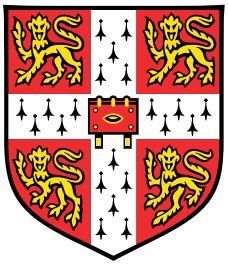
The clonal architecture and tumour microenvironment of breast cancers are shaped by neoadjuvant chemotherapy



Stephen John Sammut

University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare Hall

September 2018

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding figures, photographs, tables, appendices and bibliography.

Stephen John Sammut
September 2018

Acknowledgements

> The White Rabbit put on his spectacles.
> "Where shall I begin, please your Majesty?" he asked.
> "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."
>
> *Alice's Adventures in Wonderland*
> *Lewis Carrol*

Reflecting on this journey has allowed me to appreciate how truly fortunate I was to have met and worked with so many wonderful people in such a terrifically collaborative institution.

I am forever indebted to Professor Carlos Caldas for his unwavering support and mentorship. His guidance has been paramount towards my development as a clinician-scientist, and more importantly, as a person.

I am especially thankful to Suet-Feung Chin and Oscar Manuel Rueda, both senior post-doctoral scientists in the Caldas laboratory, for selflessly guiding me from start to finish and helping me acquire the scientific skills I required to undergo this work. Never have I learnt so much in so little time.

Recruitment into this study would not have been possible without the combined efforts of the Cambridge Breast Cancer Unit, including the research team (most notably Linda Jones and Emma Harrison), the recruiting oncologists, operating surgeons and radiologists. I felt privileged to be part of a team that is truly interested in making a difference.

Helen Bardwell was single-handedly responsible for sectioning all the tumour samples in this study. Elena Provenzano meticulously reviewed each and every slide and performed all post-therapy histopathological assessments. Many would benefit from taking a leaf out of their book and adopting their work ethic and can-do attitude. Likewise, the hard-working

members of the genomics core generated the whole exome and RNA sequencing libraries used in this study. Never have I seen so many people pull one rope in the same direction and expect nothing in return.

Ali Dariush, Wei Cope and Mireia Crispín Ortuzar generated the raw digital pathology data and provided a sounding board for many of my inter-disciplinary hypotheses. Turid Torheim and Ramona Woitek generated the unprocessed MRI volumetric data. Being able to foster collaborations with people from different disciplines was truly wonderful. I also thank Marion Karniely for her advice and support through the administrative elements of my PhD and am very grateful to all the members of the Caldas laboratory for providing an outstanding working environment.

I sincerely thank all the people who selflessly donated tissue for use in this study, knowing that they would not derive any benefit. I am also forever grateful to the funding bodies who showed confidence in my work by funding my research. The Wellcome Trust funded three years of my Clinical Research Training fellowship, whilst the Cancer Research (UK) Cambridge Institute funded my final year.

Lastly, my greatest thanks go to my loving wife, Audrienne, and my parents, John and Victoria. We are all the products of our environment and upbringing, and none of this would have been possible was it not for their unwavering and unconditional love and support.

Abstract

Neoadjuvant chemotherapy has become standard practice in patients with high-risk early breast cancer as it improves rates of breast conservation surgery and enables prediction of recurrence and survival by using response to treatment as a surrogate. Previous studies have focused on generating molecular datasets to develop prediction models of response, though little is known on how tumours and their microenvironments are modulated by neoadjuvant chemotherapy.

The thesis aims at molecularly characterising tumour changes during neoadjuvant chemotherapy in a cohort of 168 patients. Serial tumour samples at diagnosis, and, when available, midway through chemotherapy and on completion of treatment were profiled by shallow whole genome sequencing, deep exome sequencing and transcriptome sequencing, resulting in the generation of an unprecedented genomics dataset with tumours in situ while patients received chemotherapy.

Molecular predictors of response to chemotherapy were inferred from the diagnostic biopsy. Several novel observations were made, including previously undescribed associations between copy number alterations, mutational genotypes, neoantigen load, HLA genotypes and intra-tumoural heterogeneity with chemosensitivity. Possible mechanisms of chemoresistance included LOH at the MHC Class I locus, decreased expression of MHC Class I and II genes and drug influx molecules, as well as increased expression of drug efflux pumps. A complex relationship between proliferation, tumour microenvironment composition (TME) and response to treatment was explored by deconvoluting bulk RNAseq data and performing digital pathology orthogonal validation.

Clonal and microenvironment dynamic changes induced by/associated with chemotherapy were then modelled. Two types of genomic responses were identified, one in which the clonal composition was stable throughout treatment and another where clonal emergence and/or extinction was evident. Validation by multi-region deep sequencing confirmed the dynamics of the clonal landscape. Clonal emergence was shown to be associated with higher proliferation and decreased immune infiltrate, with an increase in genomic instability and homologous recombination deficiency during treatment. The immune TME composition and activity mirrored response to treatment, with cytolytic activity and innate and adaptive immune infiltrates linearly correlating with the degree of residual disease remaining after chemotherapy.

Finally, the circulating tumour DNA (ctDNA) genomic landscape was explored by using shallow whole genome sequencing and targeted sequencing of plasma DNA. Tumour mutations detected on exome sequencing were also detected in ctDNA in plasma, supporting the use of liquid biopsies as a biomarker for monitoring response to therapy and detection of minimal residual disease.

Table of contents

Acronyms

AC	Doxorubicin, Cyclophosphamide
AC→T	Doxorubicin, Cyclophosphamide, followed by Docetaxel
AF	Allelic Fraction
ATP	Adenosine Tri-Phosphate
AUC	Area Under Curve
BAF	B Allele Frequency
BAM	Binary Alignment
BCR	B Cell Receptor
CCF	Cancer Cell Fraction
cDNA	Complementary DNA
CDR3	Complementarity-Determining Region 3
cfDNA	Circulating Free DNA
CI	Confidence Interval
CIN	Chromosomal Instability
CMF	Cyclophosphamide, Methotrexate and 5-Fluorouracil
CNA	Copy Number Alterations
CP	Cellular Prevalence
CTA	Cancer/Testis Antigens
CTC	Circulating Tumour Cells
ctDNA	Circulating Tumour DNA
CYT	Cytolytic activity
DCE MRI	Dynamic Contrast Enhanced MRI
DFS	Disease Free Survival
DLDA	Diagonal Linear Discriminant Analysis

DNA Deoxyribonucleic acid

E→CMF Epirubicin, followed by Cyclophosphamide, Methotrexate and 5-Fluorouracil
ER Oestrogen Receptor

FC Fold Change
FDR False Discovery Rate
FEC 5-Fluorouracil, Epirubicin, Cyclophosphamide
FEC→T 5-Fluorouracil, Epirubicin, Cyclophosphamide, followed by Docetaxel
FFPE Formalin Fixed and Paraffin Embedded
FPKM Fragment Per Kilobase Millions

GATK Genome Analysis Toolkit
GGI Genomic Grade Index
GO Gene Ontology
GSEA Gene Set Enrichment Analysis
GSVA Gene Set Variation Analysis
GTF Gene Transfer Format
GVCF Genomic VCF

H&E Haematoxylin and Eosin
HER2 Human Epidermal growth factor Receptor 2
HLA Human Leukocyte Antigen
HR Hazard Ratio
HRD Homologous Recombination Deficiency

iC Integrative Cluster
IEDB Immune Epitope Database
IPS Immunophenogram Score

LOH Loss Of Heterozygozity

Mb Megabase
MCMC Markov Chain Monte Carlo
MHC Major Histocompatibility Complex
MRI Magnetic Resonance Imaging

NICE National Institute for Health and Care Excellence
NPV Negative Predictive Value
NSC Nearest Shrunken Centroids

OR Odds Ratio
OS Overall Survival

PCR Polymerase Chain Reaction
pCR Pathological Complete Response
PFS Progression Free Survival
PPV Positive Predictive Value
PR Progesterone Receptor

RCB Residual Cancer Burden
RD Residual Disease
REC Research Ethics Council
RNA Ribonucleic acid
ROC Receiver Operating Characteristic
RPM Rotations Per Minute

SNP Single Nucleotide Polymorphism
SNV Single Nucleotide Variant
ssGSEA Single Sample Gene Set Enrichment Analysis

T→FAC Docetaxel, followed by 5-Fluorouracil, Doxorubicin, Cyclophosphamide
T→FEC Docetaxel, followed by 5-Fluorouracil, Epirubicin, Cyclophosphamide
TAC Docetaxel, Doxorubicin, Cyclophosphamide
TC Docetaxel, Cyclophosphamide
TCGA The Cancer Genome Atlas
TCH Docetaxel, Carboplatin, Trastuzumab
TCR T Cell Receptor
TMB Tumour Mutation Burden
TME Tumour Microenvironment
TMEM Tumour Microenvironment of Metastasis
TMM Trimmed Mean of M-values

TNBC Triple Negative Breast Cancer
TPM Transcripts Per Millions

VAF Variant Allele Fractions
VCF Variant Calling Format

Chapter 1

Introduction

Contents

1.1 The dawn of adjuvant chemotherapy

Give up all hope oh ye who enter. Such was inscribed on the entrance to Paul Ehrlich's laboratory, an eminent German chemist who is regarded by many to be the father of chemotherapy. Choosing an inscription found at the entrance to Hell in Dante Alighieri's *Divina Commedia* (*"Lasciate ogni speranza, voi ch'intrate"*) [13] to bear greeting to all who entered his laboratory gives us a profound insight into the pessimism that prevailed in the 1900s: early chemotherapeutic agents showed little promise in the treatment of cancer.

While the search for effective and safe chemotherapeutic options continued, surgery remained the mainstay of breast cancer treatment. This was partly driven by the eminent American surgeon William Halsted, who strongly advocated radical mutilating surgery in order to increase chances of cure. Indeed, Halsted's radical mastectomy involved extensive incisions and tissue ablation to remove the affected breast, the underlying pectoral muscles and all draining lymph nodes [112]. The popularity of this operation resulted in over 90% of patients with breast cancer in the United States undergoing this operation until the 1970s and subsequently suffering from debilitating co-morbidities including profound lymphoedema of the arm and severe paraesthesia [212]. It was only in 1971 that Bernard Fisher revealed the unnecessity of such radical surgery by publishing a trial comparing the survival following the Halsted mastectomy to a more conservative modified radical mastectomy [80].

Despite improvements in surgical techniques and radiotherapy regimens, it rapidly became apparent that cure rates were limited by micro-metastatic disease and that cytotoxic agents would be necessary to eradicate distant deposits. In 1976, the combination of cyclophosphamide, methotrexate and 5-fluorouracil (CMF) given over 12 two-weekly cycles, was the first chemotherapy regimen to show a statistically significant reduction in recurrence rates after radical mastectomy, leading to the birth of the adjuvant post-surgical setting [35]. The motivation to develop a treatment given over a shorter period that was less emetogenic led to the investigation of a regimen consisting of doxorubicin and cyclophosphamide (AC), given over two months. The results of this trial were published in 1990, and showed an identical outcome for AC and CMF chemotherapy, with significantly less nausea and a shorter duration of treatment [81].

In the next few years, various permutations of these two keystone regimens were developed in order to improve efficacy. The most notable change involved substituting methotrexate in the CMF regimen with an anthracycline (doxorubicin/epirubicin), giving rise to FAC [187] and FEC [89]. Both of these regimens showed a superior disease-free survival however this

came at a cost of increased clinical toxicity. In another study, four cycles of epirubicin were given prior to four cycles of CMF (rather than replacing methotrexate, as had been done previously). Patients treated with E→CMF had a significantly higher relapse-free and overall survival (OS) compared to the CMF group, once again at the cost of a higher incidence of adverse events [227].

The discovery of taxanes in the 1970s, together with their introduction in metastatic and early-stage breast cancer regimens, proved to be an important advance. In the United States, AC→T was shown to be more effective than the standard AC protocol, with modest additional toxicities [120, 184]. Meanwhile, in Europe, taxanes were added sequentially to the already established FEC regimen: FEC→T showed a 5% additional increase in five-year disease-free survival (DFS) [238] and is now the mainstay of treatment in node-positive breast cancer in the UK [210].

Patients with tumours exhibiting amplification of the *ERBB2* gene (described as $HER2^+$ tumours) often had a much poorer prognosis, with significantly shorter overall survival and time to relapse despite adjuvant therapies [267]. The development of trastuzumab, a monoclonal antibody targeting the extracellular domain of the HER2 protein, revolutionised the way $HER2^+$ tumours were treated, with impressive results seen in the metastatic setting, including longer time to disease progression (7.4 vs. 4.6 months), higher rate of objective response (50% vs. 32%), longer duration of response (9.1 vs. 6.1 months) and longer overall survival (25.1 vs. 20.3 months) [268]. Adjuvant trials were rapidly designed to determine whether such a benefit would also be seen in the adjuvant setting. In 2000, within the space of a few months, two very similar trials were launched: the *National Surgical Adjuvant Breast and Bowel Project (NSABP) B-31* trial and *North Central Cancer Treatment Group (NCCTG) N9831* trial, both of which assessed the improvement in survival gained by adding trastuzumab to adjuvant AC→T. A combined analysis of both of these trials [239] showed a great benefit in adding trastuzumab to chemotherapy regimens in $HER2^+$ disease, with an absolute difference in DFS of 12% in the trastuzumab group and a corresponding 33% decrease in the risk of death.

As the early clinical studies had shown that both trastuzumab and anthracyclines were cardiotoxic, the efficacy and safety of non-anthracycline regimens with trastuzumab was evaluated in the *Breast Cancer International Research Group (BCIRG)* 006 clinical trial [266], where 3,222 women with $HER2^+$ early-stage breast cancer were randomised to receive either AC→T, AC→T+trastuzumab or docetaxel, carboplatin and trastuzumab (TCH). The addition of one year of adjuvant trastuzumab confirmed the significantly improved DFS and

OS observed in the previous trials, though no difference in DFS or OS was detected between the two trastuzumab regimens, indicating equal efficiency. The non-anthracycline regimen, however, had a lower incidence of adverse events, with significantly lower rates of congestive heart failure and cardiac dysfunction.

Clearly, treatment with trastuzumab conferred an unprecedented survival benefit. However, until the late 2000s, the optimum duration of therapy was still a subject of great debate. The *HERA (BIG 1-01)* trial [269] sought to answer this by assigning 5,102 women to receive either adjuvant trastuzumab for 1 year, trastuzumab for 2 years or no anti-HER2 therapy at all. After a median follow-up of 11 years, the addition of trastuzumab was shown to significantly increase DFS and reduce the risk of death (hazard ratio: 0.74). There was no difference between the two-year and one-year arms of the study [45], and subsequently, one year of adjuvant trastuzumab became standard of care [209]. Clinical trials, including the *Persephone* trial [124], have now been set up to determine whether adjuvant exposure to trastuzumab can be reduced to six months, rather than a year.

In the post-trastuzumab era various efforts were made to create compounds that synergistically blocked the HER2 pathway. Pertuzumab, a recently introduced anti-HER2 agent which inhibits HER2 heterodimerisation with other HER receptors was shown to improve survival in the metastatic setting when administered with trastuzumab and chemotherapy [275]. The *APHINITY* trial [302] sought to determine whether adjuvant pertuzumab could also improve DFS, and randomised 4,805 patients to receiving trastuzumab and chemotherapy with or without pertuzumab. The addition of pertuzumab minimally improved the DFS rates among patients with $HER2^{+}$ tumours, with three-year rates of invasive DFS of 94.1% in the pertuzumab group and 93.2% in the non-pertuzumab cohort.

1.2 Downstaging with neoadjuvant chemotherapy

Following the successes described in the adjuvant post-surgical setting, efforts were undertaken in the 1980s to determine whether radiotherapy and chemotherapy could be administered to downstage locally advanced breast cancer, thereby rendering inoperable tumours operable. In one of the first published studies [244], 24 patients were assigned to receiving one of two chemo-radiotherapy regimens, with high objective regression rates (83% and 92%) but high subsequent relapse rates (50% and 58%). Both chemotherapy and radiotherapy allowed the downstaging of these previously inoperable tumours and were shown to provide excellent local control [129, 130, 188]. Previously inoperable tumours were now becoming operable: this was seen as a great triumph in the field.

These successes prompted the usage of pre-surgical (or neoadjuvant) chemotherapy regimens in the earlier breast cancer setting. Rather than limiting its use to patients with inoperable tumours, physicians sought to determine whether neoadjuvant chemotherapy could be used to downstage larger operable tumours. Perhaps, by administering pre-surgical chemotherapy to patients with larger tumours, less radical surgery might be required. Agents and regimens that had been explored in the adjuvant setting were rapidly investigated in the neoadjuvant setting, including CMF [232], VTMF [143], FAC [253, 254], AVCFM [27] and high dose FEC [54]. The results seen in the inoperable setting were reproduced in patients with larger operable tumours: neoadjuvant chemotherapy downstaged tumours and allowed for higher rates of breast conservation surgery.

However, neoadjuvant chemotherapy remained an experimental form of therapy until the mid-1990s [36]. The fact that neoadjuvant therapies caused a decrease in tumour bulk and allowed more breast-conserving surgery was incontrovertible, however there was no solid evidence to support the claim that primary chemotherapy was not inferior to adjuvant chemotherapy.

The publication of the *NSABP-18* trial in 1998 [82] paved the way to increased usage of neoadjuvant therapies. In this trial, 1,523 women were assigned to preoperative or postoperative AC therapy, and clinical tumour response graded as complete, partial or no response. No difference was observed in DFS and OS between the adjuvant and neoadjuvant cohorts, however breast conservation surgery was more frequently performed in the neoadjuvant group. The trial also reported similar rates of ipsilateral breast tumour recurrence (7.9% in neoadjuvant cohort and 5.8% in adjuvant cohort) and showed that response to chemotherapy could be used as a surrogate for overall outcome: women who attained pathological complete

response (pCR), that is, absence of any residual tumour cells, had relapse-free survival rates of 85.7%, compared to those with a partial clinical response (68.1%) and no response (63.9%). For all intents and purposes, the neoadjuvant setting provided many benefits, with no evidence of decreased efficacy compared to the adjuvant regimens.

Following the publication of this milestone trial, the European *EORTC 10902* trial [295] randomised 698 breast cancer patients to receive FEC either before or after surgery. After 56 months of follow-up there was no statistically significant difference in OS, progression-free survival (PFS) and time to loco-regional recurrence. Once again, neoadjuvant chemotherapy was shown to significantly increase the rate of breast conservation surgery. The publication of these landmark trials resulted in widespread acceptance of the benefits that neoadjuvant therapies conferred, and became standard of care for downstaging inoperable tumours in order to increase rates of breast-conserving surgery [210].

Neoadjuvant regimens combining both taxanes and anthracyclines rapidly became standard of care following the publication of the *NSABP B-27* clinical trial, which randomised 2,411 patients to receive either neoadjuvant AC, neoadjuvant AC→T, or four cycles of neoadjuvant AC and four cycles of adjuvant docetaxel [23, 24]. The results from this trial showed that the addition of a taxane greatly increased both clinical complete response rate (63.6% AC→T vs. 40.1% AC) and pCR rate (26.1% AC→T vs. 13.7% AC). Hence, combination chemotherapy with both a taxane and anthracycline conferred a synergistic benefit.

While the standard of care rapidly became treatment with anthracyclines and taxanes, the optimum sequencing of these therapies was unknown. In the *Neo-tAnGo* trial [73], 831 women were assigned to receive either epirubicin and cyclophosphamide then paclitaxel (with or without gemcitabine) or paclitaxel (with or without gemcitabine) then epirubicin and cyclophosphamide. Treatment with a taxane prior to an anthracycline was shown to be superior to the reversed regimen, with more patients attaining pCR if a taxane was administered first (20% vs. 15%). A subsequent meta-analysis confirmed this observation [32]. Hence not only did the agents that were delivered mattered but so did the sequence in which they were administered.

It is worth noting that, as the response to chemotherapy was often dramatic with no residual tumour detectable macroscopically, localisation at the time of surgery or during imaging was frequently difficult or even impossible. Fiduciary radio-opaque markers were developed for insertion into a tumour prior to commencing chemotherapy, making localisation at the time

of surgery much easier, especially if pCR had occurred with no evidence of any residual disease [21, 62, 109]. This has now become standard practice.

As the benefits of neoadjuvant chemotherapy were being explored in clinical trials, it rapidly became apparent that the degree of response to chemotherapy differed across breast tumours. Various meta-analyses concluded that tumours that showed high expression of the oestrogen receptor (ER^+) had very low pCR rates whilst those with low or no expression (ER^-) were 12 times more likely to attain pCR [57, 58]. Classification using the intrinsic subtypes [221] showed that basal-like and $HER2^+$ subgroups were associated with the highest rates of pCR (45% in both), whilst luminal tumours had a pCR rate of only 6% [243]. Additionally, a meta-analysis of 9,020 breast cancer patients from nine German neoadjuvant trials showed that patients with invasive lobular carcinoma had significantly lower rates of pCR (6.2% vs. 17.4% in all other histologies) with higher mastectomy rates [177]. Markers of high proliferation, including high histological grade, lymph node positive disease and lymphovascular invasion, were also associated with increased rates of pCR [84, 91, 136, 172, 243]. Younger age, as well as lower body mass index, were correlated with higher probabilities of attaining pCR [85].

Treatment optimisation for triple negative tumours

As discussed above, patients with triple negative breast cancer (TNBC, defined as having low or no expression of oestrogen and progesterone receptors (ER^-, PR^-), and lack of HER2 amplification ($HER2^-$)) have consistently showed higher rates of pCR following neoadjuvant treatment with anthracyclines, cyclophosphamide and taxanes, compared to patients with ER^+ tumours. As shown by the *GeparTrio* study that recruited 2,072 patients to evaluate treatment with six to eight cycles of TAC or two cycles of TAC followed by four cycles of vinorelbine and capecitabine, the highest pCR rate (57%) was observed in patients below the age of 40 with TNBC or grade 3 tumours [136].

Following the observation from preclinical data that triple negative tumours were increasingly sensitive to interstrand cross-linking agents such as platinum salts due to deficiencies in BRCA-associated DNA repair mechanisms, the *GeparSixto (GBG 66)* trial [303] was designed to assess whether the addition of neoadjuvant carboplatin to a regimen containing an anthracycline, a taxane, and targeted therapy improved rates of complete response in patients with TNBC and $HER2^+$ tumours. 588 patients were recruited to the study and the addition of carboplatin was shown to improve pCR rates in the TNBC cohort (53.2% vs. 36.9% without

carboplatin), but not in the HER2$^+$ group (32.8% vs. 36.8% without carboplatin). Long-term follow-up in the *GeparSixto* study also showed an increased event-free survival in TNBC patients treated with carboplatin [111].

The *Alliance (CALGB 40603)* trial [263] further confirmed the *GeparSixto* findings. 443 patients with TNBC were randomised to receive paclitaxel and dose dense AC with or without concurrent carboplatin, and the addition of carboplatin was shown to increase the rates of pCR in the breast and axilla (54% vs. 41% without carboplatin), but was associated with increased grade 3 and 4 adverse events, such as neutropenia and thrombocytopenia.

I-SPY 2 [245] assessed the benefit of adding carboplatin and the PARP inhibitor veliparib to standard of care chemotherapy (T→AC). The estimated rates of pCR in the TNBC cohort were 51% in the veliparib-carboplatin group and 26% in the control group, though significant additional toxicities were observed in the PARP/platinum group, confirming the toxicity findings observed in the *GeparSixto* and *Alliance* trials. However, it was uncertain whether the benefit observed was specifically from the PARP inhibitor or the platinum agent, and therefore the *BrighTNess* trial [176] was subsequently designed to further investigate this. 634 patients were randomly assigned to one of three groups: paclitaxel+carboplatin+veliparib, paclitaxel+carboplatin or paclitaxel only. Patients receiving paclitaxel and carboplatin had the highest pCR rate (58%), with patients receiving paclitaxel, carboplatin, and veliparib having a pCR rate of 53%, and those receiving paclitaxel alone having a pCR rate of 31%. Hence there was no superior benefit gained by the addition of veliparib to platinum-based therapies. Further trials, such as *Partner*, are now investigating whether other PARP inhibitors are more effective in this setting.

Treatment optimisation for HER2$^+$ tumours

As was repeatedly shown in the metastatic and adjuvant settings, trastuzumab revolutionised the prognosis of patients with HER2$^+$ tumours. In order to assess the efficacy of trastuzumab in the neoadjuvant setting, the *NOAH* trial [99] recruited 235 patients with HER2$^+$ breast cancer and randomised them to receive neoadjuvant doxorubicin, paclitaxel, cyclophosphamide, methotrexate, and fluorouracil with or without trastuzumab. The addition of trastuzumab was shown to significantly improve three-year event-free survival (71% in the trastuzumab group vs. 56% in the non-trastuzumab group). 38% of patients receiving trastuzumab attained pCR, compared to 19% without trastuzumab, with a similar degree of adverse events. The addition

of trastuzumab to neoadjuvant regimens rapidly became standard of care for $HER2^+$ disease [210].

Following the approval of pertuzumab in the metastatic setting [275], the *NeoSphere* trial [100, 101] randomised 417 $HER2^+$ patients between 2007-2009 to receive either trastuzumab + docetaxel (group A), pertuzumab + trastuzumab + docetaxel (group B), pertuzumab + trastuzumab (group C) or pertuzumab + docetaxel (group D). Patients in group B had a significantly improved pCR rate (45.8%), compared to those in group A (29.0%), group C (16.8%) and group D (24.0%), showing that the addition of pertuzumab to these neoadjuvant regimens did increase pCR rate considerably. Five-year follow-up data showed that patients allocated to group B had the highest five-year PFS (86%) and DFS (84%), whilst those in group A had a PFS of 81% and a DFS of 81%. Similar findings were reported a year later by the *TRYPHAENA* trial [252]: a high proportion of patients attained pCR (57.3–66.2%) when pertuzumab and trastuzumab were added to neoadjuvant chemotherapy. The impressive performance of combined pertuzumab and trastuzumab regimens rapidly led to their adoption by NICE in 2016.

Since the inclusion of anti-HER2 agents within neoadjuvant regimens showed a dramatic improvement in pCR rates, various groups hypothesised that the main benefit was derived from the targeted therapy, rather than the cytotoxic backbone. The phase III *KRISTINE* trial [137] sought to determine whether neoadjuvant chemotherapy could be replaced by neoadjuvant targeted therapy only. 444 women with $HER2^+$ cancer were randomised to receive either neoadjuvant docetaxel, carboplatin, trastuzumab and pertuzumab or trastuzumab emtansine and pertuzumab. A higher rate of pCR was seen in the docetaxel, carboplatin, trastuzumab and pertuzumab group (56%) compared to the trastuzumab emtansine plus pertuzumab group (44%), suggesting that cytotoxic chemotherapy should continue to play a key role in the treatment of $HER2^+$ disease.

Long-term follow-up

Clearly the neoadjuvant setting offers many advantages. Neoadjuvant chemotherapy can (1) render inoperable tumours operable, (2) allow conservative surgery through down-staging of the disease [163, 183], (3) improve overall survival by eradicating micrometastatic disease and (4) allow the in vivo assessment of tumour response to chemotherapy allowing speculation of prognosis.

Many trials showed very few or no drawbacks to neoadjuvant therapies when compared to adjuvant therapies, though these conclusions were made in the absence of mature long-term follow-up data. A very recent landmark meta-analysis released by the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) [74] combined clinical trial data from 4,756 women with breast cancer recruited to ten randomised trials between 1983 to 2002, with a median follow-up of 9 years, and compared neoadjuvant chemotherapy regimens with identical regimens administered after surgery. It is worth noting, however, that none of the recruited patients received anti-HER2 therapies.

28% of the patients analysed within the EBCTCG study had a complete response, 41% had a partial response, whilst 31% had stable or progressive disease. Using neoadjuvant treatment increased rates of breast conservation surgery (65% in the neoadjuvant group vs. 49% in the adjuvant chemotherapy group), and treatment with combined anthracyclines and taxanes was also associated with greater probability of less radical surgery, compared to other chemotherapy regimens. There was no statistically significant difference between neoadjuvant and adjuvant treatment in 15-year distant recurrence (38.2% vs. 38.0%), breast cancer death (34.4% vs. 33.7%) or death from any cause (40.9% vs. 41.2%), showing that neoadjuvant chemotherapy was not inferior to adjuvant treatment.

Notwithstanding these benefits, neoadjuvant chemotherapy was associated with a moderately greater incidence of local recurrence when compared to similar sized tumours treated with adjuvant treatment, which persisted for at least 10 years. The authors observed that this was due to an increased use of breast-conserving therapy in women who would have otherwise had mastectomy in the absence of a good response, and suggested that difficult tumour localisation, with lack of clearly palpable margins, as well as heterogeneous response increased the technical difficulty of surgery, making complete tumour removal challenging. Reassuringly, the authors noted that the increase in local recurrence was not associated with an increase in distant recurrence or breast cancer mortality.

1.3 Assessment of response to neoadjuvant therapies

Response to neoadjuvant chemotherapy following its introduction was often assessed clinically or by serial ultrasound assessments, with most trials classifying pathological response as a dichotomous metric, that is, complete histological response or remaining residual invasive carcinoma, with no quantification of the degree of residual disease (RD) remaining [24, 37, 54, 82, 126, 153]. The different histological appearances of response have been well described, with two general morphologies: concentric shrinkage of the tumour mass or scattered foci over an ill-defined tumour bed [40, 246]. Despite these very obvious post-therapy appearances, early classifications were agnostic to the degree of remaining RD.

Various efforts were made to develop new classification systems of histological RD and these were often based on the cellularity of a tumour at the time of biopsy. The Miller and Payne system [217], for example, graded tumours into 5 categories, with Grade 1 tumours showing no alteration in overall cellularity, and Grade 5 tumours defined as pCR, however this classification did not take into account the primary tumour bulk. Sataloff and colleagues [249] developed a histological classification with four response categories: total/near total therapeutic effect, greater than 50% therapeutic effect, less than 50% therapeutic effect and little or no effect. In the Honkoop classification [127], two categories of response were described, minimal residual disease and gross residual disease. The Kuerer classification [160] described three categories of response based on volume: pCR, less than 1 cm^3 of residual tumour macroscopically and more than 1 cm^3 of residual tumour. These classifications were alternative methods of trying to model response as a continuum, rather than a dichotomous variable, but were never formally universally adopted, with most centres preferring to use the simplistic classification adopted by the *NSABP B-18* trial described previously [82].

The Residual Cancer Burden (RCB) score was published in 2007 [279] and harmonised the way the degree of residual cancer post-therapy was categorised. This classification brought together approaches used by other efforts, such as tumour area and cellularity, and added in other previously unappreciated variables such as the total lymph node disease burden.

The numeric RCB score was defined as the sum of primary tumour bed variables as well as lymph node variables, as shown in equation 1.1:

$$RCB \;=\; 1.4(d_{prim} \times f_{inv})^{0.17} \;+\; [4(d_{met} \times (1-0.75^{LN})]^{0.17} \tag{1.1}$$

where d_{prim} (area of the tumour bed) and f_{inv} (overall cancer cellularity as a percentage of area) were tumour variables and d_{met} (diameter of the largest lymph node metastasis) and *LN* (number of axillary lymph nodes containing carcinoma) were lymph node variables.

The tumour bed area (d_{prim}) was defined as the geometric mean of the largest bidimensional measurements of the residual primary tumour bed (Equation 1.2):

$$d_{prim} = \sqrt{d_1 \times d_2} \tag{1.2}$$

Whilst the percentage area with invasive cancer (f_{inv}) was defined as shown in Equation 1.3:

$$f_{inv} = \left(1 - \frac{\%\ In\ situ\ cancer}{100}\right) \times \left(\frac{\%\ Tumour\ cellularity}{100}\right) \tag{1.3}$$

The numeric RCB score ranged from 0 (pCR) with no upper bounds. Cut-off points were selected by maximizing the profile log-likelihood of a multivariate Cox model [279]. Tumours with a score less than 1.36 (but higher than 0) were classified as having RCB-I (minimal) RD, tumours with scores between 1.36 and 3.28 were classified as having RCB-II (moderate) RD and tumours with scores above 3.28 were classified as RCB-III (extensive) RD. By providing a continuous variable, the RCB score moved away from the dichotomization of response and was able to model a broad range of actual responses from near pCR to frank resistance to therapy. Tumours attaining pCR or RCB-I disease after neoadjuvant chemotherapy were subsequently classified as chemosensitive, tumours with RCB-II and RCB-III disease classified as non-chemosensitive, and tumours with RCB-III RD classified as chemoresistant [117].

The reproducibility of the score was recently shown in a study performed by the MD Anderson Cancer Center, where five pathologists were asked to review 100 random cases and assign an RCB score without any prior coaching [224]. The overall concordance of agreement in RCB score among all five pathologists was 0.93, with an overall accuracy of 0.989, indicating high reproducibility. Since publication, the RCB score has been adopted as a primary or secondary endpoint of chemotherapy response in several major trials, including *I-SPY 1* [76, 77, 46], *I-SPY 2* [19], *GEICAM* [258], *CALGB 40601* [47], *CALGB 40603* [263] and *NSABP B41* [25].

Increasing RCB score was shown to correlate with poor prognosis and increased probability of distant relapse [279]. Patients with RCB-I RD had the same 5-year prognosis as those

with pCR, whilst extensive RD (RCB-III) was associated with poor prognosis. In a study by Symmans et al. [279], all patients with RCB-III disease after neoadjuvant chemotherapy who did not receive adjuvant hormonal therapy were noted to relapse within 3 years. 13% of patients with ER^+ tumours had RCB-III RD after chemotherapy and had a 5-year distant relapse rate of 40% despite receiving adjuvant hormonal therapies [279].

In a recent update by Symmans et al. [280], 1,158 patients treated with chemotherapy (T→FAC, FAC and trastuzumab+T→FAC) were recruited to five prospective breast cancer studies designed to assess the long-term prognostic relevance of the RCB score. The reported distribution of cases within each RCB class are shown in Table 1.1, with ER^-HER2^- or $HER2^+$ tumours most likely to attain pCR on therapy. As had been shown in the original reporting publication [279], the RCB score was prognostic across different ER and HER2 subsets, with 10-year relapse-free survival rates in the four response/disease burden categories across different phenotypes as shown in Table 1.2. Patients with ER^-HER2^- or $HER2^+$ tumours had a very poor 10-year relapse-free survival (23% and 21% respectively) if extensive residual disease was present at the time of surgery despite neoadjuvant chemotherapy, with all groups having excellent 10-year relapse-free survival if pCR was attained.

Table 1.1: Proportion of patients within each RCB category across ER/HER2 groups, as published by Symmans et al. [280]

	pCR	RCB-I	RCB-II	RCB-III
ER^- $HER2^-$	35%	15%	33%	17%
ER^+ $HER2^-$	10%	13%	60%	17%
$HER2^+$	45%	19%	29%	7%

Table 1.2: 10-year relapse-free survival rates stratified by RCB category and ER/HER2 status, as published by Symmans et al. [280]

	pCR	RCB-I	RCB-II	RCB-III
ER^- $HER2^-$	86%	81%	55%	23%
ER^+ $HER2^-$	83%	97%	74%	52%
$HER2^+$	95%	77%	47%	21%

As shown in Figure 1.1 the RCB classification was prognostic in all treatment cohorts and ER/HER2 classes [280]. The rate of pCR was highest in $HER2^+$ patients treated with trastuzumab and chemotherapy (Figure 1.1D), where 45% attained a pCR. Patients with ER^-HER2^- disease (Figure 1.1A) had the second highest rate of pCR (35% pCR), and only 10% of ER^+HER2^- patients (Figure 1.1B) attained pCR, showing that ER^+ tumours were the least chemosensitive.

Additionally, lower RCB categories were associated with longer relapse-free survival and overall survival across all ER/HER2 subgroups. The hazard ratio (HR) for overall survival in patients who attained pCR was 0.72 in $HER2^+$ disease (95% CI, 0.63–0.82), 0.76 in ER^-HER2^- disease (95% CI, 0.70–0.82) and 0.73 in ER^+HER2^- disease (95% CI, 0.67–0.78). Patients who attained pCR had excellent long-term prognoses: this was particularly evident in the ER^-HER2^- and $HER2^+$ cohorts, where the presence of RCB-III disease was associated with a significantly reduced relapse-free survival.

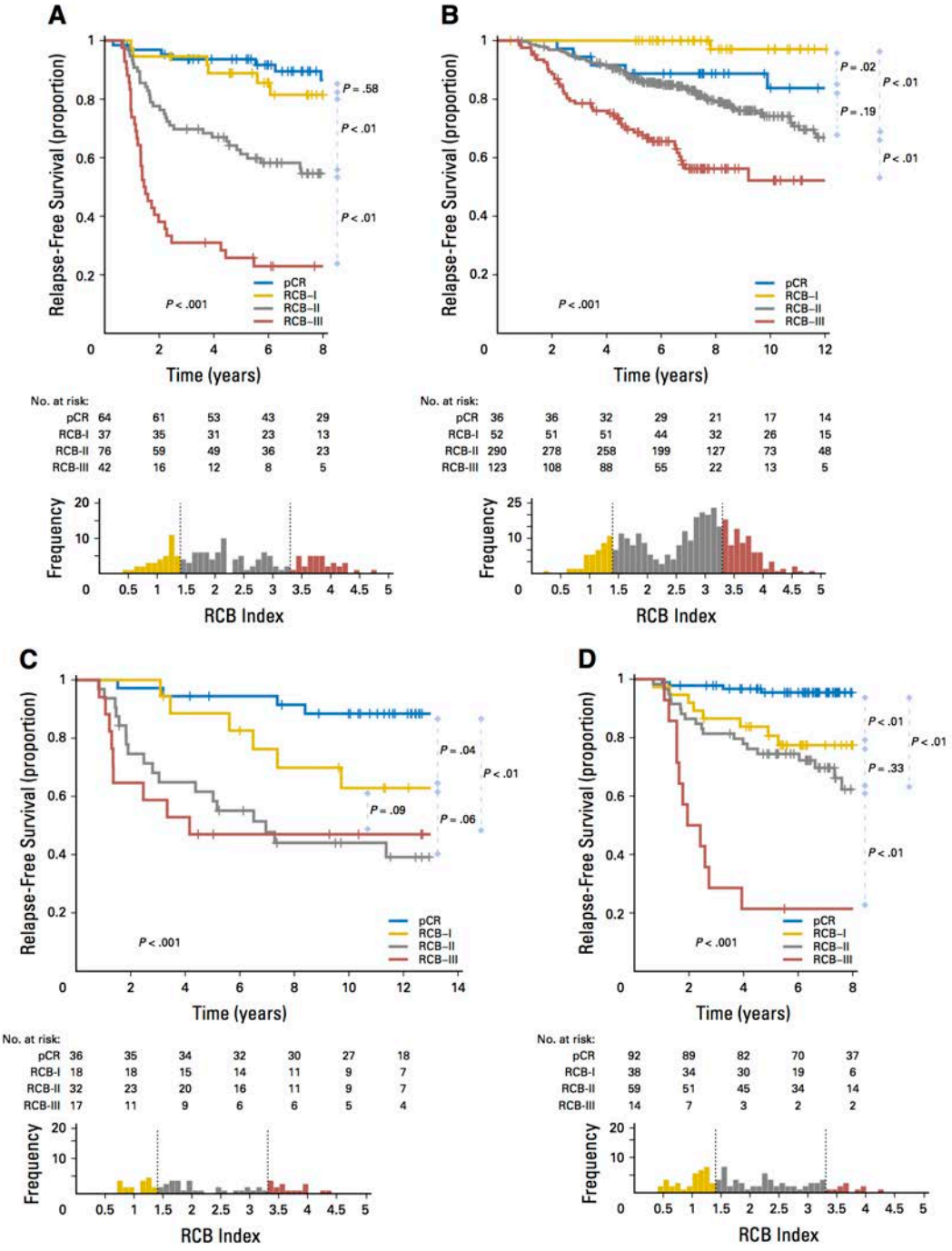


Figure 1.1: Kaplan-Meier plots of relapse-free survival according to RCB categories in (A) triple-negative, (B) $ER^{+}HER2^{-}$, (C) $HER2^{+}$, not treated with trastuzumab, (D) $HER2^{+}$ treated with trastuzumab. Reprinted with permission. © 2017 American Society of Clinical Oncology. All rights reserved. Symmans, W et al. *J Clin Oncol*, 35 (10), 2017: 1049-1060 [280].

1.4 Prediction of response to neoadjuvant therapies

Shortly following the introduction of neoadjuvant therapies, efforts were made to identify molecular predictors of response. Such predictors would allow the stratification of patients into subsets that would derive the most benefit and others that would not. Patients with chemoresistant tumours would therefore be spared neoadjuvant therapy and the adverse reactions associated with it and instead proceed with primary surgery.

1.4.1 Genomic predictors

TP53 mutations correlate with response

Sensitivity assays done on cell lines in the late 1990s showed that treatment with taxanes was more effective than treatment with anthracyclines on breast cancer cells with mutated *TP53* than in those with wild-type *TP53* [216, 304]. Reduced *TP53* was shown to correlate with increased G2/M cell cycle arrest and apoptosis during treatment with taxanes. This hypothesis was tested in the *EORTC 10994/BIG00-01* trial [38], which randomised 1,856 patients between 2001 and 2006 to receive either 6 cycles of FEC or 6 cycles of T→ET. While the study showed that *TP53* mutation status was prognostic for overall survival, it did not identify patients most likely to benefit from taxane-based chemotherapy regimens. Indeed, treatment with taxanes (T→ET) did not result in an increase in 5-year PFS in women with *TP53* mutated tumours, compared to treatment with no taxanes (FEC). In an effort to generate signatures predictive of response in these cohorts, RNA from 125 ER^- tumours (66 tumours in the FEC group and 59 tumours in the T→ET group) was extracted and analysed on Affymetrix X3P microarrays. The analysis resulted in the generation of gene expression signatures predictive of response to chemotherapy however this was later retracted. The dataset, however, remains available for mining (Gene Expression Omnibus accession number GSE6861) and has been used in various analyses described later on in this chapter.

Various other studies were set up to ascertain whether *TP53* status correlated with prognosis (discussed in Chen et al. [52]), with a meta-analysis of 26 published studies encompassing a total of 3,476 patients showing that *TP53* mutations or overexpression were associated with pCR (RR 1.37). These findings were further validated by a recent study by Wang et al. [307] in which *TP53* mutation status in 351 patients who received neoadjuvant chemotherapy was

assessed. 41% of the patients studied harboured a mutation in *TP53*, with a pCR rate of 28.6% in *TP53* mutant tumours and 7.1% in *TP53* wild-type tumours.

PIK3CA mutations correlate with resistance to therapy

In a prospective study, Yuan et al. [319] obtained tumour biopsies from 729 patients who received neoadjuvant chemotherapy and profiled *PIK3CA* mutations by cDNA polymerase chain reaction (PCR) amplification. 28.3% of patients harboured a *PIK3CA* mutation and presence of a mutation within this oncogene was associated with a lower rate of pCR (14.6% vs. 21.4%). Patients with *PIK3CA* hotspot mutations (E542, E545 and H1047) had a lower pCR rate than patients with wild-type *PIK3CA* (13.5% vs. 21.4%), whilst there was no difference in pCR rate between patients with non-hotspot mutations and wild-type *PIK3CA*. Of note, 9 patients with detectable *PIK3CA* mutations prior to commencing chemotherapy were noted to have absence of the corresponding mutation after neoadjuvant chemotherapy.

In order to assess the association between mutations within the *PIK3CA* oncogene and response, Loibl et al. [175] analysed *PIK3CA* mutations in 967 patients with $HER2^+$ tumours recruited to five neoadjuvant studies (*GeparQuattro* [293], *GeparQuinto* [292], *GeparSixto* [303], *NeoALTTO* [22] and *CHERLOB* [107]). The analysis showed that pCR rates were lower in the *PIK3CA* mutant cohort (16.2%), compared with to the wild-type cohort (29.6%), however in a subgroup analysis this was only confined to ER^+ tumours (mutant vs. wild-type pCR rate: 7.6% vs. 24.2%), and was not observed in ER^- tumours. Additionally, *PIK3CA* mutations did not confer any alteration to survival.

1.4.2 Transcriptomic predictors

TOP2A: Predicting response to anthracycline therapy

As anthracyclines formed the backbone of most neoadjuvant regimens, efforts were made to determine biomarkers associated with response to this class of cytotoxics. The prospective multi-centre neoadjuvant *Trial of Principle (TOP)* study was the first amongst such studies and enrolled 149 patients with ER^- tumours between January 2003 and June 2008. All patients were treated with single agent epirubicin monotherapy and the study set out to define gene expression signatures that predicted chemoresistance to anthracycline therapy [67]. Accrual into the study was stopped prematurely because of low recruitment as well as concern

by several investigators that the study did not contain multi-agent chemotherapy: indeed the pCR rate in this study was only 13.7% compared to 45% expected in triple negative breast cancers treated with multiple agents [243].

In the study, *TOP2A* amplification was significantly associated with pCR (OR: 18.75), and a *TOP2A* biomarker signature was defined as the averaged sum of the expression of genes in the vicinity of *TOP2A* (including: *PSMD3, CSF3, MED24, THRA, NR1D1, CASC3, RAPGEFL1, WIPF2, CDC6, RARA, GJD3, TOP2A, IGFBP4, TNS4, CCR7* and *SMARCE1*). The expression of this signature was shown to be associated with pCR in the $ER^{-}HER2^{+}$ subgroup, but not in $ER^{-}HER2^{-}$ tumours. Of note, none of the patients recruited to the study received trastuzumab. In order to further refine the metagene, previously published stromal and immune metagenes [68] were combined with the TOP2A metagene score to derive an *A-score* for $HER2^{+}$ tumours. In view of the observed lack of relationship between *TOP2A* metagene score and pCR, the A-score for $ER^{-}HER2^{-}$ cases did not include *TOP2A* metagene expression, but only contained the immune and stromal metagene scores. The A-Score was characterized by a high negative predictive value (98%) and was significantly associated with pCR status in the *TOP* trial, as well as the anthracycline-based arms (but not in the taxane/anthracycline arms) of the *EORTC 10994/BIG00-01* and *MDACC 2003-0321* study. Data from this study was deposited at the Gene Expression Omnibus with ID GSE16446.

MDA-1: Predicting response to T→FAC

Following the publication of the *NSABP B-27* clinical trial, which showed increased benefit of adding taxanes to anthracycline-based therapies [23, 24], efforts were made to characterise molecular predictors of response to T→FAC chemotherapy. Ayers and colleagues [16] prospectively enrolled 42 patients receiving neoadjuvant T→FAC chemotherapy to develop a multigene predictor to predict response to chemotherapy. Fine needle aspiration of the tumour was undertaken prior to starting chemotherapy and expression quantified using microarrays. The cohort was split into discovery (24 patients) and validation (18 patients) sets, and machine learning algorithms applied to the discovery set in order to identify a multi-gene model that was associated with response. By training a support vector machine and combining with a *k* nearest neighbours (k-NN) class prediction algorithm, a 74-gene model was derived, which had an accuracy of 78% of predicting pCR, a positive predictive value (PPV) of 100%, a negative predictive value (NPV) of 73%, sensitivity of 43% and

specificity of 100%. It is worth noting, however, that only 3 patients had attained pCR in the validation dataset.

In order to further expand this study, a further 91 patients were added to the first cohort of patients by Hess et al. [122], and a similar analysis was done: pre-treatment gene expression profiling was done using oligonucleotide microarrays on tumour samples obtained by fine-needle aspiration and predictors of pCR were derived from 82 cases and assessed on a validation set of 51 cases. 20 classifier algorithms were assessed: a Diagonal Linear Discriminant Analysis (DLDA) classifier trained on a 30-probe set gave the highest area under curve, and was validated in the dataset (accuracy: 0.76, sensitivity: 0.92, specificity: 0.71, PPV: 0.52, NPV: 0.96). Data from this study was deposited at the Gene Expression Omnibus with ID GSE20194.

Subsequently, between October 2003 and October 2006, the same group further assessed the utility of this DLDA30 classification prospectively in a clinical study that randomised patients to receive either T→FAC chemotherapy or FACx6 [282]. Two hundred and seventy-three patients were enrolled and the pCR rates were significantly higher in the T→FAC arm compared with the FAC arm (19% vs. 9%, $p<0.05$). Interestingly, the DLDA30 predictor appeared to exhibit regimen specificity, as it performed reasonably well (though less impressively compared to the original study) in predicting pCR in the T→FAC arm (PPV 38%, NPV 88%, sensitivity 63%, specificity 72%, area under curve (AUC) 0.71), and less accurately in the FAC-only arm (PPV 9%, NPV 92%, sensitivity 29%, specificity 75%, AUC 0.58), cementing the fact that metagene derivation algorithms are very specific to the chemotherapy regimens they are trained on.

I-SPY 1: Predicting response to AC±T

The *I-SPY 1* trial [76, 77] recruited 237 patients between May 2002 and March 2006, with 221 patients receiving neoadjuvant anthracycline chemotherapy and 210 subsequently receiving a taxane pre-surgery. 149 of these patients were profiled with microarrays, 171 had *TP53* mutation chip data and 153 had copy number alteration data obtained by molecular inversion probe arrays. The trial showed that pCR rates were lowest for luminal A (3%) and highest for HER2-enriched tumours (50%). *TP53* mutations and amplification at 17q also were associated with increased pCR (47% and 45% respectively). Various previously derived gene signatures that were related to aggressive disease were assessed in the dataset, and rates of pCR were higher for poor prognosis signatures, including a 70-gene high-risk signature

for relapse derived in 98 primary breast tumours (24% vs. 0%) [297], an activated wound healing signature (26% vs. 7%) [50] and a *TP53* mutation predicted by expression profiling (34% vs. 9%) [288]. Hence, more aggressive and proliferative tumours were more likely to attain pCR. Data from this study was deposited at the Gene Expression Omnibus with ID GSE25066.

DFCI study: Predicting response to Cisplatin

So as to identify biomarkers associated with response to cisplatin, investigators from the Dana Faber Cancer Institute (DFCI) recruited 28 women with $ER^{-}HER2^{-}$ tumours into a study that administered four cycles of preoperative cisplatin [264]. 22% of patients attained pCR, and it was noted that low *BRCA1* expression was associated with pCR. *BRCA1* promoter methylation was also statistically significantly correlated with response to platinum and inversely correlated to *BRCA1* mRNA expression. In addition, a significant association was observed between *TP53* truncating mutations as well as E2F3 oncogenic pathway activation with response. Data from this microarray experiment was deposited at the Gene Expression Omnibus with ID GSE18864.

Predicting response to Ixabepilone

The epothilone agent ixabepilone was initially licensed for use in taxane-resistant breast cancer [285], and several biomarkers were initially proposed to predict response, including βIII-tubulin protein and mRNA expression, *TACC3* and *CAPG* gene expression as well as the expression of two metagenes containing 20 and 26 genes respectively [128]. In view of this Horak et al. [128] designed a clinical study to assess whether predictive markers for ixabepilone could be discovered in the neoadjuvant setting. 295 patients were randomised to receive either AC→paclitaxel or AC→ixabepilone and the expression levels of *TACC3* and *CAPG*, as well as the two multi-gene models were measured using an Affymetrix gene expression profiling approach (deposited at the Gene Expression Omnibus with ID GSE41998). The pCR rate was similar in both arms of the study (ixabepilone: 24.3% vs. paclitaxel: 25.2%) and none of the predicted models were predictive of pCR between both treatment arms. No further metagenes of response to treatment were derived.

Predicting response to anti-HER2 therapies

The *NeoALTTO* neoadjuvant trial [22] randomised 455 women with HER2$^+$ cancer to receiving paclitaxel with trastuzumab or lapatinib, or both. 254 women had available RNA-seq data, and the expression of *ERBB2* was noted to be the most significant predictor of pCR (OR 3.1), followed by *ESR1* expression (OR 0.53) [90]. The Genomic Grade Index proliferation metagene was also shown to be correlated with response (OR 1.5). An immune metagene was statistically significant in a univariate model (OR 1.3), but lost significance in a model adjusted for clinicopathological parameters and treatment arm. Stromal signatures, AKT/mTOR signatures as well as an *AURKA* signature lost significance on multiple correction [90]. Hence proliferation metagenes continued to dominate the predictive landscape, with evidence of possible contribution from immune signatures.

Predicting response using the Oncotype DX® and MammaPrint® assays

FFPE samples from 95 women with breast cancer treated with doxorubicin and paclitaxel were obtained by Gianni et al. [102] and 384 candidate genes linked to various biological processes (including proliferation, invasion, apoptosis, metastasis, immune pathways, metabolism, drug resistance and DNA repair) were profiled. This was also coupled with measurement of the expression of 21 genes used in the Breast Cancer Oncotype DX® assay. 86 of the 384 genes were found to correlate with pCR, and included genes regulating proliferation (such as *CDC20, E2F1, MYBL2, TOPO2A*), immune response (including *MCP1, CD68, CTSB, CD18, ILT-2* and *HLA-DPB1*), and ER status (*ESR1, SCUBE2* and *GATA3*). Additionally, the Oncotype DX® Recurrence Score was positively associated with the likelihood of pCR (p=0.005), suggesting that the patients at greatest risk of recurrence were more likely to benefit from chemotherapy.

The predictive power of the 70-gene MammaPrint® assay was subsequently studied in a cohort of 167 patients recruited into two trials at the Netherlands Cancer Institute between 2000 and 2008 [273]. 14% of patients had a good prognosis signature and none of them attained pCR. 20% of patients in the poor prognosis signature group attained pCR, and all triple-negative tumours had a poor prognosis signature. As shown by the Oncotype DX® assay, tumours that were more aggressive were more likely to benefit from treatment with chemotherapy. These results were independently confirmed in a second combined analysis in 2013 [103].

Harmonising microarray analysis - the MAQC-II study

The MicroArray Quality control (MAQC)-II study was published in 2010 and aimed to robustly assess methods of generation of predictive models in classifying lung or liver toxicity in rodents, as well as various clinical endpoints in breast cancer, multiple myeloma and neuroblastomas in humans [260].

Rather than aiming at accruing new knowledge, the MAQC-II study was a survey of the then current practices, however it did provide the community with larger datasets that could be further analysed. The breast cancer cohort, for example, contained the dataset generated by Hess et al. [122] as a discovery set and a further 100 preoperative samples that were treated with T→FEC chemotherapy, with all the microarray data generated deposited in the Gene Expression Omnibus with ID GSE20194. Indeed this dataset has been mined by multiple groups in order to further improve machine learning methods and determine novel predictors of response to treatment [174].

Microarray dataset integration and mining

Following the publication of these individual datasets, efforts were made to integrate the data generated and harness the power of greater numbers to derive response metagenes that were closer approximations to the truth.

Hatzis et al. [117] merged data from 508 patients included within the *I-SPY 1*, the *US Oncology Protocol 02103* study [259], the *MDACC 2003-0321* trial [230] and the *LAB-03-432* study, and used the collated normalised data to generate a genomic predictor of response and survival following taxane and anthracycline-based chemotherapies. 310 patients were used in a training set, with data from the remaining 198 patients used for validation. 39 microarray probe sets were associated with chemosensitivity in ER^+ tumours and 55 probe sets in ER^- tumours: the derived metagene had a PPV of 56% and an NPV of 73% for prediction of pCR or RCB-I RD.

Callari et al. [44] integrated expression data from 4 different gene expression sets with clinical data [67, 108, 117, 128] to generate metagenes predictive of response. The work identified a *T-cell metagene* (including *CXCL13, PRF1, IRF1, IKZF1, GZMB* and *HLA-E*), which was predictive of pCR in ER^-HER2^- tumours only. Indeed tumours with high, intermediate and low expression of this metagene were associated with pCR rates of 33.7%,

35.2% and 11.6% respectively. Interestingly, the intermediate expression group had a slightly higher pCR rate than that seen in the high expression group. In ER^+HER2^- tumours, a proliferation metagene (comprising: *NCAPG, BUB1B, PRC1, CCNB2, RAD51AP1, ORC6, FANCI, UBE2C, AURKA, KIF20A*) and an ER-related metagene (comprising: *ABAT, CA12, MCCC2, SCUBE2, LRIG1, FAM63A, CCDC176, MYB, CACNA1D, GATA3*) were associated with response, with the high-risk group (proliferation high, ER low) having a pCR rate of 18.9% and the low-risk group (proliferation low, ER high) having a pCR rate of 4.4%.

Iwamoto et al. [142] integrated expression data from the MAQC-II [260] and *USO-02103* (GSE23988) studies and derived gene sets associated with response. ER^+ tumours had significant enrichment of cell cycle and proliferation Gene Ontology terms, as well as various immune GO terms, including signalling from chemokine receptors 3 and 5 and interleukin-8 in tumours that were chemosensitive. A lack of association between immune system and proliferation GO terms was noted in the ER^- cohort, with enrichment of gene sets were involved in glycolipid, sphingolipid, and fucose metabolism pathways moderately associated with pCR.

Ignatiadis et al. [139] evaluated expression data from 8 major studies (comprising 996 patients) (*TOP* [67], *EORTC 10994* [38], *MDACC* [282], *MAQC-II* [260], *I-SPY 1*, *LBJ/INEN/GEICAM*, *MAQC-III* [117] and *USO-02103* [142]) so as to determine whether response to anthracycline (with or without a taxane) was based on activation of different pathways. As opposed to observations made by Callari et al. [44] and Iwamoto et al. [142], high immune module scores were associated with increased rates of pCR in all ER and HER2 subtypes, with chromosomal instability and PTEN loss associated with increased pCR probability in $HER2^-$ but not $HER2^+$ tumours. Increased activation of the insulin-like growth factor 1 module was associated with increased pCR probability only in $HER2^-$ tumours.

Hence, all these integrative studies showed that proliferation and immune activation both played a role, to a degree, in determining response to neoadjuvant therapies. However, the analyses were confounded by different chemotherapy agents used, different exposures to chemotherapy and different data generation technologies, explaining the contradictory findings observed between studies.

Unlike previous approaches, which mined expression data for biomarker discovery and validation, Juul et al. [147] derived a metagene for response to paclitaxel in ER^- tumours by using a functional genomics approach informed by previous RNA interference screening experiments on cell lines exposed to paclitaxel. These experiments had revealed two

distinct gene sets that modulated resistance to taxanes: one involved in mitosis and mitotic spindle assembly checkpoint and another involved in the metabolism of the pro-apoptotic lipid ceramide [277]. By building on the observations of this work through using gene expression data from five neoadjuvant clinical studies, a paclitaxel sensitivity metagene was developed, composed of four genes (*BUB1B, CDK1, AURKB* and *TTK*) in a mitotic module and two genes in a ceramide module (*UGCG, COL4A3BP*) [147]. Overexpression of the mitotic module conferred sensitivity to paclitaxel, whilst overexpression of the ceramide module related to increased metabolism of ceramide to sphingomyelin (via COL4A3BP) and glucosylceramide (via UGCG) and increased resistance to taxane therapy, as shown by the above-mentioned work by Iwamoto et al. [142]. The paclitaxel response metagene was defined as the difference between the mean of the expression of the mitotic module and the ceramide module, with higher scores associated with increased response to paclitaxel treatment. To interrogate this metagene, microarray data was integrated from five studies (*MDA1* [122], *MDA/MAQC-II* [260], *TOP* [67], *EORTC 10994* [38] and a *DFCI* cohort [264]. In the T→FAC treated triple-negative cohort the response metagene was highly predictive of pCR (p=0.0039, OR 19.92), with an AUC ranging between 0.72 and 0.79.

Identification of resistance mechanisms

The neoadjuvant setting provides an unparalleled in vivo physiological system to study mechanisms of resistance to therapies. ATP-binding cassette (ABC) transporters have been long been attributed with chemoresistance [281], with *ABCB1*, *ABCC1* and *ABCG2* associated with resistance to taxanes, anthracyclines and topoisomerase inhibitors [220]. In an effort to molecularly characterise ABC mediated resistance to neoadjuvant chemotherapy, Park et al. [220] obtained tumour tissue via a core needle biopsy from 21 patients undergoing primary chemotherapy with FEC→T and profiled the expression of ABC mRNA. Tumours that attained pCR had high expression of *ABCB3, ABCC7* and *ABCF2*, whilst those with RD had a significantly higher expression of *ABCC5, ABCA12, ABCA1, ABCC13* and *ABCC11*.

In a study by Balko et al. [18], 49 triple negative tumours following neoadjuvant chemotherapy were profiled using the NanoString™ platform. *DUSP4*, a negative regulator of extracellular-regulated kinase (ERK), was shown to be a potential mediator of chemoresistance, as well as a probable tumour suppressor in these tumours. High expression of *DUSP4* was noted to increase chemotherapy-induced apoptosis, whilst low expression was associated with activation of the Ras-ERK pathway with post-therapy high Ki-67 scores and chemoresistance.

1.5 Characterising tumour changes during chemotherapy

In sharp contradistinction to the large volume of literature discussing predictive biomarkers and response to neoadjuvant chemotherapy, fewer studies have concentrated on serial analyses of tumour tissue throughout chemotherapy.

Analysis of pre and post-treatment samples

Some work has been done to describe how post-treatment tumours differ from treatment naïve tumours. Gonzalez-Angulo et al. [104] examined gene expression differences between pre- and post-neoadjuvant therapy samples in 21 patients after 4 to 6 months of chemotherapy, with solely cases that had evidence of RD analysed. Different pathways were preferentially perturbed in ER^- versus ER^+ breast cancers, with the former having increased expression of PI3K, G protein and energy metabolism pathways, and the latter having increased notch signalling and energy metabolism (including fatty acid synthesis). Immune cell signatures were decreased in both types of tumours in the residual tumour samples.

Hannemann et al. [113] compared the gene expression profile before and after chemotherapy in 48 patients and the main conclusion of the study was that chemosensitive tumours showed significant changes in the expression landscape, while chemoresistant tumours had a more stable expression landscape during treatment.

Vera-Ramirez et al. [299] performed expression analysis on 56 matched pre- and post-chemotherapy samples from 28 patients treated with anthracycline and taxane-based chemotherapy and stratified response using the Miller and Payne grading system [217]. The expression of 65 genes was significantly up-regulated after treatment with chemotherapy, with a Gene Ontology analysis revealing enrichment over extracellular matrix pathways, cell proliferation and adhesion, oxidative stress and angiogenesis. A comparison of pre and post-therapy samples showed that tumours that attained a good response (defined as grades 4 and 5 and pCR) showed changes in differential gene expression during therapy, whilst those with a poor response (grades 1 and 2) did not have any differentially expressed genes, indicating that chemoresistant tumours had a highly stable transcriptome during chemotherapy. Microarray data from this study was deposited at the Gene Expression Omnibus under accession number GSE28844.

Analysis of pre and early-treatment samples

Modlich et al. [199] analysed the expression landscape in tumour samples obtained from 25 patients taken prior to the commencement of chemotherapy and after 24 hours of treatment. The analysis showed that two genes (*CDKN1A* and *GDF15*) were up-regulated in post-treatment samples, with few other inferences made. In a similar study, Buchholz et al. [42] analysed global gene expression changes in serial tumour core biopsy specimens obtained before treatment and within 24-28 hours of the first cycle of chemotherapy in five patients: no genes were found to be differentially expressed before and during therapy, though global alterations in transcriptomic expression were noted.

Sotiriou et al. [270] analysed the correlation between tumour expression profiles of 9 patients, with samples taken via fine needle aspiration prior to commencing chemotherapy, and a second sample taken on day 21 after the first cycle of chemotherapy. Tumours exhibiting better responses had more transcriptomic changes than those with a poorer response, with 16 genes identified as being differentially expressed in good versus poor responders.

Analysis of pre, midway and post-treatment samples

As part of an effort to molecularly characterise tumours during therapy, serial tumour biopsies were obtained from a selection of patients recruited to the *I-SPY 1* trial [76] before treatment, 24–96 hours after the first dose of chemotherapy, and at the time of surgery [182]. 36 cases had matched pairs of the first two tumour biopsies, with 39 cases having matched pairs of the first and last biopsy. Expression levels were determined at all time points: an analysis comparing diagnostic and early biopsies showed profound down-regulation of proliferation and immune-related genes during anthracycline chemotherapy, with decreased expression of cell cycle inhibitors associated with poor response. Following completion of chemotherapy, increased interferon signalling (through *IFIT2, IFIT1, IFITM1, IFIH1* and *EML2*), as well as increased expression of cell proliferation genes in any remaining RD was associated with reduced recurrence-free survival [182]. Data from this study was deposited in the Gene Expression Omnibus database with ID GSE32603.

A recent study [155] described the evolution of triple negative breast cancer in patients undergoing neoadjuvant chemotherapy by using single-cell DNA and RNA sequencing as well as bulk exome sequencing. 20 patients were recruited and tumour sampling performed at three time points: pre-treatment, after two cycles of therapy (mid-treatment), and after

six cycles of chemotherapy. Whole exome sequencing showed that *TP53* mutations were identified in 60% of cases. Two types of response were described, as defined by Kim et al. [155]: clonal extinction, wherein previously observed tumour clones were completely eliminated by chemotherapy and clonal persistence, where clonal frequencies shifted but all tumour populations identified in the diagnostic sample remained present in the post-treatment samples. Using a combination of deep sequencing (mean depth 1,671,000×) and single-cell sequencing of 8 cases, the authors showed that the new tumour clones observed at the end of treatment were pre-existing and selected by chemotherapy, consistent with a punctuated model of evolution [94]. In the tumours that showed evidence of clonal extinction, chemotherapy-induced pCR, with normal diploid cell types including fibroblasts and immune cells dominating the stroma.

Further evidence of such clonal dynamics using bulk exome sequencing data was also shown by Miller et al. [197] in a cohort of patients with breast cancer receiving neoadjuvant aromatase inhibitors.

41 patients with $ER^{+}HER2^{-}$ tumours recruited to the *NeoPalAna* phase II trial [181], were included in a translational study wherein, when possible, tumour biopsies were taken at four time points during neoadjuvant treatment with palbociclib and anastrozole and at the time of surgery. Limited targeted sequencing, as well as microarray gene expression analysis, were performed on the tissues obtained. The expression of *CCND3, CCNE1*, and *CDKN2D* was noted to be persistently high in resistant cases, while the PAM50 proliferation score significantly reduced at each time point. Luminal A and B subtypes were sensitive to treatment, while non-luminal subtypes were associated with resistance and persistent E2F-target gene expression. Mutations within *PIK3CA, CDH1, PTEN, TP53, TBX3*, and *MAP3K1* predominated the somatic landscape, and there was evidence of *RB1* mutation loss in some cases, with the authors suggesting that this could be due to intra-tumoural heterogeneity, clonal evolution or sequencing artefact.

Hence all these analyses have shown that the tumour genomic and transcriptomic landscape is indeed dynamic during chemotherapy, with greater shifts seen in sensitive tumours compared to chemoresistant subtypes.

1.6 Scope of this thesis

The large number of trials performed in the neoadjuvant setting have clearly shown the benefits of pre-operative chemotherapy, ranging from enabling the downstaging of advanced disease and increasing rates of breast conservation surgery to allowing the prediction of relapse-free and overall-survival by using response to chemotherapy as a surrogate.

Despite the importance of this setting, few studies that have generated a thoroughly mined dataset comprising serial DNA and RNA sequencing data of a large cohort of patients receiving neoadjuvant chemotherapy. Most studies are either small case-series, do not have serial tumour sampling, or have limited amount of genomic or transcriptomic data. Meta-analyses have combined series with differing chemotherapy regimens, differing chemotherapy exposures, tumour samples obtained by different methodologies and sequencing or expression data acquired in non-standardised ways. Additionally, most patients with $HER2^{+}$ tumours were not exposed to trastuzumab.

The work described in this thesis strives to generate a robust molecular dataset comprising serial tumour sampling from a prospective neoadjuvant clinical study, with tumour biopsies obtained prior to starting therapy, midway through chemotherapy, and at the time of surgery. By doing so, the main aims of this work are to:

1. Describe genomic and transcriptomic features predictive of response or resistance to neoadjuvant chemotherapy in breast cancer patients
2. Describe changes in the tumour architecture and tumour microenvironment during chemotherapy and assess dynamics associated with response
3. Determine whether circulating tumour DNA can be detected both before and during neoadjuvant chemotherapy

The key strengths of this work include:

1. The use of standardised chemotherapy regimens as recommended by national guidelines, with very few local protocol deviations
2. The availability of a high quality detailed clinical database capturing patient and tumour phenotypes, chemotherapy schedules and imaging data
3. The use of standardised operating procedures to guide tissue sampling and processing in order to mitigate batch effects

4. The involvement of experienced breast radiologists and experts in breast cancer neoadjuvant pathology to ensure accurate datasets
5. The integration with a Diffusion Contrast Enhanced Magnetic Resonance Imaging (MRI) study, allowing the integration of genomic and transcriptomic data with imaging data
6. The integration with digital pathology analysis, allowing orthogonal validation of various findings discovered through the mining of the RNA-seq data

The neoadjuvant dataset generated is described in Chapter 2, which details the trial protocol, the patient population recruited, as well as the experimental and bioinformatic methods used in this thesis. Following this, molecular predictors of response and resistance to chemotherapy are described in Chapter 3, with an in-depth analysis of DNA and RNA sequencing data obtained prior to commencing chemotherapy. Chapter 4 investigates the changes that occur within the tumour clonal architecture during chemotherapy and illustrates the different manners in which chemotherapy alters tumour subclonal populations. The changes in the transcriptomic landscape are subsequently described in Chapter 5, where the tumour microenvironment is shown to be dynamic and correspond to response to chemotherapy. Finally, Chapter 6 explores the detection of circulating tumour DNA, as well as its dynamics during therapy.

Chapter 2

Generation of a comprehensive neoadjuvant dataset

Contents

2.1 Introduction

2.1.1 Establishing a neoadjuvant molecular study

The *TransNEO* translational study (Research Ethics Council (REC) registration number: 12/EE/0484) was set up at Cambridge University Hospitals, Cambridge in 2013 by ***Dr Sarah-Jane Dawson*** and ***Prof Carlos Caldas*** to prospectively profile tumour tissues and circulating nucleic acids in patients receiving neoadjuvant chemotherapy for early breast cancer.

The study objectives of TransNEO included:

1. The characterisation of expression and genome-based molecular profiles of breast cancer, with subsequent association of these profiles with clinical outcomes
2. The analysis of tumour specific circulating nucleic acids levels during neoadjuvant therapy and association of levels with clinical outcomes

Within the study protocol, tumour tissue and blood were collected serially at defined time points during neoadjuvant therapy (Figure 2.1), these being:

1. Tumour tissue collection:
   (a) At diagnosis
   (b) Midway through treatment
   (c) On completion of chemotherapy
2. Plasma sample collection:
   (a) At diagnosis
   (b) After one cycle of chemotherapy
   (c) Midway through treatment
   (d) On completion of chemotherapy
3. Peripheral blood mononucleated cells collection:
   (a) At diagnosis
   (b) At any other time point, when possible

Tumour biopsies at diagnosis were mandatory and were obtained at the time of fiduciary marker insertion. Midway biopsies were obtained if patient consent was given and there was evidence of remaining residual disease on ultrasound. Samples from the surgical resection specimen were taken if the residual tumour mass was still evident. Biopsies were always

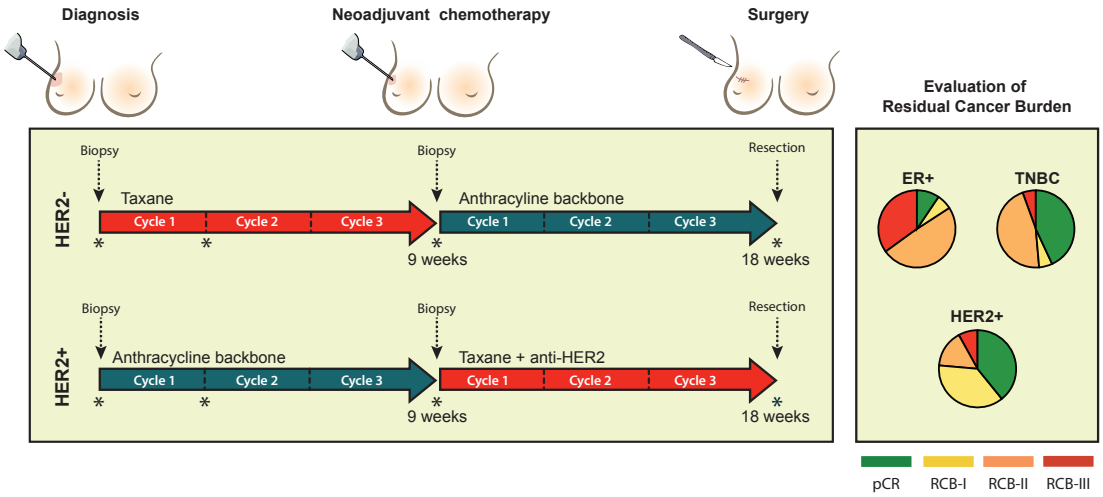


Figure 2.1: Schematic representation of the *TransNEO* trial schema detailing tumour biopsy time points, as well as plasma sampling time points (marked with *). Following completion of chemotherapy, assessment of response was performed by an experienced histopathologist in breast cancer neoadjuvant chemotherapy.

obtained at the site of the fiduciary marker in order to mitigate, as much as possible, intra-tumoural heterogeneity.

Key criteria for inclusion within the study included:

1. Histological diagnosis of invasive breast cancer
2. Planned administration of neoadjuvant therapy
3. Ability to give written informed consent
4. Eastern Cooperative Oncology Group Performance status 0 to 2
5. Absence of metastatic disease at diagnosis

The study did not impose restrictions on cytotoxic agent regimens, and therefore the decision as to which chemotherapy regimen to administer was left to the attending oncologist in order to mirror standard of care practice. All chemotherapy regimens within the Oncology Department at Cambridge University Hospitals are protocolised with regular auditing, and it is of no surprise that most patients with good performance status received one of two standard therapies (FEC$\rightarrow$T+anti-HER2 if HER2$^+$ or T$\rightarrow$FEC if HER2$^-$). Further details of all cytotoxic agents administered will be discussed further on in this chapter, and are detailed in Appendix 1, Table A.2.

Serial ultrasounds and Magnetic Resonance Imaging (MRI) scans were used to monitor disease response: these were performed by experienced breast radiologists in the Breast

Cancer Unit at Cambridge University Hospitals. In addition, as part of an effort to characterise the biology of tumours throughout neoadjuvant chemotherapy, patients recruited to the *TransNEO* trial were also encouraged to participate in companion translational studies. One such study was the *TRICKS* radiology study (REC: 13/LO/0411) led by ***Prof Fiona Gilbert*** in the Department of Radiology at Cambridge University Hospitals. Within this study, patients were asked to undergo *Dynamic Contrast Enhanced MRI* (DCE MRI) and tumour biopsies were obtained concurrently to facilitate the integration of genomic data with radiological features. 81 patients within the *TransNEO* study were also included within the *TRICKS* study.

Response to chemotherapy was assessed on the surgical resection tissue by an experienced breast pathologist and graded using the Residual Cancer Burden (RCB) scoring system described in Chapter 1 [279].

2.1.2 Patient demography

By July 2017, 180 patients were recruited to the study and had completed neoadjuvant chemotherapy, with an estimated recruitment rate of circa 45 patients per year. Twelve patients were withdrawn from the study, as detailed in Figure 2.2: a pre-therapy tumour biopsy was not obtained in six patients, three patients were diagnosed with metastatic disease shortly after recruitment, two patients were recruited to phase II trials and received experimental therapies and one patient developed severe complications during treatment and died early on during treatment.

Hence, the neoadjuvant cohort studied in this thesis comprised of 168 patients. RCB assessment was not available for seven cases: two had incomplete surgeries so full staging was not possible, two were not operated (one refused surgery, the other was too unwell to undergo treatment), one patient transferred her care to another centre, one patient developed metastatic disease late during therapy, and one died at the end of neoadjuvant therapy.

Clinical features of recruited cohort

A summary of clinical features at diagnosis is presented in Table 2.1, with a detailed table showing all features for each case in Appendix 1, Table A.1. The median age of the 168 patients recruited to the study was 51 years (range: 20-81 years), which was significantly

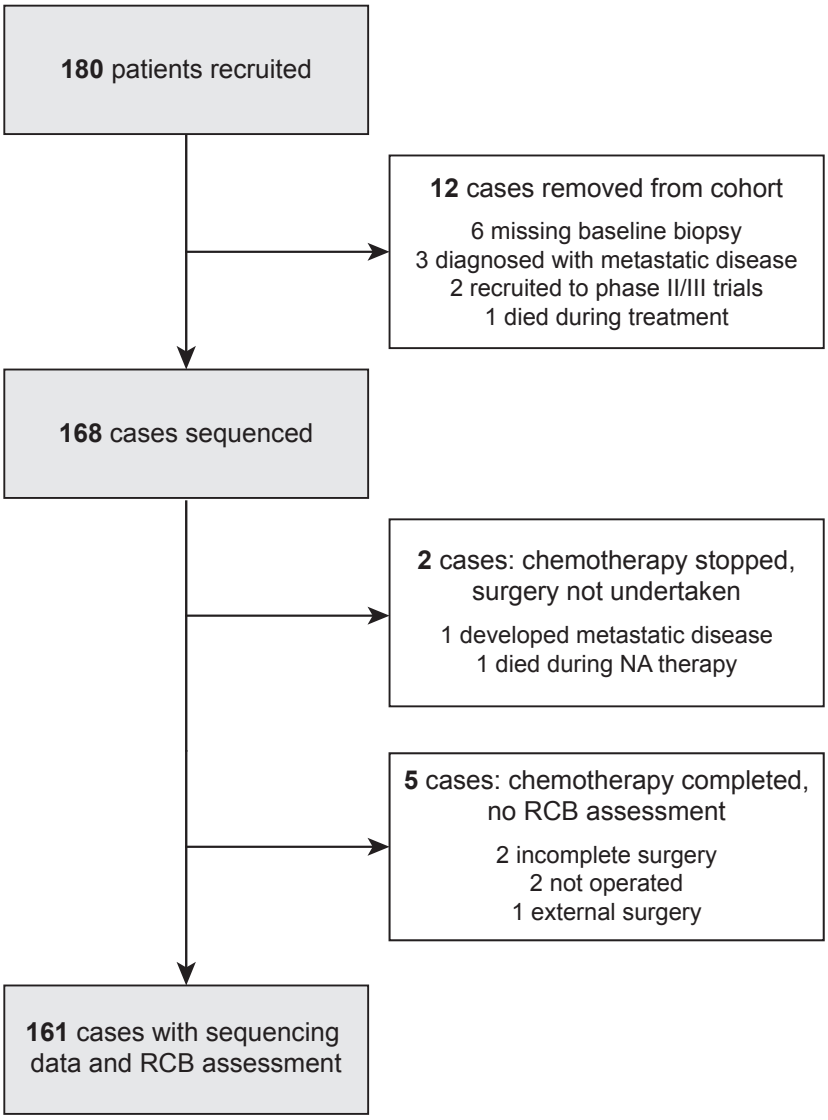


Figure 2.2: 180 patients completed treatment within the study by July 2017. 168 cases were sequenced and 161 cases had RCB assessment available

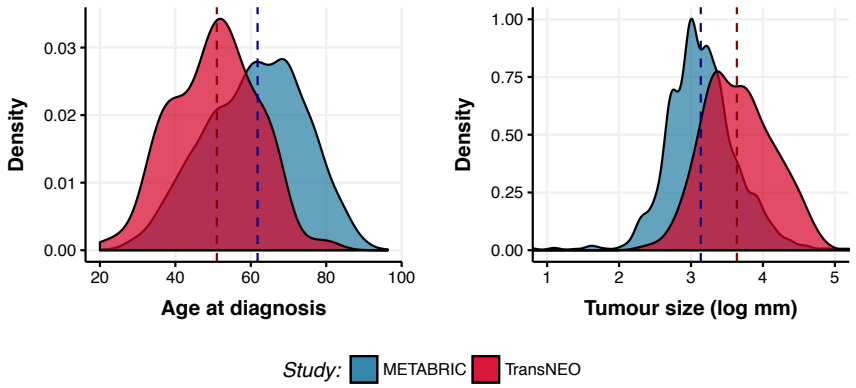


Figure 2.3: Difference in age and tumour size distribution between the METABRIC (no neoadjuvant therapies delivered) and the TransNEO neoadjuvant study

younger than the median age at which breast cancer is diagnosed in the general population, where more than 80% of breast cancer diagnoses occur in women over the age of 50 (Cancer Research UK, 2018). A comparison with the METABRIC cohort [60, 225], which recruited patients who were not treated with neoadjuvant chemotherapy, also highlighted this difference: the median age at presentation in METABRIC was 62 years (Figure 2.3). The difference in demography is perhaps unsurprising. Younger women often presented with more aggressive tumours than older women and benefited the most from down-staging surgery, with breast-conserving surgery often preferable to mastectomies.

The patients recruited to the TransNEO study presented with larger tumours (median size: 3.8cm, range 1.1-12.0cm) compared to those treated with adjuvant intent (METABRIC median tumour size 2.3cm). 61% of tumours were poorly differentiated (Grade 3) with a noticeable absence of Grade 1 tumours. Conversely, in the METABRIC study 48% of patients were diagnosed with Grade 3 tumours and 9% diagnosed with Grade 1 tumours. Hence the neoadjuvant cohort enriched for larger, higher grade tumours diagnosed in younger women.

The prevailing histology was invasive ductal (85.7%), with 4.3% of patients having an invasive lobular carcinoma and 3.6% having mixed histologies. 69% of patients had ER^+ tumours and 29% had evidence of HER2 amplification on IHC or FISH (Table 2.1). 51% of patients had histologically confirmed axillary lymph node involvement at diagnosis.

Classification of tumours into the intrinsic subtypes [221] was possible for 97% of tumours. 45.8% of all tumours were classified as Luminal (12.5% Luminal A, 33.3% Luminal B), 23.2% were Basal, 26.6% HER2 enriched and 5.4% falling within the Normal-like category.

Table 2.1: Distribution of clinical features across 168 patients at diagnosis

	% of Cases (n)
Age	
20 - 40	22.0% (37)
41 - 50	27.4% (46)
51 - 60	29.8% (50)
60 - 80	20.2% (34)
Tumour size	
T1 ($\leq$2cm)	4.8% (8)
T2 (>2 - 5cm)	63.1% (106)
T3 (>5cm)	26.2% (44)
T4d	5.3% (9)
Unevaluable	0.6% (1)
Tumour histology	
Invasive ductal	85.7% (144)
Invasive lobular	4.2% (7)
Mixed	3.6% (6)
Micropapillary	3.0% (5)
Apocrine	1.8% (3)
Medullary	1.8% (3)
ER/HER2 status	
ER^- $HER2^-$	22.6% (38)
ER^- $HER2^+$	8.3% (14)
ER^+ $HER2^-$	38.7% (65)
ER^+ $HER2^+$	30.4% (51)
Tumour grade	
2, moderately differentiated	38.7% (65)
3, poorly differentiated	61.3% (103)
Lymph node involvement	
N0	48.2% (81)
N1+	51.8% (87)
PAM50 subtype	
Luminal A	12.5% (21)
Luminal B	33.3% (56)
Basal	23.2% (39)
HER2 enriched	22.6% (38)
Normal-like	5.4% (9)
Not evaluated	3.0% (5)

Chemotherapy regimens administered

The chemotherapy regimens administered were in line with current NICE guidance [210], with anthracycline and/or taxane administration (Table 2.2, Appendix 1: Table A.2). 81% of all HER2$^-$ patients received a taxane (docetaxel or paclitaxel) for three cycles, followed by three cycles of FEC (Fluorouracil, Epirubicin, Cyclophosphamide). 79% of all HER2$^+$ patients received upfront FEC for 3 cycles, followed by three cycles of docetaxel and trastuzumab pre-operatively. 5 HER2$^-$ patients also received pertuzumab following its introduction in 2016. Choice of chemotherapy was not restricted within the study and despite this there were very few variations within local protocols (Table 2.2).

Of the 168 cases that were sequenced:

- 2 cases did not complete chemotherapy and did not have surgery
  - **T088**: died after 4 cycles of Docetaxel, Cyclophosphamide (TC) chemotherapy and 2 cycles of trastuzumab
  - **T106**: diagnosed with bone-only metastatic disease after 2 cycles of docetaxel
- 5 cases completed chemotherapy but did not have a formal surgical RCB assessment
  - **T027**: had external surgery and RCB assessment was not performed
  - **T018**: was not fit for surgery
  - **T157**: declined surgery
  - **T158** and **T165** had incomplete surgery due to intra-operative complications
- 3 cases received fewer than three cycles of neoadjuvant chemotherapy but had surgery and an RCB assessment
  - **T042**: developed severe peritonitis and bowel perforation after 2 cycles of docetaxel which necessitated a prolonged stay in intensive care. (RCB-I)
  - **T102**: diagnosed with metastatic visceral disease during treatment, chemotherapy stopped after 1 cycle and switched to letrozole. The primary tumour was subsequently removed (RCB-III)
  - **T167**: was unable to tolerate docetaxel, and the decision was taken to proceed straight to surgery after 1 cycle of treatment (RCB-III)
- 4 HER2$^+$ cases received one, rather than three, preoperative cycles of trastuzumab (T069, T077, T122, T144)

Patients with suboptimal chemotherapy exposure or who did not have a comprehensive pathological assessment of their residual disease following surgery were not included within any analyses that utilised RCB or dichotomous response (pCR/RD) as a response variable.

Table 2.2: Neoadjuvant regimens administered

	% of Cases (n)
HER2$^-$ chemotherapy regimens	
Taxane → Anthracycline	48.8% (82)
Taxane + Cyclophosphamide	5.4% (9)
Anthracycline → Taxane	4.2% (7)
Taxane + Platinum	1.8% (3)
HER2$^+$ chemotherapy regimens	
Anthracycline → Taxane + anti-HER2	31.5% (53)
Taxane + anti-HER2	5.4% (9)
Taxane + anti-HER2 → Anthracycline	1.8% (3)
Taxane + Platinum + anti-HER2	0.6% (1)
Anthracycline + anti-HER2	0.6% (1)
Number of cycles administered	
7	8.3% (14)
6	79.2% (133)
5	1.8% (3)
4	8.3% (14)
3	0.6% (1)
2 or less	1.8% (3)
Regimen delays/changes	
Dose delay only	6.0% (10)
Dose reduction/omission only	22.6% (38)
Dose delay and reduction	8.3% (14)

Assessment of post-therapy tumour pathology

Following completion of chemotherapy, pathological assessment of the surgical tissue was performed by ***Dr Elena Provenzano*** (Tables 2.3, 2.4). This included a measure of tumour diameter, tumour cellularity, percentage of *in situ* component, number of lymph nodes infiltrated with metastatic disease, as well the diameter of the largest metastasis, which were all used to compute the RCB score.

26.1% (42/161) of all cases that had RCB scoring performed attained pCR: this was higher than the reported 18.9% (range 6.6–21.5%) in pooled meta-analyses, indicating more discerning selection by physicians in selecting the patient population that would benefit most from neoadjuvant chemotherapy [131]. High rates of pCR predominated in the $ER^{-}HER2^{-}$ and $HER2^{+}$ subgroups (Table 2.3), in keeping with the published literature discussed in Chapter 1. The PAM50 intrinsic subtypes classification also reflected this, with Basal and HER2 enriched tumours also attaining high rates of complete response. Analyses of associations with response to neoadjuvant chemotherapy will be discussed in Chapter 3.

Table 2.3: Proportion of patients within each RCB class across different ER/HER2 and intrinsic PAM50 subtypes

	pCR	**RCB-I**	**RCB-II**	**RCB-III**
ER/HER status				
ER^{-} $HER2^{-}$	43.2%	5.4%	45.9%	5.4%
ER^{+} $HER2^{-}$	9.5%	6.3%	49.2%	34.9%
$HER2^{+}$	32.8%	31.1%	29.5%	6.6%
Intrinsic subtypes				
Luminal A	0.0%	15.8%	42.1%	42.1%
Luminal B	9.3%	13.0%	51.9%	25.9%
Basal	52.6%	5.3%	39.5%	2.6%
HER2	41.7%	30.6%	19.4%	8.3%
Normal	0.0%	14.3%	87.5%	0.0%

Table 2.4: Tumour pathology following neoadjuvant chemotherapy

	% of Cases (n)
Tumour size	
pCR	27.4% (46)
$_{yp}$T1 ($\leq$2cm)	39.3% (66)
$_{yp}$T2 (2 - 5cm)	17.2% (29)
$_{yp}$T3 (>5cm)	12.5% (21)
Not assessed	3.6% (6)
Tumour cellularity	
No tumour	27.4% (46)
1 - 10%	34.5% (58)
11 - 50%	21.4% (36)
>50%	13.1% (22)
Not assessed	3.6% (6)
Number of positive lymph nodes	
0	54.8% (92)
1 - 10	36.3% (61)
>10	4.7% (8)
Not assessed	4.2% (7)
Number of fibrotic lymph nodes	
0	70.8% (119)
1 - 10	25.0% (42)
Not assessed	4.2% (7)
Largest lymph node deposit	
0cm	54.7% (92)
<2cm	39.3% (66)
2 - 5cm	1.2% (2)
>5cm	0.6% (1)
Not assessed	4.2% (7)
% TILS	
0%	8.9% (15)
1 - 10%	47.6% (80)
11 - 50%	7.7% (13)
>51%	5.4% (9)
RCB category	
pCR	25.0% (42)
RCB-I	14.9% (25)
RCB-II	39.3% (66)
RCB-III	16.6% (28)
Not assessed	4.2% (7)

2.2 Tissue collection and processing

2.2.1 Plasma and buffy coat collection and processing

Whole blood from patients recruited in the TransNEO study was collected at the Oncology Department at Addenbrooke's Hospital, Cambridge in S-Monovette 7.5mL Haematology EDTA tubes, and centrifuged by the Cambridge Breast Unit research team within one hour of collection at 820$\times g$ for 10 minutes at room temperature to partition plasma, buffy coat and erythrocytes. The plasma fraction was then removed, centrifuged at 14,000 RPM for 10 minutes to pellet any remaining cellular debris and the supernatant frozen in 1ml aliquots. The buffy coat fraction obtained following the first centrifugation was resuspended in 10ml of red cell lysis buffer (comprising of 155mM NH_4Cl, 10mM $KHCO_3$, 0.1mM EDTA pH 7.4), split into two 5ml aliquots and left standing at room temperature for 10 minutes. The two aliquots were centrifuged at 3,600$\times g$ for 10 minutes at room temperature, the resultant pellets resuspended in 5ml of red cell lysis buffer, and centrifuged at 3,600$\times g$ for a further 10 minutes at room temperature. Each cell pellet was resuspended in 1ml of phosphate buffered saline, split into two 500µL aliquots and centrifuged at 10,000 RPM for 5 minutes. Two of the final four white cell pellets were suspended in 700µL of Qiazol lysis reagent each for eventual RNA extraction, and two pellets were frozen without the addition of lysis buffer for eventual DNA extraction.

2.2.2 Tumour tissue collection and processing

Tumour tissue was collected prior to the initiation of neoadjuvant chemotherapy, midway through treatment, and at the time of surgery. Biopsies at the diagnostic and midway time point were obtained by the radiology staff, whilst tissue at surgery was obtained by the operating surgeon. All samples were flash frozen in liquid nitrogen and stored at -80°C. Sectioning of the samples was performed on a cryostat (CM1520; Leica Biosystems, Germany) by ***Helen Bardwell*** in the Histopathology Core. Following an initial 6µm section taken for haematoxylin and eosin (H&E) staining, twenty 30µm sections were taken and ten sections placed in each of two tubes containing either 180µL ATL buffer or 700µL of Qiazol for DNA or RNA extraction respectively. This process was repeated if enough tissue was available, and a final 6µm section for H&E staining was taken. All sample were stored at -80°C until required for extraction. Any remaining tumour tissue was banked in a Human

Tissue Act compliant freezer. The histology slides containing the initial, midway and final tumour sections were stained with H&E, and tumour, stromal, and immune infiltrate scoring was performed by ***Dr Elena Provenzano***.

For a select number of cases, if fresh frozen tumours were not available, Formalin Fixed and Paraffin Embedded (FFPE) blocks were obtained from the Department of Pathology, sectioned by ***Helen Bardwell*** and areas suitable for coring identified by ***Dr Elena Provenzano***. 1.5mm punches were then taken using the Manual Tissue Arrayer MTA-1 (AlphaMetrix Biotech) and stored in a desiccator until required.

2.3 Nucleic acid processing

2.3.1 Tumour and buffy coat DNA extraction

Isolation of DNA from a total of 502 buffy coat and sectioned tumour tissue samples was performed using the QIAGEN DNeasy Blood and Tissue Kit (Cat No: 69506). DNA from tumour tissue was extracted using the manufacturer recommended protocol. Briefly, tissue sections suspended in 180µL ATL buffer were thawed and 20µL of proteinase K added and incubated overnight at 56°C on a thermal shaker at 650 RPM. Following this, RNA digestion was performed by adding 4µL of RNase A. 200µL of Buffer AL and 200µL of ethanol were added to each sample and the mixture pipetted to a spin column, which was spun at 6,000×*g* in a bench centrifuge. Two washes of the column membrane were subsequently performed, the first with 500µL of Buffer AW1 (with a 1 minute 6,000×*g* spin) and a second with 500µL of Buffer AW2 (with a 3 minute 20,000×*g* spin). Elution of DNA from the column membrane was performed using two 50µL Buffer AE washes, each with a 1 minute 6,000×*g* spin.

In view of an expected high DNA yield from each buffy coat sample, which would saturate the DNA spin column membrane and result in an inefficient extraction, each white cell pellet was suspended in 400µL PBS, and digested with 80µL proteinase K (at >600 mAU/ml), 8µL RNAse A (at 100 mg/ml) and 400µL AL lysis buffer for 12 hours at 56°C. Following this, the lysate was divided equally into two volumes and DNA extraction performed as described above.

DNA quantification was performed using the Qubit Fluorometer (Invitrogen). The median DNA concentration obtained from the buffy coat samples was 75ng/µL (range: 7-328ng/µL), whilst the median concentration obtained from the tumour tissues was 20ng/µL (range: 1-120ng/µL, Figure 2.4A). Tumour samples obtained prior to the commencement of chemotherapy had higher concentrations of DNA compared to those obtained at the midway and surgical time points (median DNA concentrations: diagnosis 29ng/µL, midway 12ng/µL, surgical 13ng/µL). The decrease in extracted DNA yield could be explained by a decrease in viable cellular content and an increase in necrotic tissue secondary to chemotherapy. Thus, as cytotoxic therapies were administered increasing amounts of necrotic tissue were sampled, and therefore less DNA was available for extraction.

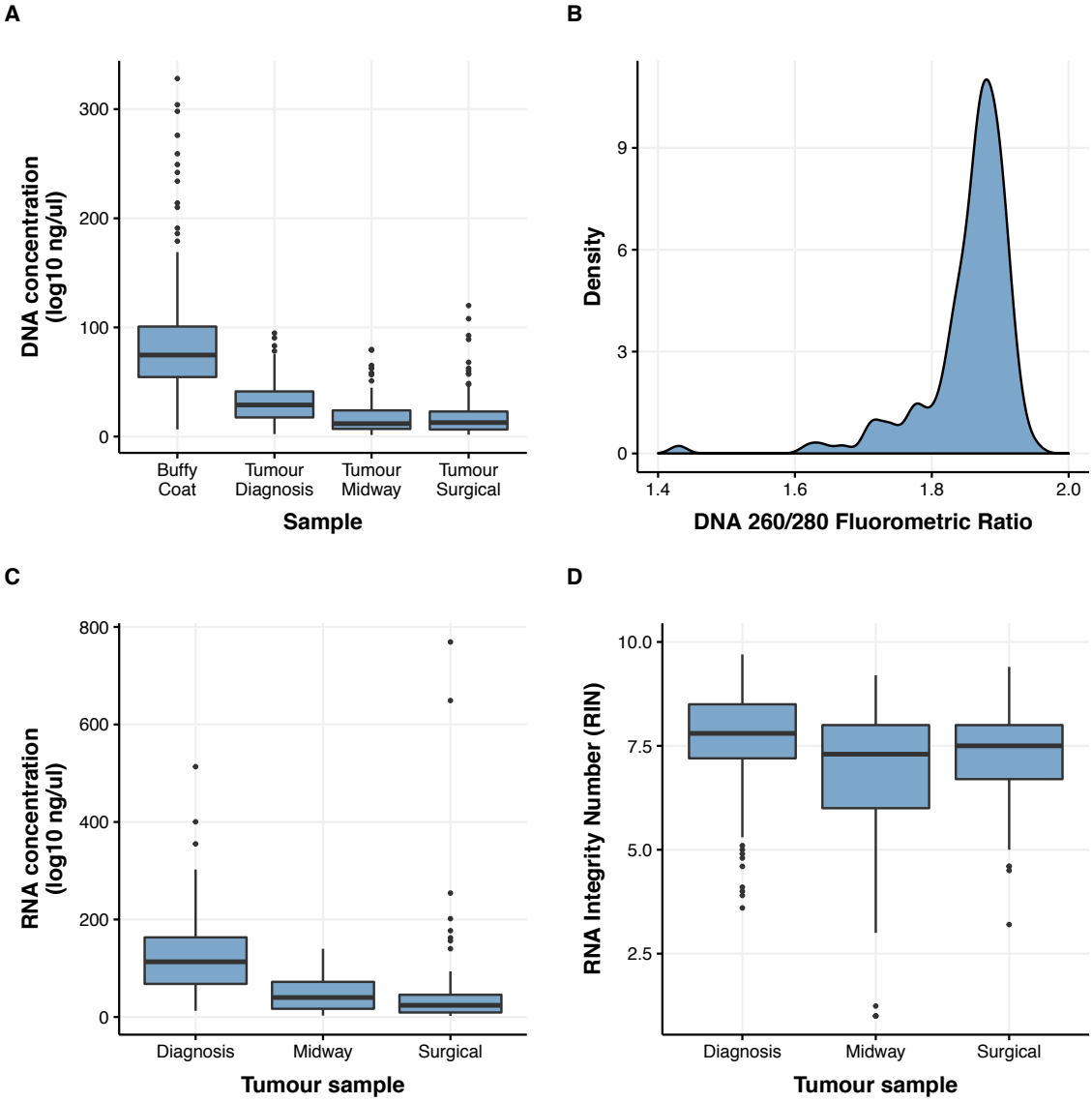


Figure 2.4: Nucleic acid quantification and assessment. (A) Box plots showing DNA concentrations obtained from normal and tumour samples. (B) Density plots showing distribution of 260/280 ratios as obtained by NanoDrop 8000. (C) Box plots showing RNA concentrations obtained from tumour samples. (D) Box plots showing RIN distribution across sampling time points.

Assessment of 260/280 fluorometric ratios was performed using the NanoDrop 8000 (ThermoFisher scientific) to assess the purity of the nucleic acids. 260/280 ratios of around 1.8 were accepted as "pure" for DNA. The median 260/280 ratio for the extracted DNA samples was 1.87 (Figure 2.4B), both showing a very low presence of any contaminants which would impede downstream genomic applications. There was no difference in median ratio across tumour samples obtained from serial time points.

DNA from FFPE tissues was extracted using the QIAGEN QIAamp DNA FFPE Tissue Kit (Cat No: 56404) using the manufacturer's recommended protocol. Briefly, 320µL of QIAGEN deparaffinization solution was added to the paraffin embedded cores and incubated at 56°C for 3 minutes. Following this, 180 µL of buffer Buffer ATL was added to each sample, mixed by vortexing and centrifuged for 1 minute at 11,000×g. Subsequently, 20µL of proteinase K was added, and the samples incubated at 56°C overnight. A final incubation at 90°C for 1 hour ensured that Buffer ATL partially reversed formaldehyde modification. DNA extraction was then performed as follows:

1. addition of 200µL Buffer AL and 200µL of 100% ethanol, transfer to a spin column, followed by a 6,000×g 1-minute spin
2. addition of 500µL Buffer AW1 followed by a 6,000×g 1-minute spin
3. addition of 500µL Buffer AW2 followed by a 6,000×g 1-minute spin
4. a 20,000×g 3-minute spin to dry the spin column membrane and prevent carry over of buffer
5. elution with 50µL of Buffer ATE and a 20,000×g 1 minute spin

DNA quantification and contaminant estimation were performed using the Qubit Fluorometer (Invitrogen) and NanoDrop 8000 (ThermoFisher scientific).

2.3.2 Tumour RNA extraction

Isolation of RNA from 329 tumour tissue samples was performed using the QIAGEN miRNeasy Mini Kit (Cat No: 217004) which purified both microRNA and total RNA. Tissue sections suspended in 700µL of Qiazol were thawed and mixed by vortexing. 140µL of chloroform was added to each sample, vortexed and transferred to a heavy phase lock tube (QIAGEN MaXtract, cat no: 129056). The samples were then spun at 12,000×g for 15 minutes at 4°C, following which the upper clear phase containing RNA was transferred to a 2ml Eppendorf tube. Subsequent extraction was then performed using the QIAGEN

QIAsymphony in order to limit batch effect across all samples. Within the QIAsymphony, further washes were done as follows:

1. addition of 525µL 100% ethanol followed by an 8,000×g 15-second spin
2. addition of 700µL Buffer RWT followed by an 8,000×g 15-second spin
3. addition of 500µL Buffer RPE followed by an 8,000×g 15-second spin
4. addition of 500µL Buffer RPE followed by an 8,000×g 2-minute spin
5. elution with 60µL of RNase free water and an 8,000×g 1-minute spin

RNA quantification was performed using the Qubit Fluorometer (Invitrogen). As observed previously for DNA, RNA yields were lower at later time points. Indeed, the median RNA concentration prior to the commencement of chemotherapy was 113ng/µL, which was significantly higher than those obtained at the midway time point (40ng/µL) and following the completion of treatment (24ng/µL) (Welch Two Sample t-test $p < 2.2e^{-16}$, Figure 2.4C).

Assessment of the RNA integrity number (RIN) was performed using the High Sensitivity RNA assays on either the Agilent 4200 TapeStation Instrument or Agilent 2100 Bioanalyzer. The RIN is a ratio of the area of the 18S and 28S rRNA peaks to the total area under the electropherogram, with higher ratios indicative of very little degradation. The median RIN for the RNA samples was 8 (Figure 2.4D), showing that the RNA extracted was of a high quality and sequencing libraries could be generated from it.

2.3.3 Cell-free DNA processing

DNA extraction from 305 plasma samples for the first 80 recruited patients was performed at the Cancer Molecular Diagnostics Laboratory at Cambridge using the QIAsymphony DSP Circulating DNA Kit (Cat No: 937556). The automated recovery method ensured a high-efficiency extraction using magnetic beads. Briefly circulating DNA was extracted using the default 4ml extraction protocol: following the addition of 220µL proteinase K, binding buffer and magnetic beads to the samples, the samples were transferred to the QIAsymphony machine, wherein magnetic separation was performed and automated wash steps undergone in order to remove non-nucleic acid contaminants. Following this, the DNA bound to the magnetic beads was eluted into 60µL of buffer. Quantification of the total concentration of DNA eluted was obtained using the Qubit Fluorometer (Invitrogen).

The median concentration of circulating DNA following extraction was 0.58ng/μL (range: 0-10.4ng/μL). The median concentration of all plasma samples at the different sampling time points was 0.46ng/μL prior to commencing treatment, 0.74ng/μL after the first cycle of treatment, 0.69ng/μL midway through the treatment cycle and 0.52ng/μL at the end of treatment.

2.4 Genomic and transcriptomic library preparation

2.4.1 Whole genome and whole exome library generation

DNA samples were normalised to a concentration of 5ng/μL through a series of serial dilutions and quantifications using the Qubit Fluorometer (Invitrogen). Samples with concentrations less than 5ng/μL were concentrated using a SpeedVac (Thermo Fischer Scientific). All normalised samples were transferred to 96 well plates: the location of each patient's samples was randomised based on ER, HER2 and pathological response (pCR vs. RD) in order to ensure that these clinical variables would not confound downstream analyses. Samples obtained to the same individual were pipetted into adjacent wells across the same row in order to ensure that they would be captured together in the same pool in later stages of the protocol, limiting intra-patient variability.

Exome libraries were prepared by the ***Genomics Core*** at the Cancer Research (UK) Cancer Institute using the Illumina Nextera Rapid Capture Exome Library Preparation kit according to the manufacturer's protocol (Illumina document number: 15037436). This kit was selected as it was able to generate whole exome libraries from as little as 50ng of DNA: this was specifically important as some of the post-chemotherapy samples had lower quantities of DNA extracted. The protocol contained three main steps: (1) Tagmentation of DNA, in which a Tn5 transposase performs simultaneous fragmentation and tagging of DNA [6], (2) amplification and addition of unique indexes and (3) exome probe hybridisation and target enrichment, as described below.

1. Tagmentation

25μL Tagment DNA Buffer (TD) and 15μL Tagment DNA Enzyme 1 (TDE1) were added to 10μL of DNA at 5ng/μL and incubated at 58°C for 10 minutes. Following this 15μL of Stop Tagmentation (ST) buffer was added to stop the tagmentation reaction. The tagmented DNA was purified by the addition of 65μL of Sample Purification Beads (SPB) and magnetic separation on a magnetic stand. The beads were then washed twice with 200μL of freshly prepared 80% ethanol and the purified tagmented DNA eluted in 22.5μL of Resuspension Buffer (RSB).

2. Amplification and index addition to generate a whole genome library

5µL of unique Index 1 (i7) and Index 2 (i5) adapters were added to 20µL of each purified tagmented DNA sample, such that each sample on the 96 well plate had a different combination of i7 and i5 indexes. 20µL of Library Amplification Mix (NLM) were added and a limited cycle polymerase chain reaction (PCR) performed using the following thermal cycling protocol:

- 72°C for 3 minutes
- 98°C for 30 seconds
- 10 cycles of:
  - 98°C for 10 seconds
  - 60°C for 30 seconds
  - 72°C for 30 seconds
  - 72°C for 5 minutes
- Hold at 10°C

The amplified DNA was purified by adding 90µL of SPB followed by magnetic separation and two washes with freshly prepared 80% ethanol. The amplified library was eluted in 27µL of Buffer RSB, and the DNA concentration measured using the Qubit Fluorometer. Fragment size distribution was assessed using the Agilent 4200 TapeStation Instrument.

The amplified library produced was, in essence, a whole genome library, and all samples per plate were pooled together at 5nM and sequenced on two lanes of an Illumina HiSeq 4000 sequencer in 50 base pair single read mode to obtain copy number profiles.

3. Exome probe hybridisation and target enrichment

500ng of library were pooled into 6-plex reactions, ensuring that all samples belonging to the same patient were always within the same capture pool. The volume of each pool of DNA was adjusted to 40µL and 50µL of Enrichment Hybridisation Buffer (EHB) and 10µL of biotinylated Rapid Capture Oligos added. Hybridisation was performed on a thermocycler programmed to cycle as follows: 95°C for 10 minutes, followed by 18 one minute cycles, starting at 94°C, then decreasing 2°C per cycle. The PCR product was incubated overnight at 58°C, and the next day the amplified product was purified by adding 250µL Streptavidin Magnetic Beads and incubated at room temperature for 25 minutes. Following magnetic

bead separation, bead washing was performed twice using 200µL of Enrichment Wash Solution (EWS). The purified enriched library was eluted in 23µL of elution buffer (28.5µL Enrichment Elution Buffer 1 and 1.5ulµL HP3), as well as 4µL Elute Target Buffer 2 (ET2).

These libraries then underwent a second hybridisation with the same capture probe set in order to increase the specificity of the captured regions and decrease off-target fragments. The final enriched libraries were amplified using 20µL of Enrichment Amplification Mix and 5µL of PCR Primer Cocktail, and a polymerase chain reaction performed using the following thermal cycling protocol:

- 98°C for 30 seconds
- 12 cycles of:
  - 98°C for 10 seconds
  - 60°C for 30 seconds
  - 72°C for 30 seconds
- 72°C for 5 minutes
- Hold at 10°C

The amplified enriched libraries were purified by adding 90µL of SPB and performing a magnetic separation, followed by two washes with freshly prepared 80% ethanol. The libraries were then eluted in 32µL of Buffer RSB. The DNA concentration of each sample was measured using the Qubit Fluorometer and the fragment size distribution assessed using the high sensitivity assay on the Agilent 4200 TapeStation Instrument. Five nanomolars of each library was prepared and 48 samples pooled per lane of sequencing on an Illumina HiSeq4000 system, with sequencing performed at 75 base pair paired-end mode.

2.4.2 Targeted sequencing library generation

Targeted sequencing was performed using the Raindance Thunderbolts NGS Target Enrichment System (Raindance Technologies). This technology enabled single molecule PCR within oil droplets, with up to 8 million PCR droplets synthesised per sample and as little as 5ng of DNA required as input. Library preparation involved the initial generation of droplets and a first round of PCR within the oil emulsion, followed by the addition of adaptor and index sequences in preparation for clustering and sequencing.

Droplet generation and amplification

A premix containing the reagents listed in Table 2.5 was assembled per sample and loaded into the Raindance Thunderbolts system. Following this, the reaction mixtures were partitioned in oil droplets, such that each droplet contained one or no DNA molecules, ensuring single molecule PCR.

Table 2.5: PCR 1 reagent components

Reagent	Volume (µL)
TaqMan Genotyping Master Mix (Life Technologies: 4371355)	20
25× (12.5%) Droplet Stabilizer	1.6
Custom Primers	4
DNA (5-75ng)	up to 14.4
Water	variable

On completion of the run, the emulsions were collected and placed on a thermocycler, and an in-droplet PCR reaction initiated using the following cycling settings:

- 94°C for 2 minutes
- 55 cycles of:
  - Ramp 1°C/second to 94°C
  - 94°C for 30 seconds
  - Ramp 1°C/second to 54°C
  - 54°C for 30 seconds
  - Ramp 1°C/second to 68°C
  - 68°C for 1 minute
- 68°C for 10 minutes
- Hold at 12°C

Following PCR amplification, the droplet emulsions were destabilised using 50µL of Droplet Destabiliser, and the aqueous layer containing the amplified product retained. A bead clean up using 2× AMPure XP beads was performed: this was followed by two washes with 180µL of freshly-made 80% ethanol and a final elution in 20µL of 10nM Tris HCl pH8.0.

Adapter and index addition

A second PCR premix containing the components listed in Table 2.6 was assembled for each reaction: this allowed the addition of Illumina P5 and P7 adapters, as well as unique barcodes that would enable the identification of each individual sample following sequencing demultiplexing.

Table 2.6: PCR 2 reagent components

Reagent	Volume (µL)
Platinum Taq DNA Polymerase High Fidelity Buffer	3.25
50 mM MgSO4 (included in Life Technologies 11304-029)	0.875
dNTP mix (10mM) (Integrated DNA Technologies)	1.125
DMSO (Sigma D8418-100ML)	1.25
10× Barcoded Illumina Primers (5µM)	2.5
10× Platinum Taq DNA Polymerase High Fidelity	0.5
1st PCR Template DNA	13

A barcoding PCR was then performed using the following settings on a thermocycler:

- 94°C for 2 minutes
- 11 cycles of:
  - Ramp 1°C/second to 94°C
  - 94°C for 30 seconds
  - Ramp 1°C/second to 56°C
  - 56°C for 30 seconds
  - Ramp 1°C/second to 68°C
  - 68°C for 1 minute
- 68°C for 10 minutes
- Hold at 12°C

The amplified products were then purified using 1.2× AMPure XP beads and washed twice using 180µL of freshly prepared 80% ethanol. 20µL of 10nM Tris HCl pH8.0 were used to elute the final products from the beads. Fragment length determination was performed using the Agilent Tapestation and quantification performed using qPCR. Following quantification

and pooling, sequencing was performed on one lane of a MiSeq sequencer, with a maximum of 24 samples sequenced per lane.

Primer selection and optimisation

Customised primers for specific regions containing mutations of interest detected on exome sequencing were designed and multiplexed together. The Raindance Manufacturer's protocol suggested using 0.8μM of primers along with up to 75ng of DNA, however at lower DNA concentrations this resulted in an abundance of primer dimers. To determine the optimal concentration of primer mix and DNA, a serial dilution experiment was designed. Library preparation was performed using four reactions in which the primer concentration was 0.04nmol and another four in which the primer concentration was lowered to 0.008nmol. For each of these two sets of reactions, input DNA quantities were varied: 25ng, 10ng, 5ng and 1ng. Following completion of library preparation, the products were analysed on an Agilent Tapestation in order to assess the product/primer dimer ratio. Of all the combinations tested, a primer concentration of 0.008nmol, together with an input DNA concentration of 25ng were shown to give rise to the least amount of primer dimers and were therefore used in this protocol.

2.4.3 Whole transcriptome library generation

RNA samples were normalised to a concentration of 10ng/μL through a series of serial dilutions and quantifications using the Qubit Fluorometer (Invitrogen). All normalised samples were transferred to 96 well plates using a randomisation methodology similar to the one described previously: the location of each patient's samples on a plate was randomised based on ER status, HER2 status and response to chemotherapy in order to ensure that these clinical variables did not confound downstream analyses.

Transcriptomic libraries were prepared using the Illumina TruSeq Stranded mRNA Library Preparation kit (Cat No: 20020595) by the ***Genomics Core*** at the Cancer Research (UK) Cancer Institute, according to the manufacturer's protocol (Illumina document number: 1000000040498). Briefly, the protocol consisted of six main steps:

1. mRNA purification and fragmentation, in which Poly-A containing mRNA molecules were purified using magnetic beads with Poly-T oligos and the isolated mRNA was fragmented using divalent cations at elevated temperatures.
2. First and second strand complementary DNA (cDNA) synthesis, where cleaved RNA fragments were converted to first strand cDNA using reverse transcriptase and the second cDNA strand was synthesised using DNA Polymerase I and RNase H.
3. 3' Adenylation and adapter ligation, where an adenine was added to the 3' ends of the blunt fragments to prevent the molecules from ligating to each other, and P5/P7 adapters and barcodes were added to each sample to allow hybridization onto a flowcell.
4. DNA enrichment and purification.

1. mRNA purification and fragmentation

50µL of RNA Purification Beads (RPB) were added to each well containing 50µL of RNA at 10ng/µL and mRNA denaturation induced by incubating at 65°C for 5 minutes. Magnetic separation was performed, the supernatant discarded and the beads washed with 200µL of Bead Washing Buffer (BWB). 50µL of Elution Buffer were added and the solution incubated at 80°C on a thermal cycler for 2 minutes, followed by a 25°C hold. Following this, the beads were washed with 50µL of Bead Binding Buffer. 200µL of BWB and 19.5µL of Fragment, Prime, Finish Mix were added to each well and the plates incubated at 94°C for 8 minutes to induce mRNA fragmentation. Following a magnetic bead separation, 17µL of the supernatant were retained for each sample.

2. First and second strand cDNA synthesis

8µL of a mixture of SuperScript II and First Strand Synthesis Act D Mix were added to each sample and the first strand cDNA synthesis performed by means of a PCR with the following temperature settings:

- 25°C for 10 minutes
- 42°C for 15 minutes
- 70°C for 15 minutes
- Hold at 4°C

The second cDNA strand was then synthesised by adding 5µL of End Repair Control solution (CTE) diluted in 1:50 RSB and 20µL of Second Strand Marking Master Mix to each well and incubating at 16°C for 1 hour. Following this the cDNA was purified using 90µL AMPure XP beads and incubated at room temperature for 15 minutes. Magnetic separation was performed, the beads washed twice with 200µL of freshly prepared 80% ethanol and the purified tagmented RNA eluted in 17.5µL of Resuspension Buffer (RSB).

3. 3' Adenylation and adapter ligation

2.5µL of A-Tailing control buffer (CTA) and 12.5µL A-Tailing Mix were added to each well and incubated at 37°C for 30 minutes and 70°C for 5 minutes.

Adapter ligation was performed by adding 2.5µL of Ligation Control Mix, 2.5µL of Ligation Mix and 2.5µL of RNA adapters to each well, and incubated at 30°C for 10 minutes. 5µL of Stop Ligation Buffer were added to halt the reaction. A clean up of the ligated fragments was done using 42µL of AMPure XP beads and two 200µL 80% ethanol washes. The fragments were eluted using 52.5µL of RSB. A second clean up was performed using 50µL of AMPure XP beads and two 200µL 80% ethanol washes and the fragments eluted in 22.5µL of RSB.

4. DNA enrichment and purification

25µL PCR Master Mix were added to each well and a PCR performed on a thermocycler using the following settings:

- 98°C for 30seconds
- 15 cycles of:
  - 98°C for 10 seconds
  - 60°C for 30 seconds
  - 72°C for 30 seconds
- 72°C for 5 minutes
- Hold at 4°C

A final clean-up was performed using 50µL of AMPure XP beads and two washes of 200µL fresh 80% ethanol. The fragments were eluted in 32.5µL of RSB and the DNA concentration of each sample quantified using qPCR. 5nM of each library was prepared and 94 samples pooled per lane of sequencing on an Illumina HiSeq4000 system (75bp paired-end mode).

2.5 Bioinformatic analysis

2.5.1 Exome and shallow whole genome analysis

Alignment and preprocessing

The GRCh37 decoy (b37) assembly of the human genome used by the 1000 Genomes Project Consortium was selected as the reference genome of choice [1, 2, 97]. This assembly contained chromosomal sequences present in the initial GRCh37 release as well as:

1. the Revised Cambridge Reference Sequence of the human mitochondrial DNA genome (NCBI Accession: NC_012920)
2. the Human herpes virus 4 type 1 genome (NCBI Accession: NC_007605)
3. decoy sequences missing from the GRCh37 assembly derived from HuRef, Human BAC and Fosmid clones and the ALLPATH-LG assembly of the Illumina NA12878 platinum genome.

The additional sequences added to the assembly have been shown to improve aligner efficiency, as reads derived from sequences present in regions captured by the decoy would be rapidly aligned confidently and therefore result in more accurate alignments as the decoy would absorb reads that would otherwise map with low quality and mismatches to the reference.

For each exome paired FASTQ file, sequencing quality metrics were generated using `FastQC` (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Alignment was performed using `Novoalign` version 3.2.13 (http://www.novocraft.com) in paired-end mode with the following parameters enabled: (1) base quality recalibration, (2) trimming of Nextera adaptor sequence `CTGTCTCTTATA`, (3) hard clipping of trailing bases with quality ≤ 20. Shallow whole genome sequencing (sWGS) data was processed in a similar manner, however `Novoalign` was run in single read mode. `Novoalign` was used as the aligner of choice as it has been shown to make the fewest mapping mistakes (0.019%) compared to BWA-MEM (0.777% of reads incorrectly mapped) and Bowtie2 (3.72% of reads incorrectly mapped) [123].

Binary aligned sequencing (BAM) file merging, coordinate sorting and PCR and optical duplicate marking were performed using `Novosort`. Local realignment around insertions and deletions was performed using the Genome Analysis Toolkit (GATK) programs

`RealignerTargetCreator` and `IndelRealigner`, using a calibration set derived from the 1000 Genomes project that has been validated to a high degree of confidence for increased accuracy [65, 198, 294].

Quality assessment

In order to assess the performance of the library preparation as well as the quality of the sequencing data, target coverage metrics within exonic regions specified by the Nextera target BED file obtained from Illumina (Manifest version 1.2) were generated using `Picard` (version 2.17.0) `CalculateHSMetrics`, whilst fragment insert size metrics were computed using `Picard CollectInsertSizeMetrics`.

All DNA samples that were sequenced generated usable exome data, with no failures across the cohort. The median number of paired-end sequencing reads per sample was 192 million (range: 73 - 481 million, Figure 2.5A). A median of 86.3% of sequenced reads aligned to the reference genome and 53.5% of these reads aligned to within the exome target regions (Figures 2.5B and C). The median PCR duplication rate was 32%: the number of PCR duplicates per sample correlated with the number of total reads in the sample and followed the law of diminishing returns (Figures 2.5D and E). The median insert size of the libraries generated was 145 bases (Figure 2.5F) and therefore there was minimal overlap between the paired-end reads, which were each at 75 base pairs. These metrics confirmed the high quality of the libraries generated.

In order to assess the performance of the exome enrichment protocol, the probe bait statistics were then analysed. The median target coverage across all samples was **160**$\times$ (range: 67 - 369$\times$, Figure 2.6A), with less than 2% of all bases within the target region having no coverage (Figure 2.6B). The exome sequencing data was therefore deep enough to enable rarer variant detection, which was especially important as somatic mutation allelic frequencies decreased during neoadjuvant therapy.

Germline variant calling and sample genotyping

Germline variants were identified using GATK `HaplotypeCaller` (version 4.0.2.1). Briefly, `HaplotypeCaller` identified regions within the genome that had significant evidence of variation and built a De Bruijn-like graph on each region so as to identify potential haplotypes.

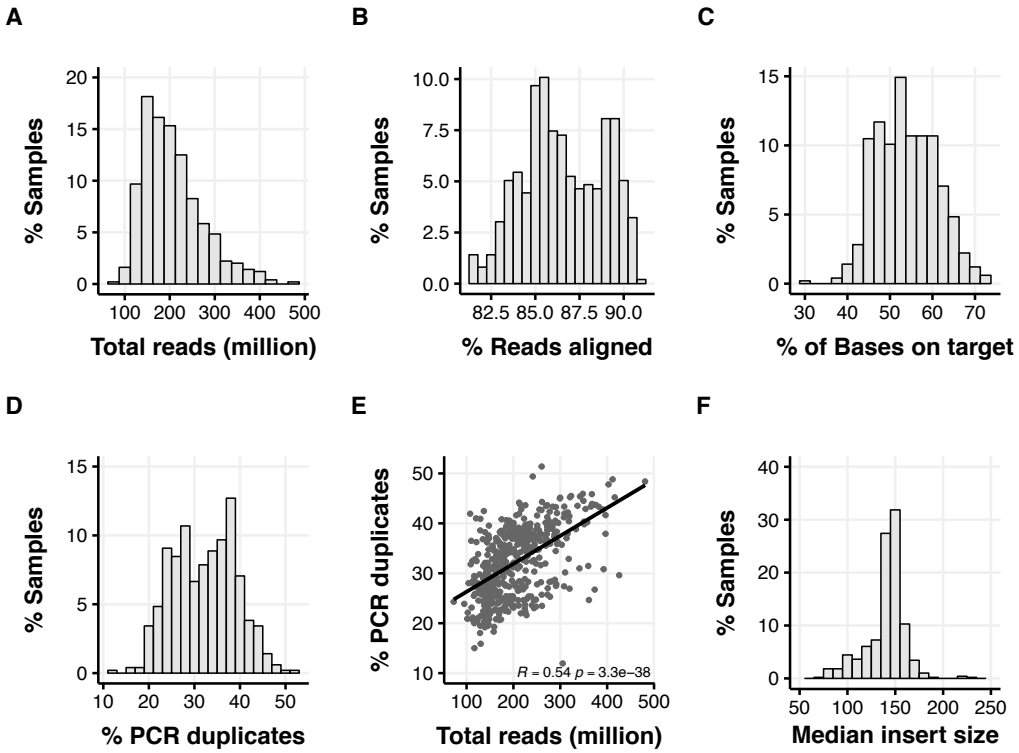


Figure 2.5: Distribution of exome sequencing metrics across all samples showing the distribution of (A) total reads per sample, (B) % of reads aligned per sample, (C) % of bases on target, (D) % of PCR duplication, (E) correlation between number of sequenced fragments and PCR duplication rate and (F) median insert size.

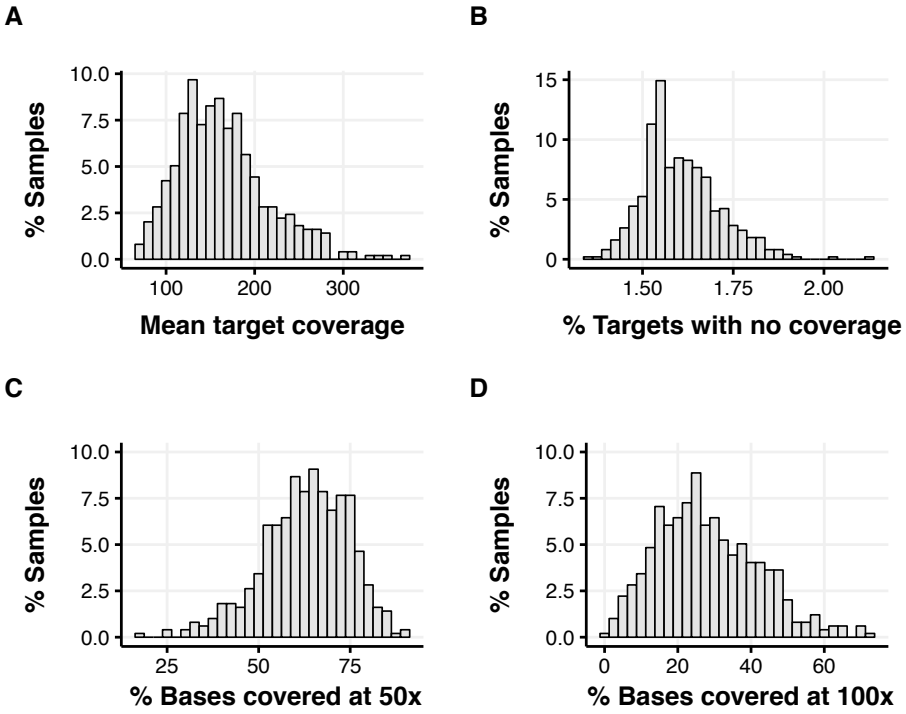


Figure 2.6: Distribution of targeted exome capture metrics across all samples showing distribution of (A) mean target coverage, (B) % of targets with no coverage, (C) % of bases covered at a minimum of 50x and (D) % of bases covered at a minimum of 100x.

A pairwise alignment of each read was then performed against each haplotype using the PairHMM algorithm, producing a likelihood matrix of haplotypes. Bayesian modelling was subsequently used to calculate the likelihoods of each genotype per sample given the read data observed and the most likely genotype assigned.

In order to leverage the large number of samples within the cohort, rather than running `HaplotypeCaller` on each sample independently and then filtering based on hard thresholds, the variant caller was run in `GVCF` mode, wherein variant calling was performed across all samples at the same time. This resulted in a combined Variant Calling Format (VCF) file comprising all variants present across all samples and was subsequently filtered using GATK `VariantRecalibrator`, which used machine learning algorithms to identify true variants from sequencing artefact. By providing `VariantRecalibrator` with highly validated training sets, including variant resources from the 1000 Genomes project and the HapMap project, the algorithm learned what the metrics for true variants were across the whole dataset and then used these to identify signal from noise. This varied significantly from the traditional way of performing variant filtration, wherein a predefined set of thresholds and hard filters (including mean depth at variant site, allelic fraction, mapping quality, strand bias) would be applied to a dataset without any machine learning. The key disadvantage of these traditional methods is that many true positives often fall below filtering thresholds and are discarded.

Following variant filtration, germline variants that were only called within tumour samples only were discarded (as these were likely to be somatic). Mutations that deviated significantly from the expected allelic fraction of 0 (homozygous reference), 0.5 (heterozygous) or 1 (homozygous alternative) were identified by using the equation:

$$I_A{}^{(x)} = \begin{cases} 1\, x \ \in A \\ 0\, x \ \notin A \end{cases} \tag{2.1}$$

$$\frac{\sum_{i=1}^{n} I^{(BAFi)}_{(0.1,0.32)\ \cup\ (0.68,0.9)}}{\sum_{i=1}^{n} I^{(BAFi)}_{(0.9,1]\ \cup\ [0,0.1)}} > 0.25 \tag{2.2}$$

This resulted in the removal of 2,156 noisy variants, with a distribution of B-allele frequency (BAF) as shown in Figure 2.7.

A median of 26,249 germline variants were called across all cases (range: 21,044 - 32,393). In order to identify potential sampling errors or sample swaps during the library preparation

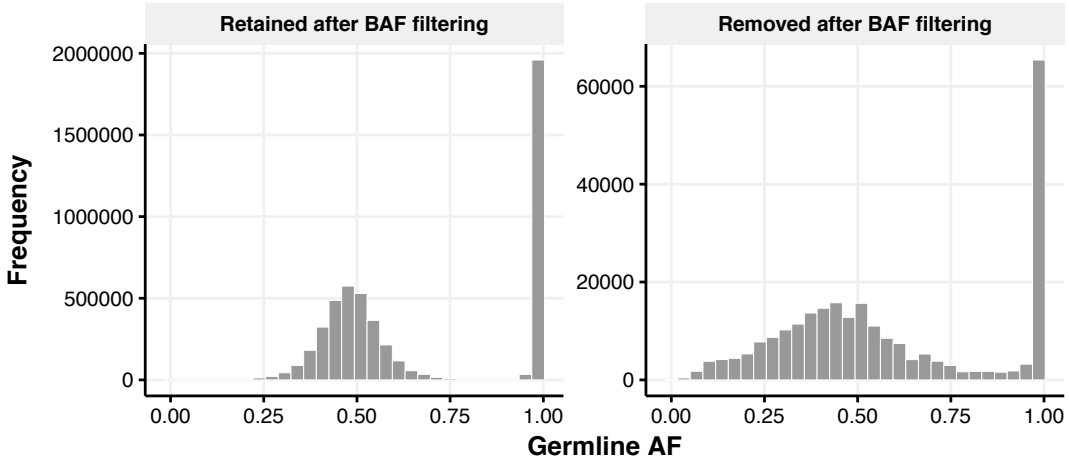


Figure 2.7: Removal of variants likely to be sequencing artefacts or due to misalignment

process, single nucleotide polymorphism (SNP) genotyping was performed across all tumour and normal samples. Homozygous SNPs were selected for comparisons in order to increase the accuracy of genotyping, as heterozygous variants can become homozygous in tumours within areas of loss of heterozygosity (LOH). The percentage of homozygous SNPs shared between a tumour and all normal samples within the cohort was subsequently computed. The percentage median concordance across samples derived from each patient was 99.9%. Unrelated samples had a median shared homozygous SNP concordance of 59.1% (Figure 2.8).

Apart from confirming the robustness of the sequencing data and pairing of samples, the genotyping analysis also showed two unexpected relationships. The tumours and normals from two sets of two cases (T004 and T118, as well as T031 and T054) showed a higher rate of concordance (75-80%) compared to the background concordance rate. This prompted a closer look at the clinical records of these four patients, which revealed two unexpected relationships: T004 and T118 were sisters while T031 and T054 were fraternal twins, explaining the greater than expected concordance in genotypes between these four cases.

All germline variants identified were then annotated using Ensembl `Variant Effect Predictor`, using Ensembl version 87 [8, 194].

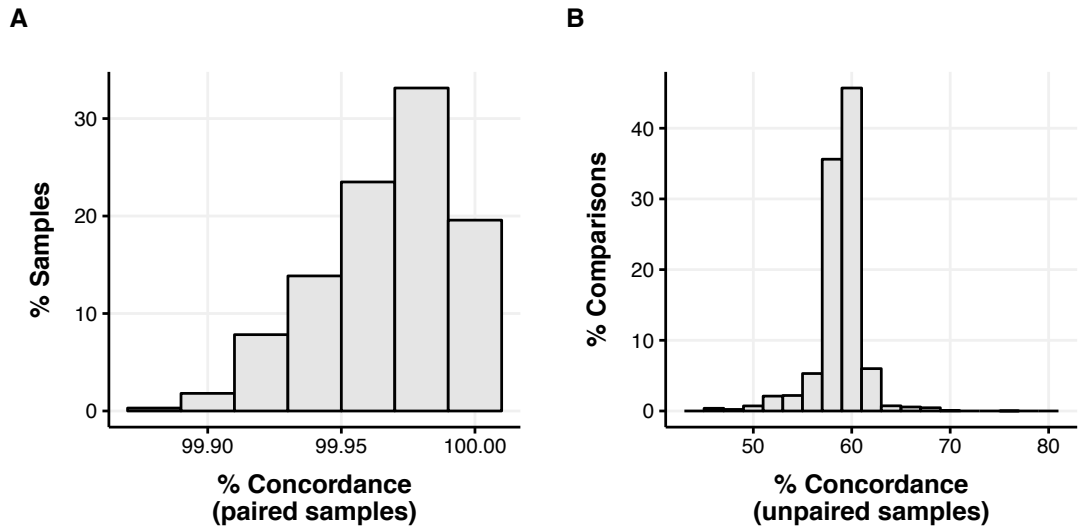


Figure 2.8: SNP genotyping across all homozygous germline mutations. Tumour and normal samples matched with over 99.9% concordance, compared to unrelated comparisons.

Somatic variant calling, filtering and annotation

Somatic variant calling was performed using `Mutect2` from the GATK4 suite. Before commencing variant calling, a panel of normals was created by running `Mutect2` in tumour only mode on all normal samples. The resulting VCF files which contained germline variants, as well as sequence artefacts, were merged using `CreateSomaticPanelOfNormals`, retaining only sites with variants present in more than one sample.

`Mutect2` was then run on each tumour/matched normal sample pair: in order to aid germline variant identification and decrease the false positive rate, the previously generated panel of normals, as well as germline variants present within the gnomAD resource, were also supplied to the variant calling engine.

Filtration for confident somatic calls was performed using `FilterMutectCalls` which applied pre-set thresholds tuned for human somatic analyses in order to separate signal from noise. `FilterMutectCalls` removed any mutations that were:

1. present in matched normal sample
2. present in panel of normals
3. present in germline resource (gnomAD)
4. of low median base quality (median quality $<$20)
5. of low mapping quality (median quality $<$30)

6. present on shorter reads (and were more likely to be due to misalignment after clipping)
7. mostly found on PCR duplicates
8. present on reads in one sequencing direction only
9. present at a very low variant allelic fraction (VAF) and likely due to be due to contamination (VAF $<$ contamination estimation)
10. present close to other mutations (clustered events)
11. present towards the end of a read (within 5 bases of read ends)
12. below the `Mutect2` statistical likelihood threshold for calling (`t_lod`, default threshold: 5.3)

After completion of variant filtering, all tumour VCF files belonging to the same patient were concatenated into one master VCF. `Haplotypecaller` was run in joint genotyping mode across all samples derived from the same patient, using as a guide the master VCF generated previously. Hence by doing so, `Haplotypecaller` determined whether the mutations observed in one sample could also be identified in another, even if that mutation had not been detected by `Mutect2`. This *variant rescuing pipeline* developed was especially important as, during chemotherapy, tumour purity was expected to decrease with subsequent decreases in VAF. By using this method, variants that had been detected prior to starting chemotherapy, but had not been called at later time points because of low VAF, would be 'rescued' and retained.

In order to further fine-tune the variant call set, a *compositional filter* was developed to remove variants falling within regions of low complexity within the human genome. Previous work has shown that variants called within these regions often were secondary to misalignment [170]. In order to generate such a filter, areas of low complexity regions were identified by:

1. selecting regions within `Low_complexity`, `Satellite` and `Simple_repeat` regions as annotated by `Repeat Masker`
2. selecting regions identified as having low complexity by the `mDUST` algorithm (hosted at: ftp://occams.dfci.harvard.edu/pub/bio/tgi/software/seqclean/), which is a stand-alone implementation of the DUST algorithm first used by BLAST
3. writing a custom high throughput C script that identifies homopolymeric regions, defined as any region with 6 or more iterations of the same base.

Variants falling within these low complexity regions were removed and excluded from the analysis.

A final round of filtering was performed by trying to determine the false positive rate of the dataset and identify any unique features. 29 samples that were exome sequenced did not have any tumour cells within them on histopathological review and therefore these were used in order to benchmark the false positive rate, as any mutations called within these samples are artefacts. Indeed, without any further filtering, the false positive rate across the dataset was 0.39 mutations per megabase of exon: most of these mutations had allelic fractions of less than 5%, with genotype quality scores less than 30 and coverage of less than 25× (Figures 2.9A, B and C).

In view of this, samples were filtered further by applying additional hard thresholds. Variants passing these criteria were retained:

1. Genotype quality > 30
2. Coverage > 25×
3. VAF in normal sample < 0.05
4. Minimum tumour VAF 0.05
5. Variants that are present in gnomAD in < 1% of the population
6. passed OxoG filtering

OxoG artefacts induced during the library preparation process were detected using the tool `FilterByOrientationBias`. Oxidation of DNA, causing the conversion of guanine to 8-oxoG, which then pairs with both cytosine and adenosine during PCR leading to C>A/G>T is commonly seen in sequencing data [59] and results in a high false positive rate. The addition of these filters **decreased the false positive rate to 0.08 mutations per megabase of exon and increased specificity by an additional 80%** (Figures 2.9D and E).

Variant annotation was performed using Ensembl `Variant Effect Predictor`, using Ensembl version 87 [8, 194].

Copy number calling

Genome binning and segmentation on low pass shallow whole genome sequencing BAM files was performed using the R package `QDNAseq` [251]. Binning was performed across 100kb windows and counts corrected for GC-rich regions as well as poorly mappable regions. Normal sequencing data was used to correct for technical and germline artefacts. Segmentation was then performed using the Circular Binary Segmentation algorithm implemented in

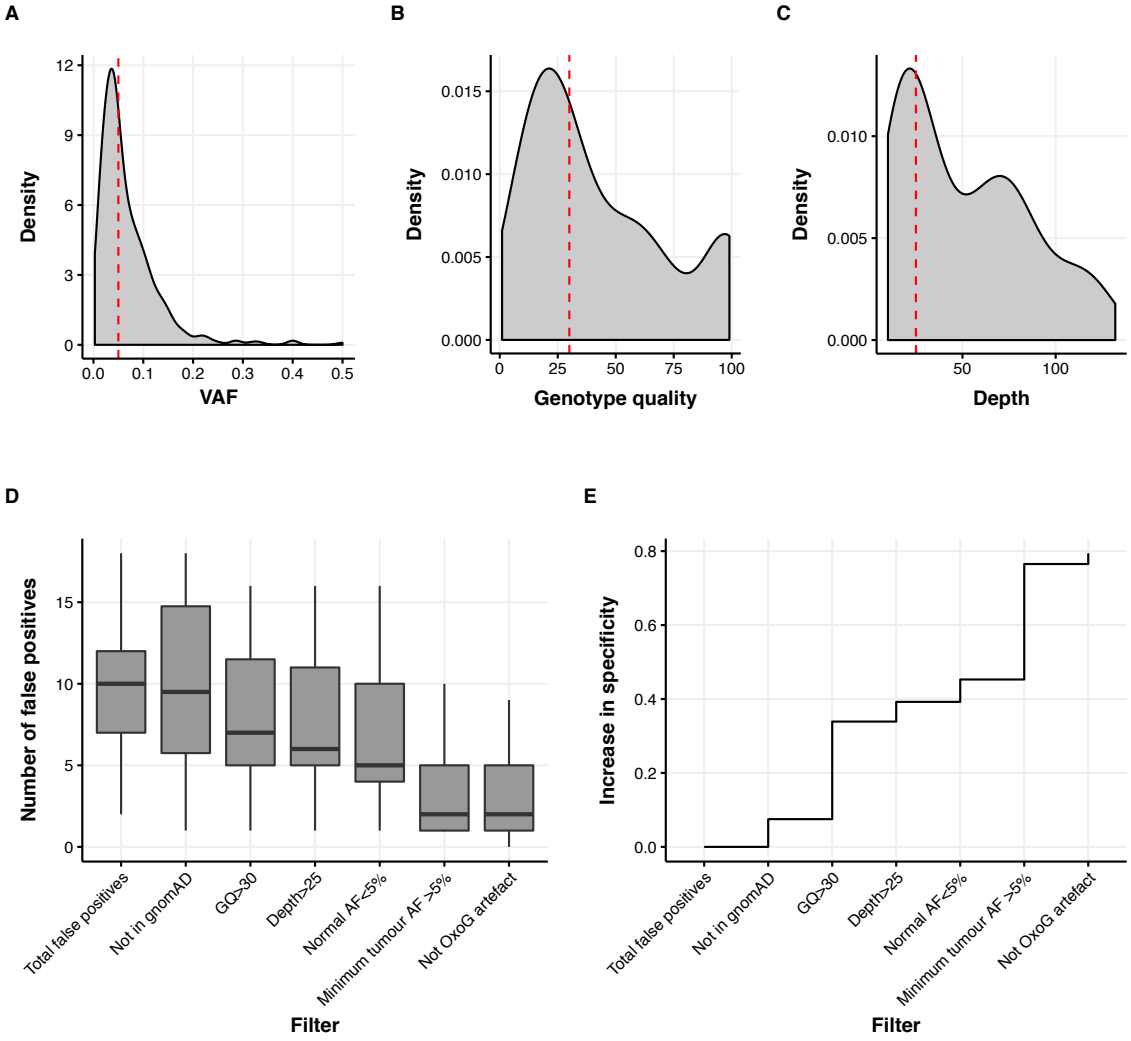


Figure 2.9: Identifying features of false positive calls. Artefacts had: (A) lower VAFs, (B) lower genotype qualities, and (C) were often called in regions of lower coverage. Red dashed line indicates filtering cut-offs selected. (D, E) Gain in specificity with additional hard threshold filtering

the R package DNAcopy [218]. The output from this algorithm was the averaged log ratio of each 100kb bin within the genome (Figure 2.10A).

Parental copy number quantification, as well as computational estimation of tumour purity and ploidy were then performed using ASCAT version 2.5.1 [296], which performed joint segmentation and copy number calling across related samples. Log ratios derived from QDNAseq, as well as a common set of germline SNPs for all normal and tumour samples obtained from the same patient were supplied to ASCAT. As recommended by the authors, the technology parameter gamma was set to 1 for exome sequencing SNP data. Additionally, the MAXPLOIDY parameter was increased to 8 (from 5.5), the MINRHO cellularity parameter decreased to 0.05 (from 0.2) and the MINGOODNESSOFFIT parameter decreased to 60 (from 80). The tuning of parameters allowed ASCAT to estimate copy number at lower cellularities, which was especially important as tumour purity often was less than 20% as response to chemotherapy occurred. Hence, for each sample, the parental copy number was obtained per segment of genome, allowing the identification of amplifications, gains, losses and loss of heterozygosity (Figure 2.10B).

Clonal reconstruction

Clonal reconstruction was performed using the PyClone Bayesian clustering method [242]. PyClone utilised the allelic frequency and parental copy number of deeply sequenced somatic mutations, together with tumour purity, to compute cancer cellular frequencies (CCF) and subsequently cluster mutations with similar CCFs into clonal clusters. As PyClone performed better on deeper sequencing data, all somatic mutations were filtered further in order to retain high quality variants only. Only samples with a minimum median VAF of 2% were retained for the analysis, and any samples with no mutations were excluded. Mutations with a minimum coverage of at least 40$\times$ across all related samples were retained.

As recommended by the authors for whole exome sequencing data, PyClone was run using a beta-binomial statistical method. 200,000 iterations were used with a burnin of 100,000 iterations in order to ensure full convergence of the Markov Chain Monte Carlo (MCMC) algorithm. Clonal clusters containing a minimum of 3 mutations were then retained for further analyses. In cases where the CCF of the founding cluster was less than 1, scaling of all clusters within that case was performed such that the upper CCF bound of the founding cluster was 1.

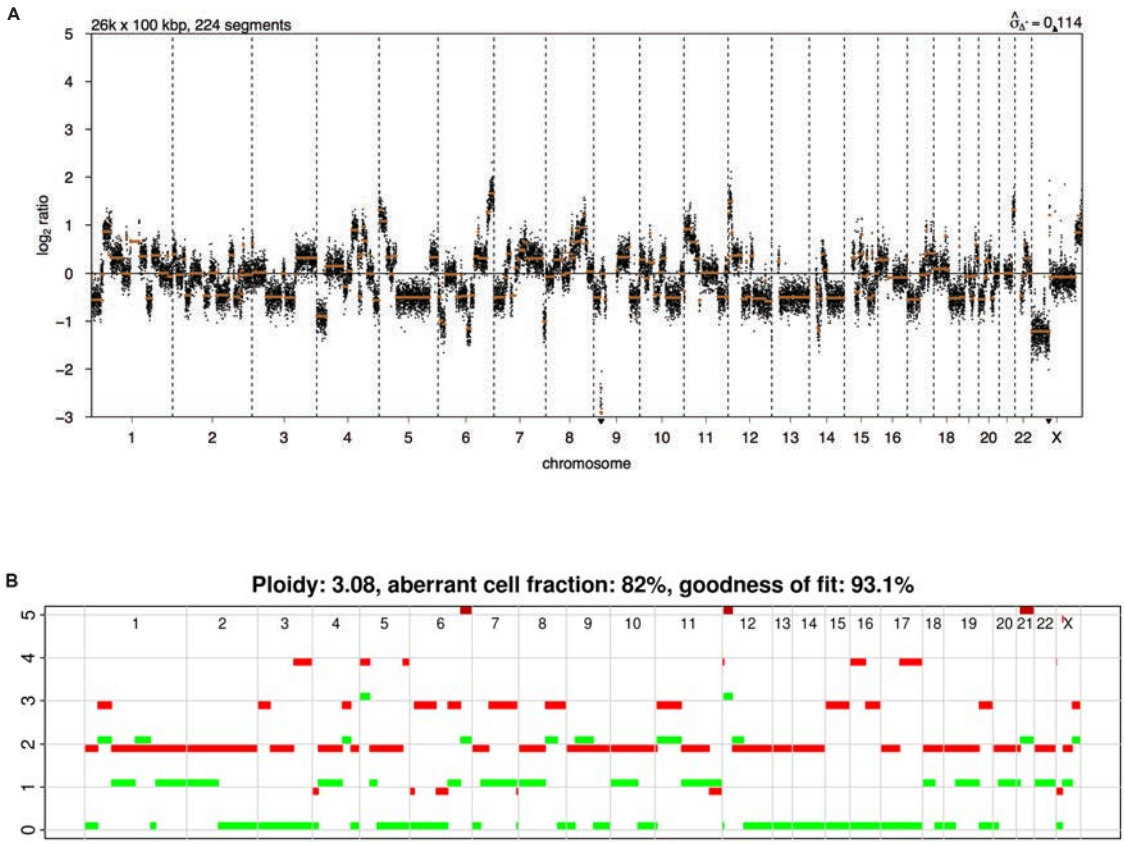


Figure 2.10: Copy number calling for case T004, post-chemotherapy tumour. (A) Copy number segmentation and log ratio estimation using `QDNAseq`. (B) Corresponding parental copy number calling using `ASCAT`

Phylogenetic tree inference was performed on the mutational clonal clusters inferred by PyClone by using LICHeE [228], which took as input the raw mutational clustering data and constructed evolutionary constraint networks to identify trees that satisfied phylogenetic constraints.

2.5.2 RNA analysis

Alignment and preprocessing

FASTQ files for each sample generated from multiple sequencing lanes were merged and aligned using STAR version 2.5.2b [70], using an index generated from the GRCh37 decoy assembly of the human genome previously described and a transcriptomic Gene Transfer Format (GTF) guide obtained from Ensembl Release 87. STAR was run in 'two-pass' mode for sensitive novel junction discovery, wherein the first pass performed a 'default' mapping, and the second pass used the splice junctions detected in the first pass to perform a further round of alignment enhancement. This STAR BAM file was subsequently used for the purposes of differential expression and counting.

In order to allow for variant calling, the BAM files generated by STAR were subsequently processed as per GATK best practices guidelines for RNA-seq variant calling. Firstly, PCR and optical duplicates were marked using Picard MarkDuplicates. Following this, the GATK tool SplitNCigarReads was used to split reads having N CIGAR elements in separate sequence reads. Local realignment around insertions and deletions was performed using RealignerTargetCreator and IndelRealigner, using a calibration set derived from the 1000 Genomes project, as described previously [65, 294, 198]. Base quality recalibration across all variant sites was then performed using BaseRecalibrator. This GATK BAM file was subsequently used for variant calling.

Gene and transcript abundance estimation

Gene counting was performed on the STAR aligned BAM file using HTSeq [14] in read strand-aware mode as the library preparation kit used retained strand information. Additionally, gene counting was performed in 'union' overlap resolution mode, where a read would only be assigned to a gene if it only overlapped within an exonic region of one gene, rather than multiple genes. As HTSeq provided absolute counts per gene within the GTF file, transcript

quantification was performed using `Salmon` version 0.9.1 [223] using default settings. The median number of reads per sample was 85.9 million reads (range: 3-220 million).

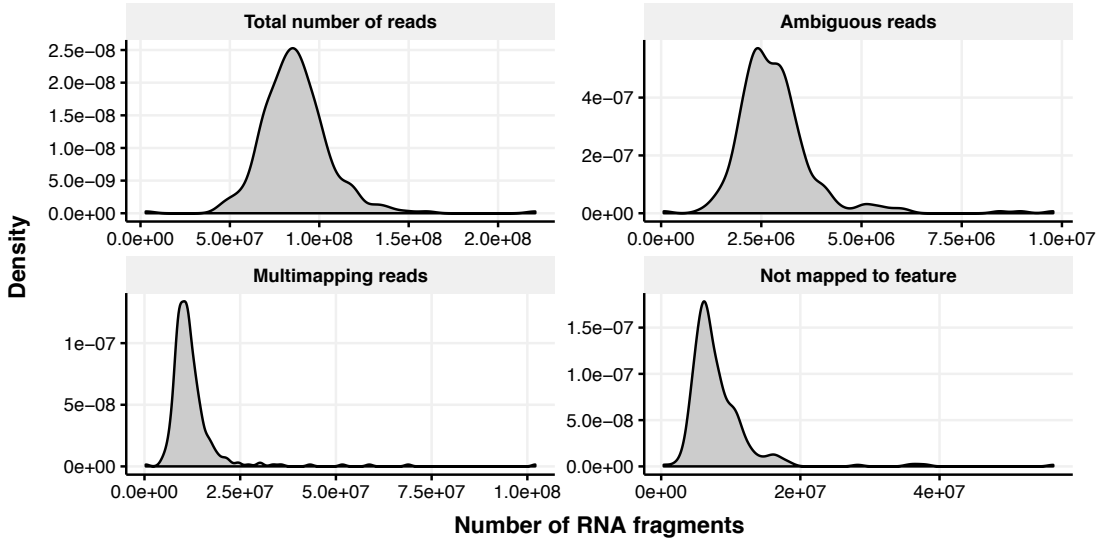


Figure 2.11: RNA sequencing metrics, showing distribution of counts across the dataset.

Counts across 57,906 regions present within the Ensembl 87 GTF in all samples were merged into one counts matrix using R, and a trimmed mean of M-values (TMM) normalization performed across all samples using the `edgeR` R package in order to correct for composition biases and make the transcript counts comparable across all samples [190, 236]. The library normalised counts were then transformed into Fragment Per Kilobase Millions (FPKMs) by using the equation:

$$FPKM_i = \frac{X_i}{\frac{l_i}{10^3} \times \frac{N}{10^6}} = \frac{X_i}{l_i N} \times 10^9 \tag{2.3}$$

where X was the number of counts falling within a gene, l was the sum of the lengths of all exons within a gene and N was the total number of reads sequenced. Effectively FPKM transformation accounted for the length of the gene and made inter-gene comparisons possible [202]. FPKM values were then scaled to a total of a million counts, changing the unit of measure to Transcripts per Million (TPM) [169]:

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j}\right) \times 10^6 \tag{2.4}$$

Differential expression

In order to determine which set of genes were highly or lowly expressed given a set of experimental conditions (such as pCR vs RD), differential expression was performed on the gene raw counts data obtained as described above, using the `edgeR` R package [190, 236], which modelled the dispersion of digital counts using a negative binomial model. Transcript count normalisation was done using the TMM method described previously, and the statistical method was run using various variations of design matrices depending on the experimental set-up discussed: details of the linear model used for each experiment is provided in the text describing each experimental variation. The output of each model was a list of differentially expressed genes, as well as an expression-ranked list of genes.

Gene set enrichment analyses (GSEA)

Following the generation of a ranked list of differentially expressed genes for any comparison of interest, gene set enrichment was performed using the `camera` statistical method in `edgeR`: in brief, this method performed a competitive gene set test accounting for inter-gene correlation and tested whether genes were highly ranked relative to other genes in terms of differential expression [311]. As input to this GSEA method, the annotated gene sets provided within the Molecular Signatures Database (MSigDB) version 6.1 were used [171, 274]. In addition, further enrichment over the Reactome database [78] was performed using the `ReactomePA` R package [317].

Immune microenvironment deconvolution from bulk RNA-seq data

Various methodologies have been employed to quantify cell populations from bulk RNA-seq data, with most methods relying on ssGSEA or non-negative matrix factorisation to deconstruct the transcriptomic data. The two key algorithms used in this work to deconvolute various cell populations were `MCPcounter` [26], which was able to enrich for 10 cell types, and `CIBERSORT` [208], which enriched for 22 cell types. The input to each method was a matrix of normalised gene expression counts.

iC10 Classification

Classification of all tumours into one of the ten iC10 clusters [60, 63] was performed using the `iC10` R package [12] (Figure 3.9C), which took, as input, (1) cellularity corrected copy number log ratios (obtained by running `QDNAseq` on the shallow whole genome sequencing files), as well as (2) `voom` normalised gene expression counts.

Quality assessment and sample genotyping

Germline variants identified on exome sequencing were filtered by removing multi-allelic variants, indels, as well as mutations for which the minimum depth was less than 30× across all samples. The remaining 14,468 germline variants were subsequently genotyped across all RNA samples and comparisons done across homozygous germline variants only. As shown in Figure 2.12, the percentage median concordance across samples derived from a matched patient was 100%, whereas unrelated samples had a median concordance of 60%.

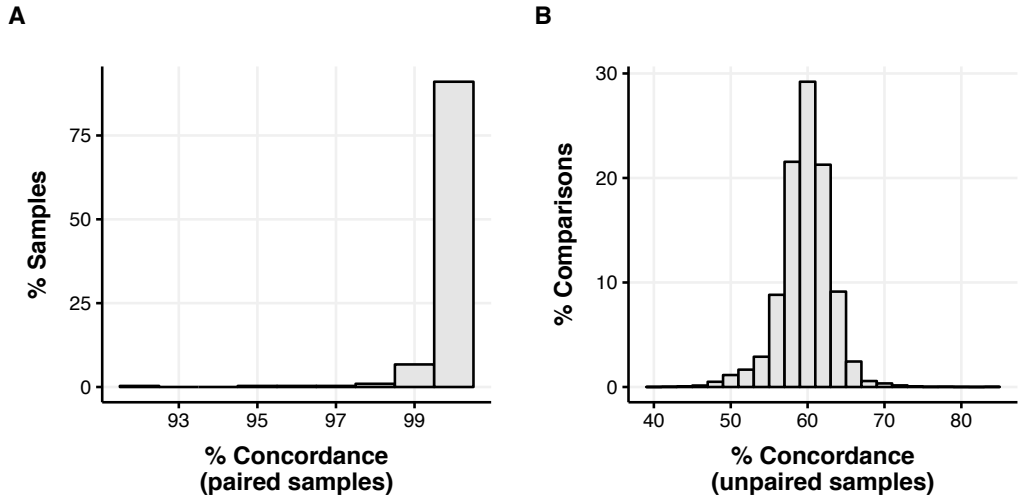


Figure 2.12: Concordance between DNA and RNA SNP mutations. Samples derived from the same patient genotyped with > 99% concordance, compared to unrelated samples for which the concordance rate was 60%

In order to ensure that the clinical ER and HER2 status matched with the RNA-seq expression data, the log distributions of TMM normalised TPMs for *ESR1* and *ERBB2* were modelled using mixed-effect Gaussian models via the R package `MClust` (Figure 2.13). For *ESR1* a mixture of two Gaussian distributions was observed to best model the data, compatible with an *ESR1* low (i.e. ER^-) and *ESR1* high (i.e. ER^+) model, with a log TPM cut-off of 1.7.

Correlation with the clinical data showed a tight correlation between *ESR1* expression on RNA-seq and ER status by IHC (Figure 2.13 E). Most ER^- tumours did indeed have an *ESR1* expression of less than 1.7 TPM, whilst most ER^+ tumours had an expression of more than 1.7 TPM, with most weak positives (i.e. Allred 3-5) falling at the boundary of these two classifications. Interestingly, 4 ER^- tumours (T056, T065, T076, T118) on IHC had *ESR1* expression higher than 1.7 TPM whilst three ER^+ tumours (T002, T007, T073) had ESR1 expressions of less than 1.7 TPM. The discrepancy could be explained by the fact that these two methods of quantification are different: IHC quantifies protein load whilst expression quantifies RNA abundance. None of these cases had mutations within the *ESR1* gene that would result in a non-functional protein.

A similar approach was used to model *ERBB2* expression. This time, however, the model best suited to characterise the distribution comprised a mixture of four Gaussian distributions (Figure 2.13C), with log TPM cut-offs of 3.83, 5.11 and 6.52. Once again, this classifier correlated with the HER2 IHC/FISH status, where strong positive tumours had an *ERBB2* expression greater than 6.52, negative tumours had an expression less than 3.83, and the intermediary subgroup had expressions between 3.83 and 6.52 (Figure 2.13F).

Variant calling

Somatic variants detected on exome sequencing were genotyped in the RNA GATK BAM by using `HaplotypeCaller` in `GENOTYPE_GIVEN_ALLELES` mode. Mutations present in all samples for one patient were concatenated together, and a VCF generated to guide `HaplotypeCaller` local reassembly and variant calling. 60% of coding exonic mutations detected in all samples on exome sequencing were also identified within the RNA, and the variant allelic fraction between both had a correlation of 0.58 (Figure 2.14).

2.5.3 Statistical testing

All statistical tests in this work have been conducted using R version 3.5.0. Multiple test correction was applied whenever more than 20 comparisons were performed, using the Benjamini–Hochberg procedure [29]. All statistical tests were two-sided unless otherwise specified in the text.

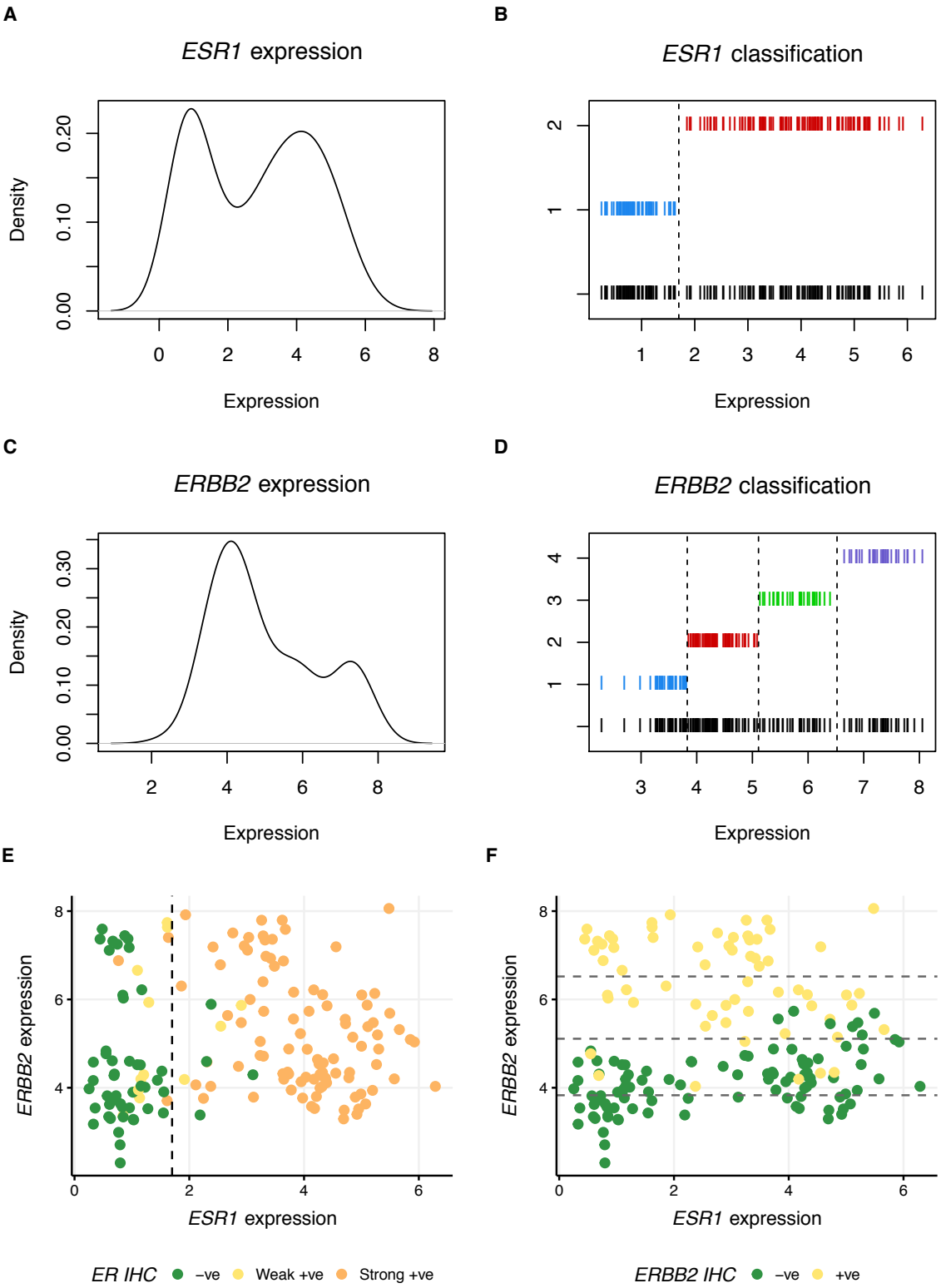


Figure 2.13: Distribution of *ESR1* and *ERBB2* expression in diagnostic samples and correlation with IHC. (A, B) *ESR1* expression was bimodal, with clear ER^- and ER^+ categories (blue and red respectively in B) that also correlated with ER IHC (E). (C, D) *ERBB2* expression was modelled as a mixture of four Gaussian distributions (low (blue), intermediate (red and green) and high (purple)), and also correlated with clinical phenotype (F).

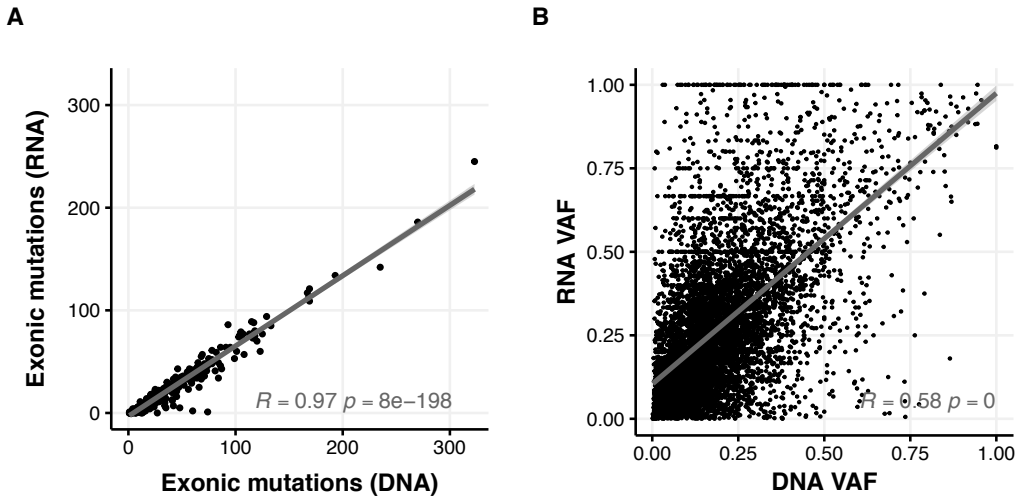


Figure 2.14: Relationship between somatic variants identified in DNA, and those identified in corresponding RNA samples. (A) Scatterplot showing number of mutations detected in both DNA and RNA sequencing data per sample. 60% of mutations found in coding regions by exome sequencing were also detected on RNA-seq, with high concordance. (B) VAF of all somatic coding mutations detected on exome sequencing and corresponding VAF seen on RNA sequencing. Note the significantly higher expression of a cohort of somatic mutations (RNA VAF of 100%)

The following convention for symbols indicating statistical significance was used in this work:

- ns: $p > 0.05$
- *: $p \leq 0.05$
- **: $p \leq 0.01$
- ***: $p \leq 0.001$
- ****: $p \leq 0.0001$

Whenever comparing variables associated with response, two main statistical comparisons were made:

1. A comparison across all RCB categories to pCR, using either **Wilcoxon rank sum tests** or **logistic regression models**. Therefore each of the RCB groups was compared to pCR individually.
2. Comparisons using ordered RCB categories, where the change of a variable was modelled across increments (or decrements) in the degree of response (i.e. pCR $>$ RCB-I $>$ RCB-II $>$ RCB-III) using **ordinal logistic regression models**.

2.6 Dataset summary

Figure 2.15 shows a summary of all samples that were exome/shallow whole genome and RNA-sequenced in this dataset.

In total, 502 samples were exome sequenced, with 168 cases having germline exome sequencing data (and 5 having multiple germline samples sequenced). A total of 329 tumour samples were DNA sequenced. 29 cases were found not to contain any tumour cells on histological analysis and were therefore removed from the analysis but used to fine tune the exome variant pipeline as described previously. This left a total of 168 cases with a tumour biopsy pre-therapy, 75 cases with a midway biopsy, 51 cases with a post-therapy biopsy, and 6 relapse tissue samples from 5 patients. Additionally, in 14 cases that did not have fresh tissue at the time of surgery, FFPE blocks were retrieved and shallow whole genome sequencing libraries generated.

Out of the 329 tumour samples that were RNA sequenced, only 314 generated usable sequencing data. This was due to (1) poor RNA extraction yield and (2) poor RIN, resulting in failure of library preparation. 24 of these cases were not taken from the tumour site and were therefore removed from the analysis but used as a source of matched normal tissue. This left a total of 163 cases with a tumour biopsy pre-therapy, 72 cases with a midway biopsy, 49 cases with a post-therapy biopsy, and 6 relapse tissue samples from 5 patients.

Circulating tumour DNA was extracted from 305 plasma samples from the first 80 recruited patients. Shallow whole genome sequencing data was generated for 96 of these samples (24 cases with four sequential time-points), whilst deep targeted sequencing data was generated for 4 cases (15 plasma samples).

Deep sequencing was also performed on the tumour tissue in four cases to validate the mutation calls as well as the clonal phylogenies observed. In one case, discussed in Chapter 4, multi-region sequencing data generated from deep sequencing of cores taken from FFPE blocks.

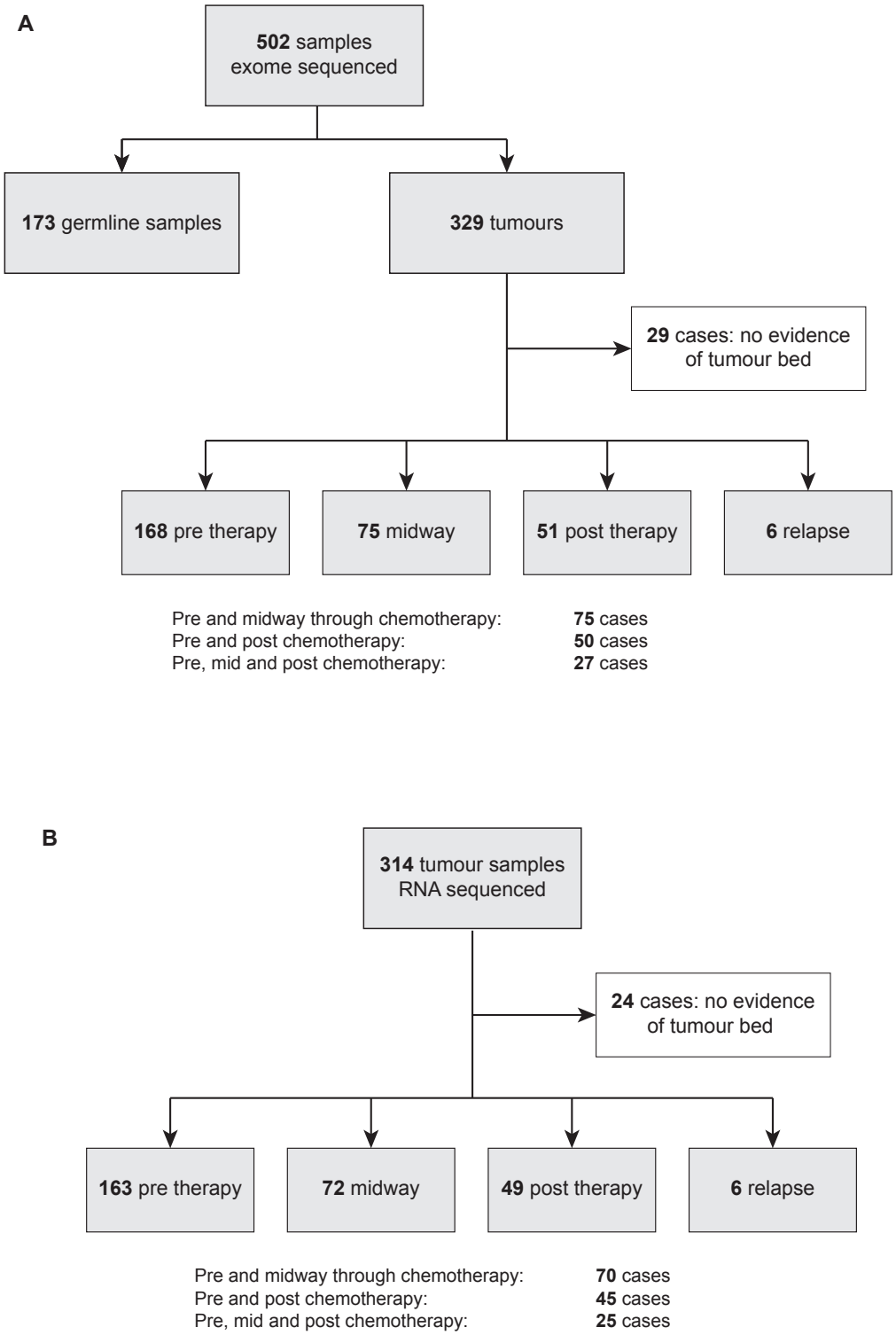


Figure 2.15: Summary of all samples with exome and RNA sequencing data used in this work

Chapter 3

Molecular predictors of response to neoadjuvant chemotherapy

Contents

3.1 Introduction

The neoadjuvant setting provides an opportunity to study tumour biology *in vivo* and identify mechanisms of resistance and response to therapy. Indeed, many neoadjuvant studies have been conducted to identify biomarkers of response to chemotherapy, with genomic and transcriptomic data being made publicly available for most datasets. Efforts have been made to integrate these datasets and use the power of larger numbers to attain closer approximations of the ground truth. However, these integrative studies have been fraught with limitations. Notably, these included: (1) merging studies with very different treatment regimens and unequal lengths of chemotherapy exposure, (2) absence of trastuzumab treatment in most $HER2^+$ patients, (3) differing source tissues analysed, with data generated from a mixture of high quality fresh tissue and lower quality FFPE tissue (4) differing methods of tumour sampling (eg needle versus core biopsies, with the former capturing tumour but not microenvironment) and (5) differing technologies and analytical methods used to generate the individual datasets, which unavoidably resulted in biases due to batch effects.

This chapter aims to describe predictors of response to chemotherapy at diagnosis in the TransNEO cohort by using high depth exome sequencing and transcriptomic profiling. The key strengths of this dataset which allowed this type of analysis included: (1) treatment with standard of care chemotherapy regimens, with strict adherence to local protocols and NICE guidance, (2) high quality clinical metadata which captured, amongst other variables, the degree of response to chemotherapy as assessed by an expert in neoadjuvant breast pathology allowing accurate assessment of biomarkers associated with response, (3) data generation using standardised state of the art sequencing technologies and analysis using a standardised set of bioinformatics pipelines to limit batch effect and generate a robust and accurate dataset.

In the first half of the chapter, genomic predictors of response to chemotherapy were assessed, including (1) total somatic mutation burden, (2) the mutational landscape, (3) mutational signatures, (4) intra-tumoural heterogeneity, (5) copy number alterations and instability, (6) HLA genotypes and (7) neoantigen load.

In the second half of the chapter, transcriptomic associations with response were investigated, with an emphasis on (1) identifying differentially expressed genes and pathways that confer increasing probability of response to neoadjuvant treatment, (2) the identification of potential mechanisms of chemoresistance and (3) the interplay between proliferation and immune activation. Finally, two published metagenes were evaluated and a novel metagene predictive of response was established using machine learning methods.

3.2 Clinical phenotypes and association with response

Phenotypes associated with response to neoadjuvant chemotherapy are well described in the literature, with features predictive of pathological complete response including ER^- status, *ERBB2* amplification, high histological grade, lymph node positive disease, lymphovascular invasion, lower body mass index and younger age, all of which are associated with a more aggressive phenotype [84, 85, 91, 136, 172, 243].

To identify clinical variables associated with response in the TransNEO cohort, the effect of lymph node status, histology, ER and HER2 status, grade, tumour size and age on response was estimated using logistic regression models. Three different response groups were defined using published definitions [117]:

1. **pCR**: comprising tumours that attained pCR, as opposed to all other tumours with remaining residual disease (RD)
2. **Chemosensitive tumours**: comprising tumours that attained pCR or RCB-I RD on completion of chemotherapy
3. **Chemoresistant tumours**: comprising tumours with RCB-III RD on completion of chemotherapy

A simple univariable logistic regression model showed that higher grade, ER^- status, younger age and lymph node positive disease were positively associated with pCR. In addition to these four variables, $HER2^+$ status and invasive ductal histology were associated with increased chemosensitivity (Figure 3.1A). Conversely, ER^+ status, low grade, $HER2^-$ status, non-invasive ductal histology, and increasing age were associated with chemoresistance, in keeping with the low proliferation ER^+ luminal A phenotype seen in older women. The clinical phenotypes corresponding with response were identical to those reported within the literature, indicating that the recruited cohort was similar to the patient population studied in other trials and was not biased towards a particular subgroup.

To correct for related phenotypic effects (such as that often seen between ER^- tumours and higher grade), a multiple logistic regression model was constructed that modelled the effects of all the above-mentioned clinical phenotypes into one linear model (Figure 3.1A). The association with pCR was maintained in ER^- tumours and younger age at presentation, with these two variables and $HER2^+$ status strongly associated with chemosensitivity. Lower grade was the only variable associated with chemoresistance and was in keeping with the

strong body of published evidence that has found a positive correlation between increased grade and response to chemotherapy [84, 172, 243].

So as to assess the performance of the multiple logistic regression model built on seven clinical variables, a Receiver Operating Characteristic (ROC) analysis was performed using the three different response groups as response variables (Figure 3.1B and C). The clinical model was able to predict pCR with an AUC of 0.82, chemosensitivity with an AUC of 0.84 and chemoresistance with an AUC of 0.92. Indeed, the clinical classifier was able to predict pCR with a reasonably high NPV (93.5%, CI: 88.4-97.7%), but lower PPV (53.6%, CI: 45.3-66.0%), with a specificity of 74.1% (CI: 64.2-85.7%) and a sensitivity of 84.6% (CI: 71.8-94.9), showing that the clinical features oncologists use to predict sensitivity to chemotherapy are robust.

These analyses confirmed that the patient cohort recruited to the TransNEO trial was very similar to that seen routinely in neoadjuvant clinics, with concordance to the published phenotypes associated with varying degrees of response. The statistics mentioned in this section should serve as a baseline to which the molecular metagenes predictive of response described later on in this work can be compared to.

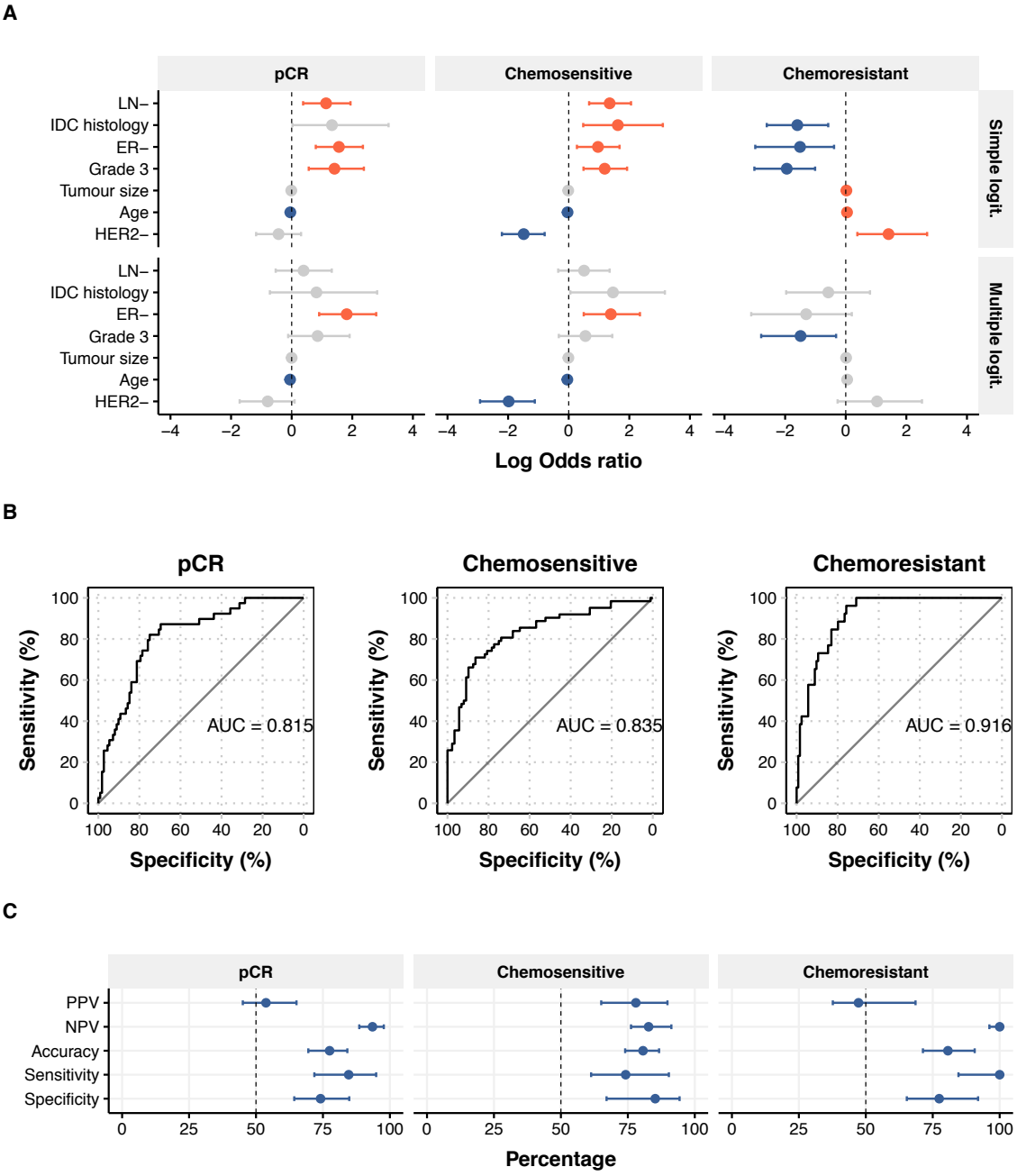


Figure 3.1: Associations between clinical phenotypes and response. (A) Clinical predictors of response on simple and multiple logistic regression, red indicates significantly increased log odds ratio, blue indicates significantly decreased log odds ratio. (B) ROC curve analysis for the ability of the multiple logistic models shown in (A) to accurately model response. (C) Sensitivity, specificity, accuracy, NPV and PPV of the multiple logistic clinical model to predict varying degrees of response.

3.3 Genomic predictors of response

3.3.1 Mutational landscape

Across the 168 patients sequenced within the cohort, a total of 12,662 somatic mutations were identified in the diagnostic tumour exome sequencing data using the bioinformatic pipeline discussed in Chapter 2. These included a total of 6,566 missense mutations, 2,383 silent mutations, 486 frameshift insertions and deletions, 476 nonsense mutations, 158 splice site mutations, 101 inframe insertions and deletions, 7 nonstop mutations and an additional 2,485 non-coding mutations falling within 3'/5' flanks, 3'/5' UTRs, intergenic regions and introns.

The median diagnostic sample mutation rate per megabase of region sequenced was 1.34 Mb^{-1} (range: 0.02-9.67 Mb^{-1}, Figure 3.2A). This was higher than reported by the TCGA (1 Mb^{-1}), however the overall distribution of mutational burden across the TransNEO cohort was not statistically significantly different from that in the TCGA (p=0.12, Kolmogorov–Smirnov test, Figure 3.2B). $ER^{-}HER2^{-}$ tumours had a higher mutation rate than $ER^{+}HER2^{-}$ tumours (1.70 vs. 0.96 Mb^{-1}, p=0.001 Wilcoxon rank sum test), though there was no significant difference in mutation burden between $ER^{+}HER2^{-}$ and $HER2^{+}$ tumours (0.96 vs. 1.4 Mb^{-1}, p=0.086) and between $ER^{-}HER2^{-}$ and $HER2^{+}$ tumours (1.7 vs. 1.4 Mb^{-1}, p=0.127).

The total mutation burden (TMB), defined as the total number of mutations per megabase (Mb) of region sequenced correlated with response to chemotherapy (Figure 3.3A): tumours that attained pCR had significantly more mutations than those with RD post-chemotherapy (2.1 vs. 1.5 Mb^{-1}, p= 0.0002 Wilcoxon rank sum test). Tumours with higher RCB scores had lower mutational burdens at diagnosis than those with less RD (p=0.004, ordinal logistic regression). On dividing the cohort by ER and HER2 status, the strongest association between TMB and response was seen in $ER^{+}HER2^{-}$ tumours (p=0.0007 Wilcoxon rank sum test), with a non-statistically significant trend in $ER^{-}HER2^{-}$ cases (p=0.054, Wilcoxon rank sum test, Figure 3.3B). There was no association between response and TMB in $HER2^{+}$ tumours ($ER^{+}HER2^{+}$ and $ER^{-}HER2^{+}$ analysed independently and when grouped) in this cohort (p=0.76), though this was probably due to the fact that most $HER2^{+}$ cases (63.9%, Table 2.3) attained pCR or very minimal RD (RCB-I) after neoadjuvant therapy.

Not all mutations fall within coding regions or are eventually transcribed into RNA molecules. In view of this, the expressed TMB was computed by retaining mutations identified on

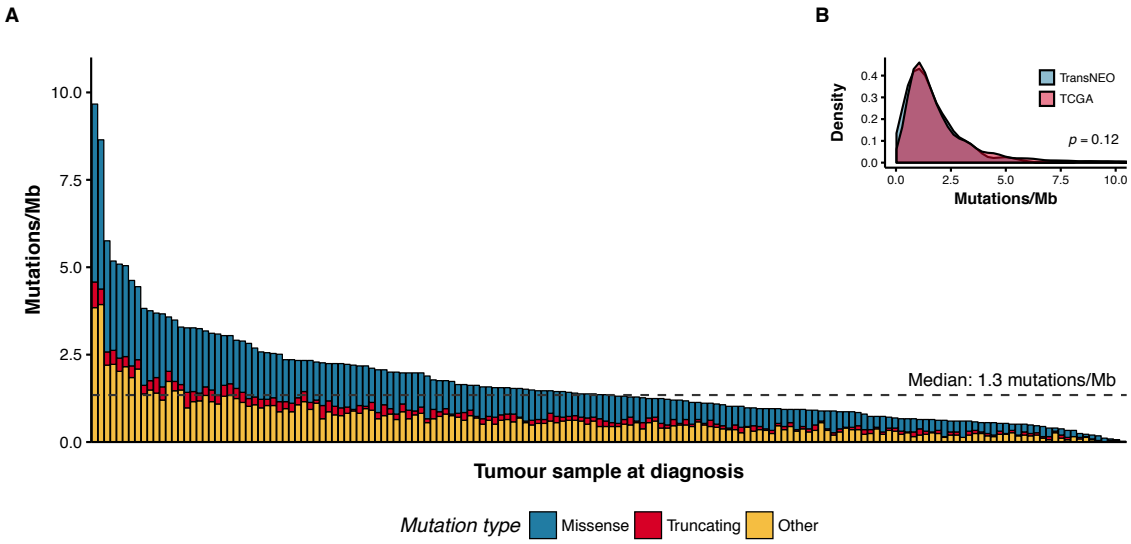


Figure 3.2: Distribution of tumour somatic mutation burden per megabase across all 168 cases in the TransNEO cohort. (A) The median mutation burden was 1.3 mutations Mb^{-1}, with missense mutations predominating the somatic landscape. (B) Distribution of tumour mutation burden across the TransNEO (blue) and the TCGA (red) studies.

exome sequencing that were also present in the RNA sequencing data (Section 2.5.2). The association between expressed TMB load and response was also significant (Figure 3.3D): tumours that attained pCR had a higher number of expressed mutations than those with remaining RD (0.73 vs. 0.47 expressed mutations Mb^{-1}, p=1.18e-05, Wilcoxon rank sum test), with this association being statistically significant in $ER^{+}HER2^{-}$ (0.93 vs. 0.29 Mb^{-1}, p=0.0003) and $ER^{-}HER2^{-}$ (0.71 vs. 0.47 Mb^{-1}, p=0.03) tumours, but not $HER2^{+}$ tumours (Figure 3.3D).

As TMB was shown to associate not only with pCR but also with the degree of RD after treatment, an ordered ordinal regression model was constructed using the `polr` function from the `MASS` R package [298] to linearly predict the probability of response to chemotherapy based on the diagnostic TMB (Figure 3.3E). As expected, the resulting prediction model consistently predicted a higher probability of attaining pCR as TMB increased, with a TMB cut-off of 3 mutations Mb^{-1} determined to be the point at which the likelihood of attaining pCR post-chemotherapy was higher than the likelihood of having remaining RD. At TMBs less than 3 Mb^{-1} it was more likely for any degree of RD to be present after neoadjuvant chemotherapy.

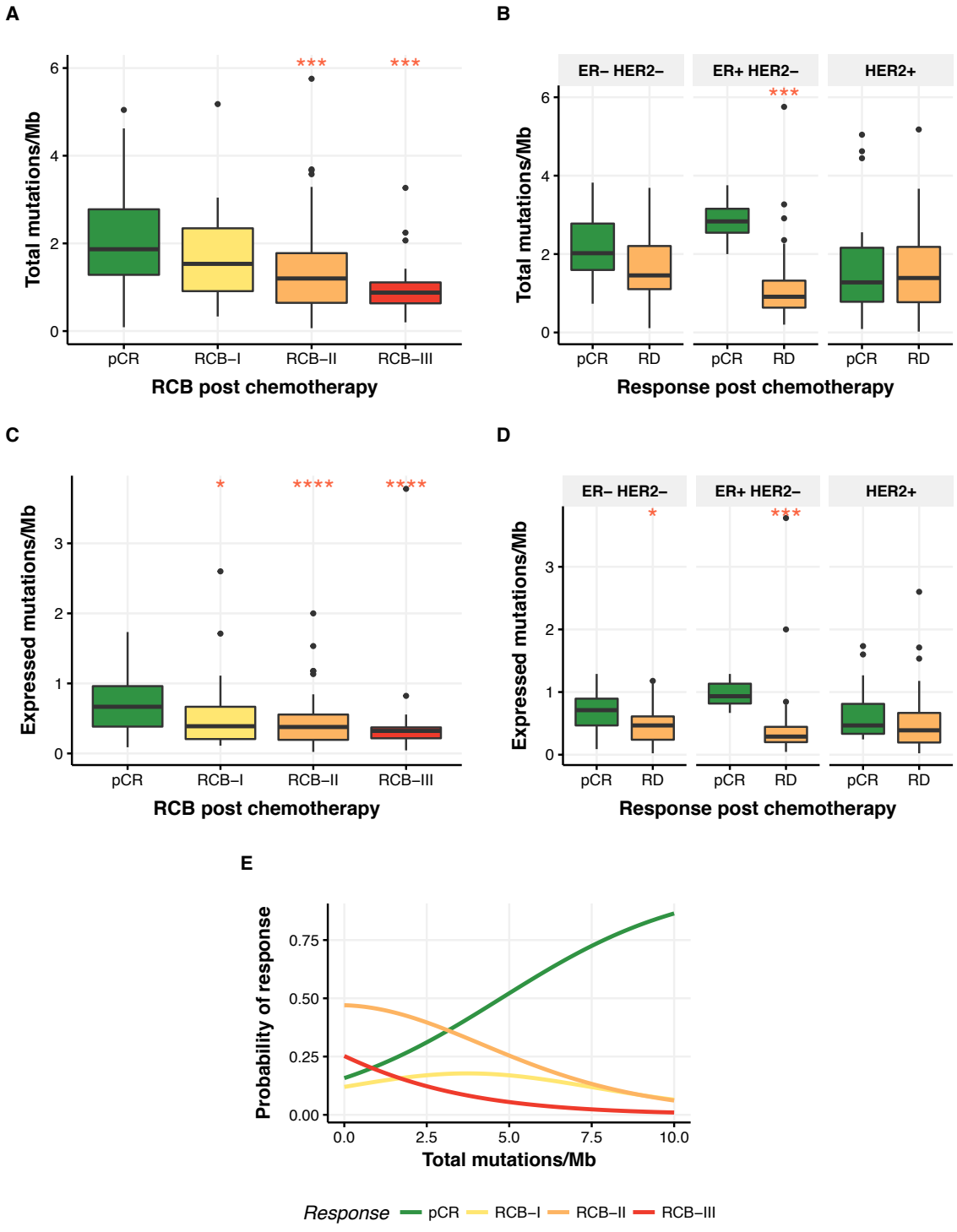


Figure 3.3: Association of diagnostic TMB with response. (A) TMB was associated with response to neoadjuvant chemotherapy: tumours that attained pCR had more mutations than those with remaining RD. (B) Sub-classification of TMB and response across ER^+/ER^- and $HER2^+/HER2^-$ subtypes: ER^+HER2^- tumours showed a strong association between TMB and response: this association was not observed in ER^-HER2^- and $HER2^+$ tumours. (C,D) Expressed TMB was also associated with response to chemotherapy, though this association was not observed in $HER2^+$ tumours. (E) Probability of response as mutation burden increased, as modelled by an ordinal logistic regression model.

Driver gene landscape

The somatic mutation landscape in breast cancer often comprises a few frequently mutated genes in a large number of samples, with a long tail of genes that are infrequently mutated and present in fewer tumour samples [96, 308]. In keeping with this known observation, the most commonly mutated genes within this dataset were driver genes, formally defined as genes in which the presence of a mutation or alteration increased net cell growth [287]. Specifically, the driver genes detected were those frequently implicated in breast cancer [225] and included, amongst others, *TP53*, *PIK3CA*, *GATA3* and *MAP3K1* (Figure 3.4A). Mutations in breast cancer driver genes were identified in 149 cases (89%), and the ability to robustly call a mutation within these genes was not correlated with the degree of tumour purity within the samples (tumour purity 45% in samples with no detectable driver mutation vs. 50% in samples with driver mutations, p=0.06, Welch Two Sample t-test). 67% of all cases had more than one detectable driver gene mutation (Figure 3.5).

TP53, a tumour suppressor gene, was the most commonly mutated breast cancer driver gene observed and a somatic mutation was present in 57% of cases sequenced (Figures 3.4A and 3.5). The prevalence of *TP53* mutations was higher than that seen in the METABRIC and TCGA studies (35.4% and 33% respectively, p=4.29e-07, Fisher's Exact Test with FDR correction, Figure 3.4C) [225]. This is perhaps, unsurprising: *TP53* regulates the control of the G1 checkpoint in the cell cycle and can induce cycle arrest and apoptosis in cases of extensive DNA damage [309], and mutations within this driver gene have been associated with more aggressive disease. Within the METABRIC study, mutations in this driver gene were associated with higher histological grade in both ER^+ (OR=3.3) and ER^- (OR=3.6) tumours, as well as worse outcome in ER^+ (HR=1.6) tumours [225]. As the neoadjuvant setting enriches for more aggressive tumours it is therefore of no surprise that *TP53* mutations would predominate within the cohort. Indeed, the presence of a high number of tumours harbouring somatic *TP53* mutations is, perhaps, a testament to the diagnostic skills of the recruiting oncologists in selecting appropriate patients who would benefit most from neoadjuvant treatment. Of note, case **T156** had a germline heterozgous deletion of *TP53* exons 2 to 6, compatible with a diagnosis of Li Fraumeni Syndrome.

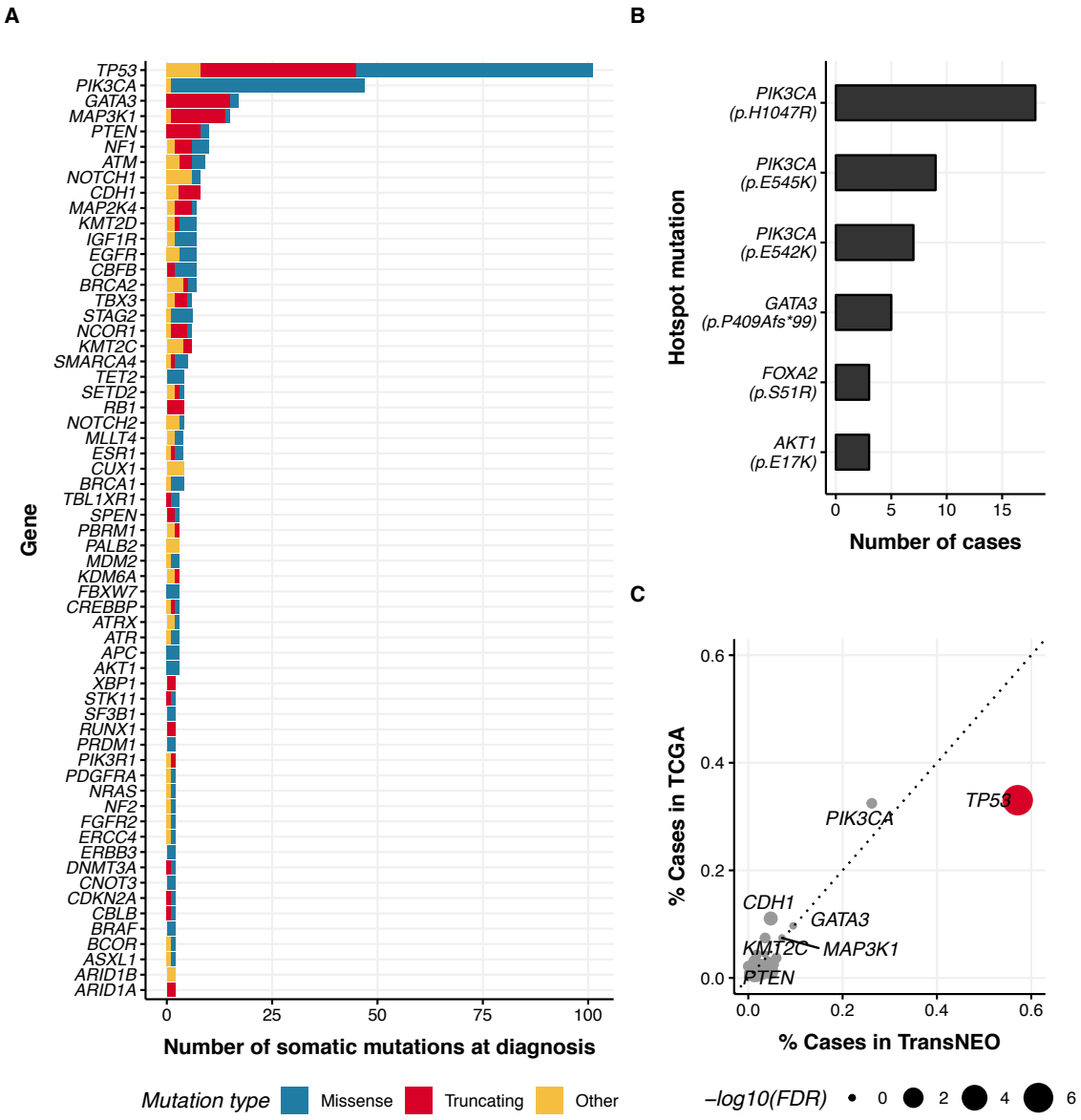


Figure 3.4: Mutations within breast cancer driver genes identified in the dataset. (A) Frequently mutated driver genes (observed in more than 1 tumour). *TP53* was the most frequently mutated gene, followed by *PIK3CA*, *GATA3* and *MAP3K1*. (B) Recurrently mutated sites within previously described hotspot locations in *PIK3CA*, *GATA3*, *FOXA2* and *AKT1*. (C) *TP53* (highlighted in red) mutations were more prevalent within the TransNEO cohort, compared to the TCGA (57% vs. 33%), in keeping with the selection of a more aggressive cohort.

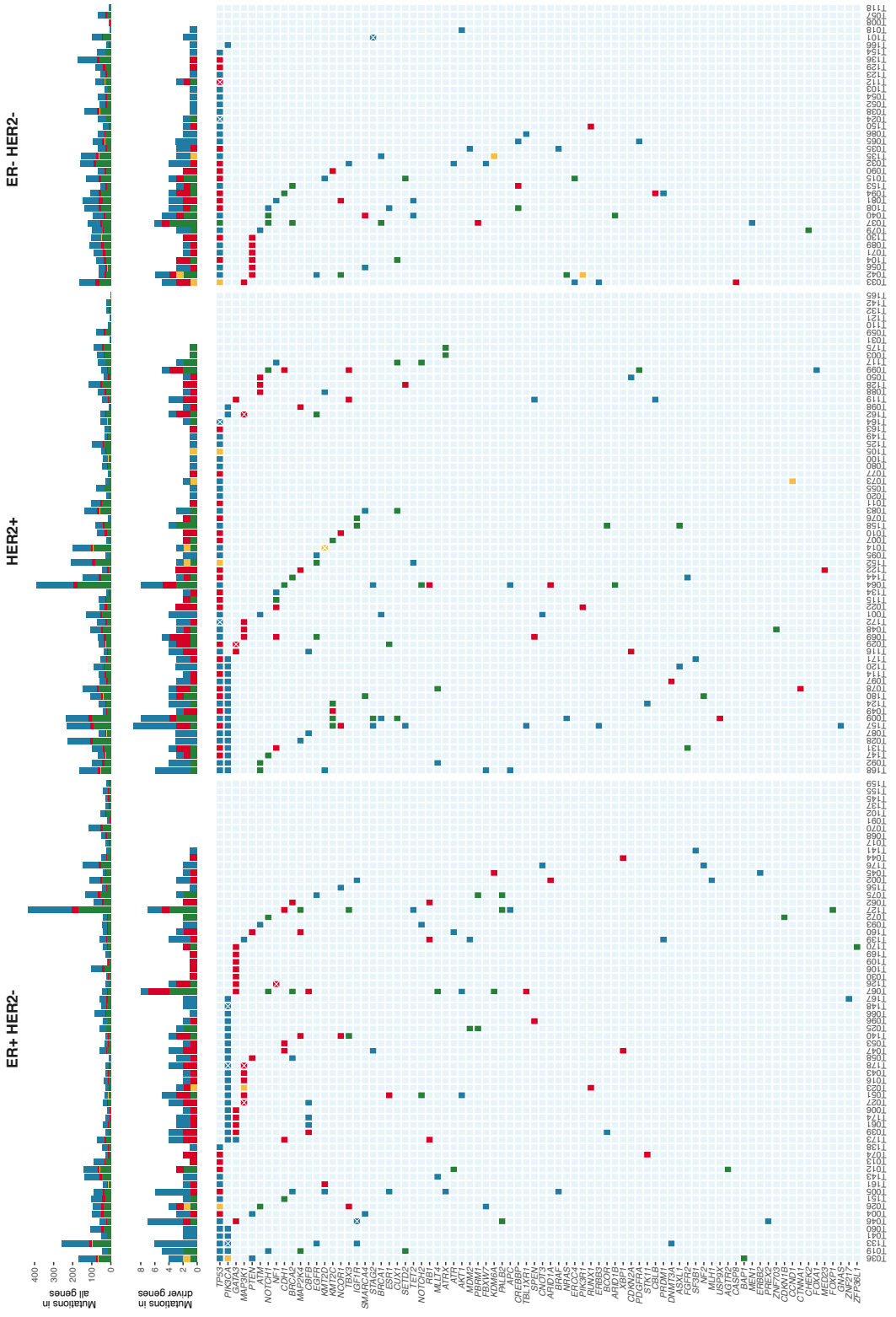


Figure 3.5: Distribution of somatic driver gene mutations across cohort, showing missense mutations (blue), truncating mutations (red), in-frame mutations (yellow) and other mutations (green). Presence of multiple mutations indicated by $\times$.

PIK3CA was the second most commonly mutated driver gene, with 26% of tumours harbouring a mutation within this oncogene. The prevalence of mutations within *PIK3CA* was slightly less than that seen in other studies (32% in the TCGA), though this was not statistically significant (p=0.75, Fisher's Exact Test with FDR correction). As described within the literature, most of the mutations within *PIK3CA* fell in 'hotspot' locations (Figure 3.4B), and included the activating mutations *p.H1047R*, *p.E545K* and *p.E542K* [17]. 80% of all *PIK3CA* mutations were harboured by ER^+ tumours.

Somatic and germline mutations within the tumour suppressor genes *BRCA1* and *BRCA2* somatic mutations were also present within the dataset. Five patients harboured a deleterious BRCA1 mutation or alteration (**T013**: p.V757Ffs*8, **T040**: exon 20 deletion, **T071**: p.Q1777Pfs*74, **T081**: p.X1799_splice, **T082**: exon 1-17 deletion) and 2 patients harboured deleterious germline BRCA2 mutations (**T017**: p.K2162Nfs*5, **T160**: p.X173_splice).

Other frequently mutated gene drivers included the tumour suppressor genes *GATA3*, *MAP3K1* and *PTEN*, which, as expected, often harboured inactivating mutations (Figure 3.4A). The prevalence of all other driver genes within this dataset did not differ when compared to the TCGA and METABRIC.

Mutations associated with response to treatment

Binomial logistic regression models were used to explore associations between all non-silent somatic mutations and response to chemotherapy, correcting for ER and HER2 status. Response was classified into three categories: pCR, chemosensitivity (pCR and RCB-I) and chemoresistance (RCB-III), as described by Hatzis et al. [117].

In the ER/HER2-corrected logistic regression model, non-silent mutations within *TP53* were strongly associated with response (Figure 3.6A). *TP53* mutations were associated with a greater likelihood of pCR (OR: 2.95, CI: 1.36-6.86, p=0.008) and chemosensitivity (OR: 3.58, CI: 1.80-7.38, p=0.0004), and a lesser likelihood of chemoresistance (OR: 0.21, CI 0.08-0.53, p=0.001). As mutations within this tumour suppressor gene resulted in more aggressive and rapidly proliferating disease, it was unsurprising that cells harbouring these mutations were more rapidly eliminated by cytotoxic chemotherapies. The association between *TP53* mutations and response to chemotherapy has already been reported in two neoadjuvant meta-analyses [52, 307] and it is reassuring that the same observation also holds true in this dataset.

Non-silent mutations within *PIK3CA* exhibited an opposite effect in the ER/HER2-corrected model and were more likely to be associated with increased chemoresistance (OR: 2.93, CI: 1.21-7.05, p=0.016) and decreased chemosensitivity (OR: 0.37, CI: 0.16-0.79, p=0.014). The association between *PIK3CA* mutation status and pCR was previously described in a study with 729 patients with breast cancer treated with neoadjuvant intent [319] and in a meta-analysis of five clinical trials with a pooled total of 967 $HER2^+$ patients [175]. However, in both studies, there were no formal associations made with chemoresistance. In colorectal cancer, the presence of *PIK3CA* mutations was associated with resistance to first-line chemotherapy through augmented PI3K/Akt signalling and a subsequent increase in LGR5+ stem cell survival and proliferation [305].

Tumours harbouring *GATA3* mutations also showed decreased chemosensitivity (OR: 0.20, CI: 0.03-0.75, p=0.04) and a non-statistically significant trend for increased chemoresistance, an observation that has not been made in the literature to date. It is worth noting, however, that all *GATA3* mutations were present in ER^+ tumours.

Amongst the non-silent mutations within non-driver genes, *AKAP6*, *KDM5C*, *MYH8* and *ZC3H6* mutations were associated with pCR across the whole cohort (Figure 3.6A). *KDM5C* (Lysine Demethylase 5C) has been attributed to playing a role in chromatin remodelling and transcription regulation and over-expression has been shown to increase proliferation and invasion in gastric cancer via a decrease in *TP53* expression [312]. There were no previously published associations between mutations within *AKAP6*, *MYH8* and *ZC3H6* and response to neoadjuvant chemotherapy.

Mutations within *C3*, *CBFB*, *DNAH7*, *MAP2K4* and *RP1* were associated with chemoresistance: none of these mutations, to date, have been associated with response to neoadjuvant chemotherapy in the literature.

In order to determine specific associations between response and ER/HER2 subtype, a similar analysis was done but applied to the ER^+HER2^-, ER^-HER2^- and $HER2^+$ cohorts individually (Figure 3.6B). There were no mutations predictive of response in the ER^-HER2^- subgroup, though there was a trend for significance for *TP53* mutations and pCR/chemosensitivity. In the $HER2^+$ subgroup, *NF1* mutations were associated with chemoresistance. In ER^+ tumours, mutations within *DNAH8*, which is involved in androgen receptor signalling, *MYO3A*, *SDK1*, *TGFBR3* and *TP53* were all associated with pCR, while non-silent mutations in *ANK2* and *DNAH8* were associated with chemosensitivity (Figure 3.6B).

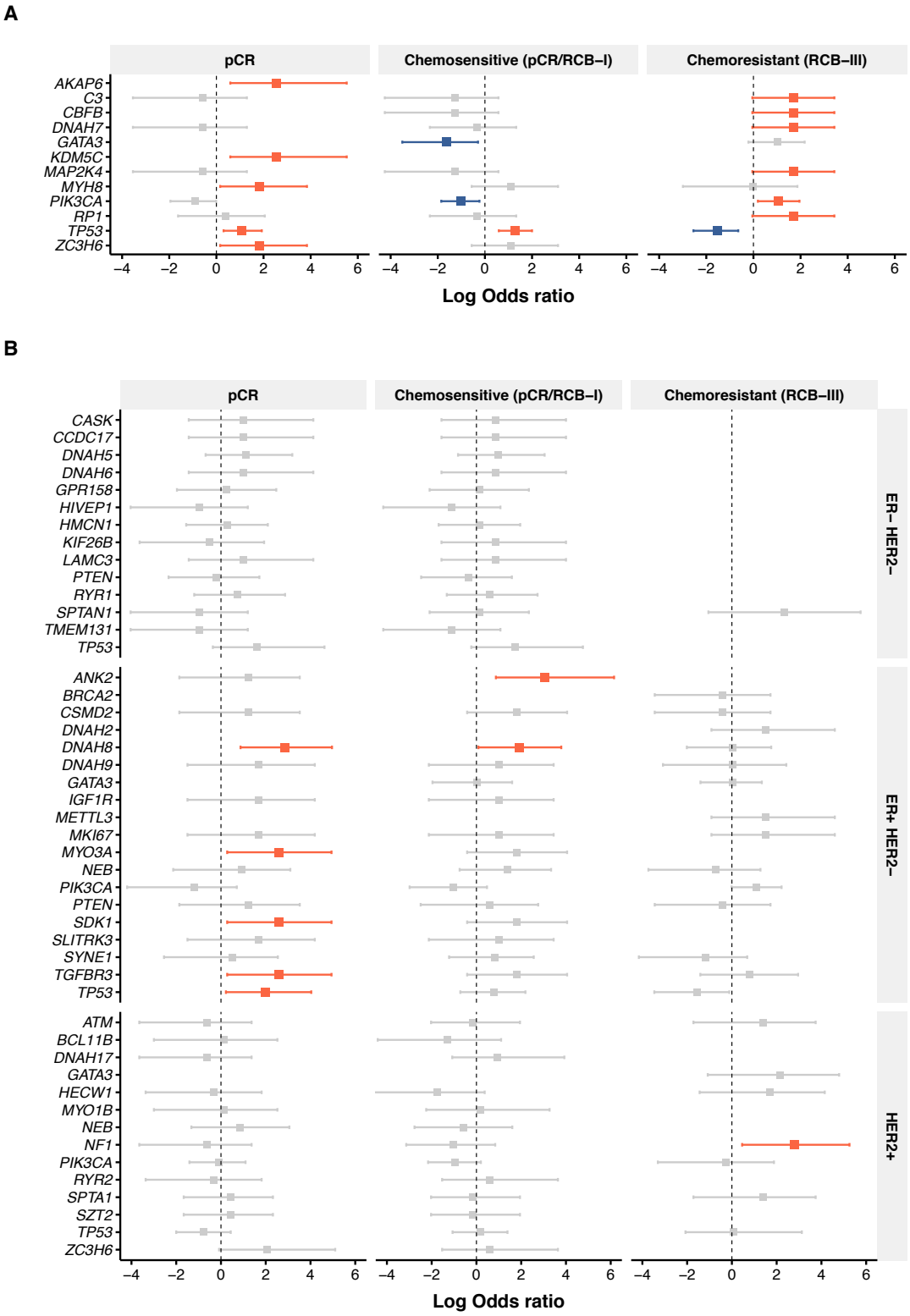


Figure 3.6: Associations between non-silent mutations within commonly mutated genes and pCR, chemosensitivity, and chemoresistance. Red indicates increased log odds, blue indicates decreased log odds. (A) Log odds ratios shown across entire cohort, model corrected for ER and HER2 status. (B) Cohort split by ER and HER2 status.

Mutational signature landscape

The somatic mutational landscape of a tumour is fashioned by processes that induce and maintain mutations within specific trinucleotide sequence contexts [9, 10, 213]. A method involving signature extraction by non-negative matrix factorisation was described in 2013 and more than 20 distinct signatures were defined using a cohort of 4,938,362 mutations derived from 7,042 cancers. These signatures included, amongst others, DNA repair deficiency, ageing, APOBEC activity signatures as well as genomic scars secondary to exposure to genotoxins [9].

In order to determine the mutational signatures landscape present in the dataset, signature decomposition from the bulk exome sequencing mutation data was performed using the `DeconstructSigs` R package [241]. This statistical method uses the Wellcome Trust Sanger Institute Mutational Signature Framework as a reference and determines the linear combination of 30 pre-defined signatures by using a multiple logistic regression model with constraints to reconstruct the mutational profile of each tumour. 160 tumour samples with at least 10 mutations were used for this analysis.

As reported in the original paper describing the mutational signatures, Signature 1, which is the result of mutational process initiated by spontaneous deamination of 5-methylcytosine and correlates with age at cancer diagnosis, was present in nearly 100% of cases at diagnosis (Figure 3.7A). Signature 3 (BRCA associated, failure of DNA double-strand break repair by homologous recombination) was the second most prevalent signature and present in more than 25% of cases. This was closely followed by two APOBEC signatures (Signatures 2 and 13) and a defective DNA repair signature (Signature 15, associated with high numbers of small insertions and deletions at homopolymeric sites). The enrichment of BRCA and APOBEC signatures could explain the higher mutation rate in this cohort when compared to the TCGA, as activation of these processes is known to contribute to an increased mutation rate [235].

An inherent limitation of these mutational signatures lies in the way they are computed. By default, the sum of the contributions of all signatures is constrained to a total of 1 (i.e. 100%), making inter-sample comparisons very challenging as the contribution of one signature is highly dependent on the contribution from all other signatures operating in a tumour. Hence estimates of the effect of each signature on response cannot be defined using this method, as adding one unit causes a relative decrease in some of the other signatures. In order to circumvent this, each signature was normalised to the presence of a signature that was

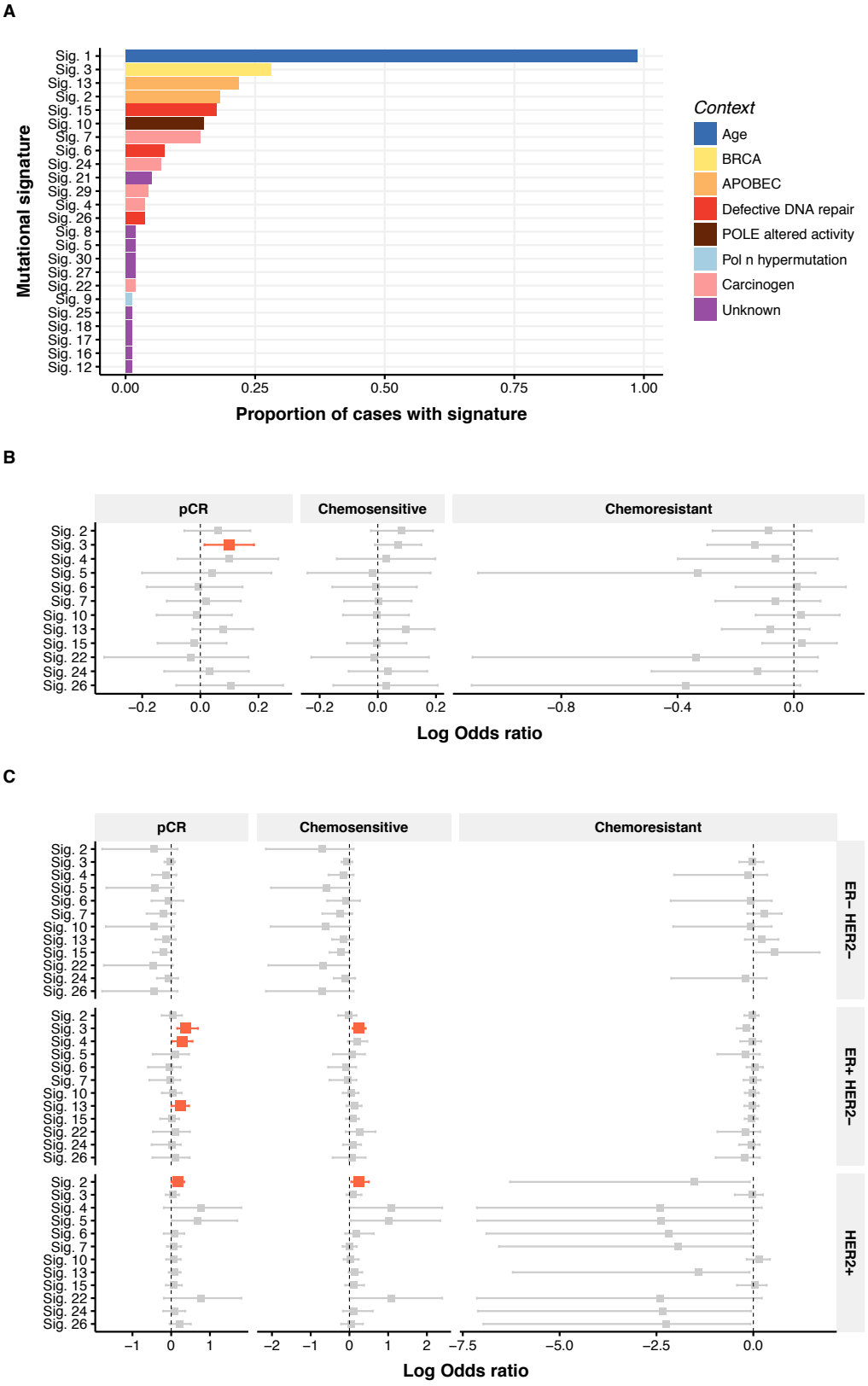


Figure 3.7: Mutational signature landscape within the cohort. (A) Prevalence of mutational signatures, with marked dominance of BRCA and APOBEC signatures. (B) The BRCA signature (Signature 3) was predictive of response across the whole cohort. (C) Association between response and mutational signatures, stratifying by ER and HER2 status.

universally present in nearly all cases, in this case, Signature 1 (Age), as shown in Equation 3.1:

$$normalised\ Sig_N = \frac{log_2\ (Sig_N + 0.001)}{log_2\ (Sig_1 + 0.001)} \tag{3.1}$$

By using this approach, the signature contribution was freed from the 0-1 constraints set upon it by the method that computed it and allowed formal comparisons of normalised signature contributions to overall response to chemotherapy.

The BRCA signature (Signature 3) was the only mutational process that was associated with pCR across the entire cohort (OR: 1.10, CI: 1.10-1.20, p=0.02, Figure 3.7B). On subdividing the cohort by ER and HER2 status, the BRCA association with pCR was maintained in the $ER^{+}HER2^{-}$ cohort, whilst an association between APOBEC activity and pCR was observed in the $ER^{+}HER2^{-}$ and $HER2^{+}$ cohorts. A carcinogenic signature (Signature 4, tobacco), which shows transcriptional strand bias for C>A mutations, was associated with pCR in $ER^{+}HER2^{-}$ tumours.

Intra tumoural heterogeneity

Tumours are genomically heterogeneous secondary to evolutionary pressures and stochasticity, with heterogeneity being a key driver of resistance to both cytotoxic and targeted therapies [61, 193, 215, 276]. Genomic alterations present in all cancer cells (i.e. having a cancer cell fraction (CCF) of 100%) are considered to be truncal as they would have been established within the very early stages of the tumour's evolutionary history and passed on to subsequent descendent cells. Conversely, mutations present in a subset of cancer cells (i.e. with CCFs of less than 100%) are considered to be subclonal and established later on during the tumour's evolutionary history. Indeed clonal mutations have often been likened to the trunk of a tree, with subclonal mutations forming its branches [276, 314].

Various methods have been proposed to compute CCF [15, 66, 169, 191, 242], with most relying on adjusting mutation allelic frequencies to the tumour purity and local copy number alterations. Following the computation of CCFs for each mutation, mixture models with differing density functions allow the inference of mutational clusters [169, 228, 242]. For example, `PyClone` [242] uses beta-binomial distributions while `LICHeE` [228] and `CHAT` [169] use Gaussian distributions.

The clonal composition of all tumours was established by running `PyClone` on the mutational and copy number data (Section 2.5.1). Tumours that had RD following chemotherapy had a higher number of predicted mutational clusters at diagnosis, compared to those that had complete response (2.25 vs. 1.79 mutational clusters, p=0.02, Wilcoxon rank sum test, Figure 3.8A), in keeping with observations that clonally diverse tumours are more likely to harbour a clone with a resistant phenotype, and therefore more likely to have a poorer response to chemotherapy. A similar observation was also made in oesophageal adenocarcinoma treated with neoadjuvant chemotherapy [206].

In order to provide further evidence that heterogeneity was associated with response, the CCF for each mutation was computed using the mathematical framework shown in Equation 3.2, derived by McGranahan et al. [191]:

$$CCF = \frac{VAF}{p} \times \Big((1-p)CN_{normal} + pCN_{tumour} \Big) \tag{3.2}$$

where VAF was the variant allele fraction for each mutation determined by exome sequencing, p was the tumour purity (computed using `ASCAT`), CN_{normal} was the germline copy number state (which was universally 2 (diploid) in this dataset) and CN_{tumour} the total copy number state at the mutant locus in the tumour (computed using `ASCAT`). Point estimates for CCF and confidence intervals were computed using a binomial distribution modelled by the `binconf` function from the `Hmisc` R package and a mutation classified as clonal if the CCF 95% confidence interval overlapped 1, with all other mutations classified as subclonal. As emphasised by McGranahan et al. [191], this was a conservative classification which was dependent on sequencing read depth, in keeping with other methods that assessed subclonality.

This second method of computing CCF revealed that tumours that attained pCR had more clonal mutations compared to those with RD post-chemotherapy (median 76.4 vs. 49.8 mutations, p=0.00001, Wilcoxon rank sum test), however there was no statistically significant difference in the number of subclonal mutations in both response groups (Figure 3.8B). Tumours with RD post-chemotherapy had an overall higher percentage of subclonal mutations than those that attained pCR (27.0% vs. 16.2%, p=0.004 Wilcoxon rank sum test, Figure 3.8C). This indicated that tumours attaining pCR had less clonally complex architectures, with most mutations being clonal, whilst tumours with RD post-chemotherapy had fewer clonal mutations, but more subclonal mutations and therefore more clonal diversity, in keeping with the findings from the `PyClone` analysis described previously.

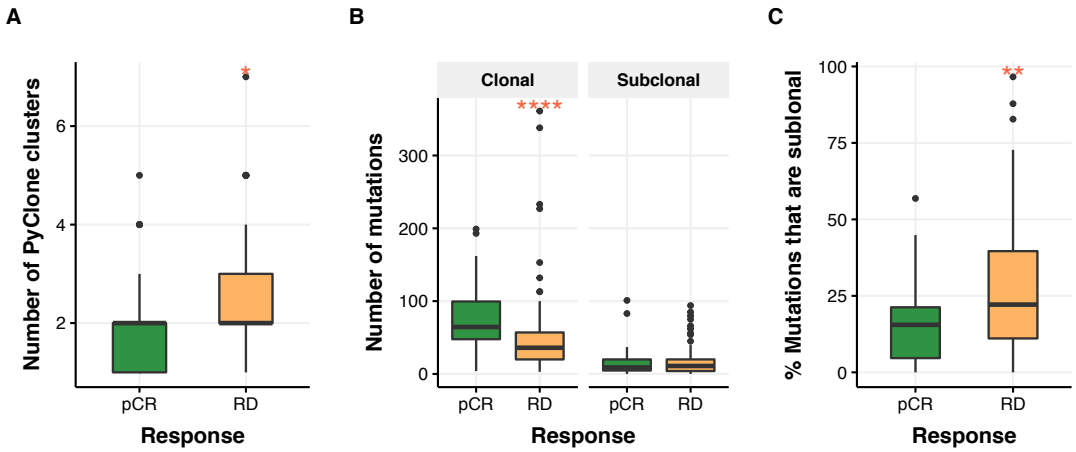


Figure 3.8: Tumour heterogeneity and resistance to treatment. (A) Tumours that attained pCR had fewer `PyClone` predicted mutational clusters compared to those with RD. (B) CCF computation using the method described by McGranahan et al. [191] shows that tumours that attained pCR had a higher number of clonal mutations than those with RD, but no difference in the number of subclonal mutations. (C) Conversely, tumours with RD post-chemotherapy had a larger proportion of subclonal mutations.

3.3.2 Copy number landscape

The iC10 classification

Breast cancer is predominantly driven by copy number rather than mutational events [56, 60]. As observed with mutation load, the degree of chromosomal instability at diagnosis also correlated with response to treatment: tumours with increasing chromosomal instability and genomic rearrangements were more likely to attain pCR (Figure 3.9A): this association was strongest in $ER^{+}HER2^{-}$ tumours (p=0.003), but was not statistically significant in $ER^{-}HER2^{-}$ and $HER2^{+}$ tumours (Figure 3.9B).

In 2012, a classification based on copy number changes was proposed through the integration of genomic and transcriptomic data from 2,000 breast tumours [60, 63]. This classification, entitled IntClust, identified 10 different subgroups, each with distinct copy number alterations and distinguishable clinical outcomes. The key strength of this classifier, amongst others, was the ability to split seemingly similar breast cancer intrinsic subtypes (as classified by the PAM50 classification [221]) into ones with very different clinical outcomes [12].

IntClust classification was performed using the `iC10` R package [12] (Figure 3.9C), using as input: (1) cellularity corrected copy number log ratios (obtained by running `QDNAseq` on

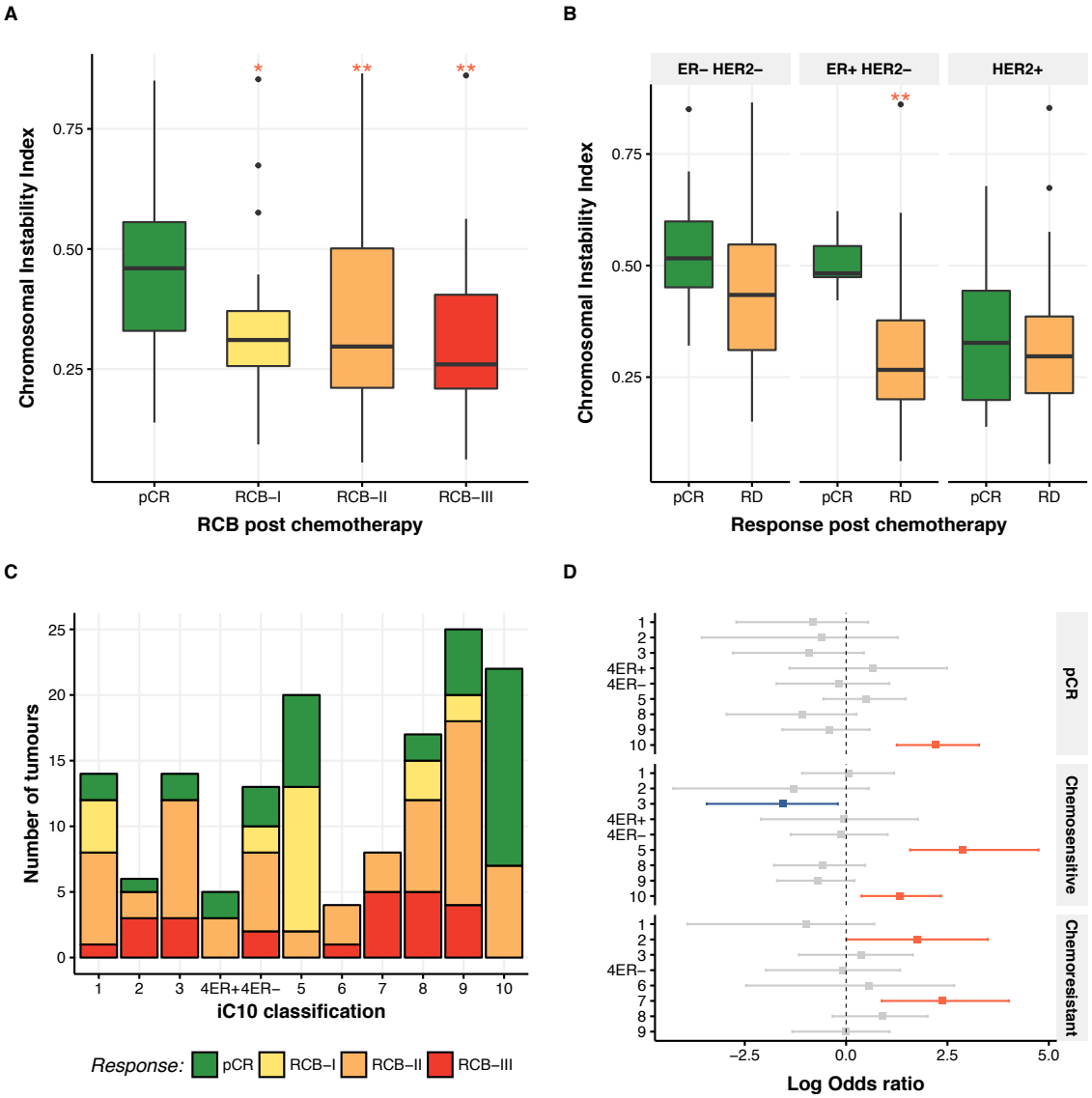


Figure 3.9: Association between copy number alterations and response to chemotherapy. (A) Increasing chromosomal instability was associated with pCR, this association was statistically significant in $ER^{+}HER2^{-}$ tumours (B). (C) Tumours classified by iC10 status. (D) Association between iC10 and response to chemotherapy.

the shallow whole genome sequencing files (Section 2.5.1), as well as (2) `voom` normalised gene expression counts. Most tumours fell into iC 9 (mostly ER$^+$ luminal B), 10 (mostly TNBC) and 5 (HER2-enriched) each with distinct copy number alterations (iC9: 8q gain, 20q amplification, iC10: 5q loss, 8q gain, 10p gain, 12p gain, iC5: *ERBB2* amplification) [63].

Associations with pCR, chemosensitivity and chemoresistance were performed using a logistic regression model accounting for ER and HER2 status (Figure 3.9D). Tumours classified in the iC10 group were most associated with obtaining pCR and RCB-I following neoadjuvant chemotherapy. iC10 tumours are mostly triple negative tumours from the PAM50 basal-like subtype, with enrichment of *TP53* mutations, intermediate levels of genomic instability and copy number alterations involving 5q loss (harbouring many DNA damage repair and apoptosis genes) and gains at 8q, 10p and 12p [60, 63].

The iC5 subgroup encompassed tumours with *ERBB2* amplification at 17q12, intermediate levels of genomic instability and a high proportion of *TP53* mutations, and tumours falling into this category were associated with increased chemosensitivity (pCR and RCB-I). This was unsurprising, given the clinical efficacy of trastuzumab.

iC3 tumours, which were often low proliferation, luminal A tumours with low genomic instability and paucity of copy number changes, as well as low prevalence of *TP53* mutations and high frequency of *PIK3CA*, *CDH1* and *RUNX1* mutations were associated with decreased chemosensitivity (i.e. more likely to result in RCB-II or III disease). Once again, the decreased proliferation, as well as negative association with *TP53* mutations and positive association with *PIK3CA* mutations explained the decreased chemosensitivity.

Tumours classified in the iC7 group, which were characterised by 16p gain, 16q loss and 8q amplification, were associated with chemoresistance (RCB-III). As with iC3 tumours, iC7 tumours were predominately ER$^+$, low proliferation, well-differentiated luminal A tumours, albeit with more genomic instability than iC3 tumours.

Finally, tumours in iC2, characterised by 11q13/14 amplification (in which the driver genes *CCND1*, *EMSY* and *PAK1* reside) and high levels of genomic instability were also associated with chemoresistance, as previously shown by Ali et al. [12]. This group comprised of luminal A and B tumours which paradoxically had the worst prognosis of all ER$^+$ tumours, with intrinsic chemoresistance despite being highly proliferative.

Hence the iC10 classification, to an extent, mirrored the observations previously made in the earlier parts of this chapter. Copy number groups associated with *TP53* mutations with

higher levels of proliferation were more likely to attain pCR or be chemosensitive, whilst iC10 groups enriching for *PIK3CA* mutations and which had decreased genomic instability were associated with chemoresistance.

GISTIC copy number analysis

Pinpointing specific regions of copy number alterations (CNA) within the genome that are associated with response is challenging in view of the breadth of the genome and the large number of samples required in order to obtain genome-wide significance. In order to overcome this, genomic regions were condensed into minimum common regions by using `GISTIC` version 2 [196] across three cohorts:

1. **Response - pCR**: `GISTIC` run separately on tumours that attained pCR and those with residual disease.
2. **Response - Chemosensitivity**: `GISTIC` run separately on chemosensitive tumours (pCR and RCB-I) and non-chemosensitive tumours (RCB-II and RCB-III).
3. **Response - Chemoresistance**: `GISTIC` run separately on chemoresistant tumours RCB-III) and non-chemoresistant tumours (pCR, RCB-I and RCB-II).

Regions of significant changes in each of these three comparative groups were merged, resulting in a candidate list of regions that were more statistically significantly altered than expected by chance. Associations between response and each of these genomic regions was assessed by using a logistic regression model. The analysis identified 9 regions associated with pCR, 13 regions associated with chemosensitivity and 3 regions associated with chemoresistance at an FDR of < 0.05 (Table 3.1).

Oncogene amplification was often associated with chemosensitivity, whilst tumour suppressor gene loss was associated with chemoresistance.

Amplification of 1q41-1q44, which codes for the DNA repair protein *PARP1* that is involved in the base excision repair pathway, was positively associated with pCR and negatively associated with chemoresistance, in keeping with previous findings [322]. Likewise, amplification of *PIK3CA* and *TBL1XR1* on 3q26, *MYC* on 8q22-8q24, *RPS6K* and *AML1* on 21q22 were associated with pCR. Deletion of *FBXW7* on 4q31-4q35, which participates within the molecular complex involved in the ubiquitination of cyclin-E, JUN, MYC and NOTCH1 [115] was associated with pCR.

Table 3.1: CNAs associated with response to neoadjuvant therapy.

Response	CNA	Cytoband	FDR	Odds Ratio	CI (Odds Ratio)
pCR	Amplification	1q41 - 1q44	0.0000000	1.84	1.54-2.26
		3q25.31 - 3q29	0.0000001	1.60	1.39-1.91
		8q22.3 - 8q24.22	0.0000000	1.34	1.22-1.48
		14q21.3 - 14q22.3	0.0223600	1.32	1.14-1.65
		17q22 - 17q24.1	0.0370086	1.18	1.09-1.37
		17q24.3 - 17q25.3	0.0032401	1.41	1.23-1.77
		21q22.12 - 21q22.3	0.0000001	1.83	1.53-2.3
	Deletion	4q31.23 - 4q35.2	0.0306197	5.21	1.66-16.14
		16q12.2 - 16q21	0.0013203	0.05	0.01-0.18
Chemosensitivity	Amplification	1p12 - 1q24.1	0.0034722	1.52	1.29-2.03
		1q31.2 - 1q44	0.0000001	1.70	1.45-2.05
		3p26.3 - 3p22.3	0.0004380	1.42	1.24-1.76
		3q25.32 - 3q29	0.0000091	1.50	1.3-1.78
		8p12 - 8p11.22	0.0379688	0.82	0.7-0.93
		8q21.3 - 8q22.2	0.0276927	1.16	1.05-1.29
		13q33.3 - 13q34	0.0247815	1.16	1.08-1.34
		14q11.2 - 14q22.3	0.0001925	1.45	1.27-1.83
		17q12 - 17q21.1	0.0028822	1.10	1.05-1.17
		17q25.3	0.0126556	1.26	1.12-1.51
		19p13.12 - 19p13.11	0.0185112	1.37	1.17-1.8
		19q13.42 - 19q13.43	0.0004380	1.37	1.2-1.63
		21q22.2 - 21q22.3	0.0000208	1.61	1.35-1.99
	Deletion	13q11 - 13q21.31	0.0126556	0.19	0.06-0.55
Chemoresistance	Amplification	1q41 - 1q44	0.0401158	0.68	0.52-0.85
	Deletion	11q14.2 - 11q25	0.0401158	7.59	2.16-26.98
		16q22.2 - 16q24.3	0.0401158	4.24	1.67-10.73

Chemosensitive tumours had amplification of breast cancer drivers *NOTCH2*, *TBL1XR1*, *PIK3CA*, *ZNF703*, *FGFR1*, *FOXA1*, *ERBB2* and *CNOT3*, and deletion of *BRCA2* and *RB1*.

Tumour suppressor genes were predominantly present within genomic regions associated with chemoresistance. Loss of the FRA16D common chromosomal fragile site at 16q was associated with chemoresistance (OR: 4.24). This region also codes for four tumour suppressor genes: *ZFHX3*, *WWOX*, *CBFA2T3* and *FBXO31*, with *WWOX* being a key player in maintaining genomic stability and regulating DNA repair by modulating the activity of *ATM* [3]. The second site associated with chemoresistance was deletion of 11q, which contains the tumour suppressor protein *YAP1*, as well as two genes involved in checkpoint mediated cell cycle arrest in response to DNA damage: *CHEK1* and *ATM*. Through the loss of key molecules involved in inducing apoptosis secondary to cytotoxic damage, these tumours were more likely to become chemoresistant.

3.3.3 Genomic immune landscape

The contribution of the immune system to the tumour ecosystem is being increasingly appreciated, with evidence rapidly accumulating that immune-mediated cell kill regulates tumour evolution [233, 234, 240, 286]. Indeed, the latest generation of anti-tumour agents are increasingly immunomodulatory, with PD-L1 inhibitors (including atezolizumab and avelumab), PD1 inhibitors (including pembrolizumab and nivolumab) and CTLA4 inhibitors (such as ipilimumab) already being used in clinic in a variety of tumour sites following impressive results in phase III clinical trials [39, 119, 125].

The foundation of the immune system is built on the ability of discriminating between 'self' and 'non-self': during the early stages of development, immune system cells that recognise self molecules as foreign are eliminated and the remaining immune cell repertoire is competent at discriminating self from non-self, thereby becoming effective in recognising and eliminating pathogens [145].

Tumour cells translate and transcribe mutant DNA sequences producing non-self peptide sequences that, if bound to MHC class I molecules (including HLA-A, HLA-B and HLA-C) and displayed on the cell surface, have the potential of being antigenic and inducing an immune response [41, 240, 315]. As the number of expressed coding mutations correlated with response to chemotherapy, the subsequent hypothesis was that this could be due to an increased number of neoantigens presented to the scanning immune cells.

In order to characterise the neoantigenic landscape of these tumours, HLA typing was performed on the matched normal tissue sequencing data using the `Polysolver` (POLYmorphic loci reSOLVER) tool [261], which inferred the 4-digit HLA type for each sample by using a Bayesian classifier to determine genotype. 182 different HLA genes across all cases were genotyped (HLA-A: 54, HLA-B: 82, HLA-C: 46, Figure 3.10A).

HLA genotypes HLA-B*38:01 and HLA-A*23:01 were strongly associated with pCR (Figure 3.10B, HLA-B*38:01: OR 34.94, CI: 4.16-786.77, p=0.004, HLA-A*23:01: OR: 9.99, CI: 1.35-92.83, p=0.030). The association with HLA-B*38:01 was maintained in the chemosensitive subtypes (OR:16.82, CI: 2.17-355.59, p=0.017), with genotype HLA-C*14:02 also conferring chemosensitivity (OR: 13.42, CI: 1.35-302.15, p=0.038). Genotypes HLA-B*18:01 and HLA-C*07:18 were both associated with chemoresistant tumours (HLA-B*18:01: OR:6.18, CI: 1.44-27.71, p=0.014, HLA-C*07:18: OR: 9.63, CI: 1.50-59.74, p=0.013).

Of note, three of these HLA alleles have been described in playing a role in autoimmunity or HIV infection:

1. The HLA-B*38:01 allele, which was associated with pCR and chemosensitivity and was present in 3.6% of the cases included within this cohort, has been shown to be the most protective MHC Class I allele against the development of type 1 diabetes, though the mechanism by which this happens has not been elucidated [132, 262].
2. HLA-C*14:02, which was associated with chemosensitivity and was present in 3.5% of cases has been shown to be significantly associated with an increased risk of severe acute graft-versus-host disease after unrelated donor bone marrow transplantation, as well as increased transplant-related mortality [201].
3. HLA-B*18:01, present in 4.4% of cases, was associated with chemoresistance and has been shown to be associated with high HIV viral replicative capacity in adults [7, 48, 168].

As LOH over the HLA class I locus has recently been described as a potential means of immuno-evasion (as it results in fewer neoantigens presented), the prevalence of LOH over the HLA class I locus was determined by using the `LOHHLA` tool [192], which took as input tumour purity from `ASCAT` [296] as well as HLA genotyping data from `PolySolver` [261]. Statistically significant HLA alleles with a copy number of less than 0.5 were assumed to be undergoing LOH. 29 cases (16%) had evidence of LOH across at least one MHC class I locus. 15 cases had LOH over one MHC Class I locus, 8 cases had LOH over two HLA

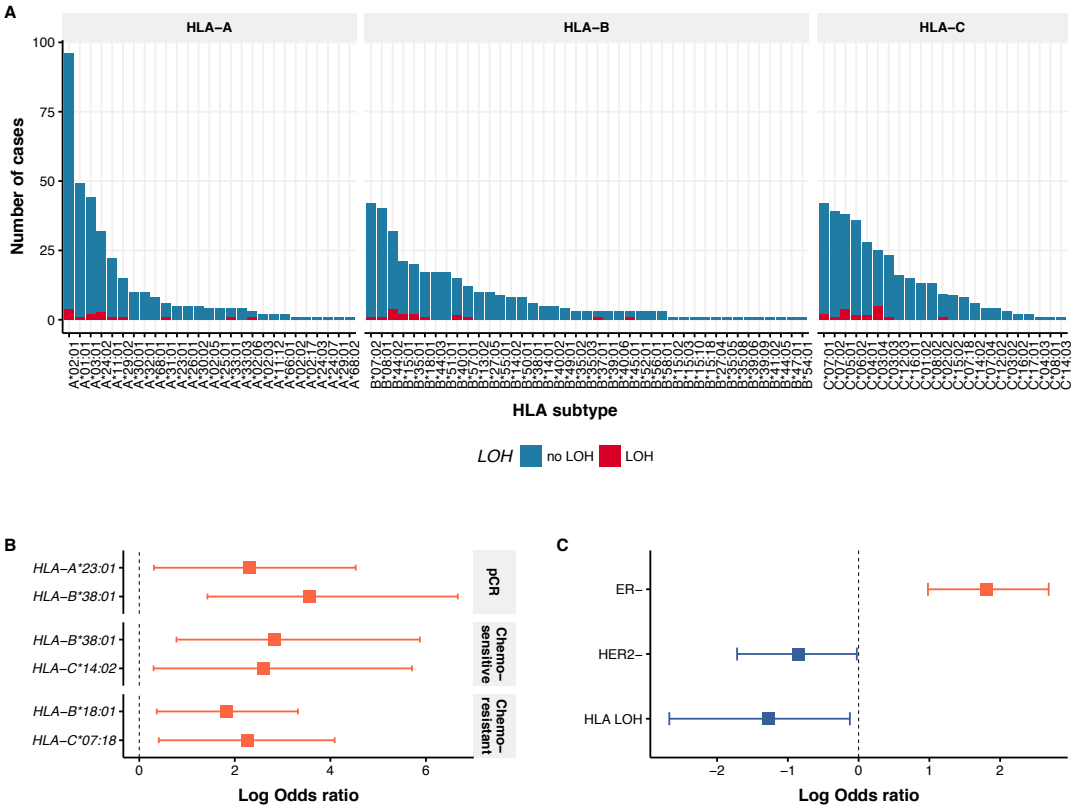


Figure 3.10: MHC Class I landscape. (A) HLA genotyping identified 182 different HLA molecules across the cohort, red bars indicate frequency of HLA molecules subject to LOH. (B) Five HLA molecules were strongly associated with response to chemotherapy, one of which (HLA-B*18:01) has been shown to be associated with poor outcome in HIV+ disease. (C) Presence of HLA LOH is associated with a lower likelihood of attaining pCR, potentially secondary to fewer neoantigens presented to the immune system.

genes and 6 cases had loss of 3 genes. HLA LOH was strongly associated with RD (OR: 0.28 CI: 0.07-0.88, p=0.04, logistic regression model factoring in ER and HER2 status, Figure 3.10C), in keeping with the hypothesis that tumours that present fewer neoantigens are less likely to elicit an immune-mediated response.

Having genotyped the HLA class I molecules and determined associations with response, putative neoantigen calling was performed by using the `pVAC-tools` cancer immunotherapy suite [135]. Mutations identified on exome sequencing were translated into corresponding mutant proteins and a list of potential neoantigenic fragments containing the mutant protein generated by using a sliding window approach across the mutated locus, retaining epitopes of lengths 8-11 amino acids. These potentially antigenic fragments were analysed for binding affinity to the HLA class I molecules using the prediction software `NetMHCPan` version 3 [211], `NetMHC` version 4 [179] and `PickPocket` version 1.1 [323] bundled within the Immune Epitope Database resource (IEDB) [300]. Neoantigens with a binding affinity score of less than 500nM and which had a higher binding affinity than the corresponding wild-type sequences were retained. Further downstream filtering was done by retaining neo-epitopes generated by transcripts that had an expression greater than 1 TPM.

After filtering, a total of 4,134 neoantigens were retained, with a median of 20 neoantigens per case (range: 0-149). 19.6% of all coding non-silent mutations generated an expressed neoantigen. The total neoantigen burden strongly correlated with the total mutation count (R=0.76, p=4e-32), Figure 3.11A. 1.4% of neoantigenic sequences were present in multiple unrelated cases and were due to identical mutations in *GATA3*, *MAP3K1*, *MT-ND4*, *PIK3CA*, *SF3B1* and *TP53*. There was no association between the presence of a particular neoantigenic sequence and response to therapy.

Tumours that attained pCR post-chemotherapy had higher neoantigen burdens at diagnosis than those with RD (31 vs. 22 neoantigens, p=0.0004, Wilcoxon rank sum test, Figure 3.11C), and there was a clear statistically significant separation of neoantigen burden distributions between chemoresistant tumours and those that attained pCR (p=0.001, Kolmogorov-Smirnov Test, Figure 3.11D). The association between neoantigen burden and response was statistically significant in ER^+HER2^- and ER^-HER2^- tumours, but not in $HER2^+$ tumours. (ER^-HER2^- p=0.009, ER^+HER^- p=0.011, $HER2^+$ p=0.219, Figure 3.11E).

Clonal neoantigens were more prevalent in tumours that attained pCR (p=0.001, Wilcoxon rank sum test), and there was no association between the number of sub-clonal neoantigens and response to treatment (p=0.640, Wilcoxon rank sum test).

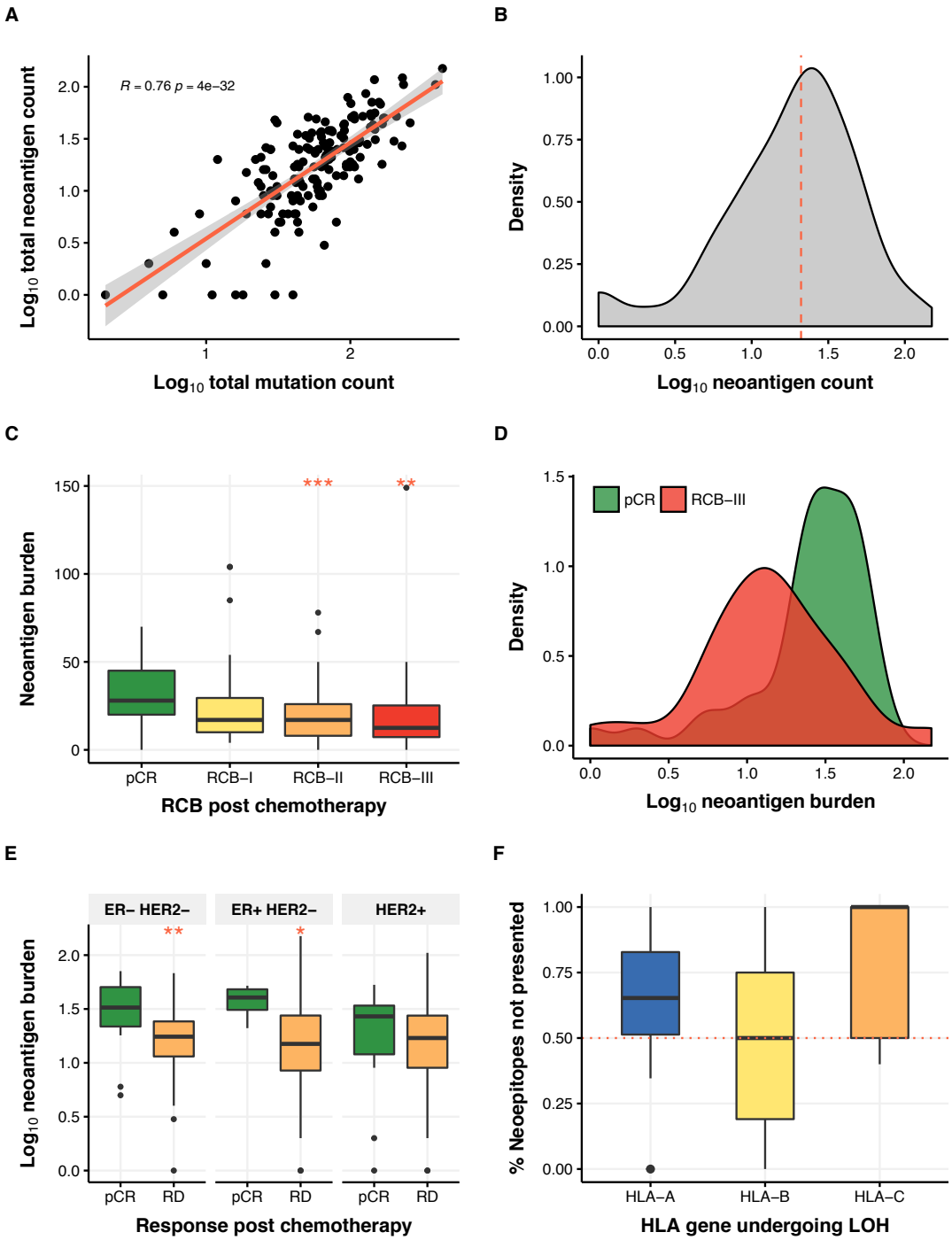


Figure 3.11: Neoantigen burden and association with response. (A) Total neoantigen burden strongly correlated with total mutation count. (B) The median neoantigen burden per sample was 25, though tumours that attained pCR had statistically significantly higher numbers of neoantigens than those with RD (C,D). This association was most clearly seen in $HER2^{-}$, but not $HER2^{+}$ tumours (E). Bar plot showing the percentage of neoantigens that were not presented after LOH: LOH preferentially occurred over the HLA molecule that presented more neoepitopes.

92.8% of all expressed neoantigens bound to a single MHC molecule, and tumours that had increasing numbers of neoantigens that bound to more than one MHC class I molecule were more likely to attain pCR (OR: 1.28 CI 1.07-1.55, p=0.009, logistic regression).

The effect of HLA LOH on the number of presented neoantigens across all 29 cases harbouring LOH was then determined. In total, these 29 cases had 419 neoantigenic peptides, out of which 233 (56%) were presented by the HLA molecule undergoing LOH. Hence, by developing HLA LOH, the tumours effectively prevented more than half of the neoepitopes generated from being presented. 73% of LOH events resulted in the loss of a molecule that presented an equal or greater number of neoepitopes than its retained alternative allele (Figure 3.11F)). Consequently, only 27% of LOH events resulted in the loss of an allele that presented fewer neoepitopes that its alternative haplotype. The distribution of neoantigen loss was significantly above 50% (p=0.003, Wilcoxon signed rank test), indicating that tumours actively lost the HLA locus that was able to present the greater number of neoantigens to the T cell receptors, and consequently were more likely to have RD post-chemotherapy.

3.4 Transcriptomic predictors of response

3.4.1 Differential gene expression analysis

An analysis of genomic predictors of response showed many important observations, including the association of CNAs and mutations in cancer driver genes with response to treatment, as well as the probable involvement of the immune system through recognition of neoantigenic peptides. The transcriptomic landscape and its association with response to chemotherapy has been explored in the literature, with various indices, including proliferation and immune infiltrate, associated with response [68, 140, 172, 182].

So as to identify genes associated with increasing probability of response, unprocessed RNA-seq counts obtained as discussed in Section 2.5.2 were normalised using trimmed mean of M-values (TMM) normalisation to mitigate sample-specific effects due to differing sequencing depths of each RNA sample as well as batch effects [237]. Differential expression using a negative binomial model was performed using the `edgeR` R package [190, 236] using the linear model:

$$\sim ER_{status} + HER2_{status} + Batch + Treatment + RCB_{score} \tag{3.3}$$

where ER_{status}, $HER2_{status}$, *Batch* and *Treatment* were categorical variables, and RCB_{score} the numeric RCB score (with 0 being pCR). Rather than modelling response as a binary variable (i.e. pCR and RD), the response variable in the differential expression model was preferentially selected to be the continuous RCB score. Indeed, as the genomic analysis data has shown, most genomic predictors of response can be modelled on a continuous scale, rather than a discrete scale (eg TMB, mutational signature contribution, neoantigen load) and correlate monotonically with degrees of response. So, for example, an increasing contribution of a factor might result in an increased degree of response: such an association would not be picked up if binary classifications were used. Hence this differential expression model was able to detect gradual gene expression changes across the four response groups, and reveal transcriptomic variables associated with improved response. By including ER, HER2, treatment and batch within the model, the resulting effect observed was adjusted for these variables.

The differential analysis based on the model described in Equation 3.3 revealed 871 lowly expressed and 1,202 highly expressed genes in patients that were more likely to attain a better response with chemotherapy (Figure 3.12A).

Chemotherapy metabolising drugs and adenosine triphosphate (ATP) binding cassette multi-drug transporters featured within the differentially expressed gene list, shedding light into mechanisms of resistance to treatment (Figure 3.12B). The lowliest expressed gene in tumours attaining a better response (and conversely the highest expressed in those with higher residual cancer burden) was *CYP2A7* which encodes a member of the cytochrome P450 super family of enzymes involved in xenobiotic metabolism, though the substrate of this enzyme is currently unknown. *CYP2C8*, which is involved in taxane and cyclophosphamide metabolism and elimination [227] was also lowly expressed in tumours attaining a better response. Hence, high expression of these xenobiotic enzymes presumably resulted in faster elimination of the cytotoxic drugs administered and therefore a lesser probability of attaining pCR.

ABCG2, an ATP binding cassette multi-drug transporter which is known to pump out various chemotherapeutic agents such as 5FU, anthracyclines and taxanes, and is associated with drug resistance [134, 320] was also highly expressed in tumours that did not respond to treatment. Indeed, high expression of this gene has already been implicated in resistance to treatment in breast cancer [320]. Similarly, *SLC29A1*, a solute carrier which is involved in the uptake of 5FU from the basolateral membrane is highly expressed in patients with pCR: a similar association was seen in pancreatic cancer [291].

Hence, the pre-treatment expression landscape showed that tumour cells were already equipped with mechanisms conferring resistance to therapy. By increased expression of xenometabolising enzymes and chemotherapy efflux molecular pumps, as well as decreased expression of chemotherapy drug influx molecular transporters, tumours were already equipped with a resistance toolkit to circumvent the cytotoxic effects of these treatments.

Breast cancer driver genes that were highly expressed in tumours attaining a better response to treatment (Figure 3.12A) included: *BRCA2*, *BUB1B*, *CBFB*, *CCND3*, *CCNE1*, *CDKN2A*, *DNMT3A*, *EGFR*, *NOTCH1* and *PIK3CA*. Interestingly, as discussed previously, *PIK3CA*, *ERBB2* and *BRCA2* copy number gains were also very strongly associated with chemosensitivity, and it is reassuring that the corresponding expression of these three breast cancer driver genes was also increased, thereby validating observations made in the copy number landscape

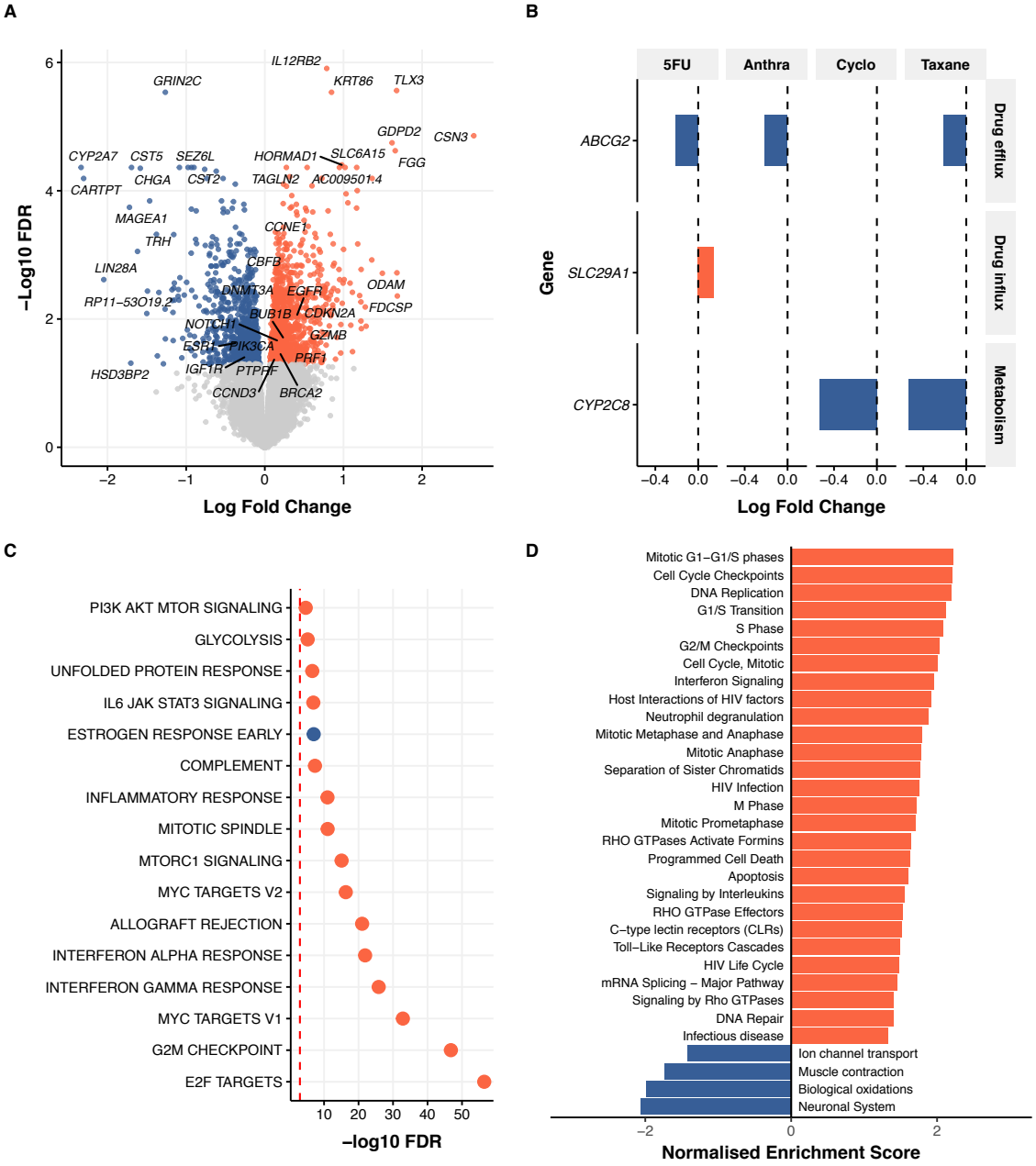


Figure 3.12: Transcriptomic differential expression. Red indicates highly expressed in pCR, blue indicates lowly expressed in pCR. (A)Volcano plot showing genes that are highly or lowly expressed in tumours that are more likely to attain a better response to chemotherapy. (B) Xenobiotic enzymes and drug efflux/influx molecular pumps associated with response. (C) GSEA over the Hallmarks MSigDB dataset, showing enrichment over proliferation and immune system pathways. (D) GSEA over the Reactome database: proliferation and immune pathways strongly correlated with response.

analysis. Higher expression of cell cycle driver genes was associated with chemosensitivity, confirming observations that increased proliferation was associated with response to therapy.

Unsurprisingly, *BCL2*, *ESR1*, *IGF1R*, *PREX2* and *STK11* were highly expressed in tumours with poor response to treatment: strongly expressing *ESR1* luminal A tumours in iCs 3, 7 and 8 were less likely to be chemoresponsive as they were of lower grade. Additionally, the increased expression of *BCL2*, a negative regulator of apoptosis, would result in the evasion of cytotoxic cell death and explain the increased expression in poor responders.

The transcriptomic landscape differences were not only limited to xenobiotic and driver gene expression. Various immune system genes were also highly expressed in tumours attaining pCR, most notably perforin (*PRF1*), a key mediator of lymphocyte mediated cytolysis [167] which forms channels in the target cell membrane and allows the influx of cytotoxic molecules such as granzyme into target cells. Indeed, perforin is a key component of the cytolytic activity score that is a surrogate for the degree of T cell killing [240]. Additionally, tumours attaining a better response after neoadjuvant therapy had higher expression of *IL12RB2*, which induces proliferation of T-cells and NK cells, thereby enhancing IFN-γ production. In keeping with this, *IFNG* was also strongly co-highly expressed in the microenvironment of these tumours. The differential analysis also revealed high expression of the cytotoxic T cell marker *CD8*, as well as *HLA-A*, *PDL1* and *PDCD1* in tumours that attained pCR, confirming the presence of increased immune cell infiltrate.

To identify transcriptomic pathways associated with better response to chemotherapy, gene set enrichment (GSEA) on the differentially expressed list of genes was performed using the `camera` statistical method in `edgeR`: in brief, this method performed a competitive gene set test accounting for inter-gene correlation and tested whether genes were highly ranked relative to other genes in terms of differential expression [311]. As input to the GSEA method, the annotated gene sets provided within the Molecular Signatures Database (MSigDB) version 6.1 were used [171, 274], specifically the Hallmark (H) gene set, Computational (C4) gene set and Gene Ontology (GO, C5) gene set.

Enrichment over the Hallmarks gene set (Figure 3.12C) showed that two key pathways were enriched: cell division/proliferation (including: E2F targets, mitotic spindle, G2M checkpoint) and immune system pathways (interferon alpha/gamma response, allograft rejection, inflammatory response, IL6 JAK STAT3 signalling). STAT3 signalling is a major pathway that induces cancer inflammation and is associated with increased proliferation, angiogenesis and metastasis, while also inhibiting anti-tumour immunity [159, 219, 248].

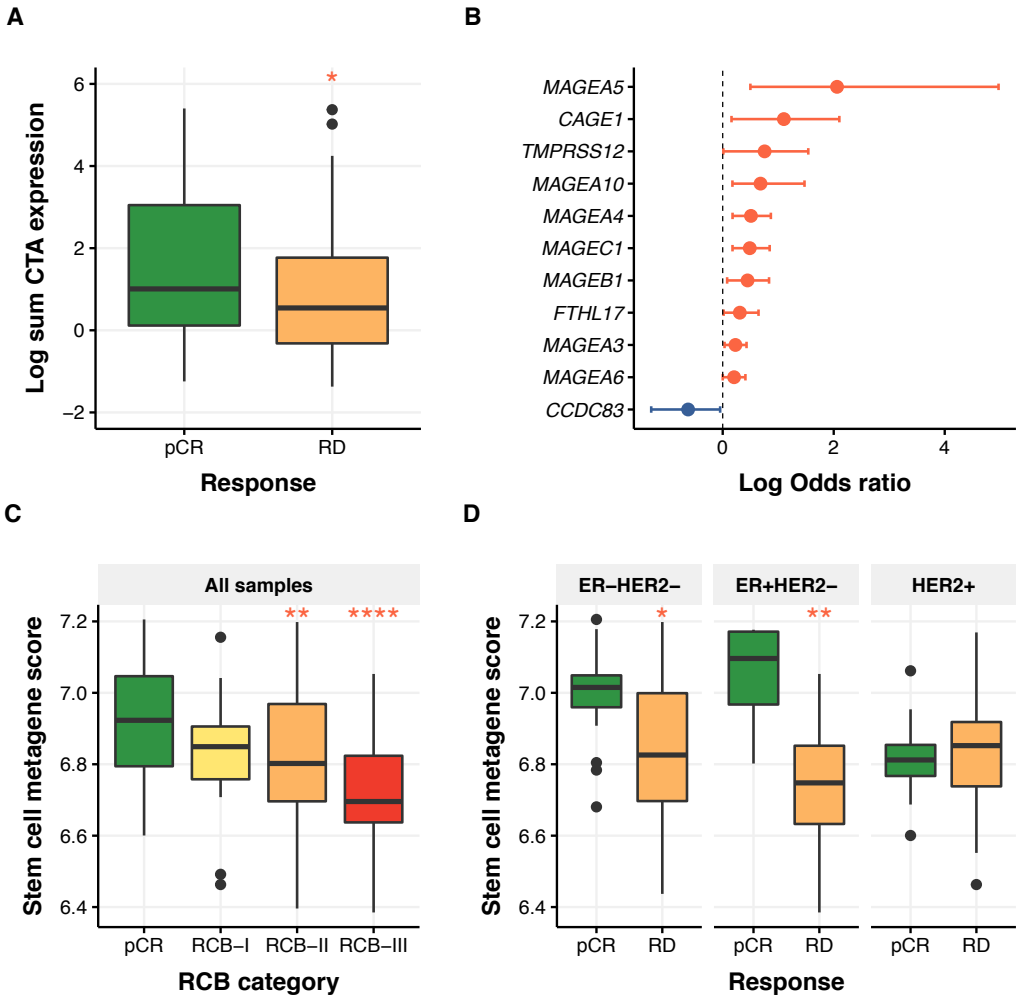


Figure 3.13: Expression of Cancer/Testis Antigens and stem cell markers at diagnosis. (A) Box plot showing difference in total CTA expression between tumours that attain pCR and those with RD. Red indicates highly expressed in pCR, blue indicates lowly expressed in pCR. (B) CTA expression was associated with response to chemotherapy. (C) Expression of a stem cell metagene was associated with response, with correlations seen in ER^+ and ER^- disease, but not $HER2^+$ tumours (D).

Indeed, STAT3 decreases anti-tumour immune responses by antagonising NF-κB and STAT1 expression of T_H1 cytokines (such as IL-12 and IFN-γ) which are necessary for innate and T cell-mediated anti-tumour immunity [318]. Perhaps the heightened expression of STAT3 (and consequent immuno-inhibition) was a direct consequence of the increased anti-tumour immune response and complemented the increased PD1 and PDL1 expression observed in tumours that attain pCR. Furthermore, as revealed by the differential expression analysis and copy number analysis, GSEA confirmed decreased *ESR1* signalling in tumours that attained pCR and increased *PIK3CA*, *AKT* and *MTOR* signalling.

GSEA on the C5 GO set revealed enrichment over 328 processes. Once again, most of the GO terms that were identified pertained to proliferation and immune activation pathways. An enrichment over the Computational C4 gene set also demonstrated a strong correlation between response and genes around the vicinity of *CCNA2*, *CDC20*, *CDK1*, *CENPF*, *PCNA* and *RRM1*, all of which were strongly associated with cell cycle and division.

Finally, to validate these observations using an independent pathway database, enrichment over the Reactome database [78] was performed using the `ReactomePA` R package [317]. This enrichment validated the previous observations on a different curated dataset: tumours that were likely to respond better had higher proliferation and greater immune activation (Figure 3.12D).

The genomic analyses showed that increased neoantigen load corresponded with response to chemotherapy, with tumours having greater neoantigen burdens more likely to attain pCR. Expression of Cancer/Testis Antigens (CTAs) is restricted to male germ cells, however their expression is often reactivated in cancer and can be immunogenic if processed peptide sequences are presented on HLA Class I molecules [250]. The expression of 59 CTA genes [240] was examined (Figures 3.13A, B): of these, 10 genes (*CAGE1*, *FTHL17*, *MAGEA10*, *MAGEA3*, *MAGEA4*, *MAGEA5*, *MAGEA6*, *MAGEB1*, *MAGEC1*, *TMPRSS12*) were strongly associated with pCR if expressed, while the expression of *CCDC83* was associated with RD (OR: 0.5, CI: 0.3-0.9, p=0.04). The combined expression of these 59 CTA genes correlated with the degree of response: high expression of CTA genes was more likely to result in pCR (OR: 1.31, CI: 1.05-1.65, p=0.02). Since CTAs are expressed as tumours become more de-differentiated, the expression of an established adult stem cell metagene [310] was assessed in the dataset (Figures 3.13C, D). Tumours that attained pCR had higher enrichment of this stem cell metagene (p=0.02, Wilcoxon rank sum test).

3.4.2 Tumour proliferation

As the GSEA showed that tumour proliferation was a key pathway that determined response to chemotherapy, the contribution of proliferation signatures to response was explored in further detail. Gene set variation analysis (GSVA) was performed on the Genomic Grade Index (GGI) gene set [271], comprising 97 genes associated with tumour histological grade, by using the `GSVA` R package [114]. GSVA was chosen to obtain a pathway activation score ranging from -1 (inactive) to 1 (highly active). As seen in Figure 3.14A, the distribution of GGI was distinctly bimodal, with high and low proliferation groups. To determine the optimum cut off score, the distribution of scores was modelled into a mixture of two Gaussian distributions by using a mixed effect model provided by the `MClust` R package [255]. A threshold score of -0.14 was used to define the cut-off point between high and low proliferation groups.

80% of tumours that attained pCR had a high proliferation score, compared to 77% of RCB-I tumours, 51% of RCB-II tumours and 29% of RCB-III tumours (Figure 3.14B). The gradual decrease in proliferation activation across the four response groups is shown in Figure 3.14C: as the degree of residual cancer increased, proliferation activation at diagnosis decreased. Similar findings have been previously reported [172]. The association between GGI and pCR was strongest in ER^+HER2^- (p=0.006) and ER^-HER2^- (p=0.004) tumours, but not $HER2^+$ tumours (p=0.964, Wilcoxon rank sum test), Figure 3.14D. Splitting the $HER2^+$ group into ER^+ and ER^- subtypes did not reveal any distinctive findings. An ordinal model (Figure 3.14E) confirmed the relationship between GGI GSVA score and increasing RD and predicted that at activation scores of 0.4, it was far more likely for a tumour to attain pCR than have RD after chemotherapy (p=2.04e-05).

Interestingly, 64% of cases with high proliferation did not attain pCR, whilst 13% of cases with low proliferation still attained pCR, showing that proliferation, by itself, was not the sole factor that determined response to chemotherapy and that other pathways were also contributing to response. To explore this further, the dataset was split into two groups: one in which tumours with high proliferation that attained pCR were compared to those with RD and a second group in which tumours with low proliferation that attained pCR were compared to those with RD. Hence, by doing so, mechanisms responsible for pCR in both highly proliferative and lowly proliferative tumours could be ascertained. In both cases, tumours that attained pCR in both proliferation high and low subgroups had highly significant enrichment over GO immune response pathways (Figure 3.15), more so in the high proliferation group than the low proliferation group, suggesting that proliferation was

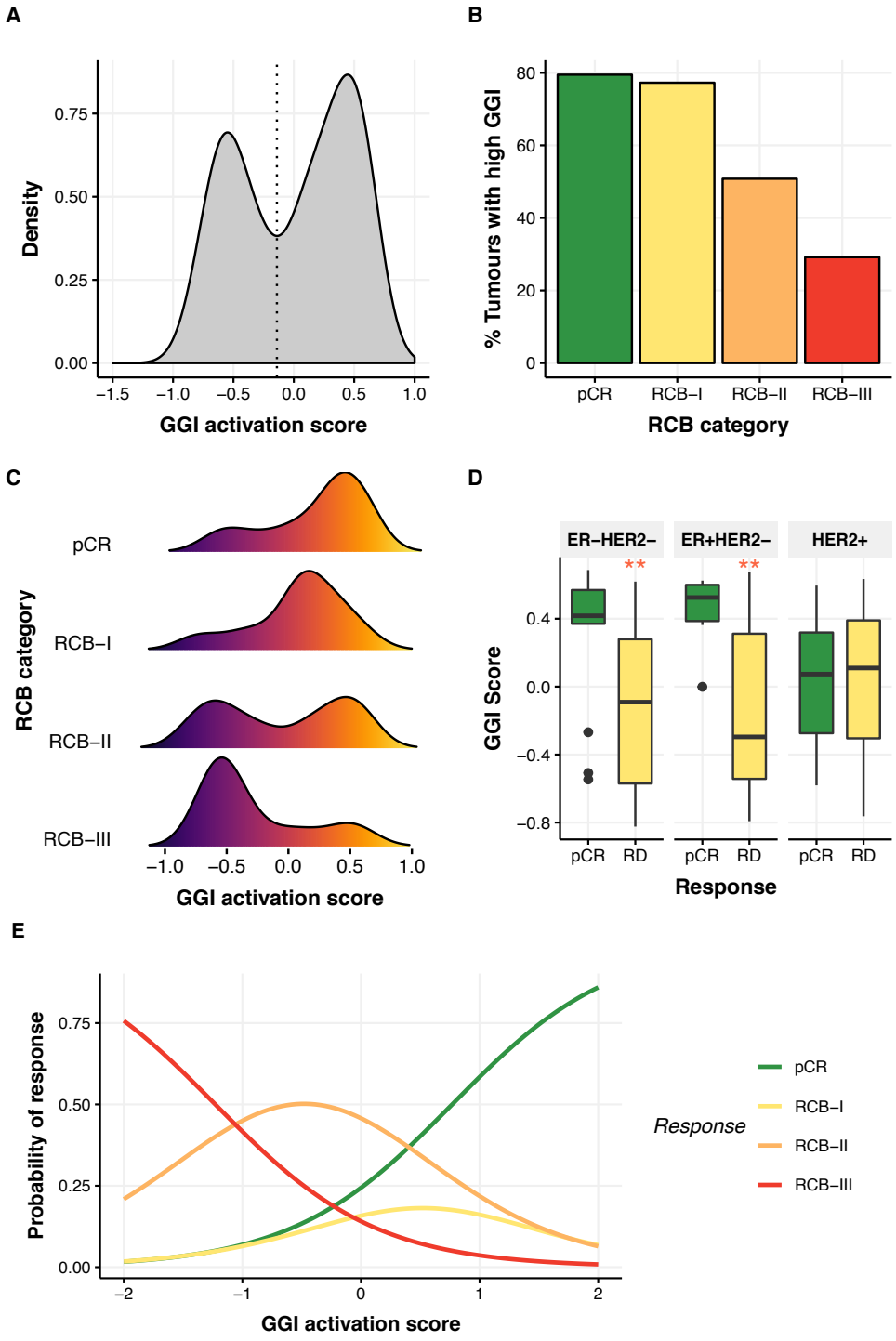


Figure 3.14: GGI score and association with response. (A) GSVA on GGI gene set over all samples at diagnosis, showing a biomodal distribution. (B) Barplot showing percentage of tumours with high GGI in each response category. (C) Shifting density plot showing increasing pathway deactivation as residual cancer burden score increases. (D) Association between GGI score and response statistically significant in HER2$^-$, but not HER2$^+$ cohorts. (E) Ordinal prediction model of response given a change in GGI GSVA score.

not the sole mediator of response and attaining pCR also required an immuno-competent tumour microenvironment. Interestingly, low proliferation tumours that attained pCR also had an increased enrichment over extracellular matrix gene sets, indicating that the non immune tumour microenvironment also played a role in fashioning response to therapy.

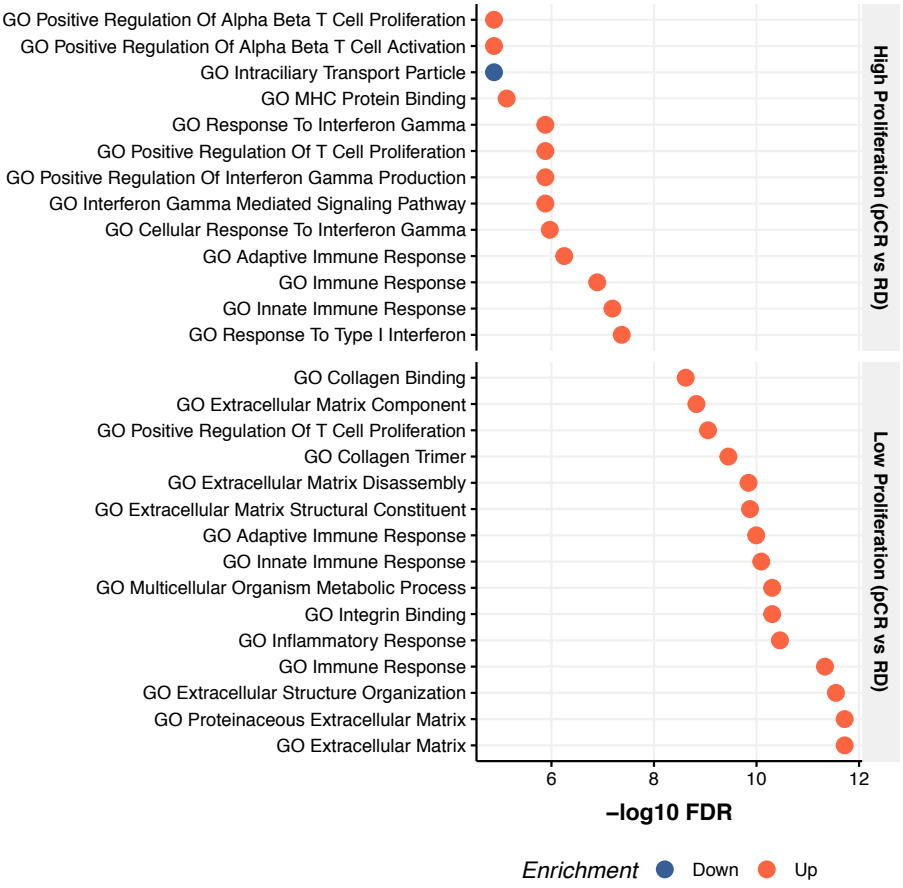


Figure 3.15: GSEA on highly proliferative (top) and lowly proliferative (bottom) tumours, pCR vs. RD. Tumours that attained pCR in both groups had higher expression of immune system pathways, compared to those that had RD post treatment, showing that chemotherapy requires a competent immune system for maximum effect.

3.4.3 Tumour immune microenvironment

In view of the key findings of the previous differential expression analysis (Section 3.4.1), as well GSEA enrichment over the GO and Reactome pathways showing a strong association between immune pathways and response to chemotherapy, the contribution of the immune system was analysed further.

The cytolytic activity (CYT) score is considered to be a surrogate measure of T-cell mediated cytolysis [240] and is defined as the geometric mean of the expression of granzyme A (*GZMA*) and perforin (*PRF1*). In view of this, the CYT score at diagnosis was computed in order to gain an understanding of the degree of immune-mediated cytolysis across the whole cohort. The previous differential expression analysis had already suggested that the expression of *PRF1* was higher in tumours that attained a better response to chemotherapy, and a comparison of CYT score across all response groups (Figure 3.16A) confirmed the association between increased immune-mediated cytolysis and improved response to chemotherapy. Tumours that attained pCR on chemotherapy had higher CYT scores at diagnosis than those with RD (median: 3.1 vs. 2.6, p=0.007, Welch Two Sample t-test), indicating higher T-cell activity in these treatment naive tumours.

ER^- tumours had higher CYT scores than ER^+ tumours (mean 3.0 vs. 2.6 p=0.02, Welch Two Sample t-test). The association between CYT score and response was only observed in ER^+HER2^- and $HER2^+$ tumours (Figure 3.16B). ER^-HER2^- tumours, despite having higher CYT scores compared to all other subgroups, did not have a statistically significant relationship between CYT score and response.

In order to validate the CYT score findings, the `ESTIMATE` R package [316], which used single-sample GSEA to compute the fraction of stromal and immune cells in tumour samples, was run on `voom` transformed RNA-seq counts. As part of its computation to infer tumour purity, `ESTIMATE` used a 141 gene immune signature to compute an *ImmuneScore*. Tumours that attained pCR had higher `ESTIMATE` ImmuneScores compared to those with RD (median score 807 vs. 186, p=0.005, Wilcoxon rank sum test, Figure 3.16C). This was statistically significant in $HER2^+$ tumours (p=0.02), with a trend for significance in ER^+HER2^- tumours (p=0.059), but not ER^-HER2^- tumours (p=0.8, Wilcoxon rank sum test), as shown in Figure 3.16D. The findings were similar to those seen between CYT score and response, although this was unsurprising, given the high correlation between CYT score and ImmuneScore (R=0.85, p=3e-43, Figure 3.16H).

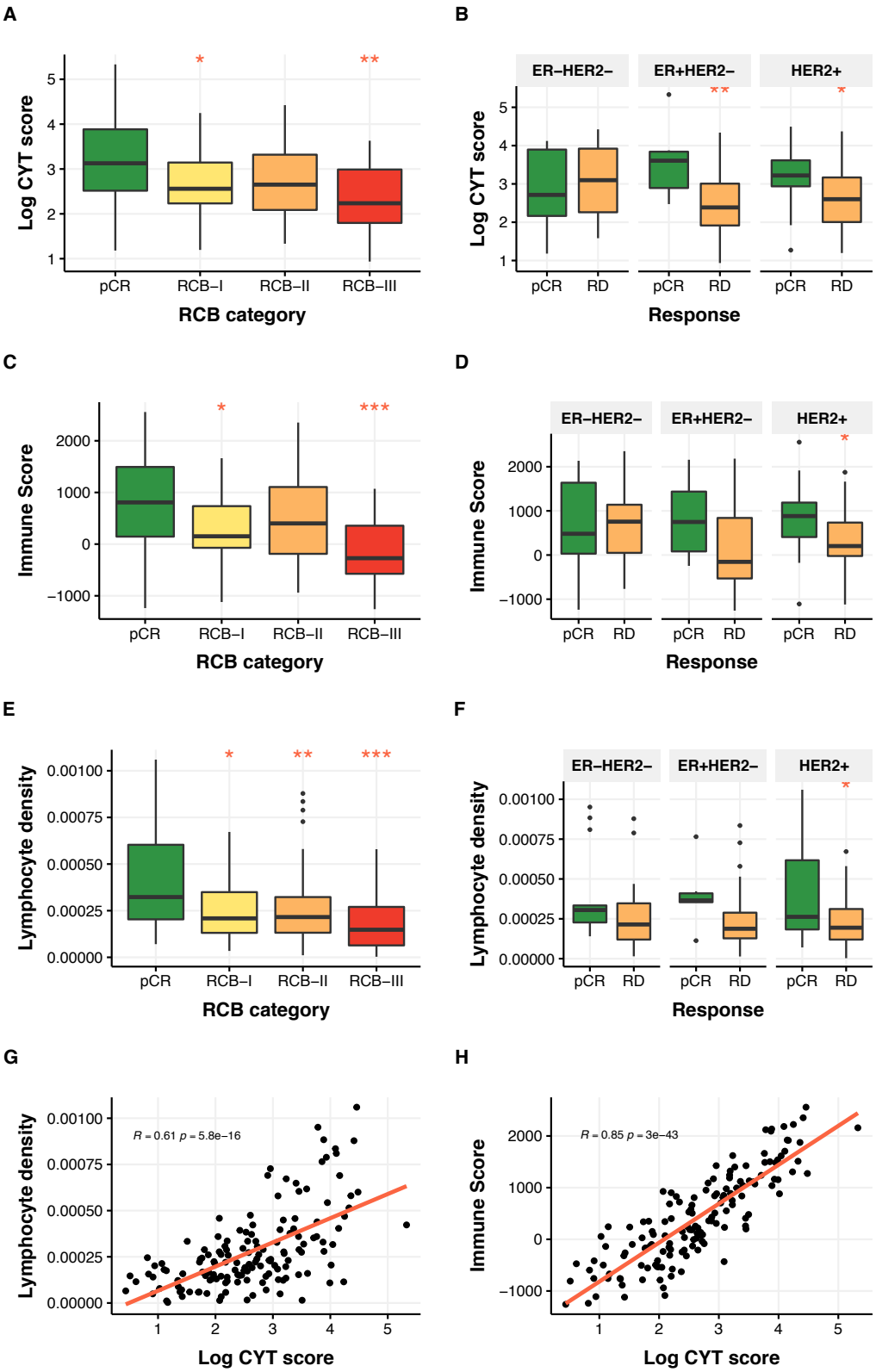


Figure 3.16: Association between metrics of immune infiltration/activation and response. (A) Box plots showing distribution of CYT score across different response categories. (B) CYT score distribution by ER and HER2 status. (C,D) `ESTIMATE` ImmuneScore and correlation with response. (E,F) Orthogonal validation using distribution of lymphocytic density across response groups. (G) Correlation between CYT and Lymphocyte density. (H) Correlation between CYT and ImmuneScore.

To validate these observations in an orthogonal way that was independent of RNA-seq enrichment strategies, lymphocyte density of the H&E sections taken from the fresh frozen biopsies that were subsequently sequenced was computed by ***Dr Ali Dariush*** using the method described by Ali et al. [11] The distance between the region analysed by digital pathology and that sequenced varied only by a few microns and hence was a reasonable approximation of the cellular populations sequenced. Lymphocyte density was higher in treatment naive tumours that attained pCR (p=0.0003, Wilcoxon rank sum test, Figure 3.16E). Once again, this observation was statistically significant in HER2$^+$ tumours, with a trend for significance in ER$^+$HER2$^-$ tumours (Figure 3.16F). Additionally, as seen with the ImmuneScore, there was a strong correlation between CYT score and lymphocyte density (Figure 3.16G), showing that most of the lymphocytes in these cases were cytotoxic CD8 T cells.

So as to define the interplay between proliferation, immunity and the four response subgroups, the STAT1 immune signature [68] was chosen as a representative signature of immune activation. *STAT1* is responsible for IFN-γ activation, which inhibits proliferation of tumour cells, and enhances their immunogenicity by increasing STAT1-dependent expression of MHC molecules [204]. GSVA scores were computed for the STAT1 immunity signature, and the correlation between GGI score and STAT1 score computed across all types of responses. The 2D density plot in Figure 3.17A shows the result of this analysis. Tumours that attained pCR had high proliferation and immune activation, with the maxima of both in the right upper proliferation^{high}/immune^{high} quadrant. As the degree of RD increased, the maxima of the density plot moved towards the lower left proliferation^{low}/immune^{low} quadrant, with tumours attaining pCR and those with RCB-III RD being diagrammatic opposites. The 2D density plot showed that the transition occurred in the RCB-II category, with some tumours in the proliferation^{high}/immune^{high} category and others in the proliferation^{low}/immune^{low} category. In order to determine whether there were different degrees of response in the RCB-II category that would explain this observation, the individual components of the RCB score were computed [279]: an association was observed between the diameter of the tumour bed at surgery and proliferation/immune interplay. RCB-II proliferation^{high}/immune^{high} tumours (29% of all cases) had smaller tumours at the time of surgery, whilst RCB-II proliferation^{low}/immune^{low} tumours (30% of all cases) were larger in size (Figure 3.17B).

To elucidate which cellular populations were responsible for the increased immune activation at diagnosis, immune and stromal cell populations were quantified from the bulk RNA-seq data using the R package `MCPcounter` [26] (Figure 3.18). The deconvolution method revealed increased T cells, NK cells, monocytes and myeloid dendritic cells in tumours

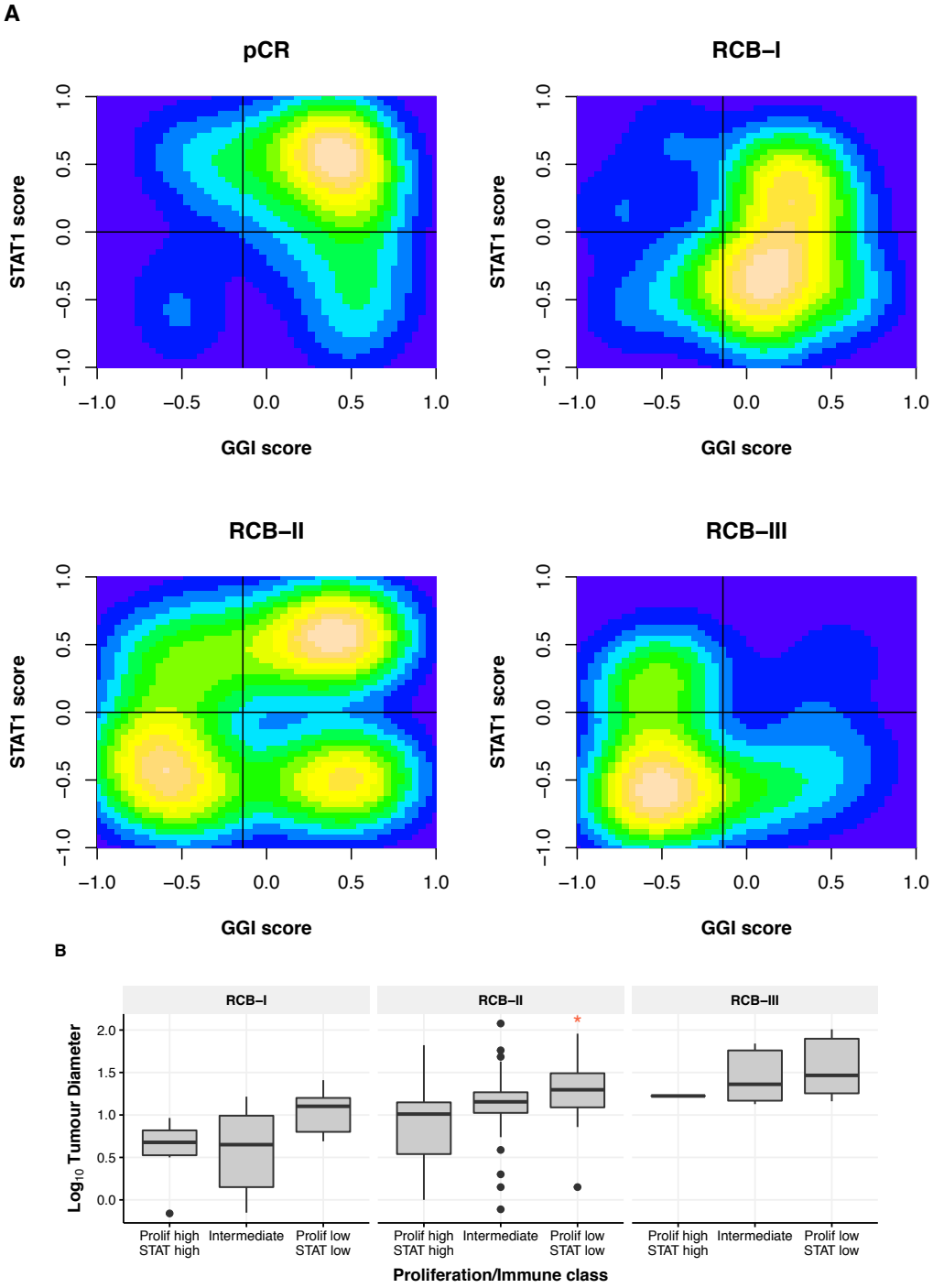


Figure 3.17: Association between proliferation (GGI) and immune (STAT1) pathways at diagnosis and response to chemotherapy. (A) Tumours that attained pCR had high levels of proliferation and immune activation, as opposed to tumours with higher burdens of RD. The yellow maxima can be seen moving from the upper right-hand corner in tumours that attained pCR, to the lower left-hand corner in tumours with high amounts of RD post-chemotherapy. (B) Log tumour diameter at surgery and correlation with proliferation and immune scores. The RCB-II transition point shown in (A) was secondary to a mixture of tumours with larger tumour diameters and low proliferation and immune infiltration, as well as smaller tumours with higher proliferation and immune activation.

that attained pCR. None of these associations were seen in ER$^-$ tumours, while all were statistically significant in HER2$^+$ tumours, with a trend for statistical significance in the ER$^+$ subgroup. To provide a more granular understanding of a larger number of immune cell populations, single sample GSEA was performed over the LM22 immune gene set [208] (Figure 3.19).The analysis confirmed enrichment of multiple immune cell types, including B cells, T cells, macrophages, neutrophils and NK cells, suggesting that most subtypes were up-regulated in tumours that attain pCR, including both immuno inhibitory and immune activating cells.

ER$^-$ tumours in all analyses did not show any differences in immune cell populations within the response groups, a finding corroborated by the lymphocyte density analysis. In order to assess whether immunoexhaustion correlated with response, a recently published signature [324] derived from deep single-cell RNA sequencing on 5,063 T cells isolated from blood, tumour, and normal tissues, was used to determine the degree of immunoexhaustion within the dataset (Figure 3.20). Tumours that attained pCR across the whole cohort had higher immunoexhaustion, probably secondary to the increased immune activation. A similar result was obtained using the immunoinhibitory STAT3 signature and immuno activatory STAT5 signature. Of note, ER$^-$ tumours did not show any difference in immunoinhibitory and immunoexhaustion signatures, suggesting that this does not play a role in explaining the absence of a relationship between immune activation and response at diagnosis.

3.4.4 Mapping transcriptomic differences across ER / HER2 subtypes

The analyses performed over the whole sample cohort revealed key differences between the different ER/HER2 subtypes. As a case in point, proliferation appeared to play a key role in response in HER2$^-$, but not HER2$^+$ tumours (Figure 3.14D). The degree of immune infiltrate correlated with response in ER$^+$HER2$^-$ and HER2$^+$ tumours, but not ER$^-$HER2$^-$ tumours (Figure 3.16B, D and F).

In order to delineate subtype specific predictors of response, differential analyses were performed separately for each of the three ER/HER2 groups.

In **ER$^+$HER2$^-$** tumours, low expression of driver genes *BCL2*, *FOXA1*, *FOXP1*, *IGF1R*, *RB1*, *RET*, *SPOP*, *XBP1* and *ZNF703* and high expression of *AMER1*, *AXIN1*, *B2M*, *BUB1B*, *CARD11*, *CCNE1*, *CDK6*, *CDKN2A*, *DNMT1*, *DNMT3A*, *EGFR*, *EZH2*, *IDH2*, *IKZF1*, *JAK3*, *MSH2*, *MSH6*, *MYC*, *MYCL*, *MYCN*, *NOTCH1*, *PRDM1*, *PTCH1*, *SKP2*, *SMARCA4*,

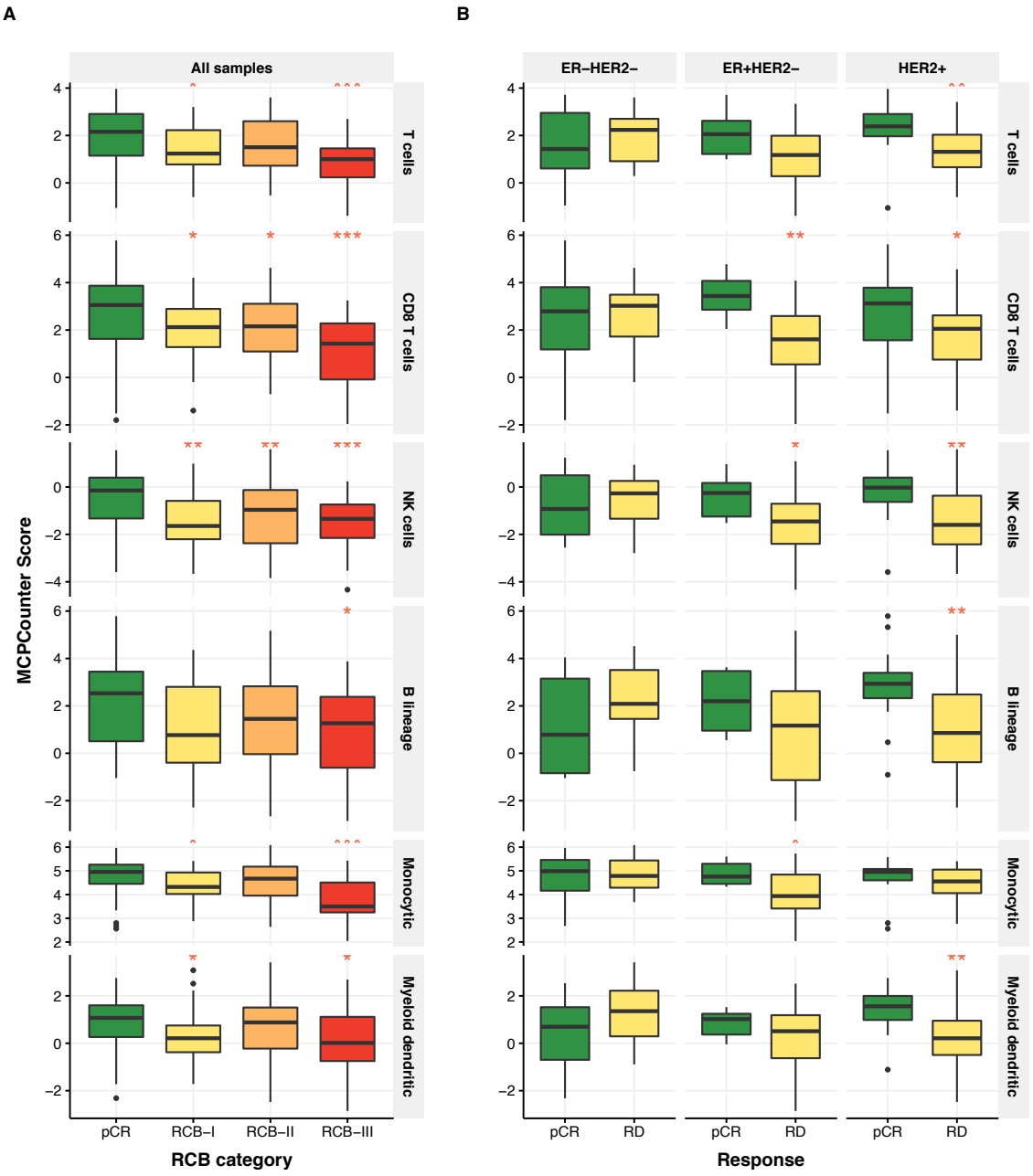


Figure 3.18: Deconvolution of immune cell populations using the `MCPcounter` R package. (A) Across the whole cohort tumours that attained pCR had higher T cells, NK cells, monocytes and myeloid dendritic cells. (B) Cohort split across ER and HER2 subtypes, showing the strongest association in $HER2^{+}$ tumours, and no association in $ER^{-}HER2^{-}$ tumours.

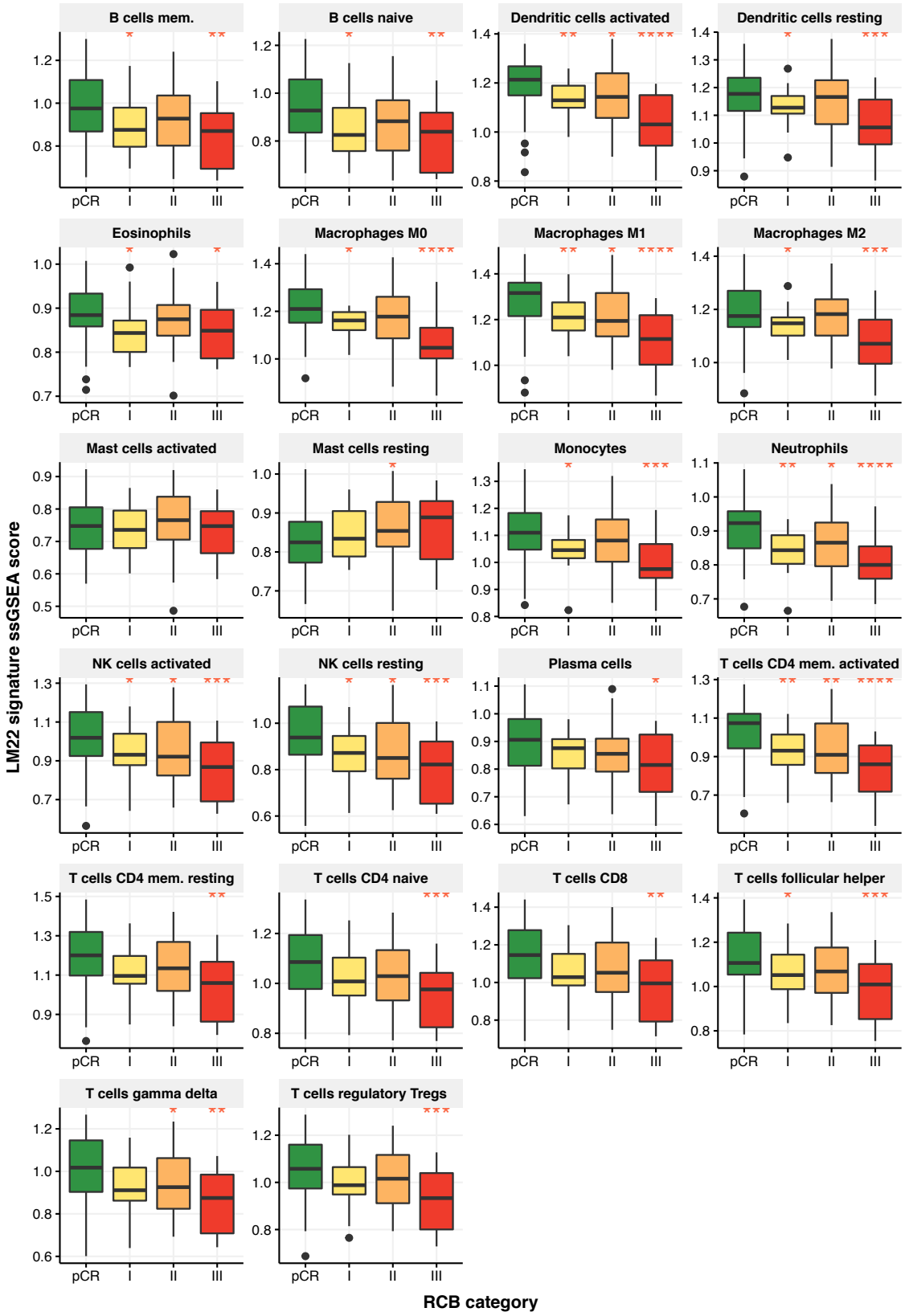


Figure 3.19: Deconvolution of immune cell populations using the LM22 gene signature set. Most immune cell populations were up-regulated in tumours that attained pCR, with a monotonic relationship observed across RCB categories.

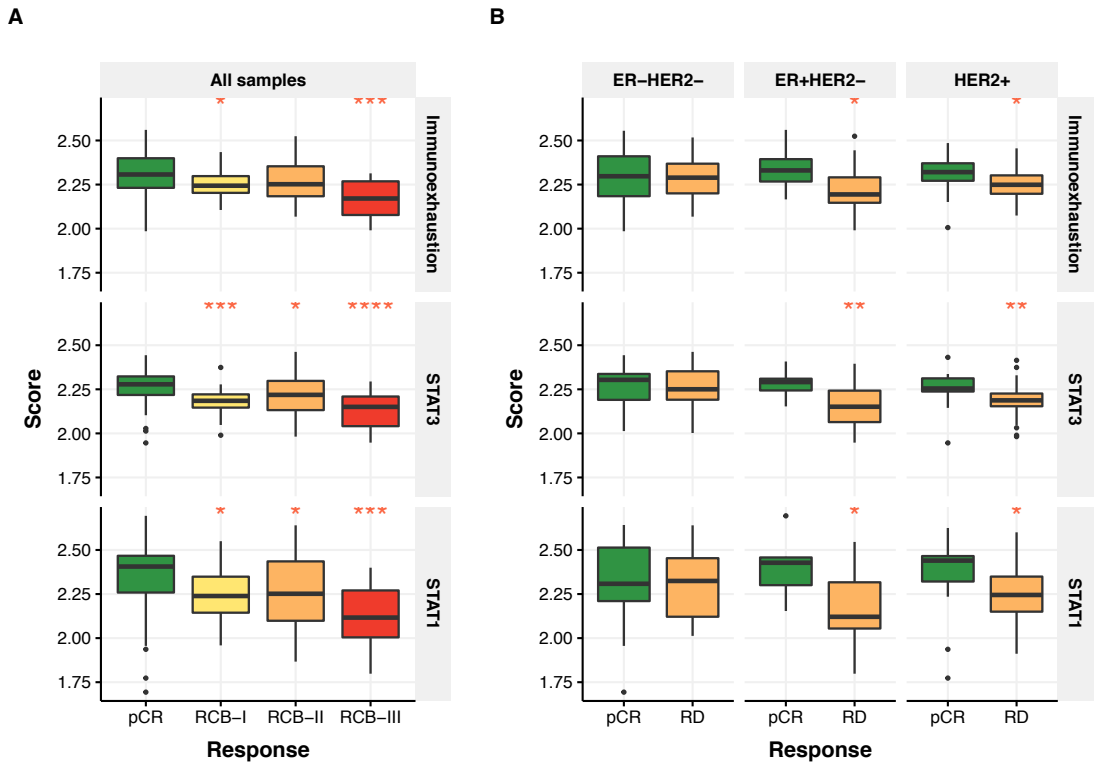


Figure 3.20: Single sample gene set enrichment over immunoexhaustive, immunoinhibitory (STAT3) and immunoactivating (STAT1) signatures across (A) whole cohort and (B) cohort subset by ER and HER2 status. ER^- tumours did not have a significant difference across the two response groups, unlike ER^+ and $HER2^+$ tumours

SMARCB1, *SMO* and *SOCS1* were associated with better response (Figure 3.21A). Additionally, ER^+HER2^- chemo-responsive tumours showed high expression of various immune genes, including both components of the cytolytic activity score (*GZMA* and *PRF1*), and increased expression of most MHC Class I and MHC Class II molecules (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DMA*, *HLA-DMB*, *HLA-DOA*, *HLA-DOB*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DPB2*, *HLA-DQA2*, *HLA-DQB2*, *HLA-DRA*, *HLA-E*, *HLA-F*, *HLA-H*, *HLA-K*, *HLA-L* and *HLA-V*, Figure 3.21B).

A GSEA run over the MSigDB GO (C5) and Hallmarks (H) gene lists revealed a strong correlation between immune pathways and response, more so that between proliferation and response (Figure 3.21C, D), indicating that in the ER^+HER2^- tumours in this cohort, the immune system played a stronger role than proliferation in determining response to treatment. A Reactome pathway enrichment (Figure 3.21E) confirmed these results, showing a very strong association between immune pathways and response.

In $\mathbf{ER^-HER2^-}$ disease, tumours that attained a better response had 43 highly expressed genes and 64 lowly expressed genes (Figure 3.22A), none of which were breast cancer driver genes. A Reactome GSEA (Figure 3.22B) showed significant enrichment over proliferative pathways, with a lesser degree of immune cell pathway activation. Enrichment over the Hallmarks and GO MSigDB datasets (Figure 3.22C and D) showed similar results. Hence, unlike ER^+HER2^- tumours, in ER^-HER2^- tumours proliferation mostly played a key role in response.

In $\mathbf{HER2^+}$ disease, tumours that attained a better response had 13 highly expressed genes and 29 lowly expressed genes. *LIN28A*, a stem cell regulator, was strongly highly expressed in tumours that attained pCR (logFC: 3.04, FDR: 0.007). *LIN28A* is known to be highly expressed in $HER2^+$ breast cancer [226] and its RNA-binding protein product binds to HER2 mRNA, leading to enhanced HER2 protein expression [79]. Hence, expression of *LIN28A* increases the amount of HER2 available on the cell surface, making tumours more sensitive to anti-HER2 therapies. A gene set enrichment analysis over the Hallmarks MSigDB dataset showed a very strong association between response and increased Interferon-α response (FDR=0.0027). In keeping with the enhanced immune response described earlier, a published T cell metagene comprising 6 immune genes [44] (*HLA-E*, *IRF1*, *CXCL13*, *GZMB*, *IKZF1* and *PRF1*) that has been shown to increase likelihood of attaining pCR also associated with response to treatment (p=0.006, Wilcoxon rank sum test).

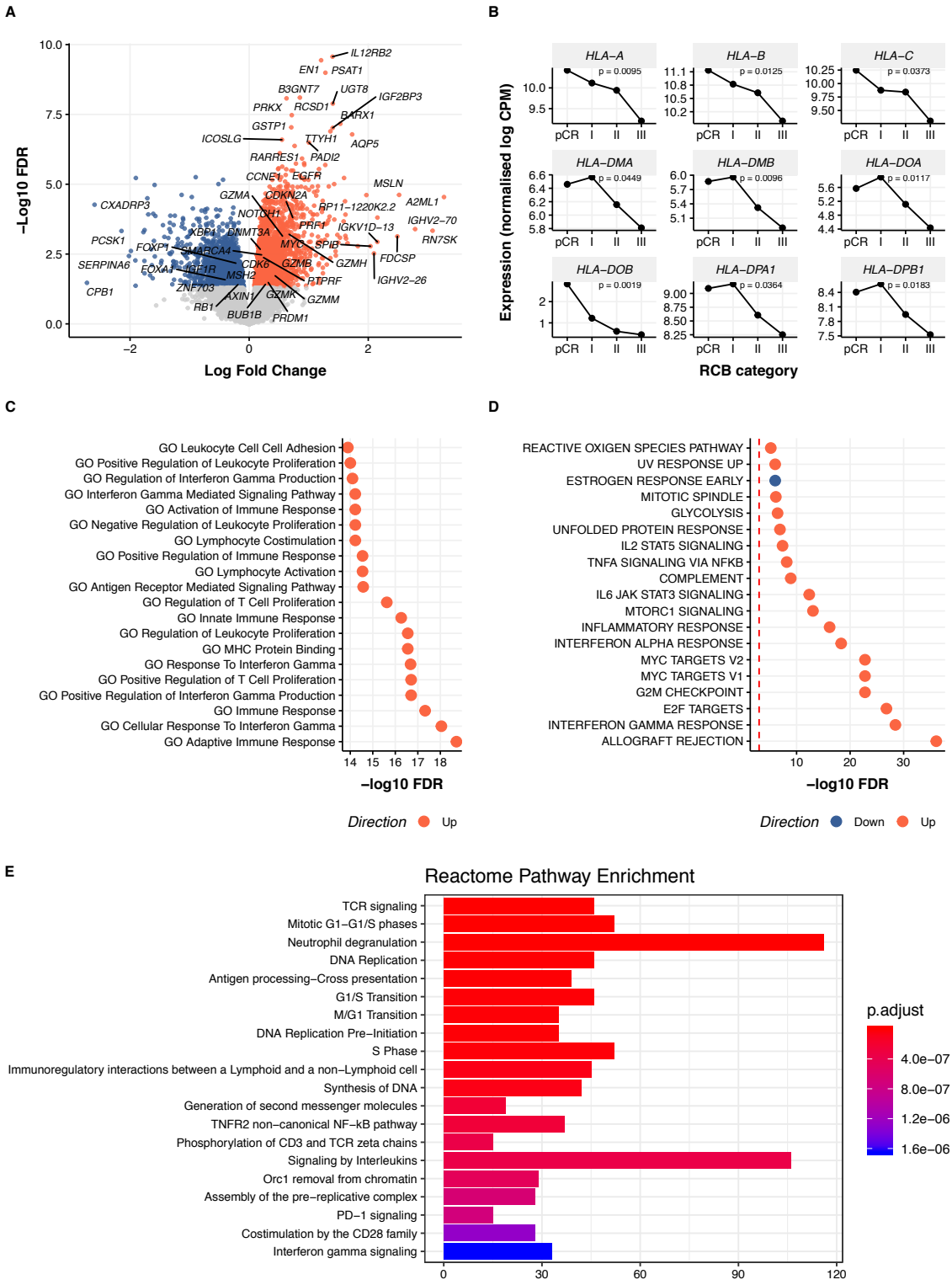


Figure 3.21: Transcriptomic differential expression in ER$^+$HER2$^-$ tumours. Red indicates highly expressed in pCR, blue indicates lowly expressed in pCR. (A)Volcano plot showing genes that were highly or lowly expressed in tumours that were more likely to attain a better response to chemotherapy. (B) HLA class I and II molecules were highly up-regulated in tumours that responded better to chemotherapy (C) GSEA over the GO MSigDB dataset, showing strong enrichment over immune system pathways. (D) GSEA over the Hallmarks MSigDB dataset, showing strong enrichment over proliferation and immune system pathways. (E) Pathway enrichment over Reactome, showing statistically significantly up-regulated pathways in tumours that attained pCR.

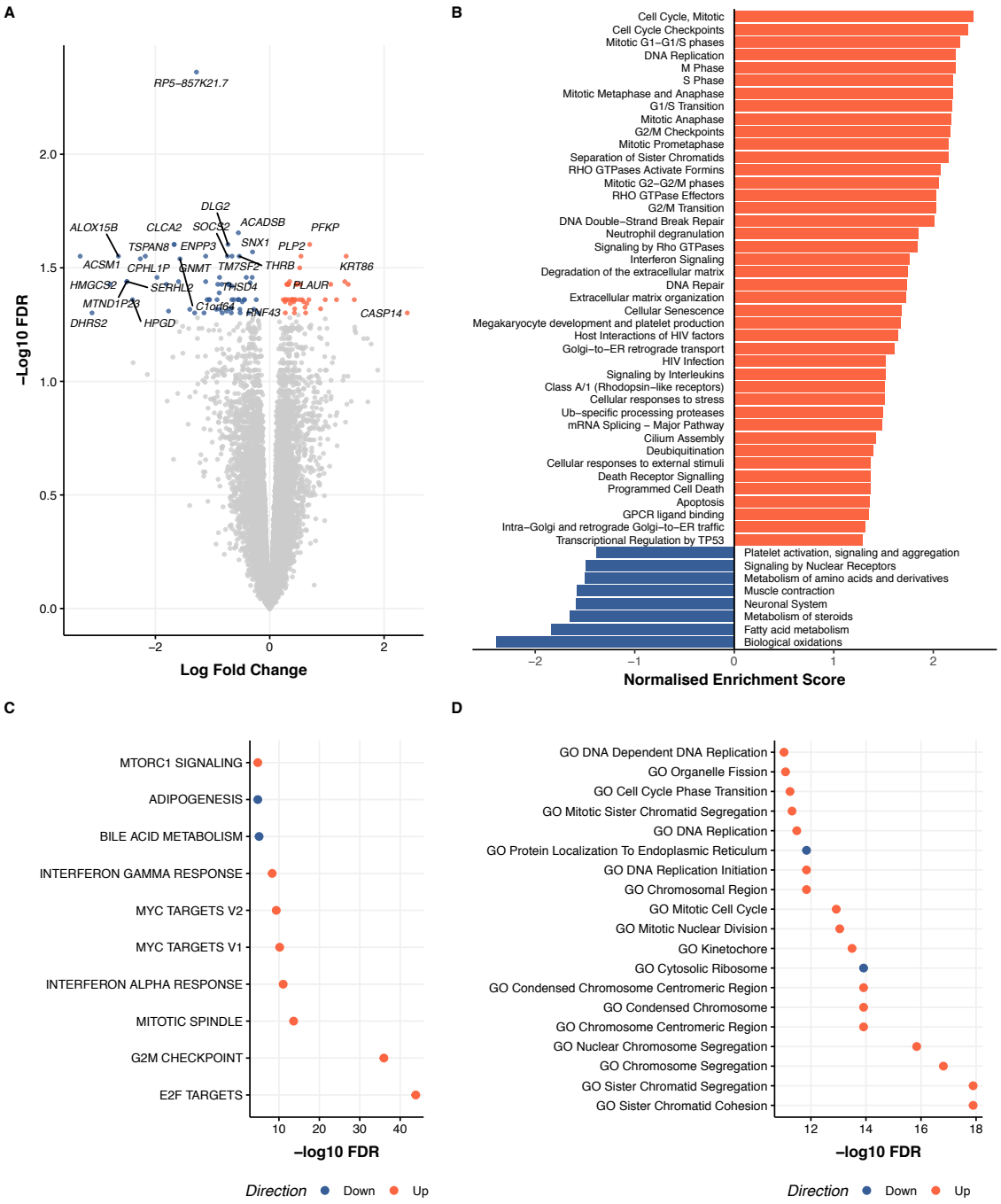


Figure 3.22: Transcriptomic differential expression in ER$^-$HER2$^-$ tumours. Red indicates highly expressed in pCR, blue indicates lowly expressed in pCR. (A)Volcano plot showing genes that are highly or lowly expressed in tumours that are more likely to attain a better response to chemotherapy. (B) Reactome GSEA, showing predominant enrichment over proliferation and *TP53* pathways. (C) GSEA over the Hallmarks MSigDB dataset, showing strong enrichment over proliferative pathways. (D) GSEA over the GO MSigDB dataset, showing strong enrichment over proliferative pathways.

3.4.5 Validation of established metagenes

As discussed in Chapter 1, various efforts have been made to identify metagenes predictive of response over the past decade. For example, based on RNAi screening experiments on cell lines exposed to paclitaxel, two gene sets were identified by Swanton et al. [278] that modulated resistance to taxanes: one involved in mitosis and mitotic spindle assembly checkpoint and another involved in the the metabolism of the pro-apoptotic lipid ceramide. This resulted in the development of a taxane response metagene, combining four genes (*BUB1B*, *CDK1*, *AURKB* and *TTK*) in a mitotic module and two genes in a ceramide module (*UGCG*, *COL4A3BP*) [148]. Overexpression of the mitotic module conferred sensitivity to paclitaxel, whilst overexpression of the ceramide module related to increased resistance to taxane therapy.

In order to assess whether the paclitaxel response metagene at diagnosis correlated to response, the metagene was derived using an approach similar to the one described in the original publication [148], where the geometric mean of the log2 expression of *BUB1B*, *CDK1*, *AURKB* and *TTK* was used to obtain a mitotic score, and the geometric mean of the log2 expression of *UGCG* and *COL4A3BP* used to obtain a ceramide score. The final response metagene was computed as the difference between both scores. In order to describe the predictive power of this metagene, the dataset was split into two groups: one in which patients received an anthracycline-based backbone in the first block of chemotherapy, followed by a taxane, and a second group which received taxanes first, followed by a switch to an anthracycline-based backbone. Any cases that did not receive block sequential treatment with an anthracycline and a taxane in any order were removed.

As shown in Figure 3.23A, expression of the ceramide, mitotic and taxane scores correlated with response in the subgroup treated with taxanes first, but not in those treated with anthracycline-based chemotherapy first. Indeed tumours treated with a taxane first that attained pCR had a much higher expression of the taxane sensitivity metagene, compared to those with RD post-chemotherapy (Figure 3.23B). A ROC analysis showed that the metagene predicted pCR with a specificity of 76%, sensitivity of 90%, accuracy of 80%, NPV of 96% and a PPV of 58% in tumours treated with a taxane first (Figure 3.23C), with a significant difference seen between tumours treated with taxanes first vs. those treated with anthracyclines first (p=0.001, DeLong's test).

The neoadjuvant Trial of Principle (TOP) [67] was designed to discover biomarkers of anthracycline response and resistance and used RNA expression data from 139 patients

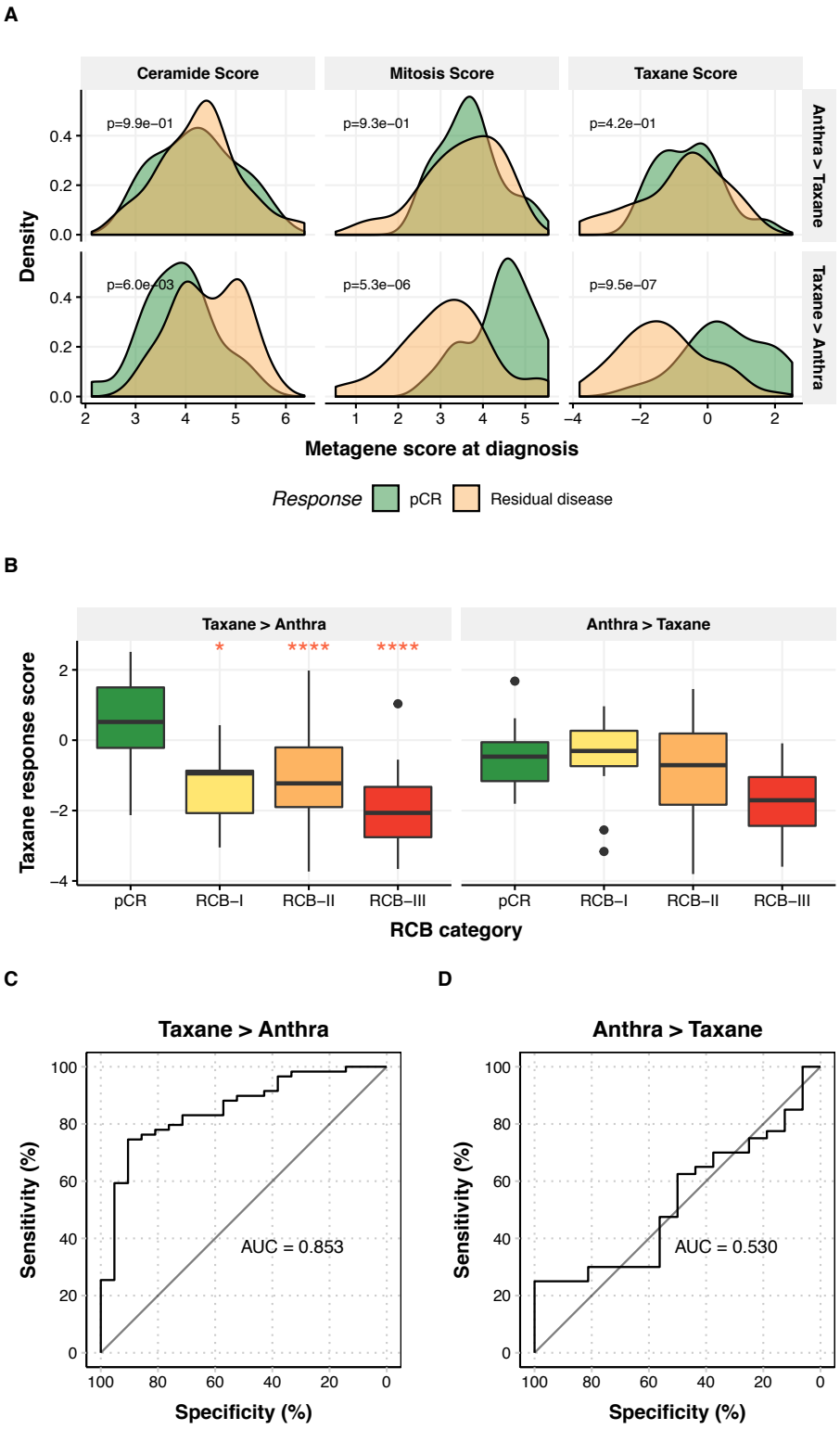


Figure 3.23: Assessment of the taxane sensitivity metagene in the TransNEO study. (A, B) Ceramide, mitotic and taxane score all correlated with response in tumours treated with upfront taxanes, but not in tumours treated with upfront anthracyclines. (C, D) ROC analysis showing the ability of the metagene to predict pCR in taxane-first and taxane-second treated tumours.

treated with epirubicin monotherapy to derive an anthracycline metagene score (TOP2A-score). This score was computed as the averaged sum of the genes on chromosome 17 from 35.37Mb to 36.06Mb (hg18 assembly), including Topoisomerase IIα, and was found to correlate with pCR in HER2$^+$ tumours. In order to assess the performance of this metagene in this dataset, the TOP2A-score was computed as described in the original publication: no association was found between TOP2A score and response. This is perhaps, unsurprising. In the TOP study patients received anthracyclines only, whilst in the TransNEO cohort most patients who received anthracyclines first had HER2$^+$ disease and therefore received sequential trastuzumab: this was not administered to patients in the TOP study. In addition, patients within the TOP trial only received single agent anthracycline, whilst none of the patients recruited to this study received single agent chemotherapy regimens.

3.4.6 Derivation of a response metagene

In order to identify a metagene that was predictive of treatment with T→FEC chemotherapy, the raw RNA-seq counts were normalised using a `voom` transformation and modelled using a nearest shrunken centroids (NSC) approach using the `voomDDA` R package [321]. Until recently, microarray technologies predominated the RNA field and most of the machine learning classifications were built to deal with the continuous data generated by these technologies. However, RNA-seq datasets are distinctly different from those generated by microarrays: counts are positive integers and should be modelled with different distributions. Hence microarray based classifiers do not perform as well on RNA-seq datasets. Various efforts have been made to develop comparable statistical methods to counteract this challenge [71, 93, 166, 325]. The `voom` normalisation method has been shown to have a better performance compared to count based methods, with classifiers using this normalisation method having a lower type-I error rate and low false discovery rate [166]. By using a sparse NSC classifier on voom normalised RNA-seq counts, [321] generated a statistical framework that has been shown to predict putative biomarkers given a set of conditions and a normalised counts expression dataset.

By using this approach, four putative biomarker genes associated with response across the whole cohort (Figure 3.24A) were identified, including:

1. *CA12* (Carbonic anhydrase XII), which encodes a zinc metalloenzyme responsible for acidification of the microenvironment of cancer cells and has been shown to correlate with ER expression [20]. Indeed, by regulating microenvironment pH, *CA12*

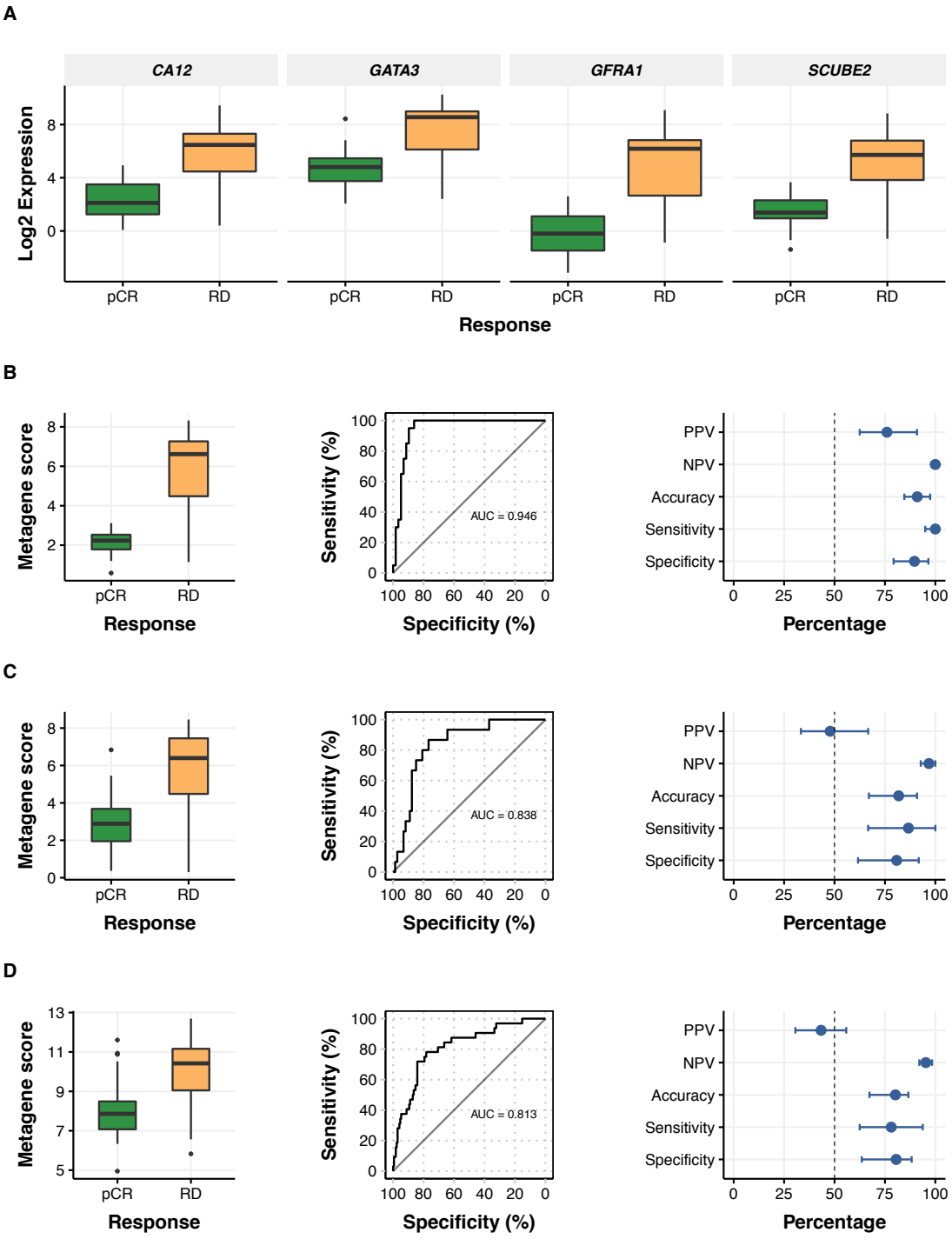


Figure 3.24: Evaluating a 4 gene metagene derived using an NSC classifier. (A) Expression of the four identified genes was consistently higher in the pCR groups. Validation of the classifier in (B) the TransNEO cohort, (C) ARTemis cohort and (D) MAQC-II microarray study.

enables tumour cell survival in hypoxic environments and inhibition results in increased apoptosis [178].

2. *GATA3*, a well characterised breast tumour suppressor gene, involved in ER signalling, although increased expression has also been observed in 43% of ER^- tumours [55].
3. *GFRA1* (GDNF Family Receptor Alpha 1), which mediates activation of the RET tyrosine kinase receptor and is often highly expressed in luminal A breast tumours [31]. Recent work has shown that it is responsible for resistance to aromatase inhibitor therapies in breast cancer [200] and to cisplatin chemotherapy in osteosarcoma [157].
4. *SCUBE2* (Signal Peptide, CUB Domain And EGF Like Domain Containing 2) is a tumour suppressor gene that inhibits tumour migration and invasion [173] and is part of the 21-gene *OncotypeDX* assay[272].

Within the TransNEO dataset, the metagene was able to predict pCR with a specificity of 89%, a sensitivity of 100%, an NPV of 100% and a PPV of 76% (Figure 3.24B).

Validation of the metagene in the ARTemis [72] neoadjuvant trial RNA-seq dataset (data generated during this doctoral work, but not shown), showed a specificity of 79%, a sensitivity of 87%, an NPV of 96% and a PPV of 47% (Figure 3.24B). Validation in a microarray dataset, the MAQC-II (GSE20194), showed a specificity of 81%, a sensitivity of 78%, an NPV of 95% and a PPV of 43% (Figure 3.24C), showing sustained high specificity and negative predictive value of the metagene in predicting response to neoadjuvant chemotherapy in a mixed population.

3.5 An integrated approach to predicting response

The work described in this chapter has identified multiple predictors of response, ranging from established clinicopathological features classically used in routine practice, to genomic and expression-based features.

While each of the identified molecular features could independently model different degrees of response, further work was undertaken to determine whether a combination of multiple features could be integrated into a model that outperformed individual predictors. All identified features associated with response were combined into one generalised linear model which was tuned over 10,000 iterations to minimise the Akaike information criterion (AIC) and thereby identify the optimum combination of features that strongly associated with response.

The model selection identified two clinical variables (age at diagnosis and number of positive lymph nodes), two genomic variables (expressed TMB and HLA-LOH) and five expression variables (GGI metagene expression, CYT score and *ESR1, ERBB2* and *LIN28A* expression) that, when combined together, were highly associated with response (Figure 3.25). The combination of these features outperformed any other combination. The choice of selected predictors was, perhaps, unsurprising. The contribution of the immune system in fashioning response to therapy was repeatedly referred to in this chapter and the inclusion of a T cell-mediated cytolysis metagene (CYT), HLA-LOH and expressed TMB in the final model reflected the importance immuno-activation and immuno-evasion played. The inclusion of a proliferation gene set (GGI) within the model was also to be expected, given that the cytotoxic agents used preferentially killed cells that were actively dividing. Additionally, *ESR1* and *ERBB2* expression was also very closely associated with response, in keeping with better responses observed in ER^- and $HER2^+$ tumours and greater degrees of RD post-therapy in ER^+HER2^- tumours. The unsupervised selection of *LIN28A* as a key marker of response within the model was unsurprising given the role it plays in inducing a stem-like phenotype and its role in stabilising HER2 mRNA, thereby playing a role in determining response to trastuzumab (Section 3.4.4).

The results of the ROC curve analysis for the derived model is shown in Figure 3.25. For each of the three response groups (pCR, chemosensitive and chemoresistant tumours) an analysis was performed initially on all tumours and subsequently across the three ER and HER2 subgroups. The model showed very good performance across the entire cohort and in the ER/HER2 stratified cohorts and outperformed the clinical model (Section 3.2). While

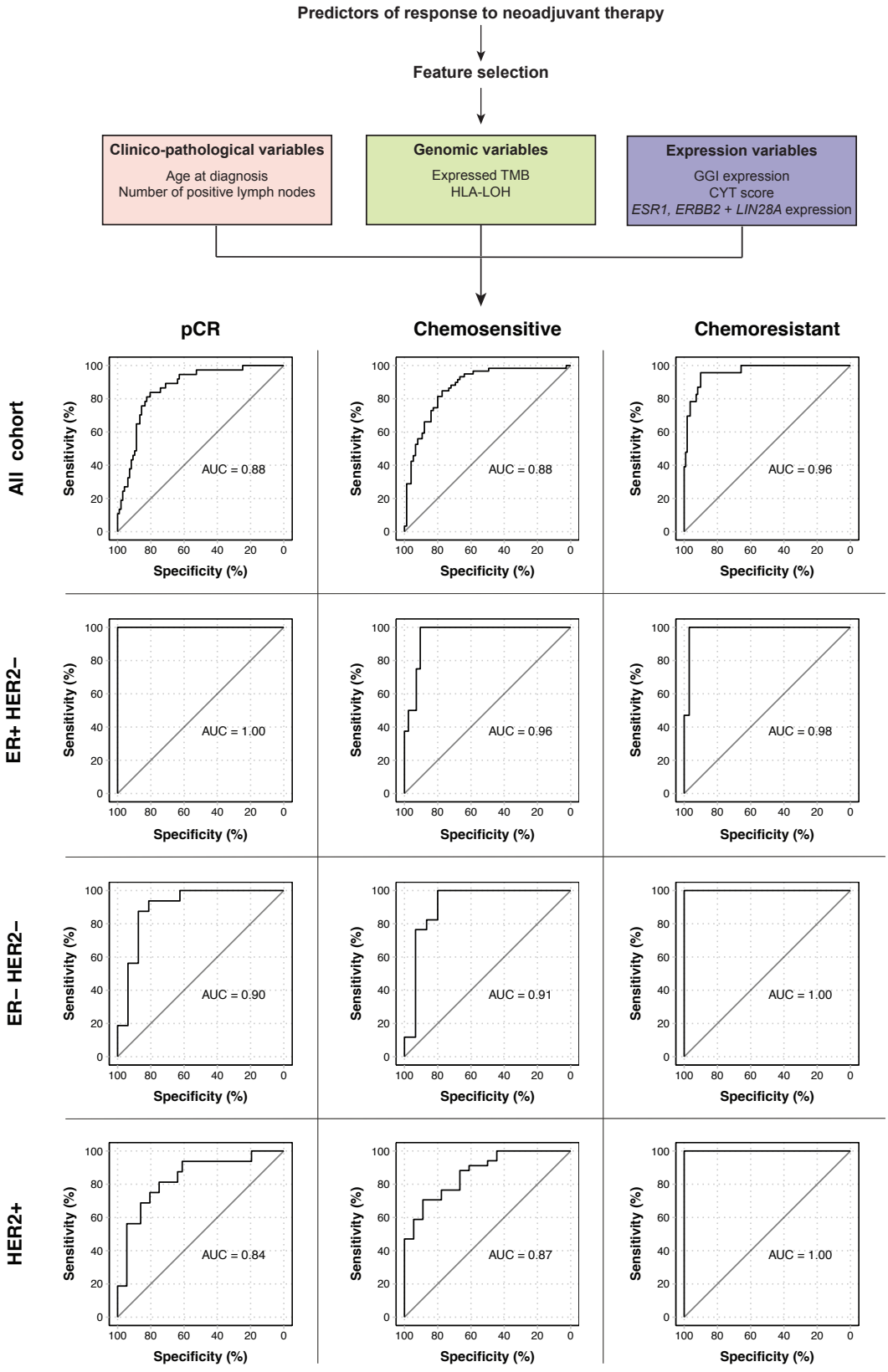


Figure 3.25: Combining clinical, genomic and expression variables predictive of response to predict pCR, chemosensitivity and chemoresistance.

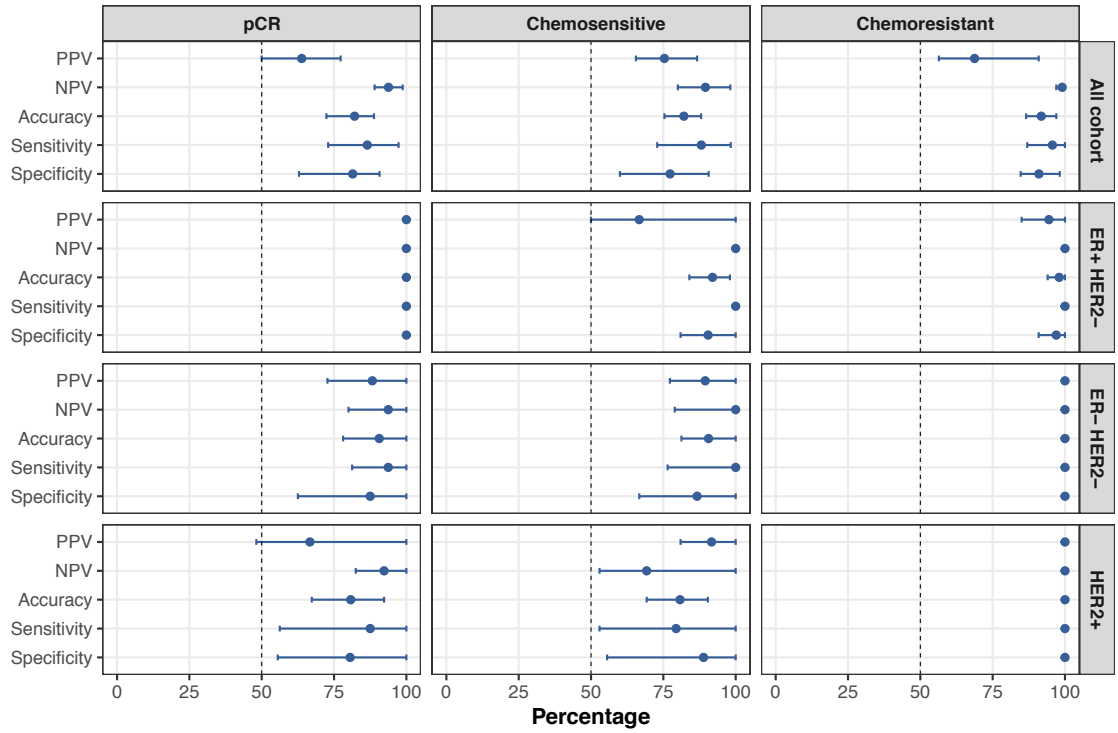


Figure 3.26: Performance metrics of the combined clinical and molecular model in predicting response to neoadjuvant therapy.

the AUC of 1 seen in the pCR $ER^{+}HER2^{-}$ model, as well as chemoresistant $ER^{-}HER2^{-}$ and $HER2^{+}$ models could be due to over-fitting secondary to the smaller number of samples in these categories, very good performance was observed in all other subgroups. Indeed, as shown in Figure 3.26, the predictor could predict pCR in the entire cohort with a sensitivity of 86.5%, specificity of 81.5% and an NPV of 93.4%, whilst it could predict chemoresistance with a sensitivity of 95.7%, specificity of 91.0% and an NPV of 97.0%.

Hence the combined predictor derived in this work was shown to be superior to the clinicopathological model described in the literature and allowed the classification of patients into those that would derive the most, as well as the least, benefit from neoadjuvant therapy. Undoubtedly, validation of this predictive model will be required: the *Personalised Breast Cancer Programme* currently being run in Cambridge and which aims to whole genome and transcriptome sequence tumours from every consenting breast cancer patient will provide the optimum dataset to test this predictor.

3.6 Discussion

As Louis Pasteur elegantly said: *"Dans les champs de l'observation le hasard ne favorise que les esprits préparés"*. Fortune does indeed favour the prepared mind: a better understanding of the biology of tumour response to chemotherapy and the identification of surrogates to predict chemosensitivity and chemoresistance allows the identification of patients who stand to benefit most, or indeed least, from treatment. Hence, subgroups that will not respond well to chemotherapy should be spared the morbidities induced by treatment if little benefit is to be gained. Additionally, knowledge acquired in the neoadjuvant setting is directly translatable to the adjuvant and metastatic settings.

The work done in this chapter adds to our understanding of the factors that are associated with response to neoadjuvant chemotherapy. In the first half of the chapter, key genes associated with response were identified. *TP53* driver mutations were not only the most prevalent across the cohort, but were also associated with increased chemosensitivity and decreased chemoresistance. An opposite relationship was seen with *PIK3CA*, where activating mutations were more likely to be associated with chemoresistance. Mutational signatures fashion the somatic landscape of tumours, and in this dataset there was significant enrichment of BRCA and APOBEC mutational signatures, with the BRCA signature predictive of response in the whole cohort analysis and ER^+ $HER2^-$ subset, and APOBEC signatures predictive in the ER^+ $HER2^-$ and $HER2^+$ cohorts.

The mutational landscape, however, was not the only genomic feature that correlated with response. Breast cancer is predominantly a copy number driven disease, and newer integrated classification systems predominantly use gene expression and copy number alteration patterns. One of these classifications, iC10, has been shown to associate very closely with overall prognosis. Tumours classified within cluster 10, typically characterised by ER^- tumours with 5q loss and gains in 8q, 10p and 12p, as well as high degrees of chromosomal instability, were more likely to attain pCR. iC5 tumours, with *ERBB2* amplification, were associated with increased chemosensitivity. Low proliferation tumours in the iC7 subtype characterised by 16p gain, 16q loss and 8q amplification were highly chemoresistant. Paradoxically, the aggressive highly proliferative ER^+ iC2 tumours harbouring amplification of 11q13 and 11q14 were associated with chemoresistance.

A targeted GISTIC analysis showed very clearly that amplification of oncogenes was associated with pCR, whilst loss of tumour suppressor genes and corresponding cell cycle arrest molecular machinery was associated with a higher likelihood of chemoresistance. *MYC,*

TBL1XR1, PARP1 and *PIK3CA* amplification were associated with pCR. It was perhaps surprising that *PIK3CA* amplification was associated with response, given that activating mutations within the gene were associated with chemoresistance. This finding was validated on transcriptomic analysis: *PIK3CA* overexpression was associated with increased chemosensitivity. Loss of tumour suppressor genes was associated with chemoresistance: this was typified by loss of FRA16D fragility site which encodes four tumour suppressor genes, including *WWOX* on which regulates the cell cycle checkpoint protein *ATM*. Additionally, loss of *ATM* and *CHEK1* on 11q also correlated strongly with chemoresistance.

By combining mutation data with copy number data, the degree of intra-tumoural heterogeneity could be determined. Increased intra-tumoural heterogeneity was noted to correlate with increasing chemoresistance: the greater the diversity of the clonal landscape, the greater the probability of a subclone harbouring a resistant phenotype. Tumours that attained pCR had more clonal mutations compared to those that had RD post-chemotherapy, but a lower percentage of subclonal mutations.

Early on within the chapter it was noted that increased tumour mutation burden was associated with response: the greater the number of mutations a tumour harboured, the more likely it was that less residual disease would be present at the time of surgery and at TMBs above 3 Mb^{-1} it was highly probable that a tumour would attain pCR. Two possible observations could explain this effect. Firstly, more proliferative tumours were more likely to have a greater number of mutations due to a higher rate of cell division and a greater chance of replication errors. As shown, highly proliferative tumours were more likely to be killed by cytotoxic chemotherapy and therefore the effect observed could be secondary to TMB being an indirect surrogate for proliferation. Secondly, an increased TMB translated into an increased neoantigen load. Indeed, there was a positive association between expressed TMB and response, as well as neoantigen load and response. The greater the number of neoantigens presented on the cell surface, the greater the probability of an immune response being mounted. In keeping with this observation, tumours that had more clonal neoantigens, as well as tumours that had neoepitopes that bound to more than one HLA allele were more likely to attain pCR. Additionally, the work done here provides evidence that LOH over the HLA locus confers resistance to treatment in breast cancer, an observation that has not, to date, been made yet. Indeed, it was very intriguing that tumours predominantly lost HLA loci that presented more neo-epitopes than the alternate allele retained. Furthermore, six HLA loci were also associated with response, with one of the HLAs associated with with chemoresistance (HLA-B*18:01) also associated with high HIV viral replicative capacity in adults and poorer overall prognosis.

The transcriptomic analysis showed distinct differences in the pathways activated as the degree of response to treatment increased. Tumours that were more likely to attain a better response had higher activation of proliferation and immune signatures, however the combination of both was required for a higher probability of attaining pCR. Indeed, high proliferation tumours that did not attain pCR often had lower degrees of immune activation, showing that an immuno-competent tumour microenvironment was required for maximum chemotherapy benefit. A detailed analysis of the immune component revealed that the cytolytic index was higher in tumours that attained pCR, and linearly decreased as the degree of RD increased. This observation was orthogonally validated by computing lymphocyte density scores using digital pathology analysis of the fresh frozen histological slides. Further deconvolution of the tumour microenvironment into constituent immune cell types showed a greater degree of T cell, B cell, NK cell and macrophage infiltration, amongst others, in tumours that eventually attained pCR, indicating that immune infiltrate at diagnosis was highly correlated with eventual response.

Following the observation that increased immune infiltrate and neoantigenic load at diagnosis correlated with therapy outcome, the expression of antigenic CTAs was explored. The expression of 11 CTAs was associated with response to treatment and the sum of all CTA gene expression strongly associated with the degree of response. As CTAs are expressed as part of a de-differentiation programme, a metagene associated with adult stem cells was assessed: once again, tumours that had a more stem-like phenotype were more likely to attain pCR during chemotherapy.

The expression data also allowed for the elucidation of mechanisms of resistance to treatment. Tumours with higher burdens of RD post-chemotherapy had higher expression of chemotherapy metabolising enzymes, as well as increased expression of drug efflux pumps and decreased expression of drug influx pumps before the commencement of any chemotherapeutic agent. Similarly, in ER^+ $HER2^-$ tumours, chemoresistance was associated with profound down regulation of most MHC class I and class II molecules, suggesting immuno-evasion as a potential mechanism of chemoresistance. In $HER2^+$ tumours, *LIN28A* expression was found to be a key mediator of chemosensitivity, with increased expression of this stem cell regulator, that binds to and stabilises HER2 mRNA, associated with decreased chemosensitivity.

An assessment of two published metagenes was undertaken. Previous work showed that ceramide metabolism played a key role in taxane resistance, and a metagene identified that correlated with response to taxane treatment. Indeed, assessment of this metagene reaffirmed its utility in predicting response to treatment, but only if taxanes were given prior

to anthracycline chemotherapy. Finally, using a machine learning approach, a four gene metagene comprising *GATA3*, *CA12*, *GFRA1* and *SCUBE2*, which was independent of ER status or proliferation and predictive of response to T→FEC treatment with a specificity of 89%, a sensitivity of 100% was derived.

Hence the work done in this chapter has described novel and established features associated with pCR and various degrees of chemosensitivity/resistance. As shown by an integrated classifier built based on all the predictive features described, the key components that were shown to accurately model response were immune-based (CYT, HLA-LOH, expressed TMB), proliferation-based (GGI) and, as expected given the body of literature, ER and HER2-based. Fitting these parameters to a generalised linear model resulted in the generation of a predictor that could accurately model response to chemotherapy: prospective validation will be required, though datasets that will allow this are already being generated through the Personalised Breast Cancer Programme being run in Cambridge.

Chapter 4

Modulation of tumour clonal architecture by neoadjuvant chemotherapy

Contents

4.1 Introduction

The somatic genomic landscape of breast cancer has been shown to be altered by neoadjuvant chemotherapy [18, 155] as well as hormonal therapy [197], with evidence of new mutations being detected during therapy, as well as the apparent loss of others.

The emergence of chemoresistant subclones and the appearance of new mutations has long been a debated subject, with studies reporting either acquired [69, 222] or adaptive [164] resistance to chemotherapy, albeit in differing types of tumours. In ovarian cancer, for example, platinum-based chemotherapy was shown to induce new somatic mutations with a distinct mutational signature and this was associated with acquired resistance [222]. On the other hand, in bladder cancer, resistance was attributed to the selection of pre-existing resistant clones [164]. A recent study [155] which described the evolution of triple negative breast cancer in patients undergoing neoadjuvant chemotherapy by using single-cell DNA and RNA sequencing showed that the newly-detected population of cells observed at the end of treatment was pre-existing and selected by chemotherapy, consistent with a punctuated model of evolution [94].

Despite the interest in the field, few studies have sought to describe, in detail, the genomic and transcriptomic landscape of breast cancer during neoadjuvant chemotherapy. The main published studies either concentrated on one subtype of breast cancer [18, 155], lacked a midway biopsy [18, 197], analysed a very small cohort of samples [155, 197], or did not have any integrated transcriptomic data [18]. Without a midway biopsy, formal characterisation of the gradual changes that occur during therapy is challenging, especially for tumours that attain pCR, wherein by the time tissue is taken at the time of surgery there is no evidence of any remaining tumour tissue, making the recapitulation of the events leading up to complete tumour death impossible. Secondly, the analysis of small cohorts makes large-scale inference of evolutionary change difficult.

The work in this chapter aims to describe, at great depth, the evolution of tumours during neoadjuvant chemotherapy through the use of sequential tumour biopsies taken prior to commencing therapy, midway through therapy, as well as on completion of chemotherapy. By integrating data from MRI imaging and digital pathology assessment, as well as through the deconvolution of tumour clonal architecture from bulk sequencing data and the construction of evolutionary phylogenies, the work in this chapter aims to recapitulate in detail the way chemotherapy sculpts the tumour's genomic architecture.

4.2 Quantification of response to treatment

Quantification of response to neoadjuvant chemotherapy is typically performed twice during treatment by means of ultrasound and/or MRI. Formal assessment of the degree of any remaining residual disease (RD) is performed on the surgical resection specimen using the RCB score (Section 1.3).

As chemotherapy was administered, a decrease in tumour bulk was to be expected as response ensued, with the degree of volume change correlating to chemosensitivity. All patients recruited to the TransNEO study had ultrasound imaging performed as part of their standard of care treatment, however tumour measurements were performed in one or two dimensions, rendering 3D volumetric analysis impossible to perform. 81 patients (45%) in the TransNEO study were also recruited to the TRICKS study led by ***Prof. Fiona Gilbert*** at the Department of Radiology, Cambridge, wherein Dynamic Contrast Enhanced (DCE) MRI of the breast tumour was performed at the same time biopsies were taken to facilitate the integration of genomic data with radiological features. For these 81 patients, three-dimensional tumour volumetric measurements at diagnosis, midway through treatment, and on completion of neoadjuvant chemotherapy were computed by ***Dr Ramona Woitek***.

By modelling three-dimensional tumour volumes on MRI, the effect of chemotherapy on breast tumour bulk could be visualised. As shown in Figure 4.1A, administration of chemotherapy resulted in a gradual decrease in breast tumour bulk, with a median volume of 3,960mm^3 at diagnosis (range: 927-48,512mm^3), down to a median volume of 1,025mm^3 at the midway time point (range: 0-26,507mm^3) and a median volume of 308mm^3 on completion of chemotherapy (range: 0-4,321mm^3). As confirmed multiple times by the numerous trials supporting the use of neoadjuvant therapy, treatment before surgery was indeed effective at reducing tumour bulk pre-operatively, though the degree of response varied greatly.

Most tumours attained a minimum volume decrease of 50% midway through treatment (Figure 4.1B), with significant shrinkage seen on completion of chemotherapy. Four tumours increased in volume shortly after commencing chemotherapy. Three of these (**T072**, **T076** and **T095**) were small at diagnosis, and an increase in volume resulted in a relatively large percentage of volume gained. **T068** increased in size by 2,450mm^3 after 9 weeks of chemotherapy, and on completion of treatment RCB-III RD was present, denoting a significant degree of chemoresistance.

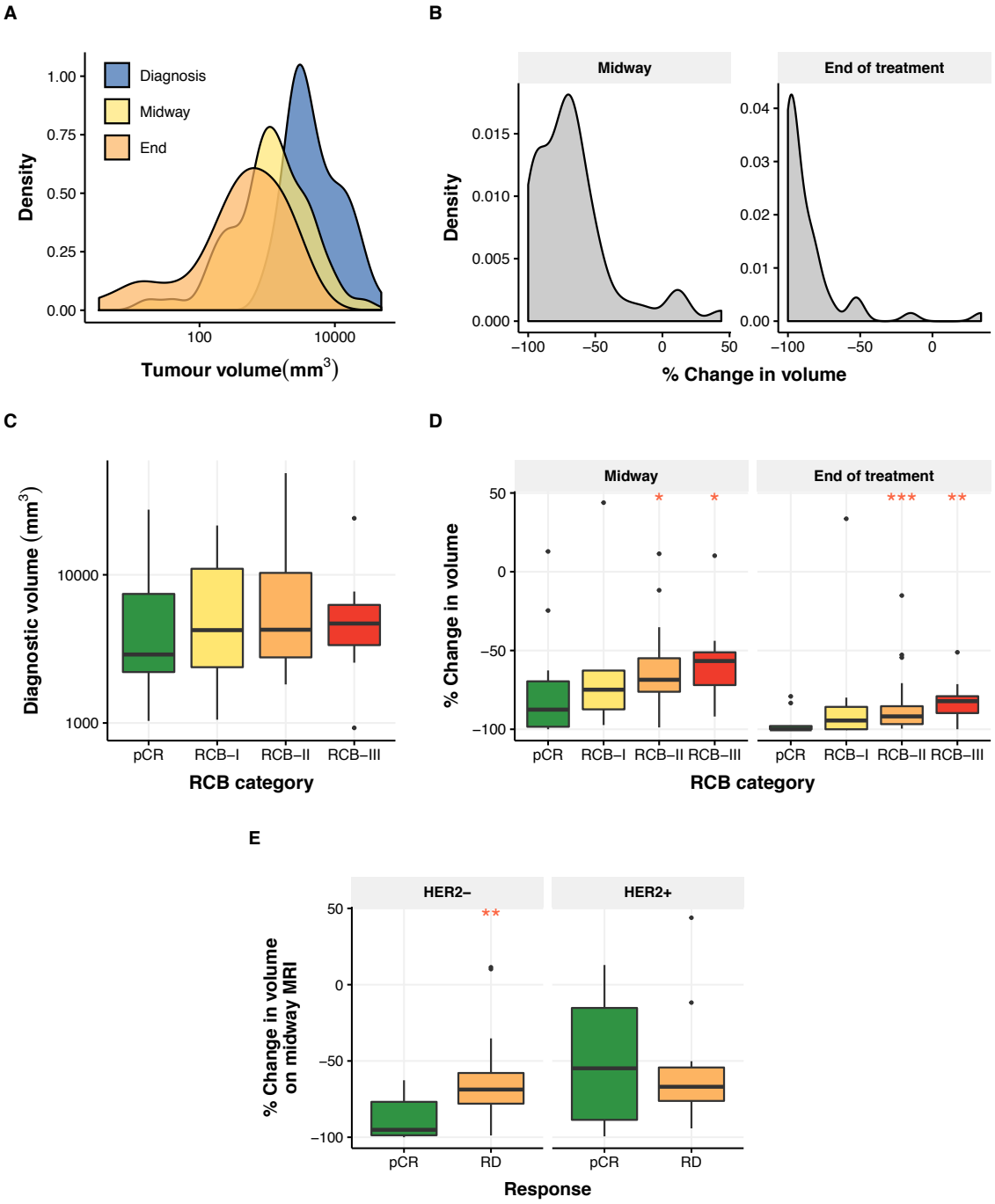


Figure 4.1: Response to chemotherapy as assessed by serial MRI scans in 81 patients. (A) Distribution of tumour volumes during chemotherapy. (B) Density plots showing percentage change in tumour volumes at the midway and pre-surgical time points. (C) Box plots showing distribution of tumour volumes at diagnosis and eventual response - no correlation observed. (D) Correlation between rate of volume change and pathological response to treatment in both the midway and surgical MRI scans. (E) The association between change in volume and pCR was seen in $HER2^{-}$ but not $HER2^{+}$ tumours.

The diagnostic tumour volume was not predictive of the degree of RD at surgery (Figure 4.1C), though by the midway MRI scan the rate of volume loss correlated with response to therapy (Figure 4.1D). Chemosensitive tumours attained a larger degree of volume loss than chemoresistant tumours: tumours that would eventually attain pCR decreased in volume by a median of 87%, as compared to: RCB-I (74%), RCB-II (68%) and RCB-III (56%) tumours. A similar trend was seen on the pre-surgical MRI scan. As the diagnostic tumour volume did not correlate with response, the effect seen was not confounded by initial tumour volume. Interestingly, the association between response and volume decrease was only observed in $HER2^{-}$ tumours. $HER2^{+}$ tumours decreased in bulk during chemotherapy, but those that attained pCR did not decrease in volume faster than those with RD at the time of surgery (Figure 4.1E), hinting that anthracycline use, and the associated absence of trastuzumab in the first part of the regimen, did not result in changes that were predictive of response in $HER2^{+}$ tumours by the midway time point.

Tumour purity provided an alternate method of assessing response to neoadjuvant chemotherapy, with decreasing purity a feature of increasing response to therapy (Figure 4.2A). Indeed, some of the earlier classifications of response, such as the Miller and Payne system [217] solely relied on tumour purity to grade response. 75 cases sequenced in the TransNEO cohort had a biopsy taken from the tumour bed halfway during chemotherapy and a further 51 cases had tissue taken from the tumour bed at the time of surgery. The median tumour purity of the sequenced samples at diagnosis, as assessed by a pathologist, was 50% (range 5-95%) and did not correlate with response to treatment. As chemotherapy was administered, by the midway time point, a global decrease in tumour purity was observed (median purity 30%), which decreased even further by the end of treatment (median purity 20% excluding samples with pCR). At the midway biopsy time point, 9 cases had tumour purities that were below 0.05: 3 of these cases eventually attained pCR (**T022**, **T037** and **T144**), 2 cases attained RCB-I RD (**T170** and **T175**), and 4 cases were low cellularity RCB-II tumours (**T056, T091, T137** and **T173**), thereby indicating that the decrease in purity was a function of response.

As observed with the MRI volumetric analysis, the greater the decrease in tumour purity by the midway and surgical time points, the lesser the degree of RD at the time of surgery (Figure 4.2B). This was statistically significant both by histopathological assessment (midway: p=0.028, surgical: p=0.011, ordinal logistic regression model) and digital pathology analysis of histological slides taken microns apart from the region that was sequenced (midway: p=0.017, surgical: p=0.02, ordinal logistic regression model).

These results showed that changes in tumour volume and purity at the midway and surgical time points both correlated with eventual RCB score. Hence, the RCB score was not merely a final snapshot of RD remaining after treatment but was also correlated with overall response to chemotherapy. Tumours with higher RCB scores were more likely to have less volume loss during treatment, an important observation that needs to be kept in mind throughout this work.

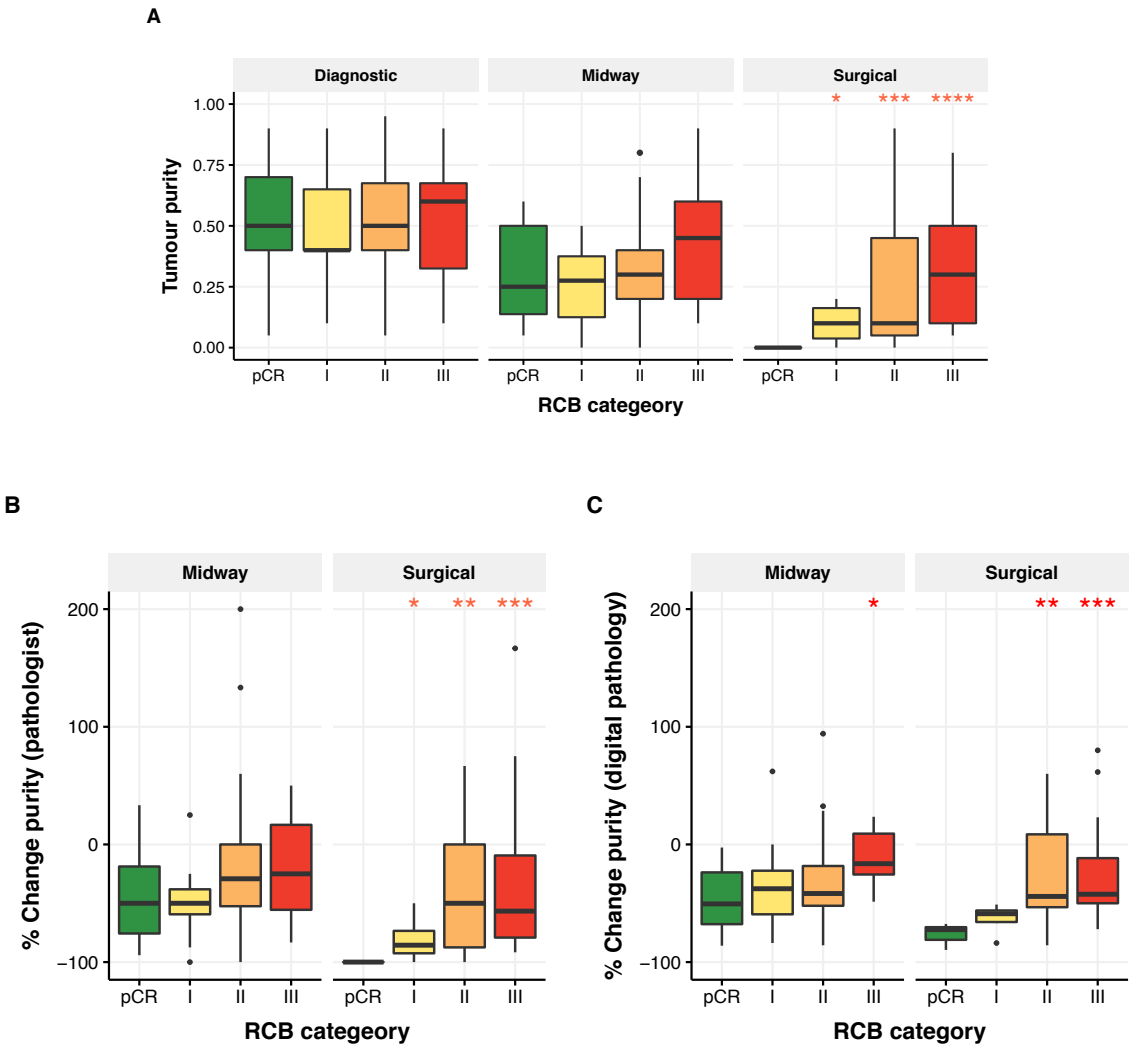


Figure 4.2: Tumour purity across the sequenced cases. (A) Tumour purity at all three sampling time points, as assessed by a histopathologist. (B, C) Change in tumour purity during treatment as assessed by an expert histopathologist and digital pathology analysis.

4.3 Early genomic landscape alterations

4.3.1 Mutation dynamics

Paired exome sequencing data at the diagnostic and midway time points were available for 75 cases. 13 of these had low cellularity at the midway time point, with median mutation allelic fractions less than 2%. 7 of these 13 cases (**T014, T022, T024, T028, T037, T104, T144**) attained pCR by the end of treatment and the decrease in tumour purity was attributed to early response to treatment. The remaining 6 cases had RCB-I (**T170, T175**) and RCB-II (**T056, T091, T137, T173**) RD post-treatment.

The tumour mutational profile was strikingly different midway through chemotherapy compared to that seen at diagnosis and showed great dynamicity, as shown in Figure 4.3. Across the 75 cases, 1,904 mutations that were present in the tumour exome sequencing data at diagnosis were no longer detectable at the midway time point (blue bars in Figure 4.3). Similarly, 499 previously unobserved mutations were detected at the midway time point (red bars) and 3,579 mutations persisted (green bars). Such variation could be due to (1) sampling bias due to tumour heterogeneity, resulting in spatially different clones being sampled at both time points [206], (2) death of chemosensitive cells carrying passenger mutations and altered selection pressures resulting in clonal expansion and increased prevalence of a pre-existing non dominant resistant clone [155] and (3) chemotherapy-induced mutagenesis [98].

The number of mutations that were present in the diagnostic biopsy but not detected at the midway time point correlated with the total mutation count at diagnosis (R=0.69, p=2.4e-11, Figure 4.4B), therefore the greater the number of mutations a tumour initially harboured, the greater the likelihood that an increased number of these mutations would not be detected at the second sampling, thereby indicating that these were probably subclonal.

In sharp contrast to chemoresistant tumours, tumours that attained pCR by the end of treatment had significantly fewer mutations detected by the midway time point (Figure 4.4A). 46% of all mutations detected at diagnosis were not detected by the midway time point in tumours that would eventually attain pCR, compared to 34% of mutations in RCB-I tumours, 22% of mutations in RCB-II tumours and 8.5% of mutations in RCB-III tumours.

The decrease in the number of mutations detected across the response subgroups could be attributed to tumour purity, as increasingly chemosensitive tumours were more likely to have a rapid decrease in tumour cellularity, and therefore as fewer tumour cells were sequenced

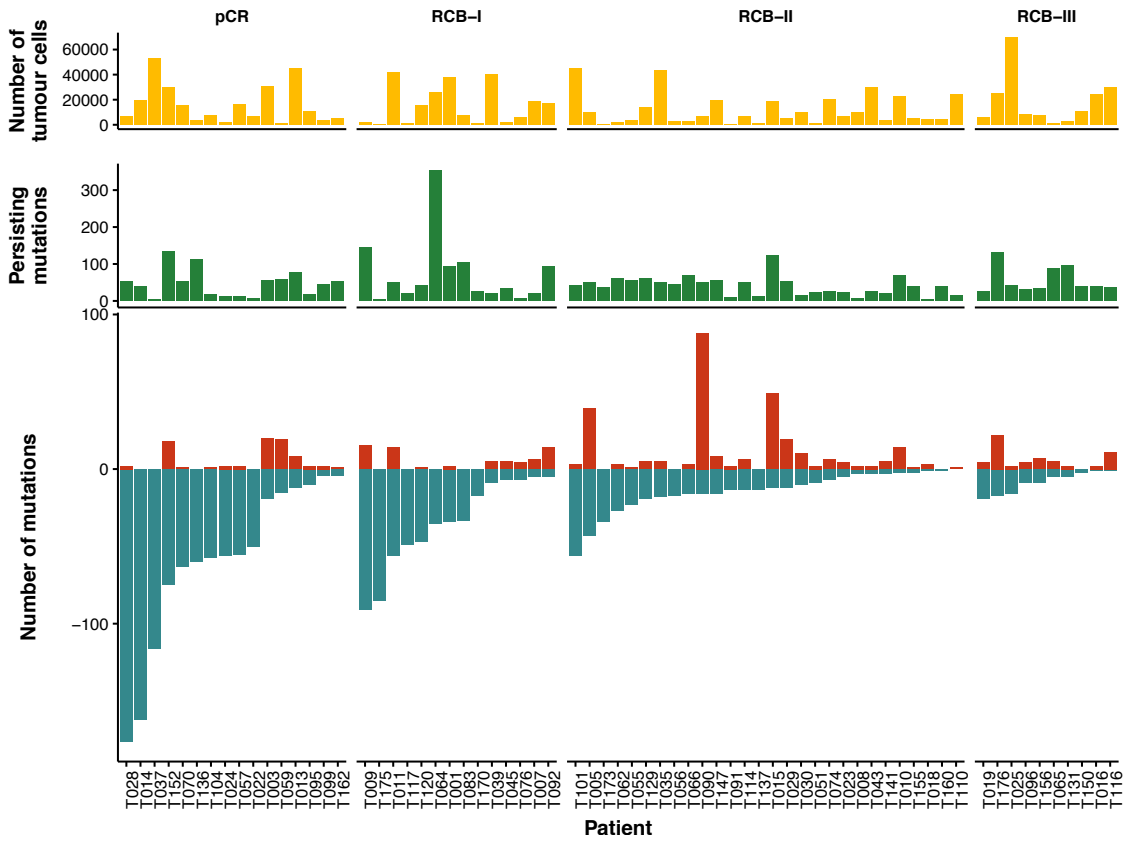


Figure 4.3: Somatic mutation landscape dynamics during chemotherapy. Tumours sequenced at the midway time point showed evidence of persisting mutations (green bars), mutations that had been detected in the diagnostic biopsy but were no longer detectable (blue bars), as well as mutations that had not been previously observed (red bars). The yellow bar plots show the distribution of the number of tumour cells sequenced at the midway time point, as computed by digital pathology analysis.

the ability to detect a mutation decreased. In order to explore this hypothesis further, the total number of tumour cells present within the sequenced samples was computed by digitally analysing the H&E slides taken a few microns adjacent to the tumour region sequenced [11] to determine whether the number of tumour cells sequenced correlated with the total number of mutations detected. The 13 cases with known low purity were removed for the purposes of this analysis. As shown in Figure 4.4D, there was no overall correlation between the number of tumour cells sequenced at the midway time point, and the number of persisting, new and undetectable mutations (R= 0.05, 0.07, 0.07 respectively). Additionally, there was no association between the number of tumour cells sequenced at diagnosis and the total mutation count (R=0.17 p=0.14, Figure 4.4C), indicating that the ability to detect a mutation was not always related to the number of cancer cells sequenced.

These results, while surprising, could be explained by a number of observations. Firstly, the exome sequencing data was relatively deep (>160×) and therefore rarer variants were more likely to be detected. Secondly, the variant rescue pipeline developed retained mutations with very low allelic fractions if they were convincingly detected at other time points. Hence, if a mutation was robustly called in the diagnostic sample and was present at very low AFs (< 1%) at another time point, that variant would be retained in both samples. Hence the detectable mutation load was not directly influenced by the number of cancer cells sequenced. Indeed, a closer look at Figure 4.3 shows that in some cases, such as **T014**, hundreds of mutations were lost and few mutations persisted, despite a large number of tumour cells being sequenced (top yellow bar graph). In some cases, such as **T005**, a large number of new mutations could be observed despite a lower number of cells sequenced: this might have been secondary to clonal replacement, with most of the sensitive cells eradicated by the midway time point, and the minor clone comprising chemoresistant cells increasing in number. Indeed, **T005** had a poor response to chemotherapy (RCB-II). In other cases, there was evidence of a large number of mutations lost despite a high number of cancer cells sequenced (eg **T011**, RCB-I).

The Jaccard coefficient was subsequently computed to determine the degree of genetic similarity between diagnostic and midway tumours. In keeping with Figure 4.3, tumours that attained better degrees of response were less similar to the originating pre-treatment tumour due to more rapid clonal eradication, whilst chemoresistant tumours remained more mutationally stable (Figure 4.4E).

While there was an association between the number of mutations lost and response to treatment, there was no statistically significant difference between the number of new somatic mutations gained and eventual response to treatment. Hence new mutations were observed at

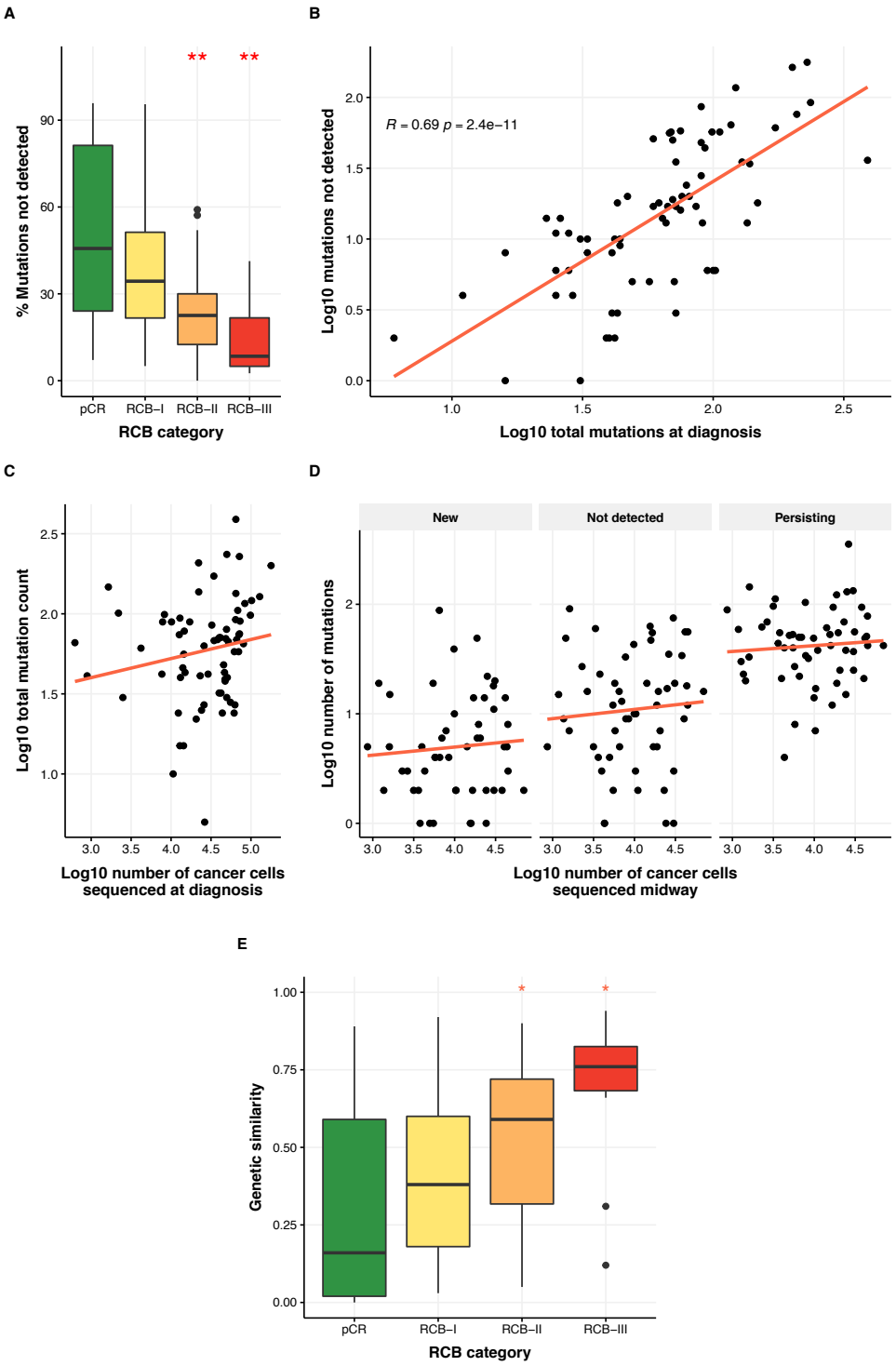


Figure 4.4: Mutational dynamics during treatment. (A) Bar plot showing the percentage of mutations that were not detected at the midway time point, compared to the diagnostic biopsy. The rate of mutation loss monotonically correlated with RCB category. (B) Scatter plot showing correlation between the total number of mutations at diagnosis, and the number of mutations that were no longer detectable in the midway biopsy. (C) Lack of correlation between number of cancer cells sequenced in diagnostic biopsy and total mutation load. (D) No correlation observed between number of tumour cells sequenced in the midway biopsy and the number of new/persisting/undetectable mutations. (E) Genetic similarity between serially sampled tumours as computed by using Jaccard coefficient. Tumours with higher degrees of RD were more genetically stable than those with less RD at the time of surgery.

the midway time point in all response groups, including in tumours that would eventually attain pCR, indicating that perhaps the chemoresistant cells harbouring these increasingly prevalent mutations were eradicated following a chemotherapy backbone switch.

In conclusion, by the midway time point, the mutational landscape showed evidence of differing dynamics that correlated with eventual response to therapy. The rate of mutation loss within 9 weeks of neoadjuvant chemotherapy was predictive of final response, with chemosensitive tumours having a greater percentage of undetectable mutations compared to chemoresistant tumours.

Subclonal dynamics

Within the 62 higher cellularity samples, 10 non-silent driver gene mutations that had been detected at diagnosis were no longer detected during therapy across 8 cases (Figure 4.5A). In 7 of these cases, the driver gene had a subclonal probability of greater than 50% (computed as described in Section 3.3.1 [191]), indicating that these were indeed subclonal driver mutations acquired later on during the tumour's evolutionary history. In one case (**T005**), the drivers were all predicted to be clonal on the diagnostic biopsy (subclonal probability < 50%). All but one (T101) tumours that lost a subclonal mutation had evidence of a concurrent clonal driver that was not eradicated by therapy (**T001**- *ATM, BRCA1, TP53*; **T005**- *BRCA2, ESR1, TP53* ; **T009**- *BRCA1, NRAS, PIK3CA,TP53*; **T039**- *CBFB, GATA3, PIK3CA*; **T062**- *BRCA2*; **T095**- *TP53*; **T152**- *TP53*). Additionally, 5 tumours had at least one newly detectable driver mutation appear during therapy (Figure 4.5B), with subclonal probabilities above 50% in 3 of these tumours, including case **T101** which had eradication of the *STAG2* subclonal driver mutation (subclonal probability: 100%) and emergence of an *ATM* driver mutation (subclonal probability: 100%). There was also evidence of the emergence of a previously unobserved *PIK3CA* driver mutation in case **T156** which was chemoresistant (RCB-III).

Similar dynamics were also observed within passenger mutations (Figures 4.5C and D). 72% of all mutations that were no longer detected on therapy had a subclonal probability greater than 50%, whilst 82% of all mutations that were newly-detected on therapy had a subclonal probability greater than 50%, confirming the robustness of the mutation calling and copy number pipelines, as well as the sampling techniques used.

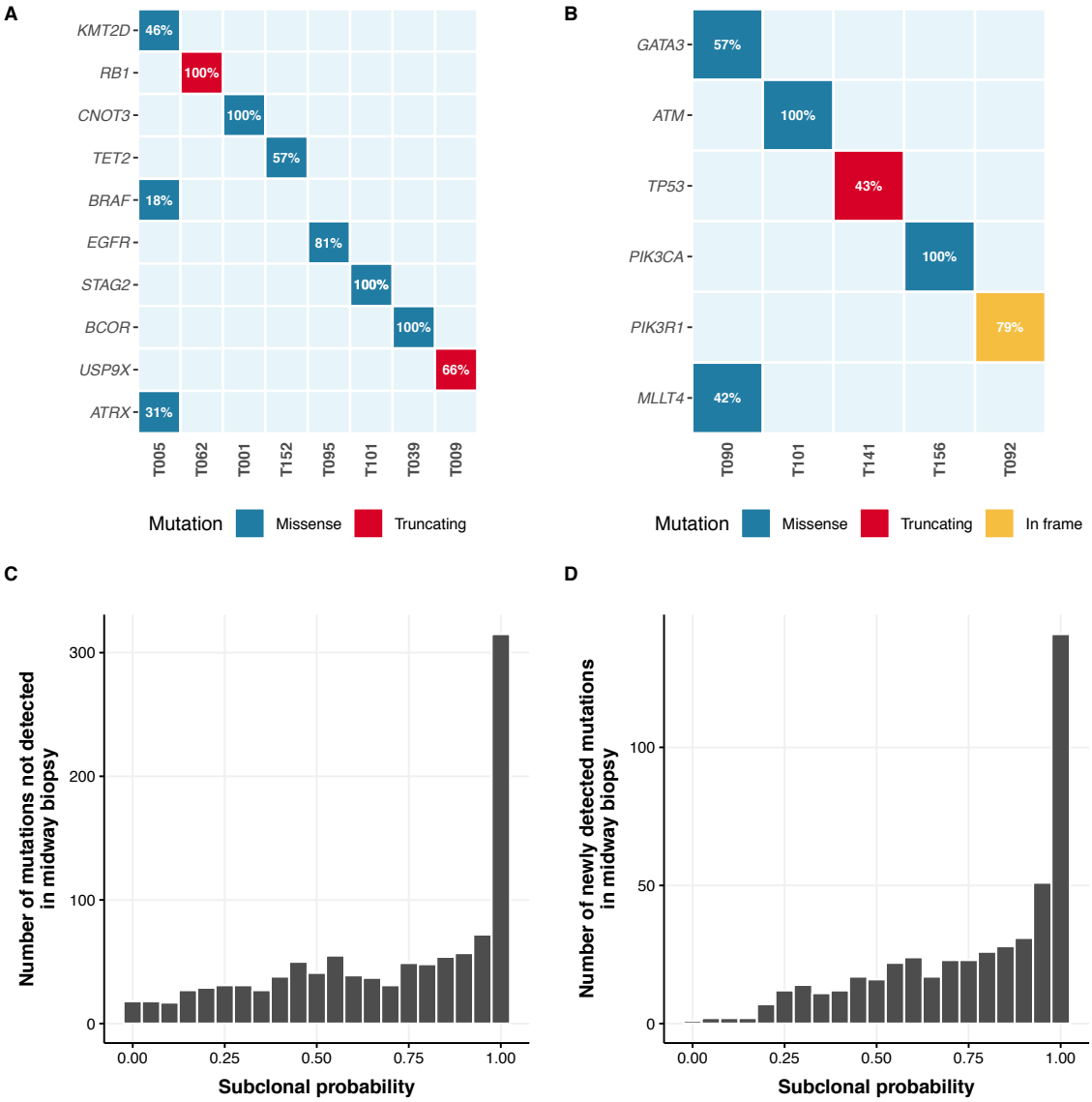


Figure 4.5: Driver and passenger mutation subclonal status. (A) Tumours with loss of non-silent driver mutations midway during therapy, with subclonal probabilities derived from diagnostic sample shown in tiles. (B) newly-detected driver mutations in 5 cases, with subclonal probabilities derived from midway sample shown. (C, D) Distribution of subclonal probability across all mutations that were no longer detected (C) or newly-detected (D) during therapy.

In summary, driver and passenger mutations showing differing dynamics during therapy were often subclonal. In all but one case, loss of a subclonal driver mutation was associated with the persistence of a clonal driver mutation.

Pathway enrichment

To elucidate whether mutations that were newly-detected or undetectable during therapy enriched for specific cellular pathways, an over-representation analysis of both gene sets was performed using the `ConsensusPathDB` tool from the Max Plank Institute [149, 150]. An enrichment analysis was done on Reactome, KEGG, PID and Biocarta, retaining pathways that had a minimum of four mutated genes, with statistical significance ascertained below an FDR cut-off of 0.05 to decrease the number of false positive hits.

An enrichment analysis over 222 non-silent coding mutations that were newly-detected at the midway time point (comprising: 21 frame-shift insertions/deletions, 10 in-frame insertions/deletions, 171 missense mutations, 15 nonsense mutations and 5 splice site mutations) showed a statistically significant over-representation of 36 biological pathways (Table 4.1) across three databases. The top pathway hits were high level immune system pathways (eg TCR signalling in BioCarta, BCR and CXCR-4 signalling in PID), proliferation pathways (cell cycle, G2/M transition/mitosis in Reactome) as well as molecules involved in solute transport (Transport of small molecules, Ion channel transport in PID). This was very intriguing, considering that proliferation, immune pathways as well as ATP binding cassette multi-drug transporter expression were all shown to play a key role at diagnosis in ascertaining response to treatment (Chapter 3). For example, *TAB1* (TGF-β Activated Kinase 1 Binding Protein 1), which harboured a newly-detected missense mutation during therapy, mediates intracellular signalling pathways induced by TGF-β, IL-1, and WNT-1, and any activating mutations would mimic the effect of increased IL-1, which is a pro-tumourigenic cytokine that promotes angiogenesis and proliferation [75].

Enrichment was also noted over biological pathways involved in genes responsible for cell motility and invasion, such as RAC1 cell motility signalling (BioCarta) and CDC42 signalling (PID). Both Rac1 and Cdc42 are Rho GTPases that participate in cell migration, with filopodia formation regulated by Cdc42 and formation of lamellipodia regulated by Rac1 [165]. Hence, if the mutations observed were activating mutations (as seen with *PIK3CA*, which was involved in both pathways) then this might suggest increased metastatic potential. Additionally, the pathway analysis showed over-representation of the *Angiopoietin receptor*

Table 4.1: Pathway enrichment on newly-detected mutations

Database	Pathway	FDR	Overlapping Genes
BioCarta	T Cell receptor signalling pathway	0.003	*ARHGAP4 ASAP1 MAPK8 PIK3CA PIK3R1 PTPRC*
	RAC1 cell motility signalling pathway	0.003	*ARHGAP4 ASAP1 PIK3CA PIK3R1 PPP1R12B*
	Phospholipids as signalling intermediaries	0.011	*PIK3C2G PIK3CA PIK3R1 SPHKAP*
	Actions of nitric oxide in the heart	0.016	*PIK3CA PIK3R1 PRKAR2A RYR2*
PID	Ephrin B reverse signalling	0.005	*FGR MAPK8 PIK3CA PIK3R1*
	Reelin signalling pathway	0.007	*MAP3K11 MAPK8 PIK3CA PIK3R1*
	N-cadherin signalling events	0.012	*GRIA2 MAPK8 PIK3CA PIK3R1*
	IL1-mediated signalling events	0.012	*MAPK8 PIK3CA PIK3R1 TAB1*
	p75(NTR)-mediated signalling	0.016	*MAPK8 PIK3CA PIK3R1 PLG TP53*
	Angiopoietin receptor Tie2-mediated signalling	0.024	*MAPK8 PIK3CA PIK3R1 PLG*
	CXCR4-mediated signalling events	0.026	*FGR PIK3CA PIK3R1 PTPRC SSH1*
	Signalling events mediated by focal adhesion kinase	0.029	*ASAP1 MAPK8 PIK3CA PIK3R1*
	BCR signalling pathway	0.035	*MAPK8 PIK3CA PIK3R1 PTPRC*
	CDC42 signalling events	0.036	*MAP3K11 MAPK8 PIK3CA PIK3R1*
Reactome	Transport of small molecules	0.029	*ABCA8 ABCF1 ATP12A ATP1A2 ATP2C1 EIF2S1 FLVCR1 PRKAR2A RYR2 SLC26A3 SLC6A15 SLC6A6 SPG7 TRPM3 TRPV3 TRPV4*
	Ion channel transport	0.029	*ATP12A ATP1A2 ATP2C1 RYR2 TRPM3 TRPV3 TRPV4*
	Cell Cycle	0.029	*CASC5 CCNA2 CKAP5 ESCO1 FBXO5 MZT1 PCNT PPP1R12B PRIM1 SGOL2 SPO11 TAOK1 TP53 TP53BP1*
	G2/M Transition	0.029	*CCNA2 CKAP5 MZT1 PCNT PPP1R12B TP53*
	Mitotic G2-G2/M phases	0.029	*CCNA2 CKAP5 MZT1 PCNT PPP1R12B TP53*
	Respiratory electron transport	0.030	*MT-CO1 MT-CO3 MT-CYB MT-ND3 MT-ND4*
	Cell Cycle, Mitotic	0.036	*CASC5 CCNA2 CKAP5 ESCO1 FBXO5 MZT1 PCNT PPP1R12B PRIM1 SGOL2 TAOK1 TP53*

Tie2-mediated signalling pathway (PID), with missense mutations in *MAPK8, PIK3CA* and *PLG* and an in-frame deletion of *PIK3R1* responsible for this signal. All of these mutations were predicted to alter protein function according to *SIFT* and *PolyPhen*. This pathway has also been associated with increased breast cancer invasion [152] and will be further explored in Chapter 5. Hence this analysis provided an insight into pathways that could potentially be associated with chemoresistance during neoadjuvant therapy.

Enrichment over mutations that were no longer detected on therapy revealed 32 pathways, although these were less specific. Pathways varied from muscle related pathways (such as muscle contraction, dilated cardiomyopathy and hypertrophic cardiomyopathy), metabolism of steroids and hormonal secretion. The lack of distinct enrichment could be attributed to the fact that most of the mutations that were no longer detected were passenger mutations, rather than oncogenic mutations, and therefore any signals originating from true pathways of chemosensitivity would be masked by the even larger number of passenger mutations.

In summary, newly-detected mutations enriched for immune system, cell cycle, solute transport and cellular motility pathways, suggesting that aberrations within these pathways could play a role in inducing resistance to neoadjuvant therapy.

Mutational signature dynamics

To investigate whether the altered mutation dynamics were driven by specific mutational signatures, the percentage change in mutational signature contribution was computed in all samples at the diagnostic and midway time points (Figure 4.6A). Across the whole cohort, there was no statistically significant change in mutational signature contribution between the first and second time points. Hence the overall mutational signature landscape remained stable. Signature 15, which was associated with defective DNA mismatch repair and enrichment of short insertions and deletions at mono/polynucleotide repeats, was the only signature that showed a trend for increased contribution during chemotherapy (p=0.06, one side Student t-test to assess difference from 0). An analysis of the expression of DNA mismatch repair genes (Figure 4.6C) showed that the expression of *MSH2*, *MSH6* and *PMS1* were significantly lower at the midway time point compared to the diagnostic time point (*MSH2*: diagnosis vs. midway: 23TPM vs. 15TPM, p=1.3e-06, *MSH6*: diagnosis vs. midway: 17TPM vs. 12TPM, p=4e-08, *PMS1*: diagnosis vs. midway: 10TPM vs. 8TPM, p=0.012). There was no difference in the expression of *MLH1, MLH3, MSH3* and *PMS2* in the diagnostic and midway tumour samples.

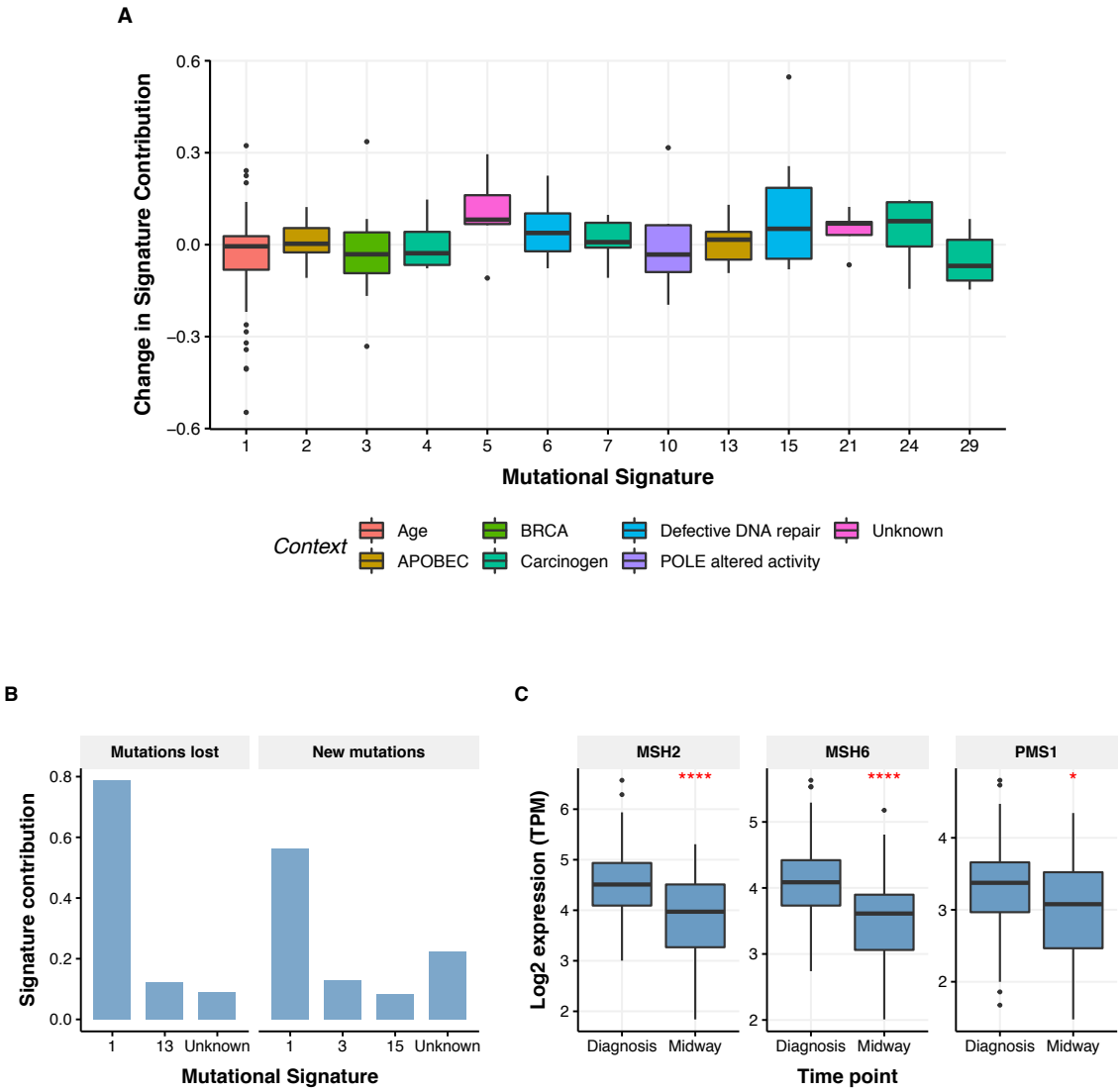


Figure 4.6: Mutational signature contribution change during treatment. (A) Percentage difference between mutational signature contribution at the midway and diagnostic time points. No statistically significant changes were observed across the cohort, indicating that the overall mutational landscape was unchanged by chemotherapy. (B) Signature deconvolution performed on all mutations that were no longer detected and newly-detected. Mutations lost were more likely to be secondary to an APOBEC signature (13), whilst those gained secondary to a BRCA (Signature 3) or defective DNA repair (Signature 15) mutational processes. (C) *MSH2, MSH6* and *PMS1* expression decreased during chemotherapy, potentially contributing to the increased number of Signature 15-related mutations observed.

In order to assess enrichment for mutational signatures specifically on mutations that were newly detectable and undetectable on therapy, signature deconvolution was performed on the grouped mutations using the `DeconstructSigs` R package (Section 3.3.1). This enabled the identification of mutational processes that were responsible for generating these newly-detected mutations, as opposed to the aforementioned analysis which measured signature change globally across each case. As shown in Figure 4.6B, the mutations that were no longer detected were mostly those generated by age (Signature 1) and APOBEC (Signature 13). On the other hand, newly-detected mutations were attributed to age (Signature 1), BRCA (Signature 3) and defective DNA mismatch repair (Signature 15) processes. This analysis showed that chemotherapy rapidly eliminated mutations induced by APOBEC processes, whilst new mutations observed were induced by BRCA and DNA mismatch repair processes. Despite the significant enrichment of BRCA generated mutations in the dataset at diagnosis (Section 3.3.1), very few of these mutations were rapidly eliminated by chemotherapy, perhaps indicating that mutations induced by this process conferred chemoresistance. Over 20% of new mutation signature contributions fell within an unclassified category, perhaps revealing previously undescribed signatures secondary to chemotherapy.

In conclusion, the mutational signature landscape remained globally stable on a per-sample basis during chemotherapy. Mutations that were no longer detectable on therapy enriched for APOBEC signatures, whilst those that were newly detectable enriched for BRCA and DNA mismatch repair signatures, with evidence of *MSH2, MSH6* and *PMS1* expression decreasing during chemotherapy.

Neoantigenic landscape dynamics

Given the dynamicity of the mutational landscape, as well as a previous observation that the tumour neoantigen burden associated with response to neoadjuvant therapy (Section 3.3.3), the change in neoantigenic landscape was explored.

To describe the change in neoantigen load during treatment a neoantigen ratio was computed by dividing the number of mutations that generated a neoantigen by the number of non-neoantigenic mutations for each sample. The change in ratio was derived as the difference between the on-therapy and diagnostic ratio (midway minus diagnostic ratio): higher ratios signified an increase in the number of neoantigens and lower ratios indicated an overall decrease in neoantigenic burden during chemotherapy. Tumours that attained pCR had a greater decrease in neoantigen ratio compared to chemoresistant tumours (Figure 4.7A) and

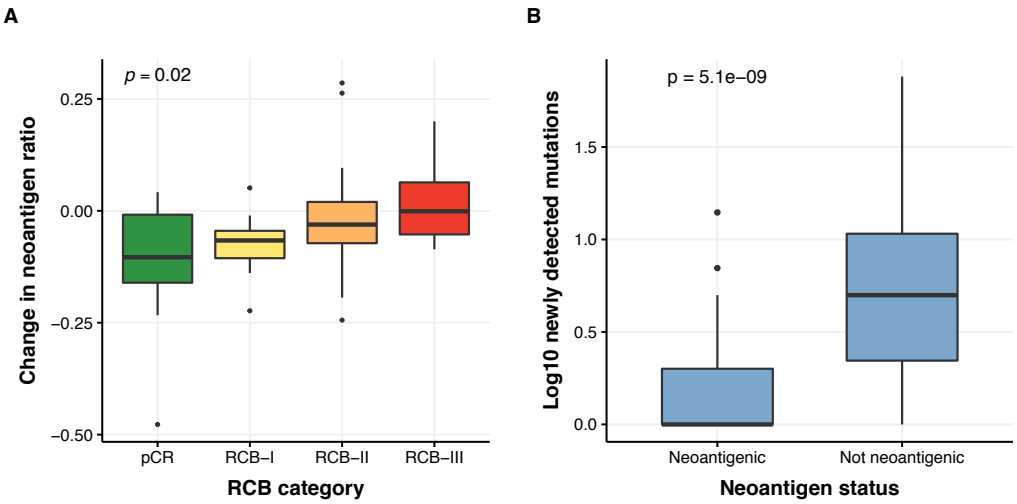


Figure 4.7: Change in neoantigen landscape during treatment. (A) Change in neoantigen ratio during treatment: monotonic change in ratio observed across different RCB categories. (B) Mutations that were newly-detected during chemotherapy were much less likely to be neoantigenic, with only 11% of newly-detected mutations generating a neoantigen.

the relationship across the RCB classes was monotonic (p=0.02, ordinal regression model). Hence, the change in ratio decreased as the degree of RD increased, with RCB-III tumours showing some evidence of an increase in neoantigen ratio during treatment.

55 cases had at least one new mutation detectable at the midway time point, and of these, 51 (harbouring a total of 462 gained mutations) had matching RNA-seq data. 51 of the 462 (11%) newly-detected mutations were neoantigenic, compared to 411 (89%) newly-detected mutations that were not neoantigenic (Figure 4.7B): this difference was highly significant (p-value = 5.127e-09, Wilcoxon rank sum), indicating that newly-detected mutations were much less likely to be antigenic and therefore less likely to be detected by the immune system.

Additionally, four cases had evidence of previously undetected HLA LOH (**T009, T090, T110, T120**) midway during treatment. None of these four cases attained pCR by the end of therapy.

In summary, tumours that attained pCR had a greater decrease in neoantigen burden during therapy, with some chemoresistant tumours showing evidence of an increased neoantigen ratio. Newly detectable mutations were much less likely to generate a neoantigen, indicating that they were less likely to elicit an immunogenic response.

4.3.2 Clonal structure deconvolution

The mutational landscape during therapy has been shown to be dynamic, with the findings so far describing changes presumed to be secondary to the selection pressures induced by chemotherapy, in keeping with observations made by Kim et al. [155]. Changes secondary to tumour heterogeneity and multi-region sampling could not be excluded, though all tumour biopsies were consistently taken close to a radio-opaque fiduciary marker inserted at the time of first sampling in order to mitigate spatial heterogeneity and random sampling.

The tumour clonal architecture was constructed using `PyClone` (Section 2.5.1): clonal clustering was successful in 57 of 75 cases with diagnostic and midway sequencing data. Failure to identify a clonal structure was due to `PyClone` being unable to identify optimal solutions due to the presence of few mutations or very low purity estimations: for example tumours with pCR by the midway time point would have no or very few somatic mutations detectable and therefore these were excluded for the purposes of this analysis as no useful clonal dynamic information would have been obtained.

On establishing the clonal architectures of these tumours and visualising the changes in clonal prevalences during treatment (Figures 4.8 and 4.9) two distinct patterns of genomic response to chemotherapy were evident:

1. **Clonally stable tumours**: defined as tumours that were characterised by mutational prevalences through chemotherapy that were minimally altered (less than a change of 0.25 CCF points during treatment).

2. **Clonally dynamic tumours**: defined as tumours that were characterised by altered clonal prevalences during treatment (more than a change of 0.25 CCF points). Tumours falling within this category were subsequently sub-classified as:
   (a) **Tumours with clonal extinction**: defined as tumours with evidence of subclonal mutations that were either decreasing in prevalence during therapy or no longer detectable
   (b) **Tumours with clonal emergence**: defined as tumours with evidence of subclonal mutations increased in prevalence during therapy, including mutations that were either present at low CCFs or not detected at diagnosis.
   (c) **Tumours with clonal emergence and extinction**: defined as tumours with evidence of mutations that were both decreasing and increasing in prevalence during therapy.

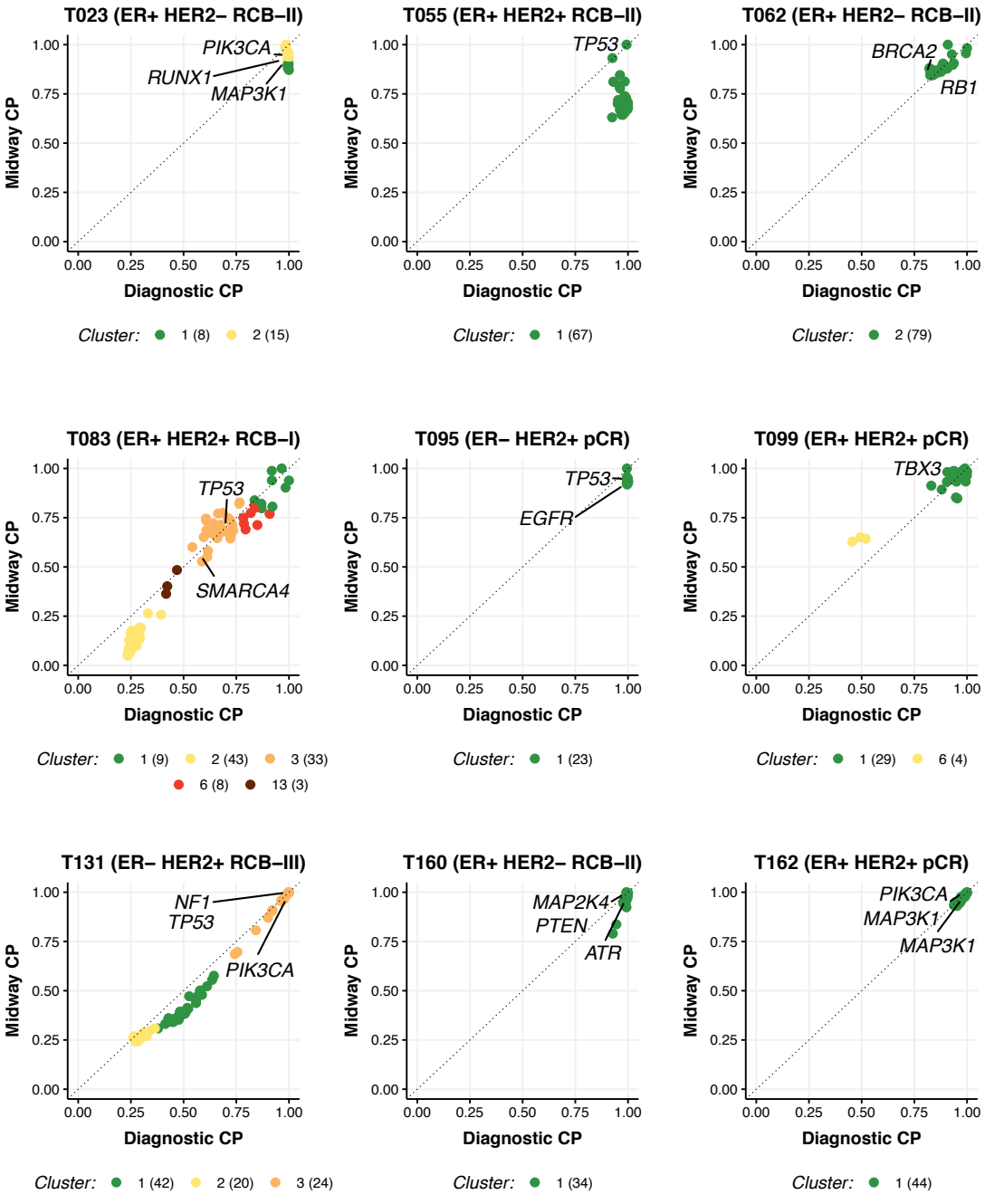


Figure 4.8: Cases showing a clonally stable architecture midway through chemotherapy.

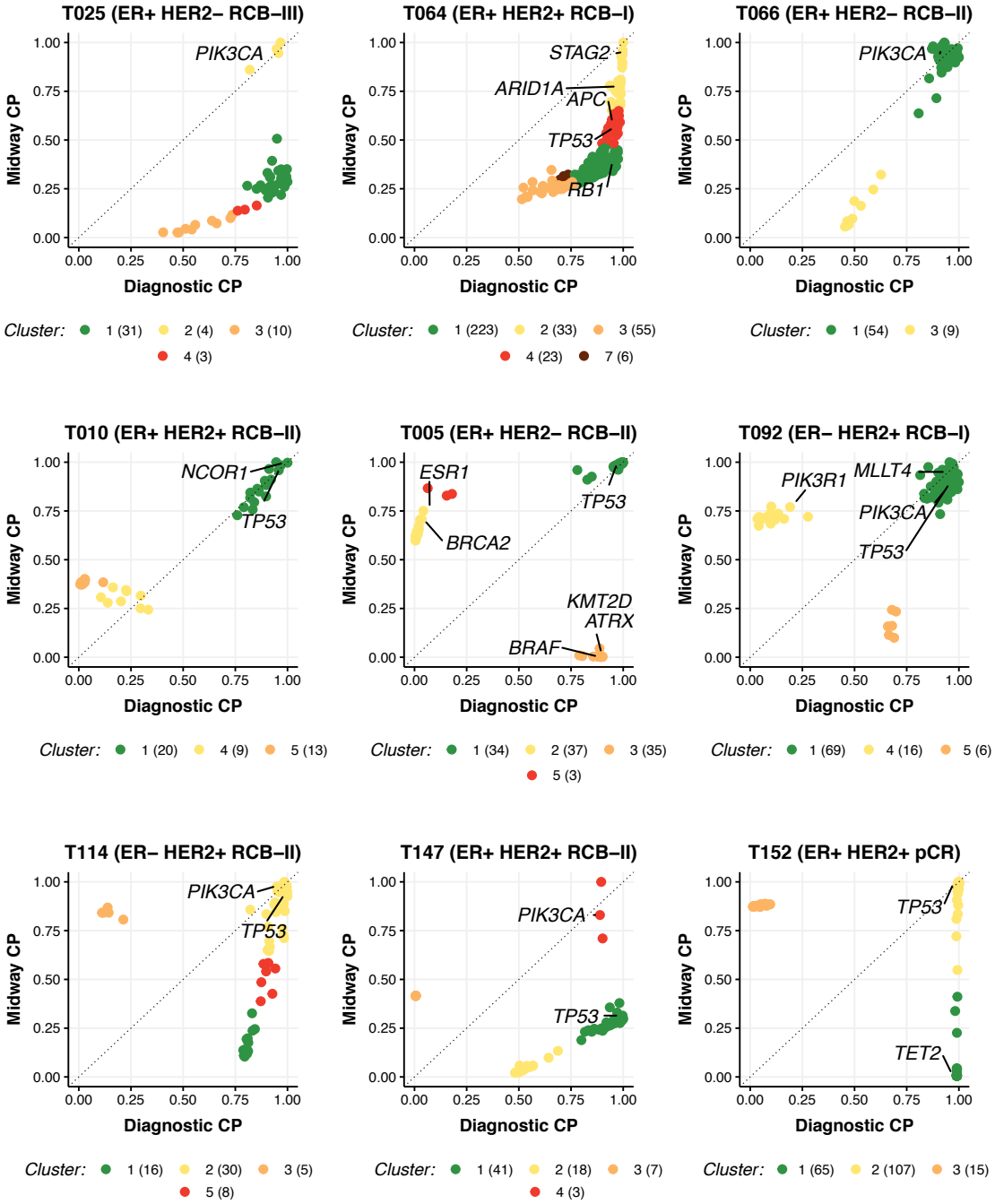


Figure 4.9: Cases showing a clonally dynamic architecture midway through chemotherapy, with evidence of clonal extinction only (T025, T064, T066), clonal emergence only (T010), and both types of genomic response.

Classification of cases into these categories was performed computationally. Only mutational clusters containing at least 3 mutations were retained for classification. In cases where the CCF of the founding clone was less than 1, scaling of all mutations was performed such that the upper CCF bound of the founding clonal cluster was at 1. If the difference in CCF at the two time points was within 0.25 CCF points then the cluster was classified as unchanged. Any deviation of more than 0.25 was assumed to represent extinction or emergence.

Clonally stable tumours

Clonally stable tumours were characterised by mutational prevalences through chemotherapy that were minimally altered, with representative examples shown in Figure 4.8. There was no evidence of appearance or disappearance of subclones during treatment.

23% of all tumours (13/57) showed this response pattern and members comprised both uni-clonal and multi-clonal tumours. 8 of these tumours were moderately differentiated (grade 2), with the remaining 5 being poorly differentiated (grade 3). There was no association between clonal stability at the midway time point and response to chemotherapy.

Tumours that were clonally stable were more likely to have a coding non-silent mutation within Clusterin (*CLU*, Testosterone-Repressed Prostate Message 2) when compared to clonally dynamic tumours (34% vs. 0%, p=0.0004, Fisher's Exact Test). *CLU* has been shown to be a key mediator in the protection against cytotoxic-induced cell death through NF-κB activation and Bcl-2 overexpression [256, 306], with knockdown studies in human cancer cells showing a significant reduction of growth and higher rates of apoptosis [289]. Additionally, in breast cancer, *CLU* overexpression has been associated with resistance to neoadjuvant chemotherapy [214].

Furthermore, clonally stable tumours also had an increased non-silent coding mutation rate in the mitochondrial protein *MT-CYB* (p=0.04, Fisher's Exact Test).

There was no association between *TP53* and *PIK3CA* mutation status and clonal stability during therapy.

Clonally dynamic tumours

Clonally dynamic tumours were characterised by a significant shift in subclonal structure (Figure 4.9): 44 of 57 cases (77%) had this type of response. Whereas the founding clone, often containing an identifiable non-silent driver mutation, was clearly observed in both initial and midway samples, there was evidence of subclonal mutations that varied significantly in prevalence between sampling time points.

Three types of differing clonal dynamic tumours were identified. The first type was predominantly characterised by clonal extinction, wherein the cellular prevalence (CP) of the tumour subclones decreased during therapy. Cases **T025**, **T064** and **T066** showed such a genomic response pattern (Figure 4.9) - the CP of the subclones appeared to dramatically reduce during therapy indicating that they were sensitive to chemotherapy. In the second type of response, clonal emergence, rare subclones that were present at diagnosis increased in prevalence during treatment. **T010** was one such case, where a rare clone increased in prevalence from less than 5% before treatment to around 40% midway through therapy. The third type of response represented a combination of clonal emergence and clonal extinction. For example, in case **T005** (Figure 4.9) there was clear evidence of a founding clone with a *TP53* mutation and a subclone containing *KMT2D*, *BRAF* and *ATRX* mutations at diagnosis. As chemotherapy was administered, the chemosensitive *KMT2D-BRAF-ATRX* subclone was no longer detectable and was replaced by a chemoresistant *ESR1-BRCA2* mutant subclone. Likewise, in case **T092**, a chemoresistant *PIK3R1* driven subclone emerged during chemotherapy. Case **T147** was characterised by a founder clone with a *PIK3CA* mutation and a subclone with a *TP53* mutation. In keeping with observations made in Chapter 3, tumours with *PIK3CA* mutations were more likely to be chemoresistant - indeed the tumour from case **T147** had RCB-III RD at the time of surgery, with the more aggressive *TP53* subclone rapidly decreasing in prevalence during treatment, leaving the founding *PIK3CA* clone that did not contain a *TP53* mutation to predominate the landscape.

All tumours that exhibited clonal emergence without clonal extinction by the midway time point had RCB-II or RCB-III RD by the end of treatment, indicating aggressive disease with more prevalent subclones arising during treatment. Dynamic clonal landscapes with combined emergence and extinction were also evident in tumours that eventually attained pCR, suggesting that the chemoresistant subclones maintained by the first block of chemotherapy were subsequently eliminated by the second block of treatment.

Association with chromosomal instability and copy number alterations

The copy number landscape also showed evidence of changes during therapy, with some tumours showing greater degrees of CNA dynamics than others. Tumours with clonal emergence only had a statistically significant increase in homologous recombination deficiency (HRD, Figure 4.10A) during therapy, compared to all other types of responses. Indeed, the change in HRD scores varied monotonically across the four different response subgroups, with a decrease in HRD score as clonal extinction predominated (p=0.01, ordinal regression model).

A whole-genome copy number landscape analysis echoed a similar finding: tumours that had evidence of clonal emergence had an increased number of CNAs during therapy and increasing chromosomal instability (Figure 4.10B). As clonal extinction predominated, the degree of copy number instability present decreased (p=0.03, ordinal regression model). Hence the genomic responses identified were consistent irrespective of whether the analysis was performed on mutational cellular prevalences or whole-genome copy number changes.

Tumours that were clonally stable also had a more stable copy number landscape compared to tumours which showed evidence of a dynamic genomic architecture (Figure 4.10C), with tumours undergoing clonal extinction being the most different from the treatment-naive tumours secondary to the loss of subclonal copy number events.

As the copy number landscape was shown to be altered during therapy, an analysis of the tumour iC10 copy number classification before and midway through therapy was performed (Section 2.5.2). 70% of all cases did not exhibit a change in the iC10 copy number classification midway through therapy. In 30% of cases, a switch to cluster 4 (normal-like) was observed: this was universally secondary to lower degrees of tumour purity. Indeed, 60% of cases with a switch to iC4 attained pCR by the end of therapy whilst 25% had evidence of low-cellularity RCB-I RD, indicating that the switch to a normal-like genotype was secondary to a decrease in tumour purity attributed to a good response to chemotherapy. More importantly, this iC10 analysis showed that although the copy number landscapes were dynamic during therapy, there was no significant change within the CNA driver amplicons that generated and maintained the tumour cells as there was no evidence of iC10 class switching in the high purity tumours.

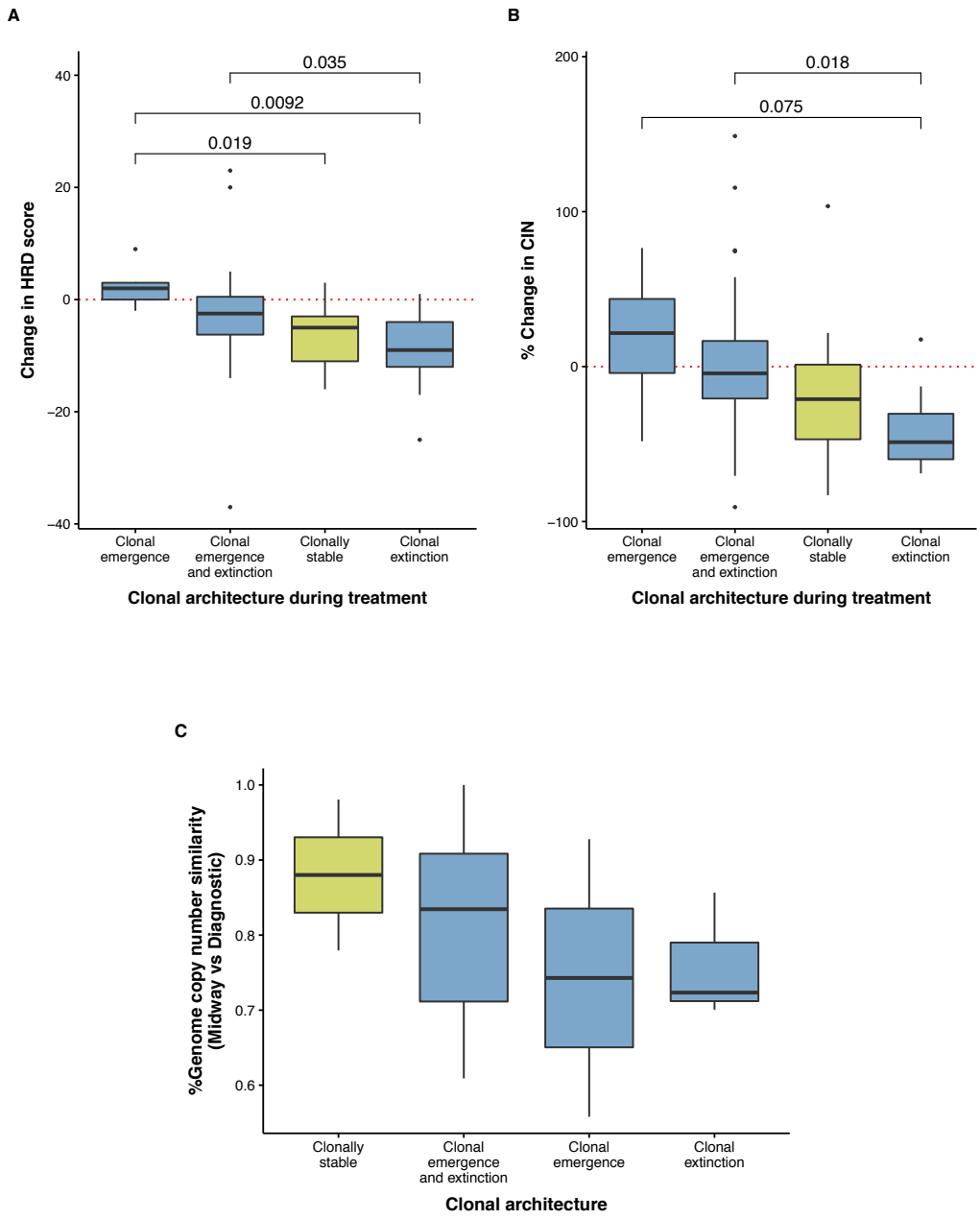


Figure 4.10: Changes in copy number landscape during therapy. (A) Change in HRD index across four genomic response groups. (B) Percentage change in copy number instability during therapy. (C) Tumour copy number landscape similarity midway through chemotherapy.

4.3.3 Clonal phylogeny reconstruction

Following the deconvolution of clonal architecture from the bulk exome sequencing data, clonal phylogeny was derived using `LICHeE` [228], which imposed evolutionarily constrained networks on the clonal prevalence data inferred by PyClone and determined optimal phylogenetic solutions. Representative phylogenies are shown in Figure 4.11.

The phylogenetic networks generated mirrored very closely previous observations, with evidence of changes in the tumour clonal hierarchy during therapy. Clonally stable tumours, as represented in Figure 4.11A, had a quasi-identical subclonal composition between sampling time points, with very little deviation of subclone prevalences and no new clones detected. In the example chosen in Figure 4.11A (**T131**), the founding clone, as well as two daughter clones were both identified in the diagnostic and midway time points, with no change in clonal prevalence.

In contrast to the clonally stable tumours, tumours classified as clonally dynamic showed distinct changes within the tumour subclonal composition, as shown in Figures 4.11B-F. Tumours undergoing solely clonal extinction showed two types of phylogenetic alterations, as depicted by cases **T025** (Figure 4.11B) and **T064** (Figure 4.11C). In case **T025**, subclones that were generated later on in the tumour's evolutionary history, and which therefore had a greater degree of mutations/CNAs, were no longer detected after commencing chemotherapy, with only evidence of the founding clone and earlier subclones detected at the second sampling time point. Perhaps by accruing new genomic aberrations daughter clones became more proliferative (and therefore more chemosensitive) or more immunogenic and were eliminated faster by cytotoxic therapies. In a second scenario, as exhibited by case **T064**, rather than exhibiting complete subclonal elimination there was a gradual regression of the dominant subclone as treatment ensued. For example, in case **T064**, the pre-treatment landscape was predominated by the subclone harbouring the 55 mutations in cluster 3 (orange, also see Figure 4.9), but as treatment was administered there was evident regression of this abundant clone, with the less chemosensitive parent subclones predominating the clonal landscape now that the dominant clone was no longer selected for. Perhaps the dynamics shown by **T064** were, in fact, a prelude to those observed in case **T025**, and with time complete regression of the dominant subclone would have occurred.

The most common pattern of clonal dynamics was one in which extinction and emergence occurred together, as shown in Figures 4.11D-F. In case **T152**, the chemoresistant subclone which harboured a *TET2* mutation and dominated the landscape was not convincingly

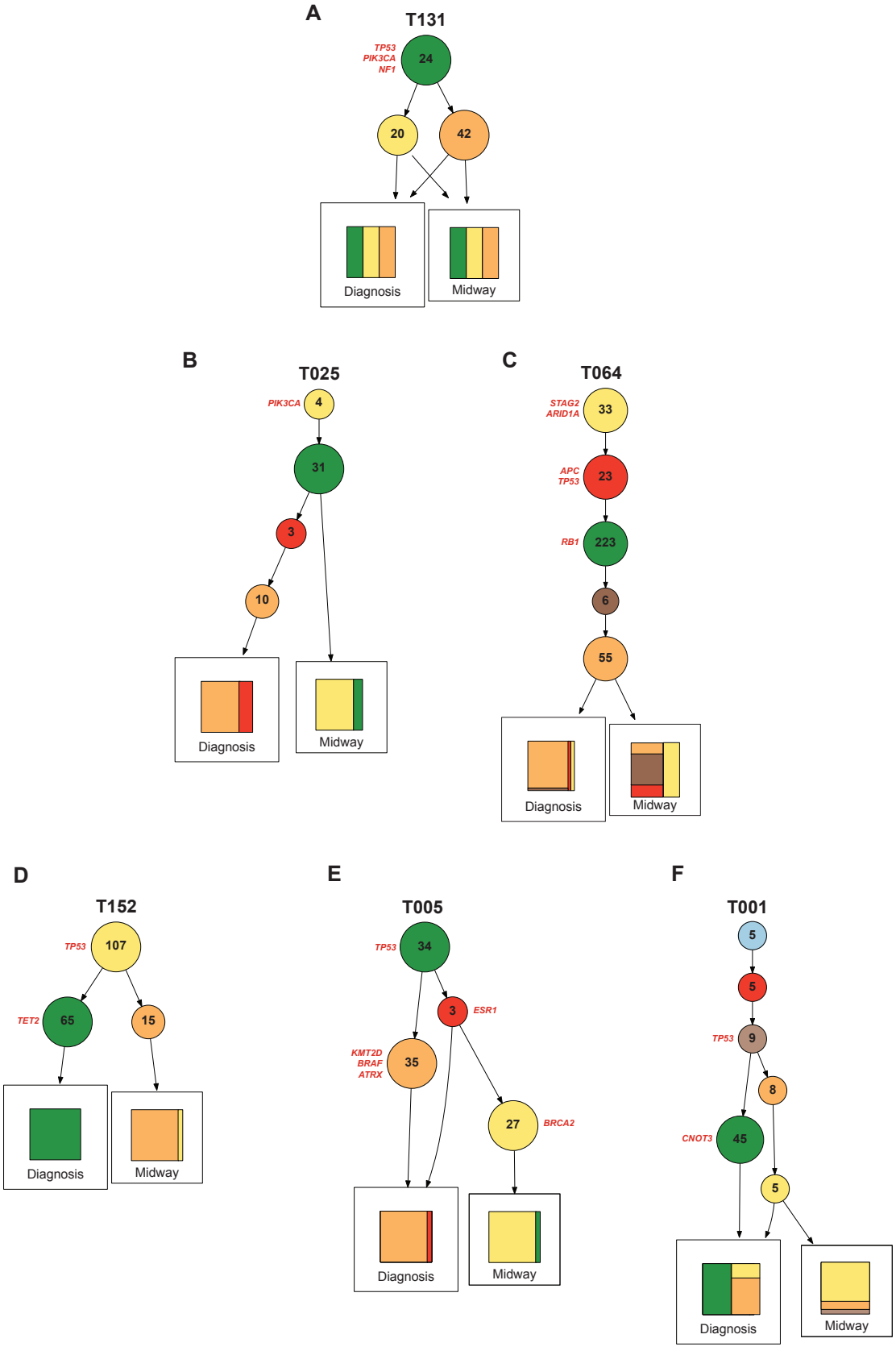


Figure 4.11: Tumour phylogeny construction and alterations induced by chemotherapy. Numbers in circles indicate number of mutations. (A) Clonally stable tumour. (B) Clonally dynamic tumours showing clonal extinction. (C-F) Clonally dynamic tumours showing clonal extinction and emergence.

observed at the midway time point (Figure 4.11D), and instead a previously unobserved subclone harbouring 15 previously undetected mutations was seen, together with the originating founding clone that harboured a *TP53* mutation (Figure 4.9). This could have occurred either due to spatial sampling or due to clonal expansion secondary to altered resistance patterns. Whereas in case **T152** the originating resistant subclone originated directly from the founding clone, in other cases the resistant cell was derived from a subclone that was generated later on in the tumour's evolutionary history (Figures 4.11E and F). In case **T005**, the *BRCA2* subclone predominating the landscape was not sampled at diagnosis, but its founder, a subclone carrying an *ESR1* mutation, was convincingly detected, albeit at very low cellular prevalences. On commencement of therapy the *BRCA2* subclone (which also contained an *ESR1* mutation), expanded and predominated the landscape. Interestingly, **T005** was an ER^+ tumour, and the *ESR1* mutation (c.600G>C, p.W200C) was predicted at a high probability to be deleterious by both SIFT (classification: `deleterious`, score 0) and PolyPhen (classification: `probably_damaging`, score 0.999) and was, perhaps, the selection of an endocrine therapy resistant subclone.

In summary, subclonal deconvolution using mutational and copy number alteration data showed evidence of tumours that remained genomically stable midway through therapy, as well as others that showed a significant change in their architecture. Clonally dynamic tumours showed evidence of subclonal mutations that were either newly-detected or no longer detected on second sampling, potentially attributed to chemosensitive or chemoresistant populations. While tumour spatial heterogeneity could also explain some of these findings, the biopsy techniques used were designed to consistently sample the same region of a tumour in order to mitigate this.

4.4 Late genomic landscape alterations

4.4.1 Mutation dynamics

50 cases had tumour sampling performed at the diagnostic and surgical time points and 26 cases had sampling at all three time points. H&E sections from all the samples that were subsequently sequenced and included within downstream analyses were examined by a pathologist and confirmed to have been taken from the tumour bed.

Figure 4.12A shows the mutation dynamics between the pre and post-therapy samples for all 50 cases. A total of 1,799 mutations observed at diagnosis were not detected by the end of therapy, whilst 224 mutations were detected in the post-therapy samples but not in the diagnostic samples. 1,028 mutations were detected at both time points.

A total of 8 cases that attained pCR were sequenced in order to identify whether any remaining mutations could be identified despite no pathological evidence of tumour cells in the resected tumour specimens. On close inspection, Figure 4.12A shows that, in these 8 cases, persisting mutations could still be observed within the exome sequencing data: in case **T136** more than 20 mutations persisted within the sequenced tissue with a median AF of 0.8%, despite a pathological assessment of pCR. The persistent detection of these mutations could either be secondary to residual tumour cells within the sequenced sample, and therefore incorrect response classification, or the presence of in-situ carcinoma harbouring these mutations. Indeed pCR in this study was defined as the absence of any tumour cells in the breast or lymph nodes, irrespective of the degree of remaining residual carcinoma *in situ*. Two cases that attained pCR (**T003** and **T059**) each had two previously unobserved mutations detected, though it was highly likely these mutations were false positives and secondary to variant calling error. As described in Chapter 2, the predicted false positive rate for the variant calling pipeline was estimated at 0.08 mutations Mb^{-1}, so around 3 mutations per sample would be expected to be false positives. This analysis also showed that chemotherapy did not induce any new detectable mutations within the germ-line normal breast tissue.

In sharp contrast to the observations made at the midway time point, significantly fewer newly-detected mutations were observed in the post-chemotherapy specimens, with only three cases having more than 20 new mutations identified (**T004**, **T090**, **T127**). This might have been secondary to longer chemotherapy exposure and/or the alteration of chemotherapy agents administered: most patients in this study received either a taxane for three cycles,

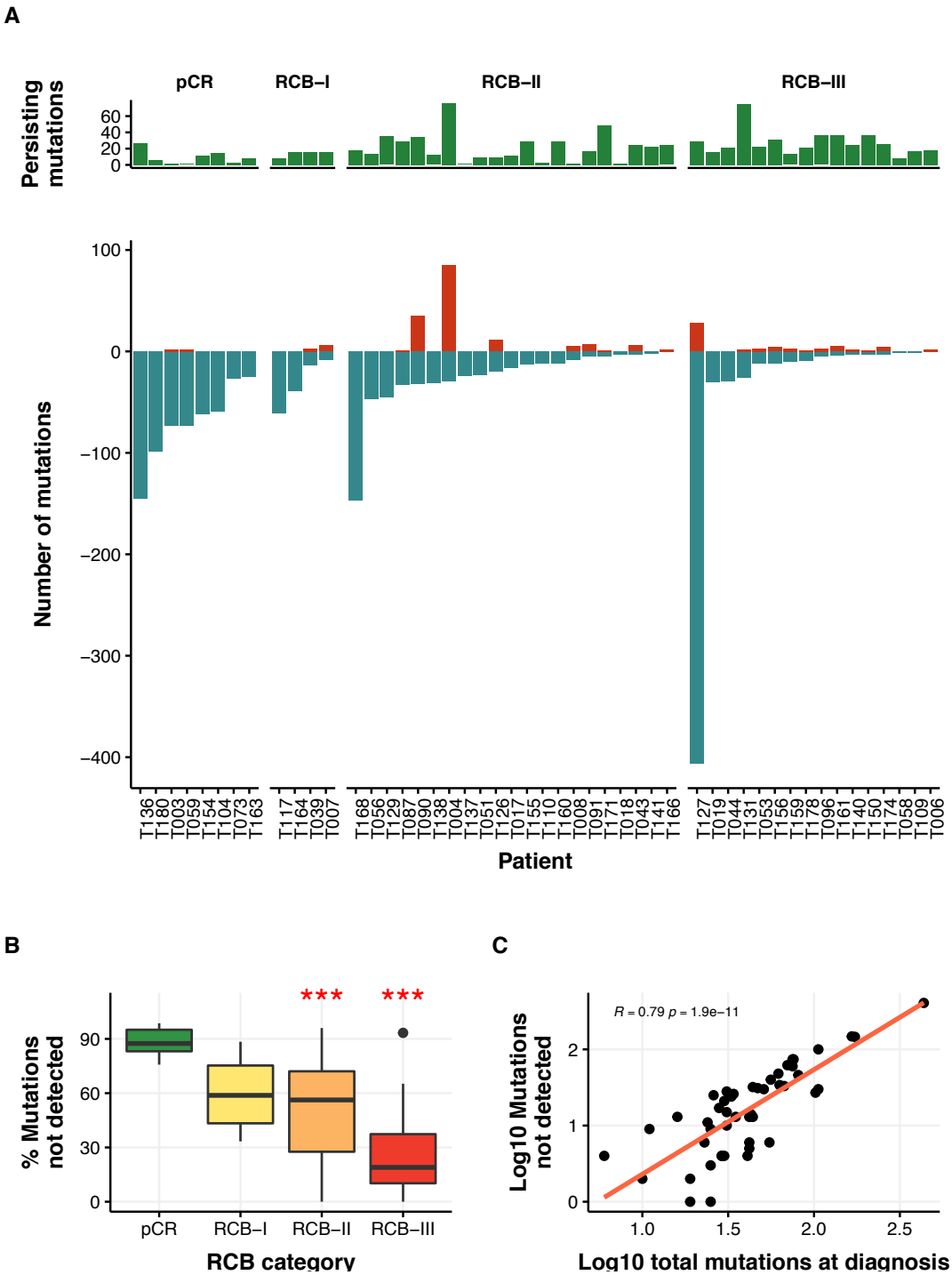


Figure 4.12: Somatic mutation landscape dynamic on completion of chemotherapy. (A) Green bars depict number of mutations detected pre and post-therapy. Red bars show number of newly-detected mutations detected on completion of chemotherapy, blue bars quantify mutations detected at diagnosis only. (B) Box plots showing percentage of mutations that were no longer detected after completion of chemotherapy across the four response groups. (C) Scatter plot showing relationship between the total number of mutations observed at diagnosis and the number of mutations that were no longer detected after completion of chemotherapy.

followed by an anthracycline for a further three cycles if HER2$^-$ while in HER2$^+$ disease the cytotoxic sequence was reversed and trastuzumab was added to the second block of the chemotherapy regimen. This suggests that by swapping chemotherapy backbone any chemoresistant subclones maintained by the first block of the chemotherapy regimen might have been effectively eliminated by the second block of treatment.

In keeping with the observations made at the midway time point, the percentage of mutations that were not detected correlated to both the response observed (Figure 4.12B) as well as the number of mutations observed at diagnosis (Figure 4.12C). Tumours that had higher degrees of RD present at the time of surgery remained genomically very stable, with most of the mutations observed at the initial time point persisting after completion of neoadjuvant chemotherapy.

Subclonal driver mutation dynamics

As observed in the midway time point, at the end of treatment the driver gene landscape was also altered, with evidence of newly detectable driver mutations, as well as the apparent eradication of others (Figure 4.13). Of all the higher purity post-treatment samples analysed, six cases had evidence of loss of a driver gene during treatment: in all cases the mutation was predicted to be subclonal on the pre-therapy tumour sample and included: *NF1, SMARCA4, SF3B1, PIK3CA, MAP3K1, ESR1* and *BCOR*. Three cases had evidence of new driver mutations emerging during therapy, and these included *GATA3, BUB1B, TP53* and *MLLT4*. In all but one case (**T004**, Figure 4.13) the subclonal probability was estimated to be 100%, indicating that these were truly not present within the founding clone. None of the cases with RCB-I RD had evidence of new driver mutations: the emergence of previously unobserved driver mutations was restricted to chemoresistant (i.e. RCB-II and RCB-III) tumours within this dataset.

Pathway enrichment and signature deconvolution

To elucidate whether the mutations that were newly-detected during treatment played a key role in cellular pathways an over-representation analysis was performed using the `ConsensusPathDB` tool [149, 150]. An enrichment analysis was conducted on the Reactome, KEGG, PID and BioCarta databases, retaining pathways that had at least four overlapping

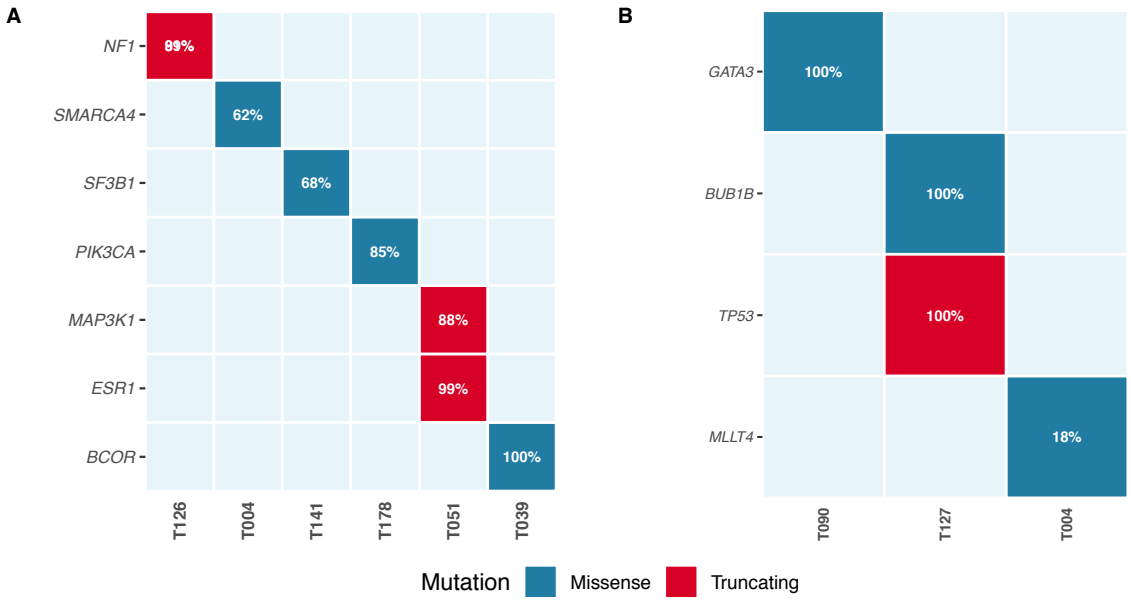


Figure 4.13: Subclonal driver mutation dynamics during treatment. (A) Tumours with loss of non-silent driver mutations during therapy, with subclonal probability shown. (B) newly-detected driver mutations observed within three chemoresistant cases.

genes, with statistical significance ascertained only below an FDR cut-off of 0.05 in order to minimise the number of false positive hits.

Pathway enrichment over newly-detected mutations revealed over-representation of two pathways, oxidative phosphorylation/respiratory electron transport and mitosis, all at an FDR of less than 0.04:

1. Reactome pathways
   (a) Respiratory electron transport (*MT-CO1, MT-ND2, MT-ND3, MT-ND4*)
   (b) G2/M Transition and Mitotic G2-G2/M phases (*CUL-1, TP53, FGFR1OP, HMMR*)
2. KEGG pathways
   (a) Oxidative phosphorylation (*MT-CO1, MT-ND2, MT-ND3, MT-ND4*)

Four genes involved in cell cycle regulation were associated with chemoresistance. One of these, *HMMR* (Hyaluronan-mediated motility receptor), codes for a protein that associates with microtubules and plays a key role in the maintenance of spindle integrity and mitosis [189]. Additionally, *HMMR* has been shown to interact with *BRCA1* and *BARD1*, with the resulting complex playing a key role in the regulation of mitosis [146, 229].

In order to assess whether the new mutations were driven by any particular mutational process, signature decomposition was performed on the aggregated new mutation data, which showed enrichment of Age (Signature 1, 69%), two APOBEC signatures (Signature 2: 11% and Signature 13: 8%) and the error-prone polymerase POLE signature (Signature 10, 6%), with only 6.6% of total signature contribution attributed to the "Unknown" category. Two observations can be made from this analysis. Firstly, only a very small proportion of mutations fell within an unknown mutational signature, showing that the overall detectable mutagenic effect of the chemotherapeutic agents used was indeed very low, with most mutations being generated by a pre-existing mutational signature. Secondly, the new mutations observed in the non-chemosensitive (RCB-II and RCB-III) tumours were secondary to hypermutation signatures (including APOBEC) that were operating during chemotherapy.

4.4.2 Clonal structure deconvolution

To delineate changes within the subclonal composition of the tumour after completion of neoadjuvant therapy, `PyClone` was run on 25 cases which did not attain pCR and had a tumour purity above 20%. In addition, 10 of these cases also had a corresponding tumour sample taken at the midway time point.

The clonal dynamics previously observed midway through chemotherapy were once again very evident, with 76% (19/25) of all cases harbouring a clonally dynamic genotype, and 24% (6/25) showing a clonally stable architecture, as defined previously. Remarkably, the ratio of tumours with a clonally dynamic versus a clonally stable genotype was very similar to that seen in the midway biopsy time point. Out of the 19 tumours with a clonally dynamic genotype, 4 had evidence of clonal emergence only, 5 had evidence of clonal eradication only and 10 had evidence of both clonal extinction and clonal emergence.

Clonally stable tumours (Figure 4.14) retained a genomically identical architecture throughout treatment, with no new subclones appearing and the overall clonal structure remaining very similar. As was observed at the midway time point, tumours with both uni-clonal and multi-clonal architectures were classified within this response group.

Clonally dynamic tumours showed significant shifts within the cellular prevalences of the mutational subclones (Figure 4.15), with evidence of new subclones identified (or increasingly prevalent ones), as well as loss of subclonal mutations (or decreasingly prevalent ones).

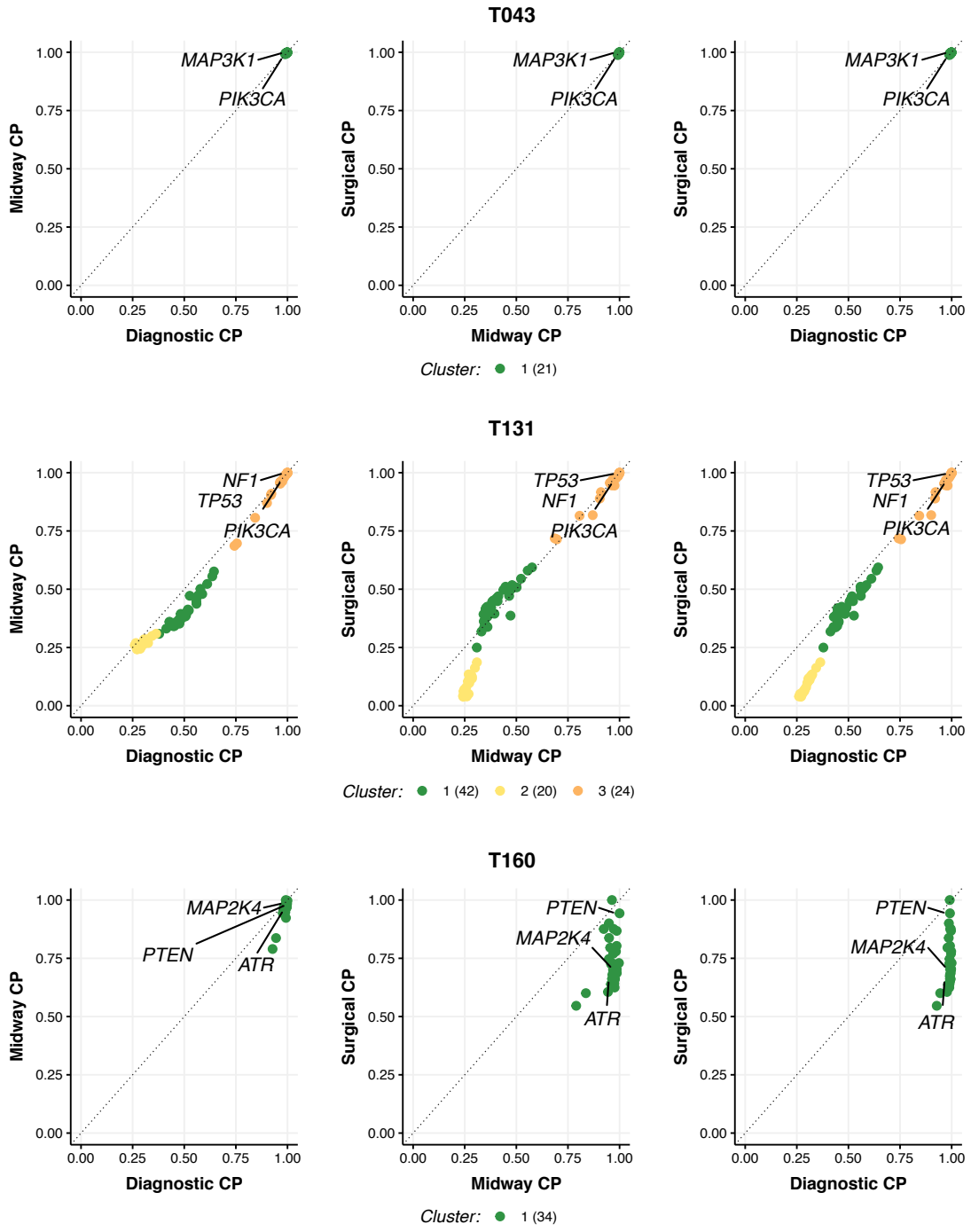


Figure 4.14: Genomic architecture of clonally stable tumours throughout treatment. Cellular prevalences compared at the diagnostic, midway and surgical time points. T131 was clonally stable between the diagnostic and midway time points, with a slight decrease in cellular prevalence of one subclone by the surgical time point.

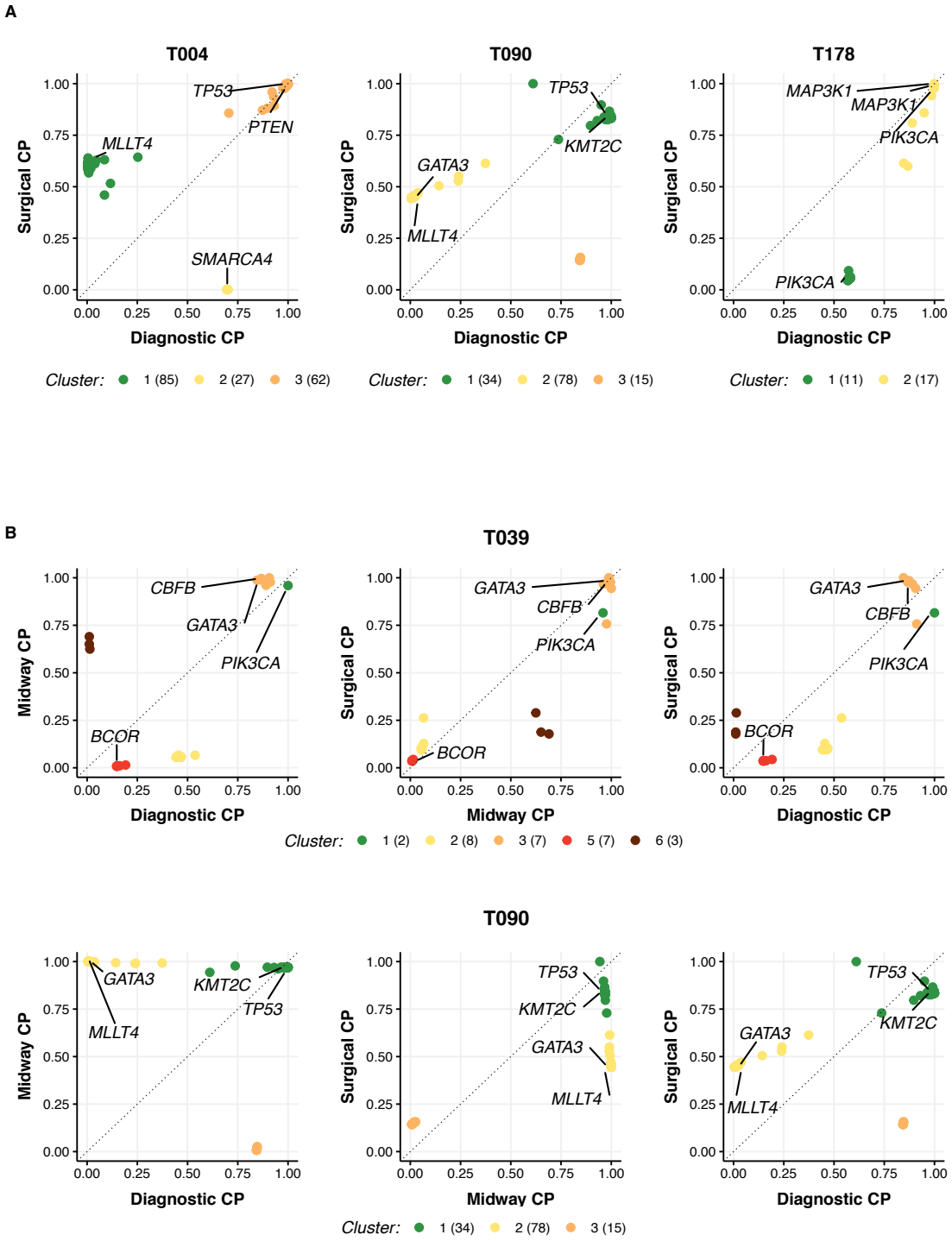


Figure 4.15: Clonally dynamic tumours. Altered clonal dynamics identified in (A) cases with diagnostic and surgical tumour samples and (B) tumours serially sampled at three time points.

Following the deconvolution of the mutational and copy number data into clonal clusters, clonal phylogenies were reconstructed using `LICHeE`, as described previously, to navigate temporal clonal evolution. Riverplots of the tumour phylogenies were generated from the `LICHeE` output using the R package `timescape` (Figure 4.16). The riverplots showed distinctive clonal pattern shift during chemotherapy. For example, case **T090** (Figure 4.16A) received treatment with T→FEC chemotherapy and showed evidence of a dominant subclone (orange) being replaced by a second subclone during treatment with docetaxel. This subclone, which became increasingly prevalent during treatment with docetaxel, was also chemoresistant to the second block of the chemotherapy backbone (FEC), thereby exhibiting a degree of multi-agent resistance. At the time of surgery, RCB-II disease was present.

Case **T007** (Figure 4.16B) demonstrated clearly the benefits of trastuzumab when commenced at the second block of treatment, During treatment with 3 cycles of FEC, a subclone (orange) gradually decreased in cellular prevalence, and was replaced by a chemoresistant subclone (red). However, on commencing trastuzumab and docetaxel, the subclone that was chemoresistant to FEC greatly decreased in prevalence, showing that a therapy switch to docetaxel and trastuzumab was beneficial. This decrease in clonal dominance was echoed by the presence of RCB-I RD at surgery.

Case **T131** was also a $HER2^+$ tumour, though treatment was instituted with four cycles of docetaxel, pertuzumab and trastuzumab with no backbone switch. Unlike the clonal dynamics observed previously, the tumour remained clonally stable after 2 cycles of treatment (Figure 4.16C), and by the end of therapy only one of the clones decreased in cellular prevalence. The genomic stability observed was in concordance with the extensive RD present at the time of surgery (RCB-III).

Case **T096**, which was treated with T→FEC, showed a very intriguing change in subclone architecture. On commencing chemotherapy with docetaxel there was quick clonal alteration with a marked decrease in the prevalence of two subclones (green/orange clones in Figure 4.16D). On switching the chemotherapy backbone to FEC, a gradual increase was observed in one of the subclones, indicating chemosensitivity to docetaxel but chemoresistance to FEC. Indeed, as therapy was switched to FEC, there was evidence of resistance to therapy and an increase in the prevalence of one of the subclones, resulting in RCB-III disease.

Case **T141** showed the transient appearance of a *TP53* driven subclone (yellow) at the midway time point during therapy with docetaxel (Figure 4.16E). This clone was no longer

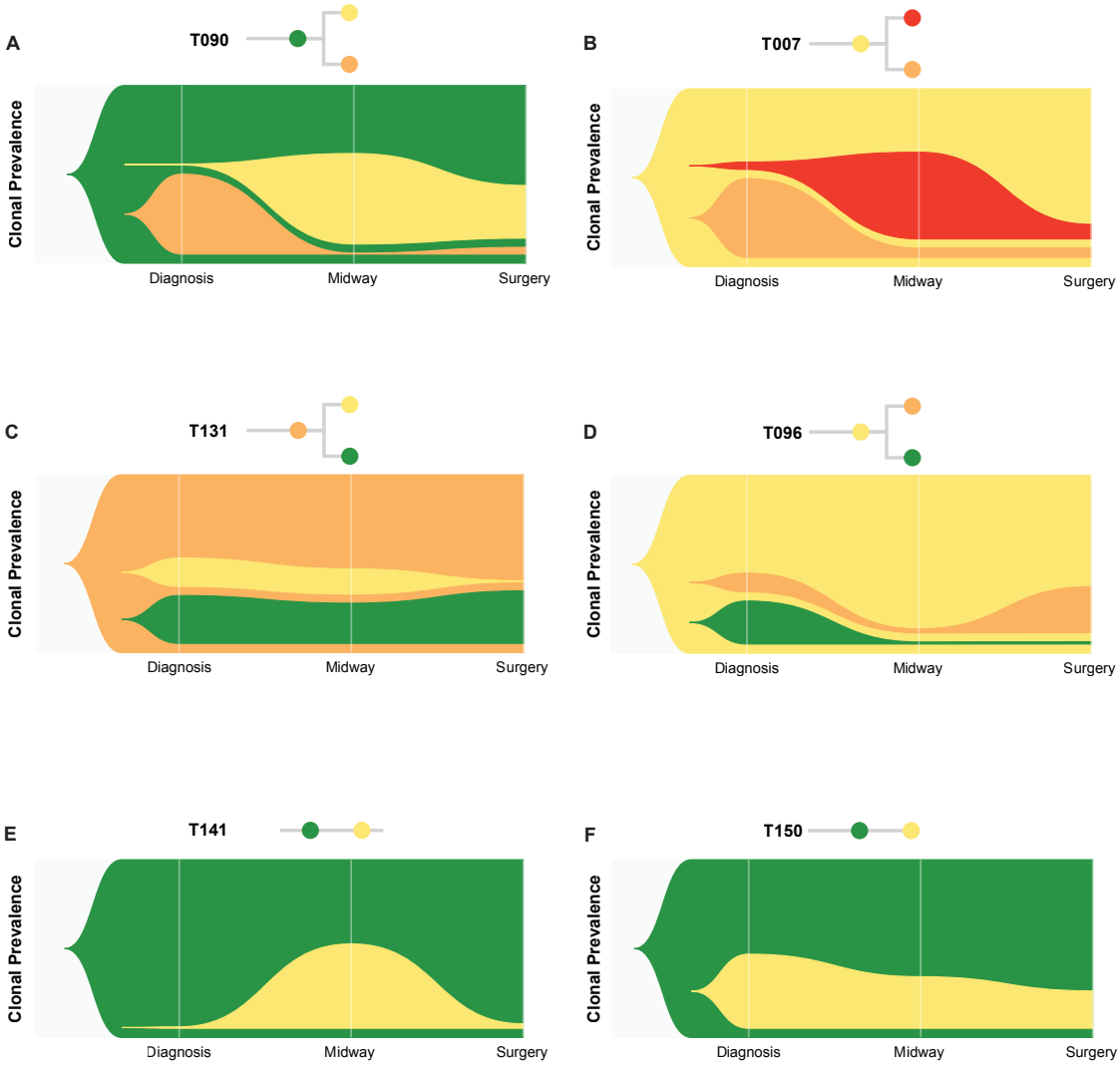


Figure 4.16: Riverplots of tumour clonal dynamics during treatment for a selection of cases with three serial sampling time points. Evidence of: (A) Emergence of a resistant subclone at the midway time point, maintained throughout chemotherapy. (B) Emergence of a resistant subclone at midway time point, sensitive to second block of chemotherapy regimen. (C) Clonally stable architecture by the midway time point, followed by decrease in cellular prevalence of one of the subclones during the second block of the chemotherapy backbone. (D) Highly sensitive daughter subclones to first part of chemotherapy backbone, with re-emergence of a chemoresistant subclone on switching therapy. (E) Evidence of a new subclone at the midway time point, sensitive to second block of therapy. (F) Suboptimal chemotherapy exposure following dose reduction and agent omissions.

detected at the time of surgery following an additional three cycles of FEC, though RCB-II disease was still present at the time of surgery.

Case **T150** showed the gradual and slow eradication of a subclone during treatment with paclitaxel→EC. This patient did not tolerate chemotherapy and was initially started on weekly paclitaxel (rather than three weekly docetaxel) and required multiple dose reductions of the taxane. Fluorouracil was removed from the second block of the chemotherapy backbone, and only received EC instead. As can be seen from the riverplot (Figure 4.16F), the suboptimal chemotherapy exposure resulted in very little change in the clonal constitution of the tumour and RCB-III RD was present at the time of surgery.

In summary, the genomic landscape changes seen at the midway time point were also observed on completion of chemotherapy, with evidence of tumours having a very stable clonal landscape and others showing greater degrees of shifts. By integrating data from three sampling time points the gradual changes in clonal prevalences could be modelled, increasing our insight into the differing genomic responses to therapy induced by backbone switches.

4.5 Deep sequencing validation

4.5.1 Validation on fresh frozen tissue

The bulk exome sequencing analysis discussed has shown evidence of changes in the clonal architecture of tumours that occurred during treatment with chemotherapy. In order to validate these findings, deep sequencing was performed on four index cases that exhibited evidence of clonal dynamics during chemotherapy (**T003**, **T004**, **T025**, **T029**). Matched normal samples were also sequenced for all four cases. Diagnostic samples were available for all four cases, midway samples available for cases **T003**, **T025** and **T029** and surgical samples available for cases **T003**, **T004** and **T029**. The aliquots of DNA used were the same as those that were used for the exome library preparation. In addition, two unrelated control samples containing normal DNA were also included within the analysis.

A total of 156 mutations were targeted across all four cases and primers designed for each region. Targeted amplification was performed using the Raindance system (Section 2.4.2), and the base statistics of the loci containing the mutations of interest generated by using GATK `HaplotypeCaller`. As the mutations targeted were exclusive to each case all other cases served as controls to allow background noise estimation.

The mean amplicon coverage across all samples was 4,596× (range 39-9,931×, Figure 4.17A). All samples had a median coverage of over 2,500×, except for the **T003** surgical sample, which had lower DNA content due to the fact that pCR had occurred and very few cells were present within the sequenced sample as most of the tumour had been replaced by fibrous tissue. Of the 156 amplicons targeted, 5 had a median coverage of less than 10x across all samples (Figure 4.17B) and were removed from the analysis. Of the remaining 151 amplicons, the median coverage was 5,144× (range 133-9890×). Based on this depth of coverage, any rare variants that were not detected on the exome sequencing data would be expected to be detected at low allelic fractions.

In order to assess for primer specificity and for inter-sample contamination, the VAFs of patient-specific mutations were assessed across all controls and cases (Figure 4.17D). Mutations present in one case were not observed in other cases, indicating high primer specificity (as opposed to low specificity, in which through binding to regions with high homology mutations due to misalignments would be expected to be present) and no cross-contamination between samples.

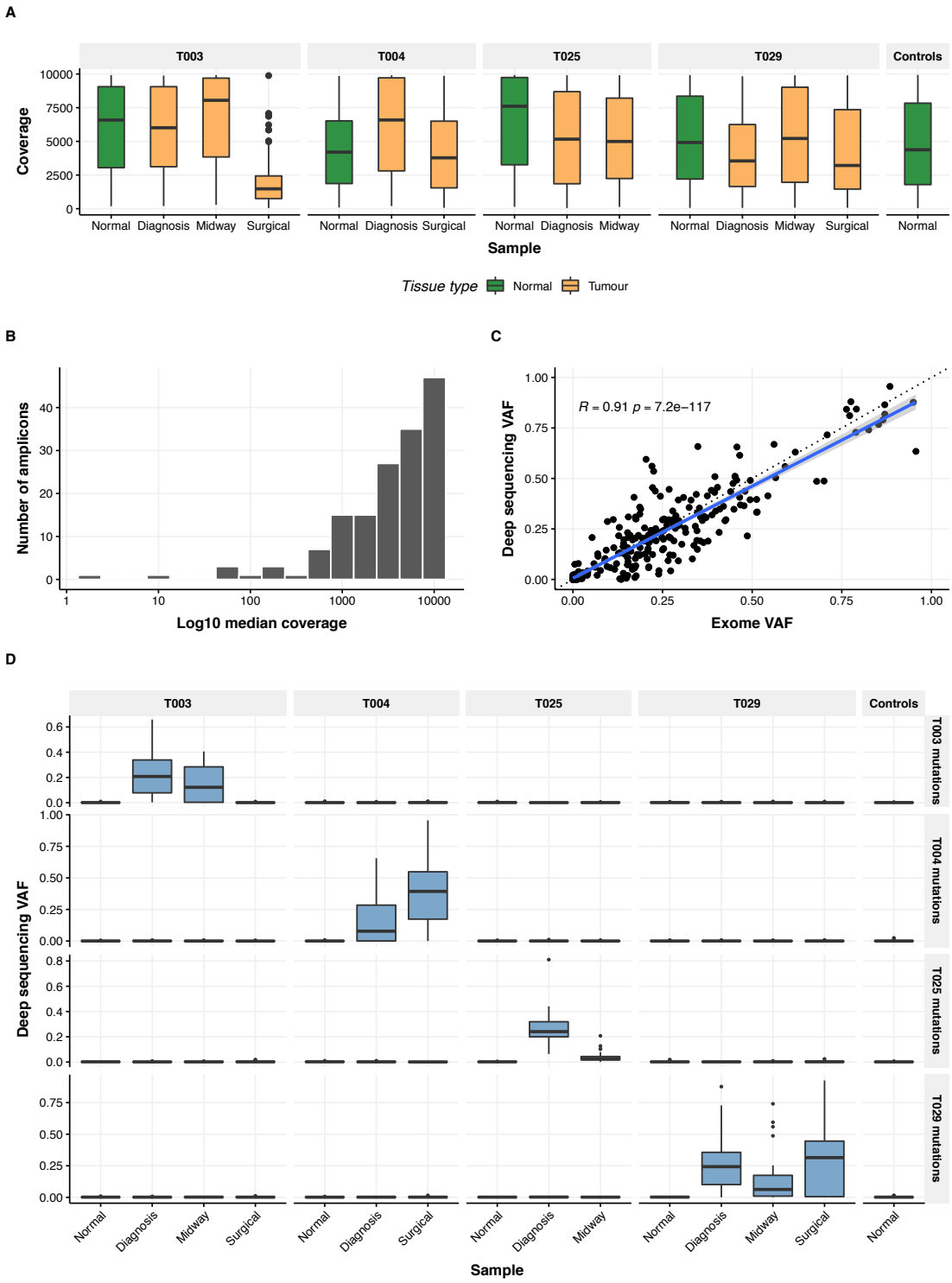


Figure 4.17: Targeted amplification metrics. (A) Distribution of coverage across all samples sequenced. (B) Median coverage distribution across the 156 amplicons sequenced. (C) Concordance between exome AF and deep sequencing AF. (D) VAF of targeted mutations across cases and controls.

A mutation was deemed to be present if the VAF of the mutation in a case was three standard deviations above that of controls. A high concordance was observed between exome VAF and deep sequencing VAF (R= 0.91, p=7.2e-117), showing the reproducibility of the exome sequencing pipeline. By orthogonally validating the selected mutations, the deep sequencing experiment allowed the characterisation of the performance of the exome pipeline, with an accuracy of 0.96, sensitivity of 0.96, specificity of 0.95 and false positive rate of 0.05. Only 8 mutations were not detected by exome sequencing but were detected by deep sequencing.

Correlation with predicted clonal structure

In Case **T003** (Figure 4.18), three mutational clusters were detected on bulk exome sequencing: one founding clone (green) and two subclonal clusters (yellow and orange). Representative mutations from each of the three groups were deep sequenced and this confirmed the presence of a founder clone present at the diagnostic and midway time points, a subclone that was no longer detectable midway through therapy, and another that appeared midway and was no longer detectable by the time of surgery. This tumour attained pCR on therapy and none of the selected mutations were identifiable. Four of the six mutations that were only present in the midway sampling time point on exome sequencing were also present at the diagnostic time point on deep sequencing, albeit at low allelic fractions, suggesting that they were not generated by chemotherapy but present prior to the commencement of treatment.

In Case **T004** three mutational clusters were detected on bulk exome sequencing, with a founder clone (orange) persisting throughout treatment, a subclone (yellow) that was detected at diagnosis but not at surgery, and a new clone emerging (green). The deep sequencing data analysis once again confirmed this clonal structure. The mutations in the founder clone were present throughout therapy, with the eradication of a subclone driven by *SMARCA4*, and the emergence of a subclone with a *POT1* mutation. Two of the variants within the subclonal population that become increasingly prevalent during chemotherapy (*RYR2* and *POT1*) were robustly identified at the diagnostic time point, albeit at low cellular allelic fractions.

In Case **T025**, four mutational clusters were present at diagnosis: three of these decreased in prevalence during treatment but remained detectable. Deep sequencing confirmed the presence of these clones, with the absence of one mutation in the clonal cluster that was predicted to be undergoing the greatest degree of clonal extinction (orange cluster in Figure 4.18).

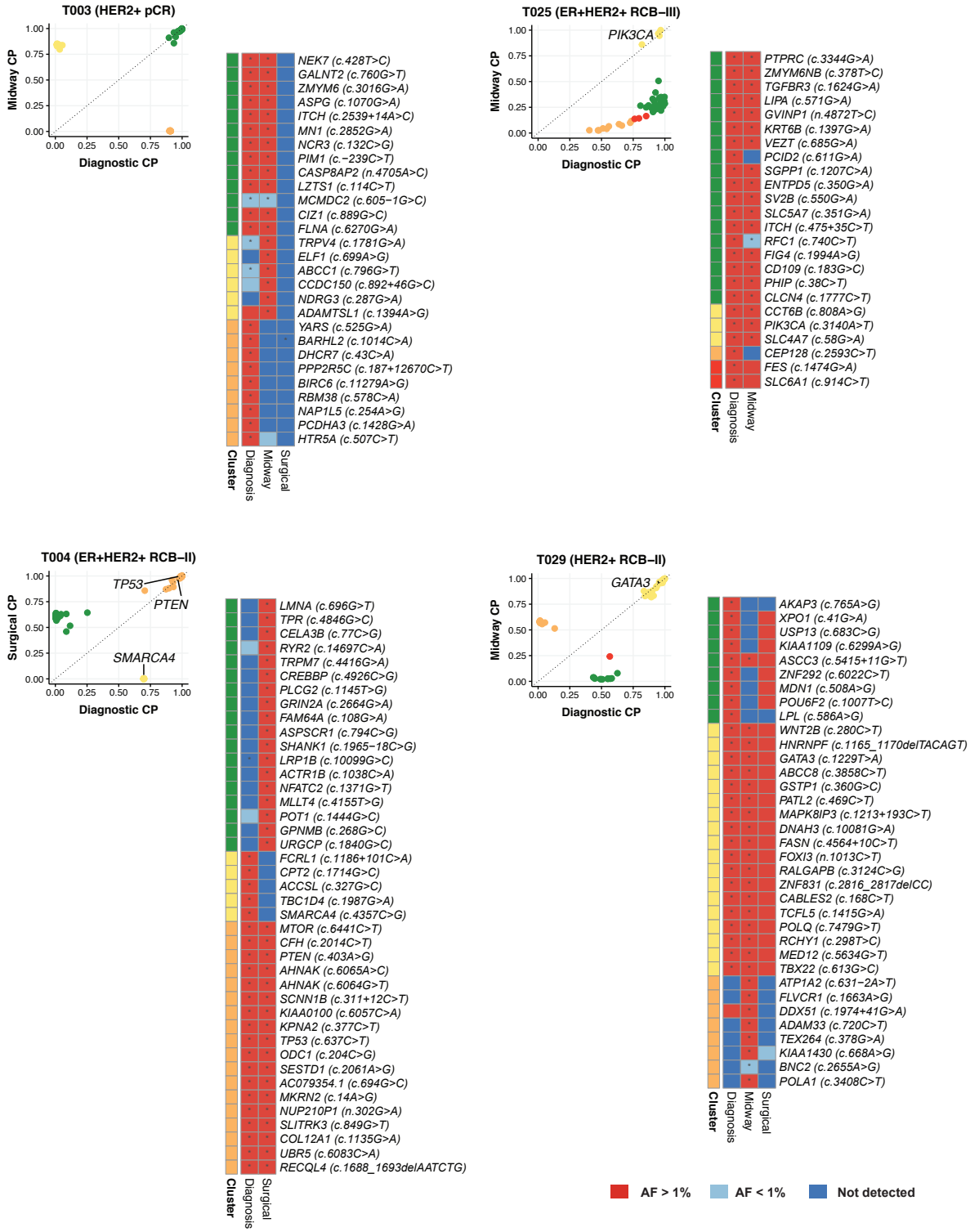


Figure 4.18: Deep sequencing of serial tumour samples from four cases that were also exome sequenced. Scatter plots show clonal structure derived from exome sequencing, heatmaps show deep sequencing results. Presence of an asterisk indicates identification of the mutation in the exome sequencing data. The surgical sample of case T029 was not exome sequenced as fresh tissue was not available.

Case **T029** had fresh tissue at two time points (diagnosis and midway) and FFPE tissue at the surgical time point: only the diagnostic and midway biopsies had been exome sequenced. The clonal structure identified by exome sequencing in this HER2$^+$ tumour was recapitulated from the deep sequencing data, though several observations can be made from this case. Midway through therapy, after three cycles of docetaxel, there was an apparent significant shift in structure, with one clone no longer detectable (green) and another emerging (orange). At the surgical time point though, the newly prevalent subclone was no longer detectable, though there was evidence of re-emergence of the previously undetectable subclone (green). This could be been either secondary to real shifts in dynamics to treatment secondary to differential chemoresistance and sensitivity to cytotoxic and targeted therapy (i.e. one clone was sensitive to docetaxel and resistant to trastuzumab and FEC, and vice versa for the second subclone), or secondary to spatial heterogeneity, as the FFPE sample was not taken from the same region the other two samples had been taken from.

In summary, the targeted deep sequencing experiment orthogonally validated the high accuracy of the exome sequencing pipeline and showed that some newly-detected mutations called at later time points in the exome data were actually detectable at very low VAFs in the diagnostic samples. Hence, these mutations were not generated by chemotherapy, but rather pre-existed as rare events within the treatment-naive tumour.

4.5.2 Multi-region deep sequencing

A recent publication has shown that clonal emergence was often secondary to pre-existing chemoresistant clones present in treatment-naive tumours, with clonal expansion occurring due to altered selection pressures induced by cytotoxic therapies [155]. In order to further determine whether mutations within clonally emergent subclones were present before commencing therapy, or whether they were induced and maintained by chemotherapy (as is the case when platinum therapy is administered in ovarian carcinoma), multi-region sequencing was performed on case **T004** (Figure 4.19). The clonal architecture during treatment, as defined by exome sequencing of the fresh frozen tumour samples, comprised three clonal clusters (Figure 4.19A), with one founding clone containing non-silent *TP53* and *PTEN* mutations, a subclone driven by *SMARCA4* that was no longer detectable during therapy, and a subclone driven by *MLLT4* that showed evidence of clonal emergence during chemotherapy.

Histopathological assessment of the diagnostic and surgical samples by *Dr Wei Cope* revealed a population of tumour cells that were very pleomorphic, with vesicular nuclei, prominent

nucleoli and frequent mitotic figures at both time points (red arrow in Figure 4.19B). However, whereas the diagnostic biopsy was relatively homogeneous for this morphology, at the surgical time point a second population of cells was identified that had not been observed in the diagnostic sample. These were less pleomorphic, smaller in size and with fewer mitoses, indicating a less proliferative phenotype (green arrow in Figure 4.19B). This apparent difference in morphology, with the identification of two distinct morphologies after chemotherapy, could have accounted for the significant genomic shift observed on exome sequencing.

Four cores were punched out of the diagnostic FFPE biopsy for this case (regions 8-11), and five cores taken from the surgical FFPE resection specimen (regions 3-7), as shown in Figure 4.19C. Two cores were taken from matched normal FFPE tissue (regions 1 and 2, not shown), and a further five cores taken from normal breast tissue from a different case to increase the number of controls for noise modelling and robust assessment of mutation status. Targeted amplification was performed using the Raindance platform.

The analysis of the deep sequencing experiment showed various salient points. Firstly, all the mutations selected within the founding cluster (cluster 3), were detected in all regions in the diagnostic and surgical samples. Secondly, most of the mutations present in the newly predominant subclone at the surgical time point (cluster 1, green) were also observed at the diagnostic time point, albeit at AFs of less than 1% in most cases, showing that these mutations were not generated by chemotherapy, but rather were already pre-existing at very low cancer cell fractions. Finally, Cluster 2, which was seen at diagnosis, but not at the surgical time point on exome sequencing, was not detected in any of the regions sequenced (except for *FCRL1*, which was seen at a low AF in one case at diagnosis). The mutations falling within these clusters were not false positives, as they were convincingly detected on deep sequencing of the fresh frozen DNA that was exome sequenced (Figure 4.18). Hence this could suggest that these mutations were not detected in the FFPE diagnostic sample due to tumour heterogeneity. Unlike the fresh frozen biopsies, which were often taken adjacent to the clip site, the diagnostic FFPE biopsy was taken before fiduciary marker insertion, and therefore was most likely taken from a different region of the tumour, explaining the difference. Hence the result of this experiment shows how important it is to consistently sample tumours from the same site, rather than performing unguided biopsies. Additionally, only a small number of mutations were targeted, and ideally a further experiment comprising more amplicons would be designed to investigate this further.

Clonal phylogeny construction using `Treeomics` (Figure 4.19D) showed the genomic relationship between the different sampled areas. The phylogenetic reconstruction mirrored

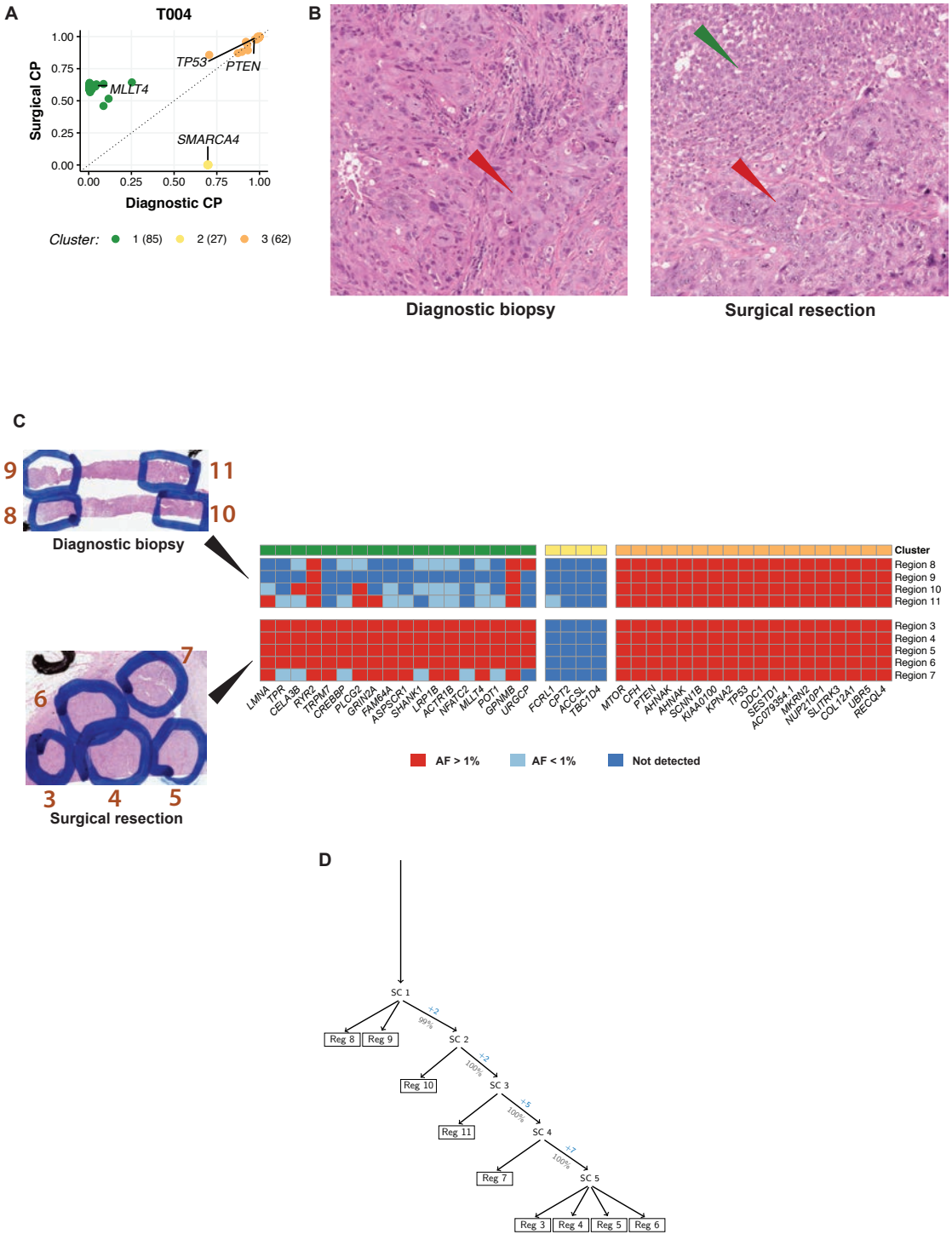


Figure 4.19: Deep multi-region sequencing of case T004. (A) Clonal architecture derived from exome sequencing data. (B) Differences in tumour morphology seen pre and post-chemotherapy, with highly mitotic and pleomorphic cells observed in the diagnostic and surgical specimens (red arrows), and less mitotic and pleomorphic cells observed only in the surgical resection (green arrow). (C) Multi-region analysis across 9 regions within the tumour, with a heatmap showing variant detection. (D) Seeding pattern across diagnostic and surgical specimens as predicted by the `Treeomics` algorithm.

very closely the anatomical regions of the specimens, with regions 8 and 9 on the diagnostic biopsy denoted as harbouring the originating tumour populations. Region 11 in the diagnostic biopsy was shown to have the greatest similarity to Region 7 in the surgical sample, which subsequently gave rise to the other populations sampled in the surgical biopsy.

4.6 Genomic architecture of relapse

Fourteen cases recruited to the TransNEO cohort relapsed by April 2018. Four cases relapsed regionally, whilst ten patients developed metastatic disease. All four regional relapses (**T020, T052, T078** and **T089**), as well as two brain metastases from case **T108** were sequenced and clonal phylogenies constructed (Figure 4.20A). Cases T020 and T052 had RCB-I RD whilst cases T078 and T089 had RCB-II RD following neoadjuvant therapy.

Of the four cases with local relapse, three showed a genomic architecture that was dynamic and different from the one sampled at diagnosis (T020, T052, T089: Figure 4.20A). This was very much akin to the dynamics observed during neoadjuvant chemotherapy, with therapy seemingly altering the clonal composition of the tumours. Indeed, in all three cases there was evidence of clonal eradication and clonal emergence perhaps secondary to the selective pressures induced by chemotherapy, with the emergent clone responsible for local relapse.

Case **T078** developed a regional relapse within the ipsilateral axilla: remarkably the clonal architecture of the relapse was identical to that seen at diagnosis, with two populations both present at identical cellular prevalences. This suggested that both tumour clones migrated together to the axilla and maintained the clonally stable architecture observed at diagnosis.

Case **T108** was the only patient with a distant relapse that was sequenced in this study. This patient was diagnosed with a lymph node negative, 30x29mm $ER^{-}HER2^{-}$ grade 3 tumour and developed respiratory failure following 4 cycles of TC chemotherapy. She was deemed unfit for surgery and instead received radiotherapy to the breast and axilla with the primary tumour left in situ. Unfortunately, 16 months after diagnosis she developed brain metastases in the right temporal lobe: two biopsies were taken from different regions of the brain tumour at the time of metastasectomy. The clonal phylogeny for the three tumours sampled (one diagnostic, two relapse) are shown in Figures 4.20A and B. The clonal structure at diagnosis was different from that in the two sampled nodules. The founding clone containing *TP53, NOTCH1* and *ESR1* mutations predominated in all samples. Cluster 4, which observed in the diagnostic biopsy, was no longer detectable in the relapsed tissue, indicating eradication by chemotherapy or failure of the clone to metastasise. However, three subclones that were not detected in the pre-therapy tumour were identified in the two brain metastases (clusters 2, 3 and 5, Figure 4.20C), with evidence of different clonal compositions in both metastatic sites (cluster 2 in relapse 1a, cluster 3 in relapse 1b). These newly-detected mutations could have been selected for by therapy or generated within the 16-month period between completion of therapy and relapse.

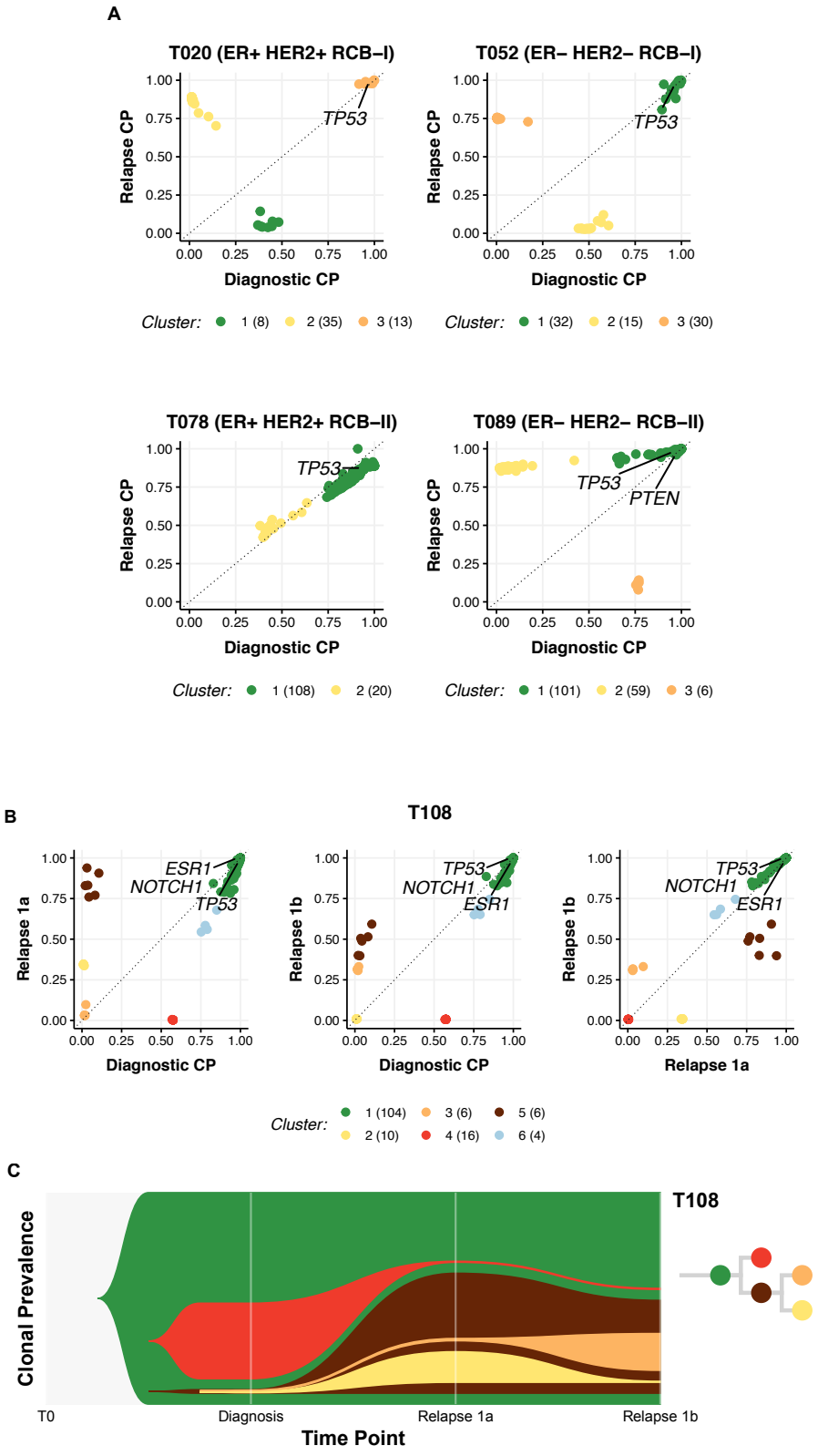


Figure 4.20: Clonal architecture of relapsed samples. (A) Regional relapses. (B) Comparison between diagnostic tumour sample and two brain metastases in case T108. (C) Riverplot showing temporal evolution in case T108.

4.7 Discussion

The work done in this chapter has described how tumour burden and genomic architectures were altered during neoadjuvant chemotherapy.

Serial MRI scans done on a subset of patients within this study showed that tumour volumes decreased during neoadjuvant chemotherapy, thereby allowing for an increased rate of breast conservation surgery. The rate of volume loss midway through treatment correlated with the degree of response observed by the end of treatment. Tumours that would eventually attain pCR decreased in volume faster than those with RD on completion of treatment, with a linear correlation observed between volume change midway through therapy and response at the end of therapy. Hence, volume change midway through chemotherapy was predictive of eventual RCB category. Additionally, an analysis of change in tumour purity based on both expert histopathological assessment and digital pathology analysis of the sequenced H&E slides showed that rate of loss of purity midway through chemotherapy also correlated with eventual response.

The tumour genomic landscape was seemingly altered by neoadjuvant chemotherapy. Through the use of serial tumour sampling it was very evident that the genomic architectures of the tumours studied changed during chemotherapy, with evidence of previously undetected mutations as well as the disappearance of mutations that had originally been detected prior to commencing therapy. An analysis done on the midway tumour samples showed that chemosensitive tumours had a greater number of subclonal mutations that were no longer detected on second sampling. Chemoresistant tumours showed very little change in mutational profile halfway through therapy. This observation was further confirmed by computing the Jaccard's coefficient of genetic similarity between the midway and diagnostic tumour samples. Serial tumours that attained pCR were much more dissimilar to each other, compared to those with higher degrees of RD, possibly secondary to the greater number of mutations that were no longer detected. By integrating a digital pathology analysis, the number of mutations that were newly-detected, eradicated, or persisting were not shown to be correlated to the number of cancer cells sequenced, showing that the dynamics observed were not confounded by tumour content.

Mutation dynamics were not limited to passenger mutations only but were also observed in non-silent mutations harboured within driver genes. Eight cases showed apparent loss of subclonal driver mutations, whilst five cases showed evidence of newly-detected subclonal driver mutations, with the computation of CCFs confirming that most of these mutations were

indeed subclonal rather than clonal. Pathway enrichment on the newly-detected mutations showed an over-representation of immune pathways (such as TCR and BCR signalling), proliferation and cell cycle pathways, as well as pathways involved in cellular motility and metastasis, with a high degree of statistical significance retained even after multiple test correction. As shown in Chapter 3, these pathways were shown to play key roles in determining response to chemotherapy, and therefore it was not surprising that mutations within these pathways were selected for in the chemoresistant tumour cells. Not only did some of these mutations possibly confer a survival advantage, but they were also less likely to be neoantigenic, with only 11% of the new (previously unsampled) mutations generating a putative neoantigen that could be presented by surface MHC class I molecules. Additionally, tumours that attained pCR had a greater decrease in neoantigen ratio during therapy, with a monotonic correlation observed across the different response groups.

The work done in this chapter described two types of genomic responses to chemotherapy: *Clonally stable tumours* were noted to harbour mutational clusters that were minimally altered during therapy. Hence the clonal composition during therapy remained remarkably stable with no evidence of cellular shift. *Clonally dynamic tumours* were characterised by altered clonal prevalences during treatment, with evidence of clonal extinction and/or clonal emergence, possibly indicating evidence of chemoresistant and chemosensitive subclones. Tumours with non-silent mutations within *CLU* and *MT-CYB* were associated with a clonally stable phenotype. Additionally, the copy number landscape reflected changes observed within the mutational landscape, with clonally stable tumours also having minimally alteration in the number of CNAs. Clonally dynamic tumours also showed evidence of CNA dynamics, increased chromosomal instability, and an increase in HRD during therapy.

By reconstructing clonal phylogenies, evidence of differential chemosensitivity and chemoresistance across different subclones was observed. The work described in this chapter mapped in detail the way chemotherapy altered the pre-existing tumour phylogeny, with multiple patterns of response observed. In some cases there was evidence of regression of aggressive tumour subclones generated later on in the tumour's evolutionary history and increasing prevalence of earlier, less proliferative, subclones. In some cases, subclones generated later on in the tumour's evolution and which were present at very low allelic frequencies at diagnosis were more chemoresistant and dominated the tumour landscape during therapy. While these observations could be due to tumour heterogeneity, it is worth noting that all tumour biopsies were taken at the site of a fiduciary marker in order to ensure consistent sampling across cases and thereby mitigate, to a degree, any spatial heterogeneity.

The serial tumour samples obtained at three time points allowed for the appreciation of clonal dynamics induced by chemotherapy backbone switch. In some cases there was evidence of differential chemosensitivity to chemotherapy agents, with evidence of clonal eradication during one half of the chemotherapy treatment, followed by increased cellular prevalence as soon as the agent was changed. Suboptimal chemotherapy exposure with multiple dose delays, reductions and omissions were also shown to have a minimal impact on the tumours genomic architecture, with extensive RD remaining at the time of surgery.

Deep sequencing and multi-region sequencing confirmed that the mutations that were becoming increasingly prevalent during therapy pre-existed prior to commencing chemotherapy and were selected for by treatment, rather than being generated by chemotherapy, as shown by a recent report by Kim et al. [155]. These mutations were present at very low allelic fractions in the diagnostic biopsy, often less than 1%.

Finally, a genomic analysis of tumours that eventually relapsed showed great diversity of the relapsed tumour compared to the primary tumour, with four of the five cases analysed harbouring different mutation profiles compared to the diagnostic sample. The patterns of genomic change echoed very closely those seen during chemotherapy and could have been indicative of the chemoresistant populations being responsible for the eventual relapse.

Chapter 5

The expression landscape during chemotherapy

Contents

5.1 Introduction

The expression landscape of breast cancer as it is subjected to neoadjuvant chemotherapy has not been very well described in the literature, with no studies attempting a comprehensive description of transcriptomic changes throughout chemotherapy.

Few studies have concentrated on attempting to describe the changes in the expression landscape between pre- and post-therapy samples. Unfortunately because of the small sample sizes in these studies very few meaningful inferences were made. It was consistently shown that chemosensitive tumours showed significant changes in the expression landscape, while chemoresistant tumours had a more stable expression landscape during treatment [113, 299], and that immune cell signatures were decreased in tumours with residual disease following chemotherapy [104]. Additionally, some studies also showed a switch in the intrinsic subtype classification of the tumours studied [104, 181].

Even fewer studies have attempted to describe the transcriptomic alterations observed using serial sampling during chemotherapy. Magbanua et al. [182] molecularly characterised tumours obtained before treatment, 24–96 hours after the first dose of chemotherapy, and at the time of surgery. 36 cases had matched pairs of the first two tumour biopsies and 39 cases had matched pairs of the first and last biopsy. Expression levels were determined at all time points and an analysis comparing diagnostic and early biopsies showed profound down-regulation of proliferation and immune-related genes during chemotherapy, with decreased expression of cell cycle inhibitors associated with poor response. Following completion of chemotherapy increased interferon signalling and increased expression of cell proliferation genes in any remaining RD was associated with reduced recurrence-free survival [182].

The work in this chapter aims to describe, at great depth, the transcriptomic changes in tumours and their surrounding microenvironment during neoadjuvant chemotherapy through the use of sequential tumour biopsies taken prior to commencing therapy, midway through therapy and on completion of chemotherapy. By integrating data from MRI imaging, transcriptomic shifts associated with response were noted, and mechanisms of chemoresistance elucidated. The involvement of the immune system in shaping tumour clonal evolution was also explored in detail, with BCR and TCR CDR3 deconvolution from the RNA-seq data used to corroborate findings. Additionally, the possible activation of pro-metastatic pathways during chemotherapy was also explored.

5.2 Early expression changes induced by chemotherapy

5.2.1 Mapping MRI dynamics to early expression changes

Numerous trials have shown that neoadjuvant chemotherapy is effective at reducing tumour bulk. In this study, the degree of volume loss by the midway time point was shown to correlate to the degree of RD at the end of neoadjuvant treatment (Figure 5.1A). Similarly, the change in tumour area on MRI (defined as the geometric mean of the largest bidimensional measurements) also correlated with the degree of RD by the end of therapy (Figure 5.1B).

The rate of disease reduction varied greatly between tumours, with some responding more rapidly to chemotherapy than others. To delineate the differences in transcriptomic pathways that were associated with the rate of tumour bulk change during therapy, RNA-seq data obtained from the midway biopsy was integrated with the change in tumour volume and area on MRI in order to answer the question: *Which pathways were responsible for greater decreases in tumour bulk on MRI?*

Twenty-two cases had matched MRI volumetric data (obtained from the TRICKS study) and RNA-seq data at the midway time point, while 32 patients had bi-dimensional measurements on MRI (acquired as part of standard of care imaging) and paired RNA-seq data at the midway time point. The larger dataset was taken forward in this analysis, however it is worth noting that very similar results were obtained when tumour volumes were used.

This differential analysis was performed using the linear model shown in Equation 5.1:

$$\sim Treatment + ER_{status} + HER2_{status} + \Delta_{area} \tag{5.1}$$

Treatment, ER_{status} *and* $HER2_{status}$ were categorical variables whilst Δ_{area} was the percentage change of the geometric mean of the two largest bidimensional measurements between the midway and diagnostic scans. The analysis revealed 30 lowly expressed genes and 346 highly expressed genes as tumour bulk decreased more rapidly (Figure 5.1C). Most notably, these included a large number of MHC class II molecules (*HLA-DMA, HLA-DMB, HLA-DOA, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, HLA-DRB1* and *HLA-DRB5*), as well as multiple members of the immunoglobulin Fc receptor gene family (*FCGR2A, FCGR2B, FCGR2C* and *FCGR3A*). Macrophages and neutrophils predominantly express these cell surface receptors, allowing the phagocytosis of opsonised antigens. Hence,

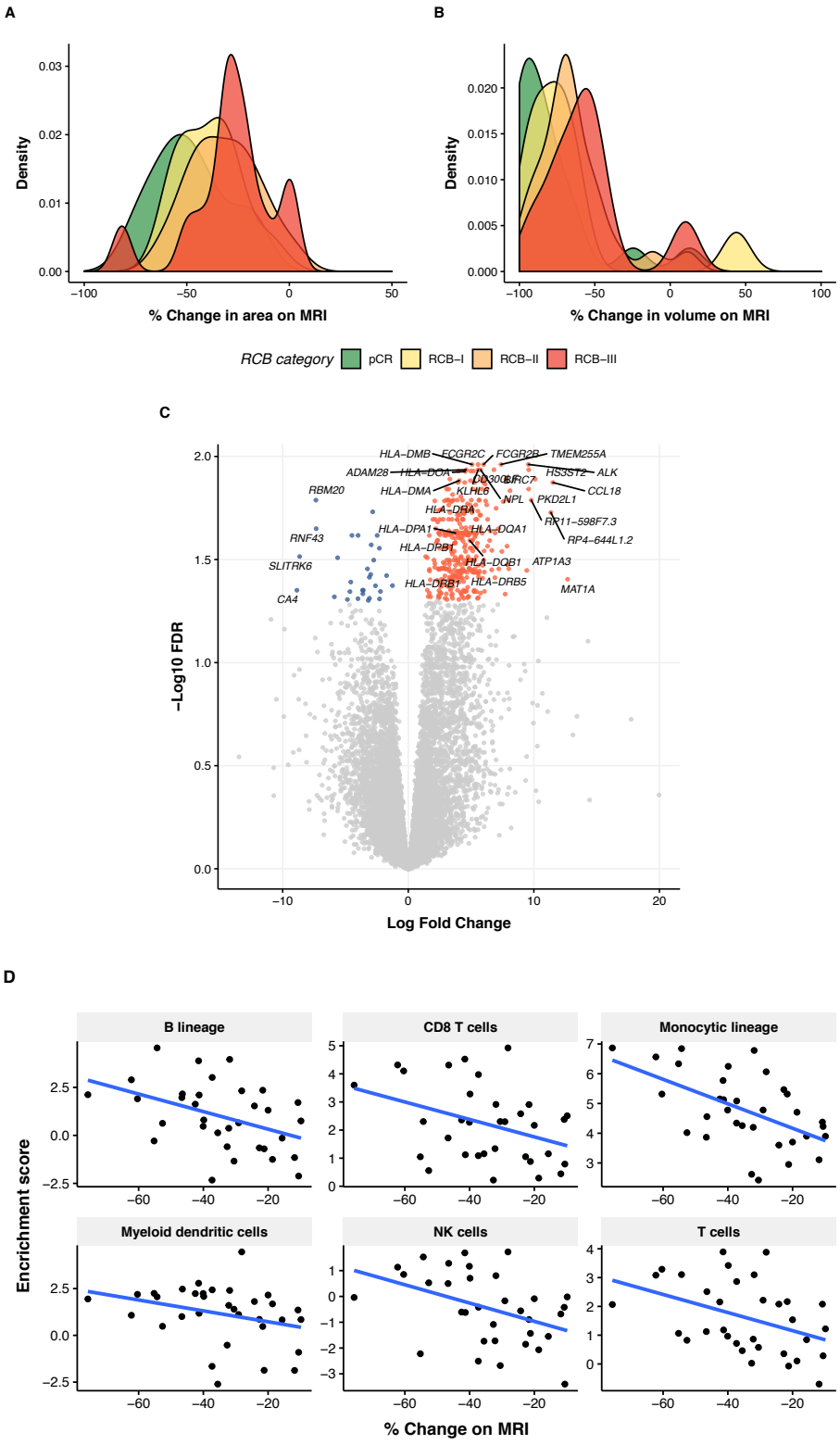


Figure 5.1: Correlating MRI data with transcriptomic signatures. Decrease in (A) tumour volume and (B) area during treatment, with tumours attaining better responses by the end of therapy decreasing in size more rapidly. (C) Differential expression volcano plot showing increased expression (in red) of multiple innate immune system genes associated with greater decrease in tumour bulk on MRI. (D) Correlation between immune infiltrate and percentage decrease in tumour bulk on MRI.

the innate immune system pathways appeared to play a greater role in determining the degree of tumour bulk reduction during neoadjuvant therapy. An enrichment analysis using `MCPcounter` [26] showed that various components of the immune infiltrate correlated with the percentage change in disease bulk (Figure 5.1D). T-cells, B-cells and NK cells, amongst others, were shown to be strongly correlated with response (p<0.05, Pearson's product-moment correlation): the higher the degree of immune infiltrate the greater the decrease in tumour bulk observed on MRI.

Gene set enrichment analysis (GSEA) and pathway enrichment using the Reactome database further supported these findings (Figure 5.2). The most up-regulated pathways pertained to immune system function: tumours exhibiting a greater degree of volume loss were more likely to have increased TCR and BCR signalling, as well as general augmentation of adaptive and innate immune pathways. In Chapter 3 it was shown that tumours with high proliferation at diagnosis were more likely to attain pCR (Section 3.4.2). At the midway time point, however, tumours that had a greater decrease in volume had significantly lower proliferation activation, in keeping with increased cell death in tumours that were responding to therapy (Figure 5.2A).

In order to orthogonally validate these observations, a similar enrichment was done on KEGG pathways, using the `gage` R package [180] with a significance cut off (q) of 0.1. Enrichment was observed over 14 KEGG pathways, of which 8 were immune system pathways, including: T cell receptor signalling (hsa04660), Antigen processing and presentation (hsa04612), Natural killer cell mediated cytotoxicity (hsa04650), B cell receptor signalling (hsa04662), Fc gamma R-mediated phagocytosis (hsa04666) and FcεRI signalling (hsa04664). Three pathways were down-regulated, including Oxidative phosphorylation (hsa00190), Cell cycle (hsa04110) and Ribosome (hsa03010). As shown in Figure 5.3, the extent of down and up-regulation of these pathways was very profound, with most members of the immune and cell cycle pathways statistically significantly altered.

In summary, the findings of this analysis continue to strengthen previous observations made in Sections 3.4.2 and 3.4.3: proliferation and immune activation played key roles in determining response to therapy. High proliferation and increased immune activation at the start of treatment were predictive of pCR and as therapy ensued an increase in immune activation and a decrease in proliferation were associated with a greater reduction in disease bulk on MRI.

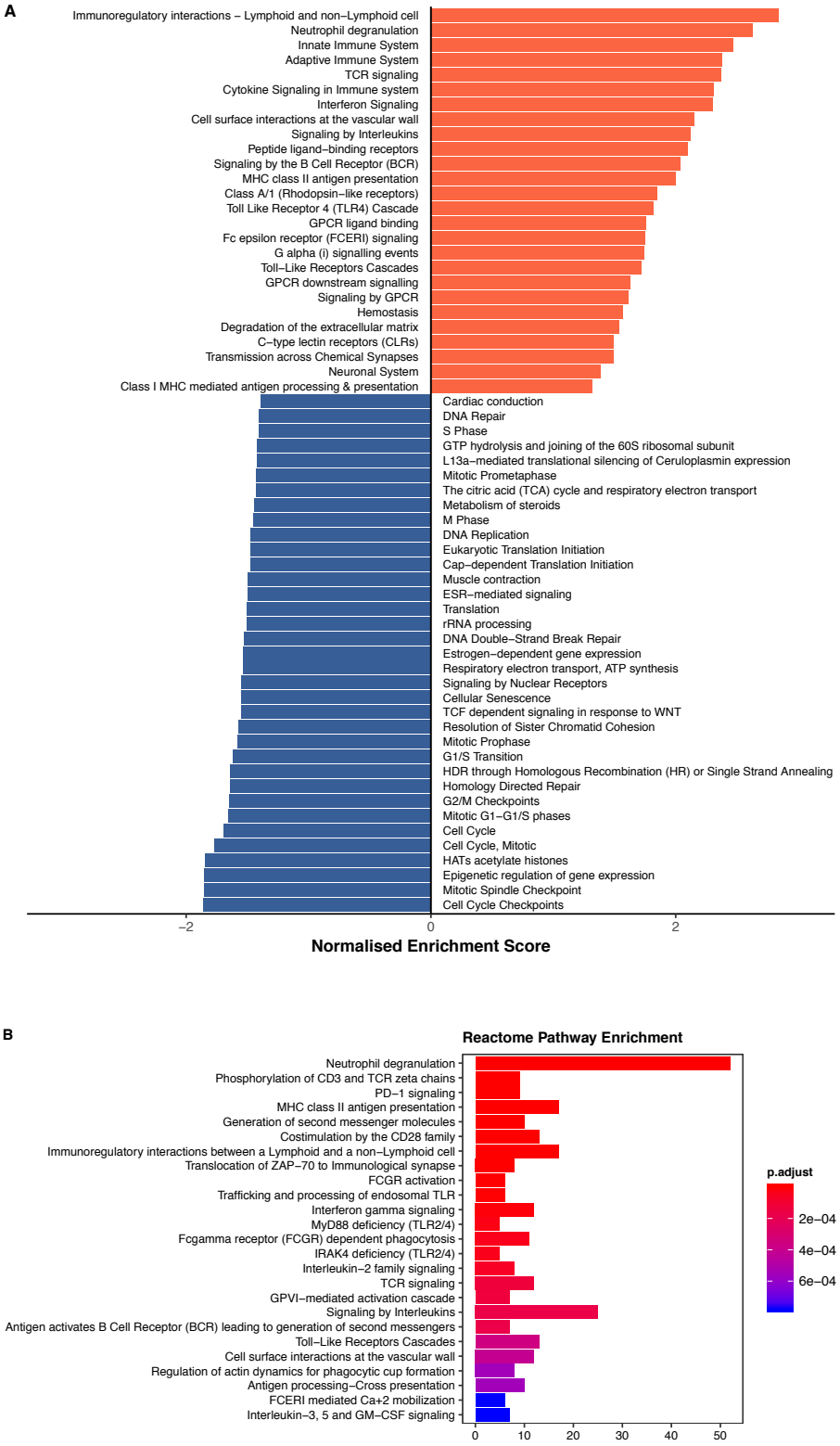


Figure 5.2: Differential expression analysis corresponding to increased degree of response on MRI. (A) GSEA over the Reactome database: proliferation and immune pathways strongly correlated with increased tumour bulk reduction on MRI. (B) Significantly overexpressed Reactome pathways associated with tumour bulk reduction on MRI.

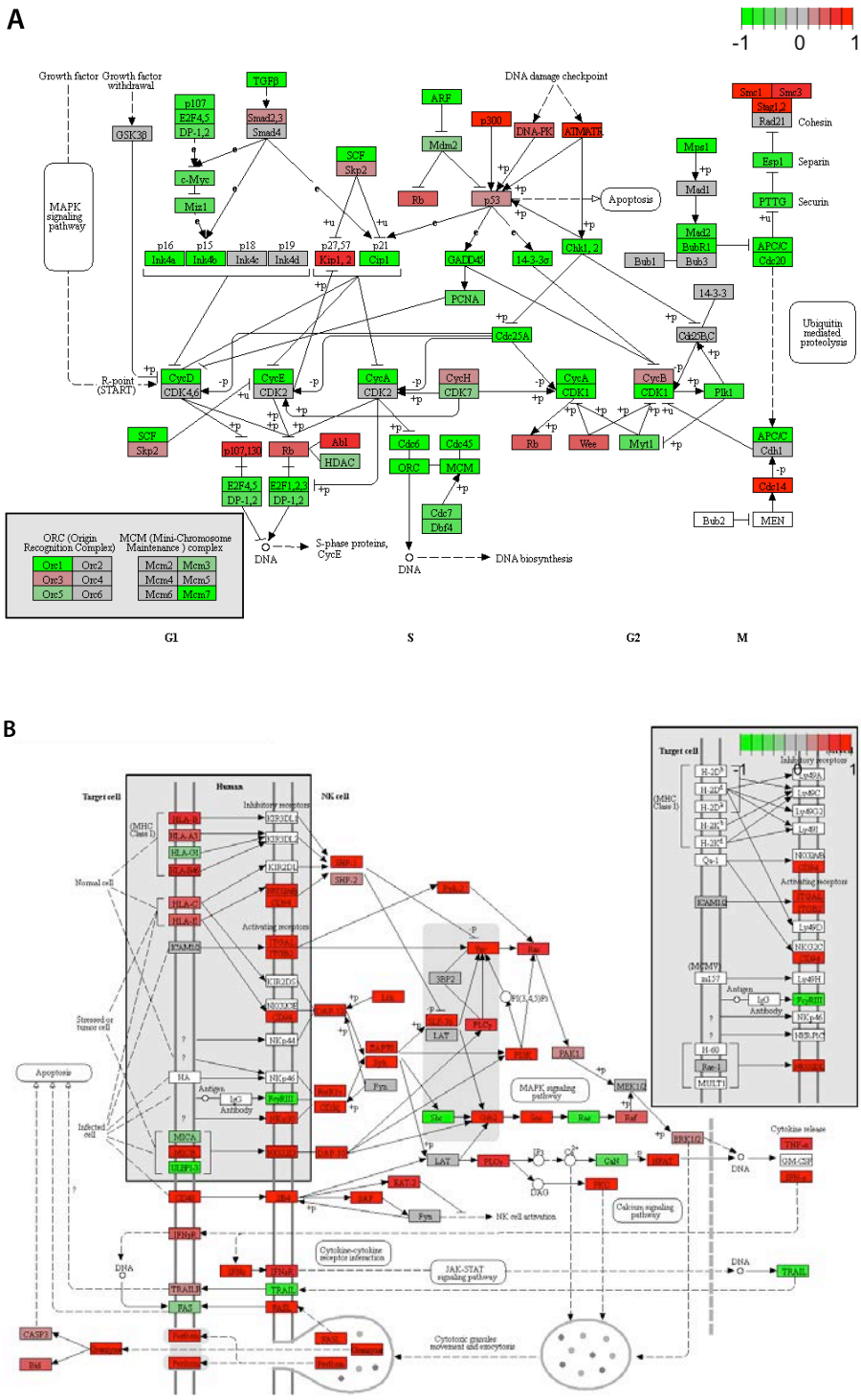


Figure 5.3: Enrichment over two KEGG pathways associated with increased disease bulk reduction on MRI: (A) Cell cycle (B) NK cell mediated cytotoxicity. Red indicates increased gene expression, green denotes decreased gene expression.

5.2.2 Transcriptomic changes associated with response

Midway expression profile analysis

The differential expression analysis in Section 5.2.1 described transcriptomic shifts associated with MRI dynamics and hence the response variable that was modelled was a snapshot of current tumour response on MR imaging. In order to determine whether the midway expression profile correlated with pathological assessment of response on completion of chemotherapy, expression data for 64 patients who had a midway biopsy was explored using a differential expression model incorporating ER, HER2 and treatment, with the response variable being the continuous RCB score. The analysis revealed 506 low expression and 404 high expression genes in tumours attaining pCR (Figure 5.4A), with MHC class II molecules being highly expressed, thereby providing further evidence that the innate immune system played a key role in determining response to therapy. The analysis also revealed that tumours that would eventually attain pCR had increased activation of cell death pathways, with increased expression of various apoptotic proteins, including *CASP1* and *CASP4* which both play a central role in the execution of apoptosis. High expression of the driver genes *ALK* and *JAK3* was associated with an increased probability of pCR, whilst high expression of *ESR1, AR, FOXA1* and *FGFR2* was associated with chemoresistance (Figure 5.4B).

A further Reactome gene set enrichment analysis showed profound enrichment over both immune and proliferation pathways, as was observed in the MRI analysis (Section 5.2.1). Tumours that attained better degrees of response by the end of therapy had higher immune activation and lower proliferation at the midway time point whilst chemoresistant tumours were more proliferative and had significantly less immune activation. To demonstrate the interplay between these two key pathways, a GSVA analysis was performed using the `GSVA` R package [114] on the Genomic Grade Index gene set [271] as well as the STAT1 immune gene set [68]. The relationship dynamics observed were strikingly different from those observed in the diagnostic biopsy (Figure 3.17). Whereas at the pre-treatment time point tumours that attained pCR had high proliferation and immune activation, at the midway time point tumours that would eventually attain pCR had lower proliferation but sustained high immune activation, whilst those with RCB-III RD had higher proliferation and sustained low immune activation (Figure 5.4C).

In order to gain an in-depth understanding of the specific immune pathways that were activated and corresponded to response to chemotherapy, ssGSEA was performed on the Reactome components of the Adaptive and Innate immune system pathways by using

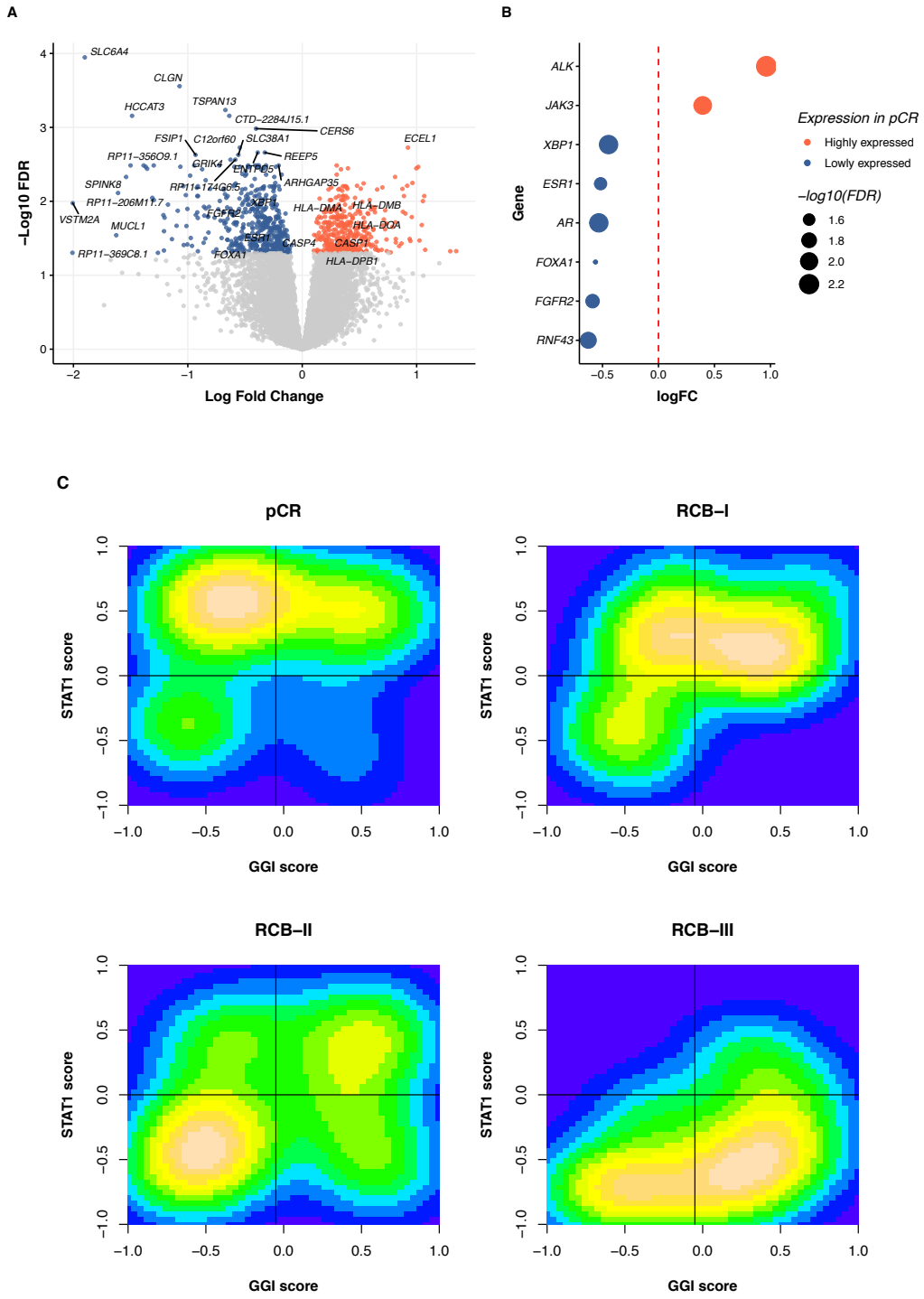


Figure 5.4: Correlation of midway transcriptomic signatures with eventual response to therapy. (A) Volcano plot showing top differentially expressed genes associated with chemosensitivity. Red indicates highly expressed in pCR, blue indicates lowly expressed in pCR. (B) Driver gene differential expression. (C) Association between proliferation (GGI) and immune (STAT1) pathways, and response to chemotherapy. Tumours that attained pCR had high levels of immune activation and low levels of proliferation, as opposed to tumours with higher burdens of RD. The yellow maxima can be seen shifting from the upper left-hand corner in tumours that attained pCR, to the lower right-hand corner in tumours with higher RD post-chemotherapy.

the GSVA R package, and the distribution of ssGSEA scores prior and midway through chemotherapy across different response groups analysed (Figures 5.5A and B). There was a remarkable correlation between the degree of enrichment of both innate and adaptive immune pathways and eventual response, more so than that seen at the diagnostic time point. This was most prominent for the innate immune pathways. In most cases (eg Neutrophil degranulation, Complement cascade, Antimicrobial peptides, TCR and BCR signalling) the relationship was monotonic, with a stepwise decrement in immune activation as the degree of residual disease post-chemotherapy increased. Unsupervised clustering over these pathways (Figure 5.5C) showed that tumours that were immunologically activated by the midway time point were much more likely to attain pCR or RCB-I RD, compared with those with significantly less activation.

In summary the expression landscape midway through chemotherapy was predictive of eventual response, with decreased proliferation and increased immune activation associated with pCR, and sustained proliferation and low immune infiltrate associated with chemoresistance.

Quantifying change in expression profiles

Computing the transcriptomic correlation between the diagnostic and midway samples using Pearson's product-moment correlation (Figure 5.6A) showed that the transcriptomic landscapes of tumours that were chemoresistant (RCB-III) was minimally changed by the midway time point (median R=0.96), compared to those that were chemosensitive (pCR R=0.90, RCB-I R=0.91), with the change in correlation related to the degree of RD post surgery (p=0.00005, ordinal logistic regression). Two archetypal examples are shown in Figures 5.6B and C: the dispersion in intra-sample correlation was significantly higher in tumours that would eventually attain pCR (such as case T136) than those that were chemoresistant (such as case T131). Hence, chemoresistant tumours were remarkably resilient and showed minimal change in the expression landscape, unlike tumours that were responding to therapy and whose transcriptomic landscape showed a greater degree of change secondary to altered tumour dynamics as well an altered microenvironment.

To highlight changes in pathways that correlated with response, a further differential expression analysis was performed that focused on changes in tumour transcriptomic pathways. Hence, rather than performing a differential analysis based on single time point samples, as had been done so far, each midway sample was normalised to its corresponding pre-therapy

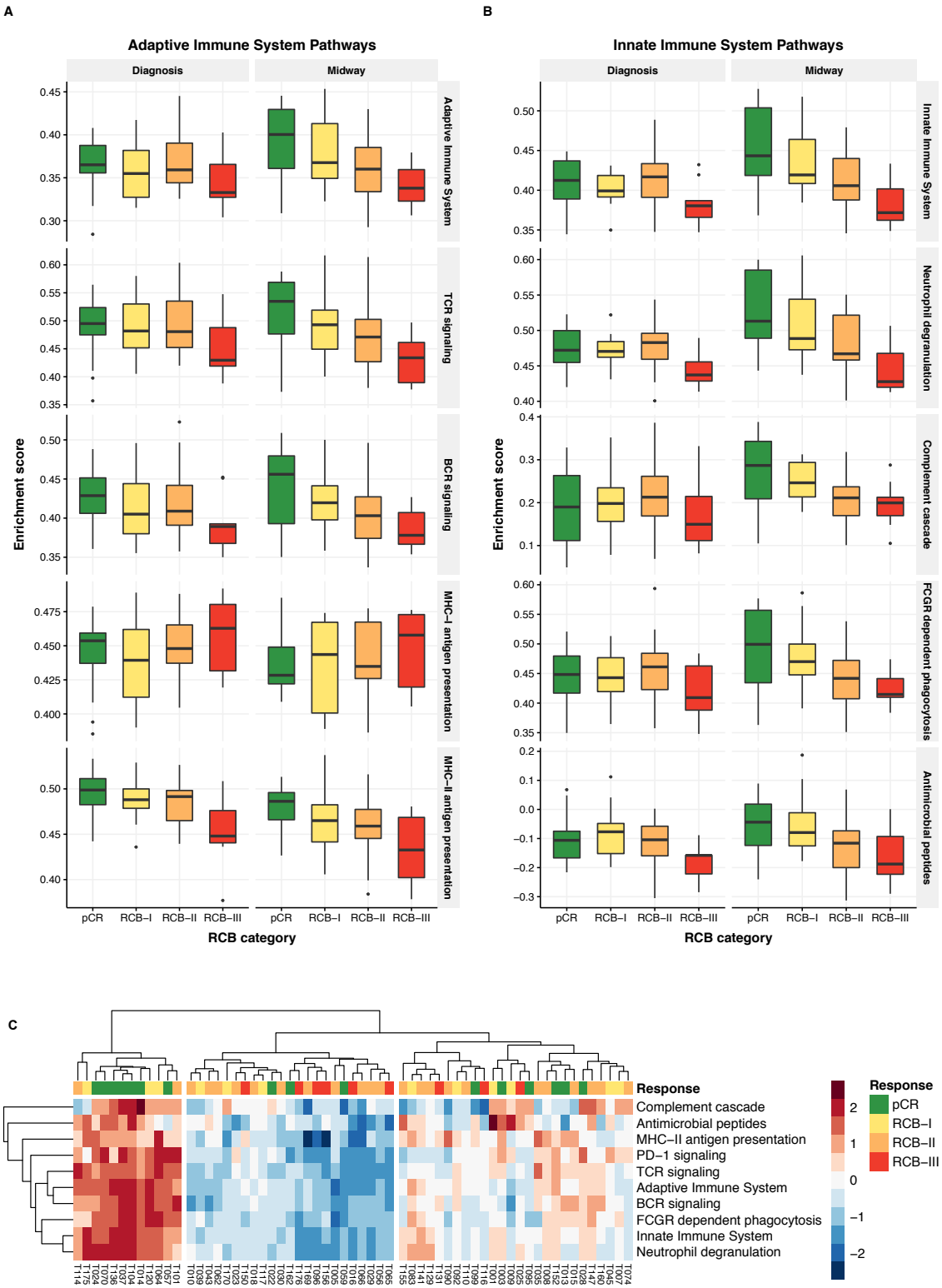


Figure 5.5: ssGSEA over Reactome immune pathways. (A, B) Distribution of ssGSEA scores for both immune and adaptive pathways at diagnosis and the midway time point. (C) Unsupervised clustering over Reactome immune pathways. Tumours with higher enrichment of immune pathways were much more likely to attain pCR or RCB-I RD.

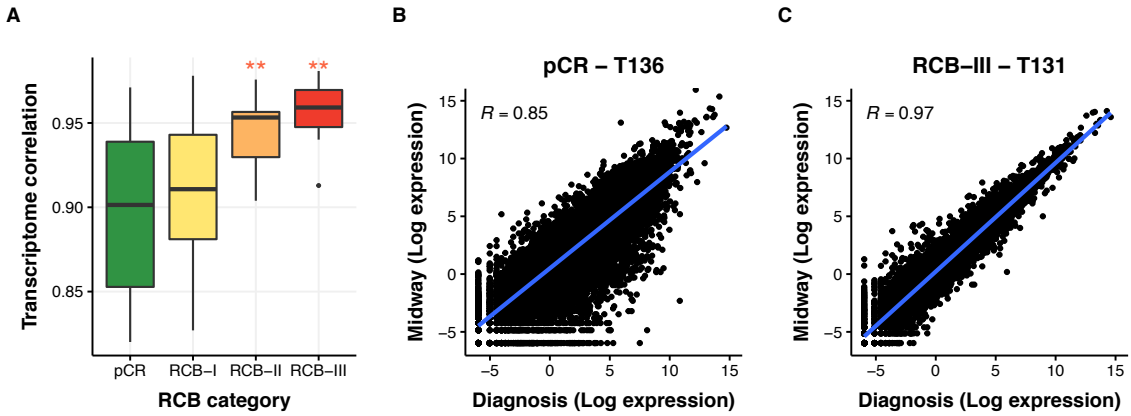


Figure 5.6: Changes in transcriptome correlation. (A) Chemoresistant tumours had a very stable transcriptome after chemotherapy was commenced, whilst chemosensitive tumours showed a greater degree of dynamics. throughout chemotherapy. Chemosensitive tumours (B) showed a greater degree of dispersion in correlation plots compared to chemoresistant tumours (C).

sample so as to identify changes in transcriptomic expression that were secondary to the therapy administered, thereby accounting for inter-tumour differences.

A differential expression analysis was performed on the midway and diagnostic samples using a nested factorial model, as shown in Equation 5.2:

$$\sim Response + Response : Patient + Response : Treatment \tag{5.2}$$

Where *Response* was the binary classification of response at surgery (pCR vs. RD), *Patient* was the patient identifier (required to identify related tumour samples) and *Treatment* was a variable containing information as to whether the tumour sample was obtained pre-chemotherapy or midway through treatment. Hence patients were nested within the response groups, and treatments nested within response groups. This differential expression model allowed the elucidation of response-specific treatment effects by normalising each midway tumour expression profile to the matched diagnostic expression profile.

This differential expression model was built using 61 cases with paired RNA-seq data at the diagnostic and midway time points and identified 975 genes that globally decreased in expression during chemotherapy and 709 genes that increased in expression in tumours that attained pCR, as shown in Figure 5.7A. Most notably, once again, these included a large number of MHC class II molecules (*HLA-DMA, HLA-DMB, HLA-DOA, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-DRB5,*

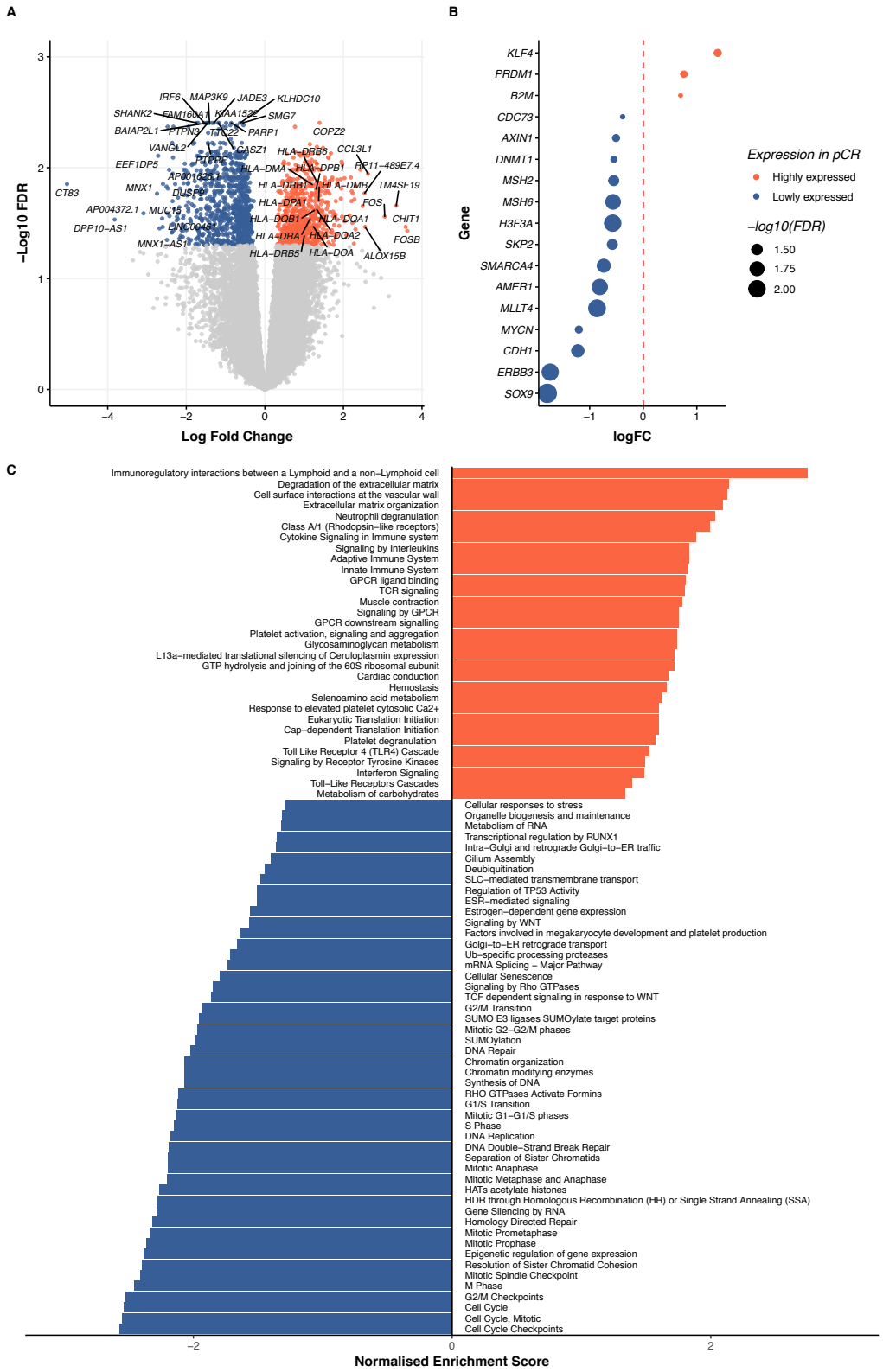


Figure 5.7: Differential expression of midway tumour samples normalised to matched diagnostic samples. Red indicates over-expressed in pCR, blue indicates under-expressed in pCR. (A)Volcano plot showing genes that were over or under-expressed in tumours that were more likely to attain a better response on chemotherapy. (B) Driver gene expression changes associated with response. (C) GSEA over the Reactome database: proliferation and immune pathways strongly correlated with response.

HLA-DRB6). $\beta 2$-microglobulin was the only MHC class I molecule that had an increased expression at the midway time point compared to the diagnostic sample. This result clearly showed that during therapy there was increased expression of MHC Class II molecules irrespective of what the baseline expression was. The increased expression of these antigen-presenting molecules was probably a consequence of increased tumour death and necrosis, with neoantigen release and increased antigen presentation by antigen-presenting cells, with expansion of the components of the innate immune system.

The key involvement of immune system components was further confirmed in a pathway enrichment over the Reactome database (Figure 5.7C). The results were highly similar to those seen in the volume analysis, which was unsurprising given that volume change on MRI was shown to correlate with final response at surgery. Tumours with increased immune activation and inflammatory infiltrate (neutrophil degranulation, platelet degranulation) attained pCR, whilst tumours that were responding to cytotoxic drugs became increasingly less proliferative, with less activation of DNA damage repair pathways compared to those with decreased response to treatment.

Interestingly, as was also seen in the MRI analysis (Figure 5.2), tumours that were chemoresistant showed increased activation of senescent profiles, perhaps showing that tumours that were not as responsive to chemotherapy were entering a state of therapy-induced senescence (TIS) secondary to cytotoxic induced stress. To validate this, a list of TIS genes were obtained from ***Dr Masashi Narita's*** research group in Cambridge and the change in expression of these genes computed. An increase in TIS signatures was observed in chemoresistant subtypes (data not shown), showing that cellular senescence was potentially a mechanism of evasion to therapy. In order to validate this orthogonally, senescence-associated beta-galactosidase immunohistochemistry could be performed on the pre and on-therapy tumour slides.

A driver gene analysis (Figure 5.7B) showed that DNA repair proteins *MSH2* and *MSH6* were down-regulated in tumours responding to chemotherapy, showing decreased activity in DNA repair pathways in tumours that would eventually attain pCR. Tumours that were sensitive to chemotherapy were less likely to activate DNA repair pathways. The most highly expressed driver gene was *KLF4*, a zinc-finger transcription factor and oncogene that has recently been associated with the promotion of cancer stem-cell like programmes [231, 283, 284] and has been known to confer worse prognosis in breast cancer [88]. Interestingly, it seemed that in this case, a stem-like pathway was being activated in those tumours that were responding to chemotherapy.

To define pathways associated with monotonic dynamics during treatment, ssGSEA was performed in all matched tumour samples over all Reactome pathways. Following this, the difference in ssGSEA scores between the initial and midway biopsies was computed and ordered ordinal models generated to predict pathways for which there was a monotonic decrease or increase across all response groups. A total of 152 pathways were identified as exhibiting trends that correlated with response (FDR < 0.01, Figure 5.8).

99 pathways showed a profound decrease in expression in tumours attaining pCR and little or increased expression in tumours with higher burdens of RD. The most statistically significant pathway in this analysis was cholesterol biosynthesis, where tumours that would eventually attain pCR had a greater likelihood of having decreased cholesterol metabolism during treatment, with the degree of change in expression compared to the diagnostic time point monotonically correlating with the degree of RD. The role of lipid metabolic reprogramming in tumour cells has been well characterised [28, 43, 265], with increased expression of lipogenic enzymes such as acetyl-CoA carboxylase and fatty acid synthase occurring in most tumours [195], and increased expression associated with poorer prognosis [161]. As described previously, DNA repair mechanisms, mitosis, and methylation pathways were greatly decreased in tumours that were attaining a better response. Interestingly, RCB-III tumours appeared to gain activation of cell motility pathways during treatment [186], perhaps showing that as chemotherapy was administered, activation of pro-metastatic pathways was induced. This phenomenon was initially observed in Chapter 4 (Table 4.1), where an enrichment of new mutations in pro-metastatic pathways was noted. This observation will be explored further later on in Section 5.3.3 where similar observations were noted using a different analytical approach.

Pathways associated with a greater decrease or smaller increase in chemoresistant tumours included apoptotic pathways (such as caspase activation) and innate immune pathway enrichment (Figure 5.8B). There was predominant enrichment of the Toll-Like Receptor signalling pathway [154], including the TLR2, TLR4, TRL7/8, TLR9 cascades, as well as increased scavenger receptor A pathways found in macrophages, showing a profound activation of innate immune system components.

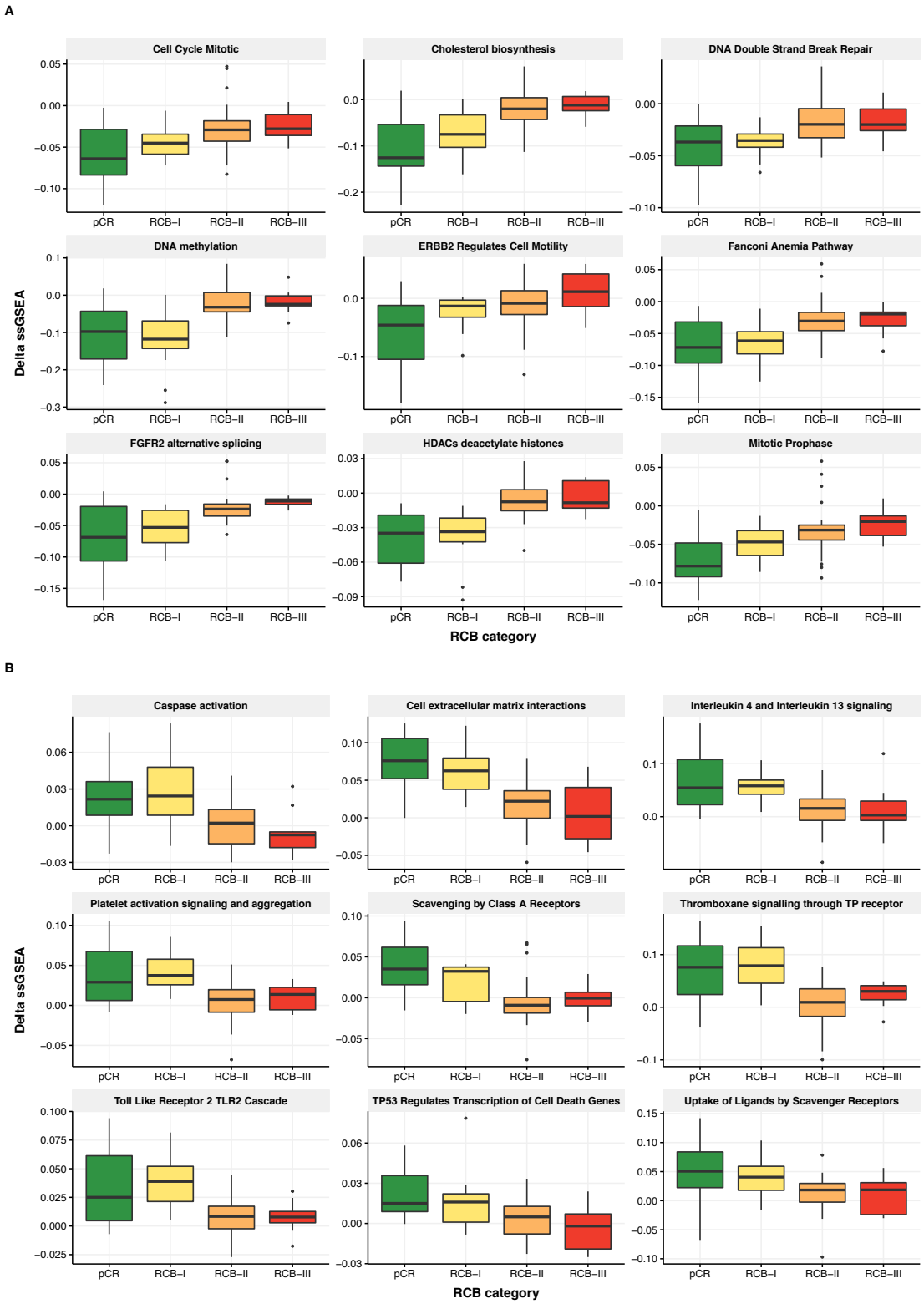


Figure 5.8: ssGSEA on Reactome pathways, showing a selection of pathways in which the rate of change was associated with response. (A) Pathways in which expression decreased much more profoundly in tumours attaining pCR than those with RD. (B) Pathways in which expression increased much more profoundly in tumours attaining pCR than those with RD.

5.2.3 Tumour immune microenvironment dynamics

T cell mediated cytolysis

To further describe the immune system dynamics, the cytolytic activity score (CYT) was computed for all paired samples (Figure 5.9A). At the diagnostic time point there was a very clear association between CYT and response at diagnosis (p=0.003) and this association strengthened during treatment (p=0.0008, ordinal regression). Hence, T cell-mediated cytolysis continued to play a key role in determining response to chemotherapy.

Taxane treated tumours, which were often HER2$^-$, exhibited a significant increase in CYT midway during chemotherapy if pCR occurred by the end of therapy (Figure 5.9B). Conversely, no statistically significant increase in CYT was evident at the midway time point if RD was present at the time of surgery, showing that early events occurring after commencing chemotherapy were predictive of response to therapy. An inverse relationship was observed in anthracycline-treated tumours, which were often HER2$^+$ (Figure 5.9B), where an increase in CYT was associated with RD and an initial decrease in CYT associated with pCR. Whether this was due to anthracycline treatment or was a biological feature of HER2$^+$ tumours, or indeed a combination of both, was difficult to untangle. Results from the *I-SPY 1* trial showed a decrease in immune activation in tumours treated with anthracyclines [182], whilst the *Neo-tAnGo* trial [73] showed that neoadjuvant treatment with a taxane prior to an anthracycline was superior to the reversed regimen, with more patients attaining pCR if a taxane was administered first (20% vs. 15%). Hence, trial evidence also suggests differences between anthracycline-first and taxane-first treated tumours.

IPS and HLA LOH dynamics

As the CYT score was a measure of the expression of two key genes involved in T cell mediated cytolysis, further evidence was sought to determine whether CYT dynamics correlated with other components of the immune system, such as those described by the immunophenogram score (IPS) [51]. As shown in Figure 5.9C, the four IPS components showed a statistically significant correlation with CYT. Tumours that had a CYT increase during chemotherapy had an associated increase in MHC molecules, effector and suppressor cells (EC, SC) as well checkpoint (CP) molecules.

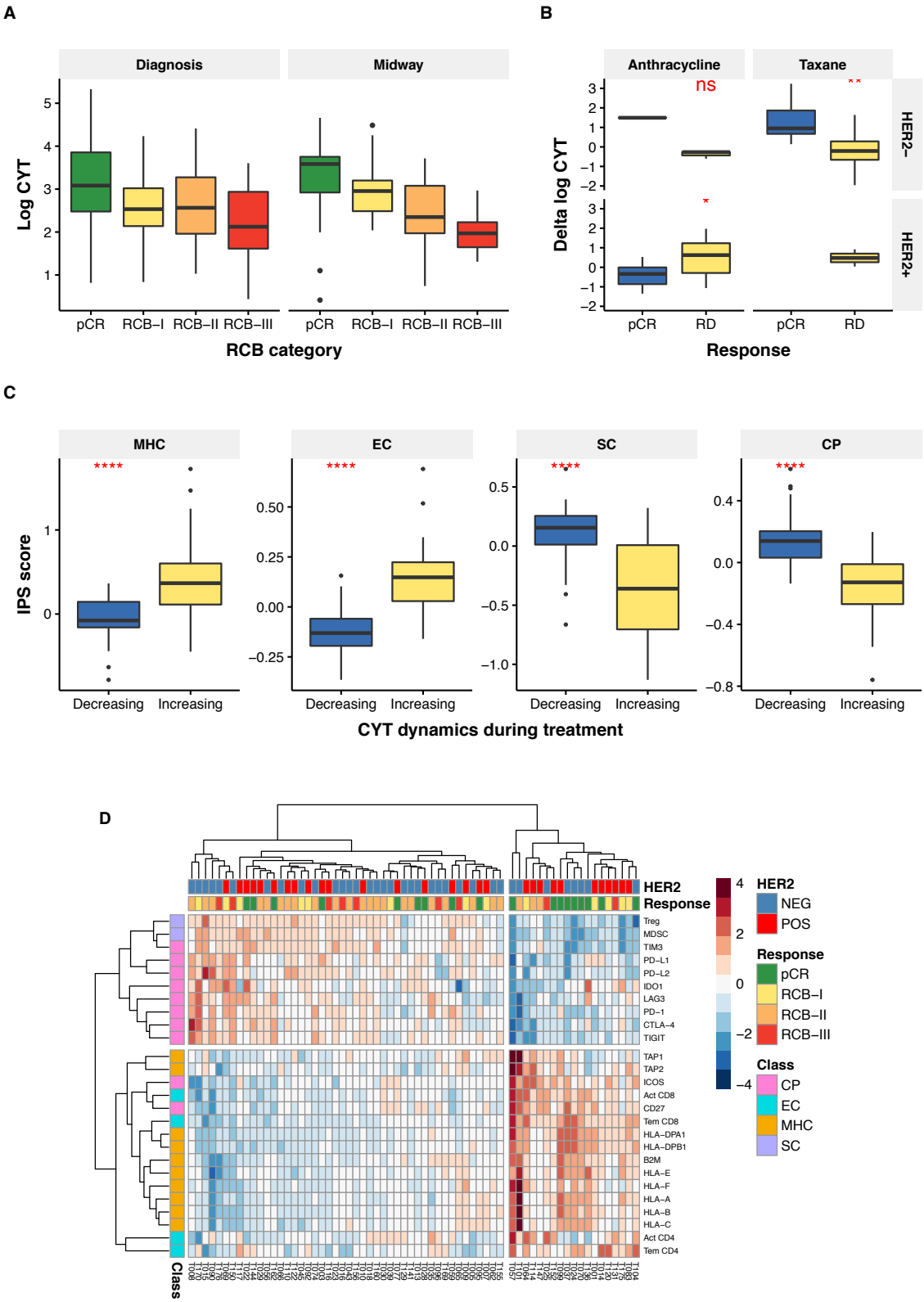


Figure 5.9: Immune dynamics during treatment. (A) Positive correlation with CYT seen at diagnostic time point more evident at midway time point. (B) Correlation between CYT and response in anthracycline and taxane treated tumours. (C) Correlation between CYT dynamics and individual components of the IPS score (EC = effector cells, SC = Suppressor cells, CP= checkpoints). (D) Unsupervised clustering of IPS components and midway samples. Taxane treated tumours that attained pCR were immunologically more active than anthracycline-treated tumours that attained pCR.

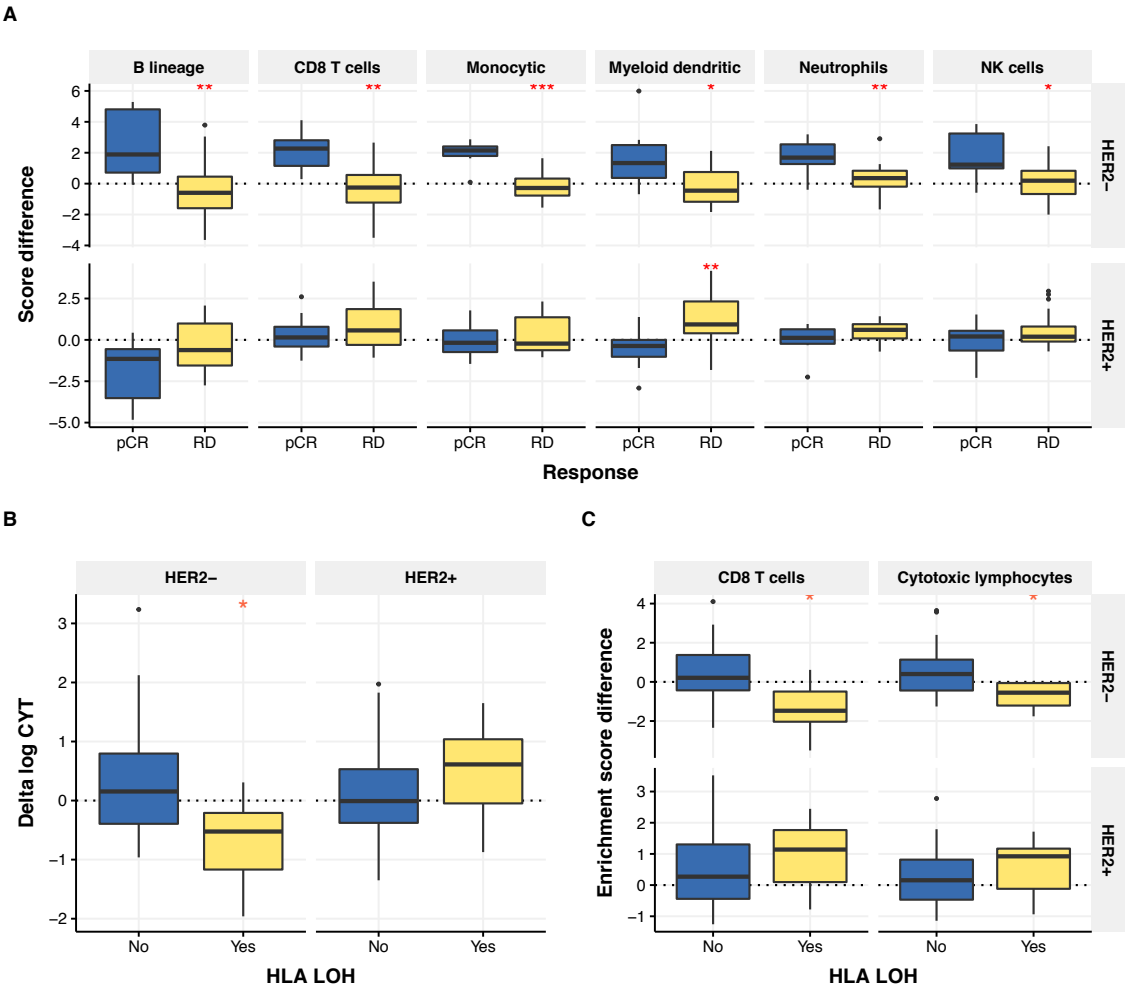


Figure 5.10: Change in immune landscape between diagnostic and midway time points. (A) HER2$^-$ taxane treated tumours that attained pCR had an increase in the degree of innate and adaptive immune cell enrichment. (B, C) Taxane-treated tumours with HLA LOH at diagnosis had a decrease in CYT during therapy, compared to tumours with no HLA LOH. A similar pattern was observed on enriching for CD8 T cells and cytotoxic lymphocytes.

Unsupervised clustering revealed that most of the taxane-treated HER2$^-$ tumours that attained pCR were immunologically active (Figure 5.9D), whilst the anthracycline-treated HER2$^+$ tumours that attained pCR clustered in the immunologically 'poor' group, providing further evidence to differing immune activation patterns possibly secondary to the different sequence of chemotherapeutic agents used.

Tumour microenvironment deconvolution using `MCPcounter` [26] also showed a shift in the immune tumour microenvironment (Figure 5.10A) and corresponded to the observations seen with the CYT metric. HER2$^-$ tumours that attained pCR had an increase in B cell, CD8 T cell, neutrophil and NK cell score during therapy. This was not observed in HER2$^+$ tumours, where an initial increase in immune infiltrate was not associated with pCR.

Interestingly, the presence of HLA LOH in the taxane treated HER2$^-$ tumours correlated with CYT dynamics. Tumours without LOH over the HLA locus had an increase in CYT score during therapy, whilst those with LOH did not (Figure 5.10B). A similar pattern was also evident on enriching for T cell populations using `MCPcounter` (Figure 5.10C).

CDR3 dynamics

The Gini coefficient is used in economics to represent the wealth distribution of a nation and is a measure of inequality. The lower boundary of the coefficient (0) indicates perfect equality, whilst the upper boundary (1) indicates inequality, with all the wealth distributed amongst a privileged few. Calculation of the Gini coefficient using the `tcr` R package [207] on the TCR and BCR CDR3 data obtained from RNA-seq deconvolution using the `MiXCR` package [34] showed that the Gini coefficient at diagnosis was associated with response (Figure 5.11A, p=0.002 ordinal regression). Tumours that had dominance of few CDR3 clonotypes were more likely to attain pCR than those with a greater variety of clonotypes in the tumour microenvironment. Additionally, this observation continued to hold true midway through chemotherapy for HER2$^-$, but not HER2$^+$ tumours (p=0.02 and p=0.15 respectively, Figure 5.11B). Hence, CDR3 clonotype dominance was positively associated with response.

This observation was further strengthened through a correlation with CYT (Figure 5.11C). The Gini coefficient was shown to positively correlate with CYT (R=0.53, p<2.2e-16), with TCR clonotype dominance associated with increased T-cell mediated cytolysis.

The CDR3 similarity between the diagnostic and midway samples in tumours that attained pCR was lower than that observed in tumours with RD post chemotherapy (Figure 5.11D).

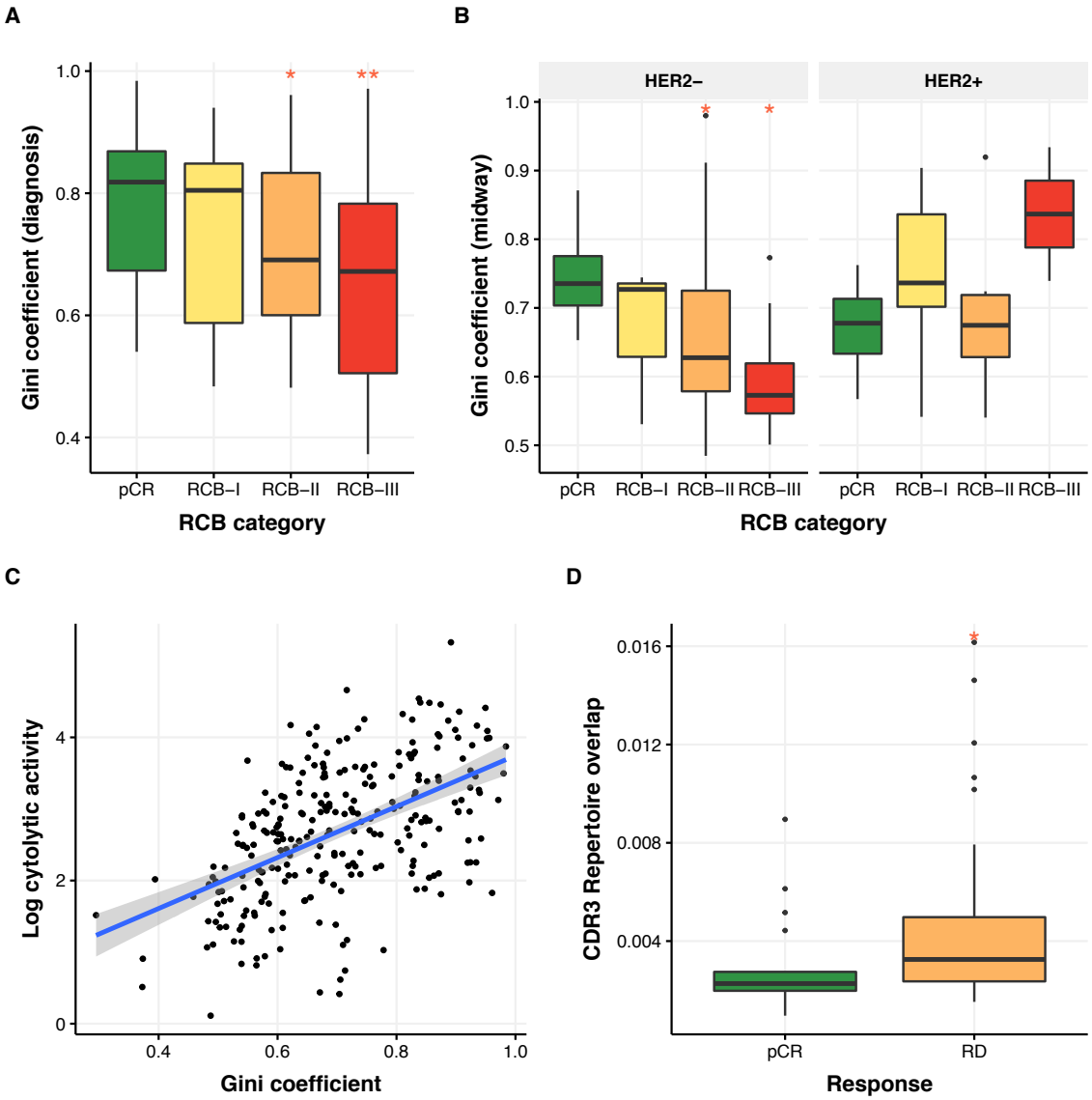


Figure 5.11: CDR3 deconvolution from bulk RNA-seq data and estimation of clonotype diversity. (A) Gini coefficient at diagnosis was associated with response. (B) At the midway time point, HER2$^-$ tumours continued to show an association between decreased CDR3 diversity and response. (C) Positive correlation observed between CYT and Gini coefficient. (D) Box plots showing distribution of CDR3 repertoire overlap between midway and diagnostic time points: tumours that attained pCR had a lower degree of overlap between time points, indicating expansion of a select few clonotypes, making rarer CDR3 sequences that were previously detected less prevalent.

This indicated that, as chemotherapy was administered, selective clonotype expansion of few T and B cells occurred in tumours responding to chemotherapy, with less diversity observed at the midway time point due to clonal expansion and preferential sequencing of the more abundant clonotypes.

In summary, the tumour immune microenvironment was dynamic during therapy, and showed changes that were associated with RCB score at the end of neoadjuvant therapy. An upregulation of immune infiltrate, T-cell mediated cytolysis as well as TCR/BCR clonotype expansion were all associated with better response to therapy in taxane-treated HER2$^-$ tumours.

5.2.4 Integration of clonal and expression dynamics

The work discussed in Chapter 4 has provided evidence that tumour clonal architecture can be altered by chemotherapy. Subclones that were sensitive to therapy decreased in prevalence, and consequently those resistant to therapy increased in prevalence.

The genomic analysis showed that mutations in *CLU* and *MT-CYB* were associated with clonally stable tumours. In order to determine transcriptomic differences between the different genomic types of response, tumours with evidence of clonal emergence only (in which subclones were becoming increasingly prevalent), and those with evidence of clonal extinction (in which subclones were decreasing in clonal prevalence) were compared to clonally stable tumours using a differential expression model where the baseline comparator comprised midway samples from clonally stable tumours.

Clonally stable tumours were less proliferative than clonally dynamic tumours (Figure 5.12). This offered an explanation as to why these tumours showed very little alteration in clonal prevalence during therapy: as these tumours were less proliferative they were less likely to be targeted by cell cycling drugs.

Key expression differences were observed between the tumours undergoing clonal extinction and those with clonal emergence (Figure 5.12). The immune system played a profound role in determining the evolutionary trajectory of a tumour: tumours with evidence of ongoing clonal extinction had significantly higher enrichment of immune pathways, perhaps indicating that a combination of T-cell induced cytolysis as well as chemotherapy-induced cell death was responsible for the elimination of these subclones. On the other hand, tumours with

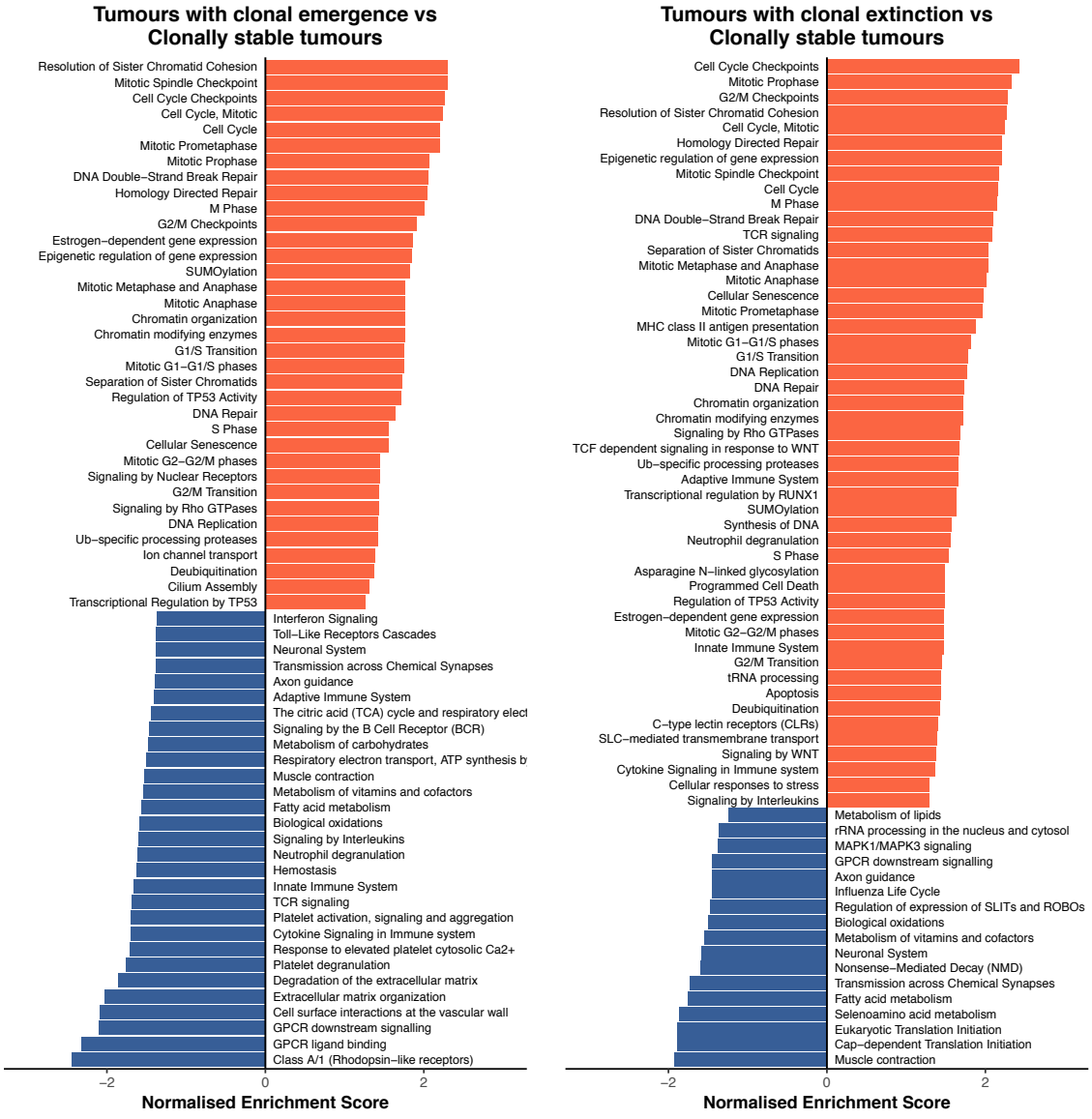


Figure 5.12: Reactome pathway enrichment on differential expression model of tumours undergoing clonal extinction or emergence versus clonally stable tumours.

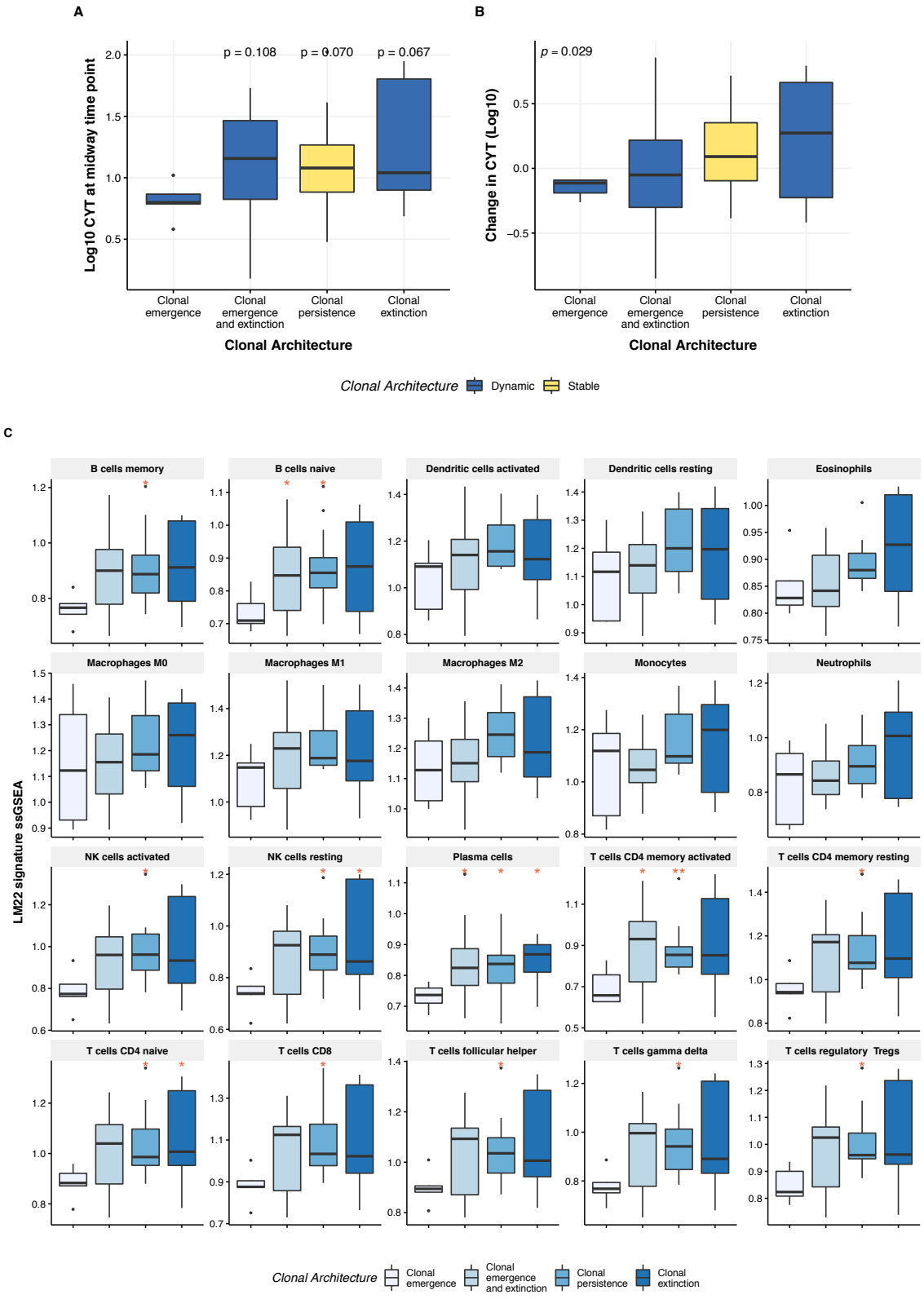


Figure 5.13: Immune dynamics in different genomic response subgroups. (A) CYT score at midway time point across different subclonal dynamics. (B) Delta CYT across different groups. (C) ssGSEA over LM22 gene set, showing a selection of 20 immune cell populations. Tumours with clonal emergence had statistically significantly lower degrees of immune cell infiltrate (marked with *) compared to tumours exhibiting other types of response.

clonal emergence had decreased TCR, BCR and interleukin signalling, as well as decreased enrichment over high level innate and adaptive immunity pathways.

Analysis of CYT scores across the different response groups showed that tumours with clonal emergence had a statistically significant lower CYT score at the midway time point compared to tumours that exhibited clonal extinction only (Figure 5.13A). The change in CYT score between diagnostic and midway time points monotonically correlated with the type of architecture observed (Figure 5.13B), with tumours showing evidence of clonal emergence having a decrease in CYT, while those with evidence of clonal extinction only having an increase in CYT. Enrichment over the LM22 gene set used by `CIBERSORT` further corroborated these findings (Figure 5.13C).

In summary, clonally stable tumours were less proliferative than clonally dynamic tumours. Tumours with clonal emergence has statistically significantly less immune activation compared to tumours with clonal extinction.

5.3 Late expression changes induced by chemotherapy

5.3.1 Mapping MRI dynamics to late expression changes

The analysis of tumour samples taken midway through therapy showed significant transcriptomic dynamics during neoadjuvant therapy, with tumours that attained better degrees of response more likely to exhibit decreased proliferation and increased immune activation.

To correlate the decrease in tumour bulk on MRI to transcriptomic pathways, the change in tumour area was computed as discussed in Section 5.2.1. This continuous response variable was then integrated into a differential expression model (Equation 5.3) to identify genes and pathways associated with increasing chemosensitivity.

$$\sim Treatment + ER_{status} + \Delta_{area} \tag{5.3}$$

ER_{status} and *Treatment* were categorical variables whilst Δ_{area} was the percentage change of the geometric mean of the two largest bidimensional measurements between the final and diagnostic scans. HER2 status was not included in this model, as all HER2$^+$ tumours used in this analysis were exposed to the same therapy sequence, and all HER2$^-$ tumours were exposed to a similar sequence, and therefore the effect of HER2 status and treatment could not be distinguished from each other.

The output from the differential expression model was used to perform an enrichment over Reactome pathways, and the analysis once again revealed that immune pathway activation corresponded with response to chemotherapy. Chemoresistant tumours had significantly less BCR, TCR, interleukin and interferon signalling compared to chemosensitive tumours, as well as having increased proliferation and steroid metabolism (Figure 5.14A). Of note, the most highly expressed pathway in chemoresistant tumours was *ESR1* signalling, which was unsurprising as ER$^+$ tumours have been shown to be more chemoresistant in multiple studies as well as in this work.

As had been seen at the midway sampling time point, the decrease in tumour bulk on MRI also corresponded to the enrichment for various immune populations (Figure 5.14B), including T cells (R= -0.56, p=0.036), CD8 T cells (R= -0.76, p=0.001), NK cells (R= -0.78, p=0.001) and B cells (R= -0.64, p=0.01). Hence the involvement of the adaptive and innate components of the immune system played a key role in determining response throughout therapy.

Figure 5.14: Differential expression analysis corresponding to increased degree of response on MRI. (A) GSEA over the Reactome database: proliferation and immune pathways strongly correlated with response. (B) Percentage decrease in tumour bulk on MRI correlated with enrichment of B cells, T cells, as well as NK cells.

5.3.2 Transcriptomic changes associated with therapy

Chemoresistance signatures in post-chemotherapy samples

In order to characterise the post-treatment expression landscape of chemoresistant tumours, expression data obtained at the end of chemotherapy from 41 cases were analysed by means of a differential expression model and a subsequent Reactome gene enrichment (Figure 5.15A). Chemoresistant tumours were more proliferative than chemosensitive tumours and interestingly also had significant enrichment of neural signatures. In a recent study [133], nerve fibres within breast tumours were associated with increased aggressiveness and associated with poor prognosis. Indeed the thickness of infiltrating nerve fibres was correlated with poor differentiation, increased probability of lymph node metastasis and shorter disease-free survival.

Chemoresistant tumours also showed increased oestrogen receptor pathway signalling and significantly less adaptive and innate immune pathway activation. ssGSEA enrichment over the MsigDB C2 curated gene sets revealed that chemoresistant tumours enriched for various experimentally derived gene sets of resistance, including cisplatin, docetaxel, doxorubicin and fluorouracil resistance (Figure 5.15B). It was very evident that the combined expression of these gene sets increased during therapy, showing either acquired resistance during therapy, or the selection and expansion of a pre-existing resistant clone that was present at lower cellular prevalences prior to commencing therapy. The enrichment analysis also showed that the post-therapy tumours had an enrichment of a tamoxifen resistance gene set, potentially also indicating the selection of an endocrine-therapy resistant population. Of note, the selection of a potentially endocrine-resistant clone was also observed in the analysis of case **T005** (Section 4.3.2), where the emergence of a deleterious *ESR1* subclonal mutation was shown to occur during therapy.

Tumours with RCB-III RD post-chemotherapy are considered to be highly chemoresistant. To identify changes in the transcriptome in RCB-III tumours during therapy, a differential expression was performed by comparing matched pre and post-therapy samples from 14 cases with RCB-III RD. This allowed the elucidation of genes and pathways that were differentially expressed in highly chemoresistant tumours after completion of chemotherapy. The differential analysis revealed 411 under expressed and 649 overexpressed genes post-therapy in chemoresistant tumours (Figure 5.16A). The most highly expressed gene was *IL6*: this pro-inflammatory cytokine has been shown to be a central player in linking chronic inflammation to cancer by driving tumour growth and metastasis, as well as increasing

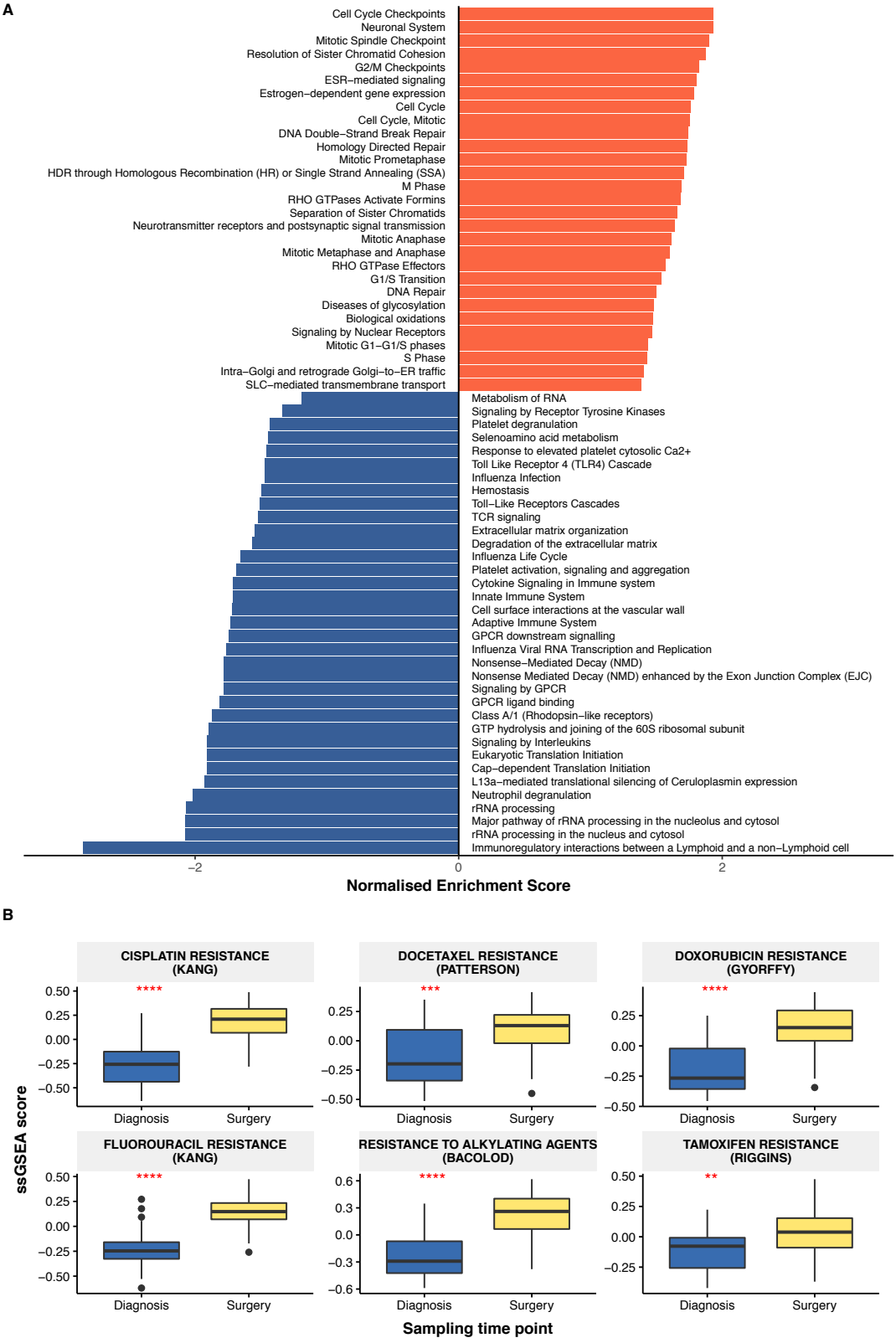


Figure 5.15: Differential expression analysis on all post-therapy samples, indicating features associated with chemoresistance. (A) GSEA over the Reactome database: proliferation and immune pathways strongly correlated with response. (B) Enrichment of resistance signatures in the post-chemotherapy samples, compared to pre-therapy tumours.

angiogenesis and modulating a highly immunosuppressive environment [83]. Additionally, recent work has also shown that IL-6 protects tumours from therapy-induced DNA damage, oxidative stress and apoptosis by facilitating the induction of antioxidant and anti-apoptotic cellular pathways [162].

The angiogenic chemokine *CXCL2* was also highly expressed post-therapy in tumours with RCB-III RD. There is published evidence that CXCL2 mediates lung metastasis and chemoresistance in breast cancer [247], with the administration of CXCR2 blockers shown to break the chemoresistance cycle induced by CXCL2 and increasing the efficacy of chemotherapy in breast cancer [4]. *CTGF*, which was the most statistically significantly highly expressed gene in chemoresistant tumours following chemotherapy, has also been shown to mediate taxane resistance in osteosarcoma when overexpressed through the up-regulation of survivin expression [290]. Additionally, recent work has also shown that *CTGF* expression also mediated resistance to fluorouracil by increasing the expression of *BCL2L1* and survivin and increased MEK/ERK signalling [313]. *CYR61* has also been associated with resistance to gemcitabine [121].

Tubulin binding cofactor C (*TBCC*), a crucial protein for the folding of α and β tubulin, was very under expressed in chemoresistant tumours, in keeping with recent reports that showed that *TBCC* overexpressing tumours displayed increased sensitivity to anti-microtubule agents through decreased microtubule dynamicity and slower passage into mitosis [110].

A driver gene analysis revealed key gene expression alterations that were associated with chemoresistance (Figure 5.16B). Increased expression of *MYC, EGFR, TNFAIP3, PDGFRA, CREBBP* and *ABL1* were all associated with chemoresistance. Decreased expression of the mitotic checkpoint gene *BUB1B*, S phase protein *SKP2* and cellular senescence gene *EZH2* was associated with chemoresistance. Indeed, low expression of *EZH2* has been associated with the induction of cellular senescence [141, 257], showing that these tumours were activating senescent profiles following cytotoxic induced stress, further supporting the observations made earlier in this chapter.

Enrichment over the MSigDB Hallmarks using the `camera` statistical method in `edgeR` showed enrichment over several pathways, including down-regulation of E2F targets, mitosis and oxidative phosphorylation, and increased activation of TNFα signalling, epithelial mesenchymal transition pathways, coagulation, angiogenesis, TGFβ signalling, and the immunosuppressive IL2-STAT5 and IL6-JAK-STAT3 pathways (Figure 5.16C).

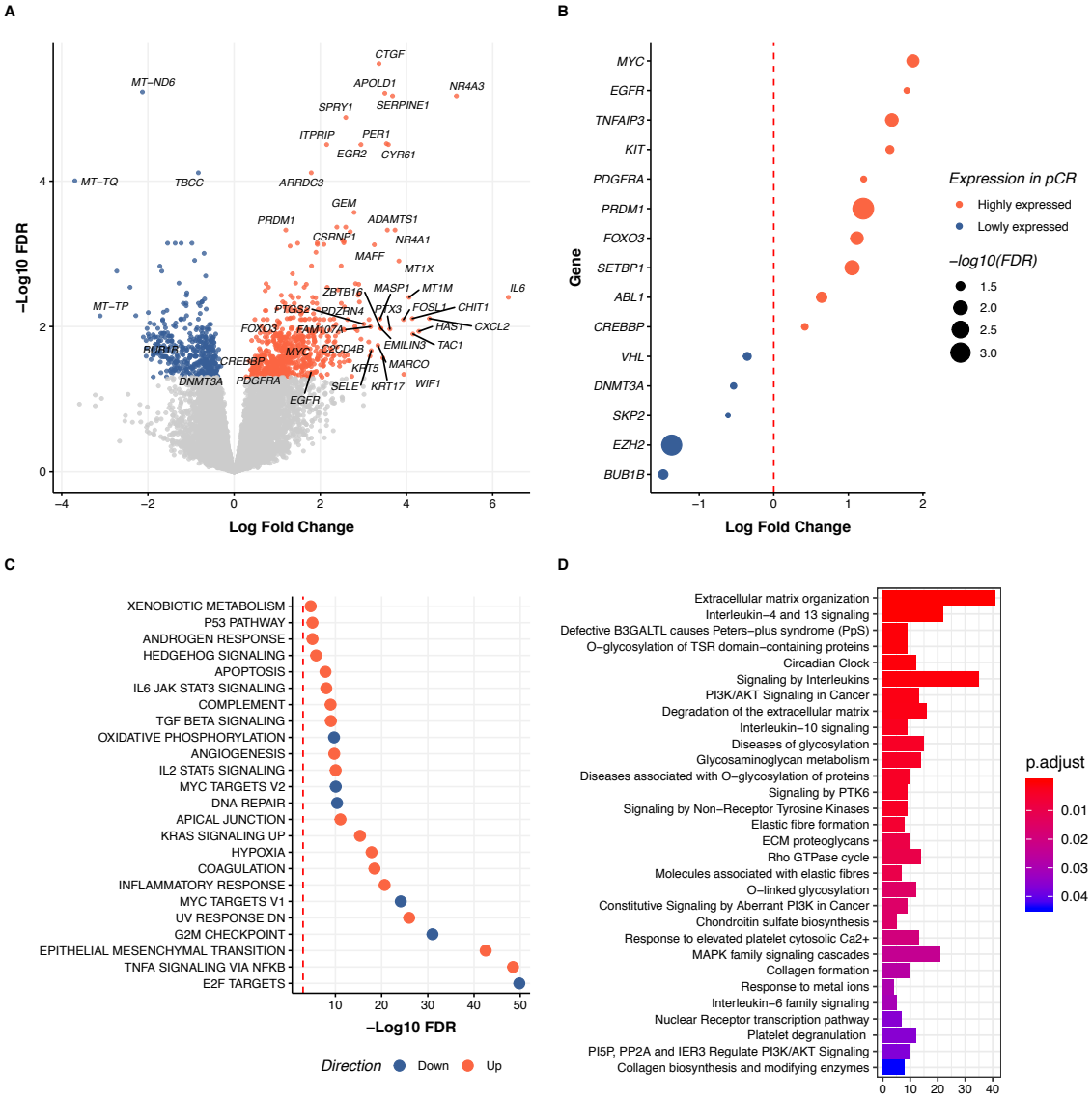


Figure 5.16: Differential expression analysis on all RCB-III post and pre-therapy samples. (A) Volcano plot showing genes that were differentially expressed in chemoresistant samples compared to the same samples prior to commencing therapy. (B) Driver gene expression in chemoresistant tumours. (C,D) Enrichment analysis over Reactome and MSigDB showing pathways differentially activated in resistant tumours.

A pathway enrichment on the Reactome database, focusing on highly expressed genes showed enrichment over various interleukin pathways, as well as strong associations with platelet function (Figure 5.16D).

In summary, chemoresistant tumours showed increased activation of various gene sets associated with resistance to cytotoxic therapies.

Quantifying change in expression profiles

To determine the degree of gene expression and pathway alteration induced by neoadjuvant chemotherapy, ssGSEA was performed on Reactome pathways using 41 matched tumour samples with paired RNA-seq data at diagnosis and on completion of therapy. As shown in Figure 5.17A, the ssGSEA scores of all samples correlated strongly between both sampling time points, with some samples showing a stronger temporal correlation than others. The correlation between ssGSEA scores calculated at both time points was then assessed using Pearson's product-moment correlation (Figure 5.17B): a statistically significant association was observed between pathway correlation and the degree of RD post-chemotherapy (p=0.001, ordinal logistic regression). Tumours that attained pCR showed the greatest degree of change secondary to tumour elimination as well as associated reprogramming of the tumour microenvironment as shown earlier in this chapter. Tumours with much higher degrees of RD, such as RCB-III tumours, had very little change in ssGSEA pathway scores, indicating that the expression landscape remained remarkably stable in these tumours despite therapy with cytotoxic and targeted agents.

Breast tumours with remaining RD post surgery were classified using the intrinsic subtype (PAM50) classification (Figure 5.18). Most tumours were noted to switch classification to the less aggressive Luminal A or normal subtype, with only two tumours retaining their original classification (**T004**, Basal and **T131**, HER2 enriched). This was in keeping with previous observations made by Gonzalez-Angulo et al. [104], where following chemotherapy most tumours were noted to switch to less aggressive subtypes.

In order to assess how the tumour bed following pCR differed from normal breast tissue, post treatment tissues taken from the tumour bed of six tumours that attained pCR were compared to ten breast normal tissues that had been taken at the time of surgery. A differential expression showed profound differences between these two tissues. Normal tissue taken from the tumour bed at the time of pCR showed a significantly increased immune infiltrate, as

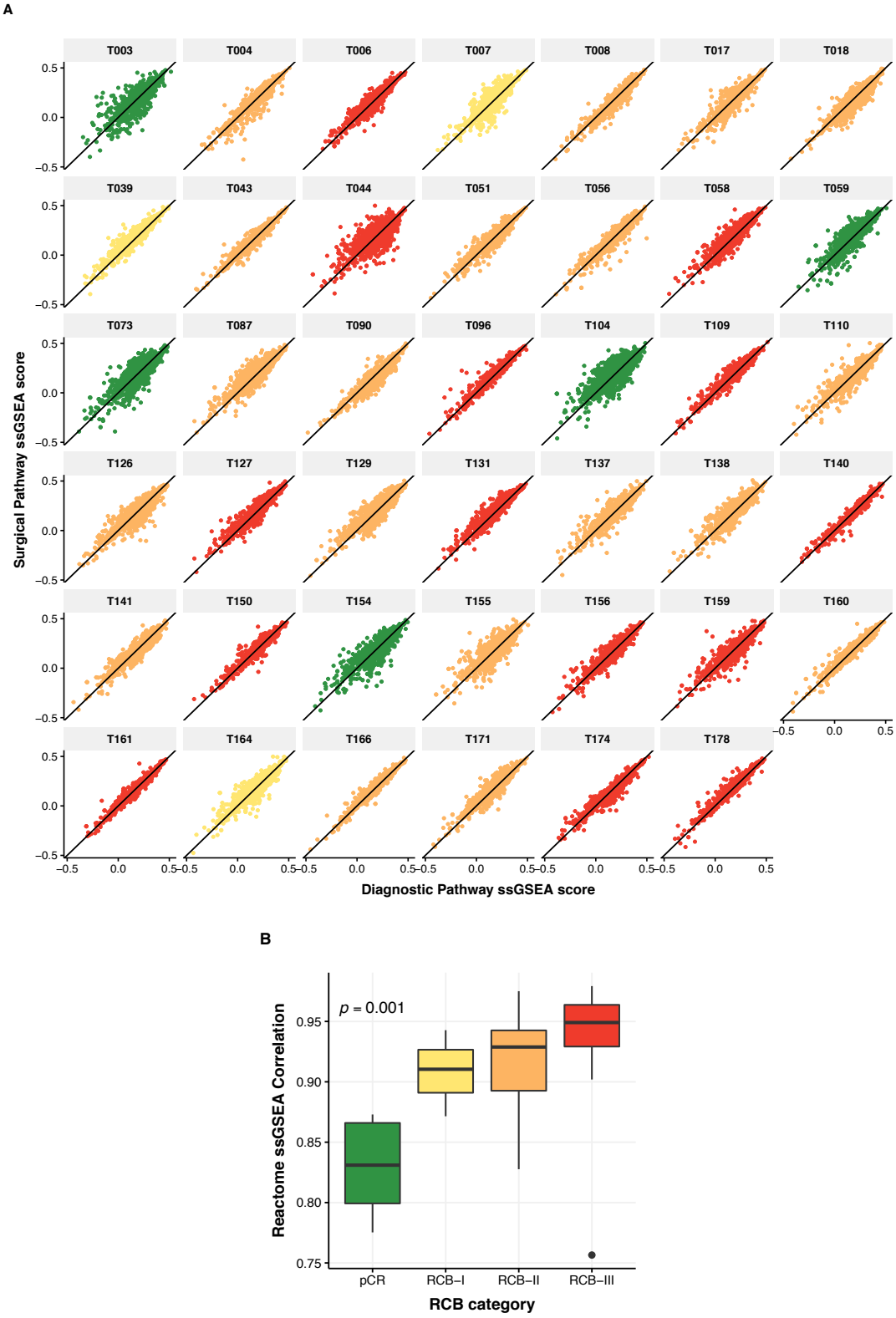


Figure 5.17: Correlations between diagnostic Reactome pathway ssGSEA scores and post-therapy scores. (A) Scatter plot showing degree of dispersion between two sampling time points (B) Box plot showing distribution of pathway correlation between diagnosis and surgical tumour samples, with higher correlations seen as the degree of RD increased.

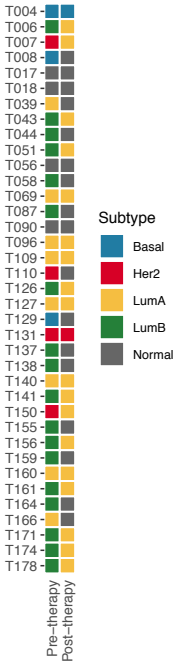


Figure 5.18: Intrinsic subtype classification of tumours pre and post neoadjuvant chemotherapy, showing class change to a less aggressive Luminal A or normal type following neoadjuvant therapy.

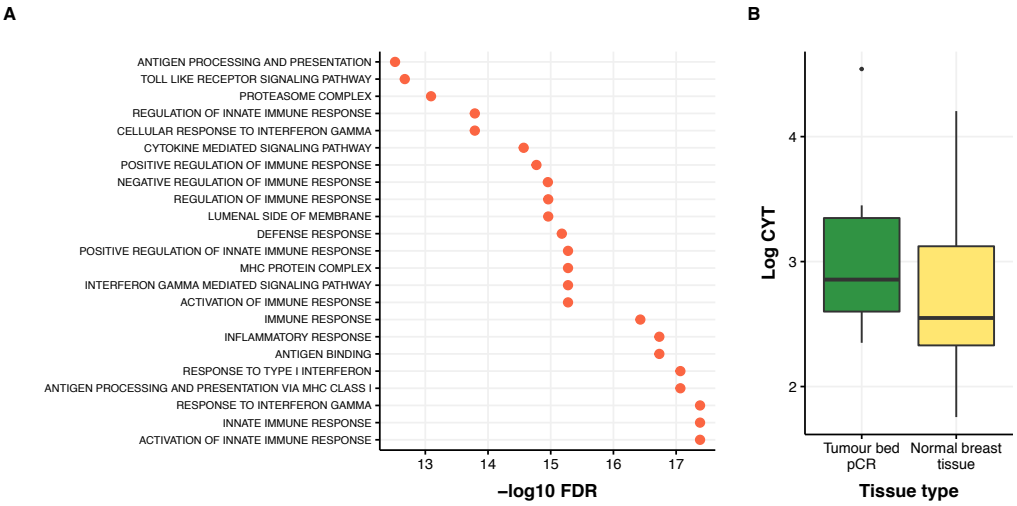


Figure 5.19: Comparison of normal tissue obtained from tumour bed following pCR and unrelated normal breast tissue. (A) Enrichment over MSigDB pathways (B) Comparison of CYT between different cohorts.

shown by a Reactome enrichment and gene ontology analysis (Figure 5.19A). There was no appreciable difference in CYT between the normal pCR tissue and the normal breast tissue, indicating that no further T cell mediated cytolysis was occurring following the complete elimination of tumour cells (Figure 5.19B).

5.3.3 Metastatic pathway reprogramming

There has been published evidence of neoadjuvant chemotherapy inducing pro-metastatic behaviour [116, 151] and the work done in this thesis has hinted that this might occasionally be the case. For example, in Chapter 4 (Table 4.1), it was noted that new mutations enriched within RAC1 cell motility and angiopoietin receptor Tie2-mediated signalling pathways, both of which are involved in metastasis and migration. In this Chapter, it was noted that ERBB2 regulated cell motility pathways increased in chemoresistant tumours, compared to a decrease in chemosensitive tumours (Figure 5.8). Additionally, the angiogenic pro-metastatic chemokine *CXCL2* [247] increased in expression during chemotherapy in tumours with RCB-III RD (Figure 5.16A).

One of the published routes of metastasis centres around the *ENAH* gene, which encodes Mena proteins (member of the enabled/ vasodilator-stimulated phosphoprotein family) involved in regulating the assembly of actin filaments and modulating cell adhesion and motility [151]. Alternate splice variants of this gene have been correlated with tumour invasiveness in breast cancer, with two main Mena splice variants described in the literature [106]:

1. Mena^{11a}, an isoform of Mena that contains an additional 21 amino acids in the EVH2 domain of Mena, which is highly expressed in primary tumour cells, but down-regulated in invasive cells, and has been shown to decrease motility and dampen invasion responses to epidermal growth factor (EGF).
2. Mena^{INV}, an isoform of Mena that has an additional 19-amino acid sequence encoded by the *INV* exon inserted between the EVH1 domain and the LERER repeats, and has been shown to promote invasion, intravasation and metastasis by sensitizing cells to EGF, subsequently allowing them to invade in response to low concentrations of growth factor.

An immunofluorescence-based quantitative method, Mena^{Calc}, defined as the difference between total Mena protein expression and expression of the non-invasive Mena protein isoform (Mena^{11a}) has recently been developed and commercialised. Studies have shown

evidence of a significant positive association between $Mena^{Calc}$ and poor disease-specific survival in breast cancer [86].

Chemotherapy has been shown to mobilise $TIE2^{HI}$ macrophages into the tumour microenvironment, which associate with newly constructed blood vessels and cause increased $Mena^{INV}$ isoform expression in tumour cells by NOTCH mediated signalling, increasing their metastatic potential [116, 151]. Indeed, these macrophages, together with associated tumour cells and endothelial cells have been shown to form anatomical structures called tumour microenvironment of metastasis (TMEM), at which breast cancer intravasation into blood vessels occurs. These TMEM sites have been shown to release circulating tumour cells (CTC) faster during chemotherapy, with a recent paper suggesting that neo-adjuvant chemotherapy can lead to increased CTC dissemination as well as increased metastatic burden [152].

In order to explore this hypothesis further, the expression of $Mena^{INV}$ and $Mena^{11a}$ was measured in all samples within the dataset. De novo transcript assembly and quantification was performed using `Salmon` as described in Chapter 2. The expression of $Mena^{11a}$ was defined as the expression of transcript ENST00000366844, the expression of $Mena^{INV}$ was defined as the expression of transcript ENST00000284563 and the total Mena expression defined as the sum of *ENAH* transcripts ENST00000366844, ENST00000284563 and ENST00000366843. Any non-coding transcripts were not included in this computation. Only high purity tumours were retained for this analysis.

Tumours with higher $Mena^{INV}$ and $Mena^{calc}$ expression at diagnosis were more likely to attain pCR during treatment (Figure 5.20A). As the expression of $Mena^{INV}$ was assumed to correlate with invasive potential and therefore aggression this observation was in keeping with previous observations made in this work, where neoadjuvant chemotherapy was shown to induce better responses in more aggressive tumours. This analysis, therefore, showed that tumours that had higher metastatic potential were more likely to be eliminated by chemotherapy.

However, as shown in Figure 5.20B, during chemotherapy, Mena expression correlated with the degree of RD at surgery. While tumours that attained pCR had very low levels of $Mena^{INV}$ detectable, tumours with RCB III RD had an increase in $Mena^{INV}$ expression during treatment, potentially providing further evidence to the findings published by Karagiannis et al. [152], where chemotherapy was shown to augment metastatic potential of tumour cells that had not responded to chemotherapy. While similar findings were seen in this work, further larger studies will be needed for validation.

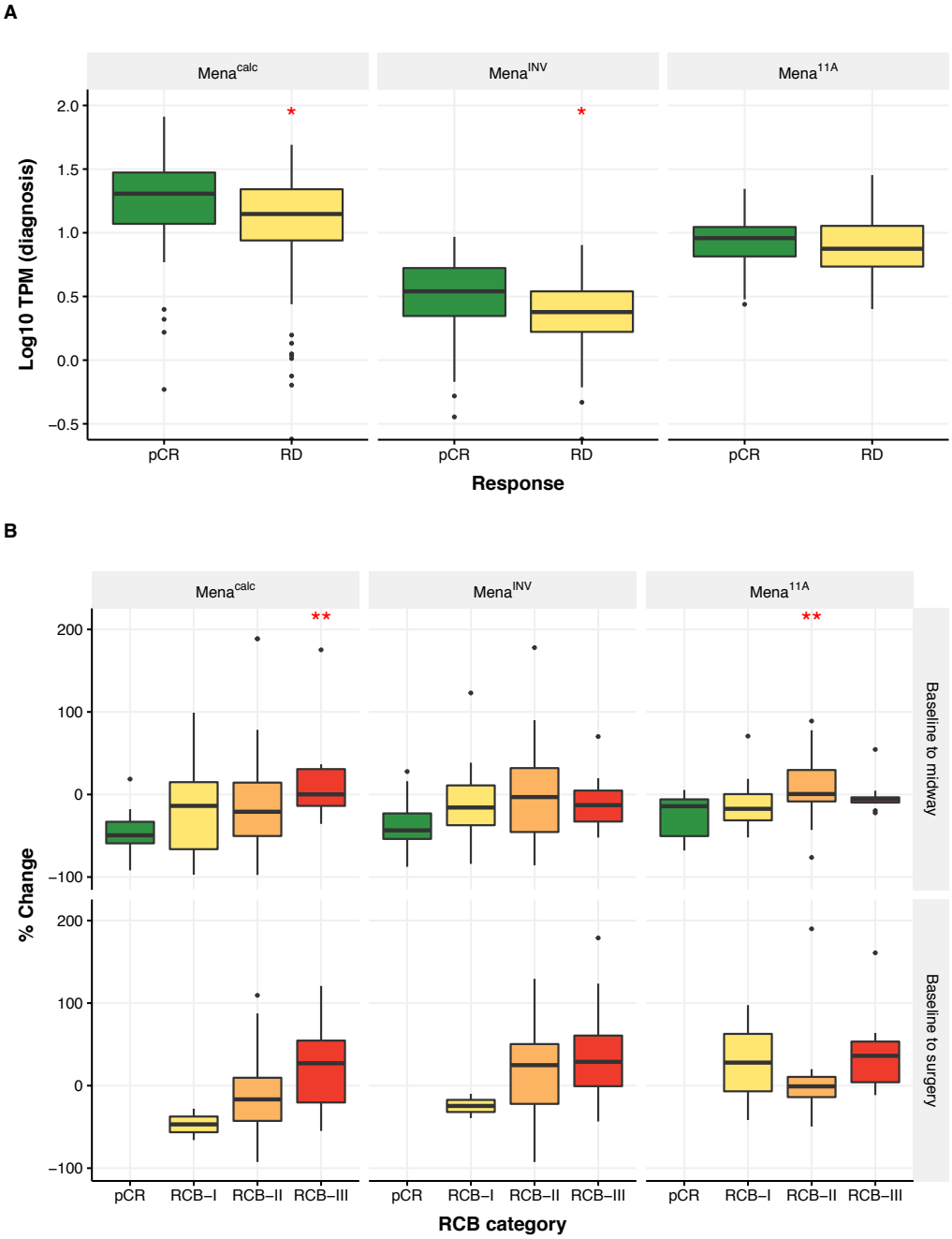


Figure 5.20: Mena dynamics during chemotherapy. (A) Tumours with higher Mena expression at diagnosis were more likely to attain pCR. (B) Change in Mena expression during treatment. Tumours with higher degrees of residual disease on completion of chemotherapy had increased expression of Mena during chemotherapy.

5.4 Discussion

The work done in this chapter has described how the expression landscape of both the tumour and its microenvironment were altered during neoadjuvant chemotherapy.

Shortly after commencing chemotherapy, transcriptomic changes were set in motion that correlated tightly with eventual response to treatment. Tumours that attained pCR by the end of therapy showed great differences in their transcriptomic landscape at the midway and post-therapy time points, compared to chemoresistant tumours which continued to have a remarkably stable expression profile despite cytotoxic therapies.

By integrating changes observed on serial MRI scans with transcriptomic data it was shown that the reduction in tumour bulk correlated with the degree of immune cell infiltrate as well as tumour proliferation. The greater the reduction in tumour volume, the higher the degree of both adaptive and innate immune activation and the greater the decrease in tumour proliferation, in keeping with increased cell death in tumours that were responding to therapy. A similar pattern was observed when a differential expression analysis was performed to see which changes in the midway biopsy associated with final response to therapy. Tumours that attained better degrees of response by the end of therapy had higher immune activation and lower proliferation at the midway time point, compared to chemoresistant tumours.

Clearly the immune system continued to play a key role throughout treatment. Indeed, the correlation between the degree of adaptive and innate immune pathway activation and eventual RCB score was monotonic at the midway time point, with a stepwise decrement in immune activation as the degree of RD post-chemotherapy increased. Additionally, a clear association was observed between the cytolytic activity score (CYT) and response, with the correlation observed being much stronger than that seen in the pre-therapy samples.

Interestingly, taxane-first treated tumours, which were mostly $HER2^-$, exhibited a significant increase in CYT midway during chemotherapy if pCR occurred by the end of therapy, with no statistically significant increase if RD was present at the time of surgery. An inverse relationship was observed in anthracycline-treated tumours, which were often $HER2^+$ (Figure 5.9B), where an increase in CYT was associated with RD and an initial decrease in CYT associated with pCR. There is a body of evidence that suggests superiority of neoadjuvant treatments with a taxane prior to an anthracycline, and perhaps this might shed light into why this is so. $HER2^-$ tumours that attained pCR had an increase in B cell, CD8 T cell, neutrophil

and NK cell enrichment during therapy; this was not observed in the anthracycline-treated $HER2^+$ tumours.

The presence of HLA LOH in the taxane treated $HER2^-$ tumours correlated with CYT dynamics. Tumours with LOH in at least one MHC class I gene had a decrease in CYT during therapy, compared to tumours without LOH which had an increase in CYT. These findings were corroborated on enriching for T cell signatures in the RNA-seq data.

By deconvoluting the TCR and BCR CDR3 clonotype sequences from the bulk RNA-seq data, various interesting observations were made. Tumours that had CDR3 clonal dominance at the diagnostic and midway sampling time points, as computed using the Gini coefficient, were more likely to attain pCR than those with a greater variety of clonotypes in the tumour microenvironment. The Gini coefficient was also shown to positively correlate with CYT, indicating that clonotype dominance was associated with increased T cell-mediated cytolysis.

The immune system was also shown to play a key role in fashioning the subclonal response to therapy and played a profound role in determining the evolutionary trajectory of a tumour. Tumours with evidence of ongoing clonal extinction had significantly higher enrichment of immune pathways whilst those with clonal emergence had a distinct paucity of enrichment over high level innate and adaptive immunity pathways, with less TCR, BCR and interleukin signalling. Analysis of CYT across the different response groups showed that tumours with clonal emergence had a statistically significant lesser CYT score at the midway time point compared to tumours that exhibited clonal extinction only, with findings corroborated by an enrichment over the LM22 immune gene set.

Chemoresistant tumours at the end of therapy showed an increased expression of various chemotherapy resistance gene sets, including cisplatin, docetaxel, doxorubicin and fluorouracil resistance metagenes. This could have been either due to acquired resistance during therapy, or the selection and expansion of a pre-existing resistant clone that was present at lower cellular prevalences prior to commencing therapy. The enrichment analysis also showed that the post-therapy tumours also had an enrichment of a tamoxifen resistance gene set, potentially also indicating the selection of an endocrine-therapy resistant subclone. Additionally, chemoresistant tumours had significant enrichment of neural signatures, as well as increased activation of senescent profiles, probably showing evidence of therapy-induced senescence.

Various genes associated with chemoresistance were identified, including the up-regulation of the pro-inflammatory cytokine *IL6*, which has been shown to be a central player in linking

chronic inflammation to cancer by driving tumour growth and metastasis, the angiogenic chemokine *CXCL2* which has been shown to mediate lung metastasis and chemoresistance in breast cancer and *CTGF*, which has been shown to mediate taxane and fluorouracil resistance in other malignancies.

Finally, in view of published observations that chemotherapy might increase the metastatic potential of tumour cells that do not respond to therapy, the expression dynamics of the pro-metastatic transcript Mena^{INV} were elucidated. Tumours that had higher metastatic potential were more likely to be eliminated by chemotherapy, however tumours with RCB III RD had an increase in Mena^{INV} expression during treatment, potentially providing further evidence to the findings published by Karagiannis et al. [152], where chemotherapy was shown to augment the metastatic potential of tumour cells.

Chapter 6

The circulating tumour genomic landscape

Contents

6.1 Introduction

The presence of circulating cell-free DNA (cfDNA) in plasma was first described more than 70 years ago [185] and is a physiological occurrence following apoptosis or necrosis of cells [105, 144]. In healthy individuals, the concentration of cfDNA is generally low (mean 13ng/ml), however the concentration increases dramatically in malignancy (mean 180ng/ml) [105], with the overall concentration correlating very closely with tumour burden and response to treatment in patients with metastatic breast cancer [64, 138].

cfDNA is highly fragmented and is typically found in fragments of 160-180 base pairs [203], reflecting degradation of the nucleic strands into nucleosomal units during apoptosis. The short fragment length, in the context of a high background level of normal DNA, renders the detection of tumour derived cfDNA (referred to as circulating tumour DNA, or ctDNA) challenging [30, 64]. Two main approaches of detecting ctDNA have been explored: the first approach, as adopted by digital PCR [301], involves identifying single mutations in the plasma with high sensitivity and specificity. This requires a priori knowledge of the mutational landscape of a tumour and is often not conducive to high throughput studies requiring the assaying of multiple mutations. In the second approach, direct plasma sequencing is performed over either targeted small panels [64, 87, 92] or over much larger panels (including whole exome sequencing) [158, 205]. Shallow whole genome sequencing has also been shown to be a reliable way of detecting ctDNA, though this often requires higher ctDNA fractions, and has shown to perform best in the presence of high tumour burdens [118].

Most ctDNA studies have been performed in the metastatic setting, where ctDNA concentrations correlated closely with the burden of disease [33, 64, 87, 118, 205], with very few studies performed in the adjuvant or neo-adjuvant setting. This was often secondary to technological limitations of detecting mutant allele signatures at very low frequencies and bioinformatic limitations of distinguishing rare mutations from technical artefacts [30]. Indeed, all the studies that have attempted to detect ctDNA during therapy either had very small numbers of recruited patients or failed to consistently detect mutations in the plasma [53, 156]. In the post neoadjuvant setting, the detection of ctDNA in plasma after completion of therapy predicted metastatic relapse with high accuracy [95].

The work described in this chapter aimed to determine whether ctDNA could be reliably detected in serial plasma samples throughout neoadjuvant therapy by using deep sequencing

as well as shallow whole genome sequencing, using recently published methods optimised for ultra-low ctDNA detection [5].

6.2 Somatic variant detection and tracking using deep sequencing

To investigate whether somatic mutations detected on exome sequencing could be reliably detected within the plasma before and during therapy, the Raindance deep sequencing primer panel discussed in Chapter 4 was used to amplify cfDNA present in sequential samples obtained from patients **T003**, **T004**, **T025** and **T029**. Cases T003, T004 and T029 had plasma samples available pre-therapy, before the commencement of cycle 2, before the commencement of cycle 4 (i.e. midway through therapy), and following completion of chemotherapy. Case T025 did not have a diagnostic plasma sample available. Library preparation was performed as discussed in Chapter 2, with an initial single molecule PCR partitioned in oil droplets, followed by sample barcoding and ligation of sequencing adapters. All 15 samples were pooled into one library and sequenced on one lane of an Illumina MiSeq sequencer.

A total of 156 mutations were targeted across all four cases, with amplification performed in 15 plasma samples. As can be seen in Figure 6.1A, over half of the regions sequenced attained a coverage of over 3,500× in most samples. 9 amplicons had a median coverage of less than 100× across all samples and were removed from the analysis. The median amplicon coverage across all samples was 4,093× (Figure 6.1B), with most samples attaining adequate amplification. The sequencing data obtained from case T004's second plasma sample had a lower median coverage, though somatic mutations were still detectable at this time point, as discussed later on.

Mutation detection was performed using the bioinformatic pipeline discussed in Chapter 4. As mutations were mutually exclusive to each case, all plasma samples that were not acquired from one case served as controls. A mutation was deemed to be present if the variant allele fraction (VAF) of the mutation in a case was three standard deviations above that of controls.

Figure 6.2 illustrates the results of this sequencing experiment. For each of the four cases, heat maps showing mutation detection in both the tumour exome sequencing data and plasma deep sequencing data are depicted, with mutations detected at AFs above 1% shown in red, rare alleles detected robustly at AFs less than 1% shown in light blue and absent mutations shown in darker blue. While mutations were readily detectable in the tumour exome data, mutations within the plasma were less frequent, and present at very low AFs, in keeping with findings from other studies [53, 156].

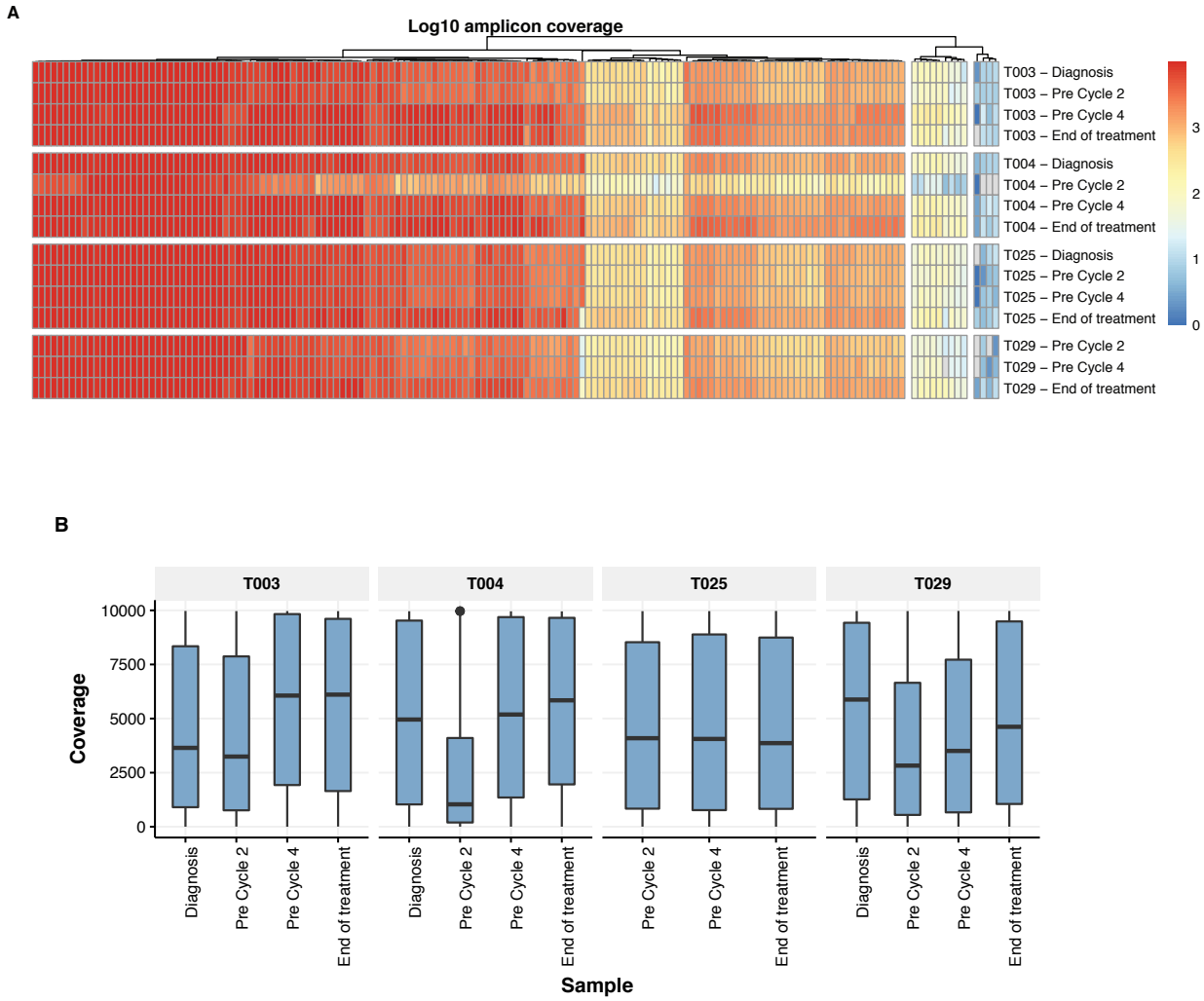


Figure 6.1: ctDNA targeted amplification metrics. (A) Distribution of coverage across all 156 amplicons sequenced. (B) Distribution of coverage across all 15 plasma samples sequenced.

ctDNA was detected at diagnosis in all four plasma samples analysed. At least one mutation derived from the founding mutational cluster was present at diagnosis, though driver gene mutations were not always detectable within the plasma (such as, for example, *PIK3CA* in T025). Additionally, ctDNA was also detectable in all three cases with RD post-chemotherapy (T004, T004 and T029), indicating the usefulness of using ctDNA to monitor residual disease. It is worth noting that as only a few mutations were detected at all time points, the usefulness of ctDNA in this setting would require targeting multiple, rather than single, mutations. The greater the number of mutations targeted, the greater the probability of detecting a tumour derived DNA molecule.

The analysis also showed that some mutations that were detected at later time points on tumour exome sequencing were actually present in the plasma at earlier time points. For example, in case T004, mutations within *CELA3B* and *CREBBP* were only detectable in the tumour after completion of chemotherapy, however there was evidence of these mutations in the plasma at diagnosis, albeit at very low allelic fractions, once again confirming the suspicions that these mutations were not generated by chemotherapy, but were already present at diagnosis.

In order to define a metric associated with total plasma disease burden, the allelic fractions of all detectable mutations at each time point were added together and used as a surrogate of total mutation load within the plasma (Figure 6.3). Cases T004 and T029 had a decrease in ctDNA burden during treatment. In case T004 a high tumour fraction was initially detectable in the plasma, however this rapidly decreased following one cycle of FEC chemotherapy. The levels of ctDNA then remained low throughout therapy, with only one mutation detectable in the plasma on completion of chemotherapy. A significant burden of disease (RCB-III) was present at the time of surgery despite the low amount of ctDNA detected. This might indicate that chemotherapy eliminated the more proliferative tumour cells and the tumour bulk present following completion of therapy was less metabolically active and released less DNA into the bloodstream. Case T009 showed a very gradual but consistent decrease in plasma ctDNA fractions during therapy, with the decrease in plasma tumour burden correlating to a corresponding tumour volume decrease seen on serial imaging.

Case T003 had 7 mutations detectable at diagnosis with a combined AF of more than 4%. Following one cycle of FEC chemotherapy ctDNA was no longer detectable, however by the midway time point an increase in tumour DNA was observed, with an associated minimal decrease in size on ultrasound, suggesting that the tumour was resistant to FEC chemotherapy.

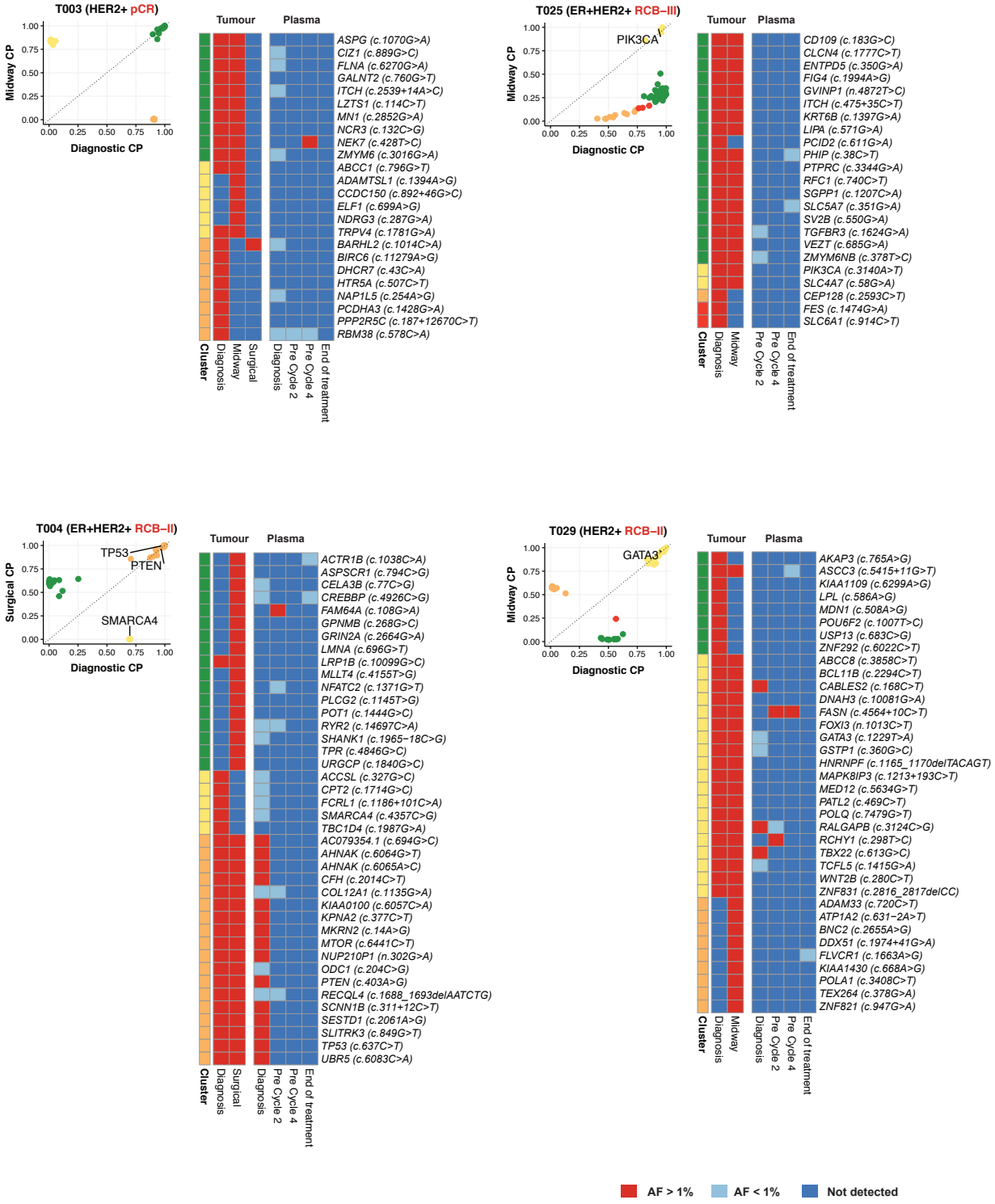


Figure 6.2: Deep sequencing of serial plasma samples from four cases, with corresponding mutations identified on tumour exome sequencing. Scatter plots show clonal structure derived from exome sequencing, plasma heatmaps show deep sequencing results.

However following a therapy backbone switch to docetaxel and trastuzumab, pCR occurred with no evidence of any mutations detected within the plasma.

Case T025 did not have a plasma sample taken at diagnosis and therefore the mutation dynamics prior to commencing chemotherapy could not be elucidated. However, following one cycle of treatment two mutations were robustly detected at allelic fractions below 1%. No ctDNA was detectable by the midway time point, though by the end of chemotherapy the total ctDNA fraction was noted to be higher than that seen prior to cycle 2 therapy (Figure 6.3). Indeed this tumour had RCB-III disease at the end of chemotherapy.

In summary, these analyses provide us with evidence to suggest that ctDNA can be robustly detected in early-stage breast cancer, with evidence of mutations present at diagnosis in all cases analysed as well as after one cycle of chemotherapy. Truncal mutations were more likely to be detected in the plasma. By using deep sequencing, the response to chemotherapy could also be monitored. Finally, this approach was also able to detect mutations in the plasma in all three cases with RD post-chemotherapy.

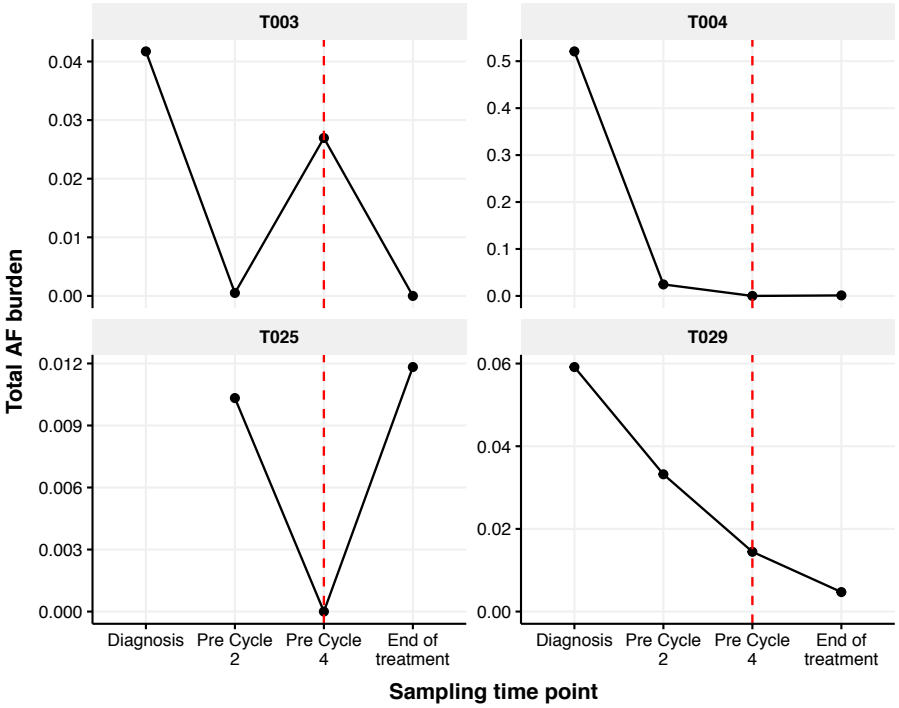


Figure 6.3: Total ctDNA burden detected in serial plasma samples across four patients (T003: pCR, T004 and T029: RCB-II, T025: RCB-III). Dotted red line indicates therapy backbone switch.

6.3 Estimating ctDNA fraction through CNA analysis

The mutational analysis described previously provided incontrovertible evidence that mutations could be detected in the plasma of patients with early-stage breast cancer prior to commencing therapy, as well as during neoadjuvant treatment. In order to determine whether copy number alterations (CNAs) could also be detected, whole genome libraries were constructed from 96 plasma samples obtained from 24 patients. All patients had plasma sampling performed pre-therapy, prior to commencement of cycle 2, prior to cycle 4 and prior to surgical intervention. The libraries were sequenced over two lanes of an Illumina HiSeq4000 in paired-end mode, resulting in a median depth of around 30 million reads per sample ($1\times$ coverage).

Copy number binning and segmentation using `QDNAseq`, as described in Chapter 2, showed evidence of detectable CNAs within the plasma in some cases, corresponding to a high ctDNA burden. One such example, case T004, had evidence of a high degree of copy number alterations detected within the plasma (Figure 6.4) at the diagnostic and second sampling time points, with no detectable alterations in the midway and surgical sampling time points. The mutational deep sequencing analysis performed on this sample had also shown evidence of a high mutation burden in the plasma prior to commencing therapy (Figure 6.3), corresponding to the results obtained in this analysis.

The previous mutational analysis showed that at most time points the tumour variant AF in the plasma was less than 10%, with increasingly lower AFs observed as therapy progressed. As the plasma tumour DNA fraction was already quite low at diagnosis and decreased even further through therapy, standard algorithms that estimated tumour purity from copy number data could not be used, as these often required tumour purities well above 25% to be able to quantify purity confidently [296]. Indeed tools such as `ABSOLUTE` [49] failed to detect any tumour content secondary to the very low tumour purity.

The `ichorCNA` R package [5] was specifically designed to estimate the fraction of ctDNA from ultra-low-pass whole genome sequencing and was tuned to detect ctDNA levels above 3%. Indeed, in benchmark experiments, `ichorCNA` had a lower limit of sensitivity for detecting the presence of tumour DNA of 3%, with a 91% specificity and 95% sensitivity. To determine the tumour fraction present in the sequencing data, `ichorCNA` was run on one megabase segmented data obtained from `QDNAseq` using the recommended settings for ultra-low detection, which served to reduce the complexity of the statistical model while increasing sensitivity. To account for technical noise and artefacts induced by library preparation, a

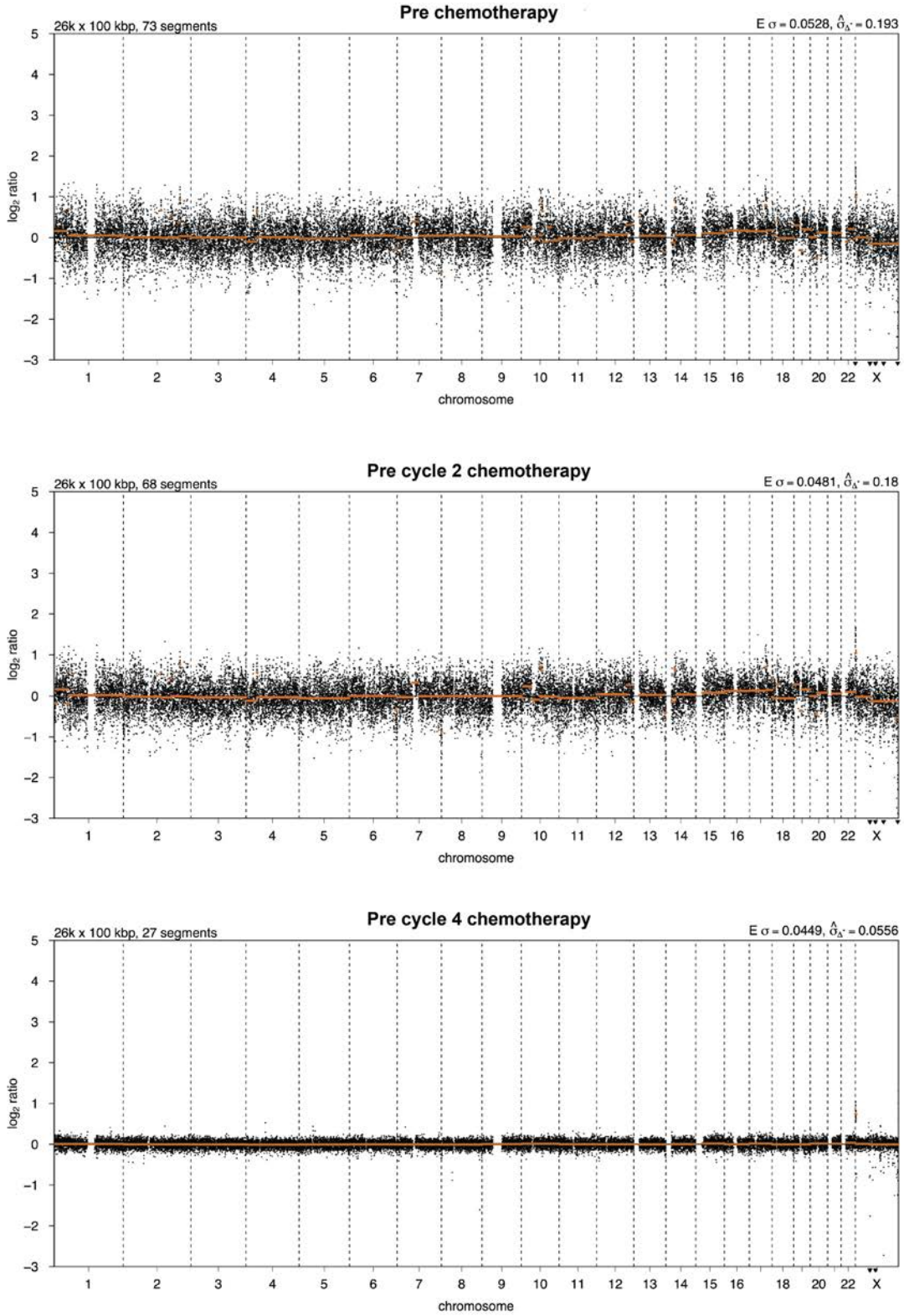


Figure 6.4: Sequential copy number profiles for case T004. Initial copy number changes observed prior to commencing therapy were no longer evident following three cycles of chemotherapy. Coverage: Pre chemotherapy: 20.3M reads, Pre cycle 2: 24.4M reads, Pre cycle 4: 35M reads.

panel of five normal pooled plasma samples that were processed and sequenced using the same technology was provided to the algorithm, thereby ensuring that no copy number alterations were called in plasma samples which did not have any evidence of tumour DNA.

Figure 6.5 shows the results of the `ichorCNA` analysis. Of the 24 cases sequenced, 19 had detectable ctDNA at diagnosis (defined as a tumour fraction higher than 3%). In keeping with the observations made with the mutational analysis, this confirmed that ctDNA was detectable in early-stage disease and that a combination of copy number analysis, as well as mutation detection, could help augment sensitivity.

Of the six cases that attained pCR, only two had undetectable ctDNA following completion of therapy. Case T003 showed an initial decrease in ctDNA after one cycle of chemotherapy, followed by an increase in ctDNA fraction midway through therapy and no detectable ctDNA on completion of treatment. These changes mirrored the pattern observed in the mutational analysis (Figure 6.3), confirming a high correlation between the dynamics observed in both the mutational and CNA analysis. Four cases that attained pCR had evidence of ctDNA in the post-therapy samples, and in one case (T067) the total tumour DNA content was above 10%.

The ctDNA analysis showed evidence of varying dynamics during neoadjuvant chemotherapy. In some cases, such as case T011 (Figure 6.3), the plasma tumour burden was shown to increase steadily throughout therapy. In other cases, an initial decrease in ctDNA levels was succeeded by a rapid rise: this often occurred following a chemotherapy backbone switch (cases: T005, T041, T053, T060, T065), potentially indicating resistance to therapy. A reversed observation was made in cases T052 and T064, where an increase in ctDNA levels were detectable prior to a chemotherapy backbone switch, with a decrease following agent alteration. While these observations are purely descriptive, they highlight the great diversity of dynamics detected during therapy.

50% (3/6) of all cases with RCB-I RD, 88% (7/8) of all cases with RCB-II RD and all (4/4) cases with RCB-III RD had evidence of ctDNA present in the plasma at the time of surgery. There was no relationship observed between the baseline ctDNA level and eventual response. Additionally, no relationship was observed between the change in ctDNA level after 3 and 9 weeks of therapy and response.

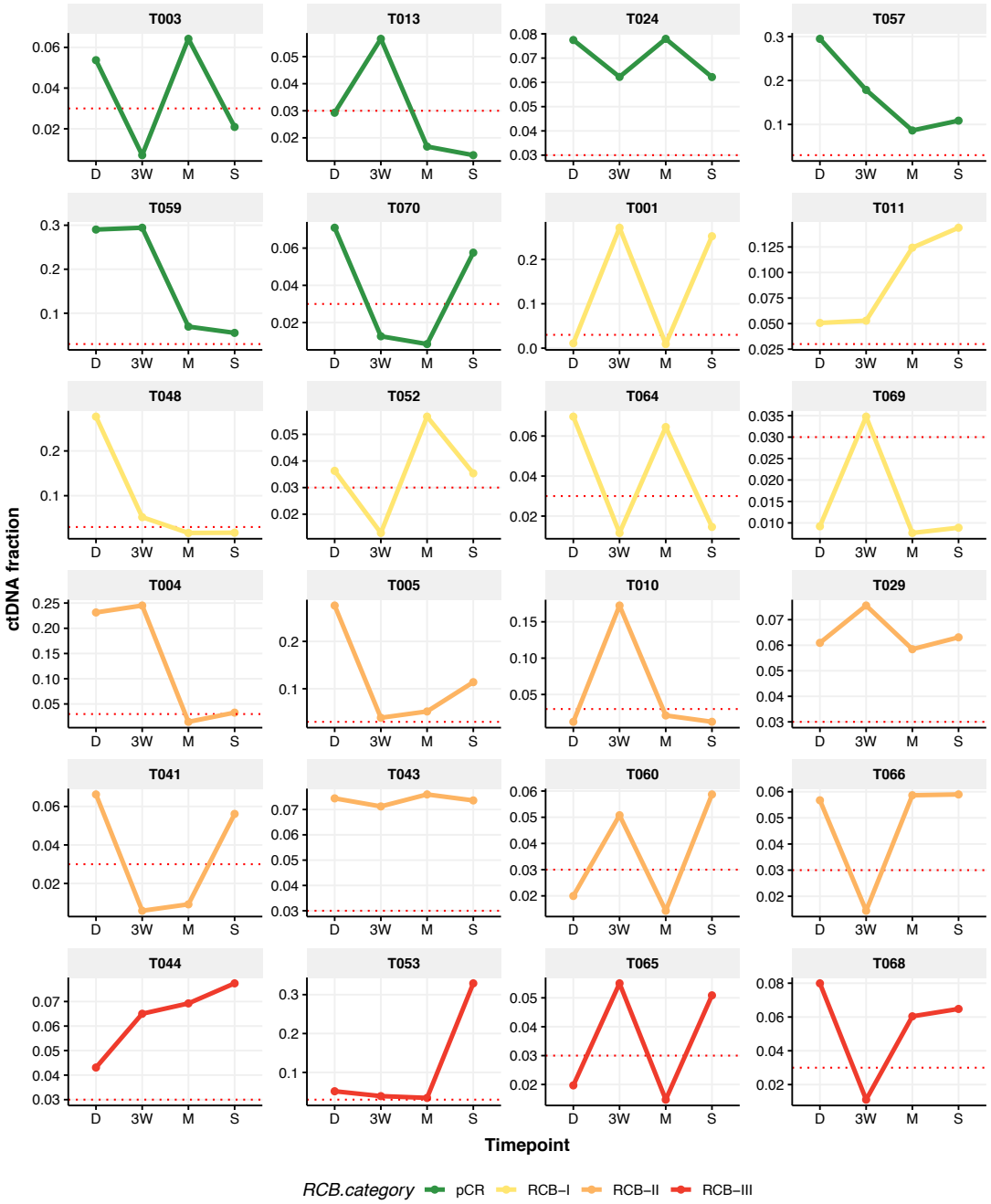


Figure 6.5: Fraction of ctDNA in serial plasma samples (Diagnosis (D), after 3 weeks (3W), Midway (M) and post therapy (S)) as estimated by `ichorCNA`. Dotted red line indicates threshold above which ctDNA fraction can be confidently detected (0.03).

6.4 Conclusion

The descriptive work done in this chapter will serve as a foundation for further ctDNA work in the neoadjuvant setting. It was reassuring that ctDNA could be used to detect disease in early-stage breast cancer (100% of samples using deep sequencing, 79% using shallow whole genome sequencing) and these findings will motivate the eventual comprehensive analysis of all the plasma samples acquired within the TransNEO study.

Targeted deep sequencing has been shown to enable the detection of ultra-rare variants. As described in Chapter 4, by using a comprehensive background noise model, mutations with allelic fractions below 1% could be robustly identified using this library preparation method. Because of the very low ctDNA mutational AF in the early breast cancer setting, a large number of mutations needed to be targeted to detect and track tumour burden. This is very unlike the metastatic setting, where ctDNA detection is possible using relatively fewer mutations because of the higher degree of disease burden. Indeed, as shown in this work, most of the mutations targeted were not detectable in the plasma, and therefore the greater the number of mutations targeted, the greater the probability of ctDNA detection.

Mutations as well as copy number alterations were detected throughout therapy and used as a surrogate of plasma disease burden. Various patterns of ctDNA dynamics were observed. In some cases (T004, T048) a gradual decrease in ctDNA burden was observed throughout therapy and assumed to be secondary to the elimination of a tumour sub-population. In other cases, there was evidence of an increase in ctDNA during the first half of the chemotherapy backbone, followed by a profound drop in ctDNA levels (eg T052, T064), perhaps showing selective chemosensitivity to the second set of agents administered. There were also cases with very low ctDNA levels during the first half of therapy, followed by increased levels as soon as the therapy backbone was altered (eg T060, T070). In a few cases, ctDNA levels remained high throughout therapy (T011, T043, T044).

It was interesting to note that ctDNA was also detectable following chemotherapy in cases with evidence of pCR. This could have been due to undetected micrometastatic disease elsewhere. While statements regarding survival and relapse cannot be made in view of the small sample size analysed, further analysis on the whole cohort, as well as more mature follow-up data, will show whether the detection of ctDNA following chemotherapy in this group that received apparently-curative treatment has any ramifications on relapse and survival.

Chapter 7

Summary and Perspective

Overview

This thesis has described the molecular profiling of a dataset comprising 168 patients with early breast cancer treated with neoadjuvant chemotherapy, with analyses performed on serial tumour samples obtained at diagnosis, midway through chemotherapy, and on completion of neoadjuvant therapy. A total of 300 tumours were available for analysis.

Through the analysis of serial tumour high depth exome and RNA sequencing, and integration with MRI and digital pathology data, this work has aimed to describe molecular predictors of response to chemotherapy and chart the changes induced in a tumour and its microenvironment throughout neoadjuvant treatment.

Molecular predictors of response to chemotherapy

The ability to accurately predict response to chemotherapy remains one of the holy grails of cancer research. If we were able to confidently stratify patients into groups that would derive the most and least benefit, therapy would be directed more appropriately and patients who would derive the least benefit would be spared systemic therapy and its associated morbidities.

Currently, very few molecular features are utilised in the clinic to guide the use of neoadjuvant therapies. A review of the current NICE breast cancer treatment guidelines confirms this [210]: with the exception of ER status, no other molecular determinant is routinely used to assess suitability. Clinical features (including tumour size) and patient preference are often the main determinants used to guide neoadjuvant therapy usage, with decisions often agnostic to molecular features.

The work done in this thesis aimed to add to our understanding of characteristics associated with response to neoadjuvant therapy. Associations with published phenotypes correlating with response (ER, HER2, histological grade and age at diagnosis) were observed in this study, indicating that the patient population recruited was similar to that observed in clinical practice.

The first striking observation was a strong association between tumour mutation burden (TMB) and the degree of residual disease (RD) post-chemotherapy. Tumours with high mutation burdens were more likely to attain a better response to therapy, with the probability of attaining pCR highest above 3 mutations Mb^{-1}. While this could have been secondary to a

moderate association between TMB and proliferation, it is likely that a stronger correlation between TMB and neoantigen burden contributed to this observation. Indeed, the expressed neoantigen burden was strongly associated with response: tumours with higher neoantigen burdens were more likely to elicit stronger immune responses. Intriguingly, tumours that expressed neoantigens presented by more than one HLA molecule were more likely to be chemosensitive.

Tumour heterogeneity played a key role in mediating resistance to therapy: unsurprisingly, tumours with a higher degree of heterogeneity at diagnosis were more likely to harbour resistant phenotypes, in keeping with observations made in the literature in other tumour sites. Tumour populations harbouring *TP53* mutations were associated with a more aggressive phenotype and increased chemosensitivity, whilst somatic mutations within *PIK3CA* and *GATA3*, which were often found in low proliferation ER^+ tumours, were associated with extensive RD post-therapy. In addition, oncogene amplification was likely to increase proliferation and was associated with chemosensitivity while deletions over tumour suppressor genes and apoptotic machinery conferred chemoresistance.

This study has comprehensively highlighted the role the genomic immune landscape played in fashioning response to therapy in breast cancer and has resulted in several novel discoveries. Both the expressed neoantigen load and HLA subtype correlated with response to chemotherapy. The presence of HLA-A*23:01 and HLA-B*38:01 were associated with increased rates of pCR, whilst the presence of HLA-C*07:18 and HLA-B*18:01 were associated with extensive RD and chemoresistance: interestingly, the latter HLA molecule has also been associated with poor prognosis in HIV infection, indicating a possible mechanism underlying immune escape. These novel observations will motivate further investigation into the role HLA subtypes play in determining response to therapy and relapse.

Another novel finding was the relationship between HLA LOH and response to therapy. Tumours with LOH of HLA class I genes were more likely to be chemoresistant: 56% of neoantigens in 29 cases that harboured HLA LOH were no longer presentable, with 73% of all LOH events resulting in the preferential loss of an HLA molecule that presented an equal or greater number of neoantigens than the retained molecule. It was evident that, by employing a further method of immuno-evasion, tumour cells were less likely to be recognised as 'non-self' by the immune system. This previously unappreciated role of HLA LOH has resulted in a collaborative effort to comprehensively characterise HLA dynamics during therapy by performing HLA class-I and II immunohistochemical staining on pre and post-therapy FFPE breast tumour and metastatic lymph node tissues. This will help ascertain

the role of HLA heterogeneity and dynamics in allowing immuno-evasion in both the primary breast tumour and the involved lymph nodes. However it is worth noting that this LOH analysis was not performed in the context of whole genome LOH normalisation and therefore more work will need to be done to determine whether this observed relationship was causal.

One of the many strengths of this study lied in the integration of genomic and transcriptomic data. An analysis of the expression landscape corroborated observations made at the genomic level: the degree of proliferation and immune activation prior to commencing chemotherapy played key roles in fashioning response to therapy. This study is the first to model the interplay between proliferation and immune infiltration across RCB categories. Tumours attaining pCR had the highest degrees of proliferation and immune activation at diagnosis, with a gradual decrease in both pathways as the degree of RD increased. This affirmed the importance of modelling response as a continuum, rather than as a dichotomized variable as has been done in the past in many studies. Digital pathology analysis orthogonally validated the association between increasing immune infiltrate and response, and through RNA-seq deconvolution the prevalence of a wide gamut of immune subpopulations, including B cells, T cells, NK cells and macrophages, was shown to associate with response.

The neoadjuvant setting provides a unique opportunity to study mechanisms of chemoresistance. In keeping with the immuno-evasive methods described previously, chemoresistant tumours also profoundly under-expressed HLA genes, thereby presenting fewer neoantigens. It is worth noting that the current generation of anti-tumour agents are increasingly immuno-modulatory and often rely on inhibiting immuno-suppressive phenotypes. However, such agents are very unlikely to provide great benefit in cases where the main method of immuno-evasion is secondary to the decreased presentation of neoantigenic peptides, and perhaps screening for HLA LOH will play an important role in the future when selecting candidates suitable for immunotherapies.

Mechanisms of chemoresistance were not solely limited to immuno-evasive pathways. Treatment-naive chemoresistant tumours also had an augmented expression of xenobiotic cytochrome P450 enzymes and ATP binding cassette multi-drug efflux transporters and solute carriers. Perhaps these resistance mechanisms were acquired throughout a lifetime of toxin exposure and might explain why patients with higher body mass indexes have a higher risk of breast cancer recurrence after chemotherapy. A novel mechanism of chemoresistance in $HER2^{+}$ tumours was shown to be mediated by the stem-cell regulator *LIN28A* which has been shown to stabilise HER2 mRNA, resulting in increased protein expression. Through profound down-regulation of *LIN28A*, fewer HER2 molecules would be available on the

cell surface, rendering these tumour cells less sensitive to anti-HER2 therapies. Future work will involve the analysis of these chemoresistance signatures and genes within independent adjuvant datasets, such as METABRIC, in order to determine whether an association can be observed between relapse following adjuvant chemotherapy and chemoresistant gene expression.

One key aim of this work was to generate a classifier that accurately predicted response to therapy. By integrating features best posed to predict response to therapy, a model that incorporated two clinical variables (age at diagnosis and number of positive lymph nodes), two genomic variables (expressed TMB and HLA-LOH) and five expression variables (GGI metagene expression, CYT score and *ESR1, ERBB2* and *LIN28A* expression) was shown to be highly predictive of response. Indeed, this model was able to predict pCR in the entire cohort with a sensitivity of 86.5%, specificity of 81.5% and an NPV of 93.4%, whilst it could predict chemoresistance with a sensitivity of 95.7%, specificity of 91.0% and an NPV of 97.0%, outperforming the clinical variable-only model.

The work described in the first section of the thesis has shed a great deal of light on predictors and mechanisms of chemosensitivity and resistance, with applications directly translatable to the clinic. While many spin-off projects and collaborative efforts have resulted to further explore aspects of the data described, the greatest benefit lies in the application of this data within clinical practice. Oncology practice is becoming increasingly data-driven, with efforts such as the *Personalised Breast Cancer Programme* (PBCP) generating tumour whole genome and transcriptome sequencing data. The PBCP has already resulted in a tangible alteration of patient management in Cambridge, and plans are in motion to include the features identified in this work within the molecular reports. Hence, through the additional data provided, physicians and patients will be in a better position to decide whether neoadjuvant therapies will provide the desired benefit.

Genomic and transcriptomic changes seemingly induced by chemotherapy

A further aim of this study was the molecular analysis of serial tumour samples obtained throughout neoadjuvant therapy. Biopsies were consistently taken under ultrasound guidance from the site of a fiduciary marker to mitigate, as much as possible, intra-tumoural heterogeneity. By characterising the genomic and transcriptomic profiles throughout therapy, the work described in this thesis aimed to increase our understanding of the way tumours, and their surrounding microenvironment, were altered during therapy.

Chemotherapy seemingly altered the genomic architecture of tumours, with evidence of mutation eradication and the emergence of previously unobserved mutations during therapy. Clonal structure deconvolution revealed two main patterns of genomic response to chemotherapy: *clonally stable tumours* characterised by mutational prevalences that were minimally altered on serial sampling, and *clonally dynamic tumours* characterised by altered clonal prevalences during treatment. Clonally dynamic tumours were sub-classified as *tumours with clonal extinction* (harbouring mutations with decreasing cellular prevalences) and *tumours with clonal emergence* (harbouring mutations with increasing cellular prevalences). Deep sequencing of serial tumour samples from four cases validated the clonal structure observed on exome sequencing. One case was subjected to multi-region sequencing and the results showed that newly emergent mutations were present prior to commencing chemotherapy, albeit at very low allelic fractions, corroborating findings observed by Kim et al. [155]. Hence mutations that were 'newly-detected' during therapy had been previously unobserved due to insufficient sequencing depth. Deep sequencing on a larger cohort as well as single-cell sequencing will be required to provide additional validation and delineate the dynamics in greater detail: work to undertake this is already underway.

Hence, selection pressures introduced by chemotherapy seemingly altered the tumour's clonal architecture, with selection against chemosensitive populations. Newly observed somatic mutations were unlikely to be neoantigenic and on pathway enrichment were noted to be harboured within genes that were modulators of immune signalling, proliferation, cell motility and molecular transporters. It is worth remembering that these pathways were also noted to be predictive of resistance prior to commencing chemotherapy.

Clonally stable tumours retained a stable mutational and copy number landscape, with lower proliferation pathway activation compared to clonally dynamic tumours. Tumours with clonal extinction had a significantly higher immune infiltrate compared to all other groups, indicating that cytotoxic therapy and an augmented immune response were responsible for the subclonal extinction observed. In contrast, tumours with clonal emergence showed evidence of chemoresistance and an associated increase in copy number instability and homologous recombination deficiency during therapy, with evidence of changes within the CNA landscape. These tumours also had a significantly lower immune infiltrate when compared to all other tumours in different response groups. It was surprising how the immune infiltrate not only determined the clinical response to therapy, but also the clonal architecture of a tumour.

Another strength of this study is the integration of breast tumour MRI data with concurrently obtained expression data. The rate of tumour bulk decrease on MRI at the midway time point

correlated with pathological response to therapy and on integrating this with the expression data it rapidly became evident that the key pathways associated with a greater change in tumour bulk on MRI were proliferation and immune activation. It was highly remarkable that a linear correlation was observed between the degree of immune subpopulation enrichment at the midway time point and the percentage decrease in bulk on MRI. This continued to confirm the important role the immune system played in determining response to therapy. In view of this, collaborations have now been set up with the Department of Radiology and within the Cancer Research (UK) Cambridge Institute to integrate other radiomic features, such as tumour texture and perfusion analyses, with the expression and genomic data generated in this work. The data generated to date shows great promise.

Intriguingly, differing immune dynamics were observed at the midway time point in taxane treated $HER2^{-}$ tumours and anthracycline-treated $HER2^{+}$ tumours. In the former, an increase in CYT score, as well as an increase in B cells, T cells, NK cells and neutrophils was strongly associated with pCR. This observed increase in tumour infiltrate was secondary to decreased tumour purity and increased infiltration by immune cells, which could have been recruited secondary to increased chemotherapy-induced cell death. An inverse relationship was seen in the anthracycline-treated $HER2^{+}$ tumours. This observation has not been reported in the literature: whether this was an effect secondary to the type of chemotherapy administered, or whether this was a $HER2^{+}$-specific response was difficult to untangle, though observations made in various studies that taxanes given before anthracyclines were more effective at inducing pCR may hint that this was a chemotherapy-related event. A larger cohort will be required to validate this observation: as the TransNEO study is still actively recruiting patients plans are being made to validate these findings in the subsequent cohort of patients analysed. Further experimental work will need to be done to confirm the immune system dynamics observed.

Therapy-induced senescence, amongst other chemoresistance mechanisms, was noted to be associated with chemoresistance, with some tumour cells entering a senescent state in order to circumvent the cytotoxic effects of chemotherapy. Through collaborations with research groups in the Cancer Research (UK) Cambridge Institute, efforts have now started to use computational pathology and machine learning classifiers to identify and quantify senescent phenotypes on more than 5,000 tumour H&E slides obtained prior to commencing therapy and on completion of therapy. While this work is still in its infancy, the performance of the algorithms so far has been impressive.

Hence this study has described how tumours and their microenvironments are altered during therapy. The next stages will involve the deeper integration of genomic, transcriptomic, digital pathology and radiomics data in order to generate a comprehensive, multi-modal and integrated description of the changes chemotherapy induces within tumours and their microenvironments.

ctDNA detection

In the final section of this work, deep sequencing and shallow whole genome sequencing were used to detect ctDNA within the plasma. ctDNA was detected robustly in all diagnostic plasma samples that were deep sequenced and in 79% of cases that were shallow whole genome sequenced. The increased sensitivity and specificity of the deep sequencing technology used were attributed to the deep single molecule amplification as well as the utilisation of a robust noise model that aided sensitivity and specificity.

The work described has provided a proof-of-principle solution that allows the robust detection of ctDNA at diagnosis in early-stage disease and during neoadjuvant therapy and will motivate the eventual comprehensive analysis of the remaining plasma samples acquired in the study. With a larger number of samples it will be possible to determine whether ctDNA levels are predictive of response to therapy and whether the dynamics observed correlate with response. Indeed, if increasing ctDNA levels were consistently observed during one block of chemotherapy, then an argument could be made that this indicated chemoresistance and a switch to an alternative cytotoxic agent or earlier surgical intervention might be beneficial. Post-surgical ctDNA analysis could also be used to predict the probability of eventual relapse or guide further adjuvant therapy.

ctDNA surveillance could be used to allow the early detection of relapse, and to this effect, the *REVEAL* translational study was set up as part of this PhD. Within this study, plasma sampling is being performed at defined time points over a six-year period, and ctDNA analysis performed using ultra-sensitive techniques in order to enable early detection of relapse in patients with all stages of breast cancer.

References

[1] Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

[2] Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

[3] Abu-Odeh, M., Salah, Z., Herbel, C., Hofmann, T. G., and Aqeilan, R. I. (2014). WWOX, the common fragile site FRA16D gene product, regulates ATM activation and the DNA damage response. *Proc. Natl. Acad. Sci. U. S. A.*, 111(44):E4716–25.

[4] Acharyya, S., Oskarsson, T., Vanharanta, S., Malladi, S., Kim, J., Morris, P. G., Manova-Todorova, K., Leversha, M., Hogg, N., Seshan, V. E., Norton, L., Brogi, E., and Massagué, J. (2012). A CXCL1 paracrine network links cancer chemoresistance and metastasis. *Cell*, 150(1):165–78.

[5] Adalsteinsson, V. A., Ha, G., Freeman, S. S., Choudhury, A. D., Stover, D. G., Parsons, H. A., Gydush, G., Reed, S. C., Rotem, D., Rhoades, J., Loginov, D., Livitz, D., Rosebrock, D., Leshchiner, I., Kim, J., Stewart, C., Rosenberg, M., Francis, J. M., Zhang, C.-Z., Cohen, O., Oh, C., Ding, H., Polak, P., Lloyd, M., Mahmud, S., Helvie, K., Merrill, M. S., Santiago, R. A., O'Connor, E. P., Jeong, S. H., Leeson, R., Barry, R. M., Kramkowski, J. F., Zhang, Z., Polacek, L., Lohr, J. G., Schleicher, M., Lipscomb, E., Saltzman, A., Oliver, N. M., Marini, L., Waks, A. G., Harshman, L. C., Tolaney, S. M., Van Allen, E. M., Winer, E. P., Lin, N. U., Nakabayashi, M., Taplin, M.-E., Johannessen, C. M., Garraway, L. A., Golub, T. R., Boehm, J. S., Wagle, N., Getz, G., Love, J. C., and Meyerson, M. (2017). Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.*, 8(1):1324.

[6] Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., and Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.*, 11(12):R119.

[7] Adland, E., Paioni, P., Thobakgale, C., Laker, L., Mori, L., Muenchhoff, M., Csala, A., Clapson, M., Flynn, J., Novelli, V., Hurst, J., Naidoo, V., Shapiro, R., Huang, K.-H. G., Frater, J., Prendergast, A., Prado, J. G., Ndung'u, T., Walker, B. D., Carrington, M., Jooste, P., and Goulder, P. J. R. (2015). Discordant Impact of HLA on Viral Replicative Capacity and Disease Progression in Pediatric and Adult HIV Infection. *PLoS Pathog.*, 11(6):e1004954.

[8] Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Girón, C. G., Gordon, L., Hourlier, T.,

Hunt, S. E., Janacek, S. H., Juettemann, T., Keenan, S., Laird, M. R., Lavidas, I., Maurel, T., McLaren, W., Moore, B., Murphy, D. N., Nag, R., Newman, V., Nuhn, M., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Wilder, S. P., Zadissa, A., Kostadima, M., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Cunningham, F., Yates, A., Zerbino, D. R., and Flicek, P. (2017). Ensembl 2017. *Nucleic Acids Res.*, 45(D1):D635–D642.

[9] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013a). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21.

[10] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, 3(1):246–59.

[11] Ali, H. R., Dariush, A., Provenzano, E., Bardwell, H., Abraham, J. E., Iddawela, M., Vallier, A.-L., Hiller, L., Dunn, J. A., Bowden, S. J., Hickish, T., McAdam, K., Houston, S., Irwin, M. J., Pharoah, P. D. P., Brenton, J. D., Walton, N. A., Earl, H. M., and Caldas, C. (2016). Computational pathology of pre-treatment biopsies identifies lymphocyte density as a predictor of response to neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res.*, 18(1):21.

[12] Ali, H. R., Rueda, O. M., Chin, S.-F., Curtis, C., Dunning, M. J., Aparicio, S. A., and Caldas, C. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.*, 15(8):431.

[13] Alighieri, D., Medici, L., and Galizzi, G. B. (2001). *La Divina Commedia*. Bolis.

[14] Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–9.

[15] Andor, N., Harness, J. V., Müller, S., Mewes, H. W., and Petritsch, C. (2014). EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60.

[16] Ayers, M., Symmans, W. F., Stec, J., Damokosh, A. I., Clark, E., Hess, K., Lecocke, M., Metivier, J., Booser, D., Ibrahim, N., Valero, V., Royce, M., Arun, B., Whitman, G., Ross, J., Sneige, N., Hortobagyi, G. N., and Pusztai, L. (2004). Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer. *J. Clin. Oncol.*, 22(12):2284–93.

[17] Bachman, K. E., Argani, P., Samuels, Y., Silliman, N., Ptak, J., Szabo, S., Konishi, H., Karakas, B., Blair, B. G., Lin, C., Peters, B. A., Velculescu, V. E., and Park, B. H. (2004). The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biol. Ther.*, 3(8):772–5.

[18] Balko, J. M., Giltnane, J. M., Wang, K., Schwarz, L. J., Young, C. D., Cook, R. S., Owens, P., Sanders, M. E., Kuba, M. G., Sánchez, V., Kurupi, R., Moore, P. D., Pinto, J. A., Doimi, F. D., Gómez, H., Horiuchi, D., Goga, A., Lehmann, B. D., Bauer, J. A., Pietenpol, J. A., Ross, J. S., Palmer, G. A., Yelensky, R., Cronin, M., Miller, V. A., Stephens, P. J., and Arteaga, C. L. (2014). Molecular profiling of the residual disease of triple-negative breast cancers after neoadjuvant chemotherapy identifies actionable therapeutic targets. *Cancer Discov.*, 4(2):232–45.

[19] Barker, A. D., Sigman, C. C., Kelloff, G. J., Hylton, N. M., Berry, D. A., and Esserman, L. J. (2009). I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin. Pharmacol. Ther.*, 86(1):97–100.

[20] Barnett, D. H., Sheng, S., Charn, T. H., Waheed, A., Sly, W. S., Lin, C.-Y., Liu, E. T., and Katzenellenbogen, B. S. (2008). Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. *Cancer Res.*, 68(9):3505–15.

[21] Baron, L. F., Baron, P. L., Ackerman, S. J., Durden, D. D., and Pope, T. L. (2000). Sonographically guided clip placement facilitates localization of breast cancer after neoadjuvant chemotherapy. *AJR. Am. J. Roentgenol.*, 174(2):539–40.

[22] Baselga, J., Bradbury, I., Eidtmann, H., Di Cosimo, S., de Azambuja, E., Aura, C., Gómez, H., Dinh, P., Fauria, K., Van Dooren, V., Aktan, G., Goldhirsch, A., Chang, T.-W., Horváth, Z., Coccia-Portugal, M., Domont, J., Tseng, L.-M., Kunz, G., Sohn, J. H., Semiglazov, V., Lerzo, G., Palacova, M., Probachai, V., Pusztai, L., Untch, M., Gelber, R. D., Piccart-Gebhart, M., and NeoALTTO Study Team (2012). Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): a randomised, open-label, multicentre, phase 3 trial. *Lancet*, 379(9816):633–40.

[23] Bear, H. D., Anderson, S., Brown, A., Smith, R., Mamounas, E. P., Fisher, B., Margolese, R., Theoret, H., Soran, A., Wickerham, D. L., Wolmark, N., and National Surgical Adjuvant Breast and Bowel Project Protocol B-27 (2003). The effect on tumor response of adding sequential preoperative docetaxel to preoperative doxorubicin and cyclophosphamide: preliminary results from National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *J. Clin. Oncol.*, 21(22):4165–74.

[24] Bear, H. D., Anderson, S., Smith, R. E., Geyer, C. E., Mamounas, E. P., Fisher, B., Brown, A. M., Robidoux, A., Margolese, R., Kahlenberg, M. S., Paik, S., Soran, A., Wickerham, D. L., and Wolmark, N. (2006). Sequential preoperative or postoperative docetaxel added to preoperative doxorubicin plus cyclophosphamide for operable breast cancer:National Surgical Adjuvant Breast and Bowel Project Protocol B-27. *J. Clin. Oncol.*, 24(13):2019–27.

[25] Bear, H. D., Tang, G., Rastogi, P., Geyer, C. E., Liu, Q., Robidoux, A., Baez-Diaz, L., Brufsky, A. M., Mehta, R. S., Fehrenbacher, L., Young, J. A., Senecal, F. M., Gaur, R., Margolese, R. G., Adams, P. T., Gross, H. M., Costantino, J. P., Paik, S., Swain, S. M., Mamounas, E. P., and Wolmark, N. (2015). Neoadjuvant plus adjuvant bevacizumab in early breast cancer (NSABP B-40 [NRG Oncology]): secondary outcomes of a phase 3, randomised controlled trial. *Lancet. Oncol.*, 16(9):1037–1048.

[26] Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W. H., and de Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.*, 17(1):218.

[27] Bélembaogo, E., Feillel, V., Chollet, P., Curé, H., Verrelle, P., Kwiatkowski, F., Achard, J. L., Le Bouëdec, G., Chassagne, J., and Bignon, Y. J. (1992). Neoadjuvant chemotherapy in 126 operable breast cancers. *Eur. J. Cancer*, 28A(4-5):896–900.

[28] Beloribi-Djefaflia, S., Vasseur, S., and Guillaumond, F. (2016). Lipid metabolic reprogramming in cancer cells. *Oncogenesis*, 5:e189.

[29] Benjamini, Yoav and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, 57(1):289–300.

[30] Bettegowda, C., Sausen, M., Leary, R. J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B. R., Wang, H., Luber, B., Alani, R. M., Antonarakis, E. S., Azad, N. S., Bardelli, A., Brem, H., Cameron, J. L., Lee, C. C., Fecher, L. a., Gallia, G. L., Gibbs, P., Le, D., Giuntoli, R. L., Goggins, M., Hogarty, M. D., Holdhoff, M., Hong, S.-M., Jiao, Y., Juhl, H. H., Kim, J. J., Siravegna, G., Laheru, D. a., Lauricella, C., Lim, M., Lipson, E. J., Marie, S. K. N., Netto, G. J., Oliner, K. S., Olivi, A., Olsson, L., Riggins, G. J., Sartore-Bianchi, A., Schmidt, K., le Ming Shih, Oba-Shinjo, S. M., Siena, S., Theodorescu, D., Tie, J., Harkins, T. T., Veronese, S., Wang, T.-L., Weingart, J. D., Wolfgang, C. L., Wood, L. D., Xing, D., Hruban, R. H., Wu, J., Allen, P. J., Schmidt, C. M., Choti, M. a., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., Papadopoulos, N., and Diaz, L. a. (2014). Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.*, 6(224):224ra24.

[31] Bhakta, S., Crocker, L. M., Chen, Y., Hazen, M., Schutten, M. M., Li, D., Kuijl, C., Ohri, R., Zhong, F., Poon, K. A., Go, M. A. T., Cheng, E., Piskol, R., Firestein, R., Fourie-O'Donohue, A., Kozak, K. R., Raab, H., Hongo, J.-A., Sampath, D., Dennis, M. S., Scheller, R. H., Polakis, P., and Junutula, J. R. (2018). An Anti-GDNF Family Receptor Alpha 1 (GFRA1) Antibody-Drug Conjugate for the Treatment of Hormone Receptor-Positive Breast Cancer. *Mol. Cancer Ther.*, 17(3):638–649.

[32] Bines, J., Earl, H., Buzaid, A. C., and Saad, E. D. (2014). Anthracyclines and taxanes in the neo/adjuvant treatment of breast cancer: does the sequence matter? *Ann. Oncol.*, 25(6):1079–85.

[33] Board, R. E., Wardley, A. M., Dixon, J. M., Armstrong, A. C., Howell, S., Renshaw, L., Donald, E., Greystoke, A., Ranson, M., Hughes, A., and Dive, C. (2010). Detection of PIK3CA mutations in circulating free DNA in patients with breast cancer. *Breast Cancer Res. Treat.*, 120(2):461–7.

[34] Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., and Chudakov, D. M. (2015). MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, 12(5):380–1.

[35] Bonadonna, G., Brusamolino, E., Valagussa, P., Rossi, A., Brugnatelli, L., Brambilla, C., De Lena, M., Tancini, G., Bajetta, E., Musumeci, R., and Veronesi, U. (1976). Combination chemotherapy as an adjuvant treatment in operable breast cancer. *N. Engl. J. Med.*, 294(8):405–10.

[36] Bonadonna, G. and Valagussa, P. (1989). Systemic therapy in resectable breast cancer. *Hematol. Oncol. Clin. North Am.*, 3(4):727–42.

[37] Bonadonna, G., Valagussa, P., Brambilla, C., Ferrari, L., Moliterni, A., Terenziani, M., and Zambetti, M. (1998). Primary chemotherapy in operable breast cancer: eight-year experience at the Milan Cancer Institute. *J. Clin. Oncol.*, 16(1):93–100.

[38] Bonnefoi, H., Piccart, M., Bogaerts, J., Mauriac, L., Fumoleau, P., Brain, E., Petit, T., Rouanet, P., Jassem, J., Blot, E., Zaman, K., Cufer, T., Lortholary, A., Lidbrink, E., André, S., Litière, S., Lago, L. D., Becette, V., Cameron, D. A., Bergh, J., Iggo, R., and EORTC 10994/BIG 1-00 Study Investigators (2011). TP53 status for prediction of sensitivity to taxane versus non-taxane neoadjuvant chemotherapy in breast cancer (EORTC 10994/BIG 1-00): a randomised phase 3 trial. *Lancet. Oncol.*, 12(6):527–39.

[39] Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D. R., Steins, M., Ready, N. E., Chow, L. Q., Vokes, E. E., Felip, E., Holgado, E., Barlesi, F., Kohlhäufl, M., Arrieta, O., Burgio, M. A., Fayette, J., Lena, H., Poddubskaya, E., Gerber, D. E., Gettinger, S. N., Rudin, C. M., Rizvi, N., Crinò, L., Blumenschein, G. R., Antonia, S. J., Dorange, C., Harbison, C. T., Graf Finckenstein, F., and Brahmer, J. R. (2015). Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. *N. Engl. J. Med.*, 373(17):1627–39.

[40] Bossuyt, V., Provenzano, E., Symmans, W. F., Boughey, J. C., Coles, C., Curigliano, G., Dixon, J. M., Esserman, L. J., Fastner, G., Kuehn, T., Peintinger, F., von Minckwitz, G., White, J., Yang, W., Badve, S., Denkert, C., MacGrogan, G., Penault-Llorca, F., Viale, G., Cameron, D., and Breast International Group-North American Breast Cancer Group (BIG-NABCG) collaboration (2015). Recommendations for standardized pathological characterization of residual disease for neoadjuvant clinical trials of breast cancer by the BIG-NABCG collaboration. *Ann. Oncol.*, 26(7):1280–91.

[41] Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., and Holt, R. A. (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res.*, 24(5):743–50.

[42] Buchholz, T. A., Stivers, D. N., Stec, J., Ayers, M., Clark, E., Bolt, A., Sahin, A. A., Symmans, W. F., Hess, K. R., Kuerer, H. M., Valero, V., Hortobagyi, G. N., and Pusztai, L. (2002). Global gene expression changes during neoadjuvant chemotherapy for human breast cancer. *Cancer J.*, 8(6):461–8.

[43] Buchwald, H. (1992). Cholesterol inhibition, cancer, and chemotherapy. *Lancet*, 339(8802):1154–6.

[44] Callari, M., Cappelletti, V., D'Aiuto, F., Musella, V., Lembo, A., Petel, F., Karn, T., Iwamoto, T., Provero, P., Daidone, M. G., Gianni, L., and Bianchini, G. (2016). Subtype-Specific Metagene-Based Prediction of Outcome after Neoadjuvant and Adjuvant Treatment in Breast Cancer. *Clin. Cancer Res.*, 22(2):337–45.

[45] Cameron, D., Piccart-Gebhart, M. J., Gelber, R. D., Procter, M., Goldhirsch, A., de Azambuja, E., Castro, G., Untch, M., Smith, I., Gianni, L., Baselga, J., Al-Sakaff, N., Lauer, S., McFadden, E., Leyland-Jones, B., Bell, R., Dowsett, M., Jackisch, C., and Herceptin Adjuvant (HERA) Trial Study Team (2017). 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. *Lancet*, 389(10075):1195–1205.

[46] Campbell, J. I., Yau, C., Krass, P., Moore, D., Carey, L. A., Au, A., Chhieng, D., Giri, D., Livasy, C., Mies, C., Rabban, J., Sarode, V. R., Singh, B., Esserman, L., and Chen, Y.-Y. (2017). Comparison of residual cancer burden, American Joint Committee on Cancer staging and pathologic complete response in breast cancer after neoadjuvant chemotherapy: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res. Treat.*, 165(1):181–191.

[47] Carey, L. A., Berry, D. A., Cirrincione, C. T., Barry, W. T., Pitcher, B. N., Harris, L. N., Ollila, D. W., Krop, I. E., Henry, N. L., Weckstein, D. J., Anders, C. K., Singh, B., Hoadley, K. A., Iglesia, M., Cheang, M. C. U., Perou, C. M., Winer, E. P., and Hudis, C. A. (2016). Molecular Heterogeneity and Response to Neoadjuvant Human Epidermal Growth Factor Receptor 2 Targeting in CALGB 40601, a Randomized Phase III Trial of Paclitaxel Plus Trastuzumab With or Without Lapatinib. *J. Clin. Oncol.*, 34(6):542–9.

[48] Carlson, J. M., Listgarten, J., Pfeifer, N., Tan, V., Kadie, C., Walker, B. D., Ndung'u, T., Shapiro, R., Frater, J., Brumme, Z. L., Goulder, P. J. R., and Heckerman, D. (2012). Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *J. Virol.*, 86(9):5230–43.

[49] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., and Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, 30(5):413–21.

[50] Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelink, H., van de Rijn, M., Brown, P. O., and van de Vijver, M. J. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. U. S. A.*, 102(10):3738–43.

[51] Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H., and Trajanoski, Z. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.*, 18(1):248–262.

[52] Chen, M.-B., Zhu, Y.-Q., Xu, J.-Y., Wang, L.-Q., Liu, C.-Y., Ji, Z.-Y., and Lu, P.-H. (2012). Value of TP53 status for predicting response to neoadjuvant chemotherapy in breast cancer: a meta-analysis. *PLoS One*, 7(6):e39655.

[53] Chen, Y.-H., Hancock, B. A., Solzak, J. P., Brinza, D., Scafe, C., Miller, K. D., and Radovich, M. (2017). Next-generation sequencing of circulating tumor DNA to predict recurrence in triple-negative breast cancer patients with residual disease after neoadjuvant chemotherapy. *NPJ breast cancer*, 3:24.

[54] Chevallier, B., Roche, H., Olivier, J. P., Chollet, P., and Hurteloup, P. (1993). Inflammatory breast cancer. Pilot study of intensive induction chemotherapy (FEC-HD) results in a high histologic response rate. *Am. J. Clin. Oncol.*, 16(3):223–8.

[55] Cimino-Mathews, A., Subhawong, A. P., Illei, P. B., Sharma, R., Halushka, M. K., Vang, R., Fetting, J. H., Park, B. H., and Argani, P. (2013). GATA3 expression in breast carcinoma: utility in triple-negative, sarcomatoid, and metastatic carcinomas. *Hum. Pathol.*, 44(7):1341–9.

[56] Ciriello, G., Miller, M. L., Aksoy, B. A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, 45(10):1127–33.

[57] Colleoni, M., Bagnardi, V., Rotmensz, N., Gelber, R. D., Viale, G., Pruneri, G., Veronesi, P., Torrisi, R., Cardillo, A., Montagna, E., Campagnoli, E., Luini, A., Intra, M., Galimberti, V., Scarano, E., Peruzzotti, G., and Goldhirsch, A. (2009). Increasing steroid hormone receptors expression defines breast cancer subtypes non responsive to preoperative chemotherapy. *Breast Cancer Res. Treat.*, 116(2):359–69.

[58] Colleoni, M., Viale, G., Zahrieh, D., Bottiglieri, L., Gelber, R. D., Veronesi, P., Balduzzi, A., Torrisi, R., Luini, A., Intra, M., Dellapasqua, S., Cardillo, A., Ghisini, R., Peruzzotti, G., and Goldhirsch, A. (2008). Expression of ER, PgR, HER1, HER2, and response: a study of preoperative chemotherapy. *Ann. Oncol.*, 19(3):465–72.

[59] Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., Fostel, J. L., Friedrich, D. C., Perrin, D., Dionne, D., Kim, S., Gabriel, S. B., Lander, E. S., Fisher, S., and Getz, G. (2013). Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.*, 41(6):e67.

[60] Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–52.

[61] Dagogo-Jack, I. and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.*, 15(2):81–94.

[62] Dash, N., Chafin, S. H., Johnson, R. R., and Contractor, F. M. (1999). Usefulness of tissue marker clips in patients undergoing neoadjuvant chemotherapy for breast cancer. *AJR. Am. J. Roentgenol.*, 173(4):911–7.

[63] Dawson, S.-J., Rueda, O. M., Aparicio, S., and Caldas, C. (2013a). A new genome-driven integrated classification of breast cancer and its implications. *EMBO J.*, (January):1–12.

[64] Dawson, S.-J., Tsui, D. W. Y., Murtaza, M., Biggs, H., Rueda, O. M., Chin, S.-F., Dunning, M. J., Gale, D., Forshew, T., Mahler-Araujo, B., Rajan, S., Humphray, S., Becq, J., Halsall, D., Wallis, M., Bentley, D., Caldas, C., and Rosenfeld, N. (2013b). Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.*, 368(13):1199–209.

[65] DePristo, M. a., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. a., del Angel, G., Rivas, M. a., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498.

[66] Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, 16:35.

[67] Desmedt, C., Di Leo, A., de Azambuja, E., Larsimont, D., Haibe-Kains, B., Selleslags, J., Delaloge, S., Duhem, C., Kains, J.-P., Carly, B., Maerevoet, M., Vindevoghel, A., Rouas, G., Lallemand, F., Durbecq, V., Cardoso, F., Salgado, R., Rovere, R., Bontempi, G., Michiels, S., Buyse, M., Nogaret, J.-M., Qi, Y., Symmans, F., Pusztai, L., D'Hondt, V., Piccart-Gebhart, M., and Sotiriou, C. (2011). Multifactorial approach to predicting resistance to anthracyclines. *J. Clin. Oncol.*, 29(12):1578–86.

[68] Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.*, 14(16):5158–65.

[69] Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., McMichael, J. F., Wallis, J. W., Lu, C., Shen, D., Harris, C. C., Dooling, D. J., Fulton, R. S., Fulton, L. L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Wendl, M. C., Heath, S., Watson, M. A., Link, D. C., Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., Graubert, T. A., Mardis, E. R., Wilson, R. K., and DiPersio, J. F. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–10.

[70] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

[71] Dong, K., Zhao, H., Tong, T., and Wan, X. (2016). NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. *BMC Bioinformatics*, 17(1):369.

[72] Earl, H. M., Hiller, L., Dunn, J. A., Blenkinsop, C., Grybowicz, L., Vallier, A.-L., Abraham, J., Thomas, J., Provenzano, E., Hughes-Davies, L., Gounaris, I., McAdam, K., Chan, S., Ahmad, R., Hickish, T., Houston, S., Rea, D., Bartlett, J., Caldas, C., Cameron, D. A., Hayward, L., and ARTemis Investigators (2015). Efficacy of neoadjuvant bevacizumab added to docetaxel followed by fluorouracil, epirubicin, and cyclophosphamide, for women with HER2-negative early breast cancer (ARTemis): an open-label, randomised, phase 3 trial. *Lancet. Oncol.*, 16(6):656–66.

[73] Earl, H. M., Vallier, A.-L., Hiller, L., Fenwick, N., Young, J., Iddawela, M., Abraham, J., Hughes-Davies, L., Gounaris, I., McAdam, K., Houston, S., Hickish, T., Skene, A., Chan, S., Dean, S., Ritchie, D., Laing, R., Harries, M., Gallagher, C., Wishart, G., Dunn, J., Provenzano, E., Caldas, C., and Neo-tAnGo Investigators (2014). Effects of the addition of gemcitabine, and paclitaxel-first sequencing, in neoadjuvant sequential epirubicin, cyclophosphamide, and paclitaxel for women with high-risk early breast cancer (Neo-tAnGo): an open-label, 2×2 factorial randomised phase 3 trial. *Lancet. Oncol.*, 15(2):201–12.

[74] Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (2018). Long-term outcomes for neoadjuvant versus adjuvant chemotherapy in early breast cancer: meta-analysis of individual patient data from ten randomised trials. *Lancet. Oncol.*, 19(1):27–39.

[75] Elaraj, D. M., Weinreich, D. M., Varghese, S., Puhlmann, M., Hewitt, S. M., Carroll, N. M., Feldman, E. D., Turner, E. M., and Alexander, H. R. (2006). The role of interleukin 1 in growth and metastasis of human cancer xenografts. *Clin. Cancer Res.*, 12(4):1088–96.

[76] Esserman, L. J., Berry, D. A., Cheang, M. C. U., Yau, C., Perou, C. M., Carey, L., DeMichele, A., Gray, J. W., Conway-Dorsey, K., Lenburg, M. E., Buxton, M. B., Davis, S. E., van't Veer, L. J., Hudis, C., Chin, K., Wolf, D., Krontiras, H., Montgomery, L., Tripathy, D., Lehman, C., Liu, M. C., Olopade, O. I., Rugo, H. S., Carpenter, J. T., Livasy, C., Dressler, L., Chhieng, D., Singh, B., Mies, C., Rabban, J., Chen, Y.-Y., Giri, D., Au, A., Hylton, N., and I-SPY 1 TRIAL Investigators (2012a). Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res. Treat.*, 132(3):1049–62.

[77] Esserman, L. J., Berry, D. A., DeMichele, A., Carey, L., Davis, S. E., Buxton, M., Hudis, C., Gray, J. W., Perou, C., Yau, C., Livasy, C., Krontiras, H., Montgomery, L., Tripathy, D., Lehman, C., Liu, M. C., Olopade, O. I., Rugo, H. S., Carpenter, J. T., Dressler, L., Chhieng, D., Singh, B., Mies, C., Rabban, J., Chen, Y.-Y., Giri, D., van 't Veer, L., and Hylton, N. (2012b). Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL–CALGB 150007/150012, ACRIN 6657. *J. Clin. Oncol.*, 30(26):3242–9.

[78] Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, 46(D1):D649–D655.

[79] Feng, C., Neumeister, V., Ma, W., Xu, J., Lu, L., Bordeaux, J., Maihle, N. J., Rimm, D. L., and Huang, Y. (2012). Lin28 regulates HER2 and promotes malignancy through multiple mechanisms. *Cell Cycle*, 11(13):2486–94.

[80] Fisher, B. (1977). United States trials of conservative surgery. *World J. Surg.*, 1(3):327–30.

[81] Fisher, B., Brown, A. M., Dimitrov, N. V., Poisson, R., Redmond, C., Margolese, R. G., Bowman, D., Wolmark, N., Wickerham, D. L., and Kardinal, C. G. (1990). Two months of doxorubicin-cyclophosphamide with and without interval reinduction therapy compared with 6 months of cyclophosphamide, methotrexate, and fluorouracil in positive-node breast cancer patients with tamoxifen-nonresponsive tumors: results from t. *J. Clin. Oncol.*, 8(9):1483–96.

[82] Fisher, B., Bryant, J., Wolmark, N., Mamounas, E., Brown, A., Fisher, E. R., Wickerham, D. L., Begovic, M., DeCillis, A., Robidoux, A., Margolese, R. G., Cruz, A. B., Hoehn, J. L., Lees, A. W., Dimitrov, N. V., and Bear, H. D. (1998). Effect of preoperative chemotherapy on the outcome of women with operable breast cancer. *J. Clin. Oncol.*, 16(8):2672–85.

[83] Fisher, D. T., Appenheimer, M. M., and Evans, S. S. (2014). The two faces of IL-6 in the tumor microenvironment. *Semin. Immunol.*, 26(1):38–47.

[84] Fisher, E. R., Wang, J., Bryant, J., Fisher, B., Mamounas, E., and Wolmark, N. (2002). Pathobiology of preoperative chemotherapy: findings from the National Surgical Adjuvant Breast and Bowel (NSABP) protocol B-18. *Cancer*, 95(4):681–95.

[85] Fontanella, C., Lederer, B., Gade, S., Vanoppen, M., Blohmer, J. U., Costa, S. D., Denkert, C., Eidtmann, H., Gerber, B., Hanusch, C., Hilfrich, J., Huober, J., Schneeweiss, A., Paepke, S., Jackisch, C., Mehta, K., Nekljudova, V., Untch, M., Neven, P., von Minckwitz, G., and Loibl, S. (2015). Impact of body mass index on neoadjuvant treatment outcome: a pooled analysis of eight prospective neoadjuvant breast cancer trials. *Breast Cancer Res. Treat.*, 150(1):127–39.

[86] Forse, C. L., Agarwal, S., Pinnaduwage, D., Gertler, F., Condeelis, J. S., Lin, J., Xue, X., Johung, K., Mulligan, A. M., Rohan, T. E., Bull, S. B., and Andrulis, I. L. (2015). Menacalc, a quantitative method of metastasis assessment, as a prognostic marker for axillary node-negative breast cancer. *BMC Cancer*, 15(1):483.

[87] Forshew, T., Murtaza, M., Parkinson, C., Gale, D., Tsui, D. W. Y., Kaper, F., Dawson, S.-J., Piskorz, A. M., Jimenez-Linan, M., Bentley, D., Hadfield, J., May, A. P., Caldas, C., Brenton, J. D., and Rosenfeld, N. (2012). Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.*, 4(136):136ra68.

[88] Foster, K. W., Frost, A. R., McKie-Bell, P., Lin, C. Y., Engler, J. A., Grizzle, W. E., and Ruppert, J. M. (2000). Increase of GKLF messenger RNA and protein expression during progression of breast cancer. *Cancer Res.*, 60(22):6488–95.

[89] French Adjuvant Study Group (2001). Benefit of a high-dose epirubicin regimen in adjuvant chemotherapy for node-positive breast cancer patients with poor prognostic factors: 5-year follow-up results of French Adjuvant Study Group 05 randomized trial. *J. Clin. Oncol.*, 19(3):602–11.

[90] Fumagalli, D., Venet, D., Ignatiadis, M., Azim, H. A., Maetens, M., Rothé, F., Salgado, R., Bradbury, I., Pusztai, L., Harbeck, N., Gomez, H., Chang, T.-W., Coccia-Portugal, M. A., Di Cosimo, S., de Azambuja, E., de la Peña, L., Nuciforo, P., Brase, J. C., Huober, J., Baselga, J., Piccart, M., Loi, S., and Sotiriou, C. (2016). RNA Sequencing to Predict Response to Neoadjuvant Anti-HER2 Therapy: A Secondary Analysis of the NeoALTTO Randomized Clinical Trial. *JAMA Oncol.*

[91] Gajdos, C., Tartter, P. I., Estabrook, A., Gistrak, M. A., Jaffer, S., and Bleiweiss, I. J. (2002). Relationship of clinical and pathologic response to neoadjuvant chemotherapy and outcome of locally advanced breast cancer. *J. Surg. Oncol.*, 80(1):4–11.

[92] Gale, D., Lawson, A. R. J., Howarth, K., Madi, M., Durham, B., Smalley, S., Calaway, J., Blais, S., Jones, G., Clark, J., Dimitrov, P., Pugh, M., Woodhouse, S., Epstein, M., Fernandez-Gonzalez, A., Whale, A. S., Huggett, J. F., Foy, C. A., Jones, G. M., Raveh-Amit, H., Schmitt, K., Devonshire, A., Green, E., Forshew, T., Plagnol, V., and Rosenfeld, N. (2018). Development of a highly sensitive liquid biopsy platform to detect clinically-relevant cancer mutations at low allele fractions in cell-free DNA. *PLoS One*, 13(3):e0194630.

[93] Gallopin, M., Rau, A., and Jaffrézic, F. (2013). A hierarchical poisson log-normal model for network inference from RNA sequencing data. *PLoS One*, 8(10):e77503.

[94] Gao, R., Davis, A., McDonald, T. O., Sei, E., Shi, X., Wang, Y., Tsai, P.-C., Casasent, A., Waters, J., Zhang, H., Meric-Bernstam, F., Michor, F., and Navin, N. E. (2016). Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.*, 48(10):1119–30.

[95] Garcia-Murillas, I., Schiavon, G., Weigelt, B., Ng, C., Hrebien, S., Cutts, R. J., Cheang, M., Osin, P., Nerurkar, A., Kozarewa, I., Garrido, J. A., Dowsett, M., Reis-Filho, J. S., Smith, I. E., and Turner, N. C. (2015). Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci. Transl. Med.*, 7(302):302ra133.

[96] Garraway, L. A. and Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, 153(1):17–37.

[97] Genovese, G., Handsaker, R. E., Li, H., Kenny, E. E., and McCarroll, S. A. (2013). Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. *Am. J. Hum. Genet.*, 93(3):411–21.

[98] Gercel-Taylor, C., Scobee, J. J., and Taylor, D. D. (2005). Effect of chemotherapy on the mutation frequency of ovarian cancer cells at the HPRT locus. *Anticancer Res.*, 25(3B):2113–7.

[99] Gianni, L., Eiermann, W., Semiglazov, V., Manikhas, A., Lluch, A., Tjulandin, S., Zambetti, M., Vazquez, F., Byakhow, M., Lichinitser, M., Climent, M. A., Ciruelos, E., Ojeda, B., Mansutti, M., Bozhok, A., Baronio, R., Feyereislova, A., Barton, C., Valagussa, P., and Baselga, J. (2010). Neoadjuvant chemotherapy with trastuzumab followed by adjuvant trastuzumab versus neoadjuvant chemotherapy alone, in patients with HER2-positive locally advanced breast cancer (the NOAH trial): a randomised controlled superiority trial with a parallel HER. *Lancet*, 375(9712):377–84.

[100] Gianni, L., Pienkowski, T., Im, Y.-H., Roman, L., Tseng, L.-M., Liu, M.-C., Lluch, A., Staroslawska, E., de la Haba-Rodriguez, J., Im, S.-A., Pedrini, J. L., Poirier, B., Morandi, P., Semiglazov, V., Srimuninnimit, V., Bianchi, G., Szado, T., Ratnayake, J., Ross, G., and Valagussa, P. (2012). Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. *Lancet. Oncol.*, 13(1):25–32.

[101] Gianni, L., Pienkowski, T., Im, Y.-H., Tseng, L.-M., Liu, M.-C., Lluch, A., Starosławska, E., de la Haba-Rodriguez, J., Im, S.-A., Pedrini, J. L., Poirier, B., Morandi, P., Semiglazov, V., Srimuninnimit, V., Bianchi, G. V., Magazzù, D., McNally, V., Douthwaite, H., Ross, G., and Valagussa, P. (2016). 5-year analysis of neoadjuvant pertuzumab and trastuzumab in patients with locally advanced, inflammatory, or early-stage HER2-positive breast cancer (NeoSphere): a multicentre, open-label, phase 2 randomised trial. *Lancet. Oncol.*, 17(6):791–800.

[102] Gianni, L., Zambetti, M., Clark, K., Baker, J., Cronin, M., Wu, J., Mariani, G., Rodriguez, J., Carcangiu, M., Watson, D., Valagussa, P., Rouzier, R., Symmans, W. F., Ross, J. S., Hortobagyi, G. N., Pusztai, L., and Shak, S. (2005). Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. *J. Clin. Oncol.*, 23(29):7265–77.

[103] Glück, S., de Snoo, F., Peeters, J., Stork-Sloots, L., and Somlo, G. (2013). Molecular subtyping of early-stage breast cancer identifies a group of patients who do not benefit from neoadjuvant chemotherapy. *Breast Cancer Res. Treat.*, 139(3):759–67.

[104] Gonzalez-Angulo, A. M., Iwamoto, T., Liu, S., Chen, H., Do, K.-A., Hortobagyi, G. N., Mills, G. B., Meric-Bernstam, F., Symmans, W. F., and Pusztai, L. (2012). Gene expression, molecular class changes, and pathway analysis after neoadjuvant systemic therapy for breast cancer. *Clin. Cancer Res.*, 18(4):1109–19.

[105] Gormally, E., Caboux, E., Vineis, P., and Hainaut, P. (2007). Circulating free DNA in plasma or serum as biomarker of carcinogenesis: practical aspects and biological significance. *Mutat. Res.*, 635(2-3):105–17.

[106] Goswami, S., Philippar, U., Sun, D., Patsialou, A., Avraham, J., Wang, W., Di Modugno, F., Nistico, P., Gertler, F. B., and Condeelis, J. S. (2009). Identification of invasion specific splice variants of the cytoskeletal protein Mena present in mammary tumor cells during invasion in vivo. *Clin. Exp. Metastasis*, 26(2):153–159.

[107] Guarneri, V., Frassoldati, A., Bottini, A., Cagossi, K., Bisagni, G., Sarti, S., Ravaioli, A., Cavanna, L., Giardina, G., Musolino, A., Untch, M., Orlando, L., Artioli, F., Boni, C., Generali, D. G., Serra, P., Bagnalasta, M., Marini, L., Piacentini, F., D'Amico, R., and Conte, P. (2012). Preoperative chemotherapy plus trastuzumab, lapatinib, or both in human epidermal growth factor receptor 2-positive operable breast cancer: results of the randomized phase II CHER-LOB study. *J. Clin. Oncol.*, 30(16):1989–95.

[108] Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., Bertrand, F., Bonnet, F., Martin, A. L., Feugeas, J. P., Bièche, I., Lehmann-Che, J., Lidereau, R., Birnbaum, D., Bertucci, F., de Thé, H., and Theillet, C. (2012). A refined molecular taxonomy of breast cancer. *Oncogene*, 31(9):1196–206.

[109] Guenin, M. A. (2000). Clip placement during sonographically guided large-core breast biopsy for mammographic-sonographic correlation. *AJR. Am. J. Roentgenol.*, 175(4):1053–5.

[110] Hage-Sleiman, R., Herveau, S., Matera, E.-L., Laurier, J.-F., and Dumontet, C. (2010). Tubulin binding cofactor C (TBCC) suppresses tumor growth and enhances chemosensitivity in human breast cancer cells. *BMC Cancer*, 10:135.

[111] Hahnen, E., Lederer, B., Hauke, J., Loibl, S., Kröber, S., Schneeweiss, A., Denkert, C., Fasching, P. A., Blohmer, J. U., Jackisch, C., Paepke, S., Gerber, B., Kümmel, S., Schem, C., Neidhardt, G., Huober, J., Rhiem, K., Costa, S., Altmüller, J., Hanusch, C., Thiele, H., Müller, V., Nürnberg, P., Karn, T., Nekljudova, V., Untch, M., von Minckwitz, G., and Schmutzler, R. K. (2017). Germline Mutation Status, Pathological Complete Response, and Disease-Free Survival in Triple-Negative Breast Cancer: Secondary Analysis of the GeparSixto Randomized Clinical Trial. *JAMA Oncol.*, 3(10):1378–1385.

[112] Halsted, W. S. (1907). The Results of Radical Operations for the Cure of Carcinoma of the Breast. *Ann. Surg.*, 46(1):1–19.

[113] Hannemann, J., Oosterkamp, H. M., Bosch, C. A. J., Velds, A., Wessels, L. F. A., Loo, C., Rutgers, E. J., Rodenhuis, S., and van de Vijver, M. J. (2005). Changes in gene expression associated with response to neoadjuvant chemotherapy in breast cancer. *J. Clin. Oncol.*, 23(15):3331–42.

[114] Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14:7.

[115] Hao, B., Oehlmann, S., Sowa, M. E., Harper, J. W., and Pavletich, N. P. (2007). Structure of a Fbw7-Skp1-cyclin E complex: multisite-phosphorylated substrate recognition by SCF ubiquitin ligases. *Mol. Cell*, 26(1):131–43.

[116] Harney, A. S., Arwert, E. N., Entenberg, D., Wang, Y., Guo, P., Qian, B.-Z., Oktay, M. H., Pollard, J. W., Jones, J. G., and Condeelis, J. S. (2015). Real-Time Imaging Reveals Local, Transient Vascular Permeability, and Tumor Cell Intravasation Stimulated by TIE2hi Macrophage-Derived VEGFA. *Cancer Discov.*, 5(9):932–43.

[117] Hatzis, C., Pusztai, L., Valero, V., Booser, D. J., Esserman, L., Lluch, A., Vidaurre, T., Holmes, F., Souchon, E., Wang, H., Martin, M., Cotrina, J., Gomez, H., Hubbard, R., Chacón, J. I., Ferrer-Lozano, J., Dyer, R., Buxton, M., Gong, Y., Wu, Y., Ibrahim, N., Andreopoulou, E., Ueno, N. T., Hunt, K., Yang, W., Nazario, A., DeMichele, A., O'Shaughnessy, J., Hortobagyi, G. N., and Symmans, W. F. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA*, 305(18):1873–81.

[118] Heidary, M., Auer, M., Ulz, P., Heitzer, E., Petru, E., Gasch, C., Riethdorf, S., Mauermann, O., Lafer, I., Pristauz, G., Lax, S., Pantel, K., Geigl, J. B., and Speicher, M. R. (2014). The dynamic range of circulating tumor DNA in metastatic breast cancer. *Breast Cancer Res.*, 16(4):421.

[119] Hellmann, M. D., Ciuleanu, T.-E., Pluzanski, A., Lee, J. S., Otterson, G. A., Audigier-Valette, C., Minenza, E., Linardou, H., Burgers, S., Salman, P., Borghaei, H., Ramalingam, S. S., Brahmer, J., Reck, M., O'Byrne, K. J., Geese, W. J., Green, G., Chang, H., Szustakowski, J., Bhagavatheeswaran, P., Healey, D., Fu, Y., Nathan, F., and Paz-Ares, L. (2018). Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *N. Engl. J. Med.*, 378(22):2093–2104.

[120] Henderson, I. C., Berry, D. A., Demetri, G. D., Cirrincione, C. T., Goldstein, L. J., Martino, S., Ingle, J. N., Cooper, M. R., Hayes, D. F., Tkaczuk, K. H., Fleming, G., Holland, J. F., Duggan, D. B., Carpenter, J. T., Frei, E., Schilsky, R. L., Wood, W. C., Muss, H. B., and Norton, L. (2003). Improved outcomes from adding sequential Paclitaxel but not from escalating Doxorubicin dose in an adjuvant chemotherapy regimen for patients with node-positive primary breast cancer. *J. Clin. Oncol.*, 21(6):976–83.

[121] Hesler, R. A., Huang, J. J., Starr, M. D., Treboschi, V. M., Bernanke, A. G., Nixon, A. B., McCall, S. J., White, R. R., and Blobe, G. C. (2016). TGF-β-induced stromal CYR61 promotes resistance to gemcitabine in pancreatic ductal adenocarcinoma through downregulation of the nucleoside transporters hENT1 and hCNT3. *Carcinogenesis*, 37(11):1041–1051.

[122] Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, T., Gómez, H. L., Hortobagyi, G. N., and Pusztai, L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J. Clin. Oncol.*, 24(26):4236–44.

[123] Highnam, G., Wang, J. J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N., and Mittelman, D. (2015). An analytical framework for optimizing variant discovery from personal genomes. *Nat. Commun.*, 6:6275.

[124] Hiller, L., Dunn, J. A., Loi, S., Vallier, A.-L., Howe, D. L., Cameron, D. A., Miles, D., Wardley, A. M., and Earl, H. M. (2018). Adjuvant trastuzumab duration trials in HER2 positive breast cancer - what results would be practice-changing? Persephone investigator questionnaire prior to primary endpoint results. *BMC Cancer*, 18(1):391.

[125] Hodi, F. S., O'Day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C., Akerley, W., van den Eertwegh, A. J. M., Lutzky, J., Lorigan, P., Vaubel, J. M., Linette, G. P., Hogg, D., Ottensmeier, C. H., Lebbé, C., Peschel, C., Quirt, I., Clark, J. I., Wolchok, J. D., Weber, J. S., Tian, J., Yellin, M. J., Nichol, G. M., Hoos, A., and Urba, W. J. (2010). Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.*, 363(8):711–23.

[126] Honkoop, A. H., Pinedo, H. M., De Jong, J. S., Verheul, H. M., Linn, S. C., Hoekman, K., Wagstaff, J., and van Diest, P. J. (1997). Effects of chemotherapy on pathologic and biologic characteristics of locally advanced breast cancer. *Am. J. Clin. Pathol.*, 107(2):211–8.

[127] Honkoop, A. H., van Diest, P. J., de Jong, J. S., Linn, S. C., Giaccone, G., Hoekman, K., Wagstaff, J., and Pinedo, H. M. (1998). Prognostic role of clinical, pathological and biological characteristics in patients with locally advanced breast cancer. *Br. J. Cancer*, 77(4):621–6.

[128] Horak, C. E., Pusztai, L., Xing, G., Trifan, O. C., Saura, C., Tseng, L.-M., Chan, S., Welcher, R., and Liu, D. (2013). Biomarker analysis of neoadjuvant doxorubicin/cyclophosphamide followed by ixabepilone or Paclitaxel in early-stage breast cancer. *Clin. Cancer Res.*, 19(6):1587–95.

[129] Hortobagyi, G. N. (1990). Comprehensive management of locally advanced breast cancer. *Cancer*, 66(6 Suppl):1387–91.

[130] Hortobagyi, G. N., Ames, F. C., Buzdar, A. U., Kau, S. W., McNeese, M. D., Paulus, D., Hug, V., Holmes, F. A., Romsdahl, M. M., and Fraschini, G. (1988). Management of stage III primary breast cancer with primary chemotherapy, surgery, and radiation therapy. *Cancer*, 62(12):2507–16.

[131] Houssami, N., Macaskill, P., von Minckwitz, G., Marinovich, M. L., and Mamounas, E. (2012). Meta-analysis of the association of breast cancer subtype and pathologic complete response to neoadjuvant chemotherapy. *Eur. J. Cancer*, 48(18):3342–54.

[132] Howson, J. M. M., Walker, N. M., Clayton, D., Todd, J. A., and Type 1 Diabetes Genetics Consortium (2009). Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. *Diabetes. Obes. Metab.*, 11 Suppl 1:31–45.

[133] Huang, D., Su, S., Cui, X., Shen, X., Zeng, Y., Wu, W., Chen, J., Chen, F., He, C., Liu, J., Huang, W., Liu, Q., Su, F., Song, E., and Ouyang, N. (2014). Nerve fibers in breast cancer tissues indicate aggressive tumor progression. *Medicine (Baltimore).*, 93(27):e172.

[134] Huisman, M. T., Chhatta, A. A., van Tellingen, O., Beijnen, J. H., and Schinkel, A. H. (2005). MRP2 (ABCC2) transports taxanes and confers paclitaxel resistance and both processes are stimulated by probenecid. *Int. J. cancer*, 116(5):824–9.

[135] Hundal, J., Carreno, B. M., Petti, A. A., Linette, G. P., Griffith, O. L., Mardis, E. R., and Griffith, M. (2016). pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.*, 8(1):11.

[136] Huober, J., von Minckwitz, G., Denkert, C., Tesch, H., Weiss, E., Zahm, D. M., Belau, A., Khandan, F., Hauschild, M., Thomssen, C., Högel, B., Darb-Esfahani, S., Mehta, K., and Loibl, S. (2010). Effect of neoadjuvant anthracycline-taxane-based chemotherapy in different biological breast cancer phenotypes: overall results from the GeparTrio study. *Breast Cancer Res. Treat.*, 124(1):133–40.

[137] Hurvitz, S. A., Martin, M., Symmans, W. F., Jung, K. H., Huang, C.-S., Thompson, A. M., Harbeck, N., Valero, V., Stroyakovskiy, D., Wildiers, H., Campone, M., Boileau, J.-F., Beckmann, M. W., Afenjar, K., Fresco, R., Helms, H.-J., Xu, J., Lin, Y. G., Sparano, J., and Slamon, D. (2018). Neoadjuvant trastuzumab, pertuzumab, and chemotherapy versus trastuzumab emtansine plus pertuzumab in patients with HER2-positive breast cancer (KRISTINE): a randomised, open-label, multicentre, phase 3 trial. *Lancet. Oncol.*, 19(1):115–126.

[138] Ignatiadis, M. and Dawson, S.-J. (2014). Circulating Tumor Cells and Circulating tumor DNA for precision medicine: Dream or reality? *Ann. Oncol.*, (October):2304–2313.

[139] Ignatiadis, M., Singhal, S. K., Desmedt, C., Haibe-Kains, B., Criscitiello, C., Andre, F., Loi, S., Piccart, M., Michiels, S., and Sotiriou, C. (2012). Gene modules and response to neoadjuvant chemotherapy in breast cancer subtypes: a pooled analysis. *J. Clin. Oncol.*, 30(16):1996–2004.

[140] Issa-Nummer, Y., Darb-Esfahani, S., Loibl, S., Kunz, G., Nekljudova, V., Schrader, I., Sinn, B. V., Ulmer, H.-U., Kronenwett, R., Just, M., Kühn, T., Diebold, K., Untch, M., Holms, F., Blohmer, J.-U., Habeck, J.-O., Dietel, M., Overkamp, F., Krabisch, P., von Minckwitz, G., and Denkert, C. (2013). Prospective validation of immunological infiltrate for prediction of response to neoadjuvant chemotherapy in HER2-negative breast cancer–a substudy of the neoadjuvant GeparQuinto trial. *PLoS One*, 8(12):e79775.

[141] Ito, T., Teo, Y. V., Evans, S. A., Neretti, N., and Sedivy, J. M. (2018). Regulation of Cellular Senescence by Polycomb Chromatin Modifiers through Distinct DNA Damage- and Histone Methylation-Dependent Pathways. *Cell Rep.*, 22(13):3480–3492.

[142] Iwamoto, T., Bianchini, G., Booser, D., Qi, Y., Coutant, C., Shiang, C. Y.-H., Santarpia, L., Matsuoka, J., Hortobagyi, G. N., Symmans, W. F., Holmes, F. A., O'Shaughnessy, J., Hellerstedt, B., Pippen, J., Andre, F., Simon, R., and Pusztai, L. (2011). Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. *J. Natl. Cancer Inst.*, 103(3):264–72.

[143] Jacquillat, C., Weil, M., Baillet, F., Borel, C., Auclerc, G., de Maublanc, M. A., Housset, M., Forget, G., Thill, L., and Soubrane, C. (1990). Results of neoadjuvant chemotherapy and radiation therapy in the breast-conserving treatment of 250 patients with all stages of infiltrative breast cancer. *Cancer*, 66(1):119–29.

[144] Jahr, S., Hentze, H., Englisch, S., and Hardt, D. (2001). DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res.*, pages 1659–1665.

[145] Janeway, C. A. and Medzhitov, R. (2002). Innate immune recognition. *Annu. Rev. Immunol.*, 20:197–216.

[146] Joukov, V., Groen, A. C., Prokhorova, T., Gerson, R., White, E., Rodriguez, A., Walter, J. C., and Livingston, D. M. (2006). The BRCA1/BARD1 heterodimer modulates ran-dependent mitotic spindle assembly. *Cell*, 127(3):539–52.

[147] Juul, N., Szallasi, Z., Eklund, A. C., Li, Q., Burrell, R. A., Gerlinger, M., Valero, V., Andreopoulou, E., Esteva, F. J., Symmans, W. F., Desmedt, C., Haibe-Kains, B., Sotiriou, C., Pusztai, L., and Swanton, C. (2010a). Assessment of an RNA interference screen-derived mitotic and ceramide pathway metagene as a predictor of response to neoadjuvant paclitaxel for primary triple-negative breast cancer: a retrospective analysis of five clinical trials. *Lancet. Oncol.*, 11(4):358–65.

[148] Juul, N., Szallasi, Z., Eklund, A. C., Li, Q., Burrell, R. A., Gerlinger, M., Valero, V., Andreopoulou, E., Esteva, F. J., Symmans, W. F., Desmedt, C., Haibe-Kains, B., Sotiriou, C., Pusztai, L., and Swanton, C. (2010b). Assessment of an RNA interference screen-derived mitotic and ceramide pathway metagene as a predictor of response to neoadjuvant paclitaxel for primary triple-negative breast cancer: a retrospective analysis of five clinical trials. *Lancet. Oncol.*, 11(4):358–65.

[149] Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., and Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, 39(Database issue):D712–7.

[150] Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.*, 37(Database issue):D623–8.

[151] Karagiannis, G. S., Goswami, S., Jones, J. G., Oktay, M. H., and Condeelis, J. S. (2016). Signatures of breast cancer metastasis at a glance. *J. Cell Sci.*, 129(9):1751–8.

[152] Karagiannis, G. S., Pastoriza, J. M., Wang, Y., Harney, A. S., Entenberg, D., Pignatelli, J., Sharma, V. P., Xue, E. A., Cheng, E., D'Alfonso, T. M., Jones, J. G., Anampa, J., Rohan, T. E., Sparano, J. A., Condeelis, J. S., and Oktay, M. H. (2017). Neoadjuvant chemotherapy induces breast cancer metastasis through a TMEM-mediated mechanism. *Sci. Transl. Med.*, 9(397).

[153] Kaufmann, M., Hortobagyi, G. N., Goldhirsch, A., Scholl, S., Makris, A., Valagussa, P., Blohmer, J.-U., Eiermann, W., Jackesz, R., Jonat, W., Lebeau, A., Loibl, S., Miller, W., Seeber, S., Semiglazov, V., Smith, R., Souchon, R., Stearns, V., Untch, M., and von Minckwitz, G. (2006). Recommendations from an international expert panel on the use of neoadjuvant (primary) systemic treatment of operable breast cancer: an update. *J. Clin. Oncol.*, 24(12):1940–9.

[154] Kawasaki, T. and Kawai, T. (2014). Toll-like receptor signaling pathways. *Front. Immunol.*, 5:461.

[155] Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., and Navin, N. E. (2018). Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell*, 173(4):879–893.e13.

[156] Kim, J.-Y., Park, D., Son, D.-S., Nam, S. J., Kim, S. W., Jung, H. H., Kim, Y. J., Park, G., Park, W.-Y., Lee, J. E., and Park, Y. H. (2017a). Circulating tumor DNA shows variable clonal response of breast cancer during neoadjuvant chemotherapy. *Oncotarget*, 8(49):86423–86434.

[157] Kim, M., Jung, J.-Y., Choi, S., Lee, H., Morales, L. D., Koh, J.-T., Kim, S. H., Choi, Y.-D., Choi, C., Slaga, T. J., Kim, W. J., and Kim, D. J. (2017b). GFRA1 promotes cisplatin-induced chemoresistance in osteosarcoma by inducing autophagy. *Autophagy*, 13(1):149–168.

[158] Klevebring, D., Neiman, M., Sundling, S., Eriksson, L., Darai Ramqvist, E., Celebioglu, F., Czene, K., Hall, P., Egevad, L., Grönberg, H., and Lindberg, J. (2014). Evaluation of exome sequencing to estimate tumor burden in plasma. *PLoS One*, 9(8):e104417.

[159] Kortylewski, M., Xin, H., Kujawski, M., Lee, H., Liu, Y., Harris, T., Drake, C., Pardoll, D., and Yu, H. (2009). Regulation of the IL-23 and IL-12 balance by Stat3 signaling in the tumor microenvironment. *Cancer Cell*, 15(2):114–23.

[160] Kuerer, H. M., Newman, L. A., Smith, T. L., Ames, F. C., Hunt, K. K., Dhingra, K., Theriault, R. L., Singh, G., Binkley, S. M., Sneige, N., Buchholz, T. A., Ross, M. I., McNeese, M. D., Buzdar, A. U., Hortobagyi, G. N., and Singletary, S. E. (1999). Clinical course of breast cancer patients

with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. *J. Clin. Oncol.*, 17(2):460–9.

[161] Kuhajda, F. P. (2006). Fatty acid synthase and cancer: new application of an old pathway. *Cancer Res.*, 66(12):5977–80.

[162] Kumari, N., Dwarakanath, B. S., Das, A., and Bhatt, A. N. (2016). Role of interleukin-6 in cancer progression and therapeutic resistance. *Tumour Biol.*, 37(9):11553–11572.

[163] Kümmel, S., Holtschmidt, J., and Loibl, S. (2014). Surgical treatment of primary breast cancer in the neoadjuvant setting. *Br. J. Surg.*, 101(8):912–24.

[164] Kurtova, A. V., Xiao, J., Mo, Q., Pazhanisamy, S., Krasnow, R., Lerner, S. P., Chen, F., Roh, T. T., Lay, E., Ho, P. L., and Chan, K. S. (2015). Blocking PGE2-induced tumour repopulation abrogates bladder cancer chemoresistance. *Nature*, 517(7533):209–13.

[165] Lauffenburger, D. A. and Horwitz, A. F. (1996). Cell migration: a physically integrated molecular process. *Cell*, 84(3):359–69.

[166] Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, 15(2):R29.

[167] Law, R. H. P., Lukoyanova, N., Voskoboinik, I., Caradoc-Davies, T. T., Baran, K., Dunstone, M. A., D'Angelo, M. E., Orlova, E. V., Coulibaly, F., Verschoor, S., Browne, K. A., Ciccone, A., Kuiper, M. J., Bird, P. I., Trapani, J. A., Saibil, H. R., and Whisstock, J. C. (2010). The structural basis for membrane binding and pore formation by lymphocyte perforin. *Nature*, 468(7322):447–51.

[168] Leslie, A., Matthews, P. C., Listgarten, J., Carlson, J. M., Kadie, C., Ndung'u, T., Brander, C., Coovadia, H., Walker, B. D., Heckerman, D., and Goulder, P. J. R. (2010). Additive contribution of HLA class I alleles in the immune control of HIV-1 infection. *J. Virol.*, 84(19):9879–88.

[169] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500.

[170] Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20):2843–51.

[171] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, 1(6):417–425.

[172] Liedtke, C., Hatzis, C., Symmans, W. F., Desmedt, C., Haibe-Kains, B., Valero, V., Kuerer, H., Hortobagyi, G. N., Piccart-Gebhart, M., Sotiriou, C., and Pusztai, L. (2009). Genomic grade index is associated with response to chemotherapy in patients with breast cancer. *J. Clin. Oncol.*, 27(19):3185–91.

[173] Lin, Y.-C., Lee, Y.-C., Li, L.-H., Cheng, C.-J., and Yang, R.-B. (2014). Tumor suppressor SCUBE2 inhibits breast-cancer cell migration and invasion through the reversal of epithelial-mesenchymal transition. *J. Cell Sci.*, 127(Pt 1):85–100.

[174] Liu, Q., Sung, A. H., Chen, Z., Liu, J., Huang, X., and Deng, Y. (2009). Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *PLoS One*, 4(12):e8250.

[175] Loibl, S., Majewski, I., Guarneri, V., Nekljudova, V., Holmes, E., Bria, E., Denkert, C., Schem, C., Sotiriou, C., Loi, S., Untch, M., Conte, P., Bernards, R., Piccart, M., von Minckwitz, G., and Baselga, J. (2016). PIK3CA mutations are associated with reduced pathological complete response rates in primary HER2-positive breast cancer: pooled analysis of 967 patients from five prospective trials investigating lapatinib and trastuzumab. *Ann. Oncol.*, 27(8):1519–25.

[176] Loibl, S., O'Shaughnessy, J., Untch, M., Sikov, W. M., Rugo, H. S., McKee, M. D., Huober, J., Golshan, M., von Minckwitz, G., Maag, D., Sullivan, D., Wolmark, N., McIntyre, K., Ponce Lorenzo, J. J., Metzger Filho, O., Rastogi, P., Symmans, W. F., Liu, X., and Geyer, C. E. (2018). Addition of the PARP inhibitor veliparib plus carboplatin or carboplatin alone to standard neoadjuvant chemotherapy in triple-negative breast cancer (BrighTNess): a randomised, phase 3 trial. *Lancet. Oncol.*, 19(4):497–509.

[177] Loibl, S., Volz, C., Mau, C., Blohmer, J.-U., Costa, S. D., Eidtmann, H., Fasching, P. A., Gerber, B., Hanusch, C., Jackisch, C., Kümmel, S., Huober, J., Denkert, C., Hilfrich, J., Konecny, G. E., Fett, W., Stickeler, E., Harbeck, N., Mehta, K. M., Nekljudova, V., von Minckwitz, G., and Untch, M. (2014). Response and prognosis after neoadjuvant chemotherapy in 1,051 patients with infiltrating lobular breast carcinoma. *Breast Cancer Res. Treat.*, 144(1):153–62.

[178] Lounnas, N., Rosilio, C., Nebout, M., Mary, D., Griessinger, E., Neffati, Z., Chiche, J., Spits, H., Hagenbeek, T. J., Asnafi, V., Poulsen, S.-A., Supuran, C. T., Peyron, J.-F., and Imbert, V. (2013). Pharmacological inhibition of carbonic anhydrase XII interferes with cell proliferation and induces cell apoptosis in T-cell lymphomas. *Cancer Lett.*, 333(1):76–88.

[179] Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O., and Nielsen, M. (2008). NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.*, 36(Web Server issue):W509–12.

[180] Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10:161.

[181] Ma, C. X., Gao, F., Luo, J., Northfelt, D. W., Goetz, M., Forero, A., Hoog, J., Naughton, M., Ademuyiwa, F., Suresh, R., Anderson, K. S., Margenthaler, J., Aft, R., Hobday, T., Moynihan, T., Gillanders, W., Cyr, A., Eberlein, T. J., Hieken, T., Krontiras, H., Guo, Z., Lee, M. V., Spies, N. C., Skidmore, Z. L., Griffith, O. L., Griffith, M., Thomas, S., Bumb, C., Vij, K., Bartlett, C. H., Koehler, M., Al-Kateb, H., Sanati, S., and Ellis, M. J. (2017). NeoPalAna: Neoadjuvant Palbociclib, a Cyclin-Dependent Kinase 4/6 Inhibitor, and Anastrozole for Clinical Stage 2 or 3 Estrogen Receptor-Positive Breast Cancer. *Clin. Cancer Res.*, 23(15):4055–4065.

[182] Magbanua, M. J. M., Wolf, D. M., Yau, C., Davis, S. E., Crothers, J., Au, A., Haqq, C. M., Livasy, C., Rugo, H. S., I-SPY 1 TRIAL Investigators, Esserman, L., Park, J. W., and van 't Veer, L. J. (2015). Serial expression analysis of breast tumors during neoadjuvant chemotherapy reveals changes in cell cycle and immune pathways associated with recurrence and response. *Breast Cancer Res.*, 17:73.

[183] Makris, A., Powles, T. J., Ashley, S. E., Chang, J., Hickish, T., Tidy, V. A., Nash, A. G., and Ford, H. T. (1998). A reduction in the requirements for mastectomy in a randomized trial of neoadjuvant chemoendocrine therapy in primary breast cancer. *Ann. Oncol.*, 9(11):1179–84.

[184] Mamounas, E. P., Bryant, J., Lembersky, B., Fehrenbacher, L., Sedlacek, S. M., Fisher, B., Wickerham, D. L., Yothers, G., Soran, A., and Wolmark, N. (2005). Paclitaxel after doxorubicin plus cyclophosphamide as adjuvant chemotherapy for node-positive breast cancer: results from NSABP B-28. *J. Clin. Oncol.*, 23(16):3686–96.

[185] Mandel, P. and Metais, P. (1948). Les acides nucléiques du plasma sanguin chez l'homme. *C. R. Seances Soc. Biol. Fil.*, 142(3-4):241–3.

[186] Marone, R., Hess, D., Dankort, D., Muller, W. J., Hynes, N. E., and Badache, A. (2004). Memo mediates ErbB2-driven cell motility. *Nat. Cell Biol.*, 6(6):515–22.

[187] Martin, M., Villar, A., Sole-Calvo, A., Gonzalez, R., Massuti, B., Lizon, J., Camps, C., Carrato, A., Casado, A., Candel, M. T., Albanell, J., Aranda, J., Munarriz, B., Campbell, J., Diaz-Rubio, E., and GEICAM Group (Spanish Breast Cancer Research Group), S. (2003). Doxorubicin in combination with fluorouracil and cyclophosphamide (i.v. FAC regimen, day 1, 21) versus methotrexate in combination with fluorouracil and cyclophosphamide (i.v. CMF regimen, day 1, 21) as adjuvant chemotherapy for operable breast cancer: a. *Ann. Oncol.*, 14(6):833–42.

[188] Mauriac, L., Durand, M., Avril, A., and Dilhuydy, J. M. (1991). Effects of primary chemotherapy in conservative treatment of breast cancer patients with operable tumors larger than 3 cm. Results of a randomized trial in a single centre. *Ann. Oncol.*, 2(5):347–54.

[189] Maxwell, C. A., Keats, J. J., Crainie, M., Sun, X., Yen, T., Shibuya, E., Hendzel, M., Chan, G., and Pilarski, L. M. (2003). RHAMM is a centrosomal protein that interacts with dynein and maintains spindle pole stability. *Mol. Biol. Cell*, 14(6):2262–76.

[190] McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, 40(10):4288–97.

[191] McGranahan, N., Favero, F., de Bruin, E. C., Birkbak, N. J., Szallasi, Z., and Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.*, 7(283):283ra54.

[192] McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., Birkbak, N. J., Veeriah, S., Van Loo, P., Herrero, J., Swanton, C., and TRACERx Consortium (2017). Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*, 171(6):1259–1271.e11.

[193] McGranahan, N. and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, 168(4):613–628.

[194] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.*, 17(1):122.

[195] Menendez, J. A. and Lupu, R. (2007). Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat. Rev. Cancer*, 7(10):763–77.

[196] Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, 12(4):R41.

[197] Miller, C. A., Gindin, Y., Lu, C., Griffith, O. L., Griffith, M., Shen, D., Hoog, J., Li, T., Larson, D. E., Watson, M., Davies, S. R., Hunt, K., Suman, V. J., Snider, J., Walsh, T., Colditz, G. A., DeSchryver, K., Wilson, R. K., Mardis, E. R., and Ellis, M. J. (2016). Aromatase inhibition remodels the clonal architecture of estrogen-receptor-positive breast cancers. *Nat. Commun.*, 7:12498.

[198] Mills, R. E., Pittard, W. S., Mullaney, J. M., Farooq, U., Creasy, T. H., Mahurkar, A. A., Kemeza, D. M., Strassler, D. S., Ponting, C. P., Webber, C., and Devine, S. E. (2011). Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, 21(6):830–9.

[199] Modlich, O., Prisack, H.-B., Munnes, M., Audretsch, W., and Bojar, H. (2004). Immediate gene expression changes after the first course of neoadjuvant chemotherapy in patients with primary breast cancer disease. *Clin. Cancer Res.*, 10(19):6418–31.

[200] Morandi, A., Martin, L.-A., Gao, Q., Pancholi, S., Mackay, A., Robertson, D., Zvelebil, M., Dowsett, M., Plaza-Menacho, I., and Isacke, C. M. (2013). GDNF-RET signaling in ER-positive breast cancers is a key determinant of response and resistance to aromatase inhibitors. *Cancer Res.*, 73(12):3783–95.

[201] Morishima, S., Kashiwase, K., Matsuo, K., Azuma, F., Yabe, T., Sato-Otsubo, A., Ogawa, S., Shiina, T., Satake, M., Saji, H., Kato, S., Kodera, Y., Sasazuki, T., Morishima, Y., and Japan Marrow Donor Program (2016). High-risk HLA alleles for severe acute graft-versus-host disease and mortality in unrelated donor bone marrow transplantation. *Haematologica*, 101(4):491–8.

[202] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–8.

[203] Mouliere, F. and Rosenfeld, N. (2015). Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc. Natl. Acad. Sci.*, 112(11):3178–3179.

[204] Muhlethaler-Mottet, A., Di Berardino, W., Otten, L. A., and Mach, B. (1998). Activation of the MHC class II transactivator CIITA by interferon-gamma requires cooperative interaction between Stat1 and USF-1. *Immunity*, 8(2):157–66.

[205] Murtaza, M., Dawson, S.-J., Pogrebniak, K., Rueda, O. M., Provenzano, E., Grant, J., Chin, S.-F., Tsui, D. W. Y., Marass, F., Gale, D., Ali, H. R., Shah, P., Contente-Cuomo, T., Farahani, H., Shumansky, K., Kingsbury, Z., Humphray, S., Bentley, D., Shah, S. P., Wallis, M., Rosenfeld, N., and Caldas, C. (2015). Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat. Commun.*, 6:8760.

[206] Murugaesu, N., Wilson, G. A., Birkbak, N. J., Watkins, T., McGranahan, N., Kumar, S., Abbassi-Ghadi, N., Salm, M., Mitter, R., Horswell, S., Rowan, A., Phillimore, B., Biggs, J., Begum, S., Matthews, N., Hochhauser, D., Hanna, G. B., and Swanton, C. (2015). Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov.*, 5(8):821–831.

[207] Nazarov, V. I., Pogorelyy, M. V., Komech, E. A., Zvyagin, I. V., Bolotin, D. A., Shugay, M., Chudakov, D. M., Lebedev, Y. B., and Mamedov, I. Z. (2015). tcR: an R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics*, 16:175.

[208] Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., Hoang, C. D., Diehn, M., and Alizadeh, A. A. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–7.

[209] NICE (2009). Early and locally advanced breast cancer: Diagnosis and treatment [CG80].

[210] NICE (2018). Early and locally advanced breast cancer: diagnosis and management. NICE guideline [NG101]. Technical report.

[211] Nielsen, M. and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.*, 8(1):33.

[212] NIH (1992). Consensus statement: treatment of early-stage breast cancer. National Institutes of Health Consensus Development Panel. *J. Natl. Cancer Inst. Monogr.*, (11):1–5.

[213] Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K. W., Mudie, L. J., Varela, I., McBride, D. J., Bignell, G. R., Cooke, S. L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P. S., Davies, H. R., Papaemmanuil, E., Stephens, P. J., McLaren, S., Butler, A. P., Teague, J. W., Jönsson, G., Garber, J. E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerød, A., Tutt, A., Martens, J. W. M., Aparicio, S. A. J. R., Borg, Å., Salomon, A. V., Thomas, G., Børresen-Dale, A.-L., Richardson, A. L., Neuberger, M. S., Futreal, P. A., Campbell, P. J., Stratton, M. R., and Breast Cancer Working Group of the International Cancer Genome Consortium (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–93.

[214] Niu, Z.-H., Wang, Y., Chun, B., Li, C.-X., and Wu, L. (2013). Secretory clusterin (sCLU) overexpression is associated with resistance to preoperative neoadjuvant chemotherapy in primary breast cancer. *Eur. Rev. Med. Pharmacol. Sci.*, 17(10):1337–44.

[215] Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–8.

[216] O'Connor, P. M., Jackman, J., Bae, I., Myers, T. G., Fan, S., Mutoh, M., Scudiero, D. A., Monks, A., Sausville, E. A., Weinstein, J. N., Friend, S., Fornace, A. J., and Kohn, K. W. (1997). Characterization of the p53 tumor suppressor pathway in cell lines of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Res.*, 57(19):4285–300.

[217] Ogston, K. N., Miller, I. D., Payne, S., Hutcheon, A. W., Sarkar, T. K., Smith, I., Schofield, A., and Heys, S. D. (2003). A new histological grading system to assess response of breast cancers to primary chemotherapy: prognostic significance and survival. *Breast*, 12(5):320–7.

[218] Olshen, V. E. S. and Adam (2018). DNAcopy: DNA copy number data analysis.

[219] Ostrand-Rosenberg, S. and Sinha, P. (2009). Myeloid-derived suppressor cells: linking inflammation and cancer. *J. Immunol.*, 182(8):4499–506.

[220] Park, S., Shimizu, C., Shimoyama, T., Takeda, M., Ando, M., Kohno, T., Katsumata, N., Kang, Y.-K., Nishio, K., and Fujiwara, Y. (2006). Gene expression profiling of ATP-binding cassette (ABC) transporters as a predictor of the pathologic response to neoadjuvant chemotherapy in breast cancer patients. *Breast Cancer Res. Treat.*, 99(1):9–17.

[221] Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, 27(8):1160–7.

[222] Patch, A.-M., Christie, E. L., Etemadmoghadam, D., Garsed, D. W., George, J., Fereday, S., Nones, K., Cowin, P., Alsop, K., Bailey, P. J., Kassahn, K. S., Newell, F., Quinn, M. C. J., Kazakoff, S., Quek, K., Wilhelm-Benartzi, C., Curry, E., Leong, H. S., Australian Ovarian Cancer Study Group, Hamilton, A., Mileshkin, L., Au-Yeung, G., Kennedy, C., Hung, J., Chiew, Y.-E., Harnett, P., Friedlander, M., Quinn, M., Pyman, J., Cordner, S., O'Brien, P., Leditschke, J., Young,

G., Strachan, K., Waring, P., Azar, W., Mitchell, C., Traficante, N., Hendley, J., Thorne, H., Shackleton, M., Miller, D. K., Arnau, G. M., Tothill, R. W., Holloway, T. P., Semple, T., Harliwong, I., Nourse, C., Nourbakhsh, E., Manning, S., Idrisoglu, S., Bruxner, T. J. C., Christ, A. N., Poudel, B., Holmes, O., Anderson, M., Leonard, C., Lonie, A., Hall, N., Wood, S., Taylor, D. F., Xu, Q., Fink, J. L., Waddell, N., Drapkin, R., Stronach, E., Gabra, H., Brown, R., Jewell, A., Nagaraj, S. H., Markham, E., Wilson, P. J., Ellul, J., McNally, O., Doyle, M. A., Vedururu, R., Stewart, C., Lengyel, E., Pearson, J. V., Waddell, N., DeFazio, A., Grimmond, S. M., and Bowtell, D. D. L. (2015). Whole-genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–94.

[223] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419.

[224] Peintinger, F., Sinn, B., Hatzis, C., Albarracin, C., Downs-Kelly, E., Morkowski, J., Gould, R., and Symmans, W. F. (2015). Reproducibility of residual cancer burden for prognostic assessment of breast cancer after neoadjuvant chemotherapy. *Mod. Pathol.*, 28(7):913–20.

[225] Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S.-J., Tsui, D. W. Y., Liu, B., Dawson, S.-J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., McKinney, S., Oloumi, A., Shah, S., Rosenfeld, N., Murphy, L., Bentley, D. R., Ellis, I. O., Purushotham, A., Pinder, S. E., Børresen-Dale, A.-L., Earl, H. M., Pharoah, P. D., Ross, M. T., Aparicio, S., and Caldas, C. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.*, 7:11479.

[226] Piskounova, E., Polytarchou, C., Thornton, J. E., LaPierre, R. J., Pothoulakis, C., Hagan, J. P., Iliopoulos, D., and Gregory, R. I. (2011). Lin28A and Lin28B inhibit let-7 microRNA biogenesis by distinct mechanisms. *Cell*, 147(5):1066–79.

[227] Poole, C. J., Earl, H. M., Hiller, L., Dunn, J. A., Bathers, S., Grieve, R. J., Spooner, D. A., Agrawal, R. K., Fernando, I. N., Brunt, A. M., O'Reilly, S. M., Crawford, S. M., Rea, D. W., Simmonds, P., Mansi, J. L., Stanley, A., Harvey, P., McAdam, K., Foster, L., Leonard, R. C. F., Twelves, C. J., and NEAT Investigators and the SCTBG (2006). Epirubicin and cyclophosphamide, methotrexate, and fluorouracil as adjuvant therapy for early breast cancer. *N. Engl. J. Med.*, 355(18):1851–62.

[228] Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.*, 16:91.

[229] Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W. M., Rual, J.-F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Solé, X., Hernández, P., Lázaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat. Genet.*, 39(11):1338–49.

[230] Pusztai, L., Symmans, F. W., and Hortobagyi, G. N. (2005). Development of pharmacogenomic markers to select preoperative chemotherapy for breast cancer. *Breast Cancer*, 12(2):73–85.

[231] Qi, X.-T., Li, Y.-L., Zhang, Y.-Q., Xu, T., Lu, B., Fang, L., Gao, J.-Q., Yu, L.-S., Zhu, D.-F., Yang, B., He, Q.-J., and Ying, M.-D. (2018). KLF4 functions as an oncogene in promoting cancer stem cell-like characteristics in osteosarcoma cells. *Acta Pharmacol. Sin.*

[232] Ragaz, J., Baird, R., Rebbeck, P., Coldman, A., and Goldie, J. (1985). Neoadjuvant-preoperative-chemotherapy for breast cancer–preliminary report of the Vancouver trial. *Prog. Clin. Biol. Res.*, 201:77–87.

[233] Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., Hodi, F. S., Martín-Algarra, S., Mandal, R., Sharfman, W. H., Bhatia, S., Hwu, W.-J., Gajewski, T. F., Slingluff, C. L., Chowell, D., Kendall, S. M., Chang, H., Shah, R., Kuo, F., Morris, L. G. T., Sidhom, J.-W., Schneck, J. P., Horak, C. E., Weinhold, N., and Chan, T. A. (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell*, 171(4):934–949.e15.

[234] Ribas, A. (2015). Adaptive Immune Resistance: How Cancer Protects from Immune Attack. *Cancer Discov.*, 5(9):915–9.

[235] Roberts, S. A. and Gordenin, D. A. (2014). Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer*, 14(12):786–800.

[236] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40.

[237] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25.

[238] Roché, H., Fumoleau, P., Spielmann, M., Canon, J.-L., Delozier, T., Serin, D., Symann, M., Kerbrat, P., Soulié, P., Eichler, F., Viens, P., Monnier, A., Vindevoghel, A., Campone, M., Goudier, M.-J., Bonneterre, J., Ferrero, J.-M., Martin, A.-L., Genève, J., and Asselain, B. (2006). Sequential adjuvant epirubicin-based and docetaxel chemotherapy for node-positive breast cancer patients: the FNCLCC PACS 01 Trial. *J. Clin. Oncol.*, 24(36):5664–71.

[239] Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Geyer, C. E., Davidson, N. E., Tan-Chiu, E., Martino, S., Paik, S., Kaufman, P. A., Swain, S. M., Pisansky, T. M., Fehrenbacher, L., Kutteh, L. A., Vogel, V. G., Visscher, D. W., Yothers, G., Jenkins, R. B., Brown, A. M., Dakhil, S. R., Mamounas, E. P., Lingle, W. L., Klein, P. M., Ingle, J. N., and Wolmark, N. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N. Engl. J. Med.*, 353(16):1673–84.

[240] Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*, 160(1-2):48–61.

[241] Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.*, 17:31.

[242] Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, 11(4):396–8.

[243] Rouzier, R., Perou, C. M., Symmans, W. F., Ibrahim, N., Cristofanilli, M., Anderson, K., Hess, K. R., Stec, J., Ayers, M., Wagner, P., Morandi, P., Fan, C., Rabiul, I., Ross, J. S., Hortobagyi, G. N., and Pusztai, L. (2005). Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin. Cancer Res.*, 11(16):5678–85.

[244] Rubens, R. D., Sexton, S., Tong, D., Winter, P. J., Knight, R. K., and Hayward, J. L. (1980). Combined chemotherapy and radiotherapy for locally advanced breast cancer. *Eur. J. Cancer*, 16(3):351–6.

[245] Rugo, H. S., Olopade, O. I., DeMichele, A., Yau, C., van 't Veer, L. J., Buxton, M. B., Hogarth, M., Hylton, N. M., Paoloni, M., Perlmutter, J., Symmans, W. F., Yee, D., Chien, A. J., Wallace, A. M., Kaplan, H. G., Boughey, J. C., Haddad, T. C., Albain, K. S., Liu, M. C., Isaacs, C., Khan, Q. J., Lang, J. E., Viscusi, R. K., Pusztai, L., Moulder, S. L., Chui, S. Y., Kemmer, K. A., Elias, A. D., Edmiston, K. K., Euhus, D. M., Haley, B. B., Nanda, R., Northfelt, D. W., Tripathy, D., Wood, W. C., Ewing, C., Schwab, R., Lyandres, J., Davis, S. E., Hirst, G. L., Sanil, A., Berry, D. A., Esserman, L. J., and I-SPY 2 Investigators (2016). Adaptive Randomization of Veliparib-Carboplatin Treatment in Breast Cancer. *N. Engl. J. Med.*, 375(1):23–34.

[246] Sahoo, S. and Lester, S. C. (2009). Pathology of breast carcinomas after neoadjuvant chemotherapy: an overview with recommendations on specimen processing and reporting. *Arch. Pathol. Lab. Med.*, 133(4):633–42.

[247] Sarvaiya, P. J., Guo, D., Ulasov, I., Gabikian, P., and Lesniak, M. S. (2013). Chemokines in tumor progression and metastasis. *Oncotarget*, 4(12):2171–85.

[248] Sasaki, K., Zhao, X., Pardee, A. D., Ueda, R., Fujita, M., Sehra, S., Kaplan, M. H., Kane, L. P., Okada, H., and Storkus, W. J. (2008). Stat6 signaling suppresses VLA-4 expression by CD8+ T cells and limits their ability to infiltrate tumor lesions in vivo. *J. Immunol.*, 181(1):104–8.

[249] Sataloff, D. M., Mason, B. A., Prestipino, A. J., Seinige, U. L., Lieber, C. P., and Baloch, Z. (1995). Pathologic response to induction chemotherapy in locally advanced carcinoma of the breast: a determinant of outcome. *J. Am. Coll. Surg.*, 180(3):297–306.

[250] Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J., and Chen, Y.-T. (2002). Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol. Rev.*, 188:22–32.

[251] Scheinin, I., Sie, D., Bengtsson, H., van de Wiel, M. A., Olshen, A. B., van Thuijl, H. F., van Essen, H. F., Eijk, P. P., Rustenburg, F., Meijer, G. A., Reijneveld, J. C., Wesseling, P., Pinkel, D., Albertson, D. G., and Ylstra, B. (2014). DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.*, 24(12):2022–32.

[252] Schneeweiss, A., Chia, S., Hickish, T., Harvey, V., Eniu, A., Hegg, R., Tausch, C., Seo, J. H., Tsai, Y.-F., Ratnayake, J., McNally, V., Ross, G., and Cortés, J. (2013). Pertuzumab plus trastuzumab in combination with standard neoadjuvant anthracycline-containing and anthracycline-free chemotherapy regimens in patients with HER2-positive early breast cancer: a randomized phase II cardiac safety study (TRYPHAENA). *Ann. Oncol.*, 24(9):2278–84.

[253] Scholl, S. M., Asselain, B., Palangie, T., Dorval, T., Jouve, M., Garcia Giralt, E., Vilcoq, J., Durand, J. C., and Pouillart, P. (1991). Neoadjuvant chemotherapy in operable breast cancer. *Eur. J. Cancer*, 27(12):1668–71.

[254] Scholl, S. M., Fourquet, A., Asselain, B., Pierga, J. Y., Vilcoq, J. R., Durand, J. C., Dorval, T., Palangié, T., Jouve, M., and Beuzeboc, P. (1994). Neoadjuvant versus adjuvant chemotherapy in premenopausal patients with tumours considered too large for breast conserving surgery: preliminary results of a randomised trial: S6. *Eur. J. Cancer*, 30A(5):645–52.

[255] Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2017). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.*, 8(1):205–233.

[256] Sensibar, J. A., Sutkowski, D. M., Raffo, A., Buttyan, R., Griswold, M. D., Sylvester, S. R., Kozlowski, J. M., and Lee, C. (1995). Prevention of cell death induced by tumor necrosis factor alpha in LNCaP cells by overexpression of sulfated glycoprotein-2 (clusterin). *Cancer Res.*, 55(11):2431–7.

[257] Sha, M.-Q., Zhao, X.-L., Li, L., Li, L.-H., Li, Y., Dong, T.-G., Niu, W.-X., Jia, L.-J., Shao, R.-G., Zhen, Y.-S., and Wang, Z. (2016). EZH2 mediates lidamycin-induced cellular senescence through regulating p21 expression in human colon cancer cells. *Cell Death Dis.*, 7(11):e2486.

[258] Sharma, P., López-Tarruella, S., Garcia-Saenz, J. A., Khan, Q. J., Gomez, H., Prat, A., Moreno, F., Jerez-Gilarranz, Y., Barnadas, A., Picornell, A. C., Del Monte-Millán, M., Gonzalez-Rivera, M., Massarrah, T., Pelaez-Lorenzo, B., Palomero, M. I., González Del Val, R., Cortes, J., Fuentes Rivera, H., Bretel Morales, D., Marquez-Rodas, I., Perou, C. M., Lehn, C., Wang, Y. Y., Klemp, J. R., Mammen, J. M., Wagner, J., Amin, A., O'Dea, A. P., Heldstab, J., Jensen, R. A., Kimler, B. F., Godwin, A. K., and Martín, M. (2018). Pathological response and survival in triple-negative breast cancer following neoadjuvant carboplatin plus docetaxel. *Clin. Cancer Res.*

[259] Shen, K., Qi, Y., Song, N., Tian, C., Rice, S. D., Gabrin, M. J., Brower, S. L., Symmans, W. F., O'Shaughnessy, J. A., Holmes, F. A., Asmar, L., and Pusztai, L. (2012). Cell line derived multi-gene predictor of pathologic response to neoadjuvant chemotherapy in breast cancer: a validation study on US Oncology 02-103 clinical trial. *BMC Med. Genomics*, 5:51.

[260] Shi, L., Campbell, G., Jones, W. D., Campagne, F., Wen, Z., Walker, S. J., Su, Z., Chu, T.-M., Goodsaid, F. M., Pusztai, L., Shaughnessy, J. D., Oberthuer, A., Thomas, R. S., Paules, R. S., Fielden, M., Barlogie, B., Chen, W., Du, P., Fischer, M., Furlanello, C., Gallas, B. D., Ge, X., Megherbi, D. B., Symmans, W. F., Wang, M. D., Zhang, J., Bitter, H., Brors, B., Bushel, P. R., Bylesjo, M., Chen, M., Cheng, J., Cheng, J., Chou, J., Davison, T. S., Delorenzi, M., Deng, Y., Devanarayan, V., Dix, D. J., Dopazo, J., Dorff, K. C., Elloumi, F., Fan, J., Fan, S., Fan, X., Fang, H., Gonzaludo, N., Hess, K. R., Hong, H., Huan, J., Irizarry, R. A., Judson, R., Juraeva, D., Lababidi, S., Lambert, C. G., Li, L., Li, Y., Li, Z., Lin, S. M., Liu, G., Lobenhofer, E. K., Luo, J., Luo, W., McCall, M. N., Nikolsky, Y., Pennello, G. A., Perkins, R. G., Philip, R., Popovici, V., Price, N. D., Qian, F., Scherer, A., Shi, T., Shi, W., Sung, J., Thierry-Mieg, D., Thierry-Mieg, J., Thodima, V., Trygg, J., Vishnuvajjala, L., Wang, S. J., Wu, J., Wu, Y., Xie, Q., Yousef, W. A., Zhang, L., Zhang, X., Zhong, S., Zhou, Y., Zhu, S., Arasappan, D., Bao, W., Lucas, A. B., Berthold, F., Brennan, R. J., Buness, A., Catalano, J. G., Chang, C., Chen, R., Cheng, Y., Cui, J., Czika, W., Demichelis, F., Deng, X., Dosymbekov, D., Eils, R., Feng, Y., Fostel, J., Fulmer-Smentek, S., Fuscoe, J. C., Gatto, L., Ge, W., Goldstein, D. R., Guo, L., Halbert, D. N., Han, J., Harris, S. C., Hatzis, C., Herman, D., Huang, J., Jensen, R. V., Jiang, R., Johnson, C. D., Jurman, G., Kahlert, Y., Khuder, S. A., Kohl, M., Li, J., Li, L., Li, M., Li, Q.-Z., Li, S., Li, Z., Liu, J., Liu, Y., Liu, Z., Meng, L., Madera, M., Martinez-Murillo, F., Medina, I., Meehan, J., Miclaus, K., Moffitt, R. A., Montaner, D., Mukherjee, P., Mulligan, G. J., Neville, P., Nikolskaya, T., Ning, B., Page, G. P., Parker, J., Parry, R. M., Peng, X., Peterson, R. L., Phan, J. H., Quanz, B., Ren, Y., Riccadonna, S., Roter, A. H., Samuelson, F. W., Schumacher, M. M., Shambaugh, J. D., Shi, Q., Shippy, R., Si, S., Smalter, A., Sotiriou, C., Soukup, M., Staedtler, F., Steiner, G., Stokes, T. H., Sun, Q., Tan, P.-Y., Tang, R., Tezak, Z., Thorn, B., Tsyganova, M., Turpaz, Y., Vega, S. C., Visintainer, R., von Frese, J., Wang, C., Wang, E., Wang, J., Wang, W., Westermann, F., Willey, J. C., Woods, M., Wu, S., Xiao, N., Xu, J., Xu, L., Yang, L., Zeng, X., Zhang, J., Zhang, L., Zhang, M., Zhao, C., Puri, R. K., Scherf, U., Tong, W., Wolfinger, R. D., and MAQC Consortium (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, 28(8):827–38.

[261] Shukla, S. A., Rooney, M. S., Rajasagi, M., Tiao, G., Dixon, P. M., Lawrence, M. S., Stevens, J., Lane, W. J., Dellagatta, J. L., Steelman, S., Sougnez, C., Cibulskis, K., Kiezun, A., Hacohen, N., Brusic, V., Wu, C. J., and Getz, G. (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.*, 33(11):1152–8.

[262] Sidney, J., Schloss, J., Moore, C., Lindvall, M., Wriston, A., Hunt, D. F., Shabanowitz, J., DiLorenzo, T. P., and Sette, A. (2016). Characterization of the peptide binding specificity of the HLA class I alleles B*38:01 and B*39:06. *Immunogenetics*, 68(3):231–6.

[263] Sikov, W. M., Berry, D. A., Perou, C. M., Singh, B., Cirrincione, C. T., Tolaney, S. M., Kuzma, C. S., Pluard, T. J., Somlo, G., Port, E. R., Golshan, M., Bellon, J. R., Collyar, D., Hahn, O. M., Carey, L. A., Hudis, C. A., and Winer, E. P. (2015). Impact of the addition of carboplatin and/or bevacizumab to neoadjuvant once-per-week paclitaxel followed by dose-dense doxorubicin and cyclophosphamide on pathologic complete response rates in stage II to III triple-negative breast cancer: CALGB 40603. *J. Clin. Oncol.*, 33(1):13–21.

[264] Silver, D. P., Richardson, A. L., Eklund, A. C., Wang, Z. C., Szallasi, Z., Li, Q., Juul, N., Leong, C.-O., Calogrias, D., Buraimoh, A., Fatima, A., Gelman, R. S., Ryan, P. D., Tung, N. M., De Nicolo, A., Ganesan, S., Miron, A., Colin, C., Sgroi, D. C., Ellisen, L. W., Winer, E. P., and Garber, J. E. (2010). Efficacy of neoadjuvant Cisplatin in triple-negative breast cancer. *J. Clin. Oncol.*, 28(7):1145–53.

[265] Simigdala, N., Gao, Q., Pancholi, S., Roberg-Larsen, H., Zvelebil, M., Ribas, R., Folkerd, E., Thompson, A., Bhamra, A., Dowsett, M., and Martin, L.-A. (2016). Cholesterol biosynthesis pathway as a novel mechanism of resistance to estrogen deprivation in estrogen receptor-positive breast cancer. *Breast Cancer Res.*, 18(1):58.

[266] Slamon, D., Eiermann, W., Robert, N., Pienkowski, T., Martin, M., Press, M., Mackey, J., Glaspy, J., Chan, A., Pawlicki, M., Pinter, T., Valero, V., Liu, M.-C., Sauter, G., von Minckwitz, G., Visco, F., Bee, V., Buyse, M., Bendahmane, B., Tabah-Fisch, I., Lindsay, M.-A., Riva, A., Crown, J., and Breast Cancer International Research Group (2011). Adjuvant trastuzumab in HER2-positive breast cancer. *N. Engl. J. Med.*, 365(14):1273–83.

[267] Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., and McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235(4785):177–82.

[268] Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.*, 344(11):783–92.

[269] Smith, I., Procter, M., Gelber, R. D., Guillaume, S., Feyereislova, A., Dowsett, M., Goldhirsch, A., Untch, M., Mariani, G., Baselga, J., Kaufmann, M., Cameron, D., Bell, R., Bergh, J., Coleman, R., Wardley, A., Harbeck, N., Lopez, R. I., Mallmann, P., Gelmon, K., Wilcken, N., Wist, E., Sánchez Rovira, P., Piccart-Gebhart, M. J., and study team, H. (2007). 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. *Lancet*, 369(9555):29–36.

[270] Sotiriou, C., Powles, T. J., Dowsett, M., Jazaeri, A. A., Feldman, A. L., Assersohn, L., Gadisetti, C., Libutti, S. K., and Liu, E. T. (2002). Gene expression profiles derived from fine needle aspiration correlate with response to systemic chemotherapy in breast cancer. *Breast Cancer Res.*, 4(3):R3.

[271] Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M. J., Bergh, J., Piccart, M., and Delorenzi, M. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.*, 98(4):262–72.

[272] Sparano, J. A. and Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *J. Clin. Oncol.*, 26(5):721–8.

[273] Straver, M. E., Glas, A. M., Hannemann, J., Wesseling, J., van de Vijver, M. J., Rutgers, E. J. T., Vrancken Peeters, M.-J. T. F. D., van Tinteren, H., Van't Veer, L. J., and Rodenhuis, S. (2010). The 70-gene signature as a response predictor for neoadjuvant chemotherapy in breast cancer. *Breast Cancer Res. Treat.*, 119(3):551–8.

[274] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–50.

[275] Swain, S. M., Baselga, J., Kim, S.-B., Ro, J., Semiglazov, V., Campone, M., Ciruelos, E., Ferrero, J.-M., Schneeweiss, A., Heeson, S., Clark, E., Ross, G., Benyunes, M. C., Cortés, J., and CLEOPATRA Study Group (2015). Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. *N. Engl. J. Med.*, 372(8):724–34.

[276] Swanton, C. (2012). Intratumor heterogeneity: evolution through space and time. *Cancer Res.*, 72(19):4875–82.

[277] Swanton, C., Marani, M., Pardo, O., Warne, P. H., Kelly, G., Sahai, E., Elustondo, F., Chang, J., Temple, J., Ahmed, A. A., Brenton, J. D., Downward, J., and Nicke, B. (2007a). Regulators of mitotic arrest and ceramide metabolism are determinants of sensitivity to paclitaxel and other chemotherapeutic drugs. *Cancer Cell*, 11(6):498–512.

[278] Swanton, C., Marani, M., Pardo, O., Warne, P. H., Kelly, G., Sahai, E., Elustondo, F., Chang, J., Temple, J., Ahmed, A. A., Brenton, J. D., Downward, J., and Nicke, B. (2007b). Regulators of mitotic arrest and ceramide metabolism are determinants of sensitivity to paclitaxel and other chemotherapeutic drugs. *Cancer Cell*, 11(6):498–512.

[279] Symmans, W. F., Peintinger, F., Hatzis, C., Rajan, R., Kuerer, H., Valero, V., Assad, L., Poniecka, A., Hennessy, B., Green, M., Buzdar, A. U., Singletary, S. E., Hortobagyi, G. N., and Pusztai, L. (2007). Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J. Clin. Oncol.*, 25(28):4414–22.

[280] Symmans, W. F., Wei, C., Gould, R., Yu, X., Zhang, Y., Liu, M., Walls, A., Bousamra, A., Ramineni, M., Sinn, B., Hunt, K., Buchholz, T. A., Valero, V., Buzdar, A. U., Yang, W., Brewster, A. M., Moulder, S., Pusztai, L., Hatzis, C., and Hortobagyi, G. N. (2017). Long-Term Prognostic Risk After Neoadjuvant Chemotherapy Associated With Residual Cancer Burden and Breast Cancer Subtype. *J. Clin. Oncol.*, 35(10):1049–1060.

[281] Szakács, G., Annereau, J.-P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K. J., Reinhold, W., Guo, Y., Kruh, G. D., Reimers, M., Weinstein, J. N., and Gottesman, M. M. (2004). Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells. *Cancer Cell*, 6(2):129–37.

[282] Tabchy, A., Valero, V., Vidaurre, T., Lluch, A., Gomez, H., Martin, M., Qi, Y., Barajas-Figueroa, L. J., Souchon, E., Coutant, C., Doimi, F. D., Ibrahim, N. K., Gong, Y., Hortobagyi, G. N., Hess, K. R., Symmans, W. F., and Pusztai, L. (2010). Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clin. Cancer Res.*, 16(21):5351–61.

[283] Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–72.

[284] Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–76.

[285] Thomas, E., Tabernero, J., Fornier, M., Conté, P., Fumoleau, P., Lluch, A., Vahdat, L. T., Bunnell, C. A., Burris, H. A., Viens, P., Baselga, J., Rivera, E., Guarneri, V., Poulart, V., Klimovsky, J., Lebwohl, D., and Martin, M. (2007). Phase II clinical trial of ixabepilone (BMS-247550), an epothilone B analog, in patients with taxane-resistant metastatic breast cancer. *J. Clin. Oncol.*, 25(23):3399–406.

[286] Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., Ziv, E., Culhane, A. C., Paull, E. O., Sivakumar, I. K. A., Gentles, A. J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J. S., Mose, L. E., Vo, N. S., Liu, J., Liu, Y., Rader, J., Dhankani, V., Reynolds, S. M., Bowlby, R., Califano, A., Cherniack, A. D., Anastassiou, D., Bedognetti, D., Rao, A., Chen, K., Krasnitz, A., Hu, H., Malta, T. M., Noushmehr, H., Pedamallu, C. S., Bullman, S., Ojesina, A. I., Lamb, A., Zhou, W., Shen, H., Choueiri, T. K., Weinstein, J. N., Guinney, J., Saltz, J., Holt, R. A., Rabkin, C. E., Cancer Genome Atlas Research Network, Lazar, A. J., Serody, J. S., Demicco, E. G., Disis, M. L., Vincent, B. G., and Shmulevich, L. (2018). The Immune Landscape of Cancer. *Immunity*, 48(4):812–830.e14.

[287] Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., and Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U. S. A.*, 113(50):14330–14335.

[288] Troester, M. A., Herschkowitz, J. I., Oh, D. S., He, X., Hoadley, K. A., Barbier, C. S., and Perou, C. M. (2006). Gene expression patterns associated with p53 status in breast cancer. *BMC Cancer*, 6:276.

[289] Trougakos, I. P., So, A., Jansen, B., Gleave, M. E., and Gonos, E. S. (2004). Silencing expression of the clusterin/apolipoprotein j gene in human cancer cells using small interfering RNA induces spontaneous apoptosis, reduced growth ability, and cell sensitization to genotoxic and oxidative stress. *Cancer Res.*, 64(5):1834–42.

[290] Tsai, H.-C., Huang, C.-Y., Su, H.-L., and Tang, C.-H. (2014). CTGF increases drug resistance to paclitaxel by upregulating survivin expression in human osteosarcoma cells. *Biochim. Biophys. Acta*, 1843(5):846–54.

[291] Tsujie, M., Nakamori, S., Nakahira, S., Takahashi, Y., Hayashi, N., Okami, J., Nagano, H., Dono, K., Umeshita, K., Sakon, M., and Monden, M. (2007). Human equilibrative nucleoside transporter 1, as a predictor of 5-fluorouracil resistance in human pancreatic cancer. *Anticancer Res.*, 27(4B):2241–9.

[292] Untch, M., Loibl, S., Bischoff, J., Eidtmann, H., Kaufmann, M., Blohmer, J.-U., Hilfrich, J., Strumberg, D., Fasching, P. A., Kreienberg, R., Tesch, H., Hanusch, C., Gerber, B., Rezai, M., Jackisch, C., Huober, J., Kühn, T., Nekljudova, V., von Minckwitz, G., German Breast Group (GBG), and Arbeitsgemeinschaft Gynäkologische Onkologie-Breast (AGO-B) Study Group (2012). Lapatinib versus trastuzumab in combination with neoadjuvant anthracycline-taxane-based chemotherapy (GeparQuinto, GBG 44): a randomised phase 3 trial. *Lancet. Oncol.*, 13(2):135–44.

[293] Untch, M., Rezai, M., Loibl, S., Fasching, P. A., Huober, J., Tesch, H., Bauerfeind, I., Hilfrich, J., Eidtmann, H., Gerber, B., Hanusch, C., Kühn, T., du Bois, A., Blohmer, J.-U., Thomssen, C., Dan Costa, S., Jackisch, C., Kaufmann, M., Mehta, K., and von Minckwitz, G. (2010). Neoadjuvant treatment with trastuzumab in HER2-positive breast cancer: results from the GeparQuattro study. *J. Clin. Oncol.*, 28(12):2024–31.

[294] Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.*, 43:11.10.1–33.

[295] van der Hage, J. A., van de Velde, C. J., Julien, J. P., Tubiana-Hulin, M., Vandervelden, C., and Duchateau, L. (2001). Preoperative chemotherapy in primary operable breast cancer: results from the European Organization for Research and Treatment of Cancer trial 10902. *J. Clin. Oncol.*, 19(22):4224–37.

[296] Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L., and Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.*, 107(39):16910–5.

[297] van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6.

[298] Venables, W. and Ripley, B. (1997). *Modern applied statistics with S-Plus.* Springer.

[299] Vera-Ramirez, L., Sanchez-Rovira, P., Ramirez-Tortosa, C. L., Quiles, J. L., Ramirez-Tortosa, M., and Lorente, J. A. (2013). Transcriptional shift identifies a set of genes driving breast cancer chemoresistance. *PLoS One*, 8(1):e53983.

[300] Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A., and Peters, B. (2015). The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, 43(Database issue):D405–12.

[301] Vogelstein, B. (1999). Digital PCR. *Proc. Natl. Acad. Sci.*, 96(August):9236–9241.

[302] von Minckwitz, G., Procter, M., de Azambuja, E., Zardavas, D., Benyunes, M., Viale, G., Suter, T., Arahmani, A., Rouchet, N., Clark, E., Knott, A., Lang, I., Levy, C., Yardley, D. A., Bines, J., Gelber, R. D., Piccart, M., Baselga, J., and APHINITY Steering Committee and Investigators (2017). Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. *N. Engl. J. Med.*, 377(2):122–131.

[303] von Minckwitz, G., Schneeweiss, A., Loibl, S., Salat, C., Denkert, C., Rezai, M., Blohmer, J. U., Jackisch, C., Paepke, S., Gerber, B., Zahm, D. M., Kümmel, S., Eidtmann, H., Klare, P., Huober, J., Costa, S., Tesch, H., Hanusch, C., Hilfrich, J., Khandan, F., Fasching, P. A., Sinn, B. V., Engels, K., Mehta, K., Nekljudova, V., and Untch, M. (2014). Neoadjuvant carboplatin in patients with triple-negative and HER2-positive early breast cancer (GeparSixto; GBG 66): a randomised phase 2 trial. *Lancet. Oncol.*, 15(7):747–56.

[304] Wahl, A. F., Donaldson, K. L., Fairchild, C., Lee, F. Y., Foster, S. A., Demers, G. W., and Galloway, D. A. (1996). Loss of normal p53 function confers sensitization to Taxol by increasing G2/M arrest and apoptosis. *Nat. Med.*, 2(1):72–9.

[305] Wang, Q., Shi, Y.-L., Zhou, K., Wang, L.-L., Yan, Z.-X., Liu, Y.-L., Xu, L.-L., Zhao, S.-W., Chu, H.-L., Shi, T.-T., Ma, Q.-H., and Bi, J. (2018). PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer. *Cell Death Dis.*, 9(7):739.

[306] Wang, Y., Wang, X., Zhao, H., Liang, B., and Du, Q. (2012). Clusterin confers resistance to TNF-alpha-induced apoptosis in breast cancer cells through NF-kappaB activation and Bcl-2 overexpression. *J. Chemother.*, 24(6):348–57.

[307] Wang, Y., Xu, Y., Chen, J., Ouyang, T., Li, J., Wang, T., Fan, Z., Fan, T., Lin, B., and Xie, Y. (2016). TP53 mutations are associated with higher rates of pathologic complete response to anthracycline/cyclophosphamide-based neoadjuvant chemotherapy in operable primary breast cancer. *Int. J. cancer*, 138(2):489–96.

[308] Wheeler, D. A. and Wang, L. (2013). From human genome to cancer genome: the first decade. *Genome Res.*, 23(7):1054–62.

[309] Wiesmüller, L. (2001). Genetic Stabilization by p53 Involves Growth Regulatory and Repair Pathways. *J. Biomed. Biotechnol.*, 1(1):7–10.

[310] Wong, D. J., Liu, H., Ridky, T. W., Cassarino, D., Segal, E., and Chang, H. Y. (2008). Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell*, 2(4):333–44.

[311] Wu, D. and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, 40(17):e133.

[312] Xu, L., Wu, W., Cheng, G., Qian, M., Hu, K., Yin, G., and Wang, S. (2017). Enhancement of Proliferation and Invasion of Gastric Cancer Cell by KDM5C Via Decrease in p53 Expression. *Technol. Cancer Res. Treat.*, 16(2):141–149.

[313] Yang, K., Gao, K., Hu, G., Wen, Y., Lin, C., and Li, X. (2016). CTGF enhances resistance to 5-FU-mediating cell apoptosis through FAK/MEK/ERK signal pathway in colorectal cancer. *Onco. Targets. Ther.*, 9:7285–7295.

[314] Yap, T. A., Gerlinger, M., Futreal, P. A., Pusztai, L., and Swanton, C. (2012). Intratumor heterogeneity: seeing the wood for the trees. *Sci. Transl. Med.*, 4(127):127ps10.

[315] Yarchoan, M., Johnson, B. A., Lutz, E. R., Laheru, D. A., and Jaffee, E. M. (2017). Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer*, 17(9):569.

[316] Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., Carter, S. L., Getz, G., Stemke-Hale, K., Mills, G. B., and Verhaak, R. G. W. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, 4:2612.

[317] Yu, G. and He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.*, 12(2):477–9.

[318] Yu, H., Pardoll, D., and Jove, R. (2009). STATs in cancer inflammation and immunity: a leading role for STAT3. *Nat. Rev. Cancer*, 9(11):798–809.

[319] Yuan, H., Chen, J., Liu, Y., Ouyang, T., Li, J., Wang, T., Fan, Z., Fan, T., Lin, B., and Xie, Y. (2015). Association of PIK3CA Mutation Status before and after Neoadjuvant Chemotherapy with Response to Chemotherapy in Women with Breast Cancer. *Clin. Cancer Res.*, 21(19):4365–72.

[320] Yuan, J.-H., Cheng, J.-Q., Jiang, L.-Y., Ji, W.-D., Guo, L.-F., Liu, J.-J., Xu, X.-Y., He, J.-S., Wang, X.-M., and Zhuang, Z.-X. (2008). Breast cancer resistance protein expression and 5-fluorouracil resistance. *Biomed. Environ. Sci.*, 21(4):290–5.

[321] Zararsiz, G., Goksuluk, D., Klaus, B., Korkmaz, S., Eldem, V., Karabulut, E., and Ozturk, A. (2017). voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. *PeerJ*, 5:e3890.

[322] Zhai, L., Li, S., Li, X., Li, H., Gu, F., Guo, X., Liu, F., Zhang, X., and Fu, L. (2015). The nuclear expression of poly (ADP-ribose) polymerase-1 (PARP1) in invasive primary breast tumors is associated with chemotherapy sensitivity. *Pathol. Res. Pract.*, 211(2):130–7.

[323] Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics*, 25(10):1293–9.

[324] Zheng, C., Zheng, L., Yoo, J.-K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J. Y., Zhang, Q., Liu, Z., Dong, M., Hu, X., Ouyang, W., Peng, J., and Zhang, Z. (2017). Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell*, 169(7):1342–1356.e16.

[325] Zwiener, I., Frisch, B., and Binder, H. (2014). Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One*, 9(1):e85150.

Appendix A

Clinical Tables

Table A.1: Tumour pathology and stage at diagnosis

Trial ID	Age	T stage	N stage	Grade	Histology	ER/HER2 status	PAM50
T001	51	T2	N0	3	IDC	ER+ HER2+	Her2
T002	42	T4d	N1+	2	IDC	ER+ HER2-	Basal
T003	36	T2	N0	2	IDC	ER+ HER2+	LumB
T004	33	T3	N0	3	IDC	ER+ HER2-	Basal
T005	49	T3	N1+	3	IDC	ER+ HER2-	LumB
T006	46	T2	N1+	3	IDC	ER+ HER2-	LumB
T007	42	T2	N0	2	IDC	ER+ HER2+	Her2
T008	50	T3	N1+	3	IDC	ER- HER2-	Basal
T009	53	T2	N1+	2	IDC	ER+ HER2+	Her2
T010	56	T2	N0	2	Mixed	ER+ HER2+	LumB
T011	53	T3	N1+	3	IDC	ER+ HER2+	Her2
T012	33	T4d	N1+	3	IDC	ER+ HER2-	Basal
T013	33	T2	N0	3	IDC	ER+ HER2-	Basal
T014	40	T2	N0	3	IDC	ER+ HER2+	Her2
T015	57	T2	N0	3	IDC	ER- HER2-	Basal
T016	56	T3	N1+	2	ILC	ER+ HER2-	LumA
T017	46	T3	N1+	3	IDC	ER+ HER2-	Normal
T018	47	T3	N1+	2	IDC	ER- HER2-	Normal
T019	59	T2	N1+	2	IDC	ER+ HER2-	NA
T020	34	T1	N0	3	IDC	ER+ HER2+	LumB
T022	42	T2	N1+	2	IDC	ER- HER2+	Her2
T023	39	T2	N1+	2	IDC	ER+ HER2-	LumA
T024	27	T2	N0	3	IDC	ER- HER2-	Basal
T025	62	T2	N1+	2	IDC	ER+ HER2-	LumA
T026	32	T2	N0	3	Medullary	ER+ HER2-	Basal
T027	52	T2	N0	2	IDC	ER+ HER2-	LumA
T028	37	T3	N1+	3	IDC	ER+ HER2+	Her2
T029	53	T2	N0	3	IDC	ER+ HER2+	LumB
T030	48	T3	N0	2	IDC	ER+ HER2-	LumA
T031	51	T1	N0	3	IDC	ER- HER2+	Her2
T032	69	T2	N0	3	IDC	ER- HER2-	Basal
T033	49	T2	N0	3	IDC	ER- HER2-	Basal
T035	33	T2	N0	3	IDC	ER- HER2-	Basal
T036	54	T2	N1+	3	IDC	ER+ HER2-	Her2
T037	49	T2	N0	3	IDC	ER- HER2-	Basal
T038	41	T2	N0	2	IDC	ER- HER2-	Basal
T039	51	T2	N1+	2	IDC	ER+ HER2-	LumA

Table A.1 – continued from previous page

Trial ID	Age	T stage	N stage	Grade	Histology	ER/HER2 status	PAM50
T040	34	T2	N0	3	IDC	ER- HER2-	Basal
T041	44	T2	N0	3	IDC	ER+ HER2-	LumB
T042	57	T3	N0	3	IDC	ER- HER2-	Basal
T043	54	T2	N1+	2	IDC	ER+ HER2-	LumB
T044	40	T2	N1+	3	IDC	ER+ HER2-	LumB
T045	41	T2	N1+	3	IDC	ER+ HER2-	LumA
T046	65	T2	N1+	3	IDC	ER+ HER2-	LumB
T047	58	T3	N0	2	IDC	ER+ HER2-	LumA
T048	66	T2	N0	2	IDC	ER+ HER2+	LumB
T049	37	T2	N1+	3	IDC	ER+ HER2+	Her2
T050	60	T3	N1+	2	IDC	ER+ HER2+	LumB
T051	46	T2	N0	2	IDC	ER+ HER2-	LumB
T052	35	T2	N0	3	IDC	ER- HER2-	Basal
T053	46	T3	N1+	2	ILC	ER+ HER2-	NA
T054	51	T2	N0	3	IDC	ER- HER2-	Basal
T055	63	T2	N1+	3	IDC	ER+ HER2+	NA
T056	55	T2	N0	2	IDC	ER- HER2-	Normal
T057	42	T2	N0	3	IDC	ER- HER2-	Basal
T058	24	T3	N1+	2	IDC	ER+ HER2-	LumB
T059	52	T1	N0	3	IDC	ER+ HER2+	Her2
T060	67	T3	N1+	3	IDC	ER+ HER2-	LumB
T061	53	T2	N1+	2	IDC	ER+ HER2-	LumA
T062	64	T3	N1+	3	IDC	ER+ HER2-	LumB
T064	61	T2	N0	3	IDC	ER+ HER2+	Her2
T065	63	T3	N1+	2	ILC	ER- HER2-	Basal
T066	52	T4d	N1+	2	IDC	ER+ HER2-	LumB
T067	48	T2	N1+	2	IDC	ER+ HER2-	LumA
T068	61	T4d	N1+	2	IDC	ER+ HER2-	LumA
T069	65	T3	N1+	3	Mixed	ER+ HER2+	LumA
T070	63	T2	N1+	3	IDC	ER+ HER2-	Basal
T071	35	T2	N1+	3	IDC	ER- HER2-	Basal
T072	53	T2	N1+	2	Mixed	ER+ HER2-	LumA
T073	38	T3	N0	3	IDC	ER+ HER2+	Her2
T074	50	T2	N1+	2	IDC	ER+ HER2-	LumB
T075	52	T2	N1+	2	IDC	ER+ HER2-	NA
T076	54	T2	N0	2	IDC	ER- HER2+	Normal
T077	48	T2	N0	3	IDC	ER+ HER2+	NC
T078	50	T3	N0	3	IDC	ER+ HER2+	Her2

Table A.1 – continued from previous page

Trial ID	Age	T stage	N stage	Grade	Histology	ER/HER2 status	PAM50
T079	59	T2	N0	2	IDC	ER- HER2-	Basal
T080	50	T3	N1+	3	IDC	ER+ HER2+	Her2
T081	39	T2	N0	3	IDC	ER- HER2-	Basal
T083	48	T3	N0	3	IDC	ER+ HER2+	LumB
T086	63	T2	N0	3	IDC	ER- HER2-	Her2
T087	56	T2	N1+	3	IDC	ER+ HER2+	LumB
T088	67	T2	N0	2	IDC	ER+ HER2+	LumB
T089	33	T4d	N1+	3	IDC	ER- HER2-	Basal
T090	49	T2	N0	3	IDC	ER- HER2-	Normal
T091	45	T2	N0	3	IDC	ER+ HER2-	NA
T092	53	T2	N1+	3	IDC	ER- HER2+	Her2
T093	48	T2	N0	3	IDC	ER+ HER2-	LumB
T094	35	T2	N1+	3	IDC	ER- HER2-	Basal
T095	53	T2	N0	2	Apocrine	ER- HER2+	Her2
T096	59	T2	N1+	2	ILC	ER+ HER2-	LumA
T097	56	T1	N0	3	IDC	ER- HER2+	Her2
T098	37	T2	N0	2	IDC	ER+ HER2+	LumA
T099	62	T3	N0	2	IDC	ER+ HER2+	LumB
T100	57	T3	N0	3	IDC	ER- HER2+	Her2
T101	74	T3	N1+	3	IDC	ER- HER2-	Basal
T102	65	T2	N1+	3	IDC	ER+ HER2-	LumB
T103	55	T2	N0	3	Medullary	ER- HER2-	Basal
T104	53	T2	N1+	3	IDC	ER- HER2-	Her2
T105	43	T3	N0	3	IDC	ER+ HER2+	Her2
T106	39	T4d	N1+	3	IDC	ER+ HER2-	LumB
T108	66	T2	N0	3	IDC	ER- HER2-	Basal
T109	39	T3	N1+	2	IDC	ER+ HER2-	LumA
T110	42	T4d	N1+	2	Micropapillary	ER+ HER2+	Her2
T112	38	T2	N0	3	IDC	ER- HER2-	Basal
T114	62	T2	N1+	3	IDC	ER- HER2+	Basal
T115	56	T3	N1+	3	Mixed	ER+ HER2+	Her2
T116	66	T2	N1+	3	IDC	ER+ HER2+	LumB
T117	59	T2	N0	3	IDC	ER+ HER2+	Her2
T118	44	T3	N1+	2	IDC	ER- HER2-	Normal
T119	52	T3	N1+	2	Mixed	ER+ HER2+	LumB
T120	61	T2	N0	3	IDC	ER- HER2+	Her2
T121	50	T2	N1+	2	Apocrine	ER+ HER2+	Normal

Table A.1 – continued from previous page

Trial ID	Age	T stage	N stage	Grade	Histology	ER/HER2 status	PAM50
T122	55	T2	N1+	2	IDC	ER+ HER2+	LumB
T123	53	T2	N1+	3	IDC	ER- HER2-	Basal
T124	50	T3	N0	2	IDC	ER+ HER2+	Her2
T125	56	T2	N0	3	IDC	ER+ HER2+	LumB
T126	32	T2	N0	2	IDC	ER+ HER2-	LumB
T127	65	T2	N1+	2	ILC	ER+ HER2-	LumA
T128	55	T2	N1+	3	IDC	ER- HER2+	Her2
T129	43	T2	N0	3	IDC	ER- HER2-	Basal
T130	68	T2	N1+	3	IDC	ER- HER2-	Basal
T131	81	T4d	N1+	2	Apocrine	ER- HER2+	Her2
T132	47	T3	N0	2	IDC	ER+ HER2+	Her2
T133	61	T3	N1+	3	Micropapillary	ER+ HER2-	LumB
T134	36	T3	N0	3	IDC	ER+ HER2+	Her2
T135	41	T3	N0	3	IDC	ER- HER2-	Basal
T136	67	T3	N0	3	IDC	ER- HER2-	Basal
T137	45	T2	N1+	2	IDC	ER+ HER2-	LumB
T138	38	T2	N1+	3	Micropapillary	ER+ HER2-	LumB
T139	46	T2	N0	3	IDC	ER+ HER2-	LumB
T140	47	T3	N1+	2	IDC	ER+ HER2-	LumA
T141	39	T2	N0	3	IDC	ER+ HER2-	LumB
T142	42	T2	N1+	3	IDC	ER+ HER2+	LumB
T143	49	T2	N0	3	IDC	ER+ HER2-	Basal
T144	58	T2	N0	3	IDC	ER+ HER2+	LumB
T145	62	T2	N1+	2	IDC	ER+ HER2-	LumB
T147	57	T2	N1+	2	IDC	ER+ HER2+	LumB
T148	38	T1	N1+	3	IDC	ER+ HER2-	LumB
T149	28	T2	N0	3	IDC	ER+ HER2+	LumB
T150	60	T3	N1+	2	ILC	ER- HER2-	Her2
T151	68	T2	N1+	3	IDC	ER+ HER2-	LumB
T152	36	T2	N0	3	IDC	ER+ HER2+	Her2
T153	64	T3	N1+	3	IDC	ER- HER2-	Basal
T154	62	T2	N0	3	IDC	ER- HER2-	Basal
T155	58	T4d	N1+	3	Micropapillary	ER+ HER2-	LumB
T156	67	T2	N1+	3	IDC	ER+ HER2-	LumB
T157	79	T2	N1+	3	IDC	ER+ HER2+	Her2
T158	46	T1	N0	3	IDC	ER+ HER2+	Her2
T159	36	T3	N1+	2	IDC	ER+ HER2-	LumB
T160	50	T3	N0	2	IDC	ER+ HER2-	LumA

Table A.1 – continued from previous page

Trial ID	Age	T stage	N stage	Grade	Histology	ER/HER2 status	PAM50
T161	55	T3	N1+	3	IDC	ER+ HER2-	LumB
T162	55	T1	N0	3	IDC	ER+ HER2+	LumB
T163	39	T2	N0	3	IDC	ER+ HER2+	NA
T164	44	T2	N0	2	IDC	ER+ HER2+	LumB
T165	20	T2	N0	3	IDC	ER+ HER2+	LumA
T166	48	T1	N0	2	IDC	ER- HER2-	LumA
T167	59	T2	N1+	2	ILC	ER+ HER2-	LumB
T168	50	T2	N0	3	Medullary	ER- HER2+	Basal
T169	35	NA	N1+	2	IDC	ER+ HER2-	LumB
T170	66	T2	N1+	2	Micropapillary	ER+ HER2-	LumB
T171	68	T3	N0	3	IDC	ER+ HER2+	LumB
T172	54	T2	N1+	3	IDC	ER- HER2+	Her2
T173	68	T2	N0	2	IDC	ER+ HER2-	LumB
T174	50	T2	N1+	2	IDC	ER+ HER2-	LumB
T175	48	T3	N1+	3	IDC	ER+ HER2+	Her2
T176	39	T3	N1+	2	IDC	ER+ HER2-	LumB
T178	56	T2	N1+	2	Mixed	ER+ HER2-	LumB
T180	52	T3	N0	3	IDC	ER- HER2+	Her2

Table A.2: Neoadjuvant regimens administered

Trial ID	Regimen name	Regimen classification	Total cycles
T001	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T002	T-FEC	Taxane > Anthracycline	6
T003	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T004	FEC-T	Anthracycline > Taxane	6
T005	FEC-T	Anthracycline > Taxane	6
T006	T-FEC	Taxane > Anthracycline	6
T007	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T008	T-FEC	Taxane > Anthracycline	6
T009	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T010	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T011	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T012	T-FEC	Taxane > Anthracycline	6
T013	T-FEC	Taxane > Anthracycline	6
T014	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T015	T-FEC	Taxane > Anthracycline	6
T016	T-FEC	Taxane > Anthracycline	6
T017	T-FEC	Taxane > Anthracycline	6
T018	T-FEC	Taxane > Anthracycline	6
T019	T-FEC	Taxane > Anthracycline	6
T020	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T022	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T023	T-FEC	Taxane > Anthracycline	6
T024	T-FEC	Taxane > Anthracycline	6
T025	T-FEC	Taxane > Anthracycline	6
T026	T-FEC	Taxane > Anthracycline	6
T027	TC	Taxane, no anthracycline	4
T028	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T029	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T030	T-FEC	Taxane > Anthracycline	6
T031	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T032	T-FEC	Taxane > Anthracycline	6
T033	T-FEC	Taxane > Anthracycline	6
T035	T-FEC	Taxane > Anthracycline	6
T036	T-FEC	Taxane > Anthracycline	6
T037	T-FEC	Taxane > Anthracycline	6
T038	T-FEC	Taxane > Anthracycline	6
T039	T-FEC	Taxane > Anthracycline	6

Table A.2 – continued from previous page

Trial ID	Regimen name	Regimen classification	Total cycles
T040	T-FEC	Taxane > Anthracycline	6
T041	T-FEC	Taxane > Anthracycline	6
T042	T	Taxane, no anthracycline	2
T043	T-FEC	Taxane > Anthracycline	5
T044	T-FEC	Taxane > Anthracycline	6
T045	T-FEC	Taxane > Anthracycline	6
T046	T-EC	Taxane > Anthracycline	6
T047	T-FEC	Taxane > Anthracycline	6
T048	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T049	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T050	EC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T051	T-FEC	Taxane > Anthracycline	6
T052	T-FEC	Taxane > Anthracycline	6
T053	T-FEC	Taxane > Anthracycline	6
T054	TC	Taxane, no anthracycline	4
T055	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T056	T-FEC	Taxane > Anthracycline	6
T057	T-FEC	Taxane > Anthracycline	6
T058	T-FEC	Taxane > Anthracycline	6
T059	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T060	T-FEC	Taxane > Anthracycline	6
T061	FEC-T	Anthracycline > Taxane	6
T062	T-FEC	Taxane > Anthracycline	6
T064	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T065	T-Carboplatin	Taxane + Platinum	6
T066	T-FEC	Taxane > Anthracycline	6
T067	T-FEC	Taxane > Anthracycline	6
T068	T-FEC	Taxane > Anthracycline	6
T069	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	5
T070	FEC-T	Anthracycline > Taxane	6
T071	T-FEC	Taxane > Anthracycline	6
T072	T-FEC	Taxane > Anthracycline	6
T073	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T074	FEC-T	Anthracycline > Taxane	6
T075	T-FEC	Taxane > Anthracycline	6
T076	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T077	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T078	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6

Table A.2 – continued from previous page

Trial ID	Regimen name	Regimen classification	Total cycles
T079	T-FEC	Taxane > Anthracycline	6
T080	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T081	T-FEC	Taxane > Anthracycline	6
T083	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T086	T-FEC	Taxane > Anthracycline	6
T087	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T088	TC + Trastuzumab	Taxane + anti-HER2	4
T089	EC-T	Anthracycline > Taxane	5
T090	T-FEC	Taxane > Anthracycline	6
T091	T-FEC	Taxane > Anthracycline	6
T092	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T093	T-FEC	Taxane > Anthracycline	6
T094	T-FEC	Taxane > Anthracycline	6
T095	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T096	T-FEC	Taxane > Anthracycline	6
T097	TC + Trastuzumab	Taxane + anti-HER2	4
T098	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T099	FEC-T + Trastuzumab + Pertuzumab	Anthracycline > Taxane + anti-HER2	6
T100	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T101	TC	Taxane, no anthracycline	4
T102	TC	Taxane, no anthracycline	1
T103	T-FEC	Taxane > Anthracycline	6
T104	T-FEC	Taxane > Anthracycline	6
T105	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T106	T	Taxane, no anthracycline	3
T108	TC	Taxane, no anthracycline	4
T109	T-FEC	Taxane > Anthracycline	6
T110	FEC-T + Trastuzumab + Pertuzumab	Anthracycline > Taxane + anti-HER2	6
T112	T-FEC	Taxane > Anthracycline	6
T114	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T115	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T116	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T117	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T118	P-FEC	Taxane > Anthracycline	7
T119	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T120	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T121	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6

Table A.2 – continued from previous page

Trial ID	Regimen name	Regimen classification	Total cycles
T122	EC + Trastuzumab	Anthracycline, no taxane + anti-HER2	4
T123	P-FEC	Taxane > Anthracycline	7
T124	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T125	TC + Trastuzumab	Taxane + anti-HER2	4
T126	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T127	T-FEC	Taxane > Anthracycline	6
T128	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T129	P-FEC	Taxane > Anthracycline	7
T130	P-EC	Taxane > Anthracycline	7
T131	T + Pertuzumab + Trastuzumab	Taxane + anti-HER2	4
T132	TC + Trastuzumab	Taxane + anti-HER2	4
T133	T-EC	Taxane > Anthracycline	6
T134	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T135	P-FEC	Taxane > Anthracycline	7
T136	P-EC	Taxane > Anthracycline	7
T137	T-FEC	Taxane > Anthracycline	6
T138	T-EC	Taxane > Anthracycline	6
T139	T-EC	Taxane > Anthracycline	6
T140	T-FEC	Taxane > Anthracycline	6
T141	T-FEC	Taxane > Anthracycline	6
T142	T-FEC + Trastuzumab + Pertuzumab	Taxane + anti-HER2 > Anthracycline	7
T143	T-FEC	Taxane > Anthracycline	6
T144	T-FEC + Trastuzumab	Taxane + anti-HER2 > Anthracycline	6
T145	T-EC	Taxane > Anthracycline	6
T147	FEC-T + Trastuzumab + Pertuzumab	Anthracycline > Taxane + anti-HER2	7
T148	T-FEC	Taxane > Anthracycline	6
T149	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T150	P-EC	Taxane > Anthracycline	7
T151	T-FEC	Taxane > Anthracycline	6
T152	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T153	P-EC	Taxane > Anthracycline	6
T154	P-FEC	Taxane > Anthracycline	7
T155	EC-T	Anthracycline > Taxane	6
T156	TC	Taxane, no anthracycline	4
T157	T + Pertuzumab + Trastuzumab	Taxane + anti-HER2	7
T158	TC + Pertuzumab + Trastuzumab	Taxane + anti-HER2	6
T159	T-FEC	Taxane > Anthracycline	6
T160	T-FEC	Taxane > Anthracycline	6

Table A.2 – continued from previous page

Trial ID	Regimen name	Regimen classification	Total cycles
T161	T-EC	Taxane > Anthracycline	6
T162	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T163	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T164	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T165	TC + Trastuzumab	Taxane + Platinum + anti-HER2	6
T166	P-Carboplatin	Taxane + Platinum	4
T167	T	Taxane, no anthracycline	1
T168	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T169	T-EC	Taxane > Anthracycline	7
T170	T-FEC	Taxane > Anthracycline	6
T171	TC + Trastuzumab	Taxane + anti-HER2	4
T172	FEC-T + Trastuzumab	Anthracycline > Taxane + anti-HER2	6
T173	P + Trastuzumab	Taxane + anti-HER2	4
T174	T-EC	Taxane > Anthracycline	6
T175	T-FEC + Trastuzumab + Pertuzumab	Taxane + anti-HER2 > Anthracycline	7
T176	T-Carboplatin	Taxane + Platinum	6
T178	T-FEC	Taxane > Anthracycline	6
T180	FEC-T + Trastuzumab + Pertuzumab	Anthracycline > Taxane + anti-HER2	7

Table A.3: Assessment of post-therapy tumour pathology

Trial ID	Tumour dim. (mm)	% Cellularity	% CIS	Positive LN	Largest LN met (mm)	RCB score	RCB category
T001	3 x 2	5	0	0	0	0.98	RCB-I
T002	0	0	0	0	0	0	pCR
T003	0	0	0	0	0	0	pCR
T004	16 x 10	80	0	0	0	2.075	RCB-II
T005	15 x 9	5	0	5	5	2.866	RCB-II
T006	18 x 13	50	0	3	7	3.584	RCB-III
T007	21 x 12	5	30	0	0	1.267	RCB-I
T008	6 x 0.1	1	0	4	2	1.948	RCB-II
T009	0	0	0	1	0.1	0.676	RCB-I
T010	16 x 14	30	0	0	0	1.807	RCB-II
T011	8 x 5	1	0	0	0	0.876	RCB-I
T012	0	0	0	0	0	0	pCR
T013	0	0	0	0	0	0	pCR
T014	0	0	0	0	0	0	pCR
T015	5 x 3	40	0	0	0	1.508	RCB-II
T016	180 x 44	15	0	2	9	3.773	RCB-III
T017	70 x 24	0.1	0	5	5	2.403	RCB-II
T018	50 x 25	1	0	3	4	2.633	RCB-II
T019	100 x 45	5	0	5	6	3.359	RCB-III
T020	8 x 4	0.1	0	0	0	0.581	RCB-I
T022	0	0	0	0	0	0	pCR
T023	22 x 11	2	0	2	3	2.474	RCB-II
T024	0	0	0	0	0	0	pCR
T025	15 x 14	50	0	6	12	3.828	RCB-III
T026	3 x 2	80	0	0	0	1.57	RCB-II
T027							NA
T028	0	0	0	0	0	0	pCR
T029	27 x 10	60	0	0	0	2.066	RCB-II
T030	33 x 16	10	0	0	0	1.613	RCB-II
T031	0	0	0	0	0	0	pCR
T032	9 x 5	50	0	0	0	1.72	RCB-II
T033	1 x 1	95	0	0	0	1.388	RCB-II
T035	3 x 2	90	0	0	0	1.601	RCB-II
T036	0	0	0	0	0	0	pCR
T037	0	0	0	0	0	0	pCR
T038	0	0	0	0	0	0	pCR

Table A.3 – continued from previous page

Trial ID	Tumour dim. (mm)	% Cellularity	% CIS	Positive LN	Largest LN met (mm)	RCB score	RCB category
T039	6 x 4	20	40	0	0	1.279	RCB-I
T040	0	0	0	0	0	0	pCR
T041	14 x 11	80	0	0	0	2.068	RCB-II
T042	4 x 0.5	5	0	0	0	0.892	RCB-I
T043	18 x 17	40	0	2	13	1.949	RCB-II
T044	18 x 10	40	50	11	10	3.514	RCB-III
T045	27 x 9	0.1	0	0	0	0.69	RCB-I
T046	5 x 4	25	0	2	5	2.873	RCB-II
T047	41 x 37	30	2	1	0.1	2.795	RCB-II
T048	18 x 15	5	0	0	0	1.354	RCB-I
T049	0	0	0	0	0	0	pCR
T050	26 x 10	5	0	11	32	3.615	RCB-III
T051	31 x 15	10	0	0	0	1.595	RCB-II
T052	0	0	0	1	0.7	0.941	RCB-I
T053	34 x 30	10	0	4	9	3.429	RCB-III
T054	0	0	0	0	0	0	pCR
T055	6 x 4	5	0	1	4	2.368	RCB-II
T056	12 x 7	70	0	0	0	1.92	RCB-II
T057	0	0	0	0	0	0	pCR
T058	100 x 36	50	0	12	3	4.013	RCB-III
T059	0	0	0	0	0	0	pCR
T060	3 x 2	10	0	3	0.1	1.882	RCB-II
T061	15 x 12	70	5	1	5	3.346	RCB-III
T062	32 x 23	10	5	4	6	3.253	RCB-II
T064	1 x 0.8	90	0	0	0	1.349	RCB-I
T065	55 x 43	10	0	6	4	3.382	RCB-III
T066	25 x 11	1	0	9	7	2.77	RCB-II
T067	15 x 13	10	2	7	3	2.966	RCB-II
T068	130 x 48	5	5	20	9	3.591	RCB-III
T069	30 x 16	1	10	0	0	1.062	RCB-I
T070	0	0	0	0	0	0	pCR
T071	0	0	0	0	0	0	pCR
T072	70 x 26	10	10	2	4	3.152	RCB-II
T073	0	0	0	0	0	0	pCR
T074	17 x 11	15	40	0	0	1.45	RCB-II
T075	3 x 1	20	0	0	0	1.169	RCB-I
T076	1 x 0.5	10	0	0	0	0.892	RCB-I

Table A.3 – continued from previous page

Trial ID	Tumour dim. (mm)	% Cellularity	% CIS	Positive LN	Largest LN met (mm)	RCB score	RCB category
T077	0	0	0	0	0	0	pCR
T078	0	0	0	2	6	1.491	RCB-II
T079	0	0	0	0	0	0	pCR
T080	0	0	0	7	7	1.72	RCB-II
T081	0	0	0	0	0	0	pCR
T083	10 x 2	10	99	0	0	0.558	RCB-I
T086	2 x 1	10	0	1	2	2.129	RCB-II
T087	50 x 31	15	5	1	0.1	2.553	RCB-II
T088							NA
T089	100 x 44	20	2	1	1	3.165	RCB-II
T090	14 x 8	10	0	0	0	1.414	RCB-II
T091	14 x 10	50	10	0	0	1.86	RCB-II
T092	12 x 4	5	0	0	0	1.169	RCB-I
T093	6 x 5	80	0	2	0.2	2.636	RCB-II
T094	0	0	0	0	0	0	pCR
T095	0	0	0	0	0	0	pCR
T096	65 x 38	20	0.5	21	8	3.869	RCB-III
T097	0	0	0	0	0	0	pCR
T098	30 x 21	30	20	1	3	3.105	RCB-II
T099	0	0	0	0	0	0	pCR
T100	2 x 1	5	0	0	0	0.892	RCB-I
T101	17 x 12	40	30	1	9	3.225	RCB-II
T102	19 x 10	60	1	2	12	3.679	RCB-III
T103	0	0	0	0	0	0	pCR
T104	0	0	0	0	0	0	pCR
T105	11 x 6	2	0	0	0	1.028	RCB-I
T106							NA
T108							NA
T109	19 x 17	90	0	8	11	4.116	RCB-III
T110	95 x 35	5	0	2	2	2.913	RCB-II
T112	0	0	0	0	0	0	pCR
T114	14 x 9	40	0	0	0	1.807	RCB-II
T115	140 x 25	20	0	6	4	3.68	RCB-III
T116	20 x 14	70	5	1	4	3.374	RCB-III
T117	8 x 2	2	0	0	0	0.911	RCB-I
T118	13 x 4	40	70	3	9	3.041	RCB-II
T119	28 x 14	5	0	1	1	2.398	RCB-II

Table A.3 – continued from previous page

Trial ID	Tumour dim. (mm)	% Cellularity	% CIS	Positive LN	Largest LN met (mm)	RCB score	RCB category
T120	10 x 1	5	0	0	0	1.023	RCB-I
T121	90 x 53	10	5	4	1	3.114	RCB-II
T122	41 x 24	5	0	1	11	3.015	RCB-II
T123	15 x 7	75	0	0	0	1.98	RCB-II
T124	83 x 37	10	50	3	3	3.055	RCB-II
T125	17 x 5	5	40	0	0	1.125	RCB-I
T126	8 x 7	40	0	0	0	1.687	RCB-II
T127	19 x 15	10	20	26	21	3.597	RCB-III
T128	0	0	0	0	0	0	pCR
T129	20 x 15	20	0	0	0	1.729	RCB-II
T130	4 x 1	5	0	1	0.1	1.623	RCB-II
T131	28 x 22	60	0	4	14	4.074	RCB-III
T132	18 x 17	10	0	0	0	1.54	RCB-II
T133	15 x 13	90	5	1	2	3.259	RCB-II
T134	40 x 24	10	30	2	2	2.834	RCB-II
T135	0	0	0	0	0	0	pCR
T136	0	0	0	0	0	0	pCR
T137	16 x 10	1	0	4	5	2.668	RCB-II
T138	2 x 1	5	0	1	0.1	1.568	RCB-II
T139	19 x 18	60	30	1	1.5	3.055	RCB-II
T140	17 x 12	80	0	1	18	3.753	RCB-III
T141	27 x 18	20	1	0	0	1.799	RCB-II
T142	0	0	0	0	0	0	pCR
T143	0	0	0	0	0	0	pCR
T144	0	0	0	0	0	0	pCR
T145	140 x 55	5	0	10	12	3.712	RCB-III
T147	14 x 12	10	60	1	0.6	2.16	RCB-II
T148	18 x 13	10	40	3	8	3.022	RCB-II
T149	20 x 8	3	20	0	0	1.143	RCB-I
T150	48 x 22	20	0	22	10	3.796	RCB-III
T151	16 x 3	5	0	3	5	2.685	RCB-II
T152	0	0	0	0	0	0	pCR
T153	18 x 11	40	0	0	0	1.878	RCB-II
T154	0	0	0	0	0	0	pCR
T155	250 x 57	3	0	9	2	3.144	RCB-II
T156	45 x 19	30	5	3	8	3.65	RCB-III

Table A.3 – continued from previous page

Trial ID	Tumour dim. (mm)	% Cellularity	% CIS	Positive LN	Largest LN met (mm)	RCB score	RCB category
T157							NA
T158							NA
T159	96 x 50	10	0	2	7	3.477	RCB-III
T160	60 x 39	30	5	0	0	2.187	RCB-II
T161	70 x 40	30	10	5	9	3.956	RCB-III
T162	0	0	0	0	0	0	pCR
T163	0	0	0	0	0	0	pCR
T164	12 x 8	1	0	0	0	0.943	RCB-I
T165	7 x 7	60	40				NA
T166	10 x 7	60	0	0	0	1.842	RCB-II
T167	152 x 54	20	20	21	60	4.743	RCB-III
T168	4 x 3	80	0	0	0	1.665	RCB-II
T169	152 x 54	10	0	0	0	2.036	RCB-II
T170	0.8 x 0.6	10	0	0	0	0.899	RCB-I
T171	21 x 11	60	0	1	2	3.164	RCB-II
T172	0	0	0	0	0	0	pCR
T173	13 x 9	15	0	0	0	1.52	RCB-II
T174	22 x 12	50	1	1	6	3.352	RCB-III
T175	33 x 20	2	20	0	0	1.204	RCB-I
T176	130 x 80	30	20	5	12	4.255	RCB-III
T178	34 x 18	30	10	5	5	3.523	RCB-III
T180	0	0	0	0	0	0	pCR