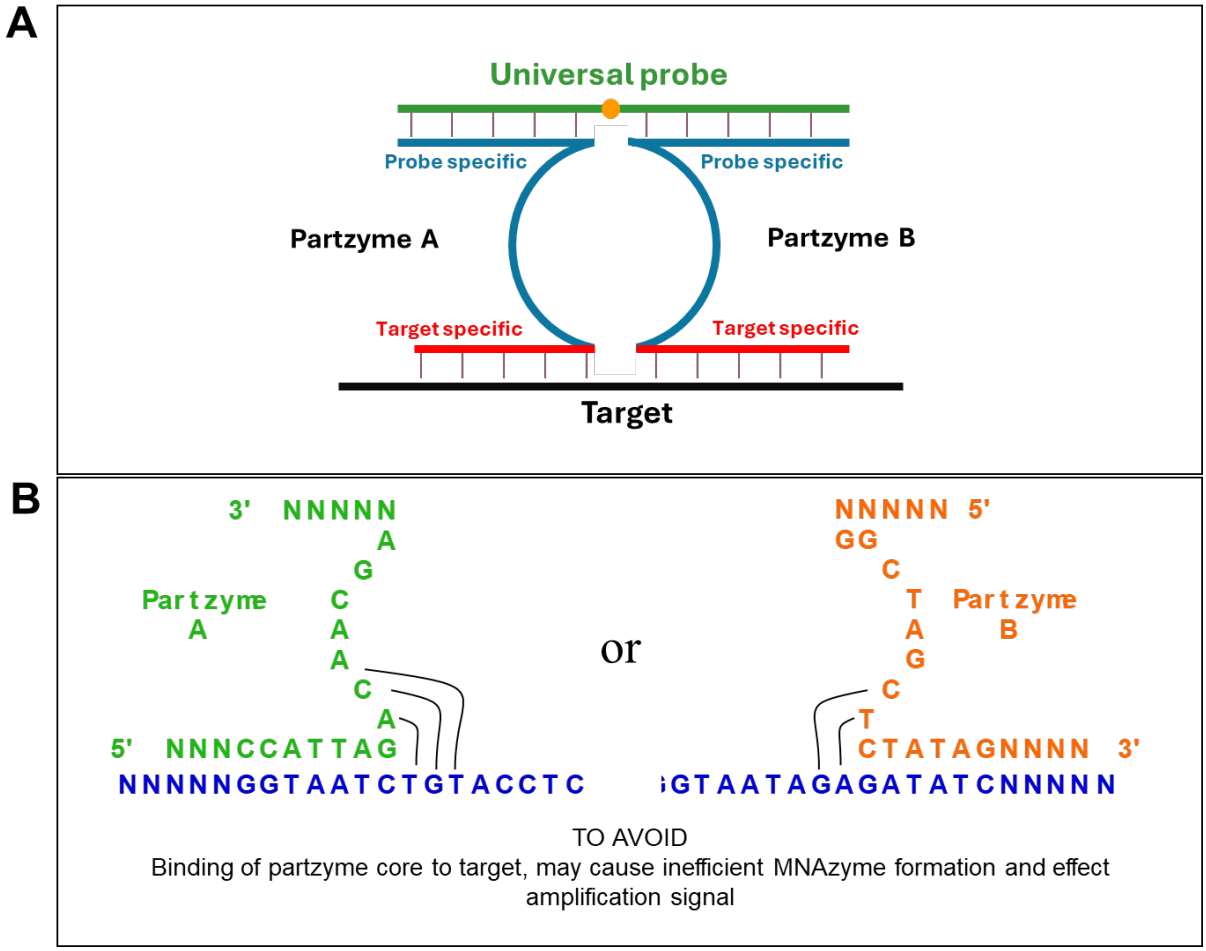


Figure 3.5: *PlexZyme* assay design. Modified image obtained from SpeeDx. (A) Schematic diagram of the *PlexZyme* assay components. (B) Illustration showing examples of excessive partzyme core to the amplicon binding denoted by the black connecting lines.

3.4.3.2 Oligonucleotide selection process

A structured primer design approach (**Figure 3.6**) was established, which allowed a systematic workflow in designing oligonucleotides for each target within the assay. As illustrated in **Figure 3.6**, the optimisation process ensured that these oligonucleotides specifically detect their target genes. Computational tools including Primer-BLAST, FastPCR and OligoAnalyzer, were utilised to confirm the specificity of the oligonucleotides and to assess potential issues such as primer dimerisation and secondary structure.

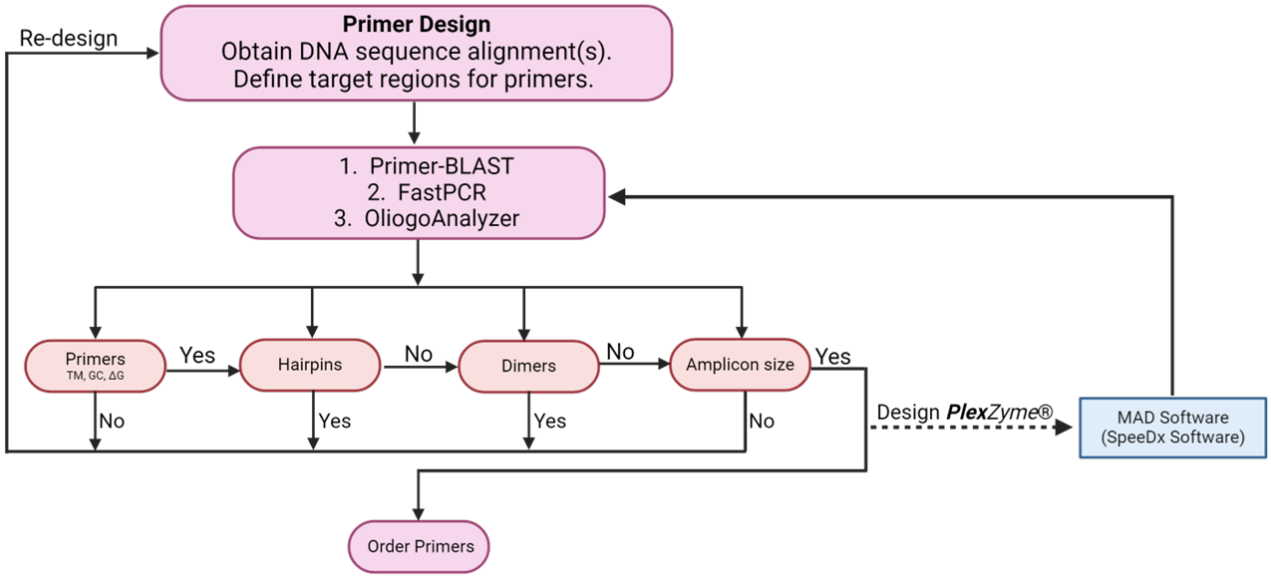


Figure 3.6: Multiplex PCR Primer and *PlexZyme* Design workflow. The primer design process began with identifying a list of sequence regions using the National Center for Biotechnology Information database. Each primer pair was then evaluated for specificity, dimer formation, and hairpins using software tools such as Primer-BLAST, FastPCR, and OligoAnalyzer. If any primer pair failed to meet these criteria, they were rejected. Next, *PlexZymes* were designed using MAD software, developed by SpeeDx. These specific oligonucleotides proceeded to the second step in the workflow. For multiplex PCR, the cycle was repeated, returning to the first step for the next target gene. This iterative process ensured that each designed primer pair was suitable for multiplexing, providing reliable and specific PCR results. Figure created in Biorender.com.

3.4.3.3 qPCR assay optimisation process

The initial step in optimising the multiplexed PCR RNA and TNA assays involved confirming that each target assay was effectively amplified in a single-plex setting. Next, the process moved to 2-plex assays, where two target assays were combined in a single reaction to assess their compatibility. This was followed by 3-plex, 4-plex, and finally 5-plex reactions. If at any point in this optimisation process, the reactions were a failure and resulted in ineffective compatibility, the assay design process would be recommenced from the first step. This iterative process resulted in

testing various assay designs. However, this chapter focuses primarily on the effective assay designs and their performance.

3.5 Results

3.5.1 Performance of each qPCR assay in a single plex

The single plex qPCR assays of the PROSPER targets and the normalising genes were tested in low and high copy numbers of TNA samples extracted from flash-frozen T-47D cells. Fluorescence channels FAM, CY5, HEX and TX4 were used for Gene 1, Gene 2, Gene 3, and Gene 4 respectively. These single plex assays showed a consistent increase in fluorescence intensity over time for each target. The amplification curves exhibited a sigmoid-shaped curve with a defined exponential phase, and no amplification was recorded for the no template controls. The amplification curves for Gene 1 and Gene 4 demonstrated a flatter geometric amplification slope, suggestive of lower efficiency of amplification compared to the other targets. (**Figure 3.7**)

As illustrated in **Figure 3.8 A and B,** single plex assays conducted in the low copy number sample yielded later Cq values for all targets. All the targets yielded earlier Cq values in the high copy number sample. Gene 4 had the lowest Cq value in both low and high copy number samples. The fold differences in Cq values, between the low and high copy number inputs, are shown in **Table 3.1**. The results closely aligned with the expected amplification patterns with the expected 25-fold difference (1000 transcript copies/40 transcript copies), while RNA HK Gene 1 and NED Gene 1 showed deviations.

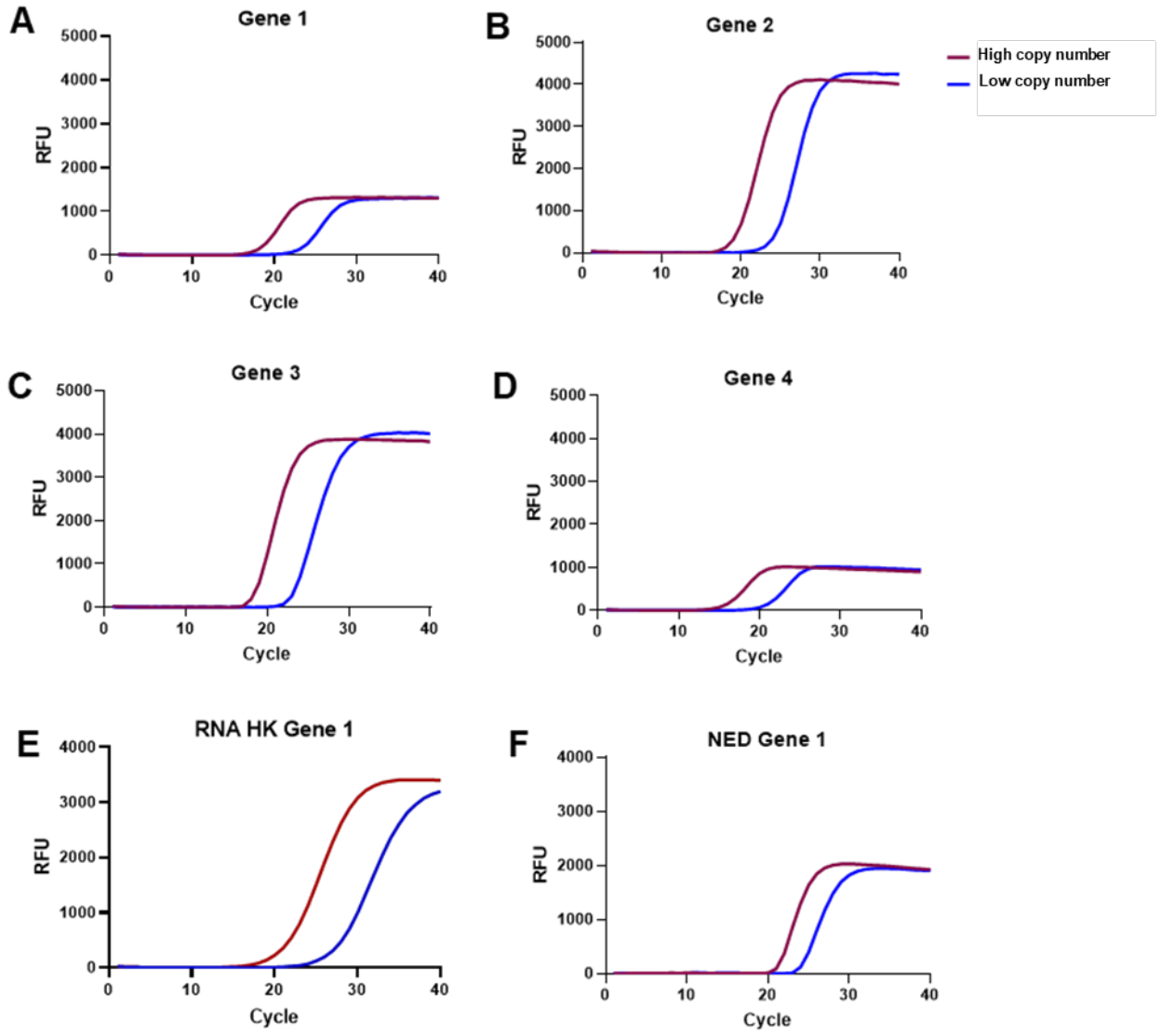


Figure 3.7: Single plex amplification plots of the PROSPER targets. Single plex assays were tested on TNA samples extracted from flash-frozen T-47D cells. (A-F) Amplification plots of the PROSPER targets, RNA HK Gene 1 and NED Gene 1. High and low copy number TNA were inputted for the PCR reactions. PROSPER targets were tested individually in duplicates.

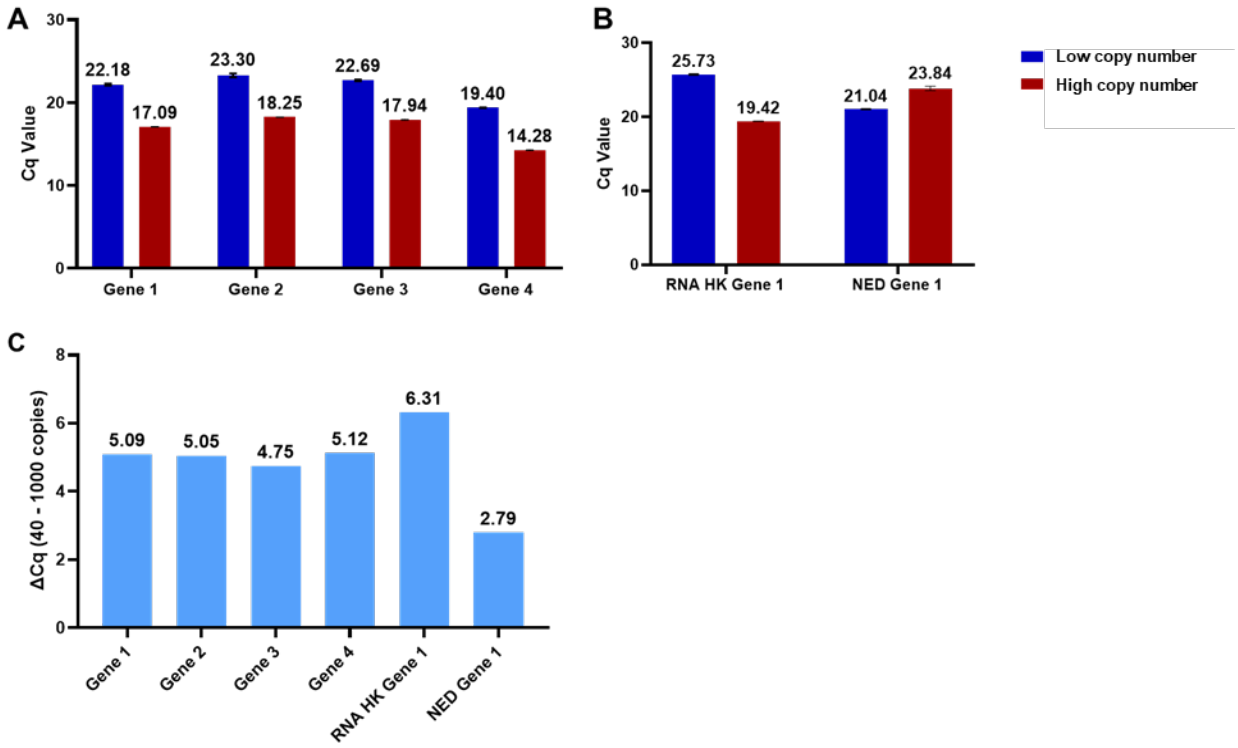


Figure 3.8: Cq values and ΔCq differences of single plex assays. (A and B) Bar charts of the average Cq values from the amplification plots of the single plex assay of all targets in low and high copy number samples. (C) Bar chart illustrating the ΔCq of the targets between the low and high copies.

Table 3.1 Fold difference in single plex assay

	ΔCq	Fold difference
Gene 1	5.09	34.3
Gene 2	5.05	34.0
Gene 3	4.75	27.0
Gene 4	5.12	34.5
RNA HK Gene 1	6.31	81.0
NED Gene 1	2.79	6.9

ΔCq = Cq values of 40 transcript copies – Cq values of 1000 transcript copies; Fold difference = $2^{\Delta Cq}$

3.5.2 Performance of assay in 2-plex and 3-plex format

The 2-plex and 3-plex qPCR assays of the PROSPER targets and the normalising genes were tested with low and high copy number TNA samples extracted from flash-frozen T-47D cells. The 2-plex assay of the targets showed consistent and defined sigmoidal PCR curves (results not shown). In the 3-plex assays, the fluorescence intensity increased over time for each target in both low and high copy number samples. Fluorescence channels FAM, CY5 and TX4 were used for Gene 1, Gene 2 and Gene 3, respectively. The amplification curves displayed the expected sigmoid-shaped curve with a defined exponential phase, and no amplification was recorded for the no template controls (**Figure 3.9 A and B**). As observed before, the slope of the geometric amplification phase of the PCR reactions for Gene 1 were flatter than for Gene 2 and Gene 3, suggesting a lower efficiency of amplification and potentially delaying the Cq value for that target. However, given that the same amplification characteristics were consistently seen for this gene, the results obtained could be compared between samples.

In the low and high copy number samples, the mean Cq values for the single-plex assays were higher than those in the 3-plex reactions. Consistent Cq values were seen in a single target and 3-plex assays for the three genes tested (**Figure 3.9 C and D**). This was true for both the low copy number and high copy number template input (**Figure 3.9 C and D).**

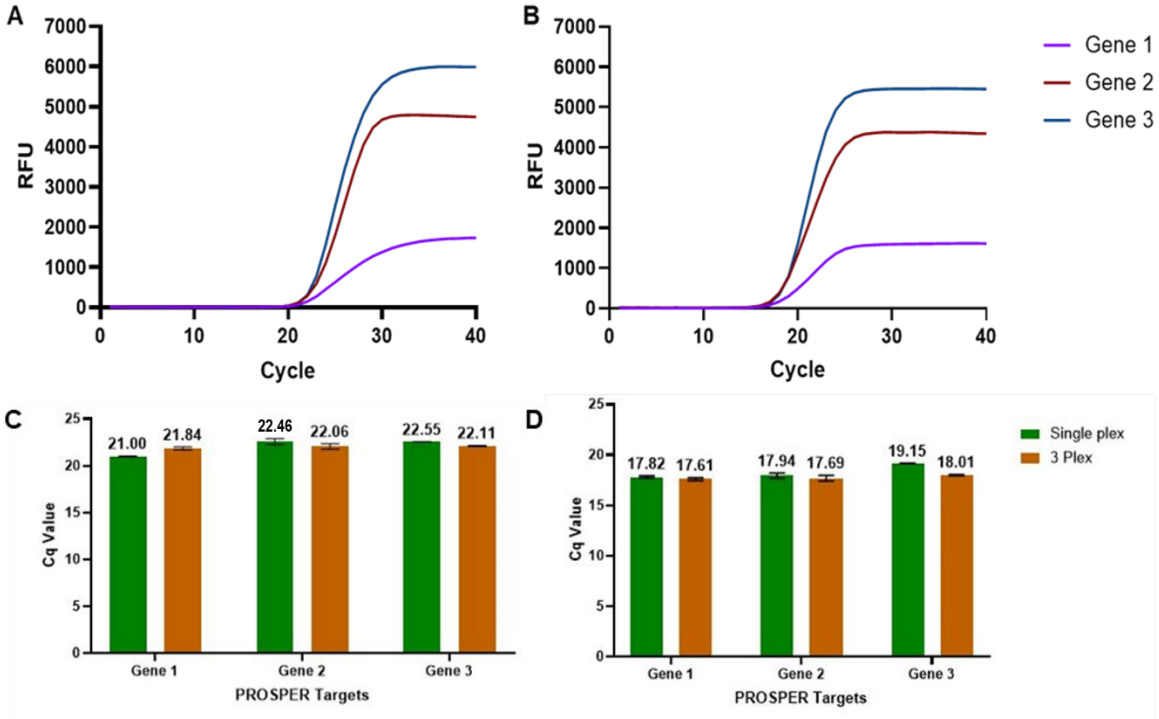


Figure 3.9: Compatibility of the 3-plex assay on TNA samples from flash-frozen T-47D cell lines. Amplification plots for (A) low and (B) high TNA copy numbers respectively, were inputted for the PCR reactions respectively. Mean Cq values comparing single plex and 3-plex assay in (C) low and (D) high copy number sample. Each assay ran in duplicate.

3.5.3 Performance of assay in 4-plex format

The 4-plex qPCR assays of the PROSPER targets and the normalising genes were tested with low and high copy number TNA samples extracted from flash-frozen T-47D cells. Fluorescence channels FAM, CY5, TX4 and HEX were used for Gene 1, Gene 2, Gene 3, and Gene 4 respectively. The 4-plex assays showed a consistent increase in fluorescent intensity over time for all 4 targets and reached exponential amplification in both low and high copy number samples (**Figure 3.10 A and B**). For the no template control, no amplification was recorded.

In the low and high copy number samples, the mean Cq values for the single-plex assays were higher than those in the 4-plex reactions (**Figure 3.10 C and D**). Consistent Cq values were seen

in a single target and 3-plex assays for the four genes tested. This was true for both the low copy number and high copy number template input (**Figure 3.10 C and D**).

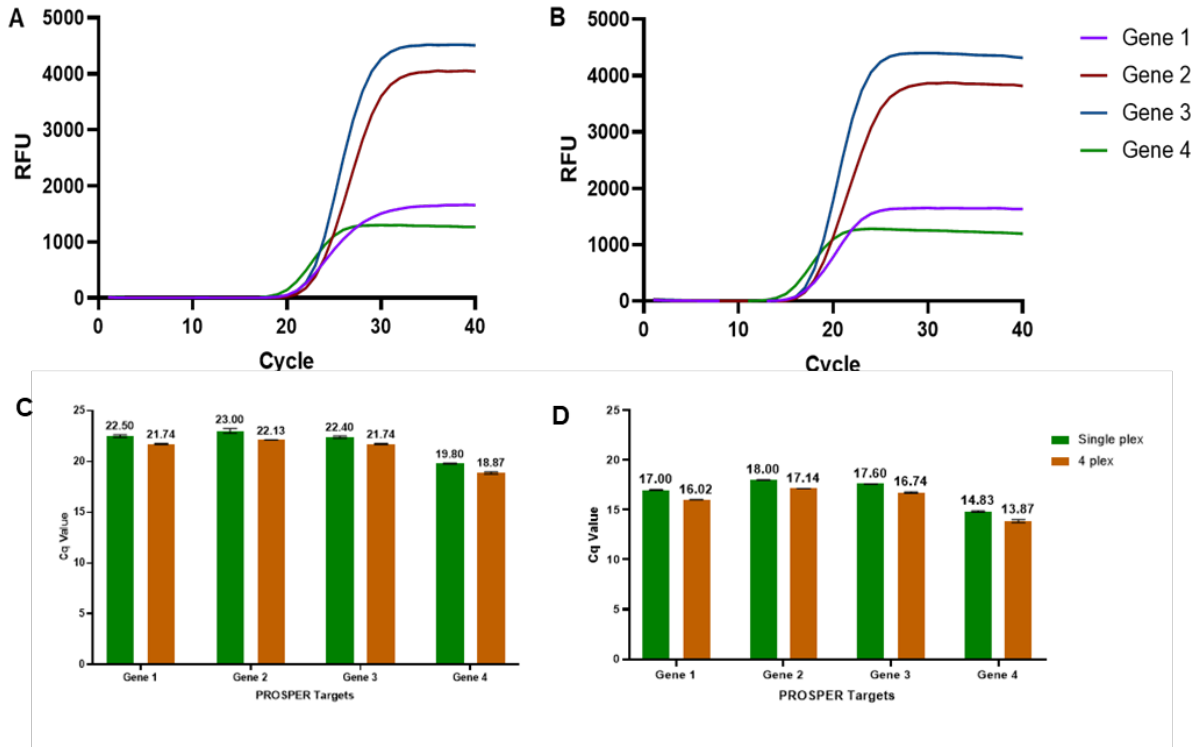


Figure 3.10: Compatibility of the 4-plex assay on TNA samples from flash-frozen T-47D cell lines. Amplification plots for (A) low and (B) high TNA copy number qPCR assays. Mean Cq values comparing single and 4-plex assay in (C) low and (D) high copy number samples. Each assay ran in duplicate.

3.5.4 Performance of assay in 5-plex format

The 5-plex assay with the PROSPER target and either RNA HK Gene 1 or NED Gene 1 was optimised in two separate reactions on TNA samples from flash-frozen T-47D cells **(Figure 3.11**). Fluorescence channels FAM, CY5, Quasar and HEX were used for Gene 1, Gene 2, Gene 3, and Gene 4 respectively. Fluorescence channel TX4 was used for RNA HK Gene 1 and NED Gene 1. Both 5-plex assays exhibited a consistent increase in fluorescence intensity over time for all five targets in both low and high copy numbers suggesting exponential amplification of target genes at both low and high abundance. For the no template control, no amplification was recorded.

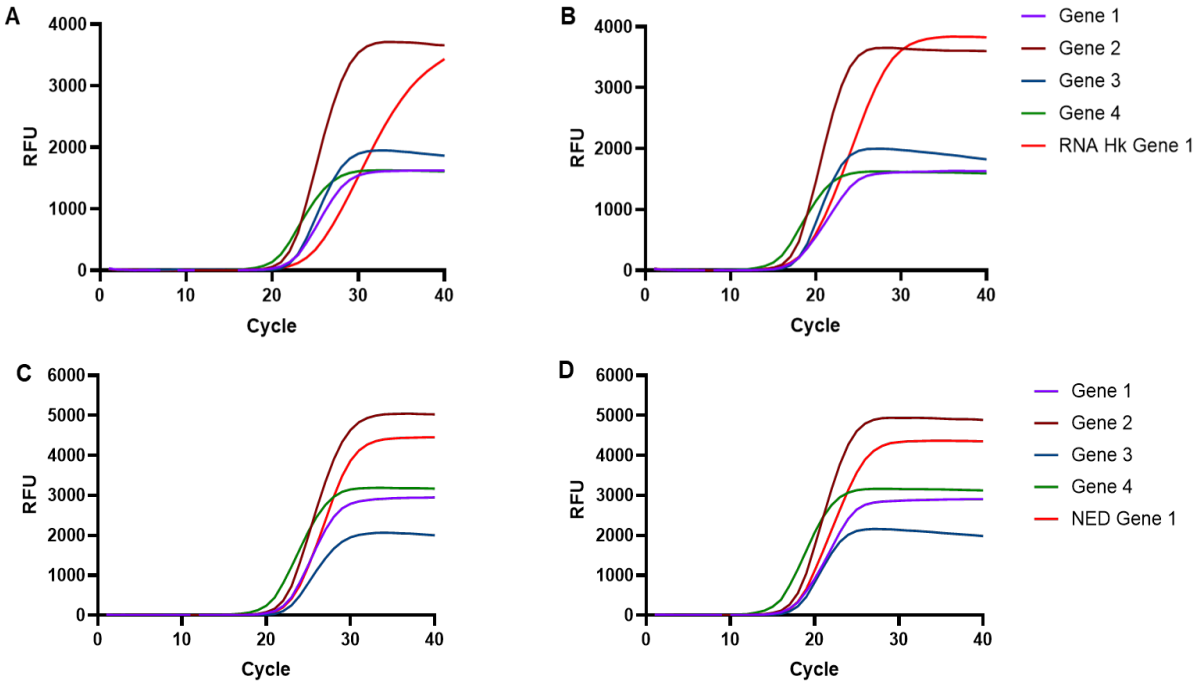


Figure 3.11: Compatibility of 5-plex assay on TNA samples from flash-frozen T-47D cells. Amplification plots RNA-based assay for (A) low and (B) high copy number. Amplification plots TNA-based assay for (C) low and (D) high copy number. Each assay ran in duplicate.

3.5.4.1 Performance of RNA-based assay

Figure 3.12 compares the mean Cq values of the RNA-based assay. Overall, low and high copy number samples yielded later and earlier Cq values of all targets, respectively. In both copy number samples, the mean Cq values of the PROSPER targets slightly decreased from a 4-plex assay to a 5-plex assay **(Figure 3.12 A and C)**, suggesting that the inclusion of the RNA HK Gene 1 affected the efficiency of PROSPER targets. However, the difference was less than one cycle difference. Similarly, RNA HK Gene 1 yielded slightly decreased Cq in the 5-plex assay compared to the single-plex assay, suggesting a reduced amplification efficiency (**Figure 3.12 B and D)**. However, this reduction was minimal.

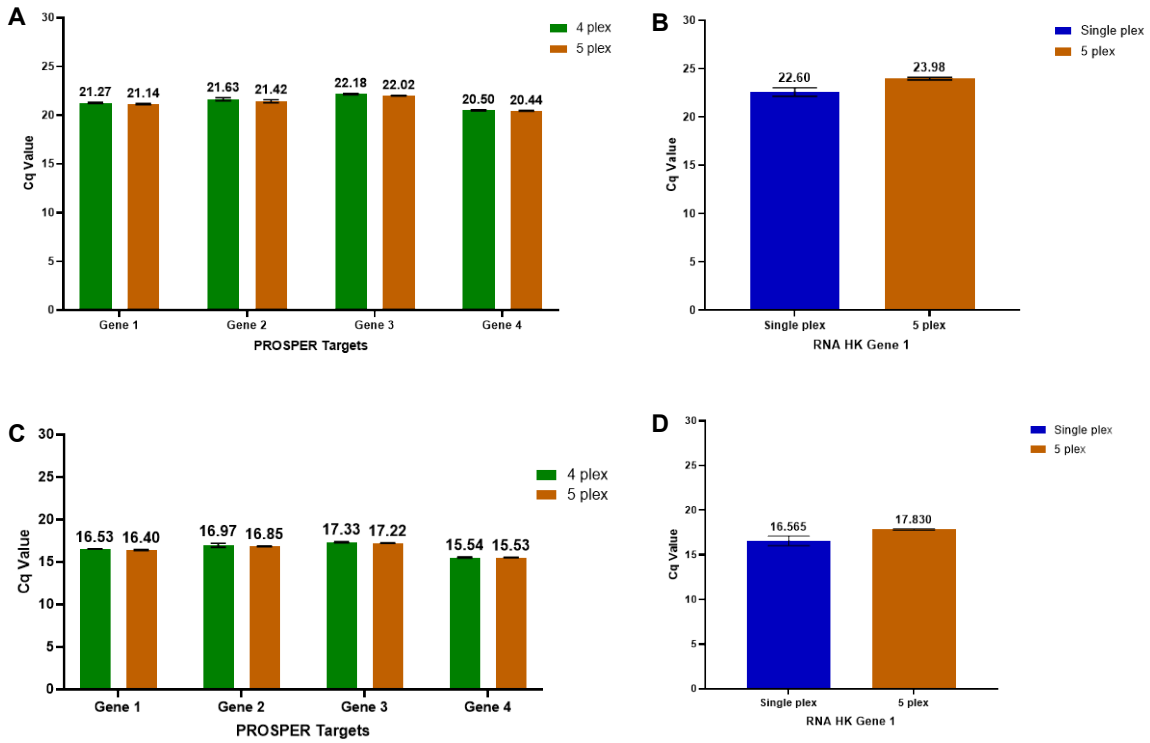


Figure 3.12: Comparison of Cq values of RNA-based assay. Bar charts illustrating the mean Cq values of the (A) PROSPER targets in 4-plex and 5-plex assays and of (B) RNA HK Gene 1 in a single plex and 5-plex assays in low copy number samples. Bar charts illustrating the mean Cq values of the (C) PROSPER targets in 4-plex and 5-plex assays and of (D) RNA HK Gene 1 in a single plex and 5-plex assays in high copy samples.

3.5.4.2 Performance of TNA-based assay

Figure 3.13 compares the mean Cq values of the TNA-based assay in 4-plex and 5-plex assays. The results were very similar to the RNA-based assay. Overall, low and high copy number samples yielded later and earlier Cq values of all targets, respectively. In both copy number samples, the mean Cq values of the PROSPER targets slightly decreased from a 4-plex assay format to a 5-plex assay format (**Figure 3.13 A and C**), indicating that the inclusion of the NED Gene 1 affected the efficiency of PROSPER targets. However, the difference was less than one cycle difference. Similarly, NED Gene 1 yielded slightly decreased Cq values in a 5-plex assay format compared to a single-plex assay format indicating a reduced amplification efficiency (**Figure 3.13 B and D)**. However, this reduction was minimal.

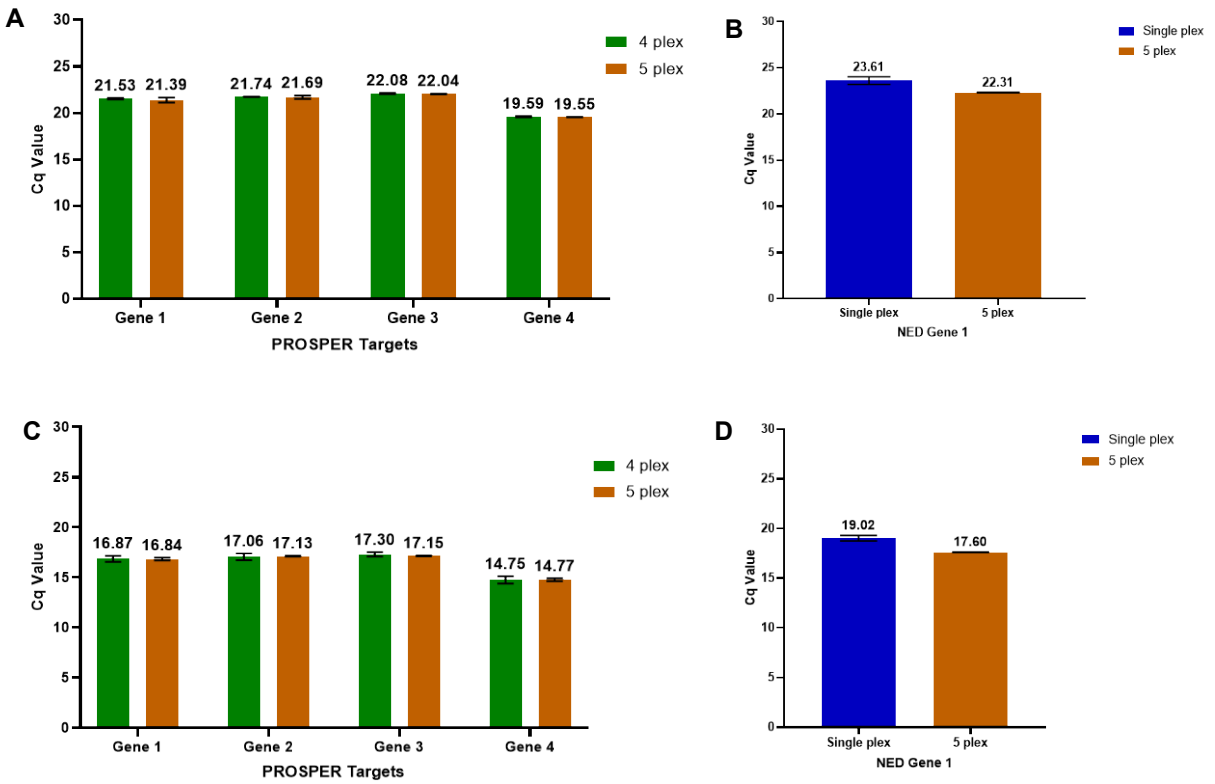


Figure 3.13: Comparison of Cq values of TNA-based assay. Bar charts illustrating the mean Cq values of the (A) PROSPER targets in 4-plex and 5-plex assays and of (B) NED Gene 1 in single plex and 5-plex assay in low-copy number samples. Bar charts illustrating the mean Cq values of the (C) PROSPER targets in 4-plex and 5-plex assays and of (D) NED Gene 1 in a single plex and 5-plex assays in high copy samples.

3.5.5 Measurement of DNA and RNA in the PROSPER gene signature

To ensure the measurement of DNA and RNA, Cq values were compared in RNA-based and TNA-based assays with and without reverse transcription (RT) on a TNA sample extracted from flash-frozen T-47D cells. The difference in Cq values, denoted as ΔCq, was used to estimate the level of RNA present in the sample. An amplification plot without RT reactions demonstrated that the assays were detecting both DNA and RNA.

In both copy number reactions, reactions performed without addition of reverse transcriptase enzyme resulted in delayed Cq values for all targets compared to reactions performed with reverse

transcriptase in both RNA (**Table 3.2**) and TNA-based (**Table 3.3**) assays. The ΔCq results of the RNA-based qPCR assay illustrated in **Figure 3.14A** showed that RNA HK Gene 1 demonstrated the highest transcript expression, followed by Gene 4, Gene 1, Gene 2, and Gene 3. Gene 3 had less than one ΔCq value difference, indicating low RNA expression of this target in this sample. Similarly, the TNA-based qPCR assay **(Figure 3.14B)** demonstrated that Gene 4 exhibited the highest RNA transcription, followed by Gene 1, Gene 2, and Gene 3 in both low and high copy number samples. NED maintained a very low ΔCq value of 0.53 and 0.32 in low and high copy number samples, respectively. This is expected as NED was designed only to target DNA levels.

Table 3.2 ΔCq values from 5-plex PCR assay for the detection of RNA transcript levels in the PROSPER gene signature with RNA HK Gene 1

		(+) RT Mean Cq	(-) RT Mean Cq	ΔCq	
				Avg	Std dev
Low copy number	Gene 1	19.50	22.12	2.62	0.13
	Gene 2	20.27	21.43	1.16	0.05
	Gene 3	21.24	22.14	0.90	0.17
	Gene 4	17.53	20.94	3.41	0.50
	RNA Hak Gene 1	18.765	25.19	6.42	0.02
High copy number	Gene 1	15.16	17.35	2.20	0.01
	Gene 2	15.89	17.10	1.21	0.04
	Gene 3	16.535	17.39	0.85	0.02
	Gene 4	12.83	16.16	3.33	0.01
	RNA Hk Gene 1	13.295	19.73	6.44	0.02

(+) RT = with reverse transcriptase; (-) RT = without reverse transcriptase; ΔCq = (-) RT - (+) RT

Table 3.3 ΔCq values from 5-plex PCR assay for the detection of RNA and DNA transcription in the PROSPER gene signature with NED Gene 1

		(+) RT Mean Cq	(-) RT Mean Cq	ΔCq	
				Avg	Std dev
Low copy number	Gene 1	19.42	22.12	2.70	0.04
	Gene 2	19.99	21.49	1.50	0.01
	Gene 3	21.35	21.88	0.53	0.02
	Gene 4	17.38	20.60	3.22	0.12
	NED Gene 1	22.48	22.81	0.33	0.01
High copy number	Gene 1	15.19	17.28	2.26	0.04
	Gene 2	15.655	17.04	1.39	0.02
	Gene 3	16.61	17.38	0.76	0.01
	Gene 4	12.895	16.02	3.12	0.09
	NED Gene 1	17.88	18.20	0.32	0.01

(+) RT = with reverse transcriptase; (-) RT = without reverse transcriptase; ΔCq = (-) RT - (+) RT

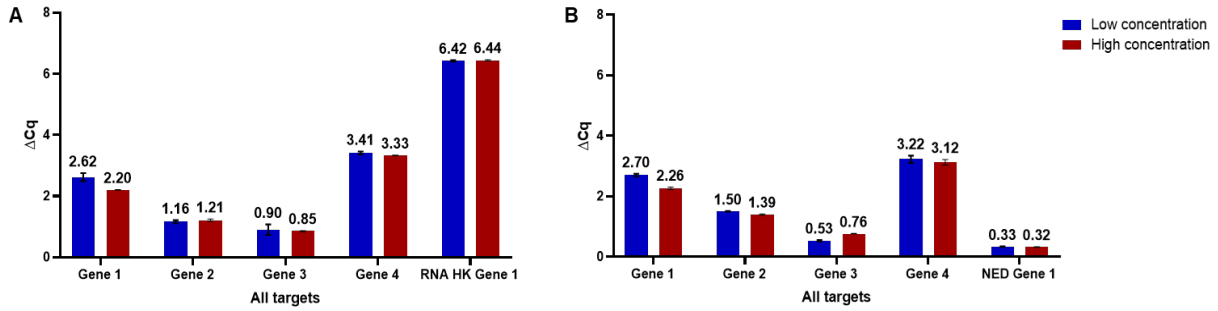


Figure 3.14: Comparison of ΔCq with and without reverse transcriptase reactions. Bar charts representing the ΔCq values of the (A) RNA-based and (B)TNA-based assay on reaction with and without reverse transcriptase. The experiment was conducted on low and high copies of TNA samples from flash-frozen T-47D cell lines.

3.5.6 Validation of the 5-plex assay in cell line TNA samples

The two 5-plex assays with RNA HK Gene 1 and NED Gene 1 were tested on various samples to assess the robustness and broader applicability of the assay. Fluorescence channels FAM, CY5, Quasar, HEX and TX4 were used for Gene 1, Gene 2, Gene 3, Gene 4 and RNA HK Gene 1/NED Gene 1 respectively. Both assays were tested on undiluted TNA extracted from human blood, Hep G2 and K562 cell lines (**Figure 3.15**). Both 5-plex assays exhibited a consistent increase in fluorescence intensity over time for all 5 targets in both low and high copy number reactions across all the samples. As typical, a sigmoid-shaped curve with a defined exponential phase was displayed. For the no template control, no amplification was recorded. A greater degree of variability was observed in the RNA HK Gene 1 than in the DNA control, consistent with the fact that the housekeeping gene was selected based on its low variance of expression between breast cancer tissues, however its expression could be expected to vary more between differing tissue types than between samples of the same tissue type, whereas genomic DNA levels would be expected to remain constant.

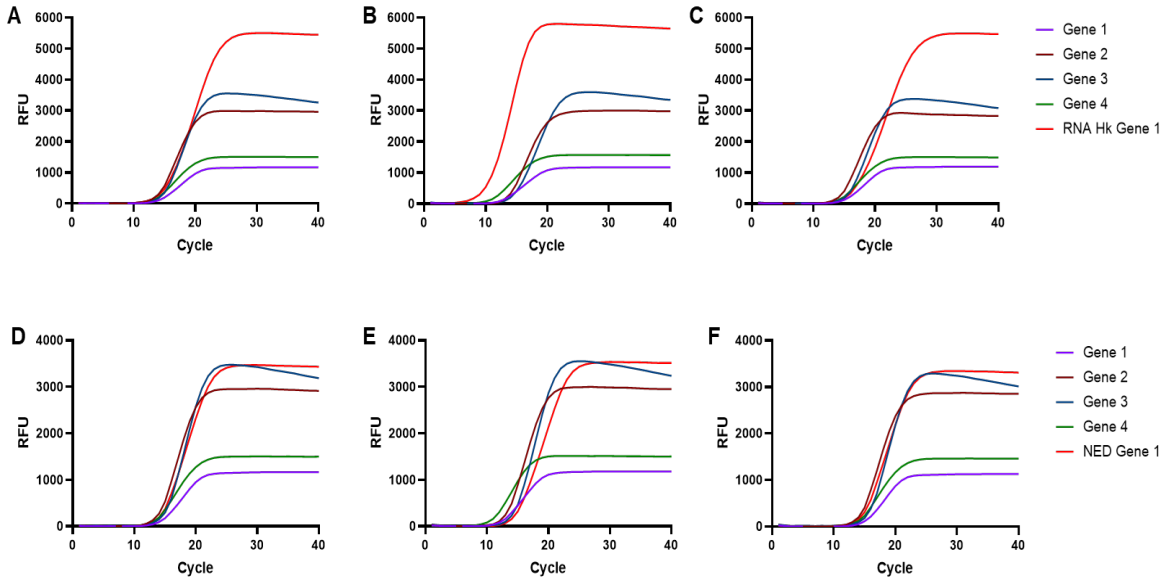


Figure 3.15: Performance of the 5-plex assay across various sample types.(A – C) Amplification plots of the RNA-based assay tested on undiluted TNA samples from human blood, Hep G2 and K562 cell lines respectively, provided by SpeeDx. (D – F) Amplification plots of the TNA-based assay were tested on the same types of samples. Each assay ran in duplicate.

Next, the multiplexed qPCR assays were tested on the ORG2058-treated T-47D cell line to evaluate the performance of the PROSPER gene signature. ORG2058 is a synthetic progestin, which is known to be growth inhibitory in T-47D cells at 48h treatment (118). The PROSPER index was calculated based on the combined estimates of expression of the five genes. The precise algorithm is subject to intellectual property protection and not able to be disclosed. Given that the PROSPER signature is made up of proliferation-associated targets, it was expected that ORG2058 would decrease the PROSPER signature in these cells. The results indicated that the assay can detect differential expression of the PROSPER genes between treated and untreated cells. In both TNA-based and RNA-based assays, a significantly reduced PROSPER index was recorded for the

drug-treated samples (Unpaired two-tailed t-test, $p < 0.0001$, **Figure 3.16**). The range of values of the PROSPER index returned after normalising with the RNA HK Gene 1 was different to those yielded after normalising to NED, as would be expected given their different abundance levels. However, both demonstrated a relative decrease in the PROSPER index in progestin treated cells compared to control. This demonstrated robust performance of the assay with both normalisation methods but indicates that the relationship between numerical PROSPER index value and risk of recurrence in ER+ breast tumour tissues will be different for the two methods.

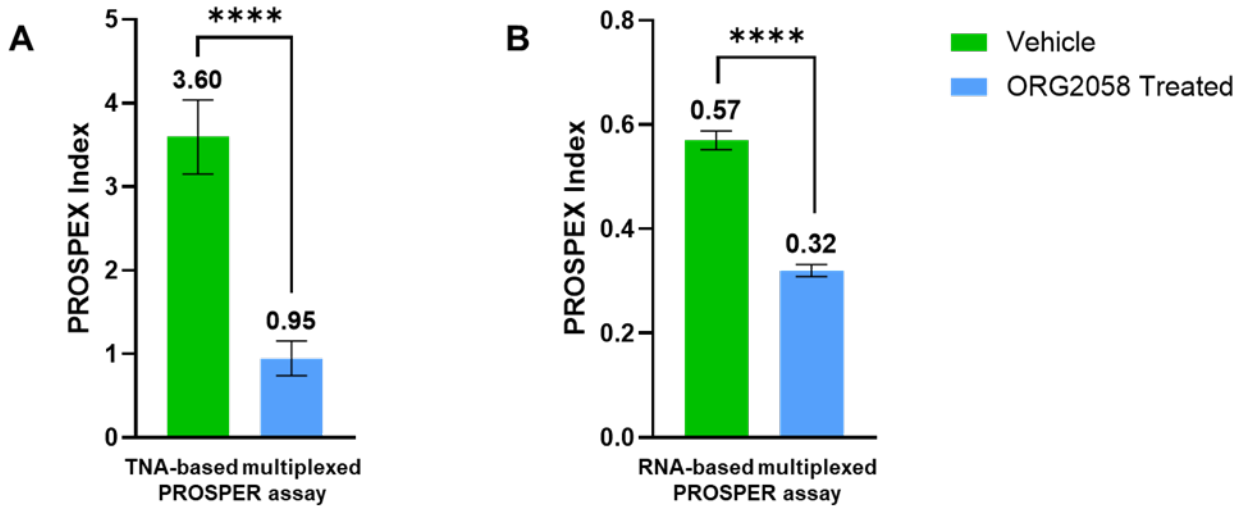


Figure 3.16: Validation of PROSPER multiplexed qPCR assay. The multiplexed assay was tested on duplicate experiments on ORG2058-treated T-47D cells. Bar charts illustrating the PROSPER index computed from the expression obtained from the PROSPER multiplexed TNA-based (A) and RNA-based (B) qPCR assay. TNA and RNA were extracted from the cells and the undiluted sample was assessed using the multiplexed qPCR (n = 3). The data were analysed using the Unpaired two-tailed t-test. **** indicate $p<0.0001$.

3.6 Discussion

3.6.1 Summary of Results

In this chapter, a 5-plex real-time qPCR assay was successfully developed to detect and quantify the transcription of the genes in the PROSPER gene signature. Since SpeeDx specialises in TNA-based qPCR assays, a TNA-based assay, as well as an RNA-based assay was developed to compare the traditional housekeeping RNA-based normalisation method to the DNA-based normalisation method used in the TNA-based qPCR assay. This was further discussed in **Chapter 4**.

The assay compatibility was assessed from single plex to 5-plex assays by comparing and analysing the amplification plots and Cq values during different transitions of the assay. The single-plex Cq values were generally higher than the 5-plex Cq values, indicating a more efficient amplification when multiple targets were combined in a single reaction. However, the differences in Cq values were small between the single plex and the 5-plex assays. These results suggested that the 5-plex assay provided accurate quantification. When the RNA HK Gene 1 or the NED Gene 1 were incorporated as the 5th target to the qPCR assay, a minimal reduction in the Cq values of the PROSPER targets was observed compared to the 4-plex assay, demonstrating that the inclusion of the normalising gene targets did not compromise the performance of the other gene targets in the 5-plex assay format.

The transcriptional analysis using the ΔCq method revealed that the RNA-based assay exhibited varying levels of RNA expression. Gene 4 and the RNA HK Gene 1 showed the highest transcription, while Gene 2 and Gene 3 had the lowest in both low and high copy number samples. The TNA-based assay showed similar trends, with Gene 4 exhibiting the highest RNA transcription, followed by Gene 1, Gene 2, Gene 3, and NED Gene 1 (which targets only DNA).

Finally, the 5-plex assays were validated on various sample types, including cell lines and human blood. The assays demonstrated consistent and robust performance across all samples, confirming their suitability for clinical use.

3.6.2 Computational tools for optimising primer design for multiplex qPCR assays

A successful PCR relies on the carefully designed synthetic oligonucleotide primers component of the assay (119). Primer design parameters including primer sequence, length, and melting temperature play a crucial role in the efficiency and specificity of the PCR reaction. Multiplexing PCR assays, where multiple target sequences are simultaneously amplified and detected in a single reaction, can be challenging (120). Multiplexing of PCR amplification requires that several primer pairs be included in a single reaction mix and optimally bind to their respective target region. Incorrect design of primers can lead to mispriming of the target DNA or the formation of primer dimers and cause nonspecific amplification (120). Additionally, it is important that the amplification of two or more targets does not preferentially amplify one of the targets.

Consequently, it is important to utilise computational tools as it aids in designing primers with high specificity to the target sequence, minimising non-specific amplification (121). Furthermore, utilising computational tools such as Primer-BLAST (116) and FastPCR (122) allow for rapid screening of primer pairs from vast databases of genetic sequences. Specifically, they greatly assist in the design of primers for multiplex PCR reactions, where multiple target sequences are amplified simultaneously, which is complex and would be an exceedingly difficult and time-consuming process without computational assistance (121).

In this study, extensive *in silico* primer design and optimisation were performed to develop the robust 5-plex assays. A comprehensive primer design workflow was devised and implemented, which incorporated computational tools, which were utilised over a range of settings from single plex to multiplex qPCR assays. This aligns with the sentiments expressed in the literature that PCR primer design is a critical step that requires significant optimisation and experimental validation, and that there is no one-size-fits-all solution. The Primer-BLAST algorithm was utilised to design primer sequences that can specifically bind to target genes (122). However, Primer-BLAST is incapable of assessing primer dimer formation in a multiplexed reaction and it does not check for thermodynamic stability and other parameters known to influence primer performance (123). Therefore, FastPCR was also implemented in the primer design workflow. FastPCR software allows simultaneous primer pair design and evaluation of their suitability for multiplex PCR, taking into consideration factors like primer self-complementarity, melting temperature, and potential for cross-reactivity, thereby optimising the multiplexed assays before experimental validation (122, 124). The optimal oligonucleotides were further assessed in OligoAnalyzer which provides a more comprehensive analysis for secondary structures, self-dimerisation and cross-dimerisation (123, 125). Overall, the primer design workflow employed in this study was well-aligned with the best practices recommended in the literature.

3.6.3 Experimental validation of *in silico* designed primers

While computational tools such as FastPCR and OligoAnalyzer are powerful tools in designing and analysing PCR primers, it is important to ensure assay performance by conducting validation and optimisation experiments. This is because, in experimental conditions, the assay performance can vary depending on the extraction methods used to purify the templates and the reagents used

for the PCR reaction (119). Therefore, it is essential to experimentally validate and consequently optimise the *in silico* designed primers.

In this research, all primers were designed *in silico,* and they were rigorously evaluated, including their performance in single-plex, duplex, and 5-plex qPCR assays. It is generally reported in the literature that the multiplexed PCR assays may display low amplification efficiency compared to single plex, this is because the presence of multiple primer sets in the reaction can result in competition of reaction components such as dNTPs, polymerase enzymes, and buffer components (126-128). However, in this research, lower mean Cq values were consistently observed when transitioning from single plex to 3-plex, 4-plex and finally 5-plex qPCR assays. These results were true for both high and low copy number samples. This is mainly because a thorough optimising of the primer design and the enhanced reaction conditions minimised the cross-reactivity and potential inhibitory interactions among amplicons (128). Furthermore, incorporating ***Plex****Zyme* technology which has already undergone extensive optimisation and validation (115) allowed to overcome the typical challenges associated with traditional multiplexed PCR such as competition for reagents and differential amplification rates among targets. In contrast, changes to individual reaction components such as magnesium concentration, enzyme choice or concentration of reaction mix, had a limited impact on product yields and for that reason, were not included in the results. A recent study by Park et al. (2020) demonstrated that a thorough systematic step-by-step process allowed for comprehensive optimisation of primer design and reaction conditions which led to the development of a successful multiplexed assay that performed better than in single plex qPCR (129). In addition, it is possible that the presence of multiple primer sets in the same reaction can have synergistic effects, where the amplification of one target facilitates the amplification of another. In some cases, simultaneous amplification may reduce the overall effect of certain

inhibitors present in the reaction, resulting in a more efficient amplification. It should be noted that the amplification efficiency of each of the target genes was not estimated from standard curves, either separately or in multiplex combinations. Comparison of the amplification curves demonstrated a flatter geometric amplification slope for Gene 1 and Gene 4, suggesting a lower amplification efficiency than the other targets. This was consistent in all experiments. So, although this difference could affect Cq estimates, provided that the same estimates are consistently obtained for each of the targets and taken into account in the coefficients used to estimate the PROSPER index, the same level of accuracy could be expected for predicting risk of recurrence. However, a formal estimate of efficiency should be performed in future.

3.6.4 Detection of nucleic acids: TNA-based and RNA-based assays

The detection of the appropriate nucleic acid was also a crucial objective for this study as two types of multiplexed assays were developed: TNA-based and RNA-based assays. The ΔCq between (+) and (-) RT reactions was calculated to ensure the expected detection of both RNA and DNA within the relevant assays. The results revealed that the RNA-based test effectively identified RNA transcription for all examined genes. Gene 1 exhibited the highest level of transcription, followed by Gene 4, Gene 2, and then Gene 3. This variation in transcription activity emphasized the RNA-based assay's precision and accuracy in measuring RNA levels from the PROSPER gene signature. The same trend was observed for the PROSPER target using the TNA-based test. Additionally, the ΔCq values of Gene 3 for both RNA-based and TNA-based assays were less than one, indicating that there was very low expression RNA for this target. This was true for both low and high copy number samples. The presence of an amplification in (-) RT reaction demonstrated that the TNA-based assay also detected DNA. As expected, NED Gene1 which targets non-expressed DNA displayed no significant transcription with the ΔCq values of less than 1 in both

copy number samples, demonstrating its specificity for detecting only DNA. These results indicated that both RNA and TNA-based assays can reliably detect RNA and DNA targets to assess gene transcription within the PROSPER gene signature, as this is important for the normalising method to commute the PROSPER index (Discussed in **Chapter 4**). Furthermore, when both assays were tested against different sample types other than T-47D cell lines, such as human blood, Hep2 and K562 cell lines, the assays performed efficiently indicating the broad applicability of the designed qPCR.

In conclusion, the RNA and TNA-based multiplexed PCR assays developed in this study demonstrate robust and reliable detection of the PROSPER gene signature, with the ability to detect RNA and DNA targets.

3.6.5 Assay validation with synthetic progestin treatment

Breast cancer growth inhibition by progestin in tissue culture has been well documented, and this is due to the reduced rate of cell entry into the S phase, which is where the DNA replication occurs in the cell (130). Since the PROSPER signature consists of proliferation genes, T-47D cells were treated with synthetic progestin ORG2058 for 48 hours to validate the multiplexed qPCR assays' performance and the ability to differentiate the difference proliferation levels. The results demonstrated a significant decrease in the expression of the PROSPER genes in the drug-treated samples, confirming the assays' ability to estimate proliferation levels. This validation was critical since the assay will be utilised to stratify patients with ER+ breast cancer into low and risk groups which is primarily driven by the proliferation status of the tumours.

3.6.6 Limitations and future directions

While the multiplexed qPCR assays developed in this study have shown promising results, some limitations need to be addressed in future work. Firstly, a qualitative assessment of the qPCR amplification was performed based on the analysis of amplification curves. However, conducting a quantitative PCR efficiency test is critical for critical for validating the performance of the assay to assess the accuracy and reproducibility of the amplification on all five targets. Therefore, to further strengthen the finding from this chapter, future studies should include PCR efficiency experiments which will provide a quantitative evaluation of assay performance and its potential clinical utility. Additionally, the assays were validated only in cell line models and not in clinical samples. While cell line models provide a controlled environment for assay development, their biological relevance to patient-derived samples has limitations. Specifically, cell lines, when cultured in vitro, are not a true representative of the tumour heterogeneity and the complex tumour microenvironment found in patient samples. This is due to the lack of influence of cytokines and other cell signalling pathways (131). To address this limitation, future studies should focus on validating the multiplex assays in well-annotated clinical cohorts of breast cancer patients with long-term clinical follow-up data. This was explored in **Chapter 4**.

A key objective in the development of the PROSPER multiplexed assay is its translation to clinical application. For this objective to be met, the assay must be easily and reproducibly performed, incorporating as much automation as possible. Streamlining the assay protocols and automation of sample processing steps will further enhance the assay to be implemented for routine clinical use (132). Additionally, a thorough evaluation of the long-term stability and reproducibility of the assays is necessary to ensure consistent performance over an extended period. Assessing the

stability of reagents, including primers, probes, and enzymes, under various storage conditions, and evaluating assay reproducibility across different operators and laboratories will be crucial and needs to be thoroughly investigated before the assay can be introduced into routine clinics (121, 133).

3.6.7 Conclusion

In summary, the development and validation of multiplexed PCR assays targeting the PROSPER gene signature in this chapter demonstrated the potential for a feasible and affordable test for risk categorisation for ER+ breast cancer. Once the limitations of this chapter are addressed, this multiplexed PROSPER test can be implemented as a valuable test in routine clinical practice. The comprehensive evaluation of these assays for specificity, sensitivity, and reproducibility underscores their potential clinical utility. By leveraging qPCR technology (121), a widely adopted diagnostic tool in the clinical setting, the PROSPER signature can be translated into a cost-effective and rapid clinical test. The streamlined protocols associated with qPCR enable efficient processing, allowing for timely results that are compatible with breast cancer patient treatment pathways.

The incorporation of ***Plex****Zyme* technology by SpeeDx (115) reduced the challenges typically encountered in traditional multiplexed PCR assays and facilitated the optimisation of four PROSPER genes within a single multiplexed assay alongside a normalising gene. This innovative approach enhanced the feasibility and practicality of implementing the PROSPER signature in clinical practice.

Moving forward, future validation studies in clinical samples(131), assessment of long-term stability (121), and exploration of additional gene targets will further refine the utility and

robustness of these multiplex PCR assays. Ultimately, the integration of the PROSPER signature into routine clinical practice has the potential to improve patient outcomes by enabling low-cost personalised treatment strategies in ER+ breast cancer management.

Chapter 4

Validation of PROSPER multiplexed qPCR assay in FFPE tissues

4.1 Introduction

Chapter 3 demonstrated the feasibility of integrating the PROSPER gene signature into a multiplexed PCR test where all the targets were assessed in a single reaction. Two assays were developed, TNA and RNA-based assays which employed the NED Gene 1, a non-transcribed DNA and the RNA HK Gene 1, a housekeeping transcript, as their normalisation controls, respectively. However, optimisation and validation of the multiplexed assay within FFPE samples are required to determine the performance and the clinical utility of the assay for the assessment of the PROSPER index.

The importance of comprehensive evaluation for multigene tests is emphasised by their profound impact on clinical decision-making and the shaping of public health policy. Commercially available precision medicine tests are significantly more accurate than traditional histopathological approaches, but they come at a considerable cost to the patient and healthcare system. The reliability and accuracy of these tests are crucial as they directly influence adjuvant treatment recommendations. False-negative results from these assays could delay crucial therapeutic interventions, such as treatment escalation to adjuvant chemotherapies, thus impacting patient outcomes. Similarly, false positives could lead to the administration of unnecessary and ineffective treatments, highlighting the vital need for rigorous analytical performance evaluation before precision medicine tests can be clinically implemented. Such evaluations are paramount in

validating the reliability and effectiveness of prognostic tests like PROSPER, which seeks to contribute significantly to personalised patient care.

Once clinically available, the intended application of the PROSPER assay will be with routinely prepared FFPE tissues. Globally, FFPE tissue samples are the most widely available and commonly used source of biological material for molecular applications, from exploratory research to diagnostics (134-136). These blocks are valued for their ability to preserve the cellular, architectural, and morphological characteristics of the biospecimen over extended periods of time at room temperature, offering a durable resource for mutation analysis as well as validation of molecular tests (134, 135). Furthermore, the extensive FFPE archives available in hospitals, private pathology facilities, biobanks, and tertiary academic pathology departments represent a valuable resource for molecular testing, offering a vast collection of samples ready for analysis. The compatibility with FFPE samples positions the PROSPER assay within a conventional diagnostic framework, making the most of the prevalent use and historical significance of FFPE tissues to ensure its applicability for cost effective, standardised assay for routine clinical utility.

Unfortunately, utilising FFPE specimens in genomic assays is difficult due to the degraded and limited yield of nucleic acids extracted from FFPE blocks. While FFPE sections provide an excellent medium for the assessment of cellular proteins by immunohistochemical tests, DNA and RNA are poorly preserved and more challenging to evaluate (137, 138). During the fixation process, DNA undergoes chemical modifications, entrapment, and fragmentation due to extensive cross-linking with protein, and the paraffin in DNA extracts can inhibit PCR amplification reactions (139). RNA is even more adversely affected by formalin fixation, undergoing extensive degradation, chemical alterations, damage to the poly-A tail, and covalent modifications that can

impact the reverse transcription from mRNA to cDNA and thereby significantly impact PCR amplification.

Despite these limitations, advancements in PCR techniques over the past decades have demonstrated the potential for quantifying nucleic acids in FFPE tissues. These advanced PCR methods offer sensitivity, precision, and the capability to analyse multiple targets, thus making them suitable for broader application in clinical research, diagnostic and prognostic assays(139).

This chapter explored the optimisation of methods for extraction of nucleic acids (TNA and RNA) from FFPE samples for application to measuring the PROSPER index as a clinically applicable assay. The role of normalisation in ensuring the accuracy of gene expression analysis in a multiplexed qPCR assay is also investigated.

4.2 Aims and hypotheses

4.2.1 Aims

1) To identify and assess the effectiveness of commercially available extraction kits for isolating TNA and RNA from FFPE tissue.

2) To evaluate the performance of the multiplexed PCR assays when applied to TNA and RNA extracted from FFPE samples.

3) To validate the PROSPER index obtained from the multiplexed PCR assays against other established markers such as the EndoPredict EPclin, the PROSPER index measured on the NanoString platform, and proliferation marker Ki-67.

4) To conduct a comparative analysis of the PROSPER index using RNA HK Gene 1 and NED Gene 1 as normalisation controls, to determine the most reliable method for accurate gene expression analysis in a multiplexed setting.

4.2.2 Hypotheses

It was hypothesised that both TNA and RNA-based PROSPER multiplexed assays will perform well on TNA and RNA extracted from FFPE tissues, respectively. The PROSPER index obtained from both multiplexed assays will be correlated well to the PROSPER index measured on the NanoString platform and Ki-67. Additionally, between the TNA and RNA-based PROSPER tests, the TNA-based assay will be the most robust and reliable method for accurate assessment of PROSPER gene expression.

4.3 Approach

The primary focus of this chapter was the optimisation of the performance of the PROSPER RNA and TNA-based multiplexed PCR assay on FFPE samples using a subset of FFPE samples from Cohort 1 and Cohort 2 to assess its effectiveness in a real-world setting. The PROSPER targets were detected on FAM, CY5, HEX, and Quasar 705 fluorescence channels. The normalising controls were detected on the Texas Red fluorescence channel. The optimisation experiments were conducted on multiple representative patient samples from Cohort 1 and 2 to ensure that the observed results were consistent across different patient samples. Due to the limited availability of these precious samples, direct replicates were not feasible at this stage of the experimental work. However, following successful optimisation, the optimised multiplexed assay was conducted on the full patient cohort in which three replicates were performed.

The initial step involved identifying and refining the optimal methods for extracting TNA and RNA from FFPE samples. The optimisation process of the FFPE extraction method involved the examination of key factors such as the initial quantity of material and the incubation times needed to increase the nucleic acid yield. TNA extraction kits, the MasterPure Complete DNA and RNA Purification Kit (MasterPure kit) from Lucigen and the Quick-DNA/RNA FFPE Kit (Quick kit) from Zymo Research were evaluated. The workflow of the kits is briefly illustrated in (**Figure 4.1**)

The subsequent phase involved evaluating and improving the multiplex assays' performance with FFPE tissue samples. This step involved adjusting gene target concentrations, reevaluating primer designs, and ensuring consistent amplification of the PROSPER targets and the normalising control targets within a multiplexed reaction.

Finally, a comparative analysis was conducted to evaluate the PROSPER index derived from a subset of Cohort 1 samples. This analysis was aimed at understanding the impact of using either RNA HK Gene 1 or NED Gene 1 as normalisation controls on the computed PROSPER index.

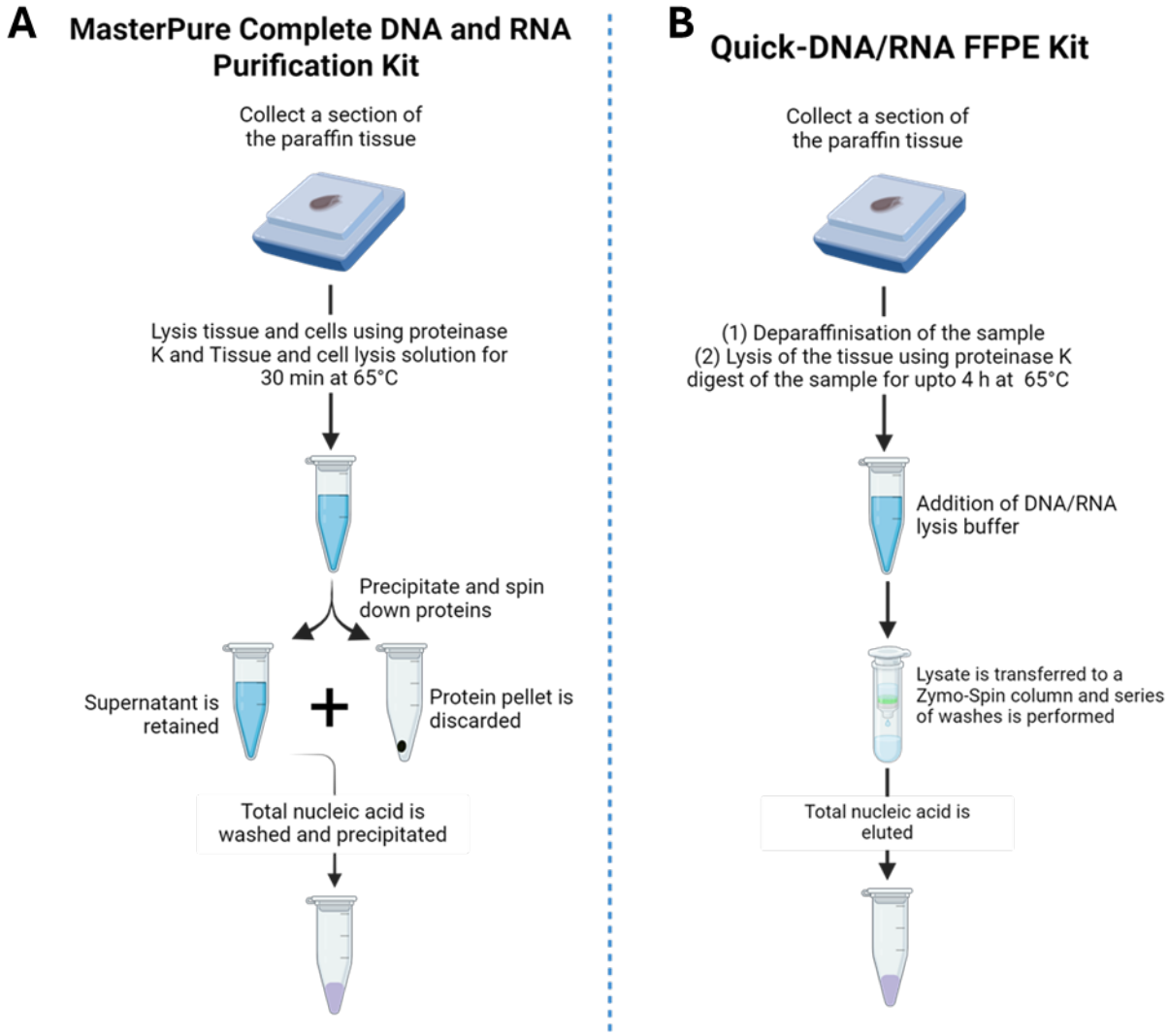


Figure 4.1: Workflow of the two commercial FFPE extraction kits. (A)The MasterPure Complete DNA and RNA kit follows a precipitate-based extraction method where the pelleted protein is discarded and from the supernatants, the TNA is washed and precipitated. (B) The Quick-DNA/RNA FFPE kit utilises the spin column methodology to extract the TNA. Figure created with BioRender.com.

4.4 Results

4.4.1 Selection of appropriate nucleic acid extraction kits

4.4.1.1 Selection of TNA extraction kit for FFPE sample

To validate the multiplexed assay on patient tumours, an appropriate FFPE TNA extraction kit needed to be identified and evaluated. Two extraction kits were selected for evaluation: the MasterPure Kit and the Quick Kit (**Table 4.1**).

The MasterPure Kit is compatible with a broader range of sample inputs such as FFPE tissue, cells, serum, and buffy coat with a capacity to process 10-50 mg of tissue of 10 to 35 μm thick paraffin sections. The Quick Kit, designed specifically for FFPE samples, is suitable for up to 25 mg of tissue or up to 4 tissue sections of ≤ 20 μm thick. Additionally, the isolation technologies used by the two kits varied: the Quick Kit uses a column-based isolation method, while the MasterPure Kit uses a salt-precipitation method. Both kits produce RNA, DNA, and TNA with similar processing times of approximately 2 hours. However, the MasterPure Kit requires an external pre-treatment for the paraffin removal process, whereas the Quick Kit already includes a deparaffinisation step.

Table 4.1 Specifications of the two TNA extraction kits

	MasterPure Complete DNA and RNA Purification Kit	Quick-DNA/RNA FFPE Kit
Manufacturer	Lucigen	Zymo Research
Sample Format	FFPE tissue, cells, serum, buffy coat	FFPE sample
Sample Source	10-50 mg of 10 to 35 µm thick paraffin sections	≤ 25 mg paraffin tissue or up to 4 tissue sections (≤ 20 µm thick)
Isolation Technology	Salt-precipitation isolation	Column-based isolation
Final Product Type	RNA, DNA, TNA	RNA, DNA, TNA
Processing Time	2 h	2 h
Elution Volume	35 µl	50 µl
Adaptations for FFPE Samples	Optional pre-treatment for paraffin removal	Included deparaffinisation step

DNA = deoxyribonucleic acid; RNA = ribonucleic acid; TNA = total nucleic acid; FFPE = formalin-fixed paraffin-embedded

4.4.1.2 Identification of optimal TNA and RNA extraction kit

To identify the most effective TNA input for subsequent qPCR validations, a series of diluted and undiluted TNA samples extracted using the Quick and MasterPure kit were assessed (**Figure 4.2**). FFPE patient tumour blocks from Cohort 1 were utilised for this optimisation. The input for the TNA extractions was two FFPE tissue sections of 10 µm thick per sample and a range of qPCR input amounts were used to assess the amplification of the PROSPER targets in FFPE patient tumour samples.

Undiluted TNA extracted using the Quick kit produced inefficient amplification plots for all targets indicating the presence of inhibitors in these TNA samples affecting the qPCR reaction. Shallow and atypical sigmoidal-shaped amplification curves were recorded for these targets (**Figure 4.2 A-E).** For Gene 4, no amplification was recorded from the 1:2 TNA dilution, however, a positive amplification was recorded for this target at 1:5 and 1:10 TNA dilutions. Overall, samples at 1:5 and 1:10 dilutions of TNA consistently yielded efficient amplification for all PROSPER targets in the TNA-based qPCR assay. For the no template controls, no amplification signals were recorded. When TNA extracted from the MasterPure Kit was assessed as undiluted and 1:10 dilutions, inefficient amplification plots were observed for most PROSPER and NED Gene 1 targets (**Figure 4.2 F-J).** However, Gene 2 performed better relative to all other targets in the undiluted TNA (**Figure 4.2 G**). The Cq values of these targets were later compared to the Cq values of the targets tested on TNA at the same dilutions extracted from the Quick Kit indicating poorer TNA yield obtained from the MasterPure Kit. For the no template controls, no amplification signals were recorded.

Further assessment of the performance of the multiplex assay on the TNA extracted using the Quick Kit at 1:5 and 1:10 dilutions was done to identify the optimal dilution of TNA for subsequent validation experiments. At 1:5 TNA dilution, no Cq values were recorded for two out of the four patient samples. However, the 1:10 TNA dilution demonstrated consistent Cq values across all 4 samples for all the PROSPER targets and NED Gene 1 **(Figure 4.3)**. Therefore 1:10 TNA dilution was chosen for subsequent optimisation experiments.

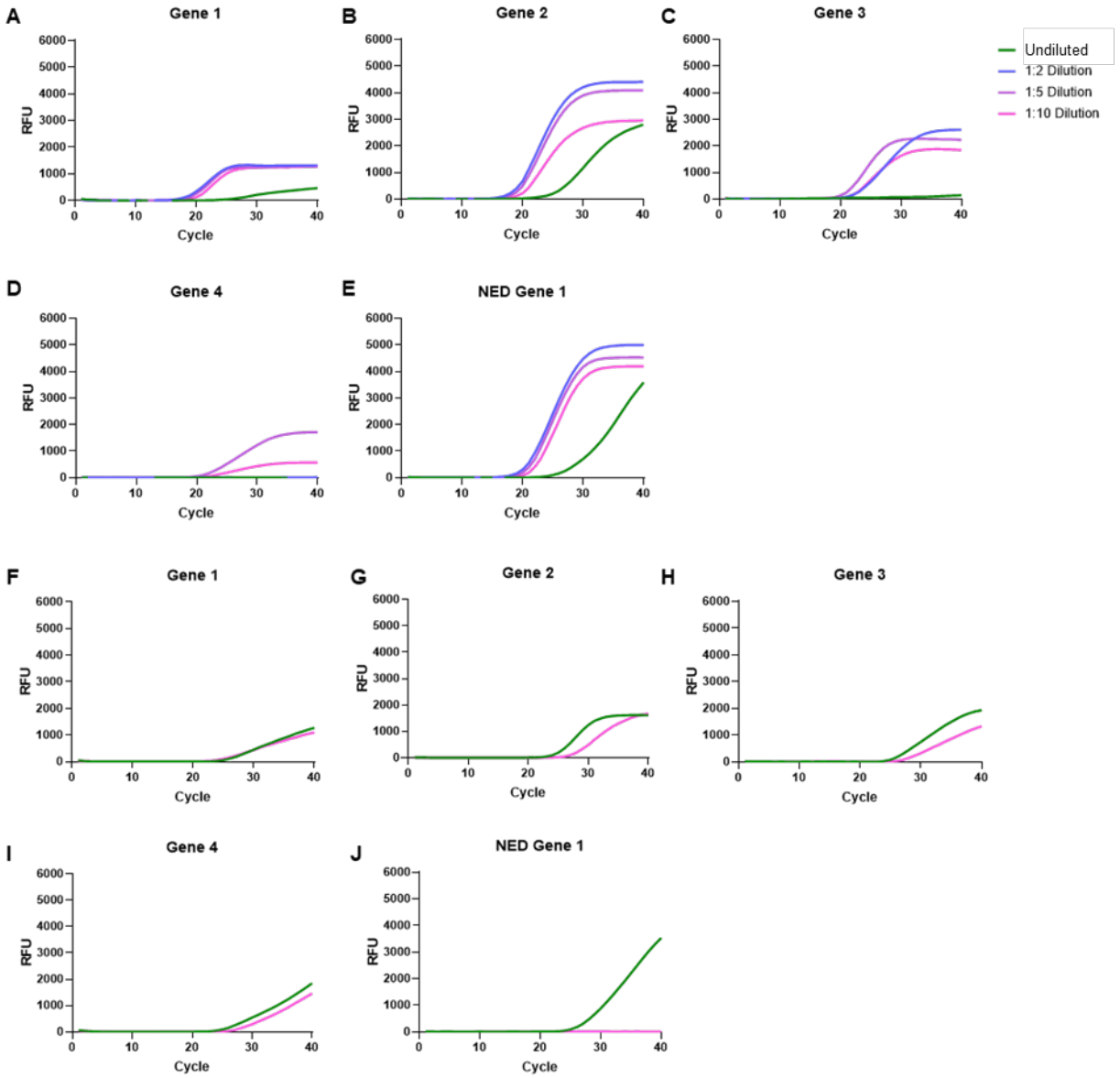


Figure 4.2: TNA-based qPCR assay performance on TNA extracted using the two kits in the FFPE patient sample from Cohort 1. The amplification plots of the multiplexed PCR assay with PROSPER targets and NED Gene 1 were tested on various dilutions of TNA extracted using the (A-E) Quick Kit and the (F-J) MasterPure Kit. RFU = relative fluorescence unit.

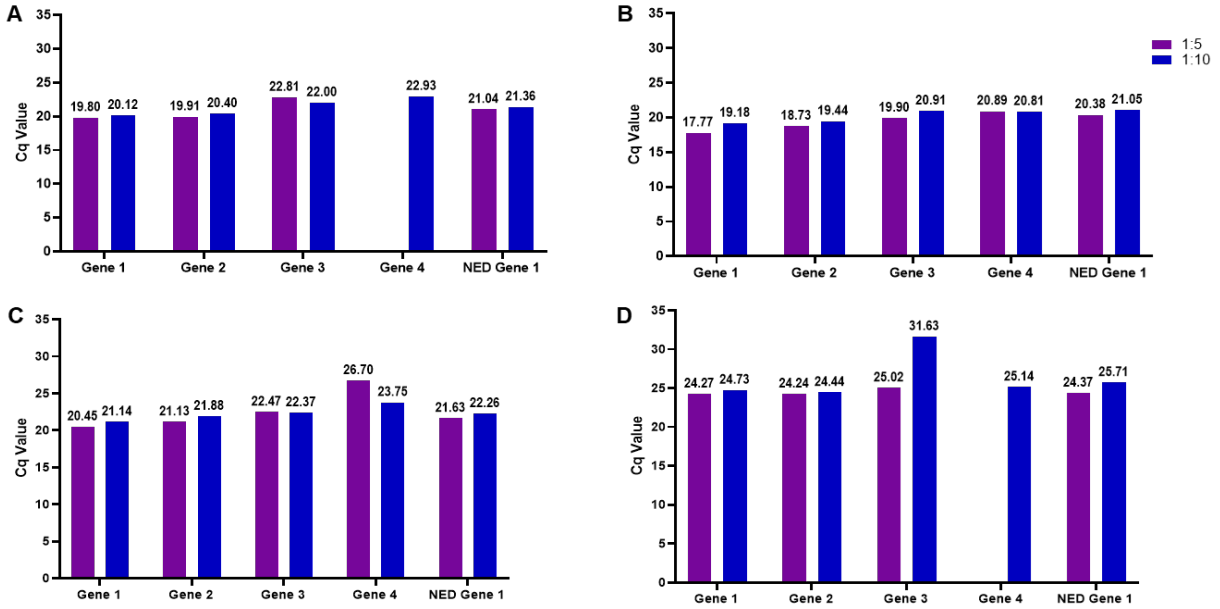


Figure 4.3: TNA-based assay performance on TNA extracted from the Quick Kit in FFPE samples from Cohort 1. Bar charts display the Cq values obtained from the multiplexed PCR assay tested on 1:5 and 1:10 dilutions of four patient samples: (A) #016, (B) #022, (C) #031 and (D) #069. TNA from these samples were extracted using the Quick Kit.

4.4.1.3 Assessing the performance of the two extraction kits on fresh FFPE sample

After establishing the PCR assay input parameters from TNA extracts from the MasterPure and Quick kits, both extraction kits were used to extract TNA from recently processed FFPE patient samples. Undiluted and 1:10 dilutions of extracted TNA were tested on the multiplexed PCR assay. This was to assess if the age of the samples was affecting the extraction process of the kits.

For TNA samples extracted using the Quick Kit, the undiluted samples exhibited inefficient amplification plots for Gene 4 (**Figure 4.4A)** No amplification curve was present and a delayed Cq value of 39.59 was recorded for this target (**Figure 4.4C**), which suggested the presence of inhibitors in the undiluted samples. However, at 1 in 10 TNA dilutions, both PROSPER and NED Gene 1 targets showed positive amplification plots and early Cq values (**Figure 4.4 A-C)**. All

PROSPER genes demonstrated amplification curves and Cq values that were comparable in the 1 in 10 diluted fresh FFPE sample, to the results seen in previous optimizations of the assay. This demonstrated the suitability and performance of the Quick kit for measuring the PROSPER signature in freshly prepared FFPE tissue samples. Conversely, despite using fresh FFPE samples, suboptimal amplification plots were observed with both undiluted and 1:10 diluted TNA extracted using the MasterPure Kit. The amplification curves were wider and almost flat for most PROSPER targets, with considerably later Cq values observed which indicated that a poor yield of TNA was being extracted from this kit. Moreover, the observation that Cq values were considerably later in undiluted samples compared to 1:10 diluted TNA prepared using the MasterPure Kit suggested there was inhibition of the PCR due to contaminants carried over in the extraction (**Figure 4.4 D and E**).

Since the Quick Kit was also able to extract RNA from FFPE samples, this kit was selected for further optimisation and validation.

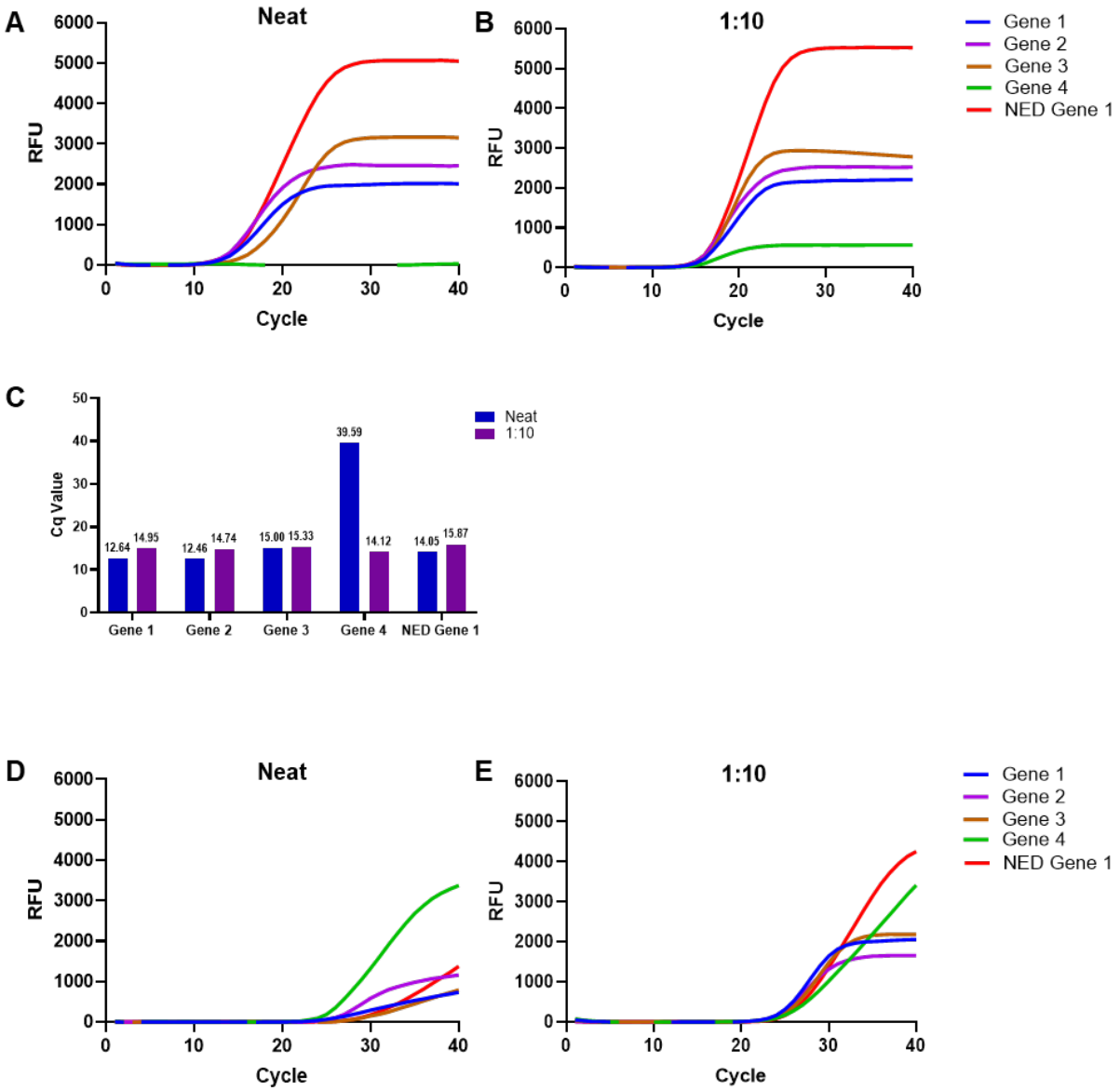


Figure 4.4: Comparison of TNA extracted from commercial TNA extraction kits in a fresh FFPE patient sample. Panels (A & B) illustrate the amplification plots of PROSPER and NED Gene 1 targets using undiluted (Neat) and 1:10 diluted TNA extracted from the Quick kit. Panels (D & E) illustrate the amplification plots of PROSPER and NED Gene 1 targets using TNA extracted from the MasterPure Kit. Undiluted and 1:10 TNA dilutions were inputted for PCR reactions. Panel (C) shows the Cq values obtained from the PCR reaction on samples extracted using the Quick Kit. RFU = relative fluorescence unit, Cq = quantitation cycle.

4.4.2 Evaluation of initial sample preparation for TNA yield

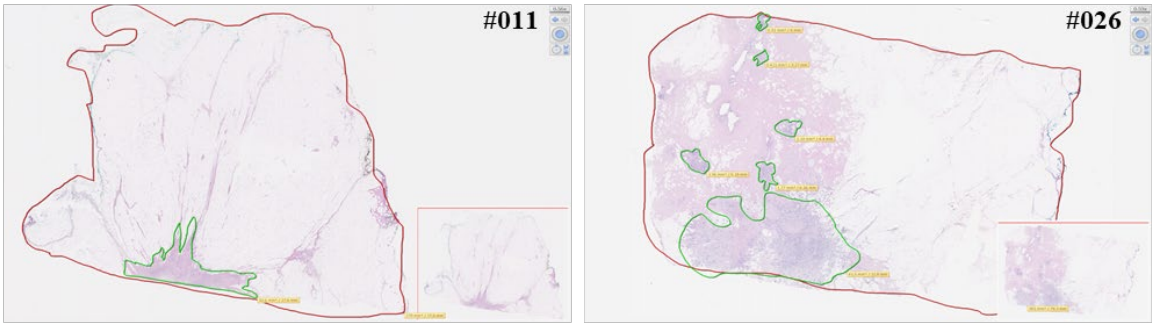
As discussed before, although high amplification efficiencies could be achieved with a 1:10 diluted TNA sample, the high Cq values suggested that there was a low nucleic acid yield, leading to low sample input for PCR reactions. Therefore, refinements of the protocol were attempted to enhance the quality of TNA extraction from FFPE patient samples. Firstly, the approach involved reducing the sample input amount, to avoid exceeding the input limits of the extraction kit. According to the manufacturer's instructions, the recommended sample input was up to 25 mg tissue from a paraffin block or a maximum of four tissue sections (≤ 20 µm thick in total), with a combined surface area not exceeding 20mm^2. In instances where sample sizes were larger than 20mm^2, a strategy of splitting a single 10 µm section across two separate extraction columns and combining the resulting eluates, was employed to increase TNA yield, and consequently improve the delayed Cq values. Approximate tissue areas were estimated from digital scans of an adjacent H&E-stained tissue section for two tissue samples (**Figure 4.5 A&B**). As before, PCR amplification of the PROSPER signature targets and NED Gene 1 control were performed with undiluted and 1:10 diluted sample inputs. A no template control was included in each condition.

As shown in **Figure 4.5 C**, positive amplification plots were observed for Gene 1, Gene 2, Gene 3, and NED Gene 1 and a suboptimal amplification curve was observed for Gene 4 in undiluted extracted TNA from FFPE patient samples, #011 and #026 after employing this strategy. However, in the 1:10 TNA dilutions for both patient samples, most targets exhibited inefficient amplification curves characterised by a flattened slope indicating less than a doubling of product with each amplification round. Specifically, for both samples, the amplification plots resembled straight lines with no plateau phase for Gene 1, Gene 2, and Gene 3. No amplification was recorded for Gene 4

in the 1:10 TNA dilutions of both samples. For the no template control, no amplification was observed.

The Cq values remained high in both samples which indicated that this strategy did not improve the TNA yield (**Figure 4.6A and B**). In the 1:10 dilutions, there was an increase in Cq values across all targets compared to undiluted TNA, with no Cq values recorded for Gene 4 for both samples.

A



B

	Tumour size (mm²)	Input (n x 10 µm)*	Further optimisation steps
#011	10.60	2	N/A
#026	46.41	1	Lysed sample was split into 2 tubes

C

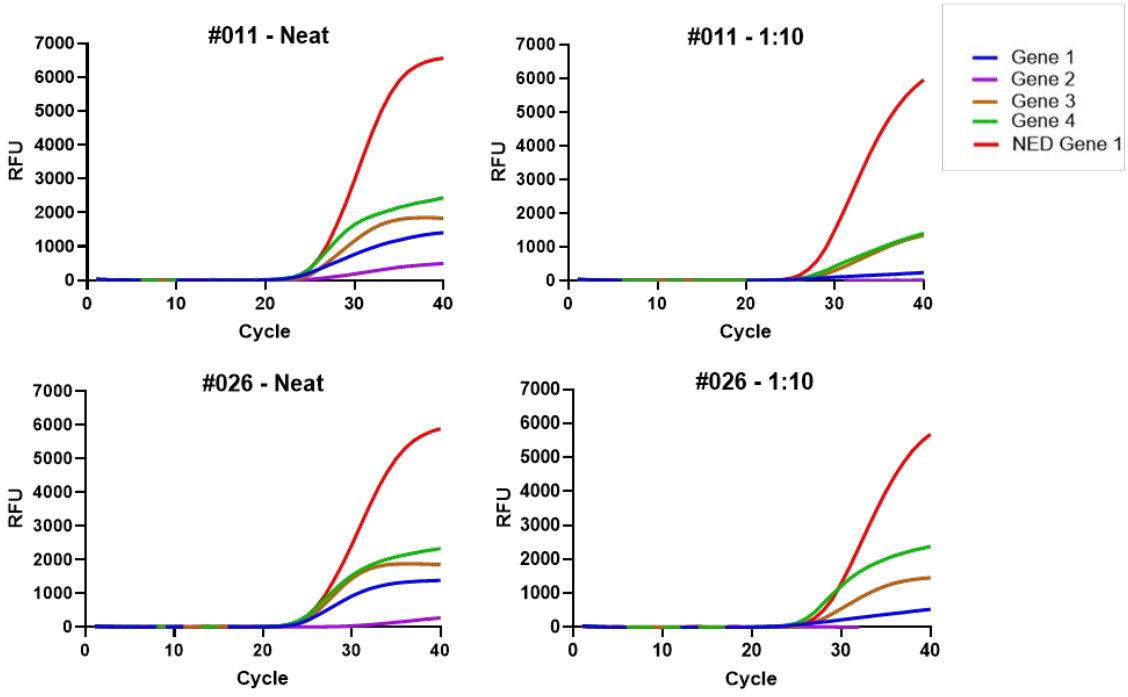


Figure 4.5: H&E staining and annotation of Cohort 1 FFPE patient samples, along with the TNA-based assay performance. Panel (A) illustrates the annotated H&E slides, which were scanned and analysed to estimate tissue area for two FFPE patient samples. Panel (B) lists the tumour sizes of the respective samples and outlines the optimization steps using the Quick Kit. The asterisk (*) denotes the number of 10 µm sections used for extraction in these samples. Panel (C) shows the subsequent

amplification plots of TNA extracted using the optimized extraction kit method on the two patient samples. Undiluted samples and 1:10 dilutions were used for PCR reactions. RFU = relative fluorescence unit.

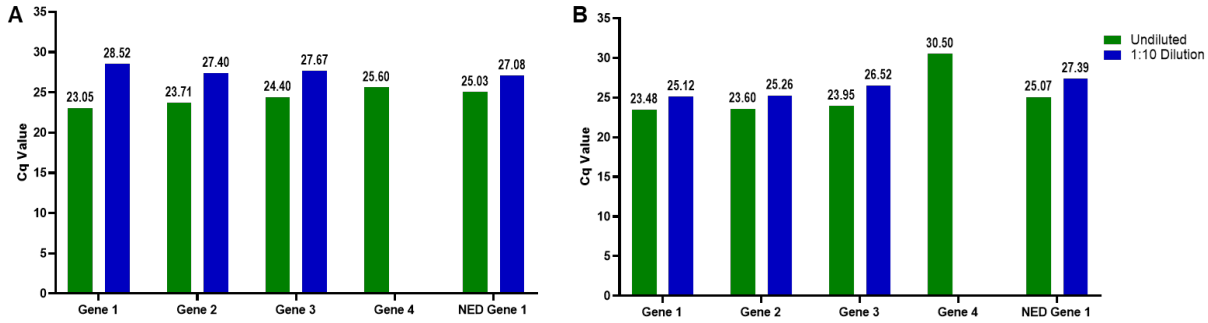


Figure 4.6: Cq values from the optimised TNA extraction kit on FFPE patient samples in Cohort 1. Bar charts show the Cq values from the multiplexed PCR assay tested on two patient samples: (A) #011 and (B) #026. TNA was extracted using the optimised Quick kit.

4.4.3 Evaluation of effect of incubation time on TNA yield

Further refinements were directed towards optimising the duration of tissue digestion in proteinase K to improve the yield of TNA from FFPE tissues. The manufacturers' recommendation for complete FFPE tissue digestion involved incubating the tissue for 4h in proteinase K. However, it is possible that this duration of incubation could lead to loss of the RNA fraction of the TNA due to degradation. Therefore, the aim was to determine if shorter incubation time led to better TNA yield.

Incubation times of 1h, 2h and 4h were evaluated. After completing the proteinase K digest at these times each of the samples was processed using the standard protocol and 5 µL undiluted TNA for each sample was assessed for expression of the PROSPER genes and NED Gene 1 control. This revealed that there was minimal impact from shortening the proteinase K digestion time in both samples (**Figure 4.7**). It was concluded that no improvement in sample yield could be achieved by shortening the incubation time.

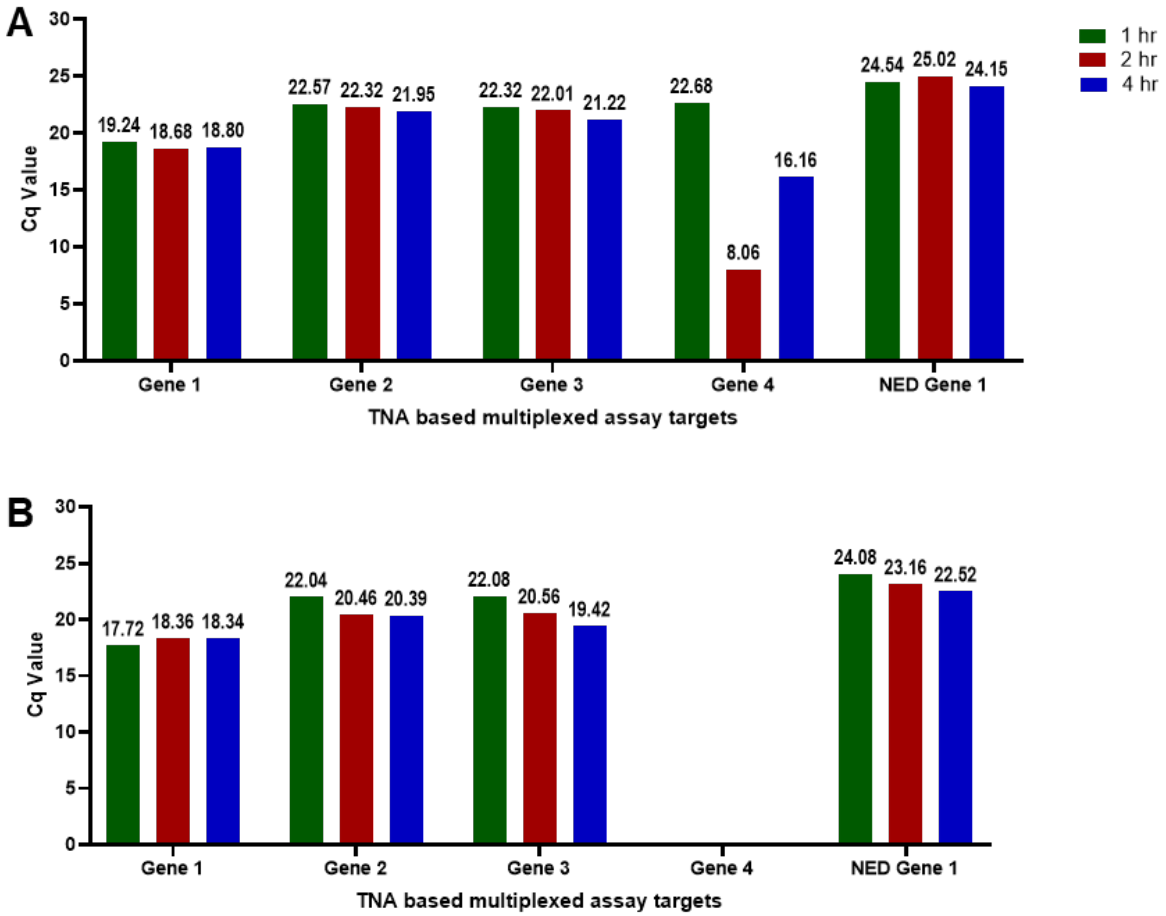


Figure 4.7: TNA-based assay performance on TNA extracted using different incubation times from FFPE patient samples in Cohort 1. Bar charts illustrating the Cq values obtained from the multiplexed PCR assay for two patient samples: (A) #011 and (B) #026. TNA was extracted using the Quick Kit.

4.4.4 Enhancement of the multiplexed PCR assay

Given suboptimal amplification curves and inconsistent amplification of Gene 4 in the previous experiments, optimisation of the multiplexed assay was explored. Due to the limited availability of archived human FFPE tissue, the optimisation experiments were performed on FFPE samples derived from cell lines. Improvements observed in these preliminary tests were then applied to archival FFPE patient samples.

4.4.4.1 Specific Optimisation for Gene 4 performance

Three conditions explored were: (i) doubling (2x) the concentration of both probe and partzymes, (ii) 2x the concentration of just the probe and (iii) the standard concentration of both probe and partzymes. The standard concentration of all components was 200nM.

Specifically, the objective of this experiment was to achieve an RFU threshold of at least 1000 and improve the amplification of this target. For this optimisation, TNA extracted from fresh T-47D breast cancer cells was used. Increasing the probe concentration from 200nM to 400nM alone led to an increase in RFU output from 700 to 1700. An increase in both probe and partzyme concentrations from 200nM to 400nM yielded an increase in RFU from 700 to 1400. For no template control, no amplification signals were observed **(Figure 4.8A**).

Furthermore, the Cq values of 2x probe only and 2x probe and partzymes conditions showed significantly lower Cq values compared to the standard condition (ordinary one-way ANOVA test, $p < 0.01$, **(Figure 4.8B)**. There was no significant difference between the 2x probe only and the 2x probe and partzymes conditions. Therefore, for subsequent experiments, the 2x probe-only concentration condition was used for the Gene 4 assay.

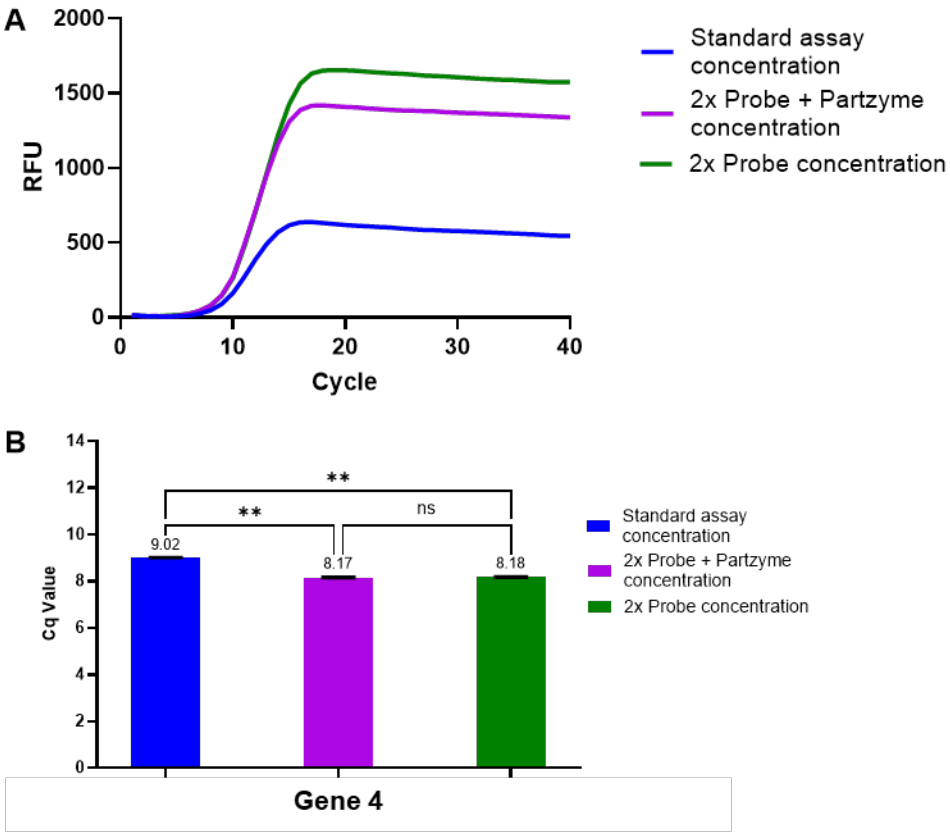


Figure 4.8: Gene 4 assay optimisations on TNA extracted from fresh T-47D cell pellet. (A) amplification plots and (B) Cq values of different Gene 4 assay conditions. Optimisations were performed in duplicates. The data was analysed using ordinary one-way ANOVA test. Asterisks (**) and (ns) represent $p < 0.01$ and not significant respectively. RFU = relative fluorescence unit, Cq = quantitation cycle

.

4.4.4.2 Evaluation of the optimised multiplexed assay in FFPE samples

The optimised TNA extraction and assay conditions were applied to FFPE patient samples from Cohort 1, to evaluate their performance in archival samples. PCR reactions were performed in duplicate, using undiluted TNA from samples #011 and #026. Upon testing the optimised conditions, there were no significant improvements observed in terms of Cq value or amplification of Gene 4.

The amplification plots revealed that only Gene 2 and NED Gene 1 exhibited well-defined sigmoidal curves in both tested samples, indicating effective amplification under the optimised

conditions. However, Gene 1 and Gene 3 displayed wider and nearly flat amplification plots for both samples (**Figure 4.9 A)**. No amplification signal and no Cq value were recorded for Gene 4 in both samples (**Figure 4.9 C**). These results indicated that the optimised conditions did not improve the PCR performance on FFPE samples.

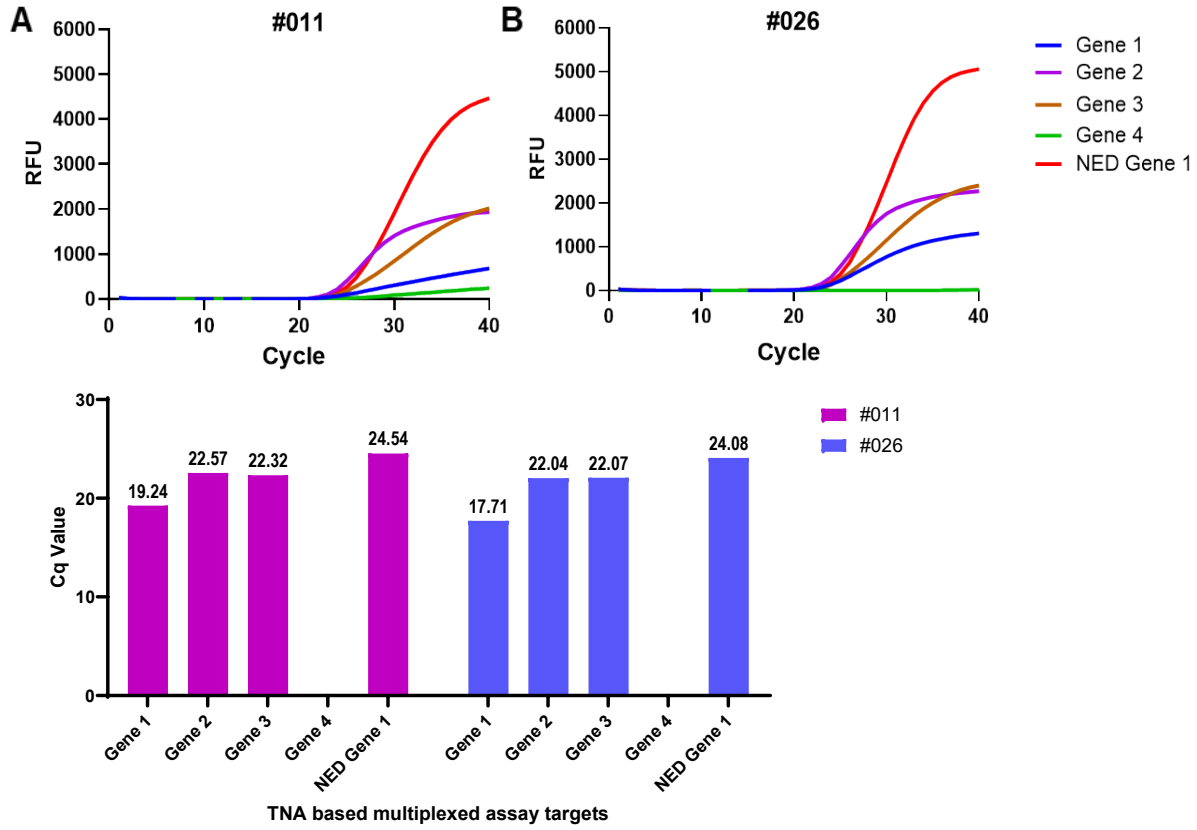


Figure 4.9: Assessment of the optimised TNA extraction and multiplexed PCR assay performance on FFPE samples from Cohort 1. Amplification plots of TNA-based assay tested on TNA from two patient samples, (A) #011 and (B) #026. The TNA was extracted using the optimised protocol of the Quick Kit and optimised qPCR assay. C shows the Cq values obtained from the qPCR assay. RFU = relative fluorescence unit, Cq = quantitation cycle.

4.4.5 Evaluation of the optimised multiplexed assay in stored RNA

There are inherent challenges posed by the high level of fragmentation of nucleic acids in long-term stored FFPE samples. Therefore, using the stored residual RNA eluates which had been extracted from freshly prepared FFPE patient tissues from Cohort 1, the PROSPER and RNA HK Gene 1 target were assessed using the RNA-based multiplexed qPCR assay.

The amplification plots revealed that the PROSPER targets and the RNA HK Gene 1 exhibited a consistent increase in fluorescent intensity over time. A typical sigmoid-shaped amplification curve with a defined exponential phase was displayed for all targets in all four samples from Cohort 1 **(Figure 4.10 A-D).** For the no template control, no amplification was recorded. All samples showed relatively early Cq values for all targets indicative of good RNA yield in these stored samples **(Figure 4.10 E-H)**. Consequently, the assessment of the PROSPER signature on all the stored residual RNA eluate from Cohort 1 was performed in duplicate using the RNA-based multiplexed assay.

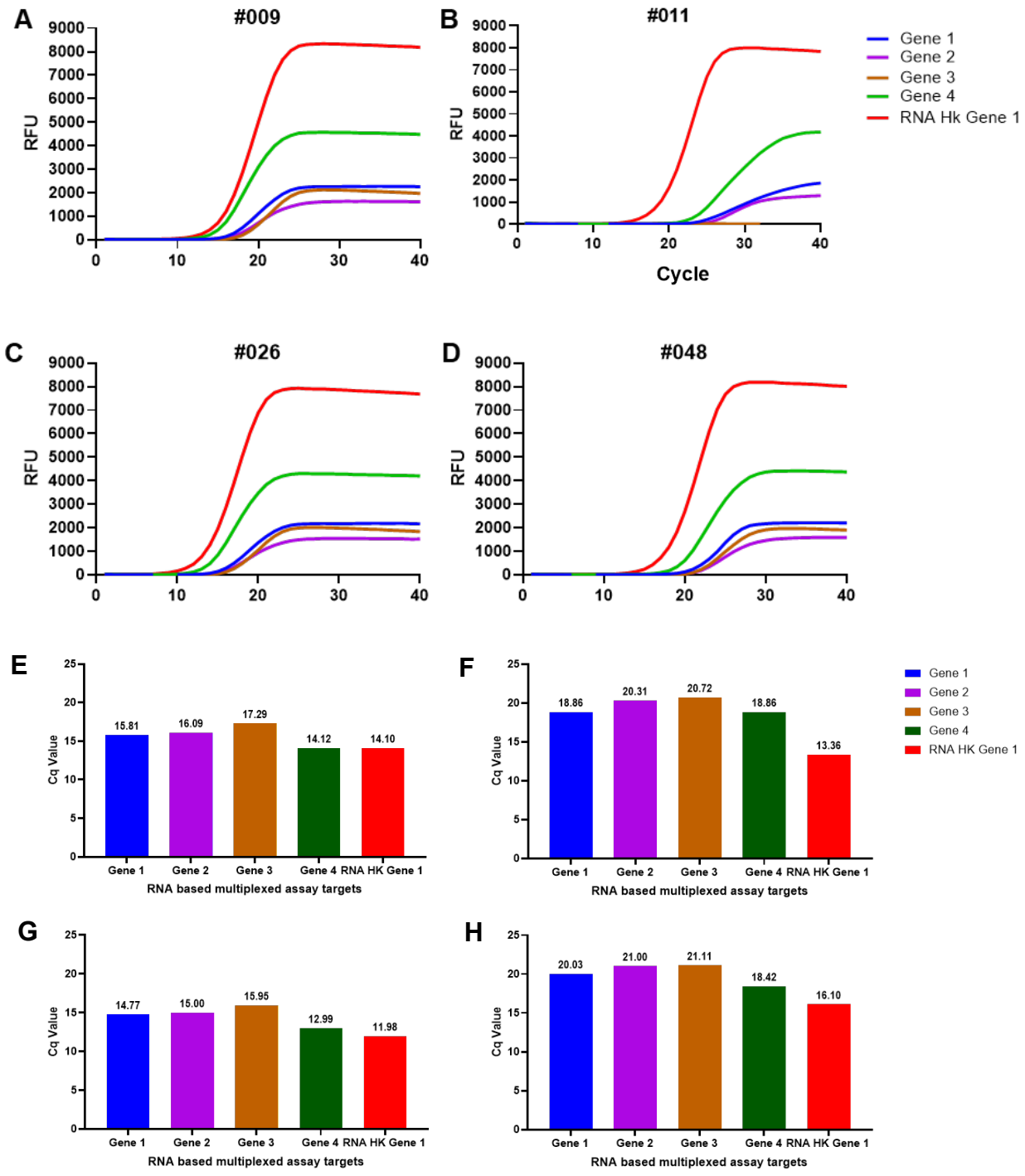


Figure 4.10: Assessment of the RNA-based assay performance on stored residual RNA from Cohort 1. Amplification plots of RNA-based assay tested on stored RNA from four representative patient samples: (A) #009, (B) #001, (C) #026 and (D) #048. Panel (E-H) shows the Cq values from the qPCR reaction. The

reactions were performed using undiluted RNA samples. RFU = relative fluorescence unit, Cq = quantitation cycle.

4.4.6 Redesign and optimisation of the PROSPER multiplexed PCR assay

Due to the high degree of fragmentation of the nucleic acids present, particularly RNA, in long-term stored FFPE samples, the qPCR assay was fundamentally redesigned. The amplicon sizes of the previous PROSPER qPCR assay components ranged from 120bp to 150bp. This redesign focused on the gene-specific primer sequences, aiming to produce amplicons of approximately 100bp. Additionally, due to the consistently suboptimal performance of Gene 4 in FFPE samples, it was decided to replace this target with Gene 5 from the original PROSPER assay. Consequently, the multiplexed assay underwent a comprehensive redesign and optimisation process, guided by the methodologies outlined in **Chapter 3.** The original PROSPER signature measured on NanoString has included 5 genes. However, Gene 5 had been removed from the algorithm for the PCR assay, after extensive in silico evaluation, due to assay limitations. So, the original Gene 5 was able to be substituted for the poorly performing Gene 4, without complete signature redesign. However, a comprehensive re-evaluation of RT-qPCR assay performance was conducted with the new 4-gene combination, including single and multiplex combination assays, to validate the inclusion of the new target in the assay. Only the final multiplex assay results are shown, in the interest of conserving space.

4.4.6.1 Detection of the transcription of DNA and RNA in the PROSPER gene signature

The transcription of DNA and RNA was assessed on TNA extracted from FFPE and fresh samples derived from T-47D cells using the smaller amplicon qPCR assay. The Cq values were compared

between reactions with and without reverse transcription. The difference in Cq values, denoted as ΔCq, was used to estimate the level of RNA transcription. The signal for all (-) RT reactions demonstrated that the assays were detecting DNA.

For both FFPE and fresh samples, ΔCq results in **Figure 4.11** indicated that Gene 5 demonstrated the highest RNA transcription followed by Gene 1, Gene 2, and Gene 3. In both fresh and FFPE samples, NED had a ΔCq of 0.19 and 0.57 (**Figure 4.11**). For all targets and NED Gene 1, Cq values were recorded in (-) RT reactions.

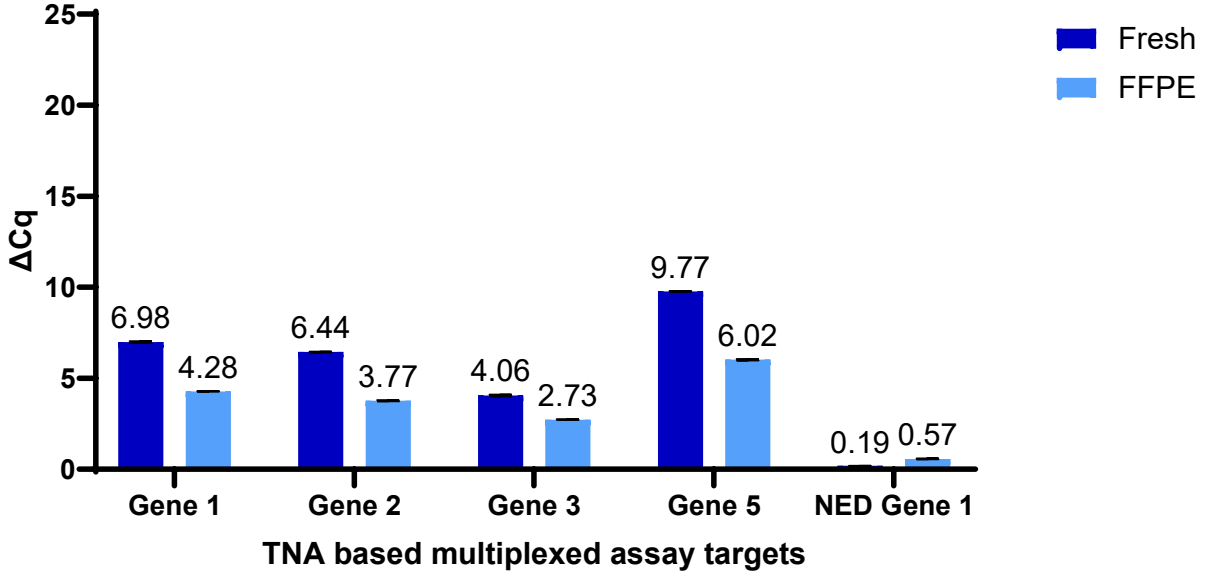


Figure 4.11: Transcription of TNA using the smaller amplicon assay. Bar charts showing the ΔCq of the PROSPER targets and NED gene 1 with and without RT in a qPCR reaction. The experiment was conducted in duplicates of TNA extracted from flash-frozen and FFPE blocks of T-47D cell lines. ΔCq = Cq RT (-) – Cq RT (+). Cq = quantitation cycle, RT (-) = without reverse transcriptase, RT (+) = with reverse transcriptase

4.4.6.2 Testing the multiplexed assay with the NED Gene 1

The smaller amplicon assay was tested on TNA extracted from the FFPE tissue block from Cohort 1, samples #011 and #026. The amplification plots revealed that the PROSPER targets and the

NED Gene 1 exhibited a consistent increase in fluorescent intensity over time, with a slight inefficiency observed in the amplification curve of Gene 5 for both samples. For the no template control, no amplification was recorded (**Figure 4.12**).

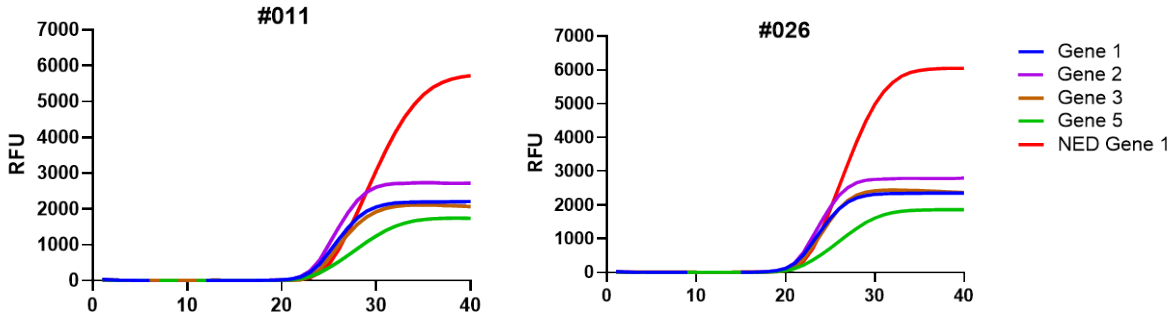


Figure 4.12: Assessment of the smaller amplicon assay multiplexed PCR assay performance on TNA extracted from FFPE patient samples from Cohort 1. Panels (A and B) are amplification plots of the PROSPER target and NED tested on TNA extracted using the optimised Quick Kit and the smaller amplicon assay for two representative patient samples. PCR reactions were conducted using undiluted TNA. RFU = relative fluorescence unit.

The smaller amplicon assay of the PROSPER targets with NED Gene 1 was tested on a set of eight additional FFPE patient samples from Cohort 1. The Cq values were compared with those from the original assay design. In the original assay, the Cq values for Gene 4 were consistently high or not recorded **(Figure 4.13**). With the smaller amplicon assay, Gene 5 was detected with considerably lower Cq values compared to the performance of Gene 4 in the original PROSPER assay design. For other gene targets, the Cq values from the smaller amplicon assay also showed lower Cq values compared to the original assay (**Figure 4.13)**.

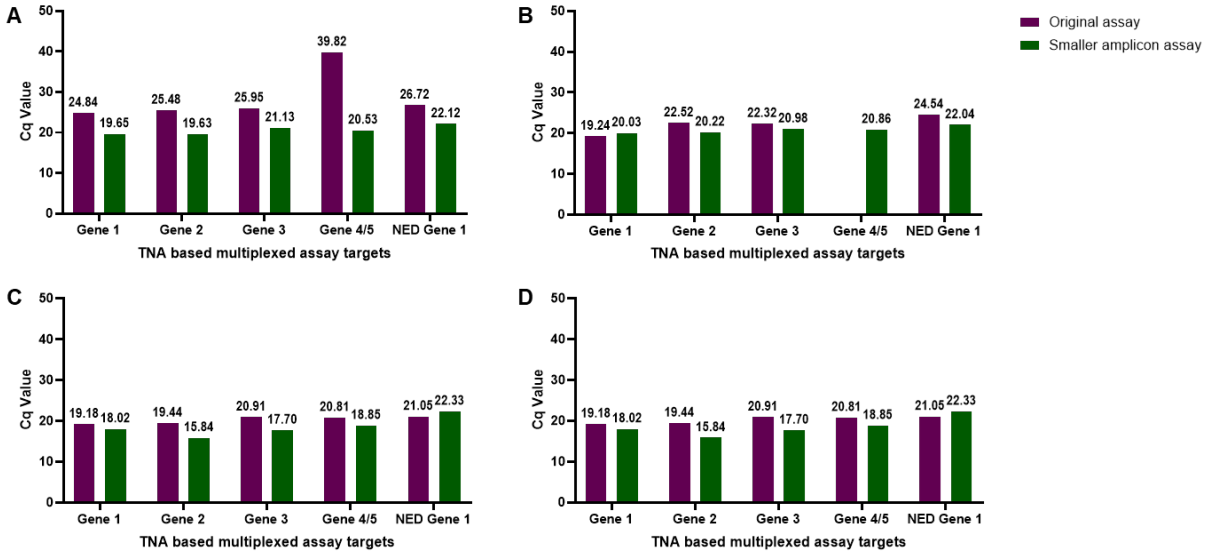


Figure 4.13: Performance of smaller amplicon TNA-based assay on Cohort 1 samples. Comparison between the Cq values obtained from the original qPCR assay and the smaller amplicon qPCR assay on four representative patient samples: (A) #010, (B) #011, (C) #022 and (D) #026 samples. The TNA was extracted using the optimised Quick kit method and qPCR reactions were tested on undiluted samples. Cq = quantitation cycle.

4.4.6.3 Testing the multiplexed assay with the RNA HK Gene 1 on Cohort 1

Using the Quick Kit, RNA was extracted, and the new smaller amplicon RNA-based assay was tested on a subset of Cohort 1 samples. Samples #009, #027, #034 and #048 were tested. The amplification plots revealed that the PROSPER targets and the RNA HK Gene 1 exhibited a consistent increase in fluorescence intensity over time (**Figure 4.14**). Lower Cq values were observed for these samples (**Table 4.2**).

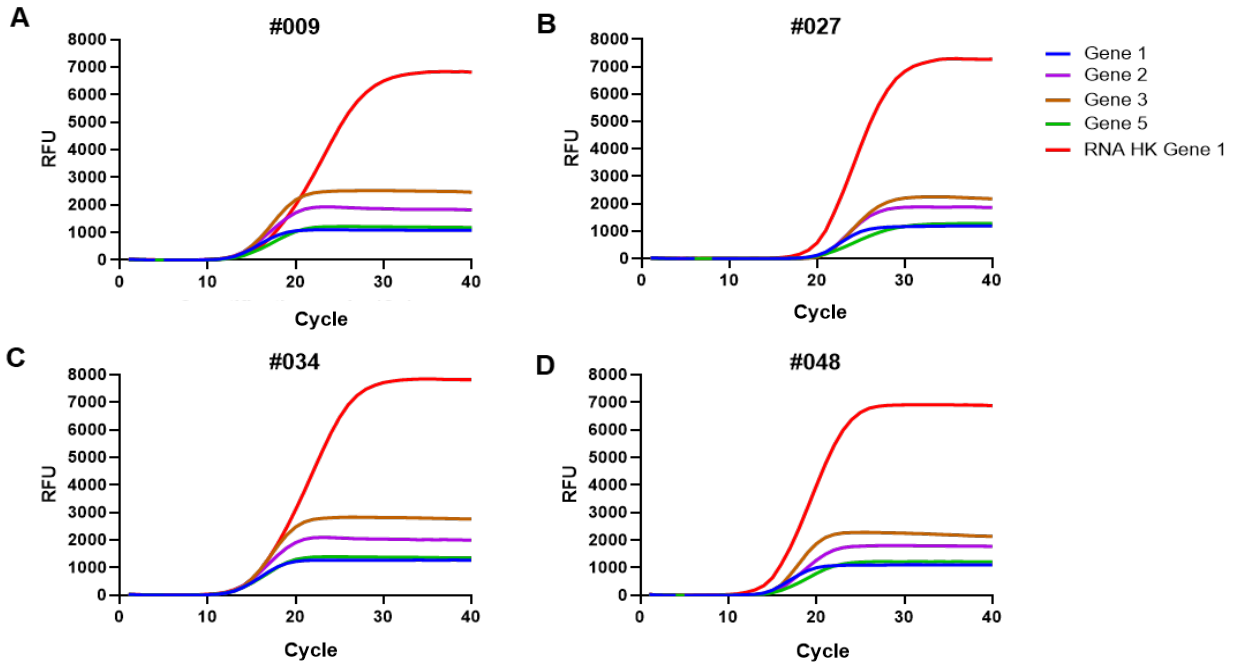


Figure 4.14: Assessment of the smaller amplicon multiplexed PCR assay performance on RNA extracted from FFPE blocks from Cohort 1. Panels (A - D) are amplification plots of the PROSPER targets and RNA HK Gene 1, tested on RNA extracted using the optimised Quick kit and assessed using the smaller amplicon qPCR on four patient samples. PCR reactions were carried out using undiluted samples. RFU = relative fluorescence unit.

Table 4.2 Cq values were obtained from the amplification of RNA utilising the smaller amplicon multiplexed assay on Cohort 1 samples.

	Cq				
	Gene 1	**Gene 2**	**Gene 3**	**Gene 5**	**RNA HK Gene 1**
#009	11.85	12.5	12.83	12.61	15.19
#027	19.04	20.19	20.62	19.49	19.15
#034	11.76	12.28	12.7	11.8	14.05
#048	13.45	14.93	14.6	14.47	13.99

Cq = Quantification cycle

1.1.1.1 Performance of the PROSPER assay with RNA HK Gene 1 on Cohort 2 RNA

The smaller amplicon assay with the PROSPER targets and the RNA HK Gene 1 was tested on a subset of RNA samples extracted from Cohort 2 using the Quick Kit. The subset included samples

#198, #185, #215 and #219. The amplification curve was poorly defined for all these samples as shown in (**Figure 4.15**). For samples #185 and #198, no exponential phase was recorded on the amplification curve (**Figure 4.15 A and B**) and with late Cq values and no Cq value was recorded for Gene 1 for sample #198 (**Figure 4.15 F**).

For samples #215 and #219, the amplification curve exhibited a small sigmoidal shape (**Figure 4.15 C and D**). However, for sample #215, the amplification curve for RNA HK Gene 1 was not recorded and had a high Cq value (**Figure 4.15 G**). Additionally, for sample #219, RNA HK Gene 1 had an amplification curve that resembled a straight line with no exponential amplification or plateau stage. These results demonstrated that, while the redesigned assay demonstrated improved performance in newer FFPE-derived samples, the older archival FFPE samples yielded nucleic acids that were too fragmented to assess by this approach.

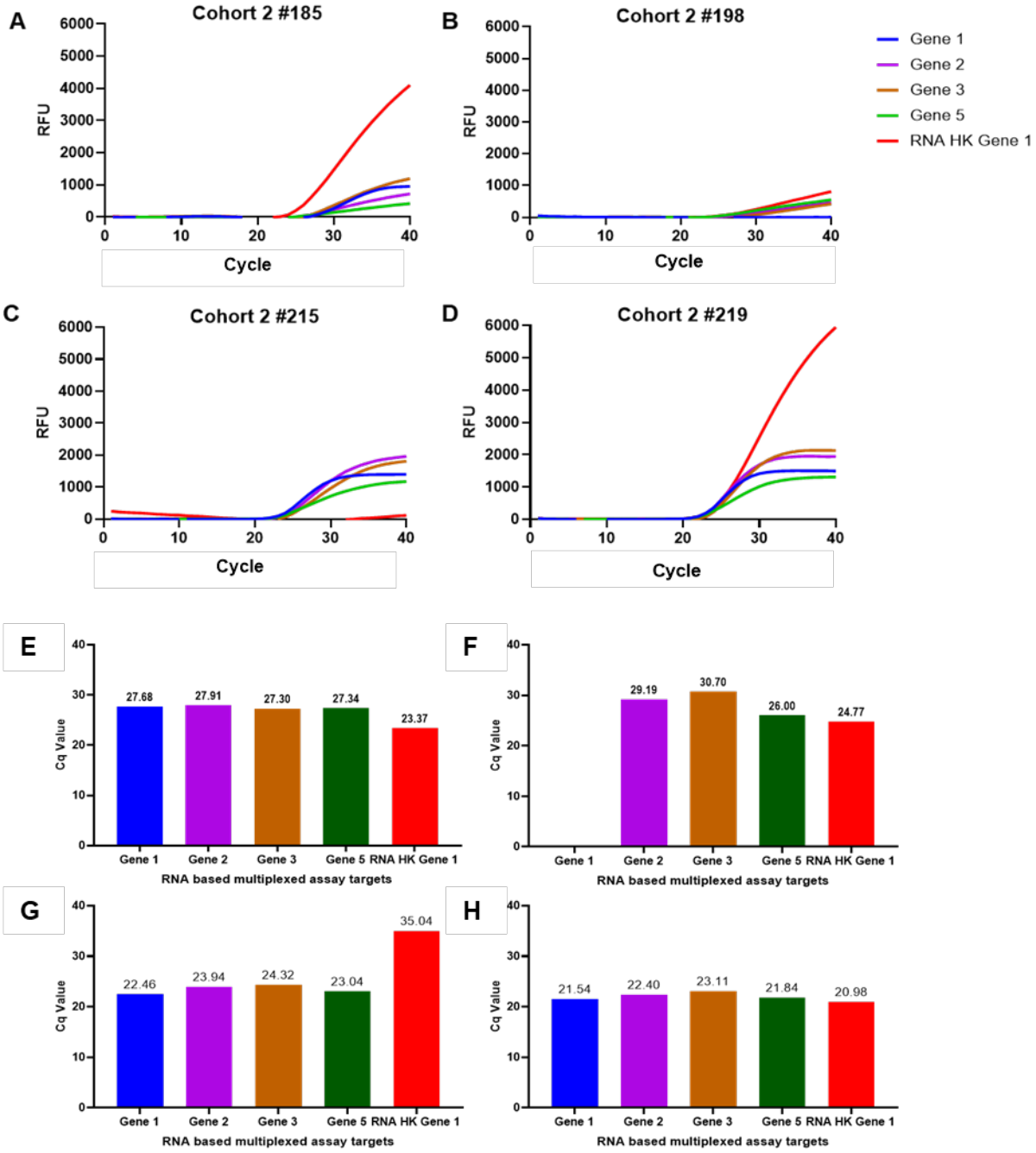


Figure 4.15: Assessment of the smaller amplicon multiplexed PCR assay performance on residual RNA from archival FFPE samples in Cohort 2. Panels (A-D) are amplification plots of PROSPER targets and NED tested on stored residual RNA extracted using the High Pure FFPET RNA Isolation Kit (Roche, Bassel, Switzerland) using the smaller amplicon multiplexed PCR assay on four patient samples. Bar charts illustrating the Cq values from the qPCR reaction for samples (E) #185, (F) #198, (G) #215 and (H) #219. Undiluted TNA was inputted for PCR reactions. RFU = relative fluorescence unit, Cq = quantification cycle.

4.4.7 Comparative analysis of PROSPER index

4.4.7.1 Comparative analysis of PROSPER index using the RNA HK Gene 1

Using the RNA-based assay, all residual RNA eluates from Cohort 1 (n=123) were tested and the PROSPER index was computed. The results were then compared to the PROSPER index computed using the NanoString assay, which utilised the expression of the same targets as the qPCR assay. The correlation between the multiplexed PCR and the NanoString Assay was high. with an r^2 value of 0.85 (**Figure 4.16A**). When the multiplexed assay results were compared to EPclin and Ki-67 scores, the r^2 values were 0.42 and 0.52 respectively (**Figure 4.16 B and C**), demonstrating a weak relationship with EPclin and a more significant correlation with the proliferation readout. These results are consistent with what was seen when the PROSPER assay was estimated using the NanoString platform with these samples, which revealed an r^2 value of 0.54 for comparisons to both the EPclin and Ki-67 scoring estimates (**Figure 4.16 D and E).**

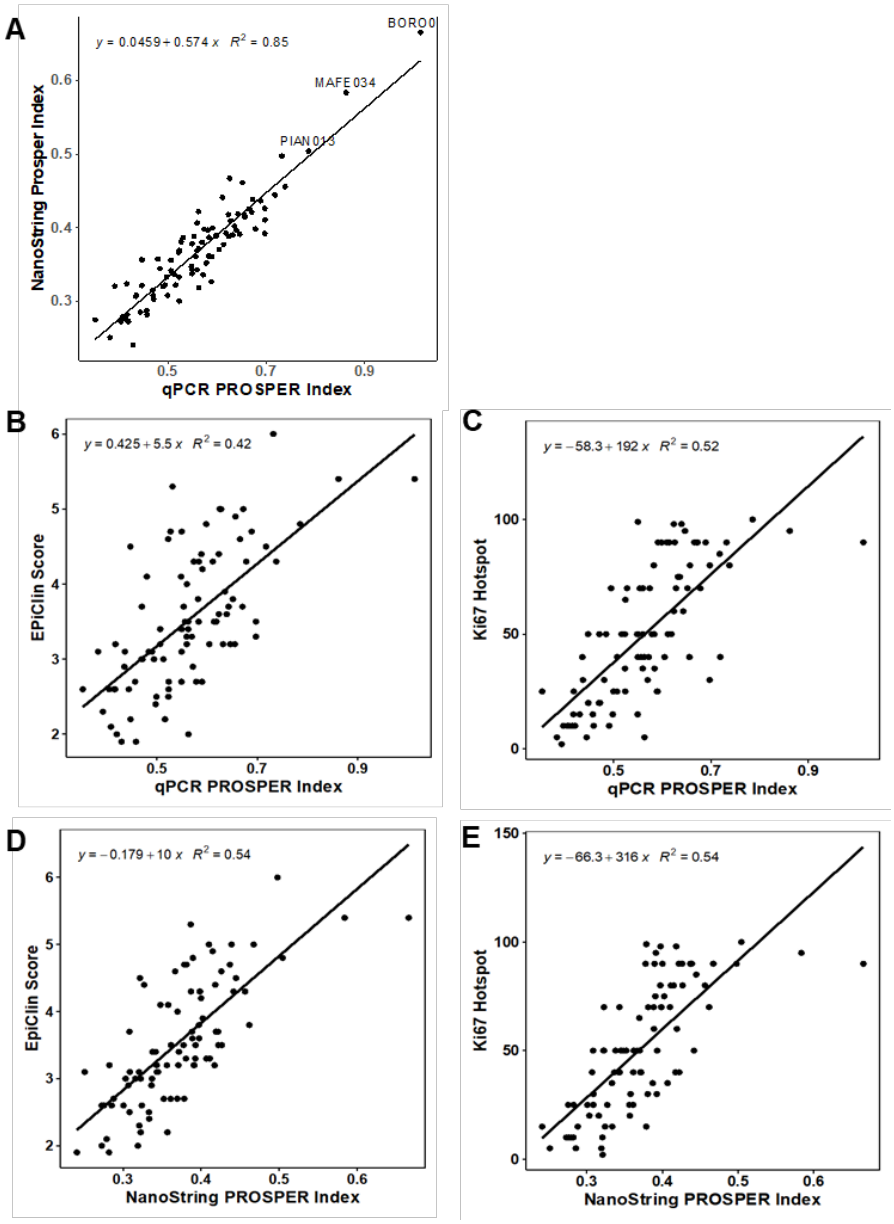


Figure 4.16: Comparison of PROSPER index measurements from Cohort 1 obtained from the multiplexed RNA-based PCR across different assays and clinical parameters. Scatter plots depict the correlation between PROSPER index values obtained from the multiplexed RNA PCR assay and those measured by (A) NanoString assay ($r^2 = 0.85$), (B) EPclin score (r2 = 0.42) and (C) Ki-67 score ($r^2 = 0.52$). Panel D and E show the correlation between PROSPER index values obtained from the NanoString assay with EPclin score ($r^2 = 0.54$), and Ki-67 score ($r^2 = 0.54$) respectively. Each data point represents a patient sample from Cohort 1. The assessment of the PROSPER targets using the RNA-based assay was performed in duplicate.

4.4.7.2 Comparative analysis of PROSPER index using the RNA HK Gene 1

Using the smaller amplicon multiplexed PROSPER assay with NED Gene 1, fresh TNA was extracted from all aged FFPE tissue blocks of Cohort 1 samples and tested with the PROSPER TNA PCR assay. The results were then compared to the PROSPER index measured using the NanoString assay, utilising the expression of the same subset of targets as the qPCR assay. The correlation between the multiplexed PCR and NanoString assay was low (r^2 = 0.32) (**Figure 4.17A**). There was also a relatively poor correlation between the EPclin values and Ki-67 scoring estimates for these samples (**Figure 4.17 B and C).** In comparison, the results obtained from the NanoString assay showed a stronger correlation with EPclin and Ki-67 scores (**Figure 4.17 D and E).**

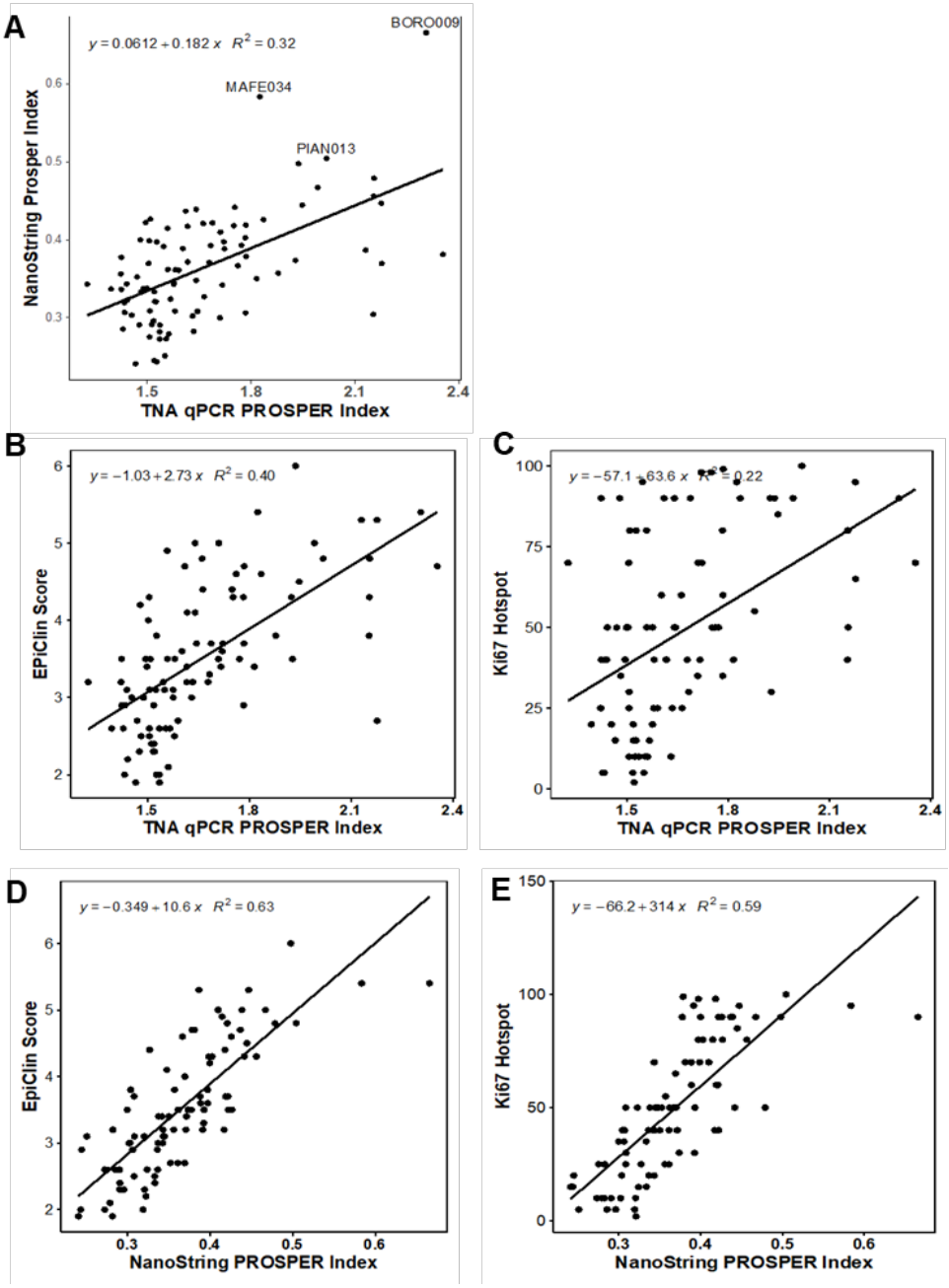


Figure 4.17: Comparison of PROSPER index measurements obtained from the multiplexed TNA-based PCR across different assays and clinical parameters. Scatter plots depict the correlation between PROSPER index values obtained from the multiplexed TNA PCR assay and those computed from (A) NanoString assay ($r^2 = 0.32$), (B)EPclin score ($r^2 = 0.40$) and (C) Ki-67 score ($r^2 = 0.22$). Panel D and E are also scatterplots illustrating the correlation between PROSPER index values obtained from NanoString assay to EPclin score ($r^2 = 0.64$) and KI-67 score ($r^2 = 0.59$) respectively. Each data point represents a patient sample from Cohort 1. The assessment of the PROSPER targets using the RNA-based assay was performed in duplicates.

To investigate whether the variation in PROSPER index correlation levels between TNA and RNA PCR assays, when measured in archival FFPE samples, resulted from the age of tissue blocks or differences in sample quality, the PROSPER index was directly compared. Specifically, the PROSPER index estimated from stored RNA (created during FFPE block processing) was compared with the index values obtained from both RNA and TNA samples derived from FFPE tissues stored for 4 to 5 years. This comparison was performed for a subset of cases from Cohort 1.

As demonstrated by **Figure 4.18**, PROSPER indices obtained from using RNA HK Gene 1 or NED Gene 1 have very similar PROSPER indices indicated by the red and pink dots, respectively. Indices produced from stored freshly extracted RNA (green) had very similar PROSPER indices to the results obtained from the NanoString assay in the same samples (**Figure 4.18**). Indices produced from nucleic acids extracted from long-term stored tissue blocks had higher PROSPER indices compared to those produced from freshly extracted nucleic acid, red/pink vs green/blue respectively (**Figure 4.18**), demonstrating that the driver of poor amplification was the age of the FFPE tissues rather than the difference between extract types. Furthermore, indices obtained using the RNA HK GENE 1 and NED Gene 1 were very similar in the long-term stored samples, indicating that a TNA-based assay could be used to assess the expression of the PROSPER signature.

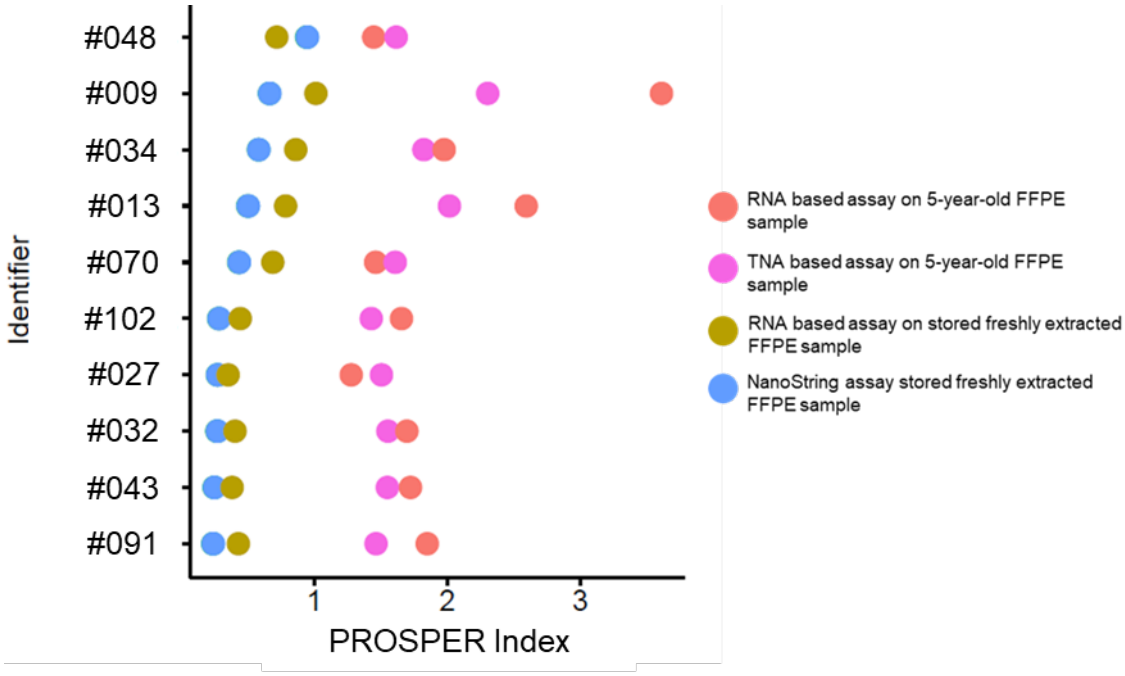


Figure 4.18: Comparison of PROSPER index measurements obtained from different assays tested on a subset of samples from Cohort 1. Blue data points are PROSPER index obtained from the NanoString assay. Dark green and pink data points are PROSPER index obtained from multiplexed RNA and TNA assays respectively.

4.5 Discussion

4.5.1 Summary of results

In this chapter, the objective was to assess the effectiveness of different nucleic acid extraction methods from FFPE tissue for optimisation of the PROSPER index test as a PCR assay. Two extraction kits were evaluated: the Lucigen MasterPure Complete DNA and RNA Purification Kit (MasterPure Kit) and the Zymo Quick-DNA/RNA FFPE Kit (Quick Kit). The MasterPure Kit employs a salt-precipitation method, making it versatile and compatible with various sample types. In contrast, the Quick Kit is specifically designed for FFPE samples and uses a column-based isolation method. Both kits were used to extract TNA from five-year-old FFPE samples and freshly made FFPE samples. Efficient amplification of the PROSPER and normalising gene targets was observed in TNA extracted from aged and new FFPE blocks at 1:10 dilution using the Quick Kit. However, TNA extracted using the MasterPure Kit consistently showed suboptimal amplification characterised by wide, flat curves with delayed Cq values compared to those obtained with the Quick Kit. Therefore, the Quick Kit was selected for further optimisation and validation steps for extracting RNA and TNA from FFPE samples.

FFPE samples pose a significant challenge for downstream applications such as multiplexed PCR assays due to high levels of RNA and DNA degradation, poor overall yield and carry-over of inhibitory contaminants leading to suboptimal RT-qPCR outcomes. Therefore, it was reasoned that poor results in the PROSPER qPCR could reflect low quality and yield of RNA and DNA from these samples, and potential protocol improvements were investigated to address this. Adjustments to the Quick Kit extraction methodology, including reducing sample input and optimising incubation times, had limited impact in improving TNA yield and amplification

efficiency. Splitting large sample sections before extraction did not significantly improve the amplification efficiency. Incubation times did not influence Cq values since shorter incubation times yielded similar results to the standard four-hour incubation. Gene 4 demonstrated suboptimal amplification in all conditions compared to other target genes.

Given the limitations observed in improving sample yield and quality within the Quick Kit protocol, optimisation was performed to improve the sensitivity and specificity of the PROSPER PCR reaction. A key objective was to enhance the amplification of Gene 4 given its poor performance in comparison to the other targets. Adjustments to the concentration of qPCR assay components improved amplification efficiency and reduced Cq values for all gene targets amplified from TNA extracted from FFPE cell line pellets. Experimenting with different concentrations of assay components specific to Gene 4 further enhanced its performance. Doubling only the probe concentration resulted in a more than two-fold increase in RFU and lower Cq values compared to the standard Gene 4 assay concentration on TNA extracted from FFPE cell pellet samples. However, these adjustments did not improve the performance of samples extracted from aged FFPE tissue samples as Gene 4 failed to amplify or amplified poorly in these samples. Nevertheless, when tested on fresh FFPE patient samples, consistent and efficient amplification of all gene targets, including Gene 4, was observed with the optimised multiplexed qPCR assay. This optimised assay with the RNA HK Gene 1 was then used to validate the performance of the PROSPER signature on freshly extracted stored RNA eluates from Cohort 1, where it displayed good performance with high amplification efficiency and low Cq values for all gene targets.

Since one of the objectives was to compare the RNA and TNA-based assays, optimisation of the qPCR assay was continued. Given the consistent suboptimal results observed in aged FFPE samples, the qPCR assay was redesigned by replacing Gene 4 with Gene 5 and the qPCR assay

was modified to produce shorter amplicons. These modifications resulted in significantly improved amplification efficiency and lower Cq values for all gene targets, including Gene 5. The smaller amplicon TNA assay with NED Gene 1 was then tested on TNA-extracted Cohort 1 samples. The results showed consistent and efficient amplification of all gene targets, demonstrating the suitability of TNA as the starting material for the PROSPER assay. Furthermore, the smaller amplicon assay demonstrated its ability to differentiate between DNA and RNA accurately in freshly processed FFPE samples, providing accurate gene expression estimates. Overall, the results demonstrated the successful optimisation of the multiplexed qPCR assay for amplifying PROSPER targets in patient FFPE samples. However, the performance of the smaller amplicon assay on a subset of RNA extracted from the archival FFPE Cohort 2 samples, which were 20 years old, was less successful. The Cq values for target genes were consistently high with poor-quality amplification curves, indicating suboptimal amplification. This result is potentially due to degraded RNA, or inhibitory contaminants present in archival FFPE samples.

There is considerable interest in the application of the PROSPER test as a TNA assay due to the simplicity of TNA extraction and testing, and its applicability to automated methods. Therefore, comparisons between the performance of RNA-based and TNA-based PROSPER assays were a key focus. However, a confounder in this study was the availability of cohorts with freshly processed FFPE tissue from which TNA could be prepared, whereas RNA extracted from fresh FFPE tissue from Cohort 1 was available since these RNA samples had been prepared at the time of patient recruitment and stored frozen for future use. The comparisons that were able to be made between the available samples and the assays performed are summarised in **Table 4.3** below:

Table 4.3 Summary of the comparison of the PROSPER index against markers.

Measure 1	Measure 2	Same extract or independent extracts	R^2 value
PROSPER RNA PCR	PROSPER RNA NanoString	Same	0.85
	EPclin	Same	0.42
	Ki-67	N/A	0.52
PROSPER TNA PCR	PROSPER RNA NanoString	Independent	0.32
	EPclin	Independent	0.40
	Ki-67	N/A	0.22
PROSPER RNA NanoString	EPclin	Same	0.54
	Ki-67	N/A	0.54

Using the results from Cohort 1, a direct comparison was possible between the PCR-based and NanoString-based PROSPER RNA assay results derived from the same set of stored RNA samples. The PROSPER PCR index was also compared with the EndoPredict EPclin result and Ki-67 proliferation marker in this cohort. The results of the PROSPER index computed from the optimised RNA-based assay revealed a strong correlation with the result obtained from the NanoString platform (**Table 4.3**, r^2=0.85). This validated the RNA-based multiplexed qPCR assay as a reliable alternative to NanoString for measuring the PROSPER index. Furthermore, analysis of the PROSPER index obtained from this optimised multiplexed qPCR assay showed comparable correlations with both the EPclin score (r^2=0.42) and Ki-67 score (r^2=0.52) to those computed from the NanoString assay (r^2=0.54 for both). This correlation further confirmed its validity for evaluating gene expression. It is important to note that there was a slight decrease in r^2 value between the PROSPER index from NanoString which utilised 5 genes and multiple housekeeping controls (Figure 1.5A, R^2=0.61) vs the NanoString that utilised 4 genes and one control (Figure 4.16D, R^2=0.54). However, as mentioned previously, most hospitals don't have specialised platforms like NanoString, however qPCR machines are widely available in clinical labs, consequently the panel had to be streamlined to five genes enabling a qPCR format.

In contrast to the performance of the PROSPER RNA assay in RNA from fresh FFPE tissues, a moderate correlation was observed between the PROSPER index obtained from the redesigned TNA-based assay and NanoString assay (r^2=0.32), and between PROSPER TNA and EPclin (r^2=0.40), while a poor correlation was seen to Ki-67 (r^2=0.22). While the lower correlation to the NanoString PROSPER assay and EPclin values could be partly attributed to the fact that measurements were made on independent samples, unlike the RNA PCR assay that was performed in exactly the same samples, the comparison to Ki-67 is independent of sample type since Ki-67 is an IHC test and was also independent of the RNA assay from fresh samples. This suggested two potential reasons for the poorer performance of the TNA-based PROSPER assay compared to the RNA-based PROSPER PCR assay. Firstly, the TNAs were extracted from considerably older FFPE tissues, suggesting that tissue age is a key determinant of sample integrity. Secondly, the extraction kit used to prepare the stored RNA was different from the reagents used to prepare the TNA from the archival FFPE tissues. Since the stored RNA had been the residual RNA saved after performing the commercial EndoPredict assay, RNA was extracted using Siemens Healthcare reagents that incorporate a silicon-coated magnetic bead extraction method. In contrast, the TNA samples prepared specifically for this study were made using the Zymo Quick RNA/DNA Kit, which uses a column-based isolation. It is possible that, while sample age is probably an important factor influencing assay performance, the choice of extraction method was also an important determinant of the correlation between the PROSPER assay outcome and other predictive markers in Cohort 1. While not significant due to the small number of samples tested, the good correlation between the RNA-based and TNA-based PROSPER results in archival samples extracted using the Zymo Quick kit suggests that the two PROSPER PCR assays are likely to perform equally well

in freshly prepared FFPE tissues, but that both may correlate more poorly to samples extracted using different reagents.

4.5.2 Integrating the multiplexed qPCR assay for laboratory diagnostics

To implement the multiplexed qPCR assay for routine laboratory practice, it is essential to validate the assay performance on TNA and RNA samples from FFPE tissues. This validation process will ensure the accuracy and reliability of the assay for the diagnostic setting. In this study, the two commercial nucleic acid extraction kits assessed were the MasterPure Kit and the Quick Kit. The MasterPure Kit utilised a salt precipitation method, while the Quick Kit used a column-based method for nucleic acid extraction. While the salt-precipitation methods (140) for nucleic acid extraction are efficient (141), they are laborious and time-consuming, leading to the adoption of column-based methods, which became popular for general nucleic acid extraction due to their convenience and ease of use (141). Nevertheless, the salt-precipitation method was being reported as economically more viable, allowing more control over the parameters involved, and reportedly giving better yields than commercial kits (141). It was found that the TNA extracted from stored and fresh FFPE tissues using the column-based isolation method showed consistent and efficient PCR amplification, whereas the samples extracted using the salt precipitation method exhibited lower PCR amplification efficiency and reduced assay sensitivity. This is because salt precipitation methods rely on the addition of high-concentration salt solutions to precipitate nucleic acids, which can also co-precipitate contaminants such as proteins, polysaccharides, and other impurities (142). These impurities can interfere with downstream applications such as PCR, leading to lower amplification efficiency and reduced assay sensitivity. Meanwhile, column-based methods like the Quick Kit utilise silica-based columns that effectively bind and separate nucleic acids from

contaminants, resulting in higher DNA and RNA purity (143, 144), enhancing the qPCR assay sensitivity and specificity, making them more suitable for diagnostic applications.

4.5.3 Optimising the extraction protocol and qPCR multiplexed assay

Nucleic acid extraction from FFPE samples typically involves the following steps: Firstly, after the FFPE curls are cut from the tissue block, they are deparaffinised using xylene and ethanol (145). A proteinase K digest follows this deparaffinisation step to break down the proteins in the sample and release the nucleic acids (146). After this step, formaldehyde crosslinks are reversed through a heating step and then TNA is purified, using a commercial kit. To enhance the yield of TNA and improve the purity of nucleic acids extracted from FFPE samples, different proteinase K incubation times were assessed. Contrary to existing studies suggesting longer incubation times, such as 24 hours, resulting in higher nucleic acid yield (147-149), there was no significant difference between one-hour and four-hour proteinase K incubations in terms of TNA yield. These results suggest that a shorter incubation time of one hour may be sufficient for extracting an optimal yield of TNA using the chosen extraction method. Recently, several commercial FFPE nucleic acid isolation kits support these findings by demonstrating efficient extraction with just one hour of proteinase K incubation time, suggesting longer periods are unnecessary (150). A potential reason for this could be that the shorter incubation time reduces the risk of over-digestion and degradation of nucleic acids that are already fragmented due to the fixation process. Nevertheless, this finding is important because it suggests that the extraction protocol can be optimised to save time and maximise efficiency without compromising the yield or quality of the nucleic acids extracted from FFPE samples.

Due to the consistent suboptimal amplification efficacy of Gene 4, the concentration of Gene 4 assay components was increased to improve the amplification efficacy and its overall performance in the multiplexed qPCR assay. An increase in just the probe concentration doubled the RFU and resulted in improved Cq values and significant improvement in the amplification efficacy of Gene 4 in cell line derived FFPE samples. However, when this improved assay was tested on FFPE patent samples, inefficient amplification was still observed in some cases, indicating that factors intrinsic to FFPE samples might be contributing to the suboptimal performance of the assay.

FFPE samples are known to be challenging for nucleic acid extraction and amplification due to the degradation and crosslinking of DNA and RNA caused by formalin fixation (151). Recently studies have investigated the impacts of storage conditions and times on nucleic acid extracted from FFPE samples (152-154), providing insights into the observed results. Groelz et al. (2018) demonstrated that the storage time and temperature of FFPE samples can have a significant impact on the quality and amplification of nucleic acids (153). They stored FFPE tissues at room temperature and 4°C for up to seven years. Subsequently, RNA and DNA were extracted from these tissues stored under different conditions and they were examined for integrity and usability in qPCR assays. Their results showed that RNA degradation starts immediately at room temperature, with significant declines in RNA integrity within 6-12 months, reaching the lowest integrity levels after this period. Furthermore, at room temperature Cq values between the FFPE tissue stored for 1 month and 72 months differed by 14. However, storing samples at lower temperatures, such as 4°C, -20°C, or -80°C, significantly slowed down the degradation process, thereby maintaining higher nucleic acid integrity over time and minimal changes in Cq values were observed between 1 month and 72 months. Similarly, Yi et al. (2020) found that long-term storage of FFPE samples at room temperature resulted in significantly decreased DNA and RNA quality

and fragment size, leading to reduced amplification efficiency in downstream qPCR assays (154). These findings were consistent with the observation that the optimised assay performed successfully with high amplification efficacy on stored freshly extracted RNA from the FFPE sample. Unfortunately, FFPE tissue blocks that are prepared for clinical diagnostic purposes are routinely stored at room temperature. Therefore, the degradation and fragmentation of nucleic acids over time likely contributed to the suboptimal amplification performance in the aged FFPE samples.

Since one of the objectives of this study was to compare the results obtained from TNA-based and RNA-based qPCR assays, the multiplexed assay was redesigned to accommodate the extraction of nucleic acids from the archival FFPE blocks that were available for this work. Due to the high degradation and fragmentation observed in stored FFPE block, it was determined that a redesign of the assay was needed to specifically target shorter amplicons to improve the amplification efficiency for older FFPE samples (155, 156). Additionally, due to the consistently suboptimal performance of Gene 4, it was replaced with Gene 5 after determining that there was a low impact from this change to the signature when assessed in existing datasets. The redesigned assay was tested on the FFPE tissues of Cohort 1. The redesigned multiplexed assay, specifically targeting shorter amplicons, showed improved amplification efficacy when tested on FFPE tissue samples compared to the previous assay, therefore the TNA-based assay was tested using the newly designed multiplexed assay on TNA extracted using the Quick Kit from Cohort 1 FFPE tissue samples. Given the age of the available samples and evidence that RNA and DNA fragments become progressively smaller and more degraded over time, the improved performance of the redesigned PROSPER assay suggests that a considerable proportion of the fragments present in

these older tissue samples were smaller than the target amplicon size of the original assay but detectable with the new assay resulting in a lower Cq value for the gene targets.

To further validate both the newly designed TNA- and RNA-based assay, the assay was tested on Cohort 2 FFPE tissue samples. However, this Cohort was more than 20 years old, with archival FFPE blocks that had been stored at room temperature. Inefficient amplification efficacy of both the TNA-based and RNA-based assays was observed when tested on the Cohort 2 samples. To date, limited knowledge is present regarding the usability of paraffin samples that have been preserved for more than 10 years in qPCR assay. Therefore, further investigation is required to assess the nucleic acid integrity of these samples and further optimisation of nucleic acid extraction needs to be investigated.

4.5.4 Implementation of PROSPER signature as multiplexed qPCR assay

As mentioned in **Chapter 2,** Cohort 1 samples were prospectively collected between the years 2017-2018. At the time of collection, RNA was extracted from freshly prepared FFPE tissue sections and stored at -80°C. The FFPE tissue blocks were subsequently stored at room temperature.

The samples were assessed for their EndoPredict EPclin score and Ki-67 score. The PROSPER index was estimated using the NanoString nCounter platform. Previous experiments from the laboratory have validated that the PROSPER index obtained from the NanoString platform correlated highly with the Ki-67 score, EPclin score and patient outcome. However, while the NanoString platform has been applied for other *in vitro* diagnostic tests, its use would add considerable cost to the PROSPER assay, which is incompatible with the objective of developing

an alternative low-cost test to commercial alternatives. The stored RNA samples from Cohort 1 were assessed for their PROSPER index using the optimised RNA-based multiplexed qPCR assay. The PROSPER index obtained from the RNA-based multiplexed qPCR assay showed a strong correlation with the PROSPER index obtained from the NanoString platform in the same store RNA sample. These results validated the use of the multiplexed qPCR assay as an alternative method for assessing the PROSPER index in FFPE samples and suggested its potential feasibility for implementation in the routine clinical setting.

The proliferative marker, Ki-67, along with ER, PR, and HER2 status, is routinely used in the clinical setting to classify breast cancer subtypes and guide treatment decisions. However, one of the primary limitations of Ki-67 is the lack of standardised methodologies for its assessment. The interpretation of Ki-67 immunohistochemical staining can be subjective, leading to variations in scoring and reporting among different laboratories and pathologists. Additionally, the cutoff values used to define high versus low Ki-67 expression have not been consistently defined, with different studies using varying thresholds (157, 158). This limits the widespread adoption of Ki-67 as a reliable prognostic marker in clinical practice.

Since the PROSPER signature consists of proliferation-based gene targets, the correlation between the Ki-67 scores and the PROSPER index was assessed. A high correlation was observed between the Ki-67 and the PROSPER index obtained from the RNA-based multiplexed qPCR assay. A similar correlation was observed between the PROSPER index obtained from the NanoString Platform and the Ki-67 score. These findings support the use of the multiplexed qPCR assay as a reliable and feasible method for assessing proliferation in FFPE samples, potentially overcoming the challenges associated with the standardisation and variability of the KI-67 scoring system. Additionally, a similar correlation between the PROSPER index from the NanoString Platform

and the EPclin score was observed from the PROSPER index from multiplexed qPCR assay and the EPclin score. These findings suggest that the multiplexed qPCR assay has the potential to serve as a valuable routine clinical tool for assessing both proliferation and clinical risk stratification in FFPE samples, providing clinicians with a comprehensive and reliable assessment of luminal breast patient prognosis.

However, these high correlation results did not translate with the PROSPER index obtained from the TNA-based multiplexed assay to the PROSPER index obtained from NanoString Platform, EPclin score or KI-67 scores. This can be potentially due to the age of the FFPE blocks when the TNA was extracted. This was confirmed when a distinct separation was observed in the PROSPER index obtained from the RNA-based assay performed on fresh samples versus older samples.

As previously mentioned, due to the simplicity of TNA extraction and their ease of applicability into an automated platform, there is significant interest in the use of TNA-based assays for the PROSPER assay for routine clinical use. However, due to the lack of availability of a matched fresh FFPE tissue cohort, a full comparative analysis between RNA and TNA-based PROSPER assays could not be conducted. Nevertheless, using a small subset of cohort 1 archival samples on both the RNA and TNA-based assays, a good correlation between the PROSPER index was observed despite the varying correlation with the established markers. These results indicate that the two assays are likely to perform equivalently and correlate well to established markers when optimal extraction methods and fresh samples are utilised.

The results from this study have successfully implemented the PROSPER signature into a multiplexed qPCR assay, demonstrating its potential as an alternative method for assessing proliferation and clinical risk stratification in FFPE samples of luminal breast cancer patients.

4.5.5 Study strengths, limitations and future directions

The work in this chapter demonstrated many strengths and limitations. One strength of this study is the use of multiple commercial kits to thoroughly evaluate different nucleic acid extraction approaches and compare their performance from both aged and fresh FFPE samples. Another strength is the use of multiple validation methods, including correlation analysis with other established assays such as Ki-67 and EPclin scores, to assess the reliability and accuracy of the PROSPER multiplexed qPCR assay. One limitation of this study is the potential degradation or loss of nucleic acid integrity in older FFPE samples, which may have affected the correlation between the PROSPER index and established assays. In addition, the difference in nucleic acid extraction methods between samples, with the use of Siemens Healthcare reagents incorporating a silicon-coated magnetic bead extraction for stored RNA and the use of column-based extraction for the older FFPE samples, hinders accurate and direct comparisons between the PROSPER index. Furthermore, the study's conclusions are based on a relatively small sample size.

Future research should aim to replicate these findings in a larger cohort of luminal breast cancer patients. Moreover, conducting a long-term stability study on freshly extracted and stored nucleic acid utilising the same extraction methodology would be highly beneficial to better understand the impact of storage duration and conditions on assay performance.

4.5.6 Conclusions

qPCR assays are used in a range of routine clinical diagnostic settings. Several commercial breast cancer tests utilise qPCR as their foundation technology, highlighting their importance and reliability in clinical applications. Additionally, qPCR tests are cost-effective and can be performed in-house, allowing for quicker turnaround times, and potentially reducing healthcare costs. The lab developed a prognostic gene signature for patients with ER+ breast cancer called the PROSPER gene signature. *In silico* studies demonstrated that the PROSPER gene signature can perform as well or better than existing commercial assays in predicting disease-free and overall survival. In this research, the PROSPER signature was successfully integrated into a multiplexed qPCR assay. The performance of the PROSPER multiplexed qPCR assay was validated in FFPE tissue samples against the established markers EndoPredict EPclin and Ki-67 scores. The results of this study demonstrate the successful implementation of the PROSPER signature into a multiplexed qPCR assay allowing for affordably assessing proliferation and clinical risk stratification for ER+ breast cancer patients.

Chapter 5

In silico Discovery and Validation of a New Prognostic Gene Signature for Triple Negative Breast Cancer

5.1 Introduction

Triple-negative breast cancer (TNBC) is a highly heterogeneous histological subtype of breast cancer characterised by the lack of expression of ER and PR, and non-amplified HER2 (159, 160). The subtype accounts for 10-20% of all breast cancers and is more prevalent in younger women. At diagnosis, TNBC tumours are often associated with lymph node involvement, are of higher grade and larger size, and are biologically more aggressive (159). Although TNBC patients have higher rates of clinical response to neoadjuvant chemotherapy, reflecting their generally higher proliferative rate, they also have a higher rate of distant recurrence and a poorer prognosis than patients with other breast cancer subtypes (161, 162). Nearly 70% of metastatic TNBC patients do not survive until 5 years after initial diagnosis, and nearly all die from their disease despite receiving chemotherapy which is the primary form of treatment (162).

Large-scale gene expression profiling studies have aided in our understanding of the molecular heterogeneity within TNBC (68, 69, 71, 163). Lehmann et al. (2011) reported that TNBC could be classified into biologically different molecular subtypes, which they initially referred to as basal-like 1 (BL1), basal-like 2 (BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), and luminal androgen receptor (LAR) (68). While these subtypes were biologically distinct, they were not clinically different and did not provide guidance on the risk of recurrence

or response to treatment. It was later recognised that the immunomodulatory subtype reflected immune infiltration rather than an intrinsic property of the tumour cells. Similarly, the mesenchymal stem-like subtype was recognised as being driven by the presence of tumour-associated stromal cells. In a refined schema, the group proposed four subtypes: BL1, BL2, M and LAR (69). In this schema, BL1 cases demonstrated a significantly better outcome than BL2. LAR and M both experience universally poor outcomes. Burstein et al. (2015) expanded further on the four-subtype classification: designating TNBC cases as AR-positive, mesenchymal, basal-like immune suppressed, and basal-like immune activated (71). This stratification recognised the importance of the immune milieu as a marker of treatment response. In particular, high immune infiltration is a marker of benefit from immune checkpoint inhibitors, which are increasingly used for the second-line treatment of TNBC (164). Moreover, high tumour infiltrating lymphocyte levels are also demonstrated as a good prognostic marker for benefit from cytotoxic chemotherapies (164, 165). Despite the demonstrated potential of these subtypes in predicting response to both targeted and untargeted treatments in TNBC, they are not routinely used in the clinical setting where chemotherapy continues to be the primary treatment choice for all patients regardless of prognosis. However, the distinct biology of these molecular subtypes suggests that the different genetic profiles within TNBC may be individually exploited to predict disease progression, providing a strong rationale for developing TNBC subtype-specific prognostic models that could offer more precise predictions of patient outcomes. In an unrelated study from our lab that explored proteomic profiles in a cohort of 340 TNBC tumours (manuscript in preparation), a sequential classification tool was developed that used a minimal target list to accurately stratify cases into biologically different subtypes. This approach reduces the heterogeneity within each of the subgroups by grouping them together based on their expressed

biological properties rather than the features that they lack (ER, PR and HER2). So, it was reasoned that employing this sequential classification tool and separately discovering predictive signatures in independent subsets of TNBC cases could produce more robust prognostic tools. In this study, both TNBC subtype-specific and non-subtype-specific prognostic gene signatures are identified and their prognostic utility in predicting TNBC patient outcomes is explored.

5.2 Aims and hypotheses

5.2.1 Aims

1) To assemble and prepare a comprehensive dataset of gene expression profiles from TNBC tumours.
2) Discovery of subtype-specific and non-subtype-specific prognostic gene signatures.
3) To validate the prognostic utility of the identified gene signatures using an independent TNBC dataset.

5.2.2 Hypothesis

It was hypothesised that subtype-specific prognostication, based on gene signatures derived after stratification of TNBC cases into specific subtypes, will offer a more accurate and clinically useful prediction of patient outcomes compared to traditional non-subtype-specific prognostication.

5.3 Approach

Since TNBC represents just 15 to 20% of breast cancer cases, most TNBC-focused datasets comprise relatively small cohorts. Moreover, to-date there is a paucity of RNA sequencing datasets from TNBC cohorts that could be utilised for this study. Therefore, to maximise cohort size, this study was confined to Affymetrix array-based gene expression datasets, which represent the greatest resource of TNBC gene expression data in the public domain. A meta data set was assembled from 3 public datasets (GSE58812, GSE21653 and GSE65194). After stepwise subtype classification, the cohort was separated into training and validation sets, ensuring equal distribution of TNBC subtypes between the two sets. Prognostic gene signatures were discovered in the training dataset by first conducting univariate Cox regression to identify individual genes that were significantly associated with outcome. The dataset was then either analysed as one set or separated into TNBC subtypes. Elastic net regularisation was then applied to the list of genes that were significantly associated with the outcome to determine a minimal gene set that was predictive of patient survival. A graphical summary of this chapter is depicted in **Figure 5.1.**

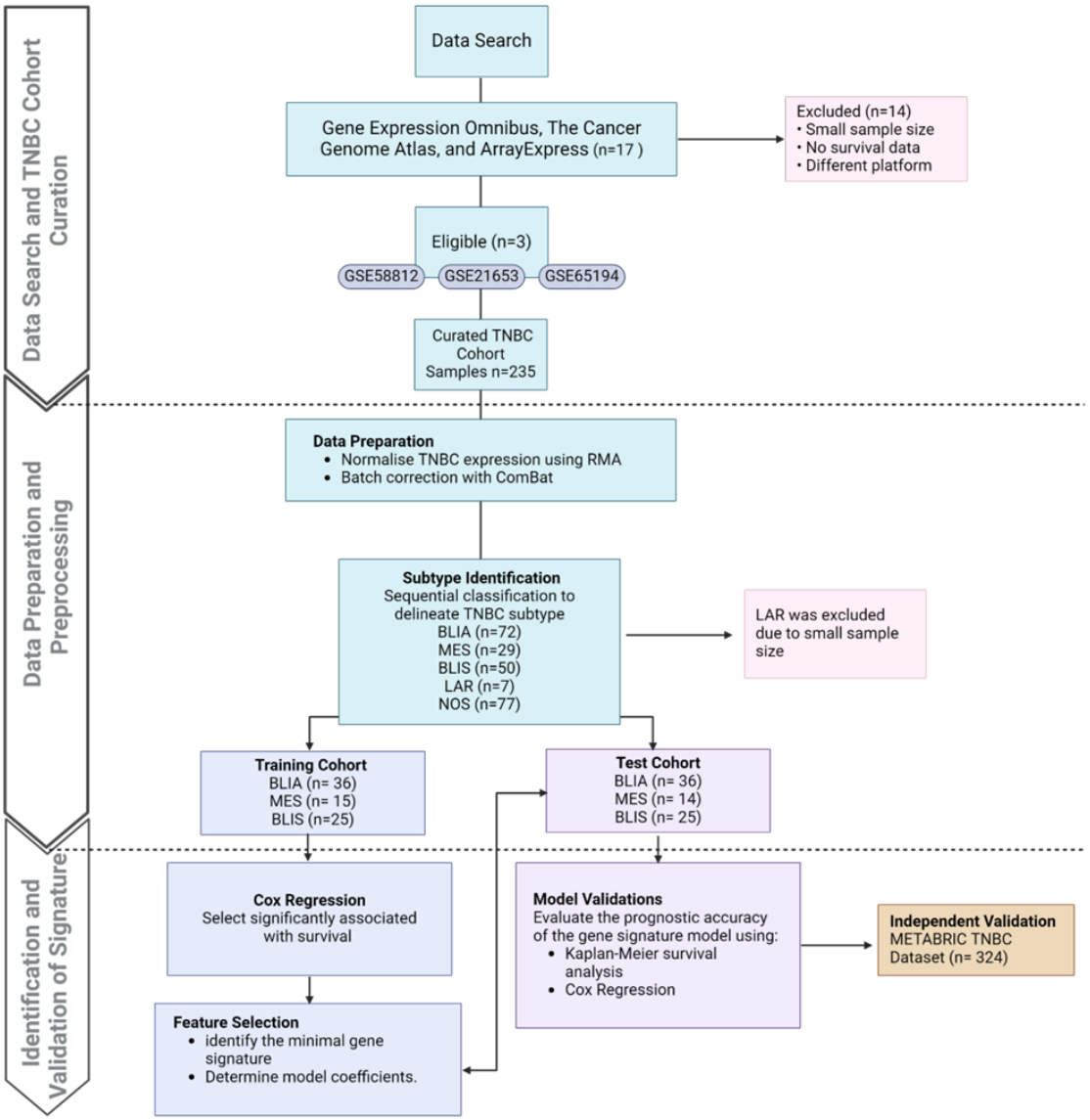


Figure 5.1: Graphical overview of chapter five. A data search was conducted to curate breast tumour gene expression profiling data from an assembled cohort of cancer TNBC patients. Using this cohort, prognostic signatures associated with disease outcomes were identified through various modelling approaches. An independent TNBC gene expression dataset with annotated survival data was utilised to validate these prognostic signatures. Figure created with BioRender.com.

5.4 Results

5.4.1 Data search and selection

A systematic literature and database search was executed to identify relevant studies containing tumour transcriptomic profiling and survival data from breast cancer patients. This search spanned multiple databases, including PubMed, Science Direct, and the Gene Expression Omnibus (GEO). The search terms “human”, “triple-negative breast” or “TNBC”, and “cancer” were used to identify eligible datasets. Only studies that provided comprehensive tumour transcriptomic expression data alongside immunohistochemistry results of ER and PR, and in situ hybridisation for HER2 were selected. Since TNBC cases represent a minority of breast cancers in archival datasets, it was necessary to assemble a meta-dataset comprising just the TNBC cases from multiple published breast cancer gene expression studies. Since the majority of archival gene expression data derive from Affymetrix microarrays, the meta dataset was assembled exclusively from studies utilising the Affymetrix Human Genome U133 Plus 2.0 Array platform to avoid the need to combine data from non-matching probe sequences for individual genes. Furthermore, only studies with at least 40 TNBC cancer patients that included clinical survival information were considered.

Altogether 17 datasets were identified with at least some clinical information for the included patients. All datasets are listed in **Table 5.1**. Datasets that lacked survival data, conducted on different platforms, or had fewer than 40 TNBC cases were ineligible. A total of 3 eligible datasets representing 235 unique TNBC patients were finally consolidated into a unified gene expression and clinical dataset, the TNBC Cohort, for further analysis.

5.4.2 Assembly of the TNBC cohort

The TNBC Cohort was established by compiling gene expression profiles from 235 TNBC patients, each accompanied by detailed clinical information. The profiles were sourced from GEO, furthermore, immunohistochemistry staining results for ER and PR, and in situ hybridisation for HER2 provided in each dataset verified the TNBC status for each sample confirming their eligibility for inclusion in this TNBC-focused investigation.

The TNBC cohort included 148 cases annotated with overall survival and 235 cases had relapse-free survival information with a majority (70.64%) of patients aged 50 years and above, while 29.36% were younger than 50. Tumour size varied, with 10.64% of cases exhibiting sizes less than 2.0 cm, 30.64% falling between 2.0 and 5.0 cm, and another 10.64% larger than 5.0 cm. Notably, a significant proportion (48.09%) had missing data for tumour size. Tumour grade distribution revealed 2.98% classified as grade 1, 7.66% as grade 2, and 25.96% as grade 3, with the majority (63.40%) having missing grade information. Node status indicated that 21.70% of patients had no affected nodes, 14.47% had one affected node, and 63.83% had missing data (**Table 5.2**)

Table 5.1 Overview of triple-negative breast cancer microarray studies.

GEO ID	Platform used	Sample size	Survival data	Receptor status	Clinicopathological data	Reason for exclusion	Eligible	Reference
GSE226289	Affymetrix Human Gene 2.0 ST Array	147	Not available	Available	Available	No survival data	No	(166)
GSE57297	Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray 039381	3	Available	Available	Not Available	Sample size too small	No	(167)
GSE38959	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	30	Not available	Available	Available	Sample size too small	No	(168)
GSE43358	Affymetrix Human Genome U133 Plus 2.0 Array	17	Not available	Available	Available	Sample size too small	No	(169)
GSE76275	Affymetrix Human Genome U133 Plus 2.0 Array	198	Available	Available	Available	Different platform	No	(170)
GSE27447	Affymetrix Human Gene 1.0 ST Array	14	Not available	Available	Not Available	Sample size too small	No	(171)
GSE76250	Affymetrix Human Transcriptome Array 2.0	165	Not available	Not available	Available	No survival data	No	(172)
GSE64790	Agilent-062918 OE Human lncRNA Microarray V4.0	3	Not available	Available	Not Available	Sample size too small	No	(173)
GSE31519	Affymetrix Human Genome U133A Array	579	Available	Not available	Available	Different platform	No	(174)
GSE18864	Affymetrix Human Genome U133 Plus 2.0 Array	38	Not available	Available	Available	Sample size too small	No	(175)
GSE58812	Affymetrix Human Genome U133 Plus 2.0 Array	107	Available	Available	Available	-	Yes	(163)
GSE83937	Affymetrix Human Genome U133 Plus 2.0 Array	131	Not available	Available	Not Available	No survival data	No	(163)
GSE76124	Affymetrix Human Genome U133 Plus 2.0 Array	198	Not available	Available	Available	No survival data	No	(71)
GSE95700	Affymetrix Human Genome U133 Plus 2.0 Array	57	Not available	Available	Not Available	No survival data	No	(176)
GSE106977	Affymetrix Human Transcriptome Array 2.0	119	Not available	Available	Not Available	No survival data	No	(177)
GSE21653	Affymetrix Human Genome U133 Plus 2.0 Array	87	Available	Available	Available	-	Yes	(178)
GSE65194	Affymetrix Human Genome U133 Plus 2.0 Array	41	Available	Available	Available	-	Yes	(179)

Table 5.2 Clinical characteristics of the TNBC patients.

Characteristics	n	%
Number of samples	235	-
Age		
< 50	69	29.36
≥ 50	166	70.64
Menopausal status		
Pre	15	6.38
Post	26	11.06
Missing	194	82.55
Tumour size (cm)		
<2.0	25	10.64
2.0-5.0	72	30.64
>5.0	25	10.64
Missing	113	48.09
Tumour grade		
1	7	2.98
2	18	7.66
3	61	25.96
Missing	149	63.40
Node Status		
0	51	21.70
1	34	14.47
Missing	150	63.83
Overall Survival		
Months - median	63.15	
Missing	87	37.02
Disease-Free Survival		
Months - median	66.05	
Missing	0	0.00

n = number of samples; % = percentage of samples

5.4.3 Data preparation and preprocessing

The raw microarray expression (CEL) files were downloaded from GEO. The Affymetrix U133 PLUS 2.0 CEL files from GSE21653, GSE58812 and GSE65194 were processed and stored in an expressionSet object in R studio. The expressionSet, comprising 54,613 probes was then normalised using the robust multi-array average (RMA) algorithm (180) in R. To mitigate potential batch effects between gene expression data produced in different laboratories using the sample platform, we applied the "ComBat" function in the "sva" package (181) for batch correction. This approach entailed the fitting of a linear model to the expression data, followed by the application of ComBat to adjust for batch-induced variations while preserving the intrinsic biological variability.

5.4.4 TNBC subtype classification and dataset segmentation

The subtyping of the TNBC samples was performed using a sequential classification approach developed based on proteomic data in a cohort of 312 TNBCs (manuscript under review). In this approach, lasso regression with 10-fold cross-validation was used to train logistic models to classify samples into distinct molecular subtypes. Each subtype was characterized by a specific set of subtype-defining proteins identified through the analysis. The classification process comprised several steps to ensure accurate subtype assignment. In the first step, the LAR subtype was distinguished from others. Subsequently, samples classified as non-LAR underwent a secondary classification step to discriminate between the basal-like breast cancer (BLBC) and non-BLBC subtypes. Within the BLBC subset, another model identified basal-like immune-activated (BLIA)

subtype samples, while non-BLBC samples were subjected to a final classification step to identify mesenchymal (MES1) subtype samples.

The logistic model was applied to the transcriptomic TNBC meta-dataset compiled for this study to classify samples into 5 subtypes: BLIA, basal-like Immune Suppressed (BLIS), luminal androgen receptor (LAR), MES1 and not otherwise specified (NOS).

The distribution of the TNBC cohort across the identified subtypes is summarised in **Table 5.3**. Among the TNBC subtypes, the NOS subtype constituted the largest proportion, comprising 32.77% (n = 77) of the cohort. Following NOS, the BLIA subtype was the next most prevalent, representing 30.64% (n = 72) of the samples. The BLIS subtype accounted for 21.28% (n = 50) of the cohort, the LAR subtype exhibited a minimal representation, comprising only 0.03% (n = 7) of the samples. Lastly, the MES1 subtype constituted 12.34% (n = 29) of the TNBC cohort (**Table 5.3**). Due to the low number of samples in the LAR subtype, further investigation of this subtype was not pursued.

Table 5.3 Classification of samples from the TNBC cohort into TNBC subtypes

TNBC Subtypes	n	%
BLIA	72	30.64
BLIS	50	21.28
LAR	7	0.03
MES1	29	12.34
NOS	77	32.77

TNBC = Triple Negative Breast Cancer; BLIA = Basal-like Immune Activated; BLIS = Basal-like Immune Stimulated; LAR = Luminal Androgen Receptor; MES Mesenchymal; NOS = Non-Otherwise Specified

5.4.5 Feature selection and prognostic signature identification

The subtype annotated TNBC dataset was randomly divided into training and test sets to identify both subtype-specific and non-subtype-specific prognostic gene signatures. This process involved

randomising the samples while maintaining the ratio of surviving and deceased outcomes to ensure a balanced representation of survival events in both sets. The randomise function in R studio was utilised for this purpose, which randomly selected samples from each subtype while considering the distribution of survival events. Specifically, the function assigned a proportionally matched number of samples with good and bad outcomes to the training and validation sets. After randomisation, the training set was used to identify prognostic gene signatures associated with each TNBC subtype through a combination of univariate Cox regression to identify survival-associated genes followed by feature selection. The randomisation process was performed five times, followed by Cox regression analysis; only the genes that were significantly associated with survival across all five iterations were then selected for feature selection analysis. The feature selection analysis used elastic net regularisation (182) to identify a minimal prognostic gene signature. The test set was used to validate the chosen prognostic markers.

5.4.5.1 Univariate Cox regression analysis

Firstly, to investigate the prognostic value of individual genes within subtype-specific and non-subtype-specific TNBC cancers, a univariate Cox proportional hazards regression analysis was performed. This statistical method was utilized to examine the association between the expression levels of all genes and disease-free survival (DFS) across patients within the training cohort of each subtype. For each gene represented in the dataset, the expression data were merged with clinical data. Each gene was checked for sufficient expression levels across samples; genes expressed in fewer than 50% of cases were excluded from further analysis to maintain robustness. In the TNBC subtypes, BLIS, BLIA and MES1, 64, 34 and 33 genes respectively were found to be significantly associated with survival ($p < 0.05$, **Table 5.4**). For the non-subtyped cohort, 2808

genes were significantly associated with DFS ($p<0.05$). However, due to the large number of genes returned, only those genes (n = 86) that were significantly associated with DFS with a $p<0.01$ were selected for feature selection for non-subtype-specific prognostic markers (**Table 5.5**).

Table 5.4 Significantly associated ($p < 0.04$) gene from the TNBC subtypes obtained from univariate Cox regression.

Entrez Gene ID	Gene Symbol	Gene name	Subtype
56895	AGPAT4	1-acylglycerol-3-phosphate O-acyltransferase 4	BLIA
492	ATP2B3	ATPase plasma membrane Ca2+ transporting 3	BLIA
84529	C15orf41	CDAN1 interacting nuclease 1	BLIA
246184	CDC26	cell division cycle 26	BLIA
92211	CDHR1	cadherin related family member 1	BLIA
1052	CEBPD	CCAAT enhancer binding protein delta	BLIA
26190	FBXW2	F-box and WD repeat domain containing 2	BLIA
113510	HELQ	helicase, POLQ like	BLIA
23308	ICOSLG	inducible T cell costimulator ligand	BLIA
11202	KLK8	kallikrein related peptidase 8	BLIA
9663	LPIN2	lipin 2	BLIA
29088	MRPL15	mitochondrial ribosomal protein L15	BLIA
91775	NXPE3	neurexophilin and PC-esterase domain family member 3	BLIA
5228	PGF	placental growth factor	BLIA
54542	RC3H2	ring finger and CCCH-type domains 2	BLIA
25942	SIN3A	SIN3 transcription regulator family member A	BLIA
6496	SIX3	SIX homeobox 3	BLIA
26050	SLITRK5	SLIT and NTRK like family member 5	BLIA
23514	SPIDR	scaffold protein involved in DNA repair	BLIA
257397	TAB3	TGF-beta activated kinase 1	BLIA
6934	TCF7L2	transcription factor 7 like 2	BLIA
148022	TICAM1	TIR domain containing adaptor molecule 1	BLIA
7086	TKT	transketolase	BLIA
348825	TPRXL	tetrapeptide repeat homeobox like	BLIA
22954	TRIM32	tripartite motif containing 32	BLIA
7541	ZBTB14	zinc finger and BTB domain containing 14	BLIA
57684	ZBTB26	zinc finger and BTB domain containing 26	BLIA
132625	ZFP42	ZFP42 zinc finger protein	BLIA
84307	ZNF397	zinc finger protein 397	BLIA
163081	ZNF567	zinc finger protein 567	BLIA

115509	ZNF689	zinc finger protein 689	BLIA
7789	ZXDA	zinc finger X-linked duplicated A	BLIA
38	ACAT1	acetyl-CoA acetyltransferase 1	BLIS
133	ADM	adrenomedullin	BLIS
203	AK1	adenylate kinase 1	BLIS
54890	ALKBH5	alkB homolog 5, RNA demethylase	BLIS
8539	API5	apoptosis inhibitor 5	BLIS
81873	ARPC5L	actin related protein 2/3 complex subunit 5 like	BLIS
285973	ATG9B	autophagy related 9B	BLIS
664	BNIP3	BCL2 interacting protein 3	BLIS
28970	C11orf54	chromosome 11 open reading frame 54	BLIS
801	CALM1	calmodulin 1	BLIS
317762	CCDC85C	coiled-coil domain containing 85C	BLIS
10344	CCL26	C-C motif chemokine ligand 26	BLIS
8812	CCNK	cyclin K	BLIS
1056	CEL	carboxyl ester lipase	BLIS
1057	CELP	carboxyl ester lipase pseudogene	BLIS
138009	DCAF4L2	DDB1 and CUL4 associated factor 4 like 2	BLIS
57062	DDX24	DEAD-box helicase 24	BLIS
1729	DIAPH1	diaphanous related formin 1	BLIS
54878	DPP8	dipeptidyl peptidase 8	BLIS
84985	FAM83A	family with sequence similarity 83 member A	BLIS
23307	FKBP15	FKBP prolyl isomerase family member 15	BLIS
55338	FLJ11292	long intergenic non-protein coding RNA 1949	BLIS
80078	FLJ13744	lung cancer associated lncRNA 1	BLIS
285759	FLJ34503	long intergenic non-protein coding RNA 2880	BLIS
23360	FNBP4	formin binding protein 4	BLIS
2632	GBE1	1,4-alpha-glucan branching enzyme 1	BLIS
2733	GLE1	GLE1 RNA export mediator	BLIS
9950	GOLGA5	golgin A5	BLIS
9555	H2AFY	macroH2A.1 histone	BLIS
3156	HMGCR	3-hydroxy-3-methylglutaryl-CoA reductase	BLIS
3605	IL17A	interleukin 17A	BLIS
3570	IL6R	interleukin 6 receptor	BLIS
128239	IQGAP3	IQ motif containing GTPase activating protein 3	BLIS
57501	KIAA1257	cilia and flagella associated protein 92	BLIS
81606	LBH	LBH regulator of WNT signaling pathway	BLIS
730202	LOC730202	PAPOLA divergent transcript	BLIS
5599	MAPK8	mitogen-activated protein kinase 8	BLIS

2847	MCHR1	melanin concentrating hormone receptor 1	BLIS
65220	NADK	NAD kinase	BLIS
4686	NCBP1	nuclear cap binding protein subunit 1	BLIS
57486	NLN	neurolysin	BLIS
140688	NOL4L	nucleolar protein 4 like	BLIS
10914	PAPOLA	poly	BLIS
5210	PFKFB4	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4	BLIS
123	PLIN2	perilipin 2	BLIS
114971	PTPMT1	protein tyrosine phosphatase mitochondrial 1	BLIS
10111	RAD50	RAD50 double strand break repair protein	BLIS
6009	RHEB	Ras homolog, mTORC1 binding	BLIS
NA	RP11-73M18.7	Kinesin light chain 1 (KLC)	BLIS
10802	SEC24A	SEC24 homolog A, COPII coat complex component	BLIS
6461	SHB	SH2 domain containing adaptor protein B	BLIS
9121	SLC16A5	solute carrier family 16 member 5	BLIS
6518	SLC2A5	solute carrier family 2 member 5	BLIS
55671	SMEK1	protein phosphatase 4 regulatory subunit 3A	BLIS
6819	SULT1C2	sulfotransferase family 1C member 2	BLIS
6850	SYK	spleen associated tyrosine kinase	BLIS
55761	TTC17	tetratricopeptide repeat domain 17	BLIS
7416	VDAC1	voltage dependent anion channel 1	BLIS
7422	VEGFA	vascular endothelial growth factor A	BLIS
664701	ZNF826P	zinc finger protein 826, pseudogene	BLIS
54753	ZNF853	zinc finger protein 853	BLIS
119	ADD2	adducin 2	MES1
57538	ALPK3	alpha kinase 3	MES1
344905	ATP13A5	ATPase 13A5	MES1
54880	BCOR	BCL6 corepressor	MES1
56905	C15orf39	chromosome 15 open reading frame 39	MES1
55765	C1orf106	innate immunity activator	MES1
843	CASP10	caspase 10	MES1
835	CASP2	caspase 2	MES1
57545	CC2D2A	coiled-coil and C2 domain containing 2A	MES1
54480	CHPF2	chondroitin polymerizing factor 2	MES1
56521	DNAJC12	DnaJ heat shock protein family	MES1
728489	DNLZ	DNL-type zinc finger	MES1
8291	DYSF	dysferlin	MES1
151354	FAM84A	LRAT domain containing 1	MES1
2177	FANCD2	FA complementation group D2	MES1

53833	IL20RB	interleukin 20 receptor subunit beta	MES1
54891	INO80D	INO80 complex subunit D	MES1
4094	MAF	MAF bZIP transcription factor	MES1
4619	MYH1	myosin heavy chain 1	MES1
123	PLIN2	perilipin 2	MES1
5817	PVR	PVR cell adhesion molecule	MES1
6009	RHEB	Ras homolog, mTORC1 binding	MES1
7355	SLC35A2	solute carrier family 35 member A2	MES1
55343	SLC35C1	solute carrier family 35 member C1	MES1
91683	SYT12	synaptotagmin 12	MES1
6920	TCEA3	transcription elongation factor A3	MES1
222865	TMEM130	transmembrane protein 130	MES1
10758	TRAF3IP2	TRAF3 interacting protein 2	MES1
7336	UBE2V2	ubiquitin conjugating enzyme E2 V2	MES1
29062	WDR91	WD repeat domain 91	MES1
331	XIAP	X-linked inhibitor of apoptosis	MES1
653121	ZBTB8A	zinc finger and BTB domain containing 8A	MES1
162967	ZNF320	zinc finger protein 320	MES1
91661	ZNF765	zinc finger protein 765	MES1

Table 5.5 Genes significantly associated ($p < 0.01$) with DFS from the non-subtype specific TNBC cohort obtained from univariate Cox regression.

Entrez Gene ID	Symbol	Gene Name
176	ACAN	aggrecan
92949	ADAMTSL1	ADAMTS like 1
133	ADM	adrenomedullin
158798	AKAP14	A-kinase anchoring protein 14
51129	ANGPTL4	angiopoietin like 4
8862	APLN	apelin
361	AQP4	aquaporin 4
51130	ASB3	ankyrin repeat and SOCS box containing 3
467	ATF3	activating transcription factor 3
27241	BBS9	Bardet-Biedl syndrome 9
623	BDKRB1	bradykinin receptor B1
624	BDKRB2	bradykinin receptor B2
387763	C11orf96	chromosome 11 open reading frame 96
125988	MICOS13	mitochondrial contact site and cristae organizing system subunit 13 (C19orf70)
84886	C1orf198	chromosome 1 open reading frame 198
254158	CXorf58	chromosome X open reading frame 58
8701	DNAH11	dynein axonemal heavy chain 11
51277	DNAJC27	DnaJ heat shock protein family
1809	DPYSL3	dihydropyrimidinase like 3
NA	DQ570096	Hs piRNA piR-30208 (DQ570096)
11072	DUSP14	dual specificity phosphatase 14
2012	EMP1	epithelial membrane protein 1
57488	ESYT2	extended synaptotagmin 2
2252	FGF7	fibroblast growth factor 7
2632	GBE1	1,4-alpha-glucan branching enzyme 1
84656	GLYR1	glyoxylate reductase 1 homolog
1839	HBEGF	heparin binding EGF like growth factor
3269	HRH1	histamine receptor H1
10808	HSPH1	heat shock protein family H
10994	ILVBL	ilvB acetolactate synthase like
54928	BPNT2	3'(2'), 5'-bisphosphate nucleotidase 2 (IMPAD2)
3678	ITGA5	integrin subunit alpha 5
85450	ITPRIP	inositol 1,4,5-trisphosphate receptor interacting protein
3745	KCNB1	potassium voltage-gated channel subfamily B member 1
26524	LATS2	large tumor suppressor kinase 2

157381	LINC00964	long intergenic non-protein coding RNA 964
100507537	STRIT1	small transmembrane regulator of ion transport 1 (LOC100507537)
NA	LOC101927705	uncharacterized LOC101927705
101928728	LOC101928728	uncharacterized LOC101928728
729966	CIST1	LOC729966. Long non-coding RNA. Colon, intestine and stomach enriched 1
5469	MED1	mediator complex subunit 1
92140	MTDH	metadherin
10397	NDRG1	N-myc downstream regulated 1
64579	NDST4	N-deacetylase and N-sulfotransferase 4
50507	NOX4	NADPH oxidase 4
27253	PCDH17	protocadherin 17
27295	PDLIM3	PDZ and LIM domain 3
5228	PGF	placental growth factor
8073	PTP4A2	protein tyrosine phosphatase 4A2
5990	RFX2	regulatory factor X2
6004	RGS16	regulator of G protein signaling 16
80196	RNF34	ring finger protein 34
NA	RP11-274H2.5	ENSG00000240032
6224	RPS20	ribosomal protein S20
55511	SAGE1	sarcoma antigen 1
6391	SDHC	succinate dehydrogenase complex subunit C
5054	SERPINE1	serpin family E member 1
144195	SLC2A14	solute carrier family 2 member 14
6548	SLC9A1	solute carrier family 9 member A1
84679	SLC9A7	solute carrier family 9 member A7
284297	SSC5D	scavenger receptor cysteine rich family member with 5 domains
54434	SSH1	slingshot protein phosphatase 1
7024	TFCP2	transcription factor CP2
148022	TICAM1	TIR domain containing adaptor molecule 1
257000	TINCR	TINCR ubiquitin domain containing
84548	TMEM185A	transmembrane protein 185A
54210	TREM1	triggering receptor expressed on myeloid cells 1
125488	TTC39C	tetratricopeptide repeat domain 39C
55075	UACA	uveal autoantigen with coiled-coil domains and ankyrin repeats
7332	UBE2L3	ubiquitin conjugating enzyme E2 L3
65110	UPF3A	UPF3A regulator of nonsense mediated mRNA decay
7422	VEGFA	vascular endothelial growth factor A

331	XIAP	X-linked inhibitor of apoptosis
7534	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta

5.4.5.2 Minimal gene set identification through elastic net regularisation modelling

To identify a prognostic gene signature, we employed elastic net regularisation, a machine-learning technique that combines ridge and lasso regression methods (182). Ridge regression moderates the influence of closely related features, thereby enhancing the model's stability and preventing overreliance on any group of characteristics (183). Conversely, lasso regression (184) identifies and retains only the most impactful features, eliminating those deemed less relevant. Utilising elastic net regularisation allowed for the selection of the most influential genes while mitigating the risk of overfitting, a common pitfall encountered in complex model development.

By adjusting a parameter known as alpha (α) from 0.2 to 0.9, which controls the ratio of contribution of ridge and lasso regression components to the final result (182), a spectrum of models ranging from predominantly ridge to predominantly lasso regression was explored. The objective was to determine the optimal α value that yielded a gene signature strongly associated with patient survival while minimising the number of genes included in the signature. The glmnet package (185) in R was employed for this analysis, with a cross-validation approach to select the optimal lambda parameter for each alpha setting.

To select the optimal α, the resulting model for each α was applied to generate a score for each patient sample. The sample score is the weighted mean of the expression values of the signature genes, with each gene's expression value weighted by its respective model coefficient. Cox regression was used to test the association of these scores against patient survival. The α with the minimal number of genes and with the best Cox regression p-value was selected as the final model.

The point at which the curve ceases to show a marked decrease in gene list size with increasing alpha (**Figure 5.2**), while maintaining a significant Cox regression p-value, was selected as the ideal model (**Figure 5.3**).

The analysis encompassed three TNBC subtypes: BLIA, BLIS, and MES1, as well as the non-subtyped TNBC group. In the BLIA subtype, an α value of 0.6 with a gene list of 18 and a p-value of 0.00085 was chosen. For the BLIS subtype, an α value of 0.3 with a gene list of 25 and a p-value of 0.00075 was chosen. For the MES1, an α value of 0.4 with a gene list of 15 and a p-value of 0.0099 was chosen. For the non-subtyped TNBC group, an α value of 0.4 with a gene list of 49 and a p-value of 0.025 was chosen.

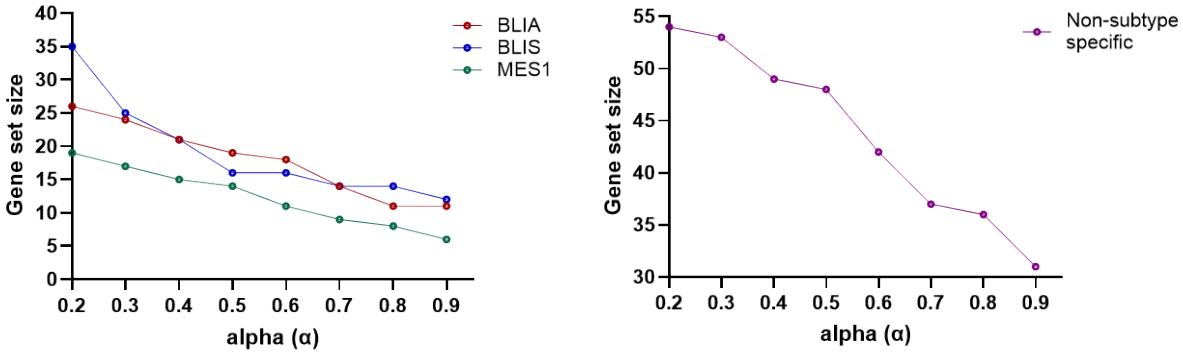


Figure 5.2: Gene set sizes across different alpha values obtained from elastic net regularisation modelling. Line plots demonstrating the relationship between the alpha parameter and the number of genes in the resulting gene signatures for both (A) subtype-specific and (B) non-subtype-specific groups within the TNBC cohort.

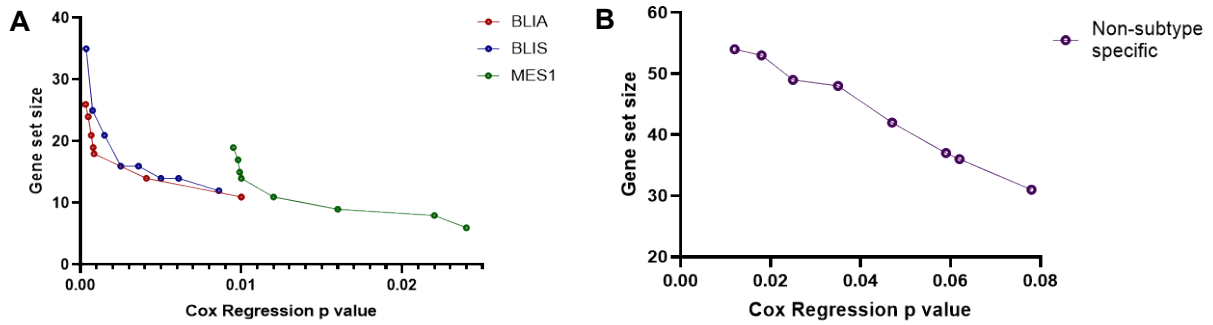


Figure 5.3 Gene set sizes and their associated Cox regression p-values obtained from elastic net regularisation modelling. Line plots depict the p-value and the number of genes in the resulting gene signatures for (A) subtype-specific and (B) non-subtype-specific groups within the TNBC cohort.

5.4.6 Prognostic utility of the models

To evaluate the predictive ability and the prognostic utility of the gene signatures, Cox regression and Kaplan Meier analysis were used. The validation cohort was stratified into three risk groups, high, medium, and low, based on tertiles of the distribution of the sample scores. Additionally, a binary stratification was also performed where patients were categorized into two groups based on the median score: above median being higher risk and below median being lower risk.

Kaplan-Meier survival plots were generated to visually assess the DFS probabilities over 10 years across the different risk groups. These plots were annotated with Cox regression p-values and log-rank test p-values to indicate the statistical significance of the observed differences in DFS and statistical differences in survival between risk groups, respectively.

Both the subtype-specific and non-subtype-specific prognostic gene signatures demonstrated significant value in predicting DFS outcomes after stratifying into binary or tertiary groups (**Figure 1.3**). The subtype-specific signatures demonstrated the strongest power in discriminating the risk groups, with the BLIS signature demonstrating the strongest performance (Cox regression p value

= 0.00075) (**Figure 5.4**). The non-subtype specific signature also showed statistically significant discrimination in both the binary and tertiary stratified validation cases although the level of significance was lower for both log-rank and Cox regression p values (**Figure 5.5**).

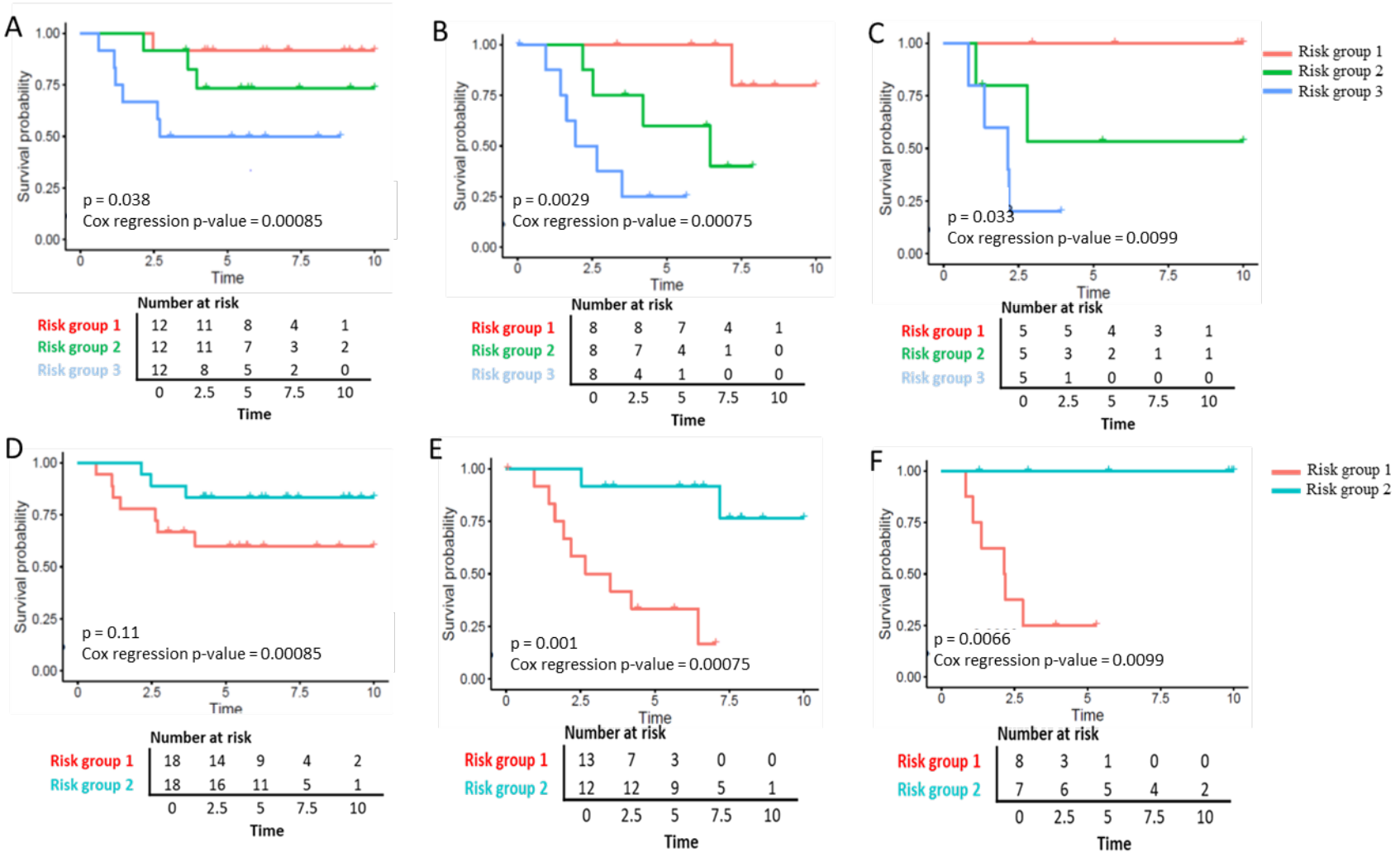


Figure 5.4: Performance of subtype specific gene signatures in discriminating patient samples according to risk of recurrence. Kaplan-Meier survival curves demonstrate the DFS survival probabilities over 10 years in tertiary stratified (A) BLIA, (B) BLIS and (C) MES1 and binary stratified (D) BLIA, (E) BLIS and (F) MES1 TNBC groups. Log-rank and Cox regression p values are shown.

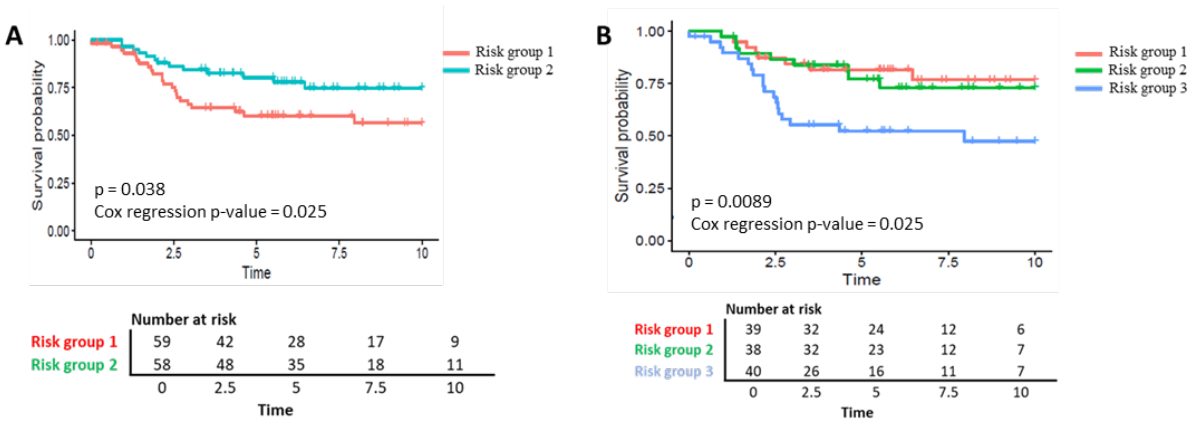


Figure 5.5: Performance of the non-subtype specific gene signature in discriminating patient samples according to risk of recurrence. Kaplan-Meier survival curves of (A) binary and (B) tertiary stratified patient samples based on risk predicted using the non-subtype specific gene signature over 10 years in the TNBC validation cohort.

5.4.7 Independent validation of the prognostic gene signatures

To independently validate the newly identified TNBC subtype-specific and non-subtype-specific prognostic gene signatures, their performance was tested in the 324 TNBC cases contained within the METABRIC gene expression dataset (114) for which detailed clinical annotation was also available. The same analysis method as before was applied to this independent cohort. The signatures performed poorly in this cohort, with only the MES1 signature achieving a significant Cox regression p-value among the subtype-specific signatures (**Figure 5.6**) and the non-subtype-specific gene signature showing weak significance (p = 0.012) (**Figure 5.7**).

For binary stratification, non-significant log-rank test p-values were observed for the subtype-specific gene signatures and a significant log-rank test p-value was observed for the non-subtype-specific gene signature (p = 0.027). For tertiary stratification, non-significant log-rank test p values were observed for all prognostic gene signatures.

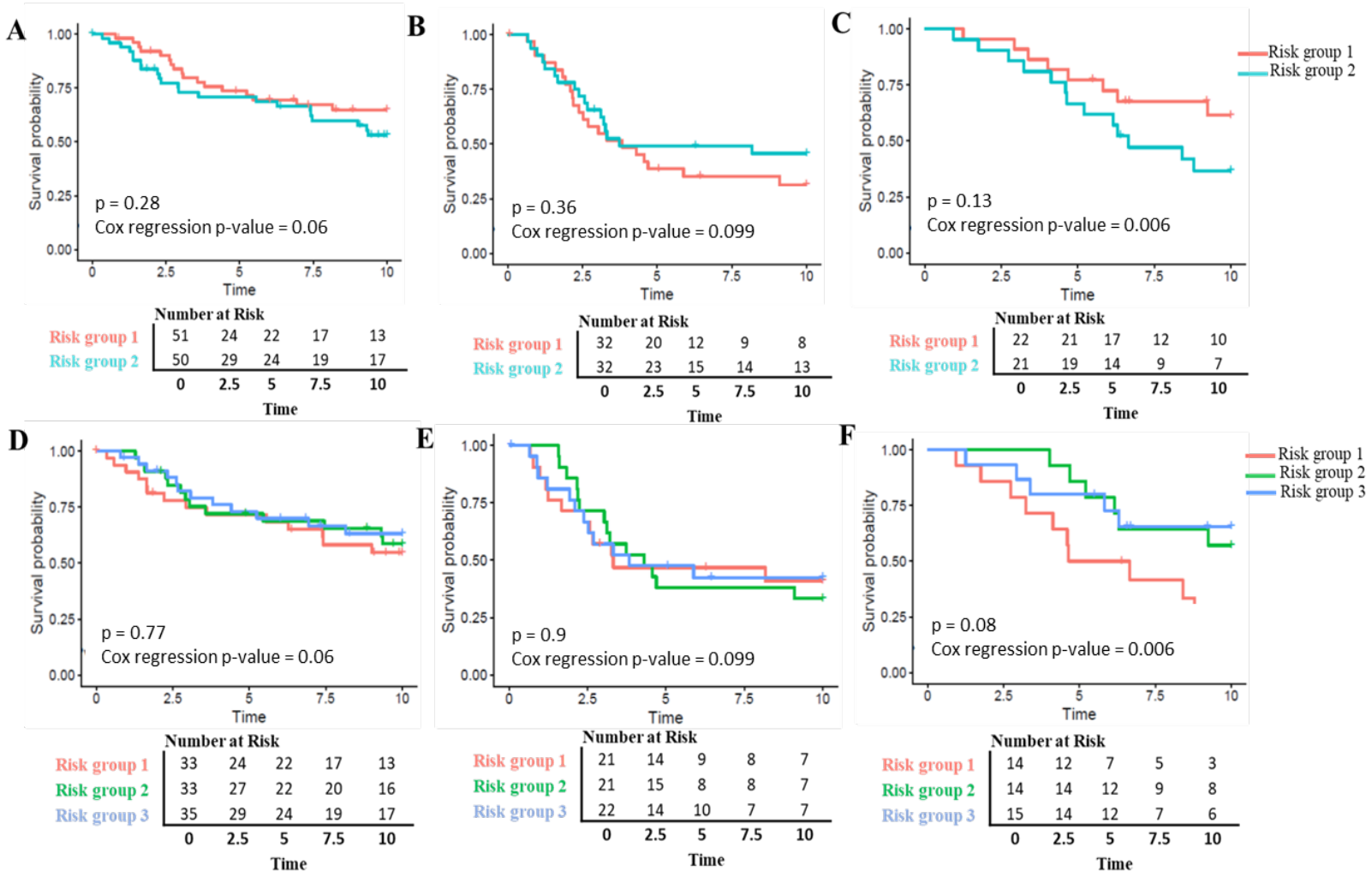


Figure 5.6: Performance of subtype-specific gene signatures in discriminating patient samples according to the risk of recurrence in the METABRIC TNBC validation cohort. Kaplan-Meier survival curves of binary stratified (A) BLIA, (B) BLIS and (C) MES1, and tertiary stratified (D) BLIA, (E) BLIS and (F) MES1 TNBC groups based on predicted risk of recurrence over 10 years.

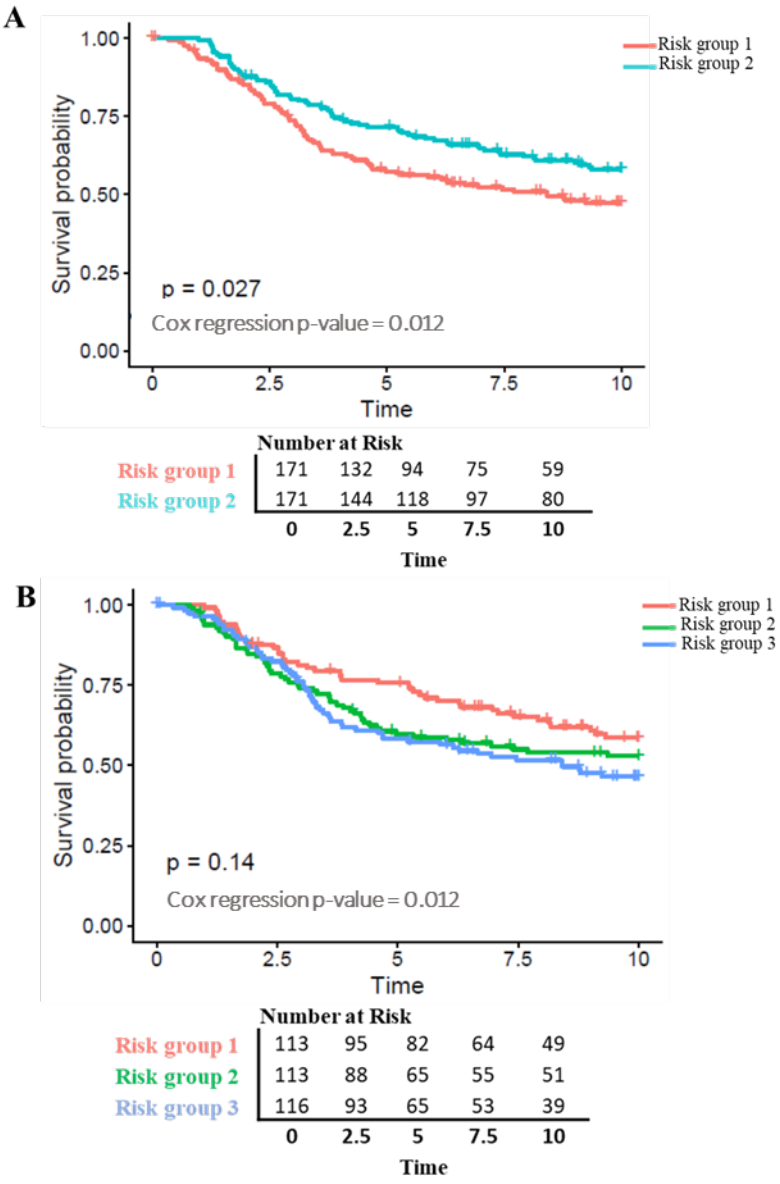


Figure 5.7: Performance of non-subtype specific gene signatures in discriminating patient samples according to the risk of recurrence in the METABRIC TNBC cohort. Kaplan-Meier survival analysis in (A) binary and (B) tertiary stratified cases based on non-subtype specific signature-based risk prediction over 10 years.

5.5 Discussion

5.5.1 Summary of results

In this chapter, an objective was to develop and validate new TNBC subtype-specific and non-subtype-specific prognostic gene signatures using *in silico* analysis. It was hypothesized that in comparison to the non-subtype-specific group, subtype-specific prognostication would provide a more accurate and clinically relevant prediction of patient outcomes. The analysis in this study focused on three TNBC subtypes: BLIA, BLIS, and MES1, as well as the non-subtyped TNBC group. These three groups are the largest and most uniform of the gene signature defined TNBC subgroups. The LAR group, while quite homogeneous and usually representing some of the poorest outcome cases, is generally found to be the smallest subset of TNBC cases (186). The NOS group is a heterogeneous group classified mainly by its lack of classification elsewhere, suggesting that the prospect of discovering a well-conserved predictive signature in this group would be very low.

First, a TNBC cohort was established through a systematic search and selection of datasets that contained transcriptomic and survival data. Out of 17 identified datasets, three major datasets were selected for further analysis based on their sample size and availability of clinical data, resulting in a cohort of 235 samples. Subtyping this cohort revealed that the majority of the samples were BLIA, followed by BLIS. The LAR subtype represented only 0.03% of the entire cohort and due to its minimal representation, this subtype was excluded from further investigation.

The TNBC cohort was divided into training and test sets, and univariate Cox regression was performed on the training set to identify survival-associated genes. For the TNBC subtypes, BLIS, BLIA and MES1 had 64, 34 and 33 significant genes respectively, whereas the non-subtyped

cohort had 86 significant genes. Utilising the elastic net regularisation approach (185), the optimal α was determined to identify the minimal gene sets for the identification of prognostic signatures. For BLIA, α = 0.6 yielded 18 genes; for BLIS, α = 0.3 yielded 25 genes; for MES1, α = 0.4 yielded 15 genes; and for non-subtyped TNBC, α = 0.4 yielded 49 genes. Interestingly, there was no overlap between the five genes utilized in the PROSPER signature for ER+ breast cancer and any of the TNBC signatures discovered, underscoring the striking difference between these breast cancer types.

These gene signatures were then validated in the test cohort, and the results revealed the subtype-specific prognostic gene signatures allowed for more precise survival outcome prediction compared to a non-subtype-specific gene signature. The patients were stratified into different risk groups based on the sample scores generated from the gene signatures. Kaplan-Meier survival analysis revealed significant differences in survival probabilities across groups stratified based on predicted risk of recurrence. Tertiary stratification yielded statistically significant log-rank p values for the BLIA, BLIS, and MES1 prognostic gene signatures, indicating their potential as prognostic markers for TNBC. In contrast, the two-risk stratification method was only significant for the BLIS and MES1 gene signatures, suggesting that these signatures may be more informative in predicting patient survival in TNBC. Although the non-subtype-specific TNBC gene signature also showed prognostic utility with statistically significant p-values for both the binary and tertiary stratifications, its performance was less robust compared to the subtype-specific signatures.

In the validation analysis of the TNBC subtype-specific and non-subtype-specific prognostic gene signature in an independent cohort, it was found that the subtype-specific signatures for BLIA and BLIS did not show significant associations with survival. However, the MES1 subtype prognostic signature showed a weak but significant association with survival, indicating the potential for this

signature to be utilised as a predictive tool. Additionally, the non-subtype specific prognostic gene signature showed statistical power in the Cox regression analysis and a significant differentiation in survival was observed with binary stratification based on the Kaplan-Meier analysis, though no significance was observed with tertiary stratification. This implies that while the signature may have some predictive value, its effectiveness could depend on context and patient stratification methods, impacting the outcome of the prognostic assessment and limiting its predictive value.

5.5.2 Preliminary exploration of subtype-specific prognostic signatures in TNBC

TNBC is a highly heterogeneous subtype of breast cancer, characterised by the absence of ER and PR, and non-amplified HER2 (187). Unlike other breast cancer subtypes, durable targeted therapy is still not available for TNBC and chemotherapy remains the primary systemic treatment in both early and advanced stages of the disease (188). Furthermore, TNBC patients typically experience a relapse within five years after surgery, facing a much poorer overall prognosis compared to other breast cancer subtypes (189, 190).

These characteristics of TNBC necessitate precise subtyping to improve individualised treatment planning. Several studies have identified molecular subtypes of TNBC based on gene expression profiles (69, 71). The refined signature by Lehmann et al. classified TNBC into four subtypes (BL1, BL2, M and LAR) (69). Burstein et al. and Lie et al. each proposed similar four-subtype models for TNBC with distinct prognoses: LAR, M, BLIS and BLIA subtypes (71). Recently, more studies have examined TNBC classification based on immune signature and microenvironment characteristics (191). However, the variability in sample quality and methodological differences among studies pose a significant challenge to establishing standardised subtyping guidelines for TNBC, impeding consistent application in research and clinical settings.

There are several challenges in attempting to discover stable prognostic signatures for TNBC. Due to the method of grouping TNBC based on features that they lack rather than properties they share, the TNBC subgroup is highly biologically heterogeneous. As a result, gene signatures that perform well in one cohort might not perform well in validation due to this highly variable nature. In addition, the TNBC group is a minority in a highly prevalent cancer type (187). So, most assembled cohorts contain only small subsets of TNBC cases. This creates significant challenges for the assembly of robust and uniform TNBC cohorts. Coupled with the fact that clinical annotation is not universally available on public cohorts, this creates considerable obstacles to obtaining robust and uniform TNBC cohort datasets with high quality clinical annotation.

To address the challenge of heterogeneity, a subtyping approach was employed, which exploited work that had been performed in an unrelated project to develop a sequential strategy to separate TNBC cohorts into uniform subsets that are consistent with the numerous published TNBC subtypes. The rationale for this approach was that more robust predictive tools could be built within these more homogeneous subgroups. These could then be applied in a stepwise fashion, where cases would first be classified into TNBC subtypes, and then the relevant predictive tool could be employed.

Despite the rarity of TNBC and the challenges faced by studies conducted on various gene expression platforms, an extensive search to assemble a combined TNBC cohort, of gene expression profiles and survival outcomes was performed. Three datasets were identified that met the criteria for inclusion, resulting in a dataset comprising 235 TNBC samples with associated clinical outcomes, all performed using the same gene expression platform, Affymetrix Human Genome U133 Plus 2.0 Array. This allowed a preliminary exploration of subtype-specific prognostic gene signatures.

5.5.3 The impact of TNBC subtype-specific prognostication on treatment

Numerous studies have demonstrated that subtype-specific prognostication has improved clinical practice and guided patient-specific treatment planning for various diseases (192-194). Specifically in the context of breast cancer, the disease can be categorised into hormone receptor-positive, HER2+, and triple-negative subtypes, each with distinct treatment strategies (195). Consequently, hormone receptor-positive tumours are commonly treated with hormone receptor-targeted therapies like the anti-estrogen tamoxifen or aromatase inhibitors that block the synthesis of estrogen, and HER2+ cancers can be treated with HER2-targeted therapies such as trastuzumab, significantly improving the prognosis for many of these patients. However, due to the high heterogeneity and aggressive nature of TNBC, treatment options are limited, and therapeutic response is unpredictable (196, 197). However, studies that have further subtyped TNBC to identify specific molecular features could aid in prognosis and treatment selection. In this study, these subtyping methods were leveraged to identify TNBC subtype-specific prognostic signatures, as they acknowledge and utilize the inherent molecular heterogeneity within this breast cancer subtype.

Combining bioinformatics and statistical analysis, potential prognostic gene signatures were identified for BLIS, BLIA and MES1 TNBC subtypes and compared with a signature for non-subtype specific TNBC. Using Cox regression and Kaplan-Meier survival analyses, it was found that the subtype-specific prognostic gene signatures demonstrated a statistically significant correlation with overall survival in their respective TNBC subtypes in the test cohort. The performance of the signatures was evaluated after stratification into two or three risk categories. Assessing binary (high and low risk) and tertiary (high, intermediate and low risk) categorisation is important due to its practical utility in guiding treatment decisions. With better risk

categorisation, clinicians can optimise treatment decisions, where high-risk patients can be recommended more aggressive treatment and low-risk patients can potentially avoid overtreatment (198). In the case of TNBC patients, low-risk patients might be recommended dual anthracycline and cyclophosphamide treatment as standard of care, whereas patients who are considered to be at elevated risk of recurrence, are commonly recommended escalation to taxanes, platinum-based treatments and potentially immune checkpoint inhibitors, in addition to the standard anthracycline-cyclophosphamide combination (199). These treatments have devastating side-effect profiles, and limiting the spectrum of patients recommended for treatment escalation is a strong imperative (199, 200).

This study demonstrated that TNBC subtype-specific gene signatures showed a statistically significant difference in DFS survival probabilities in both groups of risk stratification, with a stronger statistically significant difference with tertiary stratification. In contrast, the non-subtype-specific gene signature also showed statistical differences in DFS survival probabilities across both levels of risk stratification; however, these associations were less robust compared to those of the subtype-specific gene signatures. These findings suggest that the identified subtype-specific prognostic gene signatures had a stronger association with DFS survival outcomes within their respective TNBC subtypes compared to non-subtype-specific gene signatures. Despite the poor validation in the independent TNBC dataset, the findings provided insight into the value of molecular subtyping in TNBC for improving risk prediction in this challenging breast cancer subtype (198). Due to the significant variability in transcriptomic and genomic profiles, TNBC represents a highly heterogeneous disease. This has resulted in a lack of universally suitable molecular targets and limited therapeutic benefit (198, 201). Therefore, categorising TNBC based on biologically and clinically relevant subtypes may aid in disease management and treatment

planning. Specifically, studies have demonstrated that a subtype-specific approach can have strong prognostic value and can guide clinicians in their prognostic assessments, leading to more accurate predictions of disease progression and survival outcomes (192, 194). By avoiding a "one-size-fits-all" approach, subtype-specific prognostication may lead to a more judicious use of healthcare resources, by aligning treatments with those most likely to benefit from them and avoiding unnecessary or less effective interventions for others. Results from the test cohort validation confirmed the literature where subtyping-specific prognostic gene signature design provided a more refined and accurate tool for predicting patients' DFS outcomes compared to signatures that were not specific to individual TNBC subtypes.

5.5.4 Validation of the TNBC subtype-specific prognostic gene signature

To validate the prognostic gene signatures identified in the study, an independent validation was performed using the publicly available gene expression subset of TNBC patients from the METABRIC dataset of gene expression profiles that were generated using the Illumina HT-12v4 microarray platform (114). Kaplan-Meier survival analysis for prognostic gene signature for BLIA, BLIS, MES1 TNBC subtypes and non-subtype specific groups were analysed with both binary and tertiary risk stratification. The performance of the prognostic gene signatures in the independent cohort varied but was overall diminished compared with the dichotomised discovery and test dataset.

After binary risk stratification, the prognostic gene signatures for the BLIA and BLIS subtypes did not show significant results, indicating limited prognostic utility in this risk stratification. The MES1-specific prognostic gene signature showed modest prognostic results, with significant

differences in DFS survival outcomes overall, but a non-significant difference in the two-risk stratification.

The validation after tertiary risk stratification provided mixed outcomes. The gene signatures for BLIA and BLIS were not significantly different between risk groups, indicating no strong prognostic value. On the other hand, the MES1-specific prognostic gene signature showed a significant association with DFS survival outcomes and a marginal significance in the tertiary risk stratification. Conversely, the non-subtype-specific prognostic gene signature showed modest but significant prognostic value in predicting DFS survival outcomes and a significant difference in the two-risk group stratification. While all signatures lost strength in the independent validation cohort, they provided encouraging insight into the value of subtype-focused risk prediction in this challenging and heterogeneous breast cancer subtype.

These results demonstrate the importance of validating prognostic gene signatures in independent datasets. They also highlight the variability in prognostic performance across different subtypes of TNBC. While the MES1-specific signature demonstrated the most robust prognostic utility, particularly with the three-risk group stratification, further refinement and validation are required for the BLIA and BLIS subtypes. Even though these findings did not agree with the study hypothesis, these preliminary findings further support the notion that TNBC is a heterogeneous disease and further emphasise the need for in-depth molecular subtyping strategies required for such a highly heterogeneous disease.

5.5.5 Study strengths, limitations and future directions

This study highlighted several strengths in the methodological approach and provided valuable insights into the superior predictive power of subtype-specific gene signatures for TNBC. The systematic approach of dataset selection for initial *in silico* analysis was a key strength, as this ensured a comprehensive and replicable analysis. This approach also enhanced understanding of TNBC prognostic gene signatures and can be utilised as a foundation for future TNBC studies. Additionally, implementing elastic net regularisation for feature selection analysis strengthened this study (202). By combining lasso and ridge regression principles, elastic net regularisation identified a small gene set with the highest predictive power while reducing the overfitting and false positives for TNBC. It developed a multivariate model and integrated cross-validation analysis to further optimise model parameters, improving the predictive accuracy of the model thereby resulting in a more robust gene signature (185).

This approach, ultilising elastic net regularisation, could be applied to other heterogeneous cancer types, such as lung or ovarian cancer where there are distinct molecular subtypes that may be benefit from targeted treatments. Moreover, this strategy can also be applied to non-cancerous diseases, especially those with underlying molecular heterogeneity, such as autoimmune disorders or metabolic diseases. In these cases, identifying specific gene or molecular pathways associated with disease progression or treatment response could lead to the discovery of new therapeutic targets, improving the overall precision of medical interventions.

However, there are several limitations to the study that should be acknowledged. Firstly, as the analysis was based on publicly available TNBC datasets, it was limited by the relatively low availability of well-annotated TNBC gene expression data. As a result, the sample size of the

assembled TNBC cohort was relatively small due to the relatively low prevalence of this subtype within breast cancer cohorts. This small sample size reduced the statistical power of analysis, potentially making it difficult to draw robust conclusions (203). Moreover, the small number of cases in any individual gene expression dataset necessitated the merging of TNBC subsets from unrelated studies, potentially increasing the heterogeneity and reducing the reproducibility of the analyses performed with this composite dataset. Additionally, utilising publicly available datasets reduces the control over data quality and consistency because the data comes from multiple sources with potentially different data collecting and processing methods (204). Steps were taken to address potential sources of variability by selecting cases derived from the same Affymetrix platform. This ensured consistency in data quality and reduced variability that could arise from using different platforms.

In the curated dataset, there was limited LAR subtype representation and so this subtype was excluded from this study. Therefore, this study was unable to provide insights into prognostic markers for the LAR subtype. The limited sample size in the study explains why a statistically significant association was observed in the non-subtyped group where the cohort was analysed as a whole since the larger dataset size would provide great statistical power in this instance (203). This highlights the challenge of achieving sufficient statistical power in analyses with smaller, subtype-specific cohorts (203). One approach that could be taken to directly compare the subtype-specific and non-specific predictive signature performance would be to down-sample the non-subtype-specific dataset to determine its performance in a similarly sized cohort as the TNBC subtype groups.

Future studies should focus on increasing the sample size to involve larger and more diverse datasets, which will enhance the statistical power and provide a more robust and comprehensive

analysis and explore the prognostic utility of subtype-specific signatures (203). One way this can be done is by collaborating with other research institutions and databases. Alternatively, initiating prospective studies to collect gene expression profiles in additional TNBC samples would enhance the sample size. However, this option would represent a considerably more costly strategy and it would not be practical to generate these data using microarray platforms since most are now discontinued or deprecated. Future directions could include further investigation into the functional significance of the genes within the signature and exploring their potential as therapeutic targets. Additionally, it would be valuable to evaluate the gene signatures' performance in predicting response to specific treatments or in combination with other clinical factors.

While the data presented in this chapter identified certain genes associated with better overall survival, the association does not imply causation. It is possible that the genes are potentially linked to pathways involving tumour progression, response to therapy or reflecting the underlying biological state of the cancer. Furthermore, to establish a causal relationship, functional experiments would be required to demonstrate this. Functional enrichment analysis of the subtype-specific signatures discovered in this chapter did not reveal strong associations with any specific functional characteristics. However, a deeper subset analysis or functional testing of individual genes of interest could lead to the identification of new markers for classifying TNBC cases. The prognostic signatures described here could represent a starting point for such an approach.

5.5.6 Conclusion

In conclusion, *in silico* discovery and validation of novel prognostic gene signatures for TNBC showed variable performance in predicting survival outcomes depending on the subtype-specific approach and stratification method used. Notably, the MES1 TNBC subtype-specific prognostic gene signature identified in this study demonstrated a significant association with DFS survival in

TNBC patients. Although further validation in larger cohorts is needed, these findings provide valuable insights into the potential application of gene signatures for predicting outcomes and consequently tailoring personalised treatment methods in TNBC. This study highlights the clinical value of subtype-specific prognostication in TNBC and provides a foundation for further research in this area. TNBC is a highly heterogeneous disease with substantial diversity within and between tumours, both from a biological and clinical perspective (195). TNBC typically exhibits more aggressive behaviour and quicker metastasis compared to other breast cancer subtypes (195). These challenges highlight the importance of large-scale multicentre studies to develop and validate prognostic and predictive gene signatures. This study adds to the growing body of evidence supporting the potential use of gene signatures in TNBC prognosis, but further research and validation are necessary before such signatures can be integrated into clinical practice.

Chapter 6

Conclusion

6.1 Final Remarks

Breast cancer is the most diagnosed cancer in women with an estimated 685,000 deaths reported in 2020 (205). Current projections indicate that by 2030, the global annual number of new breast cancer cases will reach 2.7 million (206). However, improvements in early detection and better-targeted treatments have raised the 5-year survival rate to over 90%, (206).

Traditional breast cancer prognostication and treatment selection have been guided mainly by clinicopathological factors such as tumour size, nodal status, histological grade, and hormone receptor/HER2 status (4). Yet, the heterogeneous nature of the disease has driven an increased focus on molecular profiling to improve accuracy to identify individuals at high risk of recurrence, guide treatment decisions, and improve overall outcomes. The standard clinicopathological factors are adequate to define clinically useful groups such as TNBC, HER2+, and hormone receptor-positive/HER2+ tumours for which treatment recommendations are seldom controversial. It is within the subset of patients with intermediate grade luminal disease, defined by the presence of ER and/or PR, negative HER2, but moderate proliferation, that uncertainty regarding the optimal treatment most commonly arises, as clinicians seek to avoid over-treatment and under-treatment (4).

In the last decade, the emergence of multi-gene assays such as Oncotype DX (38), Prosigna (112) and EndoPredict (113) have provided valuable genomic insight into the underlying biology of

breast tumours and have been incorporated into clinical treatment pathways to guide adjuvant chemotherapy recommendations for patients with ER+ breast cancer. These gene expression tests can identify those with a high risk of recurrence who may benefit from chemotherapy and those with a low risk who can be spared the toxicity of chemotherapy (4). In the USA, 90% of private health insurers subsidise the cost of these tests to facilitate their widespread adoption. However, here in Australia, the high cost of these molecular tests and the lack of support from health insurers remains a significant barrier to their use. Moreover, most are not offered locally, and patient samples must be sent overseas for testing, adding further costs and delays in the critical next steps in patient care. Therefore, the discovery and development of an affordable alternative prognostic test to predict the risk of recurrence in ER+ breast cancer cases will be a breakthrough in Australia.

Our lab developed an alternative proliferation-based gene signature called PROSPER. In the preliminary validation studies assessing the performance of the PROSPER signature, it was demonstrated that PROSPER performed as well or better than all signatures tested when cases were stratified into predicted low and high-risk groups. When PROSPER results were compared to the results of the EndoPredict commercial test in samples, PROSPER was strongly correlated with EndoPredict suggesting that PROSPER can predict the risk of recurrence. Furthermore, PROSPER was measured in a retrospective cohort of 142 archival breast cancer with known long-term outcomes. The results indicated that PROSPER can accurately stratify cases into good and poor outcomes with a high PROSPER index associated with poor disease-free survival. These findings indicate that the PROSPER gene signature can be a valuable tool in predicting breast cancer recurrence and stratifying patients based on their risk, thereby aiding in personalized treatment decisions for ER+ breast cancer patients. However, measurements of PROSPER in the preliminary studies were conducted using the NanoString nCounter platform (79, 80) . While the

NanoString platform is suitable in the context of a private diagnostic laboratory, the high cost of the equipment and consumables and their ambiguous analysis prevent the technology from being implemented in smaller local facilities for routine use.

Therefore, in this research, the translation of PROPSER into an easily applied diagnostic laboratory test was conducted. qPCR assays are commonly used in a range of routine clinical diagnostic settings. Both Oncotype Dx and EndoPredict utilise qPCR as their foundation technology for their test. The assay is cost-effective and has high sensitivity due to the exponential amplification of the template sample. Moreover, they can be completed in one working day providing rapid results (207, 208). Being a qPCR-based assay, it is estimated to cost under $50 for reagents and consumables. Additionally, the platform required for qPCR assays are readily available in NSW Health Pathology. Furthermore, the assay will be comparable to Ki67 in labour demand and given that the IHC tests for Ki67 costs less than $200, the PROSPER assay will be significantly more affordable than the current multigene tests. Hence, in this research, one of the main objectives was optimizing PROSPER in a qPCR assay for routine clinical application.

SpeeDx proprietary ***Plex****Zyme* technology was employed (96, 115) throughout the thesis. The PlexZyme technology carries several advantages over existing multiplex qPCR approaches. In particular, the nested probe design provides much greater flexibility in target sequence selection and maximises affordability through greater reporter compatibility. **Chapter 3** demonstrated the successful integration of the PROSPER signature from a single plex assay into a multiplexed qPCR assay using the ***Plex****Zyme* assay. The robustness of the multiplexed assay was assessed and validated in different sample types. Furthermore, the performance of the PROSPER signature as a multiplexed qPCR was assessed in progestin treated T-47D breast cancer cells. Given that the PROSPER signature is made up of proliferation-associated targets, it was expected that ORG2058

would decrease the PROSPER signature in these cells. The results indicated that the assay can detect differential expression of the PROSPER genes between treated and untreated cells. This validation was critical since the assay will be utilised to stratify patients with ER+ breast cancer into low and high risk groups which is primarily driven by the proliferation status of the tumours.

Globally FFPE samples are routinely utilised in diagnostic workflow of breast cancer and is the type of samples that is most commonly available and readily accessible in clinical practice (134, 135). Therefore, to ensure the seamless integration of the PROSPER test into routine clinical workflows, FFPE tissues was selected as the primary sample type for conducing PROSPER testing. To determine the clinical utility of the assay for the assessment of the PROSPER index, optimisation, and validation of the assay within FFPE samples was critical. This was conducted in **Chapter 4**.. Typically, utilising FFPE in genomic assays is difficult due to the degraded and limited yield of nucleic acids extracted from FFPE blocks. During the fixation process, the nucleic acids within the samples undergo chemical modifications, entrapment, and fragmentation due to extensive cross-linking with protein, potentially reducing amplification efficiency, and paraffin in extracts can inhibit PCR amplification reactions (137, 138). Additionally, long storage time can further compromise the quality of the nucleic acid within FFPE blocks. Despite these limitations, advances in PCR techniques over the past decades have improved the potential for quantifying nucleic acids in FFPE tissues. These advanced PCR methods offer sensitivity, precision, and the capability to analyse multiple targets in a single reaction, thus making them suitable for broader application in clinical research, and routine diagnostic and prognostic assays (139). **Chapter 4** of this thesis demonstrated successful performance of the PROSPER multiplexed qPCR assay on nucleic acids extracted from FFPE tissues using the Quick Kit which employs a column-based nucleic acid extraction.

The PROSPER signature was successfully integrated into a qPCR assay using ***Plex****Zyme* technology. The assay was optimised and validated to perform well in FFPE samples in the ***In****Signia* method which required TNA samples for the qPCR reaction. The ***In****Signia* method was developed by industry partner SpeeDx and requires co-amplification of both DNA and RNA. The method allows for the detection of three nucleic acid species within a sample, (i) a gene, (ii) transcripts expressed from that gene and iii) a non-transcribed DNA sequence (NED). A strength of this approach is that, by estimating both DNA and RNA levels for test targets, a non-transcribed DNA region can act as a robust normalisation control. This allows for a single-sample estimate of the test signature without the need for external standards. This method eliminates the need for control for DNA contamination in RNA samples, since both nucleic acids form part of the method. However, in this study, both an ***In****Signia* method (TNA-based assay) and an RNA-based multiplexed assay were developed to determine the most reliable approach for accurate PROSPER signature gene expression analysis in a multiplexed setting. The results demonstrated that the PROSPER index obtained using both normalizing methods was similar, indicating a potential for the ***In****Signia* method to be employed in assessing the PROSPER signature. Adopting the innovative ***In****Signia* method will allow for streamlined protocols and assay, resulting in reduced sample processing time and increased throughput. Streamlined protocols will lead to more consistent and reliable results and furthermore, the insignia assay will require fewer reagents and consumable, lowering the overall cost of testing, making the PROSPER test beneficial for routine application. However, it is important to acknowledge that further validation of both normalizing methods on a larger and more recent sample size cohort is necessary to strengthen this finding.

The primary aim of **Chapter 4** was to validate the PROSPER index obtained from the TNA and RNA-based multiplexed assays against established markers such as the EndoPredict EPclin score,

the PROSPER index measured on the NanoString platform, and proliferation marker Ki-67. The PROSPER index obtained from the TNA-based assay did not correlate well with any of these established markers, whereas a significant correlation was seen between the RNA-based PCR assay and all other assays. However, this poor correlation was likely due to the difference in the sample age at the times when TNA and RNA samples were extracted for validation of the two PROSPER qPCR assays. The PROSPER RNA PCR assay was validated using the residual RNA that had been made freshly from new FFPE blocks, for measuring both EndoPredict and the PROSPER NanoString assay at the time of patient recruitment. In contrast, for the validation of the ***In****Signia* TNA-based assay, the sample preparation was performed after the FFPE blocks were stored for 5 years. The age of the blocks, coupled with the fact that the assay was not validated in identical samples guaranteed that the level of correlation would be lower than seen with RNA. Given that a strong correlation was seen between the NanoString and PCR-based PROSPER RNA assays, and the RNA and TNA samples correlated well when tested in matched samples, it is reasonable to conclude that the TNA assay will likely perform considerably better in freshly prepared FFPE tissue sample. To robustly demonstrate the clinical utility of the PROSPER test, prospective recruitment of a new cohort will be required. A potential approach could involve implementing the PROSPER test in a clinical trial through the Westmead Breast Cancer Institute. ER+ patients can be invited to participate in this study where they would be offered the EndoPredict test at no cost, alongside consenting to the measurement of PROSPER test. This approach will allow for the true comparison of the PROSPER and EndoPredict results. The Multidisciplinary team can then review the outcomes, offering critical feedback on the delivery and the utility of the PROSPER test. Additionally this approach will allow to assess the feasibility of the clinical implantation and evaluate potential economic impact on the NSW pathology. Despite the limitations in this research, the findings of this study suggest that the multiplexed

qPCR assay can serve as a valuable routine clinical tool for assessing both proliferation and clinical risk stratification in FFPE samples, providing clinicians with a comprehensive and reliable assessment of ER+ breast patient prognosis.

An increasingly popular diagnostic method is liquid biopsy. This technique uses genomic and proteomic analysis to examine circulating tumours, circulating tumour DNA, methylation signatures, and microRNAs in bodily fluids. Liquid biopsy offers a less invasive option, ease of access and reproducibility compared to traditional tissue biopsies. It makes conducting repeat testing and monitoring disease progression easier through real-time surveillance (209-211). Recently, the application of liquid biopsy in clinical settings has been increasingly recognised for its benefits in advanced-stage non-small cell lung cancer and metastatic diseases, especially for patients who cannot undergo conventional tissue biopsies due to individual conditions or the location of the tumour. While traditional tissue procurement methos are essential for diagnosis and treatment planning, they are often highly invasive and come with risks, which can be especially problematic for patients with poor health status. Therefore, advanced molecular characterisation technologies and the application of machine learning models have led to extensive research on the effectiveness of liquid in identifying diagnostic, prognostic, and predictive biomarkers for early-stage NSCLC (209, 212).

However, despite these advantages of liquid biopsy, the widespread adoption of liquid biopsy in clinical practice faces several challenges. One major limitation is sensitivity, particularly in early-stage, non-metastatic cancers where tumour-derived materials in the bloodstream are less abundant. Furthermore, the lack of standardization in existing methods and the absence of high-throughput techniques that can be easily integrated into routine clinical settings limits the broader implementation of liquid biopsy (213). While liquid biopsy holds great potential, these limitations

must be addressed before it can become a standard tool in the clinical management of lung and prostate cancer.

In the context of breast cancer, while biopsies are a standard practice and essential for staging cancer, the same tissue sample obtained during the biopsy can be used to test the PROSPER gene signature, eliminating the need for additional invasive procedures. This dual use of the biopsy sample streamlines the diagnostic process, reducing patient discomfort and the risks associated with multiple biopsies. Additionally, as a qPCR-based test, PROSPER provides a straightforward and efficient analysis, making it highly suitable for routine clinical implementation. Unlike liquid biopsies, which face challenges in standardization and sensitivity, the PROSPER test can be seamlessly integrated into existing clinical workflows, offering a reliable and practical solution for breast cancer management.

The main goal of the development of the PROSPER multiplexed assay is its translation to clinical application. Therefore, the next step of this research will be to replicate these findings in a larger prospective cohort of ER+ breast cancer patients from which freshly prepared FFPE tissues can be tested. Once the performance of the multiplexed assay is validated, then streamlining the assay protocols and incorporating automation of sample processing will further enhance the assay to be implemented for routine clinical assay. Ultimately the PROSPER multiplexed assay developed in this research allows for an alternative and affordable assessment of tumour proliferation and clinical risk stratification, to accurately guide treatment selection for ER+ breast cancer patients.

The development and application of a variety of omics techniques have greatly enhanced breast cancer research. This is particularly true for ER+ breast cancers, where the emergence of multigene prognostic tests to guide appropriate treatment selection has improved risk stratification and tailored treatment pathways. In parallel, efforts have been made to better understand the biology

of TNBC through interrogation of the gene expression landscape (4). Even though TNBC only represents 10-20% of all breast cancers, TNBC patients have a higher rate of distant recurrence and poorer prognosis than patients with other breast cancer subtypes (4). Furthermore, there is still a lack of durable targeted therapy for TNBC, and chemotherapy remains the primary systemic treatment of the disease (214). Therefore, a test that could predict the risk of recurrence and stratify TNBC cases according to benefit from specific treatments could substantially improve outcomes for TNBC patients.

A number of subtyping systems for TNBC have been investigated, potentially identifying targetable markers to guide treatment for TNBC (68, 71, 163). However, many clinical trials failed to demonstrate a significant improvement in patient outcomes by targeting these pathways (215). Moreover, while these classification tools have stratified cases into biologically different groups, they have not been able to separate them into different clinical risk categories. The distinct biology of these molecular subtypes suggests that the different genomic profiles within TNBC may be individually exploited to predict disease progression. Therefore, in **Chapter 5**, TNBC subtype-specific prognostic gene signatures were discovered using *in silico* methodologies. It was hypothesised that subtype-specific prognostication, based on gene signatures derived after stratification of TNBC cases into specific subtypes, would offer a more accurate and clinically useful prediction of patient outcomes compared to traditional non-subtype-specific prognostication.

A sequential classification tool developed from our lab, was employed to subtype TNBC cases, and subtype-specific prognostic gene signatures were identified using elastic net regularisation modelling (202) in these subtyped cases. The results from this chapter demonstrated variable performance in predicting DFS in TNBC patients. In the test cohort, the TNBC subtype-specific

prognostic gene signatures allowed for a more precise DFS outcome prediction compared to a non-subtyped specific prognostic gene signature. However, when the subtype-specific gene signatures were validated in an independent cohort, only the MES1 TNBC subtype-specific prognostic gene signature demonstrated a significant association with DFS in TNBC patients. Even though, these findings did not agree with the study hypothesis, these preliminary findings further support the notion that TNBC is a heterogeneous disease and further emphasise the need for in-depth molecular subtyping strategies that can account for that heterogeneity.

In the context of cancer diagnostics, prognostication and treatment, precision medicine, also known as personalized medicine, involves tailoring treatment strategies to the unique genetic and molecular characteristics of each individual patient's tumour. This approach has become increasingly important in breast cancer management, given the inherent heterogeneity of the caner. Advances in high-throughput molecular profiling technologies have led to comprehensive characterisation of the genomic landscape of breast tumours by identifying key drivers and stratifying distinct molecular subtypes. This has resulted in the development of several personalised multigene tests for guiding the treatment of ER+ breast cancers. However, the high cost and lack of accessibility of these tests remain barriers to their widespread clinical implementation in Australia. Therefore, the prosed assay described in this thesis, PROSPER multigene test for ER breast cancer which assess a smaller panel of genes, provides a more cost-effective and practical alternative for risk stratification and treatment selection. Preliminary studies from our lab have demonstrated that PROSPER gene signature can accurately predict the risk of recurrence in ER+ breast cancer and guide treatment decisions, such as the need for adjuvant chemotherapy. Additionally, the optimisation and validation of the PROSPER gene signature into a multiplex qPCR format showcased the potential for broader and affordable clinical adoption of

this precision medicine tool in Australia. In summary, the work described in this thesis has made major contributions to the discovery and development of an affordable and routinely accessible precision medicine for breast cancer. The ability to comprehensively profile the molecular landscape of breast tumours has enabled clinicians to identify key genetic drivers and tailor therapies accordingly. Specifically, the multiplexed qPCR assay developed in this research can be used as an affordable multigene test that can accurately predict the risk of recurrence in ER+ breast cancer, thereby guiding appropriate treatment recommendations for these patients. Furthermore, the subtype-specific prognostication explored in this research emphasises and provides a foundation for the potential application of gene signatures for predicting outcomes and consequently tailoring personalised treatment methods in TNBC.

Chapter 7

References

1. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100(1):57-70. DOI: 10.1016/s0092-8674(00)81683-9
2. Hassanpour SH, Dehghani M. Review of cancer from perspective of molecular. Journal of Cancer Research and Practice. 2017;4(4):127-9. DOI: https://doi.org/10.1016/j.jcrpr.2017.07.001
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin. 2023;73(1):17-48. DOI: 10.3322/caac.21763
4. Nolan E, Lindeman GJ, Visvader JE. Deciphering breast cancer: from biology to the clinic. Cell. 2023;186(8):1708-28. DOI: https://doi.org/10.1016/j.cell.2023.01.040
5. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2021;71(3):209-49.
6. Ferlay J. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2018). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. 2020.
7. Porter P. ``Westernizing"Women's Risks? Breast Cancer in Lower-Income Countries. New England Journal of Medicine. 2008;358(3):213.
8. Hankinson SE, Colditz GA, Willett WC. Towards an integrated model for breast cancer etiology: the lifelong interplay of genes, lifestyle, and hormones. Breast Cancer Research. 2004;6:1-6.
9. Saha T, Lukong KE. Breast cancer stem-like cells in drug resistance: a review of mechanisms and novel therapeutic strategies to overcome drug resistance. Frontiers in oncology. 2022;12:856974.
10. Fu NY, Nolan E, Lindeman GJ, Visvader JE. Stem Cells and the Differentiation Hierarchy in Mammary Gland Development. Physiol Rev. 2020;100(2):489-523. DOI: 10.1152/physrev.00040.2018

11. Gotzsche PC, Jorgensen KJ. Screening for breast cancer with mammography. Cochrane Database Syst Rev. 2013;2013(6):CD001877. DOI: 10.1002/14651858.CD001877.pub5

12. Morrow M, Waters J, Morris E. MRI for breast cancer screening, diagnosis, and treatment. The Lancet. 2011;378(9805):1804-11.

13. Akram M, Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. Biol Res. 2017;50(1):33. DOI: 10.1186/s40659-017-0140-9

14. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19(5):403-10. DOI: 10.1111/j.1365-2559.1991.tb00229.x

15. van Dooijeweert C, van Diest PJ, Ellis IO. Grading of invasive breast carcinoma: the way forward. Virchows Arch. 2022;480(1):33-43. DOI: 10.1007/s00428-021-03141-2

16. Meyer JS, Alvarez C, Milikowski C, Olson N, Russo I, Russo J, et al. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: reproducibility of grade and advantages of proliferation index. Mod Pathol. 2005;18(8):1067-78. DOI: 10.1038/modpathol.3800388

17. Tawfik O, Kimler BF, Davis M, Stasik C, Lai SM, Mayo MS, et al. Grading invasive ductal carcinoma of the breast: advantages of using automated proliferation index instead of mitotic count. Virchows Arch. 2007;450(6):627-36. DOI: 10.1007/s00428-007-0400-0

18. Ginter PS, Idress R, D'Alfonso TM, Fineberg S, Jaffer S, Sattar AK, et al. Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy. Mod Pathol. 2021;34(4):701-9. DOI: 10.1038/s41379-020-00698-2

19. Le Doussal V, Tubiana-Hulin M, Friedman S, Hacene K, Spyratos F, Brunet M. Prognostic value of histologic grade nuclear components of Scarff-Bloom-Richardson (SBR). An improved score modification based on a multivariate analysis of 1262 invasive ductal breast carcinomas. Cancer. 1989;64(9):1914-21. DOI: 10.1002/1097-0142(19891101)64:9<1914::aid-cncr2820640926>3.0.co;2-g

20. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nature Reviews Disease Primers. 2019;5(1):66. DOI: 10.1038/s41572-019-0111-2

21. Dowsett M, Nielsen T, A'Hern R, Bartlett J, Coombes R, Cuzick J, et al. International Ki-67 in Breast Cancer Working Group Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. J Natl Cancer Inst. 2011;103(22):1656-64.

22. Dowsett M, Nielsen TO, A'Hern R, Bartlett J, Coombes RC, Cuzick J, et al. Assessment of Ki67 in breast cancer: recommendations from the International Ki67 in Breast Cancer working group. J Natl Cancer Inst. 2011;103(22):1656-64. DOI: 10.1093/jnci/djr393

23. Makki J. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. Clin Med Insights Pathol. 2015;8:23-31. DOI: 10.4137/CPath.S31563

24. Rakha EA, Tse GM, Quinn CM. An update on the pathological classification of breast cancer. Histopathology. 2023;82(1):5-16. DOI: 10.1111/his.14786

25. Bonotto M, Gerratana L, Poletto E, Driol P, Giangreco M, Russo S, et al. Measures of outcome in metastatic breast cancer: insights from a real-world scenario. Oncologist. 2014;19(6):608-15. DOI: 10.1634/theoncologist.2014-0002

26. Koh J, Kim MJ. Introduction of a New Staging System of Breast Cancer for Radiologists: An Emphasis on the Prognostic Stage. Korean J Radiol. 2019;20(1):69-82. DOI: 10.3348/kjr.2018.0231

27. Hortobagyi GN, Edge SB, Giuliano A. New and Important Changes in the TNM Staging System for Breast Cancer. Am Soc Clin Oncol Educ Book. 2018;38:457-67. DOI: 10.1200/EDBK_201313

28. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.

29. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001;98(19):10869-74.

30. Kittaneh M, Montero AJ, Glück S. Molecular profiling for breast cancer: a comprehensive review. Biomarkers in cancer. 2013;5:BIC. S9455.

31. Prat A, Cheang MC, Martin M, Parker JS, Carrasco E, Caballero R, et al. Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. J Clin Oncol. 2013;31(2):203-9. DOI: 10.1200/JCO.2012.43.4134

32. Rouzier R, Perou CM, Symmans WF, Ibrahim N, Cristofanilli M, Anderson K, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clinical cancer research. 2005;11(16):5678-85.

33. Hortobagyi GN. Treatment of Breast Cancer. New England Journal of Medicine. 1998;339(14):974-84. DOI: 10.1056/nejm199810013391407

34. Lee KK, Chng WJ, Jha S. Prognostic Biomarkers for Breast Cancer Metastasis. Cancer Metastasis: IntechOpen; 2018.

35. Weigelt B, Peterse JL, Van't Veer LJ. Breast cancer metastasis: markers and models. Nature reviews cancer. 2005;5(8):591-602.

36. Soliman NA, Yussif SM. Ki-67 as a prognostic marker according to breast cancer molecular subtype. Cancer biology & medicine. 2016;13(4):496.

37. Kayl AE, Meyers CA. Side-effects of chemotherapy and quality of life in ovarian and breast cancer patients. Curr Opin Obstet Gynecol. 2006;18(1):24-8. DOI: 10.1097/01.gco.0000192996.20040.24

38. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351(27):2817-26. DOI: 10.1056/NEJMoa041588

39. Sheri A, Smith IE, Johnston SR, A'Hern R, Nerurkar A, Jones RL, et al. Residual proliferative cancer burden to predict long-term outcome following neoadjuvant chemotherapy. Ann Oncol. 2015;26(1):75-80. DOI: 10.1093/annonc/mdu508

40. Sun L, Wu A, Bean GR, Hagemann IS, Lin C-Y. Molecular Testing in Breast Cancer: Current Status and Future Directions. The Journal of Molecular Diagnostics. 2021;23(11):1422-32. DOI: https://doi.org/10.1016/j.jmoldx.2021.07.026

41. Paoletti C, Hayes DF. Molecular Testing in Breast Cancer. Annual Review of Medicine. 2014;65(Volume 65, 2014):95-110. DOI: https://doi.org/10.1146/annurev-med-070912-143853

42. Prat A, Ellis MJ, Perou CM. Practical implications of gene-expression-based assays for breast oncologists. Nat Rev Clin Oncol. 2011;9(1):48-57. DOI: 10.1038/nrclinonc.2011.178

43. Gyorffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. Multigene prognostic tests in breast cancer: past, present, future. Breast Cancer Res. 2015;17(1):11. DOI: 10.1186/s13058-015-0514-2

44. Vieira AF, Schmitt F. An Update on Breast Cancer Multigene Prognostic Tests-Emergent Clinical Biomarkers. Frontiers in medicine. 2018;5:248-. DOI: 10.3389/fmed.2018.00248

45. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol. 2006;24(23):3726-34. DOI: 10.1200/JCO.2005.04.7985

46. Goldstein LJ, Gray R, Badve S, Childs BH, Yoshizawa C, Rowley S, et al. Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features. J Clin Oncol. 2008;26(25):4063-71. DOI: 10.1200/JCO.2007.14.4501

47. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. N Engl J Med. 2018;379(2):111-21. DOI: 10.1056/NEJMoa1804710

48. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. N Engl J Med. 2015;373(21):2005-14. DOI: 10.1056/NEJMoa1510764

49. Sparano JA. TAILORx: Trial Assigning Individualized Options for Treatment (Rx). Clinical Breast Cancer. 2006;7(4):347-50. DOI: https://doi.org/10.3816/CBC.2006.n.051

50. Kalinsky K, Barlow WE, Gralow JR, Meric-Bernstam F, Albain KS, Hayes DF, et al. 21-Gene Assay to Inform Chemotherapy Benefit in Node-Positive Breast Cancer. N Engl J Med. 2021;385(25):2336-47. DOI: 10.1056/NEJMoa2108873

51. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res. 2011;17(18):6012-20. DOI: 10.1158/1078-0432.Ccr-11-0926

52. Martin M, Brase JC, Calvo L, Krappmann K, Ruiz-Borrego M, Fisch K, et al. Clinical validation of the EndoPredict test in node-positive, chemotherapy-treated ER+/HER2- breast cancer patients: results from the GEICAM 9906 trial. Breast cancer research : BCR. 2014;16(2):R38-R. DOI: 10.1186/bcr3642

53. Zhang X. Molecular Classification of Breast Cancer: Relevance and Challenges. Archives of Pathology & Laboratory Medicine. 2022;147(1):46-51. DOI: 10.5858/arpa.2022-0070-RA

54. Sestak I, Filipits M, Buus R, Rudas M, Balic M, Knauer M, et al. Prognostic value of endopredict in women with hormone receptor–positive, her2-negative invasive lobular breast cancer. Clinical cancer research. 2020;26(17):4682-7.

55. Azim HA, Jr., Michiels S, Zagouri F, Delaloge S, Filipits M, Namer M, et al. Utility of prognostic genomic tests in breast cancer practice: The IMPAKT 2012 Working Group Consensus Statement. Ann Oncol. 2013;24(3):647-54. DOI: 10.1093/annonc/mds645

56. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC medical genomics. 2015;8(1):54.
57. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160-7. DOI: 10.1200/JCO.2008.18.1370
58. Dowsett M, Cuzick J, Wale C, Forbes J, Mallon EA, Salter J, et al. Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a TransATAC study. J Clin Oncol. 2010;28(11):1829-34. DOI: 10.1200/JCO.2009.24.4798
59. Gnant M, Filipits M, Greil R, Stoeger H, Rudas M, Bago-Horvath Z, et al. Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. Ann Oncol. 2014;25(2):339-45. DOI: 10.1093/annonc/mdt494
60. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology. 2009;27(8):1160.
61. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. Clin Cancer Res. 2010;16(21):5222-32. DOI: 10.1158/1078-0432.CCR-10-1282
62. Filipits M, Nielsen TO, Rudas M, Greil R, Stoger H, Jakesz R, et al. The PAM50 risk-of-recurrence score predicts risk for late distant recurrence after endocrine therapy in postmenopausal women with endocrine-responsive early breast cancer. Clin Cancer Res. 2014;20(5):1298-305. DOI: 10.1158/1078-0432.CCR-13-1845
63. Dowsett M, Sestak I, Lopez-Knowles E, Sidhu K, Dunbier AK, Cowens JW, et al. Comparison of PAM50 risk of recurrence score with onco type DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. Journal of Clinical Oncology. 2013;31(22):2783-90.
64. Stein RC, Makris A, MacPherson IR, Hughes-Davies L, Marshall A, Dotchin G, et al. Optima: Optimal personalised treatment of early breast cancer using multi-parameter analysis, an

international randomized trial of tumor gene expression test-directed chemotherapy treatment in a largely node-positive population. Wolters Kluwer Health; 2021.

65. Collignon J, Lousberg L, Schroeder H, Jerusalem G. Triple-negative breast cancer: treatment challenges and solutions. Breast Cancer: Targets and Therapy. 2016:93-107.

66. Li Y, Zhang H, Merkher Y, Chen L, Liu N, Leonov S, et al. Recent advances in therapeutic strategies for triple-negative breast cancer. Journal of Hematology & Oncology. 2022;15(1):121. DOI: 10.1186/s13045-022-01341-0

67. Lee JS, Yost SE, Yuan Y. Neoadjuvant treatment for triple negative breast cancer: recent progresses and challenges. Cancers. 2020;12(6):1404.

68. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011;121(7):2750-67. DOI: 10.1172/JCI45014

69. Lehmann BD, Jovanovic B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PLoS One. 2016;11(6):e0157368. DOI: 10.1371/journal.pone.0157368

70. Zhao S, Zuo W-J, Shao Z-M, Jiang Y-Z. Molecular subtypes and precision treatment of triple-negative breast cancer. Annals of Translational Medicine. 2020;8(7):499.

71. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua SA, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin Cancer Res. 2015;21(7):1688-98. DOI: 10.1158/1078-0432.CCR-14-0432

72. Yin L, Duan JJ, Bian XW, Yu SC. Triple-negative breast cancer molecular subtyping and treatment progress. Breast Cancer Res. 2020;22(1):61. DOI: 10.1186/s13058-020-01296-5

73. Jézéquel P, Loussouarn D, Guérin-Charbonnel C, Campion L, Vanier A, Gouraud W, et al. Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. Breast Cancer Research. 2015;17:1-16.

74. Jézéquel P, Kerdraon O, Hondermarck H, Guérin-Charbonnel C, Lasla H, Gouraud W, et al. Identification of three subtypes of triple-negative breast cancer with potential therapeutic implications. Breast Cancer Research. 2019;21:1-14.

75. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011;7(10):e1002240. DOI: 10.1371/journal.pcbi.1002240

76. Dinh P, Graham JD, Elder EN, Kabir M, Doan TB, French J, et al. Impact of the EndoPredict genomic assay on treatment decisions for oestrogen receptor-positive early breast cancer patients: benefits of physician selective testing. Breast Cancer Research and Treatment. 2022;191(3):501-11. DOI: 10.1007/s10549-021-06456-5

77. Kulkarni MM. Digital multiplexed gene expression analysis using the NanoString nCounter system. Curr Protoc Mol Biol. 2011;Chapter 25:Unit25B.10. DOI: 10.1002/0471142727.mb25b10s94

78. Tsang HF, Xue VW, Koh SP, Chiu YM, Ng LP, Wong SC. NanoString, a novel digital color-coded barcode technology: current and future applications in molecular diagnostics. Expert Rev Mol Diagn. 2017;17(1):95-103. DOI: 10.1080/14737159.2017.1268533

79. Eastel JM, Lam KW, Lee NL, Lok WY, Tsang AHF, Pei XM, et al. Application of NanoString technologies in companion diagnostic development. Expert Rev Mol Diagn. 2019;19(7):591-8. DOI: 10.1080/14737159.2019.1623672

80. Reis PP, Waldron L, Goswami RS, Xu W, Xuan Y, Perez-Ordonez B, et al. mRNA transcript quantification in archival samples using multiplexed, color-coded probes. BMC Biotechnol. 2011;11:46. DOI: 10.1186/1472-6750-11-46

81. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol. 2008;26(3):317-25. DOI: 10.1038/nbt1385

82. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Medical Genomics. 2015;8(1):54. DOI: 10.1186/s12920-015-0129-6

83. Caballero-Solares A, Hall JR, Xue X, Rise ML. Reverse Transcription-Quantitative Real-Time Polymerase Chain Reaction (RT-qPCR) for Gene Expression Analyses. In: Christian SL, editor. Cancer Cell Biology: Methods and Protocols. New York, NY: Springer US; 2022. p. 319-40.

84. Klein D. Quantification using real-time PCR technology: applications and limitations. Trends in Molecular Medicine. 2002;8(6):257-60. DOI: https://doi.org/10.1016/S1471-4914(02)02355-9

85. Williams M. Real-time polymerase chain reaction. In: Bustin SA, editor. The PCR Revolution: Basic Technologies and Applications. Cambridge: Cambridge University Press; 2009. p. 3-11.

86. Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonák J, Lind K, et al. The real-time polymerase chain reaction. Molecular Aspects of Medicine. 2006;27(2):95-125. DOI: https://doi.org/10.1016/j.mam.2005.12.007

87. Delidow BC, Lynch JP, Peluso JJ, White BA. Polymerase chain reaction : basic protocols. Methods Mol Biol. 1993;15:1-29. DOI: 10.1385/0-89603-244-2:1

88. Kubista M, Stalberg A, Bar T, editors. Light-up-probe-based real-time Q-PCR. Genomics and proteomics technologies; 2001: SPIE.

89. Jia Y. Chapter 3 - Real-Time PCR. In: Conn PM, editor. Methods in Cell Biology. 112: Academic Press; 2012. p. 55-68.

90. Navarro E, Serrano-Heras G, Castaño M, Solera J. Real-time PCR detection chemistry. Clinica chimica acta. 2015;439:231-50.

91. Nolan T, Hands RE, Bustin SA. Quantification of mRNA using real-time RT-PCR. Nat Protoc. 2006;1(3):1559-82. DOI: 10.1038/nprot.2006.236

92. Kuang J, Yan X, Genders AJ, Granata C, Bishop DJ. An overview of technical considerations when using quantitative real-time PCR analysis of gene expression in human exercise research. PLoS One. 2018;13(5):e0196438. DOI: 10.1371/journal.pone.0196438

93. Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol. 2000;25(2):169-93. DOI: 10.1677/jme.0.0250169

94. Mokany E, Todd AV. MNAzyme qPCR: A Superior Tool for Multiplex qPCR. In: Kolpashchikov DM, Gerasimova YV, editors. Nucleic Acid Detection: Methods and Protocols. Totowa, NJ: Humana Press; 2013. p. 31-49.

95. Mokany E, Bone SM, Young PE, Doan TB, Todd AV. MNAzymes, a Versatile New Class of Nucleic Acid Enzymes That Can Function as Biosensors and Molecular Switches. Journal of the American Chemical Society. 2010;132(3):1051-9. DOI: 10.1021/ja9076777

96. Mokany E, Tan YL, Bone SM, Fuery CJ, Todd AV. MNAzyme qPCR with superior multiplexing capacity. Clin Chem. 2013;59(2):419-26. DOI: 10.1373/clinchem.2012.192930

97. Gerasimova YV, Yakovchuk P, Dedkova LM, Hecht SM, Kolpashchikov DM. Expedited quantification of mutant ribosomal RNA by binary deoxyribozyme (BiDz) sensors. Rna. 2015;21(10):1834-43.

98. Tan LY, Walker SM, Lonergan T, Lima NE, Todd AV, Mokany E. Superior multiplexing capacity of PlexPrimers enables sensitive and specific detection of SNPs and clustered mutations in qPCR. PLoS One. 2017;12(1):e0170087.

99. Tabrizi SN, Tan LY, Walker S, Twin J, Poljak M, Bradshaw CS, et al. Multiplex assay for simultaneous detection of Mycoplasma genitalium and macrolide resistance using PlexZyme and PlexPrime technology. PLoS One. 2016;11(6):e0156740.

100. Silver N, Cotroneo E, Proctor G, Osailan S, Paterson KL, Carpenter GH. Selection of housekeeping genes for gene expression studies in the adult rat submandibular gland under normal, inflamed, atrophic and regenerative states. BMC Mol Biol. 2008;9:64. DOI: 10.1186/1471-2199-9-64

101. Suzuki T, Higgins PJ, Crawford DR. Control selection for RNA quantitation. Biotechniques. 2000;29(2):332-7. DOI: 10.2144/00292rv02

102. Keydar I, Chen L, Karby S, Weiss FR, Delarea J, Radu M, et al. Establishment and characterization of a cell line of human breast carcinoma origin. Eur J Cancer (1965). 1979;15(5):659-70. DOI: 10.1016/0014-2964(79)90139-7

103. Boyages J, Taylor R, Chua B, Ung O, Bilous M, Salisbury E, et al. A risk index for early node-negative breast cancer. British Journal of Surgery. 2006;93(5):564-71. DOI: 10.1002/bjs.5207

104. Pathmanathan N, Balleine RL, Jayasinghe UW, Bilinski KL, Provan PJ, Byth K, et al. The prognostic value of Ki67 in systemically untreated patients with node-negative breast cancer. Journal of Clinical Pathology. 2014;67(3):222. DOI: 10.1136/jclinpath-2013-201793

105. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001;98(19):10869-74. DOI: doi:10.1073/pnas.191367098

106. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009;27(8):1160-7. DOI: 10.1200/jco.2008.18.1370

107. Howell A, Cuzick J, Baum M, Buzdar A, Dowsett M, Forbes JF, et al. Results of the ATAC (Arimidex, Tamoxifen, Alone or in Combination) trial after completion of 5 years' adjuvant treatment for breast cancer. Lancet. 2005;365(9453):60-2.

108. Paoletti C, Hayes DF. Molecular testing in breast cancer. Annu Rev Med. 2014;65:95-110. DOI: 10.1146/annurev-med-070912-143853

109. Oyama T, Ishikawa Y, Hayashi M, Arihiro K, Horiguchi J. The effects of fixation, processing and evaluation criteria on immunohistochemical detection of hormone receptors in breast cancer. Breast cancer. 2007;14(2):182-8.

110. Leong TY-M, Leong AS-Y. Controversies in the assessment of HER-2: more questions than answers. Advances in anatomic pathology. 2006;13(5):263-9.

111. Cheang MCU, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 Index, HER2 Status, and Prognosis of Patients With Luminal B Breast Cancer. Journal of the National Cancer Institute. 2009;101(10):736-50. DOI: 10.1093/jnci/djp082

112. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics. 2015;8:54. DOI: 10.1186/s12920-015-0129-6

113. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clin Cancer Res. 2011;17(18):6012-20. DOI: 10.1158/1078-0432.CCR-11-0926

114. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346-52. DOI: 10.1038/nature10983

115. Mokany E, Bone SM, Young PE, Doan TB, Todd AV. MNAzymes, a versatile new class of nucleic acid enzymes that can function as biosensors and molecular switches. J Am Chem Soc. 2010;132(3):1051-9. DOI: 10.1021/ja9076777

116. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134. DOI: 10.1186/1471-2105-13-134

117. Kalendar R, Khassenov B, Ramankulov Y, Samuilova O, Ivanov KI. FastPCR: An in silico tool for fast primer and probe design and advanced sequence analysis. Genomics. 2017;109(3-4):312-9. DOI: 10.1016/j.ygeno.2017.05.005

118. Koga M, Musgrove EA, Sutherland RL. Modulation of the Growth-inhibitory Effects of Progestins and the Antiestrogen Hydroxyclomiphene on Human Breast Cancer Cells by Epidermal Growth Factor and Insulin1. Cancer Res. 1989;49(1):112-6.

119. Bustin S, Huggett J. qPCR primer design revisited. Biomol Detect Quantif. 2017;14:19-28. DOI: 10.1016/j.bdq.2017.11.001

120. Delghandi M, Delghandi MP, Goddard S. The Significance of PCR Primer Design in Genetic Diversity Studies: Exemplified by Recent Research into the Genetic Structure of Marine Species. Methods Mol Biol. 2022;2392:3-15. DOI: 10.1007/978-1-0716-1799-1_1

121. Hays A, Wissel M, Colletti K, Soon R, Azadeh M, Smith J, et al. Recommendations for Method Development and Validation of qPCR and dPCR Assays in Support of Cell and Gene Therapy Drug Development. AAPS J. 2024;26(1):24. DOI: 10.1208/s12248-023-00880-9

122. Kalendar R, Lee D, Schulman AH. FastPCR software for PCR primer and probe design and repeat search. Genes, genomes and genomics. 2009;3(1):1-14.

123. Kalendar R, Lee D, Schulman AH. Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. Genomics. 2011;98(2):137-44. DOI: https://doi.org/10.1016/j.ygeno.2011.04.009

124. Kalendar R. A Guide to Using FASTPCR Software for PCR, In Silico PCR, and Oligonucleotide Analysis. Methods Mol Biol. 2022;2392:223-43. DOI: 10.1007/978-1-0716-1799-1_16

125. Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, et al. IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res. 2008;36(Web Server issue):W163-9. DOI: 10.1093/nar/gkn198

126. Markoulatos P, Samara V, Siafakas N, Plakokefalos E, Spyrou N, Moncany ML. Development of a quadriplex polymerase chain reaction for human cytomegalovirus detection. J Clin Lab Anal. 1999;13(3):99-105. DOI: 10.1002/(sici)1098-2825(1999)13:3<99::aid-jcla2>3.0.co;2-e

127. Markoulatos P, Siafakas N, Moncany M. Multiplex polymerase chain reaction: a practical approach. J Clin Lab Anal. 2002;16(1):47-51. DOI: 10.1002/jcla.2058

128. Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE. Multiplex PCR: optimization and application in diagnostic virology. Clin Microbiol Rev. 2000;13(4):559-70. DOI: 10.1128/CMR.13.4.559

129. Park M, Won J, Choi BY, Lee CJ. Optimization of primer sets and detection protocols for SARS-CoV-2 of coronavirus disease 2019 (COVID-19) using PCR and real-time PCR. Exp Mol Med. 2020;52(6):963-77. DOI: 10.1038/s12276-020-0452-7

130. Musgrove EA, Lee CS, Sutherland RL. Progestins both stimulate and inhibit breast cancer cell cycle progression while increasing expression of transforming growth factor alpha, epidermal

growth factor receptor, c-fos, and c-myc genes. Mol Cell Biol. 1991;11(10):5032-43. DOI: 10.1128/mcb.11.10.5032-5043.1991

131. Mirabelli P, Coppola L, Salvatore M. Cancer Cell Lines Are Useful Model Systems for Medical Research. Cancers (Basel). 2019;11(8). DOI: 10.3390/cancers11081098

132. Gunson RN, Collins TC, Carman WF. Practical experience of high throughput real time PCR in the routine diagnostic virology setting. J Clin Virol. 2006;35(4):355-67. DOI: 10.1016/j.jcv.2005.12.006

133. Waugh C, Clark G. Factors affecting test reproducibility among laboratories. Rev Sci Tech. 2021;40(1):131-43. DOI: 10.20506/rst.40.1.3213

134. van Deventer BS, du Toit-Prinsloo L, van Niekerk C. Practical tips to using formalin-fixed paraffin-embedded tissue archives for molecular diagnostics in a South African setting. Afr J Lab Med. 2022;11(1):1587. DOI: 10.4102/ajlm.v11i1.1587

135. Sarnecka AK, Nawrat D, Piwowar M, Ligeza J, Swadzba J, Wojcik P. DNA extraction from FFPE tissue samples - a comparison of three procedures. Contemp Oncol (Pozn). 2019;23(1):52-8. DOI: 10.5114/wo.2019.83875

136. Gaffney EF, Riegman PH, Grizzle WE, Watson PH. Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. Biotech Histochem. 2018;93(5):373-86. DOI: 10.1080/10520295.2018.1446101

137. Gilbert MT, Haselkorn T, Bunce M, Sanchez JJ, Lucas SB, Jewell LD, et al. The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? PLoS One. 2007;2(6):e537. DOI: 10.1371/journal.pone.0000537

138. Lewis F, Maughan NJ, Smith V, Hillan K, Quirke P. Unlocking the archive--gene expression in paraffin-embedded tissue. J Pathol. 2001;195(1):66-71. DOI: 10.1002/1096-9896(200109)195:1<66::AID-PATH921>3.0.CO;2-F

139. Zhang P, Lehmann BD, Shyr Y, Guo Y. The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies. Int J Genomics. 2017;2017:1926304. DOI: 10.1155/2017/1926304

140. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16(3):1215. DOI: 10.1093/nar/16.3.1215

141. Poh JJ, Gan SK. Comparison of customized spin-column and salt-precipitation finger-prick blood DNA extraction. Biosci Rep. 2014;34(5). DOI: 10.1042/BSR20140105

142. Greco M, Saez CA, Brown MT, Bitonti MB. A simple and effective method for high quality co-extraction of genomic DNA and total RNA from low biomass Ectocarpus siliculosus, the model brown alga. PLoS One. 2014;9(5):e96470. DOI: 10.1371/journal.pone.0096470

143. Tan SC, Yiap BC. DNA, RNA, and protein extraction: the past and the present. J Biomed Biotechnol. 2009;2009:574398. DOI: 10.1155/2009/574398

144. Ali N, Rampazzo RCP, Costa ADT, Krieger MA. Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics. Biomed Res Int. 2017;2017:9306564. DOI: 10.1155/2017/9306564

145. Steiert TA, Parra G, Gut M, Arnold N, Trotta JR, Tonda R, et al. A critical spotlight on the paradigms of FFPE-DNA sequencing. Nucleic Acids Res. 2023;51(14):7143-62. DOI: 10.1093/nar/gkad519

146. Oba U, Kohashi K, Sangatsuda Y, Oda Y, Sonoda KH, Ohga S, et al. An efficient procedure for the recovery of DNA from formalin-fixed paraffin-embedded tissue sections. Biol Methods Protoc. 2022;7(1):bpac014. DOI: 10.1093/biomethods/bpac014

147. Frazer Z, Yoo C, Sroya M, Bellora C, DeWitt BL, Sanchez I, et al. Effect of Different Proteinase K Digest Protocols and Deparaffinization Methods on Yield and Integrity of DNA Extracted From Formalin-fixed, Paraffin-embedded Tissue. J Histochem Cytochem. 2020;68(3):171-84. DOI: 10.1369/0022155420906234

148. Janecka A, Adamczyk A, Gasinska A. Comparison of eight commercially available kits for DNA extraction from formalin-fixed paraffin-embedded tissues. Anal Biochem. 2015;476:8-10. DOI: 10.1016/j.ab.2015.01.019

149. Senguven B, Baris E, Oygur T, Berktas M. Comparison of methods for the extraction of DNA from formalin-fixed, paraffin-embedded archival tissues. Int J Med Sci. 2014;11(5):494-9. DOI: 10.7150/ijms.8842

150. Seiler C, Sharpe A, Barrett JC, Harrington EA, Jones EV, Marshall GB. Nucleic acid extraction from formalin-fixed paraffin-embedded cancer cell line samples: a trade off between quantity and quality? BMC Clin Pathol. 2016;16:17. DOI: 10.1186/s12907-016-0039-3

151. Masuda N, Ohnishi T, Kawamoto S, Monden M, Okubo K. Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. Nucleic Acids Res. 1999;27(22):4436-43. DOI: 10.1093/nar/27.22.4436

152. Sanchez I, Betsou F, Culot B, Frasquilho S, McKay SC, Pericleous S, et al. RNA and microRNA Stability in PAXgene-Fixed Paraffin-Embedded Tissue Blocks After Seven Years' Storage. Am J Clin Pathol. 2018;149(6):536-47. DOI: 10.1093/ajcp/aqy026
153. Groelz D, Viertler C, Pabst D, Dettmann N, Zatloukal K. Impact of storage conditions on the quality of nucleic acids in paraffin embedded tissues. PLoS One. 2018;13(9):e0203608. DOI: 10.1371/journal.pone.0203608
154. Yi QQ, Yang R, Shi JF, Zeng NY, Liang DY, Sha S, et al. Effect of preservation time of formalin-fixed paraffin-embedded tissues on extractable DNA and RNA quantity. J Int Med Res. 2020;48(6):300060520931259. DOI: 10.1177/0300060520931259
155. Kong H, Zhu M, Cui F, Wang S, Gao X, Lu S, et al. Quantitative assessment of short amplicons in FFPE-derived long-chain RNA. Sci Rep. 2014;4:7246. DOI: 10.1038/srep07246
156. Abrahamsen HN, Steiniche T, Nexo E, Hamilton-Dutoit SJ, Sorensen BS. Towards quantitative mRNA analysis in paraffin-embedded tissues using real-time reverse transcriptase-polymerase chain reaction: a methodological study on lymph nodes from melanoma patients. J Mol Diagn. 2003;5(1):34-41. DOI: 10.1016/S1525-1578(10)60449-7
157. Davey MG, Hynes SO, Kerin MJ, Miller N, Lowery AJ. Ki-67 as a Prognostic Biomarker in Invasive Breast Cancer. Cancers (Basel). 2021;13(17). DOI: 10.3390/cancers13174455
158. Luporsi E, Andre F, Spyratos F, Martin PM, Jacquemier J, Penault-Llorca F, et al. Ki-67: level of evidence and methodological considerations for its role in the clinical management of breast cancer: analytical and critical review. Breast Cancer Res Treat. 2012;132(3):895-915. DOI: 10.1007/s10549-011-1837-z
159. Pinilla K, Drewett LM, Lucey R, Abraham JE. Precision Breast Cancer Medicine: Early Stage Triple Negative Breast Cancer-A Review of Molecular Characterisation, Therapeutic Targets and Future Trends. Front Oncol. 2022;12:866889. DOI: 10.3389/fonc.2022.866889
160. Penault-Llorca F, Viale G. Pathological and molecular diagnosis of triple-negative breast cancer: a clinical perspective. Ann Oncol. 2012;23 Suppl 6:vi19-22. DOI: 10.1093/annonc/mds190
161. Li Y, Zhang H, Merkher Y, Chen L, Liu N, Leonov S, et al. Recent advances in therapeutic strategies for triple-negative breast cancer. J Hematol Oncol. 2022;15(1):121. DOI: 10.1186/s13045-022-01341-0
162. Lee JS, Yost SE, Yuan Y. Neoadjuvant Treatment for Triple Negative Breast Cancer: Recent Progresses and Challenges. Cancers (Basel). 2020;12(6). DOI: 10.3390/cancers12061404

163. Jezequel P, Loussouarn D, Guerin-Charbonnel C, Campion L, Vanier A, Gouraud W, et al. Gene-expression molecular subtyping of triple-negative breast cancer tumours: importance of immune response. Breast Cancer Res. 2015;17:43. DOI: 10.1186/s13058-015-0550-y

164. Lv Y, Lv D, Lv X, Xing P, Zhang J, Zhang Y. Immune Cell Infiltration-Based Characterization of Triple-Negative Breast Cancer Predicts Prognosis and Chemotherapy Response Markers. Front Genet. 2021;12:616469. DOI: 10.3389/fgene.2021.616469

165. Voorwerk L, Slagter M, Horlings HM, Sikorska K, van de Vijver KK, de Maaker M, et al. Immune induction strategies in metastatic triple-negative breast cancer to enhance the sensitivity to PD-1 blockade: the TONIC trial. Nat Med. 2019;25(6):920-8. DOI: 10.1038/s41591-019-0432-4

166. Lee M, Yoo TK, Chae BJ, Lee A, Cha YJ, Lee J, et al. Luminal androgen receptor subtype and tumor-infiltrating lymphocytes groups based on triple-negative breast cancer molecular subclassification. Sci Rep. 2024;14(1):11278. DOI: 10.1038/s41598-024-61640-z

167. Fu J, Khaybullin R, Zhang Y, Xia A, Qi X. Gene expression profiling leads to discovery of correlation of matrix metalloproteinase 11 and heparanase 2 in breast cancer progression. BMC Cancer. 2015;15:473. DOI: 10.1186/s12885-015-1410-y

168. Komatsu M, Yoshimaru T, Matsuo T, Kiyotani K, Miyoshi Y, Tanahashi T, et al. Molecular features of triple negative breast cancer cells by genome-wide gene expression profiling analysis. Int J Oncol. 2013;42(2):478-506. DOI: 10.3892/ijo.2012.1744

169. Fumagalli D, Blanchet-Cohen A, Brown D, Desmedt C, Gacquer D, Michiels S, et al. Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. BMC Genomics. 2014;15(1):1008. DOI: 10.1186/1471-2164-15-1008

170. den Hollander P, Rawls K, Tsimelzon A, Shepherd J, Mazumdar A, Hill J, et al. Phosphatase PTP4A3 Promotes Triple-Negative Breast Cancer Growth and Predicts Poor Patient Survival. Cancer Res. 2016;76(7):1942-53. DOI: 10.1158/0008-5472.CAN-14-0673

171. Yang L, Wu X, Wang Y, Zhang K, Wu J, Yuan YC, et al. FZD7 has a critical role in cell proliferation in triple negative breast cancer. Oncogene. 2011;30(43):4437-46. DOI: 10.1038/onc.2011.145

172. Zhou YF, Xiao Y, Jin X, Di GH, Jiang YZ, Shao ZM. Integrated analysis reveals prognostic value of HLA-I LOH in triple-negative breast cancer. J Immunother Cancer. 2021;9(10). DOI: 10.1136/jitc-2021-003371

173. Wang L, Shen X, Xie B, Ma Z, Chen X, Cao F. Transcriptional profiling of differentially expressed long non-coding RNAs in breast cancer. Genom Data. 2015;6:214-6. DOI: 10.1016/j.gdata.2015.09.020

174. Rody A, Karn T, Liedtke C, Pusztai L, Ruckhaeberle E, Hanker L, et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res. 2011;13(5):R97. DOI: 10.1186/bcr3035

175. Juul N, Szallasi Z, Eklund AC, Li Q, Burrell RA, Gerlinger M, et al. Assessment of an RNA interference screen-derived mitotic and ceramide pathway metagene as a predictor of response to neoadjuvant paclitaxel for primary triple-negative breast cancer: a retrospective analysis of five clinical trials. Lancet Oncol. 2010;11(4):358-65. DOI: 10.1016/S1470-2045(10)70018-8

176. Tseng LM, Chiu JH, Liu CY, Tsai YF, Wang YL, Yang CW, et al. A comparison of the molecular subtypes of triple-negative breast cancer among non-Asian and Taiwanese women. Breast Cancer Res Treat. 2017;163(2):241-54. DOI: 10.1007/s10549-017-4195-7

177. Santonja A, Sanchez-Munoz A, Lluch A, Chica-Parrado MR, Albanell J, Chacon JI, et al. Triple negative breast cancer subtypes and pathologic complete response rate to neoadjuvant chemotherapy. Oncotarget. 2018;9(41):26406-16. DOI: 10.18632/oncotarget.25413

178. Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, et al. A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Res Treat. 2011;126(2):407-20. DOI: 10.1007/s10549-010-0897-9

179. Maire V, Nemati F, Richardson M, Vincent-Salomon A, Tesson B, Rigaill G, et al. Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. Cancer Res. 2013;73(2):813-23. DOI: 10.1158/0008-5472.CAN-12-2633

180. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010;11(2):242-53. DOI: 10.1093/biostatistics/kxp059

181. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3. DOI: 10.1093/bioinformatics/bts034

182. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2005;67(2):301-20. DOI: 10.1111/j.1467-9868.2005.00503.x

183. Marquardt DW, Snee RD. Ridge Regression in Practice. The American Statistician. 1975;29(1):3-20. DOI: 10.1080/00031305.1975.10479105
184. Ranstam J, Cook JA. LASSO regression. Br J Surg. 2018;105(10):1348-. DOI: 10.1002/bjs.10895
185. Tay JK, Narasimhan B, Hastie T. Elastic Net Regularization Paths for All Generalized Linear Models. J Stat Softw. 2023;106. DOI: 10.18637/jss.v106.i01
186. Vtorushin S, Dulesova A, Krakhmal N. Luminal androgen receptor (LAR) subtype of triple-negative breast cancer: molecular, morphological, and clinical features. J Zhejiang Univ Sci B. 2022;23(8):617-24. DOI: 10.1631/jzus.B2200113
187. Zagami P, Carey LA. Triple negative breast cancer: Pitfalls and progress. NPJ Breast Cancer. 2022;8(1):95. DOI: 10.1038/s41523-022-00468-0
188. Geyer FC, Pareja F, Weigelt B, Rakha E, Ellis IO, Schnitt SJ, et al. The spectrum of triple-negative breast disease: high-and low-grade lesions. The American journal of pathology. 2017;187(10):2139-51.
189. Bianchini G, Balko JM, Mayer IA, Sanders ME, Gianni L. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nat Rev Clin Oncol. 2016;13(11):674-90. DOI: 10.1038/nrclinonc.2016.66
190. Malorni L, Shetty PB, De Angelis C, Hilsenbeck S, Rimawi MF, Elledge R, et al. Clinical and biologic features of triple-negative breast cancers in a large cohort of patients with long-term follow-up. Breast Cancer Res Treat. 2012;136(3):795-804. DOI: 10.1007/s10549-012-2315-y
191. Yu X, Liu Y, Chen M. Reassessment of Reliability and Reproducibility for Triple-Negative Breast Cancer Subtyping. Cancers (Basel). 2022;14(11). DOI: 10.3390/cancers14112571
192. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350-6. DOI: 10.1038/nm.3967
193. Bramsen JB, Rasmussen MH, Ongen H, Mattesen TB, Orntoft MW, Arnadottir SS, et al. Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. Cell Rep. 2017;19(6):1268-80. DOI: 10.1016/j.celrep.2017.04.045
194. Ayton SG, Pavlicova M, Robles-Espinoza CD, Tamez Pena JG, Trevino V. Multiomics subtyping for clinically prognostic cancer subtypes and personalized therapy: A systematic review and meta-analysis. Genet Med. 2022;24(1):15-25. DOI: 10.1016/j.gim.2021.09.006

195. Smolarz B, Nowak AZ, Romanowicz H. Breast Cancer-Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). Cancers (Basel). 2022;14(10). DOI: 10.3390/cancers14102569

196. Orrantia-Borunda E, Anchondo-Nunez P, Acuna-Aguilar LE, Gomez-Valles FO, Ramirez-Valdespino CA. Subtypes of Breast Cancer. In: Mayrovitz HN, editor. Breast cancer. Brisbane (AU)2022.

197. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321(3):288-300. DOI: 10.1001/jama.2018.19323

198. Nassar SF, Raddassi K, Ubhi B, Doktorski J, Abulaban A. Precision Medicine: Steps along the Road to Combat Human Cancer. Cells. 2020;9(9). DOI: 10.3390/cells9092056

199. Bianchini G, De Angelis C, Licata L, Gianni L. Treatment landscape of triple-negative breast cancer - expanded options, evolving needs. Nat Rev Clin Oncol. 2022;19(2):91-113. DOI: 10.1038/s41571-021-00565-2

200. Obidiro O, Battogtokh G, Akala EO. Triple Negative Breast Cancer Treatment Options and Limitations: Future Outlook. Pharmaceutics. 2023;15(7). DOI: 10.3390/pharmaceutics15071796

201. Zhao S, Zuo WJ, Shao ZM, Jiang YZ. Molecular subtypes and precision treatment of triple-negative breast cancer. Ann Transl Med. 2020;8(7):499. DOI: 10.21037/atm.2020.03.194

202. Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Res. 2015;43(12):e79. DOI: 10.1093/nar/gkv229

203. Serdar CC, Cihan M, Yucel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochem Med (Zagreb). 2021;31(1):010502. DOI: 10.11613/BM.2021.010502

204. Sielemann K, Hafner A, Pucker B. The reuse of public datasets in the life sciences: potential risks and rewards. PeerJ. 2020;8:e9954. DOI: 10.7717/peerj.9954

205. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49. DOI: 10.3322/caac.21660

206. Lukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanislawek A. Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. Cancers (Basel). 2021;13(17). DOI: 10.3390/cancers13174287

207. Caballero-Solares A, Hall JR, Xue X, Rise ML. Reverse Transcription-Quantitative Real-Time Polymerase Chain Reaction (RT-qPCR) for Gene Expression Analyses. Methods Mol Biol. 2022;2508:319-40. DOI: 10.1007/978-1-0716-2376-3_21

208. Klein D. Quantification using real-time PCR technology: applications and limitations. Trends Mol Med. 2002;8(6):257-60. DOI: 10.1016/s1471-4914(02)02355-9

209. Lone SN, Nisar S, Masoodi T, Singh M, Rizwan A, Hashem S, et al. Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments. Mol Cancer. 2022;21(1):79. DOI: 10.1186/s12943-022-01543-7

210. Pantel K, Alix-Panabieres C. Circulating tumour cells in cancer patients: challenges and perspectives. Trends Mol Med. 2010;16(9):398-406. DOI: 10.1016/j.molmed.2010.07.001

211. Crowley E, Di Nicolantonio F, Loupakis F, Bardelli A. Liquid biopsy: monitoring cancer-genetics in the blood. Nat Rev Clin Oncol. 2013;10(8):472-84. DOI: 10.1038/nrclinonc.2013.110

212. Ko J, Baldassano SN, Loh PL, Kording K, Litt B, Issadore D. Machine learning to detect signatures of disease in liquid biopsies - a user's guide. Lab Chip. 2018;18(3):395-405. DOI: 10.1039/c7lc00955k

213. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic - implementation issues and future challenges. Nat Rev Clin Oncol. 2021;18(5):297-312. DOI: 10.1038/s41571-020-00457-x

214. Collignon J, Lousberg L, Schroeder H, Jerusalem G. Triple-negative breast cancer: treatment challenges and solutions. Breast Cancer (Dove Med Press). 2016;8:93-107. DOI: 10.2147/BCTT.S69488

215. Ensenyat-Mendez M, Llinas-Arias P, Orozco JIJ, Iniguez-Munoz S, Salomon MP, Sese B, et al. Current Triple-Negative Breast Cancer Subtypes: Dissecting the Most Aggressive Form of Breast Cancer. Front Oncol. 2021;11:681476. DOI: 10.3389/fonc.2021.681476