Understanding the clinical implications of the evolution of breast cancers from primary to metastatic disease using next generation sequencing

Peter Stephen Savas

0000-0001-5999-428X

Doctor of Philosophy – Medicine, Dentistry & Health Sciences

February 2020

Sir Peter MacCallum Cancer Department of Oncology

University of Melbourne

Submitted in total fulfilment of the degree of Doctor of Philosophy

Abstract

Precision oncology refers to the use of sophisticated assays to tailor therapy to an individual patient. The feasibility of implementing such a program in a comprehensive cancer centre in Australia is unknown. The utility of precision oncology also depends on understanding how genomic profiles may evolve over time and differ between tumours in the same patient.

A precision oncology program called SEGMENT was designed and implemented in a single centre. The program was popular, recruiting well over the study period. Timely acquisition of samples for sequencing was suboptimal from external pathology providers, and proved increasingly expensive during the study period. Delivery of results in a manner where they could be utilized by the patient was challenging in cases where patients were referred late in their natural history. A custom hybrid capture panel worked reliably. A total of 300 patients were recruited to the study, of which 288 had at least one sample received. Accounting for attrition, 214 patients or 71% went onto the main study. The spectrum of mutations and copy number alterations found in this study was similar to published cohorts. There were few differences between primary and metastatic lesions on average. Paired primary and metastatic samples however displayed discordance for both copy number and mutations. This was the case for actionable alterations in *ESR1*, *ERBB2* and very rarely PIK3CA. Approximately 50% of patients had an actionable alteration. Of these 14% of the overall cohort received a therapy matched to their genomic profile. Five patients received matched therapy off trial and 26 received matched trial therapy. Three were zero responses

in the off-trial group, and a response rate of 27% in the matched trial therapy group.

To explore genomic heterogeneity in greater resolution, 4 patients with advanced breast cancer underwent rapid autopsies to collect large numbers of metastatic samples. Whole exome sequencing was performed on multiple lesions per patient which allowed inference of the subclonal structure. All patients displayed a monophyletic architecture, with truncal driver alterations giving rise to subclones with differing genomic profiles. One patient with a long disease free interval from primary to metastasis showed the acquisition of a new driver in the metastatic lesion. Driver alterations appeared to shape subsequent evolution.

Declaration

(i) the thesis comprises only my original work towards the degree of Doctor of Philosophy except where indicated in the preface;

(ii) due acknowledgement has been made in the text to all other material used;

(iii) the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Peter Stephen Savas

Preface

The work described in this thesis is my own, save for the following:

1. The SEGMENT protocol was written by my supervisor Prof Sherene Loi and myself in collaboration.
2. The Molecular Genomics Core facility at the Peter MacCallum Cancer Centre conducted the majority of sequencing library preparation and sequencing.
3. Archival tissue processing and DNA extraction was performed by the Department of Pathology at the Peter MacCallum Cancer Centre.
4. Copy number analyses using CONTRA were run by Dr Jason Li.
5. SEGMENT study procedures including tissue requests, collection and entering of clinical data, curation of variants and generation of reports, were conducted in part by Courtney Van Geelen assisted by Keilly Kuykhoven.
6. CASCADE staff Heather Thorne, Dr Kathryn Alsop and Lisa Devereux collected tissue from autopsies conducted on 4 patients.
7. Zhi Ling Teo selected appropriate samples collected at autopsy and extracted DNA for sequencing.
8. Autopsy sequencing library preparation and sequencing was performed by contracted service providers.
9. Dr Christophe Lefevre conducted the majority of subclonality analyses in the CASCADE autopsy cases.

Chapter 6 consists of the peer-reviewed publication **The subclonal architecture of metastatic breast cancer: Results from a prospective community-based rapid autopsy program "CASCADE"**, which was published in PLoS Medicine on December 27, 2016. The author contributions are reproduced here from the publication:

Conceptualization: Peter Savas, Charles Swanton, Sherene Loi, David D. Bowtell.

Data curation: Peter Savas, Christophe Lefevre, Franco Caramia, Christoffer Flensburg, Maria Joao Silva, Sarah-Jane Dawson, Andrew Rowan, Nnennaya Kanu.

Formal analysis: Peter Savas, Christophe Lefevre, Christoffer Flensburg, Franco Caramia, Andrew Rowan, Terence P. Speed.

Funding acquisition: Heather A. Thorne, David D. Bowtell, Charles Swanton, Terence P. Speed, Sherene Loi.

Investigation: Zhi Ling Teo, Franco Caramia, Kathryn Alsop, Mariam Mansour, Heather A. Thorne, Maria Joao Silva, Sarah-Jane Dawson, Christophe Lefevre, Christoffer Flensburg, Andrew Rowan, Nnennaya Kanu, Sherene Loi, Prudence A. Francis, Heather A. Thorne, Stephen Fox.

Methodology: Peter Savas, Christophe Lefevre, Christoffer Flensburg, Charles Swanton, Andrew Rowan, Heather A. Thorne, Kathryn Alsop, Michelle Dietzen, Maik Kschischo.

Project administration: Heather A. Thorne, Kathryn Alsop, David D. Bowtell, Sherene Loi.

Resources: Zhi Ling Teo, Christoffer Flensburg, Mariam Mansour, Prudence A. Francis, Heather A. Thorne, Kathryn Alsop.

Software: Christophe Lefevre, Christoffer Flensburg, Terence P. Speed.

Supervision: Sherene Loi, Charles Swanton.

Validation: Christophe Lefevre, Franco Caramia, Christoffer Flensburg, Peter Savas, Nnennaya Kanu, Michelle Dietzen, Maik Kschischo, Terence P. Speed.

Visualization: Peter Savas, Christophe Lefevre, Franco Caramia, Christoffer Flensburg.

Writing – original draft: Peter Savas, Sherene Loi.

Writing – review & editing: Peter Savas, Zhi Ling Teo, Christophe Lefevre, Christoffer Flensburg, Franco Caramia, Kathryn Alsop, Mariam Mansour, Prudence A. Francis, Heather A. Thorne, Maria Joao Silva, Nnennaya Kanu, Michelle Dietzen, Andrew Rowan, Maik Kschischo, Stephen Fox, David D. Bowtell, Sarah-Jane Dawson, Terence P. Speed, Charles Swanton, Sherene Loi.

Funding Acknowledgement

Personal financial support:

- Shields Research Entry Scholarship, Royal Australian College of Physicians
- National Health and Medical Research Council and National Breast Cancer Foundation Co-funded Postgraduate Scholarship
- Cancer Therapeutics Cooperative Research Centre Clinician Researcher PhD Top Up Scholarship

Project funding:

- Australia & New Zealand Breast Cancer Trials Group Seed Grant

Acknowledgements

I would like to thank my supervisors Prof. Sherene Loi and Prof Terry Speed for making this project possible, and for the uncountable occasions they provided sound advice, support and encouragement.

I would like to thank A/Prof. Linda Mileshkin for her mentorship and support.

Many people were generous with their time and patience during this PhD. I would like to thank the following people in particular:

1. Andrew Fellowes, for his help and expertise in setting up the sequencing workflow;
2. Stephen Fox and the Department of Pathology at the Peter MacCallum Cancer Centre,
3. Tim Semple, for his expertise in the laboratory, dedication to troubleshooting and many hours of bench-side discussions about sequencing and life in general;
4. The CASCADE team, Heather Thorne, Lisa Devereux and Kath Alsop for the extraordinary effort and professionalism required to collect rapid autopsy tissues;
5. Courtney Van Geelen, for much hard work and assistance in the clinical sequencing project;
6. Christophe Lefevre, for his analysis of the autopsy exome data and many useful discussions over the years.
7. Joyce Teo for introducing me to the laboratory and always setting a fine example of a professional scientist at work.

I would also like to think the many other people who made this work possible (in no particular order):

Sathana Dushyanthen
Stephen Luen
Balaji Virassamy
Gisela Mir Arnau
Sreeja Gadipally
Michael Henderson
Bruce Mann
David Speakman
Sophie Nightingale
Catherine Poliness
Anita Skandarajah
Anannya Chakrabarti
Chantel Thornton
David Gyorki
Jocelyn Lippey
Kyle Kowalewski
Prue Francis
Kelly Phillips
Sarah Pratt
Lisa Sheeran
Marisa Grossi
Rhiannon Koirala
Courtney Savill
Sue Sturrock
Connie Mascarenhas
Christopher Angel
Danielle Jensen
Toni-Maree Rogers
Jia-Min Pang
Sarah-Jane Dawson
Chen-Fang Weng
Maria Joao Silva
Kate Moodie
Karda Cavanagh
Franco Caramia
Ken Doig
Maria Doyle
Jason Ellul
David Goode
Jason Li
Richard Lupat
Aparna Rao
Andrew Colebatch.

Table of Contents

List of tables

List of figures

Chapter 1: Introduction

1.1 Breast Cancer in Australia

It is estimated that breast cancer will be the most frequently diagnosed malignancy in 2018 in Australia, in both women and the general population [1]. Projected incidence in 2018 is 18,087 new cases in women, over twice the incidence of colorectal cancer, the second most diagnosed cancer. This reflects the high life-time risk of developing breast cancer in Australian women, which in 2018 was 1 in 8, up to the age of 85. As of 2013, there were 202, 577 women with a prior diagnosis of breast cancer in the last 30 years, amounting to approximately 2% of the female population of Australia. The projected number of deaths from breast cancer in 2018 are 3,157 with a life-time risk of dying of breast cancer of 1 in 44 for Australian women, second only to lung cancer with a risk of death of 1 in 28. Of note, breast cancer is the leading cause of death in women aged 30 to 59 [2]. Breast cancer is also the leading cause of non-fatal disease burden in women [2]. In 2014-15 there were at least 89,000 intravenous chemotherapy episodes related to breast cancer.

Detailed data on the population prevalence of metastatic breast cancer are not available. In 2011 approximately 5% of new breast cancer diagnoses were stage 4 [3]. Typical rates of developing metastatic disease in the 5 years after diagnosis of localised (node-negative) or regional (node-positive or locally advanced) early stage disease are approximately 5% and 18% respectively [4]. These data suggest that a significant proportion of Australian women with breast cancer are living with advanced disease at any given time.

1.2 Contemporary management of breast cancer

Many years of large, carefully conducted clinical trials have resulted in detailed treatment algorithms for early and advanced stage breast cancer [5,6]. The principles of managing early stage disease are:

Achieving optimal locoregional control of disease with surgery and radiotherapy.

Clinicopathological staging to determine risk of recurrence, supplemented with gene expression-based risk of recurrence scores, where appropriate.

Selection for adjuvant chemotherapy treatment based on estimated recurrence risk and breast cancer subtype, determined by oestrogen receptor (ER), progesterone receptor (PR) and HER2 (human epidermal growth factor receptor 2) status, and patient comorbidities.

Treatment with adjuvant endocrine therapy in hormone receptor (ER/ PR) - positive disease.

In addition to the above, chemotherapy may be administered in the neoadjuvant setting, with the goal of improving the probability of breast conserving surgery. More recently, the use of neoadjuvant chemotherapy permits the identification of patients who fail to achieve a complete pathological response as candidates for further post-operative chemotherapy [7]. Intensification or de-escalation of systemic therapy is introduced in certain situations. For example, shorter duration and less toxic chemotherapy in early stage HER2-amplified disease may be employed for node-negative tumours less than 3cm [8]. In young women with high risk hormone receptor-positive disease who receive adjuvant chemotherapy,

ovarian ablation confers added benefit over tamoxifen or aromatase inhibitor alone [9].

At present, besides routine haematoxylin and eosin staining to determine tumour size, grade and histological subtype, a limited number of biomarkers are assessed in early stage breast cancer. These have undergone extensive testing to provide high analytic and clinical validity. Oestrogen receptor and progesterone receptor expression are determined using semi-quantitative immunohistochemical assays, and tumours are scored as either hormone receptor positive or negative [10]. The epidermal growth factor receptor, HER2, is assessed with immunohistochemistry and fluorescence in situ hybridisation (FISH) [11]. More recently, multigene recurrence score assays have found a role in further differentiating recurrence risk in cancers where standard clinicopathological markers are imprecise. The 21-gene recurrence score assay, Oncotype DX, and the 70-gene assay, MammaPrint, both aim to identify patients at higher or lower risk of recurrence, particularly in the hormone receptor-positive, HER2-negative, node-negative population. Recent large randomised controlled prospective trials have proven the utility of these assays in sparing some patients from receiving adjuvant chemotherapy when there is little or no benefit [12,13].

In the advanced disease setting, oestrogen and progesterone receptors expression and HER2 status are critical for determining optimal therapy [6]. Germline *BRCA1* or *BRCA2* mutation have recently emerged as another biomarker in the metastatic setting, predicting benefit from carboplatin chemotherapy and the PARP (poly-ADP ribose polymerase) inhibitor, olaparib [14,15]. Despite these agents

providing response rates of 60-70% in cancers of germline *BRCA1* or *BRCA2* carriers, the advantage over standard of care chemotherapy was small, with a progression-free survival benefit of 2 to 3 months, and as yet, no evidence of improvement in overall survival.

Regardless of the breast cancer subtype defined by hormone receptor status and HER2, multiple lines of therapy including chemotherapy are usually administered in the advanced disease setting, with the aim of controlling disease and maintaining quality of life. Nevertheless, each additional line of therapy provides a smaller likelihood of treatment benefit and a higher risk of toxicity.

1.3 The Cancer Genome

Long before the advent of molecular techniques or the discovery of DNA, the observation via light microscopy that abnormal chromosomes were present in dividing malignant cells led to the hypothesis that cancer was a clonal proliferation caused by heritable material [16]. In 1960, it was shown that a particular chromosomal abnormality was strongly associated with chronic myelogenous leukaemia but not with other forms of leukemia [17]. It was also found that oncogenic viruses contained DNA that was very closely related to DNA in normal cells, indicating that abnormal or unregulated expression of existing genes had a role in malignancy, thus introducing the concept of oncogenes [18]. The definitive proof that cancer is a disease arising from aberrant genetic material came in the early 1980s. Elegant experiments demonstrated that transfer of nuclear material from human cancers into an immortalised murine fibroblast cell line precipitated malignant transformation [19,20]. Following this, in 1982, two

independent groups determined that a single nucleotide variant in the *HRAS* gene was sufficient to transform immortalised recipient fibroblasts [21,22]. Subsequently, more refined experimental systems clarified that multiple genetic lesions, such as introduction of a *RAS* family gene and *MYC* were required to transform normal cells which had not been immortalised [23]. These powerful demonstrations that malignant phenotypes are closely linked to genetic aberrations have motivated the application of DNA sequencing to the understanding and treatment of cancers in the modern era.

1.4 Heterogeneity and clonal evolution

The study of heterogeneity in cancer has a long history predating modern genomics, though it was apparent to early researchers that cancer heterogeneity had a genetic component. Levan and Hauschka noted in 1953 that "A neoplasm is genetically comparable with a heterogeneous clonal aggregate of asexually reproducing microorganisms…" [24]. Observing the variable growth potential of cells isolated from malignant ascites, Klein & Klein were particularly perspicacious in their findings, reporting: "It would appear that mammalian tumors are not what they were often considered to be: clearly delineated clones of one cell type, but rather complex populations in which a broad range of variation permits the continuous operation of selective forces..." [25]. Intra-tumour heterogeneity was noted at the cytogenetic level in mouse tumours [26]. Further mouse experiments demonstrated that precise phenotypes such as the acquired resistance of leukaemic cells to folic acid derivates would develop in small populations of cells, and that this characteristic was heritable and irreversible, strongly suggesting a genetic basis [27]. Small case series first noted

inter-tumour heterogeneity, obtaining variable karyotypes between different solid tumours [28]. More refined experiments in melanoma animal models established that even single cells from solid tumours exhibit significant phenotypic heterogeneity [29,30].

In the same period, evidence had accumulated that most human malignancies were of a clonal origin, developing from a single cell [31]. It is remarkable that such evidence was deduced in human tumour samples at the time, without the use of DNA sequencing. In haematological malignancies which produced immunoglobulins such as multiple myeloma or chronic lymphocytic leukaemia, it was found that all cells produced exactly the same immunoglobulin which could only be explained by a clonal origin [32]. In solid tumours, X-inactivation mosaicism was used as a cell of origin marker. This refers to the random inactivation of the paternal or maternal X chromosome in females early in embryonic development, a decision which is then maintained in all daughter cells. Normal tissues are therefore mosaic for expression of genes on paternal and maternal X chromosomes. In appropriately selected individuals who carry differing alleles for gluose-6-phosphate dehydrogenase (G-6-PD) on each X chromosome, the distinct electrophoretic properties of the G-6-PD enzyme encoded by each allele allows the active X chromosome to be distinguished with a protein-based assay. All normal tissues studied in this manner are mosaic for X-inactivation. A study of uterine leiomyomas found all tumours to display uniform expression of only one allele, indicating a clonal origin for this benign smooth muscle tumour [33]. Similarly, multiple benign proliferative lesions, carcinomas and haematological malignancies displayed uniform expression of one allele

proving their clonal origin [31]. However, in these early studies the clonal origin of breast cancer could not be reliably determined due to contamination by normal stromal and immune cells which obscures the tumour X-inactivation signal [31].

The previously proposed dichotomy between polyclonal normal tissue and clonal malignant tissue is since disproven, by the observation of clonal normal haematopoiesis [34]. Strikingly, a high prevalence of well-known cancer driver mutations have been found in normal tissues including haemopoietic cells [35,36], skin [37] and oesophageal mucosa [38], indicating that non-malignant clonal proliferations of physiologically normal cells is a widespread phenomenon.

In a review in 1976, Nowell proposed a prescient model of the clonal evolution of tumours [39]. In this model, malignancies begin from a single cell, and due to inherent genetic instability progressively diverge from the genome of normal cells. This machine of genomic diversity is constrained by the selective pressures of metabolic, nutritional and immune factors present in the host. Along the way, many non-viable or less fit clones fail to survive, with the successful clones having an accumulation of advantages which elevate their proliferative rate over others in the given conditions. Exposure to cancer therapy is another particularly potent selective pressure which drives evolution in the tumour cell population. Although this Darwinian model of malignant evolution is now thought to be oversimplified, the contemporary study of tumour heterogeneity and evolution through the high-powered lens of modern genomics has largely affirmed it to be correct. The significance of these profound aspects of malignant biology – clonal evolution and cellular heterogeneity – were appreciated by these early researchers. They

surmised that in the context of billions of phenotypically heterogenous cells present in clinically detectable tumours, and the tendency of the malignant process to generate genomic diversity fuelling tumour evolution, cure of a well-established malignancy would be a challenging undertaking [39,40]. Nowell also recognised that the clonal evolution model predicted that each tumour would be genomically and phenotypically unique, and thus would require individualised therapy, pre-empting the contemporary concept of precision oncology [39].

1.5 The breast cancer genome

As outlined above, the value of studying the cancer genome was in some way well-established 20 years prior to our ability to do so. As genomic technologies have become available, the breast cancer genome has been unravelled in increasing detail, in parallel with a rapid escalation in the number of samples interrogated. Early studies applied cytogenic methods to human breast tumours and found that large chromosomal aberrations were common. These findings are well known, including gain of 1q, loss of 16q, and a homogenously staining region (HSR) on 8p indicative of intra-chromosomal amplification [41,42]. Double minute chromosomes (DMIN) were also seen, showing that extra-chromosomal DNA amplification was a prevalent mechanism [43]. Beyond cytogenetics, amplification of specific oncogenes was also demonstrated, the classic example being the first report of ERBB2 amplification (or HER-2/*neu* as it was called at the time) in a cohort of 103 primary breast cancers using Southern blot analysis [44]. This study also showed that *ERBB2* amplification was associated with an adverse prognosis.

Better resolution for detecting copy number changes across the whole genome was afforded by comparative genomic hybridisation (CGH). CGH labels tumour and normal DNA with different fluorescent dyes and hybridises each to genomic reference DNA, with the ratio of the intensity of the fluorescent dyes providing a measure of relative copy number [45]. These early studies in human breast cancers confirmed that certain chromosomal aberrations were recurrent, particularly gain of 1q and 8q, and began to identify that the genomic profile was different between benign breast lesions and tumours, and between clinicopathological groups based on receptor status and grade [46,47,48]. It was also understood at the time that within copy number aberrant regions, some genes displayed high level amplification and were more likely to be of clinical significance compared to others [46,48]. Microarray CGH methods allowed highly amplified regions to be defined more precisely. This led to additional findings that there was significant fine-grained variations in copy number within large amplified genomic segments between different tumours [49], and that copy number amplification had a large effect on gene expression [50]. More accurate genetic maps enabled the aggregation of results from multiple studies, showing that recurrent loss of heterozygosity (LOH) events were also prevalent and recurrently deleted genes such as *PTEN* were identified [51–53].

The advent of single nucleotide polymorphism arrays allowed detection of copy-neutral LOH as an additional mechanism that could alter gene function in breast cancers [54,55]. It was also appreciated that oestrogen receptor-positive and negative breast cancers had distinctive copy number profiles [54,56]. Larger sample sizes in individual studies enabled attempts to correlate copy number

patterns which were agnostic of the amplified or deleted genes with survival outcomes [57].

The accumulation of knowledge regarding the breast cancer genome was summarised, along with other cancer types, by Futreal et al constructing a census of genes implicated in the development of cancer [58]. In the census, single nucleotide variants, insertions and deletions, chromosomal translocations and copy number alterations were considered. A restricted set of several hundred genes were found to harbour recurrent alterations, and these genes typically contained domains acting as protein kinases, transcriptional regulators or components of DNA repair machinery. The census also highlighted the difficulty in distinguishing truly significant alterations (since termed 'drivers') from passenger or bystander alterations. The census was biased by the high proportion of haematological malignancies and the lack of comprehensive genomic assays. Since its publication, whole genome (WGS) and whole exome sequencing (WES) in many thousands of tumours have largely confirmed the findings of the Futreal et al census, and have further expanded the number of cancer-related genes and the breadth of pathogenic alterations [59,60]. In addition, beyond cataloguing recurrent alterations, an increasing number of putative driver alterations have undergone experimental verification of their driver status [59].

The difficulties in determining which mutations are significant was further elaborated by Sjöblom et al and Wood et al, two companion studies which undertook high-throughput capillary based sequencing of coding region and whole genome sequences in 11 breast cancers [61,62], together with a larger but

less comprehensively sequenced validation cohort. It was clear from these studies that the more of the genome that is sequenced, the more mutations were found. These studies employed the concept that the frequency with which a gene is altered is useful in determining if it is a driver, showing that this framework was of use even in small numbers of samples if correction for background mutation rates was performed. The mechanistic basis for this concept is that alterations in driver genes will be both detectable and recurrent as opposed to passenger mutations. Firstly, by definition driver genes confer a proliferative advantage to affected cells whereas passengers do not, and therefore cells with drivers expand to a detectable prevalence in the whole tumour population. Secondly, drivers affect cell physiology through a restricted set of common pathways capable of causing malignant phenotypes in different cell types, and therefore will occur recurrently both within and across tumour types [63].

Sjöblom et al and Wood et al found a small set of driver genes that are recurrently altered in line with the cancer gene census. Concurrently many new genes with mutations of uncertain significance were also found, highlighting the necessity to use comprehensive genomic assays in 'discovery' studies to avoid missing potentially relevant alterations. Greenman et al undertook a similar study using capillary sequencing of coding regions in a larger and more diverse cohort of cancers which included 16 breast cancers [64]. This study noted the highly contrasting base substitution spectra across different tumour types, a prelude to extensive work on mutational signatures which would come later. An analysis of mutational selection pressure was also developed borrowing methods from evolutionary biology [65], where the ratio of non-synonymous and synonymous

mutations can provide evidence of a selective advantage for a particular mutation. Tumours typically showed an elevated selection pressure indicative of the presence of driver genes, as would be expected. Application of this method on a gene by gene basis also enabled an estimate of the total number of driver and passenger mutations, leading to the observation that the majority of mutations in a given cohort were likely to be passengers.

These studies set the scene for translational genomics over the next 5 to 10 years, which have since employed whole exome or whole genome approaches in large cohorts. The aims of these studies have been: to achieve sufficient sample size to reliably detect less prevalent alterations; use this wealth of data to distinguish driver and passenger alterations through increasingly sophisticated models of cancer development, mutagenesis and protein function; and associate genomic profiles with clinicopathological groups, prognosis and treatment outcome.

2012 was a pivotal year in realising these aims for breast cancer, with the publication of whole genome and whole exome results for 2,881 primary tumours in the space of two months (see Table 1). The almost simultaneous arrival of all these studies reflects the rapid adoption of 'next-generation` sequencing at the time, consisting mainly of the Illumina sequencing by synthesis technology (deployed in the Illumina Genoma Analyzer IIx and HiSeq machines), but also Applied Biosystems sequencing by oligonucleotide ligation and detection (SOLiD).

Subtype	N	Platform	Reference
All subtypes	21	WGS DNASeq	[66]
All subtypes	2000	Whole genome SNP array	[67]
All subtypes	507	Whole genome SNP array	[68]
		WES DNASeq	
Triple negative	104	Whole genome SNP array	[69]
		WGS/WES DNASeq	
All subtypes	103	WES DNASeq	[70]
All subtypes	100	WES DNASeq	[71]
ER positive	46	WGS DNASeq	[72]
	31	WES DNASeq	

Table 1.1. Large cohort studies comprehensively characterising the breast cancer genome.

Nik-Zainal et al performed whole genome sequencing on 21 primary breast cancers, which were a mixture of ER-positive, HER2-positive, triple negative and *BRCA1* or *BRCA2* mutant cases [66]. This paper highlighted that short-read massively parallel sequencing of the sort which is commonplace today was able to not only catalogue the mutations, copy number alterations and structural variants in a tumour, but also permitted inference of the evolutionary history of each cancer. The major conclusions from this study were that: there is significant evolution over time even in these untreated cancers; significant numbers of mutations were acquired prior to large scale chromosomal alterations; whole genome duplication was a common phenomenon; high level amplification of genes

such as *ERBB2* and *MYC* appeared to happen in a stepwise rather than saltatory fashion. This work also noted that patterns in the allelic frequencies of mutations could be utilised to disembroil the presence of distinct subclones in a tumour. Although it was no surprise that tumours were comprised of subpopulations with distinct genotypes and phenotypes as discussed in the preceding section, this study drove home the notion that any detailed understanding of the breast cancer genome required attention to the subclonal architecture.

Curtis et al presented what remains to this day one of the largest cohorts of a single tumour type to undergo genomic characterisation in the METABRIC cohort [67]. This study was exceptional for careful clinical annotation and long survival follow up. Genome wide SNP arrays and gene expression microarrays were applied to untreated, fresh frozen primary tumours. The chief outcome of this study was that copy number alterations and gene expression could be integrated to develop clusters which partitioned tumours into groups with different biological and clinical significance.

Studies by the Cancer Genome Atlas consortium, Banerji et al and Stephens et al sequenced a total of 703 primary breast cancers, selected from all subtypes. The most important result of these studies was the systematic determination of the driver genes in primary breast cancer (Figure 1). Furthermore, the frequency distribution of the mutations could be seen to follow a consistent pattern, with a small number of genes being mutated in the majority of samples, followed by a 'long tail' of low frequency mutations. This long tailed distribution is characteristic

of most tumour types [73,74]. Copy number alterations display a similar albeit flatter distribution [68].

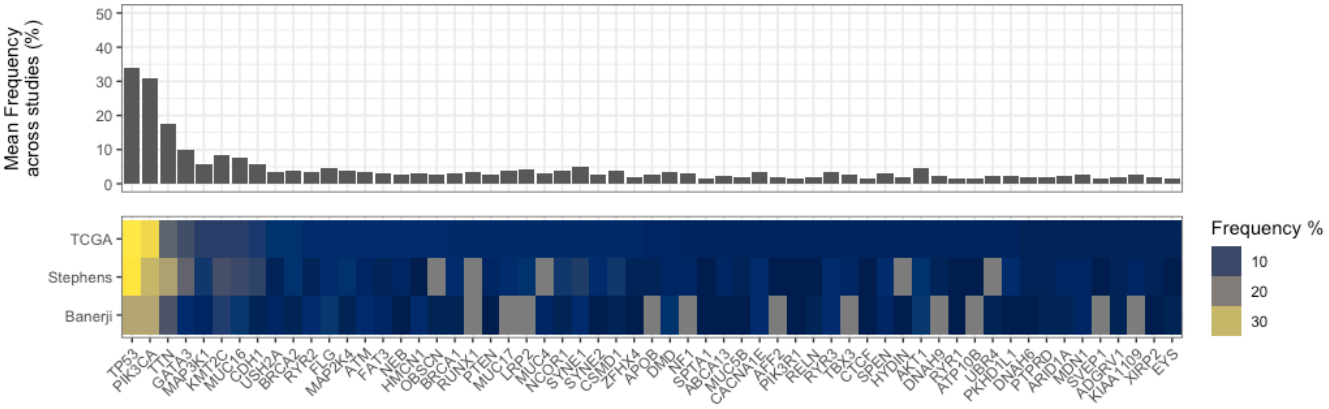


Figure 1.1. Summary of mutated genes in the TCGA, Stephens et al and Banerji et al studies. Data obtained from www.cbioportal.org

The TCGA study also highlighted the genomic differences between breast cancer subtypes. *PIK3CA* mutations were most common in ER-positive Luminal A tumours, whereas *TP53* alterations, *MYC* amplification, *PTEN* and *RB1* deletions were more common in triple negative tumours.

Shah et al focused on triple negative breast cancers (TNBC). They found that even within one subtype, the mutation burden varied widely between tumours, from a few somatic mutations to several hundred. The genes *TP53, PIK3CA, RB1, PTEN, MYO3A* and *GH1* were found to be mutated at a frequency significantly above that expected based on a calculated background mutation rate. Similar to Nik-Zainal et al, subclonal architecture was inferred by clustering mutations based on allele frequency, in this case using the copy number state of the region with the mutation to estimate the proportion of malignant cells harbouring the mutation, termed the 'cancer cell fraction' of a mutation. The co-occurrence of different mutations in the

same cancer cell can be inferred from alignment of the corresponding cancer cell fractions. It is generally assumed that since cancer develops from a single cell, lower cancer cell fractions are 'nested' within higher fractions, indicative of evolutionary branching from a 'most-recent common ancestor'. As expected, mutations in driver genes are found at higher cancer cell fractions than non-drivers, however late emergence of driver mutations, particularly in *TP53*, was also seen.

Ellis et al undertook a smaller cohort which was more clinically focussed, performing WGS on 46 and WES on 31 primary hormone receptor-positive HER2-negative breast cancers that underwent neoadjuvant aromatase inhibition[72]. Cases were defined as either aromatase inhibitor responders or non-responders based on the reduction in Ki67 index, a proliferation marker, during treatment. As well as identifying a spectrum of driver genes in line with the studies mentioned above, it was noted that non-responding tumours had a higher mutation and structural variant burden compared to responders. The association of genomic alterations with clinical behaviour was noted: *TP53* mutations were more prevalent in non-responders, and tumours with higher histological grade; *MAP3K1* mutations were more prevalent in grade 1 and lowly-proliferative tumours; there was a trend towards better aromatase inhibitor response in tumours with *GATA3* mutations.

Nik-Zainal et al and Ellis et al both conducted whole genome sequencing and thus were able to examine structural variants. Both studies did not detect evidence of recurrent structural variants such as translocation or gene fusion events aside

from chromosomal arm gains and losses. Overall however structural variants were found to be common, as seen in other studies [75]. In both studies evidence of chromothripsis, as was found. Chromothripsis refers to the occurrence of multiple genomic rearrangements in a single event occurring in an isolated genomic region, typically one chromosome [76] . It is not generally associated with high level gene amplifications. The significance of this finding in breast cancer is that large scale genomic changes may occur on short time scales, providing another mode of genomic evolution during malignant transformation beyond the stepwise accumulation of driver events.

Nik-Zainal et al performed whole genome sequencing on 560 primary breast cancers [77]. Recurrent non-coding mutations were found in the promoter regions of *PLEKHS1*, *TBC1D1* and *WDR74* in 4%, 1% and 3% of tumours respectively. More frequent alterations were seen in two long non-coding RNAs *MALAT1* (8%) and *NEAT1* (8%). The significance of these findings to breast cancer biology is unclear, and it is possible their recurrent nature merely reflects susceptibility to mutation, but overall it appears that there is a low frequency of non-coding mutations in primary tumours. Conversely, frequent disruptions of tumour suppressor genes due to structural variants was seen, with affected genes including *PTEN* and *TP53*. A subsequent study on the same cohort explored a particular rearrangement class known as tandem duplications [78]. Tandem duplications consist of the duplication and insertion of a genomic segment adjacent to the original [79] with a segment width spanning around 10 kilobases up to several hundred [80]. It has previously been noted as the most common class of rearrangement in breast cancer [75]. The phenomenon is best detected with

whole genome sequencing and has been observed in 50% of TNBC and to a lesser extent in other subtypes, with a predilection to occur in regions containing driver gene alterations or super-enhancer regulatory elements [78,80]. Tandem duplications may cause amplification of driver genes which are wholly contained within the duplicated segment, but may also disrupt tumour suppressor genes if the breakpoint of the tandem duplication (which refers to the junction between the end of the duplicated segment and the start of the next) occurs within the gene [81]. Short tandem duplications of approximately 10 kilobases are associated with deficiency of *BRCA1* but not *BRCA2* [78,80]. Suppression of replication restart at stalled replication forks appears to be a BRCA1-mediated mechanism preventing tandem duplications [82].

As to the question of whether all relevant driver genes have been discovered in breast cancer, multiple lines of evidence suggest the answer is 'No'. Firstly, although down-sampling analyses have shown that the current number of primary breast cancers sequenced should provide 90% power to detect driver alterations down to a frequency of 1%, this assumes there are no differences between breast cancer subtypes, which clearly does not hold [83]. Secondly, more sophisticated and careful modelling of mutational processes results in new drivers being uncovered by reanalysing existing data [84,85]. Thirdly, there are thousands of genes in regions of altered copy number which are of uncertain significance. Compared to single nucleotide variants and indels, it is much more difficult to determine which genes with recurrent copy number alterations are divers, since genes in the same amplicon are usually altered to the same extent *en masse* [86,87]. The most extreme version of this problem is whole genome

doubling, where gene dosage of all genes is increased [88], but even in recurrent chromosome arm level alterations, the exact relevance of each gene is largely unknown. The use of *in vitro* screens testing the proliferative effects of a gene's expression or depletion in breast cancer models has led to identification of candidate drivers from the large pool of genes with low level copy number alterations, implying that more bona fide copy number altered drivers exist [89,90].

1.6 The breast cancer genome in metastatic disease.

Although many treatment-naïve primary tumours have been profiled, relatively few metastatic breast cancers have undergone comprehensive genomic profiling. Metastatic disease is a far more diverse entity than primary disease due to the additional complexity introduced by treatment. It is assumed that this expanded phenotypic topography will further expand the number of undiscovered or under-appreciated driver genes.

Considerable efforts have failed to identify genomic features associated with the development of metastatic disease in breast cancer or other cancer types [91]. Reiter et al also showed that there were no important differences in drivers between primary and metastatic disease, although this is with the crucial caveat that these cancers were not treated in the interim. Since almost all primary breast cancers are now treated with systemic therapies, understanding the genomic profile of metastatic disease is expected to be of considerable importance. The canonical example of this is the development of ligand binding domain mutations in *ESR1* which cause resistance to endocrine therapies such as tamoxifen and

aromatase inhibitors [92–94]. These mutations are very rare in primary breast cancers but are present in around 30% of hormone receptor-positive cancers treated with endocrine therapies. As discussed previously, tumour heterogeneity and clonal evolution are central to the biology of cancer and are of particular importance when considering the genomic profile of metastatic disease under the influence of treatment.

In one of the earliest studies of clonal evolution and heterogeneity, Kuukasjärvi et al performed CGH on 29 primary tumours and associated metachronous metastases [95]. They found that although most tumours had a very similar profile to their metastases, 30% showed major differences. Using FISH probes for metastasis specific alterations, they were unable to detect rare cells with these alterations in the associated primary tumour, although the sensitivity of this approach is limited. X-inactivation studies nevertheless showed that in all cases the metastatic lesion and primary tumour derived from a common progenitor clone. A variety of other CGH and array CGH studies showed similar findings, with broadly concordant copy number profiles that infrequently diverge [96–102].

In a pioneering study utilising massively parallel sequencing, Shah et al performed whole genome sequencing on a sample of malignant pleural effusion obtained from a patient with metastatic lobular breast cancer, as well as the primary tumour excised 9 years prior to sampling the effusion [103]. The metastatic sample had 30 mutations in total, 19 of which were not present in the primary, albeit at a relatively low sequencing depth by modern standards. Some mutations that were of low allele frequency in the primary were enriched in the metastatic

lesion, hinting at the underlying clonal dynamics. Ding et al reported a similar study on a case of TNBC which had a an interval of < 12 months between the primary diagnosis and development of metastatic disease, performing whole genome sequencing on the primary tumour, a brain metastasis and a primary-derived xenograft [104]. In this case, the genomic profiles of the primary and brain metastasis were more closely related than was seen in the study of Shah et al, with only 1 additional high-confidence mutation in the metastasis. The range of allele frequencies in the primary tumour was wider than what was seen in the brain metastasis, which also had more detectable copy number changes. These findings were consistent with the clonal evolution model, where differences from primary to metastatic disease arise from ongoing genomic instability, and the changes in allele frequency spectrum represent subclonal expansion. These early studies also provided evidence that profiling metastatic disease was useful, if not necessary, to understand clonal evolution and the breast cancer genome of primary versus metastatic breast cancer.

Gerlinger et al performed multi-region exome sequencing on two primary renal cell carcinomas and associated metastatic sites [105]. Unprecedented mutational heterogeneity was found between different regions of a primary tumour, and between metastatic sites and the primary tumour, and between metastatic sites themselves, all in a single patient. The mutational profiles across tumours and metastatic sites allowed construction of a phylogenetic tree linking different lesions and allowing the order of mutations to be established. The distance between branches of the tree provided evidence for the extensive evolution that occurs as part of the malignant process and in response to treatment. This study

illustrated clearly the consequences of clonal evolution and heterogeneity in cancer, which had been explored in various forms since the 1950s and is remarkable for triggering a wave of pessimism regarding cancer therapy in the face of rampant genomic complexity, but also in catalysing the study of genomic heterogeneity. One concrete principle that this study highlighted was that so-called 'truncal' mutations which are found in all samples and therefore annotated in the trunk of the phylogenetic tree are likely to be superior therapeutic targets compared to mutations that occur in the branches.

In two large and comprehensive studies focusing on primary and metastatic disease, Yates et al examined in detail the clonal evolution and genomic heterogeneity in primary and metastatic breast cancers, in well annotated clinical cohorts [106,107]. Sequencing platforms and mutation calling pipelines were orthogonally validated to confirm the requisite sensitivity and specificity for studying heterogeneity. Regarding intra-tumour heterogeneity in primary lesions, in 12 primary tumours undergoing targeted sequencing of 8 different tumour regions, the majority displayed regional variation in genomic profiles in a similar although less dramatic fashion to what was seen in renal cell carcinoma [105]. This occurred in two modes: islands of mutations corresponding to a subclone occupying a portion of the tumour, and interweaving of different subclones across a wide area. Subclonal driver mutations were present, but all tumours had one or more driver alterations which were ubiquitous. In 18 cases, biopsies before and after neo-adjuvant chemotherapy were sequenced. 5 cases showed emergence of a subclonal after chemotherapy which was not present at diagnosis. Phylogenetic analysis suggested this subclone was present in the pre-treatment tumour, which

expanded when the dominant tumour mass was eliminated by chemotherapy, confirming the profound effects of treatment on the subclonal architecture of breast cancer.

In the Yates et al 2015 study, there were two cases where breast cancer metastases and primary lesions underwent whole genome sequencing. In one case with a lung metastasis, the metastatic subclone was present in the primary tumour but not in the residual disease after neo-adjuvant chemotherapy, and in another case with a regional lymph node metastasis, the metastatic subclone was derived from a branch of the evolutionary tree suggesting late acquisition of metastatic potential. This apparent variation in the relationship between tumour evolution and metastatic potential was studied in greater detail in a subsequent study [107]. Whole genome sequencing of primary and distant or regional axillary lymph node metastases in 17 cases showed in general that synchronous lymph node metastases are closely related to the genomic profile of the primary tumour. Distant metastatic sites and local recurrences showed more divergence, with 1 or 2 additional driver alterations present in the metastatic disease. Metastatic and recurrent lesions had a differing mutation profile, with more somatic mutations in *ESR1, STAT3, AKT1, ARID1A, BRCA1, NF1* and *TP53* compared to primary lesions. In addition to the expected finding that truncal driver mutations are present in primary and metastatic lesions, it was also seen that the mutational processes active in the tumour trunk are ongoing and contribute to mutagenesis in metastatic disease. Cytotoxic chemotherapy was not seen to contribute to mutational processes, but radiation treatment did in one case.

Lefebvre et al performed whole exome sequencing in 216 metastatic breast cancer patients [108]. Genes found to be significantly altered were *TP53, PIK3CA, GATA3, ESR1, MAP3K1, CDH1, AKT1, MAP2K4, RB1, PTEN, CBFB* and *CDKN2A*. *ESR1* and *RB1* were found altered at a higher rate in metastatic tumours compared to primary tumours from TCGA. The APOBEC3B mutational signature was more prevalent in metastatic hormone receptor-positive cancers when compared to TCGA, but the work of Yates et al discussed above would suggest that this is a feature of primary tumours destined to metastasize rather than a specific feature of metastatic disease [107].

The studies discussed thus far provide substantial evidence that clonal evolution and heterogeneity are prevalent in breast cancer and contribute to clinically meaningful disease patterns. These principles have been corroborated with great clarity by single cell sequencing studies in breast cancer [109–114]. Single cell genomes provide the ultimate view on subclonal architecture, although remain limited by significant technical challenges. These studies have confirmed unequivocally that subclones exist, can be distinguished by their genomic profiles and that multiple subclones may exist in a tumour. Systemic therapy was seen to dramatically alter the subclonal architecture of a tumour, extinguishing some subclones such that the tumour is repopulated by others. These studies also support an evolutionary model where copy number alterations are acquired early in cancer development in a punctuated manner, whereas mutations may accrue more gradually [110,111].

Related to the above considerations is the concordance of genomic alterations between primary and metastatic sites in larger and more varied clinical cohorts profiled with massively parallel sequencing technologies on FFPE tissues. The question here is less a matter of the biological significance of heterogeneity and more about the analytical validity of genomic assays. Jensen et al profiled 104 paired primary tumours and asynchronous metastatic lesions, using a hotspot mutational assay for *PIK3CA* at 3 commonly mutated loci [115]. A multiplexed fluorescent capillary electrophoretic assay was used, which does not have high sensitivity for low frequency variants. Of note, *PIK3CA* mutational status was discrepant between primary tumour and synchronous axillary node metastasis in 13% of 46 paired cases. Between primary and asynchronous metastases, 21% of 100 cases gained a *PIK3CA* mutation, and 11% lost a *PIK3CA* mutation. Schleifman et al performed a similar study using a more sensitive quantitative PCR based assay for *PIK3CA* and *AKT1* mutational status [116]. *PIK3CA* and *AKT1* concordance was 90% and 99% respectively. Meric-Bernstam et al used a hybrid capture panel covering 182 cancer related genes in 33 patients with paired primary and metastatic tissues [117]. 86.6% of mutations were concordant from primary to metastasis. Two cases displayed discordance for *PIK3CA* mutations, which were present in the primary but lacking in the recurrent disease. In both cases the variant allele frequency was ≤ 10% implying these were not clonal alterations. Fumagalli et al profiled 88 primary-metastatic pairs from hormone receptor positive cancers, using a PCR-based mutational hotspot assay covering 11 genes and a copy number assay for 29 genes, as well as a droplet digital assay for *ESR1* [118]. Genomic alterations were concordant for the most part, with *PIK3CA* mutations discordant in approximately 14% of cases. In 5 cases it was

present in the primary only, and in 4 cases in the metastasis only. *ESR1* mutations on the other hand displayed 100% discordance as expected. It is thus possible to conclude that with sensitive next generation sequencing (NGS) assays, clonally dominant truncal driver mutations are likely to be concordant.

1.7 Precision Oncology

Following the first report of the association of a genetic lesion with chronic myelogenous leukaemia, the lesion was characterised as a translocation between chromosomes 9 and 22 fusing the Bcr gene with the Abl tyrosine kinase [119]. This early finding bore great clinical significance in that 40 years after that first report, a new class of therapy targeting the aberrant kinase protein delivered spectacular, practice changing results [120,121]. This first and most clear example encompasses the aims and promise of precision oncology, to link cancer specific features with highly effective targeted therapies. We will begin with a general discussion of precision oncology before considering existing data in breast cancer.

1.7.1 Approaches to genomic profiling for clinical use

Currently, the most common approach to cancer genomic profiling employed in the translational setting are targeted exome panels. These largely use hybrid capture of genomic regions of interest which are enriched for recurrent and clinically significant alterations. Recently the FDA has approved two such assays, the FoundationOne CDx and MSK-IMPACT. These assays report single nucleotide variants, indels, copy number alterations and a limited number of fusions across 324 and 341 genes respectively. Both somatic and germline events are reported. More recently, tumour mutation burden and mutational signatures are also

reported by these assays. A key aspect of these assays which has allowed their development into approved diagnostics is the reporting strategy focused on clinical relevance and potential clinical trial options.

Whole exome sequencing employs the same technology as smaller targeted panels, extended to all coding genes, and is applicable to archival formalin fixed paraffin embedded samples. Whereas small panels may cover between 300 – 350 genes approximating 2Mb of sequence, commercially available whole exome kits cover 45 – 64 Mb [122–124]. This increase in targeted regions necessitates a larger number of sequenced reads to achieve adequate sensitivity for variant detection. Clinical samples may have variable tumour purity, sample quality, and available DNA. Additionally, coverage with hybrid capture methods may be non-uniform. High sequencing coverage is required to mitigate sample level variability and ensure a minimum level of coverage in regions that are captured inefficiently. Acceptable detection power for low frequency variants requires coverage exceeding 300-fold, which at this time is prohibitively expensive for widespread whole exome sequencing. Typical whole exome coverage of 150 fold has been shown to miss important variants in clinical samples [125]. These factors explain the preference for smaller targeted panels over whole exomes in research and commercial precision oncology application.

In addition to the recurrent nature of cancer genome alterations, the focus on targeted panels is partly due to relative ease with which exonic small somatic variants and copy number changes can be determined with short-read sequencing. Unbiased detection of structural variants and non-coding variants are

only accessible with whole genome sequencing [77]. Other long-read approaches currently in development may become useful in future for detection of structural variants. The true prevalence of important structural variants and non-coding variation in solid tumours is largely unknown, but is assumed to be much less than small variants and copy number changes in coding genes [126]. Important examples of the utility of whole genome sequencing include: sensitive determination of mutational signatures [127]; copy number based subtype classification [128]; and determination of tandem duplicator phenotype associated with benefit from immunotherapy [129]. There is also evidence to suggest that whole exome sequencing may be inferior to whole genome sequencing for variant detection, although this difference may be negligible in tumours [130].

Gene expression profiling has been less commonly employed in precision oncology as a stand-alone assay. A large proportion of the variance in gene expression between tumours is explained by tissue of origin [131–133]. This property has resulted in gene expression based assays which may identify tissue of origin in carcinomas of unknown primary [134] but contributes little in cases where the diagnosis is clear.

Recent work has suggested that machine learning approaches using gene expression could result in clinically useful predictors of drug sensitivity. From *in vitro* high throughput drug screens paired with comprehensive genomic profiling of cell lines, gene expression has proven as the most useful feature to predict drug sensitivity [132,133]. It is important to note here that tissue of origin is almost as

good a predictor as gene expression, and that the two are highly correlated [133]. This strong dependence of drug response on disease subtype has been noted in haematological cancers [135]. When considering predictors of drug sensitivity within a given tumour type, such as breast or colorectal cancer, genomic features (that is, mutations and copy number alterations) are the best predictors of drug sensitivity rather than gene expression based profiles [133].

1.7.2 Determining clinically relevant alterations

As described in the previous section, genomic assays typically discover tens to thousands of alterations in a tumour, but only a small proportion of these are likely to be clinically relevant. The first challenge is determining which alterations are drivers, such that they have a role in the biology of the cancer, primarily through conferring a growth advantage to cells that possess it. The second challenge is determining which of these drivers can be used to alter management. This has 4 components: firstly, preferring or rejecting certain therapies on the basis of the alteration; secondly, determining prognosis; thirdly detecting an unexpected germline alteration; fourthly, clarifying the cell of origin of a malignancy when this may otherwise be in doubt. For the purposes of precision oncology, the first and third components are commonly collated under the rubric of 'actionability'. Drivers and actionable alterations do not overlap completely. Some drivers may not be targetable, and some actionable alterations detected with genomic assays, such as mutational signatures of homologous recombination deficiency or tumour mutation burden, may lead to a change in therapy without meeting the definition of a driver.

The definition of actionability is a point of contention. High quality evidence from randomised controlled trials demonstrating that a particular alteration is a valid biomarker is rarely available. Actionability is generally defined as a scale with different levels of evidence. This 'hierarchy of evidence' borrows from the approach of ranking study quality used in evidence-based medicine. A variety of individual projects [74,136], precision oncology programs [137], precision oncology knowledge bases [138,139] and professional working groups [140,141] have devised hierarchies of evidence for interpreting the results of genomic assays. Unlike evidence-based medicine hierarchies, some of these actionability scales mix clinical trial and experimental evidence with regulatory approval status and vary in complexity from scales with 4 levels up to 8 levels.

The utility of these scales as research tools or in clinical practice remains to be seen. For patients with advanced cancer, it is rarely the case that the clinician is required to distinguish between treatment options with subtle differences in supporting evidence. In contrast to a clinical trial outcome, regulatory approval of a therapy incorporates many factors beyond efficacy, may be withdrawn or modified over time, and in any case usually requires a companion diagnostic assay which is not a multi-gene panel assay. In addition, the dynamic nature of evidence in precision oncology means that the actionability rate reported in one study will not be comparable to another study even if they both employ the same actionability scale. Notably, one of the most widespread commercial precision oncology assays from Foundation Medicine, which is FDA approved, does not use any such scale and merely describes genomic alterations as having therapies with 'clinical benefit' or not, either in the same tumour type or other tumour

types [142]. This highlights that a pragmatic definition of actionability is paramount, and more complicated evidence hierarchies may lack real world application.

1.7.3 Motivation for Precision Oncology

The central concept of precision oncology is to undertake additional testing to determine features of a patient's malignant disease that may inform treatment decisions. There are many existing diagnostic assays which form part of the standard diagnostic work up for various malignancies, such as oestrogen receptor staining in breast cancer. These assays have a crucial role in defining subtypes of a cancer and thus the associated natural history (termed prognostic biomarkers), and/or preference for or against treatment (termed predictive biomarkers) according to subtype. In general, the value of these markers exists within the context of a particular malignancy. Genomic biomarkers paired with targeted therapies have proved highly efficacious in some cases, such as HER2 amplification and trastuzumab in breast cancer [143], EML4-ALK fusions and crizotinib in lung adenocarcinoma [144], and BRAF mutation and combined BRAF and MEK inhibition in melanoma [145]. These examples have motivated the precision oncology project to extend determination of genomic biomarkers to all malignant disease. As discussed above, currently this has centred on the use of DNA-based assays of several hundred genes simultaneously. Meta-analyses comparing personalised treatments with non-personalised treatments in the clinical trial setting utilising these assays have demonstrated that targeted therapies are associated with a higher response rate and better progression-free survival [146,147].

It is worth highlighting the particular advantages of a tumour-agnostic gene panel approach to precision oncology with associated caveats, as discussed below.

Treatment-biomarker pairings may be agnostic to tumour type and can provide an indication for treatment without the need for traditional levels of evidence. This is perhaps seen as the prime advantage of genomic precision oncology. With this understanding, a small number of cases exhibiting clinical activity of a drug may be highly informative, as was the case with the determination that gefitinib was active in *EGFR*-mutant lung cancer [148]. Additionally, strong evidence for a treatment-biomarker pair obtained in one cancer type may be transferable to other cancers, for example vemurafenib in *BRAF* V600E-mutant non-melanoma cancers [149]. This transferability may be most pertinent in translating pre-clinical data to the clinic, where treatment-biomarker pairings discovered in high throughput mutational screens, mouse models, cell lines, patient derived cell cultures and patient derived xenografts may lead to treatment recommendations for a patient.

Treatment-biomarker pairings are not guaranteed to transfer results from one tumour type to another, the canonical example being the absence of activity of BRAF inhibition in *BRAF*-mutant colorectal cancer [150]. It should be noted that the mechanism of this failure was determined to be due to bypass EGFR signalling in colorectal cancer, and a subsequent trial which investigated combined EGFR and BRAF inhibition showed promising efficacy [151]. This suggests that identifying cancers with particular genomic alterations remains a useful

enterprise, affording in this case a potentially effective therapy for the otherwise difficult to treat population of *BRAF*-mutant colorectal cancer. It also cautions that the tissue of origin remains an important factor in delivering targeted therapy.

Precision oncology platforms can serve simultaneously as both a clinical and research tool. Resistance to targeted and other cancer therapies often operates at the genomic level. Key examples include the *EGFR* T790M mutation and resistance to EGFR inhibitors in *EGFR*-mutant lung cancer [152], *ESR1* mutations in oestrogen receptor positive breast cancer treated with endocrine therapy [92,93] and reversions of pathogenic mutations in *BRCA1* or *BRCA2* deficient cancers treated with PARP inhibitors [153]. These resistance mechanisms may be rapidly captured using repeat testing with precision oncology genomics, and in turn identify avenues for further treatment without the need to develop new procedures or assays in some cases. This benefit can be seen to occur in real time when a new effective therapy enters clinical practice, such as recent studies identifying genomic markers of benefit and primary resistance to immune checkpoint blockade [154], including both mutation burden and specific gene alterations. Unlike traditional biomarkers therefore, genomic based assays have the ability to increase in value over time without the need to develop an entirely new test.

Genomic platforms as they currently exist in the translational setting are focused on DNA sequencing. This is due to rapid advances in the technical feasibility and reliability of this type of sequencing, the fact it produces a 'binary' result and the notable successes of DNA biomarkers paired with treatment. However DNA

sequencing alone is not sufficient to provide treatment recommendations or explain mechanisms of resistance in all cases. Only approximately 30% of endocrine resistant breast cancers display an *ESR1* mutation for example.

As with any diagnostic assay, the process of ensuring analytical validity and consistency is ongoing. Although short-read sequencing has rapidly declined in cost and in principle is highly accurate, implementation of sequencing comes in a plethora of forms which may provide inconsistent results [124,155–158]. Ensuring the clinical validity of sequencing results, particularly in a rapidly changing field remains an ongoing challenge with no universal solution.

1.7.4 Implementations of Precision Oncology

One of the first precision oncology studies was performed by Von Hoff et al[159]. Of note, this study did not utilize a typical modern DNA gene panel, but rather employed gene expression microarray supplemented by IHC and FISH. Treatments were assigned based on a consistent rationale rather than pre-clinical or clinical evidence and included many chemotherapy agents. Treatment selection was open ended, in that there was no pre-selected panel of available therapies. Overexpression of a particular gene compared to normal tissue was considered a biomarker suggesting that gene should be targeted. The study met its primary endpoint, with 44% of patients having a PFS ratio > 1.3 (calculated as the ratio of the progression-free survival (PFS) on the prior therapy to the assigned therapy). Although the mechanism of treatment assignment was atypical, the study was non-randomised and highly biased, it established the feasibility of a precision

oncology program, where patients are tested and allocated to therapies in real time. Simultaneously, this study also highlights the many challenges in testing precision oncology approaches: appropriate patient selection; selection of pragmatic endpoints for demonstrating efficacy; establishing a biomarker strategy which is broadly applicable. These challenges remain in the current era.

There have been relatively few stand-alone clinical trials testing an end to end precision oncology strategy from performing an assay to delivering targeted therapies and assessing clinical outcomes. The primary difficulty of these studies is in pre-specifying the therapies that patients will receive based on genomic findings. By the necessity of making such a specification, these trials are testing one highly specific treatment algorithm. A prime example of this difficulty was the SHIVA trial [160]. This trial had a selection of treatments matched to genomic and immunohistochemistry findings. Many of the specified treatments were not particularly effective agents however, such as giving tamoxifen to cancers with hormone receptor expression that had received prior hormonal therapy. The results of SHIVA showed no advantage of this strategy compared to a standard of care control arm.

A more common approach is using an open-ended treatment allocation strategy, and evaluating the feasibility, response rate or a survival end point such as the PFS ratio. Response rate may be compared to non-personalised clinical trial therapies or standard of care. One of the largest studies was the MOSCATO 01 trial, which prospectively enrolled 1035 advanced cancer patients [161], of which 843 patients had a molecular profile generated, with 411 having an actionable finding

and 199 receiving matched therapy (19% of consented patients). Patients received matched therapies in the context of phase 1 & 2 trials which were not directly linked to the testing protocol. The primary end point was the PFS ratio, comparing the PFS on matched therapy to the PFS on the most recent therapy during which the patient experienced progression. A ratio greater than 1.3 was defined as a meaningful improvement and occurred in 33% of patients receiving matched therapy. In the overall consented cohort therefore, approximately 6% of patients experienced a benefit from molecular profiling. Although biased in many ways, a clear advantage of the approach in MOSCATO 01 was that following the return of molecular findings, patients could be allocated to clinical trials based on clinical judgment and discussion in a molecular tumour board. The low number of patients that received matched therapy also highlights the difficulties in readily accessing appropriate therapies.

To improve the proportion of patients accessing personalised therapies, more recent efforts such as the National Cancer Institute Molecular Analysis for Therapy Choice study (NCI-MATCH) [162] and Targeted Agent and Profiling Utilization Registry (TAPUR) [163] have adopted a "basket trial" strategy. These trials for patients with advanced cancers have broad inclusion criteria and are open at many sites across the United States. The initiation of testing and decision to participate in a matched therapy trial is led by the patient's clinician. Testing is not centralised, and results from any appropriately certified platform may be used. Both of these studies have multiple treatment arms with matched therapies, but also specify eligible tumour types for all treatment arms in the case of TAPUR, and some treatment arms in NCI-MATCH.

Interim results from NCI-MATCH showed a relatively low rate of accrual to therapy [164]. Of 645 patients successfully sequenced, 9% had a mutation matching an available therapy, and 2.5% received treatment. The proportion of patients matching a therapy is expected to improve as more treatment arms open. NCI-MATCH is planned to have over 20 treatment arms. The very low rate of patients in this study going onto therapy likely reflects the very broad inclusion criteria which do not place any restriction on lines of therapy or other clinical context. Nevertheless, NCI-MATCH and TAPUR have both demonstrated the popularity and feasibility of an inclusive genomic testing platform with subsequent basket trial allocation.

The chief message from these studies, and more restrictive protocols such as MOSCATO 01, is that improving the proportion of patients that might benefit from molecular profiling is the key challenge. Access to appropriate drug therapies is an ongoing problem which is highly dependent on the local treatment site. Pragmatic and judicious selection of patients for a precision oncology program will remain crucial to efficient utilisation of resources.

1.7.5 Precision Oncology in Breast Cancer

Arnedos et al undertook a pilot study in patients with metastatic breast cancer, using array CGH and Sanger sequencing of mutational hotspots in *PIK3CA* and *AKT1* [165]. Archival tissue of any sort was allowed. There were no therapies attached to the project. 108 patients were recruited over 3 years. 50% of patients had an actionable alteration, and 16% received a matched therapy. Of those that

did not receive a matching therapy, the common reasons were lack of drug access (21%), ineligibility for a trial (14%), death prior to receiving genomic results (5%), patient or physician choice (22%) and good disease control on an existing therapy (38%). 18 patients received targeted therapies, and 50% achieved a partial response or stable disease for ≥ 16 weeks.

The larger multi-centre SAFIR01/UNICANCER study employed the same technology as Arnedos et al, but performed fresh biopsies for all patients [166]. Results were reviewed in a tumour board prior to making treatment recommendations, and patients only switched therapy if they developed progressive disease. Of 423 patients enrolled, 407 (96%) had biopsies. The biopsy tissue was inadequate for 32% of patients, with bone biopsies most likely to fail. The protocol was subsequently amended to exclude patients with bone only disease. 46% of the whole cohort had at least one actionable alteration. Common alterations were *PIK3CA* mutations (25% of cases with actionable alteration), *CCND1* amplification (18%)and *FGFR1* amplification (12%). With the benefit of hindsight, *CCND1* amplifications are not actionable with current generation therapies, but were classified as such in SAFIR01 as it was unclear at the time if CDK inhibitors would be effective in cases with this alteration [167]. Fifty-five patients or 13% of the whole cohort received targeted therapy at the time of reporting. Of 43 evaluable patients, 13 (30%) had a partial response or stable disease. These results were notable, considering that only 2 genes had mutational assays. The high rate of enrolment on clinical trials with targeted therapy reflects the presence of a strong early phase clinical trial program and highly motivated patients and clinicians.

The AURORA study is a pan-European molecular screening program, aiming to enrol 1300 newly diagnosed metastatic breast cancer patients [168]. Tissue samples were processed and sequenced at a central facility. A 41- patients pilot phase has been reported. Amplicon sequencing with IonTorrent, a hybrid capture panel of 371 genes and SNP arrays were performed on each sample. Similar to SAFIR01, 63% of patients successfully underwent sequencing. 88% of patients had an actionable alteration as defined by the authors. AURORA found that IonTorrent sequencing had advantages in ease of use and rapid turn-around, but was inadequate for copy number.

These studies demonstrate that real time profiling of metastatic breast cancer is possible, with around two thirds of patients having a result returned successfully. The logistics of enrolling patients, delivering results and acting on them is a significant challenge, and there are a variety of trade-offs regarding the choice of technology platform.

1.8 Thesis aims

The aims of this work are as follows:

1. Implementation of a targeted sequencing panel for detection of targetable genomic alterations in advanced breast cancer (Chapters 3,4,5)
2. Analysis of tumour genomic heterogeneity in advanced metastatic disease using the CASCADE rapid autopsy program (Publication as Chapter 6)

Chapter 2: Methods – Clinical Sequencing Protocol

This chapter describes the methods applied to the SEGMENT study, a precision oncology program for patients with advanced breast cancer. An overview of the study design and rationale is presented in the next chapter.

2.1 Ethics approval

The SEGMENT study was approved by the Peter MacCallum Cancer Centre Human Research Ethics Committee. The study protocol and patient information and consent forms were submitted for approval by the Committee, which included a clinical geneticist. The consent form was modified in accordance with the recommendations of the Ethics Committee regarding the possibility of detecting germline findings. The project was approved on December 18, 2013. Any patient enrolling in SEGMENT was required to be reviewed by a clinician in person at the Peter MacCallum Cancer Centre.

2.2 Acquisition of samples for sequencing

Once a patient has signed the informed consent and information form, their medical records are accessed to catalogue all archival tissues that are available. Pathology reports are obtained and appropriate samples chosen based on the amount of tissue likely to be available and the clinical circumstances. The study cover letter explaining the purpose of the study and the tissue processing requirements is then sent along with the pathology report and remittance details to the pathology provider where the tissue is housed. If tissue has not been

received within 2 weeks, the pathology provider is contacted to follow up on the status of the tissue request.

Tissue receipt from pathology providers and processing for DNA extraction was performed by the Department of Pathology at the Peter MacCallum Cancer Centre. Tissue samples are given a unique processing identification number and diverted for sectioning and slide preparation if formalin-fixed paraffin-embedded (FFPE) tissue blocks have been sent, or directly for slide preparation if unstained slides have been sent. Standard processing includes cutting 20 5μm sections, staining and cover slipping one section with haematoxylin and eosin (H&E) and staining the remaining sections with methyl green. H&E stained slides are reviewed by an anatomical pathologist to confirm the presence of cancer and to identify areas of highest tumour density for microdissection. If a sample is pauci-cellular with less than 5% tumour cells on review the sample is discarded at this stage and alternative blocks or specimens sought. The molecular pathology laboratory performs microdissection of methyl green stained sections guided by the areas identified on the H&E slide. Microdissected tissue then undergoes DNA extraction with the QIAAMP DNA Blood Mini Kit (Qiagen). This involves digestion with proteinase K, then binding of DNA to a spin column followed by elution. DNA is typically eluted in 50μl of DNase-free water.

Extracted DNA is quantified by Qubit Fluorometric Quantitation (Life Technologies, Carlsbad, CA), using the dsDNA HS assay kit (Invitrogen, Eugene, OR). 1μl of extracted DNA solution is added to 199μl of buffer and fluorescent dye

solution. The fluorometer is calibrated with the manufacturer supplied standards with each batch of DNA quantification.

2.3 HaloPlex workflow

Sequencing library preparation and hybrid capture was performed in the Molecular Genomics Core of the Peter MacCallum Cancer Centre. A custom amplicon panel was designed covering genes of interest curated from the literature. The online SureDesign tool (Agilent Technologies, Santa Clara, CA) is used to specify genes of interest. Probes covering target regions are generated automatically by the tool. These probes are manually reviewed to ensure coverage of the desired regions. The amplicon panel is also reviewed by Agilent staff. Following approval, a batch of probe sets is ordered, along with library preparation reagents.

Target enrichment for sequencing with HaloPlex proceeds as follows. Extracted DNA is digested with a mix of restriction enzymes incubated at 37°C. Successful digestion is verified using the 2200 TapeStation automated electrophoresis device, with the expected appearance of 3 distinct bands. Digested DNA is then hybridised to HaloPlex target probes. The circularised probe-DNA hybrids are pulled down from solution using streptavidin-coated magnetic beads. A ligation step is performed to resolve gaps in circular DNA. The ligated probe-DNA hybrids are eluted from the beads and amplified with polymerase chain reaction (PCR) for the number of cycles recommended for each probe batch, followed by clean-up with AMPureXP beads. Successful target enrichment produces amplicons with a specific range of lengths, and this is verified using the 2200 TapeStation which

should produce a characteristic electrophoretogram. Prepared DNA is sequenced on an Illumina MiSeq.

Sequencing intensity data in 'bcl2' format were demultiplexed per sample and converted to paired end reads as 'FASTQ' files using bcl2fastq software (Illumina, CA, USA). The Agilent SureCall software (version 2.8, Agilent Technologies, Santa Clara, CA) aligns sequencing reads to the reference genome (GRCh37), and calls variants and copy number alterations. SureCall also excludes sequencing artefacts at the ends of the amplicon introduced by the library preparation steps.

2.4 SureSelect library preparation, hybrid capture and sequencing

A custom hybrid capture panel on the SureSelectXT Target Enrichment System was designed to cover genes of interest, updated from the first version of the panel on the HaloPlex platform, using the SureDesign tool. The capture probes are biotinylated RNA baits and arrive in a single tube master mix. This is kept at -80°C, and aliquoted into 2µl volumes for use in each hybrid capture reaction. Prior to hybrid capture, a DNA library is prepared using the KAPA Hyper Prep Kit (Kapa Biosystems, Inc, Wilmington, MA), which is compatible with SureSelectXT.

DNA for each sample was fragmented using a Covaris sonicator, using the settings recommended in the SureSelectXT protocol. In the initial phases of the library preparation the fragmentation profile was inspected by running DNA on the Agilent TapeStation (Agilent Technologies, Santa Clara, CA). This step was later omitted to conserve DNA when it was observed the fragmentation profile was

consistent. Reference fresh frozen samples were prepared in parallel to FFPE samples, using standard settings on the Covaris sonicator.

Following fragmentation, DNA end repair and A-tailing was performed with the KAPA Hyper reagents. SureSelectXT adapters were then ligated to DNA fragments, followed by clean-up with Agencourt AMPureXP beads (Beckman Coulter, Brea, CA). Pre-capture amplification is then performed with the KAPA HiFi polymerase and SureSelectXT primers, followed by clean-up with AMPureXP beads and quantification of the post-amplification library with the Qubit fluorometer. The target pre-capture DNA yield was 750ng. At this point, if the amplification was suboptimal and the library yield was inadequate it was discarded to avoid wasting capture reagents. In the pilot phase of the study, the number of PCR cycles was optimised to avoid over-amplification which results in a high duplication rate in subsequent sequencing. The stringency of the bead purification was also relaxed to avoid denuding highly fragmented FFPE DNA.

As the DNA input amounts were often low, the concentration of the prepared library was commonly overly dilute. A vacuum concentrator was used to increase the concentration prior to hybrid capture, by lyophilizing the sample at 45°C and resuspending in nuclease-free water. Capture probes were thawed on ice before mixing with the concentrated library and capture reagents on a PCR plate maintained at 65°C. The hybridisation mixture was incubated for 16 hours at 65 °C with a heated lid at 105 °C, ensuring PCR tubes were tightly capped and sealed to avoid desiccation. Following incubation, DNA bound to biotinylated RNA probes was pulled down using streptavidin-coated magnetic beads. On-bead PCR was

then performed to add index tags, before elution of captured DNA and clean-up with AMPureXP beads. Indexed libraries were pooled for sequencing a MiSeq, NextSeq500 and in some cases a HiSeq2000 (Illumina, San Diego, CA). After a pilot phase, the multiple pipetting steps required for capture were automated on the Agilent Bravo liquid handling robot. It was noted that less experienced users produced sequencing libraries of inferior quality. It was therefore decided that the Molecular Genomics Core staff Timothy Semple and Sreeja Gadipally would perform library preparation going forward.

2.5 Generation and processing of sequencing data

Sequencing intensity data in 'bcl2' format were demultiplexed per sample and converted to paired end reads as 'FASTQ' files using bcl2fastq software (Illumina, CA, USA). This was conducted on the high-performance computing cluster. FASTQ files were aligned to the human reference genome (version Genome Reference Consortium GRCh37) using the BWA-MEM algorithm as implemented in the bwa software package [169]. Default settings were used, resulting in a binary alignment/map (BAM) file for each sequenced sample.

BAM files were pre-processed prior to variant calling. Duplicate reads were marked and removed with the 'MarkDuplicates' function in Picard tools [170]. Duplicate reads result from PCR amplification of limited amounts of DNA during library preparation, or misregistration of amplification clusters during sequencing, the latter termed 'optical duplicates'. Duplicate reads are identified by reads that have the same starting positions and identical sequences. Reads with low mapping quality were also removed. When a read maps equally well to

different regions of the genome, the default behaviour of the read aligning algorithm BWA-MEM is to randomly allocate the read to one position and assign mapping quality zero. Some genes have pseudogenes, which are non-functional versions of the protein coding gene that have high degrees of sequence homology [171]. Depending on the sequencing read length and length of the DNA fragments, a read may map to the targeted gene and one or more of its pseudogenes in an ambiguous fashion. The net result includes elevated coverage of such genes, and the presence of spurious variants. Such ambiguous or misaligned pseudogene sequences may be identified by low mapping quality scores. BWA-MEM assigns a mapping quality score to each read. This score is higher if there is a unique optimal alignment, lower if there are more suboptimal alignments, and lower if the individual base quality scores in the read are lower. The score is calculated as -10log10 of the probability the mapping is wrong. A mapping quality of 10 or more translates to a probability of 90% or more that a read is mapped correctly. Using the samtools software [172], reads that were marked as duplicates or had a mapping quality less than 10 were removed from each BAM file. BAM files were then co-ordinate sorted and indexed with samtools.

Sequencing metrics were generated for each BAM file using the 'CollectHsMetrics' function in Picard tools. The advantage of this tool is that it calculates coverage which does not count overlapping reads twice. This phenomenon arises when highly fragmented FFPE DNA is shorter than the read length of the sequencer. Reads from each end of the DNA fragment thus overlap in the middle of the read. This overlap is considered redundant, since it is sequencing the same unique DNA fragment twice. Naïve coverage calculations by counting all reads overlapping a

region may therefore overestimate the true coverage significantly for a highly fragmented sample, obscuring compromised coverage. CollectHsMetrics avoids this problem by counting overlapping reads once only. Key metrics include the duplication rate, defined as the proportion of reads that are duplicates, the on-target rate, or the proportion of reads overlapping with target regions, and mean coverage of target regions.

2.6 Variant calling and curation

Variants were called with VarDict [173]. The input to VarDict is the BAM file and the capture regions. As well as calling single nucleotide variants, VarDict performs local realignment of reads to aid in calling indels. If a read contains an indel at the end or beginning of the read, the read undergoes 'soft-clipping' during alignment, where the aligner effectively ignores the mismatch introduced by the insertion or the lack of matching sequence caused by the deletion. The soft-clipped ends of the read are left as unaligned overhanging sequence. VarDict rescues the soft-clipped sequence and utilises it to assist in calling an indel. This approach also enables the calling of indels which are longer than the read length and facilitates calling of substitutions, where a portion of sequence is replaced with a different sequence which may be longer or shorter in length. In addition, VarDict correctly handles overlapping ends of the read pair as discussed above. This would otherwise spuriously elevate the reads in support of a variant and introduce a dependency between the allele frequency and the fragment size distribution in a sample.

VarDict was run with default settings, but with a loose allele frequency filter down to 1%. Variant calls were annotated with ANNOVAR [174]. The function of

ANNOVAR is to assign the variant to a gene and categorise the nature of the variant as either a synonymous SNV, a nonsynonymous SNV, a nonsense or stopgain variant, a splicing variant, a frameshift indel, a frameshift substitution, a nonframeshift indel or a nonframeshift substitution. Other categories such as those affecting promoter sequences are ignored. ANNOVAR also determines the transcripts associated with the gene which are affected by the variant and infers the resultant protein alteration. The variant is searched for in various databases including the 1000 Genomes project [175], the Exome Sequencing Project [176], the Exome Aggregateion Consortium (ExAC) [177], Single Nucleotide Polymorphism Database (dbSNP) [178], COSMIC [179]and ClinVar [180]. As an additional step beyond ANNOVAR, the uniqueness of the genomic region containing the variant is determined. Uniqueness refers to how unique a 35bp sequence is across the whole genome. Sequences which occur in multiple places in the genome are a source of artefactual variants due to mapping errors. Pre-calculated uniqueness scores are obtained from the Duke Uniqueness track available at the University of California Santa Cruz Genome Browser [181].

Variant curation was conducted according to the following principles:

1. Variants with an allele frequency less than 10% are removed, except variants that are found in COSMIC [179] are retained down to 5%.
2. Variants with a detectable prevalence in the 1000 Genomes database or the Exome Sequencing Project are removed, as these are presumed to be germline variants.
3. Synonymous variants are removed.

4. Variants in intronic or intergenic regions which do not affect splicing are removed.
5. Variants with a frequency of > 0.1 % in ExAC are removed (spurious variants). This is to capture germline variants but also to remove variants which are likely to be sequencing artefacts, as variants in ExAC are not in general curated or reviewed.
6. Flanking sequences around variant are reviewed, and the variant excluded if there are homopolymers of 4 or more bases or repetitive trinucleotides or dinucleotides.
7. Variants in regions of the genome with low mappability are excluded.
8. VarDict metrics including the signal to noise ratio and strand bias test are examined for outliers. A low signal to noise ratio flags that reads across the variant position have a high number of mismatches, potentially indicating a problematic genomic region or sequencing errors. Strand bias refers to asymmetry in the representation of a variant in forward and reverse reads, and may indicate PCR errors.

2.7 Copy number calls

CopywriteR (version 2.12.0) was used for final copy number calls [182]. Bam files which had duplicates removed and reads with low mapping quality removed were used as input. Self-normalisation was used for all FFPE samples. CopywriteR was run with the default settings with sliding bin widths ranging from 50 kilobases to 1000 kilobases. Segmentation was performed using CopywriterR, which employs circular binary segmentation as implemented in the DNACopy R package on the log ratio values [183]. Default segmentation parameters were used. The final

output for each sample is a tabular file with the genomic coordinates of each segment and an associated mean log ratio. Whole genome copy number plots were examined to detect excessively noisy samples or aberrant calls. CopywriteR also reports the median absolute deviation of whole genome log ratios as a measure of noise, where a higher MAD signifies higher noise levels.

To generate gene level copy number data, genes were associated with each copy number segment according to the genomic coordinates. In some cases a gene spanned multiple segments with different log ratios. Examination of these cases revealed this was invariably due to slightly different log ratios for adjacent segments, where the difference between log ratios was < 10%. In these cases, a log ratio value for the gene was calculated based on the mean of the log ratio of each segment. The Ensembl release 75 gene models for protein coding genes were used, based on the Genome Reference Consortium GRCh37 reference genome. Genes with a mean log ratio of greater than or equal to 0.7 were classified as amplified, and genes with a mean log ratio of less than or equal to -0.7 were classified as deletions.

The alternative copy number calling method CONTRA was run on pilot samples using default settings [184]. This analysis was performed by Dr Jason Li (Bioinformatics, Peter MacCallum Cancer Centre). Process matched fresh frozen diploid samples were used as reference samples.

2.8 Determining the significance of sequencing results

For each mutation and copy number alteration, the first consideration was whether the alteration is known to be significant. This is surveyed by checking for the variant in the following databases and knowledge bases:

1. COSMIC [179].
2. OncoKB, a curated knowledge base created by experts with clinical evidence [138].
3. CivicDb, a crowd sourced knowledge base with clinical evidence [139].
4. Reported as a recurrent alteration in the largest breast cancer sequencing datasets from TCGA and METABRIC accessed via cbioportal [185,186].
5. Reported as a recurrent alteration in Pan-cancer datasets from TCGA accessed via cbioportal.

The appearance of a variant as recurrent is conducted by examining its frequency in relation to other mutations in the gene. Usually this is not a large signal, with a particular variant appearing in <10 cases out of several thousand. Statistical testing and control for factors which impact the chance of a recurrent mutation are not taken into account at this screening step.

If a variant is flagged as a known alteration through this process, it is retained for further analysis and reporting. If not a more involved curation process is conducted. Firstly, it is considered whether the variant in question was likely to impact gene function. For mutations, frameshift indels and substitutions and stop mutations are most likely to affect gene function, however the position in the gene remains important, since truncation of a gene close to the end of the amino acid

sequence is less deleterious. For non-frameshift indels and substitutions, and non-synonymous single nucleotide variants it is less clear that there will be an impact on gene function. For mutations, the category of the mutation in relation to the known gene function is also considered. Thus, a truncating mutation early in the sequence of a known oncogene, such as *PIK3CA*, is unlikely to be significant, whereas the same mutation in a gene such as *BRCA1* is likely to be significant. For copy number changes, the direction of change in copy number is considered in relation to gene function. For example, amplification of oncogenes and deletions of tumour suppressor genes are more retained for further analysis, but the converse, such as amplification of *TP53*, are discarded.

For mutations where the impact of the mutation is not clear, further corroborating information is sought. Mutation Aligner is a method which can provide additional power to detect recurrent variants [187]. This is achieved through the understanding that functional domains of proteins are re-used across different genes. Although the overall sequence of two genes may be very different, they may each contain highly conserved functional domains. By aligning these functional domains, more data is accrued on the potential impact of mutations at specific sites within the domain. In addition to this, the literature is surveyed for cases where in vitro studies have examined particular mutations through mutagenesis and assays of gene function and cell phenotypes with the mutant proteins[188]. High throughput mutagenesis screens, where a large number of mutations are introduced into a gene and functional impact is determined with an assay in a high throughput fashion are another source of information which may assist in determining the significance of a mutation [189,190].

The final step of determining the clinical significance of an alteration is the most challenging due to sparse data. The gold standard level of data, which involves a randomised trial with prospective determination of treatment groups according to the genomic alteration, and randomisation of patients with and without the alteration to a particular treatment, is rarely available. Less robust clinical trial data include planned and unplanned subgroup analyses of the association between genomic findings and treatment outcomes in randomised trials, or similar analyses in single arm trials where objective response rate is the outcome. Uncontrolled studies where clinical outcomes are compared to historical controls are another less robust source of clinical annotation, and may include single arm studies where patients are selected based on the genomic alteration, case series or individual case reports. Ideally any of these forms of evidence will be available for breast cancer, but if this is not available then data from other tumours may also be useful.

As has been noted by many other groups, even allowing for these less robust sources of evidence, many genomic alterations remain of uncertain significance. In these cases, experimental data was examined for evidence which could contribute to clinical annotation. Gao et al performed a study with approximately 1000 patient derived xenografts from multiple cancer types which had comprehensive genomic characterisation and were treated with a variety of targeted and untargeted anti-cancer compounds, effectively performing an *in vivo* clinical trial [191]. The results for each xenograft along with the genomic profile are publicly available, which may be used to suggest that a specific genomic

alteration is associated with response or resistance to a particular drug target. Forty-three breast cancer xenografts were screened with 21 treatments, included monotherapy and combination therapies. Bruna et al conducted a similar study using 83 breast cancer patient derived xenografts screened with 104 compounds, although only 20 xenografts with full genomic profiles and drug screening data are available [192]. Finally, projects which have screened cancer cell lines that have had genomic characterisation with multiple compounds are an additional resource which may provide data to support clinical annotation, although the applicability to human disease is questionable and the compounds are often far from clinically useful and data are sparse [133,193,194].

2.9 Collection of clinical data

Clinical data relating to participants was collected from the electronic medical record and imported documents. Data were stored in a password protected restricted access Microsoft Access Database (Microsoft Corporation, Seattle, WA). Follow up was performed at 3 monthly intervals or until the death of the patient. Collected data included demographic information, details of the breast cancer diagnosis, surgery, radiotherapy and systemic therapy, pathology and imaging results.

2.10 Statistics and analysis software

All data analyses including statistical tests were performed using the R Language and Environment for Statistical Computing [195]. The following table lists R packages used.

Package	Version
ggplot2 [196]	3.2.1
dplyr [197]	0.8.3
tidyr [198]	1.0.0
magrittr [199]	1.5
lubridate [200]	1.7.4
ROCR [201]	1.0-7
readr [202]	1.3.1
stringr [203]	1.4.0
GenomicRanges [204]	1.36.0
GenomicFeatures [204]	1.36.4
GenomicAlignments [204]	1.20.1
ensembldb [205]	2.8.0
rtracklayer [206]	1.44.2
CopywriteR [182]	2.12.0

Chapter 3: SEGMENT Design and Pilot Phase

3.1 Overview of project

The clinical sequencing portion of this project was conceived as a prospective study, where patients would consent to the acquisition of tumour samples which would undergo next generation sequencing. These data would be analysed in real time and returned to the treating clinician with a view to informing future therapy. The key elements of the design will be reviewed in this chapter, and are summarised here and in Figure 3.1:

1. The patient population targeted for recruitment;
2. The use of FFPE or fresh frozen tissue;
3. The sequencing platform;
4. The choice of genes;
5. Whether tumour-only or paired tumour-normal sequencing would be performed;
6. The software pipeline used to analyse sequencing data;
7. Whether clinical trials would be included as part of the protocol;
8. The sample size;
9. The selection of study endpoints.

Patient population

As has been discussed in chapter 1, breast cancer is a common condition and the number of patients with incurable metastatic disease is considerable. The breast cancer phenotype is of primary importance to the natural history and treatment

options available to the patient, as well as the availability of clinical trials. The target population was chosen to be patients with metastatic incurable breast cancer. These are the patients where precision oncology is likely to play a role, since they receive a variety of therapies of varying but inevitably temporary effectiveness. It was acknowledged that particularly at the beginning of the study, there would be a delay in how soon results could be generated, and thus patients should be enrolled when they are relatively well prior to the development of major organ dysfunction such that life expectancy was likely to be limited. It was also expected that HER2-positive patients would be less suitable candidates in general since they already have a highly effective anti-HER2 therapies available including the anti-HER2 monoclonal antibody, trastuzumab, and other related treatments such as pertuzumab, lapatinib and trastuzumab emtansine.

On a related point, it was considered if enrolled patients should be required to have a new biopsy or if archival tissue could be used. New biopsies have the advantage of being current and would allow sequencing to be performed on the most clinically relevant disease. Nevertheless, they require an invasive procedure, and many breast cancer patients do not have tissue suitable for biopsy. Since the study had a strong focus on feasibility and real-world application, it was decided that a new biopsy was not mandatory but could be performed if clinically indicated and safe to do so.

The use of FFPE or frozen tissue

It is clear that fresh frozen tumour tissue is the preferable source of DNA for sequencing. At the conception of SEGMENT, it was known that formalin fixation

had various effects on DNA which introduced noise and reduced the likelihood of sequencing success. However, targeted sequencing had been successfully performed by other groups on FFPE material using a variety of platforms including Sanger sequencing, amplicon-based sequencing and hybrid capture methods, albeit with certain caveats. These included the introduction of sequencing errors at low allele frequency and reduction in sequencing efficiency. In this project, allowing the use of FFPE material had one crucial advantage in that existing biopsies and surgical specimens collected prior to enrolment on the study could be used for sequencing without the need for new biopsies. It was considered that this was a sufficient practical advantage that FFPE material should be permitted. It also enabled the secondary aim relating to the concordance of genomic findings across multiple samples from the same patient. Additionally, it was known from the work of others that genomic profiling of FFPE samples was feasible [123,207].

The sequencing platform

Although at the current time circa 2018, the platform of choice for sequencing of either the whole exome or a targeted subset has focussed around hybrid capture technology, at the inception of this project there were several competing platforms. With the focus on actionable variants and knowledge from the large sequencing projects discussed in the introduction, it was clear that sequencing a small panel of genes rather than a whole exome or whole genome was the preferred approach. This requires a strategy for enriching DNA from the regions of interest, which was the major decision in designing the study. Available enrichment strategies included PCR enrichment strategies and sequence specific

selection. PCR enrichment strategies, also known as amplicon strategies, such as Illumina's TruSeq product or RainDance Technology's RainStorm enrich target sequences using flanking PCR primers and employ a variety of techniques to enable multiplexing. Enriched sequences are then sequenced with the method of choice. Amplicon strategies have the advantage of requiring relatively low amounts of input DNA, as low as less than 50ng, and thus have the ability to produce data from samples which are not suitable for other approaches due to poor quality or low input. They have several disadvantages particularly with regards to FFPE tissues, however. As is evident from the enrichment strategy, a large number of PCR cycles are used, which worsens the propensity of formalin fixation to introduce artefactual changes [208–212]. Amplicon panels are a 'hotspot' technology, effectively requiring selection of mutant loci prior to conducting the assay, which is less desirable for projects where the expected mutations are incompletely known *a priori*. The focal enrichment with amplicon technology is also less accurate for copy number compared to approaches that target the whole gene.

The IonTorrent platform is a related amplicon-based sequencing method [213,214]. Amplicons of interest are amplified with PCR, but sequencing is performed on the IonTorrent instrument rather than sequencers employing optical detection such as Illumina machines. IonTorrent employs a specialised transistor to detect the addition of bases during 'sequencing by synthesis'. It has the advantages and pitfalls of amplicon methods as described above. It is generally lower cost and less labour intensive compared to other amplicon methods which

are typically sequenced on Illumina machines, but suffers from inaccurate calling of homopolymer runs [213].

Amplicon methods in general have limitations on the number of targets due to the difficulties in multiplexing a large number of different PCRs [215,216]. HaloPlex is a method designed to overcome this limitation by avoiding the use of selective PCR for enrichment. In this method, DNA is fragmented using restriction enzymes, generating DNA fragments with predictable sequence characteristics at the head and tail of the fragment. These fragments are then captured using sequence specific probes that bridge the ends of the fragment, producing a circularised DNA fragment which is then purified from the total DNA pool [217,218]. Universal PCR primers incorporated in the probe allow generic amplification of any DNA enriched in this manner. HaloPlex thus permits the low input requirements of amplicon sequencing but enables more targets to be enriched such that all exons of a gene may be covered. Later iterations of HaloPlex underwent modification to improve performance in FFPE tissues by capturing sense and anti-sense strands for each region [219].

The final category of sequencing methods under consideration is hybrid capture [220]. A pool of capture probes are designed to hybridise to the DNA regions of interest. These probes commonly consist of biotinylated RNA, and during the library preparation process are hybridised to randomly fragmented DNA from a sample. Pull down of the biotinylated RNA probes also delivers the DNA from regions of interest, which is then sequenced. Compared to amplicon methods, fewer PCR cycles are required and the sequencing error rate thus lower. Hybrid

capture is however relatively labour intensive and involves multiple library preparation steps where some DNA is lost, thus requiring higher input DNA. It also the most expensive due to the relatively high cost of biotinylated RNA probes. In addition, hybrid capture is less selective than amplicon methods, and the sequenced library contains significant amounts of DNA outside the regions of interest [220,221]. Sequencing this irrelevant DNA translates into higher overall sequencing costs to achieve the necessary coverage. This problem is more prevalent when the region targeted for enrichment is < 1 megabase.

Since a larger number of targets and more accurate copy number were considered desirable for this study, HaloPlex was chosen as the sequencing platform in the first iteration of the study as it had cost advantages over hybrid capture. HaloPlex has the purported advantage that a custom gene panel will work off the shelf, whereas this is not guaranteed with custom multiplex PCR amplicon panels where optimisation of primer design is often necessary.

The choice of genes

Selection of target genes was determined according to the existing literature at the time, derived mostly from the series of large-scale whole exome studies published in 2012. The list was refined to focus on genes where actionable alterations were considered more prevalent. The initial panel was influenced heavily by studies in primary tumours due to the availability of the data at the time. A notable omission in light of this was *ESR1*, which is almost exclusively seen in metastatic hormone-positive breast cancers. The breast cancer driver list was supplemented with known germline susceptibility genes and drivers from other cancer types as an

exploratory exercise. The total target region was also constrained by the tiered pricing of HaloPlex custom panels. The list of 69 genes is shown in Table 3.1, with breast cancer drivers or germline susceptibility genes underlined.

The HaloPlex platform provided tool SureDesign takes genes of interest as input and designs an amplicon panel to target them. The resulting panel contained 29,632 amplicons covering 188,496 bases across 1506 exons in the target genes. 164 bases could not be covered by the design, but these regions were not in eloquent regions of the target genes.

Whether tumour-only or paired tumour-normal sequencing would be performed

The advantages of tumour-only sequencing are clear in that the cost per patient is halved and collecting germline samples is not required. The disadvantage comes from introducing additional complexity to the analysis such that somatic mutations must be distinguished from germline mutations. Obtaining allele specific copy number also requires a matched germline sample. Since this study was focussed on detecting actionable genomic alterations, which are largely known prior to conducting the assay and very rarely appear in the germline setting, it was felt that distinguishing somatic from germline variants would be less of a problem. The exceptions are *BRCA1* and *BRCA2* mutations which do occur in both the germline and somatic setting, however since somatic mutations in these genes is rare in breast cancer, the discovery of any such mutation will automatically result in diagnostic quality germline testing in any case. With the

plans for a relatively small panel of genes, calling allele-specific copy number was also less relevant.

The software pipeline used to analyse sequencing data

At the time that HaloPlex was chosen as the sequencing platform, it was relatively new and there was not a well-defined open source software pipeline for calling variants and copy number changes. Vendor supplied software called SureCall was therefore used for this purpose. The variant caller used in SureCall is SNPPET [222]. HaloPlex introduces recurrent artefacts at the ends of reads, and SureCall accounts for this fact when calling variants. SureCall also implements alignment of sequencing reads to the reference genome using BWA.

Whether clinical trials would be included as part of the protocol

It was decided to structure SEGMENT as a sequencing protocol alone, not including clinical trial therapies. To do so would dramatically increase the complexity of the project and was not considered appropriate as the sequencing technology had not yet been piloted.

The sample size

The initial sample size was set at 80 patients, to be enrolled over an 18 month period. The aim was for all patients to have a primary and metastatic lesion available for sequencing. This relatively small number was justified by the uncertain performance of the technology platform at the outset, and the available funding and staff at the time. It was expected that approximately 30% of cases

would contain a *PIK3CA* mutation, and that this would allow the primary-metastatic concordance of this mutation to be assessed.

The selection of study endpoints

Given the decision that clinical trial therapy would not be included as part of the protocol, it was deemed that clinical outcomes would not be a primary or secondary endpoint in the trial. The primary objective of the study was to determine the feasibility of characterising genomic alterations in real-time in FFPE samples obtained from patients. The primary and secondary endpoints were thus conceived as follows:

Primary endpoint:

Percentage of successful genomic profiles completed for primary and metastatic lesions. The number of patients signing consent will serve as the denominator.

Secondary endpoints:

1. Tissue acquisition feasibility
2. Sequencing analysis feasibility on DNA derived from FFPE tumour specimens
3. Concordance between primary and metastatic lesions
4. Number of patients that go on to receive targeted therapies guided by genomic analyses, stratified by stage of disease.

3.2 Evaluation of the HaloPlex sequencing platform

The protocol incorporated a pilot phase of 25 patients recruited on the SEGMENT study, designed to assess technical and logistic feasibility of the sequencing

platform. In addition to this cohort, high quality fresh frozen reference DNA was also used. This was obtained from the Genome in a Bottle (GIAB) consortium [223]. These DNA samples originate from immortalised lymphoid cells which are diploid and non-neoplastic and have been extensively characterised with whole genome sequencing. Finally, a separate cohort of 16 primary and metastatic breast cancers was obtained from a collaborator (D Fumagalli, Institut Jules Bordet, Belgium). These samples had already undergone mutational analysis, and 13 had multigene FISH for copy number.

Sequencing metrics for these three cohorts are presented in Table 3.2. Sequencing was successful for all samples, achieving the target of approximately 4 million reads. Sequencing efficiency was better for the GIAB DNA, with a higher mean and median read depth than the FFPE samples. The fragment length, calculated as the number of bases from the 5' end of the first read of a paired end read to the 3' end of the second read, was shorter from FFPE samples as expected. The target region for all samples appeared well covered on average. One major difference between the GIAB DNA and FFPE DNA was the percentage of 'sequenceable' regions with zero coverage, which was around 1% in GIAB, 36% in the collaborator cohort and 15% in the SEGMENT cohort. In the parlance of the HaloPlex protocol, 'sequenceable' refers to genomic regions which are covered by amplicons. The target regions are a subset of these sequenceable regions. This metric therefore represents a measure of amplicon failure where an amplicon has failed to produce any DNA during library preparation. It is known that FFPE samples have high rates of amplicon failure, raising concerns about whether mutation hotspots could have poor coverage.

To investigate this further, the depth of coverage in the most commonly mutated loci in breast cancer found in the TCGA were examined [68]. In an amplicon panel, the desirable read depth is in the 500 – 1000 read range. As can be seen in Figure 3.2, even in high quality reference GIAB DNA, many mutation hotspots of the targeted genes were covered well below this level. The locus corresponding to the 3rd most common mutation, *PIK3CA* E542K is poorly covered, as are many others. In FFPE tumour samples the problem is even more pronounced, as shown in Figure 3.2B, where coverage of *PIK3CA* E542K drops to zero in some samples.

Looking at the helical domain of *PIK3CA* in more detail, screenshots of Integrated Genomics Viewer are presented in Figure 3.3 for GIAB DNA, and in Figure 3.4 for a FFPE breast cancer sample. It is clear that despite multiple amplicons spanning the region, there is a steep decline in reads at the E542 codon position in the reference sample. In the FFPE sample, coverage of the is locus is entirely absent.

These problems may arise because the HaloPlex technology relies on circularisation probes that target DNA fragments with probes that overlap short segments of the 5' and 3' ends of the fragment generated by restriction enzyme digestion of genomic DNA. If the DNA is already fragmented such that a restriction enzyme site is missing or incomplete, a valid digested fragment will not be generated and captured by this technology.

Further evidence for this can be seen in examining the copy number data. Agilent supplies proprietary software which relies on a reference diploid sample to infer

copy number gains or deletions by comparing the normalised read depth between the reference and the tumour sample. Figure 3.5 shows the correlation between copy number determined by HaloPlex and copy number from FISH for FFPE breast tumour samples. The genes on the HaloPlex panel that were tested via FISH are *AKT1, AKT3, AR, CCND1, EGFR, ERBB2, FGFR1, FGFR2, INPP4B, MAP2K4, MET, PIK3CA* and *RB2*. It is clear that the results are roughly correlated for amplifications, but HaloPlex calls many spurious deletions in genes that are diploid or show gains by FISH. This 'gene dropout' suggests that target regions in FFPE samples are recovered poorly in some cases, consistent with impaired amplification of fragmented DNA.

Towards the end of 2015 the suboptimal performance of HaloPlex on FFPE tissues was noted by the research group that originated the method [219]. A number of modifications to probe design and library preparation were introduced to improve performance. These modified protocols were not available at the time the pilot failed.

3.3 Redesigned platform

The experience with HaloPlex highlighted that there existed significant variability in sample quality, and that the sequencing technology should be robust across a wide range of sample quality. Sequencing reagent costs were also falling with the introduction of the Illumina NextSeq. Hybrid capture was selected as the replacement for HaloPlex, using the Agilent SureSelect platform which allows custom panel design. This opportunity was taken to revise the gene list with the most important edition being *ESR1*, which is commonly altered in metastatic

hormone receptor-positive cancers that have received endocrine therapy. The performance of hybrid capture panels degrades with smaller capture regions, so the gene list was supplemented with rare inherited susceptibility genes. The redesigned panel contained 363,233 bases across 2610 exons in 107 target genes. 2694 bases in discontiguous regions could not be covered by the design. The revised gene list is shown in Table 3.3.

As before, a pilot cohort of samples incorporating FFPE breast tumours and GIAB reference DNA was run using the redesigned panel. The first run included 4 fresh frozen diploid DNA samples (2 germline samples and 2 GIAB samples). The next 3 runs included an additional 2 GIAB DNA samples, 1 germline DNA sample, and 28 FFPE tumour samples. Sequencing metrics by sample type are presented in Table 3.4. The FFPE samples were sequenced at higher depth. Notable is the very low rate of targets with zero-coverage.

Key performance metrics are presented in Figure 3.6. Duplicate reads are PCR copies of the original DNA fragment and therefore non-unique. They can be determined by having identical start and end coordinates in the genome, which is otherwise a vanishingly rare event with ultrasonic fragmentation protocols. A high duplicate rate reflects input DNA with a low number of unique DNA fragments which successfully undergo PCR. This is commonly related to low input amounts, or input DNA which is damaged and resistant to polymerase activity. Duplicate rate is shown in Figure 3.6A, and although variable shows acceptable levels. The capture efficiency can be measured using the proportion of bases in target regions compared to non-target regions. Figure 3.6B shows bases in the

exons of target genes. Figure 3.6C shows bases in the regions covered by baits, which necessarily extend beyond exons. The relatively low capture efficiency for target regions is a limitation of small capture panels. Considering that the capture region of some 300 kilobases is approximately 0.01% of the total genomic DNA, an on-target rate of around 10% represents a 1000-fold enrichment.

Coverage of hotspot loci in the redesigned panel are shown in Figure 3.7. Samples in each run were sequenced to different depths, hence there is some variation in coverage. These results are more acceptable than HaloPlex, with no zero-coverage loci. *CDH1* and *ERBB2* loci are covered poorly in some cases, which originate from samples sequenced before the protocol was optimised.

On the basis of these pilot samples, the manufacturer's protocol was modified (Figure 3.8). Firstly, the settings on the Covaris Ultrasonicator device were relaxed to prevent over-fragmentation of FFPE DNA. Secondly, an alternate pre-capture library preparation kit was used (KAPA Hyper) which is optimised for FFPE. KAPA Hyper involves fewer bead purification steps and results in less DNA loss compared to the Agilent protocol. This permits lower input DNA, with as low as 50ng being possible. Another modification to deal with more fragmented FFPE DNA was altering the Solid Phase Reversible Immobilisation (SPRI) bead concentration to avoid discarding smaller DNA fragments.

3.4 Variant calling

When the panel was redesigned, an appropriate variant calling method was also reconsidered. The vendor supplied SNPPET caller software was poorly

documented, infrequently updated, not open source, and also required the use of a desktop program that did not allow batched processing of large numbers of samples. It was thus abandoned. In considering other options, there were in fact a paucity of variant callers tailored to tumour only sequencing involving small panels in late 2014 or early 2015. The must commonly used variant caller in tumour studies at the time was MuTect, but it has no native tumour-only mode and is not adapted for higher depth sequencing [224]. It can be used without a matched normal, but requires a pool of normal samples for estimating noise, which requires a number of processed matched normal samples and also introduces the problem of fresh frozen reference samples being used to estimate parameters applied to FFPE samples. VarScan2 was another commonly used tumour variant caller [225], but suffers from poor sensitivity at lower allele fractions, which is a disadvantage in real world tumour samples [226,227]. These variant callers also do not call indels, which necessitates the use of a separate tool such as pindel [228]. VarDict is a combined SNV and indel caller designed for higher coverage small panels [229]. It performs local realignment of reads around indels to improve sensitivity and is compute and memory efficient. Benchmarking studies using samples with known variants and *in silico* 'virtual tumours' found that it was robust at different levels of coverage and maintained reasonable sensitivity for SNVs in increasingly impure tumour samples, while at the time performing best in class for indel detection [229,230].

To validate the variant calling pipeline involving VarDict as implemented in the SEGMENT pipeline (see methods section), the panel was run on 19 FFPE ductal carcinoma in situ samples provided by a collaborators (Dr Jia-Min Pang and

Professor Stephen Fox, Peter MacCallum Cancer Centre). Dr Jia-Min Pang kindly performed Sanger sequencing of 35 distinct mutations detected in breast cancer related genes on the redesigned SEGMENT panel. As some mutations were recurrent, validation was performed for 42 mutations across samples, and included 10 indels and 32 single nucleotide variants, with a variant allele frequency down to 10%. 16 of 42 (38%) mutations had an allele frequency less than 30%. 41 of 42 mutations were validated, a rate of 97.6%. Validated mutations included a 25 base pair deletion in *TP53* and an 11 base pair insertion in *GATA3*, showing that larger indels are detectable with VarDict. Validated mutations are shown in Table 3.5.

The mutation which failed to validate was in *FGFR1* with an allele frequency of 21% and consisted of a thymine to cytosine transition. The variant was reviewed in Integrated Genomics Viewer and was free from extensive sequence artefacts or low mapping quality reads and was not in a region with repetitive sequences, homopolymers or sequence motifs associated with sequencing errors. It does not appear that the unvalidated *FGFR1* mutation was due to an error in the software pipeline. T to C transitions are not typical of formalin induced artefacts, but PCR amplification of a formalin induced artefact remains a possibility [231]. It is also possible that the Sanger sequencing result was a false negative, a phenomenon which has been noted by others [232].

3.5 Copy Number

Copy number analysis for targeted sequencing falls into two broad categories: On-target depth of coverage approaches and off -target depth of coverage approaches.

On-target approaches count reads in target regions and compare to a diploid reference, which may be a matched or unmatched normal sample or a pool of unmatched normals. As well as copy number of the gene in question, the depth of coverage depends on the amplification characteristics of the sample (i.e. the number of PCR cycles pre capture in particular), the efficiency of hybrid capture in the sample, and GC bias and AT dropout. In practice, it can be difficult to account for all these factors in a useful manner. In the case of a reference sample, usually these diploid samples have both higher quality and quantity of input DNA, and have quite different sequencing characteristics as seen in the preceding sections. The use of such a diploid reference also requires the choice of a sample to use for the reference, and it is not clear to what extent the choice of a particular sample may influence the results. By the same token, it would be non-sensical to use a FFPE reference sample, since this effectively adds more noise to the system which will not improve normalisation.

The SureCall software implemented a copy number method comparing on-target reads to a diploid reference. The other copy number method that was tested was CONTRA [184], which uses a similar approach but includes a statistical test for significant alterations based on the variance of the log ratio across the target regions. SureCall was abandoned due to the fact the software is not open source and is poorly documented and thus difficult to optimise for this project.

The second approach is to use off-target reads, as implemented by the CopywriteR method[182]. This method seeks to avoid the biases introduced by the capture process and PCR by discarding reads in target regions and using the off-target

depth of coverage to estimate copy number. Unlike on-target depth of coverage methods, off-target methods provide whole genome copy number, including chromosomal arm-level changes. As well as providing information for many more genes than on-target approaches, data which spans the whole genome allows more accurate segmentation, for which there exist robust methods [183]. It also makes use of the off-target reads for a small panel as was used here, comprising a significant proportion of the reads, if not the majority, in most cases.

Allele specific methods are less relevant to a small panel without a matched normal. It is not clear that calling allele-specific copy number is in fact feasible. Allele specific methods have the added advantage of detecting copy neutral copy number events such as copy neutral LOH, and also allowing estimation of tumour purity and ploidy. Without a matched normal, existing knowledge of common germline variants is used as a surrogate for germline sequencing. One such method is PureCN [233].

3.5.1 Development of CopywriteR

A number of steps and validation was undertaken to make CopywriteR suitable for application in this study. CopywriteR attempts to remove reads from on-target regions according to the panel design. This is however insufficient, as over-representation of reads still occurs in many regions of the genome not targeted for hybrid capture. This seems to be mainly due to pseudogenes with sequence homology to targeted genes. To overcome this problem CopywriteR uses a peak calling algorithm to detect these artefactual regions of high depth. This can be performed on a diploid reference sample, or by using the sample itself. In this

study, the only reference samples available were fresh frozen GIAB samples. When using these samples in the CopywriteR workflow, results were inferior to using the sample itself, with higher noise levels observed. This was despite the GIAB sample being processed in the same way and in the same batch as the tumour samples. This is likely due to the complication of using a fresh frozen reference for an FFPE sample. Each sample was therefore used as its own reference.

CopywriteR has a single parameter choice, the sliding bin size. This determines the width of the genomic region over which reads are counted and averaged. The off-target reads provide a noisy signal. A larger bin size smooths this signal by averaging read depths over a larger region at the cost of diminishing the signal. This may obscure focal alterations. The effect of differing bin size is shown for one sample in Figure 3.9. Each plotted point is the mean coverage in a bin, and the red lines show the segmented copy number derived using the circular binary segmentation algorithm. At a bin width of 50 kilobases, there are many focal deep deletions, which is an implausible scenario. These disappear at higher bin sizes. At 1000 kilobases, there is over-smoothing with a general suppression of signal towards zero, as can be seen for chromosome 6 where focal copy number gains are lost.

A series of samples in the pilot group with copy number assayed via FISH were analysed with different bin sizes to determine the impact on performance. An immediate issue that arises is that FISH copy number is reported as mean copies per cell, and the read out from CopywriteR is reported as a log ratio. The log ratio is the logarithm base 2 of the copy number of the gene in question divided by the

whole genome copy number which we refer to as ploidy. In practice the read depth in the region of a gene is assumed to be proportional to the true copy number. The ploidy is usually '2' representing the diploid genome, but tumours may be triploid (ploidy of 3) or tetraploid (ploidy of 4) or rarely higher. Tumours may also be hypodiploid (ploidy of 1). The log ratio value is therefore a relative measure of copy number compared to some reference genome which is assumed to be unaltered. CopywriteR allows the use of a known diploid reference such as a non-tumour sample, however as discussed above such a reference gives suboptimal results. Alternatively, the tumour sample itself can be used as a reference. Whether using a non-tumour reference or the sample itself, CopywriteR and read depth-based copy number methods in general cannot detect whole genome doubling, as this is cancelled out in calculating a log ratio. Nevertheless, in general we are interested in the extent to which a gene's copy number exceeds the background copy number of the whole genome. A similar standard is employed in routine clinical in situ hybridisation testing for amplification of *ERBB2*, where the ratio of the *ERBB2* copy number and a control probe on the same chromosome is taken. As such, an inability to account for increased ploidy is not a major impediment in determining the copy number of a gene.

If the FISH assay reports a control probe, this can be used to calculate the log ratio for the sample. If this is not the case, then the log ratio can be calculated with the assumption that the reference copy number is diploid. An additional complication is tumour purity. In FISH assays, the operator identifies tumour cells morphologically and counts probes in only those cells. In sequencing assays, the tumour cells and contaminating diploid non-malignant cells are assayed together.

FISH results will therefore not be directly comparable to read depth copy number assays, with the later generally reporting an attenuated signal. Conversely, the log ratio read out of CopywriteR can be converted into a total copy number by raising 2 to the power of the log ratio, and multiplying by the ploidy. As the ploidy is unknown, we set the whole genome copy number to always be 2, which results in a 'ploidy adjusted' total copy number, equivalent to subtracting the ploidy from the copy number of the gene. This adjustment is similar to the routine diagnostic practice of taking a ratio of the *ERBB2* copy number and the chromosome 17 copy number, to avoid over-estimating copy number in tumours with increased ploidy.

Figure 3.10 shows a comparison of the FISH determined copy number and CopywriteR result at different bin sizes. To make results comparable, log ratios were calculated from FISH results. For *ERBB2* in particular, it is clear that larger bin sizes produce highly inaccurate results. The correlation with FISH results was highest for both genes at the 50kb bin size (Pearson's product moment correlation coefficient for *FGFR1* ~ 0.90 [p = 0.002], *ERBB2* ~ 0.92 [$p = 7\times10^{-8}$]). Notably, the calling of spurious deletions that occurred with HaloPlex (Figure 3.5) is not seen here.

As has been discussed above, the relationship between log ratio and the true copy number state of the tumour is influenced by the purity of the sample and the ploidy. It is not possible to determine these variables for each sample using read depth copy number approaches alone. Selection of appropriate log ratio cut-offs to call copy number amplifications or deletions therefore poses a challenge. Simulation of various combinations of ploidy and purity were performed (Figure

3.11). Here, the expected log ratio of a gene LR_g is calculated from the copy number of the gene CN_g, the tumour ploidy equivalent to whole genome copy number CN_w and the tumour purity p:

$$LR_g = log2\left(\frac{\left(p \times CN_g\right) + \left((1-p) \times 2\right)}{\left(p \times CN_w\right) + \left((1-p) \times 2\right)}\right)$$

This is an oversimplification, not accounting for subclonality, and also assumes that there is a linear relationship between the amount of DNA from a genomic region and the sequenced read depth in that region. Figure 3.11 shows that there is no single cut-off which performs well in all scenarios. However, taking into account that lower ploidies are more common, breast tumour samples are rarely highly pure and accuracy for calling amplifications is a priority over deletions as they are more likely to be targetable, a log ratio of ≥ 0.7 for amplifications and ≤ -0.7 for deletions was chosen. The relationship between derived copy number and tumour purity for these cut-offs at various ploidies is shown in Figure 3.12. This highlights that purity has a significant impact on the detection of significant amplifications and alterations, and the inability to estimate purity using read depth copy number methods is therefore an important limitation. These plots also show that the minimum magnitude of a copy number alteration may be estimated by calculating the total copy number from the log ratio assuming a completely pure tumour.

3.5.2 Evaluation of copy number calling methods

CONTRA and CopywriteR were chosen as candidate methods for calling copy number in the pilot hybrid capture samples. CONTRA was run using a GIAB sample that was in the same batch as the tumour samples. Results from these two methods were compared to FISH results for *FGFR1* in 8 samples and *ERBB2* in 14 samples. The correlations of the gene-level log ratio with the log ratio derived from FISH are shown for both genes in Figure 3.13. The correlation for both methods is high. As described previously, CONTRA includes a statistical test for significant alterations. After adjusting for multiple testing, CONTRA fails to call significant amplifications for 3 of 14 pilot samples with known *ERBB2* amplification by FISH. Without adjustment for multiple testing, it still fails to call significant amplifications for 2 samples. Alternatively, CopywriteR as developed in the previous section succeeds in calling all known *ERBB2* amplifications. Since CopywriteR affords a simpler workflow without the need to select a diploid reference sample, and the statistical model employed by CONTRA appears ill-suited to the small panel in use, CopywriteR was used for calling copy number going forward.

3.6 Mutation burden

Mutation burden, or the number of nonsynonymous mutations found in a sample, has emerged as a potential biomarker of response to immunotherapy in several tumour types [234–236]. It may also identify samples with hypermutation, which in turn can lead to diagnosis of germline disorders such as defective mismatch repair. Initial work on the relationship between mutation burden and immune therapy was performed using whole exome sequencing [235,237]. To explore

whether tumour mutational burden could be estimated using the redesigned panel, a simulation was performed using primary breast tumours that had undergone whole exome sequencing in The Cancer Genome Atlas (TCGA). The whole exome mutation burden for each tumour reported in the literature was compared to the number of mutations that would have been found using only the genes on the panel. This simulation represents a best-case scenario for predicting mutation burden, since the TCGA samples were fresh frozen, relatively pure and had variants called with a matched normal. Figure 3.14A shows the correlation between the number of mutations on the simulated panel and the whole exome. Although there is a correlation over several orders of magnitude, it is clear that much of the variation in whole exome mutation burden is lost using the panel. The performance of the panel in predicting which TCGA tumours have a whole exome burden greater than 100 mutations was assessed using receiver operator characteristic curves (Figure 3.14B). The performance for breast cancer specific genes on the panel (61), all genes on the panel (107) and 50, 250 and 1000 randomly selected genes was compared. The panel performs relatively poorly as an estimator of whole exome burden, even under these ideal conditions, with an area under the curve (AUC) of 0.8 for 61 genes and 0.86 for 107 genes. 1000 random selections were performed of 50, 250 and 1000 genes, showing median AUC of 0.7, 0.87 and 0.97 respectively. As can be seen in the figure, the performance of 107 genes on the SEGMENT panel is on par with 250 genes, but nevertheless still not satisfactory to be used as an estimator.

3.7 DNA Extraction

The amount of DNA obtained from a sample is a key step in the delivery of sequencing results. As noted previously, failure rates due to inadequate DNA are expected to be in the order of 30% according to prior studies [166]. The workflow used for this study involved the Molecular Pathology laboratory at the Peter MacCallum Cancer Centre. To recap what has been detailed in the methods chapter, sections from the tumour sample in question are obtained. One section is stained with haematoxylin and eosin for review by a pathologist to confirm cancer and mark regions with malignant cells. The remaining sections are stained with methyl green then marked tumour regions are microdissected off the slides for DNA extraction. As microdissection is a laborious process, it is desirable to microdissect the minimum number of sections necessary. In general this minimum number is judged by the molecular pathology technician. In 102 samples that had DNA extraction, 66 were considered to have sufficient DNA defined as at least 0.25µg, or a concentration of 5ng/µl in the standard elution volume of 50ml, such that 36% of samples had low DNA. A histogram of the extracted DNA concentrations is shown in Figure 3.15A. Of the 36 samples with low DNA, 10 were still sequenced (28%), all of which were successful albeit with high number of duplicate reads.

In an effort to reduce samples that do not progress to sequencing due to inadequate DNA, an analysis of the relationship between tumour area, the number of sections extracted and the resulting DNA was undertaken. A series of representative H&E slides were scanned and the area of the tumour areas marked by the pathologist determined using ImageJ. A total extracted area could then be

calculated per sample by multiplying this area by the number of extracted sections. The strong correlation between extracted DNA and extracted area is shown in Figure 3.15B, suggesting other factors such as cellularity and tissue type are largely irrelevant. On the basis of these results, a guide to how many sections to extract based on the tumour area was created (Figure 3.16) and supplied to the Molecular Pathology technicians. Additionally, the protocol for sequencing library preparation was improved as described in section 1.3, such that lower input DNA became viable.

3.8 Reporting of results

One of the core objectives of the SEGMENT study was to return results to referring clinicians. At the time the project was conceived there was no codified manner in which to report genomic profiles. As the assay remains a research instrument, co-opting existing procedures and templates within the diagnostic pathology department for reporting genomic findings was not possible. In any case, genomic profiles of tumours have unique characteristics which require attention in the reporting of results. Firstly, the results of genomic profiling entail high degrees of uncertainty and at times nuance in interpretation. There are numerous examples: a driver alteration in one cancer may not have the same therapeutic implications in another type of cancer; it is often difficult to ascertain the relevance of missense mutations in tumour suppressor genes such as *PTEN*; the mere presence of a mutation is necessary but not sufficient to imply significance, and an understanding of the concepts of allele frequency and subclonality is required. In relation to this, the clinicians receiving the report are unlikely to be trained in precision oncology or genomics as it is a relatively new field. Secondly, the

significance of the results of a genomic profile may change over time. A mutation of uncertain significance at the time of reporting may be relevant to a new therapy which arrives a year or more later. Thirdly, the recommendations for therapy arising from a genomic profile may be quite specific to the patient, their geographic location and the time at which the report is released. A clinical trial with a targeted agent may only be available at a certain time and place. In Australia, there are relatively few clinical trial centres and major cities, and it is thus possible for a small team to be aware of all relevant trial options.

As discussed in chapter 1, there is no gold standard manner to report on the significance of a mutation. An arbitrary scale has the disadvantage of creating another unfamiliar construct in the presence of an already complex report. With these considerations in mind the following principles were used to guide reporting:

1. Sufficient detail regarding the mutation should be included such that if a clinician returns to the report in the future they can re-interpret the results in light of new developments. This is achieved by including the gene symbol and protein alteration, along with a RefSeq gene identifier and the relevant coding transcript identifier.
2. A pragmatic classification scale for the actionability of mutations should be used which does not require the use of an intermediary scale and tries to show the evidence as transparently as possible.
3. The results are presented together with a recommendation for management if appropriate. Thus if a clinical trial option is available then

this is included as a recommendation in the report. Off-label drug use is also included as appropriate.

Actionable genes with the most clinical relevance are shown in Table 3.6, along with their significance and supporting evidence.

Table 3.1 Genes included in version 1 of the panel. Genes of particular relevance in breast cancer are underlined.

ABL1	CDH1	FBXW7	IDH2	MPL	RUNX1
AKT1	CDKN2A	FGFR1	INPP4B	NOTCH1	SMAD4
AKT2	CSF1R	FGFR2	JAK2	NPM1	SMARCB1
AKT3	CTNNB1	FGFR3	JAK3	NRAS	SMO
ALK	EGFR	FGFR4	KDR	PDGFRA	SRC
APC	ERBB2	FLT3	KIT	PIK3CA	STK11
AR	ERBB3	GNA11	KRAS	PIK3R1	TP53
ATM	ERBB4	GNAQ	MAP2K1	PIK3R3	VHL
BRAF	FANCA	GNAS	MAP2K2	PTEN	WT1
BRCA1	FANCC	HNF1A	MAP2K4	PTPN11	
BRCA2	FANCF	HRAS	MET	RB1	
CCND1	FANCG	IDH1	MLH1	RET	

Table 3.2. Sequencing metrics in three cohorts of samples run on the HaloPlex target enrichment platform. 'sequenceable' refers to genomic regions covered by amplicons, which is distinct from target regions.

	GIAB (Frozen, n = 10)	Collaborator Cohort (FFPE, n = 16)	SEGMENT Cohort (FFPE, n = 24)
Mean reads passing quality filters	4009259	4579557	4103589
Mean % reads on target	94 (94 - 94)	99 (99 - 99)	98 (95 - 99)
Mean % of target regions with 1000 reads	82 (74 - 89)	75 (60 - 85)	75 (47 - 86)
Mean % of target regions with 500 reads	92 (89 - 96)	84 (74 - 93)	85 (57 - 94)
Mean % of target regions with 100 reads	99 (99 - 99)	94 (87 - 98)	95 (75 - 99)
Mean % of target regions with 10 reads	99 (99 - 99)	98 (94 - 99)	98 (84 - 99)
Mean % of target regions with 1 read	100 (100 - 100)	98 (97 - 99)	99 (96 - 100)
Mean % of sequenceable regions with zero coverage	1.3 (0.98 - 1.7)	36 (21 - 48)	15 (1.5 - 46)
Mean read depth in target regions	1467 (990 - 2247)	1400 (791 - 2396)	1358 (860 - 1698)
Mean fragment length	115 (114 - 118)	77 (71 - 84)	90 (64 - 117)

Table 3.3. Genes included in version 2 of the panel. Genes of particular relevance in breast cancer are underlined.

AKT1	BRCA1	CHEK2	EXT1	GATA3	MEN1	NOTCH2	PTPN11	SMO
AKT2	BRCA2	CTNNB1	FANCA	HRAS	MET	NOTCH4	RAD51C	STK11
AKT3	BRIP1	DDB2	FANCC	IGF1R	MLH1	NRAS	RAD51D	TBX3
ALK	BUB1B	EGFR	FANCG	INPP4B	MLL2	PALB2	RB1	TMEM127
APC	CASP8	ERBB2	FBXW7	KIT	MLL3	PDGFRA	RECQL4	TP53
ARID1A	CBFB	ERBB3	FGFR1	KRAS	MSH2	PIK3CA	RET	TSC1
ATM	CCND1	ERBB4	FGFR2	MAP2K1	MSH6	PIK3R1	RUNX1	TSC2
AXIN2	CCND2	ERCC2	FGFR3	MAP2K4	MUTYH	PIK3R3	SDHA	VHL
BAP1	CDC73	ERCC3	FGFR4	MAP3K1	NBN	PMS2	SDHB	WT1
BLM	CDH1	ERCC4	FH	MAX	NCOR1	PRKAR1A	SDHC	XPA
BMPR1A	CDK4	ERCC5	FLCN	MCL1	NF2	PTCH1	SDHD	XPC
BRAF	CDKN2A	ESR1	FOXA1	MDM2	NOTCH1	PTEN	SF3B1	

Table 3.4. Sequencing metrics in three cohorts of samples run on the SureSelect target enrichment platform.

	GIAB/Germline (Frozen, n=7)	SEGMENT Cohort (FFPE, n = 28)
Mean reads passing quality filters	13,444,003	37,752,008
Mean % bases on target	18 (10 - 30)	20 (5 - 37)
Mean % of target regions with 100 reads	96 (84 - 99)	92 (39 - 99)
Mean % of target regions with 50 reads	98 (89 - 100)	97 (69 - 100)
Mean % of target regions with 20 reads	99 (93 - 100)	99 (91 - 100)
Mean % of target regions with 1 read	99 (98 - 100)	99 (99 - 100)
Mean % targets with zero coverage	1 (0.7 - 1)	0.8 (0.3 - 1)
Mean read depth in target regions	653 (252 - 2671)	1195 (88 - 3061)
Mean fragment length	184 (157 - 205)	151 (96 - 204)

Table 3.5. Validation by Sanger sequencing of mutations called with VarDict. VAF is variant allele fraction.

Validated	Gene	Reference Allele	Alternate Allele	Protein	VAF	Variant Type
yes	ARID1A	C	T	p.Q1363X	0.21	stopgain
yes	BRCA2	A	G	p.N108S	0.43	nonsynonymous SNV
yes	CBFB	T	G	p.V106G	0.56	nonsynonymous SNV
yes	CDH1	G	A	p.V242I	0.34	nonsynonymous SNV
yes	CDKN2A	C	A	p.G16C	0.46	nonsynonymous SNV
yes	ERBB2	G	A	p.R985H	0.89	nonsynonymous SNV
no	FGFR1	T	C	p.Q89R	0.21	nonsynonymous SNV
yes	FGFR3	G	A	p.G137R	0.39	nonsynonymous SNV
yes	FGFR4	C	T	p.R54C	0.10	nonsynonymous SNV
yes	GATA3	-	ACACCACCCCT	p.H434Tfs*46	0.12	frameshift insertion
yes	GATA3	-	C	p.L301Pfs*3	0.22	frameshift insertion
yes	GATA3	CA	-		0.30	splicing
yes	GATA3	CA	-		0.36	splicing
yes	GATA3	A	C		0.41	splicing
yes	GATA3	CA	-		0.50	splicing
yes	NF2	C	T	p.R291C	0.47	nonsynonymous SNV
yes	PIK3CA	A	C	p.E545A	0.10	nonsynonymous SNV
yes	PIK3CA	G	A	p.E726K	0.13	nonsynonymous SNV
yes	PIK3CA	A	T	p.E542V	0.17	nonsynonymous SNV
yes	PIK3CA	A	C	p.E545A	0.19	nonsynonymous SNV
yes	PIK3CA	A	G	p.H1047R	0.22	nonsynonymous SNV
yes	PIK3CA	G	A	p.E542K	0.23	nonsynonymous SNV
yes	PIK3CA	A	C	p.E545A	0.23	nonsynonymous SNV

Table 3.5 continued. Validation by Sanger sequencing of mutations called with VarDict.

Validated	Gene	Reference Allele	Alternate Allele	Protein	VAF	Variant Type
yes	PIK3CA	C	T	p.H1047Y	0.26	nonsynonymous SNV
yes	PIK3CA	A	G	p.H1047R	0.32	nonsynonymous SNV
yes	PIK3CA	G	A	p.E545K	0.37	nonsynonymous SNV
yes	PIK3CA	A	G	p.H1047R	0.39	nonsynonymous SNV
yes	PIK3CA	A	G	p.H1047R	0.55	nonsynonymous SNV
yes	PIK3R1	A	-	p.K245Nfs*15	0.33	frameshift deletion
yes	PIK3R1	CCTT	-	p.D299Efs*4	0.37	frameshift deletion
yes	RUNX1	G	C	p.R139G	0.28	nonsynonymous SNV
yes	RUNX1	-	G	p.Q237Sfs*336	0.34	frameshift insertion
yes	SF3B1	G	C	p.A881G	0.35	nonsynonymous SNV
yes	STK11	A	C	p.E2A	0.46	nonsynonymous SNV
yes	TBX3	-	T	p.V202Rfs*38	0.26	frameshift insertion
yes	TP53	TTAGTGCTCCCTGGGGGCAGCTCGT	-	p.H297Rfs*40	0.16	frameshift deletion
yes	TP53	G	A	p.P278L	0.41	nonsynonymous SNV
yes	TP53	A	G	p.L145P	0.52	nonsynonymous SNV
yes	TP53	T	G		0.55	splicing
yes	TP53	C	A		0.60	splicing
yes	TSC2	C	T	p.A1467V	0.45	nonsynonymous SNV
yes	TSC2	G	A	p.G1421E	0.47	nonsynonymous SNV

Table 3.6. Core actionable alterations.

Gene	Example alterations	Significance	Highest level of evidence
PIK3CA	H1047R E545K, E542K	Benefit from alpha subunit specific PIK3CA inhibitors.	Survival and response rate from phase 3 randomised controlled trials [238].
ERBB2	V777L	Benefit from ERBB2 tyrosine kinase inhibitor neratinib.	Response rate from uncontrolled trials [239,240].
ERBB2	Amplification	Hallmark of HER2+ breast cancer subtype. Predicts benefit from anti-HER2 therapy.	Standard of care, survial and response rate from multiple phase 3 randomised controlled trials.
RB1	R130Q LOF	Associated with lack of benefit and secondary resistance to CDK inhibitor class drugs.	Retrosepctive analysis of randomised controlled trial [241].
BRCA1 **BRCA2**	LOF	Common germline susceptibility gene. Predicts benefit from PARP inhibitors and platinum based chemotherapy.	Survival and response rate from phase 3 randomised controlled trials, primarly for germline alterations [14,15].
PTEN	LOF	Loss of function predicts lack of benefit and resistance to alpha subunit specific PIK3CA inhibitors, benefit from AKT inhibitors.	Retrospective data from controlled trials [242].
ESR1	D538G Y537C	Predicts resistance to endocrine therapies, including fulvestrant.	Retrospective analysis of randomised controlled trial [243].
AKT1	E17K	Predicts benefit from AKT inhibitors.	Response rate from phase I trials [244].
BRAF	V600E	Predicts benefit from BRAF inhibitors.	Survival and response rate from phase 3 randomised controlled trials in other cancer types [149].

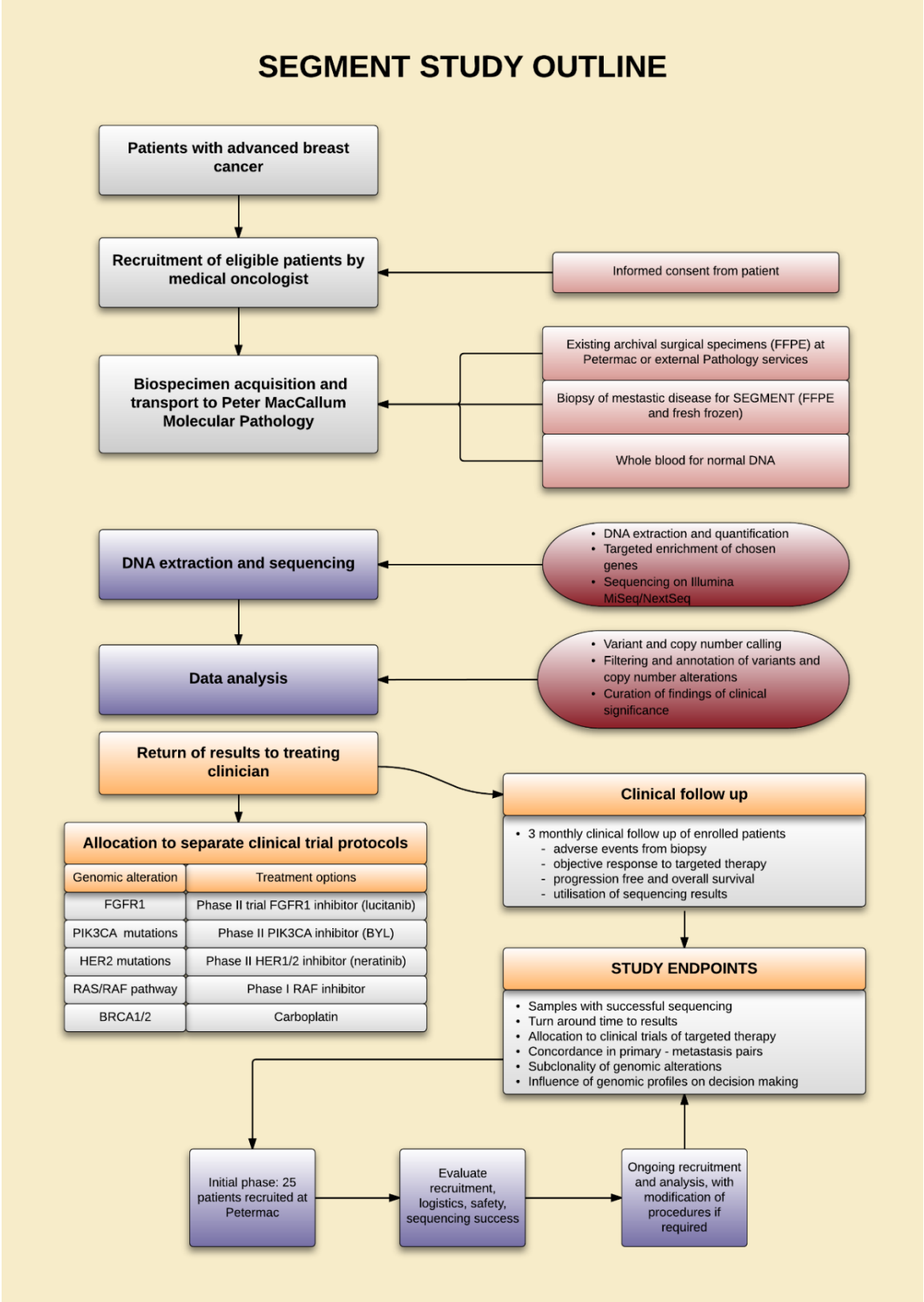


Figure 3.1. Overview of the SEGMENT study.

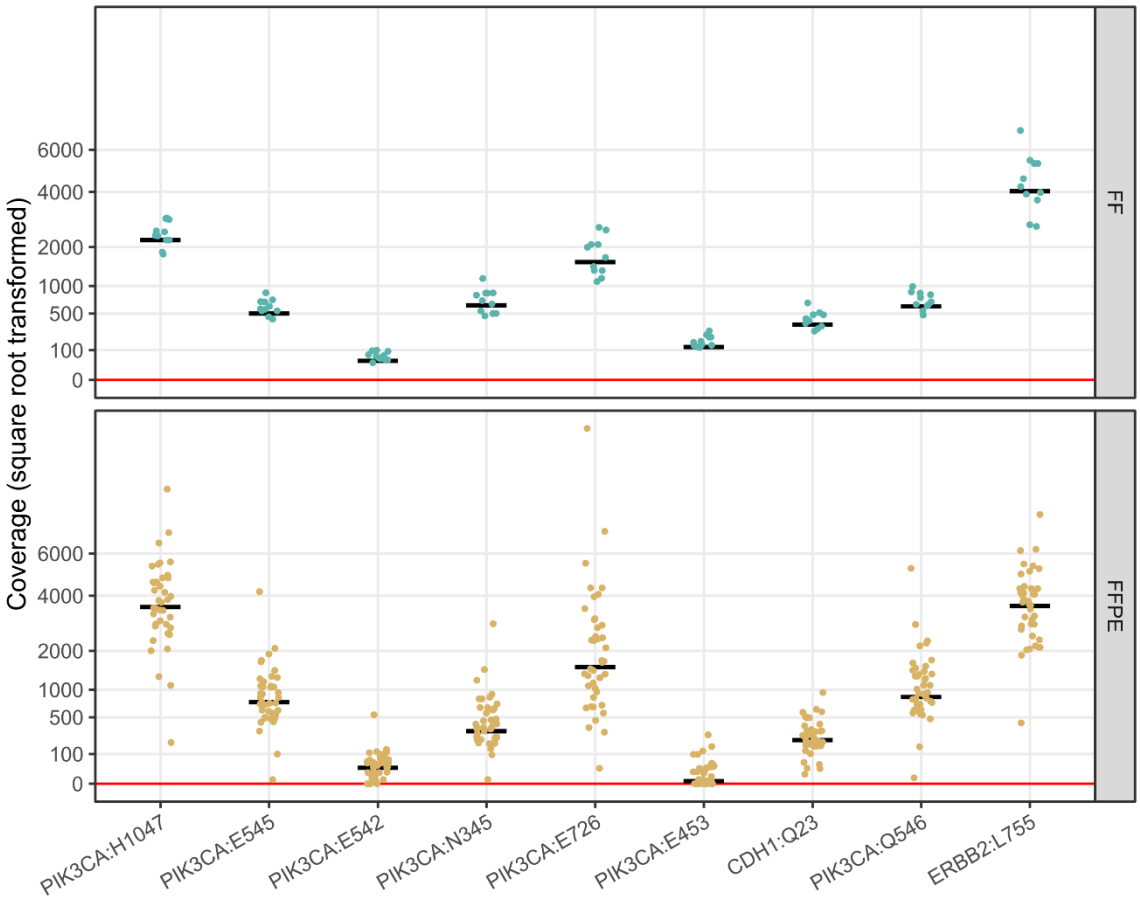


Figure 3.2. Coverage of commonly mutated loci in breast cancer in A) high quality fresh frozen GIAB DNA (FF) and breast tumour FFPE samples. Loci ranked on x-axis left to right by decreasing frequency. Y-axis is square root transformed.

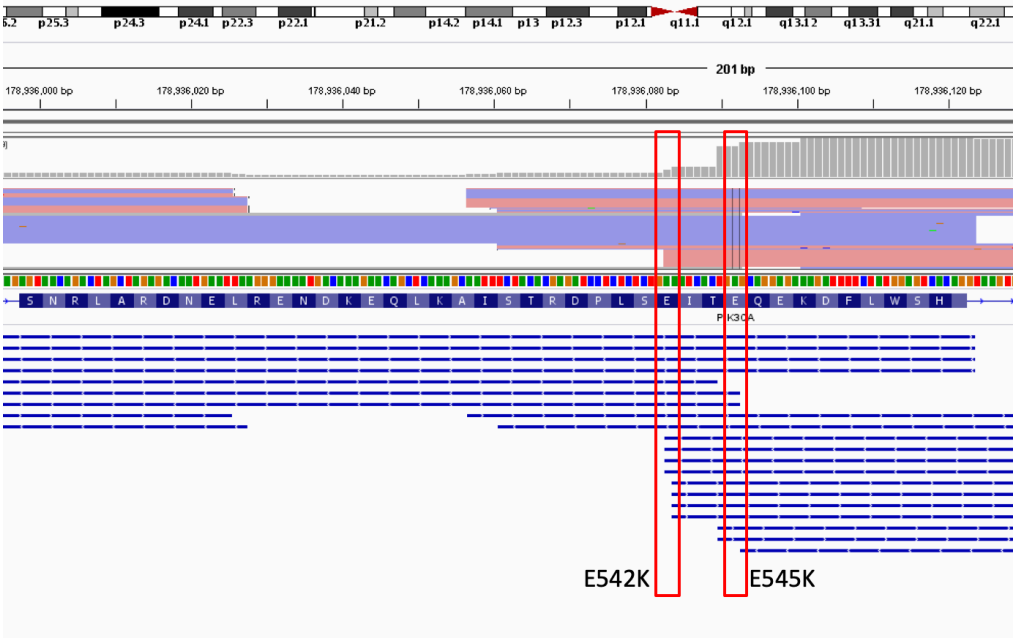


Figure 3.3. Coverage of *PIK3CA* helical domain mutation hotspots using Integrative Genomics Viewer in a high-quality reference sample.

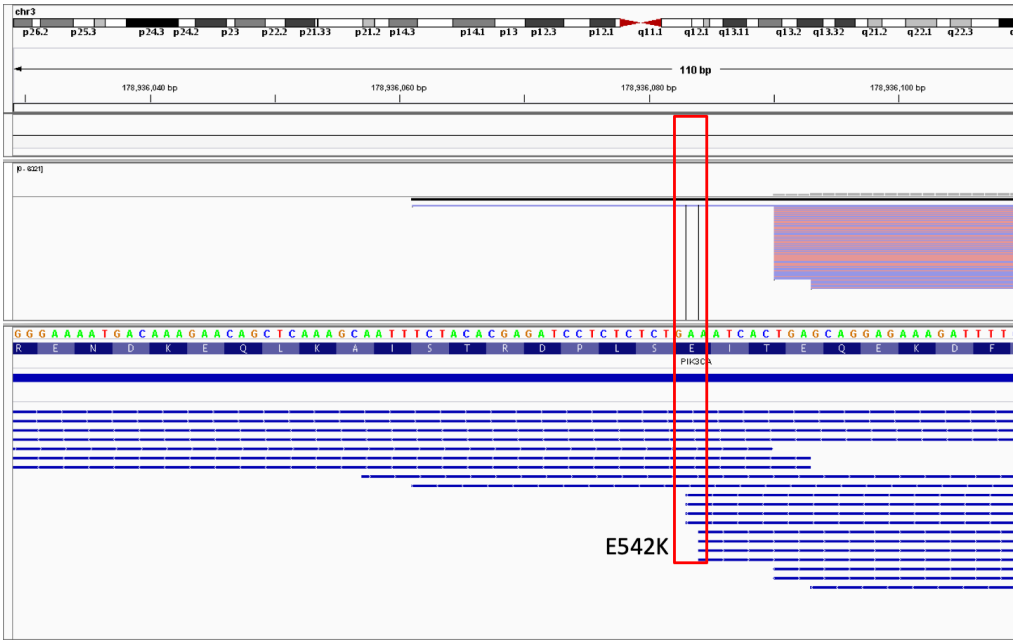


Figure 3.4. Coverage of *PIK3CA* helical domain mutation hotspots using Integrative Genomics Viewer in a typical FFPE sample.

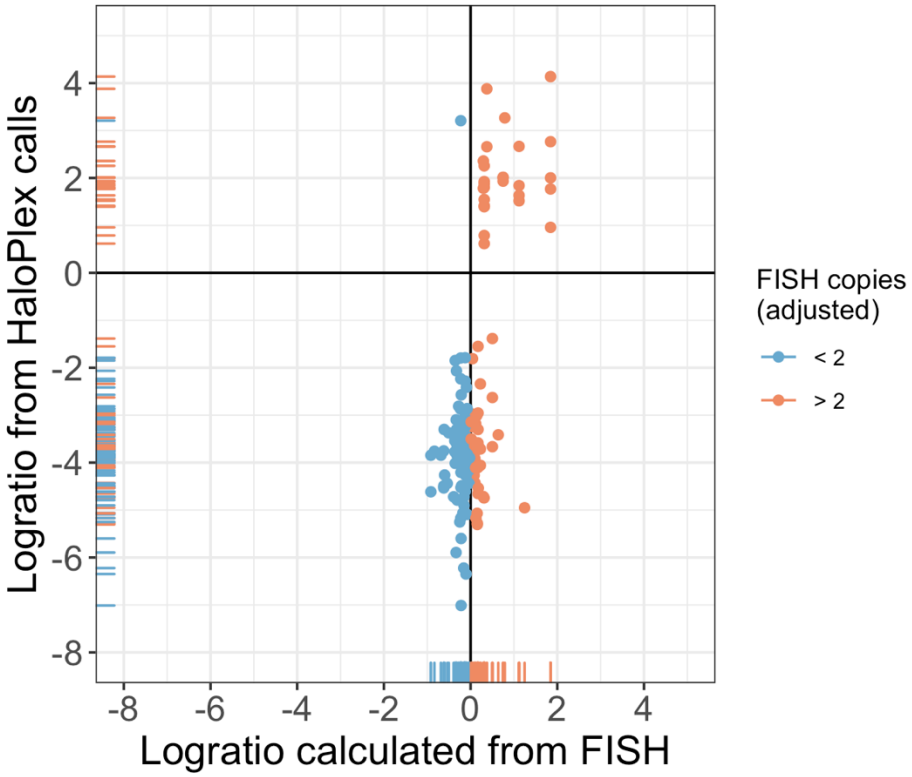


Figure 3.5. Hybrid scatter plot/rug plot showing correlation between copy number calls from HaloPlex data and FISH for FFPE tumours. FISH copy number is adjusted by subtracting mean FISH copy number for each sample. Data for 14 genes in 13 breast tumour samples is shown.

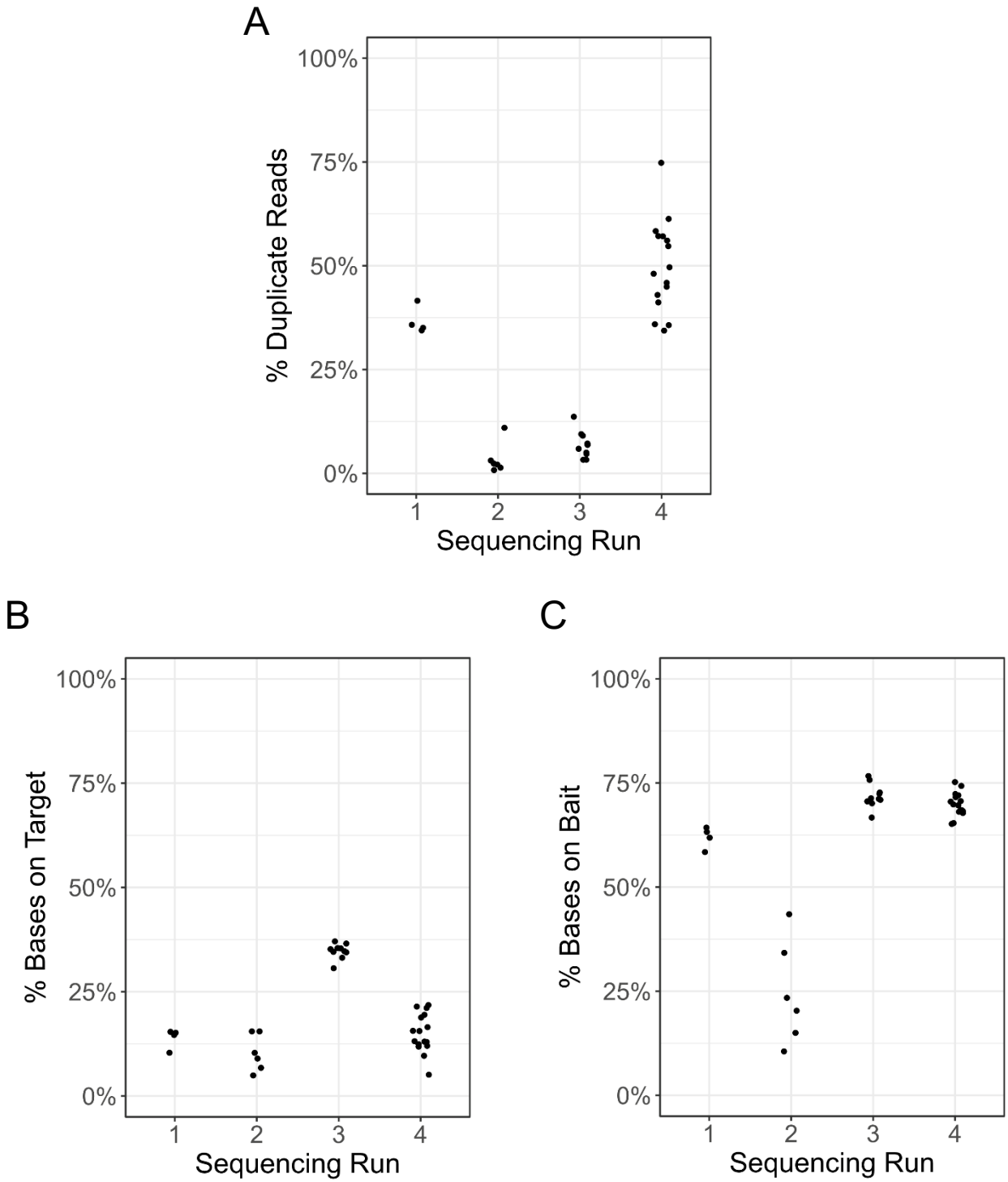


Figure 3.6. Key sequencing metrics in 4 pilot runs on the redesigned SureSelect platform. A) Percentage of duplicate reads. B) Percentage of bases in exons of panel genes. C) Percentage of bases in regions covered by baits.

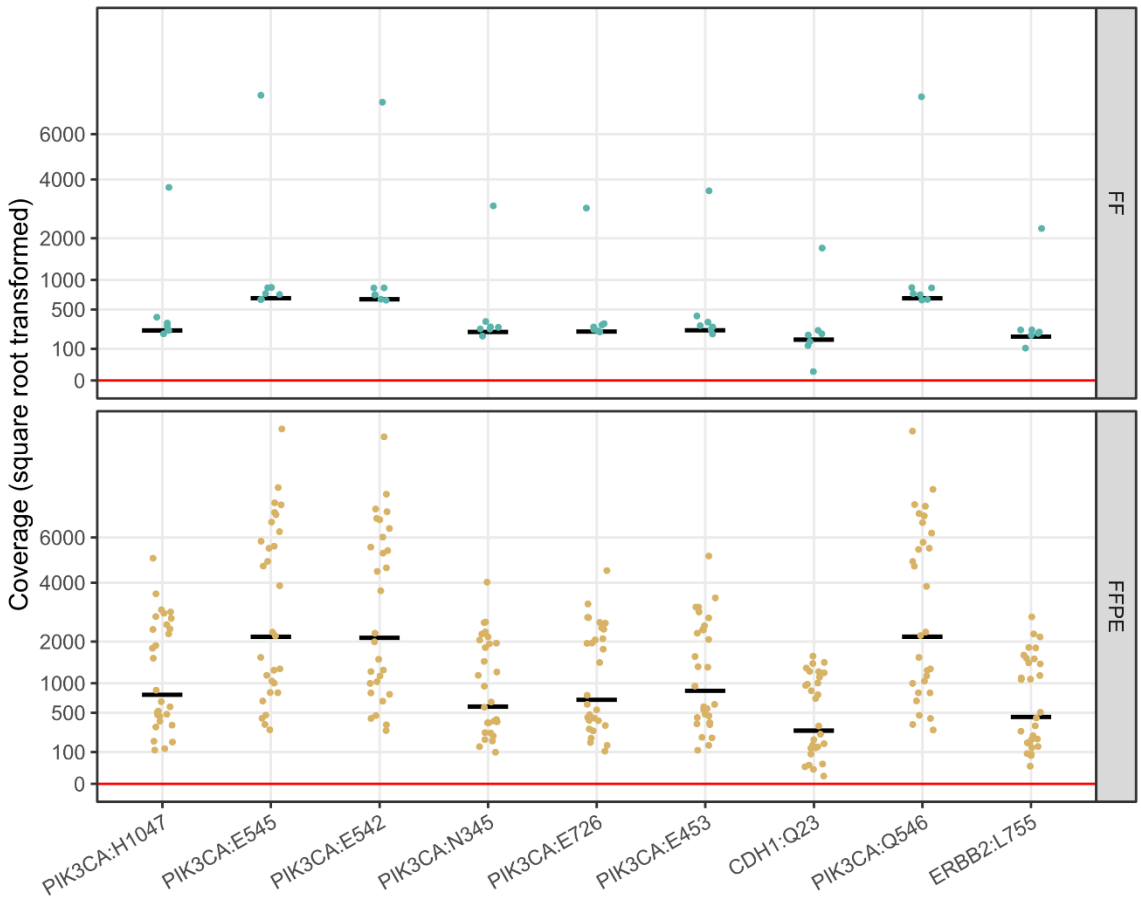


Figure 3.7. Coverage of commonly mutated loci in breast cancer in high quality fresh frozen GIAB/germline DNA (FF) and breast tumour FFPE samples. Loci ranked on x-axis left to right by decreasing frequency. Y-axis is square root transformed.

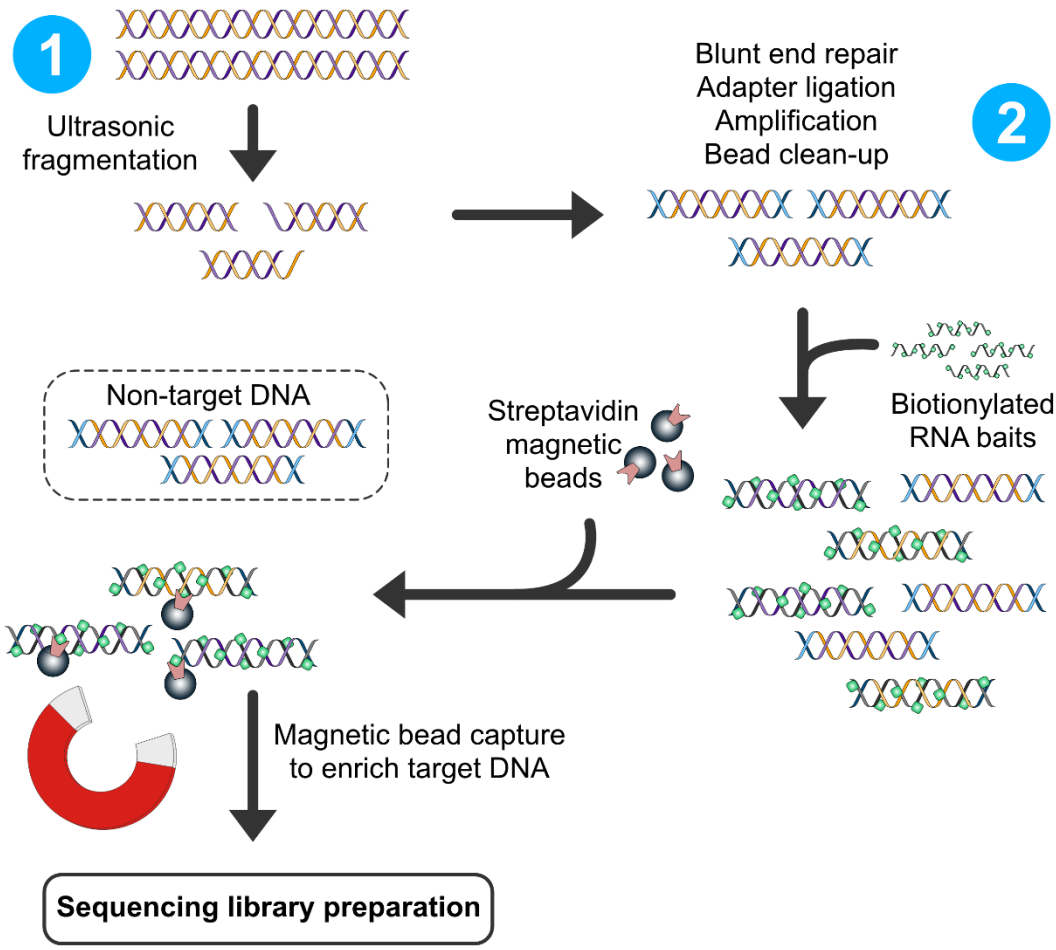


Figure 3.8. Hybrid Capture methodology. Manufacturer's protocol was modified at (1) fragmentation step and (2) pre-capture preparation (see text). This figure is an original work adapted from the Agilent SureSelect protocol.

Figure 3.9. Effect of different bin widths on off-target read depths. Log-ratios plotted on y-axis, chromosomal position on x-axis. Red lines represent segmented values. Red arrow heads are deep deletions.

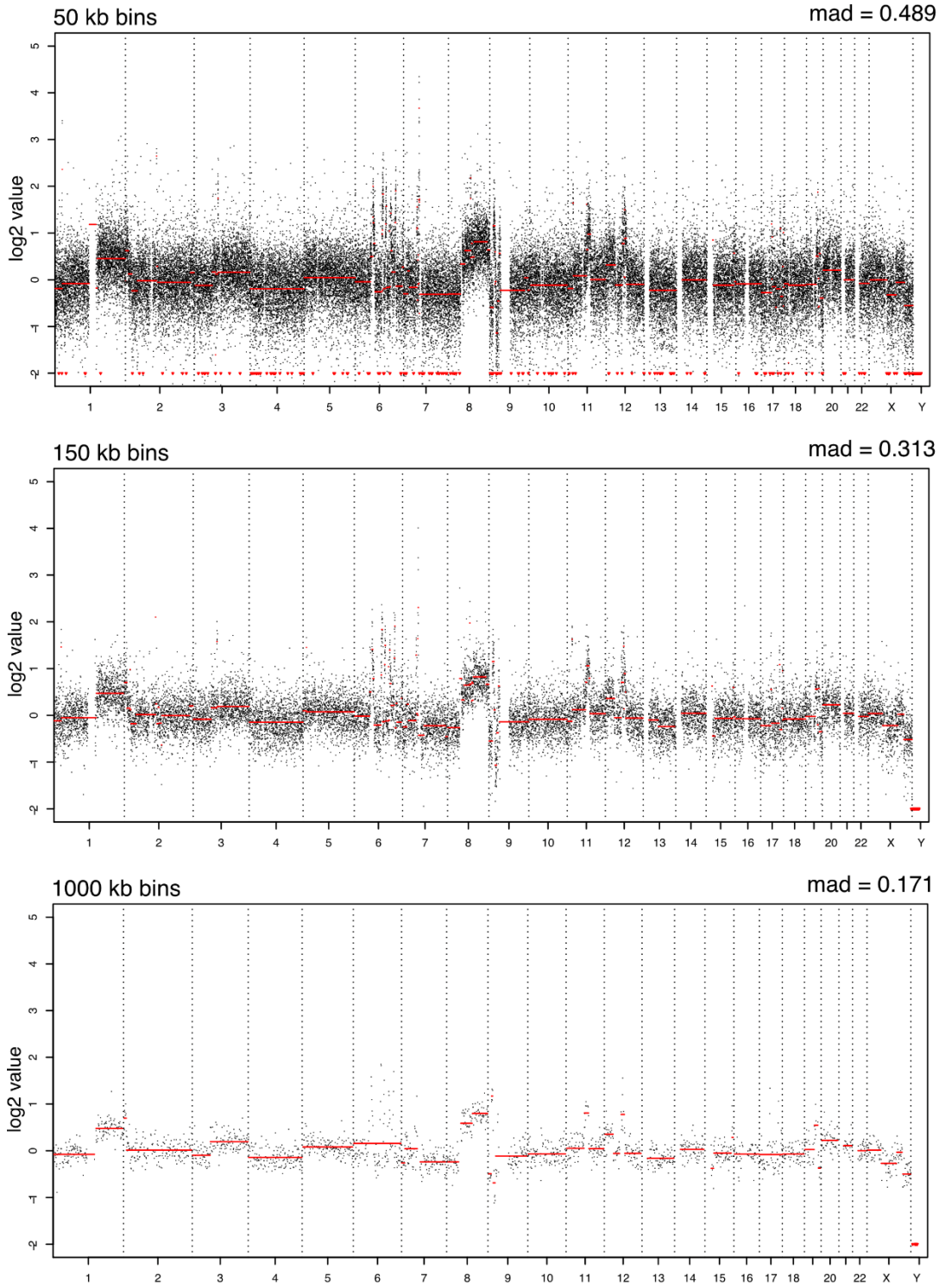
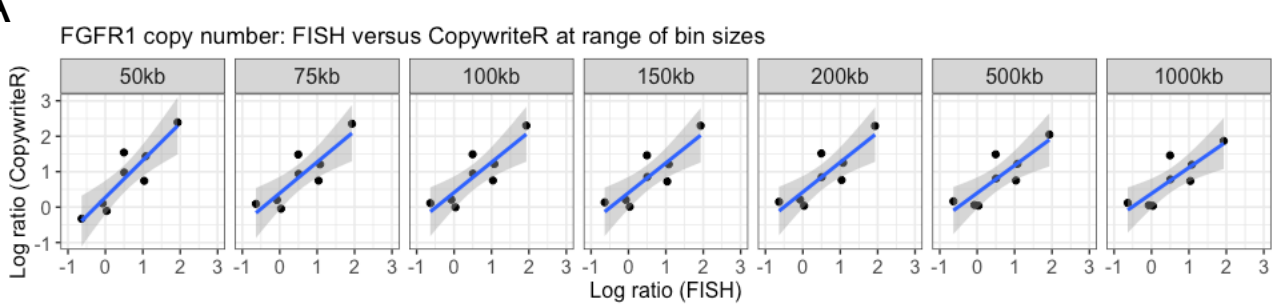


Figure 3.10. Comparison of log ratios from FISH and CopywriteR across different bin sizes. A) Data for *FGFR1* in 8 tumours. FISH log ratio was calculated using control probe to set ploidy. Pearson's product-moment correlation coefficient >0.85 and p-value <0.05 for all bin sizes. B) Data for *ERBB2* in 17 tumours. As control probes unavailable, ploidy set to 2. Linear models fitted, with grey areas showing 95% confidence interval. Pearson's product-moment correlation coefficient >0.84 and p-value <0.05 for bin sizes 50kb-200kb. For 500kb and 1000kb <0.2 and not significant.

A



B

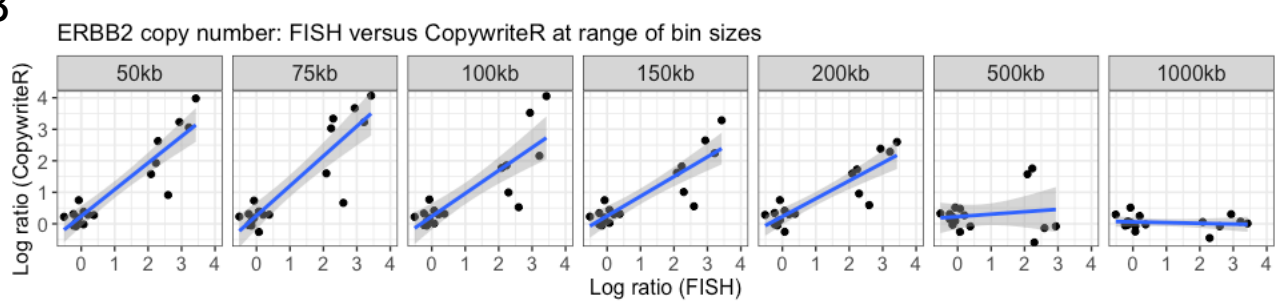


Figure 3.11. Log ratios have been simulated for different total copy numbers of a gene. Each line represents a different level of tumour purity. Results are shown for ploidy 2 -4. Horizontal red lines show total amplification (copy number 8 and 4) and deletion (copy number 1 and 0). Blue vertical lines are at pragmatic cut-offs for log ratios to call amplification (≥0.7) and deletion (≤-0.7).

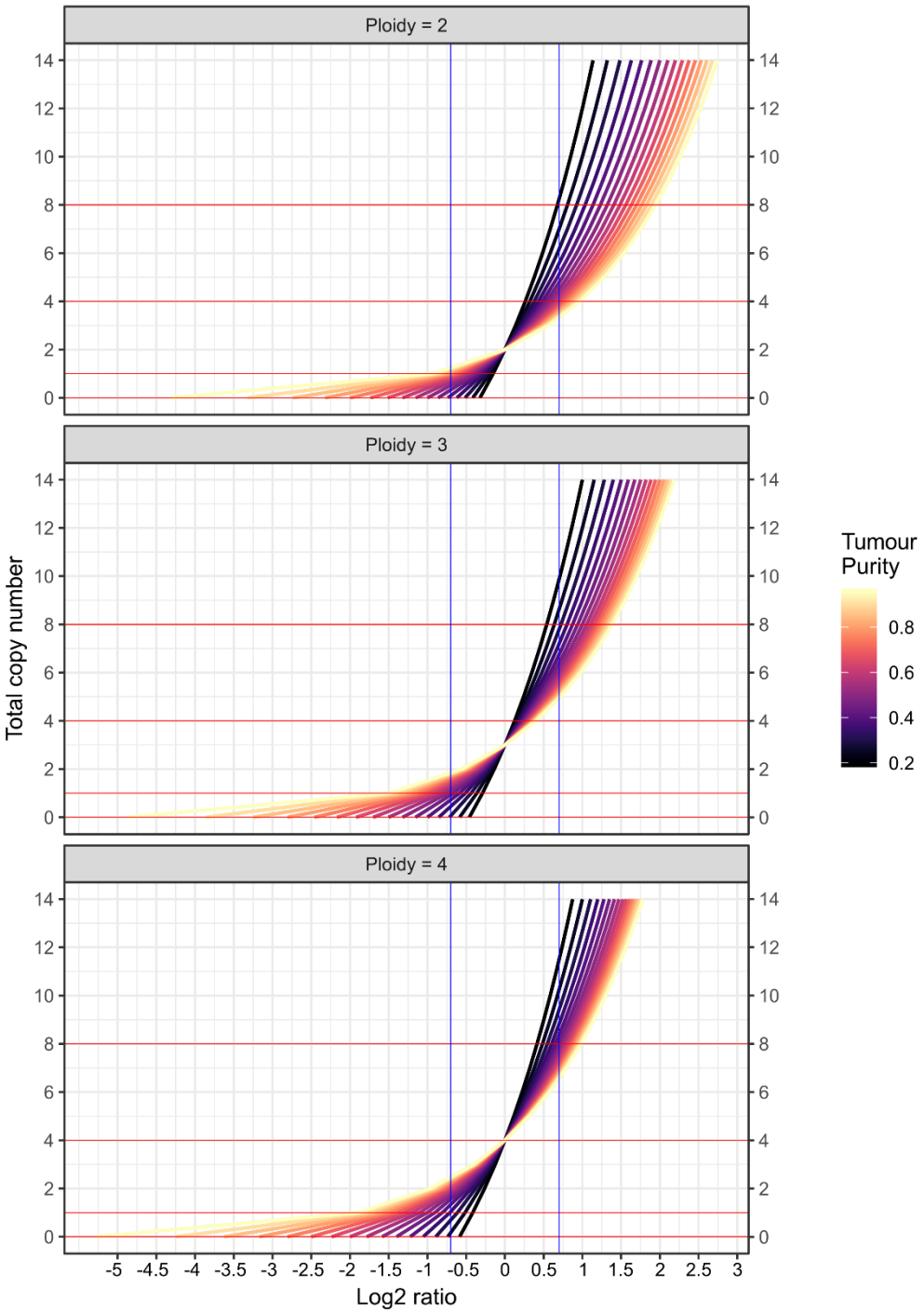


Figure 3.12. Simulated total copy number at log ratio cut-offs for amplification (A) and deletion (B).

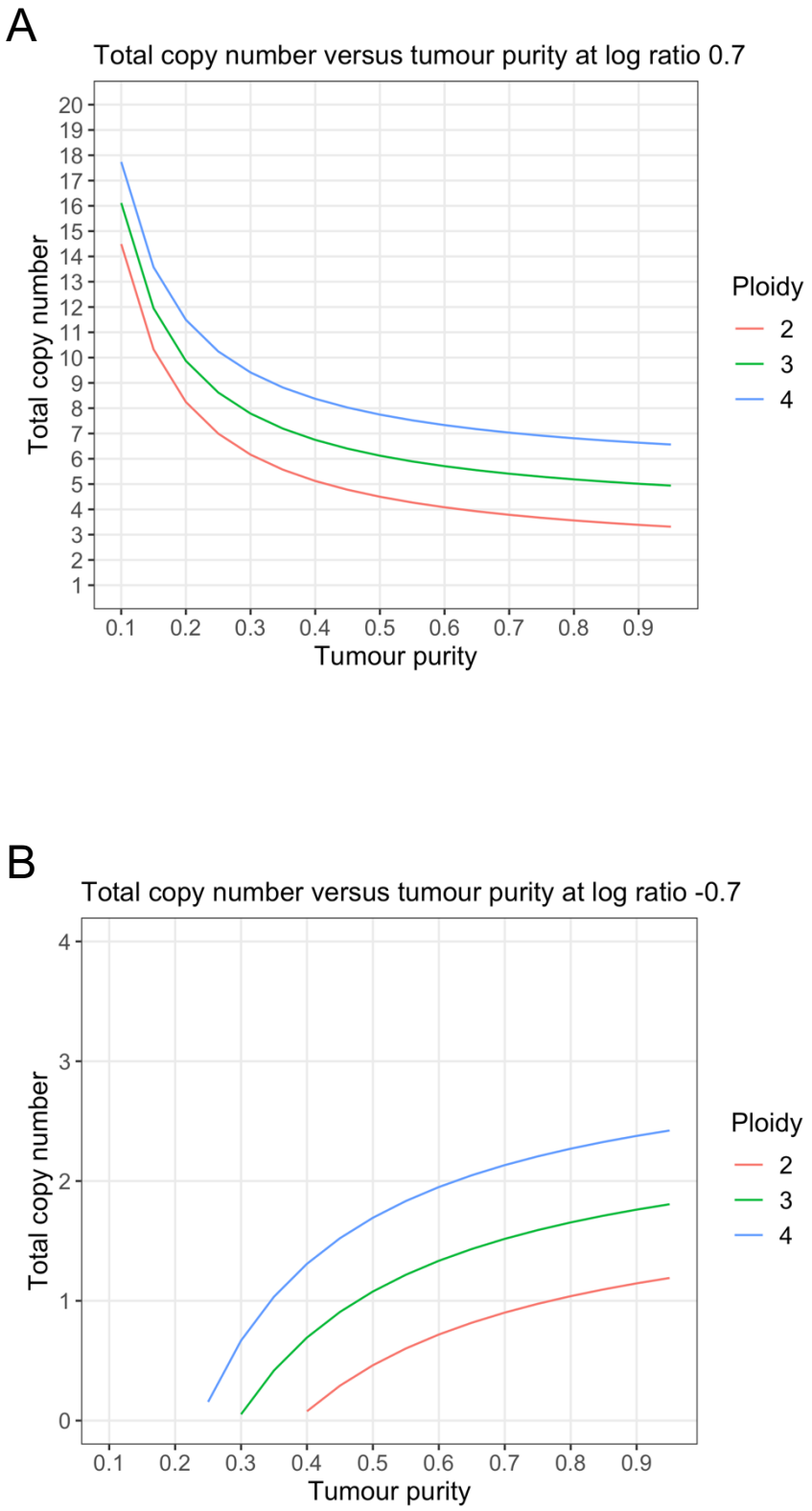
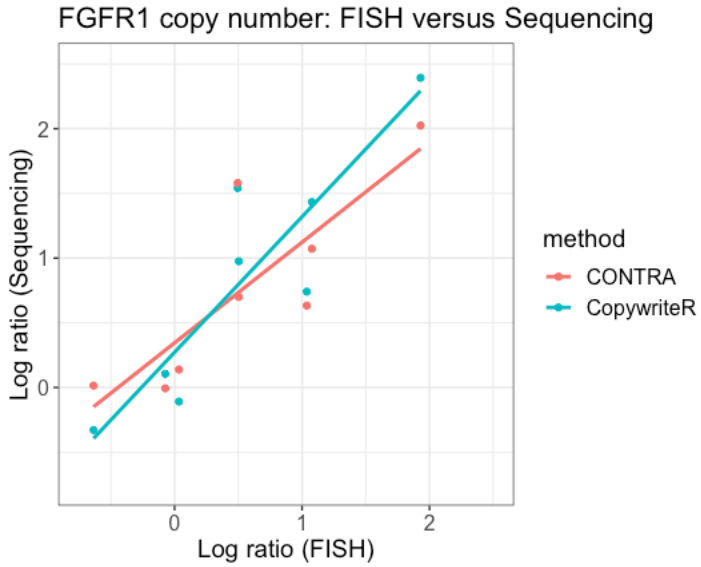


Figure 3.13. Correlation of different copy number calling methods with FISH. A) Data for *FGFR1* in 8 tumours. Pearson's product-moment correlation coefficient >0.8 and p-value <0.05 for CONTRA and CopywriteR. B) Data for *ERBB2* in 14 tumours. Linear model fitted. Confidence intervals not shown for clarity. Pearson's product-moment correlation coefficient >0.85 and p-value <0.05 for CONTRA and CopywriteR.

A



B

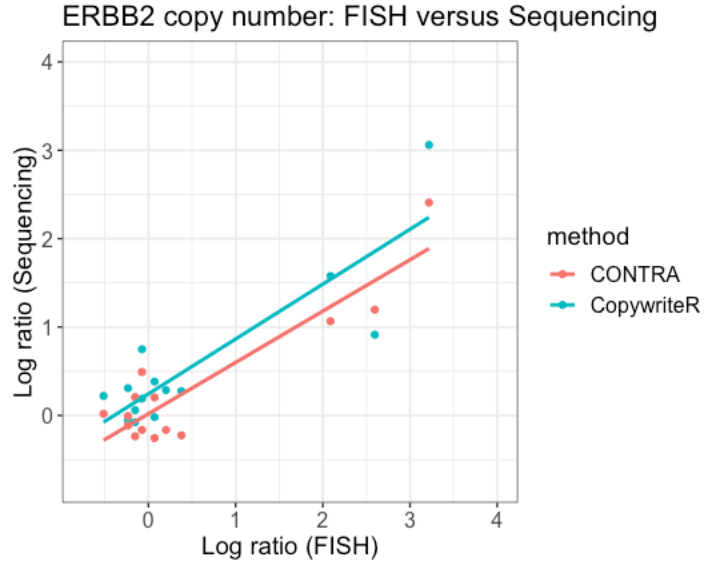


Figure 3.14. Simulated performance of mutation burden estimation using SEGMENT panel genes. A) Correlation between whole exome mutation burden and panel burden. B) Receiver operating characteristic curves for estimators of whole exome mutation burden. For random gene panels, 1000 simulated estimators are shown.

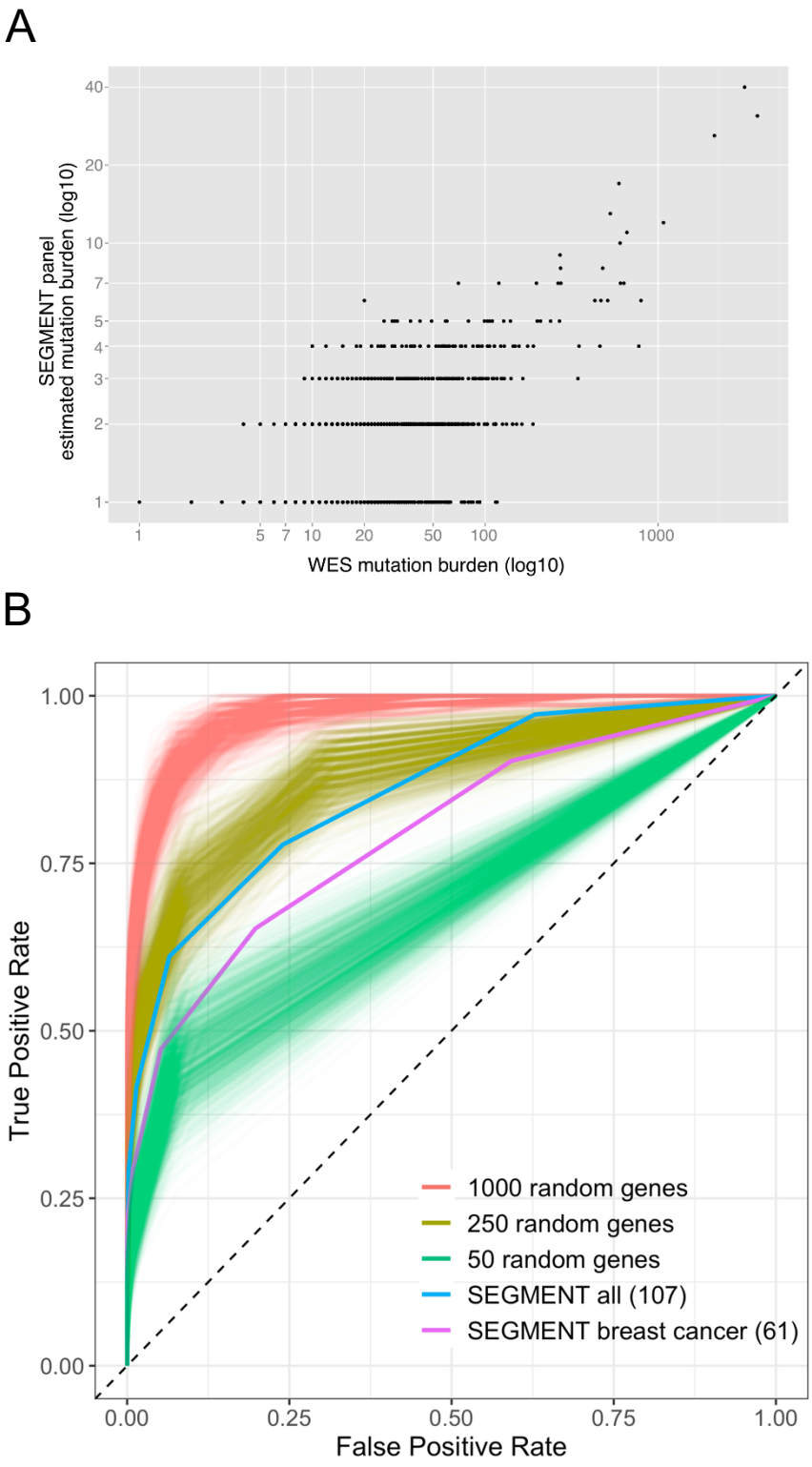


Figure 3.15. A) Histogram of extracted DNA concentration. DNA is eluted in 50 µl. A desirable input for sequencing is 0.25µg, or a concentration of 5 ng/µl, which is shown by the vertical green bar. B) Relationship between tissue area undergoing DNA extraction and final DNA concentration, with linear model fitted. Pearson's product-moment correlation coefficient 0.82, p-value < 0.05.

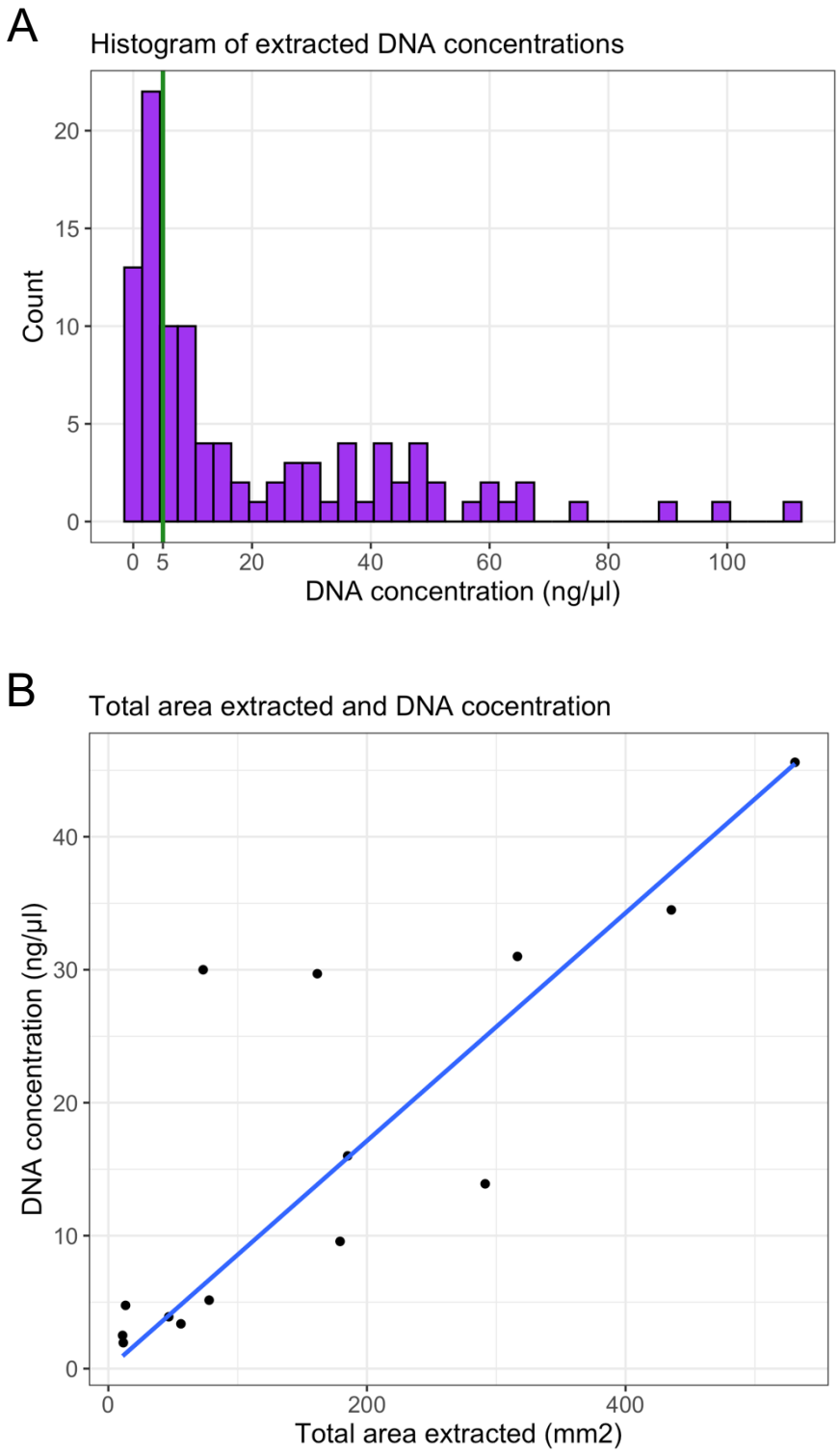
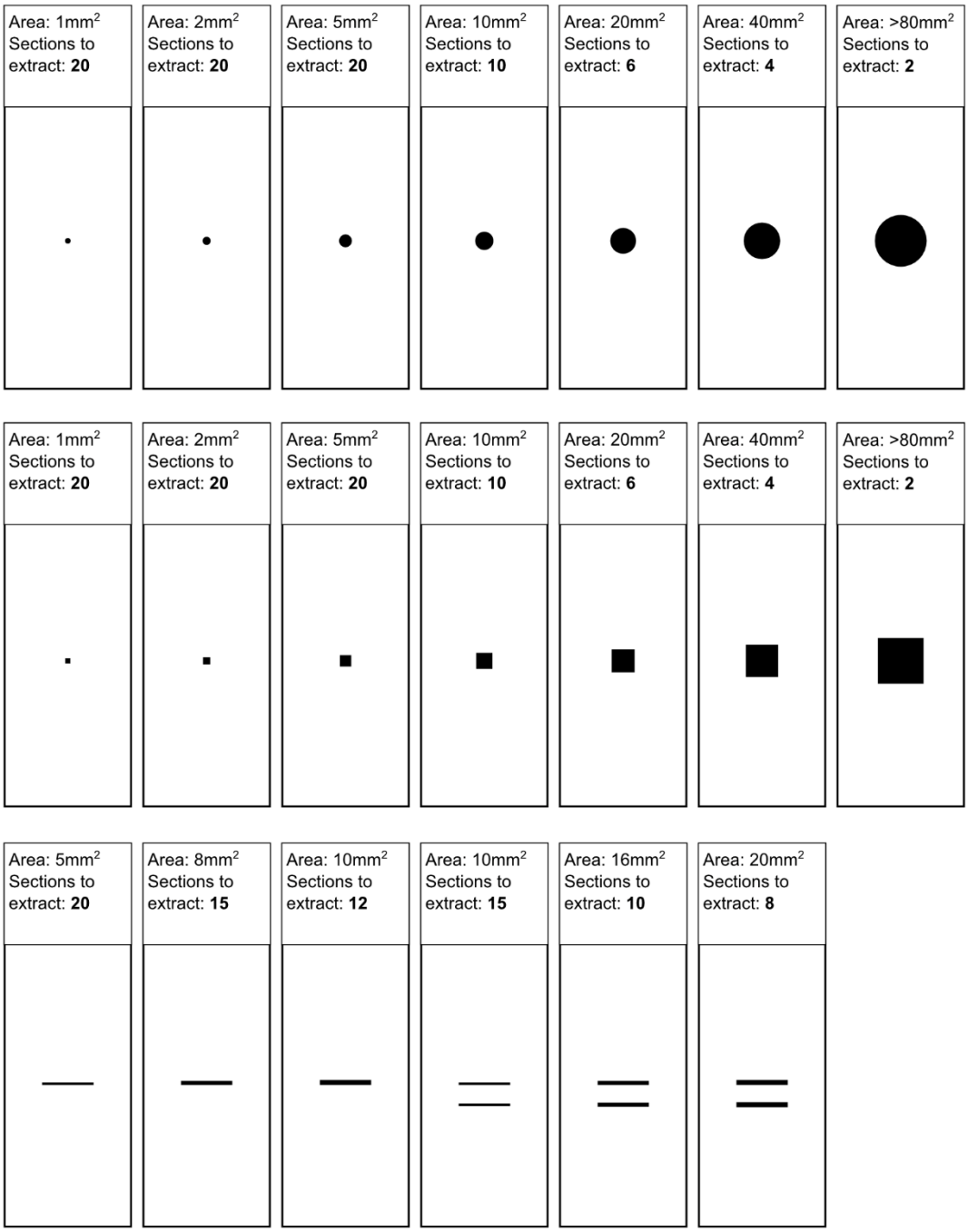


Figure 3.16. DNA extraction guide devised to improve DNA extraction yields (not to scale)



Chapter 4: Feasibility of SEGMENT

A major aim of this work was to assess the feasibility of implementing a precision oncology program for patients with metastatic breast cancer. Feasibility was assessed on technical, logistical and clinical dimensions. Following the pilot phase described in the previous chapter, the tested workflow was applied to samples prospectively. The data presented here relate to patients that were enrolled during a period of 41 months, from December 2013 to May 2018, and had samples sequenced during a period of 44 months from December 2013 to August 2018. Technical and logistical data are presented for the entire cohort including those sequenced with the HaloPlex platform and the pilot phase. Sequencing data will only be presented for the redesigned platform subsequent to HaloPlex.

Recruitment to the program was active and stable throughout the duration, as shown in Figure 4.1. A total of 300 patients were recruited, amounting to approximately 7 patients per month. No specific advertising for the study was undertaken, and patients or patient advocacy groups were not informed directly about the study.

4.1 Sample handling and DNA extraction

A total of 658 samples were requested from pathology providers for 300 patients. Of the samples requested from pathology providers, 570 were received (87%). The reasons for failure to receive 88 samples are expounded in Table 4.1. The most common cause was no response from the provider, despite multiple attempts at communication. An additional 16 samples were obtained through other means,

usually delivered by the patient themselves, giving a total of 586 samples received. 288 of 300 patients (96%) had at least one sample received.

The time from tissue request to receipt is shown in Figure 4.2, for the first tissue request made for each patient. The median time to receive samples was 12 days after request, and the mean time was 28 days. A small proportion of samples took considerably longer to arrive, the longest taking 8 months. As can be seen in the histogram, although the median time to receipt was similar between providers, the time to receipt was longer for external providers with a larger number of samples taking over one month to arrive (p value = 0.003, Wilcoxon rank sum test). In Figure 4.2B, the time from request to receipt generally improved over time regardless of provider. This is likely explained by the use of the parent institution's pathology administrative service to handle tissue requests from 2016 onwards.

The practice of external pathology providers changed during the study. Initially most providers would send a tissue block which was then sectioned at the parent institution. Generally, there was no cost for this service. In mid 2015 external providers declined to send blocks, choosing to section tissue and send unstained slides instead at the same time as charging a fee for this service. The fee varied from approximately $50 to $200 per sample. The higher fee charged by some providers necessitated that tissue held by these providers not be requested as part of the study.

The minimum recommended DNA input for SureSelect library preparation is considered to be 0.25μg. 66% of samples had greater than or equal to this amount of DNA extracted, which is in line with the literature. In a subset of samples with complete data, biopsy samples had less DNA (median 0.19μg, n = 200) than surgical specimens (median 0.75μg, n = 317) as expected ($p < 0.0001$, Wilcoxon rank sum test). There was considerable variability in DNA amounts across metastatic sites (Figure 4.3). Notably, some of the most common metastatic sites such as lymph nodes, liver, chest wall and bone all had median DNA amounts of approximately 0.25μg. As such 50% of samples will have less than the recommended amount of DNA available. For patients where there was < 0.25μg but no other samples were available, sequencing was attempted down to 0.1μg with varying success as will be discussed in the next section. 82% of samples had at least 0.1μg of DNA available post-extraction.

4.2 Sequencing logistics

The cumulative sequencing curves for all samples, and the first sample for each patient are shown in Figure 4.4A. The number of samples sequenced per month is shown in Figure 4.4B. Samples with sufficient DNA were sequenced as soon as possible, with the occasional exception if clinical urgency required some samples to be prioritised over others or the referring clinician deferred sequencing at their discretion. The median time from DNA extraction to sequencing was 76 days. This was driven by the availability of sequencing resources and the necessity of batching samples. Sequencing was performed in batches ranging in size from 3 to 67, with a median size of 13 samples. There were several reasons why batches of this size were used. It is most efficient to perform library preparation on over 10

samples, since the incremental increase in labour for each additional sample is small compared to that required for one sample. The second reason is efficient use of sequencing reagents. Consumable reagents used in sequencing afford a fixed number of reads per batch which cannot be adjusted up or down as needed. A mid output NextSeq flow cell for example has the capacity to sequence around 50 - 60 libraries with the capture size of the SEGMENT panel. There was not a sufficient number of samples in this study to regularly utilise a whole flow cell. Although in theory a single or small number of prepared libraries may be pooled with other libraries from different projects, in practice this is difficult and risks suboptimal performance. Aggregating samples into batches was therefore necessary.

Of 378 samples submitted for sequencing, 340 samples were successful while 38 samples failed to produce a result (10%). One sample had a library prepared but resulted in sequencing with coverage that was too low to be useable. One sample was lost during library preparation due to accidental expulsion from a benchtop micro-centrifuge with a faulty lid. Library preparation failed for 36 samples due to poor PCR yield prior to hybrid capture. As can be seen in Figure 4.5, failed samples had less starting DNA. Beyond low input DNA, it is also a well-known phenomenon that the DNA from formalin fixed samples may be resistant to PCR. Mechanisms found to contribute to this phenomenon include inhibition of DNA polymerase by short DNA fragments and extensive protein-DNA cross-linking which prevents polymerase engagement [245,246].

The site of origin of the tissue showed some association with failure rates, being highest for metastases in the GI tract (25%), which may reflect less favourable

fixation conditions. Bone metastases, which included bone associated soft tissue lesions and sclerotic metastases showed a lower failure rate of 4%. This is a surprising result as traditionally bone metastases are associated with higher sequencing failure rates, due to the use of decalcifying agents which destroy DNA [166]. The use of gentler decalcification protocols in routine diagnostic pathology is likely to have improved the yield here. Samples that failed were also older, with a median sample age of 1800 days for failed samples versus 588 days for successful samples ($p = 0.005$, Wilcoxon rank sum test).

The failure rate seen here begs the question if such failures could have been predicted or prevented. It is possible to assess DNA fragment size distribution prior to performing library preparation, using gel electrophoresis. Similarly, the efficiency of PCR on a sample may be determined quickly with real-time PCR. In practice, it was found that these methods entail the same labour and cost as the library prep itself while depleting the DNA available for sequencing. It was therefore decided to attempt library prep on every sample without any additional quality control measures aimed at determining samples likely to fail. Notably, only one sample that underwent successful library preparation produced suboptimal results after library capture and sequencing.

Of 300 patients recruited, 288 had at least one sample received during the study period (96% of original). Of those 288 patients, one patient died before a sample could be received and 241 had at least one sample with sufficient DNA to proceed to sequencing (80% of original). 214 patients had a sample sequenced successfully (71% of original).

Another key question is to what extent sequencing results became available before the death of the patient. It is acknowledged that the period of time during which a sequencing result may be useful to clinical decision making is not synonymous with the time the patient is alive, however survival can be more easily assessed. 27 of 214 patients with successful sequencing were deceased prior to sequencing results becoming available. The contributing factors for this outcome were as follows. For 14 patients, their survival after consenting to SEGMENT was unexpectedly less than 90 days, in contravention of one of the inclusion criteria. For another 5 patients, it took longer than 90 days to receive a tissue sample from the time of consent. For another 5 patients sequencing was initially deferred on clinical grounds introducing a delay. For another single patient, the first attempt at sequencing failed and obtaining and sequencing another sample could not be performed prior to death. For the remaining 2 patients, death occurred 1 day and 29 days before sequencing was completed.

Excluding these patients that died before sequencing was completed, the median time between completion of sequencing and death was 551 days or approximately 18 months, with a minimum of 38 days. Figure 4.6 shows the time to complete sequencing of at least one sample and the overall survival for each patient, calculated from the time of consent to SEGMENT. This includes patients sequenced on both platforms. It is evident that there is a fine margin between delivery of sequencing results and death for a small proportion of patients. 8% of patients had sequencing delivered within 6 months of death.

4.3 Sequencing performance

The computational requirements for processing sequencing data were easily met by the high-performance computing cluster at the parent institution. The mean time for alignment of sequencing reads was 368 seconds. Pre-processing of aligned reads took a mean of 129 seconds. Copy number analyses with copywriter were completed in a mean of 75 seconds per sample. Variants were called in a mean of 32 seconds. The time for performing the above analysis and curating the results was 2 weeks or less.

All samples were assed for the key metrics of the duplicate read rate, and reads on target. Data for all tumour samples are shown in Figure 4.7. The first sequencing run presented in the pilot phase is excluded, as it contained only fresh frozen diploid samples. Typical duplication rates of 25 – 60% were seen, but a considerable number of samples showed higher duplication rates up to 75%. Typical on-target rates of 10 – 25% were seen, which is largely expected for a small hybrid capture panel. As discussed in the prior section, off-target reads have utility in the chosen method for calling copy number.

The relationship between duplicate rate and DNA availability is shown in Figure 4.8. The duplication rate relates to how many unique DNA fragments exist in the sample, also termed the complexity of the DNA pool. The more unique fragments there are, the fewer duplicates will exist in the final library. There was some association between baseline DNA amount and the duplication rate, with enrichment for high duplication rates at low DNA levels (Figure 4.8A). It should be

noted that for samples with plentiful DNA, only around 0.25 to 0.5μg of DNA is used for library preparation. The numbers shown here for DNA amount do not therefore represent the actual input to the library preparation process. That there remains an association with duplication rate, even at DNA amounts far exceeding that used in library preparation suggests that DNA amount is a surrogate marker for the complexity of the DNA pool.

Figure 4.9 shows the mutation hotspot coverage in the whole cohort sequenced on the SureSelect platform, including the pilot samples already shown in Figure 3.7. Adequate coverage of these loci was in general preserved for all 340 samples. The median coverage for FFPE and fresh frozen samples was similar, but a greater proportion of FFPE samples had coverage < 100 fold. For the key *PIK3CA* loci H1047, E545 and E542 and *ERBB2* locus L755 the rates of sub-100 fold coverage were 2%, 0%, 0.03% and 4% respectively. The rates of sub 50 fold coverage were 0.03%, 0%, 0% and 1% respectively. All but one sample had at least 19 reads covering all loci.

4.4 Summary

Overall these data show that patients with metastatic breast cancer successfully underwent sequencing of tumour samples prospectively. The success rates were similar to those reported by others, despite samples being sourced from wide variety of community pathology services outside a comprehensive cancer centre. Recruitment to the study was consistent despite no active advertising, suggesting that patients and clinicians are motivated to access sequencing assays.

The time to receive the sample from the pathology provider was a major bottleneck. Increasingly during the study, the fees charged by pathology providers also became an impediment to accessing tissue. Economic factors may explain some of these problems, and may reflect the desire of pathology providers to offer their own in-house sequencing assays. More worrying was the decision by some pathology providers to destroy tissue after storing it for a certain period of time, in this case 10 years. This is of particular concern in breast cancer, as metastatic disease may develop many years after initial diagnosis, and tissue in the metastatic setting is not always available.

The most important cause of attrition during the study was inadequate DNA. Hybrid capture assays have fairly strict requirements for DNA quantity as could be seen with the elevated failure rates when less DNA was available. Efforts to minimise DNA loss and maximise extraction amounts are important to avoiding unnecessary attrition. The second most important cause of attrition was failure of library preparation. This would be improved by setting more stringent cut-offs for input DNA quantity, at the cost of excluding more patients, and thus a 5 – 10% failure rate is likely unavoidable in a real-world setting with hybrid capture assays.

Finally, the number of patients who died prior to sequencing becoming available was relatively high at almost 13%. Although some of these cases had sequencing deferred on clinical grounds, for the most part death prior to sequencing was due to unexpectedly short patient survival. It would seem imperative to refer patients early for sequencing to avoid this unfortunate outcome.

Table 4.1. Reasons for lack of receipt of tumour samples requested from pathology providers for 87 cases.

Number of samples	*Reason*
48	No response from pathology provider, despite multiple attempts at communication
18	Insufficient tissue available
14	Pathology report not available to make request (almost all for samples older than 5 years)
3	Tissue sample deliberately destroyed by pathology provider
3	Tissue sample lost by pathology provider
2	Tissue sample not suitable for DNA extraction due to de-calcification procedures

Figure 4.1. Recruitment data. A) Cumulative sum of recruited patients. B) Monthly recruitment numbers. In mid 2015 the project was expanded to include circulating tumour DNA analysis, which is not the subject of this thesis.

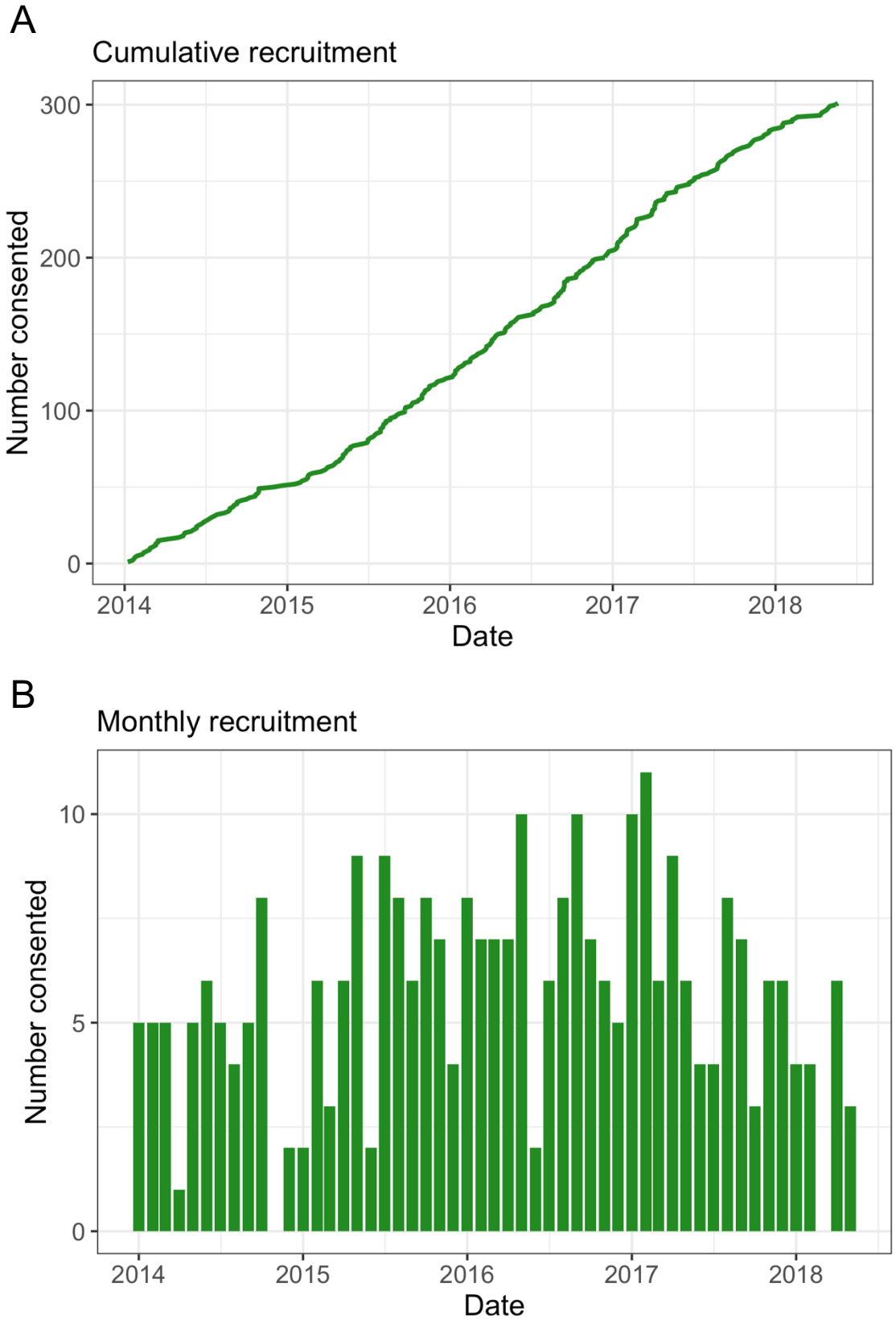


Figure 4.2. Receipt times for tissue requests. A) Stacked histogram of time to receive tissue samples from pathology providers. Some cases from the parent institution were received the day of or shortly after request. B) Time to receipt by provider across years of study. Outlier points are not shown for clarity.

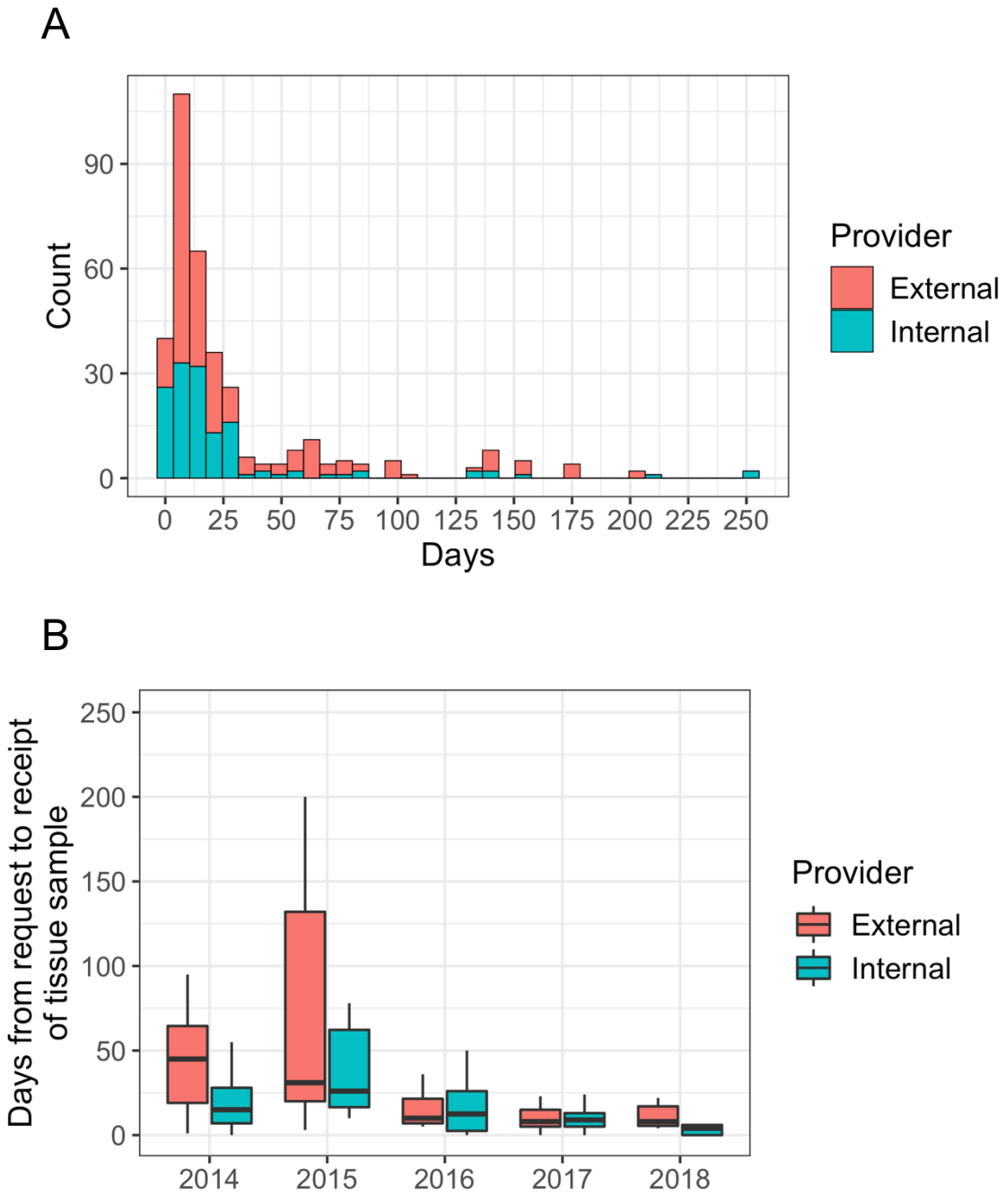


Figure 4.3. DNA extraction amount by metastatic site. 'other' category includes muscle, epidural and bone marrow metastases.

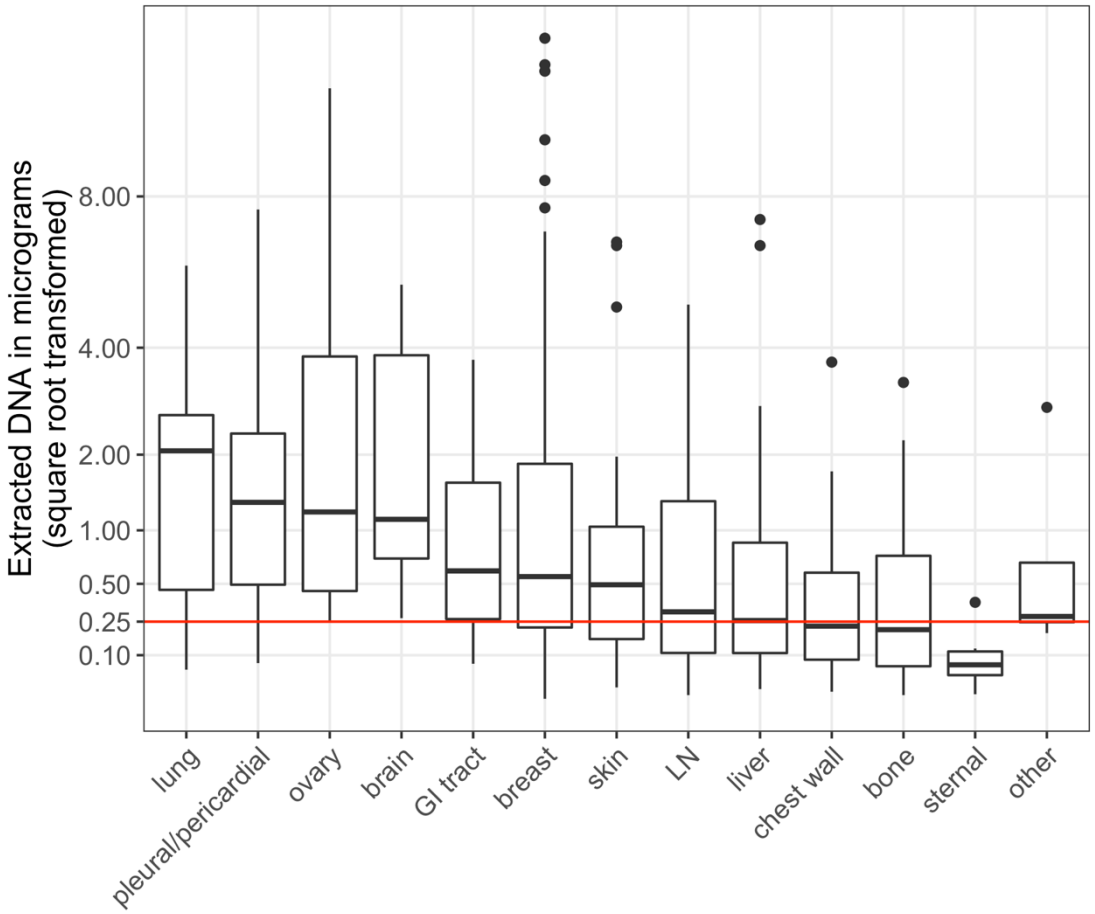


Figure 4.4. Number of samples sequenced. A) Cumulative sum for all samples (green). As there were multiple samples per patient, the cumulative sum for only the first sample for each patient is also shown (orange). B) Monthly sequencing numbers.

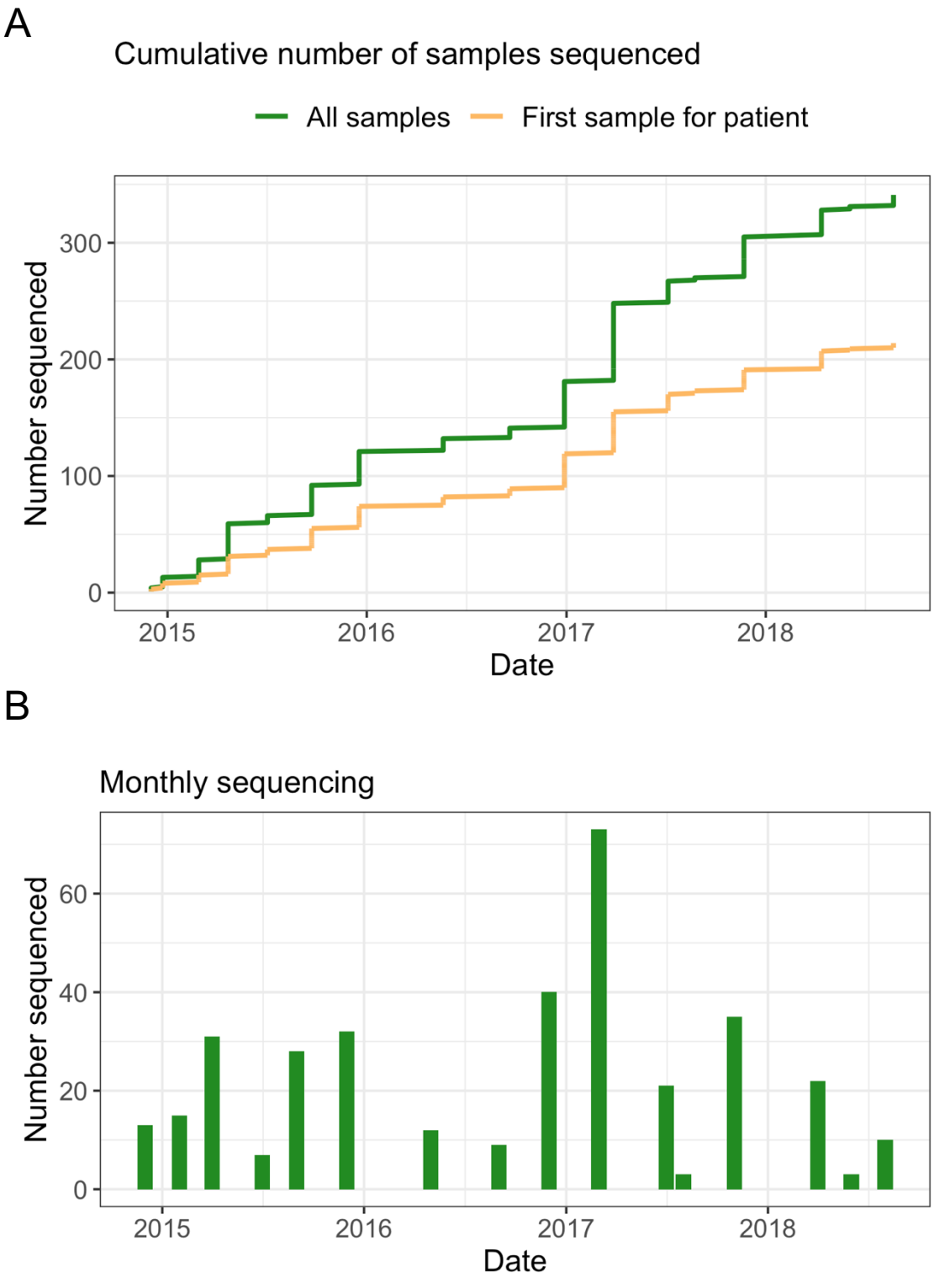


Figure 4.5. Comparison of starting DNA amounts for samples that failed or succeeded sequencing. Wilcoxon rank sum test.

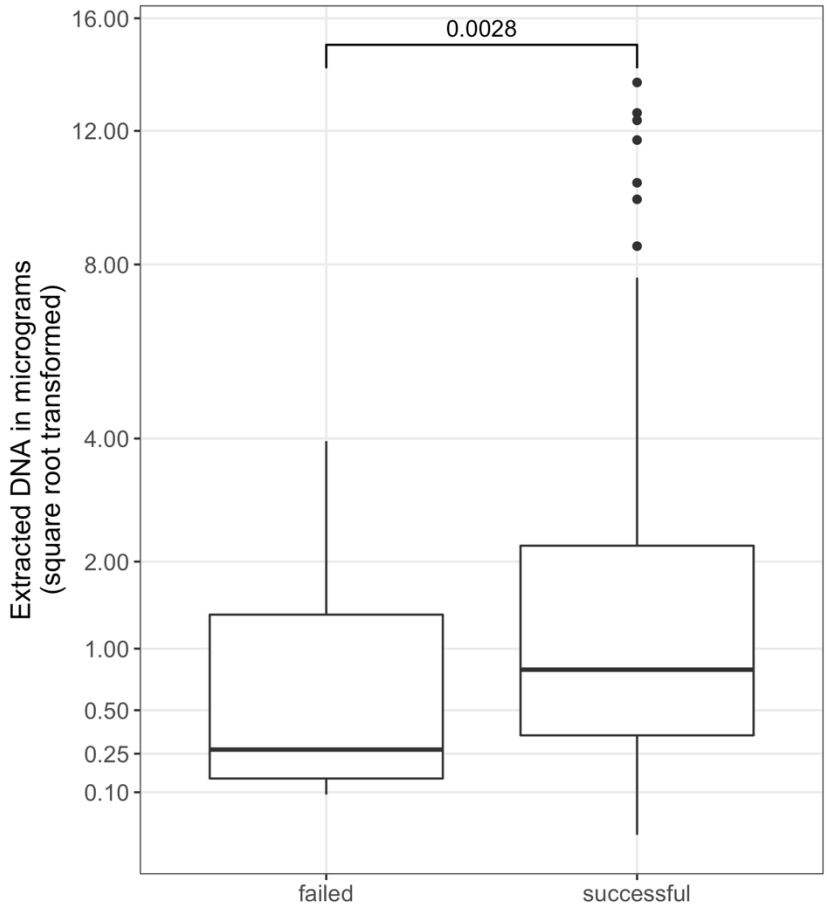


Figure 4.6. Time to sequencing of at least one sample and overall survival from the time of consent. Censored data for patients alive at the time of final follow up is shown with a black arrowhead.

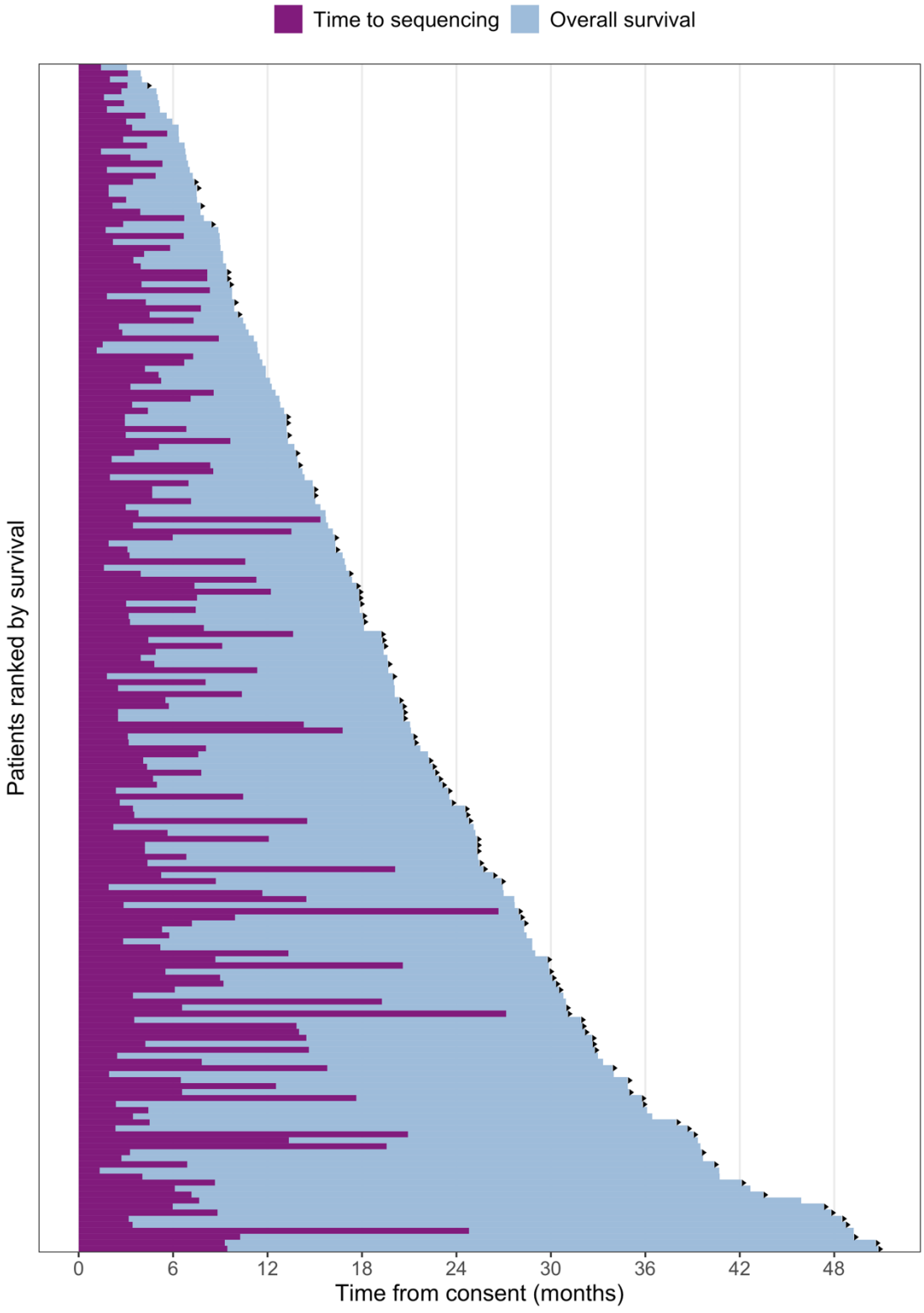


Figure 4.7. Key sequencing metrics in 4 pilot runs on the redesigned SureSelect platform. A) Percentage of duplicate reads. B) Percentage of bases in exons of panel genes. Alternating box plots are shaded different colours for clarity.

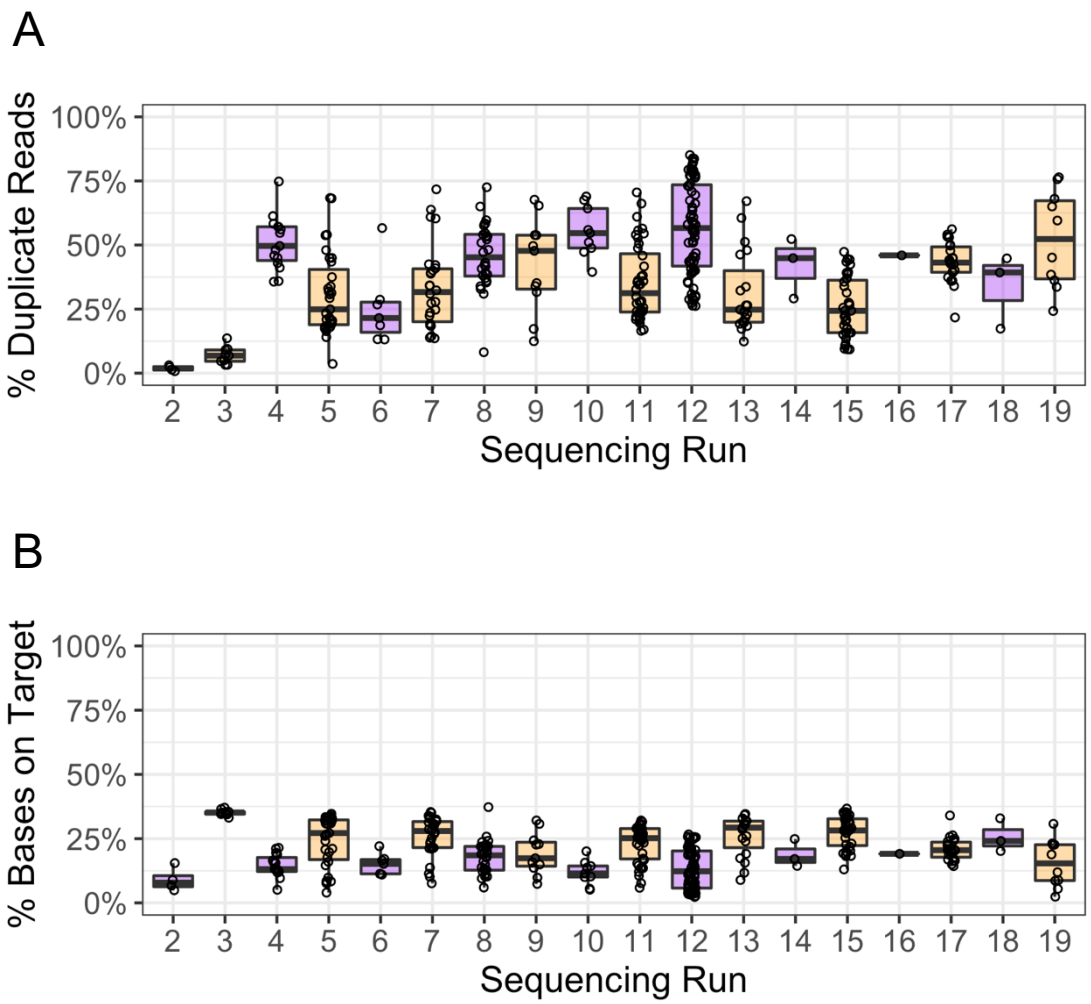


Figure 4.8. Association between DNA amount available and duplication rate.

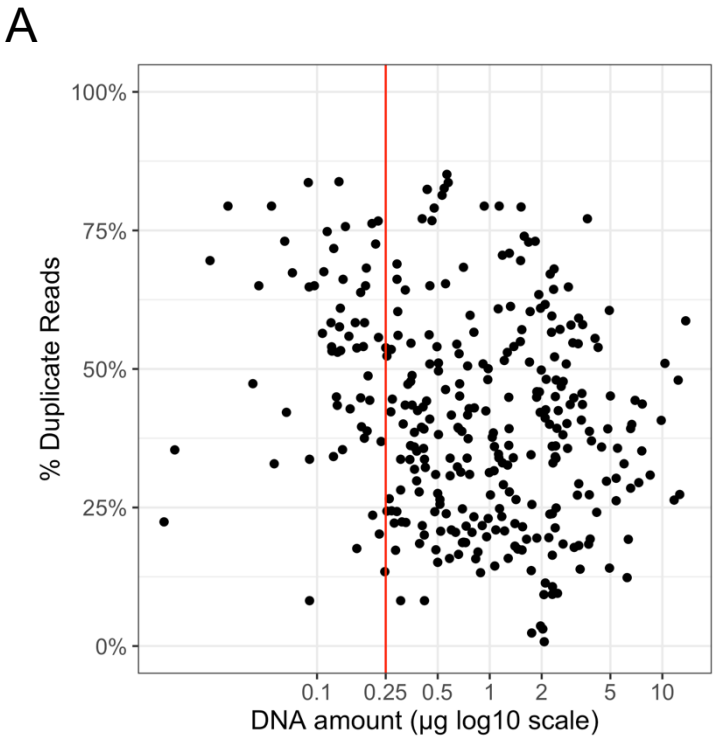
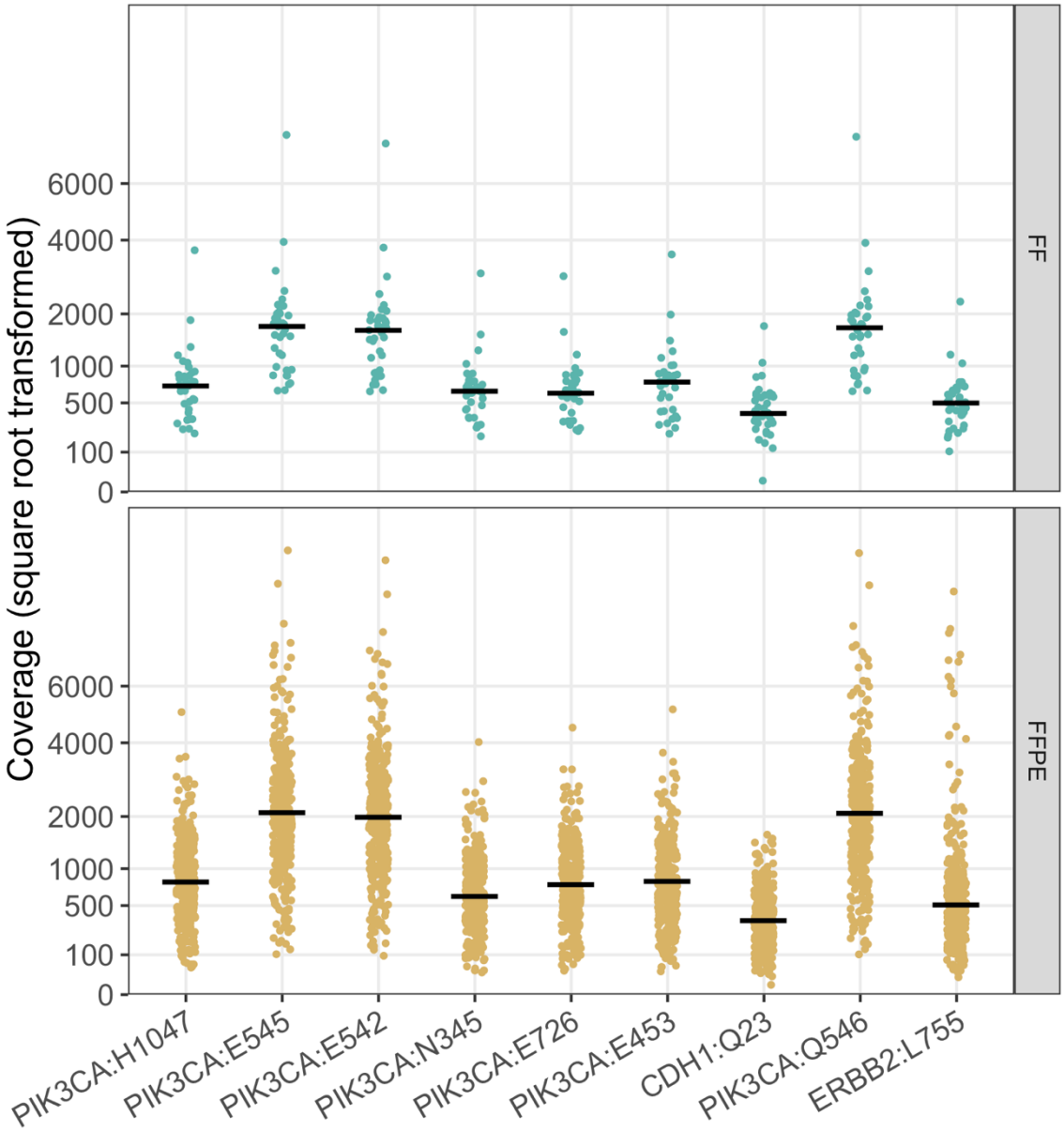


Figure 4.9. Coverage of commonly mutated loci in breast cancer in high quality fresh frozen GIAB/germline DNA (FF) and breast tumour FFPE samples from full study cohort on SureSelect platform. Includes pilot data presented in Figure 3.7. Loci ranked on x-axis left to right by decreasing frequency. Y-axis is square root transformed.



Chapter 5: SEGMENT results

5.1 Demographic variables

Breast cancer subtypes were determined on the basis of standard histopathological assessment of oestrogen receptor staining and progesterone receptor staining by immunohistochemistry, and HER2 status by immunohistochemistry and in situ hybridisation. Twelve of 214 patients displayed changes in subtype over time. Four patients had a 'late-switch' from an ER+HER2- subtype to TNBC after multiple lines of therapy. Two patients had an 'early-switch' from ER+HER2- early stage disease to TNBC at the first diagnosis of advanced disease. Two patients had metastatic lesions which were variously ER+HER2- and TNBC. Two patients had a late-switch from ER+HER2- to ER+HER2+. One patient switched from a TNBC primary to ER+ metastatic disease. One patient with ER+HER2+ early stage disease developed advanced disease which was variously ER+HER2- and TNBC. Where a switch in subtype occurred, a patient was allocated to a subtype according to the clinical judgement of the patient's treating clinician.

Table 5.1 shows the characteristics of the 213 patients that had samples sequenced successfully on the SureSelect platform. One additional patient with a malignant phyllodes tumour is not shown in these tables. Most patients had ER+HER2- disease, which will be referred to as simply ER+ from here onwards. The age of patients, and the proportion of de novo metastatic disease are typical. The majority of ER+ and TNBC patients received adjuvant or neoadjuvant chemotherapy, and a high proportion of ER+ patients had received adjuvant

endocrine therapy. Across subtypes, the median time of diagnosis of metastatic disease until enrolment ranged from 18 months for ER+, 12 months for HER2+ and 10 months for TNBC. This is noteworthy since for TNBC, overall survival is approximately 18 months, leaving little time after consent to obtain results and act on them. This is also reflected in the lines of therapy data, with patients with TNBC receiving a median of two lines of therapy, compared to 3 lines for ER+ and HER2+ disease. Patients with triple negative breast cancer had also received a median of 1 line of therapy prior to enrolment compared to 2 or more for ER+ and HER2+ disease. Figures 1 and 2 show the disposition of treatment and the time of consent to SEGMENT for patients which had data available on the timing of treatment. These figures illustrate the differing tempo of events across subtypes, with ER+ and HER2+ disease generally having long treatment histories both before and after enrolling in SEGMENT.

Table 5.2 shows the characteristics of samples that were sequenced as part of SEGMENT. Samples were a median of 18 to 24 months old at the time of sequencing. TNBC samples were less in age due to the shorter disease-free interval in this subtype. Histograms of the sample age are shown in Figure 5.3. Of the samples sequenced, 50 to 60% of samples were archival samples from before the advanced disease had been diagnosed and were primary breast tumour samples. This is to be expected as SEGMENT protocol did not mandate new biopsies on enrolment. Table 5.3 shows the timing of samples per patient. About half of all patients had a primary tumour sample sequenced which was collected prior to the diagnosis of advanced disease. Approximately one third had samples from both before and after the diagnosis of advanced disease and 16-25% had samples from

the advanced disease setting only. A median of 1 sample was sequenced per patient., and 92 (43%) patients had more than one sample sequenced. The breakdown by subtype is shown in Table 5.3, with around 40% of patients having multiple samples sequenced in all subtypes.

5.2 Overview of genomic alterations

Data here are presented for the 61 genes of interest, enriched for genes recurrently altered in breast cancer and those considered to be actionable across breast cancer and other cancer types. A median of 2 mutations was detected per sample sequenced, ranging from 0 to a maximum of 7. A median of 3 copy number alterations were detected per sample sequenced, ranging from 0 to a maximum of 12. There were no differences in the number of mutations or copy number changes between subtypes, and whether the sequenced sample was a primary tumour, a local recurrence, or a metastatic lesion.

Figure 5.4 shows the most common mutations detected in the overall cohort. Figures 5.5 and 5.6 show similar data for each subtype. Data has been aggregated at the patient level where multiple samples per patient were sequenced. The results here are broadly consistent with the known prevalence of mutations in breast cancer across the studies discussed in Chapter 1. In the small number of HER2+ cases, nearly 100% had *TP53* mutations. The prevalence of *TP53* mutations in primary HER2+ tumours in METABRIC was 52%. In two metastatic breast cancer cohorts, the prevalence of *TP53* mutations was 55% in an endocrine resistant MSK cohort from 2018 [247] and 30% in an INSERM metastatic breast cancer cohort from 2016 [108]. These differing frequencies may be due to chance

but may also reflect changing genomic profiles of advanced HER2+ disease in the era of modern anti-HER2 therapy.

Figures 5.7 and 5.8 show the prevalence of copy number alterations. Copy number alterations were curated according to their putative oncogenic role, namely amplification of an oncogene and deletion of tumour suppressor genes. These results are also broadly consistent with prior studies, with *CCND1* and *FGFR1* amplification among the most common alterations. *MCL1* amplification was commonly seen with a prevalence of 24% in the overall cohort and 47% in TNBC. Prior studies in the metastatic setting have reported *MCL1* amplification at a frequency of 3.7% (MSK) and 11.1% (INSERM). The method for calling amplifications in this study and both these published cohorts is not consistent. Re-analysing the MSK cohort using a similar log ratio threshold of 0.7 as was employed in this study, the prevalence of *MCL1* amplification returns at 11.6% in the overall cohort (n = 1891 with copy number data available) and 16% in TNBC (n = 173).

5.3 Comparing genomic profiles of primary and metastatic lesions

The frequency of mutations and copy number changes were compared in primary and metastatic lesions. As before, if a primary breast mass was biopsied after the patient had been diagnosed with metastatic disease, this was classified as a metastatic lesion. Overall there were no statistically significant differences between the primary and metastatic samples. The analysis was limited by the presence of paired samples. Numerically there were more *ESR1* mutations in the overall cohort (Figure 5.9) which was driven by ER+ samples (Figure 5.10). This

preponderance of *ESR1* mutations in advanced ER+ disease is well known, and results from the exposure of the tumour to tamoxifen or aromatase inhibitors [92–94].

A similar analysis for copy number alterations revealed amplification of *CCND1* was significantly more frequent in metastatic samples (Fisher's exact test, Benjamini-Hochberg adjusted p-value of 0.02). Frequencies are shown in Figure 5.11 for the overall cohort, and Figure 5.12 shows data for ER+ cancers only where *CCND1* was more prevalent in metastatic disease but the difference was not significant. Copy number data for TNBC is shown in Figure 5.13, where non-significant trends are seen to more *MAP2K4* and *CDKN2A* deletions in metastases at low frequency.

5.4 Heterogeneity of genomic profiles

Seventy-four cases had a primary sequenced and at least one metastatic lesion. An additional 10 cases had at least 2 metastatic lesions sequenced, without a primary. Twenty-seven cases had more than one metastatic lesion sequenced, with or without a primary tumour. The focus of this analysis will be comparing primary and metastatic tumours. As seen in Table 5.2, primary tumours are the chief source of tumour material for sequencing in this real-world cohort where new biopsies are not performed routinely for the purpose of sequencing. The concordance of genomic profiles between primary and metastatic tumours is therefore pertinent.

Figure 5.14 presents concordance data for mutations detected in cases with at least one primary and at least one metastasis sequenced. Sixty-seven cases are included here, with 7 excluded due to a paucity of mutations in genes of interest. It is apparent that for common driver alterations, the levels of concordance are high. Out of 175 mutations across all samples, 52 were discordant between primary and metastasis, a discordance rate of 29%. Thirty-four mutations were present in the metastasis but not the primary (19% of total), and 18 were present in the primary and not the metastasis (10% of total). Restricting the denominator to actionable mutations only (58 mutations), 44 were concordant, 12 were metastasis only (21%) and 2 were primary only (3%). The most common discordant actionable mutations were in *ESR1*, with 5 out of 7 mutations absent in the primary lesion (71% discordance), followed by *PIK3CA* with 2 of 23 absent in the primary (9%) and *ERBB2* with 2 of 5 absent in the primary (40%). Primary-metastasis discordance rates for all mutations across subtypes differed (ER+ 24%, HER2+ 0% and TNBC 13%), but the number of mutations and cases are too small to draw definitive conclusions.

ERS1 and *ERBB2* are both known to develop in ER+ cancers exposed to therapy [92,248]. The discordance rate for *PIK3CA* although low is unexpected, as it is considered a truncal driver that is mutated early in the evolution of the tumour. The first case to display discordance for *PIK3CA* developed an E542K mutation in a chest wall metastasis. The cancer was ER+. The primary breast cancer was resected 20 years prior the biopsy of the chest wall metastasis. An earlier biopsy of the chest wall disease at first recurrence 7 years from resection of the primary also lacked the *PIK3CA* E542K mutation. Also notable is that this ER+ cancer had

a *PTEN* frameshift deletion which was present in all lesions. The allele frequency of the *PTEN* frameshift was persistently high in all lesions ranging from 25 to 67%, whereas the *PIK3CA* mutation had an allele frequency of 14% indicating it was likely subclonal. Thus in this case, an alternate driver in the PI3K pathway and a very long disease history likely explain the absence of the *PIK3CA* mutation in the primary tumour.

The second case with discordance was an ER+ cancer. Two years after the diagnosis of early stage disease a brain metastasis developed. This was sequenced as well as the primary cancer. There was no evidence of the *PIK3CA* mutation in the primary when examining the locus by visual inspection, to account for the possibility of a false negative variant call. The brain metastasis had in addition an *RB1* frameshift deletion, and an *AKT1* mutation of unclear significance, both of which were at high allele frequency but absent in the primary tumour, which had no mutations. Inspection of the whole genome copy number profile however showed that the primary and brain metastasis had copy number aberrations which were highly concordant. The treatment administered in this case was otherwise unremarkable consisting of surgery, adjuvant chemotherapy and radiotherapy.

Copy number alterations generally displayed more heterogeneity between primary and metastatic lesions. Of 277 curated copy number alterations in genes of interest, only 70 were fully concordant (25%). We found 148 alterations which were absent in the primary lesion compared to local recurrences or metastatic lesions (53%). Restricting the analysis to 37 actionable copy number alterations

(namely deletions of *BRCA1, BRCA2, RB1, PTEN* and amplification of *ERBB2*), 30% of alterations were concordant and 57% were absent in the primary. As can be seen in Figure 5.15, metastatic lesions appear enriched for copy number alterations in general when compared to paired primary lesions accounting for the discrepancy. Primary-metastasis discordance rates were similar across subtypes for all copy number alterations (ER+ 58%, HER2+ 53%, TNBC 52%)

5.5 Actionable alterations and clinical impact

Actionable alterations were curated as described in section 3.8 of chapter. Table 5.4 lists actionable alterations found in patients that lead to treatment recommendations. The recommendations changed throughout the study due to changing evidence. Deleterious alterations in *PTEN* were thought to predict benefit from PIK3CA inhibitors, but in fact do the opposite [242]. In TNBC however, there is evidence that loss of *PTEN* function predicts benefit from AKT inhibition together with chemotherapy [249]. In some cases, multiple actionable alterations were found in the same tumour.

The impact on the care of the 214 patients with sequencing results is shown in Table 5.5. Half of these patients did not return actionable results. An additional 28.5% did not have their care altered due to actionable results. The reasons for this were in part due to the treating clinician deciding an experimental therapy was not appropriate, however the most common reason was that patients were ineligible for available trials. Deterioration of the patient's health prior to therapy and lack of access to targeted therapies were also important causes.

Thirty-one patients (14%) received a therapy matched to genomic findings. Off-trial use of targeted therapies generally had poor outcomes, with no objective responses and one case of stable disease. Patients receiving off-trial therapy generally had higher treatment exposure and disease burden. Almost all patients receiving matched therapies on trial were treated with alpelisib in the phase 2 single arm PIKNIC study. This was an investigator lead single institution study at the parent institution. One additional patient was treated with taselisib in a phase 3 study. In this small and heterogenous group, the crude response rate was 23%, with another 43% of patients exhibiting some clinical benefit with stable disease or mixed responses. In considering the size of the original cohort of 300 patients, allowing for attrition due to sample logistics and technical factors, approximately 10% received a matched therapy and 2% achieved a response to matched therapy.

One question is whether the tumour only assay employed in SEGMENT is adequate to detect germline mutations in *BRCA1* or *BRCA2*. Although known germline mutations were detected in some cases, two cases of known *BRCA2* mutations were not detected. One germline mutation was a whole exon deletion, and another a smaller complex deletion. Considering that detection of *BRCA1* or *BRCA2* mutations is highly relevant to treatment decisions, the tumour only FFPE assay employed here should not replace clinically validated testing.

One of the hopes for precision oncology has been that driver alterations with known clinical significance in certain cancer types will be detected in other cancer types, such that targeted therapies can be applied rapidly outside of their established indication. This occurred in one patient with a triple negative

metaplastic breast cancer that had a *BRAF* V600E mutation detected in multiple samples. Treatment with the BRAF inhibitor, vemurafenib, in combination with the EGFR inhibitor, erlotinib, was accessed via a clinical trial, but no objective response was detected, and clinical benefit was lost after a few months of treatment. The patient also had a *PIK3CA* mutation H1047R and did achieve a partial response to the alpha-subunit specific PIK3CA inhibitor, alpelisib. Although there is strong, high quality evidence for the benefits of BRAF inhibitors in melanoma with V600E mutations [250], the difficulties of transferring this to other cancer types with BRAF mutations is well known [149]. This case also highlights the complexity of treating tumours where there is more than one driver mutation. The variant allele frequencies of the *BRAF* and *PIK3CA* mutations were very similar in both samples sequenced (30 – 40%), suggesting they are within the dominant clone. The prevalence of *BRAF* V600E mutations in breast cancers is 3 of 5762 cases or 0.05% in a combined cohort from TCGA, METABRIC and 3 metastatic breast cancer cohorts [67,68,108,247]. Two of these 3 cases were metastatic.

5.6 Summary

The SEGMENT cohort is representative of the contemporaneous management of advanced breast cancer in Australia. All patients had received or were receiving prior therapies at the time of recruitment and continued to receive treatment after recruitment. Breast cancer is in general a highly treatable malignancy, and thus one of the challenges to implementing a precision oncology program is interdigitating sequencing and targeted therapy recommendations with existing therapies. This is most evident in HER2+ disease, where anti-HER2 therapy is

administered continuously throughout the treatment history, and the use of novel therapies necessitates adding on to the anti-HER2 backbone.

Although many of the sequenced tumours were primary tumours, it was possible to see evolutionary forces imposed by therapy at work with emergence of *ESR1* mutations and to a lesser extent *ERBB2* mutations. These trajectories were particularly evident with the examination of paired primary and metastatic samples. An important finding in this cohort is that driver mutations may not always be present in primary tumours. This was a rare event for *PIK3CA* and was predictable for one case with a very long natural history and alternative drivers. There may be technical confounding variables related to sample characteristics and sequencing performance contributing to these differences, however high-depth targeted sequencing should detect clinically meaningful alterations with high sensitivity. It is possible to model the sensitivity for detecting mutations in a sample using the ploidy and purity of the tumour. These values are not known, but an approach exists to infer these values [233].

The development of circulating tumour DNA assays is likely to alleviate these problems of heterogeneity, as all active disease contributes nuclear material to the circulating pool and avoids sampling error. A downside of circulating tumour DNA assays is that the most sensitive assays require knowledge of which mutations to detect. If there are high rates of novel mutations to be found, then a circulating assay would miss this potentially important information. These rare and novel mutations of clinical significance were not seen in this cohort. The one case with a rare *BRAF* mutation did not respond well to BRAF inhibition.

In this study, a reasonable proportion of patients were provided with therapeutic recommendations and were able to receive the recommended therapies. The figure of 50% for patient with actionable alterations, putting aside the lack of a consistent definition actionable, compares favourably with other studies discussed in more detail in section 1.12. Receipt of matched therapy by 10% of patients also compares favourably with prior studies. This was greatly facilitated by the presence of a single arm phase 2 study with broad inclusion criteria that provided access to alpelisib which targets one of the most commonly mutated genes in breast cancer. The rapid tempo of some patients' disease and the advanced stage at which they present for testing mean that additional delays for referral to other institutions or procession through molecular tumour boards may cause further attrition of the population of patients that benefit from precision oncology. That only around 2% of patients appeared to benefit from matched therapy is a disappointing result, which would only be marginally improved by more effective therapies for existing targets. Better opportunities for patients to receive matched therapies is likely to be crucial to improving outcomes from precision oncology, and moving testing as close as possible to the time of diagnosis is the most easily identifiable way to begin. An integrated testing and early phase clinical trial program with a plurality of targets and broad inclusion criteria would also enhance uptake, but this requires a large investment of resources.

Table 5.1. Characteristics of study patients

	ER+	HER2+	TNBC
Number of patients	148	20	45
Median age at consent [min-max]	53 [31-78]	48 [39 -76]	50 [24 - 77]
De novo metastatic disease	32 [22%]	9 [45%]	4 [9%]
Median disease-free interval from primary diagnosis (months, de novo patients excluded)	45 [1 - 277]	23 [4 - 57]	20 [1 - 201]
Median time from primary diagnosis to consent in months	70 [1-341]	48 [6-88]	27 [0-201]*
Median time from advanced diagnosis to consent in months	18 [1-168]	12 [1-74]	10 [1-74]
Adjuvant chemotherapy	65 [44%]	6 [30%]	16 [36%]
Neoadjuvant chemotherapy	21 [14%]	0	9 [20%]
Adjuvant endocrine therapy	104 [70%]	2 [10%]	2 [4%]**
Median lines of therapy in advanced disease during study	3 [1-12]	3 [1-9]	2 [1 -5]
Median lines of therapy in advanced disease at time of consent	2 [1-11]	2.5 [1-3]	1 [1-3]

* 2 patients with TNBC that were considered to have a high risk of metastatic disease due bulky disease were enrolled pre-emptively.

** 2 patients with TNBC had hormone receptor positive disease in the early stage setting and thus received adjuvant endocrine therapy.

Table 5.2. Characteristics of sequenced samples

	ER+	HER2+	TNBC
Number of samples	233	31	74
Median samples sequenced per patient	1 [1-5]	1 [1-3]	1 [1-5]
Median age of sequenced samples (days)	754 [46 - 7536]	796 [124 - 4876]	564 [60 - 5066]
Samples from before metastatic disease diagnosed	118 [51%]	16 [52%]	45 [61%]
Location of sequenced samples			
breast	125 [54%]	15 [48%]	45 [61%]
lymph node	31 [13%]	2 [6%]	7 [9.5%]
bone	20 [9%]	2 [6%]	1 [1%]
liver	18 [8%]	4 [13%]	5 [7%]
chest wall	11 [5%]	3 [10%]	5 [7%]
ovary	8 [3%]	0	0
brain	6 [2.5%]	2 [6.4%]	7 [9.5%]
lung	5 [2%]	3 [10%]	4 [5.4%]
GI tract	3 [1.3%]	0	0
skin	2 [0.9%]	0	0
ascites	1 [0.4%]	0	0
other	3 [1.3%]	0	0
Total samples	233	31	74

Table 5.3. Timing of samples aggregated on patient level, and proportion of patients with > 1 sample sequenced.

	ER+	HER2+	TNBC
Timing of sample			
Early stage disease	59 [40%]	9 [45%]	23 [51%]
Advanced stage disease	37 [25%]	4 [20%]	7 [16%]
Early & advanced stage disease	52 [35%]	7 [35%]	15 [33%]
More than one sample per patient	64 [43%]	8 [40%]	20 [44%]

Relevant Alteration	Context	Recommendation
AKT E17K	ER+, TNBC	AKT inhibitor
AKT E17K	TNBC	AKT inhibitor + chemo
BRCA1 K859E	ER+,TNBC	PARPi or platinum
BRCA1 splicing	ER+,TNBC	PARPi or platinum
EGFR L480_P486delinsS	TNBC	EGFR inhibitor
ERBB2 D769Y	ER+,TNBC	HER2 TKI
ERBB2 I767F	ER+,TNBC	HER2 TKI
ERBB2 I767M	ER+	HER2 TKI
ERBB2 L755_E757delinsS	ER+,TNBC	HER2 TKI
ERBB2 R345C	ER+,TNBC	HER2 TKI
ERBB2 V750L; PIK3CA H1047R; PTEN N212TFS*2	ER+	PIK3CA inhibitor/HER2 TKI
ERBB2 V777L	ER+,TNBC	HER2 TKI
ESR1 D538G	ER+	Potential resistance to endocrine therapy
FGFR1amp	ER+	FGFR1 inhibitor
HRAS G13R	TNBC	RAS pathway inhibitor
PIK3CA E418K	ER+	PIK3CA inhibitor
PIK3CA E542K	ER+	PIK3CA inhibitor
PIK3CA E545A	ER+	PIK3CA inhibitor
PIK3CA E545K	ER+	PIK3CA inhibitor
PIK3CA E726K; PIK3CA G1049R	ER+	PIK3CA inhibitor
PIK3CA H1047L	ER+	PIK3CA inhibitor
PIK3CA H1047Q	ER+	PIK3CA inhibitor
PIK3CA H1047R	TNBC	AKT inhibitor
PIK3CA H1047R	ER+	PIK3CA inhibitor
PIK3CA H1047R, PIK3CA N345T	ER+	PIK3CA inhibitor
PIK3CA H1047R; BRAF V600E	TNBC	BRAF inhibitor; PIK3CA inhibitor
PIK3CA M318I	ER+	PIK3CA inhibitor
PIK3CA N345K	ER+	PIK3CA inhibitor
PIK3CA N345T	ER+	PIK3CA inhibitor
PIK3R1 Q336X	ER+	PIK3CA inhibitor
PIK3R1 T213del	ER+	PIK3CA inhibitor
PTEN A126D	ER+	Potential resistance to PIK3CA inhibitors
PTEN C71Vfs*27	TNBC	AKT inhibitor + chemotherapy
PTEN I300Tfs*5	ER+	Potential resistance to PIK3CA inhibitors
PTEN I306Yfs*5	ER+	Potential resistance to PIK3CA inhibitors
PTEN K6Q	ER+	Potential resistance to PIK3CA inhibitors
PTEN N48D	TNBC	AKT inhibitor + chemotherapy
PTEN R130Q	ER+	Potential resistance to PIK3CA inhibitors
PTEN splicing	TNBC	AKT inhibitor + chemotherapy
PTEN T319fs*0	TNBC	AKT inhibitor + chemotherapy
PTEN T319fs*0	ER+	Potential resistance to PIK3CA inhibitors
PTEN T319Kfs*23	ER+	Potential resistance to PIK3CA inhibitors

Table 5.4 (previous page). Actionable alterations leading to treatment recommendations. TKI: tyrosine kinase inhibitor. PARPi: PARP inhibitor.

Table 5.5. Clinical outcomes following sequencing for 214 patients with sequencing results. Responses to therapy are shown in abbreviation: Stable disease (SD), progressive disease (PD), partial response (PR).

Referral to familial cancer clinic	**14**	**(6.5%)**
Matched therapy off trial	**5**	**(2.3%)**
ERBB2 TKI neratinib	3	(1 SD, 2 PD)
AKT inhibitor ipatasertib	1	(PD)
MEK inhibitor cobimetininb	1	(PD)
Matched therapy trial	**26**	**(12%)**
PIK3CA inhibitor apelisib	24	(5 PR, 10 SD/mixed, 9 PD)
PIK3CA inhibitor taselisib	1	(PR)
FGFR1 inhibitor TAS120	1	(SD)
Nil actionable results	**108**	**(50%)**
Nil change in care	**61**	**(28.5%)**
Known germline mutations	4	
ESR1 mutations not relevant	4	
PTEN in ER+ (resistance)	4	
ERBB2 mutation not considered relevant	1	
Therapy recommended but not received	**47**	
Not eligible for available trial		14
Clinician decided against therapy		11
Patient died before result or too unwell		7
Therapy not available on trial or by other means		6
Patient choice		4
Responding to other therapy		3
Already received therapy		1
Left country		1

Figure 5.1. Treatment history and timing of study enrolment for ER+HER2- disease. 132 cases had data for timing of treatment dates. 'X' marks date of diagnosis of metastatic disease. Thin grey lines show duration of survival where '•' indicates censoring for patients still alive at last follow up.

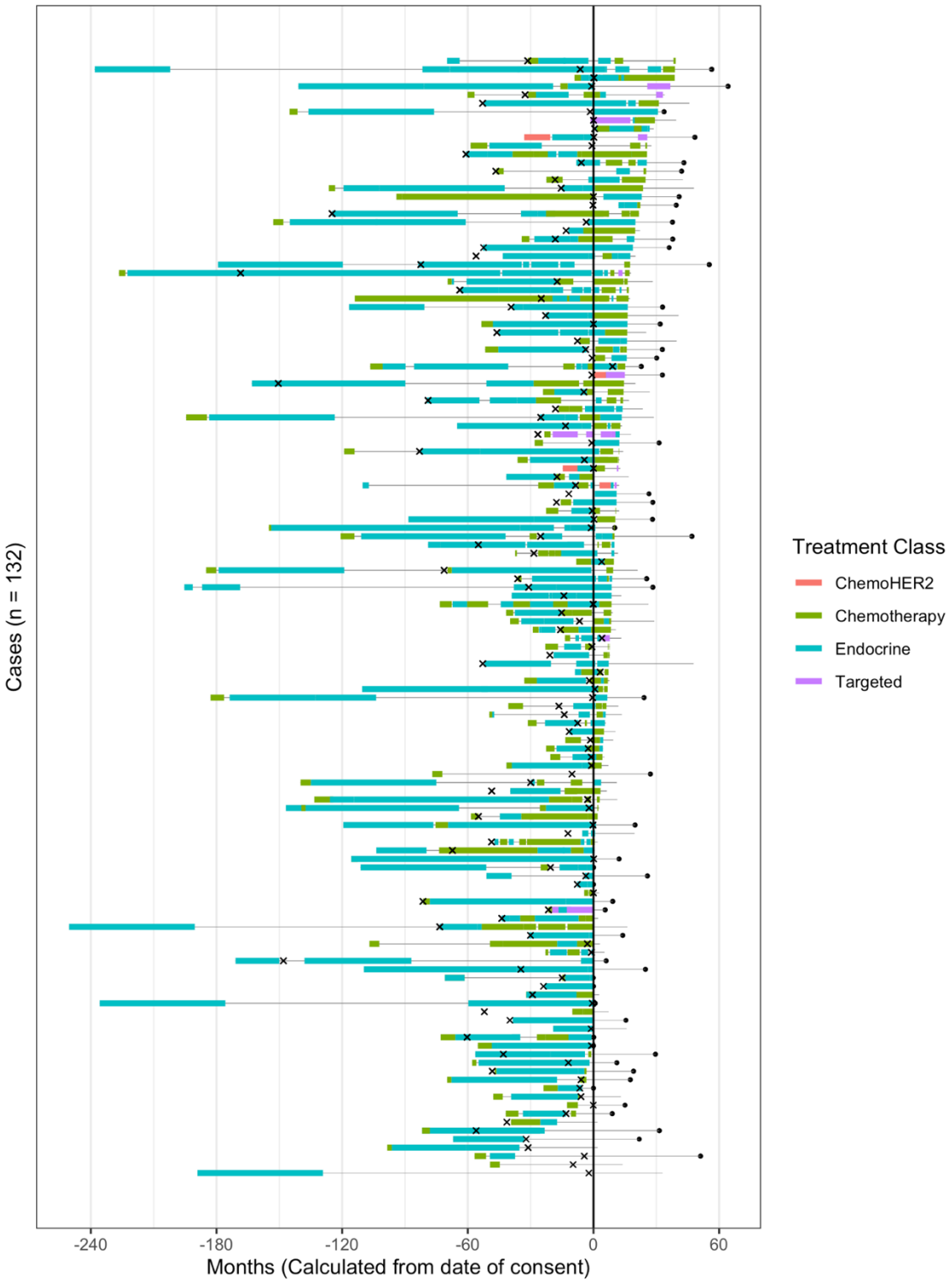
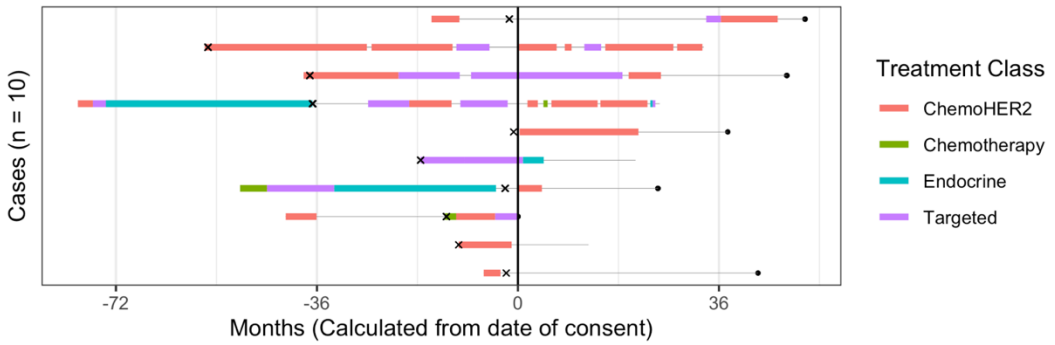


Figure 5.2. Treatment history and timing of study enrolment for A) HER2+ disease (n = 10) and B) Triple negative disease (n = 20). For HER2+ disease, 'Targeted' refers to HER2 targeted therapy without chemotherapy.

A



B

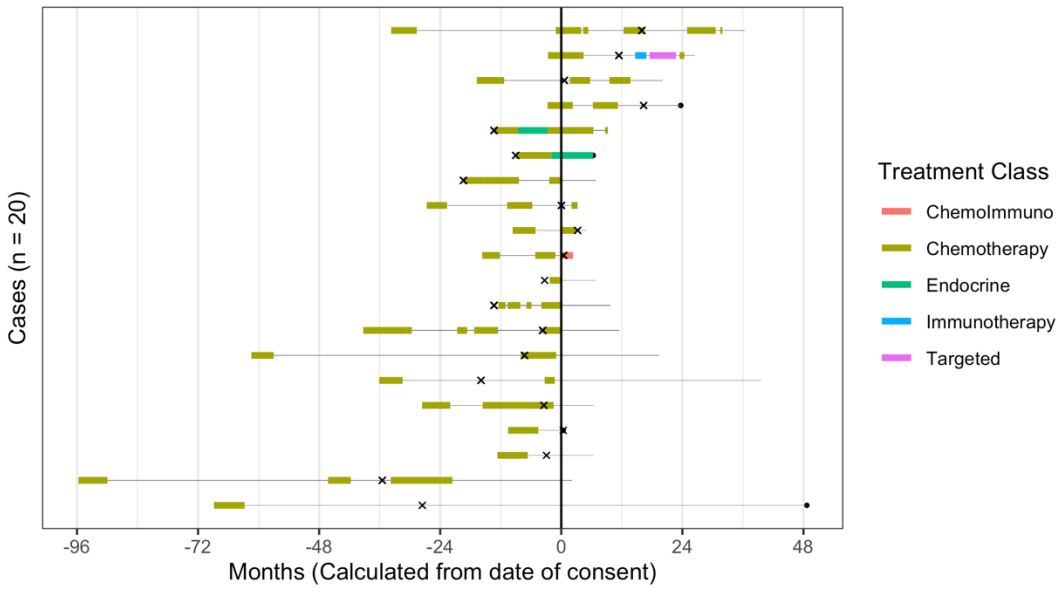


Figure 5.3. Age of samples at time of sequencing across subtypes.

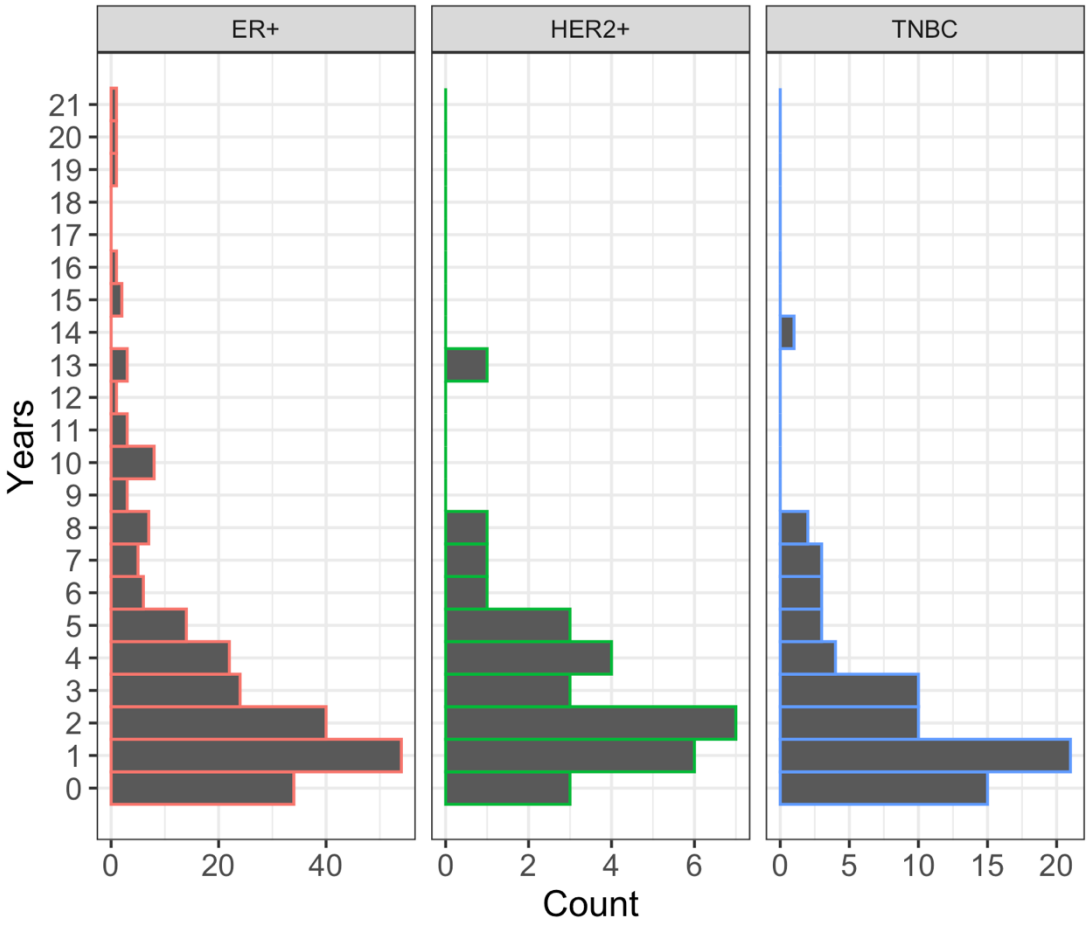


Figure 5.4. Mutation prevalence for all cases. 194 of 214 patients had sequencing which returned a mutation result.

Figure 5.5. Mutation prevalence per case for ER+ cancers.

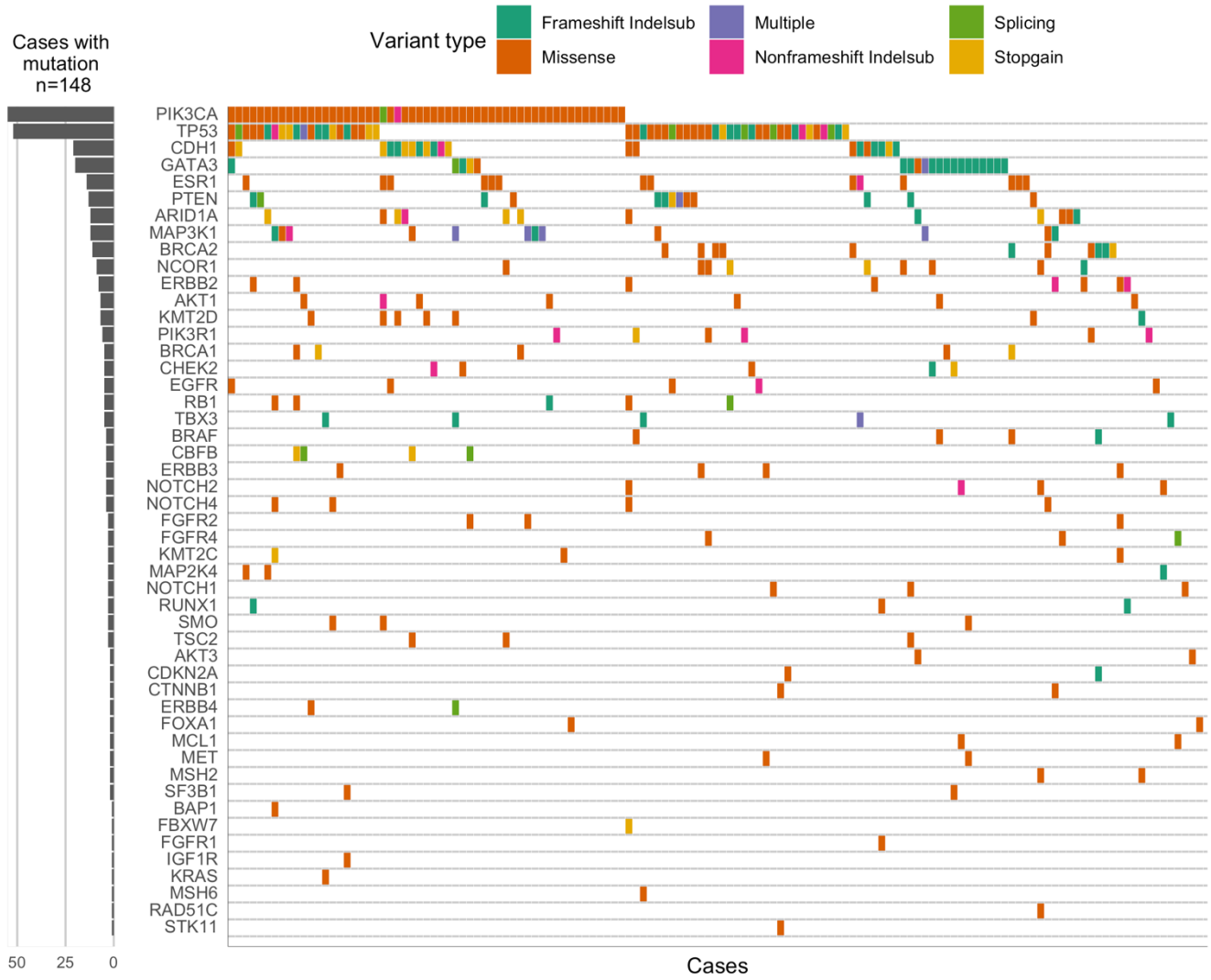


Figure 5.6. Mutation prevalence per case for A) HER2+ and B) TNBCs.

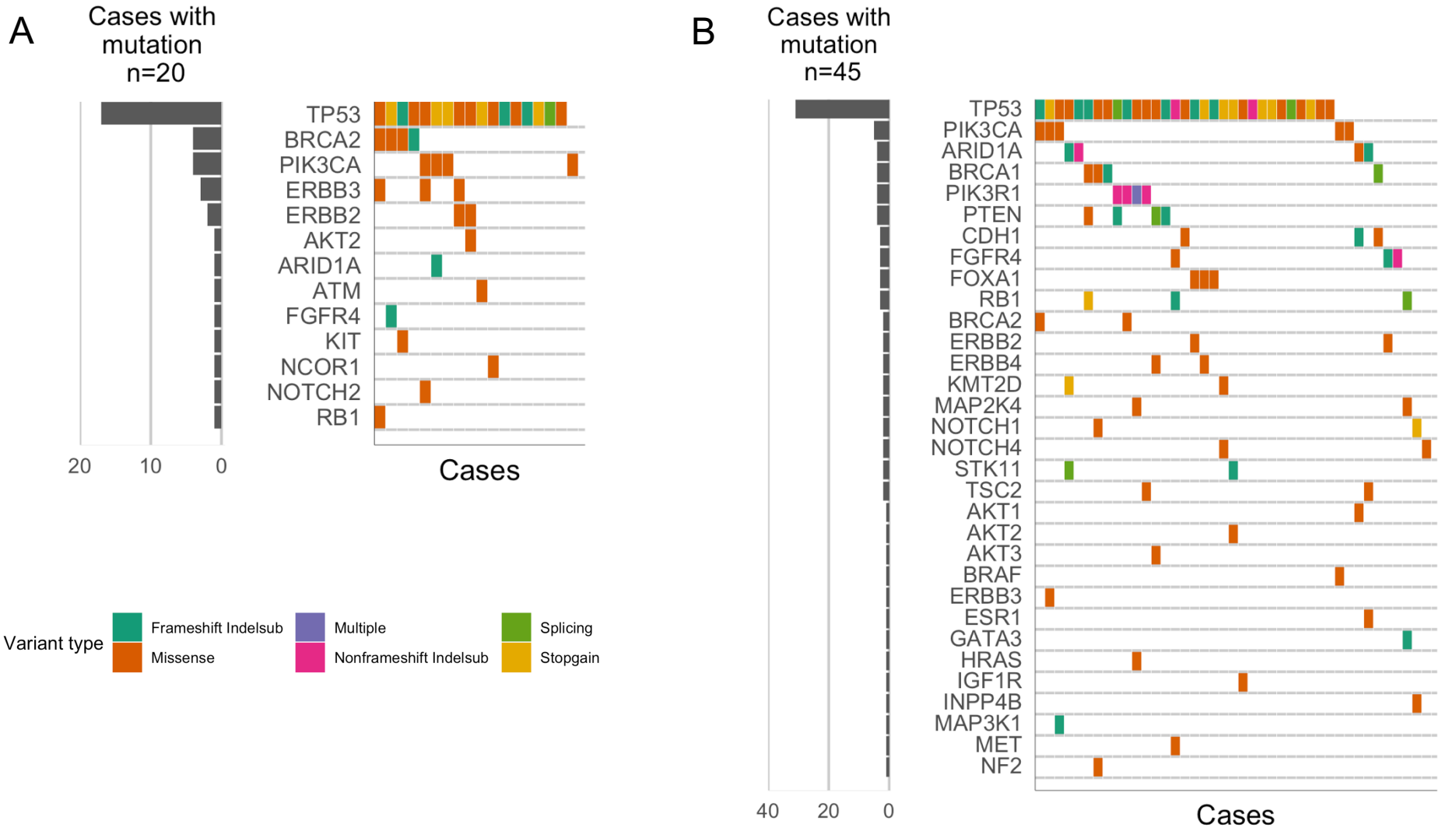


Figure 5.7. Copy number alteration prevalence per case for ER+ cancers.

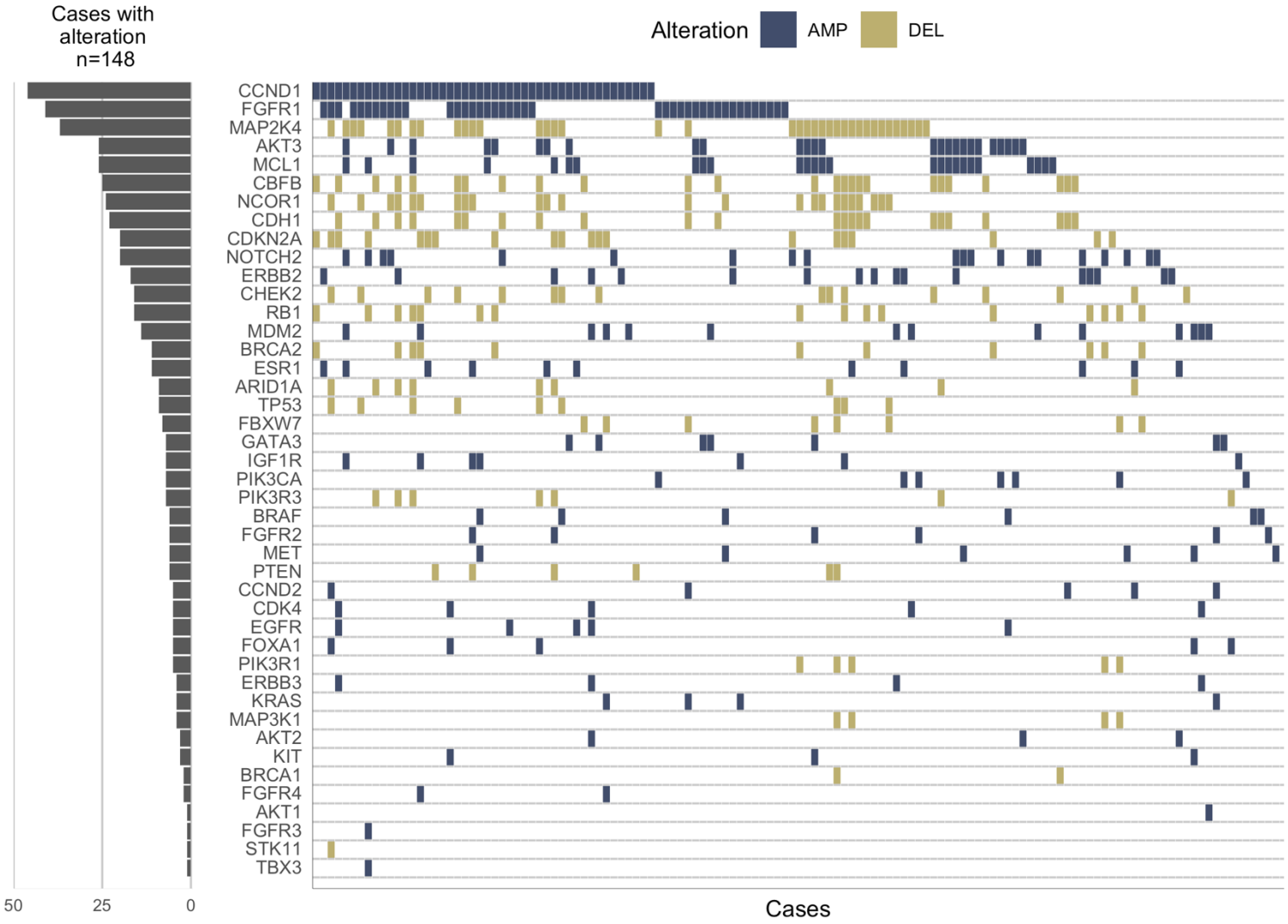


Figure 5.8. Copy number alteration prevalence per case for A) HER2+ and B) triple negative breast cancers.

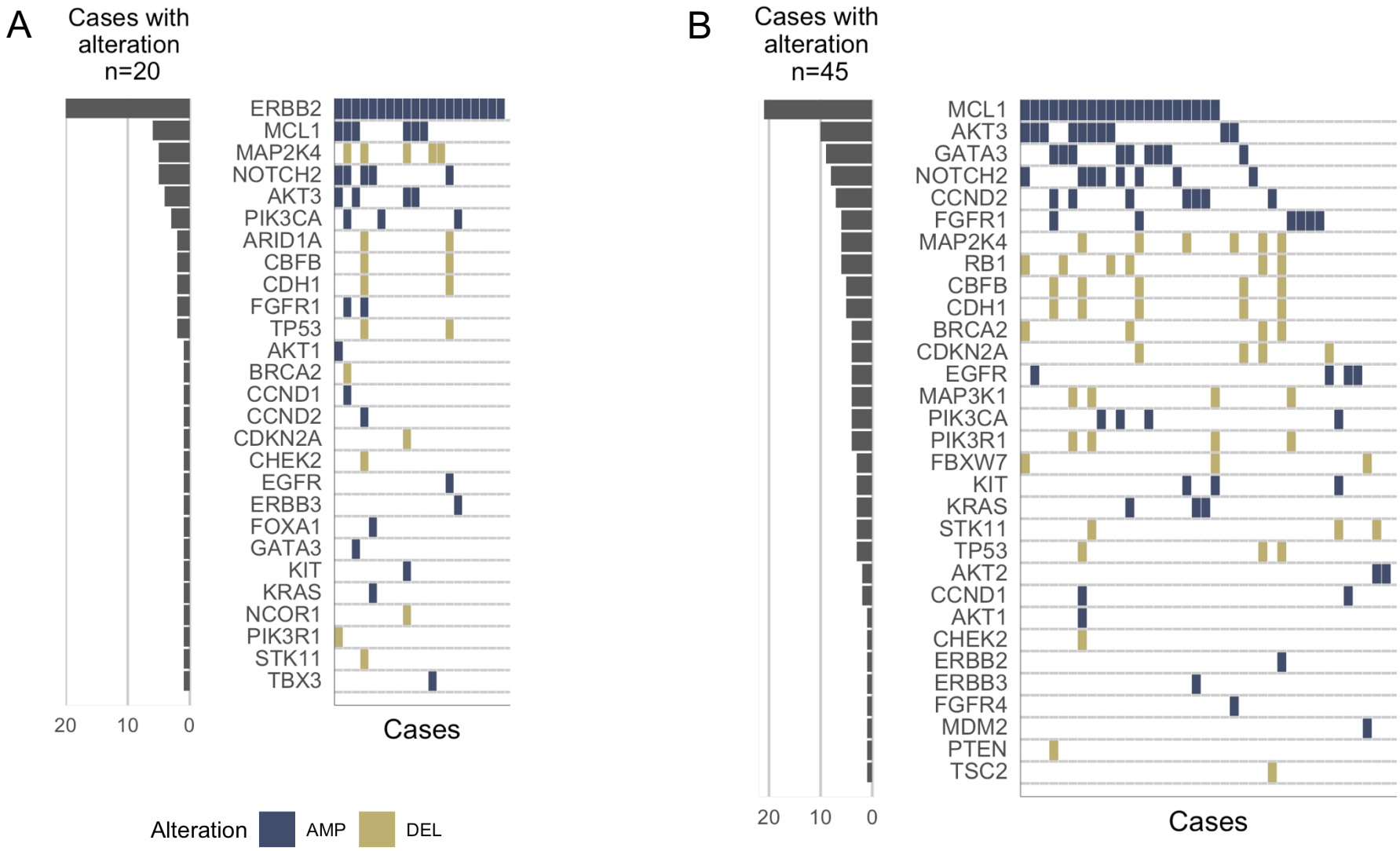


Figure 5.9. Mutation frequency in primary and metastatic samples in the whole cohort. Differences between primary and metastases were not significant (Fisher's Exact Test).

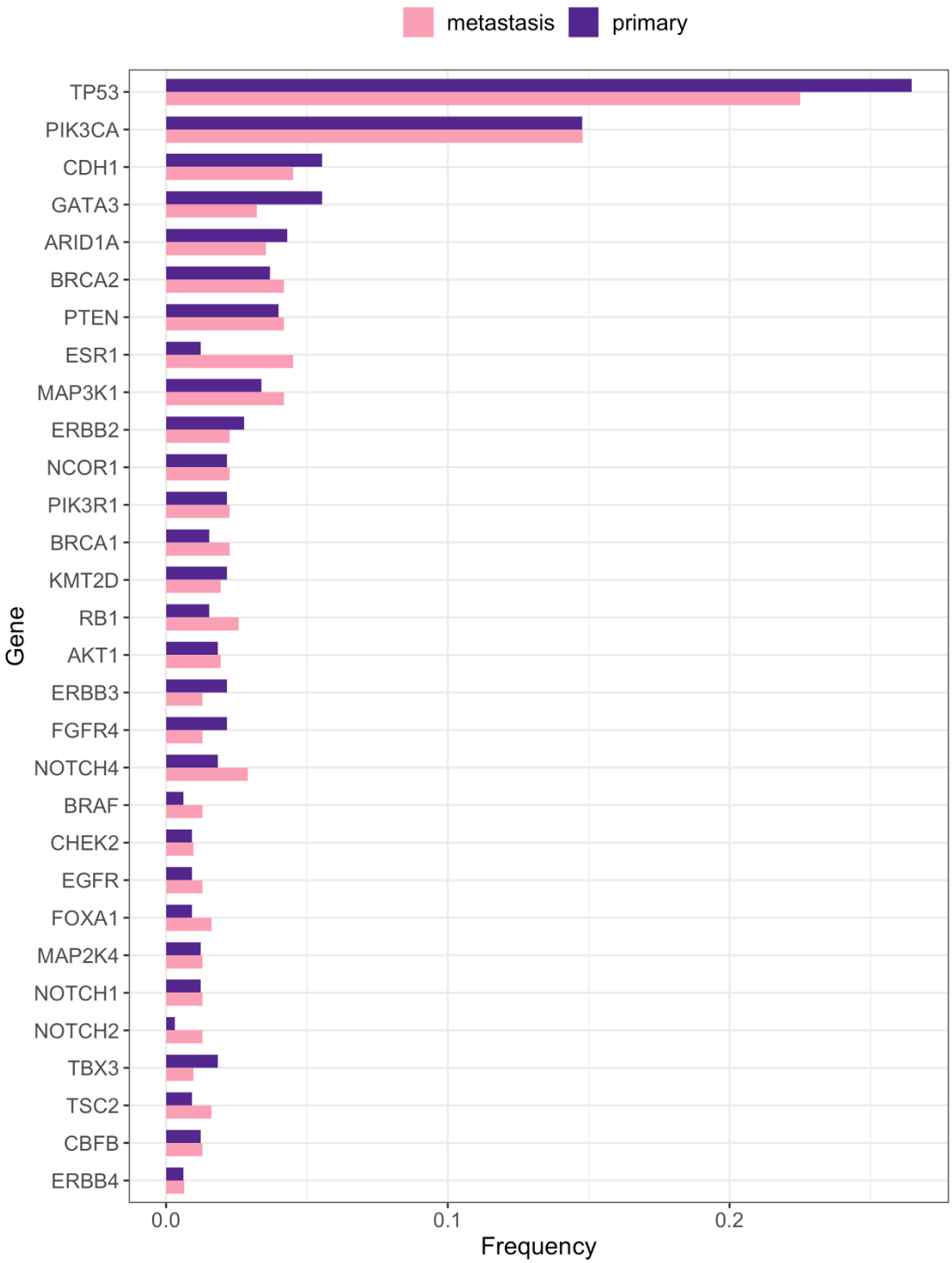


Figure 5.10. Mutation frequency in primary and metastatic samples in ER+ cases. Differences between primary and metastases were not significant (Fisher's Exact Test).

Figure 5.11. Copy number alteration frequency in primary and metastatic samples in the whole cohort. Differences with an adjusted p-value < 0.05 are marked with an asterisk (Fisher's Exact Test, p-values adjusted with the method of Benjamini-Hochberg).

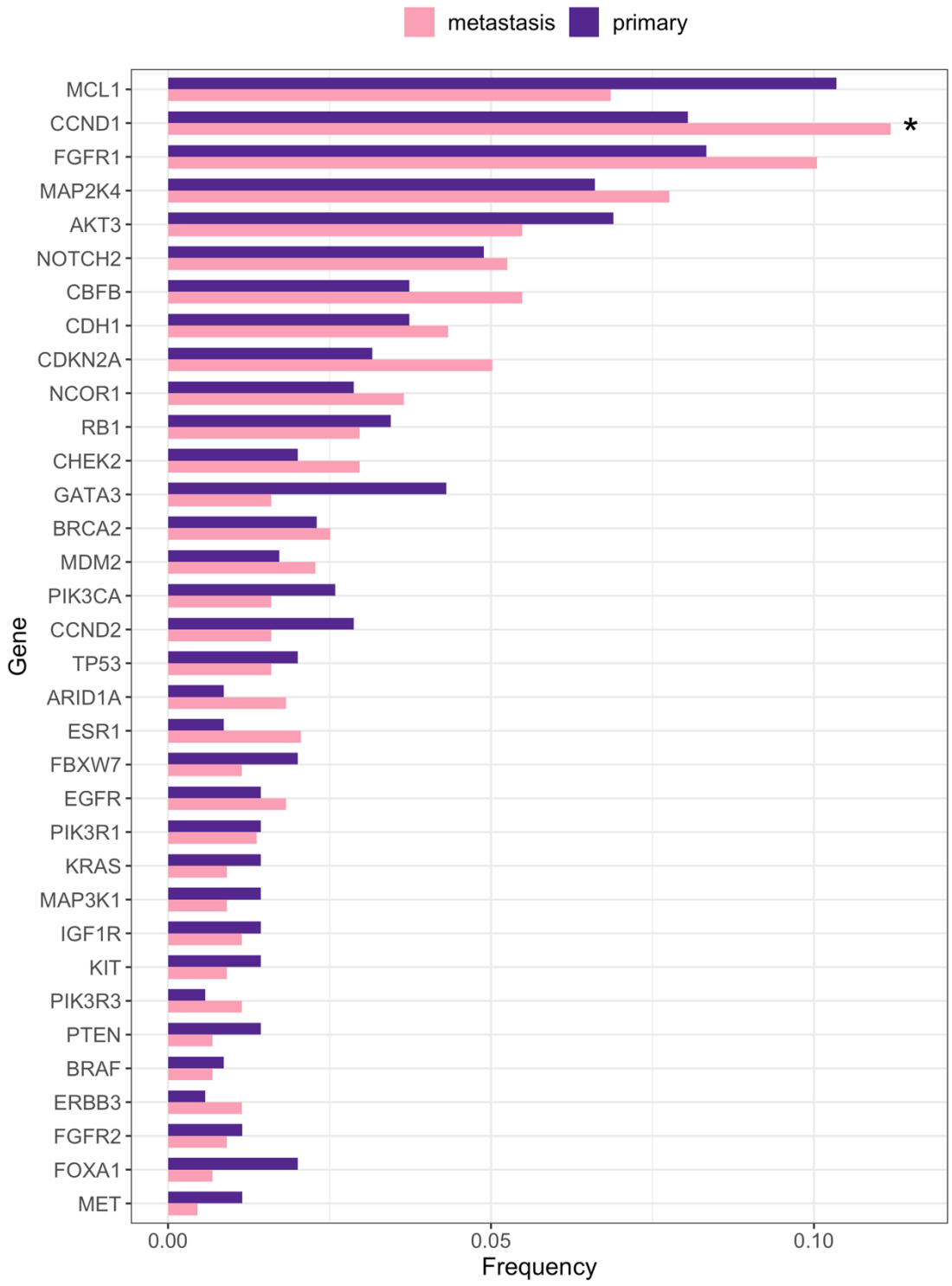


Figure 5.12. Copy number alteration frequency in primary and metastatic samples in ER+ cases. Differences between primary and metastases were not significant (Fisher's Exact Test).

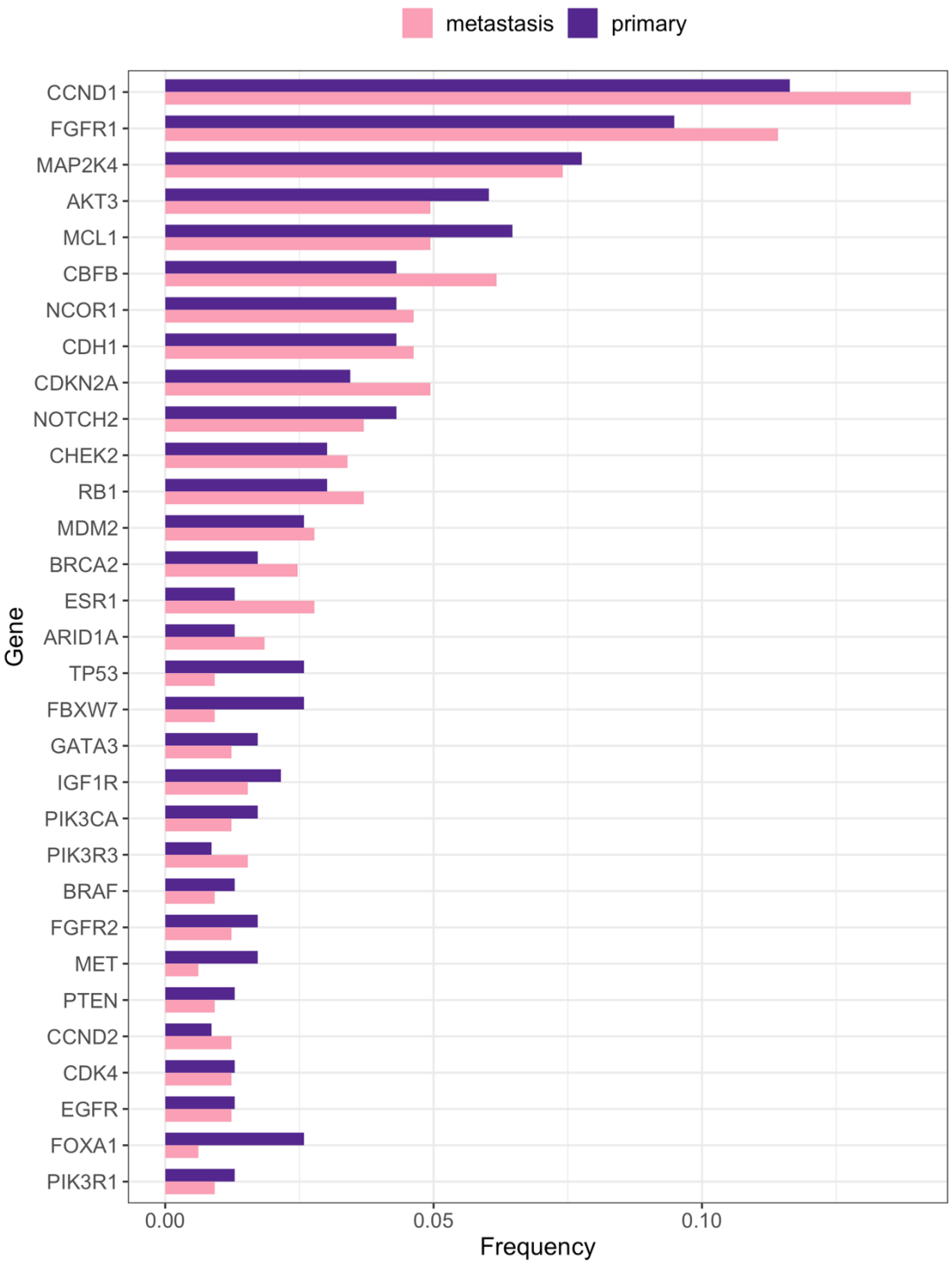
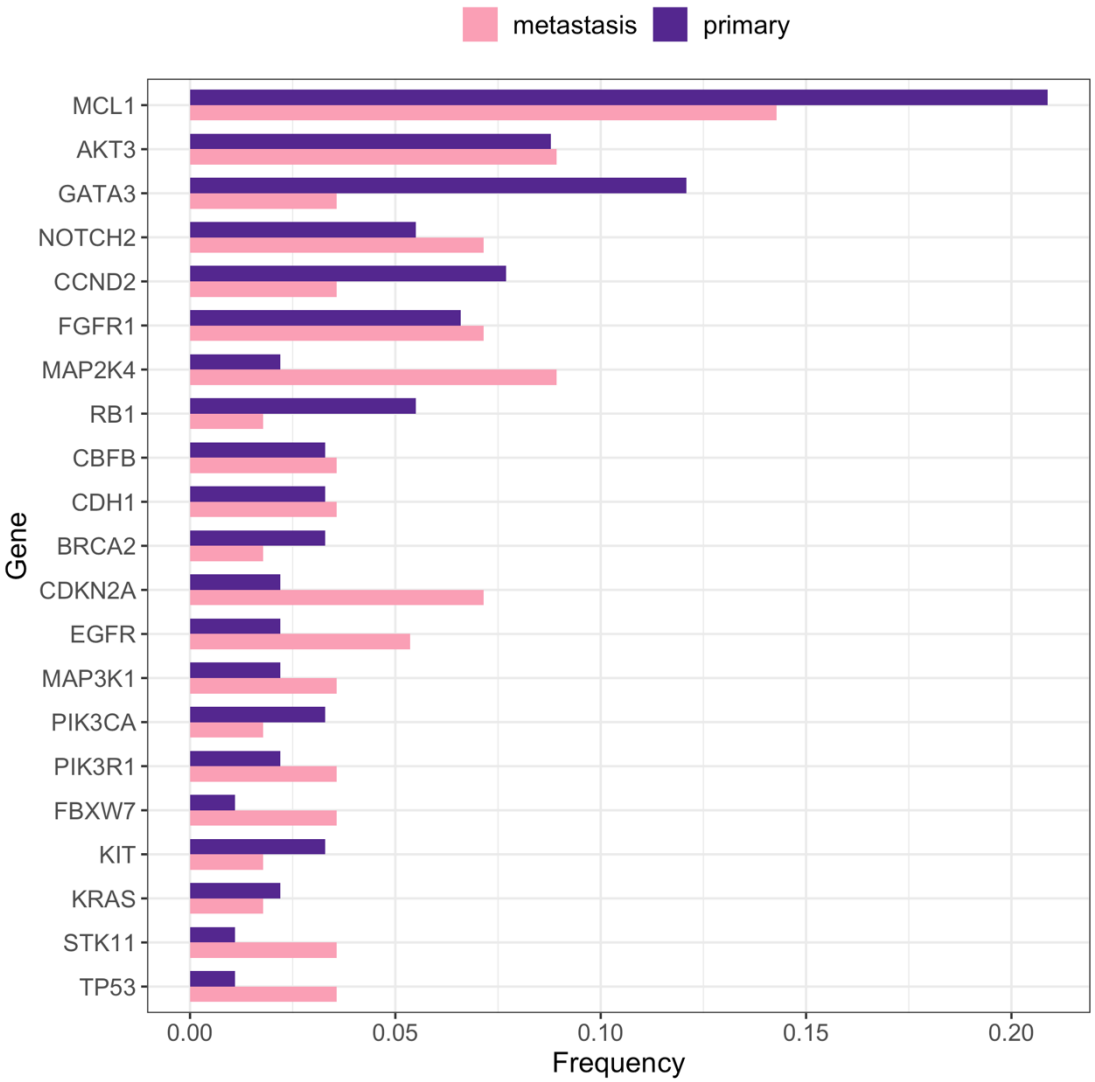


Figure 5.13. Copy number alteration frequency in primary and metastatic samples in TNBC cases. Differences between primary and metastases were not significant (Fisher's Exact Test).



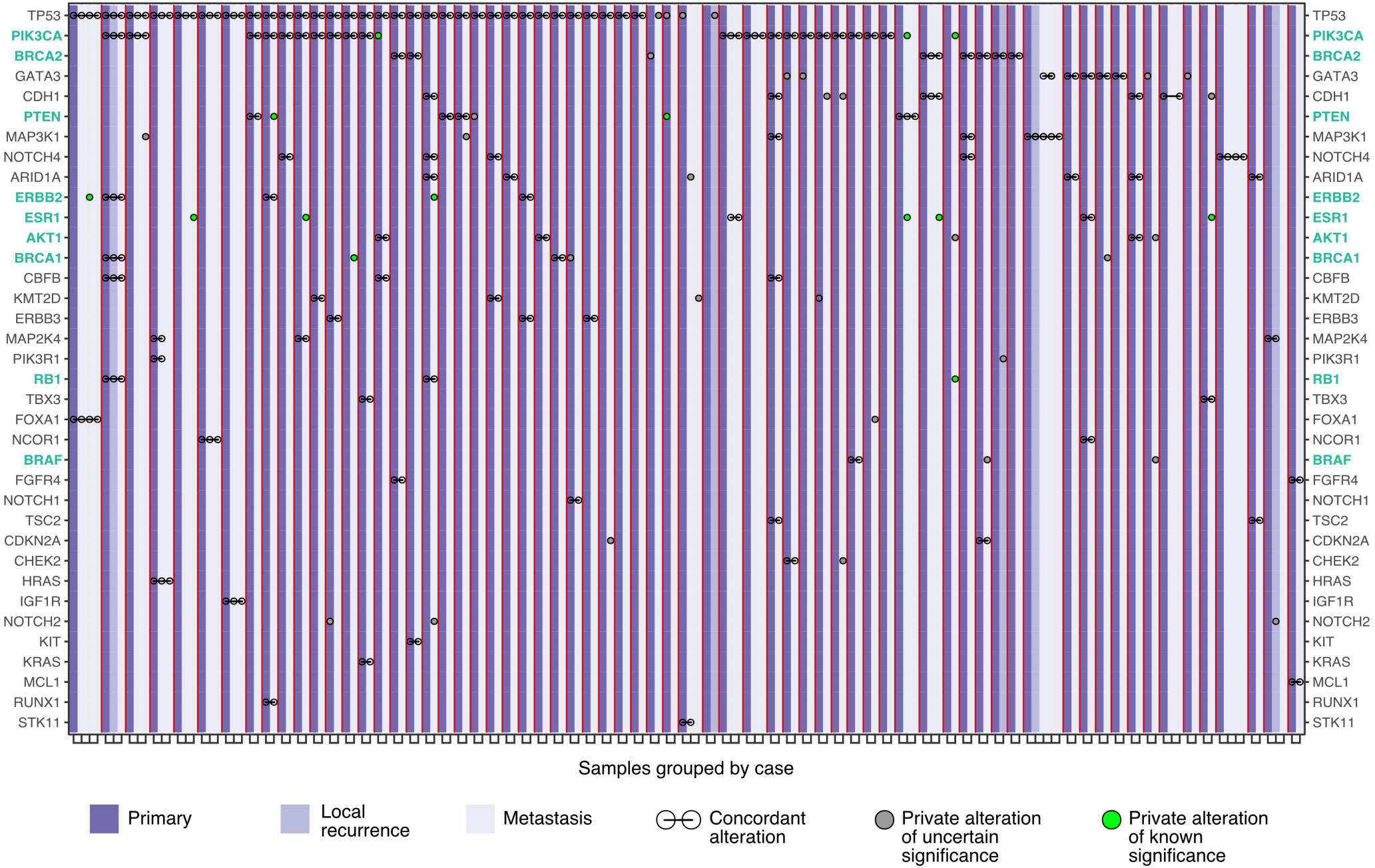

TP53
PIK3CA
BRCA2
GATA3
CDH1
PTEN
MAP3K1
NOTCH4
ARID1A
ERBB2
ESR1
AKT1
BRCA1
CBFB
KMT2D
ERBB3
MAP2K4
PIK3R1
RB1
TBX3
FOXA1
NCOR1
BRAF
FGFR4
NOTCH1
TSC2
CDKN2A
CHEK2
HRAS
IGF1R
NOTCH2
KIT
KRAS
MCL1
RUNX1
STK11
Samples grouped by case
Primary
Local recurrence
Metastasis
Concordant alteration
Private alteration of uncertain significance
Private alteration of known significance

Figure 5.14. Concordance map for mutations. Genes are in rows and each column represents one sample. Samples are grouped by cases, indicated by braces on x-axis and dividing vertical red lines. Samples are coloured according to whether they are primary tumours (dark purple), local recurrences (lighter purple) or a metastasis (lightest purple). Alterations are marked with hollow circles for concordant alterations or filled circles for private alterations. Private alterations with known significance in actionable genes are highlighted in green. Private alterations with unknown significance in actionable genes or private alterations in non-actionable genes are grey. Concordant alterations are also joined by horizontal lines across samples. Actionable genes are highlighted in teal.

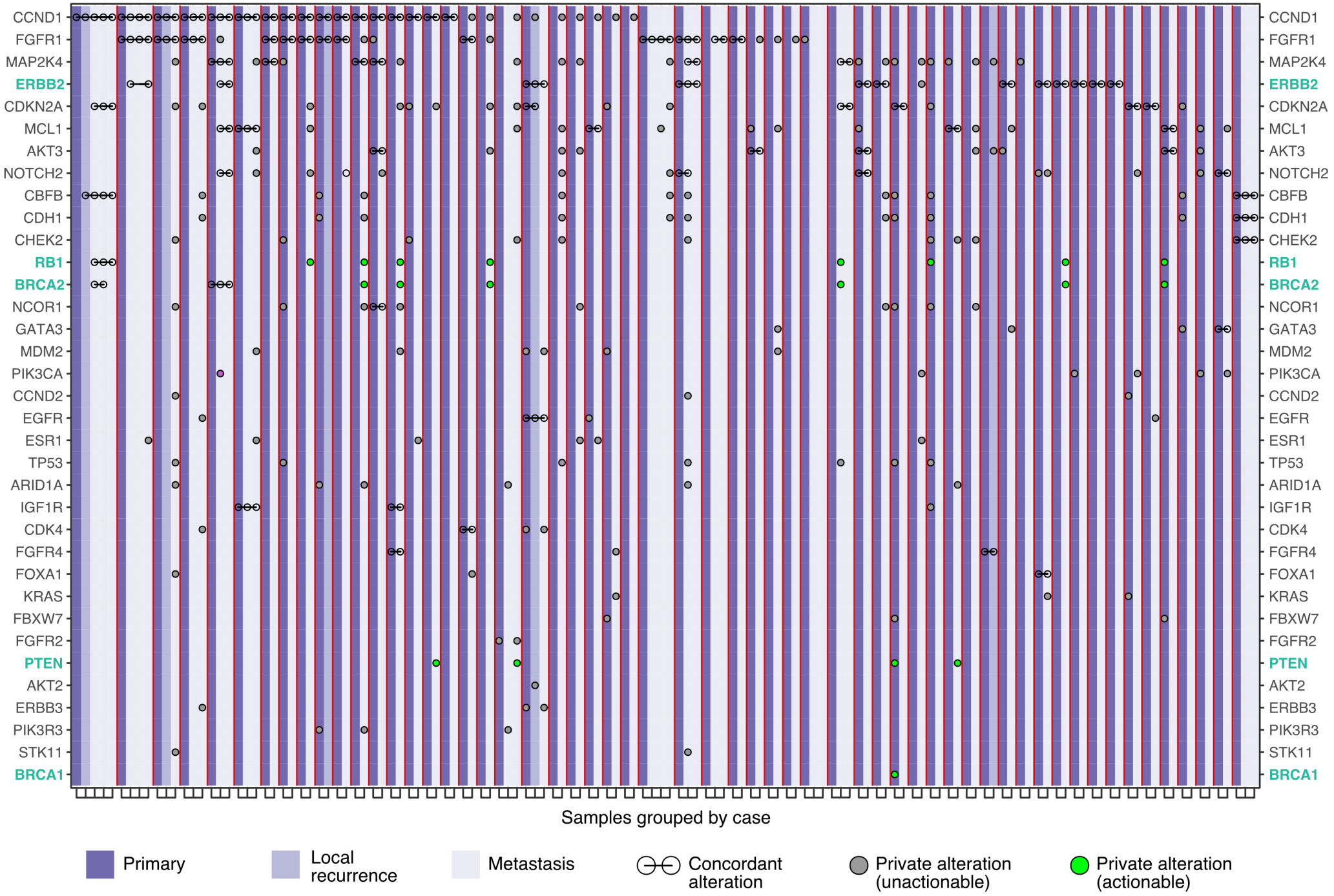

CCND1
FGFR1
MAP2K4
ERBB2
CDKN2A
MCL1
AKT3
NOTCH2
CBFB
CDH1
CHEK2
RB1
BRCA2
NCOR1
GATA3
MDM2
PIK3CA
CCND2
EGFR
ESR1
TP53
ARID1A
IGF1R
CDK4
FGFR4
FOXA1
KRAS
FBXW7
FGFR2
PTEN
AKT2
ERBB3
PIK3R3
STK11
BRCA1
Samples grouped by case
Primary
Local recurrence
Metastasis
Concordant alteration
Private alteration (unactionable)
Private alteration (actionable)

Figure 5.15. Concordance map for copy number alterations. Genes are in rows and each column represents one sample. Samples are grouped by cases, indicated by braces on x-axis and dividing vertical red lines. Samples are coloured according to whether they are primary tumours (dark purple), local recurrences (lighter purple) or a metastasis (lightest purple). Alterations are marked with hollow circles for concordant alterations or filled circles for private alterations. Private alterations with known significance in actionable genes are highlighted in green. Private alterations with unknown significance in actionable genes or private alterations in non-actionable genes are grey. Concordant alterations are also joined by horizontal lines across samples. Actionable genes are highlighted in teal.

Chapter 6: Rapid Autopsy Case Series

TITLE

The subclonal architecture of metastatic breast cancer: Results from a prospective community-based rapid autopsy program "CASCADE".

AUTHORS

Peter Savas[1], Zhi Ling Teo[1], Christophe Lefevre[4,5], Christoffer Flensburg[4,5], Franco Caramia[1], Kathryn Alsop[7], Mariam Mansour[1], Prudence A Francis[2], Heather A Thorne[6,7], Maria Joao Silva[1], Nnennaya Kanu[8], Michelle Dietzen[8], Andrew Rowan[9], Maik Kschischo[10], Stephen Fox[11], David D Bowtell[3,7], Sarah-Jane Dawson[2,3], Terence P Speed[12,13], Charles Swanton[8,9], Sherene Loi[2,3*]

AFFILIATIONS

[1]Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia.

[2]Division of Cancer Medicine, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia.

[3]Sir Peter MacCallum Department of Oncology, the University of Melbourne, Victoria 3010, Australia

[4]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia.

[5]Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Victoria 3010, Australia

[6]kConFab, Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia.

[7]Cancer Genomics Program, Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia.

[8]UCL Cancer Institute, CRUK Lung Cancer Centre of Excellence, Paul O'Gorman Building, Huntley St., London, UK.

[9]The Francis Crick Institute, Midland Road, London, UK

[10]University of Applied Sciences Koblenz, RheinAhrCampus Remagen, Department of Mathematics and Technology, Joseph-Rovan-Allee 2, D-53424 Remagen

[11]Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[12]Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia

[13]Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010.

*sherene.loi@petermac.org

ABSTRACT

Background

Understanding the cancer genome is seen as a key step in improving outcomes for cancer patients. Genomic assays are emerging as a possible avenue to personalised medicine in breast cancer. The majority of work in this area has targeted primary tumors however, and very few studies have performed comprehensive profiling of advanced disease. Evolution of the cancer genome during the natural history of breast cancer is largely unknown, as is the profile of disease at death. We sought to study in detail these aspects of advanced breast cancers that have resulted in lethal disease.

Methods and Findings

Three patients with ER-positive, HER2-negative breast cancer and one patient with triple negative breast cancer underwent rapid autopsy as part of an institutional prospective community-based rapid autopsy program. Cases represented a range of management problems in breast cancer, including late relapse after early stage disease; de novo metastatic disease; discordant disease response and disease refractory to treatment. Between 5 and 12 metastatic sites were collected at autopsy together with available primary tumors and longitudinal metastatic biopsies taken during life. Samples underwent paired tumor-normal whole exome sequencing and single nucleotide polymorphism arrays. Subclonal architectures were inferred by jointly analysing all samples from each patient. Mutations were validated using high depth amplicon sequencing.

Between cases, there were significant differences in mutational burden, driver mutations, mutational processes and copy number variation. Within each case, we found dramatic heterogeneity in subclonal structure from primary to metastatic disease and between metastatic sites, such that no single lesion captured the breadth of disease. Metastatic cross seeding was found in each case and treatment drove subclonal diversification. Subclones displayed parallel evolution of treatment resistance in some cases, and apparent augmentation of key oncogenic drivers as an alternative resistance mechanism. We also observed the key role of mutational processes in subclonal evolution.

Limitations of this study include the potential for bias introduced by joint analysis of formalin fixed archival specimens with fresh specimens, and the difficulties in resolving subclones with whole exome sequencing. Other alterations that could define subclones such as structural variants or epigenetic modifications were not assessed.

Conclusions

This study highlights the variety of mechanisms that shape the genome of metastatic breast cancer, and the value of studying advanced disease in detail. Treatment drives significant genomic heterogeneity in breast cancers which has implications for disease monitoring and treatment selection in the personalised medicine paradigm.

Author Summary

Why Was This Study Done?

- We understand very little about the genetic changes that take place in advanced breast cancers, particularly in the most advanced disease that results in death.
- Seeking this information is expected to provide important insights into cancer biology, and help us understand how advanced breast cancer evolves over time and the impact this has on the natural history of the disease.

What Did the Researchers Do and Find?

- We conducted rapid autopsies on four patients that died of advanced breast cancer, and extensively profiled multiple lesions which were compared to biopsies taken life where possible.
- We aimed to reconstruct the relationship between different metastatic lesions and understanding the tumor cell subpopulations that make up metastatic lesions.
- Our findings revealed significant heterogeneity and evolution over time, which was shape by treatment, mutational processes and the interaction of these factors with alterations in cancer promoting driver genes.

What Do These Findings Mean?

- Our findings reveal the complexity of the cancer genome in advanced disease, and highlight the importance of ongoing monitoring and

reassessment of the genetic profile of the disease if this is to be used to guide therapy.

- We also show the value of autopsy studies in providing high yield by studying small number of cases in detail.
- Detecting small populations of cancer cells from actual patient specimens is difficult, and the findings in this study could be limited by poor sensitivity to detect rare populations

INTRODUCTION

Heterogeneity in the natural history of advanced cancers has long been noted. The advent of cancer genomics has revealed that significant heterogeneity may exist both between and within lesions in the same patient. Since that time, autopsy studies have found a great variety of genomic heterogeneity in multiple cancer types. Evolution over time has also been documented, with varying influences of therapy in shaping the subclonal architecture of advanced disease. At the same time, the clinical significance of heterogeneity is yet to be established.

Patients die of metastatic disease but little is known about the biology of this late stage, lethal process. Genomics is a mature and robust platform for querying this biology. In breast cancer, although many thousands of primary tumors have been characterised in detail, such comprehensive data does not exist for metastatic disease, particularly in the most advanced and lethal disease. Shah et al studied genomic heterogeneity in a case of lobular breast cancer which recurred 9 years after initial diagnosis. This single case showed genomic evolution over time from primary to metastatic disease [1]. Ding et al analysed a chemo-resistant metastatic basal-like breast cancer, and showed low divergence between primary and metastatic lesions, with 48 of 50 mutations in common between sites [2].

More recently, the focus has shifted to understanding heterogeneity at the subclonal level. Complex but reproducible subclonal dynamics were revealed by Eirew *et al*. in a large study of treatment naïve breast cancer xenografts analysed at the single cell level [3]. Murtaza *et al*. combined multi-region sequencing at autopsy with circulating DNA (ctDNA) measurements, finding that ctDNA

captured some of the underlying subclonal dynamics [4]. Juric *et al.* used sequential biopsies and samples from autopsy to study the evolution of resistance to a PI3K inhibitor in a case of hormone positive breast cancer, finding evidence for multiple resistance mechanisms evolving simultaneously in spatially distinct sites (termed convergent evolution) [5]. Yates *et al.* performed a combination of whole genome and targeted sequencing on 50 breast cancer cases and inferred subclonal populations, finding variable intra-lesional heterogeneity [6]. Low prevalence or 'minor' subclones in primary tumors could be seen giving rise to recurrent disease, and convergent evolution where different subclones evolve similar mechanisms of resistance occurred both late and early in the disease course. The profound influence of treatment on the subclonal structure of breast cancer has also been delineated in HER2 positive tumors using a method that visualises driver mutation heterogeneity in different tumor subpopulations [7]. This study also found an association between heterogeneity and poor clinical outcome. Determining the genomic changes that define subclonal populations allows the evolutionary history of a cancer to be inferred, as was performed by Gao et al in a single cell whole genome sequencing study of triple negative breast cancers [8]. Well demarcated subclonal populations could be defined by copy number alterations, without the presence of populations showing intermediate copy number states. The presence of subclonal populations well demarcated by copy number alterations and without clones showing intermediate copy number states is most consistent with punctuated evolution of neoplastic phenotypes early in the natural history of a tumor, rather than gradual accumulation of genomic alterations. These studies paint an expanding picture of the challenges and opportunities in understanding

the complexities of advanced disease. Many questions remain about the relationship between breast cancer's capacity to evolve towards heterogeneous states and clinical outcome.

To understand the biology of lethal metastatic disease, our institution has implemented a prospective community-based rapid autopsy program (CASCADE) where patients consent whilst alive to allow tissue donation after death [9]. In this study, we attempted to understand the evolutionary history of four cases with lethal breast cancer who participated in the CASCADE program: an aggressive and treatment resistant triple negative breast cancer where the patient died 12 months from diagnosis (TN1); an ER-positive HER2-negative breast cancer with late relapse 7 years from diagnosis of early-stage disease (ER1); a *de novo* metastatic ER positive, HER2-negative breast cancer (ER2); an ER-positive HER2 negative breast cancer in a young patient with multiple instances of discordant responses to treatment between metastatic sites (ER3). By interrogating the cancer genome at multiple metastatic sites, including metastatic biopsies taken during life, we aimed to infer the subclonal structure and evolutionary history of the disease for each case as it progressed from the primary tumor to lethal metastatic dissemination.

METHODS

As breast cancer is relatively treatment responsive compared to other tumor types, we focussed our analysis on understanding subclonal composition and how this evolves over time under the influence of therapy. The first case was recruited in 2013, and at that time, methods for performing subclonal inference on multiple samples were not mature. By the time the final case was recruited in 2015, the field had advanced considerably, which made this approach feasible as will be detailed below. All four cases were analysed concurrently.

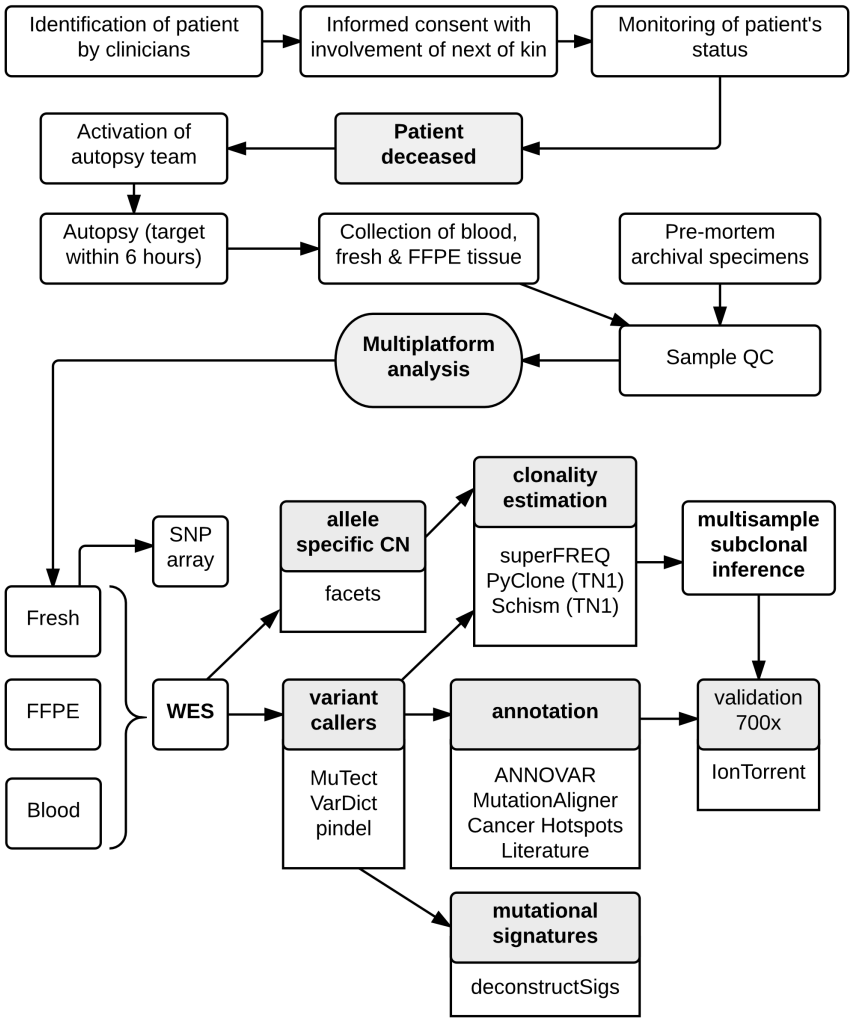
Patients were recruited by their treating clinicians to CASCADE, a prospective community-based rapid autopsy program [9], conducted in association with the Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab) and approved by the Human Research Ethics Committee of the Peter MacCallum Cancer Centre, Melbourne (HREC approval numbers: CASCADE 13/122, kConFab 92/97 and 11/102). Any patient with advanced breast cancer was eligible. All patients provided written informed consent.

Study procedures

The workflow for the CASCADE program is shown in Fig 1. Once recruited, the CASCADE coordinator communicates with the patient, their family, and the associated healthcare providers regarding the patient's status and ongoing willingness to participate. Following the patient's death, the family specifies when the body can be transferred for autopsy. Autopsies were conducted by a specialist forensic pathologist assisted by the research team. Tumor tissue was harvested from multiple representative metastatic sites. Each lesion was divided into portions that were immediately snap frozen in liquid nitrogen and also fixed

in formalin for subsequent paraffin embedding (FFPE). Frozen samples were used for sequencing where possible. Prospectively collected fresh frozen samples were available for some cases. For all fresh tissues, frozen sections were reviewed by a Pathologist to confirm the presence of tumor, quantify necrosis and estimate tumor cellularity. Pre-morbid FFPE archival samples from primary tumors and metastatic lesions were also obtained.

Fig 6.1. Workflow diagram giving overview of methodology.



Whole exome sequencing

DNA was extracted from sections cut from frozen tissues or FFPE blocks. For 3 of the cases, whole exome libraries were prepared using the Roche-NimbleGen

SeqCap EZ exome version 3. For case 1, libraries were prepared using the Illumina Nextera Rapid Exome. All libraries were sequenced on the Illumina Hiseq platform. Case 1 had DNA submitted for single nucleotide polymorphism (SNP) array analysis with the Illumina Human Omni 2.5S beadchip, and other cases on the Affymetrix Genome Wide SNP 6.0 array. Blood was available for all cases for reference germline DNA.

After adapter trimming with cutadapt [10], raw sequence data was aligned to the human reference genome GRCh37 with BWA [11]. Two samples for case TN1 exhibited low sequencing yield and a large number of likely artefactual mutations due to FFPE processing, and were excluded from further analysis. Fresh frozen samples were sequenced to a mean depth of 105, FFPE sample to a mean depth of 75. Following alignment, BAM files were processed according to GATK best practices [12]. Variants were also called using MuTect [13] with default parameters. Indels were called with VarDict [14] and pindel [15]. Called variants were annotated with ANNOVAR [16]. Variants were filtered as follows: exclusion of variants in 1000 Genomes and he Exome Sequencing Project [17,18]; variants with an allele frequency ≥ 2% or with 2 or more supporting reads in the normal; allele frequency < 5% in fresh tissues or <10% in FFPE tissues. Indels were manually reviewed and excluded if adjoining > 4 homopolymer runs or repetitive regions.

Copy number was called from whole exome (WES) data using the facets package, which also provided purity and ploidy estimates [19]. For the Illumina bead chips, raw data was tQN normalised [20] and segmented allele specific copy number called using OncoSNP [21]. For the Affymetrix SNP chips, samples were processed using the Aroma package, applying TumorBoost and PSBCS [22,23].

Output of facets was compared to the SNP chip data, and in some cases the parameters adjusted to exclude less likely combinations of copy number and purity. As SNP arrays were not available for the FFPE samples, whole exome copy number was used for all analyses. Copy number segmentation determined from SNP arrays was used as orthogonal validation of WES copy number calls from facets.

To determine the functional significance of novel nonsynonymous variants the following steps were taken: Search COSMIC [24] and TCGA [25] for previous reports (across all tumor types), prioritise variants with a high CADD score [26]; map variants to functional domains of the gene if possible; check if variant falls in a mutational hotspot using MutationAligner [27] and Cancer HotSpots [28]; review literature (using GeneRIF queried via MyGene.info [29]). Annotation of copy number alterations was limited to genes found to be recurrently altered in the literature [30,31].

Subclonal inference

superFREQ is a cancer exome clonality inference tool that takes advantage of multiple samples from the same individual by tracking both single nucleotide variants (SNV) and copy number alterations (CNA) across samples [32]. This allows the detection of highly subclonal somatic mutations that are present at higher cell fraction in other samples. An advantage of superFREQ is that it estimates and propagates statistical and systematic error sources throughout the analysis, thus decreasing the number of false positives and allowing downstream analysis of uncertainties. It also accepts aligned sequence data as input directly. Briefly, the pipeline performs the following steps: 1) GC bias correction, 2)

differential coverage analysis with limma-voom [33], 3) examine variant positions shared between individual samples, flagging variants for base quality, mapping quality, strand bias, stuttering or other artefacts detected in the pool of normal control samples, 4) summarises differential coverage and SNPs for each gene, 5) recursively clusters neighbouring genes with sufficiently similar differential coverage or SNP frequencies until a segmentation of the genome is achieved for each sample, 7) summarises consensus coverage and SNP frequencies for each segment and renormalizes taking accuracy of coverage and SNP frequency into account, 8) calls CNAs and clonality in each segment based on coverage and SNP frequency, 9) calculate clonalities of somatic SNVs using local CNA, 10) clonalities of SNVs and CNAs are tracked over multiple samples to determine if the same or different alleles are gained/lost between samples, 11) mutations (SNVs or CNAs) with similar clonalities in all samples are clustered into clones after, 12) clones are sorted into a tree structure, with smaller clones being assigned as subclones if possible when the sum of the clonalities of disjoint subclones cannot be larger than the clonality of the containing clone. Not all mutations were allocated to subclones by superFREQ, usually due to deranged copy number which does not permit the cancer cell fraction to be reliably determined. Others have noted this problem with subclonal reconstruction [6]. Case TN1 displayed a high degree of copy number variation, including large swathes of loss of heterozygosity (LOH) affecting over 50% of the genome. Accurate clonal inference with superFREQ was not successful. As an alternative, mutation clonality was determined using PyClone [34] with allele specific copy number provided by the R package facets [19]. Mutation genotypes were built using “parental_copy_number” mode. For mutations in regions of LOH, genotypes

consistent with LOH were assigned a prior weight of 10, and other genotypes down weighted to 0.1. All other genotype prior weights were left as 1. From the PyClone output, the median clonality for each mutation (or cancer cell fraction) were clustered via affinity propagation using the Schism package [35]. Mutation clusters were reviewed to separate agglomerated private subclones if needed, and then the reviewed clusters were supplied to Schism's genetic algorithm to construct a multi-sample phylogeny. To validate this approach, high depth validation sequencing of mutations found in more than one sample were subjected to the same process to check the consistency of the tree.

Validation

From each case, mutations and indels were selected for high depth validation based on their contribution to the subclonal phylogeny and biological interest. An Ion AmpliSeq custom panel was designed for the variants of interest in each case and multiplex PCR performed as per the manufacturer's protocol. Libraries were sequenced on the Ion Torrent PGM sequencer (Life Technologies) to a median depth of 770. Wild-type and mutant reads from validation loci were extracted from BAM files using the GenomicAligments R package [36]. Taking loci where there were at least 20 reads, a mutation was considered validated if there was a significant deviation from an expected error rate of 1/200 using the p-binomial test [1].

Mutational signatures

Mutational signatures were ascertained in a two- step process. Somatic mutations were filtered to have allele frequency > 10% in FFPE samples to avoid

detecting formalin fixation artefact, and >5% in fresh samples. To increase detection power and avoid spurious signature detection, all unique mutations from each patient were pooled together and used as input to deconstructSigs [37]. 'default' normalisation was used as per the authors' instructions, with a minimum signature contribution of 0.06. The 30 signatures found in the latest COSMIC classification were used [38,39]. The contribution of the restricted set of signatures found in the pooled mutation set were then examined in each sample. To test the robustness of signature detection, pooled mutations were randomly downsampled to between 10 and 90% of the original mutations and deconstructSigs was re-run on 1000 such downsampled sets.

Circulating DNA and immunohistochemistry

Circulating DNA was assayed as previously described [40]. For immunohistochemistry, heat induced antigen retrieval was performed in 1× citrate buffer (Thermo Scientific). Samples were blocked in 2% bovine serum albumin in tris-buffered saline/Polysorbate 20 solution and endogenous peroxidase inactivated in 1.5% H2O2. Samples were incubated with primary antibodies including, AGF2 (Abcam), CHD4 (Abcam) and Cullin1 (Cell Signalling Technology). Biotinylated species specific secondary antibodies were used at 1:300 (Dako) followed by the avidin-biotin-complex (ABC) method prior to visualisation with 3,3- Diaminobenzidine (DAB) chromogen (Dako). Bright field microscopy was performed on an Olympus BX-51 microscope. Murine inguinal mammary fat pads were used as normal control tissue.

RESULTS

Three ER-positive cases (denoted ER1, ER2, ER3) and one triple negative case (denoted TN1) were analysed. Cases were recruited sequentially and no cases were excluded from the analysis. In all cases, primary tumors were available as formalin-fixed, paraffin-embedded (FFPE) samples. Original pathology reports describing the macro-dissection of the primary tumors were used to select FFPE blocks in spatially distinct regions for each patient. In ER2, ER3 and TN1, metastatic biopsies were also available for analysis, as well as circulating tumor DNA for ER2. To study heterogeneity and evolution in detail, 8 (ER1), 13 (ER2), 16 (ER3) and 15 (TN1) samples were sequenced from each patient. Samples are summarised in S1 Table.

Patient TN1 is a woman diagnosed with a locally advanced triple negative breast cancer at age 39, not associated with a BRCA1 or BRCA2 germline mutation. She had a clinical response to a third-generation standard adjuvant chemotherapy regime and underwent mastectomy which showed significant residual disease. Shortly after she developed a solitary bony metastasis, followed by a liver metastasis and then multiple additional sites of metastatic disease across lung, liver and brain, before dying of liver failure less than 12 months from first diagnosis. She was treated with trastuzumab briefly when the breast primary was found to have focal HER2 positivity, which was not seen again in other lesions. In contrast, patient ER1 relapsed with metastatic disease 7 years after her primary diagnosis at age 42, and underwent a series of endocrine therapies before developing liver metastases and dying of liver failure after surviving 8 years with metastatic disease. Patient ER2 had *de novo* metastatic disease diagnosed at age 35 and survived 3 years receiving multiple lines of

chemotherapy and endocrine therapy. Patient ER3 did not receive surgical or medical treatment after initial diagnosis of early stage disease at age 36, due to personal circumstances. She presented eighteen months later with metastatic disease and displayed discordant responses to therapy during her disease course, which ran 4 years in duration before death. TN1, ER2 and ER3 all had biopsies of metastatic disease collected during their disease course.

Of 52 samples, three samples from TN1 were sequenced but not used in further analysis as they failed quality control (a breast core biopsy and biopsy of a femoral lesion taken pre-mortem, and a subcarinal lymph node from autopsy). One additional sample, a brain metastasis from TN1, failed to sequence. Quality control metrics along with per sample validation rates are shown in S2-S3 Tables. For mutations detected with an allele frequency ≥ 10% from WES, high depth validation rates for FFPE samples ranged from 95 – 100%, and for fresh samples 97 – 100%. When mutations with an allele ≥ 5% were used, 44/45 samples had validation rates of 95% or greater, with one sample from TN1 having a rate of 94%. Importantly, high depth orthogonal validation did not affect the structure of the subclonal tree. Three samples could not undergo validation due to insufficient DNA (1 pre-chemotherapy breast biopsy for ER2, and the 2 primary samples for ER3).

Subclonal architecture defines metastatic spread

Inferring subclones and constructing subclonal phylogenies is a complex problem, requiring high sequencing depth and/or multiple samples to increase sensitivity and reduce the space of possible tree reconstructions [41–43]. By utilising multiple samples from each patient, we were able to construct

consistent subclonal phylogenies using mutation and copy number information (Fig 2 – Fig 5). Private subclones (that is, subclones that are found only in one lesion) are not displayed for clarity as they do not contribute to the core tree structure, therefore all clones displayed were found in at least two spatially distinct sites.

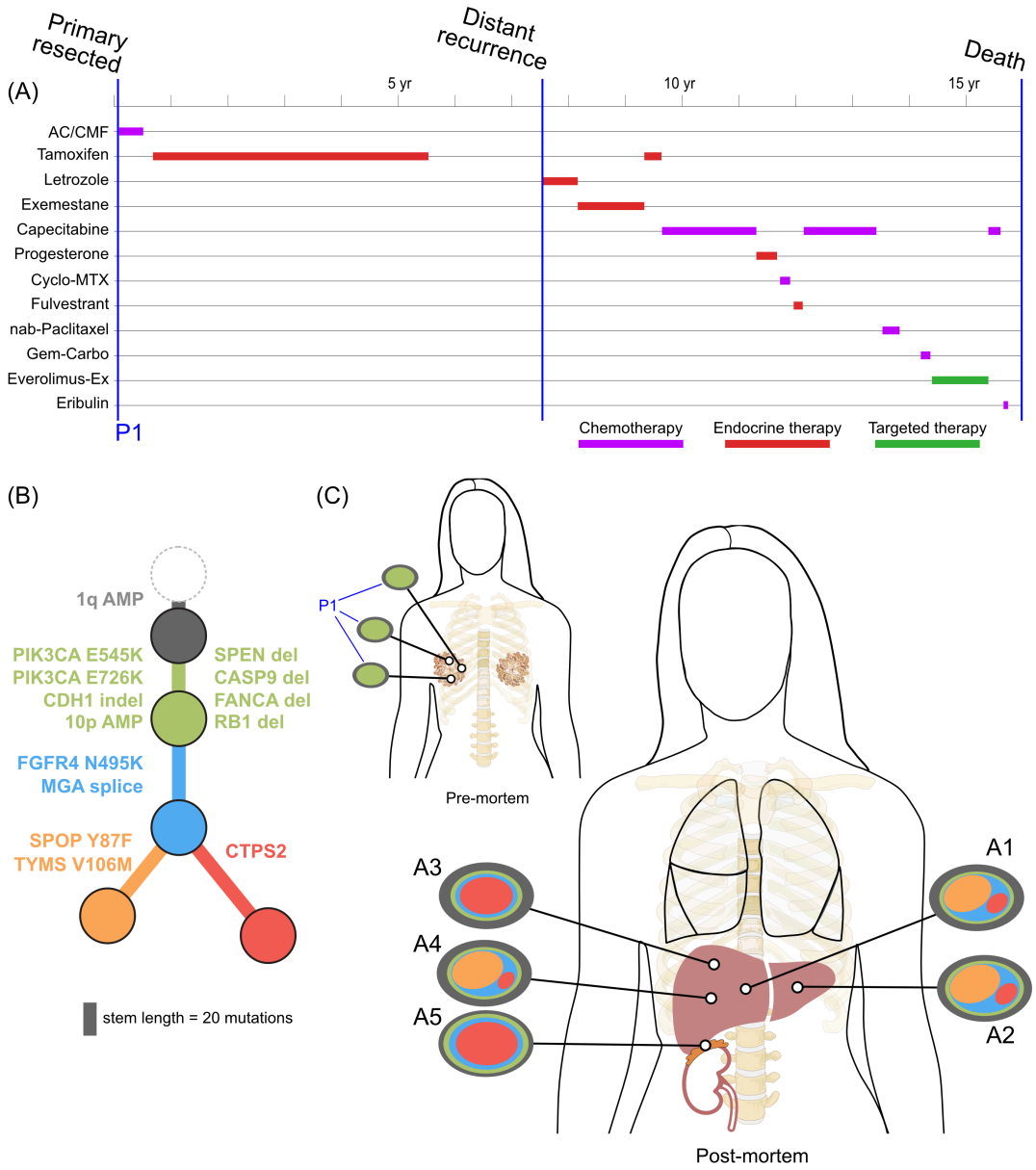


Fig 6.2. Treatment history (A) and subclonal structure (B and C) for ER1, inferred by superFREQ. Pre and post-mortem samples are shown. AC: doxorubicin, cyclophosphamide; CMF: cyclophosphamide, methotrexate, 5-

fluorouracil; Cyclo-MTX: cyclophosphamide, methotrexate; Gem-Carbo: gemcitabine, carboplatin; Everolimus-Ex: everolimus, exemestane. P1: multiple samples from archival breast primary; A1, A3, A4: right liver lobe metastases; A2: left liver lobe metastasis; A5: right adrenal lesion.

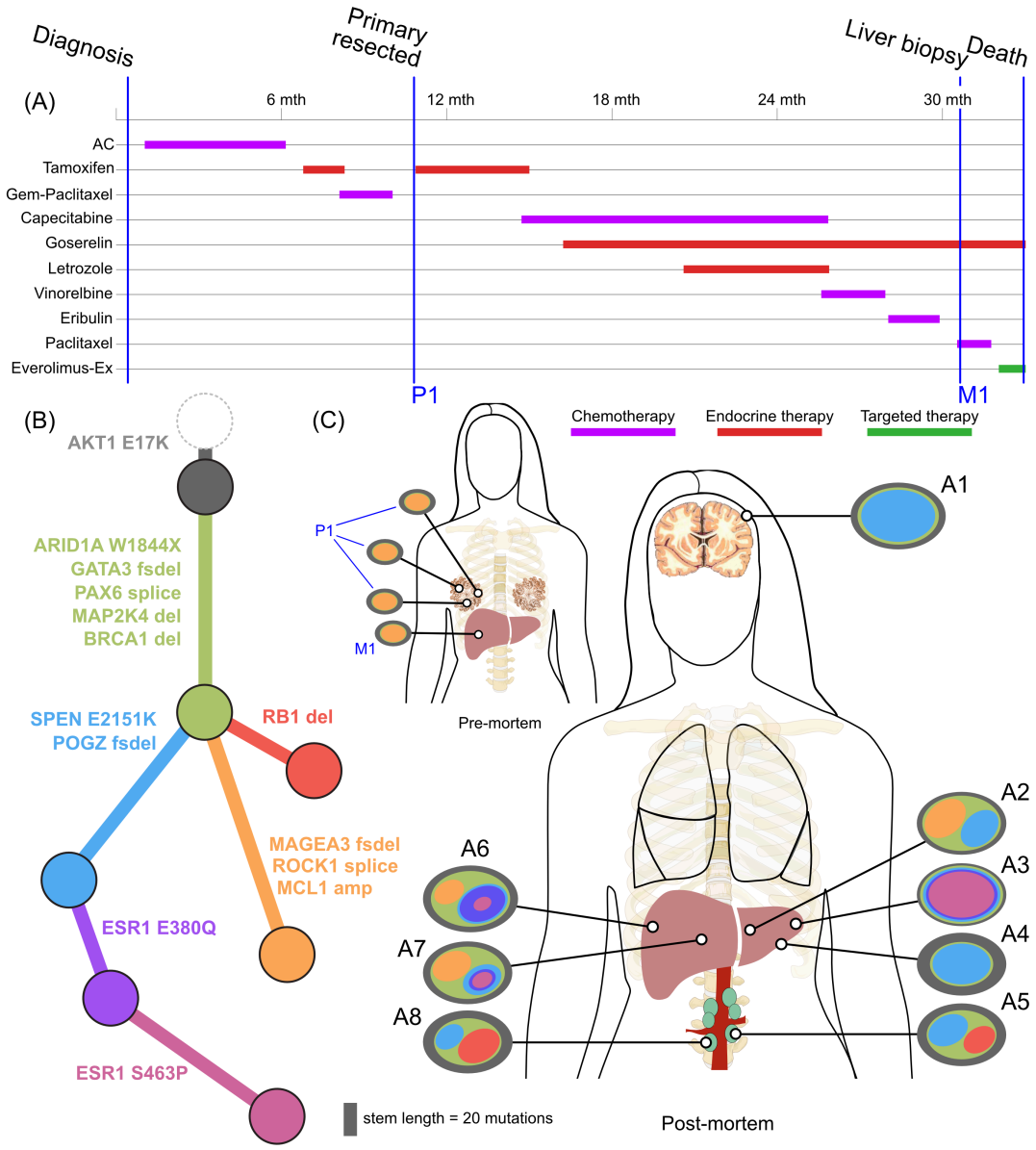


Fig 6.3. Treatment history (A) and subclonal structure (B and C) for ER2, inferred by superFREQ. Pre and post-mortem samples are shown. P1: multiple samples from archival breast primary, including pre-treatment breast core

biopsy; M1: liver core biopsy; A1: dural metastasis; A2, A3, A4: left liver lobe metastases; A5, A8: para-aortic nodal metastases; A6, A7: right liver lobe metastases.

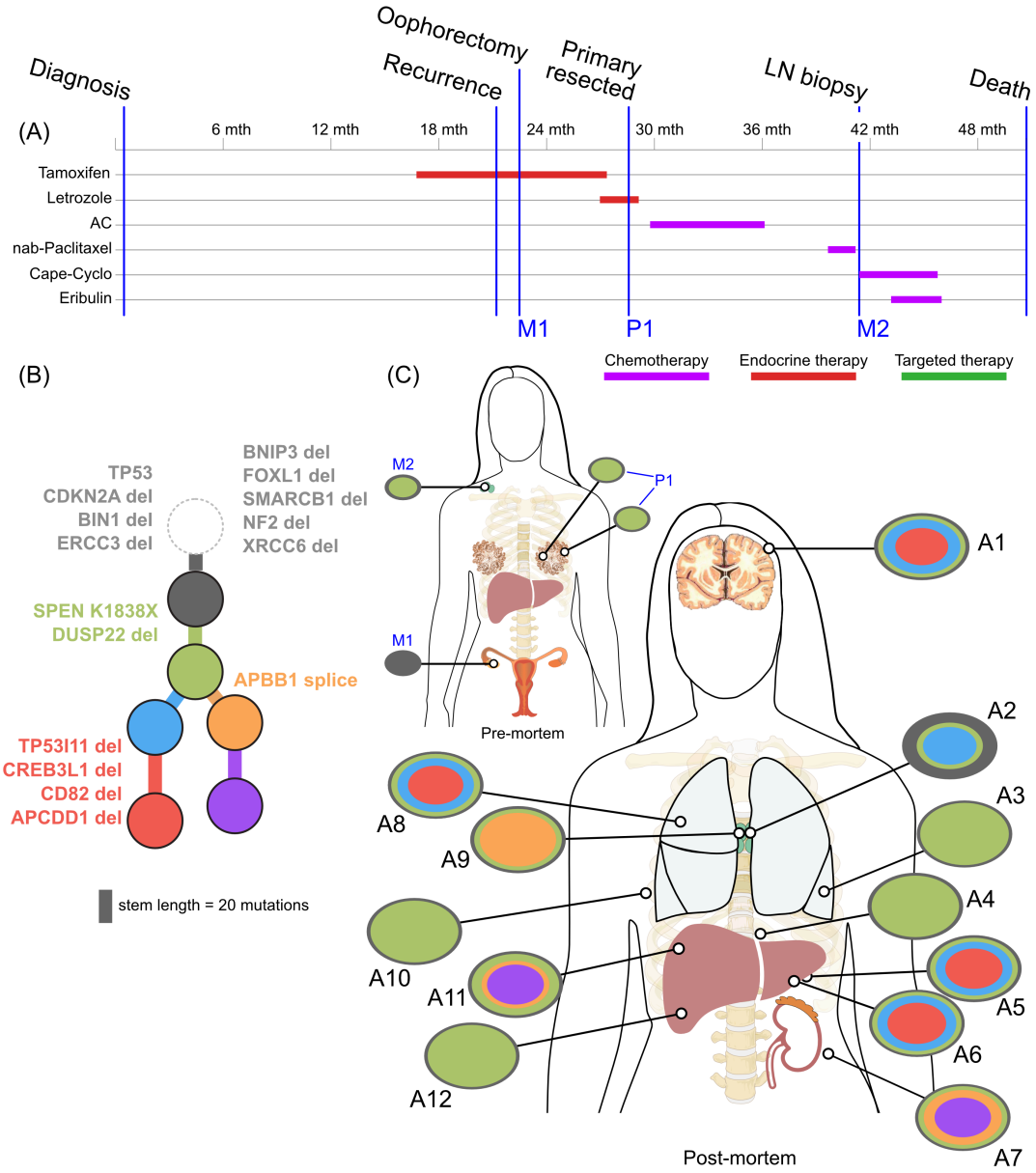


Fig 6.4. Treatment history (A) and subclonal structure (B and C) for ER3, inferred by superFREQ. Pre and post-mortem samples are shown. Cape-Cyclo: capecitabine, cyclophosphamide; P1: multiple samples from archival breast primary; M1: ovarian metastasis; M2: supraclavicular lymph node biopsy; A1:

brain metastasis; A2, A9: left and right hilar lymph nodes; A3: left lung lower lobe metastasis; A4: left parasternal soft-tissue metastasis; A5, A6: anterior and posterior left liver lobe lesions; A7: perinephric metastasis; A8: right upper lobe lesion; A10: rib metastasis; A11, A12: right liver lobe metastases.

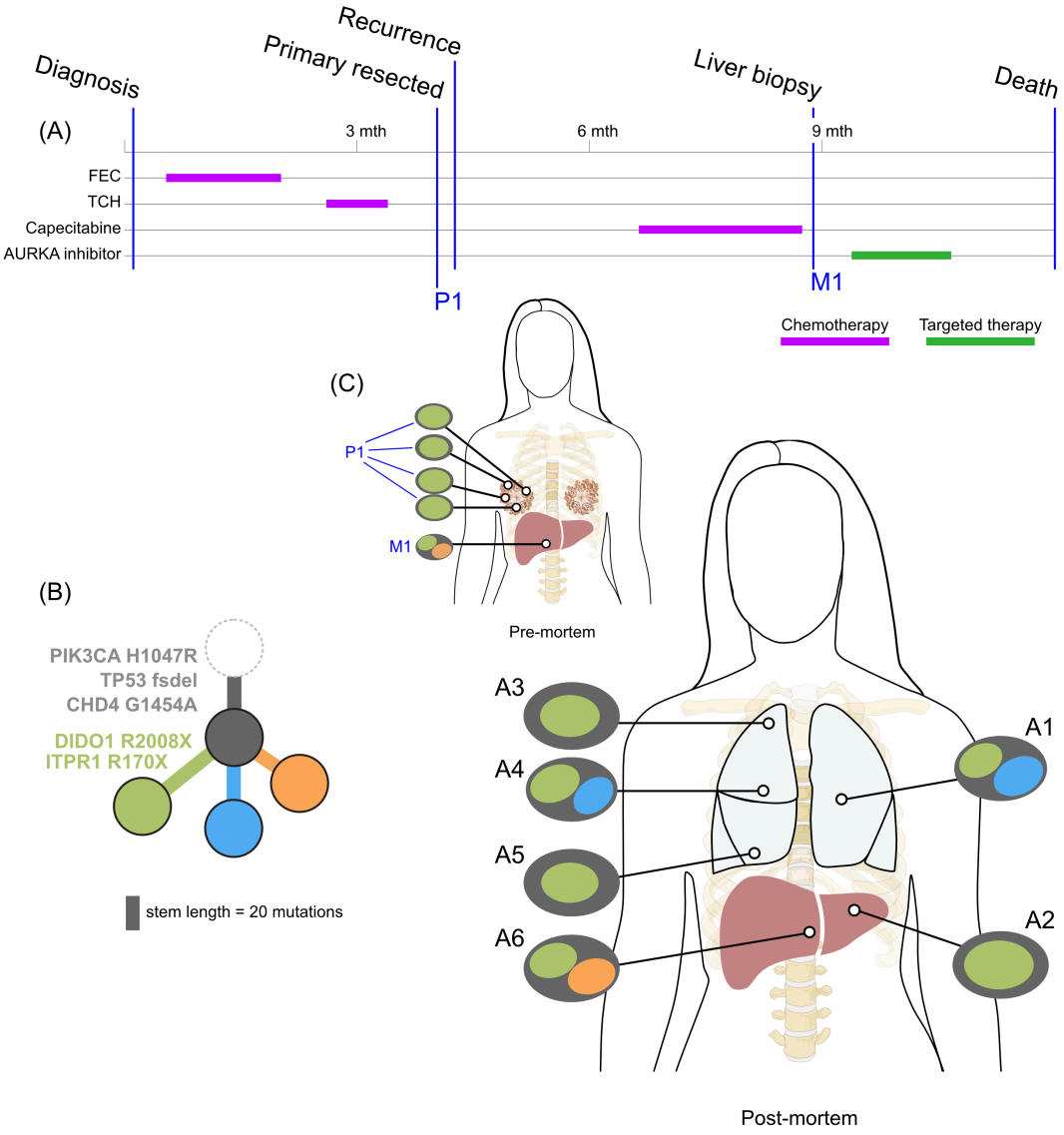


Fig 6.5. Treatment history (A) and subclonal structure (B and C) for TN1, inferred using PyClone and Schism. Pre and post-mortem samples are shown. FEC: 5-fluorouracil, epirubicin, cyclophosphamide; TCH: docetaxel, carboplatin, trastuzumab; AURKA: aurora kinase; P1: multiple samples from archival breast primary (post-neoadjuvant chemotherapy); M1: liver core biopsy; A1: left upper

lobe metastasis; A2: left liver lobe metastasis; A3, A4, A5: left lung metastases; A6: right liver lobe metastasis.

All ER-positive cases had founding clones containing known driver alterations. These founding clones developed secondary alterations before giving rise to diverse subclones across multiple metastatic sites. By way of example, in ER1 shown in Fig 2B, we inferred the presence of an intermediate blue clone not found in isolation. This follows from 1) the absence of this clone in the primary tumors, indicating the blue clone is a true subclone; 2) the absence of the orange subclone in 2 samples where the red subclone has a high clonality, indicating the red subclone did not arise in the orange subclone; 3) 2 different subclones (red and orange) with the blue clone as an ancestor. This intermediate blue clone was marked by an *FGFR4* tyrosine kinase domain mutation, and a splice site mutation in *MGA*, which counter-regulates *MYC* activity. The blue clone also contains deletions in genomic regions containing *RB1*, *SPEN*, *CASP9* and *FANCA*. This likely conferred metastatic potential, as all subsequent metastatic subclones arose from this intermediate clone as described above. The *FGFR4* mutation could not be detected in three spatially separated samples of the primary tumor, even with high depth validation sequencing, although this does not rule out very low prevalence subclones at the time of diagnosis. Emergence of an unheralded 'lethal subclone' following treatment has been documented in many tumor types including breast cancer, but here it associated with delayed disease relapse. A similar finding was noted in an autopsy of a lobular breast cancer where recurrence occurred after a similarly long period [1].

ER2 and ER3 display a different relationship between subclonal structure and metastatic disease. For ER2, a biopsy of the breast mass was the index lesion, and the breast primary was removed after chemotherapy, followed by a liver biopsy after several lines of therapy. All pre-mortem samples displayed a similar subclonal structure, with a subclone (orange) marked by a frameshift deletion in the cancer testis antigen *MAGEA3* and a splice site mutation in *ROCK1* that was subsequently detected at multiple sites in the liver at autopsy. The absence of this subclone at other metastatic sites however shows there was early divergence and parallel evolution of at least 2 other subclones (blue and red). Circulating DNA for the *AKT1* (E17K) mutation and three *ESR1* mutations (D538G, S463P, E380Q) was assayed in plasma taken at the time of liver biopsy and subsequently after death. Fig 6F shows all these mutations were detectable at both time points, even though the liver biopsy WES data did not reveal any *ESR1* mutations.

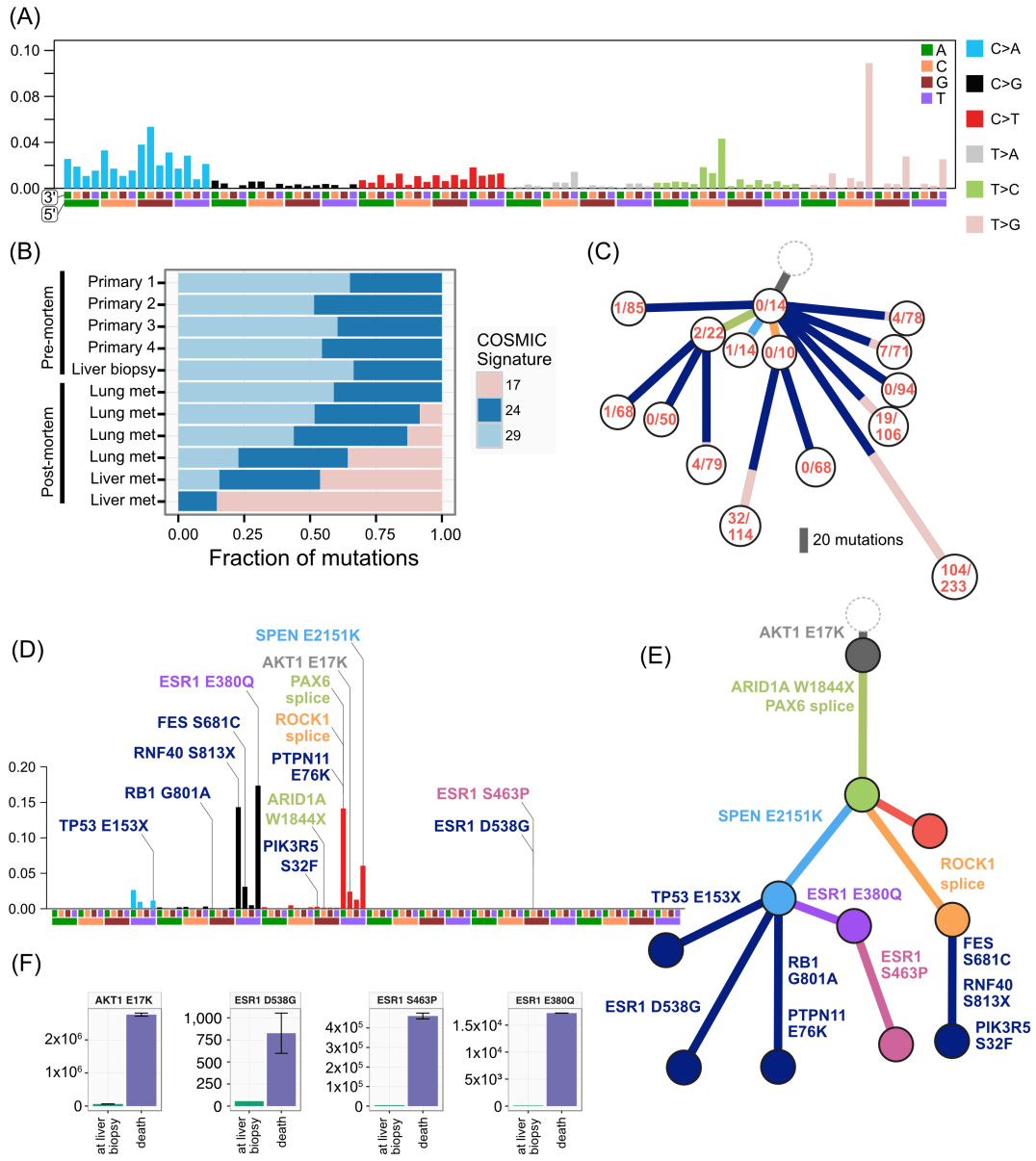


Fig 6.6. Mutational signatures in TN1 and ER2 and ctDNA assays. (A) Reconstruction of mutational context plot for TN1 using signatures 17, 24 and 29 from deconstructSigs. 5` and 3` nucleotides are indicated by colour code on X-axis. (B) Contribution of each signature per sample. (C) Subclonal phylogeny for TN1 showing private and public subclones. Private subclones in dark blue. Mutations arising from signature 17 represented in pink along branches. (D) Reconstruction of mutational context for ER2 using APOBEC signatures 2 and 13 detected with deconstructSigs, with origin of key mutations overlayed. Colour

codings as for (A). (E) Subclonal phylogeny for ER2, with private subclones in blue. (F) Bar charts show ctDNA results at time of liver biopsy and later at death. Y-axis units is copies per millilitre.

ER3 was a case of delayed presentation, with 2 years elapsing between initial detection of early stage disease and subsequent metastatic disease, without intervening treatment. No samples could be obtained from the original diagnosis. In a similar fashion to ER1, the primary subclones have a linear monophyletic relationship with the metastatic subclones. This case is distinguished however by an ovarian lesion that contains the founding clone only. This ovarian lesion was discovered incidentally during a therapeutic oophorectomy prior to any anti-cancer therapy. Hence, despite lacking the more complex subclonal structure seen in other lesions, this disease possessed the ability to metastasise very early. Whether this ovarian metastasis would have resulted in clinically significant disease is unknown. The rest of the lesions must have arisen from dissemination of a different subclone which had acquired the *SPEN* K1838X nonsense mutation and deletion of the negative modulator of estrogen signalling *DUSP22*. These alterations occurred while the patient was receiving endocrine therapy.

TN1 displayed both short linear evolution and early divergence. Fig 5 shows the subclonal structure of the primary is found in all metastatic lesions via the green clone, but the blue and orange clone diverged from a common ancestor derived from an earlier progenitor of the green clone. Fig 5C shows the private subclones, some of which also emerged from the common ancestor. It noteworthy that this case, with the most aggressive disease history with no

response to 3 different standard chemotherapy regimens, had the fewest common subclones (4). S1 Fig demonstrates that CDH4, which contains a truncal mutation, is expressed at the protein level.

A key question is whether metastatic disease from one lesion can cross-seed anatomically distant sites. This is difficult to distinguish from a subclinical low prevalence clone which may seed two different metastatic sites, a disseminative pattern rather than cross-seeding. All cases showed spatially distinct lesions with very similar subclonal structures (e.g. liver lesions in ER1, para-aortic nodal metastases in ER2; brain, lung and liver metastases in ER3; lung metastases in TN1). This is suggestive of metastatic cross seeding. In addition, ER1, ER2 and TN1 show subclonal mixtures, raising the possibility of seeding by polyclonal clusters sustained by clonal cooperation, where two subclones help maintain each other's survival and confer novel phenotypic traits [3,44]. For such patterns to arise from dissemination alone rather than cross seeding, polyclonal seeding or cooperation, multiple waves of dissemination by different subclones from a primary tumor to the same metastatic site would be required.

Copy number and subclonal evolution

In contrast to subclonal mutations, copy number changes between samples were relatively stable (Fig 7) for ER1, ER2 and TN1. Areas of LOH were particularly consistent across samples. ER1 and ER2 displayed relatively few copy number alterations, with overall 80 - 85% of coding genes unaffected. ER3 showed widespread copy number derangement affecting 40 – 60% of coding genes, particularly on chromosomes 8 and 20. The ploidy of ER3 increased from early to late stage disease - the primary, ovarian metastasis and lymph node biopsy all

displayed diploid genomes, with autopsy samples showing triploidy and tetraploidy in some cases (mean ploidy across samples 3.2). There was inter-lesion copy number heterogeneity for ER3. This is reflected in the relative abundance of subclonal copy number events in ER3 called by superFREQ, with a mean of 22 copy number events per subclone for ER3, compared to 7 and 11 for ER1 and ER2 respectively. Chromosomal instability (CIN) was therefore an ongoing process contributing to subclonal evolution in ER3. TN1 showed the most deranged genome which also had mean ploidy 3.9 consistent with genome doubling, with 40 – 50% of genes subject to LOH, and some form of copy number derangement affecting 70 – 80% of genes. Subclonal copy number could not be called satisfactorily for TN1. Although less deranged, in ER1 there was evidence that the primary event was amplification of 1p, prior to driver mutations in *PIK3CA* (Fig 2B).

Fig 6.7. Heatmap of copy number across all samples. X-axis shows all coding genes in order of chromosomal coordinate. Grey areas on map represent difficulty in calling allele specific copy number in FFPE samples. Horizontal blue lines indicate areas of LOH. Deep red is high level amplification, pale red is amplified with 4 or more copies, blue is deleted with less than 2 copies. White is 2 – 3 copies. Ploidy is displayed in last column on right. Y-axis terms correspond to Fig 2-5.

These findings are consistent with the growing body of work that extensive (CIN) takes place early in the evolution of a tumor. In particular, a variety of data from single cell sequencing of tumors and normal cells has shown tumors contain only a few copy number states, without a large number of intermediates, suggesting that large scale copy number changes appear to occur in a punctuated fashion, rather than accumulating gradually over time [8]. A limitation of studying this question in bulk sequencing is that detecting subclonal copy number changes with high sensitivity is difficult particularly in specimens with CIN and suboptimal purity.

Treatment resistance

These cases present a unique opportunity to study clonal evolution during therapy. Breast cancer is the prototypical 'treatable' solid tumor where patients receive multiple lines of therapy that usually result in a partial response or stable disease, although TN1 was treatment refractory.

There was clear evidence that treatment alters subclonal structure. Disease at autopsy showed subclones which contained mutations conferring chemotherapy resistance. One of these TYMS V106M is at a highly conserved residue in the binding site of thymidylate synthase, adjacent to two mutations known to cause resistance to 5FU in vitro (Fig 2B- yellow branch) [45].

Mutations in the ligand binding domain of *ESR1* have recently been discovered as a common mechanism of resistance to endocrine therapy in metastatic breast cancer [46]. ER2 displayed three different *ESR1* mutations all in different subclones. Furthermore, an *ESR1* S463P mutation occurred in a subclone which already harboured an *ESR1* E380Q, presumably in a different allele. As *ESR1*

mutations can be treated with alternative endocrine therapy and represent a tractable form of endocrine resistance, identifying them in patients failing endocrine therapy is of some importance [47]. ER2 had a liver biopsy while alive which did not show any *ESR1* mutations, highlighting the difficulty in managing, advanced heterogeneous disease. At the time of the liver biopsy, circulating cell free DNA in plasma showed detectable low levels of all three *ESR1* mutations (Fig 6FF), which increased dramatically by the time of autopsy.

ER3 demonstrated discordant clinical responses on several occasions, with disease in the lung responding while disease in the liver progressed, and left and right hilar lymph nodes responding differentially to chemotherapy. Fig 5C shows the divergent subclonal structure of lung, liver and hilar lymph nodes, although this may be a consequence of heterogeneous treatment responses, rather than the cause. Fig 7 also shows that differences in copy number can be discerned between sample pairs in ER3, more so than the other cases, as discussed in the previous section.

We found additional evidence of this effect when considering the evolutionary trajectory in the context of dominant oncogenic drivers. In ER2 and ER3 there were multiple subclones with alterations in genes that interact with the truncal drivers in each case, *AKT1* and *TP53* respectively, depicted in Fig 8. For ER2 alterations occurred in genes involved in the regulation of phosphatidylinositide phohsphates and PI3K signalling (*PLPP4, PNPLA6, PIK3R5*), negative regulators of *AKT1* signalling (*PLEKHO1, PHLPP1, PID1, USP12*) and downstream effectors of *AKT1* (MFN1). For ER3, alterations occurred in co-activators and facilitators of TP53 activity (*BACH2, ANKRD11, TP53I11, MYO10, WDR3, NDUFAF6*) and

negative regulators of TP53 function (*G3BP2*) and downstream effectors (*CD82*). These findings suggest that resistance to chemotherapy agents may arise from augmentation of existing oncogenic or tumor suppressor signalling pathways, rather than direct resistance through altered drug targets or drug metabolism. This could be one explanation why increasing lines of treatment become progressively less effective, despite having non-overlapping mechanisms of action.

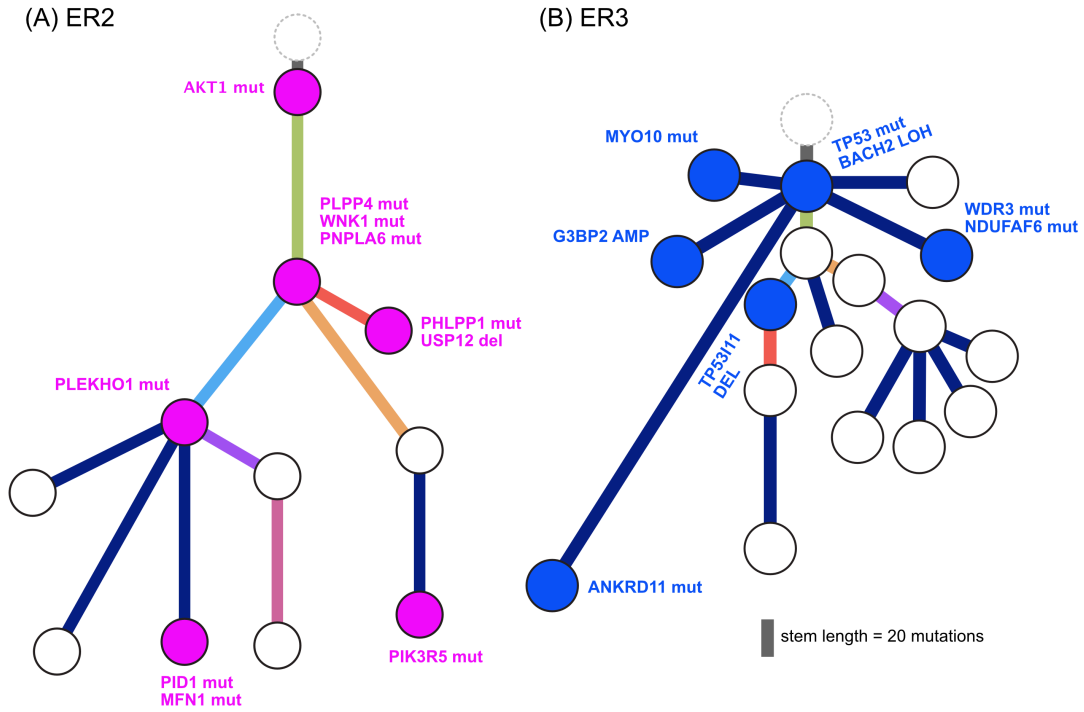


Fig 6.8. Evolution of augmented oncogenic signalling in (A) ER2 and (B) ER3. Subclonal phylogenies with private subclones are displayed with mutations that are expected to modulate oncogenic signalling.

Mutational signatures drive heterogeneity and treatment resistance.

Due to relatively low numbers of mutations, clear mutational signatures could be established only in ER2 and TN1. The pattern of mutational signatures in these

two cases were robust, and maintained a stable prevalence even when the starting mutation pool was down sampled to 10% of the total.

TN1 had an unusual composition of mutational signatures (Fig 6A-C). There was a strong signal from signature 17 in the COSMIC classification [38,39]. This signature has been reported in breast cancer as well as oesophageal, liver, lung and stomach cancers, melanoma and B-cell lymphoma [38]. The aetiology is unknown. As signature 17 produces a characteristic T>G transversion that has relatively little overlap with other signatures, the contribution of the signature to the mutations in subclones can be tracked. Mutations arising from this signature were found primarily in metastatic lesions from autopsy, comprising between 9 – 45% of the private mutations in four spatially distinct lesions in the lung and liver. The signature was not active in the primary lesions or ante-mortem liver biopsy. These private signature 17 mutations have subclonal fractions in the 0.2 – 0.6 range and clustered with other mutations to form private subclones. Assuming that it is extremely unlikely a mutational process could cause the same mutation in multiple cells within a single lesion, each private subclone with a relatively high subclonal fraction of the signature 17 mutations must have arisen from a very restricted population of single cells that expanded in each lesion. In addition, as a mutational process per se confers no proliferative advantage, the prevalence of this signature may represent cells that were able to survive a treatment bottleneck, sustaining DNA damage in the process. The only treatment the patient received after the liver biopsy (which did not contain signature 17) was a pan aurora kinase inhibitor. There was no evidence of response to this therapy. It is unclear how an aurora kinase inhibitor could cause this mutational signature. Signatures 24 and 29 also show activity in TN1 samples. Signature 29

has been associated with the mutagenic effect of tobacco chewing on the oral mucosa, and although distinct from the classic smoking associated signature 4, there is a high degree of overlap. Signature 24 is reported to occur with aflatoxin exposure. The significance of these carcinogen induced signatures in a triple negative breast cancer are unclear.

ER2 was dominated by signatures 2 and 13 in the COSMIC classification [39], which both relate to the well-known phenomenon of APOBEC deaminase activity (Fig 6D). This signature was present consistently in both early and late clones, and allows us to study how APOBEC may contribute to the evolution of breast cancer. Key driver mutations *AKT1 E17K, SPEN* and the *ESR1* E380Q mutation are consistent with those induced by APOBEC. However, the *ARID1A* and the *TP53* truncating mutations have a low probability of arising from the APOBEC signature, and the *ESR1* S463P and D538G mutations are not consistent with APOBEC. In this case therefore, APOBEC-related activity alone was able to furnish most of the putative driver mutations, and resulted in an *ESR1* mutation that causes resistance to endocrine therapy, but could not explain all of the *ESR1* mutations.

DISCUSSION

In this study we analyse four cases chosen to represent difficult clinical scenarios in the contemporary management of advanced breast cancer. We took the approach of extensively sampling metastatic disease at the time of autopsy, performing whole exome sequencing and single nucleotide polymorphism arrays coupled with high depth validation of discovered mutations. We also analysed samples taken from the primary tumor and where available, metastatic lesions biopsied while patients were still alive. In line with other work utilising the CASCADE program, we found studying advanced disease at the time of death to be highly informative [48]. We found significant heterogeneity present across multiple metastatic sites, and by performing subclonal inference, it was possible to understand the key processes that drive tumor evolution over time. We have shown several novel findings including how treatment shapes clonal evolution, the importance of mutational processes over a disease course, and augmentation of oncogenic signalling as a mechanism of treatment failure.

Late relapse is a significant problem for ER positive disease, which continues to show a decline in survival beyond 10 years from diagnosis. In the ER1 case with a long clinical latency, we observed that the subclone giving rise to metastatic disease was not detectable at diagnosis with standard sequencing and sampling measures. This has important implications for genomic determinism, or the expectation that genomic assays from a single time point are able to predict clinical outcome and guide therapy. In this case, additional oncogenic drivers from mutations and copy number alterations accumulated before metastatic disease occurred. The long time to recurrence may be because micrometastatic disease needed to acquire a driver mutation before causing symptoms, or that

the intermediate clone was present early in the natural history but was suppressed by tamoxifen. Although these possibilities cannot be distinguished here, other studies have found evidence that the primary tumor population contains low frequency subclones which may give rise to metastatic or recurrent disease [49]. Our results are similar to those obtained in another case of recurrent disease after a long latency in lobular breast cancer [1]. Assuming a model where infrequent subclones may give rise to eventual relapse 5 years or more after initial therapy, prolonging adjuvant endocrine therapy is expected to be beneficial, as has been shown in clinical trials [50,51]. Detecting and eliminating these rare subclones is a difficult proposition, but targeting known clonal driver alterations could be one strategy. These driver alterations could also be used to monitor for disease recurrence via ctDNA.

To our knowledge, ER2 is the first case of de novo metastatic ER positive/HER2 negative disease to undergo longitudinal sampling of metastases and autopsy along with ctDNA assessments. In contrast to ER1, ER2 showed less divergence between the primary and metastatic lesions. The most likely explanation for this pattern is that metastatic potential was available early in the evolutionary history of this case. Another explanation for this could be reseeding of the primary tumors from a metastatic niche, as has been reported elsewhere [52]. This case did display subclonal patterns consistent with metastatic reseeding between para-aortic nodal sites. That this mechanism exists in breast cancer highlights the difficulties in accurately profiling the cancer genome, since even a lesion once biopsied may significantly change its subclonal structure. Circulating tumor DNA is one method to avoid sampling bias, as was illustrated in this case

where biopsy of a liver lesion did not reveal any of the three *ESR1* mutations that were detectable in plasma at the time.

In addition to ESR1, all three ER positive cases showed alterations in *SPEN*, after exposure to endocrine therapy. SPEN has recently been implicated as a novel tumor suppressor that regulates cell proliferation and inhibits estrogen receptor downstream signalling with a role in resistance to tamoxifen [53]. The truncal location of *SPEN* alterations in 3 different cases suggests this is a bona fide tumor suppressor and mediator of resistance to endocrine therapy.

In ER3 there was the unexpected finding that the earliest subclone was able to metastasise to the ovary. This case is also notable for the widespread copy number derangement that was present in all samples, and similar to TN1, this may have conferred the metastatic phenotype. ER3 displayed ongoing copy number changes that segregated with subclones in our analysis. This was in contrast to the other ER positive cases, where mutational processes dominated. The presence of CIN has been shown in vitro to be associated with drug resistance, and there is some evidence that aneuploidy is an adaptive response in lower eukaryotes [54,55]. There is increasing evidence for punctuated evolution of copy number in cancers [8,56]. This fits well with our data for ER1, ER2 and TN1. The situation for ER3 is less clear cut, as there is ongoing acquisition of copy number changes beyond increases in ploidy. In some cancers therefore, 'active' CIN is likely to be an important mechanism driving evolution, treatment resistance and heterogeneity. Of note, during treatment with tamoxifen and chemotherapy, ER3 showed discordant responses between lesions in liver and lung. This could be explained by the significant copy number heterogeneity seen

in this case, although it is not possible to trace a pre-mortem lesion on imaging to a lesion at autopsy.

Understanding mutational signatures provided valuable insights for ER2 and TN1. In TN1, the late emergence of signature 17 implies rapid rescaling of the subclonal structure of several metastatic lesions. The aetiology of signature 17 is unknown, but the activity of this signature in unrelated subclones at different anatomical sites suggests it can be chemically induced. Further to this, the presence of apparent carcinogen induced mutational signatures in TN1 also raises questions about the aetiology of aggressive triple negative breast cancers.

In ER2 we were able to observe the action of APOBEC in shaping subclonal evolution, where it was responsible for many important alterations of functional significance, including the key founding drive mutation. It has been previously noted that APOBEC may give rise to clonal and subclonal driver mutations, [57,58], but we demonstrate here that this mutational activity is maintained during treatment and throughout the natural history of the disease. Arresting the activity of APOBEC may be a potential strategy to restrain progression and evolution of APOBEC enriched cancers.

We identified a potentially novel mechanism of broad treatment resistance to chemotherapy, which arises from augmentation of existing oncogenic signalling. This implies that oncogenic signalling remains essential even in heavily pre-treated disease, and raises the possibility that combining targeted therapies with chemotherapy may severely restrict the fitness landscape which a tumor can access to achieve treatment resistance. The superior efficacy of combination therapy is well known, and this approach has been used with great success in HER2 positive breast cancer for example [59]. Other studies have shown that

driver alterations influence response to chemotherapy [7]. Our findings extend these concepts to define augmented oncogenic signalling as a resistance mechanism that is widespread in advanced disease, and may be therapeutically tractable.

There are several limitations to this study. WES alone may result in poor resolution to detect subclones, which could underestimate the subclonal diversity present. In contrast to mutations, detection of subclonal copy number events remains difficult, and important subclonal amplifications or deletions may have been missed. Newer technologies, such as single cell approaches or long read sequencing will be required to overcome these limitations. In addition, we did not analyse structural variants, the transcriptome, the epigenome or the proteome which along with non-coding elements such as long non-coding RNAs or microRNAs could make important contributions to subclonal evolution and phenotypic diversity not captured by WES. Although we have analysed a small number of cases in detail, it is unclear how whether our findings are representative of the broader patient population.

In conclusion, we demonstrate the feasibility and value of subclonal inference in understanding the biology and evolutionary history of lethal breast cancer. This approach also provided insight into difficult clinical scenarios in breast cancer. Extension of the whole exome approach to whole genome, transcriptome and methylome studies, as well as more novel single cell and long read sequencing technology is expected to provide further insights, particularly when used in the setting of rapid autopsy studies which afford the unprecedented ability to sample multiple evolutionary trajectories comprehensively. It is notable that each case studied was unique in the processes that ultimately resulted in death.

It is unclear if patterns will be found that can generalize across patient subgroups: for this we will need large cohorts where we can track genetic evolution from diagnosis to death. To that end, our prospective rapid autopsy program, which continues to accrue in breast cancer and other cancer types [9], as well as other international efforts will be essential to help us understand how cancer disseminates and ultimately becomes resistant to treatment [60].

ACKNOWLEDGEMENTS

We are very grateful to the patients and their families that agreed to participate in this research. We also gratefully acknowledge the Victorian Institute of Forensic Medicine that conducted the rapid autopsies, and Tobin Brothers Funerals Melbourne for their assistance.

The Kathleen Cuningham Foundation Consortium for Research (kConfab) wishes to acknowledge Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study for their contributions to this resource, and the many families who contribute to kConFab.

REFERENCES

1. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, et al. Mutational evolution in a lobular breast tumor profiled at single nucleotide resolution. Nature. 2009;461: 809–13. doi:10.1038/nature08489
2. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature. 2010;464: 999–1005. doi:10.1038/nature08989
3. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature. 2014;518: 422–426. doi:10.1038/nature13952
4. Murtaza M, Dawson S-J, Pogrebniak K, Rueda OM, Provenzano E, Grant J, et al. Multifocal clonal evolution characterized using circulating tumor DNA in a case of metastatic breast cancer. Nat Commun. 2015;6: 8760. doi:10.1038/ncomms9760
5. Juric D, Castel P, Griffith M, Griffith OL, Won HH, Ellis H, et al. Convergent loss of PTEN leads to clinical resistance to a PI(3)Kα inhibitor. Nature. 2014;518: 240–244. doi:10.1038/nature13948
6. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015;21: 751–9. doi:10.1038/nm.3886
7. Janiszewska M, Liu L, Almendro V, Kuang Y, Paweletz C, Sakr RA, et al. In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. Nat Genet. 2015;47: 1212–9. doi:10.1038/ng.3391

8. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet. 2016;48: 1119–1130. doi:10.1038/ng.3641

9. Alsop K, Thorne H, Sandhu S, Hamilton A, Mintoff C, Christie E, et al. A community-based model of rapid autopsy in end-stage cancer patients. Nat Biotechnol. 2016;34: 1010–1014. doi:10.1038/nbt.3674

10. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.

11. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25: 1754–60. doi:10.1093/bioinformatics/btp324

12. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;43: 11.10.1-33. doi:10.1002/0471250953.bi1110s43

13. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31: 213–9. doi:10.1038/nbt.2514

14. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44: e108–e108. doi:10.1093/nar/gkw227

15. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized

insertions from paired-end short reads. Bioinformatics. 2009;25: 2865–71. doi:10.1093/bioinformatics/btp394

16. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38: e164. doi:10.1093/nar/gkq603

17. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526: 68–74. doi:10.1038/nature15393

18. NHLBI GO Exome Sequencing Project (ESP). Exome Variant Server [Internet]. [cited 18 Oct 2016]. Available: http://evs.gs.washington.edu/EVS/

19. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44: e131. doi:10.1093/nar/gkw520

20. Staaf J, Vallon-Christersson J, Lindgren D, Juliusson G, Rosenquist R, Höglund M, et al. Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. BMC Bioinformatics. 2008;9: 409. doi:10.1186/1471-2105-9-409

21. Yau C, Mouradov D, Jorissen R, Colella S, Mirza G, Steers G, et al. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biol. 2010;11: R92. doi:10.1186/gb-2010-11-9-r92

22. Bengtsson H, Neuvial P, Speed TP. TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. BMC Bioinformatics. 2010;11: 245.

doi:10.1186/1471-2105-11-245

23. Olshen AB, Bengtsson H, Neuvial P, Spellman PT, Olshen RA, Seshan VE. Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. Bioinformatics. 2011;27: 2038–46. doi:10.1093/bioinformatics/btr329

24. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015;43: D805-11. doi:10.1093/nar/gku1075

25. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015;163: 506–19. doi:10.1016/j.cell.2015.09.033

26. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–5. doi:10.1038/ng.2892

27. Miller ML, Reznik E, Gauthier NP, Aksoy BA, Korkut A, Gao J, et al. Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. Cell Syst. 2015;1: 197–209. doi:10.1016/j.cels.2015.08.014

28. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34: 155–63. doi:10.1038/nbt.3391

29. Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, et al. High-performance web services for querying gene and variant annotation. Genome Biol. 2016;17: 91. doi:10.1186/s13059-016-0953-9

30. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G.

GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12: R41. doi:10.1186/gb-2011-12-4-r41

31. Sanchez-Garcia F, Villagrasa P, Matsui J, Kotliar D, Castro V, Akavia U-D, et al. Integration of Genomic Data Enables Selective Discovery of Breast Cancer Drivers. Cell. 2014;159: 1461–75. doi:10.1016/j.cell.2014.10.048

32. Flensburg C. superFREQ [Internet]. 2016 [cited 18 Oct 2016]. Available: https://github.com/ChristofferFlensburg/superFreq

33. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15: R29. doi:10.1186/gb-2014-15-2-r29

34. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11: 396–398. doi:10.1038/nmeth.2883

35. Niknafs N, Beleva-Guthrie V, Naiman DQ, Karchin R, Nowell P, Greaves M, et al. SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. Morris Q, editor. PLOS Comput Biol. 2015;11: e1004416. doi:10.1371/journal.pcbi.1004416

36. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. Prlic A, editor. PLoS Comput Biol. 2013;9: e1003118. doi:10.1371/journal.pcbi.1003118

37. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution.

Genome Biol. 2016;17: 31. doi:10.1186/s13059-016-0893-4

38. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. 2013;500: 415–21. doi:10.1038/nature12477

39. COSMIC. Signatures of Mutational Processes in Human Cancer [Internet]. [cited 18 Oct 2016]. Available: http://cancer.sanger.ac.uk/cosmic/signatures

40. Fumagalli D, Wilson TR, Salgado R, Lu X, Yu J, O'Brien C, et al. Somatic mutation, copy number and transcriptomic profiles of primary and matched metastatic estrogen receptor-positive breast cancers. Ann Oncol Off J Eur Soc Med Oncol. 2016;27: 1860–6. doi:10.1093/annonc/mdw286

41. Griffith M, Miller CA, Griffith OL, Krysiak K, Skidmore ZL, Ramu A, et al. Optimizing Cancer Genome Sequencing and Analysis. Cell Syst. 2015;1: 210–223. doi:10.1016/j.cels.2015.08.015

42. El-Kebir M, Satas G, Oesper L, Raphael BJ, Agarwala R, Fernández-Baca D, et al. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. Cell Syst. 2016;3: 43–53. doi:10.1016/j.cels.2016.07.004

43. Gundem G, Van Loo P, Kremeyer B, Alexandrov LB, Tubio JMC, Papaemmanuil E, et al. The evolutionary history of lethal metastatic prostate cancer. Nature. 2015;520: 353–357. doi:10.1038/nature14347

44. Marusyk A, Tabassum DP, Altrock PM, Almendro V, Michor F, Polyak K. Non-cell-autonomous driving of tumor growth supports sub-clonal heterogeneity. Nature. 2014;514: 54–8. doi:10.1038/nature13556

45. Kawate H, Landis DM, Loeb LA. Distribution of mutations in human

thymidylate synthase yielding resistance to 5-fluorodeoxyuridine. J Biol Chem. 2002;277: 36304–11. doi:10.1074/jbc.M204956200

46. Toy W, Shen Y, Won H, Green B, Sakr RA, Will M, et al. ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. Nat Genet. 2013;45: 1439–1445. doi:10.1038/ng.2822

47. Fribbens C, O'Leary B, Kilburn L, Hrebien S, Garcia-Murillas I, Beaney M, et al. Plasma ESR1 Mutations and the Treatment of Estrogen Receptor-Positive Advanced Breast Cancer. J Clin Oncol. 2016;34: 2961–8. doi:10.1200/JCO.2016.67.3061

48. Patch A-M, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S, et al. Whole–genome characterization of chemoresistant ovarian cancer. Nature. 2015;521: 489–494. doi:10.1038/nature14410

49. Morrissy AS, Garzia L, Shih DJH, Zuyderduyn S, Huang X, Skowron P, et al. Divergent clonal selection dominates medulloblastoma at recurrence. Nature. 2016;529: 351–7. doi:10.1038/nature16478

50. Davies C, Pan H, Godwin J, Gray R, Arriagada R, Raina V, et al. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. Lancet. 2013;381: 805–16. doi:10.1016/S0140-6736(12)61963-1

51. Goss PE, Ingle JN, Pritchard KI, Robert NJ, Muss H, Gralow J, et al. Extending Aromatase-Inhibitor Adjuvant Therapy to 10 Years. N Engl J Med. 2016;375: 209–219.

52. Hong MKH, Macintyre G, Wedge DC, Van Loo P, Patel K, Lunke S, et al. Tracking the origins and drivers of subclonal metastatic expansion in

prostate cancer. Nat Commun. 2015;6: 6605. doi:10.1038/ncomms7605

53. Légaré S, Cavallone L, Mamo A, Chabot C, Sirois I, Magliocco A, et al. The Estrogen Receptor Cofactor SPEN Functions as a Tumor Suppressor and Candidate Biomarker of Drug Responsiveness in Hormone-Dependent Breast Cancers. Cancer Res. 2015;75: 4351–63. doi:10.1158/0008-5472.CAN-14-3475

54. Lee AJX, Endesfelder D, Rowan AJ, Walther A, Birkbak NJ, Futreal PA, et al. Chromosomal instability confers intrinsic multidrug resistance. Cancer Res. 2011;71: 1858–70. doi:10.1158/0008-5472.CAN-10-3604

55. Chen G, Bradford WD, Seidel CW, Li R. Hsp90 stress potentiates rapid cellular adaptation through induction of aneuploidy. Nature. 2012;482: 246–50. doi:10.1038/nature10795

56. Notta F, Chan-Seng-Yue M, Lemire M, Li Y, Wilson GW, Connor AA, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. Nature. 2016;538: 378–382. doi:10.1038/nature19823

57. Swanton C, McGranahan N, Starrett GJ, Harris RS. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov. 2015;5: 704–12. doi:10.1158/2159-8290.CD-15-0344

58. Henderson S, Chakravarthy A, Su X, Boshoff C, Fenton TR. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. Cell Rep. 2014;7: 1833–41. doi:10.1016/j.celrep.2014.05.012

59. Swain SM, Baselga J, Kim S-B, Ro J, Semiglazov V, Campone M, et al. Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic

Breast Cancer. N Engl J Med. 2015;372: 724–734. doi:10.1056/NEJMoa1413513

60. Erdmann J. Trace elements: Innovative biopsy programs map how cancer spreads. Nat Med. 2016;22: 963–5. doi:10.1038/nm0916-963

Supporting Information – see Appendix 1

Conclusion

Precision oncology has recently been criticised for failing to deliver on the benefits it promised, and has even been described as an illusion rather than a revolution in cancer care [251–255]. The most important treatment to emerge for solid tumours in the last 5 years is T-cell checkpoint blockade, which is not a genomic targeted therapy at all, and is administered to all suitable patients. The success of this decidedly imprecise approach is undeniable, appearing to cure a significant proportion of patients with advanced melanoma who would otherwise have a dismal prognosis [256]. Remarkably this can be achieved with as little as a few doses of therapy. In the shadow of these successes, after considerable hype and a large deployment of genomic technologies in cancer care, the right model for precision oncology still remains obscure.

The aim of this project was to implement a precision oncology program at the same time as studying why such an endeavour may fail. The latter falls under the rubric of heterogeneity and cancer evolution, which may thwart the relevance of genomic profiles taken by necessity at a particular and place. Putting these complications aside for a moment, the primary outcome from this work was that implementing a precision oncology program was feasible. This was undertaken with a small team, without a large investment in infrastructure. The concept of the program appeared to be successful in the eyes of patients and clinicians, judging by the strong and consistent recruitment. Genomic technology has matured to the point where this study could be rolled out with minimal technical challenges, notwithstanding a false start on the wrong platform. It is reasonable to conclude

therefore that the barrier to precision oncology is no longer related to sequencing tumours or generating results. With this in mind, it is likely that access to genomic assays for patients will continue to increase with time. The chief challenges encountered in this project were logistical and clinical. Acquiring tumour samples in a timely fashion remains a significant challenge. This is due in part to the attitude of pathology providers, where tumour blocks are considered their property rather than belonging to the patients. One has to question the wisdom of pathology providers keeping and storing tumour blocks for patients with metastatic disease.

Translating precision oncology into a meaningful benefit for patients remains the foremost challenge. In this study, a reasonable proportion of patients were able to access therapies matched to their genomic profiles if the most favourable denominator is chosen (namely, all patients with successful sequencing). If it is considered that many patients existed that were too unwell or too far away to consider enrolling in the study, then the rate at which patients are able to receive matched therapies drops far below 10%, and the proportion that then benefit becomes vanishingly small. The frequency with which patients received matched therapy also relies on the fact that *PIK3CA* mutations are a common driver and for the duration of SEGMENT there was clinical trial with a PIK3CA inhibitor available.

It was apparent that a small number of driver genes were responsible for all actionable alterations. In this cohort, copy number data contributed relatively little of clinical relevance. It is difficult at this time to justify the use of a larger gene panel or whole exome or even whole genome sequencing in light of this. It is not a

given that rare alterations will ever develop into biomarkers with matched therapies. Only 1 patient of 214 had the unexpected actionable BRAF mutation, indicating that detection of these unheralded alterations will remain a very rare event, not to mention the complexity in deciding on the right therapy in these circumstances. It is thus likely that truly useful genomic alterations such as *PIK3CA* mutations will be developed into individual assays that can be widely distributed to pathology providers.

The case of *ESR1* and *ERBB2* mutations provide a more compelling indication for the panel based approach employed in this study. These alterations are dynamic and emerge under the selective pressure of therapy. Detecting their presence in clinically significant disease may impact management, although drugs which restore endocrine responsiveness in the setting of *ESR1* mutations are still under development. *ESR1* mutations may evolve in parallel, as was seen in the autopsy case where multiple mutations were found in the same patient. These more complex genomic alterations that emerge during treatment may provide a niche where genomic assays may be higher yield. It was readily apparent in paired samples in the SEGMENT study and the autopsy series that treatment has a profound impact on genomic profiles to which patient and clinician remain oblivious. When a new therapy is introduced, the well established methodology of genomic profiling allows a picture of tumour evolution under fire to be rapidly developed. While it took many decades of endocrine therapy usage for *ESR1* mutations to be discovered, putative genomic biomarkers of resistance have been detected much more promptly for newer therapies, with *PTEN* resistance alterations after treatment with PIK3CA inhibitors being a prime example [242].

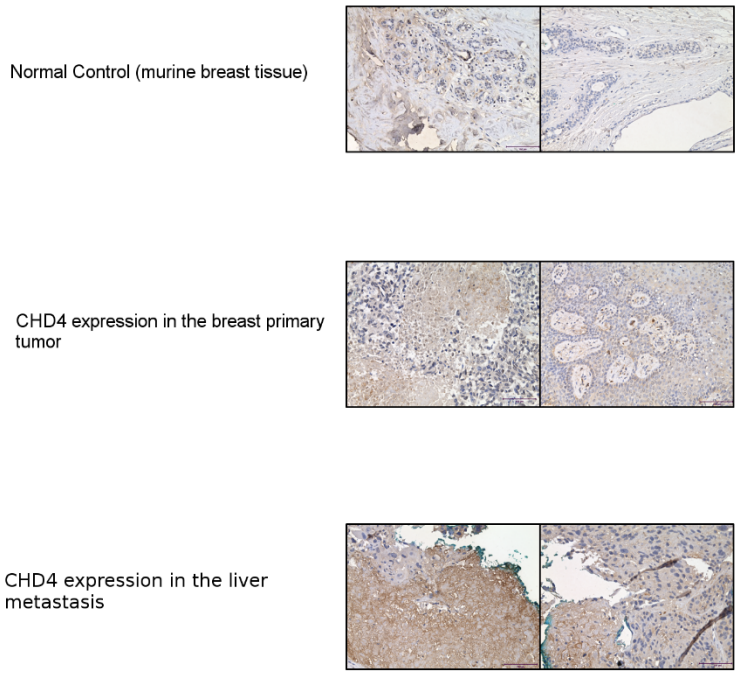
And although the success of immunotherapy is in some ways an argument against precision oncology, genomic markers that predict immunotherapy benefit such as loss of *PBRM1* [154] and markers that predict resistance such as loss of function alterations in *STK11* [257] have been quickly discovered with research focused application of precision oncology.

One of the more intractable problems in the contemporary management of ER+ breast cancer is the propensity for late relapse, 10 or 20 years beyond the original diagnosis and apparent cure. The autopsy case with delayed relapse indicates shows that the presentation of metastatic disease was associated with the acquisition of new drivers. In SEGMENT there was also a case with a very long natural history, which also displayed late acquisition of a driver *PIK3CA* mutation albeit after metastatic disease had already developed. This ability for latent driver emergence, which would be predicted to affect the tumour behaviour significantly, suggests that genomic profiles at the time of diagnosis will not assist in predicting late relapse. A focus on monitoring and early detection of late relapse, perhaps with non-invasive assays such as circulating tumour DNA, followed by steps to minimise and control disease burden may be one approach to dealing with late relapse.

The autopsy cases revealed a degree of genomic complexity which paired sample analysis in SEGMENT can only hint at. The lessons learned from studying the subclonal evolution of very advanced disease are largely sobering. It is difficult to see how any therapy could treat these high volume, heavily treated and highly diversified tumour populations which may approach half a trillion in number

[258]. Nevertheless, by providing a view of breast cancer expressing its most lethal and recalcitrant nature, some inference may be made as to the processes which lead to that point. Just as driver alterations such as *TP53* and *PIK3CA* were commonly concordant between primary and metastatic samples in SEGMENT, these alterations formed the tree of the evolutionary trunk in the autopsy cases where a much higher resolution picture of subclonal evolution was available. Further to this point, the nature of the truncal drivers appeared to influence subsequent subclonal diversification, emphasising their ongoing relevance in the natural history of the disease.

Appendix 1- Supplementary Figure and Tables for Chapter 6



S1 Fig. Immunohistochemistry for the protein product of the CDH4 gene, which contains a truncal mutation in TN1. Both primary tumor and liver metastasis show evidence of expression.

S1 Table: Description of samples.

FileID refers to naming scheme for aligned sequence data which is available from a controlled access repository (see paper).						
Case	**Sample**	**Method**	**Treatment status**	**Detail**	**Indicator**	**FileID**
ER1	Primary Tumour	Surgical excision	Treatment naïve	Primary	P1	ER1_P1A
ER1	Primary Tumour	Surgical excision	Treatment naïve	Primary	P1	ER1_P1B
ER1	Primary Tumour	Surgical excision	Treatment naïve	Primary	P1	ER1_P1C
ER1	Metastatic	Autopsy	Post all treatment	Right lobe liver anterior middle	A1	ER1_A1
ER1	Metastatic	Autopsy	Post all treatment	Left lobe liver	A2	ER1_A2
ER1	Metastatic	Autopsy	Post all treatment	Right lobe liver superior	A3	ER1_A3
ER1	Metastatic	Autopsy	Post all treatment	Right lobe liver lateral middle	A4	ER1_A4
ER1	Metastatic	Autopsy	Post all treatment	Right adrenal	A5	ER1_A5
ER2	Primary Tumour	Core biopsy	Treatment naïve	Breast core biopsy	P0	ER2_P0
ER2	Primary Tumour	Surgical excision	Post chemotherapy	Primary	P1	ER2_P1A
ER2	Primary Tumour	Surgical excision	Post chemotherapy	Primary	P1	ER2_P1B
ER2	Primary Tumour	Surgical excision	Post chemotherapy	Primary	P1	ER2_P1C

ER2	Metastatic	Core biopsy	Post chemotherapy and endocrine therapy	Biopsy of liver mass	M1	ER2_M1
ER2	Metastatic	Autopsy	Post all treatment	Dura	A1	ER2_A1
ER2	Metastatic	Autopsy	Post all treatment	Left lobe liver adjacent to falciform ligament	A2	ER2_A2
ER2	Metastatic	Autopsy	Post all treatment	Left lobe liver left outer	A3	ER2_A3
ER2	Metastatic	Autopsy	Post all treatment	Left lobe liver central	A4	ER2_A4
ER2	Metastatic	Autopsy	Post all treatment	Left para-aortic node	A5	ER2_A5
ER2	Metastatic	Autopsy	Post all treatment	Right lobe liver lateral	A6	ER2_A6
ER2	Metastatic	Autopsy	Post all treatment	Right lobe liver central	A7	ER2_A7
ER2	Metastatic	Autopsy	Post all treatment	Right para-aortic node	A8	ER2_A8
ER3	Metastatic	Surgical excision (incidental finding)	Treatment naïve	Ovarian mass	M1	ER3_M1
ER3	Primary Tumour	Surgical excision	Post endocrine therapy	Primary	P1	ER3_P1A
ER3	Primary Tumour	Surgical excision	Post endocrine therapy	Primary	P1	ER3_P1B

ER3	Metastatic	Core biopsy	Post chemotherapy	Right supraclavicular lymph node	M2	ER3_M2
ER3	Metastatic	Autopsy	Post all treatment	Brian metastasis	A1	ER3_A1
ER3	Metastatic	Autopsy	Post all treatment	Left hilar lymph node	A2	ER3_A2
ER3	Metastatic	Autopsy	Post all treatment	Left lung lower lobe	A3	ER3_A3
ER3	Metastatic	Autopsy	Post all treatment	Paraspinal tissue at T12	A4	ER3_A4
ER3	Metastatic	Autopsy	Post all treatment	Left lobe liver posterior	A5	ER3_A5
ER3	Metastatic	Autopsy	Post all treatment	Left lobe liver anterior inferior	A6	ER3_A6
ER3	Metastatic	Autopsy	Post all treatment	Left perinephric	A7	ER3_A7
ER3	Metastatic	Autopsy	Post all treatment	Right lung upper lobe	A8	ER3_A8
ER3	Metastatic	Autopsy	Post all treatment	Right hilar lymph node	A9	ER3_A9
ER3	Metastatic	Autopsy	Post all treatment	Right rib	A10	ER3_A10
ER3	Metastatic	Autopsy	Post all treatment	Right lobe liver superior	A11	ER3_A11
ER3	Metastatic	Autopsy	Post all treatment	Right lobe liver inferior	A12	ER3_A12

TN1	Primary Tumour	Surgical excision	Post neoadjuvant chemotherapy	Primary 3 o'clock	P1	TN1_P1A
TN1	Primary Tumour	Surgical excision	Post neoadjuvant chemotherapy	Primary 6 o'clock	P1	TN1_P1B
TN1	Primary Tumour	Surgical excision	Post neoadjuvant chemotherapy	Primary 9 o'clock	P1	TN1_P1C
TN1	Primary Tumour	Surgical excision	Post neoadjuvant chemotherapy	Primary 10 o'clock	P1	TN1_P1D
TN1	Metastatic	Core biopsy	Post chemotherapy	Biopsy liver mass	M1	TN1_M1
TN1	Metastatic	Autopsy	Post all treatment	Left lung upper lobe central	A1	TN1_A1
TN1	Metastatic	Autopsy	Post all treatment	Left liver lobe central	A2	TN1_A2
TN1	Metastatic	Autopsy	Post all treatment	Right lung upper lobe superior	A3	TN1_A3
TN1	Metastatic	Autopsy	Post all treatment	Right lung upper lobe inferior	A4	TN1_A4
TN1	Metastatic	Autopsy	Post all treatment	Right lung middle lobe inferior	A5	TN1_A5
TN1	Metastatic	Autopsy	Post all treatment	Right lobe liver adjacent to	A6	TN1_A6

				falciform ligament		

S2 Table: Quality control metrics for sequencing

Sample	Exome library	Mean coverage	processing	% Mapped reads duplicates	% Reads OnTarget	OnTarget properly paired	Median Fragment Length	Purity (from facets)
ER1_A1	NimbleGen	134.55	fresh	3.28	72.3	146,046,503	201	0.67
ER1_A2	NimbleGen	103.18	fresh	3.38	77	111,736,090	198	0.73
ER1_A3	NimbleGen	81.77	fresh	3.39	76.3	88,752,205	202	0.6
ER1_A4	NimbleGen	68.89	fresh	2.83	71.98	74,724,280	196	0.66
ER1_A5	NimbleGen	97.91	fresh	3.02	73.1	106,187,295	206	0.71
ER1_BP	NimbleGen	115.28	blood pellet	4.13	77.51	124,217,592	205	NA
ER1_P1A	NimbleGen	71.66	FFPE	6.39	56.09	81,607,849	143	0.51
ER1_P1B	NimbleGen	68.55	FFPE	5.68	58.84	76,709,222	150	0.6
ER1_P1C	NimbleGen	57.55	FFPE	6.47	58.04	64,261,271	151	0.61
ER2_A1	NimbleGen	106.98	fresh	3.52	75.25	116,329,306	194	0.74
ER2_A2	NimbleGen	99.5	fresh	3.12	73.43	108,980,989	212	0.49
ER2_A3	NimbleGen	99.57	fresh	3.22	74.36	108,334,943	209	0.77
ER2_A4	NimbleGen	90.94	fresh	3.16	73.22	99,026,724	200	0.87
ER2_A5	NimbleGen	92.67	fresh	3.38	76.57	101,325,897	213	0.37
ER2_A6	NimbleGen	124.08	fresh	3.9	74.79	134,241,227	196	0.69
ER2_A7	NimbleGen	167.11	fresh	4.13	74.87	181,175,477	201	0.71
ER2_A8	NimbleGen	95.36	fresh	3.36	76.55	103,241,870	208	0.57
ER2_BP	NimbleGen	126.07	blood pellet	3.19	62.47	138,277,526	210	NA
ER2_M1	NimbleGen	121.27	fresh	3.34	62.81	132,009,464	197	0.49
ER2_P0	NimbleGen	30.73	FFPE	6.23	78.46	32,137,030	195	0.36
ER2_P1A	NimbleGen	83.74	FFPE	4.38	75.18	92,631,444	171	0.26
ER2_P1B	NimbleGen	78.96	FFPE	4.36	73.48	87,609,053	168	0.33
ER2_P1C	NimbleGen	86.19	FFPE	4.28	73.8	96,039,193	164	0.47
ER3_A1	NimbleGen	165.92	fresh	8.69	66.3	163,062,652	175	0.82
ER3_A10	NimbleGen	151.55	fresh	9.12	66.63	148,590,971	174	0.83

ER3_A11	NimbleGen	132.76	fresh	9.58	66.38	129,931,829	179	0.8
ER3_A12	NimbleGen	140.59	fresh	9.45	65.06	136,998,171	182	0.75
ER3_A2	NimbleGen	154.88	fresh	6.65	65.39	148,586,201	187	0.25
ER3_A3	NimbleGen	161.41	fresh	7.29	65.95	158,716,968	176	0.81
ER3_A4	NimbleGen	179.11	fresh	9.03	67.56	171,803,202	187	0.72
ER3_A5	NimbleGen	152.83	fresh	7.82	65.91	147,854,552	181	0.8
ER3_A6	NimbleGen	157.63	fresh	9.97	65.93	151,953,818	189	0.79
ER3_A7	NimbleGen	149.9	fresh	7.65	66.02	144,979,027	184	0.77
ER3_A8	NimbleGen	171.82	fresh	6.96	64.95	162,214,361	206	0.82
ER3_A9	NimbleGen	175.27	fresh	8.44	65.93	165,464,487	201	0.86
ER3_BP	NimbleGen	138.48	blood pellet	4.57	67.27	133,801,877	189	NA
ER3_M1	NimbleGen	166.64	FFPE	17.18	69.71	183,195,632	140	0.41
ER3_M2	NimbleGen	130.07	fresh	4.14	67.01	124,571,499	188	0.57
ER3_P1A	NimbleGen	62.14	FFPE	20.84	60.14	76,479,982	116	0.87
ER3_P1B	NimbleGen	84.6	FFPE	22.6	62.33	102,089,863	121	0.81
TN1_A1	Illumina	200.49	fresh	25.98	51.33	106,101,943	122	0.71
TN1_A2	Illumina	234.46	fresh	26.78	49.78	122,572,360	134	0.74
TN1_A3	Illumina	236.3	fresh	28.08	49.67	124,032,742	131	0.69
TN1_A4	Illumina	196.26	fresh	26.04	50.09	103,285,408	130	0.8
TN1_A5	Illumina	209.24	fresh	27.29	49.88	109,473,580	129	0.58
TN1_A6	Illumina	201.34	fresh	30.88	51.62	105,438,516	131	0.57
TN1_BP	Illumina	206.4	blood pellet	22.69	51.44	108,321,468	136	NA
TN1_M1	Illumina	197.75	FFPE	31.04	54.7	106,200,808	112	0.49
TN1_P1A	Illumina	172.33	FFPE	34.43	54.68	92,223,946	109	0.58
TN1_P1B	Illumina	192	FFPE	30.14	53.84	99,962,175	129	0.39
TN1_P1C	Illumina	146.79	FFPE	34.88	55.14	79,222,786	104	0.58
TN1_P1D	Illumina	128.31	FFPE	32.29	53.91	70,671,543	96	0.49

S3 Table: Validation rates

	AF >= 0.05			AV >= 0.1		
sample	**no**	**yes**	**vrate**	**no**	**yes**	**vrate**
ER1_P1A	0	7	1.00	0	6	1
ER1_P1B	0	7	1.00	0	7	1
ER1_P1C	0	8	1.00	0	7	1
ER1_A2	0	57	1.00	0	52	1
ER1_A5	0	65	1.00	0	64	1
ER1_A4	0	50	1.00	0	49	1
ER1_A3	1	63	0.98	1	59	0.98
ER2_P1A	1	77	0.99	1	57	0.98
ER2_P1B	2	97	0.98	2	93	0.98
ER2_P1C	1	112	0.99	1	111	0.99
ER2_A8	1	172	0.99	1	162	0.99
ER2_A1	1	112	0.99	1	111	0.99
ER2_M1	1	145	0.99	1	144	0.99
ER2_A4	0	171	1.00	0	171	1
ER2_A2	1	109	0.99	1	108	0.99
ER2_A3	4	239	0.98	3	227	0.99
ER2_A5	3	165	0.98	2	100	0.98
ER2_A7	5	253	0.98	3	198	0.99
ER2_A6	4	199	0.98	0	145	1
ER3_M2	0	49	1.00	0	49	1
ER3_A6	1	59	0.98	0	55	1
ER3_A5	1	72	0.99	1	71	0.99
ER3_A2	1	51	0.98	0	44	1
ER3_A3	2	79	0.98	1	74	0.99
ER3_A7	2	78	0.98	2	77	0.97
ER3_A4	1	76	0.99	1	65	0.98
ER3_M1	1	31	0.97	1	29	0.97
ER3_A1	0	74	1.00	0	71	1
ER3_A12	4	79	0.95	1	78	0.99
ER3_A11	2	84	0.98	0	84	1
ER3_A9	1	65	0.98	0	65	1
ER3_A8	1	78	0.99	1	76	0.99
ER3_A10	3	63	0.95	1	62	0.98
TN_A2	0	181	1.00	0	156	1
TN_A1	0	49	1.00	0	45	1
TN_M1	2	37	0.95	1	35	0.97
TN_P1D	0	24	1.00	0	21	1
TN1_P1A	0	30	1.00	0	27	1
TN1_P1B	1	29	0.97	0	28	1
TN1_P1C	0	24	1.00	0	21	1
TN1_A6	1	75	0.99	0	37	1
TN1_A5	2	33	0.94	0	31	1
TN1_A4	1	57	0.98	0	43	1
TN1_A3	1	66	0.99	0	38	1

References

1 Cancer compendium: information and trends by cancer type, Summary - Australian Institute of Health and Welfare. https://www.aihw.gov.au/reports/cancer/cancer-compendium-information-trends-by-cancer/report-contents/summary (accessed 12 Sep2018).

2 Australian Institute of Health and Welfare. Cancer in Australia 2017. Canberra: : Australian Institute of Health and Welfare 2017.

3 National cancer stage at diagnosis data | National Cancer Control Indicators. https://ncci.canceraustralia.gov.au/features/national-cancer-stage-diagnosis-data (accessed 12 Sep2018).

4 Lord SJ, Marinovich ML, Patterson JA, *et al.* Incidence of metastatic breast cancer in an Australian population-based cohort of women with non-metastatic breast cancer at diagnosis. *Med J Aust* 2012;**196**:688–92. doi:10.5694/mja12.10026

5 Senkus E, Kyriakides S, Ohno S, *et al.* Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2015;**26 Suppl 5**:v8-30. doi:10.1093/annonc/mdv298

6 Cardoso F, Senkus E, Costa A, *et al.* 4th ESO-ESMO International Consensus Guidelines for Advanced Breast Cancer (ABC 4)†. *Ann Oncol* 2018;**29**:1634–57. doi:10.1093/annonc/mdy192

7 Masuda N, Lee S-J, Ohtani S, *et al.* Adjuvant Capecitabine for Breast Cancer after Preoperative Chemotherapy. *N Engl J Med* 2017;**376**:2147–59. doi:10.1056/NEJMoa1612645

8 Tolaney SM, Barry WT, Dang CT, *et al.* Adjuvant paclitaxel and trastuzumab for node-negative, HER2-positive breast cancer. *N Engl J Med* 2015;**372**:134–41. doi:10.1056/NEJMoa1406281

9 Francis PA, Pagani O, Fleming GF, *et al.* Tailoring adjuvant endocrine therapy for premenopausal breast cancer. *N Engl J Med* 2018;**379**:122–37. doi:10.1056/NEJMoa1803164

10 Hammond MEH, Hayes DF, Dowsett M, *et al.* American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol* 2010;**28**:2784–95. doi:10.1200/JCO.2009.25.6529

11 Wolff AC, Hammond MEH, Allison KH, *et al.* Human epidermal growth factor receptor 2 testing in breast cancer: american society of clinical oncology/college of american pathologists clinical practice guideline focused update. *J Clin Oncol* 2018;**36**:2105–22. doi:10.1200/JCO.2018.77.8738

12 Sparano JA, Gray RJ, Makower DF, *et al.* Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *N Engl J Med* 2018;**379**:111–21. doi:10.1056/NEJMoa1804710

13 Cardoso F, van't Veer LJ, Bogaerts J, *et al.* 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N Engl J Med* 2016;**375**:717–29. doi:10.1056/NEJMoa1602253

14 Tutt A, Tovey H, Cheang MCU, *et al.* Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT Trial. *Nat Med* 2018;**24**:628–37. doi:10.1038/s41591-018-0009-7

15 Robson M, Im S-A, Senkus E, *et al.* Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. *N Engl J Med* Published Online First: 4 June 2017. doi:10.1056/NEJMoa1706450

16 Boveri T. *Zur Frage der Entstehung Maligner Tumoren*. Gustav Fischer 1914.

17 Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst* 1960;**25**:85–109.

18 Stehelin D, Varmus HE, Bishop JM, *et al.* DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* 1976;**260**:170–3. doi:10.1038/260170a0

19 Krontiris TG, Cooper GM. Transforming activity of human tumor DNAs. *Proc Natl Acad Sci USA* 1981;**78**:1181–4.

20 Shih C, Padhy LC, Murray M, *et al.* Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* 1981;**290**:261–4.

21 Reddy EP, Reynolds RK, Santos E, *et al.* A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* 1982;**300**:149–52. doi:10.1038/300149a0

22 Tabin CJ, Bradley SM, Bargmann CI, *et al.* Mechanism of activation of a human oncogene. *Nature* 1982;**300**:143–9. doi:10.1038/300143a0

23 Land H, Parada LF, Weinberg RA. Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. *Nature* 1983;**304**:596–602. doi:10.1038/304596a0

24 Levan A, Hauschka TS. Endomitotic reduplication mechanisms in ascites tumors of the mouse. *J Natl Cancer Inst* 1953;**14**:1–43.

25 Klein G, Klein E. Conversion of solid neoplasms into ascites tumors. *Ann N Y*

Acad Sci 1956;**63**:640–61. doi:10.1111/j.1749-6632.1956.tb50883.x

26 Makino S. Further evidence favoring the concept of the stem cell in ascites tumors of rats. *Ann N Y Acad Sci* 1956;**63**:818–30. doi:10.1111/j.1749-6632.1956.tb50894.x

27 Law LW. Origin of the resistance of leukæmic cells to folic acid antagonists. *Nature* 1952;**169**:628–9. doi:10.1038/169628a0

28 Yamada K, Takagi N, Sandberg AA. Chromosomes and causation of human cancer and leukemia. II. Karyotypes of human solid tumors. *Cancer* 1966;**19**:1879–90. doi:10.1002/1097-0142(196612)19:12<1879::AID-CNCR2820191215>3.0.CO;2-Z

29 Gray JM, Pierce GB. Relationship between growth rate and differentiation of melanoma in vivo2. *JNCI: Journal of the National Cancer Institute* 1964;**32**:1201–11. doi:10.1093/jnci/32.6.1201

30 Fidler IJ, Kripke ML. Metastasis results from preexisting variant cells within a malignant tumor. *Science* 1977;**197**:893–5.

31 Fialkow PJ. The origin and development of human tumors studied with cell markers. *N Engl J Med* 1974;**291**:26–35. doi:10.1056/NEJM197407042910109

32 Klein E, Klein G, Nadkarni JS, *et al.* Surface IgM-kappa specificity on a Burkitt lymphoma cell in vivo and in derived culture lines. *Cancer Res* 1968;**28**:1300–10.

33 Linder D, Gartler SM. Glucose-6-phosphate dehydrogenase mosaicism: utilization as a cell marker in the study of leiomyomas. *Science* 1965;**150**:67–9.

34 Busque L, Mio R, Mattioli J, *et al.* Nonrandom X-inactivation patterns in

normal females: lyonization ratios vary with age. *Blood* 1996;**88**:59–65.

35 Welch JS, Ley TJ, Link DC, *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* 2012;**150**:264–78. doi:10.1016/j.cell.2012.06.023

36 Jaiswal S, Fontanillas P, Flannick J, *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* 2014;**371**:2488–98. doi:10.1056/NEJMoa1408617

37 Martincorena I, Roshan A, Gerstung M, *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 2015;**348**:880–6. doi:10.1126/science.aaa6806

38 Martincorena I, Fowler JC, Wabik A, *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* 2018;**362**:911–7. doi:10.1126/science.aau3879

39 Nowell PC. The clonal evolution of tumor cell populations. *Science* 1976;**194**:23–8. doi:10.1126/science.959840

40 Fidler IJ, Hart IR. Biological diversity in metastatic neoplasms: origins and implications. *Science* 1982;**217**:998–1003.

41 Gerbault-Seureau M, Vielh P, Zafrani B, *et al.* Cytogenetic study of twelve human near-diploid breast cancers with chromosomal changes. *Ann Genet* 1987;**30**:138–45.

42 Dutrillaux B, Gerbault-Seureau M, Zafrani B. Characterization of chromosomal anomalies in human breast cancer. A comparison of 30 paradiploid cases with few chromosome changes. *Cancer Genet Cytogenet* 1990;**49**:203–17.

43 Gebhart E, Brüderlein S, Tulusan AH, *et al.* Incidence of double minutes,

cytogenetic equivalents of gene amplification, in human carcinoma cells. *Int J Cancer* 1984;**34**:369–73. doi:10.1002/ijc.2910340313

44 Slamon DJ, Clark GM, Wong SG, *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 1987;**235**:177–82. doi:10.1126/science.3798106

45 Kallioniemi A, Kallioniemi OP, Sudar D, *et al.* Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 1992;**258**:818–21.

46 Kallioniemi A, Kallioniemi OP, Piper J, *et al.* Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc Natl Acad Sci USA* 1994;**91**:2156–60.

47 Ried T, Just KE, Holtgreve-Grez H, *et al.* Comparative genomic hybridization of formalin-fixed, paraffin-embedded breast tumors reveals different patterns of chromosomal gains and losses in fibroadenomas and diploid and aneuploid carcinomas. *Cancer Res* 1995;**55**:5415–23.

48 Tirkkonen M, Tanner M, Karhu R, *et al.* Molecular cytogenetics of primary breast cancer by CGH. *Genes Chromosomes Cancer* 1998;**21**:177–84. doi:10.1002/(SICI)1098-2264(199803)21:3<177::AID-GCC1>3.0.CO;2-X

49 Albertson DG, Ylstra B, Segraves R, *et al.* Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat Genet* 2000;**25**:144–6. doi:10.1038/75985

50 Pollack JR, Sørlie T, Perou CM, *et al.* Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA* 2002;**99**:12963–8. doi:10.1073/pnas.162471999

51 Bose S, Wang SI, Terry MB, *et al.* Allelic loss of chromosome 10q23 is associated with tumor progression in breast carcinomas. *Oncogene* 1998;**17**:123–7. doi:10.1038/sj.onc.1201940

52 Garcia JM, Silva JM, Dominguez G, *et al.* Allelic loss of the PTEN region (10q23) in breast carcinomas of poor pathophenotype. *Breast Cancer Res Treat* 1999;**57**:237–43. doi:10.1023/A:1006273516976

53 Osborne RJ, Hamshere MG. A genome-wide map showing common regions of loss of heterozygosity/allelic imbalance in breast cancer. *Cancer Res* 2000;**60**:3706–12.

54 Van Loo P, Nordgard SH, Lingjærde OC, *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* 2010;**107**.

55 Ching HC, Naidu R, Seong MK, *et al.* Integrated analysis of copy number and loss of heterozygosity in primary breast carcinomas using high-density SNP array. *Int J Oncol* 2011;**39**:621–33. doi:10.3892/ijo.2011.1081

56 Fang M, Toher J, Morgan M, *et al.* Genomic differences between estrogen receptor (ER)-positive and ER-negative human breast carcinoma identified by single nucleotide polymorphism array comparative genome hybridization analysis. *Cancer* 2011;**117**:2024–34. doi:10.1002/cncr.25770

57 Hicks J, Krasnitz A, Lakshmi B, *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* 2006;**16**:1465–79. doi:10.1101/gr.5460106

58 Futreal PA, Coin L, Marshall M, *et al.* A census of human cancer genes. *Nat Rev Cancer* 2004;**4**:177–83. doi:10.1038/nrc1299

59 Bailey MH, Tokheim C, Porta-Pardo E, *et al.* Comprehensive characterization of cancer driver genes and mutations. *Cell* 2018;**173**:371–

385.e18. doi:10.1016/j.cell.2018.02.060

60 International Cancer Genome Consortium, Hudson TJ, Anderson W, *et al.* International network of cancer genome projects. *Nature* 2010;**464**:993–8. doi:10.1038/nature08987

61 Sjöblom T, Jones S, Wood LD, *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* 2006;**314**:268–74. doi:10.1126/science.1133427

62 Wood LD, Parsons DW, Jones S, *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* 2007;**318**:1108–13. doi:10.1126/science.1145720

63 Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74. doi:10.1016/j.cell.2011.02.013

64 Greenman C, Stephens P, Smith R, *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* 2007;**446**:153–8. doi:10.1038/nature05610

65 Greenman C, Wooster R, Futreal PA, *et al.* Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* 2006;**173**:2187–98. doi:10.1534/genetics.105.044677

66 Nik-Zainal S, Van Loo P, Wedge DC, *et al.* The life history of 21 breast cancers. *Cell* 2012;**149**:994–1007. doi:10.1016/j.cell.2012.04.023

67 Curtis C, Shah SP, Chin S-F, *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012;**486**:346–52. doi:10.1038/nature10983

68 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70.

doi:10.1038/nature11412

69 Shah SP, Roth A, Goya R, *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 2012;**486**:395–9. doi:10.1038/nature10933

70 Banerji S, Cibulskis K, Rangel-Escareno C, *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* 2012;**486**:405–9. doi:10.1038/nature11154

71 Stephens PJ, Tarpey PS, Davies H, *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012;**486**:400–4. doi:10.1038/nature11017

72 Ellis MJ, Ding L, Shen D, *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* 2012;**486**:353–60. doi:10.1038/nature11143

73 Garraway LA, Lander ES. Lessons from the cancer genome. *Cell* 2013;**153**:17–37. doi:10.1016/j.cell.2013.03.002

74 Van Allen EM, Wagle N, Stojanov P, *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014;**20**:682–8. doi:10.1038/nm.3559

75 Stephens PJ, McBride DJ, Lin M-L, *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 2009;**462**:1005–10. doi:10.1038/nature08645

76 Stephens PJ, Greenman CD, Fu B, *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011;**144**:27–40. doi:10.1016/j.cell.2010.11.055

77 Nik-Zainal S, Davies H, Staaf J, *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;**534**:47–54. doi:10.1038/nature17676

78 Glodzik D, Morganella S, Davies H, *et al.* A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat Genet* 2017;**49**:341–8. doi:10.1038/ng.3771

79 Zhang J. Evolution by gene duplication: an update. *Trends Ecol Evol (Amst)* 2003;**18**:292–8. doi:10.1016/S0169-5347(03)00033-8

80 Menghi F, Barthel FP, Yadav V, *et al.* The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell* 2018;**34**:197–210.e5. doi:10.1016/j.ccell.2018.06.008

81 Menghi F, Inaki K, Woo X, *et al.* The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc Natl Acad Sci USA* 2016;**113**:E2373-82. doi:10.1073/pnas.1520010113

82 Willis NA, Frock RL, Menghi F, *et al.* Mechanism of tandem duplication formation in BRCA1-mutant cells. *Nature* 2017;**551**:590–5. doi:10.1038/nature24477

83 Lawrence MS, Stojanov P, Mermel CH, *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014;**505**:495–501. doi:10.1038/nature12912

84 Davoli T, Xu AW, Mengwasser KE, *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* 2013;**155**:948–62. doi:10.1016/j.cell.2013.10.011

85 Martincorena I, Raine KM, Gerstung M, *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* 2017;**171**:1029–1041.e21. doi:10.1016/j.cell.2017.09.042

86 Beroukhim R, Getz G, Nghiemphu L, *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci USA* 2007;**104**:20007–12. doi:10.1073/pnas.0710052104

87 Santarius T, Shipley J, Brewer D, *et al.* A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* 2010;**10**:59–64. doi:10.1038/nrc2771

88 Zack TI, Schumacher SE, Carter SL, *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 2013;**45**:1134–40. doi:10.1038/ng.2760

89 Sanchez-Garcia F, Villagrasa P, Matsui J, *et al.* Integration of genomic data enables selective discovery of breast cancer drivers. *Cell* 2014;**159**:1461–75. doi:10.1016/j.cell.2014.10.048

90 Sack LM, Davoli T, Li MZ, *et al.* Profound tissue specificity in proliferation control underlies cancer drivers and aneuploidy patterns. *Cell* 2018;**173**:499–514.e23. doi:10.1016/j.cell.2018.02.037

91 Reiter JG, Makohon-Moore AP, Gerold JM, *et al.* Minimal functional driver gene heterogeneity among untreated metastases. *Science* 2018;**361**:1033–7. doi:10.1126/science.aat7171

92 Toy W, Shen Y, Won H, *et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nat Genet* 2013;**45**:1439–45. doi:10.1038/ng.2822

93 Robinson DR, Wu Y-M, Vats P, *et al.* Activating ESR1 mutations in hormone-

resistant metastatic breast cancer. *Nat Genet* 2013;**45**:1446–51. doi:10.1038/ng.2823

94 Li S, Shen D, Shao J, *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep* 2013;**4**:1116–30. doi:10.1016/j.celrep.2013.08.022

95 Kuukasjärvi T, Karhu R, Tanner M, *et al.* Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res* 1997;**57**:1597–604.

96 Nishizaki T, DeVries S, Chew K, *et al.* Genetic alterations in primary breast cancers and their metastases: direct comparison using modified comparative genomic hybridization. *Genes Chromosomes Cancer* 1997;**19**:267–72. doi:10.1002/(SICI)1098-2264(199708)19:4<267::AID-GCC9>3.0.CO;2-V

97 Torres L, Ribeiro FR, Pandis N, *et al.* Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Res Treat* 2007;**102**:143–55. doi:10.1007/s10549-006-9317-6

98 Friedrich K, Weber T, Scheithauer J, *et al.* Chromosomal genotype in breast cancer progression: comparison of primary and secondary manifestations. *Cell Oncol* 2008;**30**:39–50.

99 Li J, Gromov P, Gromova I, *et al.* Omics-based profiling of carcinoma of the breast and matched regional lymph node metastasis. *Proteomics* 2008;**8**:5038–52. doi:10.1002/pmic.200800303

100 Santos SCL, Cavalli IJ, Ribeiro EM, *et al.* Patterns of DNA copy number changes in sentinel lymph node breast cancer metastases. *Cytogenet*

Genome Res 2008;**122**:16–21. doi:10.1159/000151311

101 Popławski AB, Jankowski M, Erickson SW, *et al.* Frequent genetic differences between matched primary and metastatic breast cancer provide an approach to identification of biomarkers for disease progression. *Eur J Hum Genet* 2010;**18**:560–8. doi:10.1038/ejhg.2009.230

102 Desouki MM, Liao S, Huang H, *et al.* Identification of metastasis-associated breast cancer genes using a high-resolution whole genome profiling approach. *J Cancer Res Clin Oncol* 2011;**137**:795–809. doi:10.1007/s00432-010-0937-1

103 Shah SP, Morin RD, Khattra J, *et al.* Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 2009;**461**:809–13. doi:10.1038/nature08489

104 Ding L, Ellis MJ, Li S, *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 2010;**464**:999–1005. doi:10.1038/nature08989

105 Gerlinger M, Rowan AJ, Horswell S, *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;**366**:883–92. doi:10.1056/NEJMoa1113205

106 Yates LR, Gerstung M, Knappskog S, *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* 2015;**21**:751–9. doi:10.1038/nm.3886

107 Yates LR, Knappskog S, Wedge D, *et al.* Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* 2017;**32**:169–184.e7. doi:10.1016/j.ccell.2017.07.005

108 Lefebvre C, Bachelot T, Filleron T, *et al.* Mutational profile of metastatic

breast cancers: A retrospective analysis. *PLoS Med* 2016;**13**:e1002201. doi:10.1371/journal.pmed.1002201

109 Navin N, Kendall J, Troge J, *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**:90–4. doi:10.1038/nature09807

110 Wang Y, Waters J, Leung ML, *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 2014;**512**:155–60. doi:10.1038/nature13600

111 Gao R, Davis A, McDonald TO, *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* 2016;**48**:1119–30. doi:10.1038/ng.3641

112 Kim C, Gao R, Sei E, *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* 2018;**173**:879–893.e13. doi:10.1016/j.cell.2018.03.041

113 Demeulemeester J, Kumar P, Møller EK, *et al.* Tracing the origin of disseminated tumor cells in breast cancer using single-cell sequencing. *Genome Biol* 2016;**17**:250. doi:10.1186/s13059-016-1109-7

114 Eirew P, Steif A, Khattra J, *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* 2015;**518**:422–6. doi:10.1038/nature13952

115 Dupont Jensen J, Laenkholm A-V, Knoop A, *et al.* PIK3CA mutations may be discordant between primary and corresponding metastatic disease in breast cancer. *Clin Cancer Res* 2011;**17**:667–77. doi:10.1158/1078-0432.CCR-10-1133

116 Schleifman EB, Desai R, Spoerke JM, *et al.* Targeted biomarker profiling of matched primary and metastatic estrogen receptor positive breast cancers.

PLoS One 2014;**9**:e88401. doi:10.1371/journal.pone.0088401

117 Meric-Bernstam F, Frampton GM, Ferrer-Lozano J, *et al.* Concordance of genomic alterations between primary and recurrent breast cancer. *Mol Cancer Ther* 2014;**13**:1382–9. doi:10.1158/1535-7163.MCT-13-0482

118 Fumagalli D, Wilson TR, Salgado R, *et al.* Somatic mutation, copy number and transcriptomic profiles of primary and matched metastatic estrogen receptor-positive breast cancers. *Ann Oncol* 2016;**27**:1860–6. doi:10.1093/annonc/mdw286

119 Daley GQ, Ben-Neriah Y. Implicating the bcr/abl Gene in the Pathogenesis of Philadelphia Chromosome-Positive Human Leukemia. Elsevier 1991. 151–84. doi:10.1016/S0065-230X(08)60998-7

120 Druker BJ, Tamura S, Buchdunger E, *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* 1996;**2**:561–6. doi:10.1038/nm0596-561

121 Druker BJ, Talpaz M, Resta DJ, *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med* 2001;**344**:1031–7. doi:10.1056/NEJM200104053441401

122 Cheng DT, Mitchell TN, Zehir A, *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* 2015;**17**:251–64. doi:10.1016/j.jmoldx.2014.12.006

123 Frampton GM, Fichtenholtz A, Otto GA, *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* 2013;**31**:1023–31. doi:10.1038/nbt.2696

124 Shigemizu D, Momozawa Y, Abe T, *et al.* Performance comparison of four commercial human whole-exome capture platforms. *Sci Rep* 2015;**5**:12742. doi:10.1038/srep12742

125 Zehir A, Benayed R, Shah RH, *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* 2017;**23**:703–13. doi:10.1038/nm.4333

126 Rheinbay E, Parasuraman P, Grimsby J, *et al.* Recurrent and functional regulatory mutations in breast cancer. *Nature* 2017;**547**:55–60. doi:10.1038/nature22992

127 Alexandrov LB, Nik-Zainal S, Wedge DC, *et al.* Signatures of mutational processes in human cancer. *Nature* 2013;**500**:415–21. doi:10.1038/nature12477

128 Macintyre G, Goranova TE, De Silva D, *et al.* Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet* 2018;**50**:1262–70. doi:10.1038/s41588-018-0179-8

129 Wu Y-M, Cieślik M, Lonigro RJ, *et al.* Inactivation of CDK12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell* 2018;**173**:1770–1782.e14. doi:10.1016/j.cell.2018.04.034

130 Belkadi A, Bolze A, Itan Y, *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 2015;**112**:5473–8. doi:10.1073/pnas.1418631112

131 Berger AC, Korkut A, Kanchi RS, *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* 2018;**33**:690–705.e9. doi:10.1016/j.ccell.2018.03.014

132 Costello JC, Heiser LM, Georgii E, *et al.* A community effort to assess and

improve drug sensitivity prediction algorithms. *Nat Biotechnol* 2014;**32**:1202–12. doi:10.1038/nbt.2877

133 Iorio F, Knijnenburg TA, Vis DJ, *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**:740–54. doi:10.1016/j.cell.2016.06.017

134 Tothill RW, Kowalczyk A, Rischin D, *et al.* An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res* 2005;**65**:4031–40. doi:10.1158/0008-5472.CAN-04-3617

135 Dietrich S, Oleś M, Lu J, *et al.* Drug-perturbation-based stratification of blood cancer. *J Clin Invest* 2018;**128**:427–45. doi:10.1172/JCI93801

136 Andre F, Mardis E, Salm M, *et al.* Prioritizing targets for precision cancer medicine. *Ann Oncol* 2014;**25**:2295–303. doi:10.1093/annonc/mdu478

137 Meric-Bernstam F, Johnson A, Holla V, *et al.* A decision support framework for genomically informed investigational cancer therapy. *J Natl Cancer Inst* 2015;**107**. doi:10.1093/jnci/djv098

138 Chakravarty D, Gao J, Phillips SM, *et al.* Oncokb: A precision oncology knowledge base. *JCO Precis Oncol* 2017;**2017**. doi:10.1200/PO.17.00011

139 Griffith M, Spies NC, Krysiak K, *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 2017;**49**:170–4. doi:10.1038/ng.3774

140 Li MM, Datto M, Duncavage EJ, *et al.* Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the association for molecular pathology, american society of clinical oncology, and college of american pathologists.

J Mol Diagn 2017;**19**:4–23. doi:10.1016/j.jmoldx.2016.10.002

141 Mateo J, Chakravarty D, Dienstmann R, *et al.* A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). *Ann Oncol* 2018;**29**:1895–902. doi:10.1093/annonc/mdy263

142 Foundation Medicine. Genomic Testing | Foundation Medicine. FoundationOne CDx. 2019.https://www.foundationmedicine.com/genomic-testing (accessed 27 Sep2019).

143 Slamon DJ, Leyland-Jones B, Shak S, *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med* 2001;**344**:783–92. doi:10.1056/NEJM200103153441101

144 Shaw AT, Kim D-W, Nakagawa K, *et al.* Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N Engl J Med* 2013;**368**:2385–94. doi:10.1056/NEJMoa1214886

145 Flaherty KT, Infante JR, Daud A, *et al.* Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N Engl J Med* 2012;**367**:1694–703. doi:10.1056/NEJMoa1210093

146 Schwaederle M, Zhao M, Lee JJ, *et al.* Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. *J Clin Oncol* 2015;**33**:3817–25. doi:10.1200/JCO.2015.61.5997

147 Jardim DL, Schwaederle M, Wei C, *et al.* Impact of a Biomarker-Based Strategy on Oncology Drug Development: A Meta-analysis of Clinical Trials Leading to FDA Approval. *J Natl Cancer Inst* 2015;**107**.

doi:10.1093/jnci/djv253

148 Lynch TJ, Bell DW, Sordella R, *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med* 2004;**350**:2129–39. doi:10.1056/NEJMoa040938

149 Hyman DM, Puzanov I, Subbiah V, *et al.* Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. *N Engl J Med* 2015;**373**:726–36. doi:10.1056/NEJMoa1502309

150 Kopetz S, Desai J, Chan E, *et al.* Phase II Pilot Study of Vemurafenib in Patients With Metastatic BRAF-Mutated Colorectal Cancer. *J Clin Oncol* 2015;**33**:4032–8. doi:10.1200/JCO.2015.63.2497

151 Kopetz S, McDonough SL, Morris VK, *et al.* Randomized trial of irinotecan and cetuximab with or without vemurafenib in BRAF-mutant metastatic colorectal cancer (SWOG 1406). *J Clin Oncol* 2017;**35**:520–520.

152 Yu HA, Arcila ME, Rekhtman N, *et al.* Analysis of tumor specimens at the time of acquired resistance to EGFR-TKI therapy in 155 patients with EGFR-mutant lung cancers. *Clin Cancer Res* 2013;**19**:2240–7. doi:10.1158/1078-0432.CCR-12-2246

153 Edwards SL, Brough R, Lord CJ, *et al.* Resistance to therapy caused by intragenic deletion in BRCA2. *Nature* 2008;**451**:1111–5. doi:10.1038/nature06548

154 Miao D, Margolis CA, Gao W, *et al.* Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* 2018;**359**:801–6. doi:10.1126/science.aan5951

155 O'Rawe J, Jiang T, Sun G, *et al.* Low concordance of multiple variant-calling

pipelines: practical implications for exome and genome sequencing. *Genome Med* 2013;**5**:28. doi:10.1186/gm432

156 Clark MJ, Chen R, Lam HYK, *et al.* Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011;**29**:908–14. doi:10.1038/nbt.1975

157 Asan, Xu Y, Jiang H, *et al.* Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* 2011;**12**:R95. doi:10.1186/gb-2011-12-9-r95

158 Chilamakuri CSR, Lorenz S, Madoui M-A, *et al.* Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 2014;**15**:449. doi:10.1186/1471-2164-15-449

159 Von Hoff DD, Stephenson JJ, Rosen P, *et al.* Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. *J Clin Oncol* 2010;**28**:4877–83. doi:10.1200/JCO.2009.26.5983

160 Le Tourneau C, Delord J-P, Gonçalves A, *et al.* Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol* 2015;**16**:1324–34. doi:10.1016/S1470-2045(15)00188-6

161 Massard C, Michiels S, Ferté C, *et al.* High-Throughput Genomics and Clinical Outcome in Hard-to-Treat Advanced Cancers: Results of the MOSCATO 01 Trial. *Cancer Discov* 2017;**7**:586–95. doi:10.1158/2159-8290.CD-16-1396

162 Coyne GO, Takebe N, Chen AP. Defining precision: The precision medicine

initiative trials NCI-MPACT and NCI-MATCH. *Curr Probl Cancer* 2017;**41**:182–93. doi:10.1016/j.currproblcancer.2017.02.001

163 Mangat PK, Halabi S, Bruinooge SS, *et al.* Rationale and design of the targeted agent and profiling utilization registry study. *JCO Precis Oncol* 2018;:1–14. doi:10.1200/PO.18.00122

164 Conley BA, Gray R, Chen A, *et al.* Abstract CT101: NCI-molecular analysis for therapy choice (NCI-MATCH) clinical trial: interim analysis: *Cancer Res* 2016;**76**:CT101-CT101. doi:10.1158/1538-7445.AM2016-CT101

165 Arnedos M, Scott V, Job B, *et al.* Array CGH and PIK3CA/AKT1 mutations to drive patients to specific targeted agents: a clinical experience in 108 patients with metastatic breast cancer. *Eur J Cancer* 2012;**48**:2293–9. doi:10.1016/j.ejca.2012.06.014

166 André F, Bachelot T, Commo F, *et al.* Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). *Lancet Oncol* 2014;**15**:267–74. doi:10.1016/S1470-2045(13)70611-9

167 Clark AS, Hong F, Finn RS, *et al.* Abstract LB-010: Molecular analysis for therapy choice (NCI-MATCH, EAY131) arm Z1B: Phase II trial of palbociclib for *CCND1, 2 or 3* amplified tumors. In: *Clinical Research (Excluding Clinical Trials)*. American Association for Cancer Research 2019. LB-010-LB-010. doi:10.1158/1538-7445.AM2019-LB-010

168 Maetens M, Brown D, Irrthum A, *et al.* The AURORA pilot study for molecular screening of patients with advanced breast cancer-a study of the breast international group. *NPJ Breast Cancer* 2017;**3**:23. doi:10.1038/s41523-017-0026-6

169 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. doi:10.1093/bioinformatics/btp324

170 Broad Institute. *Picard toolkit*. Broad Institute 2019.

171 Balakirev ES, Ayala FJ. Pseudogenes: are they "junk" or functional DNA? *Annu Rev Genet* 2003;**37**:123–51. doi:10.1146/annurev.genet.37.040103.103949

172 Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. doi:10.1093/bioinformatics/btp352

173 Lai Z, Markovets A, Ahdesmaki M, *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 2016;**44**:e108. doi:10.1093/nar/gkw227

174 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164. doi:10.1093/nar/gkq603

175 Siva N. 1000 Genomes project. *Nat Biotechnol* 2008;**26**:256. doi:10.1038/nbt0308-256b

176 NHLBI GO Exome Sequencing Project (ESP). Exome Variant Server. http://evs.gs.washington.edu/EVS/ (accessed 18 Oct2016).

177 Karczewski KJ, Francioli LC, Tiao G, *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* Published Online First: 28 January 2019. doi:10.1101/531210

178 Sherry ST, Ward MH, Kholodov M, *et al.* dbSNP: the NCBI database of

genetic variation. *Nucleic Acids Res* 2001;**29**:308–11. doi:10.1093/nar/29.1.308

179 Forbes SA, Beare D, Gunasekaran P, *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015;**43**:D805-11. doi:10.1093/nar/gku1075

180 Landrum MJ, Lee JM, Benson M, *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**:D1062–7. doi:10.1093/nar/gkx1153

181 Kent WJ, Sugnet CW, Furey TS, *et al.* The human genome browser at UCSC. *Genome Res* 2002;**12**:996–1006. doi:10.1101/gr.229102

182 Kuilman T, Velds A, Kemper K, *et al.* CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol* 2015;**16**:49. doi:10.1186/s13059-015-0617-1

183 Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 2007;**23**:657–63. doi:10.1093/bioinformatics/btl646

184 Li J, Lupat R, Amarasinghe KC, *et al.* CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012;**28**:1307–13. doi:10.1093/bioinformatics/bts146

185 Cerami E, Gao J, Dogrusoz U, *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**:401–4. doi:10.1158/2159-8290.CD-12-0095

186 Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;**6**:pl1. doi:10.1126/scisignal.2004088

187 Miller ML, Reznik E, Gauthier NP, *et al.* Pan-Cancer Analysis of Mutation Hotspots in Protein Domains. *Cell Syst* 2015;**1**:197–209. doi:10.1016/j.cels.2015.08.014

188 Bose R, Kavuri SM, Searleman AC, *et al.* Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discov* 2013;**3**:224–37. doi:10.1158/2159-8290.CD-12-0349

189 Berger AH, Brooks AN, Wu X, *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* 2016;**30**:214–28. doi:10.1016/j.ccell.2016.06.022

190 Findlay GM, Daza RM, Martin B, *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 2018;**562**:217–22. doi:10.1038/s41586-018-0461-z

191 Gao H, Korn JM, Ferretti S, *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat Med* 2015;**21**:1318–25. doi:10.1038/nm.3954

192 Bruna A, Rueda OM, Greenwood W, *et al.* A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds. *Cell* 2016;**167**:260–274.e22. doi:10.1016/j.cell.2016.08.041

193 Garnett MJ, Edelman EJ, Heidorn SJ, *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;**483**:570–5. doi:10.1038/nature11005

194 Barretina J, Caponigro G, Stransky N, *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7. doi:10.1038/nature11003

195 R Core Team. *R: A Language and Environment for Statistical Computing*.

Vienna, Austria: : R Foundation for Statistical Computing 2019.

196 Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York 2016.

197 Wickham H, François R, Henry L, *et al. dplyr: A Grammar of Data Manipulation*. CRAN 2019.

198 Wickham H, Henry L. *tidyr: Tidy Messy Data*. CRAN 2019.

199 Milton S, Wickham H. *magrittr: A Forward-Pipe Operator for R*. CRAN 2014.

200 Grolemund G, Wickham H. Dates and Times Made Easy with lubridate. *J Stat Softw* 2011;**40**. doi:10.18637/jss.v040.i03

201 Sing T, Sander O, Beerenwinkel N, *et al.* ROCR: visualizing classifier performance in R. *Bioinformatics* 2005;**21**:3940–1. doi:10.1093/bioinformatics/bti623

202 Wickham H. *readr: Read Rectangular Text Data*. CRAN 2018.

203 Wickham H. *stringr: Simple, Consistent Wrappers for Common String Operations*. CRAN 2019.

204 Lawrence M, Huber W, Pagès H, *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* 2013;**9**:e1003118. doi:10.1371/journal.pcbi.1003118

205 Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* 2019;**35**:3151–3. doi:10.1093/bioinformatics/btz031

206 Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 2009;**25**:1841–2. doi:10.1093/bioinformatics/btp328

207 Wagle N, Berger MF, Davis MJ, *et al.* High-throughput detection of

actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov* 2012;**2**:82–93. doi:10.1158/2159-8290.CD-11-0184

208 Pääbo S, Irwin DM, Wilson AC. DNA damage promotes jumping between templates during enzymatic amplification. *J Biol Chem* 1990;**265**:4718–21.

209 Wong C, DiCioccio RA, Allen HJ, *et al.* Mutations in BRCA1 from fixed, paraffin-embedded tissue can be artifacts of preservation. *Cancer Genet Cytogenet* 1998;**107**:21–7. doi:10.1016/s0165-4608(98)00079-x

210 Williams C, Pontén F, Moberg C, *et al.* A high frequency of sequence alterations is due to formalin fixation of archival specimens. *Am J Pathol* 1999;**155**:1467–71. doi:10.1016/S0002-9440(10)65461-2

211 Masuda N, Ohnishi T, Kawamoto S, *et al.* Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples. *Nucleic Acids Res* 1999;**27**:4436–43. doi:10.1093/nar/27.22.4436

212 Quach N, Goodman MF, Shibata D. In vitro mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. *BMC Clin Pathol* 2004;**4**:1. doi:10.1186/1472-6890-4-1

213 Rothberg JM, Hinz W, Rearick TM, *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011;**475**:348–52. doi:10.1038/nature10242

214 Beadling C, Neff TL, Heinrich MC, *et al.* Combining highly multiplexed PCR with semiconductor-based sequencing for rapid cancer genotyping. *J Mol Diagn* 2013;**15**:171–6. doi:10.1016/j.jmoldx.2012.09.003

215 Wang DG, Fan JB, Siao CJ, *et al.* Large-scale identification, mapping, and

genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;**280**:1077–82. doi:10.1126/science.280.5366.1077

216 Mamanova L, Coffey AJ, Scott CE, *et al.* Target-enrichment strategies for next-generation sequencing. *Nat Methods* 2010;**7**:111–8. doi:10.1038/nmeth.1419

217 Dahl F, Gullberg M, Stenberg J, *et al.* Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res* 2005;**33**:e71. doi:10.1093/nar/gni070

218 Johansson H, Isaksson M, Sörqvist EF, *et al.* Targeted resequencing of candidate genes using selector probes. *Nucleic Acids Res* 2011;**39**:e8. doi:10.1093/nar/gkq1005

219 Moens LNJ, Falk-Sörqvist E, Ljungström V, *et al.* HaloPlex Targeted Resequencing for Mutation Detection in Clinical Formalin-Fixed, Paraffin-Embedded Tumor Samples. *J Mol Diagn* 2015;**17**:729–39. doi:10.1016/j.jmoldx.2015.06.009

220 Gnirke A, Melnikov A, Maguire J, *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 2009;**27**:182–9. doi:10.1038/nbt.1523

221 Wong SQ, Li J, Salemi R, *et al.* Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours. *Sci Rep* 2013;**3**:3494. doi:10.1038/srep03494

222 Ashutosh, Devendra J, Le Cocq C, *et al.* Detection of common and low frequency variants in cancer samples with SNPPET using NGS target enrichment data. The American Society of Human Genetics 2014.

223 Zook JM, Chapman B, Wang J, *et al.* Integrating human sequence data sets

provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 2014;**32**:246–51. doi:10.1038/nbt.2835

224 Cibulskis K, Lawrence MS, Carter SL, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;**31**:213–9. doi:10.1038/nbt.2514

225 Koboldt DC, Zhang Q, Larson DE, *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;**22**:568–76. doi:10.1101/gr.129684.111

226 Wang Q, Jia P, Li F, *et al.* Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med* 2013;**5**:91. doi:10.1186/gm495

227 Xu H, DiCarlo J, Satya RV, *et al.* Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* 2014;**15**:244. doi:10.1186/1471-2164-15-244

228 Ye K, Schulz MH, Long Q, *et al.* Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;**25**:2865–71. doi:10.1093/bioinformatics/btp394

229 Lai Z, Markovets A, Ahdesmaki M, *et al.* Abstract 4864: VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Cancer Res* 2015;**75**:4864–4864. doi:10.1158/1538-7445.AM2015-4864

230 Fang LT, Afshar PT, Chhibber A, *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol* 2015;**16**:197. doi:10.1186/s13059-015-0758-2

231 Do H, Dobrovic A. Sequence artifacts in DNA from formalin-fixed tissues:

causes and strategies for minimization. *Clin Chem* 2015;**61**:64–71. doi:10.1373/clinchem.2014.223040

232 Beck TF, Mullikin JC, NISC Comparative Sequencing Program, *et al.* Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. *Clin Chem* 2016;**62**:647–54. doi:10.1373/clinchem.2015.249623

233 Riester M, Singh AP, Brannon AR, *et al.* PureCN: copy number calling and SNV classification using targeted short read sequencing. *Source Code Biol Med* 2016;**11**:13. doi:10.1186/s13029-016-0060-z

234 Le DT, Uram JN, Wang H, *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* 2015;**372**:2509–20. doi:10.1056/NEJMoa1500596

235 Rizvi NA, Hellmann MD, Snyder A, *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 2015;**348**:124–8. doi:10.1126/science.aaa1348

236 Goodman AM, Kato S, Bazhenova L, *et al.* Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol Cancer Ther* 2017;**16**:2598–608. doi:10.1158/1535-7163.MCT-17-0386

237 Van Allen EM, Miao D, Schilling B, *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 2015;**350**:207–11. doi:10.1126/science.aad0095

238 André F, Ciruelos E, Rubovszky G, *et al.* Alpelisib for PIK3CA-Mutated, Hormone Receptor-Positive Advanced Breast Cancer. *N Engl J Med* 2019;**380**:1929–40. doi:10.1056/NEJMoa1813904

239 Cocco E, Javier Carmona F, Razavi P, *et al.* Neratinib is effective in breast

tumors bearing both amplification and mutation of ERBB2 (HER2). *Sci Signal* 2018;**11**. doi:10.1126/scisignal.aat9773

240 Hyman DM, Piha-Paul SA, Won H, *et al.* HER kinase inhibition in patients with HER2- and HER3-mutant cancers. *Nature* 2018;**554**:189–94. doi:10.1038/nature25475

241 Condorelli R, Spring L, O'Shaughnessy J, *et al.* Polyclonal RB1 mutations and acquired resistance to CDK 4/6 inhibitors in patients with metastatic breast cancer. *Ann Oncol* 2018;**29**:640–5. doi:10.1093/annonc/mdx784

242 Juric D, Castel P, Griffith M, *et al.* Convergent loss of PTEN leads to clinical resistance to a PI(3)Kα inhibitor. *Nature* 2015;**518**:240–4. doi:10.1038/nature13948

243 Fribbens C, O'Leary B, Kilburn L, *et al.* Plasma ESR1 Mutations and the Treatment of Estrogen Receptor-Positive Advanced Breast Cancer. *J Clin Oncol* 2016;**34**:2961–8. doi:10.1200/JCO.2016.67.3061

244 Hyman DM, Smyth LM, Donoghue MTA, *et al.* AKT inhibition in solid tumors with AKT1 mutations. *J Clin Oncol* 2017;**35**:2251–9. doi:10.1200/JCO.2017.73.0143

245 Kuykendall JR, Bogdanffy MS. Efficiency of DNA-histone crosslinking induced by saturated and unsaturated aldehydes in vitro. *Mutat Res* 1992;**283**:131–6. doi:10.1016/0165-7992(92)90145-8

246 Dietrich D, Uhl B, Sailer V, *et al.* Improved PCR performance using template DNA from formalin-fixed and paraffin-embedded tissues by overcoming PCR inhibition. *PLoS One* 2013;**8**:e77771. doi:10.1371/journal.pone.0077771

247 Razavi P, Chang MT, Xu G, *et al.* The Genomic Landscape of Endocrine-

Resistant Advanced Breast Cancers. *Cancer Cell* 2018;**34**:427–438.e6. doi:10.1016/j.ccell.2018.08.008

248 Nayar U, Cohen O, Kapstad C, *et al.* Acquired HER2 mutations in ER+ metastatic breast cancer confer resistance to estrogen receptor-directed therapies. *Nat Genet* 2019;**51**:207–16. doi:10.1038/s41588-018-0287-5

249 Kim S-B, Dent R, Im S-A, *et al.* Ipatasertib plus paclitaxel versus placebo plus paclitaxel as first-line therapy for metastatic triple-negative breast cancer (LOTUS): a multicentre, randomised, double-blind, placebo-controlled, phase 2 trial. *Lancet Oncol* 2017;**18**:1360–72. doi:10.1016/S1470-2045(17)30450-3

250 Chapman PB, Hauschild A, Robert C, *et al.* Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N Engl J Med* 2011;**364**:2507–16. doi:10.1056/NEJMoa1103782

251 Tannock IF, Hickman JA. Limits to personalized cancer medicine. *N Engl J Med* 2016;**375**:1289–94. doi:10.1056/NEJMsb1607705

252 Prasad V, Fojo T, Brada M. Precision oncology: origins, optimism, and potential. *Lancet Oncol* 2016;**17**:e81–6. doi:10.1016/S1470-2045(15)00620-8

253 Prasad V. Perspective: The precision-oncology illusion. *Nature* 2016;**537**:S63. doi:10.1038/537S63a

254 Sánchez NS, Mills GB, Mills Shaw KR. Precision oncology: neither a silver bullet nor a dream. *Pharmacogenomics* 2017;**18**:1525–39. doi:10.2217/pgs-2017-0094

255 Brock A, Huang S. Precision oncology: between vaguely right and precisely wrong. *Cancer Res* 2017;**77**:6473–9. doi:10.1158/0008-5472.CAN-17-0448

256 Larkin J, Chiarion-Sileni V, Gonzalez R, *et al.* Five-Year Survival with Combined Nivolumab and Ipilimumab in Advanced Melanoma. *N Engl J Med* 2019;**381**:1535–46. doi:10.1056/NEJMoa1910836

257 Skoulidis F, Goldberg ME, Greenawalt DM, *et al.* STK11/LKB1 Mutations and PD-1 Inhibitor Resistance in KRAS-Mutant Lung Adenocarcinoma. *Cancer Discov* 2018;**8**:822–35. doi:10.1158/2159-8290.CD-18-0099

258 Del Monte U. Does the cell number 10still really fit one gram of tumor tissue? *Cell Cycle* 2009;**8**:505–6. doi:10.4161/cc.8.3.7608