The role of intra-tumoural heterogeneity in resistance to neoadjuvant chemotherapy in breast cancer

Waleed Said Humaid Al Amri

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Medicine and Health
Wellcome Trust Brenner Building

December, 2018

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated within the thesis. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Acknowledgements

I would like to express my sincere gratitude to my government (Sultanate of Oman) for funding me to do a PhD. My employer Ministry of Health-The Royal Hospital for supporting my application.

I would like to sincerely thank my supervisors: Dr. Thomas Hughes (Primary supervisor), Dr. Lucy Stead, Prof. Andrew Hanby and Dr. Eldo Verghese for their help, guidance and support throughout my project.

In addition, I would like to acknowledge the following people:

Lab Manager:

Mrs. Sarah Perry

Collaborators:

Ms. Stacey Jones

Dr. Sandra Bell

Dr. Abeer Shaaban

Colleagues:

Ms. Diana Baxter

Dr. Laura Wastall

Ms. Melina Teske

Ms. Shivani Shukla

Ms. Anna Whitehead

Ms. Robyn Board

Last, but not the least, my parents for their support, love and prayers.

Abstract

Breast cancer is a heterogeneous disease and accumulating evidence suggests that treatment failure may be driven by intra-tumour heterogeneity (ITH). Utilising the current protocol for neoadjuvant (pre-surgery) chemotherapy (NAC) provides the opportunity to study molecular genetic changes between pre- and post-therapy by assessing pre-therapy biopsies and post-therapy surgical resections.

Whole exome sequencing was performed on matched pre- and post-treatment cancer cells from 6 patients with oestrogen receptor positive breast cancers that showed partial responses to the chemotherapeutic combination epirubicin/cyclophosphamide. Data analysis was performed to determine differences in genetic aberrations between pre- and post-NAC, and in particular to identify evidence of consistent selection by therapy of aberrations that therefore may define chemotherapy resistance or sensitivity.

There were extensive differences in the range of genetic aberrations between pre- and post-NAC. 48 genes were identified for further study based on evidence of mutations conferring a selective advantage or disadvantage during chemotherapeutic response. The relevance of these was screened using siRNA knock-down and assessment of response to epirubicin using cell viability assays *in vitro*. Two genes were taken forward. Potential loss-of-function mutations in MUC17 were selected against during therapy in patients, and in accordance with this MUC17 knock-down was associated with increased sensitivity *in vitro*. Potential loss-of-function mutations in PCNX1 were selected for during therapy in patients, and in accordance with this PCNX1 knock-down was associated with resistance. Further work was performed to investigate mechanisms by which these genes modify chemotherapy response, by examining drug loading and ABC transporter expression levels. Data indicate that both genes impact on drug loading, potentially through modulating ABC transporter expression.

Also, MUC17 or PCNX1 protein levels were tested as prognostic and predictive markers for breast cancer clinical outcomes using tissue taken from cohorts of patients who received adjuvant chemotherapy or neoadjuvant chemotherapy. Kaplan-Meier survival analyses revealed that low MUC17 expression after neoadjuvant chemotherapy was significantly associated with longer disease free survival, which was in agreement with the selection of MUC17 mutations seen after therapy in the initial patient group, and with the *in vitro* siRNA findings concerning drug sensitivity.

I concluded that MUC17 and PCNX1 are potential markers of response to chemotherapy in breast cancer, and that therapeutic modulation of their activities could enhance chemotherapy responses.

Abbreviations

ANNOVA	Analysis of Variance
ATCC	American Type Culture Collection
BCRP	Breast Cancer Resistance Protein
BCS	Breast Conservative Surgery
Bp	Base Pair
CD	Cluster Differentiation
CFA	Colony Forming Assay
Chip-Seq	Chromatin Immuno-Precipitation Sequencing
COSMIC	Catalogue of Somatic Mutations in cancer
CSC	Cancer Stem Cells
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DPBS	Dulbecco’s Phosphate Buffered Saline
DPX	Distyrene, plasticizer, and Xylenx
dsDNA	Double-Stranded Deoxyribonucleic Acid
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
EMT	Epithelial-Mesenchymal Transition
ER	Estrogen Receptor
FACS	Flow Cytometry
FCS	Fetal Calf Serum
FdUMP	Fluorodeoxyuridine Monophosphate
FFPE	Formalin Fixed Embadded Tissue
GATK	Genome Analysis Toolkit
Gb	Gigabase
GBM	Glioblastoma Multiforme
gDNA	Genomic Deoxyribonucleic Acid
H and E	Haematoxylin and Eosin
HCL	Hydrochloric Acid
HER2	Human Epidermal Growth Factor Receptor
IDC	Infiltrating Ductal carcinoma
IF	Immuno-Fluorescence
IGF1R	Insulin-like Growth Factor-1 Receptors

IHC	Immuno-Histo-Chemistry
Indels	Small Insertions Deletions
ITH	Intra-Tumour Heterogeneity
JUN	C-Jun N-terminal kinases
Kb	Kilobase
LCM	Laser Capture Micro-Dissection
MAP3K1	Mitogen-Activated Protein Kinase Kinase Kinase 1
Mb	Megabase
MCF-7	Michigan Cancer Foundation-7
MDR1	Multidrug Resistance Protein 1
MRI	Magnetic Resonance Imaging
MRP1	Multidrug Resistance-Associated Protein 1
MTT	(3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyltetrazolium Bromide)
MUC17	Mucin 17
N	Numbers
NAC	Neoadjuvant Chemotherapy
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NSCLC	Non Small Cell Lung Carcinoma
PCNX1	Pecanx1
PCR	Polymerase Chain Reaction
pCR	Complete Pathological Response
PE	Phycoerythrin
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase
PR	Progesterone Receptors
PTEN	Phosphatase and Tensin homolog
RB	Retinoblastoma
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SIFT	Sorting Intolerant from Tolerant
siRNA	Small/Short Interference- Ribonucleic Acid
SSPS	Statistical Package for the Social Sciences
TBS	Tris-Buffered Saline
TBS-T	Tris-Buffered Saline- TWEEN20
TMA	Tissue MicroArray
TP53	Tumour Protein 53
TS	Thymidylate Synthase

VCF	Variant Calling Files
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
X	Times

Table of Contents

List of Tables

List of Figures

1. Introduction

Breast cancer is the most common cancer in women in the UK and worldwide. More than 50,000 women are newly diagnosed with invasive breast cancer in the UK every year [1]. On the positive side, there have been significant improvements in terms of survival rate for patients with breast cancer. To illustrate this, 5-year survival rates have increased from 53% during 1971-1972 to 87% during 2010-2011 in England and Wales [2]. This improvement in controlling breast cancers can be attributed to many factors, including early detection of the disease through implementation of the national screening programme, better surgical techniques, improvements in adjuvant/neoadjuvant chemotherapy, and development and evolution of targeted hormonal and biological drugs such as Tamoxifen (a partial oestrogen antagonist that inhibits function of the oestrogen receptor) and Trastuzumab (a therapy that inhibits function of the HER2 receptor), thereby decreasing mortality from this disease [3]. Despite these advances in breast cancer management and treatment, around 30% of breast cancer patients at some point will develop treatment resistance and have recurrences [4]. These treatment challenges exist mainly because of lack of precise understanding of cancer biology in the context of the heterogeneous nature of breast cancer, which not only exists between tumours but also within tumours as they progress. Accumulating evidence suggest that treatment failure is driven by intra-tumoural heterogeneity and branched tumour evolution, involving genetically distinct sub-clones [5-9]. If it were possible to define predictable molecular pathways for breast cancer evolution from initiation and the permutations that might happen during progression, particularly in the context of current treatments, the information gained may help the design of better treatment pathways with more favourable outcomes and cures.

1.1. Breast cancer classifications – tools for determining appropriate treatments

Breast cancer is routinely classified into different histopathological categories based on the cells' morphological and architectural features and growth patterns. It is broadly classified into *in situ* carcinoma and invasive (infiltrating) carcinoma. Breast carcinoma *in situ* is further sub-classified as either ductal or lobular; growth patterns and cytological features form the basis to distinguish between the two types [10].

Similarly, invasive breast carcinoma is further classified into different types including infiltrating ductal, invasive lobular, ductal/lobular, mucinous (colloid), tubular, medullary and papillary carcinomas. Of these, infiltrating ductal carcinoma (IDC) is, by far, the most common subtype accounting for 70–80% of all invasive lesions [11].

IDC is further sub-classified into 3 grades based on the levels of morphological differentiation of cellular features such as nuclear pleomorphism, glandular/tubule formation and mitotic activity; well-differentiated (grade 1), moderately differentiated (grade 2) or poorly differentiated (grade 3) [12].

In addition, staging is performed in order to provide information on the prognosis for the individual patient and to guide treatment [13]. The common staging system used is TNM classification system, which divides the tumours into stage 0–4 depending on tumour progression. The factors taken into consideration are the size of the primary tumour (T), spread to loco-regional lymph nodes (N), and distant metastasis (M). Stage 0 is non-invasive cancer, such as ductal carcinoma *in situ* and lobular carcinoma *in situ*. Stage 1–3 breast cancer (without distant metastasis) is considered curable, while stage 4 breast cancer (with distant metastasis), is considered incurable [14]. In general, women with tumours of <1 cm have very good 5-year survival, which has been reported to be as high as 99%. However, patients with 3–5

cm tumours have poorer 5-year survival, reported as 86% [15]. Also, the mean time to distant metastasis was shorter for larger tumours compared to smaller tumours [16].

The histopathological classification has served as the primary form of breast cancer classification, and has helped to guide treatment management and prognosis prediction. However, with recent advances in cancer research and an increased molecular understanding of breast cancer heterogeneity, a new form of molecular classification is also used clinically to predict responses to newer targeted therapies [10].

The molecular classification is based on the intrinsic subtypes identified by Perou and et al. based on gene expressions patterns and molecular signatures [17] however, the expression level of hormonal and HER2 receptors and proliferative marker Ki67 surrogate the intrinsic subtypes and led to establishment of the following molecular subtypes; triple negative (ER-, PR-, HER2-), HER2 subtype (ER-, PR-, HER2+), luminal A subtype (ER+, PR+, HER2-, Ki67-) and luminal B subtype (ER+, PR-/+, HER2-/+, Ki67+) [18]. These molecular subtypes are updated and recent definitions were made by Prat et al, to distinguish further luminal A and B subtypes where luminal A requires PR receptors positive, while luminal B subtype can be PR positive or negative [19]. Also, these molecular subtypes enable the prediction of overall survival and cancer progression, for example; the triple-negative ($ER^-/PR^-/HER2^-$) subtype having the shortest survival among the other subtypes [20], while stratification of the ER+ population into two subtypes (i.e., Luminal A and Luminal B) identified groups with very different clinical outcomes [21].

1.2. Treatment for breast cancer

Treatment options for primary breast cancer include surgical resection (breast conserving surgery (BCS), or partial or total mastectomy, as well as axillary surgery when required), radiotherapy, cytotoxic chemotherapy, endocrine therapies, and biological therapies. Surgery is usually the first treatment for breast cancer, with the

intention of removing the whole of the primary neoplasm. Depending on the type and completeness of excision, the surgery is often followed by chemotherapy and radiotherapy. Also, depending on the molecular profile of the excised tumour, endocrine therapies targeting oestrogen function (for example Tamoxifen or Arimidex) and/or therapies targeting HER2 (for example Trastuzumab or Lapatinib) may also be given [22]. When these therapies are used after surgery they are referred to as 'adjuvant' therapies. The main aim of these adjuvant therapies is to increase the likelihood of elimination of clinically silent micro-metastases, and thereby reduce rates of metastatic recurrences [23].

Primary breast cancer patients are typically selected for chemotherapy based on the pathology report, in which tumour is evaluated for histological grade, axillary nodal status, expression of the hormone receptors (ER receptor and PR receptors), and amplification status of the HER2 receptors. Many patients with early breast cancer are not required to receive adjuvant chemotherapy, because their chance of being cured by surgery and hormone therapy alone is high. However, chemotherapy is usually recommended for patients with triple negative subtype breast cancer, since they lack of hormonal targeted therapies and also HER2 over-expression subtype usually receive chemotherapy in addition to anit-HER2 targeted therapy. Luminal subtypes with advanced stage breast cancer (i.e. large tumour and positive lymph nodes metastasis) and those who showed resistance to targeted hormonal therapies are required to received chemotherapy in addition to surgery [24].

The choice of which chemotherapy to use depends on patients' overall health in terms of their ability to tolerate the different side-effect profiles of the agents, patient age and menopausal status, and duration and response to previous chemotherapy. It is often given for a fixed number of cycles, especially with regimens that incur toxicity such as some taxane based-chemotherapy, although some regimens may be given long term (for example, paclitaxel and capecitabine). In general, a high-dose anthracycline-based chemotherapy regimen is usually preferred for early and locally advanced breast cancer, as compared to low-dose anthracycline-based regimens or to non-anthracycline-based regimens. Adjuvant anthracycline-taxane combination chemotherapy is usually considered in patients with high-grade or stage tumours and

able to tolerate the toxicity associated with taxanes based chemotherapy. For example, patients with lymph node-positive breast cancer, certain taxane based drugs such as docetaxel can be added as part of an adjuvant chemotherapy regimen [25]. With respect to types of cytotoxic chemotherapy, a wide range of different chemotherapy drugs are available, each with different mechanisms of action. (Table 1.1) [26, 27].

Drug (Chemical name)	**Brand name**	**Mechanism of action**
Anthracycline family chemotherapy agents		
Doxorubicin	Adriamycin	Multiple mechanisms: Intercalates between DNA/RNA base pairs. Inhibits topoisomerase II enzyme activity thus blocking DNA replication. Forms free radicals destroying cell membranes [26]
Epirubicin	Ellence	
Taxane family chemotherapy agents		
Paclitaxel	Taxol	Interferes with the normal function of microtubule growth by hyper-stabilizing their structure and thus inhibiting the cells' ability to use the cytoskeleton in division [26]
Docetaxel	Taxotere	
Other chemotherapeutic drugs		
Cyclophosphamide	Cytoxan	Metabolised to form phosphoramide mustard, which forms DNA crosslinks both between and within DNA strands and leads to cell apoptosis [27]
Vinorelbine	Navelbine	Interferes with chromosomal segregation during mitosis, thus inhibiting cell growth [27]
Capecitabine	Xeloda	Thymidylate synthase inhibition which is required for DNA synthesis [27]

Table 1.1 Common cytotoxic chemotherapy agents used singly or in combination to treat early and locally advanced stage breast cancers.

1.2.1. Neoadjuvant versus adjuvant chemotherapy

Neoadjuvant chemotherapy (NAC) is given before the surgery and it has additional advantages over adjuvant chemotherapy. Firstly, when a patient presents with a breast cancer so large that mastectomy is technically not possible, neoadjuvant chemotherapy may reduce its size, making it possible to do a mastectomy with curative intent [28]. But many such patients also have distant metastases, so surgery is only performed for symptom control as it can worsen distant disease-free survival; this can be also a rare indication for neoadjuvant chemotherapy [29].

Secondly, despite there is no firm evidence demonstrating a survival benefit with NAC, it gives a major advantage for patients who want to proceed with Breast Conservative Surgery (BCS) in circumstances where the initial size of the tumour would not permit this [30]. In randomised controlled trials, BCS plus radiation has been shown to be at least equivalent, or even superior in terms of survival to mastectomy [31].

Thirdly, since the success of the treatment can be assessed by imaging (typically MRI) and by palpation, this approach has the additional benefit of permitting a switch to a different drug regime if the initial approach shows no evidence of success. However, there is no strong evidence suggest switching treatment drug regimens will work before surgery, but this has greater role following the surgery [32]. Since reporting on complete pathological response (pCR) of the resection samples following treatment correlates with disease free survival, this will enable further therapy or novel agents to be given to poor responders or avoid further adjuvant therapy if pCR is achieved. Also, whilst determining pCR is important for switching regiments and trails, this can serve as a prognostic marker of eradication of micrometastatic disease that cannot be imaged [33-35].

After NAC, if there is evidence of tumour shrinkage to a suitable level, the residual tumour or site of the tumour can be fully excised. The degree of response to the treatment and the adequacy of the excision is assessed by the pathologist thorough examination and on reporting pCR of the resection samples. In general, absence of residual invasive disease in the breast and in the axillary lymph nodes at the completion of the NAC indicates pCR. Whereas, patients show visible tumour in the breast and lymph nodes likely they have not responded to treatment and the risk of distant disease recurrence is higher [33].

It is interesting that some patients' tumours respond totally, partially, and some cases show complete resistance to chemotherapy or in some cases respond well for a period of time and then grow back. Interestingly, the molecular subtypes seems to influence the chemotherapy response for example; while the luminal A subtype is the most common in breast cancer patients, it has the highest proportion of patients who do not achieve pCR [36, 37]. However, it is important to note pCR does not reflect survival outcomes in this subpopulation of breast cancer as hormonal therapy probably equally important in disease control. This variation in response to treatment can be explained partly by tumour heterogeneity.

The availability of matched samples pre- and post-treatment following neoadjuvant chemotherapy protocol has been a valuable source for researchers in different studies [38-40]. This enabled study of the changes induced by chemotherapy and identification of chemotherapy sensitivity targets and also development of prognostic and survival predictive markers for chemotherapy response. For example in a study investigated the expression of multiple proteins between pre- and post-NAC which led to find BCRP as a predictive survival marker in breast cancer [40].

1.2.2. Mechanisms of chemo-resistance in cancer

Chemotherapeutic drug resistance has been a major obstacle in successful treatment of cancers. About 30% of patients with early-stage breast cancer have recurrent disease due to treatment failure [4]. In breast cancer, resistance to treatment is not only confined to chemotherapeutic drugs but also seen in other systemic treatment of breast cancer includes hormonal, and immunotherapeutic agents. In general, mechanisms of drug resistance can be disease specific, while others are evolutionarily conserved. Among the known conserved chemotherapeutic drug resistance mechanisms that are observed in human cancers are; drug inactivation, drug target alteration, drug efflux, DNA damage repair, cell death inhibition, and the epithelial-mesenchymal transition (EMT) [41].

Drug inactivation involves interactions between anticancer drugs and different proteins or molecules for drug metabolic activation in order to acquire clinical efficacy. However, cancer cells can also develop resistance to such treatments through decreased drug activation [42]. One example of this is observed in the treatment of acute myelogenous leukemia with cytarabine (AraC), a nucleoside drug that is activated after multiple phosphorylation events that convert it to AraC-triphosphate. Down-regulation or mutation in this pathway can produce a decrease in the activation of AraC, and this can lead to AraC drug resistance [42, 43].

A drug's efficacy is influenced by its molecular target and alterations of this target, such as mutations or modifications of expression levels [42]. For example, about 30% of prostate cancers, harbouring genemoic amplification of the androgen receptors in which androgen receptors targeted therapies such as leuprolide and bicalutamide fail to inhibit all the molecular targets present. Thereby, this leads cancers to survive and subsequently develop resistance due to androgen deficiency therapy [42, 44].

One of the most studied mechanisms of cancer drug resistance involves reducing intracellular drug accumulation by enhancing efflux. Members of the ATP-binding cassette (ABC) transporter family proteins enable this efflux and are important, well-studied regulators at the plasma membranes of healthy cells [45]. Their efflux mechanism plays an important role in preventing over accumulation of toxins within the cell [42]. Transporters-multidrug resistance protein 1 (MDR1), multidrug resistance-associated protein 1 (MRP1), and breast cancer resistance protein (BCRP) are implicated in many drug resistant cancers. For example, in tissues that do not normally express *MDR1*, such as lung, breast, and prostate cells, are often drug resistant due to the expression of the related transporters MRP1 or BCRP [42, 46].

The repair of damaged DNA has a clear role in anticancer drug resistance. In response to chemotherapy drugs that either directly or indirectly damage DNA, DNA damage response (DDR) mechanisms can reverse the drug-induced damage [42]. For example, platinum-containing chemotherapy drugs such as Cisplatin cause harmful DNA crosslinks, which can lead to apoptosis. However, resistance to platinum-based drugs often arises due to nucleotide excision repair and the primary DNA repair mechanisms involved in reversing platinum damage [42, 47].

Many chemotherapy agents lead to induction of apoptosis (cell death), however, the up-regulation of the anti-apoptotic genes such as Bcl2 and AKT, and down-regulation of pro-apoptotic genes such as Bax and Bcl_{xl} in tumour cells are associated with increased resistance to chemotherapy [48]. For example, the drug resistance can occur by the mutations in the p53 gene - a pro-apoptotic gene, which could impair the connection between DNA damage caused by chemotherapeutic agents and the activation of apoptosis [48, 49].

Epithelial to Mesenchymal Transition (EMT) is a mechanism by which cells within solid tumours can become metastatic. Several factors during EMT play significant roles in the development of drug resistance, such as cell adhesion receptors, including integrins and cadherins which are also involved with metastases development [42].

For example, in HER2 positive breast cancer, tumours that express high levels of β1 integrins develop more resistance to antibody inhibitors such as trastuzumab [42, 50].

In addition, cancer cell heterogeneity has received a lot of attention in recent years as a potential mechanism for development of chemotherapeutic drug resistance. Studies have suggested that heterogeneous populations of cancer cells could have two relevant coexisting dominant components - one being drug sensitive while the other is drug resistant. Of those resistant clones, some have stem cell properties and are usually drug resistant. The treatment of cancers, by definition, kills only drug sensitive cancer clones, and while the drug resistant cancer clones or minor clonal populations present at low frequency will survive or expand to contribute to pathology over time [42]. A clonal composition study of breast cancer revealed that breast cancers may have monogenomic or multiple genomic tumours. Polygenomic tumours contain many different types of clonal subpopulations, all of which may have different drug sensitivities and resistance characteristics [42, 51].

Taken together, the drug resistance of cancer stem cells and the acquired drug resistance of cancer cells following resistance mechanisms pose a very complex challenge for the development of better therapies to reduce the relapse of cancers. Thus, understanding cancer cells heterogeneity mechanism would help to tackle the issues related to treatments resistance.

1.3. Breast cancer heterogeneity

It is well recognised that breast cancer is a heterogeneous disease [7, 8, 52]. Heterogeneity within breast cancer could be broadly split into inter-tumour heterogeneity and intra-tumour heterogeneity. Inter-tumour heterogeneity refers to the differences between different tumours. This heterogeneity is reflected in the clinical, morphological, genetic and molecular differences that exist between different tumours. These differences form the basis of the various classifications of breast cancers that are used to guide treatment stratification (see section 1.1) [53]. Whereas, intra-tumour

heterogeneity refers to the variation between different cells of an individual tumour. Sometimes the fact that tumour contain a variety of non-tumour cell types (collectively referred to as cancer stromal cells) is included in this term, but in the main I was interested in the intra-tumour heterogeneity that is due to the underlying genetic and epigenetic differences between the individual tumour cells. These differences can be observed through cell morphology and size differences, differential expression of markers as determined by immuno-histochemistry, and even through functional characteristics such as proliferation rate, metastatic capability and sensitivity to treatment [52]. There are two common mechanistic models that are widely accepted to explain intra-tumoural heterogeneity. These are the cancer stem cell hypothesis and the clonal evolution model; each is described below.

1.3.1. Cancer stem cell hypothesis

According to the cancer stem cell hypothesis, tumours are composed of a majority tumour cell population that has limited replicative ability, and only a small sub-population drives tumour maintenance and progression, which is termed the cancer stem cells (CSC). CSCs have some properties of normal stem cells, such as unlimited replicative potential and the ability to differentiate into phenotypically diverse progeny, and are regarded as the tumorigenic cells. It is likely that epigenetic changes are associated with differentiation of the cells to form tumorigenic (CSC) and non-tumorigenic cancer cells in a single tumour mass, which results in tumour cells heterogeneity. Tumorigenic cancer stem cells, unlike non-tumorigenic cells, have the ability to drive the tumour progression and make tumours resistant to treatment [8, 9, 54] (Figure 1). Supporting evidence has been gathered based on small population of cells isolated from tumours that have the properties of normal stem cells and also express certain surface markers thought to be characteristic of stem-type cells. When these cells are injected into immunocompromised mice, it has been found that they initiate cancer, while the other isolated cells that make up the bulk of the tumours could not. Of all solid cancers, these CSCs were first isolated in breast cancer. CD44, expression of which is associated with normal breast stem cells, was used to identify breast cancer stem cells. In a study by Al-Hajj and colleagues, they showed that human breast cancer cells which have a positive expression of CD44 and negative or

low CD24 phenotype could efficiently form tumours containing an array of cell types similar to those found in the original carcinoma samples when injected into immunocompromised mice. By contrast, CD44 negative and CD24 positive cancer cells had a much lower efficiency at seeding these tumours [8, 55]. In addition, based on stem cells markers, cancer stem cells have been found in many tumour types such as lung, brain, skin, prostate, and colon [8, 54, 56-59].

1.3.2. Clonal evolution model

Originally based on the Darwinian theory of natural selection, the clonal evolution theory of carcinogenesis states that cancer cells over time acquire external various genetic and epigenetic changes leading to heterogenous population [8, 60]. According to this idea, clonal evolution takes place once multiple mutations occur in individual cells, providing them with potential selective growth advantages, for example, self-renewal ability over other neighbouring cells. As the tumour progresses, genetic instability and uncontrolled proliferation allow the production of cells with additional mutations and hence new characteristics arise. These new characteristics will be conferred to the offspring, and the new mutations may again provide growth advantages over other tumour cells, such as resistance to treatment. As a result, new subpopulations of variant cells occur, while other subpopulations may contract. The end result is that tumours are a dynamic mixture of different genetic clones, each with different characteristics that may or may not give selective growth advantages as conditions change (Figure 1.1) [7, 8, 61, 62].

Peter Nowell was the first researcher who noted the clonal evolution model of cancer in 1976, based on the observation that cells lost morphological and metabolic properties as they progressed toward malignancy, at the same time, there is genetic variation which also associated with this as the way for cells to maximize their proliferation and invasiveness [8, 63]. In breast cancer, support for the clonal evolution model was observed in many forms. For instance, in a study that looked into mechanisms of CDK4/6 inhibitor resistance in advanced estrogen receptor positive breast cancer, circulating tumour DNA sequencing was performed on paired baseline

and end of treatment samples from 195 patients in the PALOMA-3 randomized phase III trial of palbociclib plus fulvestrant versus placebo plus fulvestrant. The data analysis showed that there is clonal evolution frequently occur during treatment, reflecting substantial sub-clonal complexity in breast cancer that has progressed after endocrine therapy [64]. In addition, a study using comparative genomic hybridisation was performed on paired primary and metastatic samples from 12 patients in order to study the extent of the genetic relationship between primary tumours and regional metastases. The findings showed there is extensive clonal divergence between primary carcinomas and lymph node metastases in several cases. Also, the number of genomic imbalances in primary tumours was significantly higher in patients presenting lymph node metastases than in the patients with no evidence of disease spreading at diagnosis. This confirms the clonal evolution existence between paired primary breast tumours and lymph node metastases and clonal evolution likely occurs during tumorigenesis and may therefore continue during tumour progression [65].

1.3.1. Are the cancer stem cell and clonal evolution models mutually exclusive?

It seems that these two models that explain breast tumour heterogeneity are not mutually exclusive of each other, but it is more likely that the heterogeneity is caused by a version of the clonal evolution model that incorporates some features of the cancer stem cell hypothesis.

Findings that support this conclusion include, a study investigating the mechanism of intra-tumoural heterogeneity in breast cancer by performing a comprehensive molecular characterisation including gene expression profiles, SNP arrays and FISH on separate populations of cells that were $CD44^{+}$ $CD24^{-}$, which indicated cancer stem cells, or $CD44^{-}$ $CD24^{+}$, indicating more differentiated cells [8, 66]. The findings supported several aspects of both CSC and clonal evolutions. For example; in favour of CSC they found breast cancer stem cells (cancer CD44+ cells) may be derived from normal stem cells since they both expressed similar stem cell markers. Also, cancer CD44+ cells had a more migratory, angiogenic, and invasive phenotype than $CD44^{-}$,

indicating that these potential cancer stem cells could contribute to metastasis [8]. In contrary, some observations were made which contradict with CSC hypothesis and support clonal evolution model. For example, the finding that CD44+ (stem cells) and CD24+ (differentiated cells) cells within a tumour are clonally related but that the CD24+ cells have an additional genetic alteration shows that CD24+ cells can acquire mutations independently of CD44+ cells, which is not compatible with the CSC hypothesis. This also indicated that CD44- CD24+ cells, which were previously shown to be largely non-tumorigenic, could undergo mutations that made them more tumorigenic or promoted cancer progression. In addition, it was found metastases contained a higher frequency of CD24+ cells than did matched primary tumours, which may support clonal evolution since it could mean that cancer CD24+ cells progress in the different environments they encounter at sites of metastasis [8].

Research is needed to give a better understanding of the mechanisms of intra-tumoural heterogeneity in order to design better treatment that prevent therapy resistance and ultimately improve cancer outcomes.

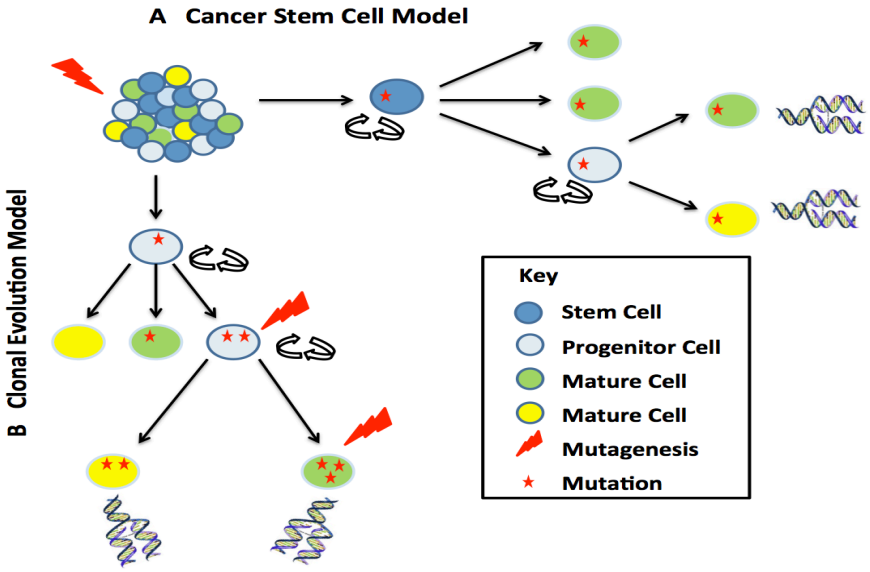


Figure 1.1 A schematic to illustrate the cancer stem cell and clonal evolution models in the context of intra-tumoural heterogeneity and therapy resistance mechanisms. (A) According to the cancer stem cell hypothesis, mutagenesis is likely to occur in normal adult stem cells, which are self-renewing cells. Based on the hierarchical organization of cells, a small subset of cells has the ability to sustain tumorigenesis and generate heterogeneity through differentiation. As in this example, normal stem acquired mutations specific to treatment resistance along with other mutations which made them tumorigenic cancer stem cells, that will be passed to progenitor and mature cells (B) The clonal evolution hypothesis states that cells acquire a varied series of mutations that can confer selective growth advantages like self-renewal ability and therapy resistance mutations. Note in this example, some cells without mutations those are original from non- mutated progenitor cells. This process repeats as tumour cells may acquire further mutations resulting in diversity of cells within the tumours. The extracted DNA from diverse cells will enable to detect clones and sub-clones mutations using NGS.

1.3.2. Implications of intra-tumoural heterogeneity (ITH) for cancer treatment

Intra-tumoural heterogeneity (ITH), arising either from the stem cell or the clonal evolution models, has been implicated in defining the tumour response to treatment. A number of studies have demonstrated that treatment can act as a selection pressure in some malignancies, driving clonal evolution and selecting for certain sub-clones. In a study in non-small cell lung carcinoma (NSCLC), the authors detected the presence

of mutations in EGFR in pre-treatment samples at low frequency (<5% of cells). These mutations are known to be associated with resistance to therapy with tyrosine kinase inhibitor (TKI), targeting EGFR. After treatment it was found that the mutation was represented in sub-clonal populations at a greater prevalence, indicating that the mutation acted as a driver of sub-clonal expansion of EGFR TKI therapy resistance phenotypes [67, 68]. It has also been suggested that cancer therapies can generate new sub-clonal drivers of therapy resistance. In glioblastoma multiforme (GBM), temozolamide is typically used to treat GBM as first-line therapy. Temozolamide induces mutations in tumour DNA: some of these are deleterious for the cells and result in death, however, others are neutral in terms of phenotype and act as passenger mutations, while still others, such as mutations in mismatch repair genes, are potentially advantageous for tumour cells allowing them to survive. It was found that some tumour cells after treatment with temozolamide harboured driver mutations in RB1 (encoding retinoblastoma 1), PIK3CA and PTEN that are the signature of temozolamide-induced resistance [67, 69]. In addition, studies have identified clonal evolution in ESR1 driving resistance to aromatase inhibitor and also of HER2 have acquired HER2-targted therapy resistance including lapatinib and trastuzumab [70, 71].

It should be noted that not all mutations in cancer genomes contribute to malignant initiation or progression, as mutational processes affect cellular functions and processes beyond those relevant to cancer development. Therefore, this illustrates the need to identify driver events and mutational processes that contribute to tumour recurrence and treatment resistance. Also, identifying the sub-clones following treatment or recurrences would direct the treatment strategies and help tailoring the treatment targeting those evolved sub-clones drivers. At this point it may be useful to define terms often applied to genomic, or even epigenetic changes in cancer cells: drivers and passengers. Mutations that provide a selective growth advantage, and thus promote cancer development, are termed driver mutations, and those that do not are termed passenger mutations. The terms driver and passenger may also be used to refer to the genes harbouring driver mutations. Genes that have been identified as drivers in at least one cancer type are described as cancer genes [72, 73].

1.4. Genomic studies into heterogeneity in breast cancer

1.4.1. Next Generation Sequencing (NGS)

Next-generation sequencing (NGS) has enabled powerful analysis of tumour evolution and has allowed improved understanding of tumour initiation and development [52]. NGS technologies (also known as massively parallel sequencing) can generate hundreds of millions of short DNA reads (usually 36- to 700-bp) that can be aligned to the reference human genome to detect focused mutations (point mutations and small insertion or deletion (indels)), copy number alterations, and structural variants, including fusion genes in cancer genomes [74]. In addition to genomic sequencing, NGS has also enabled researchers to sequence the transcriptome using RNA-seq, thereby discovering novel transcripts, RNA variants and splice sites, or quantifying genome-wide gene expression. Furthermore, epigenetic processes have been studied using genome-wide methylation analysis, or DNA-protein interactions using Chip-Seq [75].

Importantly, it is also possible to confine the sequencing to the protein-coding portion of the genome (the exome), which represents around 1.5% of human genome, but contains around 85% of most currently known diseases-causing mutations [76]. This enabled researchers to avoid whole genome sequencing, which is expensive in terms of money and time in data analysis, while focusing more effectively on regions of the genome most likely to be of interest [52]. NGS have proved to be a revolutionary tool in cancer research because it has become possible to reveal minor allelic mutations, thus enabling researchers to study the patterns of tumour heterogeneity and pathogenesis of tumour progression and treatment resistance. Therefore, a future aim is implementing NGS in the clinical practice to allow personalised medicine in which treatment will be based on the targeting of the individual tumour's genomic aberrations [52, 75, 77, 78].

Various different manufacturers have designed different NGS platforms, based on different chemistries, and these typically differ in their technical capabilities in terms of read lengths (the length of individual pieces of nucleic acid sequenced), amount of

output (the total amount of sequence produced per sample loaded), and run time; these different parameters have different suitability for specific lab settings, project designs and – importantly - finance. Table 2 provides a concise summary of current prominent commercially available NGS-platforms [75].

Company	**Platform**	**Sequencing**	**Read Length**	**Max Output**	**Run Time**	**Pros/Cons**	**Applications Highlights [79]**
Illumina	HiSeq 3000	Sequencing by synthesis using fluorescently labelled nucleotides	1 X 50 bp or 2 X 75 bp or 2 X 150 bp	750Gb	3.5 days	Pros: High output, platform has a fast run mode Cons: short reads	High-throughput applications for reference genomic lab settings. Mainly for variant discovery by WGS or WES and gene discovery in metagenomics [79]
Illumina	HiSeq 2500	Sequencing by synthesis using fluorescently labelled nucleotides	100bp X 100bp or 150bp X 150bp	600Gb	11 days	Pros: High output, platform has a fast run mode Cons: short reads	
Illumina	MiSeq*	Sequencing by synthesis using fluorescently labelled nucleotides	150bp X 150bp	2Gb	24 hours	Pros: Low error-rate, high output per run Cons: Shorter read lengths	Personal benchtop sequencer for small projects for WES and WGS [79]
Life Tech	Ion Torrent PGM*	Ion semi-conductor sequencing	Up to 400bp	Up to 1Gb	2 hours	Pros: Long reads Cons: *Homopolymer-associated indel errors are frequent	Most useful for targeted re-sequencing projects and small genome analysis. Particularly useful for clinical applications [79]
Life Tech	SOLiD 5500	Supported oligonucleotide ligation and detection	75bp X 35bp or 60bp X 60bp	~80Gb	< 7days	Pros: Low sequencing error-rate Cons: Short reads, low output	High throughput application for WGS or WES and metagenomics. Due to short reads not ideal for de novo assembly [79]
Roche	454 GS FLX	Pyro-sequencing	Up to 1kb	700Mb	23 hours	Pro: Longer reads, fast Cons: homopolymer-associated indel error are frequent, low output	Applications are mostly limited to targeted panel sequencing and small genomes sequencing like microbes [79]

Roche	454 GS Junior	Pyro-sequencing	~700bp	~35Mb	10 hours	Pros: Long reads Cons: Homopolymer-associated indel errors are frequent, low output	Benchtop system to sequence clinically relevant exons and targeted panel sequencing. Mostly used for identifying genomic variations in solid tumours treated with an antibody-based medicine [79]
Oxford Nanopore	MinION	Nanopore's current technology	Upto 60Kb	~90Mb	18 hours	Pros: very long reads, fast, machine size Cons: high error rates, low throughput	Mostly suitable for outdoor research field due to size and also for bacterial genome sequencing. Due to long reads ideal for structural variations analysis [79]
Pacific Bioscience	PacBio RS	Single-molecule Real-Time Sequencing (SMRT) cells containing 150K ZMW-wells	Average lengths of 3kb (max 15 kb)	~60Mb/ SMRT cell	< 2 hours	Pros: Very long reads, very fast Cons: High Sequencing error-rate, low-output	Particularly useful for projects involving de novo assembly of small bacterial and viral genomes as well as large genome finishing [79]

Table 1.2 A summary of the characteristics of prominent commercially-available NGS platforms. *Homopolymer = a sequence of consecutive identical bases. **~60Mb needed for to sequence all exons regions. The table was adapted from Desmedt et al. (2012) [75].

1.4.2. Genomic studies on breast cancer

There have been many studies in which the authors have attempted to characterise the heterogeneous nature of breast cancer using NGS. For example, 4 major studies that have been conducted on large-scale breast cancer cohorts [75, 80-83] - study details such as cohort size, breast cancer classification, and sequencing application are summarised in Table 1.3. The key findings from these studies are as follows. First, only the p53 and PIK3CA genes had somatic mutations in >30% of breast cancer patients, while many of the identified cancer genes thought to be potential drivers for cancer progression were mutated in less than 10% of the cohort. Secondly, there is a

large mutational landscape among different breast tumours, however, many identified mutated genes within the cohort can be grouped into the deregulation of similar pathways. For example, Stephens et al. showed that at least 6 of mutated genes were acting in the same JUN kinase pathway [82]. Similarly, Shah et al. observed that pathways involving p53, chromatin remodelling, PIK3 and ERBB signalling were over-represented in the mutated genes set [80]. This indicates that although many genes are less frequently mutated, they can be categorised into similar pathways when considering pathways that could be targeted therapeutically. Thirdly, in some tumours, there were no obvious driver mutations that could be identified, which suggested that the mechanisms for driving cancer progression in these cases may be due to non-genomic aberration such as epigenetic modifications. Fourthly, some of the mutations identified were associated with the response/resistance to treatment. An example of this point is that Ellis et al. showed that mutations in the GATA3 gene correlated with suppression of proliferation upon aromatase inhibitor treatment [81]. This demonstrates there are mutations potentially direct the treatment response and contribute to the heterogeneity landscape of tumours.

Study done by	**Number of patients**	**Type of breast cancer**	**Type of sequencing**	**Platform**	**Total number of detected mutations**
Shah et al. [80]	104	Triple Negative Breast Cancer	54 cases WES 15 cases WGS 80 cases RNA-seq	Illumina GAII SOLiD system	2414 somatic point mutations and indels
Ellis et al. [81]	77	Oestrogen receptor positive breast cancer	46 cases WGS 31 cases WES	Illumina Sequencer 454 Sequencer	Average 1,780 somatic mutations (point mutations and indels) and 16.8 somatic structural variants.
Stephens et al. [82]	100	All sub-types	All WES	Ilumina GAII HiSeq DNA	7,241 somatic point mutations
Banerji et al. [83]	108	All sub-types	17 cases WES & WGS 5 cases WGS 86 cases WES	Ilumina Sequencer	4,985 somatic substitutions and insertions/deletions

Table 1.3 Prominent genomic sequencing studies in breast cancer

There is a study by the Cancer Genome Atlas Network in which a wide range of analyses were performed, including genomic DNA copy number arrays, DNA methylation, exome sequencing, mRNA arrays, microRNA sequencing and reverse phase protein arrays on 825 patients with primary breast cancer. The data from the five platforms enabled the authors to provide key insights into previously defined gene expression subtypes (luminal A, Luminal B, triple negative, HER2 overexpression) each of which showed substantial molecular heterogeneity. For example: somatic mutations in only three genes (TP53, PIK3CA and GATA3) occurred at >10% incidence across all breast cancers; however, there were numerous subtype-associated and novel gene mutations including the enrichment of specific mutations in GATA3, PIK3CA and MAP3K1 within the Luminal A subtype. Also, there were specific signalling pathways dominant in each molecular subtype such as a HER2/p-HER2/HER1/p-HER1 signature within the HER2-Enriched expression subtype. This illustrates that there are potentially further possibility to characterise the four main breast cancer subtypes based on heterogeneity occurring within these subgroups, which could enable a better personalised treatment [84, 85].

A further example is a study conducted in a very large-scale cohort integrating analysis of copy number and gene expression in a discovery and validation set of 997 and 995 primary breast tumours, respectively, with long-term clinical follow-up; this led to revealing 10 molecular subgroups with distinct clinical outcomes. The majority of subgroups demonstrated enrichment for a number of putative driver genes, but rare to be found in other cancer types, also copy number alterations were found in a number of therapeutic targets such as amplification of IGF1R, KRAS, and EGFR. This demonstrated that the development of an integrated genome-driven classification can lead to robust breast cancer subgroups that further resolve the heterogeneity of existing classification [84, 86].

The above genomic findings assert the generic finding that breast cancer is indeed a heterogeneous disease with distinct biological and clinical features. Hence, there is a need for more characterisation of breast cancer heterogeneity in order to enable more

driver genes discovery, which will allow design of targeted therapeutic agents with superior efficacy and less toxicity, thereby improving clinical outcomes.

Despite this wealth of data, key critical questions concerning breast cancer genomics remain: for example, the functional impact of each genomic aberrations to cancer progression, and response to treatment. There is a lack of functional validation of the genomics date, which is an important step for translating these findings into new medical advances [87]. In other words, it is important to interpret the genomic data in the context of heterogeneity into meaningful findings in order to provide clinically relevant information for cancer patients. Thereby, in my study I aimed to enrich this gap in the current knowledge, by utilising genomic tools and functional *in vitro* approaches.

1.5. Study significance

Neoadjuvant chemotherapy (NAC) presents the opportunity to examine the impact of chemotherapy on the genomic clonal composition of breast cancers, via the comparison of pre-therapy samples (core biopsies) and post-therapy samples (resection). NGS of exomes should allow sufficient sequencing depth to assess changes in representation associated with therapy for even relatively rare somatic variants within the tumour. Insights gained from this approach could allow identification of genes that define chemotherapy response, and therefore – ultimately – could be used to improve the use of chemotherapy through better prediction of likely responses through biomarkers, or through novel therapies based on targeting resistance somatic mutations.

1.6. Hypothesis and study aims

My hypothesis was that chemotherapy drives changes in the genetic clonal content of breast cancers through selection of relatively resistant clones; identification of these mutations that have been selected for or against will allow identification of genes regulating chemotherapy response.

My aims were:

1. To perform whole exome sequencing on cancer cells from pre-neoadjuvant chemotherapy (NAC) core biopsies of primary breast cancers and the matched post-therapy resection samples in cases where residual (relatively-resistant) tumour is present.

2. To perform data analysis to assess the genetic heterogeneity in terms of detected mutational landscape, and to identify single nucleotide variants (SNV) and small deletion and insertion (Indels) mutations that have changed in their representation between pre-NAC and post-NAC, thereby identifying candidate regulators of chemotherapy response.

3. To investigate the functional impact of candidate regulators of chemotherapy response using cell line approaches.

4. To examine whether expression of candidate genes correlate with outcomes after chemotherapy in breast cancer patients.

2. Materials and methods

2.1. Ethics statement and sample collections

This work was conducted on breast cancer samples selected from the FFPE tissue block archive at St James's University Hospital (Leeds Teaching Hospitals NHS Trust), Leeds. Ethical approval was obtained from Leeds (East) REC reference 06/Q1206/180. Two cohorts were collected for this work, and a further cohort was available through collaboration. 1) A small cohort of patients (n=8) with samples suitable for laser micro-dissection (LCM) and genome sequencing ("LCM cohort"). 2) A much larger cohort (n=140) of patients treated with adjuvant chemotherapy ("Adjuvant cohort"). 3) A cohort of intermediate size (n=53) of post-neoadjuvant chemotherapy (NAC) samples from patients treated with NAC ("NAC cohort").

LCM cohort: Breast cancer cases that were treated with NAC using the regimen epirubicin and cyclophosphamide were identified using the NHS Trust computer systems. Cases were further selected on the basis of: showing a partial response to NAC, as defined by routine clinical MRI assessment of tumour size during treatment; ER-positive tumours, as defined by clinical assessment of ER-status as is routine for the breast cancer diagnostic pathway; presence of residual tumour cells post-NAC in the resection samples; and sufficient cells present in the pre-NAC core biopsies. Archival tissue slides (Haematoxylin and Eosin stained) and blocks (formalin-fixed paraffin-embedded) containing core biopsies (pre-NAC) and slides/blocks containing matching resection tissues were identified. Slides were reviewed to establish whether there was sufficient tumour tissue within the blocks for extraction, and whether the tumour tissue contained sufficient tumour cells so as to allow laser microdissection (LCM) of these cells. 8 cases were selected as suitable. Matching normal tissue blocks for these cases were also obtained. These cases were pseudo-anonymised, and were identified to the researcher only as case numbers 1 to 8, although the pathologist involved in their selection (Dr. Eldo Verghese, consultant histopathologist Leeds Teaching Hospitals NHS Trust, and co-supervisor for the project) was able to link these to clinical data when required.

Adjuvant cohort: A cohort of 140 patients with primary breast cancer diagnosed between 2006 and 2010 who received adjuvant chemotherapy were provided through colleague Ms Stacey Jones (Breast Surgery Clinical Research Fellow at Leeds Teaching Hospitals NHS Trust). Patients who underwent surgical resection for operable invasive breast cancer (including locally advanced) at a Tertiary Breast Centre (The Leeds Teaching Hospitals NHS Trust) were identified from Hardcopy diaries kept by the Breast Oncology Department. Exclusion criteria were; individuals who received NACT or who had not received adjuvant chemotherapy, individuals with metastatic breast cancer, individuals with a recurrence in breast cancer with primary breast cancer being diagnosed prior to 2006 and males with breast cancer. Comprehensive data including patient demographics (age at diagnosis, date of diagnosis, affected breast), pathological data (tumour type, tumour grade, receptor and nodal status), surgical data (date of surgery and surgical procedure performed), oncological data (chemotherapy regimen, radiotherapy, endocrine treatment, local and metastatic recurrence) were collected. The afore mentioned data were collected where available from pathology reports, breast multi-disciplinary team meetings, breast surgery and oncology clinical letters. Tumour slides were retrieved from the pathology files, reviewed and marked for tissue micro array (TMA) construction. The corresponding blocks of resected breast tumours were identified. Three cores of tumour were obtained from each block and placed into a new paraffin block to construct the TMAs.

NAC cohort: A cohort of intermediate size (n=53) of post-neoadjuvant chemotherapy (NAC) samples from patients treated with NAC were provided through collaboration with Dr Abeer Shaaban (consultant breast pathologist, formerly at Leeds Teaching Hospitals NHS Trust, where the patients included in this TMA were treated, but now at University Hospitals Birmingham NHS Foundation Trust). Patients who underwent NAC for primary and operable invasive carcinoma, including inflammatory and locally advanced breast cancer, at a single large United Kingdom (UK) tertiary referral breast centre (The Leeds Teaching Hospitals NHS Trust) were identified from the oncology and imaging databases. All patients undergoing NAC are routinely offered baseline

and follow-up magnetic resonance imaging (MRI) scans to assess response. Included in this cohort were patients with primary operable breast carcinoma who had received NAC between 1st January 2005 and 30th April 2013 followed by breast surgery, including patients who underwent diagnostic surgery to the axilla (sentinel node biopsy) before or after NAC. This period corresponds to the introduction of anti-HER2 therapy (from 2005 onwards). Exclusion criteria were; NAC patient who did not undergo surgery; metastatic breast cancer, and; NAC patients without MRI follow-up. Comprehensive clinical data including chemotherapy regimen, type of surgery and imaging characteristics (mammography and MRI) were collected. The following pathological data, where available, were collected from the pathology reports on both pre-treatment core biopsy sample and residual tumours: tumour type, tumour grade including individual scores for tubule formation, nuclear pleomorphism and mitoses and pathological response: classified into complete (no residual invasive carcinoma), partial (residual invasive carcinoma with histological evidence of tumour response), and no response (no evidence of tumour response). For residual invasive disease (pathological partial response and no response), tumour slides were retrieved from the pathology files, reviewed and marked for tissue microarray (TMA) construction. The corresponding tumour block was identified.

Finally, further blocks of tonsil, small intestine, ovaries and breast tissue were also collected for use in optimization work by Dr Eldo Verghese; these were fully-anonymised (i.e. once collected it was not possible to link back to patient data and no data were collected) and were identified only as numbered blocks.

2.2. Whole Exome Library construction

In this section, it covers the process involved for preparing of WES library starting from samples sectioning and LCM to data analysis. A summary flow chart is illustrated in figure 2.1.

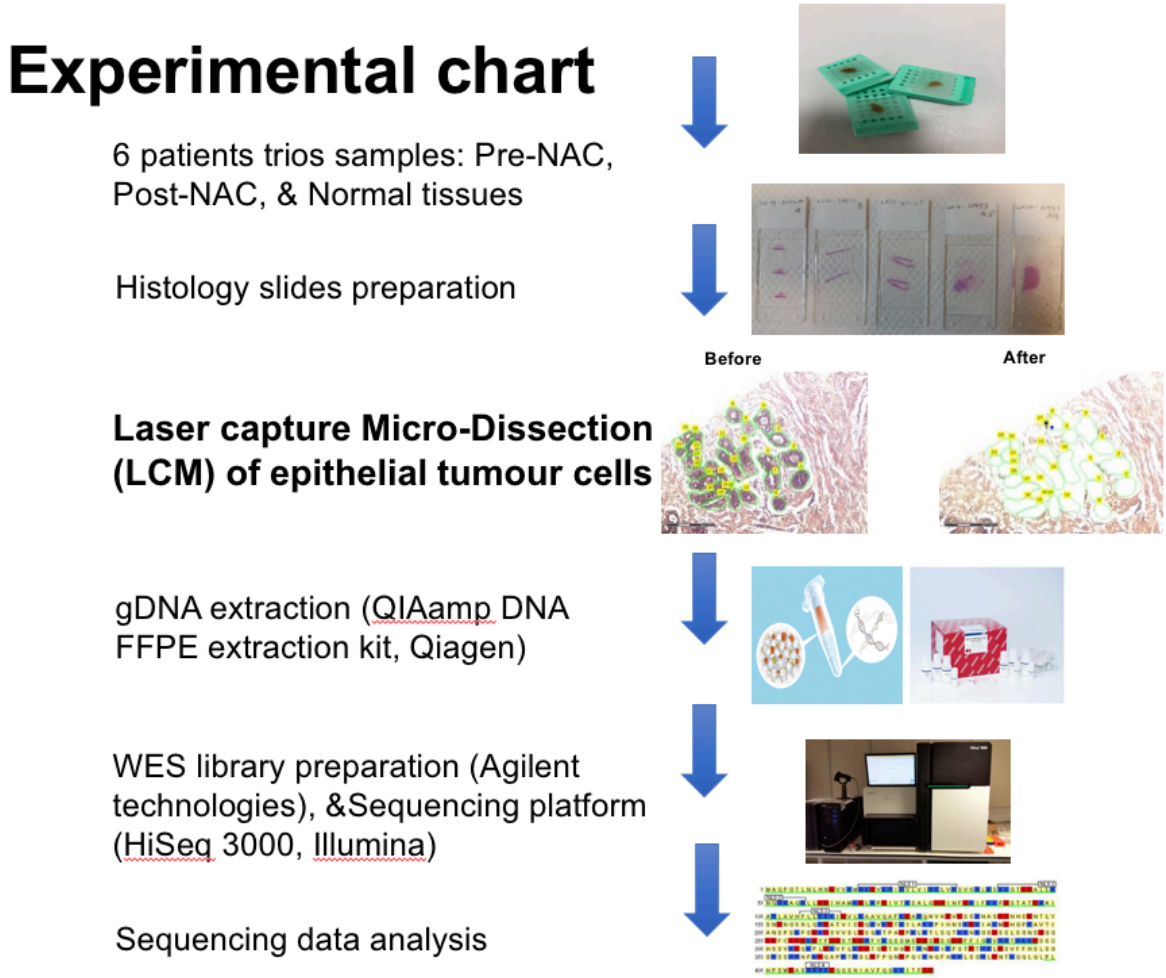


Figure 2.1 An experimental chart illustrates the process involved for WES library construction

2.2.1. Samples sectioning and staining for laser micro-dissection (LCM)

Tissue samples were sectioned at 5 or 10 microns using a microtome (Leica; Wetzlar, Germany). Sections were floated onto membrane slides, which were compatible with down-stream LCM work (Membrane Slide 1.0 PEN, Zeiss; Oberkochen, Germany) or onto standard microscopic glass slides (ThermoFisher Scientific; Massachusetts, USA). Slides were stained using Haematoxylin and Eosin (H and E) using standard protocols. Briefly, sections were dewaxed using xylene and hydrated using decreasing concentrations of ethanol. After that, slides were kept in running water for 1 minute. Next, slides were stained in Haematoxylin stock concentration (Solmedia; Shrewsbury, UK) for 30 seconds and returned to running water for another minute. Slides were then stained in 1% aqeous Eosin (Solmedia; Shrewsbury, UK) for 30 seconds and returned to running water. Next, slides were dehydrated using increasing concentrations of ethanol and dried by evaporation.

2.2.2. Laser capture micro-dissection (LCM)

Laser Capture Micro-dissection (LCM) was performed using a Zeiss/P.A.L.M. machine (P.A.L.M. Microlaser Technologies, Zeiss; Oberkochen, Germany). Tissue samples were sectioned at 5 microns thickness and stained with Haematoxylin and Eosin as described in section 2.2. The tumour cells to be dissected were marked on a reference slide with help from Dr. Eldo Verghese (consultant histopathologist and co-supervisor). Typically 5 to 10 slides were dissected depending on the availability of tumour tissue and on the density of tumour cells within the tissue. The area of dissected tumour ranged between 500,000-5,000,000μm^2 per side. The instrument's settings for UV laser cutting energy and collecting power parameters were set according to best performance on our samples and then saved. LCM work was performed in two different ways: 1) epithelial tumour cells (cells of interest) were directly dissected and collected in LCM adhesive Caps (Zeiss; Oberkochen, Germany); 2) unwanted cells (other cellular elements) were dissected and the epithelial tumour cells left on the slides and were manually dissected using a scalpel blade. Another member of the laboratory, Ms. Diana Baxter, assisted with this work. For normal tissue blocks, cellular

tissue was identified macroscopically (note, much normal breast tissue is relatively acellular), and was manually macro-dissected – LCM was not required. Representative images of these 2 ways of performing LCM are shown in Figure 2.1.

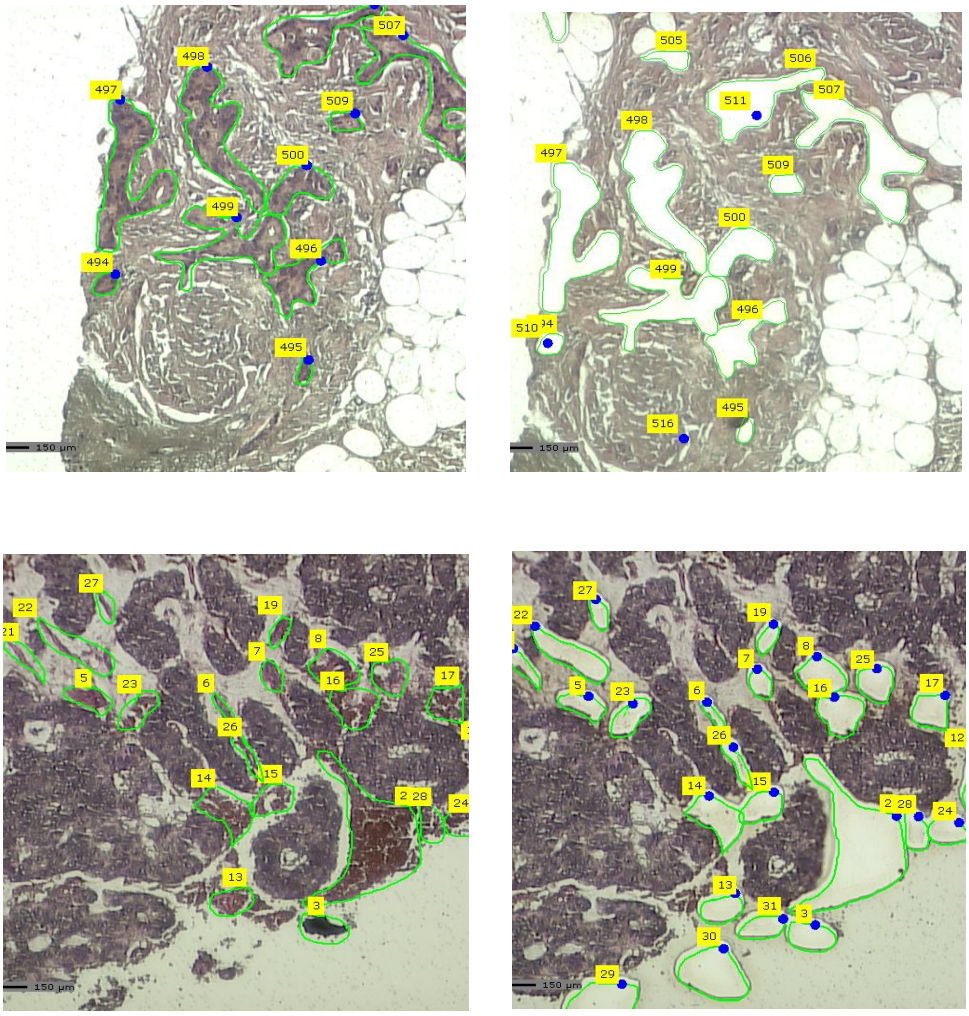


Figure 2.2 Representative images of 2 ways of performing LCM on breast tissue sections. Upper row represents epithelial tumour cells (cells of interest) were directly dissected and collected in LCM adhesive caps. Lower row represents unwanted cells (other cellular elements) were dissected and the epithelial tumour cells left on the slides and were manually dissected using a scalpel blade.

2.2.3. DNA (and RNA extraction) from FFPE samples

QIAamp MinElute Columns DNA FFPE Tissue or AllPrep DNA/RNA FFPE Kits (Qiagen; Dusseldorf, Germany) were used to extract DNA and RNA from FFPE samples following the manufacturer's reagents and protocol with some modifications. Note that although RNA was collected for some cases, this was not used in my work, and specific details relating to RNA are not included here. Briefly, for the QIAamp MinElute Columns DNA FFPE Tissue kit, 180µl of ATL buffer and 20µl of proteinase K for tissue digestion was mixed and then tissue obtained after LCM or tissue from manual micro-dissection were transferred to the mixture, vortexed vigorously and then incubated at 55°C for different periods: 1 hour, overnight, or 72 hours. After incubation, 200µl of AL buffer and 200µl of 100% ethanol were added, and then vortexted and centrifuged at 20,000 x g for 5 minutes. Next, supernatants were transferred to MinElute column in collection tubes and centrifuged at 6,000 x g for 1 minute. Next, the MinElute columns were transferred to new collection tubes and the flow through fluid was discarded. After that, washing steps with AW1 and AW2 was performed. Finally, gDNA was eluted 30 µl ATE buffer.

For the AllPrep DNA/RNA FFPE Kit, briefly, tissue obtained after LCM or tissue from manual micro-dissection was transferred to a tube containing a mixture of 150µl PKD buffer and 10µl proteinase K. The mixture was then vortexed and incubated at 56°C for 15 minutes. After that, it was incubated on ice for 3 min, followed by centrifugation for 15 minutes at 20,000x g. The supernatant was then transferred without disturbing the pellet to a new a tube for RNA purification and the pellet was kept for DNA purification. For gDNA purification, the pellet was resuspended in 180µl ATL buffer and 40µl proteinase K mixture and then mixed by vortexing and incubated at 56°C for 1 hour. Followed by incubation at 90°C for 2 hours without agitation. Next, 200µl of AL buffer and 200µl ethanol (96–100%), was added and mixed thoroughly again by vortexing. After that, the entire sample was transferred to a QIAamp MinElute spin column and placed in a 2ml collection tube and centrifuged for 1 minute at 8000 x g and then the flow-through was discarded. The QIAamp MinElute spin column was placed in a new 2 ml collection tube and a washing step with 700µl AW1 buffer was performed by centrifugation for 15 seconds at 8000 x g. Next, another washing step was performed with 700µl of AW2 buffer and with 700µl of ethanol (96–100%). The

QIAamp MinElute spin column was placed in a new 2 ml collection tube, and centrifuged at 14,000 x g. for 5 minutes. After that the collection tube with the flow-through was discarded. The QIAamp MinElute spin column was placed in a new 1.5ml collection tube and the 30–100µl ATE buffer was added directly to the spin column membrane. And then incubated for 1 minute at room temperature and centrifuged at 14,000 x g for 1 minute to elute the DNA and stored at -20°C for downstream experiments.

2.2.4. DNA Picogreen quantification assay

Extracted DNA from FFPE samples were quantified using the Quant-iT PicoGreen dsDNA reagent Kit (Thermo Fisher Scientific; Massachusetts, USA) following the manufacture's protocol. A DNA standards curve were prepared using 100µg/ml of stock standard reagent (provided by manufacturer) ranging from 0 to 100ng/µl. Samples were diluted with DNAase-free water 1 in 10. Regent and buffer mixture was prepared by making up 198ul of TE buffer with 2µl of PicoGreen dsDNA regent. Then, 198µl of the above mixture was transferred to each sample and standard wells in a 96-well black plate (Sigma-Aldrich; Missouri, USA). Next, 2µl of 1 in 10 diluted sample were added in the samples wells and 2µl of standards to each standard wells. The mixture in each well was mixed by pipetting. Next, the plate was brought to a spectrofluorometer Fluoroskan Ascent^{TM} Microplate Fluorometer (Thermo Fisher Scientific; Massachusetts, USA). Samples were excited at 480nm and the fluorescence emission intensity was measured at 520nm. Results were recorded and multiplied by the dilution factor.

2.2.5. Assessing extracted DNA quality using Agilent genomic DNA TapeStation system

The integrity of extracted DNA from FFPE samples and the quality of pre-capture library DNA was assessed using Agilent Genomic DNA ScreenTape and Agilent 2200 TapeStation system (Agilent Technologies; California, USA) using the manufacture's protocol. 10µl of genomic DNA sample buffer was added to 1µl of the genomic DNA control (10-100ng/µl), 1µl of the test sample and 3µl of genomic DNA ladder in strip tubes. Next, tubes were pulsed in a centrifuge to collect the reagents at the bottom of

the wells and vortexed, and then loaded into the TapeStation machine. Finally, the results were displayed as both electrophergrams and within tables.

2.2.6. Whole exome sequencing (WES) library construction

A modified protocol (developed by Ms. Catherine Daly, University of Leeds) for preparation of pre-capture libraries for SureSelect target enrichment for paired end sequencing of FFPE samples was used. This modified protocol uses fewer concentrating and clean-up steps, and as a result reduces DNA loss during preparation. Below I summarise the protocol, while further detail is available in the manusfacturers' standard protocols. The entire library preparation protocol was performed either by the author, or by the Leeds Genomics Core Facility (University of Leeds), depending on specific samples used and the individual library.

First, gDNA samples ranging from between 0.2µg and 1.2µg were mechanically sheared using Covaris S2 (Covaris; Massachusetts, USA) (generating fragments sizes of around 200bp). Next, end repair of DNA molecules was performed using the NEBNext Ultra DNA Library Prep Kit (New England Biolabs; Massachusetts, USA), in order to ensure that the DNA fragments contained 5' phosphate and 3' hydroxyl groups to allow ligation of the adaptors. After that, the adaptor ligation step was performed in order for the DNA fragments to bind to complementary oligos in the flow cell using the SureSelectXT Reagent Kit, HSQ (Agilent Technologies; California, USA). After that, the adaptor ligated DNA was purified using AMPure XP Beads following the manufacturer's protocol (Beckman Coulter; California, USA). Next, a PCR amplification step was necessary to amplify adapter-ligated DNA fragments to have sufficient pre-capture libraries for hybridisation and capture. It was performed using NEBNext High fidelity PCR master mix (New England Biolabs; Massachusetts, USA) and SureSelect Primer and SureSelect ILM Indexing Pre-Capture PCR reverse primers (Agilent Technologies: California, USA) using the manufacturer's standard PCR protocol for a total number of between 14 to 27 cycles. PCR products were purified using AMPure XP Beads (Beckman Coulter; California, USA). Finally, the quality of pre-capture library DNA was assessed using the Agilent 2200 TapeStation system (Agilent Technologies; California, USA) as described in section 2.6.

The SureSelect-XT Target Enrichment System for Illumina Paired-End Sequencing Library Protocol and SureSelect-XT Human All Exon V5 kit (Agilent Technologies; California, USA) were used to perform hybridisation and capture steps. Briefly, pre-capture libraries were hybridised to capture libraries containing exome biotinylated RNA library baits, and captured using the streptavidin beads. Next, index tags were added to the captured libraries through amplification in order to allow different samples to be pooled and run in one single sequencing lane reaction. 6 index assignments (A01-F01) were determined for each sample. Then PCR reaction mix was prepared using SureSelect ILM indexing Post-Capture Forward PCR primers and the appropriate indexing primers (SureSelect 8bp Indexes A01 through F01) were added to each tube containing the captured libraries accordingly. After that, tubes were transferred to a thermal cycler and run according to the manufacturer's protocol. After that, amplified captured libraries were purified using AMPure XP Beads (Beckman Coulter; California, USA). Finally, the indexed libraries were pooled and sent for sequencing using HiSeq 3000 system (Illumina Technologies, USA). The sequencing runs were set up to perform pair end reads sequencing with an 8bp index read. The read length was 2 x 150bp.

2.2.7. SureCall bioinformatics analysis

Whole exome sequencing data from tumour (pre-NAC or post-NAC) and normal samples was initially analysed using SureCall software (version 3.0.3) from Agilent (Agilent Technologies, USA). SureCall incorporates the most widely accepted open source libraries and algorithms, and augments them with tools specific for Agilent assays. The SureCall software manual was not available at the time of writing and the protocol used is described in brief here. Analyses started by uploading FASTQ files for tumour (pre-NAC or post-NAC) and matched normal into the software and then pair analysis (either tumour sample vs matched normal) was selected. Initially, reads of adaptors and low quality bases for samples were trimmed and then the alignment step performed against the *Homo Sapiens* genome reference (Hg19) using BWA-MEM to produce SAM files. SAM files were converted to the more compacted BAM files. Next, SAMTools or SNPPET was used to identify somatic variants in the tumour samples versus the matched normal, using defaults settings. Further information regarding the mutations was then aggregated from various public sources, including

NCBI, COSMIC (Catalog of Somatic Mutations in Cancer), PubMed, and Locus-Specific Databases. After that, Variant Calling Files (VCF) were generated. Finally, files were converted to an Excel format, with complete variant details and annotations from public databases if available [88].

2.2.8. Open source bioinformatics tools analysis

Whole exome sequencing data from tumour (pre-NAC or post-NAC) and normal samples was also processed to identify somatic SNVs and small indel variants using open-source bioinformatics tools by Edinburgh Genomics Laboratory (Edinburgh, UK). For each FASTQ file, adapters and primers along with poor quality bases were trimmed using cutadapt (version 1.8.3) [89]. Trimmed reads were aligned against the reference genome using BWA-MEM (version 0.7.15) [90], with parameter -M which marks split alignments as secondary and which can later be excluded by downstream tools. The reference genome used was *Homo Sapiens* Hg19. Trimmed reads were also mapped using BWA-backtrack as part of standard quality control. The alignments produced by BWA-backtrack were not used in the further analysis. PCR duplicates were marked in the BAM files using the MarkDuplicates tool from the Picard tools package (version 1.115). Base quality score recalibration (BQSR) was done using BaseRecalibrator from GATK (version 3.7) [91] in order to model any technical errors empirically and adjust the quality scores accordingly. The MuTect2 variant calling pipeline (from GATK version 3.7) was used to detect somatic variants, and the HaplotypeCaller pipeline (from GATK version 3.7) was used to detect germ-line variants. The resulting ‘g.vcf’ files were then merged into a single VCF file using the GenotypeGVCFs tool from GATK.

Annotation was added describing the likely effects of the variants. The VCF files containing the tumour samples and matched normal samples were annotated using SnpEff (version 4.3). Databases for variant annotation were obtained from the following resources: dbSNP; the dbsnp file was obtained from resource bundle supplied by GATK; COSMIC is a more highly validated resource, so it was used essentially as a whitelist to ‘rescue’ candidate mutations that would otherwise be

rejected for being in the panel of normals and/or dbSNP. COSMIC variants (version 8.0). dbNSFP; The dbNSFP file used by SnpSift was downloaded from: ftp://dbnsfp:dbnsfp@dbnsfp.softgenetics.com/dbNSFPv3.4a.zip.

Variant filtering for germ-line variants was done using the SelectVariants tool from GATK to exclude variants with a read depth of less than 5 or more than 800, or a quality phred score of at less than 30. For the somatic variants filtering, the VCF files containing the tumour samples and matched normal samples were filtered to keep only those variants considered by MuTect2 to be somatic using SelectVariants tool from GATK. Finally, for each of the samples, a TSV file containing metrics/annotations was created using snpSift tool extractFields [92].

2.3. Cell culture

MCF-7 was the cell line used in the *in vitro* studies. This line represents a breast cancer adenocarcinoma and is derived from a pleural effusion metastasis; it is characterised by expression of the oestrogen receptor alpha and the progesterone receptor and is regarded as a model cell line for luminal A breast cancers (6-7). MCF-7 cells were acquired from the American Type Culture Collection (ATCC; Manassas, USA). The cell line was propagated in Dulbecco's Modified Eagle Medium (DMEM) (ThermoFisher Scientific; Massachusetts, USA) with 10% Fetal Calf Serum (FCS) (ThermoFisher Scientific; Massachusetts, USA) in 95% air / 5% CO_2 at 37°C. Cell line identity was confirmed (STR profiles, Leeds Genomics Service) and cultures were consistently negative for mycoplasma (MycoAlert Mycoplasma detection assay, Lonza; Basal, Switzerland). The cells were sub-cultured at least once a week at ratio of up to 1:20. Sub-culturing was performed by removing the medium, rinsing in Dulbecco’s phosphate buffered saline (DPBS), and suspending the cells by incubation with 3ml of 0.5% (w/v) trypsin-EDTA (10x) (ThermoFisher Scientific; Massachusetts, USA) per T150 flask for 3-5 minutes at 37°C. After that, the trypsin was neutralised with 6ml fresh complete medium, and then an appropriate volume (depending on d FASTQ esired sub-culturing ratio) of the suspension transferred to a new flask containing 20 ml of fresh medium.

2.4. Transfection of MCF-7 cells with siRNAs

ON-TARGETplus Human siRNA-SMARTpool (Dharmacon; Colorado, USA) reagents were used to perform targeted gene knock-down screens for the 46 candidate genes. SMARTpool siRNAs comes as a mixture of 4 individual siRNA sequences provided as a single reagent. Extensive siRNA knock-down investigations were carried out on MUC17, PCNX1, TENM4 and non-targeting siRNA control: their pool target sequences are detailed below (Table 2.1).

Transfections were carried out in 96- or 6-well plates. Cells were seeded into 96-well plates at 10,000 cells/well or into 6-well plates at 500,000 cells/well and then incubated overnight. After that, siRNAs were prepared from 5μM stock solutions at either 25nM or 50nM final concentrations in serum-free medium (Opti-MEM™, ThermoFihser Scientific; Massachusetts, USA) for a total volume of 10μl (for 96-wells) or 200μl (for 6-wells) per well (tube 1). Transfection reagent DharmaFECT formula 1 (Dharmacon; Colorado, USA) was diluted in serum-free medium at concentration of 0.2μl or 4μl in a total volume of 10μl or 200μl per well, respectively (tube 2). Next, the contents of tube 1 and tube 2 were added together, mixed and incubated for 20 minutes at room temperature. 80μl or 1600μl of antibiotic-free complete medium was added to the mixtures to make up total volumes of 100μl or 2000μl per well. Culture medium was removed from wells to be transfected and 100μl (96-wells) or 2000μl (6-wells) of siRNA/transfection medium was added to each well, and then cells were cultured as normal for up to 96 hours, although after 24 hours the transfection medium was removed and replaced with complete fresh medium.

Regent name	**siRNA sequences**
ON-TARGETplus Human MUC17 (140453) siRNA-SMARTpool	Target sequence 1: GGACAAUGCCACCGAAGUA Target sequence 2: GUGCAAAACAUUACGGUGA Target sequence 3: GGACAGGUUCUGCGGCAAA Target sequence 4: GAAGAGGACUGCCGGAAGA
ON-TARGETplus Human PCNX1 (22990) siRNA-SMARTpool	Target sequence 1: GCUAAAGACACUAGAGUAU Target sequence 2: UAAGUUAGCUGCCGAGAAA Target sequence 3: AGUCUUUGAUCUUCGGAAA Target sequence 4: GCAAAUGAGUUCACGGGAUC
ON-TARGETplus Human TENM4 (26011) siRNA-SMARTpool	Target sequence 1: GGACUUUGAUCGCGUAACA Target sequence 2: GAGAGGAGAUUUCGCCUUA Target sequence 3: GGGGAGAUCUACAUGGAUA Target sequence 4: UGCCAAGGAUGCAAAGUUA
ON-TARGETplus Non-targeting Pool was used as a negative control Catalogue number: D-001810-10-05	Target sequence 1: UGGUUUACAUGUCGACUAA Target sequence 2: UGGUUUACAUGUUGUGUGA Target sequence 3: UGGUUUACAUGUUUUCUGA Target sequence 4: UGGUUUACAUGUUUUCCUA

Table 2.1 Details of targeted siRNA sequences for MUC17, PCNX1, TENM4 and non-targeting siRNA control

2.5. Epirubicin preparation and use

Epirubicin hydrochloride (Sigma Aldrich, Missouri, USA) was prepared as a 10mM stock solution in sterile dH_2O, and was stored frozen in 1ml aliquots. For use, epirubicin was diluted to working concentrations from 0.05 to 4.0 μM in fresh complete medium.

2.6. MTT assays

MTT (3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyltetrazolium Bromide) (ThermoFisher Scientific; Massachusetts, USA) assays were performed to assess cell survival after gene siRNA knock-down and/or epirubicin drug treatments over 24, 48, and 72 hours in 96-well plates. MTT working solution was prepared by adding 5mg of stock powder to 1ml Dulbecco's phosphate-buffered saline (DPBS) (Gibco, ThermoFihser Scientific; Massachusetts, USA). Medium was removed from cells in 96-well plates and wells were washed with 50μl DPBS. Next, 25μl of MTT working solution was added per well and incubated at 37°C for 2-3 hours. After that, the MTT was removed and 50μl of isopropanol was added to each well. Plates were kept on a plate shaker for 20 minutes and then colorimetric measurements were taken using a Mithras LB 940 Microplate reader (Berthold Technologies, Germany) at 570nm absorbance.

2.7. Colony forming assays

Colony forming assay was performed to assess cell survival by forming colonies after knock-down for candidate genes and/or treatment with epirubicin *in vitro*. The assays give a different read out from MTT survival assays as for cells to count as having survived in this assay they must be capable of repeated replicative division in order to form a colony, whereas for MTT assays simply surviving in the short term is sufficient. Cells were transfected with targeted siRNA and non-targeted siRNA and treated with

low doses of epirubicin (up to 100nM) for 24 hours. Medium was removed from cells in 96-well plates and wells were washed with 50µl DPBS.

After that, cells were harvested by adding 25µl of 0.5% (w/v) trypsin-EDTA (10x) for 3 minutes at 37°C and neutralised by adding 75µl of fresh media and then and counted using a haemocytometer (ThermoFisher Scientific; Massachusetts, USA). 500 or 1000 cells (depending on expected survival and to ensure a number of colonies that it was possible to count) were transferred to 10cm tissue culture dishes (Corning, Life Sciences; Massachusetts, USA) containing 11ml of fresh complete medium and cells were cultured undisturbed for 14 days (in an incubator dedicated to these assays only therefore with minimal opening/closing – this is important to minimise cells from individual colonies becoming dispersed into multiple colonies). Following incubation, the medium was removed and cells were washed with 5-10ml PBS, and then 2-3ml of fixative (Acetic acid/methanol 1:7 (vol/vol)) was added and left for 5 minutes at room temperature. Next, the fixative was removed and the plates were left to dry completely at room temperature. After that, colonies were stained with 0.5% crystal violet stain (Sigma Aldrich, Missouri, USA) for 1 minute and washed carefully with tap water until the stain background washed off. Plates were then left to dry at room temperature. Colonies were counted macroscopically taking into consideration at least 50 cells forming a colony. To validate the counting technique and reproducibility of the colonies counting scale, 10 plates with different seeding densities and drug treatments were also scored independently by my colleague Ms. Lisa Allinson (PhD student, University of Leeds). The concordance between scorers was measured using Spearmans's rank correlation coefficient statistics. It showed strong and significant inter-scorers agreement ($r=0.948$, $p<0.0001$).

2.8. Epirubicin uptake assays

The intracellular epirubicin drug uptake assay was performed using flow cytomtery on the Attune Acoustic focusing cytometer (Applied Biosystems, ThermoFisher Scientific; Massachusetts, USA) and fluorescent detection of epirubicin in the Phycoerythrin (PE) channel (BL-3). The method allowed a comparison of epirubicin intracellular drug

uptake of the cells transfected with targeted gene siRNA versus non-targeted siRNA control, by comparing the median of fluorescence intensities in the populations. Cells were seeded into 6-well plates at 500,000 cells/well and transfected as described in section 2.11. Next, cells were treated with 1uM epirubicin for 24 hours. After that, cells were detached from the wells by adding 200ul of 0.5% (w/v) trypsin-EDTA (10x) for 3 minutes at 37°C and neutralised by adding 1ml of fresh media and then centrifuged at 500x g for 10 minutes to collect the cells. Cells were resuspended in 1ml DPBS in Falcon round-bottom polypropylene tubes (Corning, Life Sciences; Massachusetts, USA) and were analysed in the standard mode at a flow rate 25ul/minute and draw volume of 100ul. The Attune Cytometric software v2.1 was set to record 10,000 events. Cells treated with epirubicin drug only (without siRNA treatment) was used to draw a gate/region to set a population or to back a population. The cells were gated on FSC/SSC for live cells with epirubicin drug intake and the median of fluorescence was measured using BL3 channel. Representative images of dot plots, density plots, and histograms are shown in Figure 2.2.

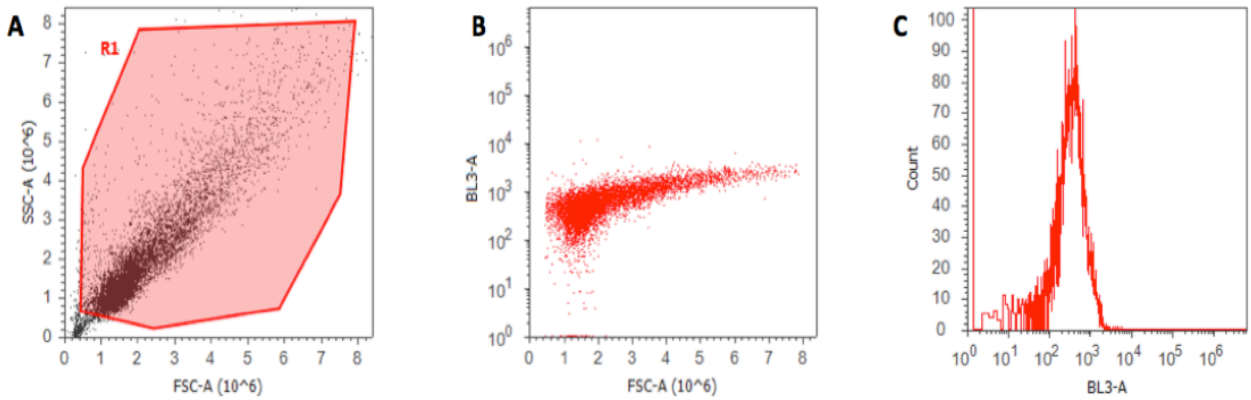


Figure 2.3 Representative images of FACS plots demonstrating the gating and strategy for quantification of epirubicin loading. (A) Representative of FSC/SCC flow cytometric dot plot of epirubicin-treated cells. R1 represents a gate for viable cells of which subsequent analyses were based. (B and C) Represents characteristics of gated cells from R1 on plots of FSC vs epirubicin fluorescence (BL3) (B), or a histogram of epirubicin fluorescence (BL3) (C).

2.9. RT-PCR

2.9.1. RNA extraction

ReliaPrep RNA cell Miniprep system (Promega; Wisconsin, USA) was used to extract total RNA from cultured cells (in 6-well plates), using the manufacturer's protocol. In brief, medium was removed from the wells and cells were washed with 1ml of ice-cold DPBS. After that, cells were harvested by pipetting up and down 10 times with 100µl of BL+TG (1-Thioglycerol) lysis buffer per well and then the lysate was transferred to sterile tubes. Next, 25µl of 100% isopropanol was added to the lysate and vortexed for 5 seconds. After that, the lysate was transferred to a Reliaprep minicolumn with collection tube and centrifuged at 13,000 x *g* for 30 seconds at room temperature. The collection tube was then discarded and the Reliaprep minicolumn was placed in a new collection tube and 500µl of RNA wash solution was added to Reliaprep minicolumn and centrifuged at 13,000 x *g* for 30 seconds at room temperature. After that, DNase I incubation mix was prepared by combining 24µl of yellow core buffer, 3µl of 0.09M $MnCl_2$ and 3µl of DNase I enzyme (30µl in total per sample tube). 30µl of the DNase I incubation mix was applied to directly to the membrane inside the column and left for incubation for 15 minutes at room temperature. After that, the column was washed with 200µl of column wash solution and twice with 500µl of RNA wash solution. Finally, RNA was eluted in 15µl of nuclease-free water, and the purified RNA was then stored at -70°C for downstream analysis.

2.9.2. Quantification of extracted total RNA

The extracted total RNA was quantified using NanoDrop 2000/2000c Spectrophotometer (ThermoFisher Scientific; Massachusetts, USA) using the manufacturer's protocol. Briefly, 1µl of nuclease-free water was pipetted onto a measurement pedestal to clean and followed by 1µl of nuclease-free water as a blank. Next, 1µl of RNA was pipetted for measurement at 280/260. The measurement is displayed in ng/µl and graph.

2.9.3. Reverse Transcription

The High-capacity cDNA reverse transcription kit (Applied Biosystems, ThermoFisher Scientific; Massachusetts, USA) was used, following the manufacturer's protocols. The following components were mixed for total of 10µl per reaction: 2ul of 10X RT buffer, 0.8ul of 25XdNTP mix (100mM), 2µl of 10X random primers, 1µl of multiScribe reverse transcriptase, 1µl of RNase inhibitor, 3.2µl of nuclease-free water. After that, 10µl of the RT mixture was transferred into PCR reactions tubes and then 10µl of RNA sample suspended in nuclease-free water (50-100ng of RNA) was pipetted into PCR reaction tube. The tubes were then centrifuged briefly and incubated as follows in a thermal cycler (see Table 2.2).

Setting	Step 1	Step 2	Step 3	Step 4
Temperature	25 °C	37 °C	85 °C	4 °C
Time	10 minutes	120 minutes	5 minutes	Infinite

Table 2.2 Thermal cycling conditions for cDNA synthesis

2.9.4. qPCR using Taqman assays

Gene expression qPCR was performed using Taqman assays for MUC17 (Assay ID: Hs00959753_s1), PCNX1 (Assay ID: Hs00900449_m1), TENM4 (Assay ID: Hs01008070_m1), ABCB1 (Assay ID: Hs00184500_m1), ABCC1 (Assay ID Hs01561483_m1), ABCG2 (Assay ID: Hs01053790_m1), RPL19 (Assay ID: Hs02338565) and ACTB (Assay ID: Hs99999903_m1) (Applied Biosystems, ThermoFisher Scientific; Massachusetts, USA). PCR reaction mixes were prepared by adding 10µl of Taqman gene expression master mix, 1µl of Taqman gene expression assay, and 9µl of cDNA diluted in nuclease-free water (50-100ng of cDNA) per reaction into 96-well PCR reaction plates. Assays were prepared in technical duplicates or triplicates. Next, the plates were vortexed and centrifuged at 400 x *g* for

5 minutes and then assayed in an Applied Biosystems real-time quantitative PCR instrument (Quant Studio 5) using the standard mode and the manufacturer's thermal cycling conditions. QuantStudio™ Design and Analysis Software (Applied Biosystems, ThermoFisher Scientific; Massachusetts, USA) was used to calculate the CTs values and the average of triplicate values was taken for each sample. Finally, comparative quantification analysis ($2^{-\Delta\Delta C}{}_{T}$ Method) was performed to calculate the fold difference in gene expression [93].

2.10. Western blots

2.10.1. Protein extraction

Total protein extraction was performed from cultured cells, typically from 6-well plates. Medium was removed from culture wells and cells were washed twice with 1ml of ice-cold DPBS. After that, 100µl of cold RIPA buffer (25mM Tris-HCl pH 7.6, 150mM NaCl, 1% NP-40, 1% sodium deoxycholate, 0.1% SDS) (ThermoFisher Scientific; Massachusetts, USA) with Halt™ Protease and Phosphatase inhibitor cocktail (ThermoFisher Scientific; Massachusetts, USA) was added to cells and harvested by pipetting up and down 10 times. Next, cells lysate was collected into sterile tubes and centrifuged at 14,000 x g for 15 minutes to pellet the cells debris. Finally, supernatants were transferred to new tubes for protein quantification or stored at -80°C for downstream analyses.

2.10.2. Protein quantification

The BCA protein assay kit (ThermoFisher Scientific; Massachusetts USA) was used to quantify extracted total protein lysates. First, a set of protein standards was prepared by serial dilution (working range 25-2,000ug/ml) using the provided albumin standard stock. The working reagent was prepared by mixing 50 parts of BCA reagent A with 1 part of BCA reagent B (50:1). Next, the microplate Thermo Scientific Pierce 96-well plate was set up and 25µl of each standard and unknown sample was pipetted

into wells in triplicate and then 200µl of working reagent was added to each well. The plate was mixed thoroughly on a plate shaker for 30 seconds and then incubated at 37°C for 30 minutes. After that, the plate was left to cool at room temperature, and the absorbance was measured at 570nm on a plate reader (Berthold Technologies; Wildbad, Germany). Finally, the analysis was performed by building the standard curve using the average blank-corrected 570nm measurement for each BSA standard versus its concentration in ug/ml and then the protein concentrations of unknown samples were determined using the standard curve.

2.10.3. Western blot analysis

20µg of total protein was to be loaded into each well of gels. Therefore 20µg samples of total protein were denatured by incubating at 70°C for 10 minutes with NuPAGE LDS sample buffer (4x), NuPAGE reducing agent (10x) and deionised water (up to 6.5µl) for a total volume of 10ul (ThermoFisher Scientific; Massachusetts USA). After denaturation, the samples were loaded onto a NuPAGE Novex (3-8%) Tris-Acetate PAGE gel (ThermoFisher Scientific; Massachusetts USA) alongside a HiMARK pre-stained protein marker (ThermoFisher Scientific; Massachusetts USA). The gel was run in an X-Cell SureLock electrophoresis tank filled with a 1 x NuPAGE Tris-Acetate SDS running buffer containing a NuPAGE antioxidant (ThermoFisher Scientific; Massachusetts USA) at a constant 150V for 60 minutes. Polyvinylidene difluoride (PVDF) membrane (ThermoFisher Scientific; Massachusetts USA) was activated by soaking in 100% methanol for 30 seconds followed by submerging in 1x NuPAGE transfer buffer (ThermoFisher Scientific; Massachusetts USA) containing 10% methanol for 5 minutes. In the X-Cell SureLock Blot module (ThermoFisher Scientific; Massachusetts USA), the SDS-PAGE gel and the activated PVDF membrane were sandwiched between blotting paper and a number of sponges soaked in 1x NuPAGE transfer buffer (ThermoFisher Scientific; Massachusetts USA) containing 10% [v/v] methanol. The module was filled with 1 x transfer buffer whilst the surrounding tank was filled with cold distilled water, and the transfer was run at a constant 30V for 60 minutes. Following transfer, the membrane was rinsed with TBS-T (Tris-buffered saline, 0.1% Tween 20), then the membrane was incubated with western blocking solution (5% [w/v] Marvel dried non-fat milk powder in TBS-T) for 1 hour at 4°C on a

rotator. The membrane was then incubated on a tube roller with the primary antibody (diluted in blocking solution) for overnight at 4°C. The antibodies dilution for PCNX1 was 1:1000 (ab220503, Abcam; USA), and 1:500 for beta tubulin (ab6046, Abcam, USA). Next, three washes with TBS-T for 20 minutes each was performed. After that, the membrane was incubated with the HRP conjugated-secondary antibody goat anti-rabbit IgG diluted in blocking solution at 1:50,000 (ab205718, Abcam, USA) for 1 hour at room temperature on a roller. After three washes of 20 minutes each in TBS-T, the membrane was then placed on an acetate sheet and incubated with Femto Super Signal West reagent (ThermoFisher Scientific; Massachusetts USA) according to the manufacturer's guidelines to develop the immune-positive bands for PCNX1, whilst SuperSignal West Pico (ThermoFisher Scientific; Massachusetts USA) was used for beta tubulin. Membranes were visualised on the ChemiDoc™ MP System with Image Lab™ software (Bio-Rad; California, USA).

2.11. Immunofluorescence (IF)

For immunofluorescence (IF) experiments, cells were seeded into 6-well plates containing glass coverslips and treated as appropriate for the experiment required (transfection for example). In order to perform the immunofluorescence assessment, medium was removed from the 6-well plate containing coverslips and cells were washed with 1ml of ice-cold DPBS. After that, 1.25ml 4% paraformaldehyde (dissolved in PBS (w/v)) was added and incubated for 15 minutes at room temperature. Next, the fixative was removed and washed with ice-cold DPBS twice and followed immediately by adding 2ml of blocking agent (0.5% [w/v] Marvel dried non-fat milk powder in TBS-T). In the meantime, primary antibody dilution for MUC17 rabbit polyclonal (ab122184, Abcam; Cambridge, UK) was prepared at 1:500 (blocking agent used as an antibody diluent). Next, the blocking agent was removed and around 100ul of diluted primary antibody was added to the coverslips (except for the no antibody negative control) and covered with parafilm and then incubated at 4°C overnight. For the no antibody negative control 0.5% skimmed milk solution was added instead. On the next day, the parafilm was removed from the coverslip and then washed with 2 ml PBS 3 times. Then, 100 ul of goat anti-rabbit IgG H&L Alexa fluor 488 fluorescent antibody

(ab122184, Abcam; Cambridge, UK) was added to all coverslips and incubated for 1 hour at room temperature in the dark. After that, coverslips were washed 3 times with PBS and then mounted with ProLong™ Gold mountant with DAPI (ThermFisher Scientific; Massachusetts, USA) by placing one drop of the mount medium on a clean glass slide and placed the coverslip with cells side on the top of the mounting medium. The coverslips were let to cure and then slides were imaged using a Zeiss fluorescent microscope (Carl Zeiss; Oberkochen, Germany).

2.12. Tissue MicroArrays (TMAs)

Two different sets of Tissue MicroArrays (TMAs) were provided by collaborators. TMAs of resection samples of 140 primary breast cancer cases that were treated with adjuvant chemotherapy were assembled by Stacey Jones (Clinical research fellow and surgical trainee, University of Leeds / Leeds Teaching Hospitals NHS Trust), comprising three independent tissue cores representing each case. Also, TMAs of resection samples from 53 primary breast cancer patients who were treated with neoadjuvant chemotherapy, therefore samples were post-chemotherapy, were obtained through collaboration with Dr Abeer Shaaban (consultant breast pathologist, formerly at Leeds Teaching Hospitals NHS Trust, where the patients included in this TMA were treated, but now at University Hospitals Birmingham NHS Foundation Trust). In brief, TMAs were constructed from the marked cores of the residual invasive carcinoma using a manual tissue microarrayer (MTA1; Beecher Instruments, USA). 2 or 3 0.6 mm core punches from the marked representative residual tumour section (central and peripheral where possible) were selected from each block and assembled into a TMA donor block with a 1.0 mm interval between cores. Of each TMA block, 3 or 5 µm sections were cut on a microtome (RM2255; Leica Wetzlar, Germany) onto Superfrost Plus slides (ThermoFisher Scientific; Massachusetts, USA) for immunohistochemistry.

2.13. ImmunoHistoChemistry (IHC)

Immunohistochemistry (IHC) was performed on TMA sections to allow examination of expression in a large cohort of tumours on only a few slides. IHC was also performed on sections of various control tissues or whole sections of single breast cancer resections. First, slides were dewaxed through xylene (3 changes, 5 minutes each) and brought through absolute ethanol (3 changes, 1 minute each), and then brought into running tap water for 5 minutes. Antigen retrieval was performed using 10mM citric acid buffer, pH 6.0 and microwave heating method as follows. 10mM citric acid buffer was prepared and the pH was adjusted using 1M NaOH. Around 750ml of the citric acid buffer was transferred into a pyrex dish and pre-warmed in a microwave for 2 minutes at high power. Then the slides were transferred to the citric buffer and heated in the microwave for 10 minutes at high power. After that, the slides were left to cool down for 20 minutes at room temperature and then transferred to running tap water for 5 minutes. Next, endogenous peroxidase was blocked by incubating the slides for 10 minutes at room temperature in 200ml methanol mixed with 2ml hydrogen peroxide (30% v/v). After that, slides were washed in running tap water for 5 minutes and followed by rinsing with Tris-Buffered Saline (TBS). Next, slides were transferred to a humidified chamber and then 100ul of antibody diluent reagent solution (Life Technologies, ThermFisher Scientific; Massachusetts, USA) was added for at least 5 minutes. After that, MUC17 antibody rabbit polycloncal (ab122184, Abcam; Cambridge, UK), 1:250 dilution or PCNX1 antibody rabbit polyclonal (ab220503, Abcam; Cambridge, UK), 1:500 dilution were added to slides to cover the entire tissue section and incubated overnight at 4°C and 2 hours at room temperature, respectively. For the no primary antibody controls, antibody diluent reagent solution was added instead. After that, slides were washed with TBS-T (TWEEN20, Sigma Aldrich; Missouri, USA) twice for 5 minutes and TBS twice for 5 minutes. Next, slides were treated with 100ul of SignalStain Boost IHC detection Reagent (HRP, Rabbit) (Cell Signaling Technology; Massachusetts, USA) for 30 minutes at room temperature. After that, the slides were washed with TBS-T and TBS exactly as in the previous step. Next, 100ul of SignalStain DAB substrate working solution (Cell Signaling Technology; Massachusetts, USA) was added for 5 minutes at room temperature. After that, slides were washed in running tap water for 5 minutes, and stained with Mayer's

Haematoxylin (counter-stain) for 1 minute, followed by washing in a running tap water for 1 minute. This was followed by incubation in by Scott's water for 1 minute and then again in running tap water for 1 minute. Next, slides were dehydrated through absolute ethanol (3 changes, 1 minute each) and through xylene (3 changes, 1 minute each). Finally, slides were mounted using coverslips with DPX (Sigma Aldrich; Missouri, USA) and then left for overnight to cure at room temperature.

2.14. Statistical analyses

Statistical analyses were performed as described in figure legends, using GraphPad Prism (version 7) for Mann Whitney tests or paired Student's T-tests or 2 ways ANNOVA. ROC curves, Spearman's rank correlation coefficient and Kaplan-Meier survival analyses were performed using IBM SSPS statistics (version 2.5). Occasional graphs were made using Microsoft Excel (version 14.6.0). Error bars represent standard error of the mean. P values of <0.05 were considered statistically significant.

3. Genomic sequencing of epithelial-enriched matched breast cancer samples taken pre- and post-neoadjuvant chemotherapy

3.1. Abstract

Many studies have shown the feasibility of performing WGS and WES on tissues in the form of FFPE samples. The aim of this chapter is to describe the process I used to obtain successful WES data from epithelial-enriched, archival FFPE samples representing matched breast cancers both pre- and post-neoadjuvant chemotherapy. This includes case selection, optimisation, and various QC steps throughout the process of exome library construction to achieve the best possible quality of sequencing data.

Initial optimisation data showed encouraging results for performing WES on FFPE breast tissue samples available to us through the hospital archive. A total 8 primary breast cancer patients were identified, with suitable clinico-pathological features, who received neoadjuvant chemotherapy, and for whom suitable tissues representing matched normal tissue and pre- and post-neoadjuvant chemotherapy tissue were available. Tissue samples from these patients were processed in two batches for WES library preparation. This involved isolation of epithelial cancer cells from the cancer samples by laser micro-dissection, extraction of genomic DNA, and library preparation itself. The quality control metrics showed variable library quality ranging from low to good. 6 out of 8 patients were taken forward for final sequencing. Samples were sequenced either once or multiple times using the HiSeq3000 Illumina platform. The sequencing data from multiple sequencings of the same library were merged when required.

The average read depth across the captured regions of the exome after duplicate removal was 20-127x. Also, the percentage of mapped reads to the target regions was of 83.4-87.0%. However, a high percentage of duplicate reads of between 35.6-92.4% was obtained. Overall, this work illustrates obstacles associated with preparing FFPE samples for WES, nevertheless, I have successfully obtained sequencing data from my target samples.

3.2. Introduction

Next generation sequencing (NGS) has been hugely helpful in cancer research to detect various patterns of mutations such as single nucleotide variants (SNV), short deletions or insertions (Indel), and copy number variations (CNV) in cancer from individual patients. The costs, in terms of both time and money, associated with this process are falling rapidly, and compare very favourably with Sanger sequencing (first generation sequencing), which was used to create the first draft of the human genome (3.2 billion bp), over a period of almost 10 years in the 1990s at vast expense [94].

Implementing NGS in the clinical setting has been challenging in terms of obtaining suitable samples. Whilst there are tissue banks that store fresh frozen tumour materials as a research resource, these are typically limited in scope, size and/or the amount of tissue banked. Therefore, archival collections of tumours fixed in formalin and embedded in wax blocks, stored in diagnostic pathology departments remain the principle repository of material available to study the molecular pathology of breast cancer [95, 96]. These archival resources represent the main realistic resource able to cover the heterogeneity of breast cancer and facilitate studies to relate their molecular profile to response to therapeutic regimes. The utilisation of formalin fixed paraffin embedded (FFPE) tissues is the most practical in the clinical laboratory setting as most histopathology departments conserve archival tissue in this manner [97, 98].

However, it is known that FFPE tissues are prone to many artefactual damages in their DNA as a consequence of sometimes-prolonged formalin exposure prior to processing. These include deamination of cytosines and DNA double-strands cross-linkage of cytosine nucleotides on either strand [95, 99, 100]. There is some predictability to these changes and such DNA alterations can be addressed bioinformatically to some degree, and as such are recognisable during the data analysis since it is likely that the damage caused by the formalin fixation is distributed evenly across the genomic reads. In addition, there are many commercial genomic DNA extractions kits available that claim using their specially optimised lysis buffer and incubation at an elevated temperature after proteinase K digestion partially

removes the formalin crosslinking of the released DNA, improving yields, as well as DNA performance in the downstream NGS assays [101]. Another challenge associated with FFPE tissue is the limited amount of DNA that can be obtained particularly in small tissue samples such as core biopsies originally taken in a diagnostic setting. However notwithstanding these concerns it is worthwhile to note that input amounts of DNA as low as to 5-500ng can still yield good library preparation results sufficient for sequencing [96, 98, 102].

In this chapter, I aimed to establish protocols to isolate tumour cells from FFPE breast tumour samples and extract sufficient masses of suitable quality DNA for whole exome sequencing (WES). I then aimed to deploy these protocols to achieve WES data representing matched normal tissue, pre-neoadjuvant chemotherapy tissue, and post-neoadjuvant chemotherapy tissue from a small cohort of primary breast cancer patients in order to allow down-stream analysis of genomic changes associated with chemotherapy treatment.

3.3. Results

3.3.1. Case selection

Breast cancer cases to be included in this project were identified using database searches of the Leeds Teaching Hospitals patient management system ("PPM"), and physical review of tissue slides and blocks. This search was led by two independent histopathologists (and supervisors for my PhD): Dr. Eldo Verghese (consultant histopathologist at Leeds Teaching Hospitals NHS Trust) and Professor Andrew Hanby (Professor of Breast Pathology at University of Leeds and Leeds Teaching Hospitals NHS Trust). The number of cases screened was 150; including patients diagnosed with primary breast cancer at Leeds Teaching Hospitals NHS Trust between the years 2010 and 2015.

The inclusion criteria were as the following:

1. A diagnosis of ER positive and HER2 negative breast cancer, as defined by clinical immuno-histochemical assessment of expression of estrogen receptor (positive) and immuno-histochemical / fluorescent in situ hybridisation assessment of her2 status (negative).
2. Treatment with neoadjvant chemotherapy (NAC) using the combination of drugs epirubicin/cyclophosamide (EC), *without* taxanes.
3. Presence of residual tumour cells post-NAC in the resection samples (indicating partial resistance to NAC).
4. Sufficient cells present in in the pre-NAC core biopsies and post-NAC resections.

Patients meeting these criteria were rare, with the need for a partial response (so sufficient tumour cells remained in the post-NAC sample) twinned with a lack of use of taxanes proving particularly limiting, since many patients showing a relatively poor response by MRI during initial cycles with epirubicin/cyclophosphamide were switched to taxanes. Furthermore, during the course of this study, use of NAC without taxanes became increasing rare. Nevertheless, it was felt that standardising the molecular subtype and the chemotherapy regimen was necessary in order to reduce likely variation in chemoresistance pathways.

Ultimately 8 patients for further examination were identified. The table below shows the patients' details for 6 of these 8 individuals, including only the 6 that were successfully sequenced in the study.

Patient identifier	Age at diagnosis	Histo-pathological diagnosis	Tumour size	Grade	Nodal metastasis	ER, PR, Her2 receptor expressions	pCR socre if available
1	46	Ductal NST	31mm	3	Yes	ER positive PR positive Her2 negative	NA
3	48	Ductal NST	20mm	2	No	ER positive PR negative Her2 Negative	NA
5	50	Ductal NST	115mm	3	Yes	ER positive PR positive Her2 Negative	NA
6	41	Ductal NST	55mm	2	Yes	ER positive PR positive Her2 Negative	NA
7	52	Ductal NST	30mm	2	No	ER positive PR positive Her2 negative	NA
8	51	Mixed ductal & lobular Carcinoma	40mm	1	No	ER positive PR negative Her2 negative	2.941/RCB-II

Table 3.1 Patient, pathological and clinical details of patients included in the sequencing study. NST= no special type, NA= not available, RCB= residual cancer burden.

3.3.2. Optimisation

A key component of my proposed experiments was to extract sufficient genomic DNA for NGS analyses from breast cancer samples using commercial column tubes extraction kit (QIAamp MinElute Columns DNA kit, Qiagen; Dusseldorf, Germany . Ideally, DNA should be from the epithelial (cancer) compartment only, and would also

be from very small breast cancer biopsies in the case of pre-NAC samples. Therefore, I needed to optimise DNA extraction protocols to maximise the amount of extracted DNA to facilitate the chances of a successful analysis. The samples used for optimisation work were breast tissues. As a consequence, my initial task was to estimate how much genomic DNA could be extracted from our FFPE samples. The initial variables that I assessed were the thickness of the sections (5 micron sections versus 10 micron sections), the areas of dissection (5 mm^2 area versus 10 mm^2 area) and the incubation period for tissue to be digested in proteinase K (1 hour, overnight and 72 hours). I concluded that the most effective combination was 5 micron sections, with as large a tissue area as possible, and overnight proteinase K digestion.

In order to enrich for epithelial cells I performed micro-dissection. I assessed various ways of doing this using a Zeiss/P.A.L.M. machine (P.A.L.M. Microlaser Technologies, Zeiss; Oberkochen, Germany) LCM in order to purify only epithelial tumour cells and avoid any other cellular elements. These were: 1) laser capture micro- dissecting epithelial cells (i.e. the cells of interest) directly, or, 2) laser capture micro-dissecting unwanted cellular elements and then manually dissecting/scrapping the tumour epithelial cells from the slides. Efficient laser capture during LCM requires use of membrane slides. However, when manually dissecting tissue from a slide that has had the unwanted elements removed by LCM, this membrane becomes a potential contaminant. Hence it was necessary to assess the effects of membrane slides on the yield of DNA extraction. I concluded that both direct LCM of epithelial cells, and LCM-based removal of non-epithelial components produced satisfactory results, with no evidence that membrane from the slides resulted in lower yields. Therefore, the LCM strategy I used was varied according to the density of epithelial tumour cells within the individual sample, with diffuse epithelial cells isolated directed by LCM, while denser epithelial cells would be collected after LCM removal of other components (see representative pictures of both these LCM approaches in section 2.3). These findings provided insights to go ahead with actual cases for the project.

3.3.3. The first batch of cases (patients 1 – 4)

3.3.3.1. Pre-NAC samples

My initial expectation was that the pre-NAC biopsies would provide the most challenging tissue from which to extract sufficient DNA. Therefore, my strategy was to purify epithelial cells from these biopsies, to extract DNA from these, and to prepare these pre-capture libraries. I would proceed with extractions from post-NAC resections and normal tissue only for cases for which the pre-NAC pre-capture libraries were successful.

I selected 4 cases at random from those available. Biopsies were sectioned and then tumour cells were purified by laser capture micro-dissection. DNA was extracted as per the optimised protocol and quantified in order to assess whether there was sufficient genomic DNA to proceed to library preparation (Table 3.2), the target being greater than 200ng as an absolute minimum. Pre-capture libraries were prepared for all the 4 samples and their fragment size distributions assessed (Table 3.3).

Case number	1	2	3	4
DNA yield (µg)	6.5	3.7	2.2	1.9

Table 3.2 DNA yields from epithelial enriched samples of pre-NAC core biopsies from the first 4 cases.

Case number	Peak DNA Fragment size (bp)	Calibrated concentration (ng/µl)
1	203	6.29
2	219	5.62
3	259	6.62
4	169	0.44

Table 3.3 Quality metrics for pre-capture libraries made from DNA from epithelial enriched samples of pre-NAC core biopsies from the first 4 cases.

The quantification of size distribution and overall amount of pre-capture libraries demonstrated notable differences between individual samples. Case 4, in particular, produced a very low yield library of fragments well below the target size range (optimal DNA distribution around 250bp). Therefore, I decided not to proceed further with this case. Of the other 3 cases, I decided to proceed with the 2 cases showing the highest library concentrations: cases 1 and 3 with 6.29ng/µl and 6.62ng/µl, respectively. I then proceeded with processing of the matched resection (post-NAC) and normal tissues for these cases.

3.3.3.2. Post-NAC and normal samples

Since case 1 and case 3 yielded satisfactory pre-capture libraries, their matched tumour resection samples and normal breast tissue samples were obtained. Tumour cells were purified by laser capture micro-dissecting tumour cells directly. Normal tissue samples were prepared by manually dissecting non-tumour tissue. DNA was extracted as per the optimised protocol and then quantified (Table 3.4), and pre-capture libraries prepared. As before, fragment sizes and overall DNA quantity for the pre-capture libraries was assessed (Table 3.5).

Case No.	**Case 1 Tumour**	**Case 1 Normal**	**Case 3 Tumour**	**Case 3 Normal**
Amount of DNA extracted in µg	3.5	1.7	2.4	12.25

Table 3.4 DNA yields from epithelial enriched samples of post-NAC resection and normal tissues for cases 1 and 3.

Case No.	Peak DNA Fragment size (bp)	Calibrated conc. (ng/µl)
Case 1 Tumour	197	5.65
Case 1 Normal	191	6.05
Case 3 Tumour	202	9.04
Case 3 Normal	195	16.7

Table 3.5 Quality metrics for pre-capture libraries for post-NAC resection and normal tissues for cases 1 and 3.

Although DNA fragments sizes for these libraries were smaller than optimal (~250bp), the concentrations were adequate therefore these were deemed to be suitable for further processing. The key result of this section is that I had prepared complete sets of pre-capture libraries for 2 cases, comprising tumour DNA pre-NAC, tumour DNA post-NAC, and matched normal DNA, ready for hybridisation and capture steps.

3.3.3.3. Indexed exon library preparation for cases 1 and 3

The 6 samples representing cases 1 and 3 were subjected to exon capture and indexed libraries were prepared. Library metrics were assessed as previously (Table 3.6) and were deemed to be adequate for sequencing. These 6 samples were pooled and sequenced in a single lane of the HiSeq 3000.

Case No.	Peak DNA fragments size (bp)	Calibrated conc. (ng/ µl)
Case 1 Pre-NAC	263	0.218
Case 1 Post-NAC	244	0.289
Case 1 Normal	244	0.167
Case 3 Pre-NAC	302	0.759
Case 3 Post-NAC	262	0.564
Case 3 Nomal	236	0.152

Table 3.6 Metrics for the indexed exon libraries from the matched triplet samples (pre-NAC, post-NAC, normal) for cases 1 and 3.

3.3.4. The second batch of cases (cases 5 to 8)

3.3.4.1. Pre-NAC samples

As for cases 1 to 4 (section 3.3.3), I initially examined pre-NAC core biopsy samples from my next 4 cases (cases 5 to 8). Tumour DNA were obtained using LCM and DNA extraction. DNA quantification is shown in Table 3.7. Although DNA yields were low, pre-capture libraries were prepared from all samples as it was not felt any samples could be excluded since more suitable cases were not available. Metrics for the pre-capture libraries are shown in Table 3.8. Case 8 showed an excellent yield of library with the optimal size peak, even though this had the lowest DNA input, however all the other cases showed relatively poor yields sometimes with substantially lower peak sizes than is ideal (notably case 7). Nevertheless, as alternative cases were not available, I proceeded with DNA extraction and preparation of libraries from the matched post-NAC and normal samples.

Case No.	Case 5	Case 6	Case 7	Case 8
Amount of DNA extracted in µg	0.63	0.38	0.36	0.23

Table 3.7 DNA yields from epithelial enriched samples of pre-NAC core biopsies from cases 5-8

Case No.	Peak DNA Fragment size (bp)	Calibrated conc. (ng/µl)
Case 5	223	0.695
Case 6	213	0.648
Case 7	195	0.624
Case 8	250	8.59

Table 3.8 Quality metrics for pre-capture libraries for post-NAC resection and normal tissues for cases 5-8.

3.3.4.2. Post-NAC and normal samples

Next, I extracted DNA from the matched post-NAC and normal samples for cases 5-8 as previously. The metrics for DNA quantification are shown in Table 3.9. I also prepared pre-capture libraries from these samples.

Case No.	Case 5 Tumour	Case 5 Normal	Case 6 Tumour	Case 6 Normal	Case 7 Tumour	Case 7 Normal	Case 8 Tumour	Case 8 Normal
Amount of DNA extracted in µg	3.78	2.94	0.290	3.04	0.610	1.12	0.390	2.02

Table 3.9 Table 3.9 DNA yields from epithelial enriched samples of post-NAC resection and normal tissues for cases 5 -8.

The initial quantification results following adaptor-ligated library preparation showed very poor yields. This was most likely due to the low initial input DNA, although other possible causes include loss of DNA during the shearing or washing steps. In order to increase yields, samples that initially gave the best yields (3.5-7.0 ng/µl) in total, which is still well below target) underwent a further 10 PCR amplification cycles, while samples that yielded even less underwent a further 13 PCR amplification cycles. However, it should be noted that additional PCR amplification increases the likelihood of PCR duplicates, which can potentially lead to reduction in detection of low frequency allele variants. To overcome this issue, I also prepared separate (second) libraries from the remaining input genomic DNA and then combined these with the previous libraries in order to increase the overall yield of pre-capture libraries. Although, this is not recommended by the manufacturer (Agilent technologies, USA), this had been done before in our facility with no detrimental effects on sequencing quality since same batch of reagents used to prepare the libraries. The combined pre-capture libraries were assessed again (Table 3.10). The results demonstrated highly variable size distributions and very variable DNA yields (ranging from 0.33 to 18.4ng/µl).

Case No.	**Peak DNA Fragment size (bp)**	**Calibrated conc. (ng/µl)**
Case 5 Post-NAC tumour	322	8.11
Case 5 Normal	332	18.4
Case 6 Post-NAC tumour	208	0.458
Case 6 Normal	330	7.46
Case 7 Post-NAC tumour	229	0.673
Case 7 Normal	259	1.78
Case 8 Post-NAC tumour	234	1.01
Case 8 Normal	219	0.325

Table 3.10 Quality metrics for combined pre-capture libraries for post-NAC resection and normal tissues for cases 5-8.

The key result of this section is that I had prepared complete sets of pre-capture libraries for cases 5 to 8, comprising tumour DNA pre-NAC, tumour DNA post-NAC, and matched normal DNA, although doubts remained about their quality for capture and sequencing.

3.3.4.3. Indexed exon library preparation for cases 5 to 8

Indexed exon libraries were prepared as previously and were assessed for the DNA fragments size distribution and concentration (Table 3.11). Concentrations were, with some exceptions, disappointingly low, and one sample was even undetectable (normal DNA for case 6) however I decided to proceed with sequencing all the samples in a single lane, with the intention that further additional lanes could be used if required using only the samples that had produced some successful data.

Case No.	**DNA Fragments size (bp)**	**Calibrated conc. (pmol/l)**
Case 5 Pre-NAC	247	257
Case 5 Post-NAC	255	14.8
Case 5 Normal	247	55.5
Case 6 Pre-NAC	241	230
Case 6 Post-NAC	232	538
Case 7 Pre-NAC	244	428
Case 7 Post-NAC	244	121
Case 7 Normal	243	144
Case 8 Pre-NAC	246	154
Case 8 Post-NAC	245	469
Case 8 Normal	242	341

Table 3.11 Metrics for the indexed exon libraries from the matched triplet samples (pre-NAC, post-NAC, normal) for cases 5-8. Note, case 6 normal sample was not quantifiable.

Unfortunately, the first attempt at sequencing failed due to a technical failure with High HiSeq 3000 Illumina platform. This led to substantial loss of samples and major delays with progress in the project. The second sequencing attempt was run exactly like the first attempt and it was successful, however, the depth of coverage was relatively low for tumour samples (ranging from 10 to 20x) and this could have resulted in unreliable conclusions from down-stream analyses. Therefore, the samples were sequenced once again and were run across two lanes. The FASTQ files from all successful runs were merged together, and this merging process was checked (see appendix 9.1).

3.3.5. Preliminary sequencing data quality control for cases 1, 3 and 5 to 8

The basic sequencing data quality within the output sequencing files, FASTQ files, were assessed using fastQC software and SureCall softwares. Files examined were the merged files, representing cases 5 to 8, or the single file, representing cases 1 and 3). Table 3.12 shows the basic quality metrics for the 6 patients who had complete sequencing data.

In summary, the merging of FASTQ files from multiple runs was successful as indicated by the total number of sequenced reads in the final merged files which corresponds to the added up total number of sequenced reads in files from first and second runs (see appendix 9.1). The quality metrics run by SureCall software showed that there were relatively high duplicate reads (all samples, except sample case 3 pre-NAC, showed above 50% of total sequenced reads were duplicates). This was partly expected due to high numbers of PCR amplification cycles during the WES library construction. Overall, all samples showed above 80% of its reads had mapped in the target regions of the genome.

Sample	Number of reads in fastq files	Duplicate reads	% of duplicate reads	Number of mapped reads in target regions	% of mapped reads	Bases on target	Mean coverage depth
1 pre	121,014,120	77,937,803	64.44%	33,307,317	83.47%	3,143,944,751	62
1 post	111,408,940	63,517,656	57.04%	38,259,419	85.58%	3,464,256,889	68
3 pre	108,054,316	38,420,225	35.58%	56,989,727	86.34%	6,436,742,015	127
3 post	116,053,836	71,663,034	61.78%	35,458,873	84.97%	3,377,940,777	67
5 pre	29,901,066	16,969,253	89.36%	9,881,199	83.42%	1,024,928,826	20
5 post	10,257,650	3,320,499	83.08%	5,540,546	85.66%	573,993,928	27
6 pre	27,261,968	16,818,092	91.56%	8,098,524	84.39%	817,842,912	30
6 post	32,165,380	24,916,043	92.38%	5,626,674	84.53%	590,740,977	32
7 pre	14,554,710	6,809,547	91.77%	6,072,702	84.56%	635,625,314	33
7 post	31,240,420	21,610,184	89.26%	7,784,227	87.03%	791,064,172	31
8 pre	16,678,481	16,678,481	87.80%	14,244,321	84.09%	1,567,446,469	41
8 post	31,035,560	18,943,570	89.33%	9,496,927	84.99%	1,012,449,896	49

Table 3.12 Quality metrics for WES for patients 1, 3 and 5 to 8 from SureCall software analysis. Duplicate reads are assumed to be the result of reading 2 or more PCR copies of the same original DNA fragment (i.e. a read). Percentage of mapped reads represent the number of reads which mapped to the reference genome by the total number of all sequenced reads that passed the mapping quality filters. The measurement of on-target bases is represented as the ratio of number of bases within a target region to total number of bases output by the sequencer, expressed as a percentage. Coverage depth represents the number of times a sequenced DNA fragment maps to a genomic target. Note – normal DNA metrics have been excluded from this table since these data from SureCall were no longer available at the time of writing; however, these metrics are summarised along with the tumour samples in a subsequent analysis in Table 4.2.

3.4. Discussion

3.4.1. The choice of ER positive and HER2 negative breast cancer treated with epirubicin/ cyclophosphamide NAC

The main criteria used to select breast cancer patients that were potentially suitable for this project was to include patients with ER positive and HER2 negative breast cancers and with tumours that showed partial resistance to the common combination of epirubicin/cyclophosphamide neoadjuvant chemotherapy regimen. The reasons for these criteria were that ER positive and HRE2 negative receptors profile is common in breast cancer patients, so that would expand the chances of finding more cases to include in the study [36, 37]. Therefore, a reasonable proportion of potential cases should have post-treatment cancer cells remaining for analysis.

In addition, I confining my patient cohort to those who received only the epirubicin/cyclophosphamide regimen. Confining my cohort to a single chemotherapy regimen was aimed to give a major advantage to address genomic aberrations relating to response to this chemotherapy specifically, based on the hypothesis that different chemotherapeutics may well have different pathways of resistance. A good example of this approach is illustrated by a study looking at the expression of let-7 miRNA from a cohort of 70 patients before NAC. All received anthracycline-based neoadjuvant chemotherapy only. It was found that lower let-7a expression was associated with epirubicin resistance in primary breast tumours. Moreover, upregulation of let-7a expression sensitized resistant breast tumour cell lines to epirubicin, by enhancing cellular apoptosis *in vitro*. Hence, the conclusion was that let-7a may be used as a therapeutic target to modulate epirubicin-based chemotherapy resistance [103].

In a contrary, a study looked at the expression level of P-glycoprotein (Pgp), Multidrug Resistance-associated Protein 1 (MRP1), and Breast Cancer Resistance Protein (BCRP), which are known to predict chemotherapy responses. Their cohort composed of pre and post-NAC samples of 45 patients where the NAC regimes were

anthracyclines with or without taxanes. Interestingly, it was found that the expression level of BCRP was a survival marker after NAC, however, this maker was not specific to anthracyclines either with or without taxanes [40].

However, a substantial disadvantage of confining the patient cohort to a single combination of NAC regimen was limiting the chance of availability of suitable cases to include in the study, since patients who appeared to be responding poorly to the initial epirubicin/cyclophosphamide regimen were very frequently switched to taxane-based regimens, making them ineligible for the study.

3.4.2. The decision to use LCM to enrich for epithelial cancer cells

The optimisation work provided helpful insights into the amount of genomic DNA that can be obtained from varying tissue thickness and areas of tissue. As it was vital to have sufficient genomic extraction for WES analysis, therefore I optimised both the DNA extractions protocol and the LCM technique. Use of LCM was a key decision as it was hoped it would provide the major advantage of having very good representation of the epithelial component and reduce the chances of sequencing unwanted cellular elements like lymphocytes and fibroblasts – thereby maximising the read depth of somatic (cancer) mutations.

Some previous studies have used LCM in combination with Next Generation Sequencing (NGS) to uncover tumour heterogeneity in FFPE samples in terms of mutant allele frequency and gene transcript expression [104-106], although such studies remain in a small minority in the field of cancer genomics. In addition, micro-dissection techniques have been used to isolate specific target cell types from within the pool of cancer cells, including cytokeratin AE1/AE3, p53, or estrogen receptor (ER) positive cells and nuclei from tissue sections with a mixed population of cells where the targets constituted only 5% of the sample. Target enrichment from this admixed cell population prior to NGS produced a minimum of 13-fold increase in mutation allele frequency detection, which reflects the robustness of this technique in detection of somatic mutations with low allele frequency [107]. Single cell-sequencing using RNA-

sequencing represents a similar, although even more highly developed, technology, based on the basic idea that analysis of pools of cells with different phenotypes must limit understanding of the biology, while such approaches have allowed insights into the fact that different phenotypically identical cells may dramatically vary with respect to transcriptomic landscape [108].

However, implementing LCM in this project was a major technical challenge in terms of the time-required for thorough isolation of tumour cells, and also led to further reductions in the amount of extracted DNA presenting technical challenges in terms of library preparation. Nevertheless, I succeeded in achieving sequencing data, and expect that the LCM greatly enriched the sequencing depth of somatic variants.

3.4.3. Use of quality control metrics during the WES protocol and for final sequence analysis

I applied different quality check steps starting from quantifying the initial input of genomic DNA (gDNA), and at may subsequent steps through to checking the quality of the final sequencing data. As a rule of thumb, the more starting gDNA material for WES the more chances for successful quality sequencing data. The recommendation for the WES enrichment reagents and protocol I used (SureSelect for Illumina Paired End Sequencing) was for a starting template mass of 3ug of high quality DNA [109]. However, I did not expect to achieve this – particularly as high quality DNA is not available when working with FFPE material. Thereby, I used a modified protocol that used fewer concentrating and clean-up steps to minimise the loss of DNA, and included the multiple check steps to assess pre-capture and post-capture library quality using the TapeStation system. Most of my samples showed library DNA fragments size distributions to peak within or close to the recommended range of the enrichment system, of 225 to 275bp after shearing DNA and adaptor ligation library, and 250 to 350bp after hybridisation and exome capture steps. Utilising the libraries QC data were helpful to decide whether to proceed with sequencing or to rectify the poor libraries for example; perform further amplification or start another library for samples that showed deficiencies in quantity or fragment size. However, due to the

fact I had a limited number of cases that fit into the study's inclusion criteria, and limited amount of material in the samples, especially in the core biopsies, the library quality was consistently a major concern and the sequencing data quality may be compromised as a result.

The quality metrics for the sequencing data showed an average eventual read depth range of between 20x and 127x, which was improved partly due to the multiple sequencing runs that were performed for samples in the second batch. However, the number duplicate reads was relatively high for most of samples (77% overall; range; 35.6-924%), an issue most likely associated with the increased number of PCR cycles used during libraries preparation. I obtained a relatively good percentage of mapped reads to the target region (i.e. the exome) in both batches of samples (83.42% targets with at least 20x coverage) which are comparable with other studies for percentage of mapped reads on target regions (for example 62.8-81.1% [95], 91-96% [110], and 98.1-98.9% [111]).

There are many studies that have suggested the feasibility of performing WES NGS on FFPE samples, [96, 98, 102, 112] and in a particular study has demonstrated that performing WES on a paired FFPE and fresh frozen (FF) tissues showed 70-80% concordance of variants detected in both types of samples stored for fewer than three years [112], which suggest NGS can be used to study FFPE archived samples. Also, another study performed successful sequencing using as low as 10ng of DNA from FFPE tissue and obtained similar quality data from this sequencing as from frozen tissue [113]. However, from the experience I have gained I conclude it is possible to sequence DNA from FFPE samples of low quantity; however, the quality of sequencing data depends heavily on the quantity and quality of initial input DNA.

3.4.4. Conclusion

The use of WES represents an extremely challenging application of this technology, since I have used FFPE samples that have been fixed and stored for many years in a clinical archive, I have started with very small samples (biopsies in some cases), and have micro-dissected specific cell types from within these samples. Despite these challenges, I have produced sequencing data with basic analysis metrics that suggest the data may be suitable for down-stream analysis. My next experimental aims were to proceed with a complete data analysis pipeline and then potentially generate a list of candidate genes that may impact on chemotherapy responses.

4. Extensive mutational differences exist between matched pre- and post-NAC samples, allowing identification of potential mediators of therapy response

4.1. Abstract

Whole exome sequencing data from trio samples (pre-NAC cancer cells, post-NAC cancer cells, and normal cells) from 6 primary breast cancer patients were obtained in the previous chapter (Chapter 3). Data were now analysed pairwise (either pre-NAC or post-NAC cancer vs. normal) to identify somatic mutations in the cancer cells, particularly single nucleotide variants and small insertions or deletions, initially in house using SureCall software (Agilent Technologies), but subsequently by an independent party using open-source bioinformatics tools (Edinburgh Genomics Laboratory). Once identified correctly, somatic mutations within matched pre-NAC and post-NAC samples were compared to determine the influence of NAC on mutation prevalence. Mutated genes were also compared between cancer cases to examine any genes or pathways influenced by NAC in the cohort overall.

Data analysis revealed substantial mutational loads in both pre-NAC samples (mean number of somatic SNVs was 398) and post-NAC samples (mean number of somatic SNVs was 112). Overlap between mutational loads in matched pre-NAC and post-NAC samples was surprisingly low (mean number of somatic SNVs was 34). Different strategies for filtering and prioritising of somatic variants were then implemented in order to generate a list of candidate genes showing evidence for these mutations influencing the response of cells to chemotherapy. In addition, functional enrichment analysis was utilised to highlight molecular pathways or biological functions that were significantly over-represented in the lists of candidate gene, to indicate any mechanisms with a role in response to NAC. Finally, a list of 46 priority candidate genes was generated, which – reassuringly - contained some genes known to be associated with chemotherapy response, for example TP53.

This list of candidate genes was suitable for further functional investigation using *in vitro* approaches in the next chapter (Chapter 5).

4.2. Introduction

Improved availability of next generation sequencing (NGS) technologies along with substantial decreases in sequencing costs have led to NGS becoming a favourable approach to apply in medicine, especially in cancer research. In addition, it is increasingly used in clinical practice for cancer diagnosis and treatment [114]. Whole exome sequencing (WES) is a popular targeted sequencing method in clinical and cancer research. Although, the exome makes up only around 1.5% of the genome, it contains around 85% of the known disease related variants [115], which makes WES a cost-effective alternative to whole genome sequencing as it requires less sequencing to achieve a required depth of coverage yet retains a strong likelihood of identifying the genomic aberrations of interest.

However, there are many challenging tasks associated with NGS practice, in particular with data management and processing, i.e. the need for advanced IT infrastructure and programming specialists to perform data analysis using the most appropriate choices among the available computational methods and analysis tools. Also, there is a need for knowledgeable specialist to interpret sequencing data into meaningful and useable information. Increasingly, it is these analysis steps that constitute substantial bottle-necks for improved understanding of the cancer genome, rather than the availability of samples or sequencing hardware [116]. Nevertheless, several programmes are available to simplify the bioinformatics analysis so that front-line “wet-lab” researchers who are not experts in bioinformatics can independently perform both the laboratory experiments and the down-stream computer-based analysis. For example, Agilent Technologies has launched a free bioinformatics tool called SureCall that is intended to provide analysis capabilities for researchers to transform raw NGS data into insightful analyses, without the need for bioinformatics training or advanced infrastructure [117].

In additions, there are several alternative commercial bioinformatics programs such as Avadis NGS (Strand Scientific Intelligence), CLC Genomics Workbench (CICbio, Qiagen), and CondonCode Aligner (CondonCode) that are said to be powerful and user-friendly bioinformatics packages for these types of analyses [118]. Moreover, there is a web-based platform for data analysis, named Galaxy, which incorporates popular open-source and community Linux command line tools into an easy to use web-based environment, thereby providing a major advantage for wet-lab biologists who are inexperienced with Unix/Linux systemBodi [119]. However, these software programmes are either expensive and/or their use is poorly represented with the literature meaning that there may be a need to validate findings from these packages by another independent means of data analysis.

On the other hand, utilising publicly available bioinformatics tools seems to be the standard and popular way of analysing WES data, as shown by the literature [120, 121]. There are a variety of programmes that have been validated to perform various bioinformatics analysis tasks and each tool has pros and cons depending on the user's specific tasks and research questions to be answered [120]. Substantial bioinformatics expertise is usually required to make appropriate choices and use of many of these programmes.

NGS has been a useful tool to investigate intra-tumour heterogeneity (ITH) because with sufficient sequencing depth it is possible to identify even minor sub-clones with genomic alterations, and such sub-clones could be related to increased tumour aggressiveness or therapy resistance. Implications of ITH in terms of responsiveness or resistance to different chemotherapeutic regimen and metastatic progression must be investigated in order to provide clinically relevant information for cancer patients. Studies have indicated that ITH influences the responsiveness to chemotherapeutic regimen: findings from ovarian, cervical, and tongue cancers suggested that specific sub-clones of tumour cells present before adjuvant chemotherapy have survived and expanded after chemotherapy [122-124].

In this chapter, I have shown the ITH findings for my matched breast cancer samples (pre-NAC and post-NAC) from both the SureCall software and data analysis using open-source bioinformatics tools. I present the issues I encountered with data analysis using the SureCall software, and the reasons for deciding to proceed with the findings from the open-source bioinformatics pipeline in order to generate my final list of genes that are candidate mediators of chemotherapy response.

4.3. Results

4.3.1. SureCall bioinformatics analyses of somatic variants in breast cancer samples

SureCall (Agilent) is a desktop application combining algorithms for end-to-end NGS data analysis from alignment to annotations of mutations. Since I used Agilent products for exome library construction, I aimed to use this software from the same company to perform the analysis of my exome sequencing data from my matched trios of samples (normal genome, pre-NAC cancer genome and post-NAC cancer genome) from 6 breast cancer patients. I initially used SureCall software to perform basic quality control analysis and these findings have already been presented in Chapter 3, section 3.3.5. After that, I performed pairwise analyses for each patient (pre-NAC versus normal, and post-NAC versus normal) to identify somatic Single Nucleotide Variants (SNV) and small insertion or deletions (indels). Note that in this chapter the patients within the sequencing study have been renumbered with sequential identifiers (i.e. 1-6), with patient 3 in Chapter 3 now designated patient 2, and patients 5-8 from Chapter 3 now designated patients 3-6.

The numbers of detected variants for each cancer sample was unexpectedly high, ranging from 31,758 variants (patient 2 post-NAC) to 49,634 (patient 2 pre-NAC), with a very substantial overlap between pre-NAC and post-NAC for each patient. Also the type of variants detected were all SNPs, with no such small insertion or deletion variants which all of these led to questioning the reliability of Surecall software data.

I performed different filtering trials on the data for patients 1 and 2 to assess the impact of changing some analysis variables on the number of somatic variants with non-synonymous effects identified, and those found to be present in both pre-NAC and post-NAC. Patients 1 and 2 were chosen as they appeared to have the highest quality data. I investigated different read depth cut-offs (>10x, low read depth; or >40x, relatively higher read depth) and whether 'overlap' between variants that are shared between pre-NAC and post-NAC samples should be regarded as requiring the same variant ID in the gene (i.e. literally exactly the same variant) or whether different variants in the same gene should be counted as 'overlap', regardless of variant ID. The purpose of the trials was to explore the size of the list of potential candidate genes based on the number of variants shared between pre-NAC and post-NAC and between patients, as those variants within the shared pool that consistently change their mutant allele frequency (MAF) after chemotherapy would be strong candidates as chemo-response mediators. The findings are summarised in Table 4.1, along with an assessment of the pros and cons of the different results.

Based on these trial findings, the final filtering strategies were set to exclude variants sequenced outside coding region of the genome (i.e. intergenic regions and introns), and read depths of <10x. In terms of overlap between pre-NAC and post-NAC variants, I required only variants within the same gene. In addition, I added further filtering criteria. I utilised a Phred quality score cut-off of 30; a Phred quality score of 30 assigned to a base means there is a probability of 1 in 1000 that the sequenced base is a sequencing error. Finally, I utilised the predictor tools SIFT (Sorting Intolerant from Tolerant) and Polyphen2 (a tool to predict how conservative or damaging the variant in the encoded protein is likely to be) to exclude likely synonymous and benign mutations as follows; SIFT: >0.05 was excluded (scores above 0.05 are predicted to be benign), and Polyphen2: benign was excluded. The numbers of variants selected using these parameters, and the overlaps between paired pre-NAC and post-NAC samples, for all six patients are represented in Venn diagrams in Figure 4.1.

Strategy	Number of variants detected shared between pre & post-NAC and shared across patients 1 and 2	Pros	Cons
Filtering at >40x read depth with exact variants ID	15	Short list, meaning experimentally tractable to investigate all further High confidence of calling variants	Missing variants with potentially relevant functions Generally detected low change in MAF between pre-NAC and post-NAC
Filtering at >40x read depth with same mutated genes, regardless of variants ID	18	Short list, meaning experimentally tractable to investigate all further High confidence of calling variants	Missing variants with potentially relevant functions Cannot calculate change in MAF for genes included with two different mutations
Filtering at >10x read depth with exact variants ID	303	Larger number of genes with known functions that appeared potentially relevant More variants with MAF change detected	Long list, meaning further prioritization required before next step of experimental testing
Filtering at >10x read depth with same mutated genes, regardless of variant ID	352	Larger number of genes with known functions that appeared potentially relevant	Even longer list Cannot calculate change in MAF for genes included with two different mutations

Table 4.1 Filtering strategies employed within SureCall to prioritise variants for further study and assessment of their impact. Numbers of variants identified in cancer samples from patients 1 and 2 were manipulated by varying analysis parameters. The impact of these variables was assessed by calculating the number of genes identified as having shared variants between pre-NAC and post-NAC samples and pros and cons of each filtering strategy are described. MAF: mutant allele frequency.

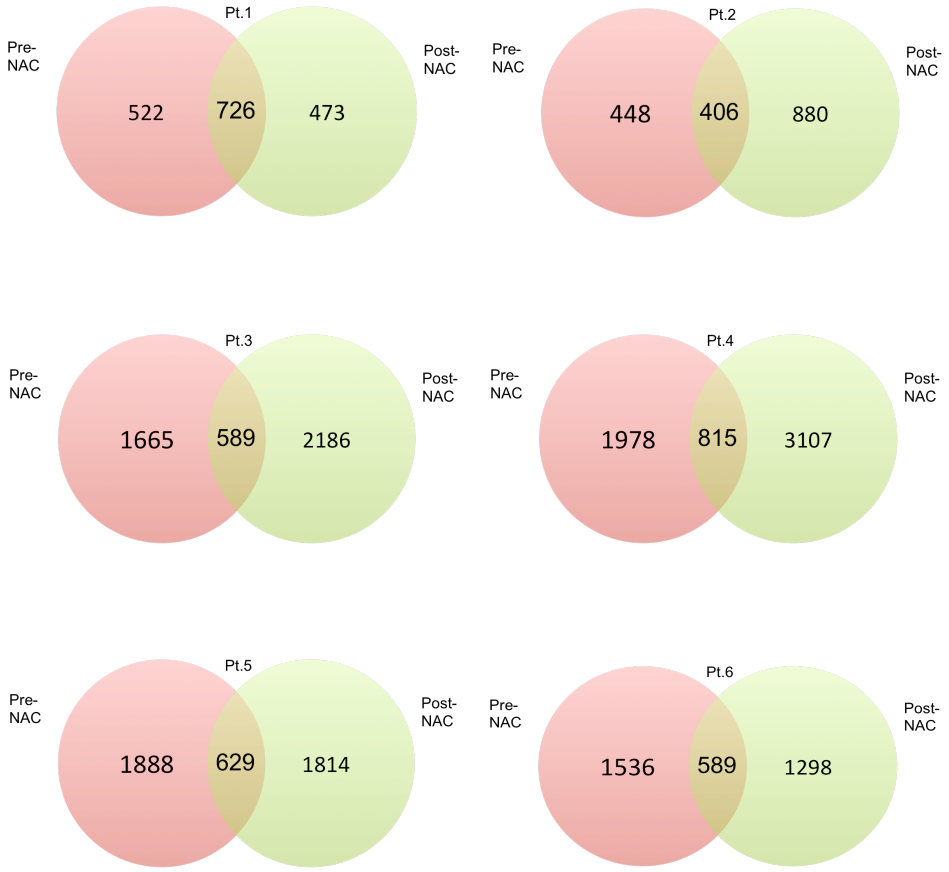


Figure 4.1 Venn diagrams to illustrate the number of somatic variants with non-synonymous effects and their overlaps in pre-NAC and post-NAC samples for patients 1-6 from SureCall analysis. Somatic variants were identified by comparison with genomes extracted from matched normal tissue, and were filtered on the basis of read depth and Phred quality score cut-offs, coding region only, and variants within same gene (but potentially different mutations) for the overlap between pre-NAC and post-NAC. Note; it was observed the poorer quality samples from the second batch patients have much higher numbers of variant calls than patients' samples 1 and 2 with SureCall analysis data, but not with the open source Edinburgh Genomics data. This provides further evidence to question SureCall data analysis reliability.

Because the number of detected variants after filtering was relatively high, the candidate genes were prioritised as following: individual variants that were shared between pre-NAC and post-NAC for each patient were selected, and then these shared mutations were filtered further for shared variants across all 6 patients requiring the same variant ID in the gene, and therefore it was possible to assess the change in

MAF after NAC. This resulted in 37 genes with somatic variants in all 6 patients that were common to pre-NAC and post-NAC samples and showed an increase in MAF after therapy suggesting that the mutations were associated with relative resistance: CACNA1S, CDCP2, CELA3B, COL16A1, CR1, CR1L, CROCC, E2F2, ERMAP, FAM178B, FAM179A, FAM63A, FBLIM1, FRMD4B, FSIP2, GBP6, HMCN1, HSPG2, IGFN1, IGSF3, MRPL9, NOTCH2NL, OR11L1, PDE4DIP, PLEK, PTGER3, RNF223, SLC35E2, TEKT4, TNN, TRIM58, UBXN11, AJAP1, DNAH14, SPEG, OBSL1, HLA-DRB1.

4.3.2. Open-source bioinformatics tools analysis findings

The findings from SureCall software were interesting, however, because of lack of intensive literature and published studies that have used this platform, and because I had some concerns about the very large number of somatic variants identified, I decided to validate the findings from SureCall by an independent party using open-source bioinformatics tools. The independent party analysis was done by the Edinburgh Genomics laboratory (UK). The overall mapping quality metrics resulting from the analysis by the Edinburgh Genomic laboratory are shown in Table 4.2. The metrics relating to coverage of the target regions do not include reads marked as PCR duplicates. Overall, the quality metrics using the open-source bioinformatics pipeline (refer to method chapter 2, section 2.9 for exact details of the pipeline) are in agreement with the quality metrics generated from the SureCall software (refer to chapter 3, section 3.3.5).

As with the SureCall analyses, some basic filtering criteria were determined to identify somatic aberrations within the cancer samples that were robust and could be considered for further analysis. Criteria were to exclude germline variants, by comparison with normal tissues, mutations with read depths of less than 5 or more than 800, and variants with a quality Phred score of less than 30. Furthermore, it should be noted that the MuTect2 tool from GATK (version 3.7) identifies variants present in the tumour and not in the matched normal sample, and a potential source of error is when a variant could be present in the normal sample, but missed in the

sequencing. Therefore, in this analysis potential somatic variants were excluded if they were identified in *any* of the other normal samples in this cohort, thereby reducing the potential errors associated with missing germline variant in matched tissues [125]. The numbers of detected variants, the types of aberration they represent (i.e. SNV or indel), and the overlaps in somatic mutations between pre-NAC and post-NAC are presented in Table 4.3 and Figure 4.2. These analyses identified on an individual patient basis genes carrying variants that were: a) unique to the pre-NAC sample (i.e. had apparently been lost or made undetectable after NAC treatment), b) unique to the post-NAC samples (i.e. had apparently been selected for by NAC and had therefore become detectable (or had been generated by NAC treatment itself), or c) common to both samples (allowing assessment of relative enrichment or depletion after treatment in terms of change in MAF). The mean number of somatic SNVs was 398 for pre-NAC, while was 122 for post-NAC, revealing substantial differences in mutational loads between pre-NAC and post-NAC (statistically significant: p=0.03; paired Student's T test). Also, the overlap between mutational loads in matched pre-NAC and post-NAC samples was surprisingly low, with the mean number of somatic SNVs in both samples being 34 (range 3-141). In addition, among the gene mutations shared between pre- and post-NAC pooled from all 6 patients, 125/209 showed increases in their MAF (60%), while 84/209 (40%) showed decreases.

Metric	**Range from 18 samples (normal, pre-NAC, post-NAC from 6 patients)**
% of mapped reads	79-98%
Duplication rate	35-92%
Bases on target	69-77%
Mean read depth	18-102
Median insert size	94-152bp
Standard deviation of insert sizes	20-52bp

Table 4.2 Overall read mapping metrics for the exome sequencing data from trios of samples from 6 breast cancer patients, as assessed using open-source bioinformatics tools by Edinburgh Genomic laboratory.

	All variants				Somatic (not detected in any normal)				Germline (detected in one or more normal samples)			
Sample ID	**SNP**	**INS**	**DEL**	**ALL**	**SNP**	**INS**	**DEL**	**ALL**	**SNP**	**INS**	**DEL**	**ALL**
Pt.1 post-NAC	80	6	7	93	36	3	6	45	44	3	1	48
Pt. 1 pre-NAC	174	9	14	197	68	7	5	80	106	2	9	117
Pt. 2 post-NAC	58	6	8	72	43	4	6	53	15	2	2	19
Pt. 2 pre-NAC	2585	53	81	2719	1355	25	54	1434	1230	28	27	1285
Pt. 3 post-NAC	385	60	20	465	112	54	14	180	273	6	6	285
Pt. 3 pre-NAC	228	80	91	399	124	76	87	287	104	4	4	112
Pt. 4 post-NAC	439	47	47	533	238	42	44	324	201	5	3	209
Pt. 4 pre-NAC	401	67	89	557	339	62	85	376	172	5	4	181
Pt. 5 post-NAC	931	38	48	1017	137	26	33	196	794	12	15	821
Pt. 5 pre-NAC	952	154	83	1189	125	135	70	330	827	19	13	859
Pt. 6 post-NAC	133	28	26	187	42	21	24	87	91	7	2	100
Pt. 6 pre-NAC	102	36	25	163	38	33	23	94	64	3	2	69

Table 4.3 Number of different types of genomic variants in each tumour sample. Variants were identified in trios of samples (normal, pre-NAC, post-NAC) from 6 breast cancer patients and were filtered based on read depth >5 and <800, and Phred score >=30. Somatic variants were determined by comparison between cancer samples and their matched normals, and then with the pooled variants across all 6 normals.

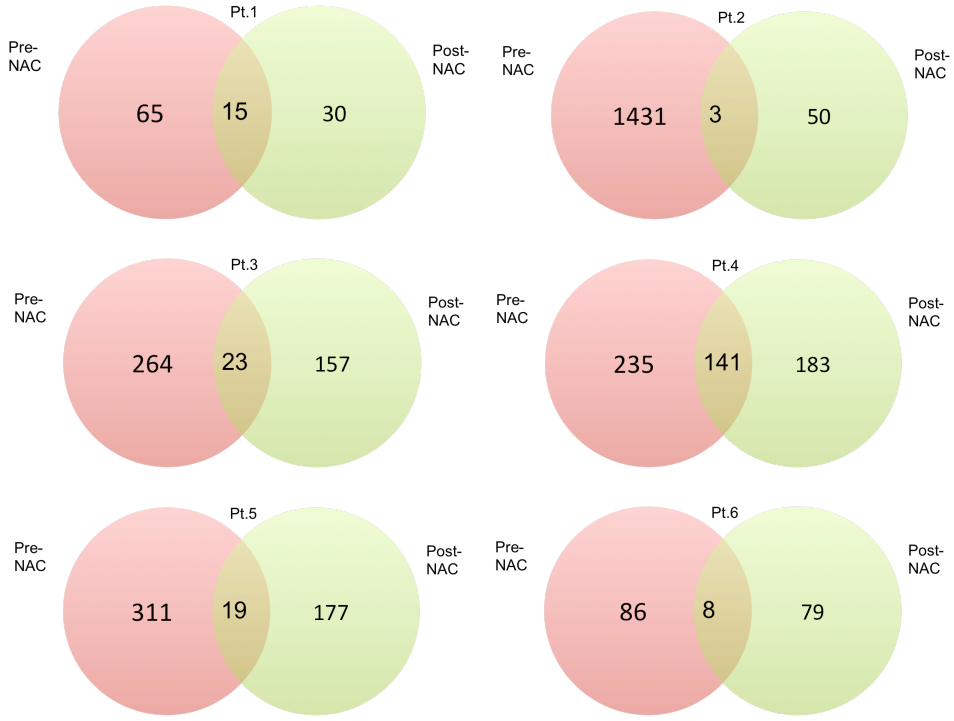


Figure 4.2 Venn diagrams to illustrate the number of somatic variants and the overlaps in paired pre-NAC and post-NAC samples for breast cancer patients 1-6 using open-source bioinformatics tools (Edinburgh Genomic laboratory, UK).

4.3.3. Comparison between SureCall and open-source analyses: troubleshooting SureCall

There was a substantial difference between the findings from SureCall software and the open-source bioinformatics analysis tools in terms of number of detected variants, and also the type of variants (all SNPs for Surecall; indels were identified using open-source analyses). Also, the results from SureCall software analysis in terms of number and type of variants was compared with different studies and found there is a high discordance with the published studies using FFPE samples (refer to discussion section 4.4.1), thereby I had to troubleshoot the SureCall software utilising the a genomic viewer. It was found that in many cases “somatic” variants in the cancer samples in fact had multiple reads aligned with high mapping quality in the matched normal samples, yet had not been called as germline variants. An example of this is

shown in Appendix 9.2. Extensive troubleshooting was performed along with the Agilent bioinformatics support team, from the initial position of Agilent denying that there was a potential problem with their software. However, after some months, Agilent acknowledged that there was a fault in the SureCall software to recognise mutations present in both tumour and reference normal samples as germline mutations; instead, they were identified (incorrectly) as somatic variants, seemingly because the variants within the normal sequence had been (incorrectly) defined as sequencing errors. The Agilent bioinformatics support team then committed to rectify this fault in their upcoming software version 4.0, although the timeline they expected for implementing this fix was many months and was therefore not useful for me to move my work forward. Considering the SureCall findings appeared to be highly unreliable, I therefore decided to proceed only with the findings from the open-source bioinformatics analyses in order to generate my final list of candidate genes for potential roles in chemo-response.

4.3.4. Functional enrichment analysis findings

Variants that were unique to either the pre-NAC sample or the post-NAC sample had potentially been selected for or against by NAC, and were therefore potential candidate mediators of chemotherapy response. However, this interpretation is susceptible to false positives for reasons including differential representation of tumour cells within the samples due to tissue sampling (see discussion section 4.4.2). Therefore, I was interested to assess whether the list of genes hosting these variants was significantly enriched for any molecular functions, as a method of increasing confidence in individual candidates. The genes were sub-categorised into 2 groups: genes with variants unique to pre-NAC and genes with variants unique to post-NAC. I assessed potential shared molecular functions within these gene lists by performing functional enrichment analysis on each list separately. I used three different enrichment analysis tools: Database for Annotation, Visualization, and Integrated Discovery (DAVID) [126], WebGestalt (GSAT) [127], and ToppGene [122]. The findings from ToppGene enrichment analysis are shown in Table 4.4, while further details of the other analyses such as statistics, scores and number of genes enriched in each term or pathway are included in Appendix Table 9.3. Various molecular

pathways were significantly over-represented among the mutated genes in each category. Of particular note was significant overrepresentation of extra-cellular matrix (ECM) molecules, collagen protein coding genes, and integrin signalling molecules, potentially hinting at roles for these pathways in chemo-response. Among the genes in the gene sets involved in the above pathways were: Mitogen-activated protein kinase genes such as MAP3K4, MAPK10, and phosphatidylinositol-4-phosphate 3-Kinase PIK3C2A. In addition, collagens such as COL9A1, COL1A1, and COL6A3, integrins alpha chains such as ITGA5, ITGA7, and ITGA9 and laminins (protein of the ECM) such as LAMA5, LAMB1, and LAMC1.

Enriched pathways for genes with variants that were unique to the pre-NAC samples	**p-value**	**Gene count in the query list**	**Number of genes defined as members of the pathway**
Genes encoding collagen proteins	2.38E-11	22	44
Collagen chain trimerization	1.24E-10	22	47
Collagen biosynthesis and modifying enzymes	1.34E-09	26	70
Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	1.34E-08	59	275
Collagen formation	1.60E-08	29	93
Extracellular matrix organization	2.66E-07	59	298
Assembly of collagen fibrils and other multimeric structures	3.67E-06	19	60

Integrin signalling pathway	7.98E-06	36	167
Genes encoding structural components of basement membranes	1.94E-05	14	40
Focal adhesion	3.52E-05	39	199
Diseases associated with O-glycosylation of proteins	5.92E-05	17	60
Diseases of glycosylation	9.42E-05	21	86
Rho GTPase cycle	9.86E-05	30	145
Degradation of the extracellular matrix	1.04E-04	25	112
Rap1 signaling pathway	1.20E-04	39	210
Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	1.40E-04	139	1028
NCAM1 interactions	1.76E-04	12	37
Protein digestion and absorption	1.88E-04	21	90
HDR through Single Strand Annealing (SSA)	2.34E-04	12	38
Nitric oxide stimulates guanylate cyclase	3.35E-04	9	24

Glutamatergic synapse	3.61E-04	24	114
Enriched pathways for genes with variants that were unique to the post-NAC samples			
Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	3.11E-05	25	275
Type II diabetes mellitus	3.87E-05	9	46
Integrin signaling pathway	4.52E-05	18	167
Genes encoding structural components of basement membranes	8.86E-05	8	40

Table 4.4 Functional enrichment analysis using ToppGene for gene lists of variants found uniquely in either pre-NAC or post-NAC breast cancer samples. Pathways are in ranked by statistical significance of over-representation.

4.3.5. Further prioritisation of somatic variants: generation of a final prioritised list of genes of interest

Next, I aimed to prioritise the candidate genes from the open-source bioinformatics findings for potential further investigation using *in vitro* studies. I applied different analysis strategies in order to prioritise candidate genes taking into consideration the frequency of mutated genes among the 6 patients, the category in which the candidate was found (i.e. variants unique to pre-NAC samples, variants unique to post-NAC samples, or variants shared between pre-NAC and post-NAC), and the extent of change in MAF for variants shared between pre-NAC and post-NAC (Table 4.5).

Analysis Type	Number of mutations detected
Genes mutated in 2 or more patients and unique to pre-NAC samples	101 genes
Genes mutated in 2 or more patients and unique to post-NAC samples	17 genes
Genes mutated in 2 or more patients with >5% change in MAF between pre-NAC and post-NAC	None
Genes with variants with >5% increase in MAF (pre-NAC to post-NAC) in 1 patient and also unique to post-NAC in another patient	1 gene
Genes with variants with >5% decrease in MAF (pre-NAC to post-NAC) in 1 patient and also unique to pre-NAC in another patient	13 genes
Total	132

Table 4.5 Different analyses of variant distributions resulting in different lists of prioritised genes for further analysis. Different gene lists are represented simply by the number of genes.

It was important to reduce the list of 132 genes in total in Table 4.5 further, since this number of genes was experimentally intractable for the planned down-stream *in vitro* investigations. Therefore, in addition, I used findings from the SnpEff annotation of the likely effects of the mutations on genes (such as amino acid changes) and the protein damaging predictor tools SIFT and Polyphen2 in order to prioritise mutations with the best chance of being functionally relevant. Furthermore, I took into consideration whether the mutated genes belonged to the enriched pathways in the functional enrichment analysis findings (section 4.3.4). The steps taken to reach the final list of candidate genes for functional analysis is illustrated in a flow chart Figure 4.3. This process of filtering and prioritising resulted in a list of the following 46 candidate genes: ABL1, AP3B1, ARAP2, CCDC88C, CENPF, CEP350, COL6A3, CRIPAK, DMBT1, EFEMP1, EGFLAM, FRYL, IGSF10, ITGA7, MUC17, MYO10, NCOA3, NLRC5, NOTCH2, PARP4, PKD2L1,

PTPN14, RPTN, S100PBP, SEZ6, SYNE1, TENM4, THADA, TP53, TPTE, TTN, ZBTB49, ZFHX4, CACNA1C, EMILIN3, FLG2, IKBKAP, PCNX1, PDGFD, ZNF853, SSPO, XDH, CEP295, KIAA1161, P2RX4, PKD2L1. Details of the variants found in each of these genes, the distributions of the variants across the samples from the 6 patients, the functional impact predictions of these variants, and whether the genes were represented in the functional enrichment analysis are presented in Tables 4.6 (mutations found only in pre-NAC samples), 4.7 (mutations found only in post-NAC samples) and 4.8 (mutations found in matched samples in one patient and in either pre- or post-NAC in a second patient).

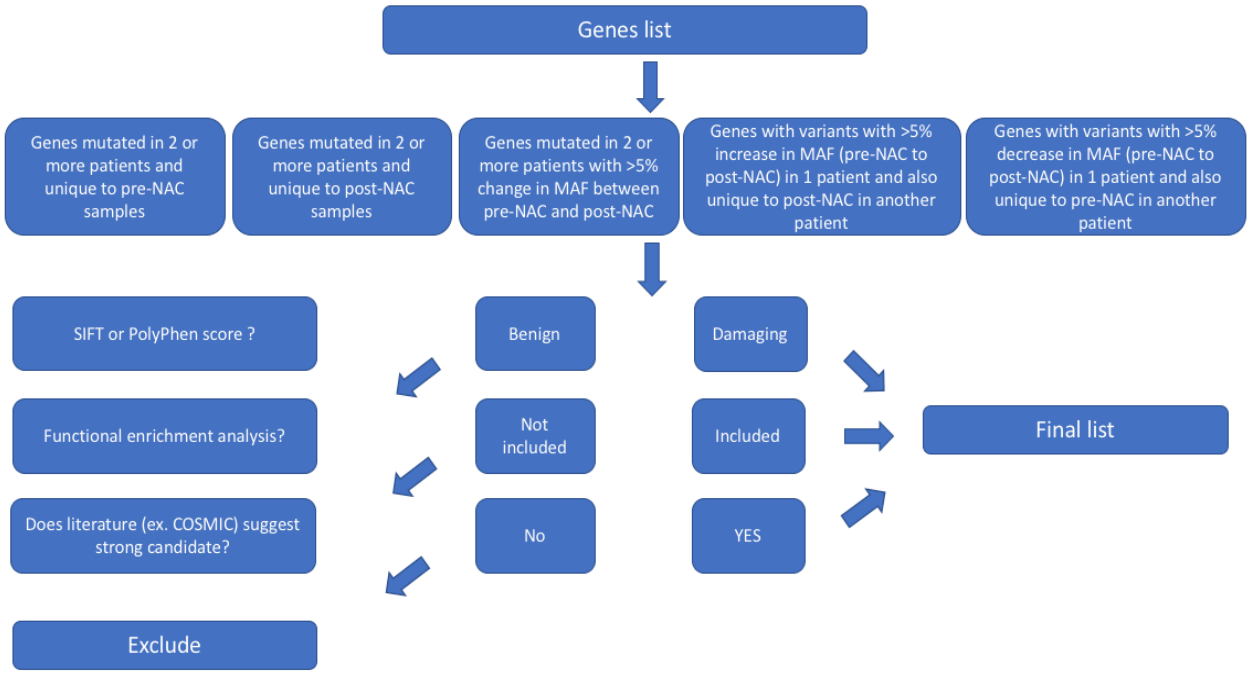


Figure 4.3 A flow chart illustrates the prioritising steps taken to generate a final list of candidate genes for functional validation using *in vitro* approaches.

GENE	Pt. ID	POS	REF	ALT	AF	DP	VARTYPE	EFFECT	IMPACT	Polyphen score	SIFT score
ABL1	2	130884719	C	T	0.629	55	SNP	Missense	Moderate	D	0.003
	5	130880070	TGG	T	0.125	30	DEL	frameshift	High	NA	NA
AP3B1	2	78294548	T	C	0.571	66	SNP	Missense	Moderate	P	0.596
	6	78216087	C	A	0.111	53	SNP	Missense	Moderate	B	0.021
ARAP2	2	36187535	T	C	0.65	40	SNP	Missense	Moderate	B	0.284
	5	36147673	T	C	0.3	60	SNP	Missense	Moderate	D	0.076
	6	36121335	G	A	0.133	29	SNP	Intron	MODIFIER	NA	NA
CCDC88C	2	91273625	A	G	0.739	118	SNP	Missense	Moderate	D	0.001
	3	91313936	CCCT	C	0.154	25	DEL	Deletion	Moderate	NA	NA
CENPF	2	214645205	C	T	0.524	371	SNP	Missense	Moderate	P	0.07
	5	214620705	C	CACTTACCTACTGTCACCCTCCAGGAGCT	0.068	83	INS	Frameshift	High	NA	NA
CEP350	1	180041193	C	T	0.034	290	SNP	Missense	Moderate	B	0.058
	4	180093972	C	ATAGAAGCCTCAGTTAATAGAAG	0.13	32	DEL	Frameshift	High	NA	NA
COL6A3	2	237365895	G	A	0.645	99	SNP	Missense	Moderate	D	0.003
	3	237372259	GTGCGAACGTACTGGAACTCAGGCCCGGCACTTTGGGACCCATCGATGAGAAAGACCACGTCCCTCTTGCCACCAACACC	G	0.033	132	DEL	Deletion	High	NA	NA
CRIPAK	3	1395743	T	TCTCTGGTCCCTGTTG	0.052	111	INS	Insertion	Moderate	NA	NA
	5	1395532	C	T	0.023	284	SNP	Missense	Moderate	D	0.059
DMBT1	2	122617175	T	C	0.679	142	SNP	Intron	Modifier	NA	NA
	3	122643175	CCCGCTTCCGGTTCAGGGCCTTCCACTTCCTGAA	C	0.019	298	DEL	Deletion	Moderate	NA	NA
	5	122585178	T	TGGTGGTGTGTGTGTGTGTG	0.133	29	INS	Intron	Modifier	NA	NA
EFEMP1	2	55876730	CATTCATTTATATCTGAAAAAAGTTTTAT	C	0.137	130	DEL	Deletion	High	NA	NA

			ATATAT ATAT								
	4	55870866	A	C	0.08	94	SNP	Missense	Moderate	D	0.005
EGFLAM	2	38406958	C	T	0.589	121	SNP	Missense	Moderate	P	0.006
	5	38305475	A	AGACC CTTTTT TTTTTT TTTTTT	0.129	62	INS	Frameshift	High	NA	NA
FRYL	2	48523013	TGGTAC TCCTGG AG	T	0.2	74	DEL	Frameshift	High	NA	NA
	4	48515114	T	TG	0.133	25	INS	Frameshift	High	NA	NA
IGSF10	2	151437365	G	C	0.231	336	SNP	Missense	Moderate	P	0.17
	3	151448270	G	A	0.382	65	SNP	Missense	Moderate	B	0.163
	5	151447249	C	T	0.06	95	SNP	Missense	Moderate	B	0.389
ITGA7	1	55701099	C	T	0.07	85	SNP	Missense	Moderate	P	0.048
	2	55688241	G	C	0.371	160	SNP	Missense	Moderate	NA	NA
MUC17	2	101042842	C	T	0.184	165	SNP	Missense	Moderate	D	0.092
	4	101033319	T	C	0.052	187	SNP	Missense	Moderate	B	0.802
	6	101038440	C	A	0.034	165	SNP	Missense	Moderate	B	1
MYO10	2	16681487	C	G	0.1	80	SNP	Missense	Moderate	D	0.011
	3	16689924	A	C	0.308	51	SNP	Intron	Low	NA	NA
NCOA3	2	47627680	C	T	0.181	356	SNP	Missense	Moderate	D	0.177
	4	47627738	AG	A	0.133	27	DEL	Frameshift	High	NA	NA
NLRC5	3	57025443	GGCAG CCCCA CGCCTT CCACCA GGTCTA TGTCCC TCCAAT CCTGC GCCGG GCCAC A	G	0.121	43	DEL	Frameshift	High	NA	NA
	4	57025887	CTA	C	0.125	32	DEL	Frameshift	High	NA	NA
NOTCH2	2	119996678	G	A	0.182	286	SNP	Stop-gained	High	NA	NA
	4	119916650	TTC	T	0.143	40	DEL	Frameshift	High	NA	NA
PARP4	2	24442624	G	A	0.121	55	SNP	Missense	Moderate	B	0.002
	5	24478255	G	T	0.167	35	SNP	Missense	Moderate	B	0.647
	6	24493683	G	A	0.214	25	SNP	Synonymous	Low	NA	NA
PKD2L1	2	100296133	C	T	0.451	191	SNP	Missense	Moderate	D	0.01
	6	100293359	G	GGCAA AATAGC TTTCTC TGCCAA AGCTTT CTGACC TTTGGC TCCATC TT	0.143	25	INS	Insertion	Moderate	NA	NA
PTPN14	2	214402898	T	C	0.486	57	SNP	Missense	Moderate	D	0.081
	5	214376293	AGG	A	0.118	32	DEL	Frameshift	High	NA	NA
RPTN	1	152154830	G	A	0.03	188	SNP	Stop_gained	High	NA	NA

	5	152156069	C	CATAGTGGGAACTCTGACCTTGTCTGTCTGGCTGACT	0.049	114	INS	Insertion	Moderate	NA	NA
S100PBP	2	32826426	G	C	0.119	95	SNP	Missense	Moderate	D	0.065
	4	32828030	G	GAA	0.143	26	INS	Frameshift	High	NA	NA
SEZ6	1	28957425	A	G	0.077	99	SNP	Missense	Moderate	B	1
	5	28979734	G	GACCAGATAGATTCACAGCCGAATTCTA	0.133	28	INS	Stop_gained	High	NA	NA
SYNE1	2	152325144	G	C	0.597	99	SNP	Synonymous	Low	NA	NA
	3	152319013	GATCT	G	0.167	23	DEL	Frameshift	High	NA	NA
	4	152399673	G	A	0.136	43	SNP	Stop_gained	High	NA	NA
TENM4	4	79064967	TAC	T	0.118	29	DEL	Frameshift	High	NA	NA
	5	78669399	C	CCTGGATCTAACGCTTTCTTTTTTTTTTT	0.12	44	INS	Frameshift	High	NA	NA
THADA	2	43485261	G	T	0.404	88	SNP	Missense	Moderate	B	0.008
	5	43574929	G	A	0.15	40	SNP	Missense	Moderate	D	0.007
TP53	4	7674945	G	T	0.571	14	SNP	Stop_gained	High	NA	NA
	5	7674220	C	A	0.211	37	SNP	Stop_gained	High	D	0.005
TPTE	2	10542449	G	A	0.14	208	SNP	Intron	High	NA	NA
	3	10569776	G	A	0.115	167	SNP	Synonymous	Modifier	NA	NA
	5	10605546	C	T	0.057	175	SNP	Synonymous	Low	NA	NA
TTN	2	178586573	G	T	0.107	142	SNP	Missense	Moderate	D	0.001
	4	178616876	T	C	0.158	37	SNP	Missense	Moderate	B	1
ZBTB49	2	4303022	G	A	0.257	57	SNP	Missense	Moderate	B	0.524
	3	4302031	TC	T	0.4	8	DEL	Frameshift	High	NA	NA
ZFHX4	2	76856262	G	A	0.299	230	SNP	Missense	Moderate	D	0.139
	5	76854316	C	CGGCCATTTTTTTTTTTTTT	0.073	81	INS	Insertion	Moderate	NA	NA

Table 4.6 Candidate genes for roles in defining chemotherapy response, identified from mutations found only in the pre-NAC samples. Genes are listed with the details of the mutations. Details include: distributions of the variants across the 6 patients, allele frequency (AF) which refers to the number of reads with mutations by the number of total of mapped reads, read depth (DP), variant type (VARTYPE), effect, mutation impact from SnpEff, Polyphen scores (B=Benign, P=Pathology, D=Damaging, NA=Not Applicable), SIFT scores (<0.05=deleterious, >0.05=Benign).

GENE	Pt. No.	POS	REF	ALT	AF	DP	VARTYPE	EFFECT	IMPACT	Polyphen2	SIFT_score
CACNA1C	2	2457610	G	A	0.057	138	SNP	Missense	Moderate	P	0.086
	4	2610680	C	CACAC ACACA CACAC ACACA CACAC ACACA CACA	0.024	241	INS	Insertion	Moderate	NA	NA
EMILIN3	4	41361974	C	T	0.282	78	SNP	NA	Moderate	B	1
	5	41361737	G	C	0.125	79	SNP	NA	Moderate	D	0.116
FLG2	3	152353023	C	T	0.065	114	SNP	Missense	Moderate	B	0.558
	5	152353444	C	T	0.018	319	SNP	Missense	Moderate	P	0.052
	6	152354489	T	A	0.022	332	SNP	Synonymous	Low	NA	NA
IKBKAP	3	108900446	C	T	0.4	19	SNP	Intron	Modifier	NA	NA
	4	108912333	G	GCCTA ACCAC CCCCC CC	0.118	32	INS	Frameshift	High	NA	NA
PCNX1	4	70978026	CACAG G	C	0.118	30	DEL	Frameshift	High	NA	NA
	6	70978204	GAT	G	0.136	45	DEL	Frameshift	High	NA	NA
PDGFD	2	104000171	G	T	0.167	34	SNP	Missense	Moderate	D	0
	6	103947667	T	C	0.294	33	SNP	Missense	Moderate	B	0.521
ZNF853	2	6621625	C	G	0.5	47	SNP	Missense	Moderate	D	0
	4	6621152	A	ACAAG CGGGC CTATGT CAGCG GCTGG TCCAC CTGCC ATCCTG CTGCTT ATGT	0.167	23	INS	Insertion	Moderate	NA	NA

Table 4.7 Candidate genes for roles in defining chemotherapy response, identified from mutations found only in the post-NAC samples. Genes are listed with the details of the mutations. Details include: distributions of the variants across the 6 patients, allele frequency (AF) which refers to the number of reads with mutations by the number of total of mapped reads, read depth (DP), variant type (VARTYPE), effect, mutation impact from SnpEff, Polyphen scores (B=Benign, P=Pathology, D=Damaging, NA=Not Applicable), SIFT scores (<0.05=deleterious, >0.05=Benign).

Mutated genes that show a >5% INCREASE in MAF between pre- and post in one patient and also present in the unique to post group in another patient											
GENE	**Pt. No.**	**POS**	**REF**	**ALT**	**AF**	**DP**	**VARTYPE**	**EFFECT**	**IMPACT**	**Polyphen2**	**SIFT_score**
SSPO	5	149789294	C	G	0.25	37	SNP	Missense	Moderate	D	NA
	3	149785168	G	A	0.155	140	SNP	Missense	Moderate	P	NA
Mutated genes that show a >5% DECREASE in MAF between pre- and post in one patient and also present in the unique to pre group in another patient											
GENE	**Pt. No.**	**POS**	**REF**	**ALT**	**AF**	**DP**	**VARTYPE**	**EFFECT**	**IMPACT**	**Polyphen2**	**SIFT_score**
XDH	2	31370399	T	C	0.539	133	SNP	Missense	High	B	0.534
	4	31388277	C	T	0.313	32	SNP	Missense	Moderate	B	0.165
CEP295	2	93684123	C	T	0.275	78	SNP	Missense	Moderate	B	0.245
	3	93698405	A	G	0.5	16	SNP	Missense	Moderate	B	1
KIAA1161	2	34372933	T	A	0.472	91	SNP	Missense	Moderate	B	NA
	4	34372875	G	C	0.455	44	SNP	Stop_gained	High	NA	NA
P2RX4	2	121228843	A	G	0.762	193	SNP	Missense	Moderate	B	0.013
	3	121222196	G	A	0.455	21	SNP	Synonymous	Modifier	NA	NA
	4	121221881	C	T	0.333	30	SNP	Synonymous	Modifier	NA	NA
PKHD1	2	52056905	A	G	0.669	265	SNP	NA	Moderate	NA	NA
	4	52046107	T	C	0.6	56	SNP	Missense	Moderate	B	0.11

Table 4.8 Candidate genes for roles in defining chemotherapy response, identified from mutations shown to change in MAF in one patient, while being unique to pre- or post-NAC in another. Genes are listed with the details of the mutations. Details include: distributions of the variants across the 6 patients, allele frequency (AF) which refers to the number of reads with mutations by the number of total of mapped reads, read depth (DP), variant type (VARTYPE), effect, mutation impact from SnpEff, Polyphen scores (B=Benign, P=Pathology, D=Damaging, NA=Not Applicable), SIFT scores (<0.05=deleterious, >0.05=Benign).

4.4. Discussion

4.4.1. Whole exome sequencing data analysis issues

I attempted to take advantage of the bioinformatics software SureCall from Agilent (Agilent Technologies, USA), which should allow researchers lacking specific expertise in bioinformatics to perform data analysis along with data quality assessment. The findings from SureCall were interesting, however, due to an unprecedentedly high number of detected mutations and a relative lack of published studies in the literature using this software, I decided to validate the findings from SureCall by an independent party using open-source bioinformatics tools.

The number of somatic variants detected using SureCall software and open-source bioinformatics tools were very different, with the open-source bioinformatics analysis findings more consistent with other published studies. For examples, the number of somatic SNVs detected on 4 FFPE gastrointestinal stromal tumours (GIST) samples were ranging between 86-766 [110], while WES on 21 fresh-frozen tumour samples from pheochromocytomas and paragangliomas allowed detection of 518-1432 somatic variants per tumour sample (filtering out low confidence somatic variants that showed MAF<0.1% or read depth <20) [128]. In comparison, my detected somatic variants from the open-source bioinformatics analysis numbered between 45 and 1434, which is very similar to above studies. After troubleshooting, I came to the realisation that there is a specific fault in the SureCall software, which has resulted in false positive calling of somatic variants. The Agilent bioinformatics support team eventually acknowledged this fault.

The quality metrics of using open-sources bioinformatics tools showed a good mapping percentage of the targeted genome region (79-98%), while there is high percentage of duplicated sequenced reads (35-92%). The overall quality metrics findings were consistent with the related SureCall findings, which was discussed in

chapter 3, demonstrating that some aspects of the SureCall pipeline appear to work correctly.

4.4.2. Interpretation of the mutational landscape

My analyses reveal that very extensive differences existed between pre- and post-NAC samples. Strikingly, the number of detected mutations in pre-NAC samples was higher than post-NAC in every patient (p=0.03; paired Student's T test), suggesting that chemotherapy reduced the genetic diversity of the tumours and this supports the clonal evolution model. In addition, around 60% of genes mutations increased in MAF in post-NAC samples, which indicates that pre-existing mutations may play a role in resistance through their expansion and adapting to the chemotherapeutic treatment [129-131].

The comparison between pre-NAC and post-NAC samples allowed me to categorise mutations into 3 main sub-categories: Sub-category 1: mutations found uniquely in pre-NAC samples. This suggests that these mutations in tumour clones were successfully treated by the chemotherapy regimen. Sub-category 2: detection of mutations unique to the post-NAC samples. This suggests that there have been new mutations that resulted in relative resistance to chemotherapy, or that very rare clones that were not detected initially have been selected for and have expanded in MAF. Sub-category 3: the mutations shared between pre-NAC and post-NAC samples. These potentially provided opportunities to identify tumour resistance driving mutations by assessing the representation (MAF) of these mutations relative to unmutated in both pre-NAC and post-NAC samples. However, these speculations are based on the assumption that mutations were not missed during tumour sampling – a particular concern given the pre-NAC samples were small core biopsies taken from a much larger tumour mass, nor undetected by sequencing and analysis due to low region coverage or low allelic frequency, as has been discussed previously [132, 133].

There are a small number of published studies where the authors have inspected the genetic profile of matched pre- and post-chemotherapy tumour samples from

individual patients and have used variety of ways of interpreting the ITH [39, 134-138]. In one such study, WES was performed on pre- and post-NAC cisplatin-based chemotherapy samples from 30 muscle-invasive bladder cancer patients. The data analysis was carried out to identify sub-clonal mutations that were unique to either of the matched pre- or post-treatment tumour samples, which they interpreted to be caused by chemotherapy-induced and/or spatial heterogeneity. In addition, different analyses such as survival and mutational signature validation analyses were carried out and these showed that greater post-treatment tumour heterogeneity predicted worse overall survival. This finding was based on patients who showed no response to treatment, rapid recurrence and short survival, as they had mutations in key genes such E2F3 and JUN (drivers of cell cycle progression) or tumour suppressor genes such as FBXW7 exclusively in the post-treatment tumour sample [135].

In the context of breast cancer, one study examined the presence or absence of 238 specific mutations in 19 cancer-related genes in paired breast tumour core biopsies obtained pre-NAC and post-first cycle doxorubicin or docetaxel in 10 treatment-naïve primary breast cancer patients. The examination was performed using a targeted panel of genotyping assays (Sequenom assays). One of the approaches for data analysis was carried out by categorising the mutational findings into 4 mutational patterns based on mutation status between pre- and post-treatment: wild-type in pre- and wild-type in post-treatment, mutant in pre- and mutant in post-treatment, mutant in pre- and wild-type in post-treatment, wild-type in pre- and mutant in post-treatment. This analysis led to the identification of *PIK3CA* as predominantly mutated in both pre-treatment samples (8/10, 80%) and post-treatment (5/10, 50%), however, no association could be made between mutational pattern category and clinicopathological feature and treatment response or survival [136] from these data alone, highlighting the importance of validating functional roles of the identified mutations (as I attempt in chapter 5 and 6).

The consensus findings from these studies are that there is substantial intra-tumoural heterogeneity in terms of number of detected mutations in pre- and post-chemotherapy samples, and substantial differences between pre- and post-

chemotherapy samples that could be a source of treatment failure through clonal evolution or expansion following selective pressures of treatment exposure. Thereby, these different approaches for analyses can lead to the identification of driver mutations that could provide potential therapeutic targets to treat resistant clones. Nevertheless, no published study exist in breast cancer using similar approaches for identifying mutated genes for chemotherapy response using whole exome level data, and therefore my study has considerable novelty.

4.4.3. Molecular pathways potentially deregulated during therapy

One way of identifying biologically important pathways that were relevant to chemo-response was to analyse the molecular pathways enriched within the lists of mutated genes. The hope was that some pathways might represent common targets for deregulation by mutations in multiple different genes - observations that would be missed when focusing on single genes only. I performed such analyses on the unique to pre-NAC sub-category and the unique to post-NAC sub-category separately. The findings suggested that there are pathways shared between the subcategories and involved in either sensitising to or resisting the epirubicin/cyclophosamide chemotherapeutic regimen. Overall, 3 common pathways that were enriched in both unique to pre-NAC sub-category and the unique to post-NAC sub-category and showed significant statistics parameters (p-value, q-value Bonferroni, q-value FDR B&H, and q-value FDR B&Y) (Appendix 9.3 for more details): extracellular matrix (ECM), collagen protein coding and integrin signalling pathways. This suggests these pathways are involved in chemotherapy response modulation and as a result has helped to focus on genes belong to these pathways to be included in the final list of candidate genes for functional validation *in vitro*.

Interestingly, ECM/integrin signalling has been identified previously as a major pathway contributing to cancer cell survival and resistance to chemotherapy by inhibiting apoptosis via beta1 integrin in solid cancers such as breast cancer and small cell lung cancer [139]. Also, involvement of alpha2 beta1 integrin with its ligand collagen I, reduced apoptosis activity in T cell acute lymphoid leukaemia (T-ALL) cell

lines and primary blasts induced by doxorubicin chemotherapeutic regimens specifically [140]. In addition, many studies have suggested that tumour microenvironment components, including both stromal cells and the non-cellular components of the ECM, play roles in development of chemo-resistance through the activation of survival pathways, such as laminins acting on survival signalling including PI3K/AKT, TP53 and MAPK [141-143]. Altogether, this appears to be concordant with my functional enrichment findings since ECM, collagen proteins and integrin signalling pathways were selected in both unique to pre-NAC sub-category and to the unique to post-NAC sub-category.

In the literature, a similar enrichment analysis was performed for high-grade serous ovarian carcinomas using data for altered gene expression patterns in matched pre- and post-NAC samples. Among the significantly enriched pathways were DNA damage repair, which was up-regulated after treatment, and MAPK signalling, cell-cycle/apoptosis, transcriptional regulation, PI3K signalling, and Notch signalling, which were down-regulated after treatment. The pathways analysis findings helped to make observatory notes such as genes involved in hereditary ovarian cancer signalling showed decreased expression in post- versus pre-NAC analysis [144]. Altogether, this depicts the robustness of functional enrichment analysis on large sets of data in identifying molecular targets for the disease in question.

4.4.4. List of candidate genes for *in vitro* validation

My analysis resulted in a list of 46 candidate genes for further assessment as mediators of chemotherapeutic response *in vitro.* Interestingly, the final list of candidate genes included the TP53 gene, which is known for its role in chemotherapeutic resistance including to anthracycline agents like doxorubicin and epirubicin [145-148]. The presence of this gene in my output lists supports the idea that my process of filtering and prioritising candidate genes has worked effectively, and the TP53 candidate gene could potentially act as a positive control for subsequent *in vitro* investigations.

A few other candidate genes represent particularly promising candidate genes as mediators of chemo-response, while others are promising simply because of the large number of them identified in relatively few overlapping molecular pathways. For example, many candidate genes in the list are from the ECM enriched pathway, such as AP3B1, COL6A3, ITGA7, and SSPO, with particular emphasis on COL6A3 and ITGA7, as they are both also involved in integrin signalling. Both pathways have been identified previously as major pathways contributing to cancer cell survival and resistance to chemotherapy [139, 140]. Furthermore, NOTCH2 appears to be a promising candidate since the Notch pathway was enriched (see Appendix 9.3), and published studies have shown this pathway to impact on chemotherapy response in cancer [149-151].

In addition, the following candidate genes are known in the literature for their roles in a variety of cellular processes that potentially relate to cancer behaviours, including cancer progression and potentially treatment resistance; ABL1, NCOA3, CCDC88C, PTPN14, S100PBP, CACNA1C. ABL1 (ABL Proto-Oncogene 1, Non-Receptor Tyrosine Kinase) is a proto-oncogene that encodes a protein tyrosine kinase and is well known for its role in the BCR-ABL fusion protein that has become a signature of chronic myeloid leukemia (CML). Also, mutations in ABL1 have been shown to confer resistance to imatinib [152, 153]. While, NCOA3 (Nuclear Receptor Coactivator 3) is known to be integrated in breast cancer pathway and EGF/EGFR signalling pathway [154, 155]. The rest of the genes in my list of 46 have relatively less literature, and therefore fewer if any functional studies conducted on them; these, therefore, have potential for high novelty in terms of potentially assigning a role in chemotherapeutic response.

4.4.5. Conclusion

In common, with a number of published studies, I have analysed the data from matched pre- and post-therapy cancer samples, which showed partial resistance to chemotherapy as a way of identifying potential mediators of therapy response. However, my study is unique, as I have assessed this in the context of breast cancer using whole exome data, and samples purified using LCM to reduce stromal contamination, which should makes the findings more relevant to the tumour resistant clones. The findings have allowed me to interpret the findings of the ITH landscape and generate a list of candidate genes that may mediate chemo-response. Interestingly, some of candidate genes appear highly promising in terms of what is known about their functional roles already, while others are novel with relatively little known. These genes have been screened for functional effects on chemo-response in data described in chapter 5.

5. MUC17 and PCNX1 are drivers of chemotherapy response *in vitro*

5.1. Abstract

46 potential mediators of chemo-response were identified in Chapter 4, using genomic analyses of breast cancers showing partial resistance to NAC. My aim in this chapter was to carry out functional screens of these genes, initially using approaches that were higher throughput but with more chance of false findings, but narrowing down on fewer genes with lower throughput approaches giving more confidence in the findings. Having identified genes with functions in defining chemotherapy response, I have attempted to investigate their mechanisms of action.

All 46 genes were tested in screens based on assessing the sensitivity of the ER-positive breast cell line MCF-7 to epirubicin treatment after siRNA knock-down of each gene individually. Based on consistent chemo-response patterns during the two rounds of siRNA screening, MUC17, PCNX1 and TENM4 were taken forward for further validation. MUC17 knock-down was associated with significantly increased cell sensitivity to epirubicin treatment *in vitro*, while PCNX1 knock-down was significantly associated with resistance. TENM4 did not demonstrate a convincing role in chemotherapy response and was excluded from subsequent analyses. Analyses were performed to investigate mechanisms by which MUC17 and PCNX1 modify cellular chemotherapy response, by examining ABC transporter expression levels and cellular loading of epirubicin. ABCB1 and ABCC1 mRNA expressions were significantly down-regulated after MUC17 knock-down, while ABCG2 mRNA gene expression was significantly up-regulated after PCNX1 knock-down. Drug loading assays indicated that MUC17 knock-down increased intracellular drug uptake, while PCNX1 knock-down decreased intracellular drug uptake. These data support a model whereby both genes impact on chemo-response through altering drug loading, potentially through modulating ABC transporter activities.

In summary, MUC17 and PCNX1 are potential drivers of response to chemotherapy in breast cancer, and therapeutic modulation of their activities could potentially enhance chemotherapy responses.

5.2. Introduction

Cancer cells contain a substantial collection of genomic mutations and epigenetic alterations that contribute to the development and individual characteristics of that cancer. Not all of the mutations, however, have important roles in tumour progression, or responsiveness or resistance to treatment; identification of the mutations that are of relevance in these terms can provide novel candidates for targeted treatments. Mutations that provide a selective survival advantage, and thus promote treatment resistance or cancer development, are termed driver mutations, and those that do not are termed passenger mutations. The terms driver and passenger may also be used to refer to the genes harbouring these mutations [73]. Hence, a goal of personalised medicine is to match patients to therapies that are specific to the oncogenic drivers in their tumours, resulting in treatments that are potentially less toxic and more effective [156]. For example, mutations in key survival pathways, such as the PI3K or p53 pathways, can confer cancer cells with different sensitivities to targeted therapy, chemotherapy, and radiation, suggesting these mutations could contribute to treatment resistance [6].

Many computational methods have been used to distinguish driver gene mutations from passenger gene mutations, for example frequency-based or function-based methods. Frequency-based approaches consider candidate driver genes to be genes mutated in a greater proportion of cancer samples than would be expected from the background mutation rate; examples of genes that would be detected by such an approach include TP53 and KRAS, which show consistently high mutation frequency rates in many cancers [73, 157]. Whilst, function-based approaches identify candidate driver mutations by their tendency to have greater impacts on protein function than passenger mutations. Two common sources for functional information are Sorting Intolerant From Tolerant (SIFT) and Polyphen, which incorporate information from sequence context, position and protein characteristics to assess the likely functional impact of mutations [73]. These bioinformatics methods do not provide definitive classification of mutations as drivers or passengers, but help to prioritise candidate driver gene mutations. It remains the case that functional validation is required to definitively classify mutations. A number of approaches have been established that

allow manipulation of the expression or function of relatively large sets of genes in cells lines, and thereby subsequent analysis of whether these manipulations influence the cellular function of interest. Examples include; short interference RNA (siRNA) library screens, cDNA library screens, or miRNA library screens [73, 158]. In the case of assessing the influence of mutations that are thought to produce a loss-of-function, siRNA is particularly appropriate, since this can reduce expression levels thereby mimicking the influence of the mutation. RNAi has become a widely used technique for target discovery, validation, and therapeutic development; and as a screening platform. It has enabled scientists to perform large-scale screens in the field of cancer genomics in order to identify novel diagnostic and drug targets for cancer [159-161].

I have carried out a siRNA screen to perform functional investigation of my candidate genes, by assessing the chemo-sensitivity of an appropriate breast cancer model cell line after siRNA-mediated loss-of-function of individual genes. I have identified MUC17 and PCNX1 as driver genes for chemotherapeutic response. Subsequently, I looked into the mechanisms by which they modulate cellular chemotherapeutic response by examining drug loading and ABC transporter expression levels.

5.3. Results

5.3.1. MCF-7 cells are an appropriate model cell line

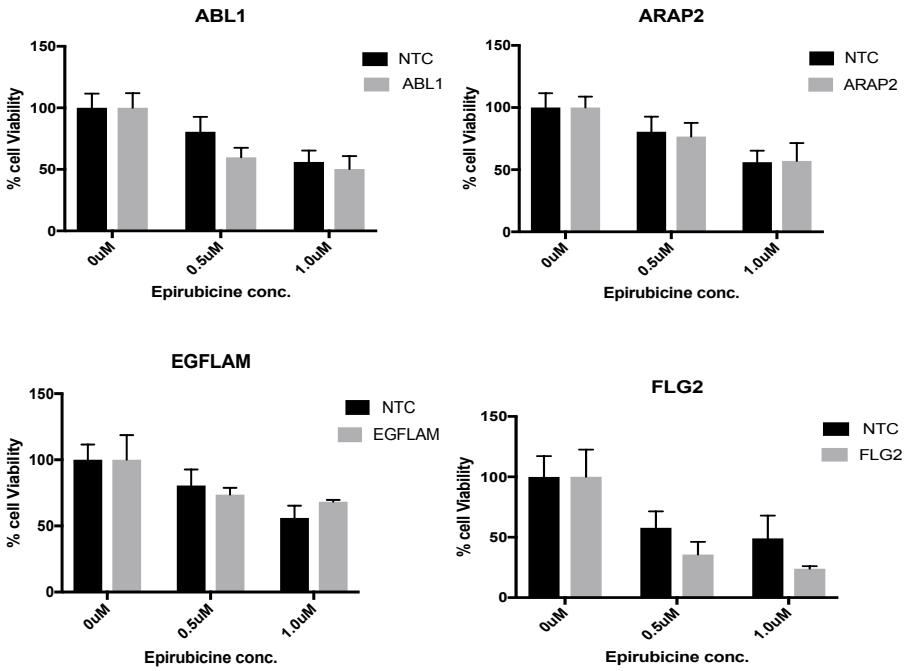
My aim was to mimic loss-of-function mutations in each of my 46 candidate genes using targeted siRNAs in an appropriate breast cancer cell line. In order to do this, I needed to identify a cell line that was representative of the same breast cancer subtype as the clinical cancers I had used in Chapter 3 (luminal A). I also needed to ensure that my 46 genes of interest were expressed in this cell line, and were genomically wild-type (therefore, presumably potentially functional). Breast cancer cell lines regarded as representative of luminal A cancers were identified in our laboratory collections, and transcriptomic expression and genomic profiles for the 46 candidate genes were examined in 3 of these (MCF-7, ZR-75-1 and T47D) using the Cancer Cell Line Encyclopedia (CCLE) dataset. I found MCF-7 was the best cell line to carry out the siRNA screen, as these cells lacked genomic aberrations in any of the candidate genes, and expression was easily detectable for every gene; this was in contrast to the other cell lines where either some data were lacking, genomic aberrations were present, or certain candidate genes were expressed at notably low relative levels.

5.3.2. SiRNA screening for functional influences of 46 genes on chemo-response

I aimed to test the impacts of siRNA knock-down of 46 separate candidate genes on chemotherapeutic drug response *in vitro* by transfecting MCF-7 cells with gene targeted siRNA, treating with epirubicin or control, and then assessing cell viability using MTT assays in comparison to cells transfected with non-targeted siRNA controls. Two separate screens were performed: first a screen of all 46 candidate genes using 2 different doses of epirubicin (0.5 μM and 1 μM), and secondly, a screen of 32 of these genes with 3 different doses of epirubicin (0.5 μM, 1 μM and 2 μM). Data are presented in Figures 5.1, 5.2 and 5.3. Figure 5.1 shows the first screen data for

14 of the genes, including only those that were not taken forward to the second screen. Figures 5.2 and 5.3 show the first and second screen data for the remaining 32 genes.

My aim was to identify siRNAs that caused consistent and notable changes in relative cell survival after treatment with epirubicin. Figure 5.1 shows the data for the 14 genes that failed to meet these criteria on the first screen. For example, transfection with siRNA targeted against ARAP2 or THADA caused almost no difference in relative survival after epirubicin treatment in comparison to the non-targeted siRNA control. By contrast, transfection with siRNA targeted against EGFLAM or KIAA1161 caused differences in relative survival after epirubicin treatment that appeared to differ in direction between the two drug doses, one appearing protective, while the other sensitised. Also, other transfections, such as against PKHD1, appeared to cause a difference in sensitivity at one drug dose only. Finally, transfection with targeted siRNA against PDGFD, SEZ6, S100PSP and ZFHX4 all had an effect at the higher dose only but were not taken forward since a pragmatic decision was made only to take the strongest candidates further. However, they were potentially still functionally relevant. Note, FLG2 represents an exception, as this was excluded from further analysis due to lack of reagents.



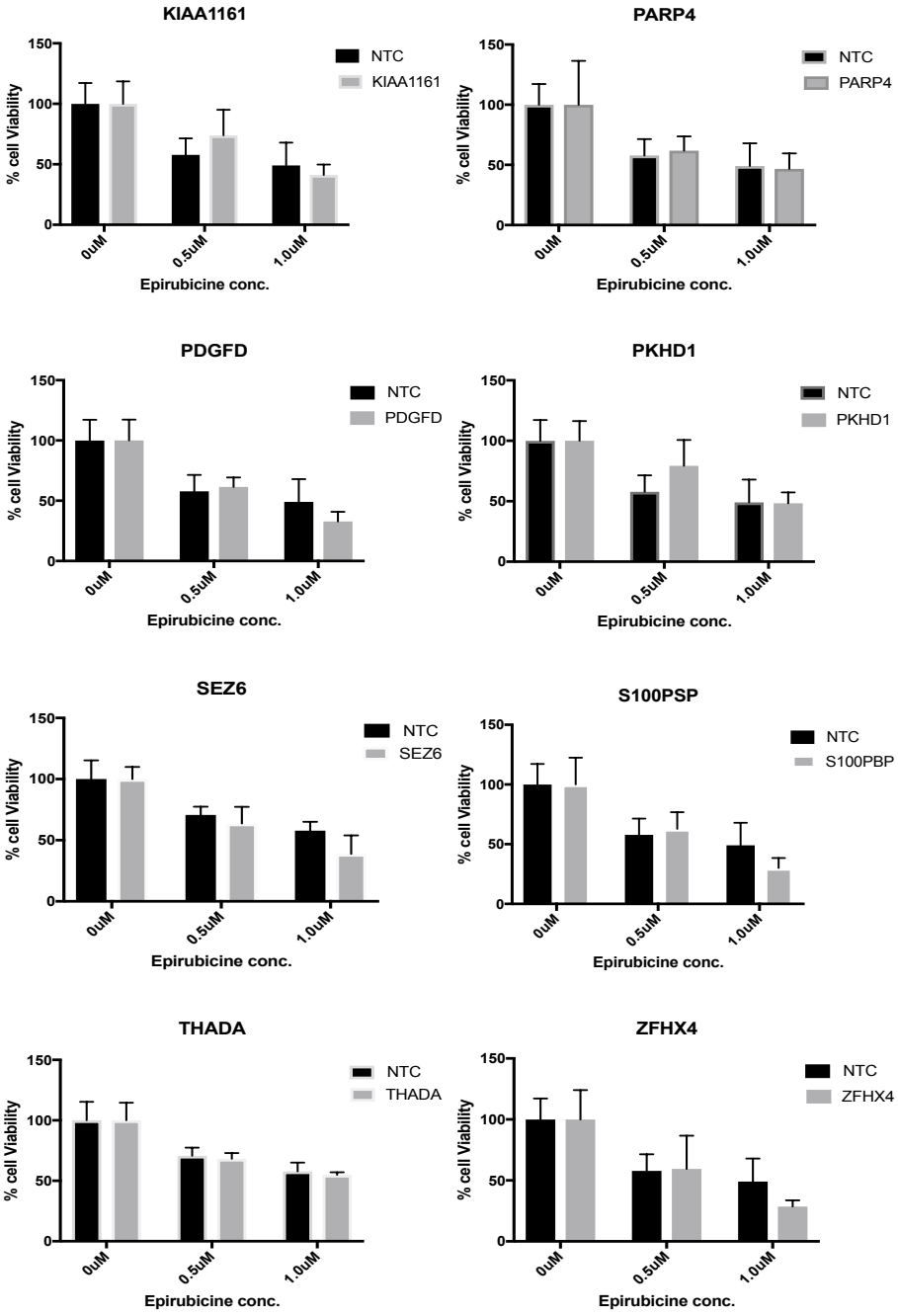
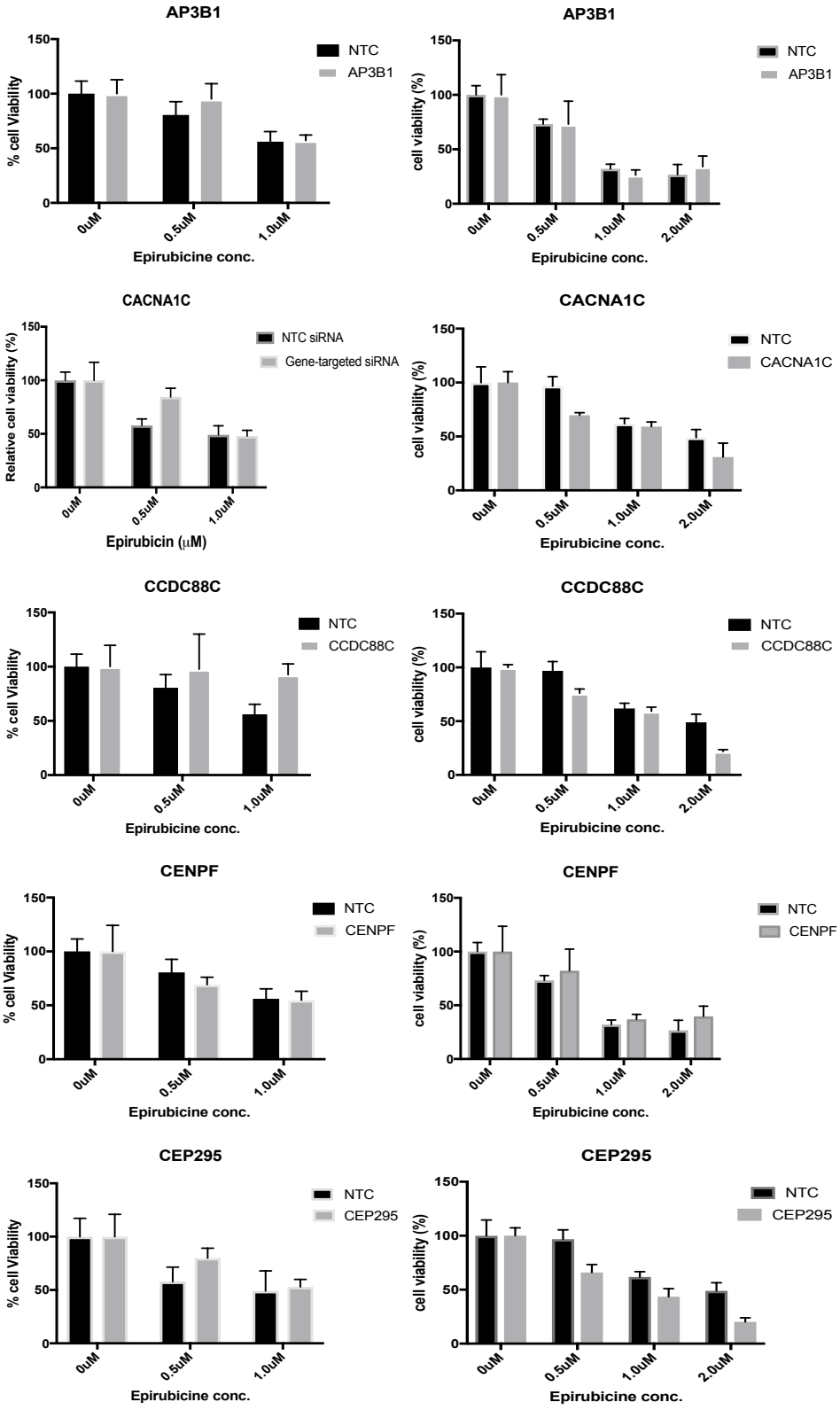
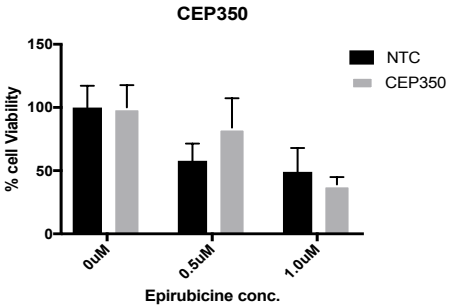


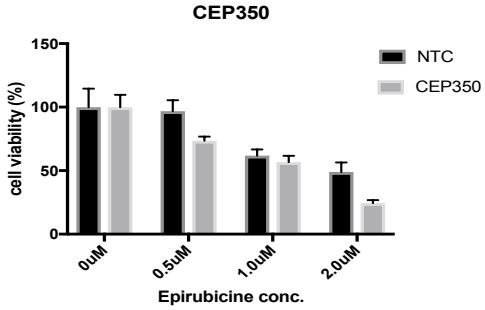
Figure 5.1 14 genes were excluded from further analysis after a first siRNA screen. MCF-7 cells were seeded in 96-wells, and transfected with 50nM of siRNA against target genes or non-targeted siRNA control. 48h later, cells were treated with either of two doses of epirubicin or were left untreated for 24h. MTT assays were performed. Data are presented as relative cell survival after epirubicin treatment by normalisation of treated values to the matched untreated values for either non-targeted control (NTC, black) or targeted siRNA (grey). Data bars represent means of 5 replicate wells from one biological experiment. Error bars represent SD. The epirubicin doses were determined based on the survival cure of MCF-7 cells treated with broad range of epirubicin doses in order to establish inhibitory concentration (IC50).

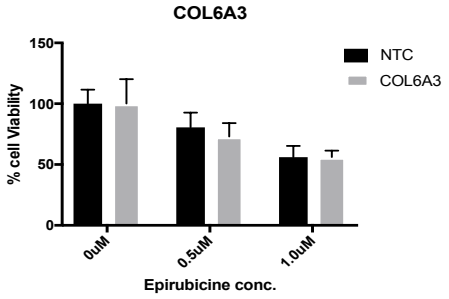
Figures 5.2 and 5.3 show data for the siRNAs that demonstrated some consistency in changes in relative cell survival within screen 1, and were then additionally used in screen 2. In addition, some siRNAs were used in screen 2 mainly based on the strength of the literature associated with the potential role of those genes in chemo-response, even though the screen 1 data showed little potential influence. Figure 5.2 shows data for genes that were not taken forward beyond screen 2, while Figure 5.3 shows data for the 3 genes selected for further analysis. Note, ZNF853 demonstrated a consistency in cells response to chemotherapeutic drug within and between screens 1 and 2, however, it was not taken forward and instead MUC17 was taken forward. Still pragmatic decision was made since both MUC17 and ZNF853 exhibited same chemo-response phenotype but MUC17 was found mutated in 3 patients as opposed to 2 patients for ZNF853.

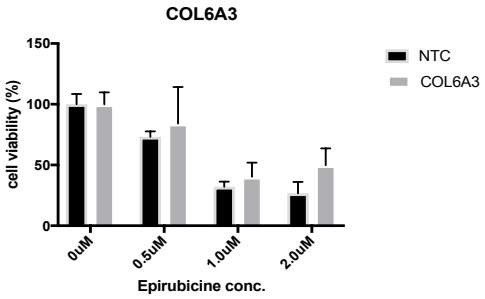
When results from the two screens were compared for each gene, it was notable how poorly concordant the data were between the screens, highlighting the importance of performing biological replicates for screening, as represented by these two independent screens, to avoid being misled by findings that are not reproducible from a single biological assessment. Transfection with siRNAs targeted against many genes, for examples CENPF or COL6A3, caused increased cell survival after epirubicin treatment in comparison to the non-targeted control in one screen, but little influence in the other. In some cases, targeted siRNAs showed increased cell survival following epirubicin treatment in one screen and then the reverse pattern in the other screen; for example, CCDC88C or CEP295.

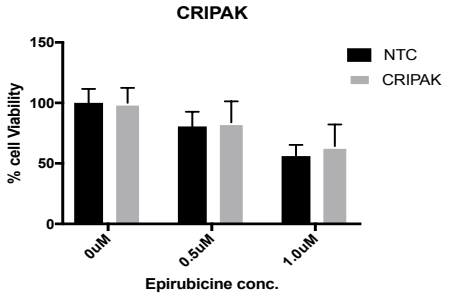

AP3B1
% cell Viability
NTC
AP3B1
0uM
0.5uM
1.0uM
Epirubicine conc.
AP3B1
cell viability (%)
NTC
AP3B1
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
CACNA1C
Relative cell viability (%)
NTC siRNA
Gene-targeted siRNA
0uM
0.5uM
1.0uM
Epirubicin (μM)
CACNA1C
cell viability (%)
NTC
CACNA1C
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
CCDC88C
% cell Viability
NTC
CCDC88C
0uM
0.5uM
1.0uM
Epirubicine conc.
CCDC88C
cell viability (%)
NTC
CCDC88C
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
CENPF
% cell Viability
NTC
CENPF
0uM
0.5uM
1.0uM
Epirubicine conc.
CENPF
cell viability (%)
NTC
CENPF
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
CEP295
% cell Viability
NTC
CEP295
0uM
0.5uM
1.0uM
Epirubicine conc.
CEP295
cell viability (%)
NTC
CEP295
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.

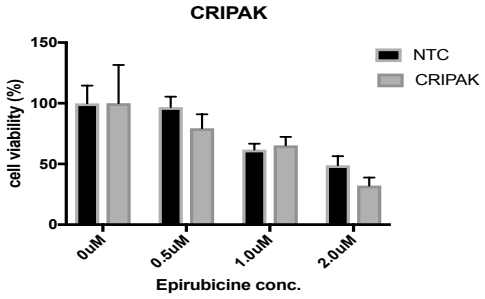

CEP350
% cell Viability
150
100
50
0
0uM
0.5uM
1.0uM
Epirubicine conc.
NTC
CEP350

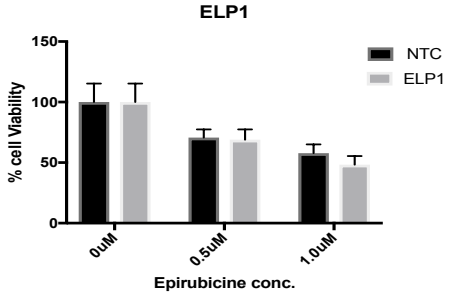

CEP350
cell viability (%)
150
100
50
0
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
NTC
CEP350

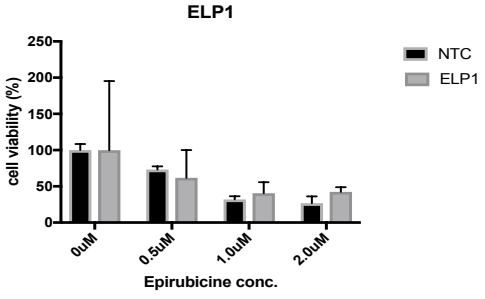

COL6A3
% cell Viability
150
100
50
0
0uM
0.5uM
1.0uM
Epirubicine conc.
NTC
COL6A3

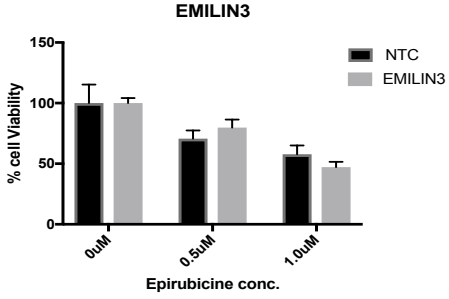

COL6A3
cell viability (%)
150
100
50
0
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
NTC
COL6A3

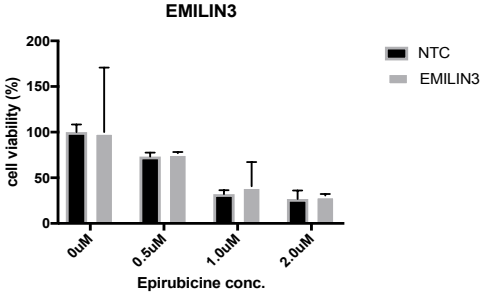

CRIPAK
% cell Viability
150
100
50
0
0uM
0.5uM
1.0uM
Epirubicine conc.
NTC
CRIPAK

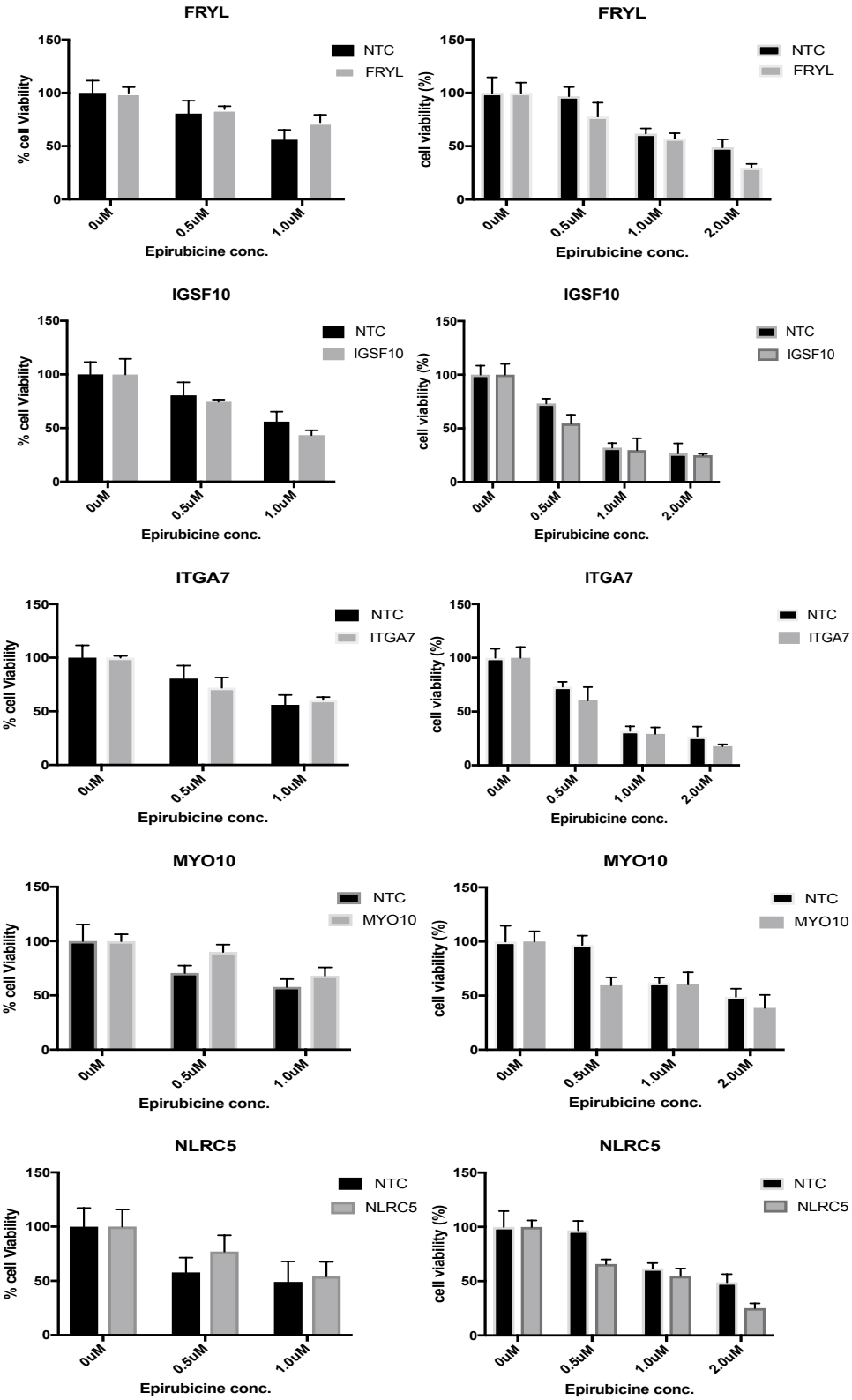

CRIPAK
cell viability (%)
150
100
50
0
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
NTC
CRIPAK

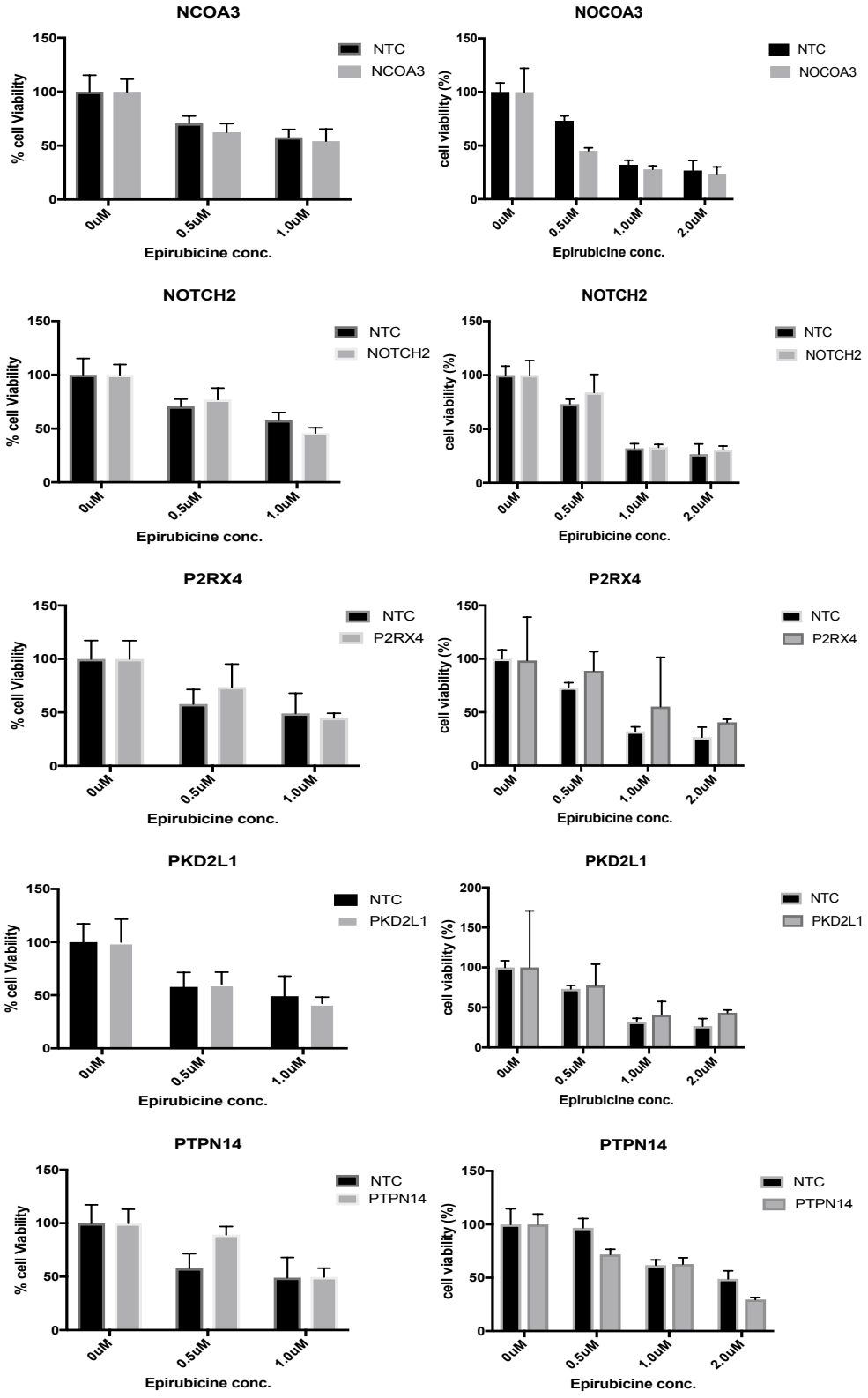

ELP1
% cell Viability
150
100
50
0
0uM
0.5uM
1.0uM
Epirubicine conc.
NTC
ELP1

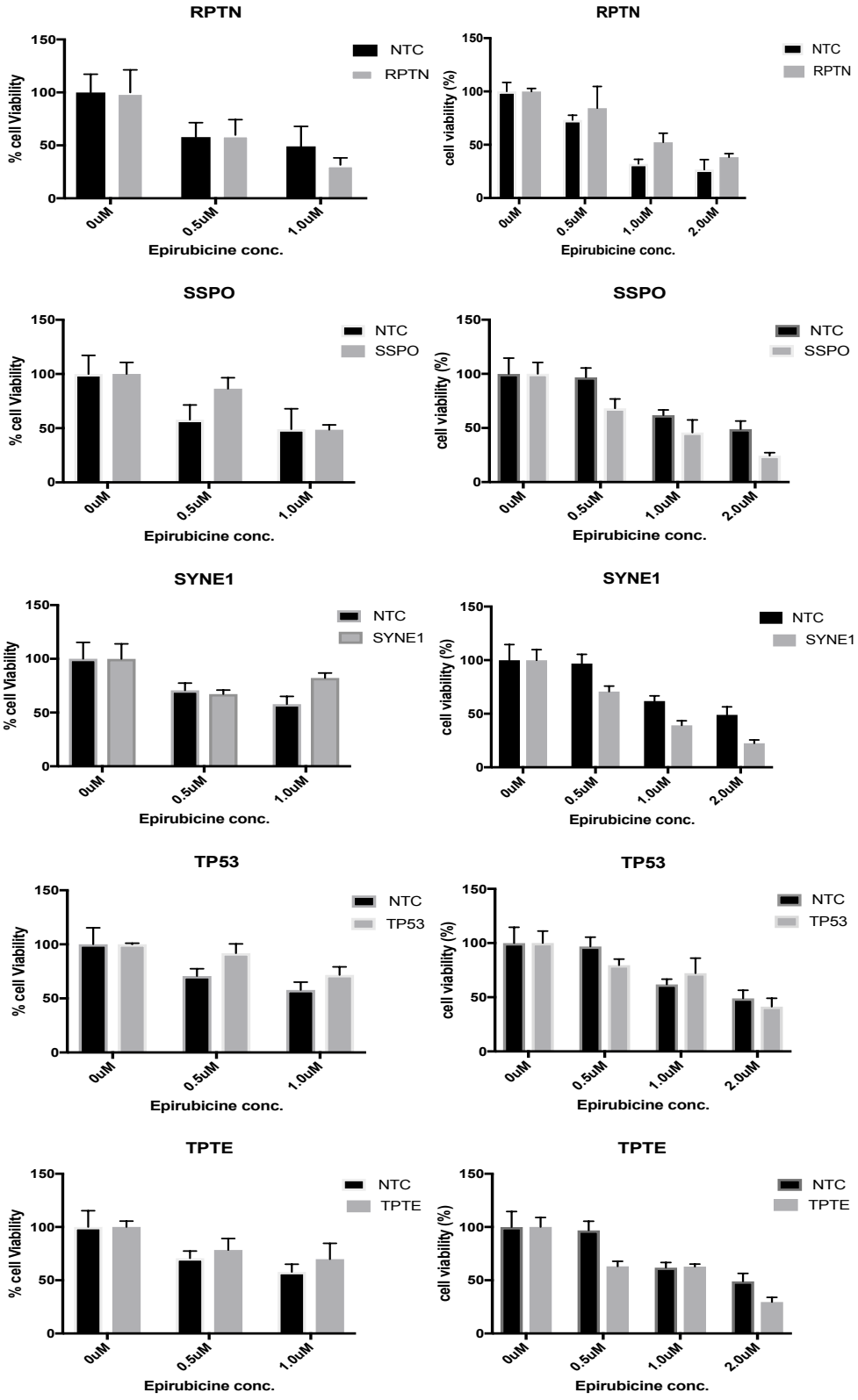

ELP1
cell viability (%)
250
200
150
100
50
0
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
NTC
ELP1


EMILIN3
% cell Viability
150
100
50
0
0uM
0.5uM
1.0uM
Epirubicine conc.
NTC
EMILIN3


EMILIN3
cell viability (%)
200
150
100
50
0
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
NTC
EMILIN3


FRYL
NTC
FRYL
% cell Viability
cell viability (%)
0
50
100
150
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
IGSF10
NTC
IGSF10
ITGA7
NTC
ITGA7
MYO10
NTC
MYO10
NLRC5
NTC
NLRC5


NCOA3
% cell Viability
150
100
50
0
0uM
0.5uM
1.0uM
Epirubicine conc.
NTC
NCOA3
NOCOA3
cell viability (%)
2.0uM
NOTCH2
P2RX4
PKD2L1
PTPN14


RPTN
% cell Viability
150
100
50
0
NTC
RPTN
0uM
0.5uM
1.0uM
Epirubicine conc.
RPTN
cell viability (%)
150
100
50
0
NTC
RPTN
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
SSPO
% cell Viability
150
100
50
0
NTC
SSPO
0uM
0.5uM
1.0uM
Epirubicine conc.
SSPO
cell viability (%)
150
100
50
0
NTC
SSPO
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
SYNE1
% cell Viability
150
100
50
0
NTC
SYNE1
0uM
0.5uM
1.0uM
Epirubicine conc.
SYNE1
cell viability (%)
150
100
50
0
NTC
SYNE1
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
TP53
% cell Viability
150
100
50
0
NTC
TP53
0uM
0.5uM
1.0uM
Epirubicine conc.
TP53
cell viability (%)
150
100
50
0
NTC
TP53
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.
TPTE
% cell Viability
150
100
50
0
NTC
TPTE
0uM
0.5uM
1.0uM
Epirubicine conc.
TPTE
cell viability (%)
150
100
50
0
NTC
TPTE
0uM
0.5uM
1.0uM
2.0uM
Epirubicine conc.

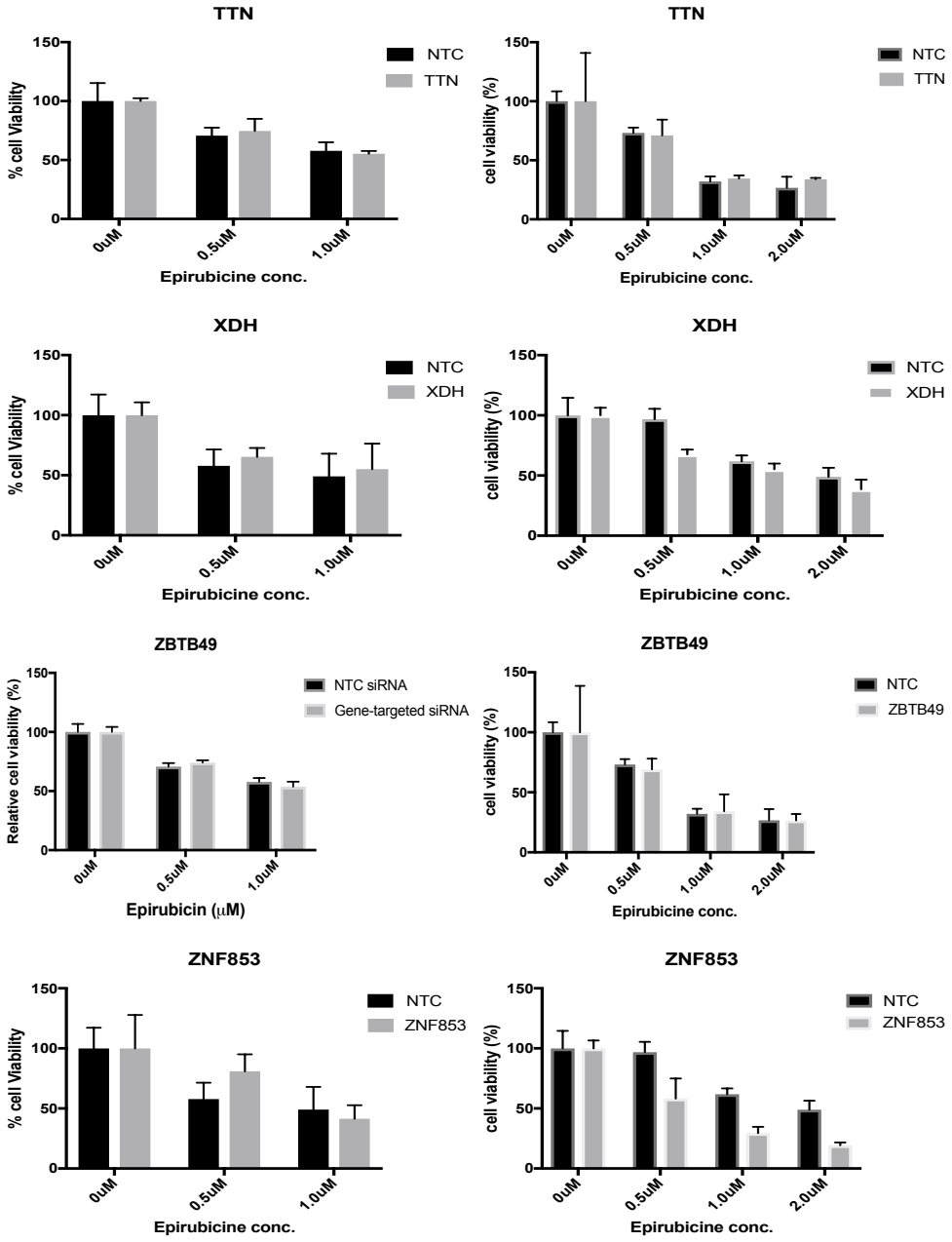


Figure 5.2 29 genes that demonstrated some consistency in changes in relative cell survival within screen 1 (left column), or were strong candidates in the literature, were then additionally used in screen 2 (right column). Cells were transfected with siRNA and treated with 2 or 3 doses of epirubicin as described for Figure 5.1. Data are presented as relative cell survival after epirubicin treatment by normalisation of treated values to the matched untreated values for either non-targeted control (NTC, black) or targeted siRNA (grey). Data bars represent means of 5 replicate wells from one biological experiment. Error bars represent SD.

Figure 5.3 shows the data for the three siRNAs I selected as representing the strongest candidate genes from these data based on the consistent chemo-response pattern during the 2 siRNA screening rounds; these were MUC17, PCNX1 and TENM4. Cells treated with siRNA against MUC17 and PCNX1 showed increased sensitivity to epirubicin treatment as compared to the non-targeting siRNA (NTC), while cells treated with siRNA against TENM4 showed increased cell survival after epirubicin treatment, equating to apparent resistance. These three genes were selected for further *in silico* and *in vitro* analyses in the next sections.

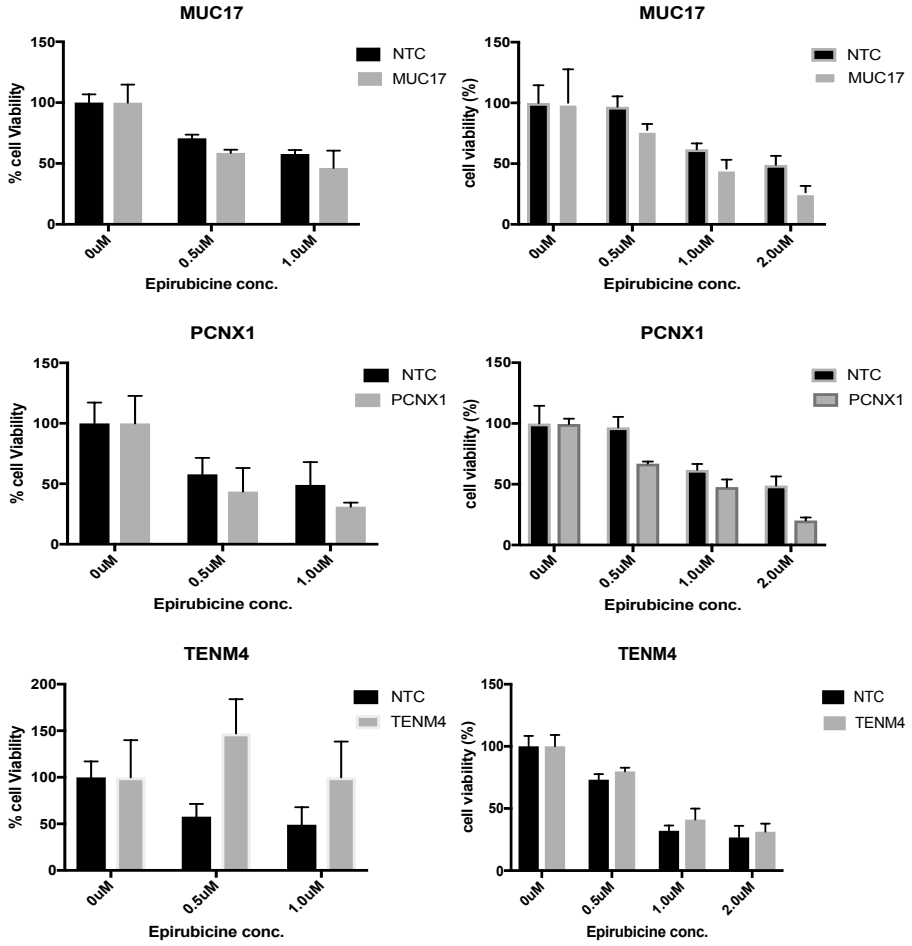


Figure 5.3 MUC17, PCNX1 and TENM4 were selected as representing the strongest candidate genes based on the consistent chemo-response pattern during 2 siRNA screens. Cells were transfected with either targeted siRNA or non-targeted siRNA control and treated with 2 or 3 doses of epirubicin as described for Figure 5.1. Data are presented as relative cell survival after epirubicin treatment by normalisation of treated values to the matched untreated values for either non-targeted control (NTC, black) or targeted siRNA (grey). Data bars represent means of 5 replicate wells from one biological experiment. Error bars represent SD.

5.3.3. Further bioinformatics analyses of the mutations in MUC17, PCNX1, and TENM4

Screening findings suggested that MUC17, PCNX1, and TENM4 were potential mediators of chemotherapy response. Before performing further wet-lab analyses, I went back to the annotation data for the genomic variants identified within these genes and also used the UniprotKB dataset, to gain potential insights into the mutational impact of these variants on the 3D protein structures and, therefore, potentially on protein function.

There were 3 variants for the MUC17 gene found in 3 patients, all of which were identified in the unique to pre-NAC sub-group, thus potentially defining cells that were successfully eradicated by chemotherapy i.e. candidate mutations for eliciting increased chemo-sensitivity. The full length of the MUC17 encoded glycoprotein is 4493 amino acids (aa) comprising 3 main regions; an extracellular portion (aa 26-4393) containing tandem repeats of a 59 residue sequence and two EGF-like repeats, one which is transmembrane (aa 4394-4414), and the other cytoplasmic (aa 4415-4493). All variants were missense mutations located in the extracellular portion of the protein and were predicted to have moderate effects on protein function. The variant p.Thr3809Met with the polyphen2 predicted damaging effect fell between the last repeat domain in the extracellular portion and the transmembrane EGF-like domain. The latter is a region of likely functional importance: EGF-like domains can associate with the EGF receptor 2 (EGFR2) on the surface of adjacent cells, an interaction that may be involved in growth signal transduction to stimulate cellular proliferation [162].

There were 2 variants for the PCNX1 gene found in 2 patients. Both were found in the unique to post-NAC sub-group, indicating the potential for the mutations to have conferred resistance to chemotherapy. Note that the data from the screen (Figure 5.3) supports the hypothesis that loss of function of this gene (mediated by siRNA) was associated with *chemo-sensitivity*; considering the mutations were found in the post-

NAC samples this could suggest that the patient mutations were gain of function mutations. The PCNX1 protein is 2341aa. Both variants, p.Thr564Fs and p.Asp623Fs, are positioned within the transmembrane domain that contains multiple transmembrane spans from residues 28 to 1312. Since both of the variants have frameshift effects in an early transmembrane helix of the protein domain, they are likely to cause loss of the majority of the encoded protein. Thereby, one might predict that the mutations are most likely to cause loss of function of the gene, which is not clearly compatible with the data above.

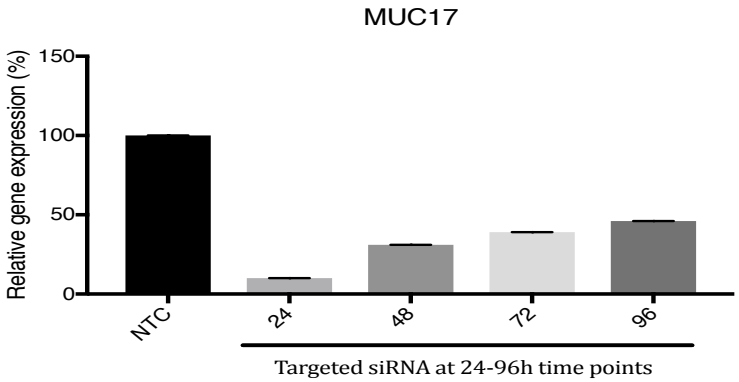
There were 2 variants for the TENM4 gene in 2 patients. Both were found in the unique to pre-NAC sub-group, suggesting that the mutations apparently define cells that were successfully eradicated by chemotherapy, therefore are relatively chemo-sensitive. Note that the screen (Figure 5.3) showed that TENM4 siRNA knock-down (loss of function) caused chemo-resistance, a finding that could be compatible, again, with the proposal that the TENM4 mutations in patients were gain of function mutations. The encoded 2,825aa protein contains a putative N-terminal signal sequence (aa 1-341) and a transmembrane domain (aa 346-366), followed by 8 EGF -like repeats (aa 562-831). One of the patient variants (p.Val88Fs) is in the Teneurin N-terminal domain and the other one (p.Asp780Fs) is in the EGF-like domain 7. Both of the variants have frameshift effects on the protein domains. Since, the N-terminal amino acid of protein is an important determinant of its half-life, it is likely that the encoded protein is degraded due to defects in this process [163]. Also, the other variant may affect cell proliferation since EGF-like domains can interact with human EGF receptor 2 (EGFR2) on the adjacent cell surface, with downstream effects on growth signalling [162]. Therefore, it might be predicted that these mutations may cause loss of function of the gene – a prediction that is again apparently not simply compatible with the data above concerning functional effect of siRNA and the mutations being unique to the pre-NAC samples.

5.3.4. MUC17, PCNX1 and TENM4 siRNAs knock-down are effective

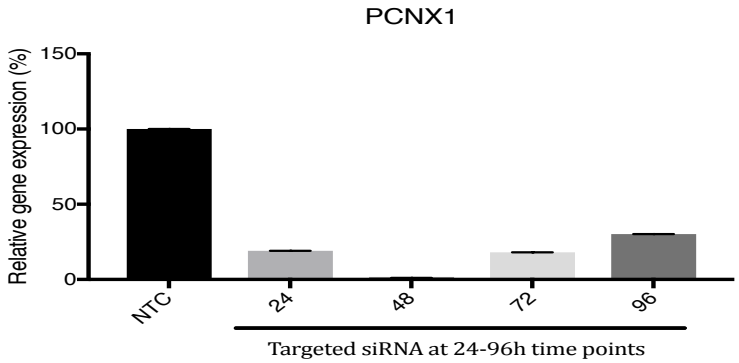
Before further investigations into MUC17, PCNX1, and TENM4 was performed, it was first critical to assess the siRNA knockdown efficiency at both mRNA and, when possible, protein levels. This was not performed at the level of the initial screens of 46 genes, simply due to the cost and work-load required for the relatively little benefit of potentially ruling out some false negative effects on chemo-response caused by poor knock-down efficiency.

MCF-7 cells were transfected with siRNAs targeted against MUC17, PCNX1, or TENM4, or with the non-targeted control, as previously. qPCR was used to quantify relative expression of these gene products over 4 different time-points (24, 48, 72 and 96 hours) post-transfection (Figure 5.4). The data demonstrate all three gene products were successfully targeted for up to 96 hours post-transfection. The highest peak of gene knock-down was found at 24 hours post-transfection for MUC17 (90%), 48 hours post-transfection for PCNX1 (99%), and 72 hours post-transfection for TENM4 (71%), with greater than 50% knock-down at all time-points for MUC17 and PCNX1. In addition, protein expression after siRNA treatment was assessed at the 48 hours post-transfection time-point using immuno-fluorescence for MUC17 (Figure 5.5) and western blots for PCNX1 (Figure 5.6). A suitable antibody was not available for TENM4, therefore TENM4 knock-down was not assessed at the protein level. Successful knock-down of MUC17 and PCNX1 protein was confirmed (~57% knock-down for MUC17 and ~50% for PCNX1).

A



B



C

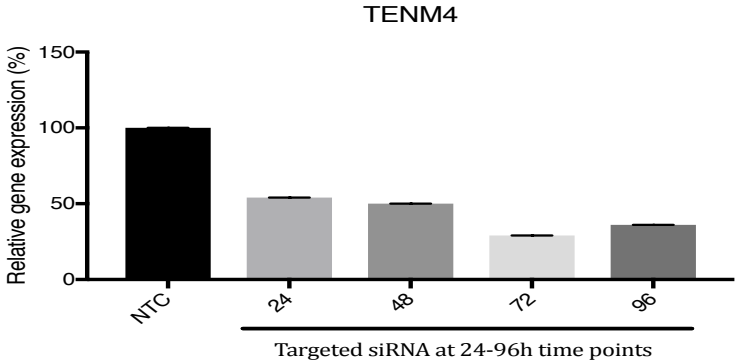


Figure 5.4 Transfection of MCF-7 cells with siRNA targeting MUC17, PCNX1 or TENM4 reduces gene expression successfully for up to 96h. MCF-7 cells were transfected with 50μM siRNA against MUC17 (A), PCNX1 (B), TENM4 (C), or non-targeted control (NTC) siRNA. RNA was prepared from the cells, and relative gene expression of MUC17, PCNX1, TENM4 was assessed at mRNA level using qPCR at 4 time points (i.e. 24, 48, 72, 96 hours post-transfection). Data are presented relative to expression with the NTC. NTC was performed for each time point and was used to normalise the expression of the targeted siRNA at each time point. Error bars represent SEM of technical replicates for one biological repeat.

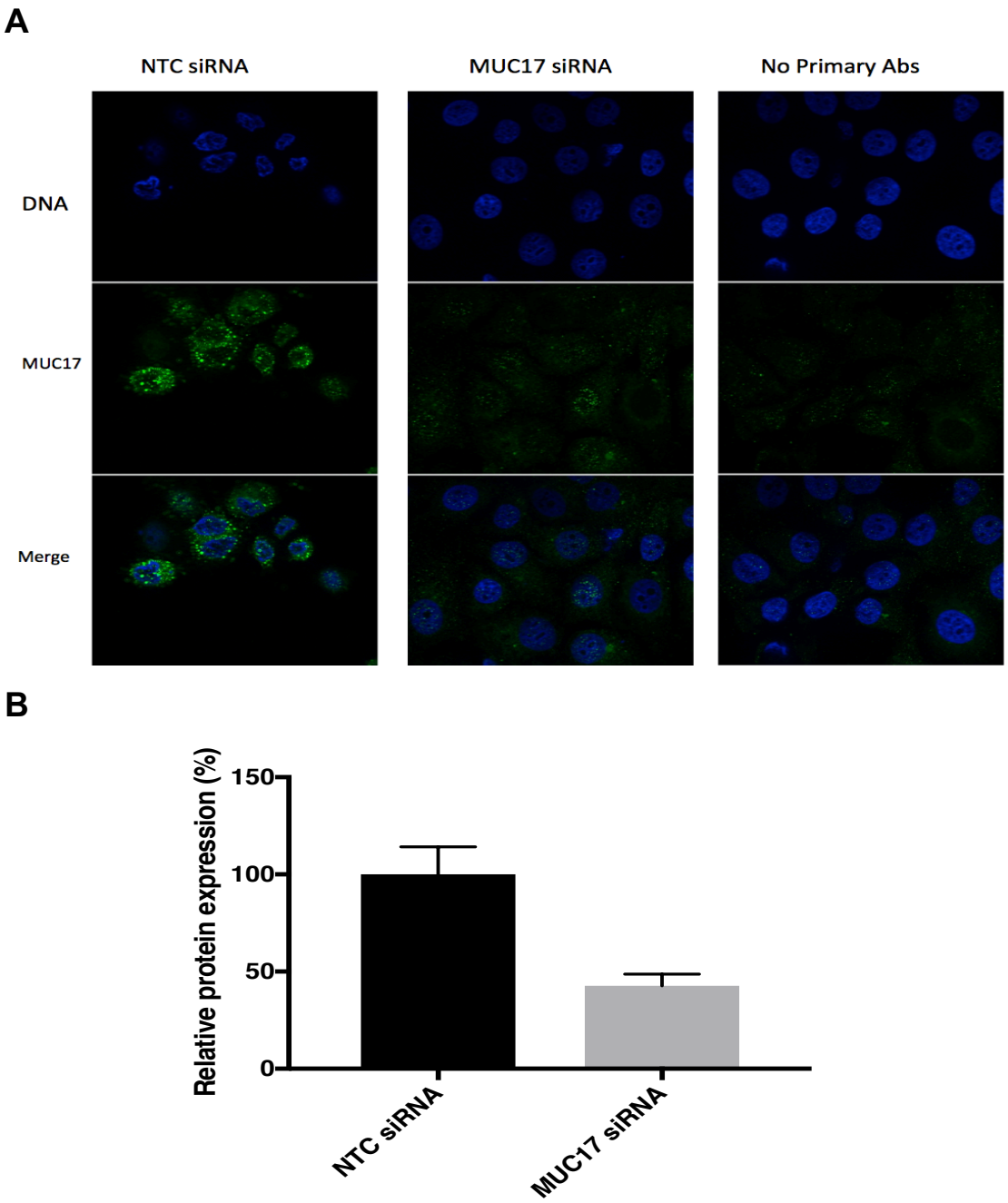


Figure 5.5 Immuno-fluorescence demonstrates successful knock-down of MUC17 protein using siRNA. MCF-7 cells were transfected with targeted MUC17 siRNA or NTC siRNA for 48h. Cells were then treated with anti-MUC17 or no primary antibodies (negative control) followed by fluorescent secondary antibodies. DAPI was used to stain the nuclear DNA. (A) Images were taken using confocal microscope. (B) Semi-quantitative densitometry analysis was performed on 10 cells to assess relative MUC17 expression using ImageJ software. The expression level of the targeted protein was normalised to the NTC. Data represent mean fluorescence levels, with error bars showing the SEM.

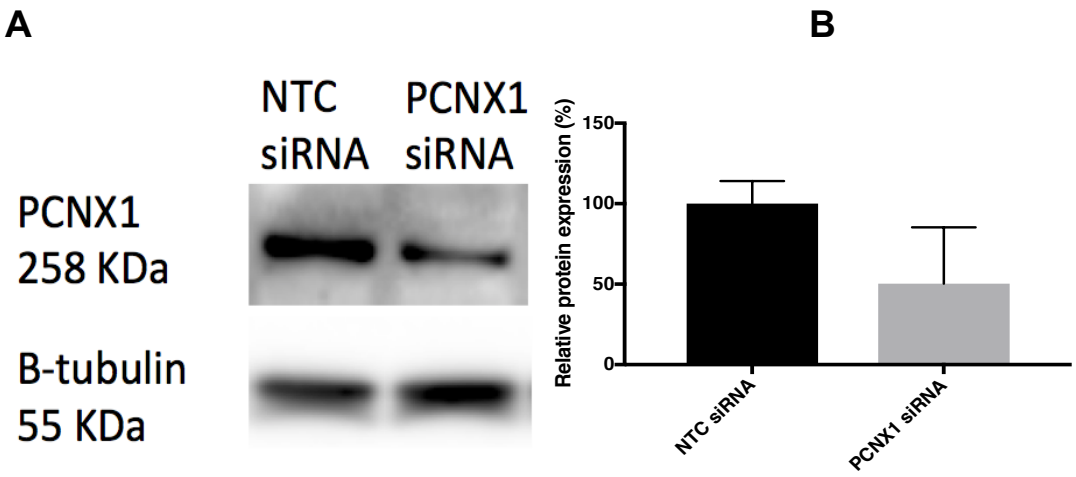


Figure 5.6 Western blot analysis of non-targeted siRNA control (NTC) and PCNX1 siRNA shows successful knock-down of PCNX1 protein. MCF-7 cells were transfected with targeted PCNX1 siRNA or NTC siRNA for 48h and then total protein extracted. Proteins were subjected to Western blot analysis for PCNX1 or B-tubulin. (A) Western blots images as labelled (B) Semi-quantitative densitometry of PCNX1 expression. The expression level of the targeted protein is normalised to the NTC. The error bars represent SEM of technical replicate measurements.

5.3.5. Expression levels of MUC17 and PCNX1, but not TENM4, influence response to epirubicin

Two further cell viability assays were performed to assess the functional impact (i.e. chemotherapeutic response) of targeted siRNA for MUC17, PCNX1 and TENM4. Firstly, a more thorough assessment was performed of short-term impacts using expanded versions of the screening assay (MTT assay), by adding a broad range of chemotherapeutic drug concentrations, testing over three time points, and performing three independent biological replicates. Secondly, an assessment of impact on chemotherapeutic response in terms of influencing longer-term survival using colony forming assays (CFA), again with three independent biological replicates.

5.3.5.1. Assessment of short-term survival influences using MTT assays

MCF-7 cells were transfected with targeted siRNA for MUC17, PCNX1 or TENM4, or with non-targeted control siRNA as described before, and then after 24h, cells were treated with a wide range of doses of epirubicin from 0 up to 4μM. MTT assays were performed after 24, 48 or 72 hours of treatment with epirubicin drug. These time points represent 48-96 hours post-transfection - time points at which suitable knock-down efficiency is maintained (Figure 5.4). As expected, MCF-7 cells showed reduced survival that was both dose-dependent and time-dependent after treatment with epirubicin. Comparison between control and targeted siRNA treatments at every individual dose or time-point revealed no significant differences (Mann–Whitney tests). Nevertheless, overall, the trends showed increased sensitivity to epirubicin treatment for MUC17 siRNA treated cells as compared to NTC siRNA (Figure 5.7; 6 separate doses at 24 hours epirubicin treatment, 2 doses at 48 hours and 4 doses at 72 hours). This trend was significant for the 72h time point, when assessed using a two-way ANOVA test (p=0.0018). Treatment with siRNA against PCNX1 or TENM4 did not show significant differences in chemotherapeutic response compared to the non-targeted control siRNA (Figures 5.8 and 5.9).

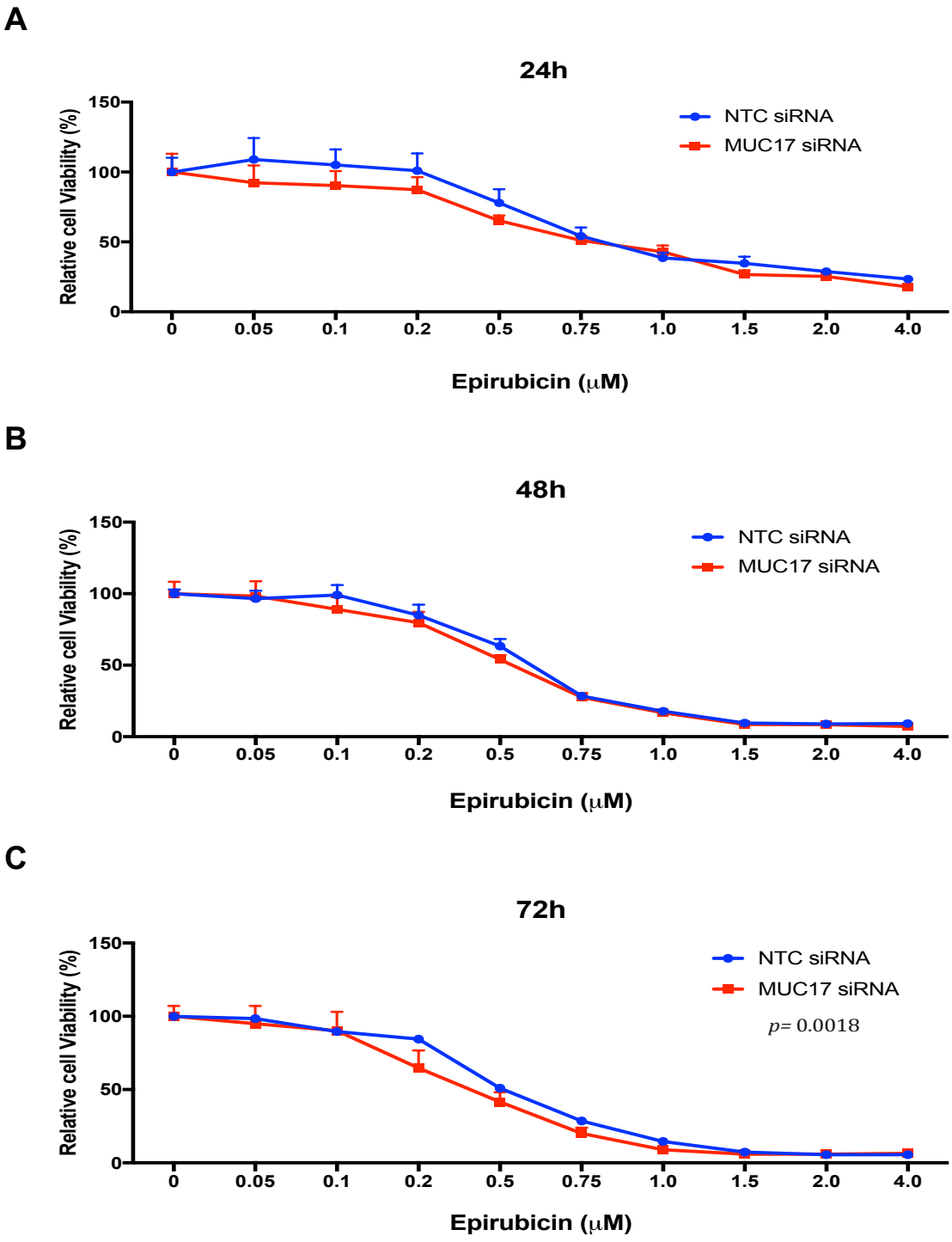


Figure 5.7 Decreased MUC17 expression leads to increased sensitivity to epirubicin. MCF-7 cells were seeded in 96-wells, and transfected with 50nM of siRNA against MUC17, or non-targeted siRNA control (NTC). 24h later, cells were treated with one of 9 different doses of epirubicin or were left untreated for 24h (A), or 48h (B) or 72h (C). MTT assays were performed. Data were normalised to 0μM epirubicin treatment. The error bars represent the SEM of 3 biological independent experiments. Trends between chemo-response after NTC or targeted MUC17 siRNA were analysed using two-way ANOVA tests and showed significant results for 72h (p=0.0018).

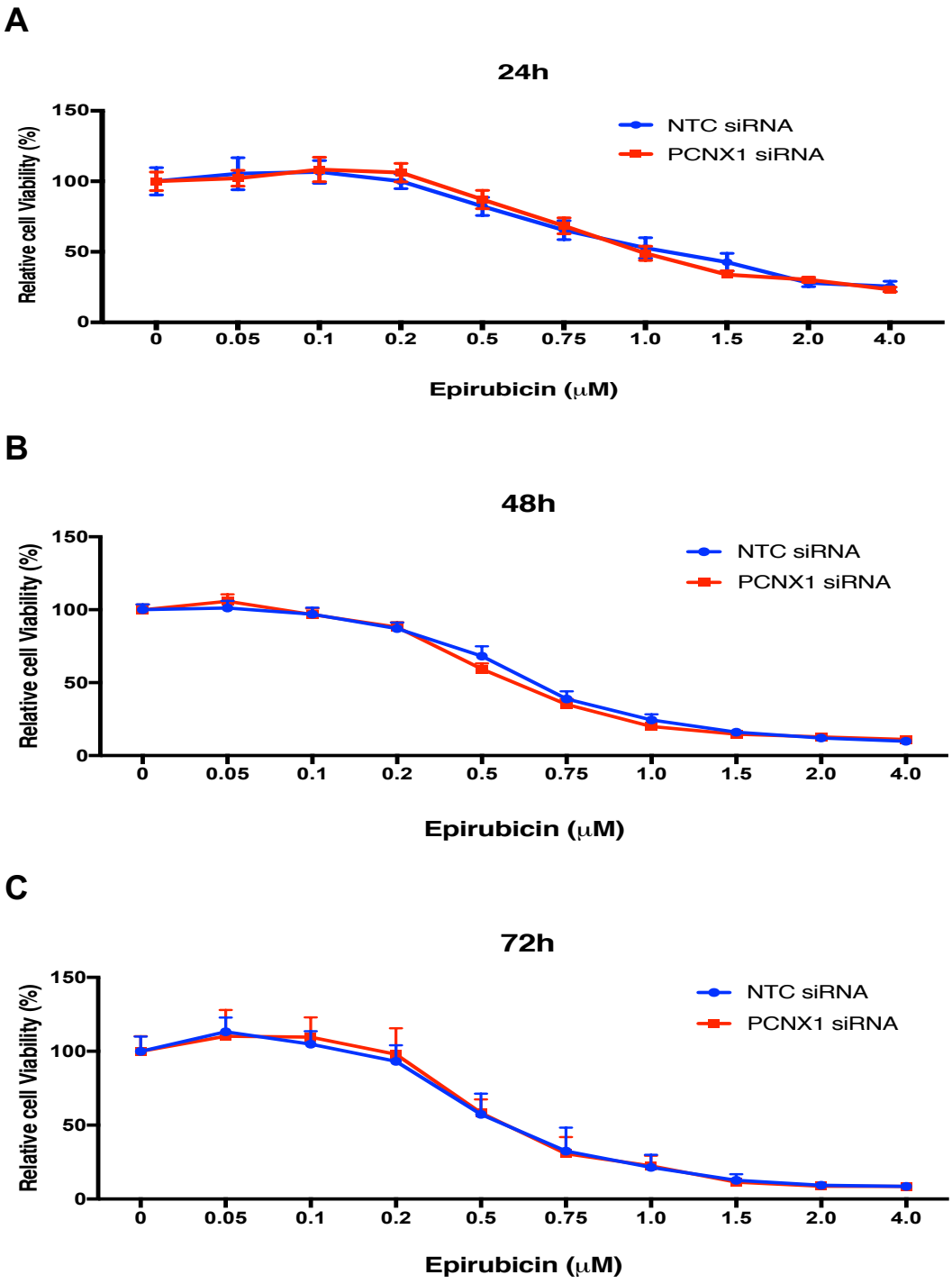


Figure 5.8 Decreased PCNX1 did not show differences in chemotherapeutic response compared to the non-targeted control siRNA. MCF-7 cells were seeded in 96-wells, and transfected with 50nM of siRNA against PCNX1, or non-targeted siRNA control (NTC). 24h later, cells were treated with one of 9 different doses of epirubicin or were left untreated for 24h (A), or 48h (B) or 72h (C). Data were normalised to 0μM epirubicin treatment. The error bars represent the SEM of 3 biological independent experiments. Trends between chemo-response after NTC or targeted PCNX1 siRNA were analysed using two-way ANOVA tests and did not show any significant results.

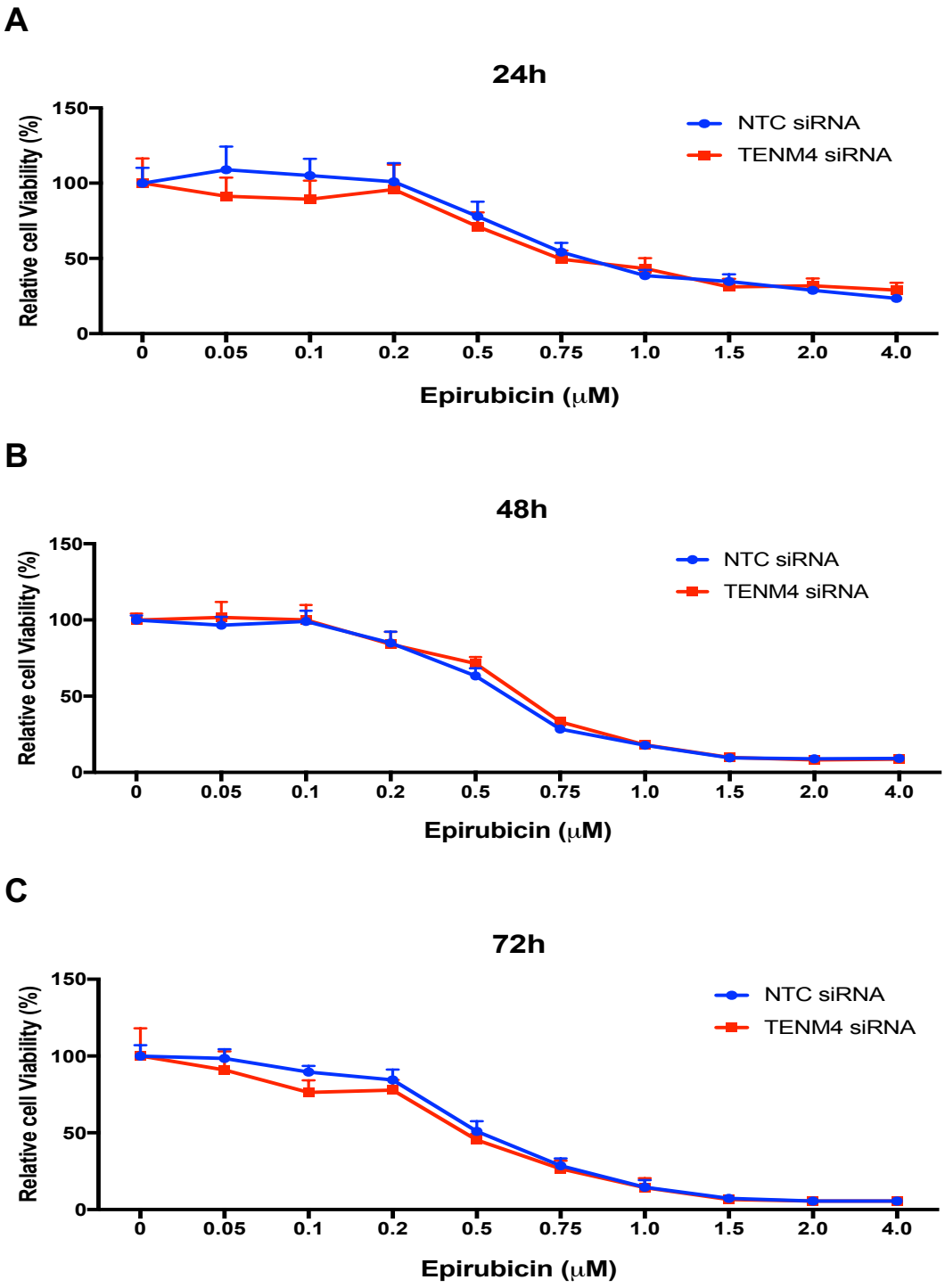


Figure 5.9 Decreased TENM4 did not show differences in chemotherapeutic response compared to the non-targeted control siRNA. MCF-7 cells were seeded in 96-wells, and transfected with 50nM of siRNA against TENM4, or non-targeted siRNA control (NTC). 24h later, cells were treated with one of 9 different doses of epirubicin or were left untreated for 24h (A), or 48h (B) or 72h (C). MTT assays were performed. Data were normalised to 0μM epirubicin treatment. The error bars represent the SEM of 3 biological independent experiments. Trends between chemo-response after NTC or targeted siRNA were analysed using two-way ANOVA tests and did not show any significant results.

5.3.5.2. Assessment of longer-term survival influences using colony forming assays (CFA)

Colony forming assays were performed to assess the influences of siRNAs targeted against MUC17, PCNX1 or TENM4 to modify the ability of cells to survive long-term after treatment with epirubicin. MCF-7 cells were transfected as before, and were then treated with 0, 10, 20, 40, 50 or 100nM epirubicin for 24h. Cells were then re-plated at low density in fresh medium (without epirubicin) to allow them to demonstrate their potential to grow into viable colonies, indicating long-term survival. Targeting MUC17 with siRNA significantly reduced survival after epirubicin treatment, whereas, targeting PCNX1 significantly increased survival (Mann Whitney tests, $p<0.05$). In addition, the overall chemo-response trend was significant for both MUC17 and PCNX1 using a two-way ANOVA test ($p<0.0001$). On the other hand, targeting TENM4 caused no significant changes in survival at any dose of epirubicin (Figure 5.10). It should be noted that the result for PCNX1 here is actually the opposite from that seen in the initial siRNA screens, in which PCNX1 knock-down appeared to be associated with increased sensitivity (Figure 5.3), although that previous screening analysis lacked sufficient experimental repeats for robust conclusions, while the analysis here in Figure 5.10 is statistically significant after three independent biological repeats. The contradictory results between the primary siRNA screen and colony forming assay for PCNX1 can be explained as following; 1) cell viability assay (MTT assay) used for initial screen with siRNA is designed to assess the cells response to chemotherapeutic drug in short term survival, whereas the colony forming assay is designed to assess cells ability to form colonies in relatively long term. 2) the initial siRNA with MTT lacks of biological repeats and was performed in a relatively lesser number of drug doses, as a result the screen is subjected to high risks for false positive/ negative phenotypes. The colony forming assay, however, was performed with 3 biological replicates and under many different drug doses which enabled to perform statistical analysis to assess the significance level of the findings.

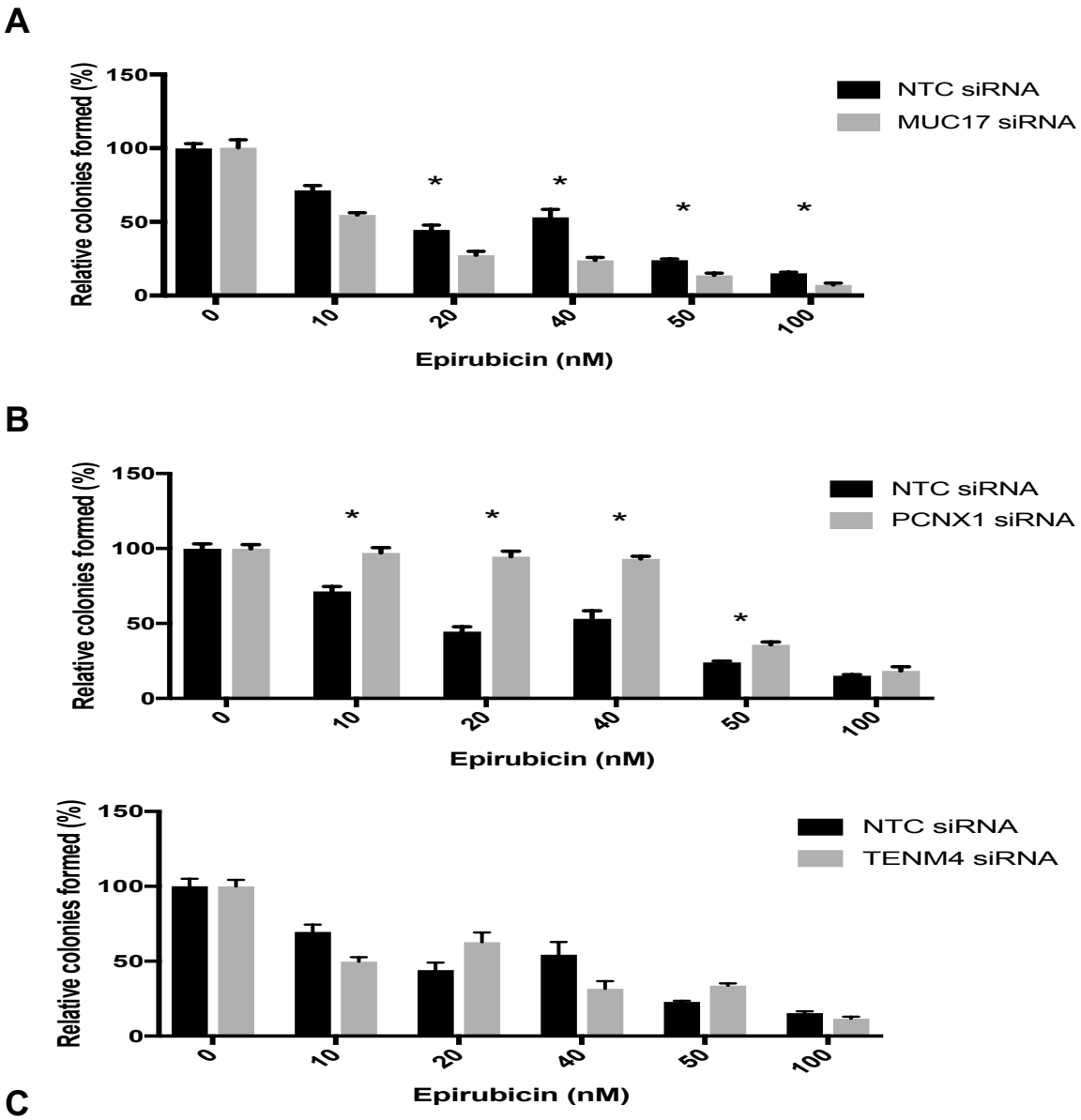


Figure 5.10 CFA for MUC17, PCNX1, and TENM4 siRNA showed significantly reduced colonies formation after epirubicin treatment for MUC17 and significantly increased colonies formation for PCNX1, while no significant change in survival for TENM4. MCF-7 cells were seeded in 96-wells, and transfected with 50nM of targeted siRNA against MUC17 (A), or PCNX1 (B), or TENM4 (C) or non-targeted siRNA control (NTC). 48h later, cells were treated with either one of 5 different doses of epirubicin or were left untreated for 24h. Cells were then re-plated at low density in fresh medium (without epirubicin) to allow them to demonstrate their potential to grow into viable colonies over 14 days incubation period. Data were normalised to 0 nM epirubicin treatment. The error bars represent SEM of 3 biological independent experiments. * Indicates statistically significant *p* value <0.05 Mann Whitney test. Trends between chemo-response after NTC or targeted siRNA were analysed using two-ways ANOVA tests and showed significant results for MUC17 and PCNX1 ($p<0.0001$).

5.3.6. MUC17 and PCNX1 are significantly up-regulated by epirubicin treatment

Having determined that expression levels of MUC17 and PCNX1 influence chemotherapy response *in vitro*, these genes were taken forward for further investigation into the mechanisms by which they modulate chemotherapeutic response. First, I aimed to investigate if their gene expression is induced or repressed in response to epirubicin treatment, since this could represent an induced survival mechanism. MCF-7 cells were treated with 1μM epirubicin for various different times up to 48h, and qPCR was used to assess relative expression of MUC17 and PCNX1. The data showed that there was dramatic, increasing, up-regulation of MUC17 expression, from 2-fold up-regulation at as early as 8 hours of treatment to more than 10-fold at 48h (Mann Whitney test, $p<0.01$). PCNX1 expression was also significantly up-regulated as early as 8h (less than 2-fold), although levels did not increase, but were maintained until 48 hours. I concluded that this up-regulation of MUC17 was compatible with a model where MUC17 is induced to provide chemo-protection, since I have previously shown higher levels to be associated with relative resistance (Figure 5.11).

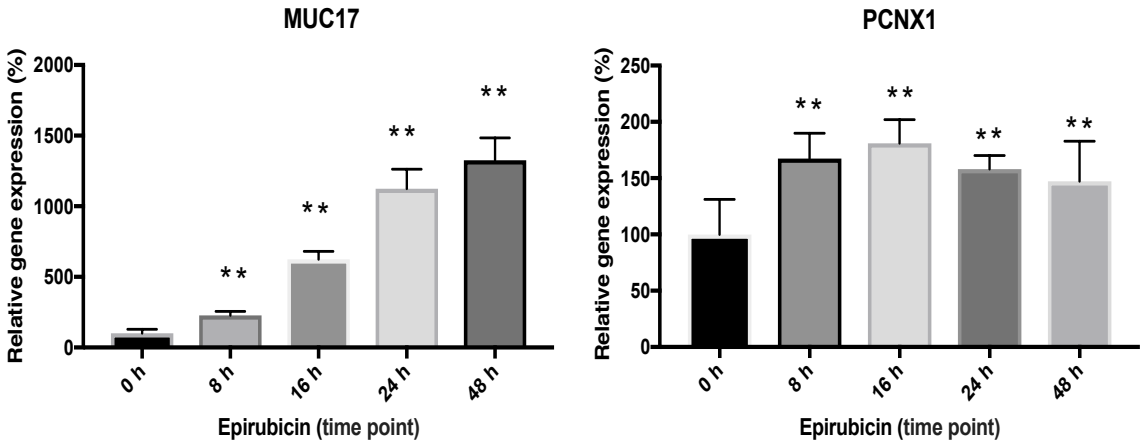


Figure 5.11 MUC17 and PCNX1 mRNA expression show significant expression up-regulation with epirubicin treatment over different time points. MUC17 and PCNX1 RNA was prepared at 5 different time points and then expression was assessed using qPCR. Expression of MUC17 and PCNX1 were determined after normalising to the endogenous gene (ACTB). Data are expressed as a percentage of expression at the 0 hour time point. The error bars represent the SEM of 3 biological independent experiments. ** Indicates statistically significant p value <0.01 Mann Whitney test.

5.3.7. Changes in epirubicin sensitivity associated with MUC17 and PCNX1 knock-down correlate with changes in drug loading and in expression of ABC transporters

One proposed mechanism by which tumour cells can be relatively resistant to chemotherapeutic drugs is via high expression of ATP-binding cassette (ABC) transporters, which can lead to reduced intracellular concentrations of active chemotherapeutics through enhancing drug efflux activity [164-166]. In order to investigate this proposed mechanism, I next measured the amount of intra-cellular accumulation of epirubicin after manipulation of MUC17 and PCNX1 expression and treatment with epirubicin. MCF-7 cells were transfected with siRNAs, as before, and were treated with 1μM epirubicin for 24 hours. Flow-cytomtery was used to assess intra-cellular epirubicin levels, taking advantage of the fact that the drug itself is fluorescent. This assay demonstrated that there was an increase in intra-cellular drug accumulation in cells treated with MUC17 siRNA by ~32% compared with NTC siRNA cells. Whereas, intra-cellular drug accumulation decreased with PCNX1 siRNA by ~16% as compared with NTC siRNA cells (Figure 5.12). However, neither observation

was statistically significant, the trend revealed here was consistent with the cell viability assays (Figures 5.9 and 5.10), since reduced MUC17 expression increased drug loading and increased sensitivity to the drug, while reduced PCNX1 expression reduced both drug loading and sensitivity.

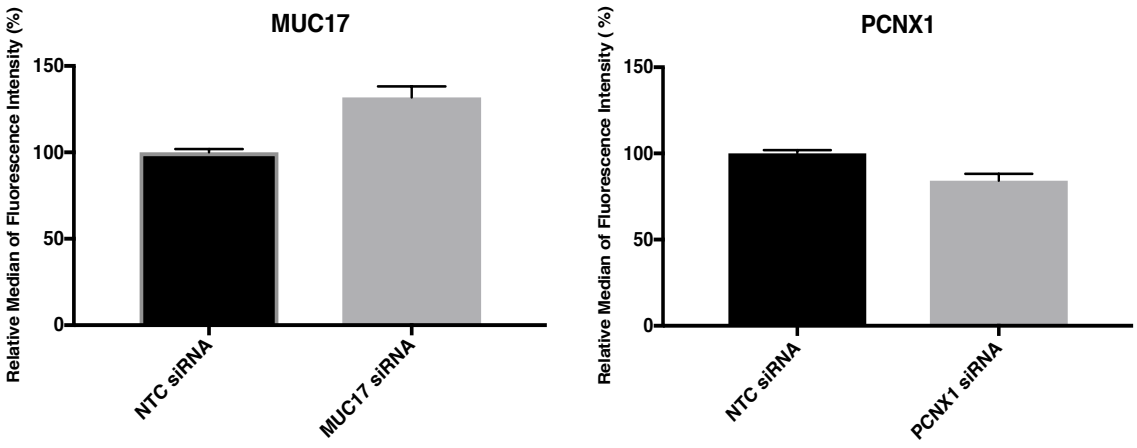


Figure 5.12 MUC17 and PCNX1 knock-down appears to alter intracellular epirubicin uptake. MCF-7 cells were transfected with 50nM of targeted siRNA against MUC17 or PCNX1 or non-targeted siRNA control (NTC). 48h later, cells were treated with 1μM epirubicin for 24 hours. The median fluorescence intensity of Epirubicin drug loading into MUC17 and PCNX1 siRNA and NTC siRNA cells was measured using flow-cytometry. Data for MUC17 and PCNX1 siRNA were normalised to NTC siRNA. The error bars represent the SEM of 3 biological independent experiments.

Changes in ABC transporter expression would provide plausible mechanisms for the apparent differences in epirubicin loading induced by siRNAs against MUC17 and PCNX1. Therefore, the influence of MUC17 and PCNX1 siRNA on the expression of three ABC drug transporters, ABCB1 (P-gp, P-glycoprotein), ABCG2 (BCRP, breast cancer resistance protein) and ABCC1 (MRP1, multidrug resistance associated protein) was assessed. These transporters were selected based on a study that found a direct association between a mucin gene family member, MUC1, and ABC transporters in development of chemotherapy resistance in pancreatic cancer [167]. MCF-7 cells were transfected with siRNAs as before and qPCR was used to quantify relative expression of the three transporters (Figure 5.13). MUC17 siRNA induced

significant down-regulation of both ABCB1 and ABCC1 transporters, although no change in ABCG2. By contrast, PCNX1 siRNA induced significant up-regulation of ABCG2.

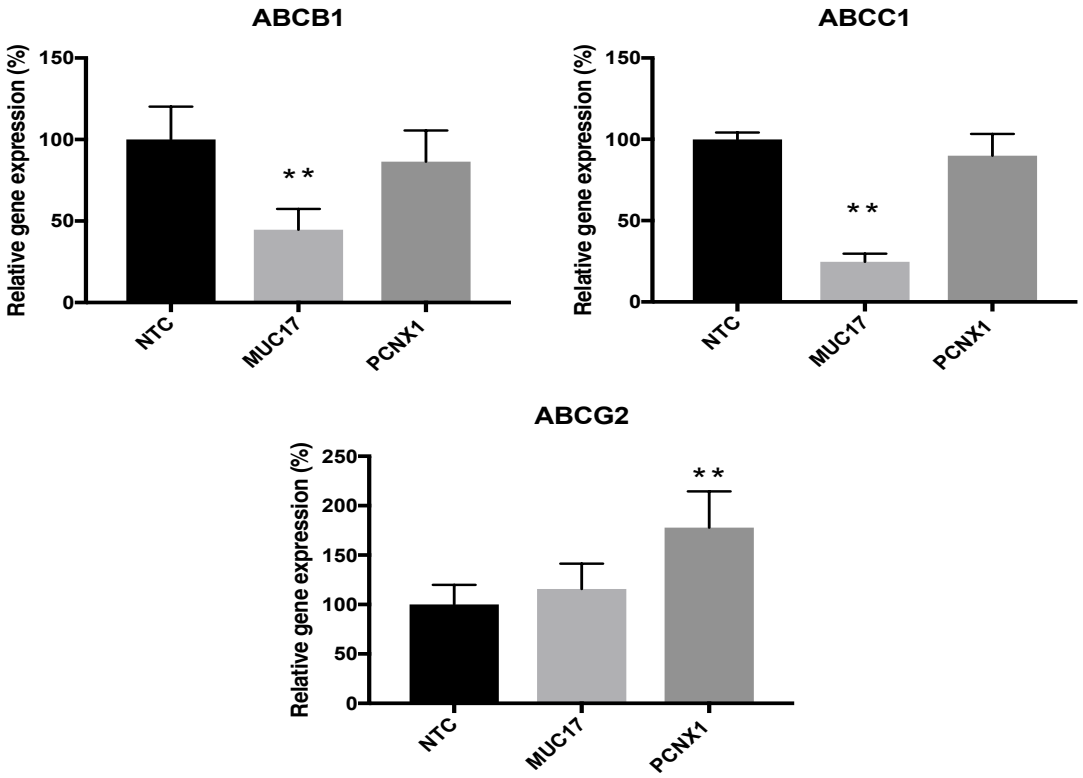


Figure 5.13 Expressions of ABCB1, ABCC1, and ABCG2 transporters are altered after knock-down of MUC17 or PCNX1. MCF-7 cells were transfected with 50nM of targeted siRNA against MUC17 or PCNX1 or non-targeted siRNA control (NTC). 48h later, the RNA was extracted. The mRNA expression of ABCB1, ABCC1 and ABCG2 were determined after normalising to the endogenous gene (ACTB), then data were normalised to NTC. The error bars represent the SEM of 3 independent biological experiments. ** Indicates statistically significant p value <0.01 Mann Whitney test.

5.4. Discussion

5.4.1. siRNA screens revealed potential chemotherapy driver gene mutations

46 candidate genes were screened using siRNA knock-downs and MTT assays to detect difference in chemotherapeutic survival after treating cells with epirubicin at 2 or 3 different concentrations. This screen was performed manually, and in relatively large wells (96-well size, not 384-well size as used for some screens) and replicate numbers (5 technical replicates within each experiment), therefore was costly in terms of time and reagents. But it is important to note that even with this investment, the initial dataset could not allow proper statistical analyses of findings to inform decisions about which genes should be studied further. Therefore, pragmatic decisions were made concerning which genes to continue with, and which to drop from the screen, even when such decisions were based on imperfect data sets. There are many limitations involved with performing this medium throughput siRNA screen. Some of these limitations are as follows.

1. The siRNA screen was designed to induce gene loss of function (LOF) and therefore this excluded proper study of mutations that cause gain of function (GOF) perturbations [168]. It should be noted the one could hope that siRNA screening might still give insight into some gene mutations that have GOF effects, by inducing the opposite phenotype as predicted from the mutation.

2. The siRNA screen was performed on individual genes separately and the influence in the chemotherapy response seen is due to dramatic gene expression reduction. However, this approach excludes multi-gene expression effects in order to produce the phenotype and also assumes that the likely substantial change in gene expression induced by siRNA functionally reproduces what could be more subtle effects of mutations.

3. The cell viability assay that was used for the siRNA screen was limited to observe short-term effects on the chemo-response. This would potentially create many false

negatives for mutations that actually do have roles in chemo-response but would only show an apparent phenotype on longer-term chemo-response.

4. There is a lack of statistical power in my approach to account for variability in chemo-response between the screens, leading to high chances of false positives and negatives [169].

5. Gene siRNA knockdown efficiency was not assessed at either the mRNA or protein level during the screen. This makes the approach susceptible to false negatives through relatively poor knock-down at the mRNA level, through ineffective siRNA function, or at the protein level, again via ineffective siRNA or owing to high stability of any existing protein. However, this can be solved by using Clustered Regularly Interspaced Short Palindromic Repeat-cas9 (CRISPR) system to induce permeant knock-out of the mutation in the cell line. In addition, off-targets effects of siRNA due to the nature of partial complementarity between an siRNA and multiple transcripts in the 3′UTR, can lead to unanticipated false positive phenotypes [168, 169].

Nevertheless, the aim of my work plan was to perform an initial screen of low confidence, with potentially high rates of false positive and negative phenotypes, to narrow down the list to fewer candidate genes with an increased chance of having true effects on chemo-response. This should lead to increased confidence in those individual genes for additional validation minimising the risks for false positive phenotype. Given that this strategy has led to the identification of two genes, MUC17 and PCNX1, that do indeed appear to act as mediators of chemotherapy response, this illustrates the successful use of this approach in terms of saving resources and time.

Many studies have implemented similar approaches to discover novel drug targets for cancer therapy. For example, two studies have utilised high-throughput RNAi (interference) (shRNA or siRNA) screens to identify candidate genes involved in paclitaxel sensitivity in triple negative breast cancer (TNBC), or in sensitivity to a range of different therapies in a panel of colorectal cancer cell lines [160, 166]. In the former study, shRNA screening was initially performed on 1,778 genes, identified as their transcripts levels significantly correlated with genome copy number and were deemed

genomically deregulated in breast cancer. The top 36 high-confidence genes from the shRNA screen were then screened again using two independent siRNA oligos into two TNBC cell lines, MDA-MB-231 and MDA-MB-468. This led to the identification of gene targets that increased paclitaxel sensitivity when knocked-down by siRNA, and that have targetable drug/chemical inhibitors currently available (for example; erlotinib for EGFR): PPM1D, CENPF, BCL2L1, FRAP1, IGF1, EGFR, ERK1, RPS6KB1, TGFB1, and SP1 [166]. In the latter study, the authors assessed sensitivity to a panel of therapeutics including cetuximab (anti-EGFR) and trametinib (anti-MEK) after siRNA targeting the human kinome as well as 95 genes commonly mutated in colorectal cancer. This led to the confirmation of at least 7 synergistic drug plus siRNA combinations involving siRNAs against PINK1, CRIM1, PIK3CA, HUNK, PIM1, CDKN2D and BRAF [160].

Interestingly, some limitations that I encountered using the RNAi screen system in my study were also addressed in these studies. For example; in the former study the variability in cells sensitivity to paclitaxel following knock-down by targeted shRNA, between different plates represented an obstacle. Also, the method of selection of high-confidence gene targets from the primary screen was complicated as the observed phenotype could be due to false positive or off-target effects of individual shRNA. The authors implemented normalisation of targeted shRNA to NTC and a bootstrap algorithm strategy and also confirmed findings with another RNAi system (i.e. individual and pooled siRNA) in order to overcome these confounding factors [166]. In comparison, I performed extensive cell viability assays on selected candidates to support that the observed phenotype was not due to false positive effects.

In the latter study, the reproducibility of phenotype between multiple screens represented an issue for the study. The rescreen of 40 genes from the primary screen led to confirmation of only 8 drug/siRNA combinations that reproduced the original synergistic phenotype in the primary screen [160]. Similar to my study, this may have led to elimination of many potential true positives, but provided high confidence and prioritised targets for future investigations. Despite the obstacles, these

publications illustrate the success of using high-throughput RNA interference genomic screens to identify cancer-relevant, druggable targets to enhance drug sensitivity, especially for patients who showed resistance to chemotherapy.

5.4.2. MUC17 and PCNX1 modulate chemotherapy response *in vitro*

Following siRNA screening of the 46 candidate genes, 3 candidates genes (MUC17, PCNX1, and TENM4) were taken forward for further functional validation, though other candidates still remain of interest for further validation. The three candidates selected were based on showing the most consistent chemotherapeutic response during two independent siRNA screens. However, TENM4 did not demonstrate a convincing role in chemotherapy response and was excluded from subsequent analyses.

MUC17 siRNA knock-down was associated with increased sensitivity to drug *in vitro*. This is in accordance with the sub-group in which the mutation was found in (unique to pre-NAC sub-group), the initial screening data, and the screening validation data. On the other hand, PCNX1 siRNA knock-down was associated with increased sensitivity to drug *in vitro* in the initial screen, which was surprising, since: 1) PCNX1 mutations were identified in unique to post-NAC sub-group, potentially indicative of chemo-resistance, yet, 2) mutations were predicated to cause LOF, as modelled by siRNA knock-down. However, the results in the validation of the function of the PCNX1 gene (Figure 5.10) differed from these screening results (Figure 5.3), and were after all in accordance with the hypothesis that LOF of PCNX1 as induced by siRNA or genomic mutations resulted in chemo-resistance. In addition, investigation of MUC17 and PCNX1 gene expression upon treatment with epirubicin at different time points revealed that there was significant induction of gene expression providing further circumstantial evidence of involvement of MUC17 and PCNX1 in responding to chemotherapy agents, perhaps as an induced survival mechanism at least for MUC17. Some published literature exist concerning MUC17 and PCNX1, although neither have been directly implicated in defining chemotherapy response previously. I review literature about these genes and their functions below.

The MUC17 gene encodes a protein belonging to the mucin (MUC) family of heavily O-glycosylated, high molecular weight, glycoproteins. Overall, family members are mainly membrane-bound in epithelial cells and function to lubricate apical mucosal epithelium surfaces, maintain luminal structures and provide signal transduction [170]. Little information on specific functions of MUC17 has been published to date, with no reports concerning MUC17 as a mediator of chemotherapy response. On the other hand, other members of mucin family, namely MUC1 and MUC4, regulate both cellular differentiation and proliferation and their aberrant expressions are known to be implicated in metastasis and tumorigenesis [171-173]. Studies have shown overexpression of MUC1 to be implicated in resistance to methotrexate in colorectal carcinoma [170]. Also, MUC4 expression modulation was found to be involved in sensitivity to gemcitabine based-chemotherapy in pancreatic cancer, using gene gain or loss of function approaches [174, 175]. Another mucin family member, MUC13, was found to activate nuclear factor-κB (NF-κB), which is a key transcription factor promoting cancer cell survival. The data showed MUC13 can be a potential therapeutic target for colorectal cancer cells via inhibition of Nuclear factor-κB (NF-κB), and this was proven upon silencing MUC13, which led to increased sensitivity of colorectal cancer cells to the chemotherapeutic fluorouracil [176]. My findings for MUC17 potentially add another member of the mucin family to the list of targets for therapy, and endorse the generic findings for mucin family members as targets for chemo-sensitizing strategies.

MUC17 was found to be expressed differentially in various diseases, such as pancreatic ductal adenocarcinoma (PDAC). In the study it was found that MUC17 was expressed differentially between PDAC with and without lymph node metastasis, which suggested that MUC17 could be used as a survival marker in patients with resected PDAC [177]. Also, potential regulatory mechanisms acting on MUC17 in PDAC were identified, acting via hypoxia and methylation status. The study findings revealed that MUC17 was significantly induced by hypoxic stimulation through a hypoxia-inducible factor 1α (HIF1α)-dependent pathway in some pancreatic cancer cells. Also, the treatment of pancreatic cells (i.e. BxPC3) with 5-aza-2'-deoxycytidine (a methylation inhibitor) resulted in the restoration of hypoxic MUC17 induction. The data suggested that the HIF1α-mediated hypoxic signal pathway

contributed to MUC17 expression, and that DNA methylation of a hypoxia responsive element (a binding site for HIF1α) could be a determinant of the hypoxic inducibility of MUC17 in pancreatic cancer cells [178]. Furthermore, it was found that histone H3-K9 (H3-K9) modification status was also closely related to MUC17 expression and hypo-methylation status was observed in patients with PDAC. This indicated that the hypo-methylation status in the MUC17 promoter could be a novel epigenetic marker for the diagnosis of PDAC [179].

In colon cells, the extracellular regions of MUC17 have been shown to contain EGF-like Cys-rich segments (CRD1 and CRD2) connected by an intervening linker domain (L). It was found that reduced expression of MUC17 in LS174T colon cells was associated with reduced cell aggregation and reduced cell-cell adherence, and reduced cell migration. Whereas, exposure of colonic cell lines to exogenous recombinant MUC17-CRD1-L-CRD2 protein significantly increased cell migration and inhibited apoptosis. This indicated there was a potential role for MUC17-CRD1-L-CRD2 recombinant protein in the treatment of mucosal diseases of the colon and, this could provide a candidate for further development as a therapeutic agent target [180, 181].

Also, up-regulation expression of MUC17 was associated with some inflammatory conditions such as rheumatoid arthritis, osteoarthritic and chronic ulcerative colitis [182, 183]. The findings suggested there was a protective effect of MUC17 and diminished expression of MUC17 resulted in inflammatory and neoplastic conditions. In addition, it was found that MUC17 polymorphisms were associated with development of endometriosis and associated with infertility. Particularly, the “A” allele at rs10953316 was found to be protective against endometriosis induced infertility (statistically significant compared with the reference GG genotype, OR = 0.45; 95 % CI: 0.29-0.7) [184].

Taken all together, studies have shown so far that MUC17 is a protein-coding gene with functions related to many cellular biological activities and to many disease

including cancer and inflammatory conditions. Nevertheless, no previously published work has suggested a role for MUC17 in chemotherapeutic resistance mechanisms.

PCNX1 (Pecanex Homolog 1) is also a protein-coding gene. The encoded protein is a conserved transmembrane protein that is similar to the pecanex homologue in *Drosophila*. In humans, less is known, although there is a study suggesting PCNX1 might be involved in endoplasmic reticulum function since the protein was found to localise there, and lack of PCNX1 in the endoplasmic reticulum resulted in an enlargement defect [185]. In addition, PCNX1 protein was found to be a component of the Notch-signalling pathway and involved in the normal development of the nervous system in *Drosophila*. That was evidenced by the fact that absence of maternal expression of the PCNX1 gene resulted in embryos with severe hyperneuralization similar to the characteristics of Notch mutant embryos [186]. Although, the Notch-signalling pathway is commonly known for its role in embryogenesis and neuronal function and development [186], there are many reports that indicated the Notch-signalling pathway is also involved in chemotherapeutic resistance [186-189]. My findings for PCNX1 support this. It is also particularly interesting to note that NOTCH2 was also included within my initial list of candidate genes. In addition, this might indicate PCNX1 potentially mediates the chemotherapy response through modulating the Notch-signalling pathway. However, further insights into PCNX1 gene function and a demonstration of its candidacy for chemotherapy modulation *in vivo* are required.

Also, PCNX1 was found expressed exclusively in the germline cells of the testis in the rat, reaching its peak at the pachytene stage of the meiotic prophase. This suggests there is a potential regulatory role of PCNX1 protein in spermatogenesis, the details of which are not yet clear [190]. The most recent functional study of PCNX1 in context of cancer showed there was an oncogenic role for PCNX1 in Non Small Cell Lung Cancer (NSCLC). PCNX1 was found to positively regulates the mRNA and protein expressions of S-phase kinase associated protein 2 (Skp2) and correlates with its activities including promoting cell growth and proliferation, accelerating cell cycle and suppressing apoptosis. The study showed that PCNX1 acts as a competitive endogenous RNA (ceRNA) for SKP2, which indicates PCNX1 has similar oncogenic

activity to SKp2. Also, miR-26, miR-182, miR-340 and miR-506 were shown as the common miRNAs shared by Skp2 and PCNX1 and was shown that increase expression of miRNA significantly attenuated the mRNA level of Skp2 and PCNX1. Furthermore, silencing PCNX1 inhibits EGF-induced Akt phosphorylation, which can be reversed by the silencing of Dicer. This suggests PCNX1 may employ its oncogenic function at least in part via mediating Akt phosphorylation in NSCLC. In addition, the PCNX1-miRNA-Skp2 regulatory pattern was established as a molecular candidate for targeted therapy in NSCLC [191].

Interestingly, my colleague showed that miRNA-26 may act as protector of cells from chemotherapeutic drug *in vitro* (unpublished work). This supports our findings of upon silencing PCNX1 using siRNA showed increased cells survival when treated with chemotherapeutic drug, which might be attributed to miRNA-26 mechanism.

Despite the fact that there are relatively few studies on PCNX1, there is enough evidence to suggest that the gene may have oncogenic properties. Herein, in this study I add a functional role of PCNX1 as chemotherapy response mediator.

5.4.3. MUC17 and PCNX1 impact on drug loading, potentially via modulating ABC transporters activities

Having determined that MUC17 and PCNX1 had roles in defining chemo-response, I was interested to identify potential mechanisms of this function. Noted from published studies, a proposed mechanism for chemotherapy modulation in the mucin family, specifically MUC1, is modulation of ABC transporters activities and subsequently changes in the intracellular accumulation of the drug through drug efflux systems [164, 167, 192]. Although, there is nothing in the literature for PCNX1 suggesting this role in chemotherapy modulation, the assays were performed for both MUC17 and PCNX1.

I showed MUC17 expression reduction led to significant down-regulation of ABCC1 and ABCB1 genes, which encode the MRP-1 and MDR1 (P-glycoprotein) proteins

respectively. While, PCNX1 inhibition expression led to significant up-regulation of ABCG2 gene, which encodes the protein BCRP. In addition, the drug uptake data showed increased intracellular accumulation of epirubicin in MUC17 siRNA cells as compared with non-targeting siRNA control, while PCNX1 siRNA cells showed decrease intracellular drug uptake. This was in agreement with the changes in ABC transporters proteins expression, however, the findings were statistically non-significant.

My finding for MUC17 is an agreement with reports concerning another mucin family member, MUC1, for which it has been shown that overexpression of MUC1 directly, increases expression of ABCC1, ABCC3, ABCC5 and ABCB1. Subsequently, this led to enhanced chemotherapeutic drugs efflux of gemcitabine and etoposide in pancreatic cancer cells, which ultimately led to developing resistance [167]. Similar reports showed that overexpression of MUC1 was found to be associated with paclitaxel resistance and increased expression of ABCB1 in cervical and lung cancer paclitaxel-resistance cell lines. Also, the reduction of MUC1 expression increased cells sensitivity to paclitaxel drug and reduction of ABCB1 expression. Since, the EGFR pathway is involved in regulation of ABCB1 expression, it was also found that MUC1 induced ABCB1 expression potentially via cooperation with EGFR pathway [193]. While there are no reports concerning PCNX1 involvement in ABC transporters activities, herein I showed that inhibition of PCNX1 led to significant up-regulation of ABCG2 transporters, which could potentially explain the decrease in epirubicin intracellular loading (non-significant).

5.4.4. Conclusions

In this chapter I have identified MUC17 and PCNX1 as potential mediators of chemotherapy response in breast cancer, initially using a medium-throughput screening approach, but later focusing on these genes alone in more detail. I also considered using published expression data to validate the role of these genes in defining chemo-response in breast cancer, however, suitable large datasets are mostly lacking. Therefore, in the next chapter I examine whether MUC17 and PCNX1 protein expression defines response to chemotherapy using novel breast cancer cohorts assembled locally.

6. MUC17 expression predicts patient survival after chemotherapy treatment

6.1. Abstract

I have identified MUC17 and PCNX1 as potential drivers of chemotherapeutic response in breast cancer. I was next interested to test whether MUC17 and PCNX1 protein levels could predict breast cancer clinical outcomes after chemotherapy.

Breast cancer tissue was available in tissue microarrays from two cohorts of patients. First, resection samples from 140 patients treated with adjuvant chemotherapy, supported by extensive clinico-pathological data, including follow up of a median of 106 months. Secondly, resection samples from 53 patients treated with neoadjuvant chemotherapy, therefore samples were post-chemotherapy treatment, supported by clinico-pathological data, including follow up of a median of 46 months. Expression of MUC17 and PCNX1 was assessed using immunohistochemistry. Scores for the two markers were obtained successfully for 133 and 135 adjuvant cases, and 47 and 40 neoadjuvant cases, respectively.

MUC17 and PCNX1 proteins were expressed in cancer cells of the majority of breast cancers. MUC17 was generally expressed homogeneously within the cancer cells of individual cases, with variation between cases from weak to strong expression. PCNX1 showed some expression variation between cells in individual cases, with patterns ranging from weak expression in a minority of cells to strong expression in essentially all cells. Neither MUC17 or PCNX1 expressions were strongly associated with histopathological features such as receptor expression or grade. Kaplan-Meier survival analyses revealed that low MUC17 expression after neoadjuvant chemotherapy was significantly associated with longer disease free survival (log rank p=0.017), and this trend was also seen in the adjuvant cohort although it did not reach statistical significance. This relationship was in accordance with *in vitro* findings for MUC17, as a driver of chemoresistance. PCNX1 expression did not show significant

relationships with survival, although the non-significant trends visible were in accordance with *in vitro* findings defining PCNX1 as a driver of chemo-sensitivity.

I concluded that MUC17 is a driver of chemo-response in breast cancer, and may have value as a predictive biomarker.

6.2. Introduction

Cytotoxic chemotherapy has been used in combination with surgery as systematic adjuvant therapy to treat breast cancer for decades [194]. Chemotherapy has shown efficacy in eradication of clinically silent micro-metastases and thereby improving rates of recurrence-free and overall survival for many patients. However, it remains the case that about 30% of these patients treated with chemotherapy still suffer recurrences [4]. To date there are no clinical-used predictive molecular markers that could be used to select patients most likely to derive benefit from chemotherapy [195]. On the other hand, there have been many research efforts to identify prognostic markers that can be used to objectively evaluate patients' overall likely outcomes, such as the probability of cancer recurrence after standard treatment [196]. For example; it was found that high expression of BRCA1 confers worse prognosis in patients with breast cancer before treatment [197]. Also, patients with ER-positive breast tumours have better survival than patients with hormonal negative tumours [198], and patients with HER2-positive breast tumours are more aggressive and have worse prognosis compared to HER2-negative tumours [199]. However, the presence or absence of these prognostic markers can be useful for the selection of patients for chemotherapy treatment, but do not directly predict the response to that treatment [196].

I have identified MUC17 and PCNX1 as potential drivers of chemotherapeutic response in breast cancer using the novel strategy of analysing changes in representation of tumoural mutations within these genes after neoadjuvant chemotherapy (see chapter 4, section 4.3.5), and by performing functional experiments in a breast cancer cell line (see chapter 5, section 5.3.4). Given this

evidence, my next interest was to investigate whether these new potential markers can act as predictive markers for chemotherapy response and subsequent overall survival. A secondary benefit would be to add weight to my *in vitro* findings.

In order to assess the potential predictive value of these two markers, I have accessed cancer samples from two separate cohorts of breast cancer patients, supported by thorough clinico-pathological data and follow up. These tissue samples have been assembled into tissue micro-arrays (TMAs), which have become a preferred method to study protein expression on large-scale pre-defined cancer cohorts. The technique has shown its effectiveness in saving reagent costs and patient material, as well as making the staining and scoring procedures less time-consuming [199]. However, there are some drawbacks with use of TMAs, for example, concerns regarding the overall representation of the heterogeneity of the targeted protein expression within the tumour. This can be alleviated by taking multiple cores of different regions of the tumour in order to expand the representation of the intra-tumour heterogeneity, as is the case for the TMAs I have used. Also, applying statistical analyses to assess the variation of expression of the targeted protein between cores can justify the method, and aid the decision about how to combine multiple core scores into a single score for each cancer case [200].

In this work, I have used samples from a cohort of patients who received adjuvant chemotherapy, therefore the resection samples studied were taken before chemotherapy, and from a cohort of patients who received neoadjuvant chemotherapy, therefore the resection samples are post-chemotherapy. The aim of including the adjuvant chemotherapy cohort in my study was to investigate whether the expression of markers MUC17 and PCNX1 could be used to direct chemotherapy in the clinical settings. Also, the post-chemotherapy cohort (post-NAC) was included since the *in vitro* findings suggest that chemotherapy induces expression of MUC17 and PCNX1 expressions (chapter 5, section 5.3.5); therefore, it was possible that this induced/selected level is the level that would define whether the cells were resistant or sensitive and therefore predict outcome, rather than the basal level before treatment. It should be noted that both cohorts are mixed of different molecular sub-types as opposed to being only luminal A (the molecular sub-type of my focus in the

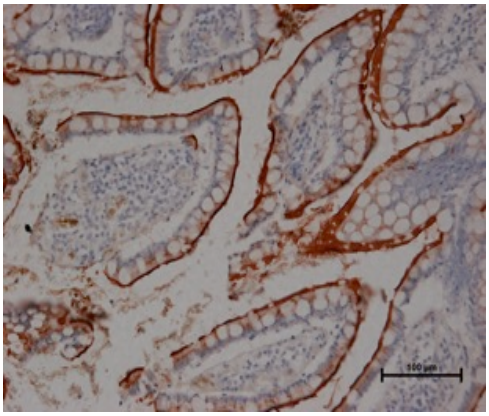
previous chapters) and received various chemotherapy regimens as opposed to only epirubicin/cyclophosamide regimen (the chemotherapy regimen of my initial cohort in Chapter 3). Nevertheless, this was a pragmatic decision based on having enough cases to study, but also based on the hope that markers that work in multiple molecular subtypes would be most use clinically.

To my knowledge this is first work of its kind in which protein levels of MUC17 and PCNX1 have been assessed as predictive markers for chemotherapy response in invasive breast cancer cohorts.

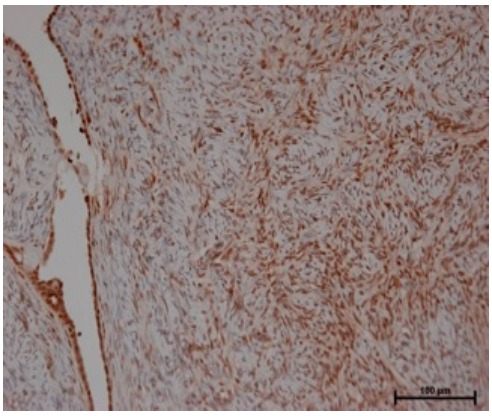
6.3. Results

6.3.1. Optimisation of immunohistochemical detection of MUC17 and PCNX1

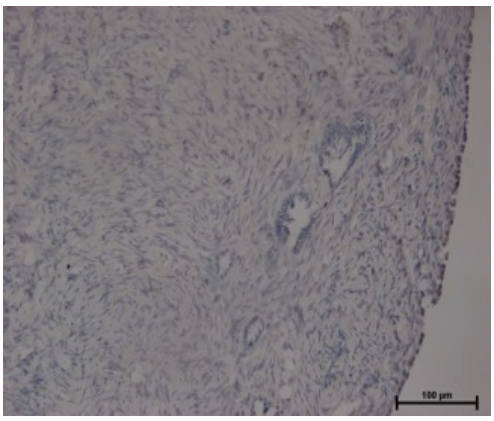
To date, there are no published studies that have used antibodies against MUC17 or PCNX1 in human invasive breast carcinoma tissues; therefore, I first had to perform optimisation steps to ensure specific staining of these proteins. Antibodies were selected that showed strong evidence of specificity from the manufacturer's data published on their website (Abcam; Cambridge, UK). In addition, the antibodies have worked successfully in different applications such as immunofluorescence for MUC17 and Western blots for PCNX1 (section 5.3.3). Based on the expression patterns for each protein shown within the Human Protein Atlas data (https://www.proteinatlas.org), I used small intestine tissue as a positive control for MUC17 expression and ovarian tissue as a positive control for PCNX1 expression. Omitting the primary antibody served as negative controls. A range of antigen retrieval, antibody concentrations, and antibody incubation times were used in order to identify conditions that allowed strong, specific-seeming staining (appropriate tissue location and low background staining) in the positive controls, and no staining in the negative controls (Figure 6.1). MUC17 staining was exclusively localised at the plasma membrane/cytoplasm of mature absorptive cells of small intestine villi, while, PCNX1 staining was confined mostly to the nucleoplasm of ovarian cells. Breast cancer tissue was also stained during optimisation: PCNX1 staining was observed but MUC17 staining was absent in the small number of cases used for optimisation (Figure 6.1).



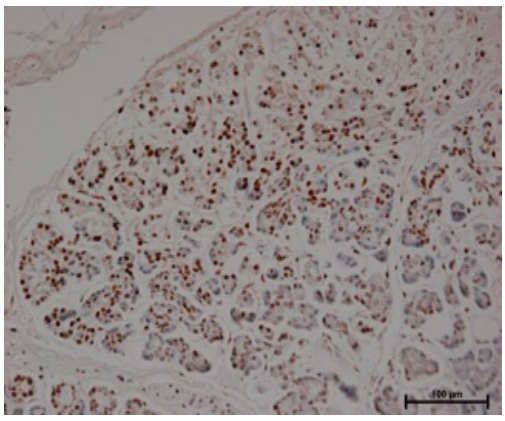
MUC17 positive control (small intestine)



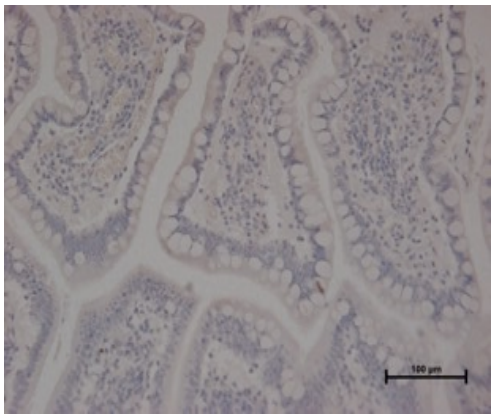
PCNX1 positive control (ovarian tissue)



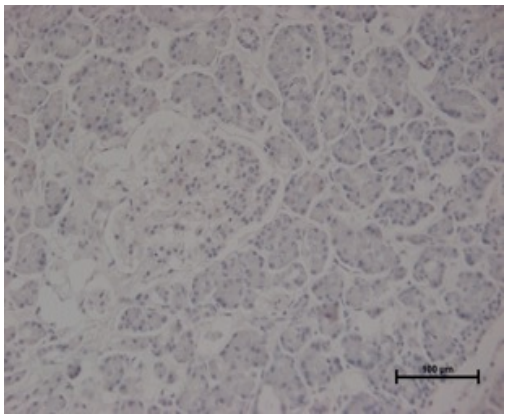
MUC17 in breast cancer



PCNX1 in breast cancer



MUC17 Negative control (small intestine)



PCNX1 negative control (ovarian tissue)

Figure 6.1 Immunohistochemical staining for MUC17 or PCNX1 in positive and negative controls, and in a single breast cancer case. The upper row shows MUC17 staining (left panel) and PCNX1 staining (right panel) in positive control tissues. The middle row shows MUC17 (left) and PCNX1 (right) staining in a breast cancer case. The lower row shows negative controls (no primary antibody in positive control tissues). Images are x20.

6.3.2. Analysis of MUC17 and PCNX1 expression in breast cancer cases treated with adjuvant chemotherapy

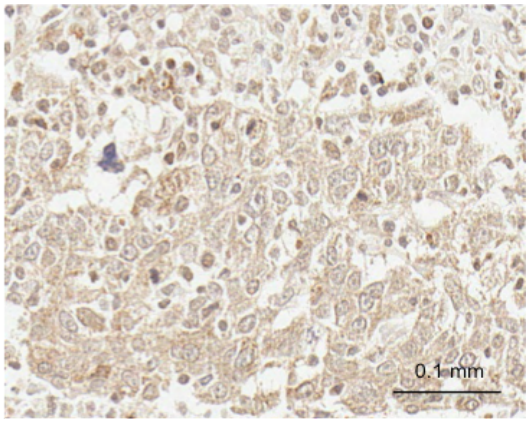
Tissue microarrays (TMAs) of resection samples from 140 primary breast cancer cases that had been treated with adjuvant chemotherapy were assembled by Stacey Jones (Clinical research fellow and surgical trainee, University of Leeds / Leeds Teaching Hospitals NHS Trust), comprising three independent tissue cores representing each case. The clinico-pathological features of this cohort are summarised in Table 6.1. I stained these TMAs for MUC17 and for PCNX1 using the optimised conditions determined above, and slides were digitally scanned for scoring (see Figures 6.2 and 6.3 for representative images). The purpose of having 3 tissue cores for each case was to take into account the possibility of intra-tumoural heterogeneity in the expression of the protein of interest. More cores also increased the chances of at least having one core successful scored for each case, since individual cores can be lost during the sectioning and staining process, which is a well-recognised issue with TMA-based studies [201].

Scoring systems for each antibody were developed in consultation with Prof. Andrew Hanby (consultant breast pathologist, Leeds Teaching Hospitals NHS Trust, and co-supervisor for the project) in order to semi-quantitatively score expression. Based on overall evaluation, it was found that MUC17 was located mainly at the plasma membrane and cytoplasm, while PCNX1 was located mainly in the nucleoplasm. For MUC17, scoring was based on cytoplasmic intensity, which was scored 0-3; 0 being negative, while 3 represented strong staining. For PCNX1, scoring was scored based on the proportion of cells showing positive nuclear staining (0-4) (0%=0, 1-5%=1, 6-25%=2, 26-75%=3, >75%=4) and the intensity of the nuclear staining (scores between 0-3), with final scores being the sum of these values. In addition, there were occasional cases that exhibited notable nuclear membranous positivity for MUC17 and cytoplasmic positivity for PCNX1, and such cases were separately reported. Half the cohort was scored by consensus between the author and Prof. Hanby together, providing robust scores based on Prof. Hanby's histopathology expertise and a

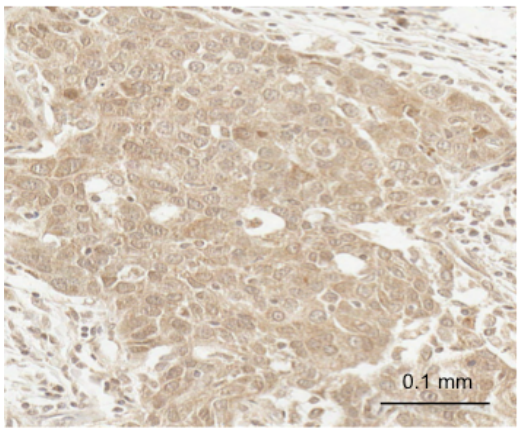
training period for the author. The second half of the cohort was scored by the author alone, and then 30% of these tissue cores were picked randomly for second (independent) scoring by Prof. Hanby to allow statistical analysis of scoring reproducibility (see section 6.3.2.1). The number of cases scored successfully, meaning at least one core was assessed, was 133/140 cases (95%) for MUC17, while for PCNX1 it was 135/140 cases (96.4%).

Characteristics	**Number (%)** **n=140**
Histological type	
Ductal no-special type	110 (78.6)
Lobular	8 (5.8)
Metaplastic	4 (2.8)
Tubular	1 (0.7)
Medullary	1 (0.7)
Mixed	16 (11.4)
Tumour grade	
1	7 (5)
2	55 (39.3)
3	78 (55.7)
Lymph node status	
At least 1 positive node	91 (65)
No positive nodes	49 (35)
ER receptors status	
Positive	101 (72.2)
Negative	39 (27.8)
PR receptors status	
Positive	72 (51.4)
Negative	58 (41.4)
Unknown	10 (7.2)
Her2 status	
Positive	24 (17.2)
Negative	116 (82.8)
Chemotherapy regimens	
Anthracycline based regimen	
Without Taxanes	66 (47.2)
With Taxanes	57 (40.7)
With carboplatin or/and Capecitabine	17 (12.1)

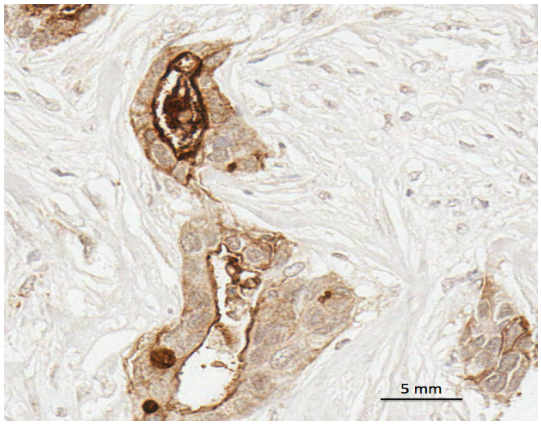
Table 6.1 Summary of the clinico-pathological features for the adjuvant chemotherapy cohort.



Score= 1

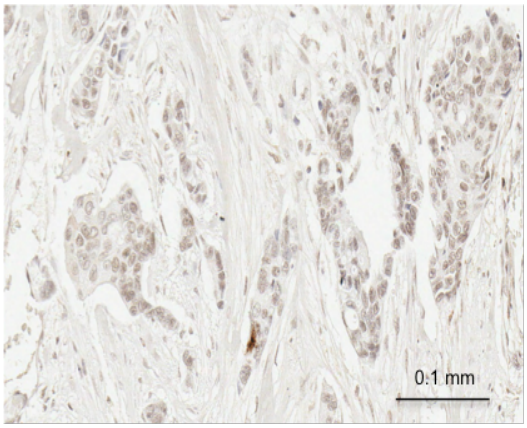


Score=3



Membrane positivity

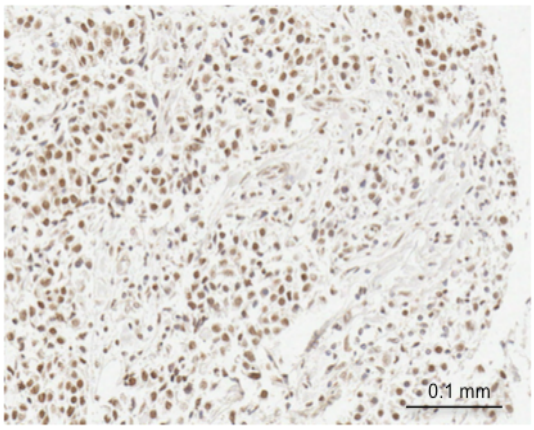
Figure 6.2 Representative images of immunohistochemical staining for MUC17 illustrating the scoring system. MUC17 scoring was based on cytoplasmic intensity (0-3), with occasional cases reported as showing membrane positivity. Images are 20x or 40x.



Intensity=2

Proportion=3

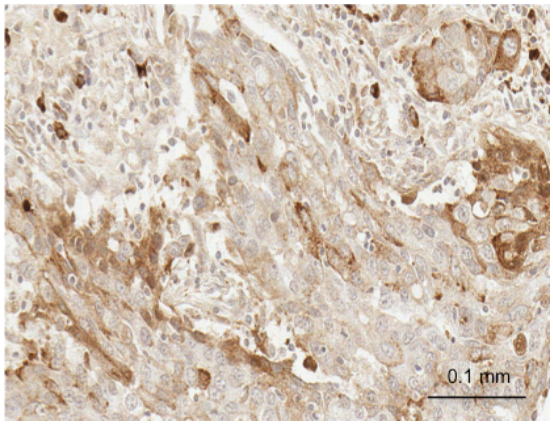
Total score=5



Intensity=3

Proportion=4

Total score=7



Cytoplasmic positivity

Figure 6.3 Representative images of immunohistochemical staining for PCNX1 in breast cancers, illustrating the scoring system. PCNX1 scoring was based on nuclear staining intensity (0-3) and proportion of positive cells (0-4); these values were summed giving a final score of 0-7 (a score of 1 is not possible). Also, occasional cases showed cytoplasmic positivity for PCNX1 were reported. Images are 20x.

6.3.2.1. Concordance between scorers was high, and core to core variability within cases was low

The Cohen's Kappa statistic was used to assess the inter-scorer concordance in the cores that were double-scored independently by the author and by Prof. Hanby. This statistical method takes into account the chances of false positive agreement between scorers due to the possibility of scorers being uncertain what to enter and simply making random guesses. Hence it represents a better statistical method than conventional methods such as percentage scoring [202, 203]. The Kappa scores showed substantial agreement between the scorers for PCNX1, while for MUC17 it was near perfect (0.60-0.79 can be regarded as substantial, while 0.80-0.90 can be regarded as near perfect (20)) (Table 6.2).

	K score	**95% confidence Interval**
MUC17	0.804	0.566-1.00
PCNX1	0.694	0.321-0.820

Table 6.2 Concordance between independent scorers was high. TMAs containing primary breast cancer cores were stained for MUC17 or PCNX1 expression and expression levels were determined manually semi-quantitatively. 30% of the cases were double-scored by independent assessors. Cohen's Kappa statistics was used to assess inter-scorer concordance and are shown, along with 95% CIs for each of MUC17 and PCNX1.

Since cases in the TMAs were mostly in triplicate or duplicate cores, there was also a need to assess the core-to-core variability in order firstly to validate the use of TMAs and secondly to aid the decision concerning the pooling method by which core scores would be combined to give scores for each case. If core-to-core variability was substantial this might suggest expression was sufficiently heterogeneous that analysis on TMAs was inappropriate. Should this not be the case, available options for pooling methods included taking lowest or highest score among cores, and taking the mean

or median of scores. Spearman's rho correlation coefficients were determined to assess correlations between scores for triplicate or duplicate scores, randomly assigned as core 1, 2 or 3 (Table 6.3). Scores were strongly and significantly correlated in all cases (rho values between 0.855 and 0.930). Thereby, I concluded that there is relatively little intra-tumoural heterogeneity in distribution of expression for these proteins, and therefore that different choices of pooling methods were unlikely to have strong impact on the final result; I selected taking the mean score for pooling.

Spearman's correlation coefficient		**Core 1**	**Core 2**	**Core 3**
MUC17	Core 1	1.0	0.924	0.893
	Core 2		1.0	0.930
	Core 3		0.930	1.0
PCNX1	Core 1	1.0	0.894	0.855
	Core 2		1.0	0.891
	Core 3		0.891	1.0

Table 6.3 Core-to-core correlation using Spearman's correlation coefficients for MUC17 and PCNX1 showed strong correlation between cores. TMAs containing primary breast cancer cores (up to three cores per case) were stained for MUC17 or PCNX1 expression and expression levels were determined manually semi-quantitatively. Expression scores for cores from the same cases were compared using Spearman's correlation analyses. All correlations were significant, *p*-value <0.001.

6.3.2.2. MUC17 protein is expressed at high levels relatively rarely, while PCNX1 protein is expressed highly more frequently in invasive breast cancers

The distribution of protein expression scores for MUC17 and PCNX1 across the breast cancer cohort is illustrated in histograms in Figure 6.4; these scores represent the mean score of all cores successfully scores for that case, rounded to the nearest whole

number for the purpose of data presentation. A range of expression levels was shown for both proteins, with some cases negative and some showing strong expression. For MUC17, a majority of cases were scored as either negative or weak (59%), while only 19% demonstrated the highest expression. As for PCNX1, the majority of cases (68%) showed high expression (scores 6 and 7), while the rest of cases spread across 0-5 scores.

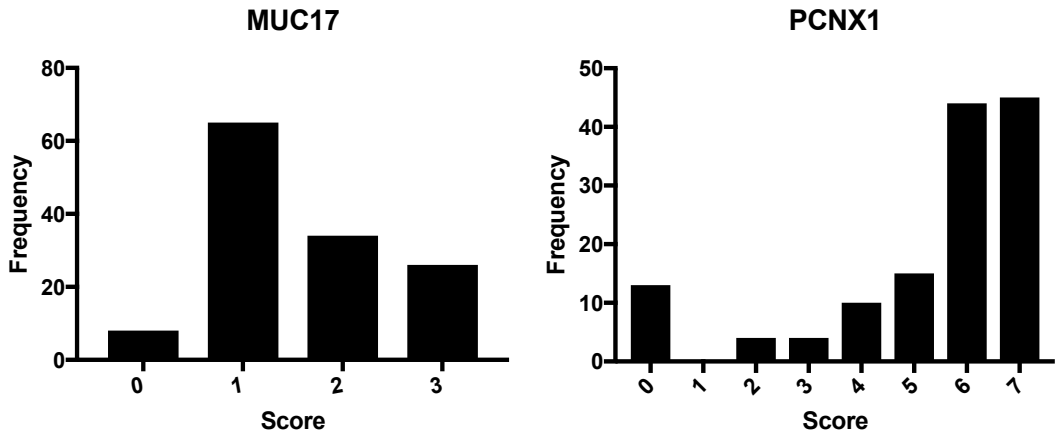


Figure 6.4 Scores distributions for MUC17 and PCNX1 in breast cancer cases treated with adjuvant chemotherapy. TMAs containing primary breast cancer cases were stained for MUC17 or PCNX1 expression and expression levels were determined semi-quantitatively. Individual tumour cores were scored to quantify expression on a scale of 0-3 (MUC17) or 0-7 (PCNX1), and the mean scores calculated for each case to combine multiple core scores. Case scores were rounded to the closest whole integer number in these distribution plots.

6.3.2.3. Protein expression of MUC17 and PCNX1 do not correlate with clinical prognostic markers in breast cancer cases treated with adjuvant chemotherapy

Correlations between MUC17 or PCNX1 protein expressions (mean score of cores for each case) and some clinical prognostic factors were tested using Spearman's rho analyses. This analysis included oestrogen receptors status (positive or negative, with >2 by Allred scoring considered positive), tumour grade (1, 2, or 3), lymph nodes metastasis (positive or negative) and molecular subtype (0 for other subtypes and 1 for triple negative subtype (TNBC)). The TNBC subtype was selected for correlation

since TNBC has high rates of metastasis capability, and evidence suggests that the risk factor profiles differ between TNBC and the more common luminal subtypes [204]. In addition, TNBC is almost always treated with chemotherapy. A summary table shows the correlations scores and the associated *p* values for these analyses (Table 6.4); there were no significant correlations between protein expression of MUC17 and PCNX1 and any of these prognostic factors, suggesting that expression of these potential markers is unrelated to these main cancer classifications that correlate with breast cancer behaviour.

		ER expression	Grade	Lymph nodes metastasis	Triple Negative sub-type
MUC17	r	0.115	0.080	-0.130	-0.103
	p	0.187	0.358	0.137	0.239
PCNX1	r	0.043	0.030	-0.16	-0.26
	p	0.618	0.724	0.858	0.768

Table 6.4 MUC17 and PCNX1 expression does not correlate with standard clinical prognostic factors. Spearman's rho analyses were performed for MUC17 and PCNX1 expression levels against the factor as shown. r indicates Spearman's rho coefficient and *p* indicates P-value.

6.3.2.4. Kaplan-Meier survival analysis revealed no significant differences in survival relating to high or low MUC17 and PCNX1 protein expression

A key objective in conducting these TMA-based studies on an adjuvant cohort was to investigate whether MUC17 or PCNX1 expression at the protein level could be used as predictive markers for survival outcomes after chemotherapy. In order to test this by Kaplan-Meier analyses it was necessary to dichotomise the cohort into patients regarded as having high expression and those with low expression using suitable cut-

off values, as this would allow comparison between the two groups. I used Receiver Operating Characteristic (ROC) curve analyses to perform this dichotomisation, which objectively determines the best balance between sensitivity and specificity for predicting the defined clinical outcome of recurrence or death from all the possible cut-off scores [205, 206]. The analysis was performed with the two different clinical outcomes, death or recurrence. Based on the area under the ROC curve, which is a measure of how well a parameter can distinguish between positive and negative groups, death status was chosen as clinical outcome to dichotomise the protein expression for MUC17 and recurrence status was chosen for PCNX1. Cut-off scores were established as follows: 1.1 for MUC17 and 5.6 for PCNX1 (more details and the ROC curve graphs can be found in Appendix Figure 9.4).

Kaplan-Meier survival analyses were then performed to assess if differential expression of MUC17 or PCNX1 proteins was associated with survival outcomes, namely, disease free survival (DFS), for which the event was recurrence, and disease specific survival (DSS), for which the event was death from cancer. Initially, I performed low expression versus high expression analyses for both DFS and DSS for both MUC17 and PCNX1 (Figure 6.5 and Table 6.5). There were no statistically significant differences between the low and high expressing groups.

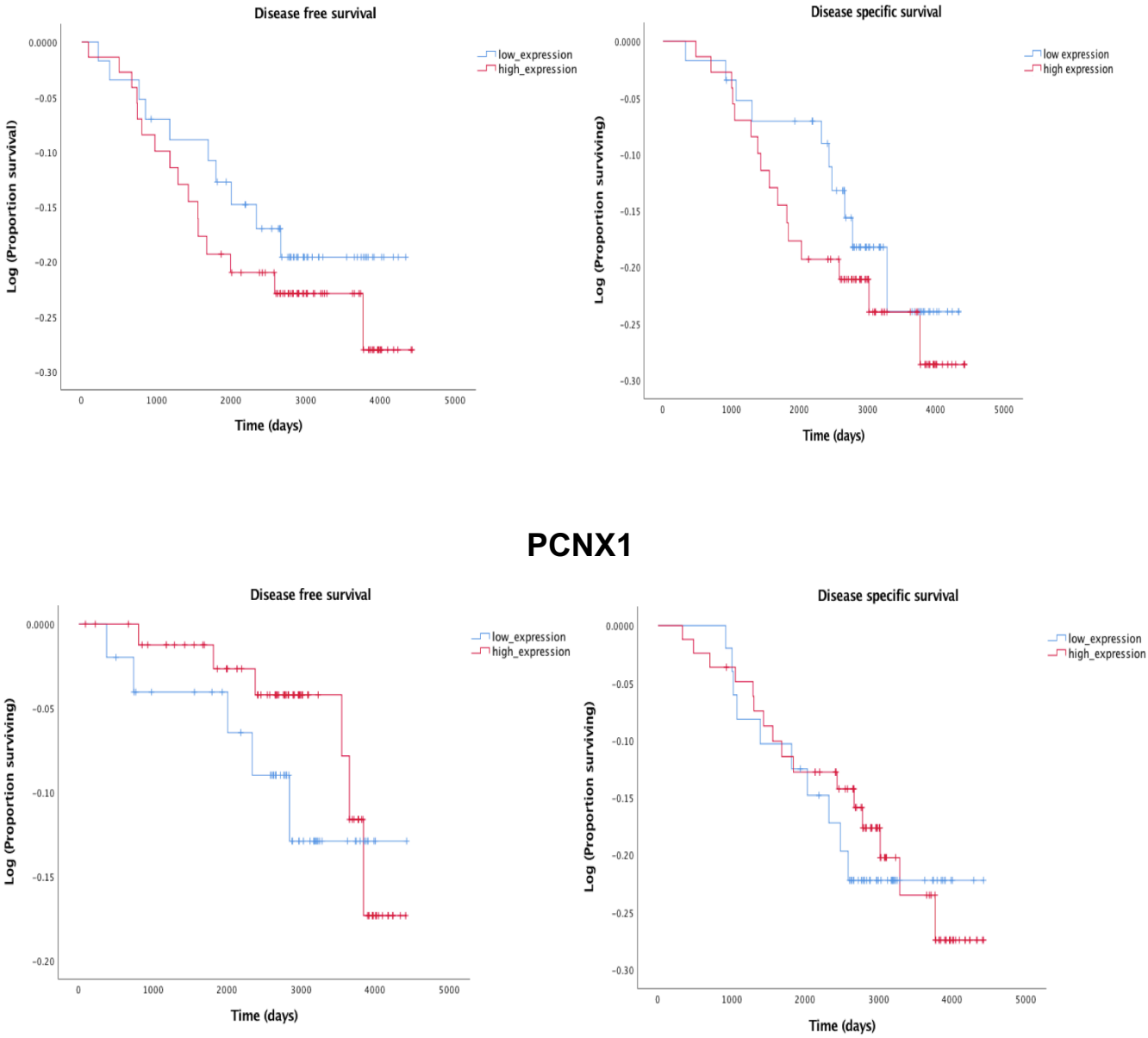


Figure 6.5 Kaplan-Meier survival analysis for breast cancer outcomes in groups with high versus low expression of either MUC17 or PCNX1. Expression of MUC17 and PCNX1 was determined in resection samples from a cohort of 140 primary breast cancers patients who were treated with adjuvant chemotherapy. The cohort was dichotomised into relatively low or relatively high expression groups using ROC analyses. Relative survival is shown for the groups, and significance was tested using the log rank test. The upper plots show data based on MUC17 expression, while the lower plots are for PCNX1. Left panels show the end point of disease free survival, while right panels show the end point of disease specific survival. The small coloured vertical lines on the plots represent the end of follow up (censor points) for individual patients.

		Mean DFS (days) (95% CI)	Log Rank	Mean DSS (days) (95% CI)	Log Rank
MUC17	Low	3693 (3375-4010)	0.623	3822 (3611-4147)	0.531
	High	3634 (3326-3941)		3737 (3520-4076)	
PCNX1	Low	4110 (3846-4374)	0.461	3875 (3563-4187)	0.860
	High	4216 (4066-4366)		3878 (3631-4126)	

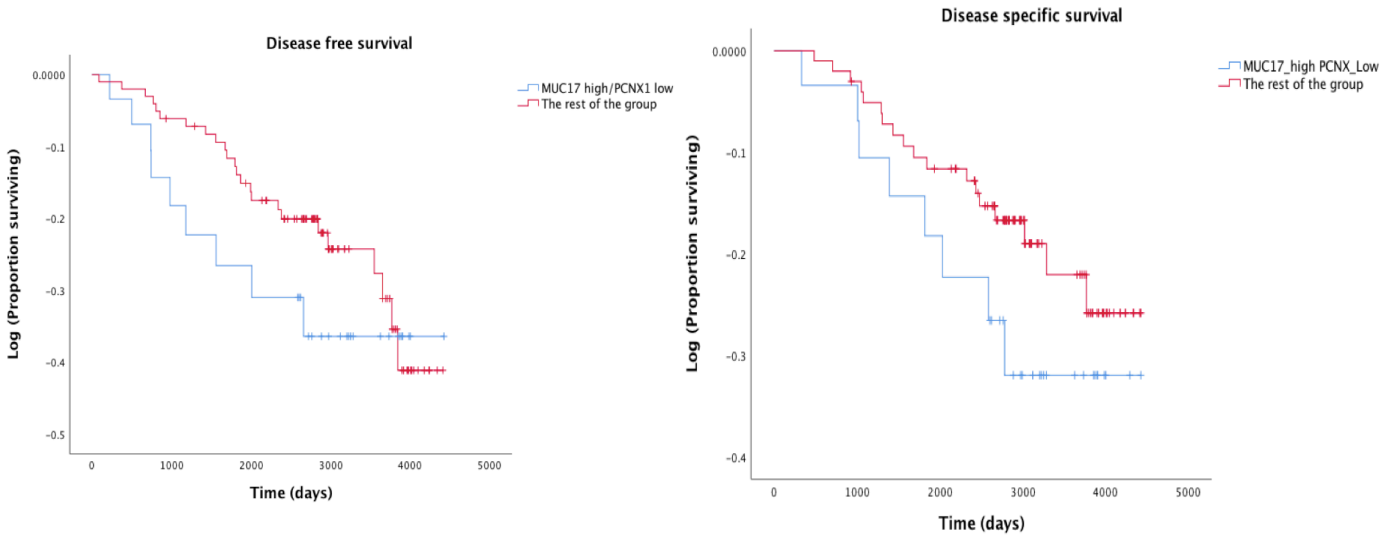
Table 6.5 Comparison of mean survival between low and high protein expression in MUC17 and PCNX1. Analyses were performed as described for Figure 6.5. Here, mean disease free survival (DFS) or disease specific survival (DSS) in the low and high expression groups are compared, showing 95% confidence intervals and log rank *p* values.

Next, I performed a combined analysis of MUC17 and PCNX1, basing my method of combining on my *in vitro* findings (chapter 5). I previously showed, *in vitro*, that low MUC17 expression (induced by siRNA treatment) was significantly associated with increased sensitivity to chemotherapy, from which I would infer low MUC17 might be associated with improved survival. For PCNX1, low expression (induced by siRNA) was associated with resistance to chemotherapy, and therefore may associate with poor survival. Thus, I identified the group of patients with both low MUC17 expression and high PCNX1 expression, both of which may be associated with improved survival.

I compared this group with the remainder of the cohort (those with high MUC17 or with low PCNX1) by Kaplan-Meier survival analysis as before (Figure 6.6A and Table 6.6). Although the patients with low MUC17 and high PCNX1 showed better survival compared to the rest, as predicted (by 347 days for DFS or 332 days for DSS), this difference was not statistically significant. In addition, I also identified the group of patients with both high MUC17 expression and low PCNX1 expression, an expression

profile that both might be associated with poor survival based on the *in vitro* findings. As before, I compared this group with the remainder of the cohort (those with low MUC17 or high PCNX1) by Kaplan-Meier survival analysis (Figure 6.6B and Table 6.6). Although patients with high MUC17 and low PCNX1 had reduced DFS and DSS (by 291 days and 247 days respectively) when compared to the remainder of the cohort, as expected, this difference was again not statistically significant.

A



B

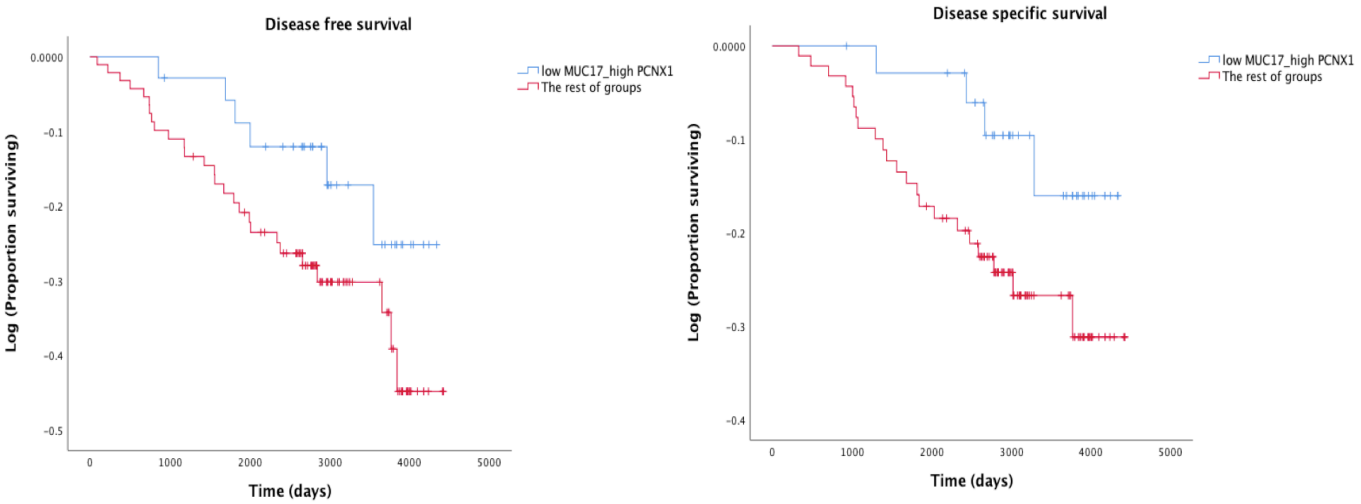


Figure 6.6 Kaplan-Meier survival analysis for breast cancer outcomes in groups with combined MUC17/PCNX1 markers. Expression of MUC17 and PCNX1 was determined in resection samples from a cohort of 140 primary breast cancers patients who were treated with adjuvant chemotherapy. The cohort was grouped into low MUC17/high PCNX1 or reminder of the cohort (A), and high MUC17/low PCNX1 or reminder of the cohort (B). Relative survival is shown for the groups, and significance was tested using the log rank test. Left panels show the end point of disease free survival, while right panels show the end point of disease specific survival. The small coloured vertical lines on the plots represent the end of follow up (censor points) for individual patients.

	Mean DFS (days) (95% CI)	**Log Rank**	**Mean DSS (days) (95% CI)**	**Log Rank**
MUC17 low/ PCNX1 high	3915 (3601-4229)	0.153	4080 (3837-4322)	0.133
The rest of the patients	3568 (3286-3850)		3748 (3494-4003)	
MUC17 high/ PCNX1 low	3440 (2887-3993)	0.459	3664 (3194-4133)	0.316
The rest of the patients	3731 (3489-3974)		3911 (3692-4130)	

Table 6.6 Comparison of mean survival between low MUC17/high PCNX1 or high MUC17/low PCNX1 and reminder of the cohort. Analyses were performed as described for Figure 6.6. Here, mean disease free survival (DFS) or disease specific survival (DSS) in the corresponding groups were compared, showing 95% confidence intervals and log rank *p* values.

Finally, I compared survival between the low MUC17 / high PCNX1 group (which should correspond to drug sensitivity) and the high MUC17 / low PCNX1 (which should correspond to drug resistance) using Kaplan-Meier analyses (Figure 6.7 and Table 6.7). There was longer survival in terms of DFS, by 475 days, and DSS, by 416 days, in the low MUC17 / high PCNX1 group as compared to the reverse group, which was in agreement with the *in vitro* findings. However, the difference remained non-significant, which may be related in part to the relatively small size of these two groups (36 and 30 patients in each, respectively).

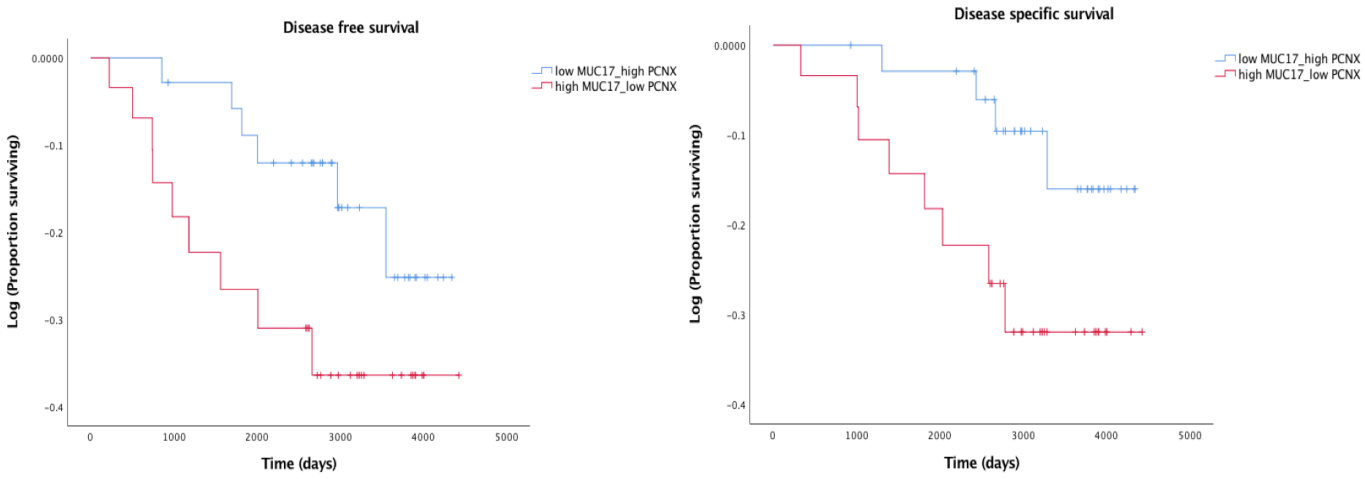


Figure 6.7 Kaplan-Meier survival analysis for breast cancer outcomes in groups with combined MUC17/PCNX1 markers. Expression of MUC17 and PCNX1 was determined in resection samples from a cohort of 140 primary breast cancers patients who were treated with adjuvant chemotherapy. The cohort was dichotomised into low MUC17 / high PCNX1 or high MUC17 / low PCNX1 based on *in vitro* findings. Relative survival is shown for the groups, and significance was tested using the log rank test. Left graph shows the end point of disease free survival, while right graph shows the end point of disease specific survival. The small coloured vertical lines on the plots represent the end of follow up (censor points) for individual patients.

	Mean DFS (days) (95% CI)	Log Rank	Mean DSS (days) (95% CI)	Log Rank
MUC17 low/ PCNX1 high	3915 (3601-4229)	0.169	4080 (3837 -4322)	0.101
MUC17 high/ PCNX1_low	3440 (2887 -3993)		3664 (3194-4133)	

Table 6.7 Comparison of mean survival between low MUC17 / high PCNX1 or high MUC17 / low PCNX1. Analyses were performed as described for Figure 6.7. Here, mean disease free survival (DFS) or disease specific survival (DSS) in the corresponding groups were compared, showing 95% confidence intervals and log rank *p* values.

6.3.3. Analysis of MUC17 and PCNX1 expression after treatment with neoadjuvant chemotherapy in breast cancer

An alternative breast cancer cohort was also available. TMAs of resection samples from 53 primary breast cancer patients who were treated with neoadjuvant chemotherapy, therefore samples were post-chemotherapy, was also assessed. This was available through collaboration with Dr Abeer Shaaban (consultant breast pathologist, formerly at Leeds Teaching Hospitals NHS Trust, where the patients included in this TMA were treated, but now at University Hospitals Birmingham NHS Foundation Trust). The TMAs comprised two independent cores representing each case. The clinico-pathological features of this cohort are summarised in Table 6.8. The cases were stained for MUC17 and PCNX1 and scored as previously. The number of cases scored successfully was 47/53 cases (89%) for MUC17, and 40/53 cases (75%) for PCNX1. The distribution of protein expression scores for MUC17 and PCNX1 across the neoadjuvant chemotherapy cohort is illustrated in Figure 6.8, using mean scores of successfully scored cores for each case. A range of expression levels was seen for both proteins, as for the adjuvant cohort. In this case, a large majority of cases were either negative or weak (89%) for MUC17, while half of cases (50%) were positive for PCNX1, which represent a similar overall pattern as for the adjuvant cohort, although overall these scores were lower.

Characteristics	Number (%) n=53
Histological type	
Ductal no-special type	39 (73.6)
Lobular	2 (3.8)
Metaplastic	1 (1.7)
Tubular	1 (1.7)
Mucinous	1 (1.7)
Mixed	6 (11.3)
Tumour grade	
1	4 (7.5)
2	25 (47.2)
3	24 (45.3)
Tumour size	
1 (<2 cm)	15 (28.3)
2 (2-5 cm)	26 (49)
3 (>5 cm)	12 (22.7)
Lymph node status	
At least 1 positive node	33 (62.3)
No positive nodes	20 (37.7)
ER receptors status	
Positive	16 (30.2)
Negative	32 (60.4)
Unknown	5 (9.4)
PR receptors status	
Positive	27 (50.9)
Negative	21 (39.6)
Unknown	5 (9.5)
Her2 status	
Positive	38 (71.7)
Negative	10 (18.9)
Unknown	5 (9.4)
Chemotherapy regimens	
Anthracycline based regimen:	
Without Taxanes	17 (32.1)
With Taxanes	27 (50.9)
With Capecitabine	9 (17)

Table 6.8 Summary clinico-pathological features for the neoadjuvant chemotherapy cohort.

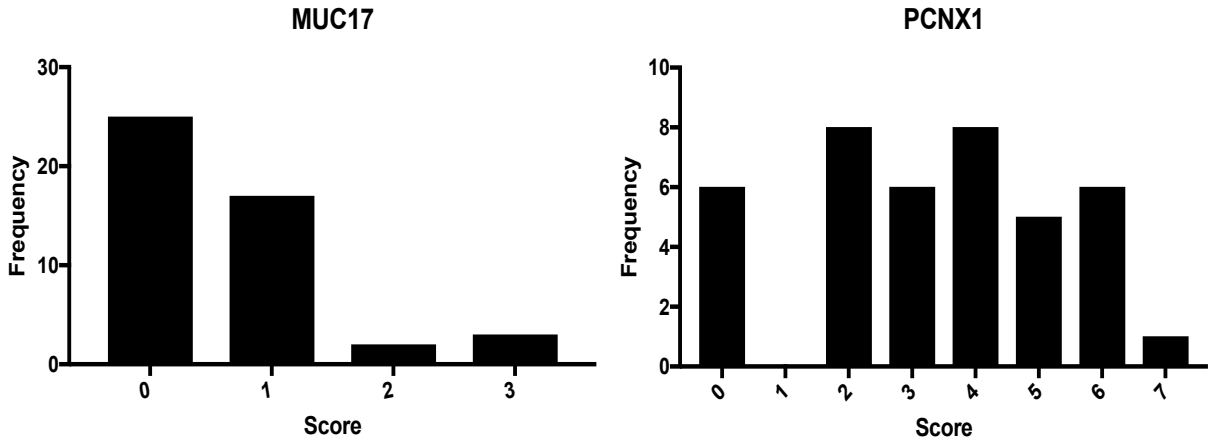


Figure 6.8 Scores distributions for MUC17 and PCNX1 in breast cancer cases treated with neoadjuvant chemotherapy. TMAs containing primary breast cancer cases were stained for MUC17 or PCNX1 expression and expression levels were determined semi-quantitatively. Individual tumour cores were scored to quantify expression on a scale of 0-3 (MUC17) or 0-7 (PCNX1), and the mean scores calculated for each case to combine multiple core scores. Case scores were rounded to the closest whole integer number in these distribution plots.

6.3.3.1. Protein expression of MUC17 and PCNX1 do not correlate with clinical prognostic markers in breast cancer cases treated with neoadjuvant chemotherapy

Expression levels of MUC17 and PCNX1 in this cohort were also tested for correlations with standard clinical prognostic factors. A summary table shows the Spearman's correlation coefficients and the associated *p*-values for these tests (Table 6.9). Only one significant correlation was noted at the threshold of $p<0.05$; this was a relatively weak negative correlation between MUC17 protein expression and cases with the triple negative molecular subtype (Spearman's rho -0.291, p =0.047). However, using a target p value adjusted for multiple testing (8 tests in this analysis, so target p value adjusted to 0.00625), this finding lost significance.

	ER expression	Grade	Lymph nodes metastasis	Triple Negative sub-type
MUC17 r	0.101	-0.048	-0.072	-0.291
	0.574	0.752	0.636	0.047
PCNX1 p	0.108	-0.102	-0.221	0.106
	0.557	0.541	0.177	0.508

Table 6.9 MUC17 and PCNX1 expression does not correlate with standard clinical prognostic factors. Spearman's rho analyses were performed for MUC17 and PCNX1 expression levels against the factor as shown. r indicates spearman's rho coefficient and *p* indicates P-value.

6.3.3.2. Low MUC17 protein expression after neoadjuvant chemotherapy is associated with longer disease free survival

As described before (section 6.3.2.4), ROC curve analysis were performed in order to determine cut-off scores for each marker, above which would be regarded as high expression, and below which would be regarded as low expression. Recurrence status was used as a clinical outcome to dichotomise the expression into low and high expression for both MUC17 and PCNX1. Cut-off scores were established based on the coordinates of the curves as follows: 0.5 for MUC17 and 3.5 for PCNX1 (more details and ROC graphs can be found in Appendix Figure 9.5).

Initially, I performed Kaplan-Meier survival analysis for low expression versus high expression for both DFS and DSS for both MUC17 and PCNX1 (Figure 6.9 and Table 6.10). PCNX1 expression did not determine significant differences in survival for either DSS or DFS, although for both end-points lower PCNX1 expression was associated

with shorter mean survival (by 594 days for DFS, and by 392 for DSS). However, for MUC17, DFS was significantly longer for patients with low MUC17 expression (by 823 days) (log rank, p=0.017), and similarly, DSS was extended (by 816 days) although this was not statistically significant. Both the significant findings, and the non-significant trends, for all of these analyses were compatible with the trends seen in the adjuvant cohort (section 6.3.2.4), and with the results of the *in vitro* studies (chapter 5, section 5.3.4).

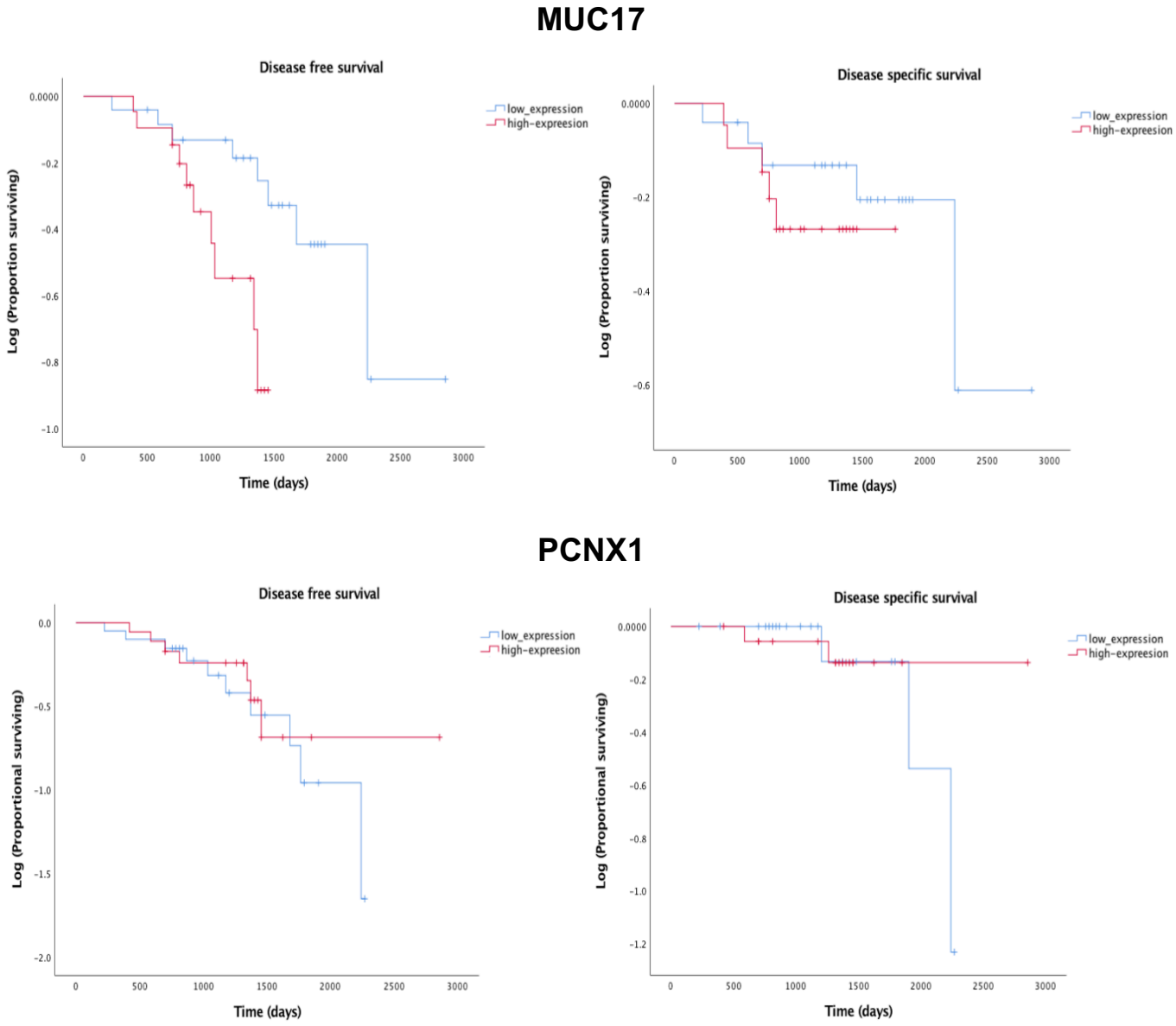


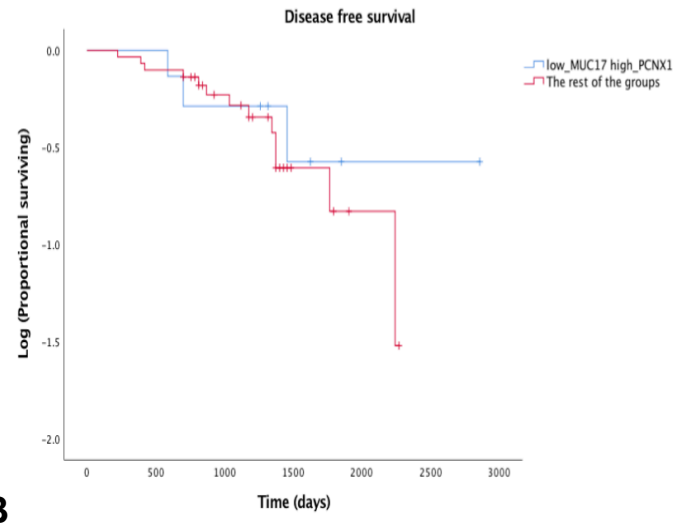
Figure 6.9 Kaplan-Meier survival analysis for breast cancer outcomes in groups with high versus low expression of MUC17 showed significantly improved DFS. Expression of MUC17 and PCNX1 was determined in resection samples from a cohort of 53 primary breast cancers patients who were treated with neoadjuvant chemotherapy. The cohort was dichotomised into relatively low or relatively high expression groups using ROC analyses. Relative survival is shown for the groups, and significance was tested using the log rank test. The upper plots show data based on MUC17 expression, while the lower plots are for PCNX1. Left panels show the end point of disease free survival, while right panels show the end point of disease specific survival. The small coloured vertical lines on the plots represent the end of follow up (censor points) for individual patients.

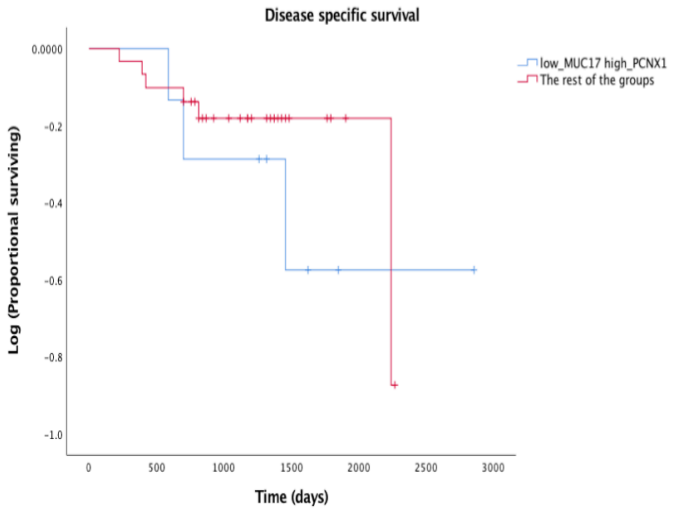
		Mean DFS (days) (95% CI)	Log Rank	Mean DSS (days) (95% CI)	Log Rank
MUC17	Low	2103 (1704-2503)	0.017*	2311 (1894-1680)	0.431
	High	1280 (1507-1344)		1495 (1286-1703)	
PCNX1	Low	2020 (1751-2290)	0.739	1576 (3563-4187)	0.655
	High	2614 (1902-2693)		1968 (1466-2470)	

Table 6.10 Comparison of mean survival between low and high protein expression in MUC17 and PCNX1. Analyses were performed as described for Figure 6.9. Here, mean disease free survival (DFS) or disease specific survival (DSS) in the low and high expression groups are compared, showing 95% confidence intervals and log rank *p* values. * indicates significant log rank test ($p<0.05$)

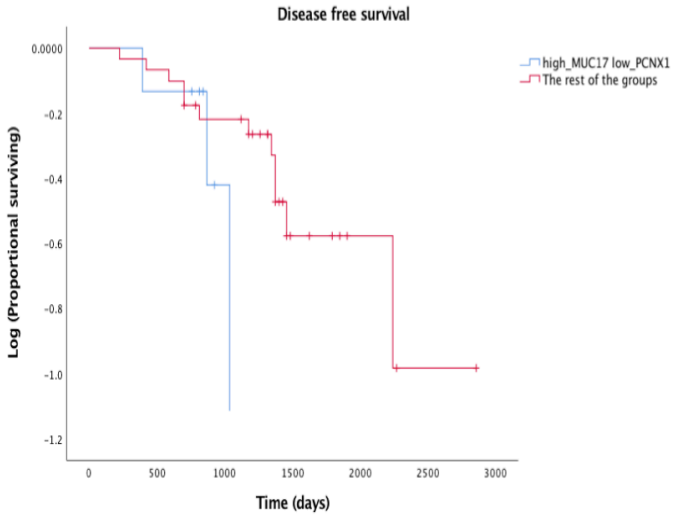
Next, I performed the same sequence of combined analyses of MUC17 and PCNX1 expression as previously for the adjuvant chemotherapy cohort (section 6.3.2.4). Analyses were performed comparing patients with low MUC17 expression / high PCNX1 expression with the remainder of the cohort (Figure 6.10A; Table 6.11), patients with high MUC17 / low PCNX1 with the remainder of the cohort (Figure 6.10b; Table 6.11), and patients with low MUC17 / high PCNX1 with patients with high MUC17 / low PCNX1 (Figure 6.11; Table 6.12). In each case, these combinations were guided by the *in vitro* findings suggesting that low MUC17 expression and high PCNX1 expression should both confer longer survival because of increased sensitivity to chemotherapy, and the reverse for the high MUC17 / low PCNX1 group. None of these analyse showed significant differences between the groups, although in every case the (non-significant) differences between the mean lengths of survival in the two groups were compatible with the hypothesis behind the combinations. In particular, for example, the low MUC17 / high PCNX1 and high MUC17 / low PCNX1 group comparison showed substantial differences in mean survival of 883 days for DFS and 448 days for DSS, although the groups were too small to allow significance (8 patients in each).

A





B



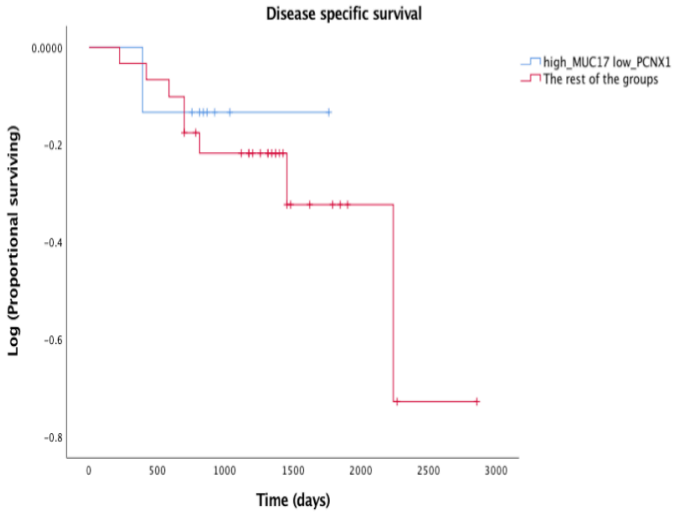


Figure 6.10 Kaplan-Meier survival analysis for breast cancer outcomes in groups with combined MUC17/PCNX1 markers. Expression of MUC17 and PCNX1 was determined in resection samples from a cohort of 53 primary breast cancers patients who were treated with neoadjuvant chemotherapy. The cohort was grouped into low MUC17 / high PCNX1 or reminder of the cohort (A) and high MUC17 / low PCNX1 or reminder of the cohort (B). Relative survival is shown for the groups, and significance was tested using the log rank test. Left panels show the end point of disease free survival, while right panels show the end point of disease specific survival. The small coloured vertical lines on the plots represent the end of follow up (censor points) for individual patients.

	Mean DFS (days) (95% CI)	Log Rank	Mean DSS (days) (95% CI)	Log Rank
MUC17 low/ PCNX1 high	2040 (1325-2755)	0.598	2040 (1325-2755)	0.502
The rest of the patients	1613 (1345-1882)		1966 (1734-2199)	
MUC17 high/ PCNX1 low	1157 (655-1659)	0.145	1592 (1278-1906)	0.202
The rest of the patients	1930 (1548-2312)		2147 (1741-2553)	

Table 6.11 Comparison of mean survival between low MUC17/high PCNX1 or high MUC17 / low PCNX1 and reminder of the cohort. Analyses were performed as described for Figure 6.10. Here, mean disease free survival (DFS) or disease specific survival (DSS) in the corresponding groups were compared, showing 95% confidence intervals and log rank *p* values.

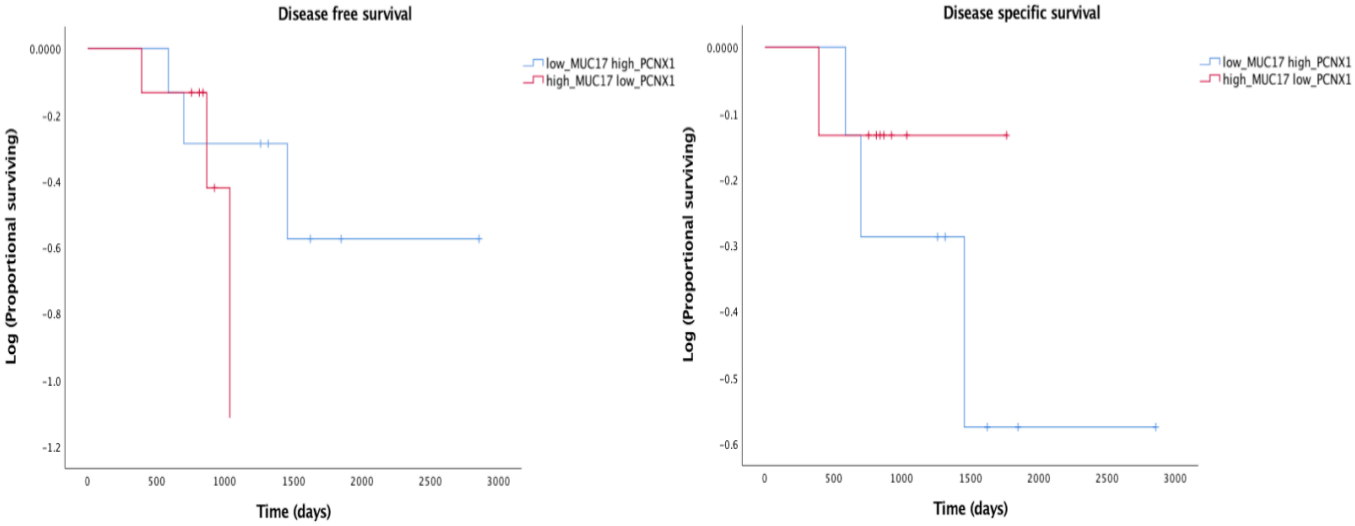


Figure 6.11 Kaplan-Meier survival analysis for breast cancer outcomes in groups with combined MUC17/PCNX1 markers. Expression of MUC17 and PCNX1 was determined in resection samples from a cohort of 53 primary breast cancers patients who were treated with neoadjuvant chemotherapy. The cohort was grouped into low MUC17/high PCNX1 or high MUC17/low PCNX1 based on *in vitro* findings. Relative survival is shown for the groups, and significance was tested using the log rank test. Left graph shows the end point of disease free survival, while right graph shows the end point of disease specific survival. The small coloured vertical lines on the plots represent the end of follow up (censor points) for individual patients.

	Mean DFS (days) (95% CI)	**Log Rank**	**Mean DSS (days) (95% CI)**	**Log Rank**
MUC17 low/ PCNX1 high	2040 (1325 -2755)	0.280	2040 (1325 -2755)	0.484
MUC17 high/ PCNX1 low	1157 (655 -1659)		1592 (1278-1906)	

Table 6.12 Comparison of mean survival between low MUC17/high PCNX1 or high MUC17/low PCNX1. Analyses were performed as described for Figure 6.11. Here, mean disease free survival (DFS) or disease specific survival (DSS) in the corresponding groups were compared, showing 95% confidence intervals and log rank *p* values.

6.4. Discussion

6.4.1. MUC17 and PCNX1 proteins are differentially expressed in breast cancer cohorts

To my knowledge, no large-scale studies have been conducted looking at protein expression of MUC17 and PCNX1 in invasive breast cancer. Herein, I showed there is a broad differential protein expression of MUC17 and PCNX1 across different invasive breast cancers. Analysing control tissue sections along with invasive breast carcinoma TMAs helped me to ensure there was specific binding of the antibodies to the targeted protein, localised to specific cellular compartments (i.e. not due to non-specific bindings to non-targeted compartments, which would likely represent false positives). I used small intestine tissue section as positive control for MUC17, in which the antibodies were expected to bind to the targeted protein at membranous cytoplasm/cytoplasmic of mature villi absorptive cells, consistent with a previous study [207]. For PCNX1, I used ovarian tissue section as a positive control in which antibodies were expected to bind to nucleoplasm cellular compartment, in concordance with a previous study [190]. Altogether, this indicates MUC17 and

PCNX1 are expressed in invasive breast carcinomas and therefore potentially play regulatory roles in the pathogenesis of the disease.

Conducting studies using samples from cohorts of patients in a TMA format presents some potential challenges, for example: inter-scorer reliability and core-to-core variability. I addressed the former by performing an assessment of the inter-scorer concordance using the Cohen's Kappa statistic, which is a recommended statistic method of choice for this kind of analysis [202, 203]. While for the latter, I performed a Spearman's correlation coefficient to assess intra-tumour heterogeneity in terms of expression of targeted proteins across different cores of tissue taken from the same tumours [208, 209]. The former showed there was substantial or near to perfect agreement between scorers for PCNX1 and MUC17, respectively, which indicated the semi-quantification method can be considered reproducible and robust. The latter showed there was a strong correlation among cores from an individual case (p<0.001), which indicates there is minimal intra-tumoural heterogeneity and the pooling method used for selecting representative scores for cases with multiple cores should not greatly influence the result. However, it should be noted minimal intra-tumour heterogeneity detected across different cores of tissue taken from the same tumours in terms of protein expression for MUC17 and PCNX1, is inconsistent with mutational landscape analysis (chapter 4, section 4.3.2) where substantial ITH was found in terms of number of detected mutations. However, this may suggest that MUC17 and PCNX1 are expressed widely across tumours and/or there is limitation due to sampling of the tumour. Thereby, there is a need to have different spatial samples of whole tumour in order to have a better representation of tumour to study ITH especially in terms of clonal evolution.

6.4.2. MUC17 is a prognostic marker in neoadjuvant chemotherapy cohort

The main objective of this chapter was to investigate whether MUC17 and PCNX1 can be used as predictive markers to direct the use of chemotherapy and subsequently improve survival outcomes. Also, to confirm the findings from my *in vitro* studies for

MUC17 and PCNX1 as drivers for chemotherapy response which could be then used as prognostic markers for treatment. Two representative cohorts were included; adjuvant chemotherapy (before treatment) and neoadjuvant chemotherapy (post-treatment). I performed Kaplan-Meier survival analysis for low versus high protein expression for MUC17 and PCNX1 individually, as well as a combination markers analysis in which patients were categorised according to the *in vitro* findings. I uncovered one key significant finding: in the neoadjuvant chemotherapy cohort relatively low MUC17 expression correlated with increased DFS compared to patients with higher expression (log rank test, p=0.017). This was compatible with the *in vitro* findings for MUC17, as siRNA targeted cells showed increased sensitivity to epirubicin treatment compared with control cells (chapter 5, section 5.3.4), and also with my original observations regarding MUC17 mutations, where cells with presumed loss-of-function mutations were apparently successfully treated by chemotherapy in breast cancer patients. This is also in agreement with a recently published paper using computational analysis to identify recurrently mutated genes (RMGs) in cancers, utilising data from The Cancer Genome Atlas (TCGA). The study revealed mucin family genes are among the RMGs, and MUC17 was one of 4 RMGs that found to be shared in many cancer types. In addition, the survival analysis showed patients with mucin-mutated cancers had significantly better overall survival compared to patients with mucin wild-type cancers in skin cutaneous melanoma and stomach adenocarcinoma [210].

The reminder of the survival analyses for both MUC17 and PCNX1 were in agreement with *in vitro* findings in terms of trends, but statistically were not significant based on log rank scores. Since, the neoadjuvant cohort is relatively small (n=47 for MUC17, n=40 for PCNX1), the statistical power could improve should the cohort size be expanded. However, published studies that have shown significant findings with similar sized cohorts in this context. For example, there is a study investigated the protein expression of stromal tumour-infiltrating lymphocytes (TILs) and programmed death ligand 1 (PD-L1) protein expression in a cohort of breast cancer patients treated with neoadjuvant chemotherapy. They compared the TIL count and PD-L1 status in paired pre-treatment and residual cancer tissues and correlated changes and baseline levels with survival in a cohort of only 58 patients assembled into TMAs. It was found

that increased stromal TILs in residual cancer compared to the pre-treatment tissue was associated with longer 5-year recurrence-free survival ($p = 0.02$, HR = 3.9, 95% CI = 1.179–15.39) [4]. Despite, the small size of this cohort, yet it yielded significant findings.

The survival analysis findings for adjuvant chemotherapy cohort were mostly in alignment with *in vitro* findings, but they were not statistically significant. It is important to note only 36 patients out of 140 (25.7%) in the adjuvant cohort had 'events' (death in case of disease specific survival and recurrence in case of disease free survival), and as a result, a high percentage of patients were censored in the Kaplan-Meier survival analyses, which ultimately led to reduced statistical analysis power. In addition, it should be noted MUC17 and PCNX1 were identified in the ER+/HER2- cohort exclusively, and since the cohorts included patients with other hormonal molecular profiles this might have let to underpower the statistical analysis in terms of specificity of targeted cohort patients.

It is interesting to compare my study to a similar study that was also conducted on two independent breast cancer cohorts. In this work, the authors investigated whether Bcl-2 can be used as a predictive and prognostic marker in triple negative breast cancer (TNBC) in both adjuvant and neoadjuvant chemotherapy settings. Patients included were 635 patients who had early primary TNBC and were treated with adjuvant chemotherapy, and 101 patients with primary locally advanced TNBC and were treated with neoadjuvant chemotherapy; both cohorts being substantially larger than mine. It was found that Bcl-2 acted as a prognostic marker, since patients with tumours that were negative for Bcl-2 had improved both breast cancer specific survival (p=0.002) and disease free survival (p=0.003) following the adjuvant chemotherapy. In addition, negative Bcl-2 expression acted as an independent predictor of pathological complete response (pCR) for primary locally advanced TNBC treated with neoadjuvant chemotherapy (p=0.008) [199]. This study showed consistent findings for Bcl-2 as prognostic and predictive marker between different cohorts. However, it should be noted this study was conducted at larger cohort scale comparatively to my

work, and relatively more patients included have had the events (either death or recurrence), which led to have a better statistical representation of the study.

Nevertheless, in my adjuvant cohort among the patients in the TMAs cohorts who have had the events (either recurrence or death) exhibited a high percentage of high MUC17 (61.1%), and low PCNX1 (58.3%) protein expression. This indicates MUC17 and PCNX1 protein expression potentially could be used as survival predictive markers before adjuvant chemotherapy in invasive breast cancer patients. Hence, further investigation by expanding the cohort and further follow-up with patients would be worthwhile.

6.4.3. Conclusion

In this chapter I have shown MUC17 is a driver for chemo-response, and as a prognostic marker for chemotherapy in breast cancer. While other analysis of MUC17 and PCNX1 were in alignment with the *in vitro* findings but statistically non-significant, the statistical power could improve should the cohort size be expanded. Further work required to confirm the findings in the context of different cohorts (i.e. of specific breast cancer molecular sub-type and other cancer types) and *in vivo* studies before the transition of the findings into clinical utilities can be made.

7. Discussion and Summary

7.1. Successful enrichment of somatic mutations attributed to chemotherapy resistance

7.1.1. Laser Capture Micro-dissection (LCM) enriches for resistance-related clones with somatic mutations

A key aim in this study was to identify the somatic mutations that are directly involved with either chemo-sensitivity or chemo-resistance, and to achieve this I took advantage of available new technologies. One of the advantages of this study is the utilisation of laser capture micro-dissection (LCM), which enabled the resistant tumour cells to be dissected directly. This expanded my chances of detecting somatic mutations, especially for those with lower allele frequency in resistant sub-clones, which would be missed in whole tissue samples. This was illustrated in a study where LCM was used to procure pure cell populations of head and neck squamous cell carcinoma in order to study the loss of heterozygosity (LOH) of chromosome 11q genomic region. This LOH analysis required pure cells of interest because the contamination by even few unwanted cells would mean the second allele lost in the cell population of interest would be amplified in the PCR reaction leading to misinterpretation of findings [211].

Furthermore, issues concerning temporal and spatial heterogeneity within the whole tumour have been highlighted previously [212-214]. To overcome these, I performed LCM on multiple regions of the tumour section. Also, since this study focused on studying genomic changes associated with chemotherapy resistance, samples were obtained between pre- and post-NAC, which should reflect the temporal aspects of tumour heterogeneity.

The utilisation of LCM in this study has enabled enrichment of somatic mutations specific to specific cell populations and thereby led to high confidence in calling somatic mutation attributed to chemotherapy resistance.

7.1.2. Integration of WES enables to detect rare somatic mutations

NGS has increasingly been used in the research setting to study tumour heterogeneity and in an effort to aid implementation of precision medicine [215]. Particularly, WES has been receiving more popularity in clinical sittings owing to its decreasing costs, and the fact that a more manageable amount of data are generated that are suitable for rapid clinical interpretation [216]. Also, WES targets the protein-coding regions which composed of approximately 1.5% of whole genome, but contains around 85% of currently known disease-relevant mutations [217]. Furthermore, WES offers a higher depth of coverage (100–150x) than WGS in up to 95% of exons and offers advantages in comparison to targeted panel sequencing of specific genes, which fails to detect complex genomic aberrations or mutations in genes outside of the preselected panel.

In my study, I performed WES in order to obtain a higher depth of coverage to detect mutations that are involved with driving treatment resistance that are at lower allele frequency. This area of research is required since there is a relative lack of profiling rare molecular alterations across different tumour entities, which presents a challenge in integrating personalised medicine [76]. This was realised in a clinical trial, the French SHIVA trial, which aimed to molecularly profile (using NGS) patients with advanced solid tumours and match them with known available molecular treatments [218]. Despite, the fact that the trial showed a successful example of integration of personalised medicine in clinical practise, there were some limitations. One of the limitations was that the trial was limited to 3 treatments, with the majority of patients receiving either hormonal therapy or mTOR inhibitors, which display only limited activity as single agents outside from their specific indications. Hence, there is a need for a more thorough characterisation of molecular drivers before assigning patients to targeted therapy in order to observe a better clinical benefit [218, 219]. The results of

this trial should also prompt a thorough reconsideration of the way clinical trials are conducted and analysed in the era of precision medicine, which require consideration of inter- and intra-tumour heterogeneity.

In my study more focus was directed to investigate those molecular drivers that are relatively rare but likely drive the treatment response in the individuals within my cohort. This aims to expand the range of genomic features that can be used to optimise treatment stratification and thereby improve responses and, hopefully, cure rates.

7.1.3. Adjacent normal tissue and pair analysis

There are many reports discussing the use of histologically normal appearing samples as the sole control tissue in cancer research, in particular relating to concerns associated with effects of field of cancerisation [220-222]. Field cancerisation is the concept that a population of daughter cells with early genetic changes (without histopathology) remain in the organ surrounding the primary tumour. The present technological advancement, including laser capture micro-dissection and high-throughput genomic technologies, and carefully designed studies using appropriate control tissue has enabled identification of important genetic alterations in transformed but histologically normal cells [220]. Hence, utilising combination of both normal adjacent tissue to the tumour which obtained similar conditions as tumour along with reference human genome would serve as better controls for tumour-specific somatic mutations induced by chemotherapeutic agent. In comparison, using both patients' blood samples and reference genome as internal controls to call for somatic mutations can lead to increased risk of calling false positive somatic mutations due to patient's specific single nucleotide polymorphisms (SNPs) [223, 224].

In addition, pairwise analysis of tumour versus matched normal was performed to call for mutations that were only present in tumours, and therefore represent somatic mutations. Furthermore, the identified somatic mutations between tumour versus normal for each individual patient, were further filtered against normal of other patients. This ensured that called somatic mutations were not missed in the sequencing of

normal tissue, which potentially represents a risk for germline mutations. Ultimately, this dual analysis ensured higher confidence for the identified somatic mutations.

7.2. Inter- and intra-tumoural heterogeneity analysis identifies common targets for chemo-response

7.2.1. Tumour heterogeneity findings have potential utility for clinical implementation

Venn diagrams in Figure 4.2 showed substantial inter- and intra-tumour heterogeneity existed in terms of detected mutations, with occasional mutated genes found in common among patients in their respective sub-categories (a maximum of 3 patients shared mutated gene between each sub-category). This was consistent with previous studies where no two patients with the same cancer type had the same collection of somatic mutations, with many pairs of tumours having no mutations in common, and a limited number of mutations appearing in a large fraction of tumours, with most genes being mutated (by SNVs or CNAs) in <5% of all patients with a given cancer type [72, 225, 226].

Nevertheless, it was possible to determine those driver gene mutations likely to modulate the chemotherapy response in this extremely heterogeneous background by implementing MAF and functional pathways analysis approaches. These approaches were also used in other studies enabling the authors similarly to identify candidate driver genes mutations, molecular signature, and de-regulated pathways [227-229]. In addition to these approaches, the design of this study has enabled categorisation of the mutational spectrum into mutations unique to pre-NAC, mutations unique to post-NAC, and mutations shared between pre- and post-NAC. This also allowed changes in MAF to be assessed for mutations shared between pre- and post-NAC and also allowed functional pathways analysis in order to determine enriched pathways involved with chemo-response modulation for each sub-category, which helped to provide candidate genes for chemotherapy response.

Since the primary goal of characterising inter- and intra-tumour heterogeneity is to understand its impact on prognosis and therapy, there have been efforts attempted to interpret tumour heterogeneity for clinical application such as whether heterogeneity-related features are associated to or predictive for some clinical outcomes, such as response to treatment and survival time [230, 231]. For example, data from network-based stratification (NBS), a method which integrates somatic tumour genomes with gene networks to allow for stratification of cancer into informative subtypes by clustering together patients with mutations in similar networks. This method enabled identification of subtypes that are predictive of clinical outcomes such as patient survival or response to treatment. Patients with the most aggressive ovarian tumour NBS subtype had a mean survival of approximately 32 months, compared to more than 80 months for those with the least aggressive NBS subtype [230]. Also, a study investigated the number of clones present at a ≥10% frequency in more than 1000 exome sequences from tumours across 12 cancer types, and assessed the association between the number of clones in a sample with overall survival outcome using computational algorithms. It was found that across cancer types, the presence of more than two clones was associated with worse overall survival as compared to tumours in which either one or two clones were detected [231].

While these methods have all demonstrated potential clinical utilities of tumour heterogeneity interpretation for clinical implementation, they lack functional validation in the context of *in vivo* or *in vitro* experiments. In contrast, in this study I validated my findings from heterogeneity landscape using *in vitro* approaches and therefore showed MUC17 to be a potential prognostic marker for chemotherapy response in breast cancer.

7.2.2. Candidate genes have profound effects on chemotherapeutic response *in vitro,* and in patients

Many review articles have indicated that there is a relative lack of studies that have combined computational analysis approaches and functional genomic biology approaches on NGS data to identify molecular targets for diseases, especially variants of unknown clinical significance or low prevalence [232, 233]. Having identified a list of candidate genes from computational analysis of the heterogeneity landscape of breast cancer patients, I took a further step to validate them using *in vitro* approaches.

Pragmatic screening using siRNA system has enabled me to shorten the list of candidate genes to focus on only those driver genes that have profound effects on chemo-response. Indeed, MUC17 and PCNX1 stood out from the screen by showing chemo-sensitivity and chemo-resistance phenotypes, respectively. The fact that MUC17 and PCNX1 have profound effects on chemo-response suggests that they play key roles in cell survival mechanisms. Such observation of profound effects of single genes was also seen in a similar study in which they integrated data from gene expression profile of primary colorectal cancer (CRC) cohort, and identified many genes that were highly expressed in the tumours. Then, the authors followed up these potential targets in 25 CRC cell lines to identify 11 genes that were consistently over-expressed in primary CRC and in CRC cell lines. After that, the 11 genes were examined for LOF effects further using siRNA and it was found that 5 candidate genes showed a 20% or greater decrease in cellular viability as compared to a control siRNA. Also, whole transcriptome expression analyses were conducted following siRNA transfected cells to identify an “RNAi signature” for each gene of interest. These RNAi signatures were defined as the genes with altered expression following transfection with targeted siRNA compared to a negative control. It was found that the RNAi signatures for some genes such as *HMGA1*, *RRM2* and *RPS2* showed that the silencing of candidate genes influence the expression of many downstream genes, which is consistent with the observation that these genes were associated with the most pronounced reduction of cell viability [234].

There are few studies in the literature that I have found that have integrated NGS and functional biology validation approaches to identify novel cancer genes for cancer progression and treatment response. One example is a study in which the authors performed WES on 13 endometrial cancers and matched normal samples and, similarly to my approaches, focused on somatic mutations utilising bioinformatics prioritisation tools and high-throughput RNAi screen system to identify 12 potential driver cancer genes. Also, the functional genomics studies led to identification of mutations in the *ARID1A* gene that co-occur frequently with mutations in PI3K pathway and were associated with PI3K pathway activation. In addition, utilising the siRNA knockdown system in endometrial cancer cell lines supported *ARID1A* as a novel regulator of PI3K pathway activity. At the time this study was published, it was first to report a novel somatic mutation in endometrial cancer and provides functional evidence of its importance [235].

In my study, I proceeded to confirm the functional genomic findings and see whether the findings from *in vitro* can be utilised for clinical application by examining protein expression of the candidate genes in clinical tissue samples. The findings from TMAs endorsed the *in vitro* findings at least for MUC17, namely that MUC17 was a predictive marker for chemotherapy prognosis. Therefore, this finding provides unbiased evidence for the robustness of this approach of charactering tumour heterogeneity for identification of novel markers for breast cancer progression and treatment response.

7.3. MUC17 and PCNX1 - a future translational pathway?

This study demonstrates successful implementation of personalised medicine in identifying molecular targets for chemo-response in cohort of patients who showed resistance to neoadjuvant chemotherapy. Analysis of tumour heterogeneity landscape of pre- and post-NAC samples followed by functional validations and relatively large-scale clinical tissue cohort studies all have led to identification of MUC17 and PCNX1 as genuine clinical targets.

Nonetheless, in order to use MUC17 and PCNX1 as targets for molecular targeted therapies requires investigating the underlying molecular mechanisms, which can be done using model systems including non-human model organisms such as mice, zebrafish, fruit-flies. Also, the candidate molecular targets need to undergo stringent biological and clinico-pathological validation i.e. multiple biological assays to validate their biological activities *in vivo* such as assessment of their signalling activity pathways. In addition, before successful translation of developed therapeutic agents in to the clinics they need to be clinically validated before they can be adapted for routine clinical practice by implementing clinical trials [236].

As far as the development of MUC17 and PCNX1 as biomarkers that can be used clinically to predict disease progression or response to therapy, they require process that involves assessment of prevalence, sensitivity, specificity and rigorous validation in multiple clinical cohorts [237].

However, there are many challenges associated with translation of genomic findings into therapies due to genetically complex nature of tumours which are characterised by many genomic alterations. Hence, targeted therapies based on the status of a single molecular alteration in patients' tumours is often not sufficient to predict therapeutic response. Also, one of the practical limitations of successful translation of

genomic findings into clinical targeted therapies is due to the lack of availability of well-defined, clinically characterised cohorts for evaluating the biomarker and lack of standardisation regarding how specimens are collected, handled, and stored. Ultimately, these issues can influence whether or not biomarkers validate in well-controlled cohorts. In addition, obtaining well-annotated clinical information regarding the cohort has been a substantial barrier in implementing representative studies for transitional therapies [238].

Despite these difficulties there have been successful examples of translating cancer genomic to targeted cancer therapy for example, identification of fusion proteins of BCR-ABL genes which mostly found in Philadelphia chromosome abnormality in chronic myeloid leukaemia (CML) that causes continuous over activation of tyrosine kinase pathway. This discovery led to development of the targeted therapy Imatinib (also commercially known as Gleevec or Glivec), which acts as a specific inhibitor of tyrosine kinase enzymes and subsequently resulted in a dramatic increase in patient response to treatment [237, 239].

Another successful example of translational genomic discoveries into targeted therapies was seen in EML4-ALK translocation genetic alteration found in around 7% of Non-Small Cell Lung Carcinoma (NSCLC). And since the ALK inhibitor is already available, the speed to translation ALK targeted therapy was relatively fast due to prior knowledge about the drug mechanism and the candidate patients harbouring the genetic alteration to include for targeted therapy [237, 240].

For MUC17 and PCNX1, there is little current literature concerning their molecular signalling pathways and oncogenic activities, and while I have shown them to be novel and promising clinical targets for chemotherapy response, a huge amount of work remains to translate these findings into better cancer outcomes. Further experiments in order to validate my findings will include the following:

- Repeat the study with expansion of the cohort to include larger number of patients for WES and also confirm the sequencing findings with targeted sequencing approaches in order to obtain higher depth of coverage to ensure the biological relevance of the candidate genes.
- Include another patient's cohort who responded to chemotherapy and the findings from both cohorts can be compared to find the exclusive mutations in both cohorts.
- For *in vitro* studies another functional technique could be included for example gain of function technique such as cDNA libraries to compare the findings from siRNA (loss of function), so the correlation assessment of phenotype of chemotherapeutic response from both techniques can me made.
- Perform CRSPR technique for permanent knockout of the genes and to overcome the limitations of above techniques.
- Validate the findings using *in vivo* approaches. For instance; following knockout of candidate genes in the cell lines, the cells can be injected into mice and treated with chemotherapy to assess their response.
- Higher number of patients should be included in TMAs cohorts in order to assess whether MUC17 and PCNX1 can be used as predictive and prognostic markers for chemotherapy response in breast cancer. Also, would be useful to have matching TMAs cohorts of the sequenced patients for validation in the TMAs cohorts.

The findings from validation studies would enable MUC17 and PCNX1 to be utilised for clinical applications such as targeted therapy (i.e. MUC17 inhibitor) to induce synergistic effect of chemotherapy especially for patients who did not show response to chemotherapy regimen at NAC sittings. In addition, MUC17 and PCNX1 would be used as predictive and prognostic markers to direct the chemotherapy for patients who would benefit from chemotherapy and also monitor the survival progression during and after the chemotherapy.

7.4. Summary

In this study I was primarily interested to investigate the role of intra-tumoural heterogeneity in resistance to Neoadjuvant Chemotherapy (NAC) in breast cancer. I sought to study the mutational landscape between pre- and post-NAC samples of within an individual and across a small group of 6 patients. The mutational analysis showed substantial inter- and intra-tumoural heterogeneity in terms of detected mutations in breast cancer patients, as is consistent with previous studies. [5-9](Campbell and Polyak 2007, Nguyen, Vanner et al. 2012, Aparicio and Caldas 2013, Rybinski and Yun 2016, McGranahan and Swanton 2017)[5-9][5-9][5-9][5-9]This presented a challenge as to identify driver gene mutations for chemo-response that were common for patients who did not respond to epirubicin/cyclophosamide chemotherapy regimen. Despite, this broad inter- and intra-tumoural heterogeneity spectrum in breast cancer patients, it was possible to find common targets and driver gene mutations for the chemo-response. In order to prove that, I endeavoured to utilise the uniqueness design of this study. The availability pre- and post-NAC samples allowed categorising mutational landscape into those selected for and selected against the chemotherapy. Also, using the adjacent normal tissue from the patients allowed to call somatic mutations that are involved in cancer and chemotherapy progression. In addition, utilising the computational functional tools, such as SIFT and polyphen2 protein damaging predictor tools, and functional pathways analysis allowed finding potential common targets in the heterogeneous breast cancer patients and common pathways shared in the mutated genes set.

The study was extended to perform functional validation of the candidate genes using *in vitro* approaches and TMAs cohort. This allowed confirming the findings from genomic sequencing data analysis and putting the findings in context of targeted therapies. This was evident based on discovering MUC17 and PCNX1 as epirubicin/cyclophosamide chemotherapy regimen response mediators. This work in essence demonstrates the successful implementation of personalised medicine to tackle issues associated with treatment ineffectiveness and treatment side effects. However, future challenges remain with translating these findings into clinical benefits.

8. List of References

1. Cancer Research UK. *Breast cancer incidence statistics*. 2015 [cited 2015 25 Feb].
2. BREASTCANCER.ORG. *U.S Breast Cancer Statistics*. 2015 [cited 2015 25. Feb].
3. Cheung, P., *Recent advances in breast cancer treatment*. 2018.
4. Pelekanou, V., et al., *Effect of neoadjuvant chemotherapy on tumor-infiltrating lymphocytes and PD-L1 expression in breast cancer and its clinical significance.* Breast Cancer Research, 2017. **19**(1): p. 91.
5. McGranahan, N. and C. Swanton, *Clonal heterogeneity and tumor evolution: past, present, and the future.* Cell, 2017. **168**(4): p. 613-628.
6. Rybinski, B. and K. Yun, *Addressing intra-tumoral heterogeneity and therapy resistance.* Oncotarget, 2016. **7**(44): p. 72322.
7. Aparicio, S. and C. Caldas, *The implications of clonal genome evolution for cancer medicine.* New England Journal of Medicine, 2013. **368**(9): p. 842-851.
8. Campbell, L.L. and K. Polyak, *Breast tumor heterogeneity: cancer stem cells or clonal evolution?* Cell Cycle, 2007. **6**(19): p. 2332-2338.
9. Nguyen, L.V., et al., *Cancer stem cells: an evolving concept.* Nature Reviews Cancer, 2012. **12**(2): p. 133-143.
10. Malhotra, G.K., et al., *Histological, molecular and functional subtypes of breast cancers.* Cancer biology & therapy, 2010. **10**(10): p. 955-60.
11. Li, C.I., D.J. Uribe, and J.R. Daling, *Clinical characteristics of different histologic types of breast cancer.* Br J Cancer, 2005. **93**(9): p. 1046-52.
12. Lester, S.C., et al., *Protocol for the examination of specimens from patients with invasive carcinoma of the breast.* Arch Pathol Lab Med, 2009. **133**(10): p. 1515-38.
13. Cao, S.-S. and C.-T. Lu, *Recent perspectives of breast cancer prognosis and predictive factors.* Oncology letters, 2016. **12**(5): p. 3674-3678.
14. O'Sullivan, B., et al., *The TNM classification of malignant tumours-towards common understanding and reasonable expectations.* Lancet Oncol, 2017. **18**(7): p. 849-851.
15. Carter, C.L., C. Allen, and D.E. Henson, *Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases.* Cancer, 1989. **63**(1): p. 181-7.
16. Koscielny, S., et al., *Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination.* British journal of cancer, 1984. **49**(6): p. 709-15.
17. Perou, C.M., et al., *Molecular portraits of human breast tumours.* nature, 2000. **406**(6797): p. 747.
18. Dai, X., et al., *Breast cancer intrinsic subtype classification, clinical use and future trends.* American journal of cancer research, 2015. **5**(10): p. 2929-43.
19. Prat, A., et al., *Clinical implications of the intrinsic molecular subtypes of breast cancer.* The Breast, 2015. **24**: p. S26-S35.
20. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
21. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.

22. National Comprehensive Cancer, N., *Breast Cancer Screening and Diagnosis Clinical Practice Guidelines in Oncology.* Journal of the National Comprehensive Cancer Network : JNCCN, 2003. **1**(2): p. 242-63.
23. Anampa, J., D. Makower, and J.A. Sparano, *Progress in adjuvant chemotherapy for breast cancer: an overview.* BMC medicine, 2015. **13**(1): p. 195.
24. Yeo, B., N.C. Turner, and A. Jones, *An update on the medical management of breast cancer.* BMJ: British Medical Journal, 2014. **348**: p. g3608.
25. Yarnold, J., *Early and locally advanced breast cancer: diagnosis and treatment National Institute for Health and Clinical Excellence guideline 2009.* Clinical Oncology, 2009. **21**(3): p. 159-160.
26. Joseph A Sparano, M. *Breast Cancer Treatment Protocols*. 2015 [cited 2015 29 Oct].
27. Wishart DS, K.C., Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* 2006 [cited 2015 29 Oct].
28. Vaidya, J.S., et al., *Rethinking neoadjuvant chemotherapy for breast cancer.* Bmj, 2018. **360**: p. j5913.
29. Badwe, R., et al., *Locoregional treatment versus no treatment of the primary tumour in metastatic breast cancer: an open-label randomised controlled trial.* The lancet oncology, 2015. **16**(13): p. 1380-1388.
30. Thompson, A. and S. Moulder-Thompson, *Neoadjuvant treatment of breast cancer.* Annals of Oncology, 2012. **23**(suppl 10): p. x231-x236.
31. Lichter, A.S., et al., *Mastectomy versus breast-conserving therapy in the treatment of stage I and II carcinoma of the breast: a randomized trial at the National Cancer Institute.* J Clin oncol, 1992. **10**(6): p. 976-983.
32. Ikeda, T., et al., *The role of neoadjuvant chemotherapy for breast cancer treatment.* Breast Cancer, 2002. **9**(1): p. 8.
33. Pinder, S.E., et al., *Macroscopic handling and reporting of breast cancer specimens pre-and post-neoadjuvant chemotherapy treatment: review of pathological issues and suggested approaches.* Histopathology, 2015.
34. Yee, D., et al., *Abstract GS3-08: Pathological complete response predicts event-free and distant disease-free survival in the I-SPY2 TRIAL*. 2018, AACR.
35. Biswas, T., et al., *The survival benefit of neoadjuvant chemotherapy and pCR among patients with advanced stage triple negative breast cancer.* Oncotarget, 2017. **8**(68): p. 112712.
36. Rouzier, R., et al., *Breast cancer molecular subtypes respond differently to preoperative chemotherapy.* Clinical cancer research, 2005. **11**(16): p. 5678-5685.
37. Kim, S.I., et al., *Molecular subtypes and tumor response to neoadjuvant chemotherapy in patients with locally advanced breast cancer.* Oncology, 2010. **79**(5-6): p. 324-330.
38. Folgueira, K., et al., *Gene expression profile of residual breast cancer after doxorubicin and cyclophosphamide neoadjuvant chemotherapy.* Oncology reports, 2009. **22**(4): p. 805-813.
39. Arend, R.C., et al., *Molecular Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma.* Molecular Cancer Research, 2018.
40. Kim, B., et al., *Neoadjuvant chemotherapy induces expression levels of breast cancer resistance protein that predict disease-free survival in breast cancer.* PLoS One, 2013. **8**(5): p. e62766.

41. Gonzalez-Angulo, A.M., F. Morales-Vasquez, and G.N. Hortobagyi, *Overview of resistance to systemic therapy in patients with breast cancer*, in *Breast Cancer Chemosensitivity*. 2007, Springer. p. 1-22.
42. Housman, G., et al., *Drug resistance in cancer: an overview.* Cancers, 2014. **6**(3): p. 1769-1792.
43. Sampath, D., et al., *Pharmacodynamics of cytarabine alone and in combination with 7-hydroxystaurosporine (UCN-01) in AML blasts in vitro and during a clinical trial.* Blood, 2006. **107**(6): p. 2517-2524.
44. Zhang, N., et al., *5-Fluorouracil: mechanisms of resistance and reversal strategies.* Molecules, 2008. **13**(8): p. 1551-1569.
45. Chang, G. and C.B. Roth, *Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters.* Science, 2001. **293**(5536): p. 1793-1800.
46. Haber, M., et al., *Association of high-level MRP1 expression with poor clinical outcome in a large prospective study of primary neuroblastoma.* Journal of clinical oncology, 2006. **24**(10): p. 1546-1553.
47. Olaussen, K.A., et al., *DNA repair by ERCC1 in non–small-cell lung cancer and cisplatin-based adjuvant chemotherapy.* New England Journal of Medicine, 2006. **355**(10): p. 983-991.
48. Mansoori, B., et al., *The different mechanisms of cancer drug resistance: A brief review.* Advanced pharmaceutical bulletin, 2017. **7**(3): p. 339.
49. De Vree, J.M.L., et al., *Mutations in the MDR3 gene cause progressive familial intrahepatic cholestasis.* Proceedings of the National Academy of Sciences, 1998. **95**(1): p. 282-287.
50. Lesniak, D., et al., *β1-integrin circumvents the Antiproliferative effects of Trastuzumab in human epidermal growth factor receptor-2–positive breast cancer.* Cancer research, 2009: p. 0008-5472. CAN-09-1591.
51. Navin, N., et al., *Inferring tumor progression from genomic heterogeneity.* Genome research, 2010. **20**(1): p. 68-80.
52. Russnes, H.G., et al., *Insight into the heterogeneity of breast cancer through next-generation sequencing.* The Journal of clinical investigation, 2011. **121**(10): p. 3810.
53. Eliyatkın, N., et al., *Molecular classification of breast carcinoma: from traditional, old-fashioned way to a new age, and a new way.* The journal of breast health, 2015. **11**(2): p. 59.
54. Shackleton, M., et al., *Heterogeneity in cancer: cancer stem cells versus clonal evolution.* Cell, 2009. **138**(5): p. 822-829.
55. Al-Hajj, M., et al., *Prospective identification of tumorigenic breast cancer cells.* Proceedings of the National Academy of Sciences, 2003. **100**(7): p. 3983-3988.
56. Singh, S.K., et al., *Identification of a cancer stem cell in human brain tumors.* Cancer research, 2003. **63**(18): p. 5821-5828.
57. Fang, D., et al., *A tumorigenic subpopulation with stem cell properties in melanomas.* Cancer research, 2005. **65**(20): p. 9328-9337.
58. Marker, P.C. *Highly purified CD44+ prostate cancer cells from xenograft human tumors are enriched in tumorigenic and metastatic progenitor cells: Patrawala L, Calhoun T, Schneider-Broussard R, Li H, Bhatia B, Tang S, Reilly JG, Chandra D, Zhou J, Claypool K, Coghlan L, Tang DG, Department of Carcinogenesis, The University of Texas MD Anderson Cancer Center, Science Park-Research Division,*

Smithville, TX. in *Urologic Oncology: Seminars and Original Investigations*. 2007. Elsevier.
59. Ricci-Vitiani, L., et al., *Identification and expansion of human colon-cancer-initiating cells.* Nature, 2007. **445**(7123): p. 111-115.
60. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306-313.
61. Sidransky, D., et al., *Clonal expansion of p53 mutant cells is associated with brain tumour progression.* 1992.
62. Siegmund, K.D., et al., *Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers.* Proceedings of the National Academy of Sciences, 2009. **106**(12): p. 4828-4833.
63. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-28.
64. O'Leary, B., et al., *The genetic landscape and clonal evolution of breast cancer resistance to palbociclib plus fulvestrant in the PALOMA-3 trial.* Cancer discovery, 2018. **8**(11): p. 1390-1403.
65. Torres, L., et al., *Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases.* Breast cancer research and treatment, 2007. **102**(2): p. 143-55.
66. Shipitsin, M., et al., *Molecular definition of breast tumor heterogeneity.* Cancer cell, 2007. **11**(3): p. 259-273.
67. Hiley, C., et al., *Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine.* Genome biology, 2014. **15**(8): p. 453.
68. Su, K.Y., et al., *Pretreatment epidermal growth factor receptor (EGFR) T790M mutation predicts shorter EGFR tyrosine kinase inhibitor response duration in patients with non-small-cell lung cancer.* J Clin Oncol, 2012. **30**(4): p. 433-40.
69. Johnson, B.E., et al., *Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma.* Science, 2014. **343**(6167): p. 189-193.
70. Fribbens, C., et al., *Tracking evolution of aromatase inhibitor resistance with circulating tumour DNA analysis in metastatic breast cancer.* Annals of Oncology, 2017. **29**(1): p. 145-153.
71. Xu, X., et al., *HER2 Reactivation through Acquisition of the HER2 L755S Mutation as a Mechanism of Acquired Resistance to HER2-targeted Therapy in HER2+ Breast Cancer.* Clinical Cancer Research, 2017. **23**(17): p. 5123-5134.
72. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719.
73. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annual Review of Pathology: Mechanisms of Disease, 2015. **10**: p. 25-50.
74. Ding, L., et al., *Analysis of next-generation genomic data in cancer: accomplishments and challenges.* Hum Mol Genet, 2010. **19**(R2): p. R188-96.
75. Desmedt, C., et al., *Next generation sequencing in breast cancer: first take home messages.* Current opinion in oncology, 2012. **24**(6): p. 597.
76. Horak, P., S. Frohling, and H. Glimm, *Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls.* ESMO open, 2016. **1**(5): p. e000094.
77. Simon, R. and S. Roychowdhury, *Implementing personalized cancer genomics in clinical trials.* Nature reviews Drug discovery, 2013. **12**(5): p. 358-369.

78. Tripathy, D., et al., *Next generation sequencing and tumor mutation profiling: are we ready for routine use in the oncology clinic?* BMC medicine, 2014. **12**(1): p. 140.
79. Reuter, J.A., D.V. Spacek, and M.P. Snyder, *High-throughput sequencing technologies.* Molecular cell, 2015. **58**(4): p. 586-597.
80. Shah, S.P., et al., *The clonal and mutational evolution spectrum of primary triple-negative breast cancers.* Nature, 2012. **486**(7403): p. 395-399.
81. Ellis, M.J., et al., *Whole-genome analysis informs breast cancer response to aromatase inhibition.* Nature, 2012. **486**(7403): p. 353-360.
82. Stephens, P.J., et al., *The landscape of cancer genes and mutational processes in breast cancer.* Nature, 2012. **486**(7403): p. 400-404.
83. Banerji, S., et al., *Sequence analysis of mutations and translocations across breast cancer subtypes.* Nature, 2012. **486**(7403): p. 405-409.
84. Curtis, C., *Genomic profiling of breast cancers.* Current opinion in obstetrics & gynecology, 2015. **27**(1): p. 34.
85. Network, C.G.A., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61.
86. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, 2012. **486**(7403): p. 346.
87. Rodenburg, R.J., *The functional genomics laboratory: functional validation of genetic variants.* Journal of inherited metabolic disease, 2018. **41**(3): p. 297-307.
88. Technologies, A., *SUreCall handnotes.* 2016. **3.0.3**.
89. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet. journal, 2011. **17**(1): p. pp. 10-12.
90. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics (Oxford, England), 2010. **26**(5): p. 589-95.
91. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nature genetics, 2011. **43**(5): p. 491-8.
92. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly, 2012. **6**(2): p. 80-92.
93. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method.* methods, 2001. **25**(4): p. 402-408.
94. Institute, N.H.G.R. 2010 [cited 2015 25 Feb].
95. Kerick, M., et al., *Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity.* BMC medical genomics, 2011. **4**(1): p. 68.
96. Schweiger, M.R., et al., *Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number-and mutation-analysis.* PLoS One, 2009. **4**(5): p. e5548.
97. Yatani, R., et al., *Latent prostatic carcinoma: pathological and epidemiological aspects.* Japanese journal of clinical oncology, 1989. **19**(4): p. 319-326.
98. Wood, H.M., et al., *Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens.* Nucleic acids research, 2010. **38**(14): p. e151-e151.
99. Aird, D., et al., *Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.* Genome biol, 2011. **12**(2): p. R18.

100. Beltran, H., et al., *Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity.* European urology, 2013. **63**(5): p. 920-926.
101. Qiagen. *QIAamp DNA FFPE Tissue Handbook*. 2012 [cited 2015 25 Oct].
102. Yost, S.E., et al., *Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens.* Nucleic acids research, 2012. **40**(14): p. e107-e107.
103. Wu, J., et al., *Reduced let-7a is associated with chemoresistance in primary breast cancer.* PloS one, 2015. **10**(7): p. e0133643.
104. Simone, N.L., et al., *Laser-capture microdissection: opening the microscopic frontier to molecular analysis.* Trends in Genetics, 1998. **14**(7): p. 272-276.
105. Shen, C.-Y., et al., *Genome-wide search for loss of heterozygosity using laser capture microdissected tissue of breast carcinoma: an implication for mutator phenotype and breast cancer pathogenesis.* Cancer research, 2000. **60**(14): p. 3884-3892.
106. Gerlinger, M., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing (vol 366, pg 883, 2012).* NEW ENGLAND JOURNAL OF MEDICINE, 2012. **367**(10): p. 976-976.
107. Rosenberg, A.Z., et al., *High-throughput microdissection for next-generation sequencing.* PloS one, 2016. **11**(3): p. e0151775.
108. Saliba, A.-E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic acids research, 2014. **42**(14): p. 8845-8860.
109. Technologies, A., *SureSelectXT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library*. 2017.
110. Astolfi, A., et al., *Whole exome sequencing (WES) on formalin-fixed, paraffin-embedded (FFPE) tumor tissue in gastrointestinal stromal tumors (GIST).* BMC genomics, 2015. **16**(1): p. 892.
111. Oh, E., et al., *Comparison of accuracy of whole-exome sequencing with formalin-fixed paraffin-embedded and fresh frozen tissue samples.* PLoS One, 2015. **10**(12): p. e0144162.
112. Hedegaard, J., et al., *Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue.* PLoS One, 2014. **9**(5): p. e98187.
113. Munchel, S., et al., *Targeted or whole genome sequencing of formalin fixed tissue samples: potential applications in cancer genomics.* Oncotarget, 2015. **6**(28): p. 25943.
114. Liu, Z.K., et al., *A three-caller pipeline for variant analysis of cancer whole-exome sequencing data.* Molecular Medicine Reports, 2017. **15**(5): p. 2489-2494.
115. Kumar, A., E. Turner, and J. Shendure. *TARGETED CAPTURE AND MASSIVELY PARALLEL SEQUENCING OF THE HUMAN EXOME*. in *JOURNAL OF INVESTIGATIVE MEDICINE*. 2010. BMJ PUBLISHING GROUP BRITISH MED ASSOC HOUSE, TAVISTOCK SQUARE, LONDON WC1H 9JR, ENGLAND.
116. Scholz, M.B., C.-C. Lo, and P.S. Chain, *Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis.* Current opinion in biotechnology, 2012. **23**(1): p. 9-15.
117. Witte, A.D. *Next-Generation Sequencing for Non-Bioinformaticians*. 2013; Available from: https://www.dddmag.com/article/2013/12/next-generation-sequencing-non-bioinformaticians.
118. Smith, D.R., *Buying in to bioinformatics: an introduction to commercial sequence analysis software.* Briefings in bioinformatics, 2014. **16**(4): p. 700-709.

119. Bodi, K., *Tools for next generation sequencing data analysis.* Journal of biomolecular techniques: JBT, 2011. **22**(Suppl): p. S18.
120. Bao, R., et al., *Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing.* Cancer informatics, 2014. **13**: p. CIN. S13779.
121. Meena, N., et al., *A Bioinformatics Pipeline for Whole Exome Sequencing: Overview of the Processing and Steps from Raw Data to Downstream Analysis.* bioRxiv, 2017: p. 201145.
122. Chen, J., et al., *Improved human disease candidate gene prioritization using mouse phenotype.* BMC bioinformatics, 2007. **8**(1): p. 392.
123. Cooke, S., et al., *Intra-tumour genetic heterogeneity and poor chemoradiotherapy response in cervical cancer.* British journal of cancer, 2011. **104**(2): p. 361.
124. Cooke, S.L., et al., *Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma.* Oncogene, 2010. **29**(35): p. 4905.
125. Teer, J.K., et al., *Evaluating somatic tumor mutation detection without matched normal samples.* Human genomics, 2017. **11**(1): p. 22.
126. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic acids research, 2008. **37**(1): p. 1-13.
127. Zhang, B., S. Kirov, and J. Snoddy, *WebGestalt: an integrated system for exploring gene sets in various biological contexts.* Nucleic acids research, 2005. **33**(suppl_2): p. W741-W748.
128. Fishbein, L., et al., *Whole-exome sequencing identifies somatic ATRX mutations in pheochromocytomas and paragangliomas.* Nature communications, 2015. **6**: p. 6140.
129. Burrell, R.A. and C. Swanton, *Tumour heterogeneity and the evolution of polyclonal drug resistance.* Molecular oncology, 2014. **8**(6): p. 1095-1111.
130. Gay, L., A.-M. Baker, and T.A. Graham, *Tumour cell heterogeneity.* F1000Research, 2016. **5**.
131. Dagogo-Jack, I. and A.T. Shaw, *Tumour heterogeneity and resistance to cancer therapies.* Nature reviews Clinical oncology, 2018. **15**(2): p. 81.
132. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* New England journal of medicine, 2012. **366**(10): p. 883-892.
133. Daber, R., S. Sukhadia, and J.J. Morrissette, *Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets.* Cancer genetics, 2013. **206**(12): p. 441-448.
134. Noorani, A., et al., *A comparative analysis of whole genome sequencing of esophageal adenocarcinoma pre-and post-chemotherapy.* Genome research, 2017. **27**(6): p. 902-912.
135. Liu, D., et al., *Mutational patterns in chemotherapy resistant muscle-invasive bladder cancer.* Nature communications, 2017. **8**(1): p. 2193.
136. Tan, S.-H., et al., *High-throughput mutation profiling changes before and 3 weeks after chemotherapy in newly diagnosed breast cancer patients.* PloS one, 2015. **10**(12): p. e0142466.
137. Murugaesu, N., et al., *Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy.* Cancer discovery, 2015.
138. Findlay, J.M., et al., *Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy.* Nature communications, 2016. **7**: p. 11111.

139. Aoudjit, F. and K. Vuori, *Integrin signaling in cancer cell survival and chemoresistance.* Chemotherapy research and practice, 2012. **2012**.
140. Naci, D., et al., *integrin promotes chemoresistance against doxorubicin in cancer cells through extracellular signalregulated kinase (ERK).* J. Biol. Chem. **287**: p. 17065-17076.
141. Senthebane, D.A., et al., *The role of tumor microenvironment in chemoresistance: To survive, keep your enemies closer.* International journal of molecular sciences, 2017. **18**(7): p. 1586.
142. Morin, P.J., *Drug resistance and the microenvironment: nature and nurture.* Drug Resistance Updates, 2003. **6**(4): p. 169-172.
143. Sato, N., et al., *Role of hyaluronan in pancreatic cancer biology and therapy: Once again in the spotlight.* Cancer Sci, 2016. **107**(5): p. 569-75.
144. Arend, R.C., et al., *Molecular Response to Neoadjuvant Chemotherapy in High-Grade Serous Ovarian Carcinoma.* Mol Cancer Res, 2018. **16**(5): p. 813-824.
145. Hientz, K., et al., *The role of p53 in cancer drug resistance and targeted chemotherapy.* Oncotarget, 2017. **8**(5): p. 8921-8946.
146. Breen, L., et al., *Investigation of the role of p53 in chemotherapy resistance of lung cancer cell lines.* Anticancer Res, 2007. **27**(3A): p. 1361-4.
147. Sturm, I., et al., *Mutation of p53 and consecutive selective drug resistance in B-CLL occurs as a consequence of prior DNA-damaging chemotherapy.* Cell Death Differ, 2003. **10**(4): p. 477-84.
148. Soussi, T., *p53 mutations and resistance to chemotherapy: A stab in the back for p73.* Cancer Cell, 2003. **3**(4): p. 303-5.
149. Kim, B., et al., *Chemotherapy induces Notch1-dependent MRP1 up-regulation, inhibition of which sensitizes breast cancer cells to chemotherapy.* Bmc Cancer, 2015. **15**.
150. Purow, B., *Notch Inhibition as a Promising New Approach to Cancer Therapy.* Notch Signaling in Embryology and Cancer, 2012. **727**: p. 305-319.
151. Liu, J.T., et al., *Blocking the NOTCH pathway can inhibit the growth of CD133-positive A549 cells and sensitize to chemotherapy.* Biochemical and Biophysical Research Communications, 2014. **444**(4): p. 670-675.
152. Marce, S., et al., *Frequency of ABL gene mutations in chronic myeloid leukemia patients resistant to imatinib and results of treatment switch to second-generation tyrosine kinase inhibitors.* Medicina Clinica, 2013. **141**(3): p. 95-99.
153. Vaidya, S., et al., *Evolution of BCR/ABL Gene Mutation in CML Is Time Dependent and Dependent on the Pressure Exerted by Tyrosine Kinase Inhibitor.* Plos One, 2015. **10**(1).
154. Wagner, M., et al., *NCOA3 is a selective co-activator of estrogen receptor α-mediated transactivation of PLAC1 in MCF-7 breast cancer cells.* BMC cancer, 2013. **13**(1): p. 570.
155. Gupta, A., et al., *NCOA3 coactivator is a transcriptional target of XBP1 and regulates PERK–eIF2α–ATF4 signalling in breast cancer.* Oncogene, 2016. **35**(45): p. 5860.
156. Mathur, S. and J. Sutton, *Personalized medicine could transform healthcare.* Biomedical reports, 2017. **7**(1): p. 3-5.
157. Korthauer, K.D. and C. Kendziorski, *MADGiC: a model-based approach for identifying driver genes in cancer.* Bioinformatics, 2015. **31**(10): p. 1526-1535.
158. Gonzalez-Perez, A. and N. Lopez-Bigas, *Functional impact bias reveals cancer drivers.* Nucleic acids research, 2012. **40**(21): p. e169-e169.

159. Gao, S., et al., *Applications of RNA interference high-throughput screening technology in cancer biology and virology.* Protein & cell, 2014. **5**(11): p. 805-815.
160. Williams, S.P., et al., *High-throughput RNAi screen for essential genes and drug synergistic combinations in colorectal cancer.* Scientific data, 2017. **4**: p. 170139.
161. Bhinder, B., D. Shum, and H. Djaballah, *Comparative analysis of RNAi screening technologies at genome-scale reveals an inherent processing inefficiency of the plasmid-based shRNA hairpin.* Combinatorial chemistry & high throughput screening, 2014. **17**(2): p. 98-113.
162. Engel, J., *EGF-like domains in extracellular matrix proteins: Localized signals for growth and differentiation?* FEBS letters, 1989. **251**(1-2): p. 1-7.
163. Kim, C.W., et al., *N-terminal domains of native multidomain proteins have the potential to assist de novo folding of their downstream domains in vivo by acting as solubility enhancers.* Protein science, 2007. **16**(4): p. 635-643.
164. Nakagawa, M., et al., *Reduced intracellular drug accumulation in the absence of P-glycoprotein (mdr1) overexpression in mitoxantrone-resistant human MCF-7 breast cancer cells.* Cancer research, 1992. **52**(22): p. 6175-6181.
165. Marin, J.J., et al., *The role of reduced intracellular concentrations of active drugs in the lack of response to anticancer chemotherapy.* Acta Pharmacologica Sinica, 2014. **35**(1): p. 1.
166. Bauer, J.A., et al., *RNA interference (RNAi) screening approach identifies agents that enhance paclitaxel activity in breast cancer cells.* Breast Cancer Research, 2010. **12**(3): p. R41.
167. Nath, S., et al., *MUC1 induces drug resistance in pancreatic cancer cells via upregulation of multidrug resistance genes.* Oncogenesis, 2013. **2**(6): p. e51.
168. Sharma, S. and A. Rao, *RNAi screening: tips and techniques.* Nature immunology, 2009. **10**(8): p. 799.
169. Mohr, S.E., et al., *RNAi screening comes of age: improved techniques and complementary approaches.* Nature reviews Molecular cell biology, 2014. **15**(9): p. 591.
170. Jonckheere, N., N. Skrypek, and I. Van Seuningen, *Mucins and tumor resistance to chemotherapeutic drugs.* Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2014. **1846**(1): p. 142-151.
171. Dahiya, R., et al., *Expression and characterization of mucins associated with the resistance to methotrexate of human colonic adenocarcinoma cell line HT29.* Cancer research, 1992. **52**(17): p. 4655-4662.
172. Lesuffleur, T., et al., *Differential expression of the human mucin genes MUC1 to MUC5 in relation to growth and differentiation of different mucus-secreting HT-29 cell subpopulations.* Journal of Cell Science, 1993. **106**(3): p. 771-783.
173. Leteurtre, E., et al., *Differential mucin expression in colon carcinoma HT-29 clones with variable resistance to 5-fluorouracil and methotrexate.* Biology of the Cell, 2004. **96**(2): p. 145-151.
174. Bafna, S., et al., *Pancreatic cancer cells resistance to gemcitabine: the role of MUC4 mucin.* British journal of cancer, 2009. **101**(7): p. 1155.
175. Mimeault, M., et al., *MUC4 down-regulation reverses chemoresistance of pancreatic cancer stem/progenitor cells and their progenies.* Cancer letters, 2010. **295**(1): p. 69-84.
176. Sheng, Y.H., et al., *MUC13 protects colorectal cancer cells from death by activating the NF-κB pathway and is a potential therapeutic target.* Oncogene, 2017. **36**(5): p. 700.

177. Hirono, S., et al., *Molecular markers associated with lymph node metastasis in pancreatic ductal adenocarcinoma by genome-wide expression profiling.* Cancer science, 2010. **101**(1): p. 259-266.
178. Kitamoto, S., et al., *Expression of MUC17 is regulated by HIF1α-mediated hypoxic responses and requires a methylation-free hypoxia responsible element in pancreatic cancer.* PloS one, 2012. **7**(9): p. e44108.
179. Kitamoto, S., et al., *DNA methylation and histone H3-K9 modifications contribute to MUC17 expression.* Glycobiology, 2010. **21**(2): p. 247-256.
180. Ho, S.B., et al., *Activity of recombinant cysteine-rich domain proteins derived from the membrane-bound MUC17/Muc3 family mucins.* Biochimica et Biophysica Acta (BBA)-General Subjects, 2010. **1800**(7): p. 629-638.
181. Luu, Y., et al., *Human intestinal MUC17 mucin augments intestinal cell restitution and enhances healing of experimental colitis.* The international journal of biochemistry & cell biology, 2010. **42**(6): p. 996-1006.
182. Volin, M.V., et al., *Expression of mucin 3 and mucin 5AC in arthritic synovial tissue.* Arthritis and rheumatism, 2008. **58**(1): p. 46-52.
183. Senapati, S., et al., *Expression of intestinal MUC17 membrane-bound mucin in inflammatory and neoplastic diseases of the colon.* Journal of clinical pathology, 2010. **63**(8): p. 702-7.
184. Yang, C.-W., et al., *Genetic variations of MUC17 are associated with endometriosis development and related infertility.* BMC medical genetics, 2015. **16**(1): p. 60.
185. LABoNNE, S.G., I. Sunitha, and A.P. Mahowald, *Molecular genetics of pecanex, a maternal-effect neurogenic locus of Drosophila melanogaster that potentially encodes a large transmembrane protein.* Developmental biology, 1989. **136**(1): p. 1-16.
186. Wang, Z., et al., *Targeting Notch signaling pathway to overcome drug resistance for cancer therapy.* Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2010. **1806**(2): p. 258-267.
187. Kim, B., et al., *Chemotherapy induces Notch1-dependent MRP1 up-regulation, inhibition of which sensitizes breast cancer cells to chemotherapy.* BMC cancer, 2015. **15**(1): p. 634.
188. Wang, Z., et al., *Cross-talk between miRNA and Notch signaling pathways in tumor development and progression.* Cancer letters, 2010. **292**(2): p. 141-148.
189. Thiery, J.P. and J.P. Sleeman, *Complex networks orchestrate epithelial-mesenchymal transitions.* Nat Rev Mol Cell Biol, 2006. **7**(2): p. 131-42.
190. Geisinger, A., et al., *The mammalian gene pecanex 1 is differentially expressed during spermatogenesis.* Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 2005. **1728**(1-2): p. 34-43.
191. Li, J., et al., *Pecanex functions as a competitive endogenous RNA of S-phase kinase associated protein 2 in lung cancer.* Cancer letters, 2017. **406**: p. 36-46.
192. Fojo, A., et al., *Reduced drug accumulation in multiply drug-resistant human KB carcinoma cell lines.* Cancer Research, 1985. **45**(7): p. 3002-3007.
193. Jin, W., et al., *MUC1 induces acquired chemoresistance by upregulating ABCB1 in EGFR-dependent manner.* Cell death & disease, 2017. **8**(8): p. e2980.
194. GROUP, E.B.C.T.C., *Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31 000 recurrences and 24 000 deaths among 75 000 women.* The Lancet, 1992. **339**(8784): p. 1-15.
195. Hassan, M., et al., *Chemotherapy for breast cancer.* Oncology reports, 2010. **24**(5): p. 1121-1131.

196. Mehta, S., et al., *Predictive and prognostic molecular markers for cancer medicine.* Therapeutic advances in medical oncology, 2010. **2**(2): p. 125-148.
197. James, C.R., et al., *BRCA1, a potential predictive biomarker in the treatment of breast cancer.* The oncologist, 2007. **12**(2): p. 142-150.
198. Wolmark, N., et al., *The prognostic significance of preoperative carcinoembryonic antigen levels in colorectal cancer. Results from NSABP (National Surgical Adjuvant Breast and Bowel Project) clinical trials.* Annals of surgery, 1984. **199**(4): p. 375.
199. Abdel-Fatah, T., et al., *Bcl2 is an independent prognostic marker of triple negative breast cancer (TNBC) and predicts response to anthracycline combination (ATC) chemotherapy (CT) in adjuvant and neoadjuvant settings.* Annals of oncology, 2013. **24**(11): p. 2801-2807.
200. Khouja, M.H., et al., *Limitations of tissue microarrays compared with whole tissue sections in survival analysis.* Oncology letters, 2010. **1**(5): p. 827-831.
201. Permuth-Wey, J., et al., *Sampling strategies for tissue microarrays to evaluate biomarkers in ovarian cancer.* Cancer Epidemiology and Prevention Biomarkers, 2009. **18**(1): p. 28-34.
202. Wong, A. and R.E. Cianciolo, *Comparison of immunohistochemistry and immunofluorescence techniques using anti-lambda light chain antibodies for identification of immune complex deposits in canine renal biopsies.* Journal of Veterinary Diagnostic Investigation, 2018. **30**(5): p. 721-727.
203. Akbar, S., et al., *Comparing computer-generated and pathologist-generated tumour segmentations for immunohistochemical scoring of breast tissue microarrays.* British journal of cancer, 2015. **113**(7): p. 1075.
204. Anders, C. and L.A. Carey, *Understanding and treating triple-negative breast cancer.* Oncology (Williston Park, NY), 2008. **22**(11): p. 1233.
205. Zlobec, I., et al., *Selecting immunohistochemical cut-off scores for novel biomarkers of progression and survival in colorectal cancer.* Journal of clinical pathology, 2007. **60**(10): p. 1112-1116.
206. Weiss, H.L., et al., *Receiver operating characteristic (ROC) to determine cut-off points of biomarkers in lung cancer patients.* Disease markers, 2004. **19**(6): p. 273-278.
207. Senapati, S., et al., *Expression of intestinal MUC17 membrane-bound mucin in inflammatory and neoplastic diseases of the colon.* Journal of clinical pathology, 2010. **63**(8): p. 702-707.
208. Graf, J.F. and M.I. Zavodszky, *Characterizing the heterogeneity of tumor tissues from spatially resolved molecular measures.* PloS one, 2017. **12**(11): p. e0188878.
209. Zhong, Q., et al., *A curated collection of tissue microarray images and clinical outcome data of prostate cancer patients.* Scientific data, 2017. **4**: p. 170014.
210. Liu, B., et al., *Genomic landscape and mutational impacts of recurrently mutated genes in cancers.* Mol Genet Genomic Med, 2018.
211. Tan, D., et al., *Definition of a region of loss of heterozygosity at chromosome 11q23. 3-25 in head and neck squamous cell carcinoma using laser capture microdissection technique.* Diagnostic Molecular Pathology, 2004. **13**(1): p. 33-40.
212. Hiley, C., et al., *Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine.* Genome Biol, 2014. **15**(8): p. 453.
213. Hao, J.-J., et al., *Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma.* Nature genetics, 2016. **48**(12): p. 1500.

214. Terra, S.B., et al., *Temporal and spatial heterogeneity of programmed cell death 1-Ligand 1 expression in malignant mesothelioma.* OncoImmunology, 2017. **6**(11): p. e1356146.
215. Gonzalez-Garay, M.L., *The road from next-generation sequencing to personalized medicine.* Personalized medicine, 2014. **11**(5): p. 523-544.
216. Esplin, E.D., L. Oei, and M.P. Snyder, *Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease.* Pharmacogenomics, 2014. **15**(14): p. 1771-1790.
217. Roychowdhury, S. and A.M. Chinnaiyan, *Translating cancer genomes and transcriptomes for precision oncology.* CA: a cancer journal for clinicians, 2016. **66**(1): p. 75-88.
218. Le Tourneau, C., et al., *Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial.* The lancet oncology, 2015. **16**(13): p. 1324-1334.
219. Tsimberidou, A.M. and R. Kurzrock, *Precision medicine: lessons learned from the SHIVA trial.* The Lancet Oncology, 2015. **16**(16): p. e579-e580.
220. Dakubo, G.D., et al., *Clinical implications and utility of field cancerization.* Cancer cell international, 2007. **7**(1): p. 2.
221. Braakhuis, B.J., C.R. Leemans, and R.H. Brakenhoff, *Using tissue adjacent to carcinoma as a normal control: an obvious but questionable practice.* The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 2004. **203**(2): p. 620-621.
222. Heaphy, C.M., J.K. Griffith, and M. Bisoffi, *Mammary field cancerization: molecular evidence and clinical importance.* Breast cancer research and treatment, 2009. **118**(2): p. 229-239.
223. Aran, D., et al., *Comprehensive analysis of normal adjacent to tumor transcriptomes.* Nat Commun, 2017. **8**(1): p. 1077.
224. Jones, S., et al., *Personalized genomic analyses for cancer mutation discovery and interpretation.* Sci Transl Med, 2015. **7**(283): p. 283ra53.
225. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nature genetics, 2013. **45**(10): p. 1127.
226. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333.
227. Liu, Y., Z. Hu, and C. DeLisi, *Mutated pathways as a guide to adjuvant therapy treatments for breast cancer.* Molecular cancer therapeutics, 2016. **15**(1): p. 184-189.
228. Chen, J., M. Sun, and B. Shen, *Deciphering oncogenic drivers: from single genes to integrated pathways.* Briefings in bioinformatics, 2014. **16**(3): p. 413-428.
229. Bailey, M.H., et al., *Comprehensive characterization of cancer driver genes and mutations.* Cell, 2018. **173**(2): p. 371-385. e18.
230. Hofree, M., et al., *Network-based stratification of tumor mutations.* Nature methods, 2013. **10**(11): p. 1108.
231. Andor, N., et al., *Pan-cancer analysis of the extent and consequences of intratumor heterogeneity.* Nature medicine, 2016. **22**(1): p. 105.
232. Middleton, F., et al., *Integrating genetic, functional genomic, and bioinformatics data in a systems biology approach to complex diseases: application to schizophrenia*, in *Neuroinformatics*. 2007, Springer. p. 337-364.

233. Greene, C.S. and O.G. Troyanskaya, *Accurate evaluation and analysis of functional genomics data and methods.* Annals of the New York Academy of Sciences, 2012. **1260**(1): p. 95-100.
234. Grade, M., et al., *A genomic strategy for the functional validation of colorectal cancer genes identifies potential therapeutic targets.* International journal of cancer, 2011. **128**(5): p. 1069-1079.
235. Liang, H., et al., *Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer.* Genome research, 2012.
236. Chin, L. and J.W. Gray, *Translating insights from the cancer genome into clinical practice.* Nature, 2008. **452**(7187): p. 553.
237. Vucic, E.A., et al., *Translating cancer 'omics' to improved outcomes.* Genome research, 2012. **22**(2): p. 188-195.
238. Liotta, L.A. and E. Petricoin, *Cancer biomarkers: closer to delivering on their promise.* Cancer cell, 2011. **20**(3): p. 279-280.
239. Druker, B.J., et al., *Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia.* New England Journal of Medicine, 2006. **355**(23): p. 2408-2417.
240. Soda, M., et al., *Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer.* Nature, 2007. **448**(7153): p. 561.

9. Appendix

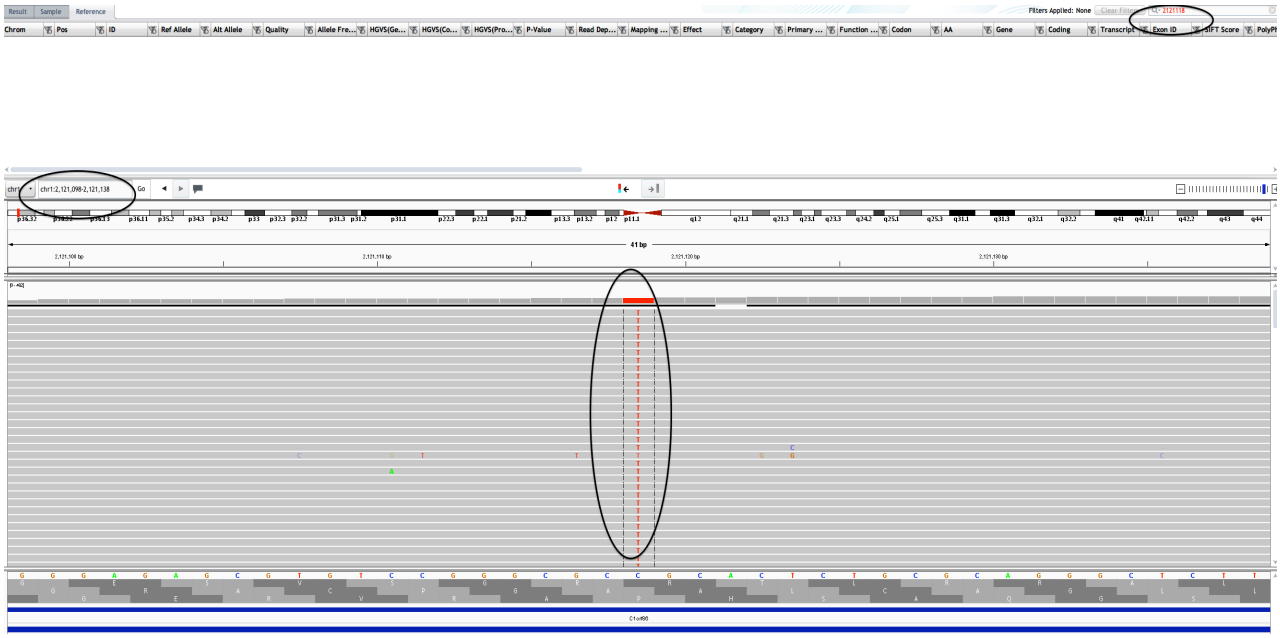
9.1 Cases 5-8 were sequenced in multiple separate reactions, and hence multiple FASTQ files were available. These were merged and checked for total number of sequenced reads in the final merged files, which corresponds to the added up total number of sequenced reads in files from first and second runs using FastQC software.

Sample ID	File No.	**First** run total reads No.	**Second** run total reads No.	**Final** merged files total reads No.
5 Pre-NACT	R1	20757727	109557256	130314983
	R2	20757727	109557256	130314983
5 Post-NACT	R1	8051567	47086305	55137962
	R2	8051567	47086305	55137962
6 Pre-NACT	R1	33232184	96097667	129329851
	R2	33232184	96097667	129329851
6 Post-NACT	R1	22598251	115852543	138450794
	R2	22598251	115852543	138450794
7 Pre-NACT	R1	13805863	127952441	141758304
	R2	13805863	127952441	141758304
7 Post-NACT	R1	20032085	73842906	93874991
	R2	20032085	73842906	93874991
8 Pre-NACT	R1	22177763	80345729	102523492
	R2	22177763	80345729	102523492
8 Post-NACT	R1	22462382	118959063	141421445
	R2	22462382	118959063	141421445

9.2 Screenshot of built-in genomic viewer in SureCall to illustrate the fault with SureCall software miscalling "somatic" variants in the cancer samples in fact had multiple reads aligned with high mapping quality in the matched normal samples, yet had not been called as germline variants. Agilent acknowledged that there was a fault in the SureCall software to recognise mutations present in both tumour and reference normal samples as germline mutations; instead, they were identified (incorrectly) as somatic variants, seemingly because the variants within the normal sequence had been (incorrectly) defined as sequencing errors. (A) shows an example of variant being called somatic under Result tab (calls only for somatic variants) in the genomic viewer (B) Same variant under Reference tab shows many reads supporting that variant in normal sample. A variant that has multiple reads, aligned with high mapping quality, all containing the same variant (with high base quality) and each aligned with a different start position (i.e. not possible to be PCR duplicates)

A

B



9.3 The findings from ToppGene enrichment analysis for further details of the other analyses such as statistics, scores and number of genes enriched in each term or pathway are included.

Enriched pathways for Pre-NAC sub-category									
ID	Name	Source	p-value	q-value Bonferroni	q-value FDR B&H	q-value FDR B&Y	Hit Count in Query List	Hit Count in Genome	Hit in Query List
M3005	Genes encoding collagen proteins	MSigDB C2 BIOCARTA (v6.0)	2.38E-11	6.76E-08	6.76E-08	5.77E-07	22	44	COL20A1,COL28A1,COL6A6,COL22A1,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL16A1,COL17A1,COL5A3,COL24A1,COL25A1,COL18A1,COL21A1
1470926	Collagen chain trimerization	BioSystems: REACTOME	1.24E-10	3.54E-07	1.77E-07	1.51E-06	22	47	COL20A1,COL28A1,COL6A6,COL22A1,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL16A1,COL17A1,COL5A3,COL24A1,COL25A1,COL18A1,COL21A1
1270246	Collagen biosynthesis and modifying enzymes	BioSystems: REACTOME	1.34E-09	3.82E-06	1.27E-06	1.09E-05	26	70	COL20A1,COL28A1,COL6A6,COL22A1,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL16A1,COL17A1,ADAMTS2,COL5A3,ADAMTS14,COL24A1,COL25A1,P3H1,COL18A1,TLL1,COL21A1
M5884	Ensemble of genes encoding core extracellular matrix including	MSigDB C2 BIOCARTA (v6.0)	1.34E-08	3.82E-05	9.12E-06	7.78E-05		275	MATN2,AEBP1,LTBP4,COL20A1,COL28A1,VWDE,BGN,COL6A6,OTOG,CHAD,COL22A1,VIT,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,CO

	ECM glycoproteins, collagens and proteoglycans								L15A1,COL16A1,COL17A1,ABI3BP,COL5A3,SNED1,VCAN,RELN,COL24A1,DMBT1,EPYC,EYS,EMILIN2,LAMC3,SVEP1,EFEMP1,SPON1,FRAS1,SLIT1,SLIT3,EGFLAM,COL25A1,ELSPBP1,SSPO,IGSF10,TINAG,NTNG2,VWA3A,COL18A1,THBS2,AGRN,USH2A,IGFBP5,COL21A1,LAMA5,LAMC1,PAPLN
1270245	Collagen formation	BioSystems: REACTOME	1.60E-08	4.56E-05	9.12E-06	7.78E-05	29	93	COL20A1,COL28A1,DST,COL6A6,COL22A1,MMP20,PLEC,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL16A1,COL17A1,ADAMTS2,COL5A3,ADAMTS14,COL24A1,COL25A1,P3H1,COL18A1,TLL1,COL21A1
1270244	Extracellular matrix organization	BioSystems: REACTOME	2.66E-07	7.56E-04	1.26E-04	1.08E-03	59	298	A2M,ACTN1,ADAM10,MMP2,MMP17,LTBP4,COL20A1,COL28A1,ADAM9,BGN,DST,COL6A6,CAPN2,CAPN3,CAST,SERPINE1,COL22A1,MMP20,PLEC,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL16A1,COL17A1,ADAMTS2,COL5A3,VCAN,PRKCA,ADAMTS14,KLK7,COL24A1,LAMC3,CAPN13,EFEMP1,COL25A1,P3H1,ADAM17,MMP25,COL18A1,TIMP1,TLL1,AGRN,TPSAB1,ITGA5,ITGA7,ITGA9,ITGAE,ITGB1,COL21A1,LAMA5,LAMC1
1270247	Assembly of collagen fibrils and other multimeric structures	BioSystems: REACTOME	3.67E-06	1.04E-02	1.49E-03	1.27E-02	19	60	DST,COL6A6,MMP20,PLEC,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL5A3,COL24A1,COL18A1,TLL1

P00034	Integrin signalling pathway	PantherDB	7.98E-06	2.27E-02	2.84E-03	2.42E-02	36	167	ACTN1,MAP3K4,COL20A1,CAV1,CDC42,ITGBL1,PIK3C2A,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL16A1,COL17A1,COL5A3,MAPK10,MAP2K3,DNAJC27,LAMC3,FLNB,ARHGAP26,TLN1,VCL,ITGA5,ITGA7,ITGA9,ITGAE,ITGB1,LAMA5,LAMC1
M5887	Genes encoding structural components of basement membranes	MSigDB C2 BIOCARTA (v6.0)	1.94E-05	5.53E-02	6.14E-03	5.24E-02	14	40	COL6A6,COL4A1,COL4A2,COL6A2,COL6A3,COL15A1,LAMC3,NTNG2,COL18A1,AGRN,USH2A,LAMA5,LAMC1,PAPLN
83067	Focal adhesion	BioSystems: KEGG	3.52E-05	1.00E-01	1.00E-02	8.54E-02	39	199	ACTN1,BIRC2,MYLK,COL6A6,CAPN2,CAV1,CDC42,CHAD,COL1A1,COL4A1,COL4A2,COL6A2,COL6A3,COL9A1,COL9A3,PRKCA,CTNNB1,MAPK10,RELN,TLN2,EGF,ROCK1,LAMC3,FLNB,FLT1,FLT4,SHC4,PDGFD,THBS2,TLN1,MYL10,VCL,ITGA5,ITGA7,ITGA9,ITGB1,PAK5,LAMA5,LAMC1
1383049	Diseases associated with O-glycosylation of proteins	BioSystems: REACTOME	5.92E-05	1.68E-01	1.53E-02	1.31E-01	17	60	MUC15,MUC3A,MUC5AC,NOTCH2,THSD7A,SEMA5A,MUC17,ADAMTS2,ADAMTS14,ADAMTSL2,MUC12,SPON1,SSPO,THBS2,ADAMTS19,MUC16,ADAMTS12
1269010	Diseases of glycosylation	BioSystems: REACTOME	9.42E-05	2.68E-01	2.11E-02	1.80E-01	21	86	MUC15,MUC3A,MUC5AC,BGN,NOTCH2,THSD7A,SEMA5A,MUC17,ADAMTS2,VCAN,CSPG4,ADAMTS14,ADAMTSL2,MUC12,SPON1,SSPO,THBS2,AGRN,ADAMTS19,MUC16,ADAMTS12
1269508	Rho GTPase cycle	BioSystems:	9.86E-05	2.81E-01	2.11E-02	1.80E-01	30	145	A2M,PREX1,ARHGAP21,FGD2,ARHGAP4,BCR,OPHN1,CDC42,ARHGAP33,CHN1,ARHGAP22,RHOF,ARHGE

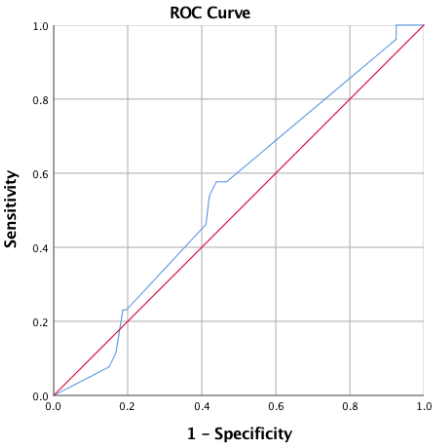
		REACTOME							F19,RHOD,ARHGEF4,DEPDC7,ARHGEF11,OBSCN,ARAP1,ARAP2,TAGAP,ARHGAP26,MCF2L,ARAP3,AKAP13,TRIO,PLEKHG2,ARHGAP40,ARHGEF5,RHOT2
1270257	Degradation of the extracellular matrix	BioSystems: REACTOME	1.04E-04	2.95E-01	2.11E-02	1.80E-01	25	112	A2M,ADAM10,MMP2,MMP17,ADAM9,CAPN2,CAPN3,CAST,MMP20,COL9A1,COL9A3,COL15A1,COL16A1,COL17A1,KLK7,CAPN13,COL25A1,ADAM17,MMP25,COL18A1,TIMP1,TLL1,TPSAB1,LAMA5,LAMC1
868086	Rap1 signaling pathway	BioSystems: KEGG	1.20E-04	3.40E-01	2.27E-02	1.94E-01	39	210	ADCY3,ADCY5,ADCY9,SIPA1L2,SKAP1,KRIT1,CDC42,MAGI1,PLCB2,CSF1R,PRKCA,PRKCI,PRKD1,CTNNB1,MAP2K3,TLN2,RALGDS,DRD2,EGF,PLCE1,MAPK12,FGF3,FGF6,FLT1,FLT4,MAGI3,SIPA1,RGS14,PDGFD,FYB1,SIPA1L3,PLCB1,GNAI3,RAPGEF4,TEK,ARAP3,TLN1,ID1,ITGB1
M5889	Ensemble of genes encoding extracellular matrix and extracellular matrix-associated proteins	MSigDB C2 BIOCARTA (v6.0)	1.40E-04	3.98E-01	2.49E-02	2.12E-01	139	1028	A2M,MATN2,ADAM10,AEBP1,MMP2,MMP17,LTBP4,SEMA7A,COL20A1,MUC15,COL28A1,BRINP2,MUC3A,MUC5AC,ADAM23,ADAM9,VWDE,BGN,COL6A6,SEMA5A,ARTN,PCSK6,SERPINE1,OTOG,PCSK5,CHAD,COL22A1,VIT,MMP20,SERPINE2,MUC17,PLAT,PLXNA1,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL16A1,COL17A1,ABI3BP,ADAMTS2,COL5A3,PPBP,CTSA,EGFL6,SNED1,VCAN,CSPG4,IL22,CST1,PLXNA4,ADAMTS14,ADAMTSL2,C1QTNF8,MASP1,RELN,TMPRSS15,COL24A1,FREM2,DMBT1,EPYC,REG1A,MUC12,SERPINA12,EYS,EGF,MEGF6,MEGF8,EMILIN2,LAMC3,F10,SVEP1,F13A

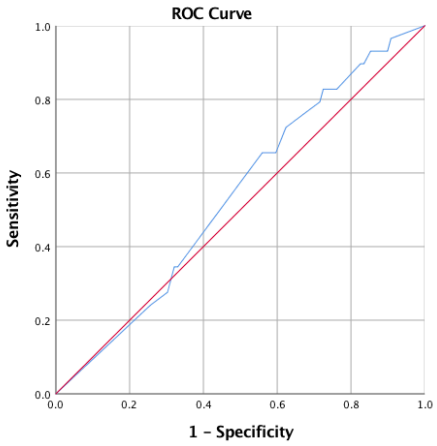
									1,EFEMP1,SPON1,FGF3,FGF6,CCL8,CLEC10A,PLXNA3,FLG,SEMA4B,FRAS1,SFTPD,PDGFD,SLIT1,SLIT3,EGFLAM,MEGF10,PLXND1,COL25A1,ELSPBP1,SSPO,FREM1,IGSF10,TINAG,NTNG2,P3H1,ADAM17,VWA3A,MMP25,TGM4,COL18A1,TGM2,THBS2,TCHH,THPO,TIMP1,TLL1,CLCF1,AGRN,TPO,ADAMTS19,PLXNB2,HRG,USH2A,IFNA21,IGFBP5,ELFN1,IL10,SEMA3G,FLG2,ITIH2,FAM20C,COL21A1,RPTN,KNG1,LAMA5,MUC16,LAMC1,PAPLN,ADAMTS12
1270313	NCAM1 interactions	BioSystems: REACTOME	1.76E-04	5.01E-01	2.95E-02	2.52E-01	12	37	CACNA1G,CACNA1C,COL6A6,ARTN,COL4A1,COL4A2,COL6A2,COL6A3,COL9A1,COL9A3,CNTN2,AGRN
172847	Protein digestion and absorption	BioSystems: KEGG	1.88E-04	5.35E-01	2.97E-02	2.54E-01	21	90	COL6A6,COL22A1,COL1A1,COL4A1,COL4A2,COL5A1,COL6A2,COL6A3,COL7A1,COL9A1,COL9A3,COL11A1,COL15A1,COL17A1,COL5A3,CPB2,COL24A1,KCNE3,SLC1A1,COL18A1,COL21A1
1309108	HDR through Single Strand Annealing (SSA)	BioSystems: REACTOME	2.34E-04	6.66E-01	3.51E-02	2.99E-01	12	38	ABL1,ATM,BARD1,BRCA1,DNA2,RAD9A,RFC2,RAD50,RPA1,RMI1,HUS1,WRN
1269343	Nitric oxide stimulates guanylate cyclase	BioSystems: REACTOME	3.35E-04	9.55E-01	4.77E-02	4.07E-01	9	24	NOS1,NOS2,NOS3,PDE2A,PDE1B,PDE11A,MRVI1,ITPR1,KCNMA1
213818	Glutamatergic synapse	BioSystems: KEGG	3.61E-04	1.00E+00	4.89E-02	4.17E-01	24	114	ADCY3,ADCY5,ADCY9,CACNA1A,CACNA1C,PLA2G4E,PLCB2,PLD1,PLD2,HOMER2,HOMER1,PRKCA,GRIN3B,

									GNB4,SLC1A1,SLC1A2,SLC1A3,GNB5,PLCB1,GNAI3,GNGT2,GRIA4,GRM8,ITPR1

Enriched pathways for Post-NAC sub-category									
ID	Name	Source	p-value	q-value Bonferroni	q-value FDR B&H	q-value FDR B&Y	Hit Count in Query List	Hit Count in Genome	Hit in Query List
M5884	Ensemble of genes encoding core extracellular matrix including ECM glycoproteins, collagens and proteoglycans	MSigDB C2 BIOCARTA (v6.0)	3.11 E-05	6.71E-02	3.25E-02	2.69E-01	25	275	MATN2,EMILIN3,FBN1,NTNG1,SSPO,BMPER,VWA3A,THBS2,OTOG,NYX,IGFBPL1,COL14A1,USH2A,COL4A6,COL6A1,COL7A1,COL13A1,COL5A3,SNED1,DMBT1,EPYC,LAMA5,LAMB1,LAMC1,PAPLN
83094	Type II diabetes mellitus	BioSystems : KEGG	3.87 E-05	8.35E-02	3.25E-02	2.69E-01	9	46	CACNA1A,CACNA1C,CACNA1E,HK2,PIK3CA,PIK3CD,IKBKB,MAPK10,INSR
P00034	Integrin signalling pathway	PantherDB	4.52 E-05	9.76E-02	3.25E-02	2.69E-01	18	167	ARFGAP1,TLN1,ITGBL1,PIK3CA,PIK3CD,COL14A1,COL4A6,COL6A1,COL7A1,COL13A1,COL5A3,MAPK10,ITGA2B,ITGAE,ITGB1,LAMA5,LAMB1,LAMC1
M5887	Genes encoding structural components of basement membranes	MSigDB C2 BIOCARTA (v6.0)	8.86 E-05	1.91E-01	4.78E-02	3.95E-01	8	40	NTNG1,USH2A,COL4A6,COL6A1,LAMA5,LAMB1,LAMC1,PAPLN

9.4 Receiver Operation Curve (ROC) analysis used to determine the cut-off scores for MUC17 and PCNX1 in Adjuvant cohort. Cut-off scores were established as follows: 1.1 for MUC17 and 5.6 for PCNX1.





Coordinates of the Curve

Test Result Variable(s): MUC17_expression

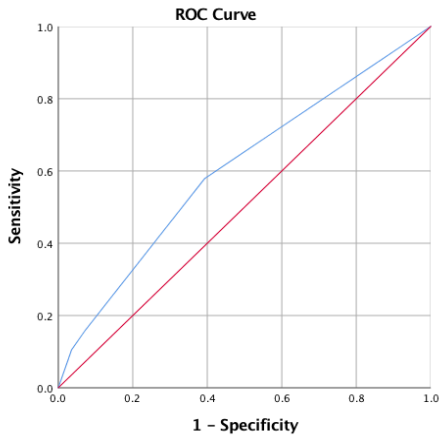
Positive if Greater Than or Equal To[a]	Sensitivity	1 – Specificity
-1.0000	1.000	1.000
.2500	1.000	.935
.6250	1.000	.925
.8750	.962	.925
1.1250	.577	.467
1.3750	.577	.439
1.6250	.538	.421
1.8750	.462	.411
2.1250	.231	.196
2.3750	.231	.187
2.6250	.115	.168
2.8750	.077	.150
4.0000	.000	.000

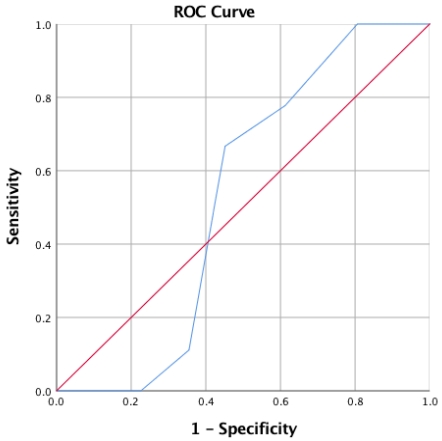
Coordinates of the Curve

Test Result Variable(s): PCNX_expression

Positive if Greater Than or Equal To[a]	Sensitivity	1 – Specificity
-1.0000	1.000	1.000
.2500	.966	.908
.9000	.931	.899
1.6500	.931	.890
2.2500	.931	.862
2.7500	.931	.853
3.2500	.897	.835
3.7500	.897	.826
4.1250	.828	.761
4.3750	.828	.743
4.6250	.828	.725
4.8750	.793	.716
5.2500	.724	.624
5.6250	.655	.596
5.8750	.655	.560
6.1250	.345	.330
6.3750	.345	.321
6.6250	.276	.303
6.8750	.241	.257
8.0000	.000	.000

9.5 Receiver Operation Curve (ROC) analysis used to determine the cut-off scores for MUC17 and PCNX1 in Neoadjuvant cohort. Cut-off scores were established as follows: 0.5 for MUC17 and 3.5for PCNX1.





Coordinates of the Curve

Test Result Variable(s): MUC17_expression

Positive if Greater Than or Equal To[a]	Sensitivity	1 – Specificity
-1.00	1.000	1.000
.50	.579	.393
1.50	.158	.071
2.50	.105	.036
4.00	.000	.000

Coordinates of the Curve

Test Result Variable(s): PCNX1_expression

Positive if Greater Than or Equal To[a]	Sensitivity	1 – Specificity
-1.00	1.000	1.000
1.00	1.000	.806
2.50	.778	.613
3.50	.667	.452
4.50	.111	.355
5.50	.000	.226
6.50	.000	.032
8.00	.000	.000