Identification of driver mutations and tumour evolution in HER2 positive breast cancer

AUTHOR(S)

Peter O'Donovan

CITATION

O'Donovan, Peter (2019): Identification of driver mutations and tumour evolution in HER2 positive breast cancer. Royal College of Surgeons in Ireland. Thesis. https://doi.org/10.25419/rcsi.10764350.v1

DOI

10.25419/rcsi.10764350.v1

RCSI

Identification of driver mutations and tumour evolution in HER2 positive breast cancer

Peter O'Donovan

Department of Physiology and Medical Physics

Royal College of Surgeons in Ireland

A thesis submitted to the School of Postgraduate Studies,

Faculty of Medicine and Health Sciences, Royal College of Surgeons in Ireland, in fulfilment of the degree of

Master of Science

October 2019

Supervision: Dr. Simon Furney, School of Physiology and Medical Physics, RCSI

Thesis declaration

I declare that this thesis, which I submit to RCSI for examination in consideration of the award of a higher degree M.Sc. is my own personal effort. Where any of the content presented is the result of input or data from a related collaborative research programme this is duly acknowledged in the text such that it is possible to ascertain how much of the work is my own. I have not already obtained a degree in RCSI or elsewhere on the basis of this work. Furthermore, I took reasonable care to ensure that the work is original, and, to the best of my knowledge, does not breach copyright law, and has not been taken from other sources except where such work has been cited and acknowledged within the text.

Signed _____

Student Number 17172004

Date 07/09/2018

Table of contents

Table of Contents

List of Abbreviations

GATK = Genome Analysis Tool Kit GUI = Graphical User Interface

ICHEC = Irish centre for High End Computing CNA = Copy Number Alteration

GFF3 = Generic Feature Format 3

ICGC = International Cancer Genome Consortium TCGA = The Cancer Genome Atlas

TCHL = Carboplatin, Docetaxel and Trastuzumab, with Lapatinib (Name of the clinical trial)

NGS = Next Generation Sequencing Hg38 = Human reference genome 38

PON = Panel of Normals (for variant calling) CGI = Cancer Genome Interpreter

DNA = DeoxyriboNucleic Acid

FGFR = Fibroblast Growth Factor Receptor

dbSNP = The Single Nucleotide Polymorphism database COSMIC= Catalogue of Somatic Mutations in Cancer NCBI = National Centre for Biotechnology Information

FACETS = Algorithm to implement Fraction and Allele specific Copy number Estimate from Tumour/normal Sequencing

CNV = Copy Number Variation

VAF = Variant Allele Frequency

ITH = Intra Tumour Heterogeneity

VCF =Variant call format (file format)

SAM = Sequence Alignment Map (file format)

BAM = Binary Alignment Map (file format) VEP = Variant Effect Predictor

FACETS = algorithm to implement Fraction And Copy number Estimate from Tumour/normal Sequencing

DSB = Double Strand Break

HER2 = Human Epidermal Growth factor receptor 2 (gene) ER= Estrogen receptor (gene family)

PR = Progesterone receptor (gene family)

EGFR = Epidermal Growth Factor Receptor (gene) TP53 = Tumour Protein p53 (gene)

BCL2 = B Cell Lymphoma 2 (gene)

MYC = MYeloCytomatosis (gene family consisting of c-myc, l-myc, and n-myc. MYC in the text refers to c-myc).

APOBEC = apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (gene family)

RAS = RAt Sarcoma (gene)

List of Figures

samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

Figure 3.5.7 - Subclonal architecture in the TCHL 29 Pre and Post treatment samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

Figure 3.5.8 - Subclonal architecture in the TCHL 32 Pre-treatment and Relapse samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

Figure 3.5.9 - Subclonal architecture in the TCHL 39 Pre and Post treatment samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

List of Tables

Summary

This thesis describes the results of a variant calling project using data from the TCHL (Trastuzumab, Carboplatin, and Docetaxel, with Lapatinib) breast cancer whole exome sequencing project. Using a data processing pipeline based on the GATK (Genome Analysis Tool Kit) Best Practices Pipeline, variants were called on a total of 34 tumour samples from 25 patients. This included 20 patients where only a pre-treatment sample was taken from that patient, and 5 patients where samples were taken from multiple timepoints across the course of therapy. The whole exome sequencing data were analysed for the presence of known or predicted driver mutations. The subclonal architecture and mutational signatures of the genomes of these samples were also analysed.

The cohort showed mutational signature patterns typical of a breast cancer cohort. The genomic landscape of the samples taken from the same patient across multiple timepoints was used to analyse the evolutionary history of the tumours in those patients and how they had evolved in response to therapy. The genomic landscape of samples from patients who showed complete response to therapy is compared to the genomic landscape of patients who did show complete response to therapy.

The results of this study highlight the following subjects as promising areas for future study with larger scale cohorts:

-The impact of the subclonal complexity of a tumour on the probability it will show complete response to therapy

-The impact of the presence of SNVs and indels in *RAD50, ARID1B, DHX9, IZF3, TNPO2, UBR5* and *TAOK1* on the probability that a tumour will show complete response to therapy

Acknowledgements

My thanks to Dr Simon Furney for acting as my supervisor during this process and to Valentina Thomas for helping me to understand how some of the software works.

My thanks to the Irish Cancer Society – Breast Predict Project for funding this research.

My thanks to my parents.

1 Introduction

1.1 The Genomics of Cancer

Cancer is the second most common cause of death in the world, killing over 8 million people every year (1). Cancer is a genetic disease (2). Out of control cellular proliferation occurs because mutations in the DNA bases of the genome and epigenetic marks of the epigenome release the usually existing constraints on cellular proliferation (3). The “genome” of an individual refers to the total DNA sequence of the DNA present in the nuclei of the cells of that individual, as well as the DNA sequence of their mitochondria (mtDNA). The “epigenome” of an individual refers to non-DNA chemical modifications on elements of the genome that affect the expression levels of genes without altering the DNA that makes up those genes directly (e.g. methyl groups attached to DNA bases, or acetyl groups attached to the histones the DNA is wound around) (4). Changes to the epigenome are referred to as “epigenetic” changes. A “mutation” in this context refers to a permanent change in the genetic sequence of the DNA in the cell(s) of that individual from the sequence that was originally there.

Later in the development of the cancer, cancerous cells spread to different parts of the body to where the cancer started, seeding new tumours in these new locations. This process is called “metastasis”, and a cancer undergoing metastasis is referred to as “metastatic” (5). The process of metastasis is estimated to be the cause of death in 90% of cancer mortalities (6). This is a clear motivation to learn as much as possible about how tumours evolve in order to understand why metastasis occurs.

The Genetic Mutations behind cancer

Genetic mutations can either be “germline” or “somatic”. Germline mutations refer to mutations that are present in the first fertilized egg cell that eventually gives rise to all of the cells in the mature individual. These mutations arise in the sperm or egg cells that form this fertilized egg, and are propagated to every cell in the body of the adult individual. In contrast, “somatic” mutations are mutations that arise at any point after the fertilized egg begins dividing. These

mutations will only be propagated to other cells or tissues that arise from that specific mutated cell (4). Cancers are caused mainly by somatic mutations, although germline mutations play a role as well in many cases (7).

It remains a point of contention whether or not it is necessary for a cell to have an elevated somatic mutation rate compared to the normal background rate for the cell to become cancerous ("mutation rate" refers to the number of mutations in a cell per unit time). Although an elevated mutation rate may not be strictly necessary for tumourigenesis to occur (8) most human cancers display an elevated mutation rate and an elevated tendency to mutate further ("mutator phenotype") compared to other somatic cells (9). The association between higher mutation rate and higher cancer risk is also seen in the fact that tissues that have more stem cell divisions over the course of a lifetime show a far higher likelihood of developing cancer than tissues with a lower rate of stem cell division (10). "Stem cells" are cells that have not yet differentiated into a particular type of cell (e.g. differentiated into a neuron or a skin cell). Stem cells within a tissue are self-renewing –they generate new stem cells along with differentiated cells to maintain that tissue (11) (12). "Stem cell divisions" in this case refers to the number of times that the stem cells in a given tissue divide, on average, over the lifetime of a human being (10).

Only a small fraction of the total mutations present in the genome of a cancer cell actually contribute to the cell's cancerous state. These mutations are referred to as "driver" mutations, in contrast to the other "passenger" mutations present in the cell, which have a neutral or even actively negative effect on the cell's survival (13). A driver mutation, by definition, has at some point during the evolutionary history of a cancer tumour been actively selected for, though it may not necessarily be necessary for the survival of the tumour at every stage in its life (14). Genes that are capable of bearing driver mutations are known as "cancer genes". The identification of cancer genes and the mutations within them that can act as drivers is a core focus of cancer research. This is because identifying which mutations are causing cancer allows the design of tailored therapeutic approaches which combat the specific effects of these mutations to attempt to cure the cancer that they cause (15).

Intratumour heterogeneity

Analysis of cancer genomics is further complicated by the phenomenon of intra tumour heterogeneity. “Intratumour Heterogeneity” (ITH) refers to the fact that cells within a cancer tumour do not share identical genomes – different cells acquire different mutations as the tumour grows (16). Some mutations in the cells in a tumour are “clonal” – this means they were present in the first progenitor cell of the tumour when it began to divide uncontrollably. “Clonal” mutations are expected to be present in every cell in a tumour, unless within that cell the genomic position of the mutation has mutated further. However, cancer tumour cells may also harbor “subclonal” mutations, which are unique to a cell or subset of cells in the tumour (17). The phenomenon of ITH appears to stand somewhat in contrast to the traditional model of tumour evolution in which cells bearing the most advantageous mutations would grow to dominate the tumour and displace all cells with less advantageous genomes (“selective sweeps”) (18). However, it is still possible that the traditional model is accurate, and that the ITH we observe in cancer genomes is due to the tumour being mid - sweep: the more beneficial mutation may not have had time to grow to ubiquity in the tumour.

ITH occurs due to the continuing action of mutagenic processes as the tumour grows, as well as the selective pressures applied by the local microenvironment around the tumour and by cancer treating drugs (19). The impact of the local microenvironment can be seen in the following phenomenon: in any sizable tumour, there are regions of low oxygen levels (hypoxia) and regions of regular oxygen levels (normoxia). Studies have shown that regions of hypoxia favour cells with mutations that cause those cells to metabolise anaerobically, whereas in regions of normoxia cells that metabolise aerobically are favoured (20) (21). This is just one example of the impact that the local microenvironment can have on the evolutionary environment of the cells in a tumour, and therefore on the amount of ITH present in that tumour.

The mutations that cause ITH can take the form of point mutations, copy number variations or differences in chromosomal structure or number (22). Point mutations are mutations that affect only a single nucleotide base in the

genetic sequence – usually this involves a nucleotide base being substituted for another one (e.g. a G > A mutation), but it sometimes involves a base being deleted from the sequence, or a new base being inserted into the sequence (4). A “copy number variation” refers to alterations in the number of copies of specific regions of DNA – the region may either be deleted or duplicated (23).

With the advent of next generation sequencing, the extent of ITH in tumours is beginning to be understood, as is the fact that some tumour types show more heterogeneity than others. Cancers that are usually preceded by prolonged exposure to a powerful external mutagen, such as lung cancer and melanoma, tend to have a much higher proportion of subclonal mutations compared to other cancer types (24).

Just as some tumour types are more likely to be heterogenous than others, some driver mutations are more likely to be subclonal than others – for example, across several cancer types subclonal mutations in genes involved the *PIK3- AKT- mTOR* pathway are more common than genes involved in the *RAS-MAPK* pathway (25). For some mutations, whether they are likely to be clonal or subclonal depends on the cancer type – for example, mutations in *TP53* are almost always clonal in ovarian cancer (26), but are often subclonal in chronic lymphocytic leukemia (27).

Intratumour heterogeneity is relevant to tumour metastasis because the metastatic site may be seeded by a cell from the original tumour harboring subclonal driver mutations (28) (29). These mutations may be crucial to allow the cell to detach from the primary tumour, survive the immune system while travelling to the site of metastasis, and survive in the microenvironment of the secondary tumour site (30).

ITH is relevant to this project because the samples being sequenced (described further below) actually represent an entire population of tumour cells. The “SciClone” R package (described further below) is used to make inferences from the mutation calling data generated about the nature and evolutionary history of the tumour cell population that that mutation calling data comes from.

1.2 Cancer evolution

Over time, a cancer evolves in a manner analogous to the evolution of an asexually reproducing, single celled species (31). Mutations accumulate in the descendant cells of the original cancer progenitor cell, and these mutations are selected in a Darwinian fashion – cells with disadvantageous variants are more likely to die, and so these are selected against, whereas cells with advantageous variants are more likely to survive and proliferate and so become more common (32). As with all evolutionary processes, tumour development is affected not just by Darwinian evolution but also by genetic drift. Genetic drift refers to the fact that genetic alleles in a population fluctuate in frequency randomly across time, and alleles may rise to ubiquity across a population ("fixation") or be driven to extinction by chance rather than because of the effect of those alleles themselves (4). This occurs because how many progeny an individual cell or organism gives rise to is not an exact function of how beneficial its genome is in its environment, but is to some extent down to random chance - in a cancer context, an example might be a driver mutation arising in a cell, but that cell being killed by unrelated bodily processes (as opposed to being targeted for destruction), before it is able to proliferate. Current evidence suggests that drift plays less of a role in tumour genome evolution than selection, but it is still a driving force that shapes tumour genomes (33).

Despite the great genomic heterogeneity seen in tumours both within and between different cancer types, the final phenotype of the tumour is always highly similar – most cancers grow and evolve in a similar way, despite the vast genomic differences they may harbor under the surface (34). We can conclude from this that there are many different genomic pathways that, when altered, can lead to tumourigenesis and eventually metastasis. This is why it is necessary to scan whole genomes of tumours to find these mutations, and why the evolution of any given tumour is such an unpredictable process - there is not one neat, simple set of genetic alterations that must occur for tumourigenesis or metastasis to occur, but instead a wide variety of possibilities that all lead to the same phenotypic outcomes (35). We also note that, as well

as being impacted by genomics, cancer biology is also impacted by the transcriptomics (36) and epigenomics (37).

As explained in section 1.1, a cell usually needs multiple driver mutations to become cancerous, with the exact number and nature of driver mutations needed varying by tissue type. Recent examination of the mutational burden of somatic tissue suggests that driver mutations are selected for in somatic tissue before the cells become fully cancerous. This, of course, increases the number of cells bearing a number of driver mutations, and so increases the chances that a driver mutation will occur in a cell that already bears other driver mutations, until an actual tumour develops when the correct combination of drivers exist in the same cell (38).

Evolution after Tumourigenesis

Evolution continues to take place after initial tumourigenesis, shaping the genomic development of the cells in the tumour over time (39). Even in the fastest growing tumours, the time taken for the tumour to double in size is vastly longer than the time taken for individual tumour cells to double. This implies either that most cells in a tumour are restrained from dividing by signals from their local micro-environment, or that the vast majority of cells produced by a tumour die before they themselves can divide. (40)

In either case, there is clearly ample scope for natural selection to occur. This may occur through the deaths of tumour cells with suboptimal genomes, or via positive selection for cells capable of dividing more rapidly, thus escaping environmental constraints on growth. The tumour may even evolve in such a way as to remodel the local microenvironment to the benefit of the tumour (41).

The most important outcome of these post tumourigenesis evolutionary processes from a therapeutic perspective is the development of metastasis. In order to become metastatic, tumour cells must develop the ability to leave the original tumour, survive in the bloodstream, and found new tumours at distant sites from the original. It remains a subject of active debate whether metastatic cells arise late in tumour development and are highly genetically distinct from the original founder cell (“linear progression”), or whether proto-metastatic cells

disseminate from the founder cell at relatively early stages in tumour development, long before the cancer can be detected, and the initial tumour develops independently from the metastasis, with the metastatic cells developing most of the features necessary for successful metastasis after leaving the initial tumour ("parallel progression") (40). Studies trying to distinguish between which of these models are correct have been inconclusive, showing some evidence for both models, sometimes even in the same patient. Thus, it is likely that there is a continuum between the two models rather than one model being entirely correct (42).

In terms of which genes specifically can trigger metastasis when mutated, the only gene consistently associated with a metastatic phenotype is *TP53* (full name: tumour protein p53) (43). The usual role of *TP53* in a healthy cell is to act as a tumour suppressor protein, preventing the cell from proliferating uncontrollably. *TP53* fulfils this role in a number of ways, including the regulation of the cell cycle, DNA repair, and cellular metabolism (44). It should be self-evident how the inactivation of a gene with this role would increase the likelihood of a cell becoming cancerous and surviving to become metastatic. Another reason for the association between *TP53* mutations and a metastatic phenotype is the association between *TP53* and chromosomal instability (45), as well as chromothripsis (46). "Chromosomal instability" refers to a state in which whole chromosomes or parts of chromosomes are duplicated or deleted. This leads to the daughter cells of a dividing tumour cells with chromosomal instability showing aneuploidy (i.e. having an incorrect number of chromosomes) (4). "Chromothripsis" refers to a phenomenon whereby chromosomes are observed to break apart and then reattach together in a single event that leads to thousands of clustered rearrangements (47) In general, metastasis seems to be associated with chromosomal instability, structural alterations and somatic copy number changes (48) (49) (50) (51). Since, as discussed above, a cell must acquire a large number of new traits to become metastatic, and these traits may not be individually beneficial for survival (e.g. it is not useful to be able to seed a new site if the cell cannot survive in the bloodstream), it would be difficult for all of these changes to occur in an evolutionary step wise manner (52) (53). Instead it is likely to be

necessary for large-scale changes to cause all of the metastasis causing mutations in a single step. This would explain why the instability-enabling *TP53* mutations are so frequently associated with metastasis. *TP53* itself also contributes to pathways that inhibit metastasis, so inactivating mutations in *TP53* represent a release of a restraint on metastasis as well as contributing to the likelihood of metastasis causing mutations occurring (43).

Studies examining metastatic genomes in breast cancer specifically support a linear progression model of metastasis in breast cancer (54). They also support the idea that chromosomal instability, *TP53* mutations and somatic CNAs are highly likely to be involved in a given case of metastasis (55) (56) (57).

The impact of therapy on Tumour Evolution

Therapy, of course, represents an environmental hazard to the cells in a cancer, and so the cells evolve in response to this constraint. Therapy is also often genotoxic, so the somatic mutation rate of the cells in the body reacting to the therapy is elevated while therapy is ongoing (58). The evolutionary environment during therapy is therefore one in which there is strong selective pressure to develop resistance and a large number of novel mutations being created as a substrate on which natural selection can act. The level of tumour genetic heterogeneity correlates with the likelihood that a tumour will survive therapy - increased genetic diversity in the tumour prior to the start of therapy raises the odds that a mutation capable of allowing the tumour to escape the attempts made by the therapy to eradicate it (59) (60) (61). Studies comparing the genetic markers of resistance in pre-treatment and post-relapse samples in cancers that eventually relapsed have shown that small populations of resistant subclones often exist prior to therapy starting, indicating that selection for pre-existent populations is the main mechanism through which a tumour becomes resistant to therapy (62).

The traditional method to learn about the evolutionary history of a tumour is to analyse the level of clonality of different mutations - is a given mutation in a tumour clonal or subclonal? If a mutation is subclonal, this implies it arose after the founder cell of the tumour originally became cancerous, whereas a clonal mutation is likely to have been present in the founder cell. The frequency of a

subclonal mutation may give us information about how advantageous that mutation is and therefore how strongly it is selected for.

1.3 Breast Cancer, HER2 status and the TCHL project

Breast cancer is the most commonly diagnosed cancer in women and the main cause of cancer-related mortality in women worldwide (63). Breast tumours are highly heterogeneous and are classified based on histology, gene expression profiles, and the expression of oestrogen (*ER*) and progesterone (*PR*) hormone receptors and *HER2* (*ERBB2*) (64). There are five broad categories in the gene expression based classification system, which is the classification system we are using in this project: Luminal A, Luminal B, basal-like, triple negative and *HER2* positive. Figure 1.1 below summarizes how the classification system works. Figure 1.2 summarizes how the classification impacts the prognosis of the patient upon diagnosis: different subtypes are associated with different survival rates. This graph looks at Overall Survival (OS) as opposed to Disease-Free Survival (OS just looks at whether a patient remains alive, Disease-Free Survival also looks at whether the patient suffered a relapse in the time frame under examination). This method of breast cancer classification was first reported by Sorlie et al in 2001 (65). These classifications are used in both research and the clinic because classification status is a significant prognostic factor (a factor that tells you the likelihood of recovering from the disease) and can guide therapy because different therapies are appropriate for different subtypes (66) (67). Another reason that these classifications are used in the clinic is that a clinic can use immunohistochemistry (IHC) staining to identify which subtype a given tumour falls into (68). This gives a doctor a convenient way to classify a tumour, rather than needing to sequence the patient's genome to classify the tumour – far more hospitals have IHC equipment and expertise than have the capacity to carry out gene sequencing.

In brief: Luminal A breast cancer is hormone receptor positive (overexpression of estrogen receptor (*ER*) and/or progesterone receptor (*PR*)), has no amplification of *HER2*, and tends to be the least aggressive and have the best prognosis (67). Luminal B breast cancers are hormone receptor positive, but also have amplification of *HER2* (68). Tumour grade does not play an explicit role in the distinction between the Luminal A and Luminal B subtypes, though

luminal B cancers have been shown to be associated with a higher grade compared to luminal A patients in a study of Pakistani patients (69). Basal like breast cancer is breast cancer with no expansion of *ER, PR*, or *HER2*, but with an expression signature of basal markers including cytokeratins 5,6 and 17sa (70). Triple negative breast cancer is similar to basal-like breast cancer in that it shows no amplification of *HER2, ER* or *PR*, but it does not show expression of the basal markers (71). It is the rarest and most aggressive form of breast cancer (72). *HER2* positive breast cancer is breast cancer in which the *HER2* gene is amplified (i.e. the gene copy number is increased) or the gene is overexpressed for other reasons, without amplification or overexpression of the hormone receptors (*ER* and *PR*) (73).

Since this project is about *HER2* positive breast cancer, this subtype is described in depth below.

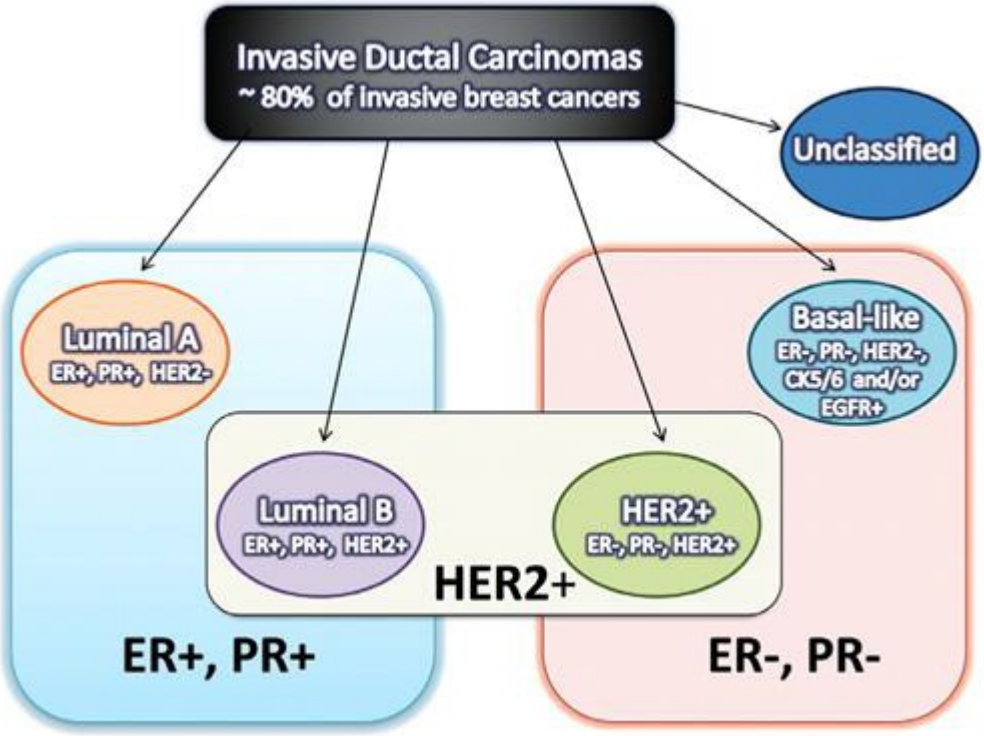
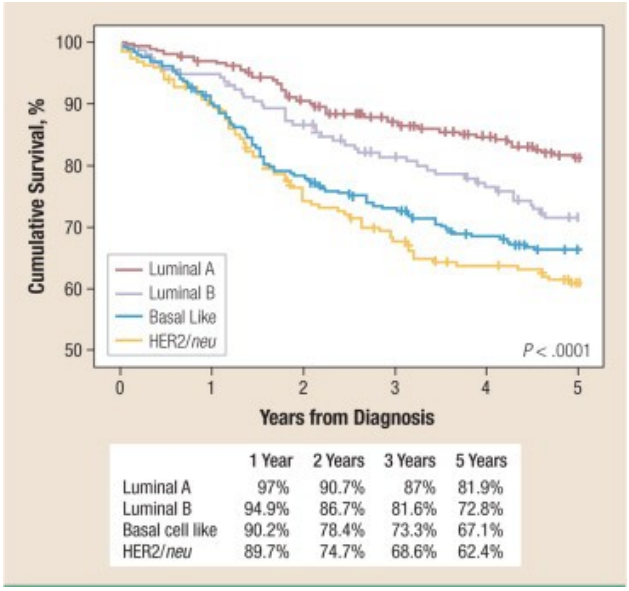


Figure 1.1 - A visual summary of the classification system for breast cancers (68)*). This diagram refers specifically to Invasive Ductal Carcinomas, but the same gene-expression based classification scheme applies to all types of breast cancers. "Invasive Ductal Carcinoma" is a classification in a different classification scheme for breast cancers, based around morphology rather than gene expression. The other classes in this classification scheme are invasive lobular carcinomas (~10% of invasive breast cancers) and a long series of other, much rarer subtypes* (68)



	1 Year	2 Years	3 Years	5 Years
Luminal A	97%	90.7%	87%	81.9%
Luminal B	94.9%	86.7%	81.6%	72.8%
Basal cell like	90.2%	78.4%	73.3%	67.1%
HER2/*neu*	89.7%	74.7%	68.6%	62.4%

Figure 1.2 – Survival rates associated with either breast cancer subtype. The data is from the cited paper, and originated from a Peruvian hospital database showing data on breast cancer patients treated between 2000 and 2002. The percentages shown refer to the overall percentage of patients with that breast cancer subtype diagnosis who were still alive that many years after the date of diagnosis (i.e. Overall Survival). The purpose of the figure is to show the impact of which subtype of breast cancer a person has on their probable survival time/ (74)

HER2 Positive Breast Cancer

HER2 stands for Human Epidermal Growth factor receptor 2 (75). It is found on the surface of all breast cells, cancerous or otherwise (76). *HER2* is one of the 4 member of the ERBB receptor tyrosine kinase family of receptors (hence why *HER2* is also known as *ERBB2*). As with all members of this family, it contains an extracellular ligand binding domain, a transmembrane domain, and an intracellular domain which can work in a ligand dependent or ligand independent manner (77). *HER2* is able to heterodimerize with any of the other family ERBB family members. Heterodimerization leads to phosphorylation of tyrosine in the intracellular domain of the receptors, activating various signaling pathways which trigger mitosis and oppose apoptosis in the cell they are active in. (78)

HER2 positive breast cancer is breast cancer in which the *HER2* gene is amplified (i.e. the gene copy number is increased) or the gene is overexpressed for other reasons, without amplification or overexpression of the hormone receptors (*ER* and *PR*). The number of *HER2* receptors correlates with the rate

of tumour growth, which makes logical sense given the function of these receptors (79). *HER2* positive cancers account for 15-20% of breast cancers in women, though they are believed to be less common in male breast cancer (80). The risk factors for getting HER2 positive breast cancer are for the most part exactly the same as the risk factors for any other given subtype of breast cancer (81). These common risk factors include age, early menarche (before 12 years old), obesity, alcohol use, physical inactivity, family history of breast cancer, X-rays and radiotherapy. Obesity ws noted in the cited paper to be slightly less common in ther HER2 positive subtype than in the other subtypes.

Though less aggressive than triple negative breast cancer, *HER2* positive breast cancer is associated with a worse prognosis than Luminal A or Luminal B breast cancer. Several targeted therapy options exist for *HER2* positive breast cancer, including trastuzumab (a.k.a herceptin). Other targeted therapies for HER2 positive breast cancer include docetaxel and carboplatin. These targeted therapies have lead to great increases in disease free survival and overall survival rates for *HER2* positive breast cancer patients, but a subset of *HER2* positive patients either fail to show any initial response to the therapy or show an initial response followed by a relapse (82). *HER2* positive breast cancers that fail to respond to targeted therapies often have mutations in the PI3 kinase pathway - either mutation in the *PIK3CA* gene or downregulation of the *PTEN* tumour suppressor gene (83).

Risk factors for disease progression to metastasis in HER2 positive breast cancer include cigarette smoking, steroid receptor status, history of cancer in the family, and being postmenopausal (84). The standard therapy for HER2 positive breast cancer worldwide has for many years been trastuzumab (explored in more detail in the next section) (79), often with platinum based chemotherapy as an adjuvant (85) (86). Modern studies are experimenting with replacing the chemotherapy with other targeted medications, as these targeted medications are likely to be more effect in treating the disease and cause fewer harmful side effects to the patient. This study is one such study. The relevant targeted therapies (docetaxel, carboplatin, lapatinib, and trastuzumab itself) are described in detail in the next section.

Targeted therapies for HER2 positive breast cancer

Trastuzumab is a humanized monoclonal antibody created from recombinant DNA (87). It binds with high affinity and selectivity to the extracellular domain of *HER2*. By binding to the *HER2* receptors, it blocks growth signals that the receptors would otherwise receive (88). Trastuzumab also plays a role in alerting the immune system to destroy the cancer cells it is attached to (89). Approximately 15% of patients who initially respond to trastuzumab eventually experience relapse. The reason for this is not yet conclusively known (90).

Docetaxel is an anti mitotic drug that is used in treatment of breast, ovarian and other cancer types (91). Docetaxel functions by binding to the β-tubulin subunit of tubulin in the microtubules. This hyper-stabilizes the structure of the microtubules and makes it impossible for the microtubule to disassemble, ruining the flexibility of the microtubules needed for mitosis to occur (92) (93). Docetaxel also binds to the anti apoptotic protein *Bcl 2* (B-cell leukemia 2), blocking its function and thereby promoting apoptosis in the cell (94).

Carboplatin is a chemotherapeutic agent most often used to treat ovarian and lung cancer (95) (96). It functions by adding platinum adducts to DNA, thereby inhibiting replication and transcription of that DNA, ultimately leading to cell death (97). Alkylating agents (e.g. cyclophosphamide) are often prescribed alongside carboplatin, as the drugs seem to be particularly effective when combined in this way (98) .

Lapatinib is a tyrosine kinase inhibitor that can block both *EGFR* (Epidermal Growth Factor Receptor) and *HER2* (99). It has been shown to have anti-tumour activity in trastuzumab resistant breast cancer (90).

The TCHL Clinical Trial

The TCHL clinical trial (clinicaltrials.gov/ct2/show/NCT01485926) assesses treatment of *HER2* positive breast cancer patients with docetaxel, carboplatin and trastuzumab, with or without lapatinib (TCH, TCHL). The trial excluded patients with concurrent therapy with any other non-protocol cancer treatment. 35 available tumour biopsies pre- and during the first treatment cycle (after 21 days), and recurrent tumours, as well as whole blood used as a ‘normal’ control,

were exome sequenced for both protein coding and selected non-coding regions. “Recurrence” can be either local (an apparently cured tumour reappears at the same site) or non-local (a tumour appears at another location in the body of a patient whose original tumour was cured, indicating that the original tumour had gone metastatic before therapy was successful). This is the dataset used in this project. The purpose of sequencing a “normal” control is that it allows us to “call” somatic mutations – germline mutations will be shared by both the normal and the tumour cell, so by subtracting any mutations seen in the normal cell from the full set of mutations seen in the tumour cell, we can identify which mutations in the tumour cell are genuinely somatic (this is a simplification of a much more complicated process – see section 1.5 for more details on how the variant calling software, Mutect2 works).

The metric used for whether therapy has been successful is whether or not pCR (pathological complete response) was achieved in the patient. pCR means that all signs of cancer are absent in tissue samples removed post-treatment. To determine pCR, a pathologist examines tissue slides under a microscope to see whether there are still any cancer cells remaining.

The results of the trial are summarized in Figure 1.3. The data analysed in this project is a subsection of the overall set of data in the overall TCHL clinical trial (100).

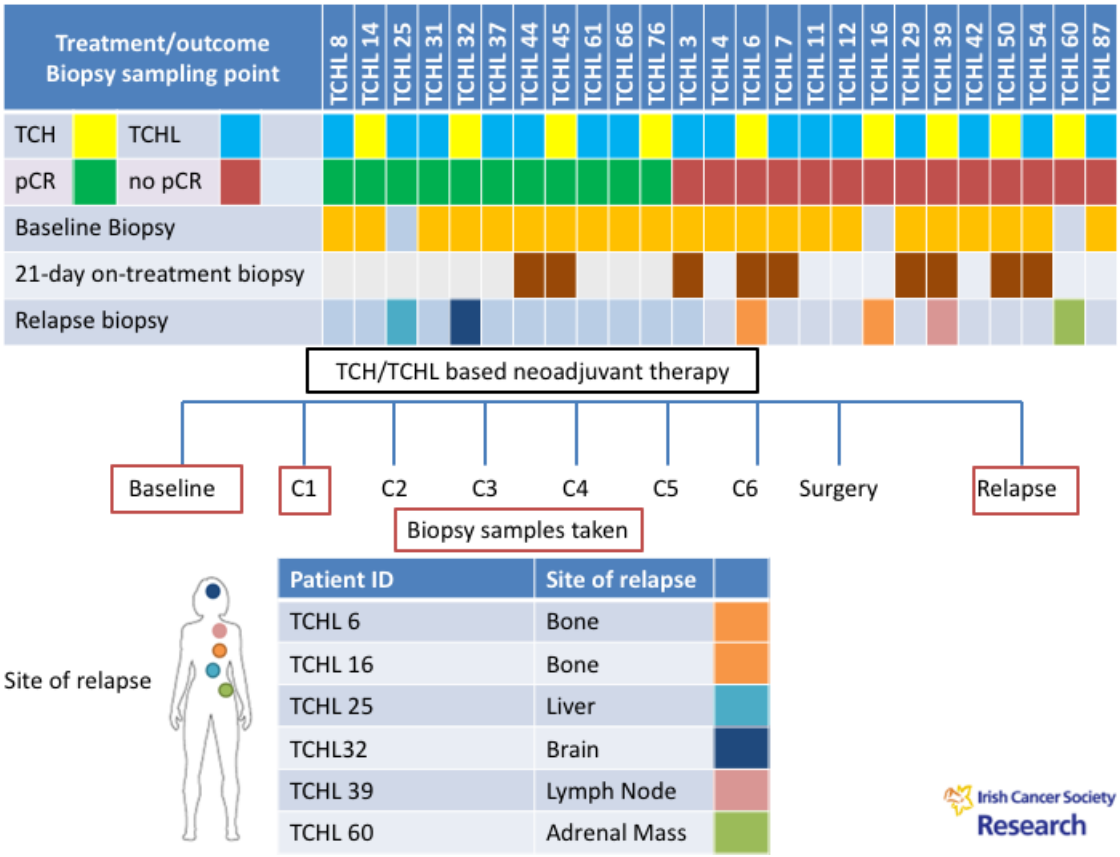


Figure 1.3 – TCHL trial summary, showing which samples were TCH and which were TCHL, which samples showed pCR and which did not, and which samples showed a relapse. The "21-days-on-treatment biopsy" row refers to whether a biopsy was taking from the patient 21 days after the treatment was started. "C1", "C2" etc. refer to being on treatment cycle 1, treatment cycle 2, and so on. It is not possible to give a citation for this image because it is not from a published paper – It is from an Irish Cancer Society Research presentation. My sincerest thanks to the Society for allowing me to include it in my thesis.

One aim of this project is to "call" mutations from the raw sequencing data and analyze these mutation calls to discover which driver mutations are present in the samples, in particular whether certain driver genes are commonly associated with those relapse samples.

1.4 Next Generation sequencing – how the data is prepared

In recent years, the speed and accuracy with which researchers can sequence genetic data has improved at a lightning pace in comparison to older Sanger sequencing technology. So called “Next Generation Sequencing” (NGS) allows for genomes, including cancer genomes, to be sequenced much more rapidly, at much higher depth, and for much lower cost than would have been possible in the past. This explosion of sequencing information has allowed researchers to scan entire exomes or genomes to look for potential driver mutations, rather than focusing on specific regions of the genome that they knew ahead of time were likely to be involved in cancer development, and to examine regions of interest at much higher “depth” of sequencing. “Depth” in this context means the number of sequencing “reads” covering the same area - the greater the depth, the more accurate the sequencing information is and the more likely it is for rare alleles to be discovered (101). Figure 1.3 below shows how much more data it has become feasible to sequence per instrument run since the year 2000.

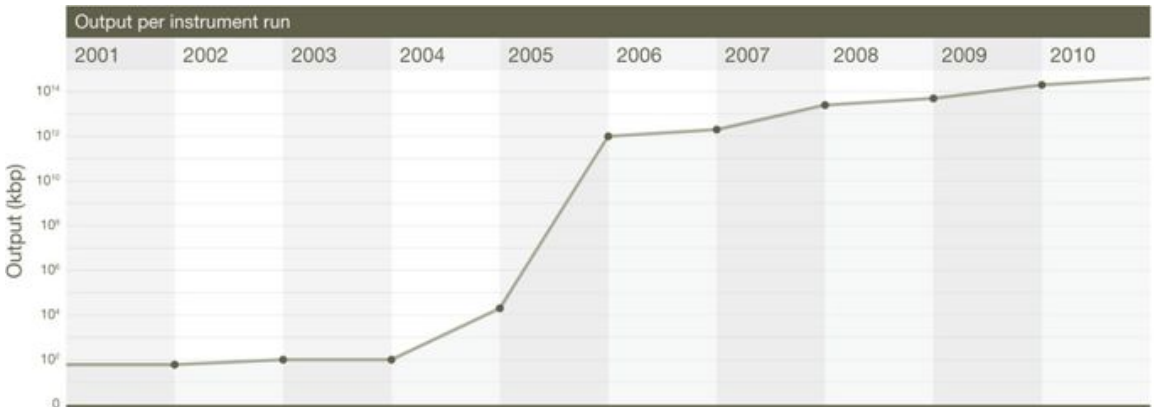


Figure 1.4- Figure 1.3 – Output in kilobases per instrument run of sequencing technology by year, demonstrating the massive improvement of sequencing technology that has taken place since the turn of the millennium (Mardis, 2011)

The Polymerase Chain Reaction

Modern sequencing technology relies on a process called the polymerase chain reaction (PCR) (NB- PCR as described here should not be confused with pCR, pathological complete response, as described above) PCR is a technique to make a very large number of copies (thousands to millions) of a single copy or a few copies of a DNA segment. Having all of these copies is necessary for the sequencing instrumentation to work (102).

Illumina Sequencing Technology

Although PCR is an extremely valuable tool without which NGS would be impossible, the sequence "copies" created by PCR contain some systematic inaccuracies (i.e. they are different from the template DNA used in systematic ways) (103). The issues created by the PCR process include preferential amplification of certain fragments, which leads to over representation of certain sequences. The factors influencing whether a fragment will be preferentially amplified include length and GC content. This issue is addressed by the "duplicate marking" stage of data processing (see section 1.6). PCR may also amplify substitution errors - if the DNA polymerase inserts the wrong base early in an early PCR cycle, the PCR process will copy this error into many read copies (104).

NGS is a catch all term for modern rapid and accurate sequencing technologies. The specific technology used to generate the sequencing data presented in this thesis is Illumina sequencing technology. Illumina sequencing technology starts with a silicon surface covered in primer sequences attached to the surface that are complementary to the adapters ligated to the DNA fragments that are to be sequenced. The DNA "library" of fragments is poured over this surface, so that the fragments bind to the adaptors. Each of the bound fragments is then amplified by a PCR process called "bridge amplification", leading to many copies of the same fragment being present all very close to each other in a "cluster" on the silicon surface. The bridge amplification process is summarized in Figure 1.5 below.

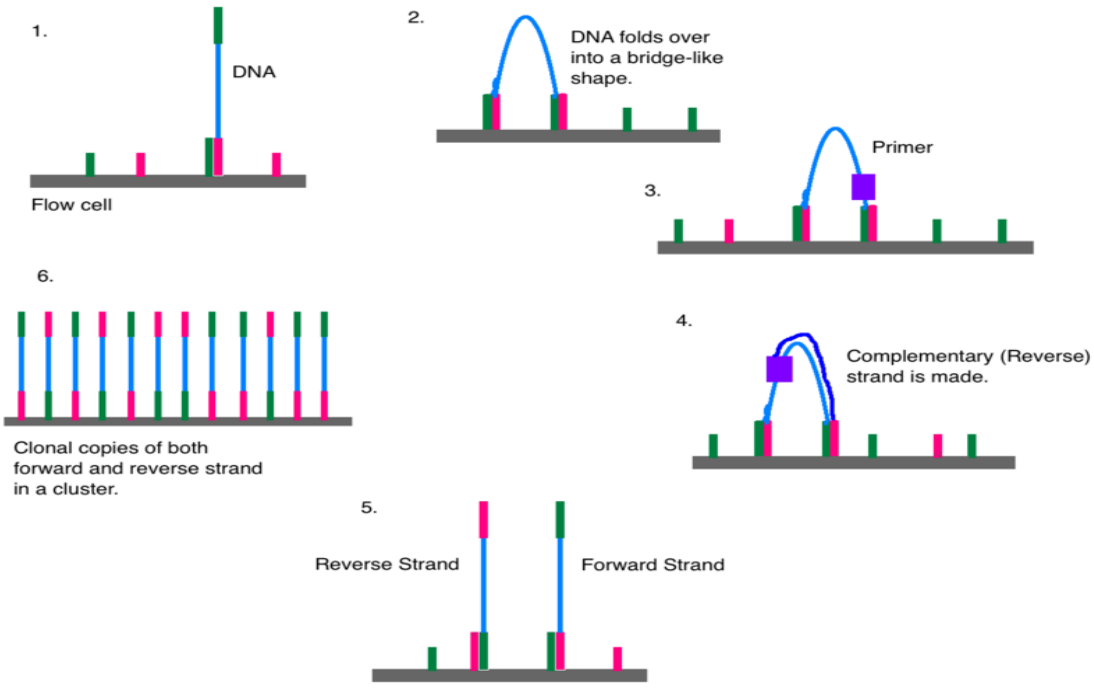


Figure 1.5 - The process of PCR bridge amplification. This process is used to create many clusters in a flow cell, each cluster corresponding to a specific fragment to be sequenced. Each read in the data output by the sequencing equipment corresponds to one cluster. Image from wikipedia, url: https://en.wikipedia.org/wiki/Illumina_dye_sequencing#/media/File:Cluster_Generation.png

With the clusters amplified, sequencing primers corresponding to the adaptors are added to the surface. These hybridize to the ends of the fragments in each cluster, in preparation for the addition of DNA bases complementary to the sequence of the fragment. The deoxyribonucleotide triphosphates (dNTPs) used to extend the primer are special in two ways - firstly, they each have an attached fluorescent molecule that, when the base is added to a growing DNA sequence, emits a color that is unique to the identity of the attached base (so there is a set color corresponding to A, another corresponding to G, and so on). The emitted fluorescence is detected by the sequencing instrumentation, and this is how the “base calls” in the “read” describing the DNA base sequence of the fragment are determined. Secondly, the dNTPs have a “blocker” molecule on the 3 prime end that blocks further synthesis until the blocker is removed. The blocker is used so that the bases are added one at a time so that fluorescence from a bases being added is gone by the time the next base is added and does not interfere with the signal from the next base, confusing the sequencing instrumentation. Despite this step, residual fluorescence does interfere with the signal from later added bases, which is one of the reasons

why base calls from later in the read (i.e. farther along the fragment) are systematically lower in accuracy than calls earlier in the read (105).

The reads produced for this project were paired end reads. This means that when sequencing is being performed, the ends of each fragment are sequenced, leading to there being 2 “mate pair” reads for each fragment. This allows for greater accuracy in alignment - because the distance between the mate pairs is known, they are easier to map to repetitive regions of the genome than solo reads would be, increasing the overall accuracy of the alignment. They are also helpful for detecting genomic rearrangements - if the mate pairs map to separate parts of the genome, it implies the presence of rearrangements (106). Figure 1.6 shows how paired end reads work and why they are helpful when mapping reads to the reference genome.

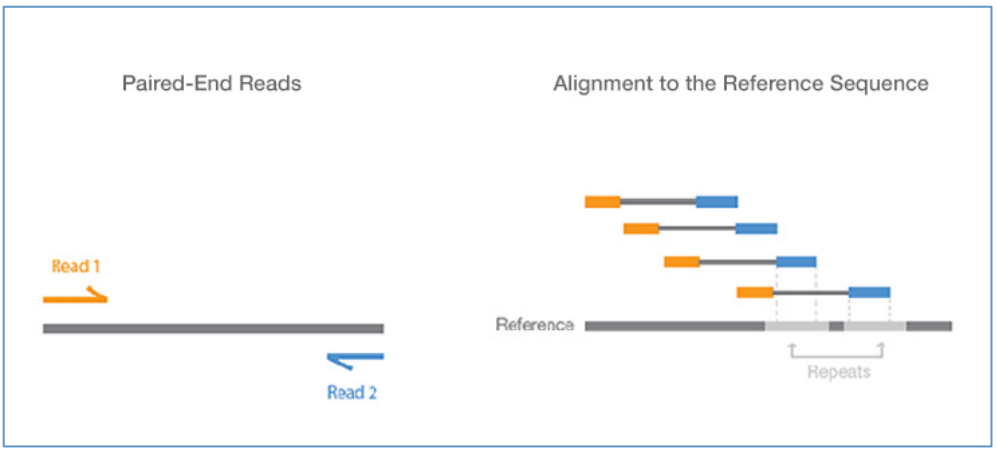


Figure 1.6 - Paired end reads and their use in mapping to the reference genome. Image from Illumina website, url: https://www.illumina.com/science/technology/next-generation-sequencing/paired-end-vs-single-read-sequencing.html

1.5 Data processing – what we do with sequencing data, and why

Many steps of the data processing in this project use tools from the Genome Analysis Toolkit (GATK) suite of programs. The GATK is a set of programs available for free for academic use from the Broad Institute. The GATK website is accessible at the following link: https://software.broadinstitute.org/gatk/. The tools used can be downloaded from the following URL: https://software.broadinstitute.org/gatk/download/. The GATK programs are all based around a MapReduce framework, in order to create programs that will work efficiently on the huge file sizes that characterize most biological data generated by NGS (107). In addition to providing these programs, the GATK website also provides a set of "Best Practices" guidelines explaining how to build a standard pipeline for variant calling in such a manner as to make the final variant calls as accurate as possible. The variant calling pipeline used in this project is based on these guidelines (108). The GATK pipeline is used because it is the industry standard for variant calling. The same rationale applies to the other non-GATK programs described below: they were chosen because they are the industry standard for the specified task. A given software is chosen as the industry standard due to having the best performance on that task in terms of accuracy, speed and computational efficiency in comparison to other available software for the same task.

The other main tool set used in this project is the Picard Tools suite of programs. This is a set of tools, again provided by the Broad Institute, designed to aid with manipulating SAM/BAM files and VCF files (explained below). These tools are called from the command line using Java (http://broadinstitute.github.io/picard/)

Introduction to data and quality control

The data used in this project started out as files in FASTQ format. FASTQ format is an extension of FASTA format. FASTA format files hold the genetic sequence of the reads as text, along with metadata about the reads in comment lines. Fastq format is similar but also includes a sequence quality score for each base in each read, showing how confident the sequence calling instruments are that that base call is accurate. These scores are important for ensuring the

accuracy of variant calling, as a low confidence base call is unlikely to represent true variation (109).

The first step of data processing for variant calling is checking the quality of the reads. In this project we used the FastQC program to generate statistics about the quality of the reads for each sample. FastQC, which can be accessed through a GUI or through the command line, generates a results readout containing information on quality scores across the reads, the GC base content across the reads, sequence duplication levels and the presence of overrepresented sequences (overrepresented sequences may represent adapter sequences mistakenly included in the reads). (110) A sample FastQC report is shown below in *Figure 1.7*

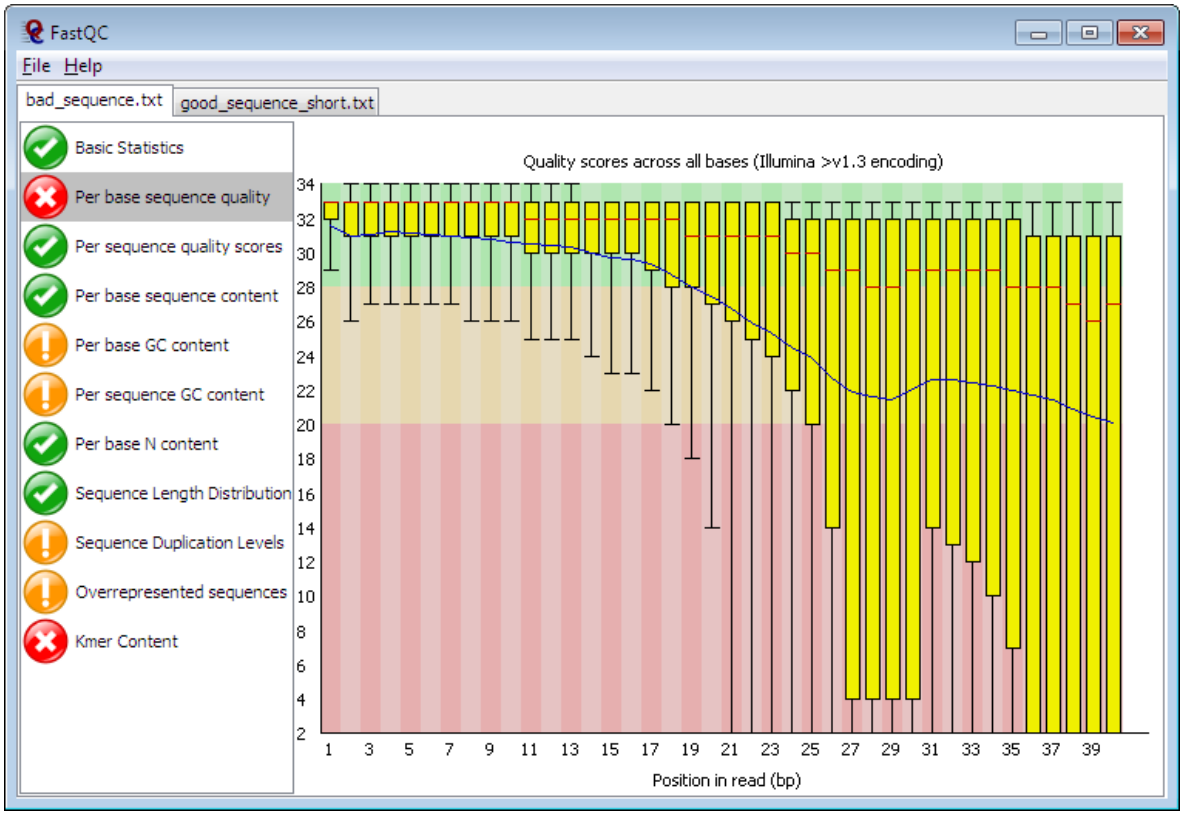


Figure 1.7 – Sample FastQC report, taken from the following URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

In order to make the reads used for alignment and variant calling as accurate as possible, we "trim" the reads produced to remove low quality bases and sequences that are likely to represent adapters rather than *bona fide* genetic data from the sample under examination. In this project we used the

"trimmomatic" program, accessible via the command line, to perform trimming on our read files. This program was used due to it being optimized to work properly with paired end Illumina read data (111).

Read Alignment to Reference Genome

Once the reads have been fully prepared, they must be "aligned" to a reference genome in order to discover what variants are present. This means that an algorithm is run to find the location in that reference genome that corresponds to the location in the genome that the read in question is most likely to have originated from. This project uses the Burrows-Wheeler Aligner, called from the command line as "bwa". This Aligner is based on the Burrows-Wheeler transform, a technique to rearrange a character string into chunks of similar characters, and the usage of this technique to match strings (112). This approach is used rather than the approach simply comparing the read to the entire genome to avoid the memory and processing costs of scanning the entire genome when many regions of the genome are likely to have no reads align to them. Along with the reads and the reference genome, the bwa program also takes as input a "read group" argument. This argument contains metadata about the reads, including what type of sequencing was used to generate them, what sample the reads were obtained from, and the sequencing library and lane they are from. This read group information is necessary for several downstream tools to work properly, and is especially important for merging BAM files that originate from the same sample together.

The bwa alignment program outputs files in the sequence alignment map (SAM) format, a format for storing genetic sequencing data aligned to a reference genome that was developed by Heng Li *et al* in 2009 in response to the need for a common alignment format to make it easier to write tools for downstream processing of alignment information (113). Each line in the SAMfile gives information about a read, where in the reference genome that read has been mapped to, the "mapping quality" (how confident the algorithm is that this mapping is correct), the base quality of the bases in the read, and the presence or absence of gaps in the alignment (if gaps are present they likely represent insertions or deletions in the genome under examination). Due to the very large

size of SAM files, they are almost always converted to the binary version, the BAM (BAM = Binary Alignment Map) file, for storage and for downstream processing. The Li et al paper also introduces the samtools suite of programs, which includes various programs to manipulate SAM format files to help with bioinformatics work. It includes utilities to sort SAM and BAM files, and to generate index files corresponding to those SAM and BAM files. These index files are necessary for several other tools (including many GATK tools) to work with the SAM/BAM files, as they are needed to allow these tools to rapidly "look up" certain parts of the file so that the programs can run in a reasonable amount of time (113).

BAM File Quality Control

Before further analysis of these sorted and indexed BAM files, it is necessary to carry out "duplicate marking" on them. "Duplicate marking" software identifies cases where there are multiple identical reads mapped to the same spot on the reference genome. While naively it might seem like each extra read covering the same location is evidence for greater sequencing depth at that location, in reality extra reads that are identical are far more likely to be PCR errors or cases where a single amplification cluster has been misidentified as multiple clusters by the sequencing instrument ("optical duplicates"). Including them in downstream analysis and in depth coverage calculations would give inaccurate information about the depth associated with variants called from those reads as a result, which would give inaccurate information about the confidence we can have in those calls. For this reason, the "MarkDuplicates.jar" tool from PicardTools was run on all BAM files before further processing. This tool flags all duplicates in the BAM/SAM and identifies what type of duplicate each identified duplicate is. Downstream tools are programmed to ignore these duplicate reads where relevant (114).

Base quality score recalibration (BQSR) is a machine learning based process that aims to enhance the accuracy of the quality score that is attached to each base in each read in an alignment (i.e. SAM or BAM file). This step in the variant calling pipeline is needed because there are systematic errors in the way that quality scores are assigned by the sequencing equipment - some due

to the chemistry of the sequencing reaction, some likely due to real life machines having manufacturing flaws and wear and tear etc and so not working exactly like theoretical ones would. Since the assigned scores in the reads we receive are always going to be imperfect due to these issues, the BQSR process gets around this by using machine learning to build a picture of which base calls are likely to be inaccurate based on empirical data (115). BQSR is a two step process. In the first step, a model of how base quality scores covary across reads and across the genomes is built by running an algorithm on the data under examination and a set of known variants. This algorithm might identify, for example, that there is a 2% higher rate of error after any ACG trinucleotide in the datasets fed into it - therefore, any base called after such a nucleotide should have its quality score reduced by 2%. This model is stored as a “recalibration table” for the corresponding BAM/SAM file. In the second step, this recalibration table is used to adjust the base quality scores in the reads in the alignment file to more accurate scores.

With the BAM files fully prepared for variant calling, there were some quality control checks we ran on the files to ensure they were appropriate for use in variant calling. First, we ran the samtools “flagstat” command, which generated for each file data on what percentage of the reads were properly paired and what percentage of reads had successfully been mapped to the reference genome. Secondly, we examined the depth of coverage of each file using the GATK DepthOfCoverage tool. This tool outputs statistics about the depth of coverage in a SAM/BAM file, which is the average number of reads mapped to a region of interest in that SAM/BAM file. In general, the higher the depth of sequencing the more confidence we can have in the variants called downstream and the greater chance of discovering rare subclonal variants in the sample. This was particularly important to examine because the samples TCHL 3,6, 12, 29, 32, 39 and 45 had been re-sequenced to achieve a greater depth of sequencing, so it was important to establish that this effort had genuinely achieved a greater depth of sequencing.

Variant calling in this project was run using the GATK 4 Mutect 2 algorithm (crucially, this is a different algorithm to GATK 3 MuTect2). This is an algorithm specifically designed to call somatic variants (as opposed to a germline variant caller, like the GATK Haplotype Caller function). Like the Haplotype Caller function, Mutect 2 re-maps reads in regions of expected variation, allowing the simultaneous discovery of indels (insertion-deletion) and SNVs (single nucleotide variants; somatic single base changes). Unlike Haplotype Caller, Mutect2 allows for varying ploidy (ploidy = the number of sets of chromosomes present in a cell), in recognition of the fact that cancer cells often have an abnormal number of chromosomes present due to chromosomal duplication or loss (aneuploidy). For each sample, the Mutect2 algorithm is given as input the tumour BAM file, a "normal" (i.e. reads derived from non-tumour tissue) BAM file from the same sample, and a file containing a set of variant calls from a set of normal samples, referred to as the Panel of Normals (PON). Sites that are unambiguously variant in the normal file compared to the reference genome are removed from the callset, as these are likely germline variants rather than somatic variants. The PON is used to filter germline sites in the same way, and is also used to identify and remove apparently variant sites that are actually artefacts of sequencing and mapping error. Removing germline variants is especially important because "tumour" samples are often contaminated with DNA from nearby non cancerous cells, so it is important to remove such variants to make the final callset as accurate as possible.

The output of the Mutect is files in the Variant Call Format (VCF files). This is a standardized format for presenting information about small scale genetic variation (i.e. genetic variation other than chromosomal alterations and CNVs) that is now used in many large scale sequencing projects such as the 1000 Genomes project and is accepted as a standard input to many bioinformatics tools. VCF files, in addition to listing each called mutation, contain information about the quality of each mutation call (i.e. likelihood that it is true), as well as further metadata about each mutation call - including, importantly for our research, the allele frequency of the reference allele and the alternate allele in the reads (116).

The initial output VCF files from Mutect2 are “raw” files that have not yet been filtered for contamination. This 3-step process uses GATK tools to annotate those variants that are judged more likely to be the result of technical errors than genuine somatic variants. The variants filtered include variants that are of too low quality to be considered genuine (either due to base call quality or due to mapping quality), as well as variants that are likely to be a sequencing artefact in the normal file. The first step is running GATK GetPileUpSummaries on each tumour BAM file, which generates a “pileup table” of the number of reads supporting the reference, the alternate and other alleles for each site. The second step is running GATK CalculateContamination on each pileup table to calculate the fraction of reads coming from cross sample contamination and generate a “contamination table” for each sample with this data. The third and final step is to run GATK FilterMutectCalls, supplying the VCF and the corresponding contamination table for that sample as inputs to annotate all of the variant calls that fail the filtration tests.

Mutect2 allows us to uncover small scale genetic variation in the samples, but we also want to analyse whether copy number alterations (CNAs) are present in the samples. To this end, we use the R package FACETS (algorithm to implement Fraction And Copy number Estimate from Tumour/normal Sequencing). First, the snp-pileup command line function is run on the tumour and normal BAM files for each sample to generate a table of the counts, for each SNP found, of the reference nucleotide and the alternate nucleotide, as well as errors and deletions. This table is then used as input to the FACETS program, which outputs data about the CNAs observed in the samples. FACETS uses an allele specific copy number analysis approach which takes into account copy-neutral loss-of-heterozygosity events, thereby allowing more comprehensive CNA discovery. (117).

Analysis of Discovered Variants

In order to analyze the likely consequences of the mutations discovered, we used the Variant Effect Predictor (VEP) tool. This is a tool provided by Ensembl and written in the Perl language that annotates variant callsets provided to it with the likely functional consequence of each mutation, along with a prediction of how impactful the mutation is likely. In this project we assume that mutations with the potential to act as driver mutations are likely to be high impact mutations, as it is unlikely that a low impact mutation would alter cell biology radically enough to be able to contribute to a cancerous phenotype (118).

The Cancer Genome Interpreter (CGI) is an online resource that takes VCF or other mutation files and returns information about which mutations are definitively known to be driver mutations, as well as predictions about which mutations are likely to act as driver mutations, and which mutations are likely to be passenger mutations. The tool is usable for free as an API or via the web interface at http://www.cancergenomeinterpreter.org. The information about known cancer driver mutations is derived from a Catalogue of Cancer Genes that the creators of the tool made using information from manually curated literature, as well as information from international initiatives like The Cancer Genome Atlas (TCGA, described further below) and the International Cancer Genome Consortium (ICGC). The predictions about whether mutations of unknown driver status are drivers or passengers is made by the "OncodriveMUT" tool. This tool uses features like whether the mutation a gene appears in is a known oncogene or tumour suppressor, the location of a mutation in a transcript and the predicted functional impact of that mutation to assess the likelihood that a given mutation is a driver mutation. It has been shown to be 86% accurate at classifying validated drivers and passengers, out-performing other currently existing tools (119). In this project, CGI is used to interpret the output from Mutect2 to predict which of the small scale mutations discovered are drivers.

To analyze which of the CNAs discovered are likely to be drivers, we use Annovar. Annovar is a Perl based command line tool that can annotate input

mutational data with information from UCSC genome browser databases, or any databases that follow Generic Feature Format 3 (GFF3) standards (120).

SciClone is a software package for the programming language R that performs clustering analysis for individual tumour samples or pairs of samples in order to infer the clonal architecture of the tumour(s), which allows us to analyse the evolutionary history of the tumour(s). As input, SciClone takes Variant allele frequency (VAF) data from the sample(s) under examination, and the analysis can be further refined by including copy number variation (CNV) data (121). In this project we use both VAF data and CNV data as inputs to SciClone to build up the most accurate model possible of the clonal architecture and evolutionary history of the samples under examination.

Genomic Data Online Resources

The freely available web resources used in this project to get publicly available genomic data were DbSNP, the 1000 Genomes Project website, COSMIC and TCGA. They are described below.

DbSNP is a database of Single Nucleotide Polymorphisms (SNPs; germline single base genetic mutations) that appear often in the human population. It was created and is maintained by the NCBI. In this project it was used as a resource for information about regularly occurring germline SNPs, so that these mutations could be filtered from the final mutation call files to avoid the risk that the mutation caller had misidentified germline variants as somatic mutations (122).

The 1000 Genomes Project is an international research project aiming to build the most detailed and complete catalogue of human genetic variation that has been created to date via deep sequencing of the genomes of 1092 anonymous individuals spanning a number of different ethnic groups (123). In this project the publicly available data from the 1000 Genomes Project was used as a resource for high confidence SNP and indel callsets for the human population.

COSMIC (The Catalogue of Somatic Mutations in Cancer) is an online database hosting data about known driver mutations (URL:

https://cancer.sanger.ac.uk/cosmic). By 2015 it listed over 500 cancer genes. The known driver mutations in these genes included over 4 million coding mutations, 10 million non-coding mutations, 1 million copy number alterations (CNAs) and over 8 million epigenetic variants (124). Mutations on COSMIC are curated in order to ensure that the database contains as few technical artefacts and other mutations not genuinely associated with cancer as possible. Mutations are also annotated with a "functional impact score" assigned by the FATHMM-MLK algorithm (more information on the algorithm available here: http://fathmm.biocompute.org.uk/). Variants with scores below 0.5 are classified as "neutral". Variants with scores between 0.5 - 0.7 are classified as "deleterious". Variants with scores of 0.7 and above are classified as "pathogenic". The COSMIC dataset is used in this project as a resource to identify known driver mutations.

The final online resource used was TCGA (The Cancer Genome Atlas). TCGA is a project funded by the US government which began in 2006 and is co-ordinated by National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute. The project studied over 20,000 cancer samples with matched normal samples across 33 cancer types at a molecular level. The data from the project is now hosted online, and is used in this project as a resource to identify known driver mutations.

The final electronic resource needed for this project was a High Performance Computing (HPC) environment capable of hosting the very large files that store genomic data, and performing the sometimes computationally intensive bioinformatics jobs that needed to be done. The Irish Centre for High End Computing (ICHEC) is an Irish HPC organization, a national body headed by Nation University of Ireland, Galway. It provides High Performance Computing services to Irish Universities. Data processing for all samples described in this thesis was performed on the now de-activated ICHEC fionn server. Some data was stored on another server in University College Dublin (UCD) called the Alpen server. The ICHEC website is available at the following URL: https://www.ichec.ie/.

1.6- Mutational signatures

Phenomena that cause mutations in the human genome do not cause mutations in a random, unpredictable pattern. Instead, there is a particular pattern of mutations associated with a given mutagenic agent, due to how the chemistry of that agent interacts with the chemistry of the DNA in a genome. By examining the patterns of mutations apparent in human cancer samples, researchers have derived a set of “mutational signatures”, each “signature” describing a particular pattern of frequency of specific SNV mutations at specific trinucleotide contexts (e.g. a given signature would have a known frequency of T > A mutations at ATG sites in the genome) (24). The cited paper was the first to define mutational signatures in cancer by looking at whole genome sequences in all cancer types as opposed to examining only mutations in driver genes to define mutational signatures, or examining whole genome sequences of one specific cancer type. As noted in the cited paper, the whole genome approach is superior because when looking at a driver gene, the signal of a mutational signature is “jumbled” by signals of positive selection for driver mutations in that driver gene. The method of the cited paper was to calculate the relative frequencies of each possible SNV (C>A, C>G, C>T, T>A, T>C, T>G) at each possible trinucleotide context (by looking at the bases 3’ and 5’ to the mutated base) across the genomes of 21 cancer types, then feed this data into an algorithm they had developed to extract mutational signatures. The benefit of this algorithm was that it allowed the disentangling of signatures in cases where multiple signatures had affected the genome of a sample under examination. The algorithm had been previously validated on beast cancer whole genome sequences. Running this algorithm on those breast cancer whole genome sequences successfully recapitulated known mutational signatures, as well as showing some novel signatures (125) (126) .

The signatures discovered by the researchers in the cited paper ((24) were developed into a curated set of signatures now hosted on the website of an organization called the Catalogue of Somatic Mutations in Cancer (COSMIC, URL here: https://cancer.sanger.ac.uk/cosmic/signatures). Note that this project was undertaken while the COSMIC “Version 2” set of 30 signatures were still the most up to date set available. There is now a version 3 set of signatures,

which can be seen by following the URL. However as I cannot redo the entire project to work with the new signatures (they came out after the original, uncorrected thesis was submitted and I no longer have access to the original data), the presented thesis is still working with the older set of signatures.

Some signatures have known causes (e.g. Signature 4 is associated with tobacco smoke), while for other signatures the cause is not yet known (127). Signature 4 is shown below in Figure 1.8 to show an example of what a mutational signature looks like.

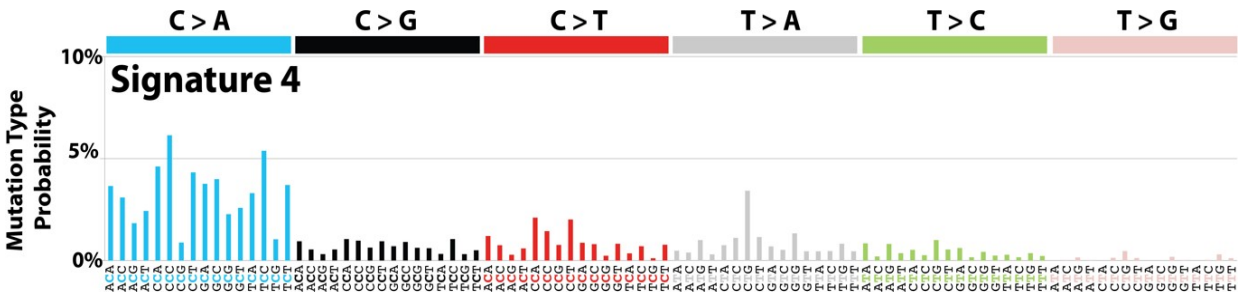


Figure 1.8 – Signature 4, the signature associated with tobacco smoke. "C>A" refers to the probability of a C base changing to an A base in the specified tri-nucleotide context. Image is from the COSMIC website, url: https://cancer.sanger.ac.uk/cosmic/signatures_v2

DeconstructSigs – The software for analyzing Mutational Signatures

In this project we used the R package deconstructSigs to find the frequency of known mutational signatures in the samples being examined. The input for DeconstructSigs is a file listing the set of somatic mutations present in a sample or samples, and the output is a table showing the extent to which the package predicts each of the known signatures contributed to causing the overall set of mutations present in that sample. To illustrate: if a sample were assigned a value of 30% for Signature 1 and 70% for Signature 7, this would mean that the overall set of somatic mutations in that sample were predicted to be 30% caused by Signature and 70% caused by Signature 7 – this would roughly correspond to 30% of the mutations being caused by Signature 1 and the other 70% being caused by Signature 7.

DeconstructSigs works by building a model of mutational frequencies by

"layering" known mutational signatures onto each other and finding the combination of mutational signatures that most closely approximates the SNV data presented to it (the package ignores non-SNV data). It is important to remember that this is how the program works when interpreting the results, as sometimes the results reflect the working of the algorithm rather than necessarily being an accurate representation of what signatures caused the mutations observed - for example, the aging based signature (signature 1) logically must have affected all of the samples as all cells have gone through the biological aging process, but DeconstructSigs sometimes assigns samples a 0% level of signature 1 because a combination of the other signatures got closer to the mutation frequencies observed rather than because the samples were actually unaffected by signature 1 (128).

In our analysis, as well as examining the prevalence of each signature individually, we also examine the results of grouping similar signatures together, under 5 groupings: Aging based signatures, APOBEC based signatures, DNA damage based signatures, Environmentally based signatures (e.g. the smoking signature, signature 4) and signatures of Unknown Aetiology (see the methods section for which signatures fall into which group). APOBEC (ApolipoPrOtein B mRNA Editing enzyme, Catalytic polypeptide-like) is a family of ssDNA cytidine deaminase proteins - they catalyse C -> U base changes via deamination of cytidine in single stranded DNA. Family members have various roles in the body, including aiding in the assembly of very low-density lipoproteins (129). The most common role shared by APOBEC proteins is protection against viral infection via alteration of the viral genome (130). However, when dysregulated, these APOBEC enzymes can cause mutations in the host DNA, thereby creating a distinct mutational signature capable of contributing to cancer development (131). "DNA damage based signatures" refers to signatures that appear in cells with deficiencies in the mechanisms for repairing DNA damage (e.g. deficiencies in double strand break (DSB) repair (132) or deficiencies in DNA mismatch repair (132). The meaning of the other signature groupings is self-explanatory.

The logic for grouping the signatures is that the signatures within a group share

the same or a similar biological cause, and since our main purpose in analyzing the signatures is to discover the biological causes of the driver mutations driving the cancers in these tumours, grouping the signatures in this way makes sense. It is also easier to interpret the relative prevalence of 5 groupings of signatures than it is to interpret the relative prevalence of 30 separate signatures.

1.7 -Objective of current study

List of core objectives:

1. To define the overall mutational landscape of a cohort of HER2 positive breast cancers taken from a population of Irish women

2. To define the mutational landscape of the Pre-treatment samples and look for commonalities in driver genes and mutational signatures prior to treatment starting

3. To compare the mutational landscape of the responders in the cohort to the non-responders

4. To investigate molecular evolution of matched pre- and post-treatment samples in patients

By calling and carefully filtering mutations from the deeply sequenced TCHL cohort, we aim to discover new driver genes and mutations that are responsible for the outcomes observed in the patients. We also aim to learn about the evolutionary history of the samples using the SciClone R package, and learn about the cause of the mutations present in the sample by examining the mutational signatures present using the deconstructSigs R package, in order to develop a holistic understanding of the biology of the samples. By learning about the driver mutations responsible for the cancers observed, the mutational signatures that caused these drivers to occur, and the way that these cells evolved through the selective pressures imposed by the therapy applied, we hope to make discoveries about cancer biology that will inform future therapeutic approaches.

2 Methods

All data processing was run on ICHEC except where otherwise indicated. The variant calling pipeline described is based on the GATK best practices pipeline from the GATK website (URL: https://software.broadinstitute.org/gatk/best-practices/). All settings used by the tools matched the settings recommended in the GATK best practice pipeline, except for specific cases which use different settings. These are described and explained below when they come up.

The GATK best practice pipeline was used as a starting point for designing the pipeline in this project because the GATK best practice pipeline is the standard in both industry and academia for how to call variants from raw sequencing data.

All data transfers between different computing environments (e.g. local drive to ICHEC) were performed using either the ftp or the sftp protocol and the accuracy of the transfer was validated using an md5 checksum each time. All of the reference files used, except where otherwise indicated, were downloaded from the GATK resource bundle.

The following steps were parallelised by multi-threading utilities built into the tools (i.e. multiple threads ran on the same task simultaneously to speed up processing):

-FastQC

-Trimmomatic

-Bwa

-Samtools “sort” command

The following steps were parallelised using the ICHEC taskfarm utility (each thread on a processor is given a separate task and all tasks are run simultaneously to speed up processing):

-MergeSamFiles (Picard Tools)

-MarkDuplicates (Picard Tools)

-CollectMultipleMetrics (Picard Tools)

-Both steps of Base Quality Score Recalibration:

    -BaseRecalibrator (GATK 4)

    -ApplyBQSR (GATK 4)

-DepthOfCoverage (GATK 3.5)

-Mutect2 (GATK 4)

-The three steps of Contamination filtering the called VCFs:

    -GetPileupSummaries (GATK 4)

    -CalculateContamination (GATK 4)

    -FilterMutectCalls (GATK 4)

2.1 TCHL sequencing data

The TCHL data analysed in this project was sequenced as paired end reads using Illumina sequencing technology by VIB Belgium. At this stage some of the samples were run in separate sequencing lanes (so that there were files with data from the same sequencing library but run on separate lanes). The data was a subsection of the overall TCHL project, introduced in the Introduction, section 1.3 (100) .

A subset of samples was selected for re-sequencing, again as paired end reads, at a higher depth by BGI China. These were from patients 3, 6, 12, 29, 32, 39, 45. Below is a table showing the samples associated with each patient included in this subset. The reason that only a subset of the patients were sequenced at a higher depth is that deeper sequencing is expensive and projects have budget constraints.

These higher depth sequencing samples were the main focus of this project, and so in-depth from these samples are analyzed in the main Results section 3.4. Analogous results for the other samples are shown in the Supplementary Materials section after the Bibliography at the end of this thesis. Table 2.1 on the next page summarizes the samples available for each of the patients whose samples were sent for higher depth sequencing, along with the relevant clinical data about these patients.

Table 2.1 – A table of the samples from the patients that were sent for re-sequencing in order to achieve higher depth of sequencing. In the “Treatment” Column, “TCHL” means that the patient was given lapatinib along with docetaxel, carboplatin and trastuzumab, while “TCH” means the patient was given docetaxel, carboplatin and trastuzumab but no lapatinib. “Non-responder” means that the patient did not respond to initial therapy.

Patient	**Samples**	**Treatment**	**Response category**
TCHL 3	AN, PRE, POST, Surgery	TCHL	Non-responder
TCHL 6	AN, PRE, POST, Relapse	TCH	Non-responder
TCHL 12	AN, PRE, POST	TCHL	Non-responder
TCHL 29	AN, PRE, POST	TCHL	Non-responder
TCHL 32	AN, PRE, Relapse	TCH	Responded to therapy initially. Later showed relapse in the brain
TCHL 39	AN, PRE, POST, Relapse	TCH	Non-responder
TCHL 45	AN, PRE	TCH	Responded to initial therapy

For all patients other than those in the table above, the data set consisted of a normal set (denoted AN) and a pre-treatment sample (denoted PRE). POST above just means a post-treatment sample. “Relapse” refers to samples taken from a patient whose cancer had relapsed after appearing to be in remission, and “Surgery” refers to samples from patients who had undergone surgery to try to remove the cancer.

The data from VIB Belgium was stored as FASTQ files on the UCD Alpen server. The data from BGI China (extra high depth sequencing samples) was received as a hard drive containing a set of FASTQ files.

Table 2.2 – Read lengths of the reads in the FASTQ files associated with each sample

Sample Name	Read length	Sample Name	Read length	Sample Name	Read length
eTCHL_11_AN	127	eTCHL_39_PRE	127	eTCHL_6_PRE	127
eTCHL_11_PRE	127	eTCHL3A_PRE	127	eTCHL_76_AN	127
eTCHL_12_AN	127	eTCHL_3_AN	127	eTCHL_76_PRE	127
eTCHL_12_POST	127	eTCHL3B	127	eTCHL_7_AN	127
eTCHL_12_PRE	127	eTCHL_3_POST	102	eTCHL_7_PRE	127
eTCHL_14_AN	127	eTCHL_42_AN	127	eTCHL_87_AN	127
eTCHL_14_PRE	127	eTCHL_42_PRE	127	eTCHL_87_PRE	127
eTCHL_20_AN	127	eTCHL_44_AN	102	eTCHL_8_AN	127
eTCHL_20_PRE	127	eTCHL_44_PRE	127	eTCHL_8_PRE	127
eTCHL_22_PRE	127	eTCHL_45_AN	127	TCHL 12 Post - Hard drive	151
eTCHL25AN	127	eTCHL_45_PRE	127	TCHL 12 Pre - Hard Drive	151
eTCHL_29_AN	127	eTCHL_4_AN	127	TCHL 29 Pre - Hard Drive	151
eTCHL_29_AN	127	eTCHL_4_PRE	127	TCHL 32C: Relapse - Hard Drive	151
eTCHL_29_POST	127	eTCHL_50_AN	102	TCHL 39 Pre - Hard Drive	151
eTCHL_29_PRE	127	eTCHL_50_PRE	127	TCHL 3A - Hard Drive	151

Sample Name	Read length	Sample Name	Read length	Sample Name	Read length
eTCHL_31_AN	127	eTCHL_54_AN	127	TCHL 3 Post Treatment - Hard Drive	151
eTCHL_31_PRE	127	eTCHL_54_PRE	127	TCHL 45 AN - Hard Drive	151
eTCHL_32_AN	127	eTCHL_61_AN	127	TCHL 6C : Relapse - Hard Drive	151
eTCHL32C	127	eTCHL_61_PRE	127	TCHL 6 Post - Hard Drive	151
eTCHL_32_PRE	127	eTCHL_66_AN	127	TCHL 6 Pre - Hard Drive	151
eTCHL_37_AN	127	eTCHL_66_PRE	127	TCHL 29 Post - Hard Drive	151
eTCHL_37_PRE	127	eTCHL_6_AN	127	TCHL 39 Post - Hard Drive	151
eTCHL_39_AN	127	eTCHL6CFF - Relapse	127		
eTCHL39C - Relapse	127	eTCHL_6_POST	102		

2.2 Pre -variant calling data processing

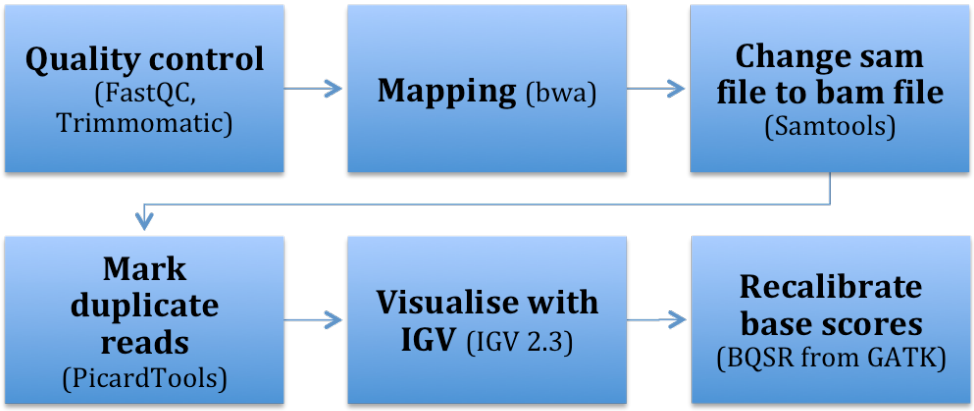


Figure 2.1 – The pre-variant calling workflow

These FASTQ files were uploaded to ICHEC. The workflow that these samples were put through is summarized above in Fig 2.1.

Data about the quality of the sequencing data in the files was obtained by running FastQC v0.10.1 on each file. To eliminate low quality base calls and adapter sequences, trimmomatic v0.27 was run on each of the FASTQ mate pairs with the following parameters:

"Phred 33, LEADING:20, TRAILING:20, SLIDINGWINDOW:4:20, MINLEN:36"

Once the reads were trimmed, FastQC was run on all of the trimmed files to examine the quality of the trimmed reads. Once the overall quality of the reads was deemed satisfactory, the reads were then aligned to Human reference genome 38 (Hg38) using bwa v0.7.5a-r405.

The resulting SAM files were processed using Samtools v1.5. The samtools "view" command with options "-b -h" was used to convert them to BAM files. These BAM files were then sorted with the samtools "sort" command and indexed with the samtools "index" command.

Duplicate marking was performed on the BAM files with the Picard Tools v1.1118 MarkDuplicates.jar command. At this stage the files that had been run on separate lanes but that were from the same original sequencing library were merged by putting all BAM files from the same sequencing library as inputs to a single MarkDuplicates run. This resulted in a single output of one duplicate marked bam file with data from all of the sequencing lanes that that sample had been run on. This also generated a duplication notes file for each BAM file. This information is shown in results section 3.1.

Base Quality Score Recalibration was then run on the Duplicate marked BAM files, using Hg 38 as the reference for all steps. This was achieved by running the following commands from GATK v4.0.4.0:

-First, RecalibrateBases was run on each BAM file to generate a recalibration table for that file. This tool was run with the following references supplied with the "--known-sites" options:

-dbsnp_146.hg38.vcf.gz

-Mills_and_1000G_gold_standard.indels.hg38.vcf.gz

-1000G_phase1.snps.high_confidence.vcf.gz

-ApplyBQSR was then run on each BAM file with the appropriate recalibration table supplied as an argument.

The final step was merging the BAM files from samples that were resequenced at a higher depth. Since these higher depth samples were sequenced entirely separately from the first run, it would not be appropriate to merge these at the marking duplicates stage, as was done for the other samples that had data split across several files. The reason for this is that higher depth resequencing involves the creation of separate sequencing libraries from separate PCR runs and matching reads from different libraries are not duplicates as a result. The appropriate BAM files were merged with PicardTools v1.118 MergeSamFile.jar command.

2.3 BAM File quality control

Once the BAM files were fully processed, the following tools were run as quality control checks:

The mapping rate of each BAM file was found by running the Samtools v1.5 “flagstat” command on each BAM file.

The sequencing depth of the files in the region of interest was found by running GATK 3.5 DepthOfCoverage on each of the processed BAM files (GATK 3.5 was used because the depth Coverage analysis tools had not yet been ported over to GATK 4). The analysis was limited to the region of interest by supplying a bed file to DepthOfCoverage with the “-L” option. The results of these depth analyses are shown in Tables 2.3 and 2.4 below. These results were placed in the Methods section rather than the Results section because the purpose of the depth information is demonstrate the validity of the data (higher depth corresponds to greater accuracy of variant calling) as opposed to the information that can be analysed to learn more about the biology of the tumours under examination.

Information from the “duplication notes” files generated for each sample when MarkDuplicates.jar was run on the samples was examined.

Table 2.3 – Mean sequencing depth of the samples from the patients from whom samples were taken at multiple timepoints after all files were merged

Patient	Pre	During	Relapse
TCHL 3	216	370	
TCHL 6	330	321	271
TCHL 12	297	226	
TCHL 29	223	357	
TCHL 32	193		227
TCHL 39	294	248	290

Table 2.4 – Mean Depth of the samples from the patients from whom samples were taken at a single timepoint after all files were merged.

Sample name	Mean Depth	Sample name	Mean Depth
eTCHL_11 PRE	84	eTCHL_50 PRE	42
eTCHL_14 PRE	148	eTCHL_54 PRE	106
eTCHL_20 PRE	88	eTCHL_61 PRE	76
eTCHL_31 PRE	83	eTCHL_66 PRE	81
eTCHL_37 PRE	73	eTCHL_7 PRE	75
eTCHL_4 PRE	59	eTCHL_76 PRE	91
eTCHL_42 PRE	89	eTCHL_8 PRE	97
eTCHL_44 PRE	114	eTCHL_87 PRE	39
eTCHL_45 PRE	167		

2.4 Variant Calling

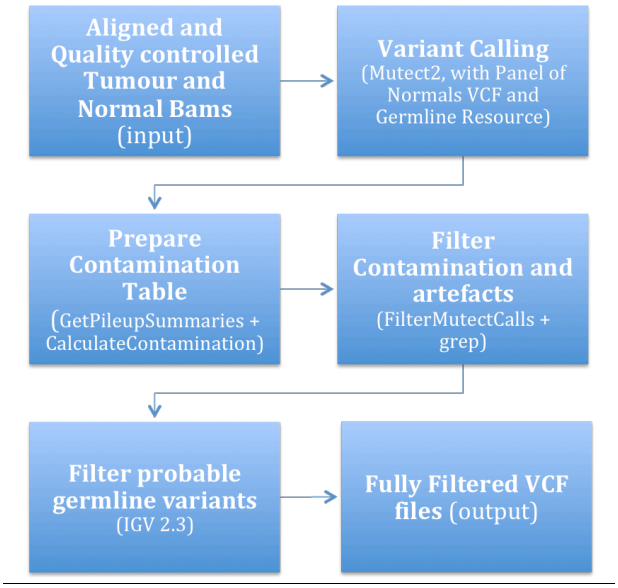


Figure 2.2 – Variant calling workflow

Figure 2.2 above summarizes the variant calling workflow we used once we had our aligned and quality controlled BAM files.

A “Panel of Normals” vcf file was generated from the data from 9 normal sample BAM files (the normal samples from TCHL 6, 12, 29, 32, 39 and 45 as well as 3 normal samples from a separate ongoing project in the same lab examining the impact of the drug Copanlisib on breast cancer). This was achieved by first using GATK 4 Mutect 2 on each of these samples in tumour-only mode with Hg38 as the reference genome. The output VCFs from this action were then used as inputs to the GATK CreateSomaticPanelOfNormals tools to generate a Panel of Normals (PON) vcf.

For each tumour sample, variants were called using GATK 4.0.4.0 Mutect 2 , inputting the corresponding AN (normal) file as the normal sample with Hg38 as reference, the above mentioned panel of normal vcf as the PON and the following additional parameters:

-Germline resource: af-only-gnomad.hg38.vcf.gz

-Allele frequency of alleles not in resource: 0.0000025 (in accordance with the Broad best practices guidelines)

To filter the files for contamination and sequencing artefacts, the following steps were run for each vcf generated in the variant calling stage (all of the following are tools from GATK 4.0.4.0):

-GetPileupSummaries was run on each tumour BAM file with “af-gnomad-only.hg38.vcf.gz” as the variant file with the common allele frequencies supplied with the “-V” option

-CalculateContamination was run on each of the pileups tables generated by GetPileupSummaries to generate a contamination table for each tumour BAM

-Finally, FilterMutectCalls was used to filter each VCF file with the contamination table from the corresponding tumour BAM file to generate a Contamination filtered vcf file. This represents the final stage of the Broad best practices pipeline. The below additional filtration steps are not part of Broad best practices and were added to be extra sure of our results.

The command line tool *grep* was then used to remove all variant calls from each file that did not pass the contamination filter (as the GATK contamination tools annotate the variant calls that fail the filter but do not actually remove them from the file).

To filter for potential germline variants, a file of common germline variants with common cancer mutations removed was prepared. The starting point for this was common_all_vcf.gz, downloaded from the NCBI dbSNP snp/organisms/human_9606/VCF directory on the 7th October 2017. To filter out mutations likely to be relevant to cancer from this file, two files were downloaded from the COSMIC website: a file containing known coding mutations relevant to cancer and a file containing known non-coding mutations relevant to cancer.

Bedtools v2.21.0 was used to filter the original “common_all_vcf.gz” file to generate the file Db_SNP_HG38_FullFiltered.vcf by using the “intersect -v command” to remove any mutations that appeared in either of the COSMIC files from the output file.

For each of the TCHL VCFs, the bedtools intersect -v command was then used to remove any mutations that appeared in the Db_SNP_HG38_FullFiltered.vcf file.

2.5 Phylogenetic analysis

The “snp-pileup” tool from the htstools library was run on each tumour-normal pairing of BAM files, with dbsnp_146.hg38.vcf.gz as the reference vcf. This generated a table of read counts for the reference and alternate allele for each position in each tumour-normal pair. These tables were used as inputs to the R package Facets v0.5.11. For each table, the Facets “emcncf” function generated files giving information about the CNV in that sample.

Python scripts were used to extract the CNV information from the Facets output and the Variant allele frequency (VAF) information from the fully processed VCFs, respectively. These data were then used as inputs for the R package SciClone v1.1.0. For each individual tumour sample, the “sciClone” function was run on that sample with the minimumDepth argument set to 50 to perform clustering analysis of the mutations present, and the “sc.plot1d()” function was used to generate a graph summarising the Variant allele frequency information about that sample. These graphs are shown in results section 3.4 for the samples from patients who had samples taken at multiple timepoints, and in the Supplementary Data for samples from the other patients.

For each patient with tumour samples from different timepoints, the “sciClone” function was also run pairwise with each pair of samples as input (e.g. for patient 3, samples PRE & POST, PRE & Surgery, and POST & Surgery were all run as pairs). The output of Sciclone in this case is data about the Variant Allele Frequencies (VAF) in the subclonal populations of cells shared by the two paired samples. The advantage of this is if the VAFs are reduced in a later sample, this suggests that therapy is successfully killing cells with that variant at a higher rate than cells with other variants, while if the VAF is increased in a later sample, this suggests that that subclonal population is surviving the therapy better than the other subclonal populations in the tumour.

For each of these sample pairing, the function “sc.plot2d” was used to generate a graph of the VAFs of the clusters identified in both samples, allowing for phylogenetic analysis. These pairs are shown and discussed in results section 3.5.

2.6 Mutational Signature analysis

To analyse the mutational signatures, analysis was performed in R 3.4.3 using the package deconstructSigs v1.8.0 . Each fully processed TCHL vcf was fed into R, then processed with the deconstructSigs functions “mut.to.sigs.input” followed by “whichSignatures” to find the mutational signatures present in that sample. The R package dplyr v0.7.4 was used to add extra columns to the dataframe generated by deconstructSigs for each sample. These extra columns represented groupings of the Signatures, as shown below:

Aging Signature = Signature 1

APOBEC Signatures = Signature 2, 13

DNA Damage Repair Mechanism Failure Signatures = Signature 3,6,9,10,15,20,26

Unknown Aetiology Signatures = Signature 5,8,12,14,16,17,18,19,21,23,25,27,28,30

Environmental Signatures = Signature 4,7,22,24,29

For each sample, the R package ggplot2 v2.2.1 was used to make a bar graph representing the proportion of each signature and each signature grouping present in the sample with the “geom_col()” function.

These plots are shown in the results section for the samples from patients who had samples taken at multiple timepoints, and in the supplementary data section for the samples from patients who had samples taken at only one timepoint.

2.7 Driver Gene Predictions

The Variant Effect Predictor (VEP) tool was downloaded from ensembl, along with “homo_sapiens_vep_93_GRCh38.tar.gz” to act as a cache directory for VEP. VEP was run on each fully processed TCHL VCF to annotate the likely effect of each variant observed in that sample.

To analyse which of the SNVs and indels present were likely to act as drivers, the Cancer Genome Interpreter (CGI) was used. Each VCF file was remapped to hg19 genomic coordinates using the crossmap tool, because the CGI only works with hg 19 files. Each vcf was submitted in turn to the CGI with the cancer type set to “Breast Adenocarcinoma (BRCA)”. Summary heatmaps of the mutations that were predicted to act as drivers are shown in the Results section. Full tables of the results from CGI are shown in the supplementary data section.

2.8 –Statistical Tests

Some statistical tests were run on the data to test whether certain factors (e.g. the presence of certain mutations) had a statistically significant effect on responder status and relapse status. Any statistical test of this type has a “null hypothesis” – essentially a hypothesis that there is no significant difference between the populations under examination (e.g. responder and non-responder samples), and that any differences observed in e.g. the frequency of mutations in certain genes between the two cohorts is due to the random chance inherent to sampling rather than genuine difference in the underlying population. The output of the tests is a “p-value” that tells you the probability of observing results at least as extreme as those observed if there is no genuine difference between the underlying populations being examined (so p =0.05 means there is a 5% chance you would see results at least as extreme as those input to the test if there was no genuine difference).

Fisher’s Exact test is a test used to examine if there is a non-random association between two categorical variables. In this Thesis, the categorical variables are presence/absence of mutations in a given gene, and Responder status, or Relapse status. The null hypothesis of Fisher’s Exact test is that the association between the two categorical variables is random (133).

The other statistical test used is the 2 sample, unpaired t-test. The null hypothesis of this t test is that the underlying populations behind the two sample populations being compared have the same mean average. It is used to compare the mean averages of two separate sets of samples (134). In this thesis it was used to compare the mean average number of subclones in the responder cohort compared to the non-responder cohort. Do note that the two sample unpaired t test assumes that both underlying populations that the samples are drawn from are normally distributed.

The final thing to note in this section is that a statistical test not reaching statistical significance (defined in this thesis as $p < 0.05$ because that is the standard in most scientific literature (135) does not mean the association that the test was investigating does not exist – it simply means that the signal of the association in this dataset was weak enough that the association seen (for example, between responder status and a certain mutation) may have been due to chance rather than due to genuine biological phenomena (135). In general, larger sample sizes have

greater power to discover statistically significant associations.

3 Results

The results section is divided into 5 parts – firstly, a summary of the mutation counts in each sample in the cohort. Secondly, heatmaps of the presence of mutational signatures and the presence of known or predicted driver mutations in specific genes in the pre treatment samples only, to allow us to see the mutational landscape of the cohort prior to the beginning of treatment. Thirdly, heatmaps of the same data for all samples divided into Responder samples and non-Responder samples, to allow comparison of the mutational landscape of samples that respond to therapy and those that do not. This section also includes a boxplot and simple statistical analysis of the mutation counts of the samples in the Responder cohort compared to the Non-Responder cohort. The fourth section consists of an in depth analysis of each of the samples from the patients that had samples taken at multiple different timepoints. This analysis is used to build a picture of what likely happened to each of these tumours over the course of therapy. The fifth and final section is the SciClone analysis – the comparison of samples taken from the same patient at different timepoints by the SciClone algorithm. This is taken in the context of the previous analysis of the mutational landscape of each of the samples to round out the predictions about what happened over the course of therapy for each sample.

3.1 Cohort summary

The table below shows the number of SNVs and indels per sample after the called variants were filtered as described in the methods section. We can see that in every sample SNVs are more frequent than indels, and that there is a wide range of possible number of mutations in a given sample (for SNVs, a minimum of 13 and a maximum of 2252, for indels a minimum of 3 and a maximum of 2170). According to the literature, a tumour with a mutation rate greater than 10 mutations per megabase has a “high” mutational burden (136), and the human exome cover 30-50 megabases. Therefore, since these tumours were exome sequenced, tumours that show >(300-500) mutations, SNVs and indels combined, are tumours with a high mutational burden. We can see that many of the tumours examined fall into this category, suggesting that this was a particularly highly mutated cohort.

SD in the following paragraph refers to Standard Deviation.

Pre treatment samples show a mean average of 693.5 SNVs (SD=560) and 287 indels (SD = 432). In comparison, post treatment samples show a mean average of 1026 SNVs (SD=592) and 749 indels (SD=605). The fact that the post treatment samples show a higher average mutation burden for both SNVs and indels suggests that cancer cells with a heavier mutation burden are more likely to survive therapy. It is also worth noting the very high standard deviation scores, which emphasizes how wide a spread there is of mutational burden in this cohort.

Table 3.1.1 – SNV, indel and sequence alteration counts per sample, based on running the Ensembl Variant Effect Predictor on each sample:

	SNVs	Indels
8 Pre-treatment	285	56
87 Pre-treatment	323	134
7 Pre-treatment	304	46
76 Pre-treatment	1750	98
66 Pre-treatment	113	24
61 Pre-treatment	395	39
54 Pre-treatment	384	71
50 Pre-treatment	290	38
4 Pre-treatment	471	296
44 Pre-treatment	600	89
42 Pre-treatment	844	142
37 Pre-treatment	647	62
31 Pre-treatment	295	64
25A Pre-treatment	13	3
20 Pre-treatment	1858	275
14 Pre-treatment	419	107
11 Pre-treatment	376	37
6 Pre-treatment	997	694
6 Post Treatment	470	563
6CFF – Relapse	998	696
45 Pre-treatment	558	66

3 Post Treatment	1052	854
3B – Surgery	570	136
3 Pre-treatment	1422	1338
39 Pre-treatment	873	809
39 Post Treatment	1356	752
39 Relapse	478	71
32 Pre-treatment	357	40
32 Relapse	1460	2170
29 Pre-treatment	1033	1565
29 Post Treatment	599	863
12 Pre-treatment	2037	805
12 Post Treatment	2252	633

3.2 Mutational landscape – Pre-treatment samples only

Below are heatmaps (Figures 3.2.1 and 3.2.2) summarizing the predicted driver SNVs and indels, as well as the mutational signature frequencies, for the pre-treatment samples only.

In regards to the putative driver mutations, TP53 and PIK3CA mutations are very common across the cohort of pre-treatment samples. Outside of those 2 genes however, there is little commonality of the known/predicted driver mutations across the cohort. Most of the known/predicted driver mutations encountered were in a gene that showed only one known/predicted driver mutation across the entire cohort.

In regards to the mutational signatures, there are only weak signals seen from any individual signature. However, when the signatures are combined into the groupings explained in the introduction, some of the samples show a strong signal from DNA-damage based signatures, and a smaller number of samples show a signal from APOBEC related signatures. This suggests that DNA-damage based processes and APOBEC dysregulation were important in the initial oncogenesis in these cancers. In contrast, there is very little signal from environmental signatures, suggesting that the environmental processes with known signatures played little to no role in driving these cancers.

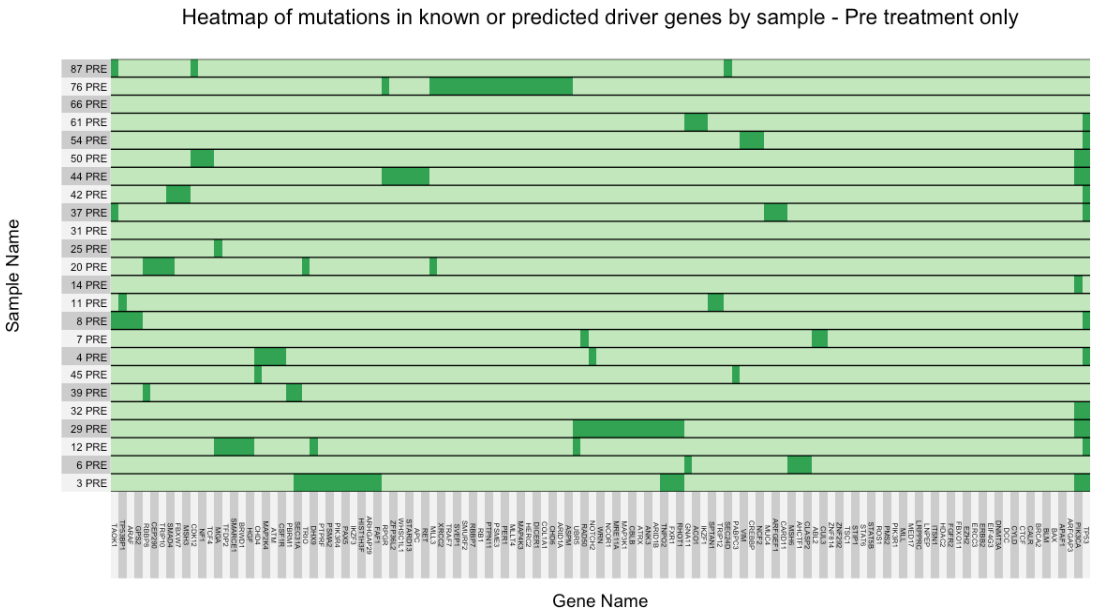


Figure 3.2.1 – Predicted or known driver indels and SNVs, across all the Pre-treatment samples. Dark green indicates that a predicted driver SNV or indel is present in that gene in that sample.

Figure 3.2.2 – Mutational signature heatmap – Pre-treatment samples only. First 30 columns are the individual signatures, final 3 columns are the signatures grouped.

3.3 Mutational landscape – Responders vs Non-Responders

Below are tables and heatmaps summarising the predicted driver SNVs, indels, as well as the mutational signature frequencies, for the full cohort, divided into responders and non-responders to initial therapy. These results allow us to see how the mutational landscape of tumours affects their likelihood of responding to TCH or TCHL therapy.

Table 3.3.1 – Table of which samples were responders and which were non-responders to therapy (i.e. which samples showed pCR and which did not)

Responders	**Non-responders**
TCHL 8	TCHL 3 (Pre, Post, Surgery)
TCHL 14	TCHL 4
TCHL 25	TCHL 6 (Pre, Post, Surgery)
TCHL 31	TCHL 7
TCHL 32 (Pre, Relapse)	TCHL 11
TCHL 37	TCHL 12 (Pre, Post)
TCHL 44	TCHL 29 (Pre, Post)
TCHL 45	TCHL 39 (Pre, Post, Relapse)
TCHL 61	TCHL 42
TCHL 66	TCHL 50
TCHL 76	TCHL 54
	TCHL 87

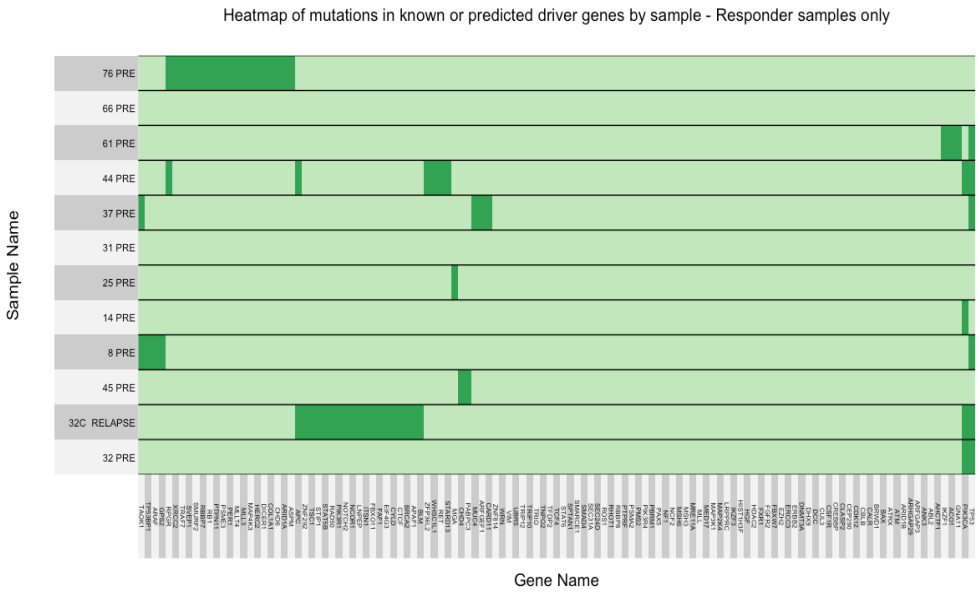


Figure 3.3.1 – Predicted or known driver indels and SNVs, across all the responder samples. Dark green indicates that a predicted driver SNV or indel is present in that gene in that sample

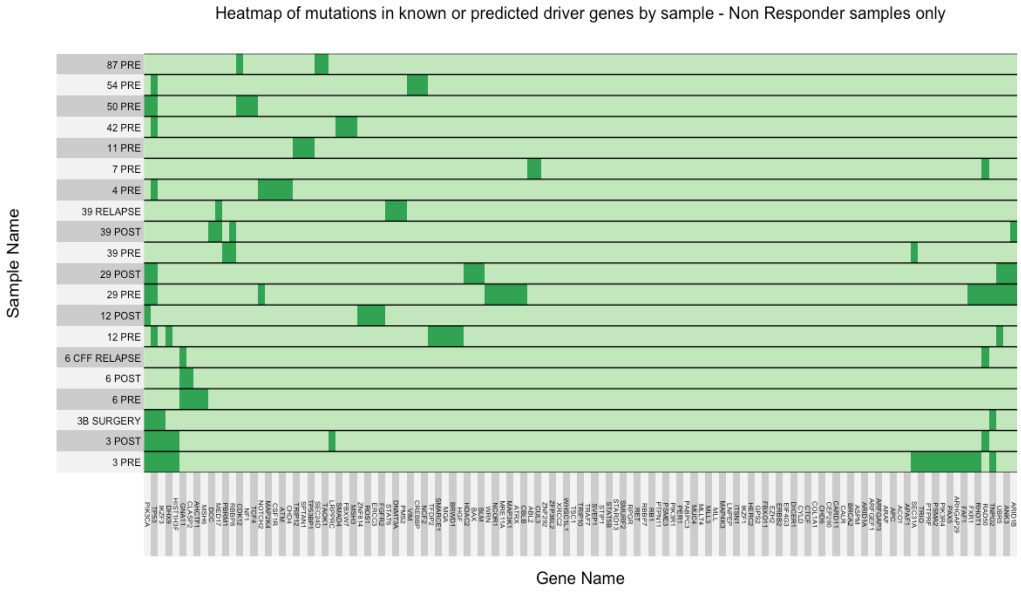


Figure 3.3.2 – Predicted or known driver indels and SNVs, across all the Non-Responder samples. Dark green indicates that a predicted driver SNV or indel is present in that gene in that sample

Below is Table 3.3.2, showing the frequency of known/predicted driver mutations by gene in the Responder and Non-Responder cohorts. The main take away is that most of the tumours, both intra-group and inter-group, are not very similar to each other in terms of the genes that known/predicted driver mutations appear in. Many of the genes appear as a potential drivers only once across the entire cohort. It is hard to build a coherent picture from this data of what the difference between driver genes in responders compared to non-responders is. This is still an informative result, because it shows us the complexity of cancer genetics in this regard.

Later on in this series there are a series of contingency tables analyzing whether mutations in any of these genes show a statistically significant association with either the Responder or non-Responder cohort.

Table 3.3.2 – Table of the frequency of known or predicted driver SNVs and indels by gene in the responder and non-responder cohorts. Frequencies were calculated by dividing the number of samples within a group with at least one mutation in that gene by the total number of samples in that group (Responder or Non-Responder). Table is split in two simply because it was impossible to fit all of the data into one page.

Gene Name	Responders	Non-Responders	Gene Name	Responders	Non-Responders	Gene Name	Responders	Non-Responders
ABL2	0	0.05	*FBXW7*	0.07692308	0	*PTPRF*	0	0.05
ACO1	0.07692308	0	*FGFR2*	0	0.05	*RAD50*	0.07692308	0.2
AHCTF1	0	0.05	*FXR1*	0	0.1	*RB1*	0.07692308	0
ANK3	0	0.1	*GNA11*	0.07692308	0.15	*RBBP7*	0.07692308	0
APAF1	0.07692308	0	*GPS2*	0.07692308	0	*RBBP8*	0	0.1
APC	0.15384615	0	*HDAC2*	0	0.05	*RET*	0.07692308	0
ARAF	0.07692308	0	*HERC2*	0.07692308	0	*RHOT1*	0	0.1
ARFGAP3	0	0	*HGF*	0	0.05	*ROS1*	0	0.05
ARFGEF1	0.07692308	0	*HIST1H3 F*	0	0.1	*RPGR*	0.15384615	0
ARHGAP29	0	0.05	*IKZF1*	0.07692308	0	*SEC24D*	0	0.05
ARID1A	0.07692308	0	*IKZF3*	0	0.15	*SEC31A*	0	0.1
ARID1B	0	0.15	*ITSN1*	0.07692308	0	*SMAD4*	0.07692308	0
ASPM	0.07692308	0	*LNPEP*	0.07692308	0	*SMARCE 1*	0	0.05
ATM	0	0.05	*LRPPRC*	0	0.05	*SMURF2*	0.07692308	0
ATRX	0	0.05	*MAP2K4*	0	0.05	*SPTAN1*	0	0.05
BAX	0	0.05	*MAP3K1*	0	0.05	*STARD13*	0.07692308	0
BLM	0.07692308	0.05	*MAP4K3*	0.07692308	0	*STAT5B*	0.07692308	0
BRCA2	0.07692308	0	*MED17*	0	0.1	*STAT6*	0	0.05
BRWD1	0	0.05	*MGA*	0.07692308	0.05	*STIP1*	0.07692308	0
CLASP2	0	0.1	*NCF2*	0	0.05	*TRAF7*	0.07692308	0

Gene Name	Responders	Non-Responders	Gene Name	Responders	Non-Responders	Gene Name	Responders	Non-Responders
CTCF	0.07692308	0	*PABPC3*	0.07692308	0	*TSC1*	0.07692308	0
CUL3	0	0.05	*PAX5*	0	0.05	*UBR5*	0	0.15
CYLD	0.07692308	0	*PBRM1*	0	0.05	*VIM*	0	0.05
DCC	0	0.05	*PER1*	0.07692308	0	*WHSC1L1*	0.07692308	0
DHX9	0	0.15	*PIK3CA*	0.30769231	0.4	*WRN*	0	0.05
DICER 1	0.07692308	0	*PIK3R1*	0.07692308	0	*XRCC2*	0.07692308	0
DNMT3 A	0	0.05	*PIK3R4*	0	0.05	*ZFP36L2*	0.07692308	0
EIF4G3	0.07692308	0	*PMS2*	0	0.05	*ZNF292*	0.07692308	0
ERCC3	0	0.05	*PSMA2*	0	0.05	*ZNF814*	0	0.05
FAF1	0.07692308	0.05	*PSME3*	0.07692308	0			
FBXO1 1	0.07692308	0	*PTPN11*	0.07692308	0			
COL1A1	0.07692308	0	*NCOR1*	0.07692308	0.05	*TRIO*	0	0.05
CREBBP	0	0.05	*NF1*	0	0.05	*TRIP10*	0	0
CSF1R	0	0.05	*NOTCH2*	0.07692308	0.1	*TRIP12*	0	0.05
CALR	0	0	*MLL*	0	0	*SVEP1*	0.07692308	0
CARD11	0.07692308	0	*MLL3*	0.07692308	0	*TAOK1*	0.15384615	0.05
CBLB	0	0.05	*MLLT4*	0.07692308	0	*TCF4*	0	0.05
CDK12	0	0.1	*MRE11A*	0	0.05	*TFDP2*	0	0.05
CEP290	0	0	*MSH3*	0.07692308	0	*TNPO2*	0	0.15
CHD4	0.07692308	0.05	*MSH6*	0	0.05	*TP53*	0.53846154	0.45
CHD6	0.07692308	0	*MUC4*	0.07692308	0	*TP53BP1*	0.07692308	0.05

Below are Figures 3.3.3 and 3.3.4, showing the mutational signature burdens found in the Responder and non-Responder cohorts respectively. In the Responder cohort most of the individual signatures show a weak presence, but when the signatures are grouped there is a significant presence of DNA damage based signatures and a smaller but still notable presence of signatures of unknown aetiology. One sample (TCHL 76) shows a very strong signal of APOBEC based signatures.

In the Non-responder cohort, there is a much greater signal from the ageing signature compared to the Responder cohort. The other notable difference is that there is a stronger signal from DNA damage based signatures in the Non-Responder cohort compared to the Responder cohort. This suggests that non-response to therapy may have been partially caused by DNA damage based mechanisms – we speculate this may be due to increased DNA damage weakening the cell's own defenses against cancer (e.g. by inactivating tumour suppressor genes), making it less likely for therapy to be effective. However for the time being this is just speculation and it would require a study focused on this question to scientifically validate this hypothesis.

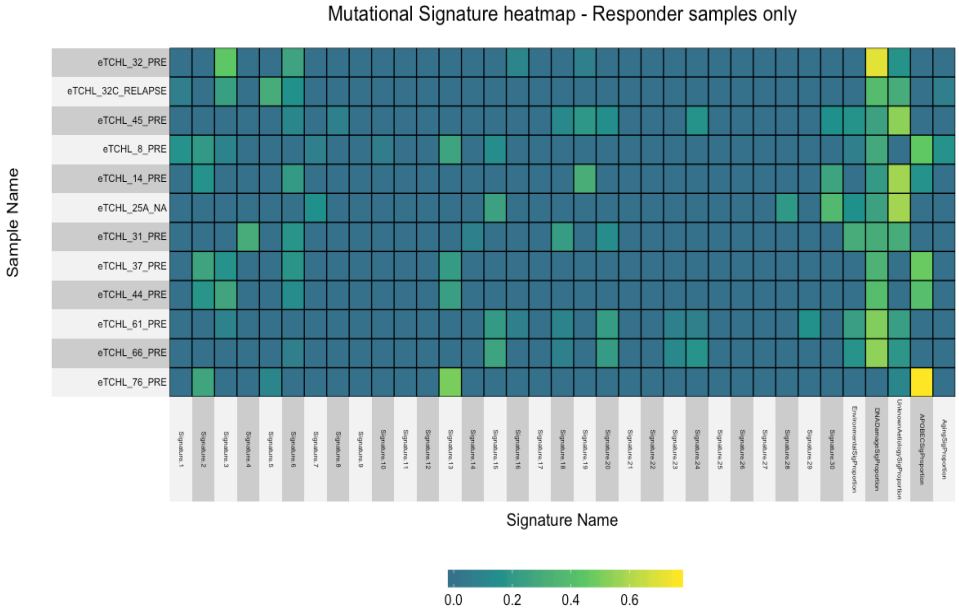


Figure 3.3.3 – Mutational signature heatmap – Responder samples only

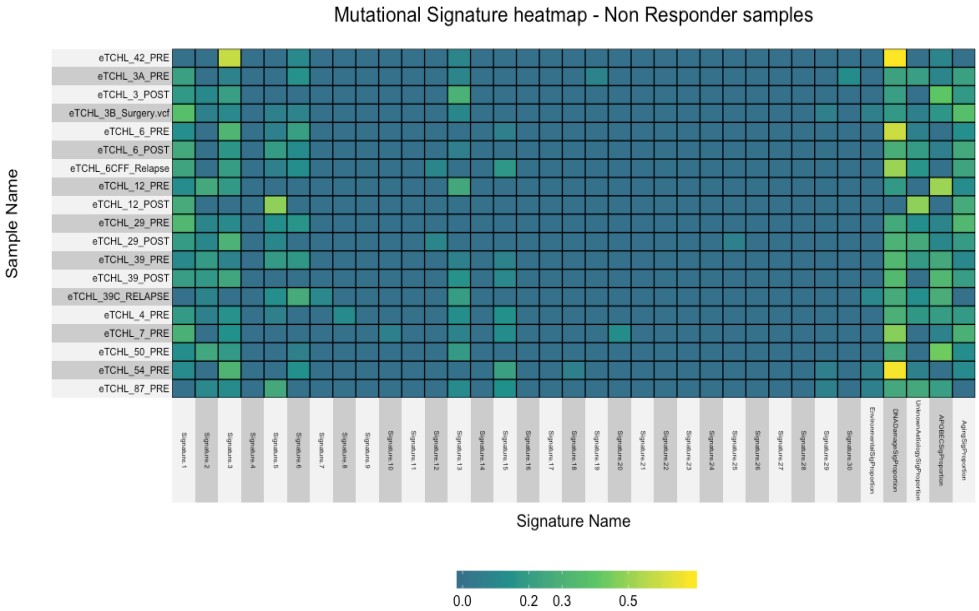


Figure 3.3.4 – Mutational signature heatmap – Non-responder samples only

Tables 3.3.3 and 3.3.4 below show the frequency with which a given gene shows a predicted/known driver amplification(s) (3.3.2) or deletion(s) (3.3.3) in each of the Responder and Non-responder cohorts, respectively. The frequencies differ little between the two cohorts, so it is unlikely that any of these mutations are responsible for the difference between the two cohorts.

Table 3.3.3 – Frequency table of the proportion of samples in each of the Responder and Non-Responder cohorts showing predicted driver amplifications in that gene.

	Responders	**Non-Responders**
AHRR	0.583333333	0.571428571
AVPR1B	0.75	0.904761905
C6orf203	0.5	0.428571429
ERBB2	0.916666667	0.952380952
FCAMR	0.75	0.904761905
GRHL2	0.75	0.80952381
NUAK2	0.75	0.904761905
PCK1	0.5	0.666666667
RNF182	0.583333333	0.714285714
RPRD2	0.75	0.904761905
STAR	0.25	0.285714286
ZNF703	0.25	0.238095238

Table 3.3.4 – Frequency table showing the proportion of samples in each of the Responder and Non-Responder cohorts showing predicted driver deletions in that gene

	Responders	**Non-Responders**
AHRR	0	0.04761905
AVPR1B	0	0
C6orf203	0	0.04761905
ERBB2	0	0
FCAMR	0	0
GRHL2	0	0.04761905
NUAK2	0	0
PCK1	0	0
RNF182	0	0
RPRD2	0	0
STAR	0.08333333	0.14285714
ZNF703	0.08333333	0.14285714

Below is Table 3.3.5, showing the number of subclonal populations in each sample (based on the number of clusters in the SciClone analysis for that sample), divided into responders and non responders. Each “subclonal population” represents a genetically distinct population of cells within the tumour – for this to happen, there must be subclonal driver mutations within the tumour (see Introduction section 1.2). These subclonal driver mutations can cause uncontrolled cellular proliferation independently from the drivers in the original tumour. This is relevant to therapy because if therapy kills cells with the original mutations, but does not eradicate cells with cancer driven by the subclonal mutations, those subclonal cancerous cells will simply colonize the space left over by the original cancer and the patient will still have cancer. In this way, subclonal populations can cause a relapse in a patient who initially appeared to be cured. The presence of subclonal driver mutations is known to adversely impact clinical outcome (137)

(Cells that show “NA” are ignored in the following analysis). The responders show a mean average of 1.75 subclones, in contrast to an average of 2.2105 in the non-responders. This is in line with the literature cited above showing that subclonal cellular populations are a poor prognostic factor. The p-value of a two sample t-test performed in R using the “t.test” function from the “stats” package in base R is 0.3566, showing that the difference between the two sample means is not statistically significant. However, this is a fairly small cohort, and the fact that the means are different at face value, along with the fact that greater genetic diversity is known to be associated with greater chance of a tumour surviving therapy (as mentioned in the Intro section), suggests that it may be worth studying the association between the number of subclones and response to therapy status in larger cohorts in future.

The relapse samples show a mean average of 3.5 subclones, while Pre treatment samples show a mean average of 2 subclones, and the Post treatment samples show a mean average of 1.8333. The relapse samples having a substantially higher number of subclones than the other groupings is in line with the logic above about subclones potentially causing relapse (the more subclones, the higher the chance that one will be capable of driving a relapse). The post treatment samples having a lower average number of subclones than post treatment samples suggests that the therapy process may, on average, lower the number of subclones in a sample (which

makes logical sense since the point of therapy is to kill the cancerous cells).

Table 3.3.5 – Table of the number of subclonal populations in each sample (based on the number of clusters in the SciClone analysis for that sample), divided into responders and non responders. Where SciClone was unable to analyse the sample, the entry is “NA”.

Responders	**Number of subclones**	**Non responders**	**Number of subclones**
32 PRE	NA	3 PRE	2
32C RELAPSE	4	3 POST	1
45 PRE	2	3B SURGERY	1
8 PRE	2	6 PRE	2
14 PRE	1	6 POST	2
25 PRE	NA	6 CFF RELAPSE	3
31 PRE	1	12 PRE	3
37 PRE	1	12 POST	3
44 PRE	1	29 PRE	2
61 PRE	NA	29 POST	1
66 PRE	NA	39 PRE	1
76 PRE	2	39 POST	3
		39 RELAPSE	NA
		4 PRE	2
		7 PRE	2
		11 PRE	1
		42 PRE	7
		50 PRE	3
		54 PRE	2
		87 PRE	1

Below are boxplots and simple statistical analysis of the numbers of SNVs and indels in each sample of the cohort, split into Non-Responders and Responders. These results should be considered with the caveat that there are more Non- Responders with higher depth sequencing, which is likely to reveal more mutations. However even with this caveat, we can see that there are substantially more mutations in the non-responder samples. This suggests that a higher mutational burden gives a tumour a greater chance of surviving therapy. A possible mechanism for this is that having more mutations raises the chance that a tumour will have a mutation that will be protective from therapy. This is in line with what we saw in Results Section 3.1, where responder samples showed a higher mutational burden. It is also in line with the Sciclone data above, as a tumour with more subclones (remember that non-Responders on average have more subclones) are logically likely to have more mutations overall than a tumour with fewer subclones (as each subclonal population will have its own distinct set of mutations). There is one outlier of an extremely highly mutated responder sample – this is sample 32C, the relapse sample from patient 32. This patient is annotated as a responder because their tumour initially responded, but unfortunately later suffered a relapse. The most likely explanation is that a mutagenic process active during therapy caused a host of new mutations that drove their cancer to relapse despite the initial therapy response – see section 3.4.5 for more details.

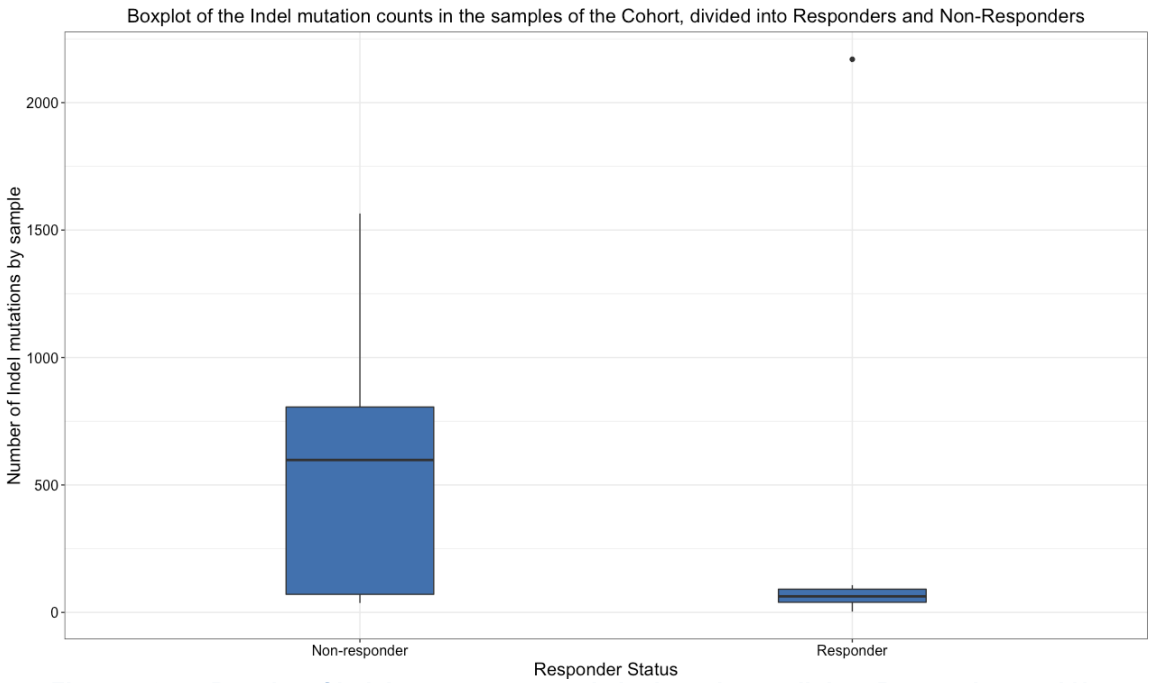


Figure 3.3.5 – Boxplot of Indel counts per sample in the cohort split into Responders and Non-Responders

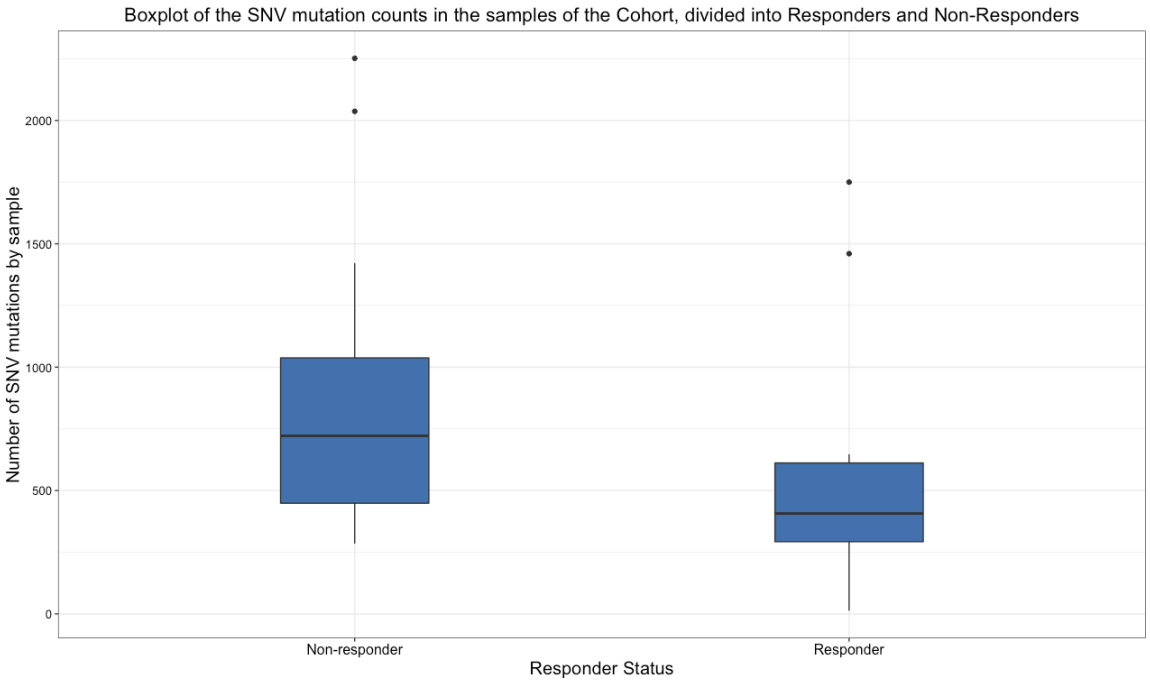


Figure 3.3.6 – Boxplot of SNV counts per sample in the cohort split into Responders and Non-Responders

Below is table 3.3.6, showing simple statistical analysis of mutation counts in the cohort, segregated into responder and non-responder samples. The results are in line with the other results examined above, in that we see a lower average mutational burden in the responder samples compared to the non-responder samples. An unpaired t test shows gives a p-value of 0.167285 for the association between SNV counts and responder status. The same test for indels gives a p value of 0.13068. So neither association reaches statistical significance at a p-value <0.05.

Table 3.3.6 – Simple statistical analysis of the mutation counts in the cohort, split into Responders and Non-Responders

	Mean	Median	Interquartile Range	Standard Deviation
Responders: SNV counts	574	407	319	519
Responders: Indel counts	235	63	51	610
Non-Responders: SNV counts	855	722	589	563
Non-Responders: Indel counts	523	598	735	455

Below are a set of contingency tables, represented on Table 3.3.7, used to run Fisher's Exact tests to test the statistical signifigance of the association between the presence/absence of predicted/known driver mutations in the gene in question and Responder status. The genes selected for this analysis were all genes that show predicted driver mutations in at least one sample in both cohorts or where genes showed over 3 predicted driver mutations in either the responder or the non-responder cohort. P-values were obtained by performing Fisher's exact test in R using the "fisher.test" function from the "stats" package from base R. The null hypothesis is that the presence or absence of mutations in a given gene is unrelated to whether a sample responds to therapy or not.

The name of the gene each table corresponds to is shown in the column on the left. Fisher's exact test is used rather than a chi square test due to the small sample sizes making tests other than Fisher's exact test inappropriate to use with this data. Genes that had identical counts were combined into the same row because it's the same calculation every time and doing it this way takes up less space. The percentages refer to what percent of that row the number represents (e.g. 50% of Responders carry a TP53 mutation).

None of the associations is significant at a p-value > 0.05. This may indicate that there is no genuine biological association between mutations in any of these genes and Responder Status, and that the differences observed are due to chance. On the other hand, it may indicate that the cohort under examination is too small to distinguish the signal from such an association from statistical noise.

Table 3.3.7 – Contingency tables used to run Fisher's exact tests.

Gene name	Response category	Samples with gene mutated	Samples without gene mutated	P-value
TP53	Responders	6 (50%)	6 (50%)	1
	Non-responders	10 (48%)	11 (52%)	
PIK3CA	Responders	4 (33.3%)	8 (66.6%)	1
	Non-responders	7 (33.3%%)	14 (66.6%)	
RAD50	Responders	1 (8.3%)	11 (91.6%)	0.6301
	Non-responders	4 (19%)	17 (81%)	
GNA11	Responders	1 (8.3%)	11 (91.6%)	1
	Non-responders	3 (14.29%)	18 (85.7%)	
ARIB1B, DHX9, IZF3, TNP02, UBR5	Responders	0 (0%)	12 (100%)	0.2841
	Non-responders	3 (14.29%)	18 (85.7%)	
NOTCH2	Responders	1 (8.3%)	11 (91.6%)	1
	Non-responders	2 (9.5%)	19 (90.5%)	
TAOK1	Responders	2 (16.6%)	10 (83.3%)	0.5831
	Non-responders	1 (5%)	20 (95%)	
BLM, CHD4, FAF1, MGA, NCOR1	Responders	1 (8.3%)	11 (91.6%)	1
	Non-Responders	1 (5%)	20 (95%)	

3.4 Matched sample In-Depth analysis

This section describes in depth each of the samples from the patients that had samples taken from multiple timepoints (e.g. Pre and Post Treatment). The mutational signatures present in each sample and the subclonal architecture of each sample are analysed. The predicted or known driver mutations in each sample are also analysed. These mutations are classified into being “known” or “predicted”, as well as being classified into “oncogenic” or “Tumour suppressing”, based on the classification assigned to them by CGI.

The CGI Catalogue of Validated Somatic Mutations is built from data in the DoCM (Database of Curated Mutations in Cancer (138)), ClinVar (an NCBI archive for clinically significant variants (139)), and OncoKB (a precision oncology knowledgebase (140)). See this URL for proof of this: https://www.cancergenomeinterpreter.org/mutations. If you would like to see the relevant literature for any gene reported in these sections, please check these databases. In cases where further claims are made about any given gene beyond the information given by CGI, these claims are supported by citations from the relevant literature. The CGI software also predicts whether the mutation is likely to cause a loss of function in that gene (denoted “LoF”) or to activate that gene in contexts it would not otherwise be active (denoted “Act”). Finally, for the mutations not conclusively known to be cancer driving, the mutation is assigned a “tier” based on how likely the mutation is to be a driver based on the databases CGI is based on (so a Tier 1 variant has more evidence supporting it being a driver than a Tier 2 variant, etc).

This data is considered holistically to build a picture of what is likely to have happened to the tumour in each patient over the course of therapy.

The equivalent data for the tumours from the patients that did not have samples taken from multiple timepoints is shown in the Supplementary Data section at the end of this thesis.

3.4.1 TCHL 3 samples

The TCHL 3 patient was given lapatinib (TCHL patient) and was a non-responder to initial therapy (no pCR).

3.4.1.1 - TCHL 3 Pre-treatment Mutational Signatures

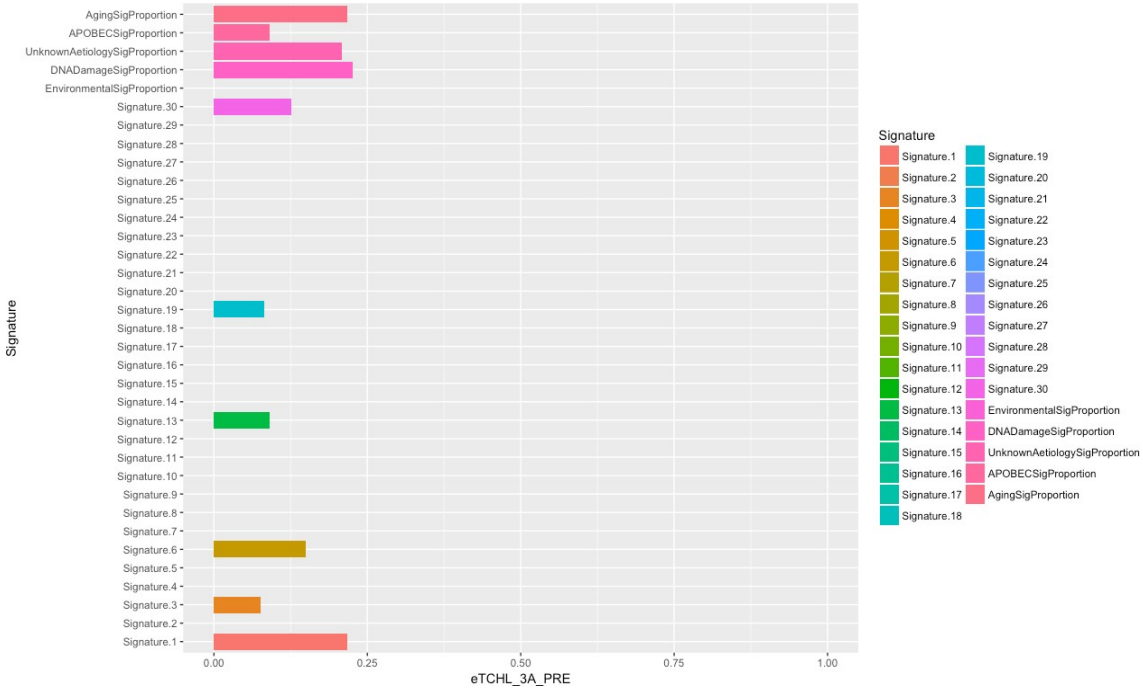


Figure 3.4.1.1 – TCHL Pre-treatment mutational signatures

TCHL 3A Pre-treatment sample shows a mutational spectrum influenced by all signature groupings other than environmental signatures. The largest individual signature is signature 1, the Aging signature, suggesting that the most impactful factor in the somatic mutations seen in this sample was naturally occurring mutations during the aging process. The presence of APOBEC related and DNA damage signatures suggests that the sample was also impacted by the dysregulation or breakdown of normally occurring processes in the cell - dysregulation of the APOBEC enzymes presumably caused the APOBEC signature, while deficiencies in DNA damage repair mechanisms presumably caused the DNA damage signature. Finally, there is a contribution from signatures of unknown aetiology (around 0.2 of the overall mutational spectrum).

SciClone Data

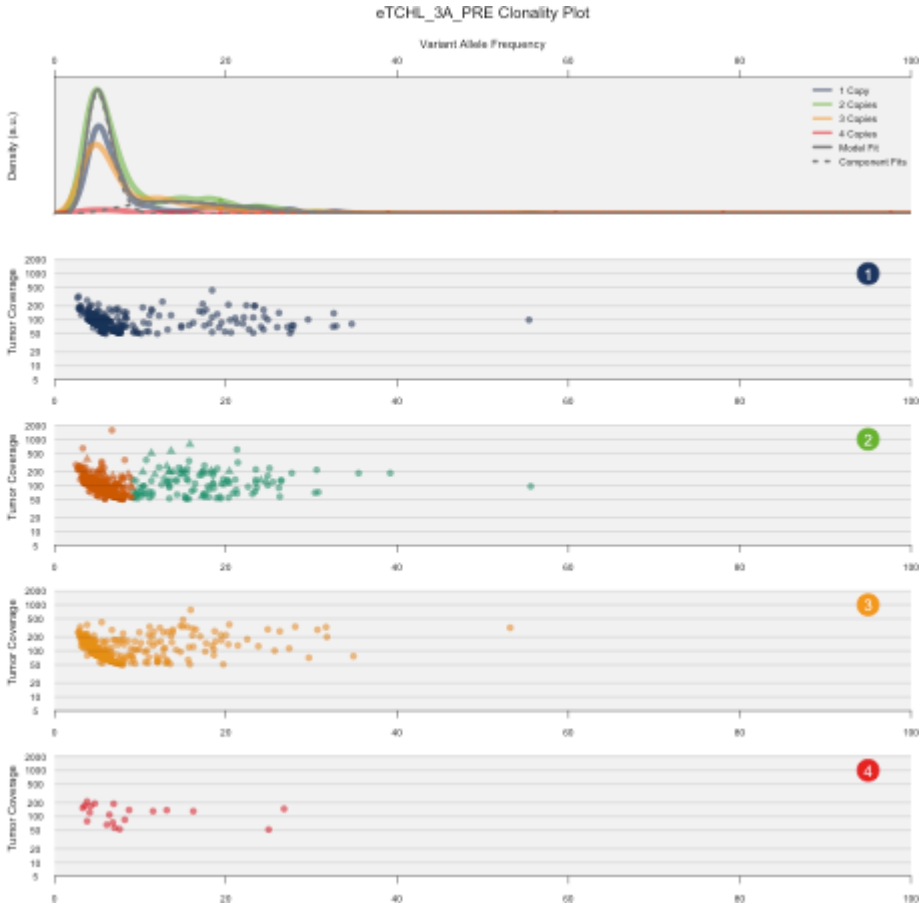


Figure 3.4.1.2 – TCHL 3 Pre-treatment subclonal architecture

The TCHL 3A Pre-treatment sample shows 2 main clusters (see the graph for copy number 2, each different colour represents a different cluster and so a different subclonal population). This suggests there is a less mutated founder clone (hence the low density of the higher VAF cluster) and a more mutated low VAF subclone.

Driver analysis

Table 3.4.1.1 – Known or predicted driver SNVs and indels for TCHL 3 Pre-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178952085|A|G	*PIK3CA*	c.3140A>G	chr3:g.178952085A>G	p.H1047R	Act	known in: COREAD;BRCA;OV;NSCL C
9|37015055|G|A	***PAX5***	**c.349C>T**	**chr9:g.37015055G>A**	**p.R117W**	**Act**	**predicted driver: tier 1**
7|148544312|T|TTAACATTATAC	*EZH2*	c.78_79insGTATAATGTTA	chr7:g.148544312_148544313insTAACATTATAC	p.R27Vfs*2	LoF	predicted driver: tier 1
6|26250542|C|A	*HIST1H3F*	c.292G>T	chr6:g.26250542C>A	p.E98*	ambiguou s	predicted driver: tier 2
5|14492844|G|A	*TRIO*	c.7801G>A	chr5:g.14492844G>A	p.A2601T	Act	predicted driver: tier 1
4|83745799|AT|A	*SEC31A*	c.3319delA	chr4:g.83745807delT	p.I1107Lfs*13	LoF	predicted driver: tier 1
3|180688118|T|TTCTGTATTATC	*FXR1*	c.1575_1576insTCTGTATTATC	chr3:g.180688118_180688119insTCTGTATTATC	p.T526Sfs*1 3	LoF	predicted driver: tier 1
3|130454792|T|TTTTAATAGGAT		c.787_788insTAGAT	chr3:g.130454792_130454793insTTT	p.Q263Lfs*2 1	ambiguou s	predicted

CTA	*PIK3R4*	CCTATTAAA	AATAGGATCTA			driver: tier 2
1|94668513|T|TAAAATATTGTCTTAACTA	*ARHGAP29*	c.914_915insTAGTTAAGACAATATTTT	chr1:g.94668513_94668514insAAAATATTGTCTTAACTA	p.K305Nfs*3	LoF	predicted driver: tier 1
1|51323662|G|GAAGTTTA	*FAF1*	c.52_53insTAAACTT	chr1:g.51323662_51323663insAAGTTTA	p.T18Ifs*6	LoF	predicted driver: tier 1
1|51323661|A|ATATCTTTAATAT	*FAF1*	c.53_54insATATTAAAGATA	chr1:g.51323661_51323662insTATCTTTAATAT	p.G19Yfs*2	LoF	predicted driver: tier 1
1|44086781|C|T	*PTPRF*	c.5533C>T	chr1:g.44086781C>T	p.R1845C	Act	predicted driver: tier 1
1|182850743|C|G	*DHX9*	c.2875C>G	chr1:g.182850743C>G	p.Q959E	Act	predicted driver: tier 2

input	gene	cdna	gdna	protein	gene_role	Driver statement
7|42966228|T|TGTAA	***PSM A2***	**c.157_158insTTAC**	**chr7:g.42966228_42966229insGTAA**	**p.K53lfs*7**	**ambiguous**	**predicted driver: tier 2**
19|12812429|G|A	*TNP O2*	c.2569C>T	chr19:g.12812429G>A	p.R857W	ambiguous	predicted driver: tier 1
17|7577556|CAGGAACTGTTACACATGT|C	*TP53*	c.707_724delACATGTGTAACAGTTCCT	chr17:g.7577559_7577576delGAACTGTTACACATGTAG	p.Y236_S241delYMCNS S	LoF	predicted driver: tier 2
17|37949145|C|G	*IKZF3*	c.205G>C	chr17:g.37949145C>G	p.E69Q	ambiguous	predicted driver: tier 2
17|37922110|C|T	*IKZF3*	c.1463G>A	chr17:g.37922110C>T	p.R488Q	ambiguous	predicted driver: tier 2
17|30521113|T|TAAAGTAA	*RHO T1*	c.857_858insAAGTAAA	chr17:g.30521114_30521115insAAGTAAA	p.Y286*fs*1	ambiguous	predicted driver: tier 2

The TCHL 3 Pre-treatment sample shows a large number of predicted driver events - 20, to be exact, which as noted in the introduction is on the higher end of the number of driver events seen in any cancer. However the only event conclusively known to be a driver is *PIK3CA* H1047R, a driver mutation known to be common in breast cancer and to be associated with trastuzumab and lapatinib resistance (141). We know this mutation is conclusively known to be oncogenic because it was marked as such by CGI, and because it appears in the literature as a validated driver mutation (142).

Mutations in the following genes are likely oncogenic (i.e. promote cellular proliferation): *PIK3CA, PAX5, TRIO,PTPRF, DHX9*.

Mutations in the following genes are likely tumour suppressor mutations (i.e. the mutations disable normal cellular breaks on unchecked cellular proliferation): *EZH2, SEC31A, FXFR1, ARHGAP29, FAF1, TP53*.

It is unclear whether mutations in the following genes are oncogenic or tumour suppressor mutations: *PSMA2, HIST1H3F, PIK3R4, TNPO2, IKZF3, RHOT1*.

3.4.1.2 TCHL 3 Post treatment

Mutational Signatures

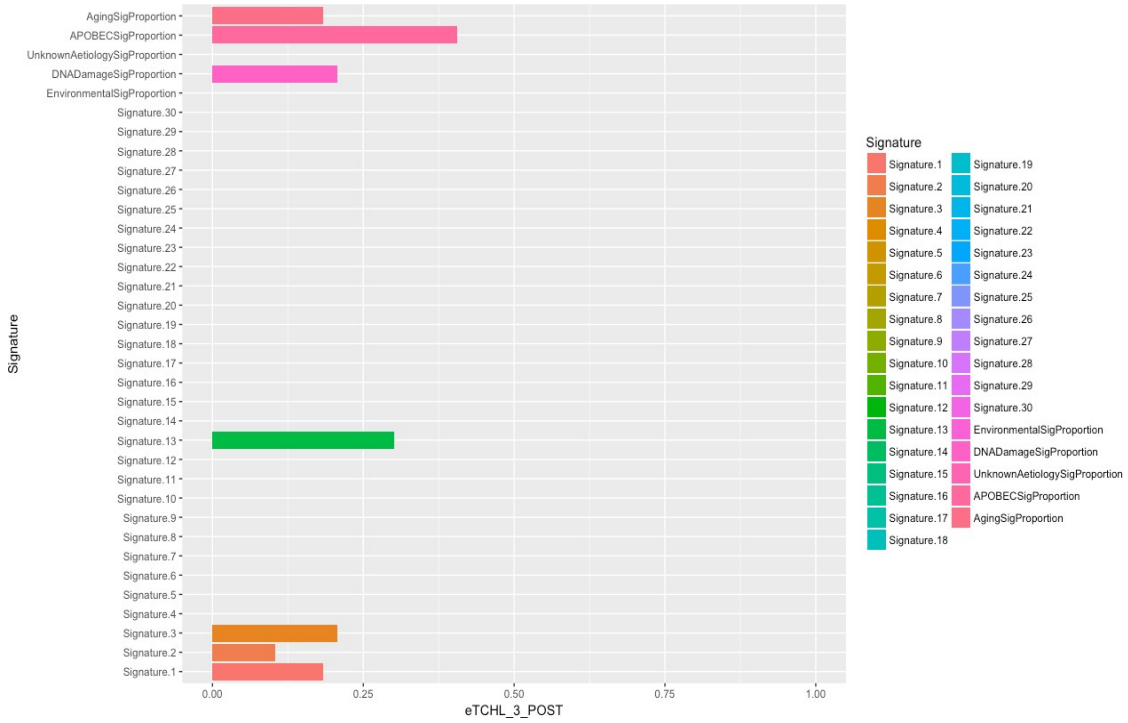


Figure 3.4.1.3 – Mutational signatures for the TCHL 3 Post treatment sample

The TCHL 3 Post treatment sample shows a mutational spectrum with a similar level of Aging signature influence and DNA damage signature influence to the corresponding Pre-treatment sample, but a much higher level of APOBEC signature proportion. Since referring to the table in section 3.5.1 shows us that the post treatment sample has fewer indels and SNVs than the Pre-treatment sample, this may be in part due to very heavily mutated cells being killed by treatment, leaving the APOBEC signature as a higher proportion of the overall mutational landscape of the sample.

SciClone Data

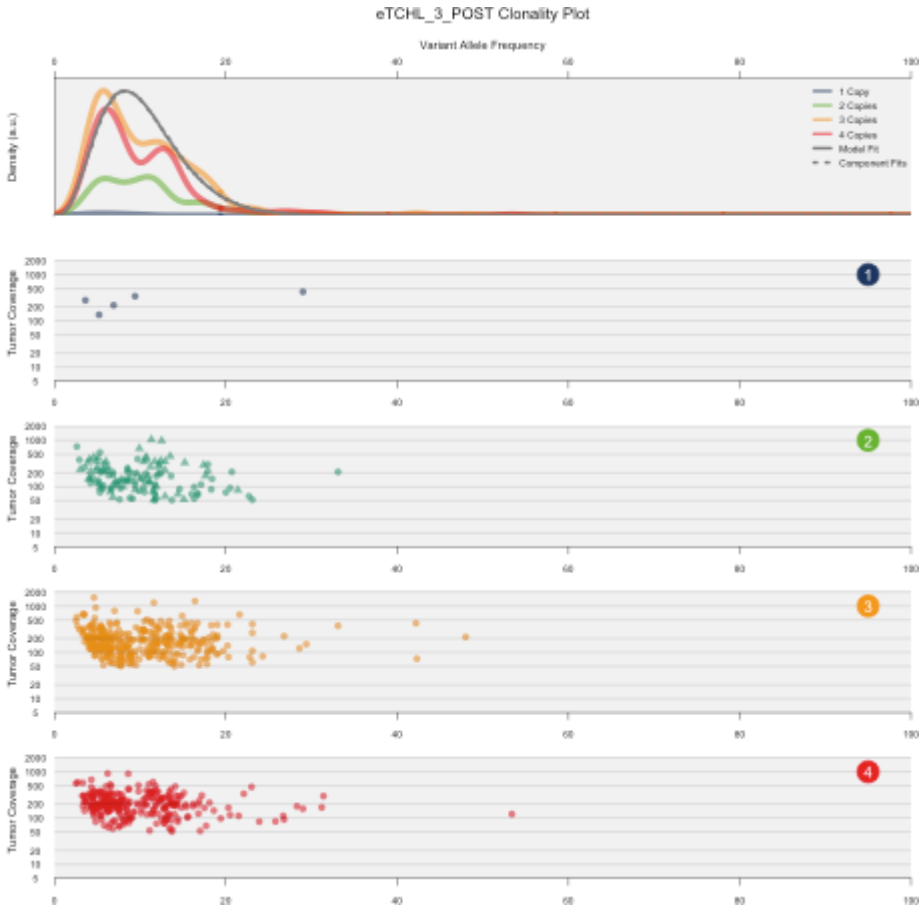


Figure 3.4.1.4 – Subclonal architecture of the TCHL 3 Post treatment sample

The clonality plot for TCHL 3 shows a single cluster per copy number, suggesting that there are no significant subclones in this sample.

Driver analysis

Table 3.4.1.2 – Known or predicted driver SNVs and indels for TCHL 3 Post-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178952085|A|G	*PIK3CA*	c.3140A>G	chr3:g.178952085A>G	p.H1047R	Act	known in: COREAD;NSCLC;BRCA;OV
6|26250542|C|A	*HIST1H3F*	c.292G>T	chr6:g.26250542C>A	p.E98*	ambiguous	predicted driver: tier 2
5|131944381|C|C A	***RAD50***	**c.2801dupA**	**chr5:g.131944389dupA**	**p.N934Kfs*10**	**ambiguous**	**predicted driver tier 2**
2|44123829|C|T	***LRPPRC***	**c.3844G>A**	**chr2:g.44123829C>T**	**p.E1282K**	**LoF**	**predicted driver tier 1**
1|182850743|C|G	*DHX9*	c.2875C>G	chr1:g.182850743C>G	p.Q959E	Act	predicted driver: tier 2
17|7577556|CAGGAACTGTTACACATGT|C	*TP53*	c.707_724delACATGTGTAACAGTTCCT	chr17:g.7577559_7577576delGAACTGTTACACATGTAG	p.Y236_S241delYMCNSS	LoF	predicted driver: tier 2
17|37949145|C|G	*IKZF3*	c.205G>C	chr17:g.37949145C>G	p.E69Q	ambiguous	predicted driver: tier 2
17|37922110|C|T	*IKZF3*	c.1463G>A	chr17:g.37922110C>T	p.R488Q	ambiguous	predicted driver: tier 2

The TCHL 3 Post treatment sample shows fewer predicted drivers than the Pre-treatment sample. Given that the Post treatment sample has fewer SNVs and

indels than the Pre-treatment sample, as mentioned in the Mutational Signatures section, this implies that more heavily mutated cells in this patient succumbed to treatment, and the surviving cells after treatment started had a lower mutational burden and fewer driver mutations. The validated driver *PIK3CA* H1047R, associated with trastuzumab and lapatinib resistance (141), is also present in the post treatment sample. The predicted driver genes in this sample are all present in the Pre- treatment sample, with the exception of the *RAD50* predicted driver mutation and the *LRPPRC* predicted driver mutation.

Mutations in the following genes are likely oncogenic: *PIK3CA, DHX9*.
Mutations in the following genes are likely tumour suppressor inactivating: *LRPPRC, TP53*.

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: *HISTH1H3F, RAD50, IKZF3*.

3.4.1.3 TCHL 3 Surgery treatment Mutational Signatures

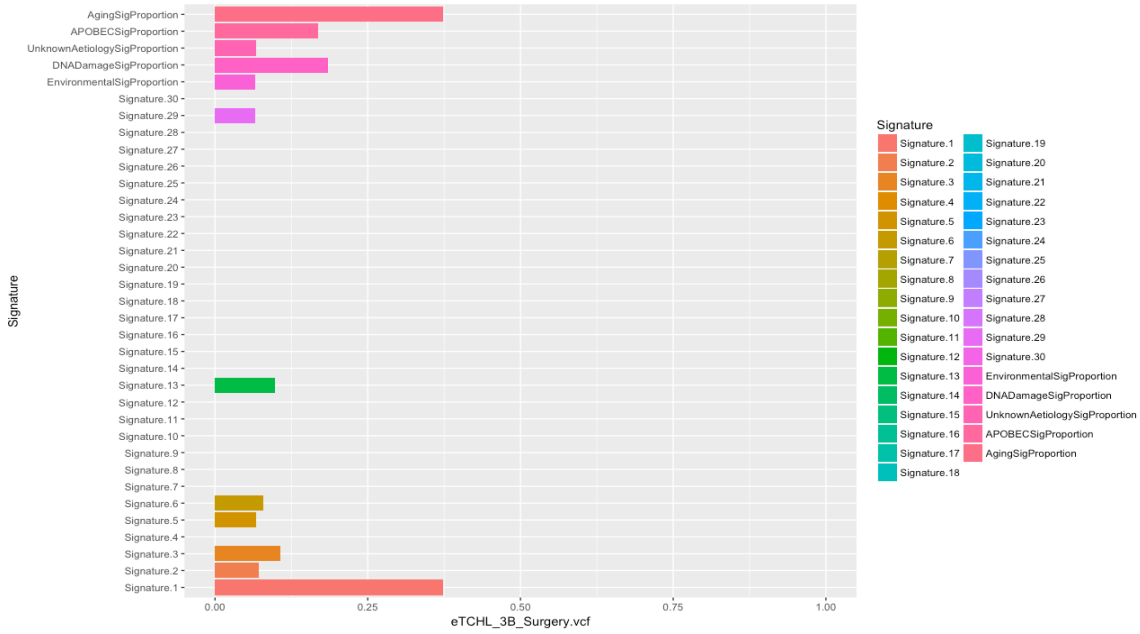


Figure 3.4.1.5 – Mutational signatures for the TCHL 3 Surgery sample

The TCHL 3B surgery sample shows a mutational spectrum dominated by the Aging signature. Since the Pre-treatment sample showed a much smaller proportion of the Aging signature, this suggests that as the treatment continued cells with a heavier mutational load were killed, leaving the Aging signature as a higher overall proportion of the mutational spectrum in the cells that survived in the tumour this sample was taken from. Referring to the table in section 3.1.1 supports this idea, as the 3B surgery sample has fewer SNVs and indels than the other samples from the TCHL 3 patient. The sample is still influenced by APOBEC, DNA damage, and unknown aetiology signatures. There is also a small environmentally associated signature, in Signature 29. Since Signature 29 is associated with tobacco chewing in mouth cancer, it is likely that this assigning this signature to this sample is an error due to the details of the deconstructSigs algorithm (see Intro section 1.6) as opposed to that signature genuinely having influenced this sample. The reasoning for this is that the mouth and lung are quite different environments biochemically – although the

exact chemical mechanism causing signature 29 is not known, it seems unlikely that the chemical reaction that causes signature 29 in a mouth environment would also be occurring in a lung environment. However, the possibility that the signature is genuinely present in this sample rather than just being an artifact of DeconstructSigs cannot be ruled out entirely.

SciClone Data

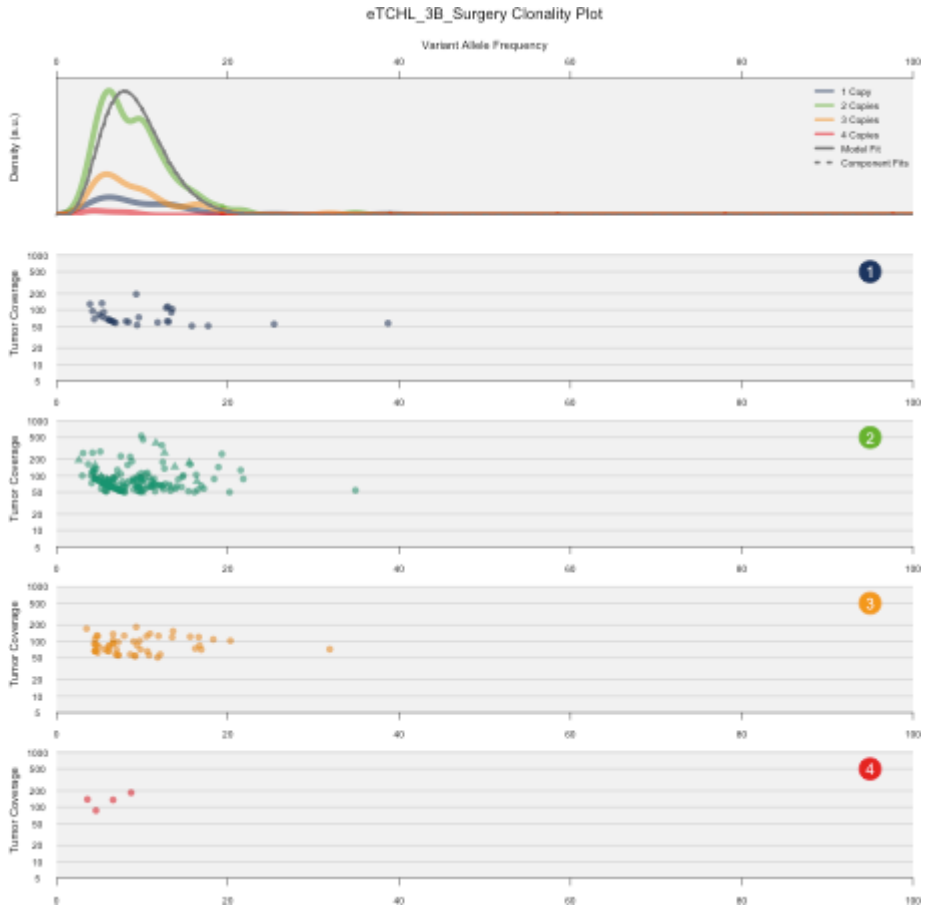


Figure 3.4.1.6 – Subclonal architecture for the TCHL 3 Surgery sample

The clonality plot for the TCHL 3 surgery sample shows a single cluster per copy number, suggesting that there are no significant subclones in this sample. The plot is centred at a lower VAF than the Post treatment sample Sciclone plot.

Driver analysis

Table 3.4.1.3 – Known or predicted driver SNVs and indels for TCHL 3 Surgery treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178952085|A|G	*PIK3CA*	c.3140A>G	chr3:g.178952085A>G	p.H1047R	Act	known in: NSCLC;BRCA;COREAD;OV
19|12814290|C|T	***TNPO2***	**c.2161G>A**	**chr19:g.12814290C>T**	**p.V721I**	**ambiguous**	**predicted driver: tier 1**
17|7577556|CAGGAACTGTTACACATGT|C	*TP53*	c.707_724delACATGTGTAACAGTTCCT	chr17:g.7577559_7577576delGAACTGTTACACATGTAG	p.Y236_S241delYMCNSS	LoF	predicted driver: tier 2
17|37949145|C|G	*IKZF3*	c.205G>C	chr17:g.37949145C>G	p.E69Q	ambiguous	predicted driver: tier 2
17|37922110|C|T	*IKZF3*	c.1463G>A	chr17:g.37922110C>T	p.R488Q	ambiguous	predicted driver: tier 2

As mentioned in the Mutational Signatures section, the surgery sample for TCHL 3 shows fewer SNVs and indels than the Pre- and Post- treatment samples. This carries through to the driver genes predicted by CGI: there are fewer driver genes here than there are in the Pre or Post treatment samples. All the predicted driver genes in the surgery sample are shared with the Pre- and Post- treatment samples, with the exception of the *TNPO2* V721I predicted driver mutation.

Mutations in the following genes are likely oncogenic: *PIK3CA*

Mutations in the following genes are likely tumour suppressor inactivating: *TP53*

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: *IKZF3, TNPO2*

3.4.2 TCHL 6 samples

The TCHL 6 patient was not given lapatinib (TCH patient). They were a non-responder to initial therapy and ultimately showed relapse in the bone.

3.4.2.1 TCHL 6 Pre-treatment

Mutational Signatures

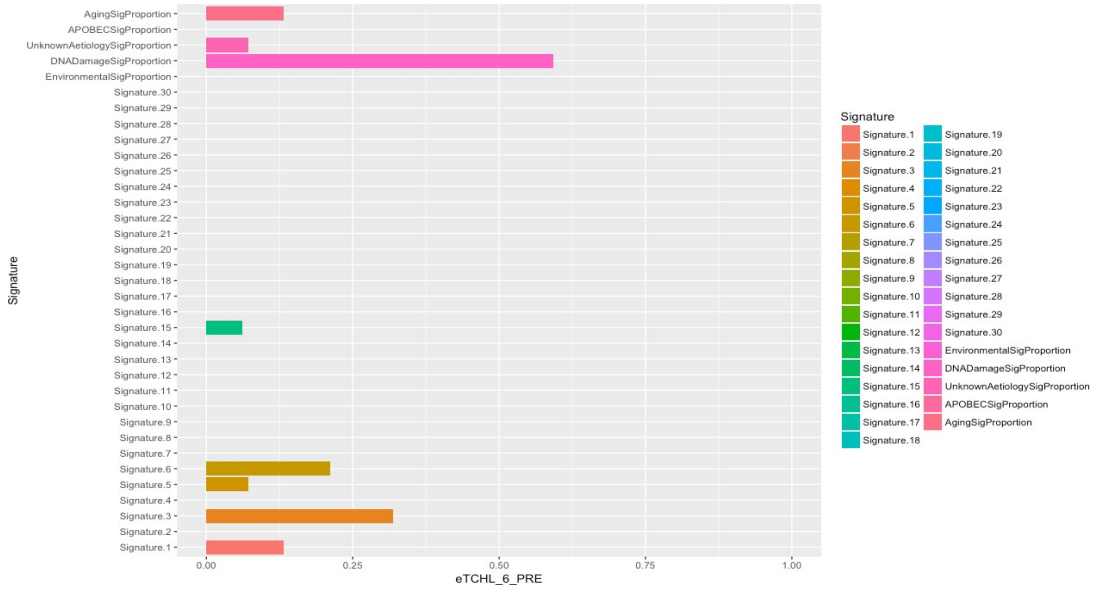


Figure 3.4.2.1 – Mutational signatures for the TCHL 6 Pre-treatment sample

The TCHL 6 Pre-treatment sample shows a mutational spectrum dominated by DNA damage based signatures, specifically signatures 3 and 6. Signature 3 is associated with failure of double-strand break (DSB) repair by homologous recombination, while signature 6 is associated with failure of DNA mismatch repair. The heavy influence of these signatures on the mutational spectrum creates a picture of a tumour in which the driver mutations have been caused primarily by a breakdown in the cellular processes usually responsible for repairing DNA damage. There is also a small contribution from the Aging signature and from a signature of unknown aetiology.

SciClone Data

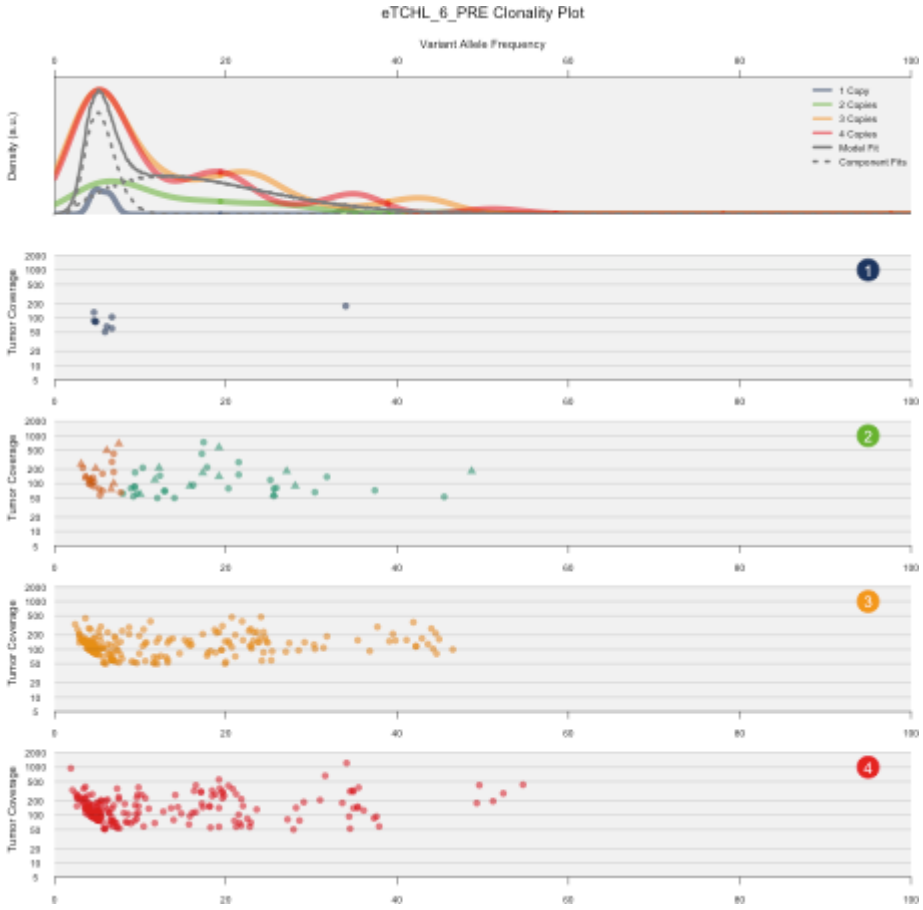


Figure 3.4.2.2 – Subclonal architecture for the TCHL 6 Pre-treatment sample

The TCHL 6 Pre-treatment sample shows a complicated clonality plot, with 2 main clusters, one at a lower VAF and much higher density than the other. This suggests the presence of a less mutated founder clone and a more mutated low VAF subclone.

Driver analysis

Table 3.4.2.1 – Known or predicted driver SNVs and indels for TCHL 6 Pre-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
3|33543233|G|A	*CLASP2*	c.4369C>T	chr3:g.33543233G>A	p.R1457W	LoF	predicted driver: tier 1
2|48033673|C|A	***MSH6***	**c.3884C>A**	**chr2:g.48033673C>A**	**p.P1295H**	**LoF**	**predicted driver: tier 2**
1|247062816|T|TA	***AHCTF1***	**c.1460-3dupT**	**chr1:g.247062824dupA**	**.**	**ambiguous**	**predicted driver: tier 2**
19|3119297|G|A	*GNA11*	c.829G>A	chr19:g.3119297G>A	p.D277N	Act	predicted driver: tier 1

The TCHL 6 Pre-treatment sample shows a small number of predicted driver mutations, with no mutations that are definitively known to act as drivers.

Mutations in these genes are found in few other samples in the cohort, suggesting that the cancer in patient 6 may have been driven by distinct biological pathways from the cancers in other individuals in the cohort. Since, as mentioned in the Intro, cancer cells usually bear around 6-7 driver mutations, the small number of drivers here may indicate that some driver mutations were filtered by the stringency tests in the variant calling pipeline (see Intro section 1.5), or were not picked up by CGI. However, it may instead indicate that the biological pathways that caused the cancer in TCHL patient 6 simply require fewer driver mutations to become cancerous.

Mutations in the following genes are likely oncogenic: *GNA11*

Mutations in the following genes are likely tumour suppressor inactivating: *MSH6*, *CLASP2*

It is ambiguous whether mutations in the following genes are oncogenic or

tumour suppressor mutations: *AHCTF1*

3.4.2.2 TCHL 6 Post treatment

Mutational Signatures

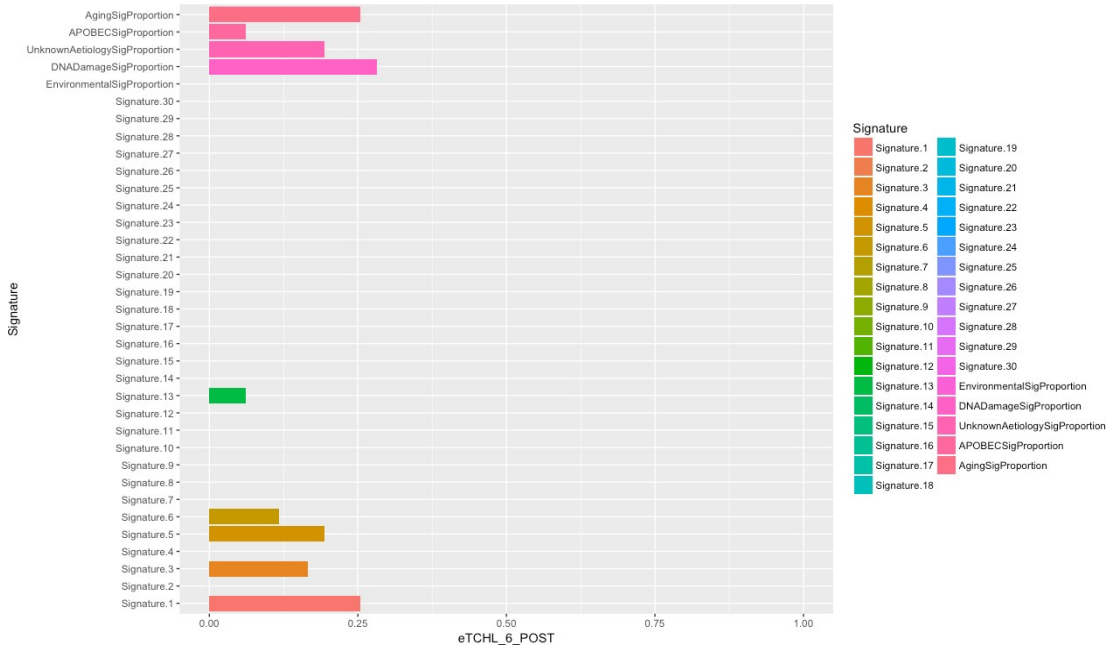


Figure 3.4.2.3 – Mutational signatures for the TCHL 6 Post-treatment sample

The TCHL 6 Post treatment sample shows a mutational landscape where DNA damage based signatures are still the largest influence, but are not nearly as high a proportion as they were in the Pre-treatment sample. Since the table in section 3.4.1 shows that the post treatment sample has fewer SNVs and indels than the Pre-treatment sample, this suggests that treatment may have killed the cells most heavily mutated by the defects in DNA damage repair mechanisms, leaving behind cells where DNA damage signatures were a smaller proportion of the overall spectrum. There is also a small proportion APOBEC based signature, which may mean that APOBEC enzymes were dysregulated in the tumour during therapy.

SciClone Data

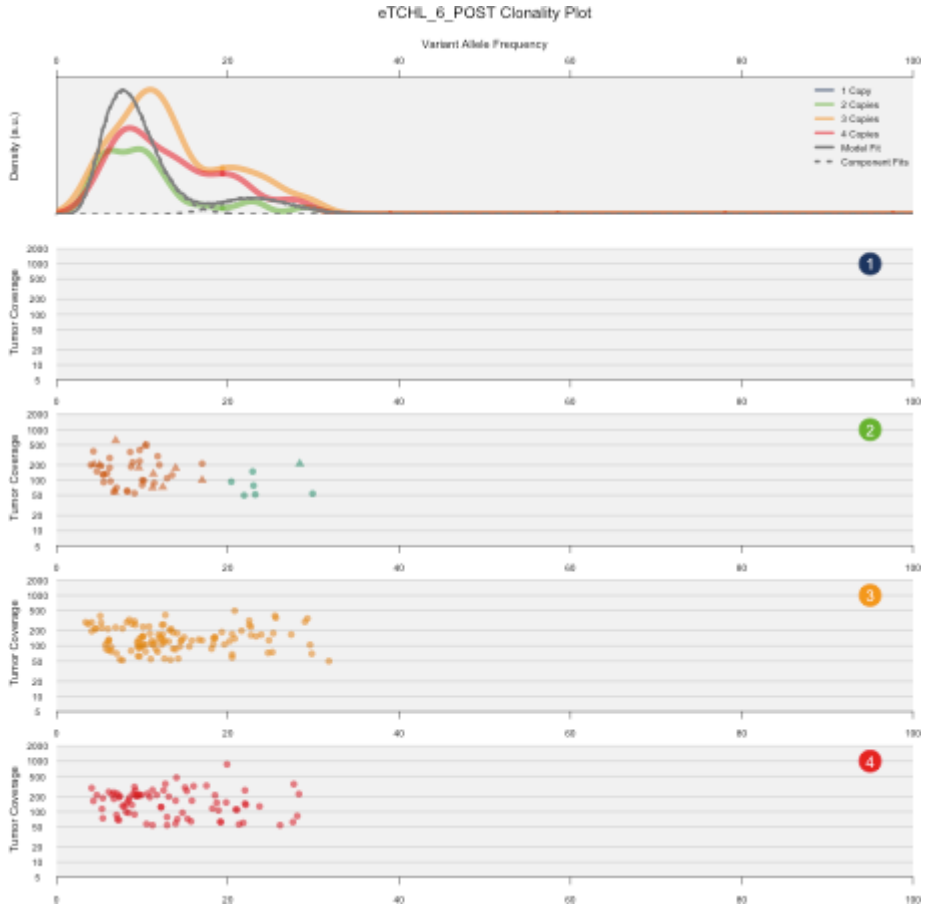


Figure 3.4.2.4 –Subclonal architecture for the TCHL 6 Post-treatment sample

The TCHL 6 Post Treatment sample shows 2 clusters, like the Pre-treatment sample. However the peaks of the clonality plot are at lower VAFs than the Pre-treatment sample, suggesting that the mutations present are being reduced in overall frequency by more mutated cells being killed during the process of therapy.

Driver analysis

Table 3.4.2.2 – Known or predicted driver SNVs and indels for TCHL 6 Post treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
3|33543233|G|A	*CLASP2*	c.4369C>T	chr3:g.33543233G> A	p.R1457W	LoF	predicted driver: tier 1
19|3119297|G|A	*GNA11*	c.829G>A	chr19:g.3119297G> A	p.D277N	Act	predicted driver: tier 1

The TCHL 6 Post treatment sample shows only 2 predicted driver mutations, both shared with the Pre-treatment sample. Since, as mentioned above, the Post treatment sample has fewer SNVs and indels overall than the post treatment sample, this likely indicates that more heavily mutated cells bearing a greater number of driver mutations were killed during the process of therapy.

Mutations in the following genes are likely oncogenic: *GNA11*

Mutations in the following genes are likely tumour suppressor inactivating: *CLASP2*

3.4.2.3 TCHL 6CFF - Relapse sample Mutational Signatures

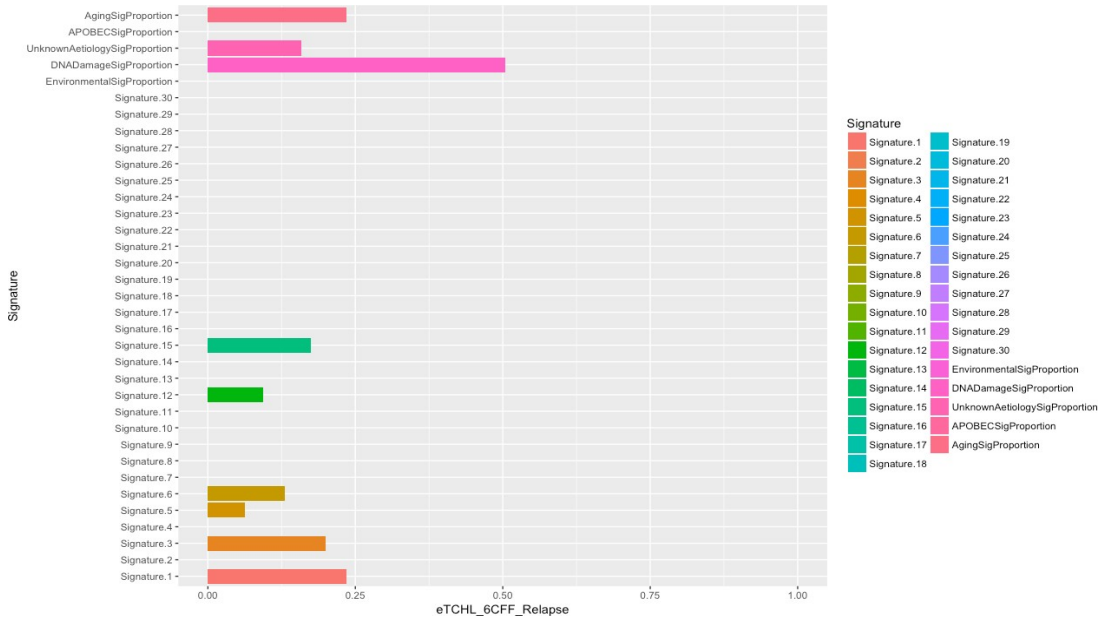


Figure 3.4.2.5 – Mutational signatures in the TCHL 6 relapse sample

The TCHL 6 CFF Relapse sample shows a signature landscape dominated by DNA damage based signatures. Interestingly, this is much closer to the Pre-treatment sample than the post treatment sample. This suggests that some cells heavily mutated by the DNA damage repair defaults may have survived subclonally during treatment, and that the mutations in these cells caused the relapse. Looking at the table in section 3.1, we can see that the number of SNVs and indels in the relapse sample is similar to the number in the Pre-treatment sample and much greater than the number in the post treatment sample, lending support to the idea that the relapse was caused by cells that existed subclonally in the Pre-treatment tumour. The fact that the Sciclone plots for the Pre and Post treatment samples both show 2 main clusters supports the idea that therapy failed to totally eradicate the subclonal population in the Pre-treatment sample, and it is likely that this subclonal population acted as a reservoir of genetic diversity that allowed ultimately allowed a relapse to occur. The Aging signature and a signature of unknown aetiology compose a similar proportion of the mutational spectrum in this sample as they do in the Pre-

treatment sample.

SciClone Data

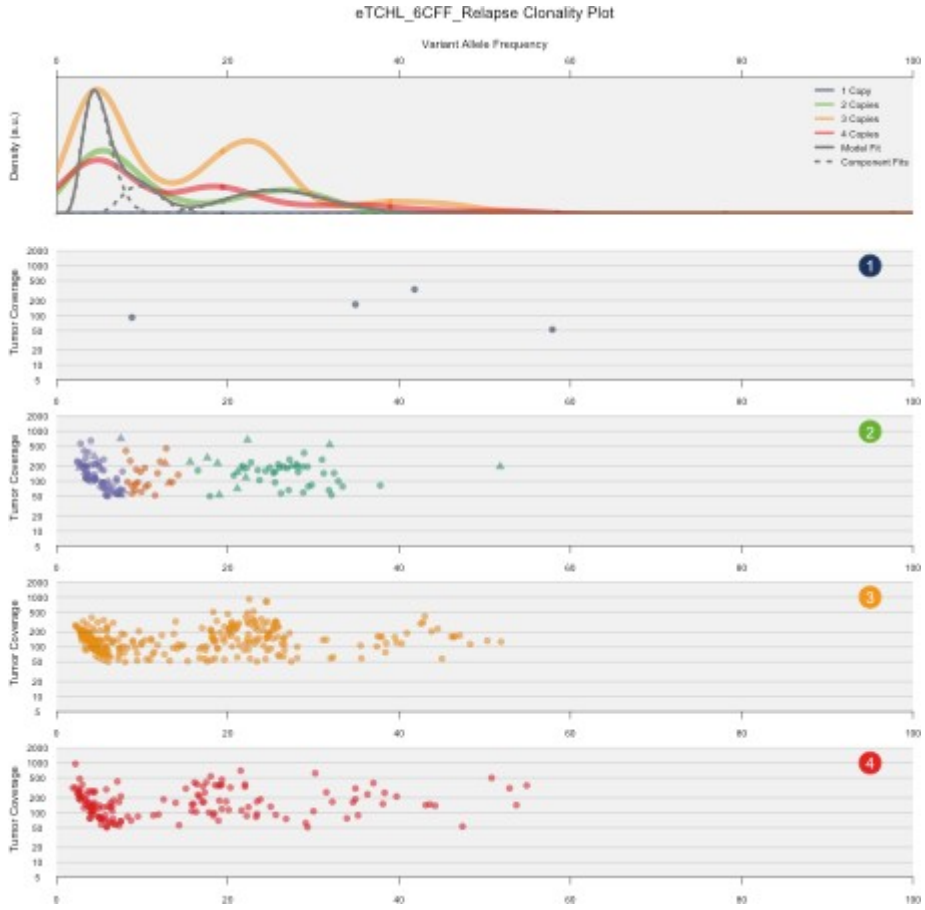


Figure 3.4.2.6 –Subclonal architecture for the TCHL 6 Relapse sample

The TCHL 6CFF relapse sample shows 3 clusters, suggesting a dominant clone centred at a VAF of 20, then a less mutated subclone centred at a VAF of 10 and a more heavily mutated subclone centred at a VAF of 5.

Driver analysis

Table 3.4.2.3 – Known or predicted driver SNVs and indels for TCHL 6 Relapse. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	driver_statement
5|131931451|TA|T	***RAD50***	**c.2165delA**	**chr5:g.131931460delA**	**p.K722Rfs*14**	**ambiguous**	**predicted driver: tier 2**
19|3119297|G|A	*GNA11*	c.829G>A	chr19:g.3119297G>A	p.D277N	Act	predicted driver: tie 1

The TCHL 6 CFF relapse sample shows only 2 predicted driver mutations. As mentioned in the Driver analysis section for the corresponding Pre-treatment sample, it is unlikely that an active cancer cell would show so few driver mutations, so we consider it possible that some driver mutations in this sample were either filtered at some point in the variant calling workflow or missed by the variant calling software entirely. We bring in the possibility of false negatives such as this because even the most advanced variant calling software workflows are known in the literature to have a non-zero false positive rate (143). Another possibility is that there are driver mutations in this sample that are not known in the CGI database and were not predicted by the CGI prediction algorithm. The presence of a *RAD50* mutation, not present in the other 2 samples from the same patient, may be notable, as there is also a *RAD50* mutation in the TCHL 39 relapse sample, as well as the TCHL 3 Post treatment sample.

Mutations in the following genes are likely oncogenic: *GNA11*

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: *RAD50*

3.4.3 - TCHL 12 samples

The TCHL 12 patient was given lapatinib (TCHL patient) and was a non-responder to initial therapy (no pCR).

3.4.3.1 - TCHL 12 Pre-treatment Mutational Signatures

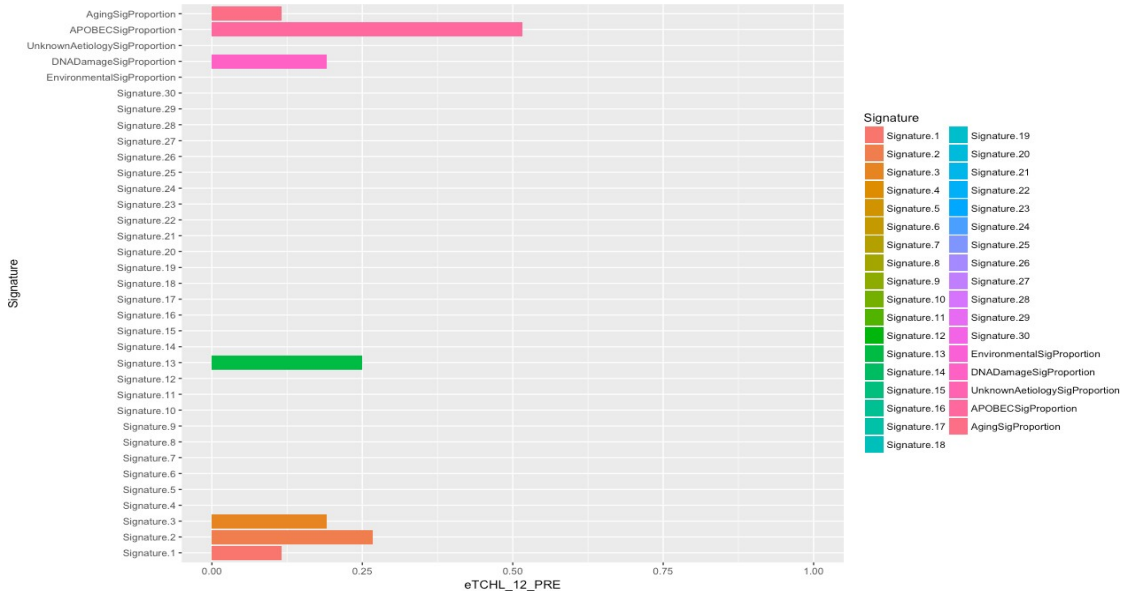


Figure 3.4.3.1 – Mutational signatures in the TCHL 12 Pre-treatment sample

The TCHL 12 Pre-treatment sample shows a mutational landscape heavily influenced by the APOBEC signature. This implies that this sample, prior to or during carcinogenesis, experienced a dysregulation of the APOBEC enzymes active in the cancer founder cell, and this faulty APOBEC activity may have contributed to the presence of the driver mutations that caused the cancer.

There is also some influence of the Aging signature, and of a DNA damage signature - specifically Signature 3, a signature associated with failure of DSB repair by homologous recombination.

SciClone Data

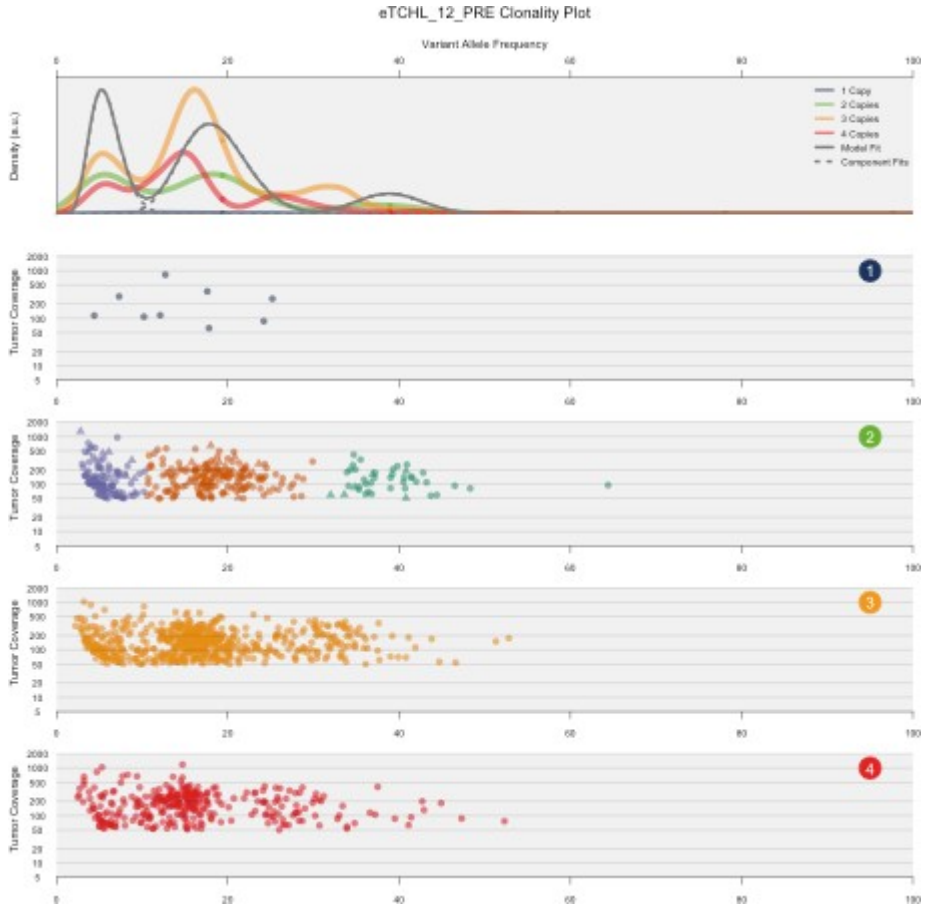


Figure 3.4.3.2 – Subclonal architecture in the TCHL 12 Pre-treatment sample

The TCHL 12 Pre-treatment sample shows 3 main clusters, suggesting a founder clone centred at a VAF of 40, then two subclonal populations with much heavier mutational burdens centred at VAFs of 5 and 15, respectively.

Driver analysis

Table 3.4.3.1 – Known or predicted driver SNVs and indels for TCHL 12 Pre-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
17|7577120|C|T	***TP53***	**c.818G>A**	**chr17:g.7577120C>T**	**p.R273H**	**LoF**	**known in: THCA;CANCER-PR;AML**
8|103311746|C|G	***UBR5***	**c.3136G>C**	**chr8:g.103311746C>G**	**p.A1046P**	**ambiguous**	**predicted driver: tier 2**
7|81372698|C|T	***HGF***	**c.836G>A**	**chr7:g.81372698C>T**	**p.R279H**	**Act**	**predicted driver: tier 2**
3|141678557|G|A	***TFDP2***	**c.1010C>T**	**chr3:g.141678557G>A**	**p.A337V**	**LoF**	**predicted driver: tier 2**
22|43213779|A|A T	***ARFGAP3***	**c.896dupA**	**chr22:g.43213787dupT**	**p.N299Kfs* 3**	**LoF**	**predicted driver: tier 1**
21|40604418|G|A	***BRWD1***	**c.2773C>T**	**chr21:g.40604418G>A**	**p.R925W**	**Act**	**predicted driver: tier 2**
1|182850527|T|C	***DHX9***	**c.2753T>C**	**chr1:g.182850527T>C**	**p.L918P**	**Act**	**predicted driver: tier 2**
17|38792691|C|T	***SMARCE 1***	**c.325G>A**	**chr17:g.38792691C>T**	**p.D109N**	**ambiguous**	**predicted driver: tier 2**
15|42003344|G|T	***MGA***	**c.2881G>T**	**chr15:g.42003344G>T**	**p.E961***	**LoF**	**predicted driver: tier 1**

The TCHL 12 Pre-treatment sample shows 9 predicted driver events, of which 1 is a validated driver: the R273H mutation in TP53, a cancer driver gene common in many cancer types and seen in many other samples in this cohort. Many of the other predicted driver mutations in this sample are in genes unique to this sample, though mutations in *UBR5, MGA* and *DHX9* do appear in some other samples in the cohort. None of the predicted driver mutations are shared

with the corresponding Pre-treatment sample.

Mutations in the following genes are likely oncogenic: *HGF, BRWD1, DHX9*
Mutations in the following genes are likely tumour suppressor inactivating: *TP53, TFDP2, AFRGAP3, MGA*

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: *UBR5, SMARCE1*

3.4.3.2 - TCHL 12 Post treatment Mutational Signatures

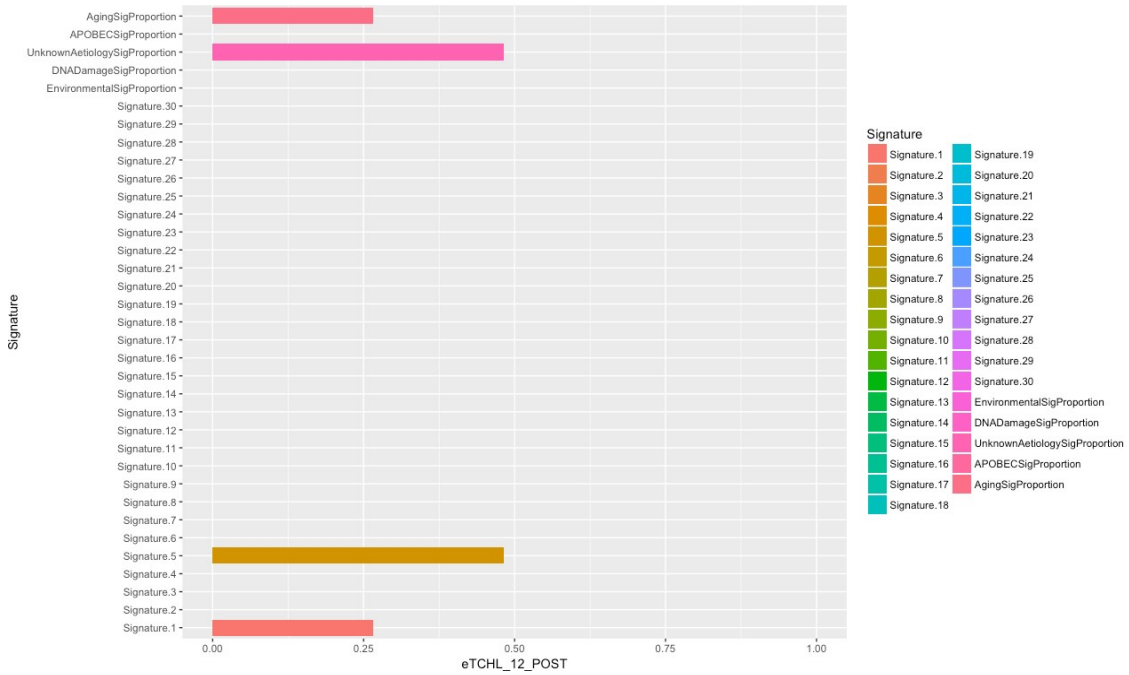


Figure 3.4.3.3 – Mutational signatures in the TCHL 12 Post treatment sample

The TCHL 12 Post treatment sample has a mutational landscape composed of only two signatures - Signature 5, a Signature of unknown aetiology found in all cancers, and signature 1, the Aging signature, present in all cancers and indeed all somatic cells. It is not clear why the APOBEC signature so influential in the Pre-treatment sample is totally absent here, as the post treatment sample has a similar number of SNVs and indels to the Pre-treatment sample (see section 3.5.1)

SciClone Data

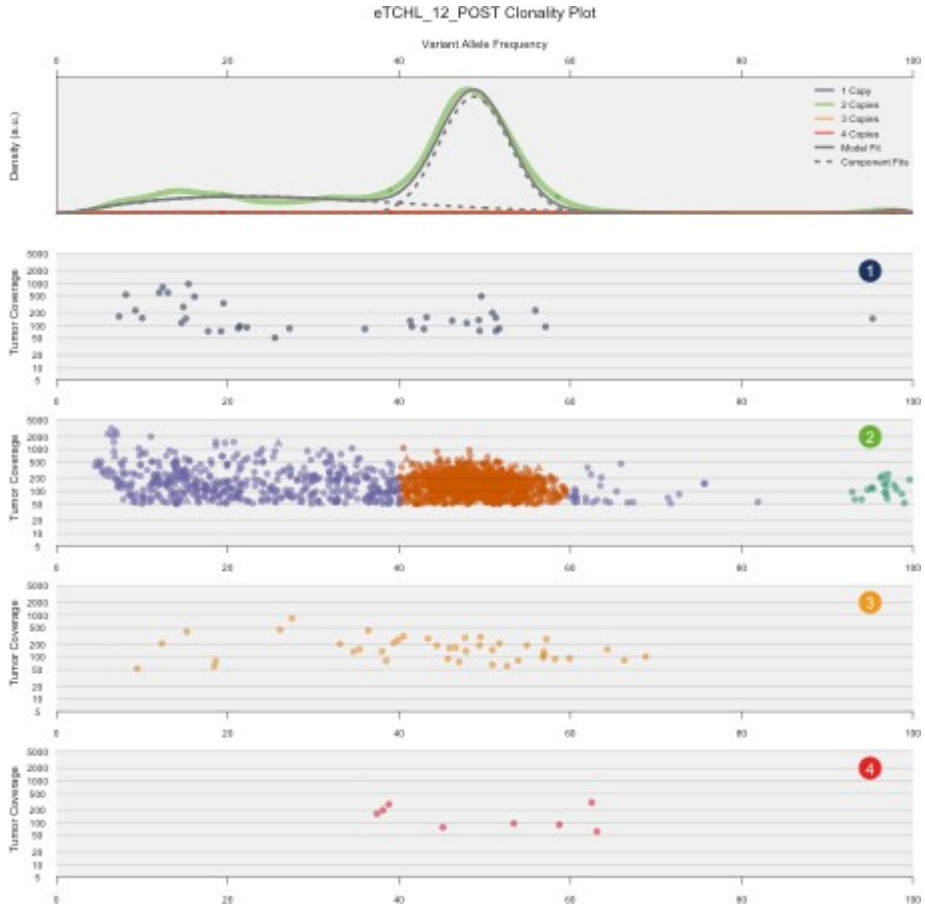


Figure 3.4.3.4 – Subclonal architecture in the TCHL 12 Post treatment sample

The clonality analysis for TCHL 12 Post Treatment sample shows 3 clusters, with the most dense by far centred at a VAF of 50. This suggests that one of the low VAF subclones in the Pre-treatment sample expanded greatly during the therapy process, while the remaining subclonal populations survived at lower frequencies.

Driver analysis

Table 3.4.3.2 – Known or predicted driver SNVs and indels for TCHL 12 Post treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178952085|A|G	***PIK3CA***	**c.3140A>G**	**chr3:g.178952085A>G**	**p.H1047R**	**Act**	**known in: NSCLC;BRCA;COREAD;OV**
10|123258034|A| T	***FGFR2***	**c.1650T>A**	**chr10:g.123258034A>T**	**p.N550K**	**Act**	**known in: EDA;ED**
6|117704649|G|A	***ROS1***	**c.2327C>T**	**chr6:g.117704649G>A**	**p.T776M**	**Act**	**predicted driver: tier 2**
2|128044322|C|G	***ERCC3***	**c.1299G>C**	**chr2:g.128044322C>G**	**p.Q433H**	**LoF**	**predicted driver: tier 2**
19|58386285|G|A	***ZNF814***	**c.473C>T**	**chr19:g.58386285G>A**	**p.A158V**	**Act**	**predicted driver: tier 1**

None of the predicted driver mutations in the TCHL 12 Post treatment sample are shared with the corresponding Pre-treatment sample. This, along with the very different mutational signature landscape in the post treatment sample (no APOBEC signature) and the clonal architecture of the the two samples, suggests the following: that the original tumour harboured subclones at a low number with a distinct biology from the rest of the tumour, and that these subclones survived therapy and expanded in the space left by the cells killed by the therapy. The presence of a PIK3CA mutation, associated with trastuzumab resistance (see Intro section 1.3 and (83)) lends credence to this hypothesis. The other validated mutation is the *FGFR2* oncogene activating Q433H mutation. Apart from the P*IK3CA* mutation, the predicted driver mutations in this sample all affect genes not predicted to act as drivers in any other sample in this cohort.

Mutations in the following genes are likely oncogenic: *PIK3CA, FGFR2, ROS1, ZNF814*

Mutations in the following genes are likely tumour suppressor inactivating:
ERCC3

3.4.4 TCHL 29 samples

The TCHL 29 sample was given lapatinib (TCHL patient) and was a non-responder to initial therapy (no pCR).

3.4.4.1 TCHL 29 Pre-treatment

Mutational Signatures

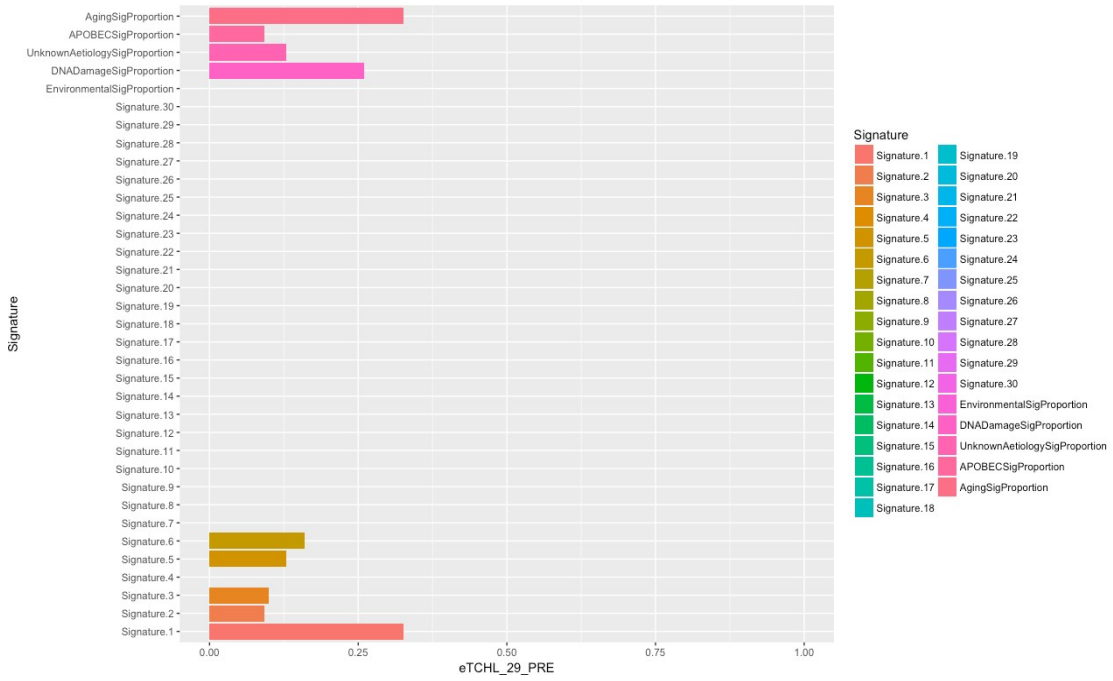


Figure 3.4.4.1 – Mutational signatures in the TCHL 29 Pre-treatment sample

The TCHL 29 Pre-treatment sample shows a mutational landscape influenced by a diverse array of signature types. The most influential is the aging signature, suggesting that a relatively high proportion of mutations in this sample were generated by the somatic mutational processes that occur in all cells as humans age and that the drivers mutations in this sample were mostly generated by this natural and unavoidable process. The second most impactful group is the DNA damage signatures, in this case Signature 3 and Signature 6. Signature 3 is associated with deficiency in DSB repair by homologous recombination, while Signature 6 is associated with defective DNA mismatch repair. There is also a small influence of the APOBEC dysregulation related Signature 2. Finally, there is a small influence of Signature 5, a signature of

unknown cause found in all cancer types.

SciClone Data

Figure 3.4.4.2 – Subclonal architecture in the TCHL 29 Pre-treatment sample

The TCHL 29 Pre-treatment shows 2 clusters, with the more dense cluster centred at a VAF of 5 and the less dense cluster centred at a VAF of 20. This suggests a founder clone at a VAF of 20 and a more mutated subclone at a VAF of 5.

Driver analysis

Table 3.4.4.1 – Known or predicted driver SNVs and indels for TCHL 29 Pre-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178917478|G|A	*PIK3CA*	c.353G>A	chr3:g.178917478 G>A	p.G118D	Act	known in: CANCER
X|76938973|G|GTGAT AAT	***ATRX***	**c.1774_1775ins ATTATCA**	**chrX:g.76938973_ 76938974insTGAT AAT**	**p.P592Hf s*8**	**LoF**	**predicted driver tier 1**
X|76849320|C|CATAT TTAT	***ATRX***	**c.5957-2_5957-1insATAAATAT**	**chrX:g.76849320_ 76849321insATAT TTAT**	**.**	**LoF**	**predicted driver tier 1**
8|30938665|T|TTCTGA AATATCCTTTA	***WRN***	**c.1122_1123ins TCTGAAATATC CTTTA**	**chr8:g.30938665_ 30938666insTCTG AAATATCCTTTA**	**p.E375Sfs *32**	**LoF**	**predicted driver tier 1**
8|103289348|C|CT	*UBR5*	c.6360dupA	chr8:g.103289356d upT	p.E2121Rf s*13	ambiguou s	predicted driver: tier 2
6|157524998|A|G	*ARID1B*	c.4895-2A>G	chr6:g.157524998 A>G	.	LoF	predicted driver: tier 1
5|56161678|A|AAGAA CACATTATAGTTT	***MAP3K1***	**c.1175_1176ins AGAACACATTA TAGTTT**	**chr5:g.56161678_ 56161679insAGA ACACATTATAGT TT**	**p.Y392*fs *1**	**LoF**	**predicted driver tier 1**
			chr5:g.131944389	**p.N934lfs**	**ambiguou s**	**predicted driver tier 2**

5|131944381|CA|C	***RAD50***	**c.2801delA**	**delA**	***6**		
3|180685928|C|T	***FXR1***	**c.1288C>T**	**chr3:g.180685928 C>T**	**p.R430***	**LoF**	**predicted driver tier 1**
3|105389149|T|TTATT CATTATATAATTTAA TGA	***CBLB***	**c.2616_2617ins TCATTAAATTA TATAATGAATA**	**chr3:g.105389149 _105389150insTA TTCATTATATAAT TTAATGA**	**p.R873Sf s*13**	**LoF**	**predicted driver tier 1**
1|120497815|A|AAACT GTAC	***NOTCH2***	**c.2066_2067ins GTACAGTT**	**chr1:g.120497815 _120497816insAA CTGTAC**	**p.N689Kf s*16**	**ambiguou s**	**predicted driver tier 2**
19|12813672|C|T	***TNPO2***	**c.2270G>A**	**chr19:g.12813672 C>T**	**p.R757Q**	**ambiguou s**	**predicted driver tier 2**

Input	gene	cdna	gdna	protein	gene_role	Driver statement
17|7577103|CA|C	*TP53*	c.834delT	chr17:g.7577104delA	p.R280Efs*65	LoF	predicted driver: tier 1
17|30521113|T|TAAAGTAA	*RHOT1*	c.857_858insAAGTAAA	chr17:g.30521114_30521115insAAGTAAA	p.Y286*fs*1	ambiguous	predicted driver: tier 2
17|30521110|G|GA	*RHOT1*	c.855dupA	chr17:g.30521112dupA	p.Y286Ifs*11	ambiguous	predicted driver: tier 2
17|15989696|G|GTAATAAATAATAA	***NCOR1***	**c.3076_3077insTTATTATTTATT A**	**chr17:g.15989697_15989698insAATAAATAATAAT**	**p.T1026Ifs*49**	**LoF**	**predicted driver tier 1**
11|94219172|A|AATATCATT	***MRE11A***	**c.231_232insAATGATAT**	**chr11:g.94219172_94219173insATATCATT**	**p.L78Nfs* 5**	**ambiguous**	**predicted driver tier 2**
10|62023637|C|T	*ANK3*	c.655G>A	chr10:g.62023637C>T	p.A219T	Act	predicted driver: tier 1

The TCHL 29 Pre-treatment sample shows 18 predicted driver mutations, a large number of driver mutations for a breast cancer sample. The only mutation definitively known to act as a driver is *PIK3CA* G118D. Several other samples in the cohort also show *PIK3CA* mutations. There is also a predicted driver mutation in *TP53*. Several other samples in the cohort bear predicted driver *TP53* mutations. The remaining predicted driver genes are either unique to the TCHL 29 samples or are only seen in one or two other samples in the cohort.

Mutations in the following genes are likely oncogenic: *PIK3CA, ANK3* Mutations

in the following genes are likely tumour suppressor: *ATRX, WRN, ARID1B, MAP3K1, CBLB, TP53, NCOR1*

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: *UBR5, RAD50, NOTCH2, TNPO2, RHOT1, MRE11A*

3.4.4.2 - TCHL 29 Post treatment

Mutational Signatures

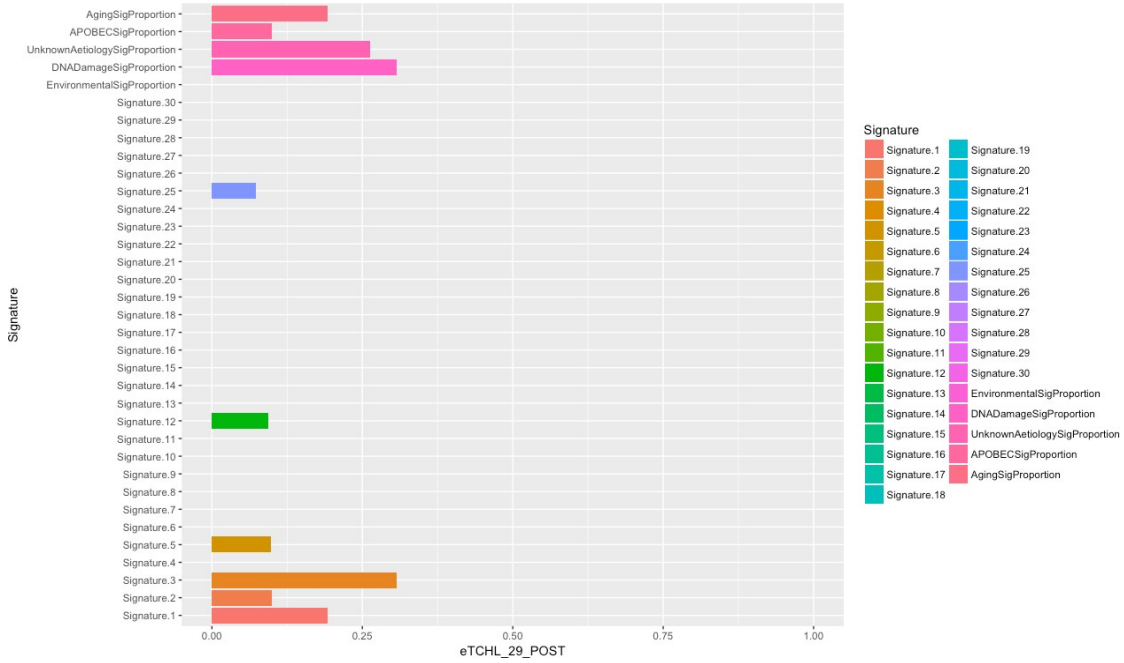


Figure 3.4.4.3 – Mutational signatures in the TCHL 29 Post-treatment sample

The TCHL 29 Post treatment sample shows a similar mutational landscape to the corresponding pre-treatment sample, though the proportion of the aging signature is reduced in comparison to that sample. It is hard to tell why this is the case, as referring to the table in section 3.5.1 shows that the Post treatment sample has far fewer SNVs and indels than the pre-treatment sample. As with the pre-treatment sample, DNA damage based signatures play a significant role, though in the post treatment sample only signature 3, the DSB repair deficiency signature, is present, and signature 6 is no longer present. The remainder of the mutational landscape in this sample is composed of APOBEC based signatures and signatures of unknown aetiology.

SciClone Data

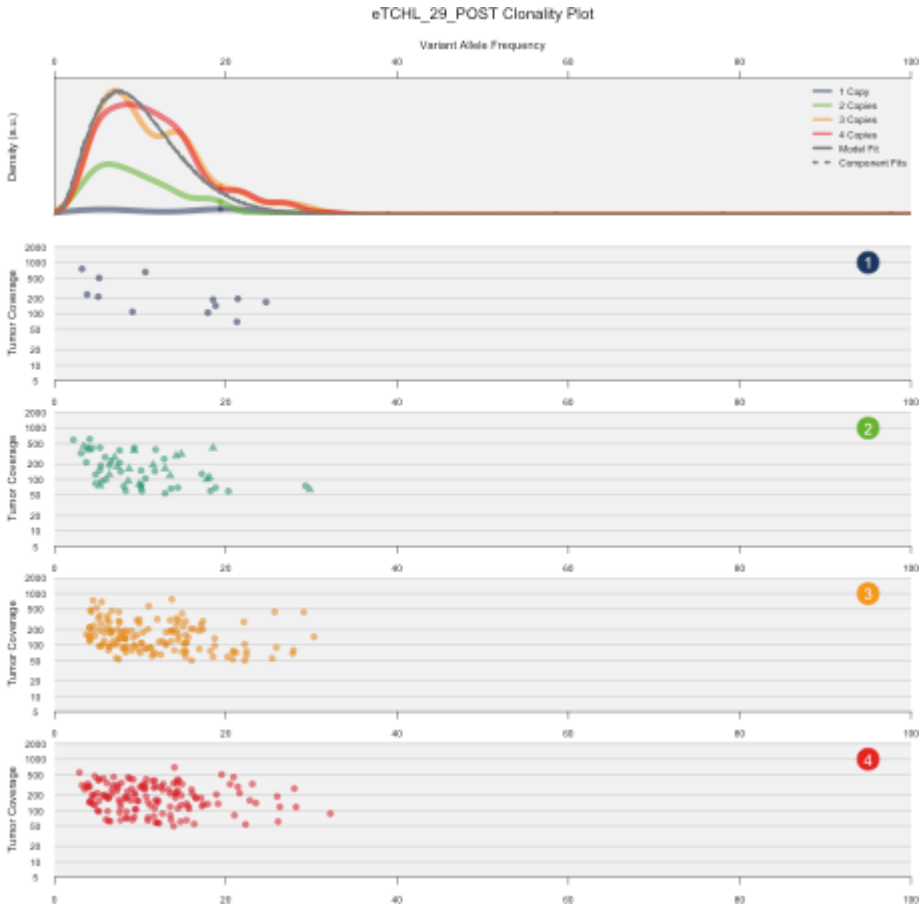


Figure 3.4.4.4 – Subclonal architecture in the TCHL 29 Post-treatment sample

The TCHL 29 Post-treatment sample shows 1 cluster at each copy number, implying there are no significant subclonal populations. This implies that only one of the subclonal populations from the pre-treatment sample survived treatment.

Driver analysis

Table 3.4.4.2 – Known or predicted driver SNVs and indels for TCHL 29 Post treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	driver_statement
3|178917478|G|A	*PIK3CA*	c.353G>A	chr3:g.178917478G>A	p.G118D	Act	known in: CANCER
8|103289348|C|CT	*UBR5*	c.6360dupA	chr8:g.103289356dup T	p.E2121Rfs*1 3	ambiguous	predicted driver: tier 2
6|157524998|A|G	*ARID1B*	c.4895- 2A>G	chr6:g.157524998A>G	.	LoF	predicted driver: tier 1
6|114292039|C|CT	***HDAC2***	**c.33dupA**	**chr6:g.114292048du pT**	**p.V12Sfs*8**	**LoF**	**predicted driver tier 1**
19|49458970|TG|T	***BAX***	**c.121delG**	**chr19:g.49458978del G**	**p.E41Rfs*19**	**LoF**	**predicted driver tier 1**
17|7577103|CA|C	*TP53*	c.834delT	chr17:g.7577104delA	p.R280Efs*65	LoF	predicted driver: tier 1
15|91304138|G|GA	***BLM***	**c.1544dupA**	**chr15:g.91304147dupA**	**p.N515Kfs*2**	**LoF**	**predicted driver tier 1**
10|62023637|C| T	*ANK3*	c.655G>A	chr10:g.62023637C>T	p.A219T	Act	predicted driver: tier 1

The TCHL 29 Post treatment sample shares the validated G118D driver mutation in PIK3CA seen in the Pre-treatment sample. Since PIK3CA mutations are associated with trastuzumab resistance, this may indicate that cells bearing this mutation were more likely to survive treatment longer than cells without this mutation in this sample. There are overall fewer predicted drivers in this sample than the corresponding Pre-treatment sample. Given that the Post treatment sample has fewer SNVs and indels than the Pre-treatment sample, as mentioned in the Mutational Signatures section, this implies that more heavily

mutated cells in this patient succumbed to treatment, and the surviving cells after treatment started had a lower mutational burden and so fewer driver mutations. Most of the predicted driver genes in this sample are shared with the Pre-treatment sample, though *HDAC2, BAX* and *BLM* are not shared. This may indicate that these predicted driver mutations occurred in cells in the tumour during therapy, possibly due to genotoxic effects of therapy.

Mutations in the following genes are likely oncogenic: *PIK3CA, ANK3* Mutations in the following genes are likely tumour suppressor inactivating: *ARID1B, HDAC2, BAX, TP53, BLM*

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: *UBR5*

3.4.5 - TCHL 32 samples

The TCHL 32 sample was not given lapatinib (TCH patient) and was a responder to initial therapy (showed pCR). The patient ultimately showed relapse in the brain.

3.4.5.1 - TCHL 32 Pre-treatment

Mutational Signatures

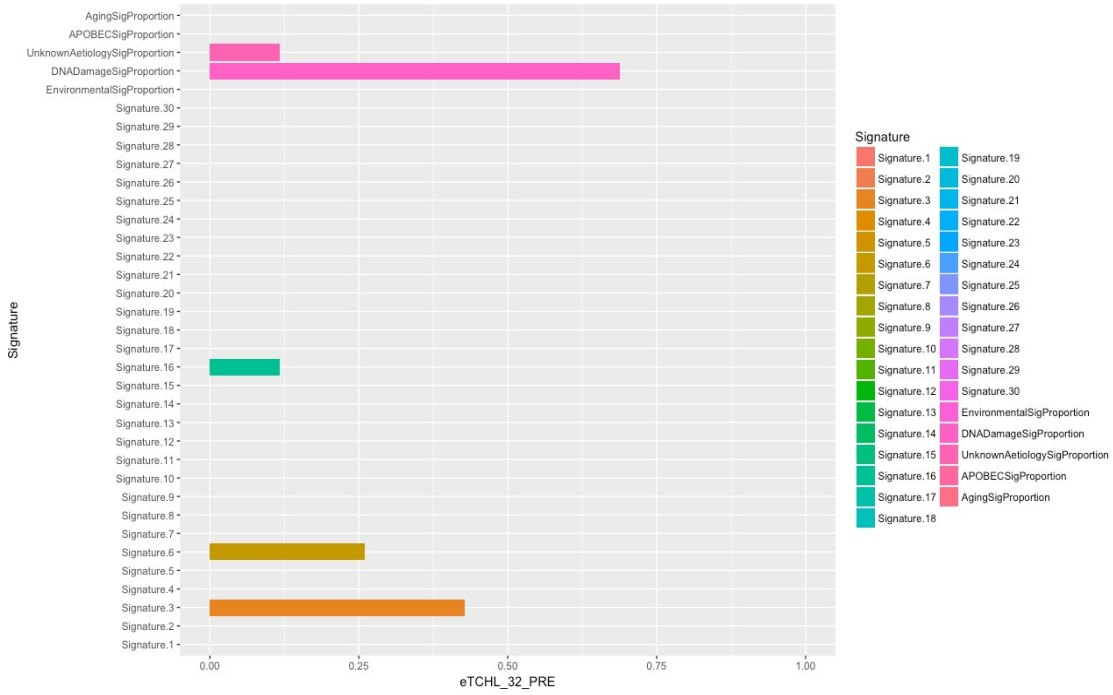


Figure 3.4.5.1 – Mutational signatures in the TCHL 32 Pre-treatment sample

The TCHL 32 Pre-treatment sample shows a mutational landscape completely dominated by the DNA damage repair deficiency signatures, specifically Signature 3 (DSB repair deficiency) and Signature 6 (DNA mismatch repair deficiency). There is also a small contribution by Signature 16, a signature of unknown aetiology. The fact that Signature 1 is entirely absent must be an error of the deconstructSigs algorithm, as Signature 1 is present in all somatic cells (see section 1.7 for an explanation of how deconstructSigs can mis-assign signatures).

SciClone Data

SciClone was unable to find any clusters in the data provided from this sample.

Driver analysis

Table 3.4.5.1 – Known or predicted driver SNVs and indels for TCHL 32 Pre-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178952085|A|G	*PIK3CA*	c.3140A>G	chr3:g.178952085A>G	p.H1047R	Act	known in: COREAD;NSCLC; OV;BRCA
17|7578212|G|A	*TP53*	c.637C>T	chr17:g.7578212G>A	p.R213*	LoF	known in: CANCER PR

The TCHL 32 Pre-treatment sample shows validated driver mutations in 2 genes that are mutated in many samples in this cohort: the H1047R mutation in *PIK3CA*, and the R123* mutation in *TP53*. This is a very low number of driver mutations for a tumour, possibly indicating that some driver mutations in this sample may have been filtered by the variant calling pipeline. As was said when a similar issue arose with the TCHL 6 relapse sample, we consider the possibility of mutations being erroneously filtered due to the fact that even the most advanced variant calling pipelines known today are known in the literature to have a non-zero false negative rate, potentially removing as high as 3% of genuine variants (144) . Also, as with TCHL 6 relapse, this result may indicate that there are driver mutations in this sample that are not known in the CGI database. All driver mutations in this sample are shared by the corresponding relapse sample.

Mutations in the following genes are likely oncogenic: *PIK3CA*

Mutations in the following genes are likely tumour suppressor inactivating: *TP53*

3.4.5.2 - TCHL 32C - Relapse sample

Mutational Signatures

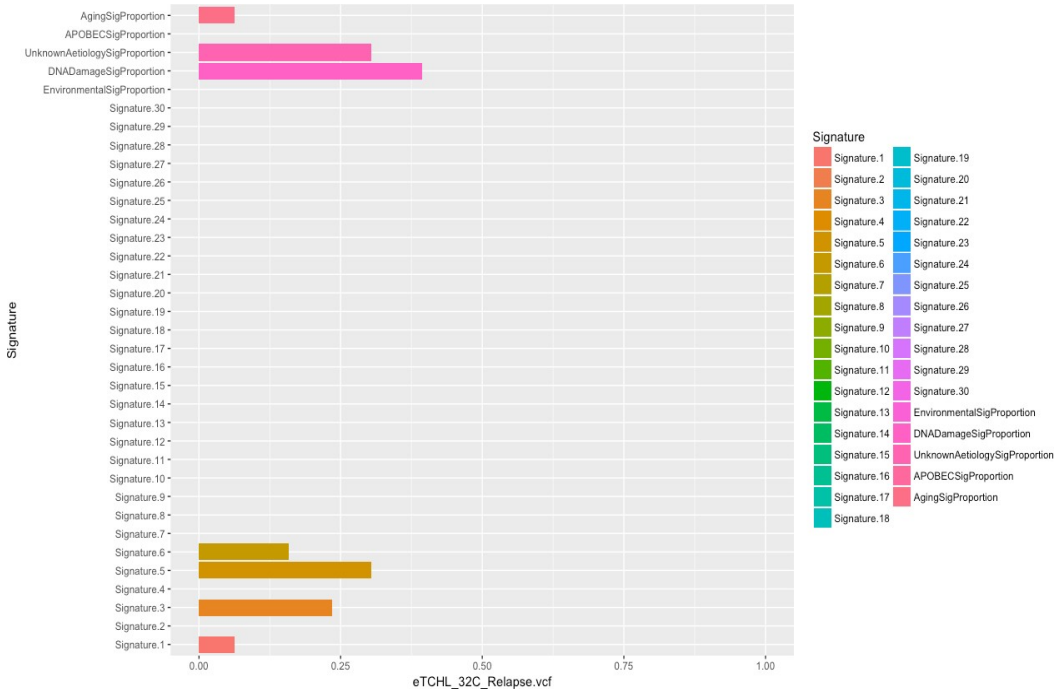


Figure 3.4.5.2 – Mutational signatures in the TCHL 32 Relapse sample

The 32C Relapse sample, like the Pre-treatment sample, shows a heavy influence of DNA damage based signatures 3 and 6. However, in the relapse sample there is a small influence of Signature 1 (which is probably also present in the Pre-treatment sample but not picked up by the deconstructSigs algorithm). The relapse sample also shows a sizable influence of Signature 5, a Signature of unknown cause found in all cancer types. These results suggest that whatever process causes signature 5 was active during the process of treatment and may have caused the driver mutations that lead to the cancer recurring. The fact that the relapse sample has many more SNVs and indels than the Pre- treatment sample (see the table in section 3.5.1) is further evidence that a mutagenic process was active between treatment starting and relapse occurring, generating a host of new mutations. These results implicate signature 5 as the likely cause of the new mutations seen in the relapse sample.

SciClone Data

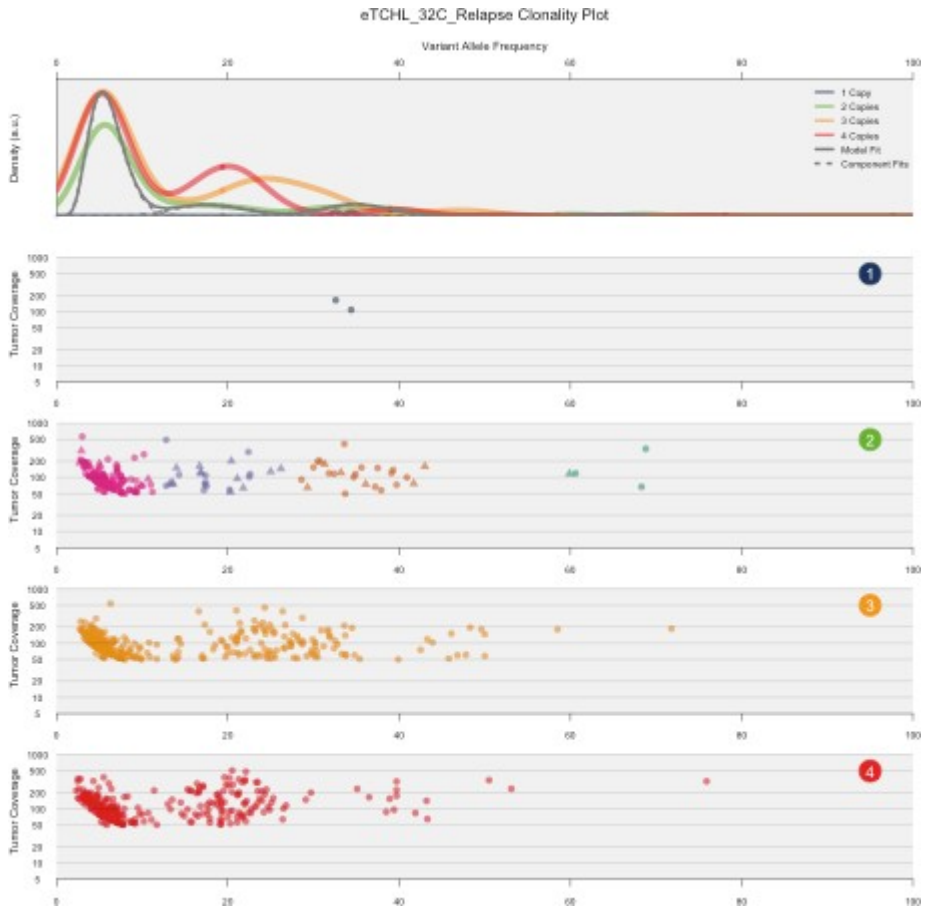


Figure 3.4.5.3 – Mutational signatures in the TCHL 32 Relapse sample

The 32C relapse sample shows a complicated subclonal architecture with 4 clusters. The graph implies a founder clone with a low mutational burden with a VAF centred at 40, then 3 subclonal populations with VAFs centred at 35, 20 and 5 respectively. The subclonal population with a VAF of 5 is the most heavily mutated part of the tumour (highest density).

Driver analysis

Table 3.4.5.2 – Known or predicted driver SNVs and indels for TCHL 32 Pre-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178952085|A|G	*PIK3CA*	c.3140A>G	chr3:g.178952085A>G	p.H1047R	Act	known in: BRCA;NSCLC;OV;CORE AD
17|7578212|G|A	*TP53*	c.637C>T	chr17:g.7578212G> A	p.R213*	LoF	known in: CANCER-PR
9|135772957|T|TATTAAGTGGAACTTC	***TSC1***	**c.2665_2666insGAAGTTCCACTTAAT**	**chr9:g.135772957_135772958insATTAAGTGGAACTTC**	**p.E889Gfs*5**	**LoF**	**predicted driver: tier 1**
6|87970733|A|ACTTAT	***ZNF292***	**c.7386_7387insCTTAT**	**chr6:g.87970733_87970734insCTTAT**	**p.R2463Lfs*24**	**LoF**	**predicted driver: tier 1**
6|87970731|T|TTATGG	***ZNF292***	**c.7384_7385insTATGG**	**chr6:g.87970731_87970732insTATGG**	**p.S2462Lfs*25**	**LoF**	**predicted driver: tier 1**
6|87970203|A|AGTTCAAATATGT	***ZNF292***	**c.6856_6857insGTTCAAATATGT**	**chr6:g.87970203_87970204insGTTCAAATATGT**	**p.K2286Sfs*5**	**LoF**	**predicted driver: tier 1**
6|87968495|A|AAGTCACTT	***ZNF292***	**c.5148_5149insAGTCACT T**	**chr6:g.87968495_87968496insAGTCACTT**	**p.A1717Sfs*6**	**LoF**	**predicted driver: tier 1**
5|96360338|A|AG		**c.2675_2676insGAAAATA**	**chr5:g.96360338_96360339insGAAAA**	**p.I893Kfs**	**ambiguou**	

AAAATATTCTG	***LNPEP***	**TTCTG**	**TATTCTG**	***4**	**s**	**predicted driver: tier 2**
5|67569784|C|CCTACTATTAA	***PIK3R1***	**c.445_446insCTACTATTAA**	**chr5:g.67569784_67569785insCTACTATTAA**	**p.L149Pfs*21**	**LoF**	**predicted driver: tier 1**
5|131930712|A|AGTTCC	***RAD50***	**c.1945_1946insGTTCC**	**chr5:g.131930712_131930713insGTTC C**	**p.I649Sfs*27**	**ambiguous**	**predicted driver: tier 2**
5|112176256|A|AGGTGGAGGTAATTT	***APC***	**c.4965_4966insGGTGGAGGTAATTT**	**chr5:g.112176256_112176257insGGTGGAGGTAATTT**	**p.S1656Gfs*7**	**LoF**	**predicted driver: tier 1**
2|48059522|C|CCAAATTATCT	***FBXO11***	**c.1363_1364insAGATAATTTG**	**chr2:g.48059522_48059523insCAAATTATCT**	**p.R455Qfs*5**	**LoF**	**predicted driver: tier 1**

Input	gene	cdna	gdna	protein	gene_role	Driver statement
21|35147307|A|AATTCCAAAGTCTTTTTCTTT	*ITSN1*	c.1491_1492insATTCCAAAGTCTTTTTCTTT	chr21:g.35147307_35147308insATTCCAAAGTCTTTTTCTTT	p.D498Ifs*13	LoF	predicted driver: tier 1
21|35094909|C|C T	*ITSN1*	c.147dupT	chr21:g.35094918dupT	p.Q50Sfs*17	LoF	predicted driver: tier 1
1|51323662|G|GAAGTTTA	*FAF1*	c.52_53insTAAACTT	chr1:g.51323662_51323663insAAGTTTA	p.T18Ifs*6	LoF	predicted driver: tier 1
1|51323661|A|ATATCTTTAATAT	*FAF1*	c.53_54insATATTAAAGATA	chr1:g.51323661_51323662insTATCTTTAATAT	p.G19Yfs*2	LoF	predicted driver: tier 1
1|21205985|A|G	*EIF4G3*	c.2303T>C	chr1:g.21205985A> G	p.L768S	Act	predicted driver: tier 2
1|120612018|C|T	*NOTCH2*	c.3G>A	chr1:g.120612018C>T	.	ambiguous	predicted driver: tier 2
17|40370235|TG| T	*STAT5B*	c.1102delC	chr17:g.40370243delG	p.Q368Rfs*2	ambiguous	predicted driver: tier 2
17|15968913|T|TTTCGTAGAATTTATTA	*NCOR1*	c.4836_4837insTAATAAATTCTACGAA	chr17:g.15968914_15968915insTCGTAGAATTTATTAT	p.I1613*fs*1	LoF	predicted driver: tier 1
16|67645475|G|GTGGC	*CTCF*	c.741_742insGGCT	chr16:g.67645476_67645477insGGCT	p.N248Gfs*2	LoF	predicted driver: tier 1

16|50813937|G|GTTTAGCTCTTC	***CYLD***	**c.1500_1501insTTTAGCTCTTC**	**chr16:g.50813937_50813938insTTTAGCTCTTC**	**p.L501Ffs*32**	**LoF**	**predicted driver: tier 1**
15|91304138|G|G	***BLM***	**c.1544dupA**	**chr15:g.91304147dupA**	**p.N515Kfs*2**	**LoF**	**predicted driver: tier 1**
13|32914971|A|ACGAGGAAGTATTTTTG	***BRCA2***	**c.6479_6480insCGAGGAAGTATTTTTG**	**chr13:g.32914971_32914972insCGAGGAAGTATTTTTG**	**p.Q2160Hfs*21**	**LoF**	**predicted driver: tier 1**
12|99060116|A|ATACATAACACCTAG	***APAF1***	**c.1343_1344insTACATAACACCTAG**	**chr12:g.99060116_99060117insTACATAACACCTAG**	**p.K448Nfs*3**	**LoF**	**predicted driver: tier 1**
11|63964969|A|AATGTAAGTCATGT	***STIP1***	**c.804_805insATGTAAGTCATGT**	**chr11:g.63964969_63964970insATGTAAGTCATGT**	**p.Y269Mfs*2**	**LoF**	**predicted driver: tier 1**

The TCHL 32C relapse sample shows 26 predicted and validated driver mutations, a far higher number of than the corresponding Pre-treatment sample. There are so many predicted driver mutations in this sample that it is

unlikely they are all genuinely contributing to the cancer. The presence of so many new predicted driver mutations in the relapse sample suggests that the tumour in this patient underwent heavy mutagenesis during the treatment process, possibly due to whatever process causes COSMIC Signature 5 (see Mutational Signatures section). The fact that there are many more SNVs and indels in this sample than in the corresponding Pre-treatment sample lends credence to this hypothesis. The *TP53* and *PIK3CA* mutations are shared with the Pre-treatment sample, and many other samples in the cohort show predicted driver mutations in these genes. The remaining predicted driver mutations in this sample are in genes that either only show predicted driver mutations in this sample, or genes that show predicted driver mutations in only one or two other samples in the cohort. It is also interesting to note that the vast majority of predicted driver mutations in this sample appear to be tumour suppressor inactivating mutations.

Mutations in the following genes are likely oncogenic: *PIK3CA, EIF4G3*
Mutations in the following genes are likely tumour suppressor inactivating: *TP53, TSC1, ZNF292, PIK3R1, APC, FBXO11, ITSN1, FAF1, NCOR1, CTCF, CYLD, BLM, BRCA2, APAF1, STIP1*

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: *LNPEP, RAD50, NOTCH2, STAT5B*

3.4.6 TCHL 39 samples

The TCHL 39 patient was not given lapatinib (TCH patient) and was a non-responder to initial therapy (no pCR). The patient ultimately showed relapse in a lymph node.

3.4.6.1 - TCHL 39 Pre-treatment

Mutational Signatures

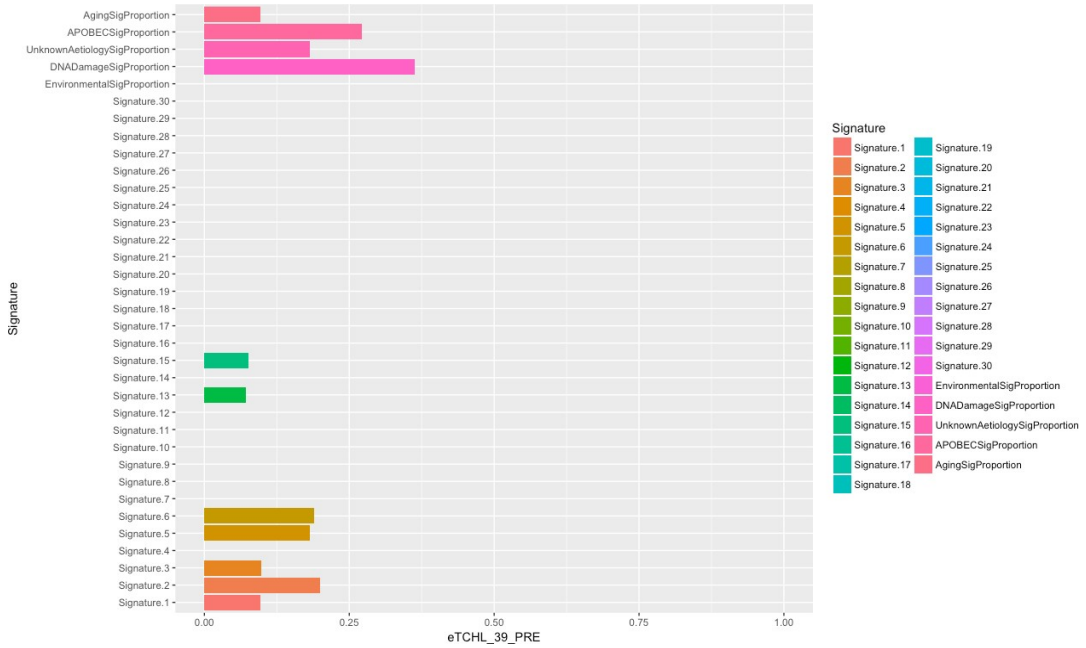


Figure 3.4.6.1- Mutational signatures in the TCHL 39 Pre-treatment sample

The TCHL 39 Pre-treatment sample shows a mutational landscape with a variety of signature groups present, though DNA damage signatures and APOBEC signatures are predominant. The DNA damage signatures present are Signature 3 (DSB repair deficiency), Signature 6 (DNA mismatch repair deficiency) and Signature 15 (DNA mismatch repair). The APOBEC signatures are Signature 2 and 13. Signatures of unknown aetiology also show an impact in the form of Signature 5 (found in all cancer types, seen in several other samples in this cohort). Finally, the Aging Signature, Signature 1, makes a minor contribution. Overall, this mutational spectrum paints the picture of a

cancer whose mutational spectrum (and therefore, probably, its driver mutations) was created mostly by the breakdown in normal cellular processes for maintaining the integrity of the DNA.

SciClone Data

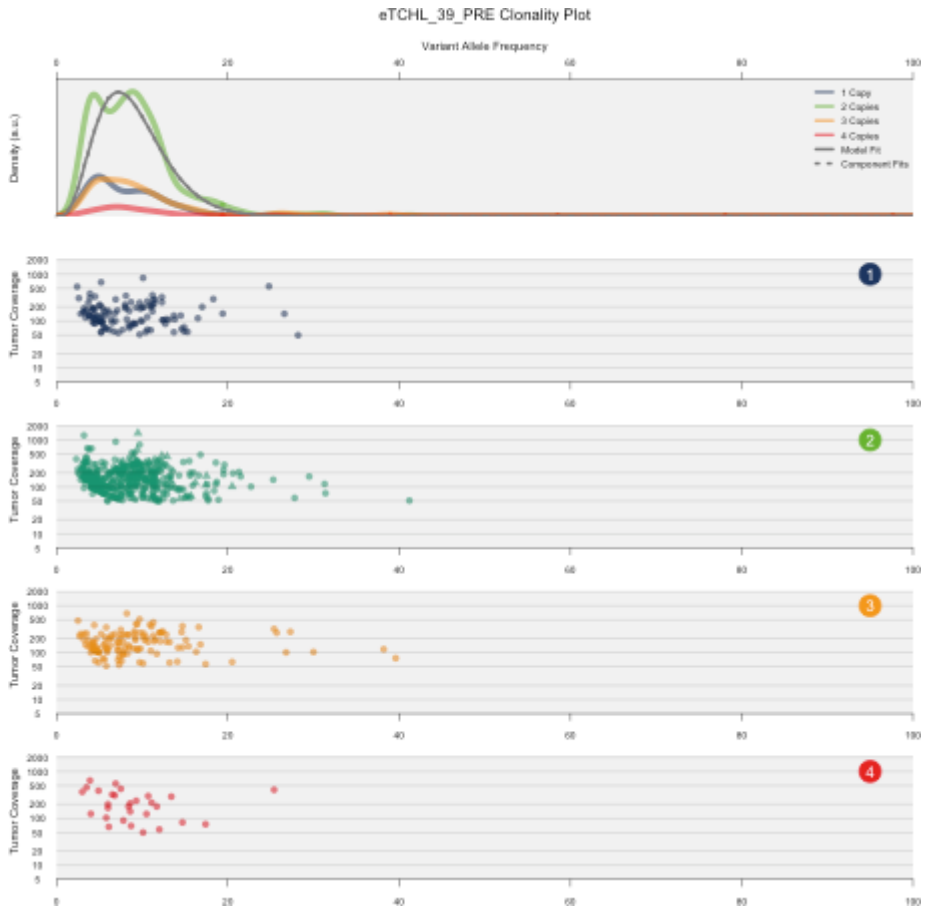


Figure 3.4.6.2- Subclonal architecture in the TCHL 39 Pre-treatment sample

The TCHL 39 Pre-treatment sample shows one cluster per copy number, implying that there are no significant subclonal populations.

Driver analysis

Table 3.4.6.1 – Known or predicted driver SNVs and indels for TCHL 39 Pre-treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
4|83785564|A|AT	***SEC31A***	**c.1384dupA**	**chr4:g.83785573dupT**	**p.I462Nfs*2**	**LoF**	**predicted driver: tier 1**
3|52582227|G|A	***PBRM1***	**c.4601C>T**	**chr3:g.52582227G>A**	**p.S1534L**	**LoF**	**predicted driver: tier 1**
18|20529651|G|A	*RBBP8*	c.223G>A	chr18:g.20529651G>A	p.E75K	LoF	predicted driver: tier 1

The TCHL 39 Pre-treatment shows 3 predicted driver mutations and no definitively known driver mutations. These mutations are all likely tumour suppressor inactivating and occur in genes that show predicted driver mutations in few other samples in the cohort.

Mutations in the following genes are likely tumour suppressor inactivating: *SEC31A, PBRM1, RBBP8*

3.4.6.2 - TCHL 39 Post treatment

Mutational Signatures

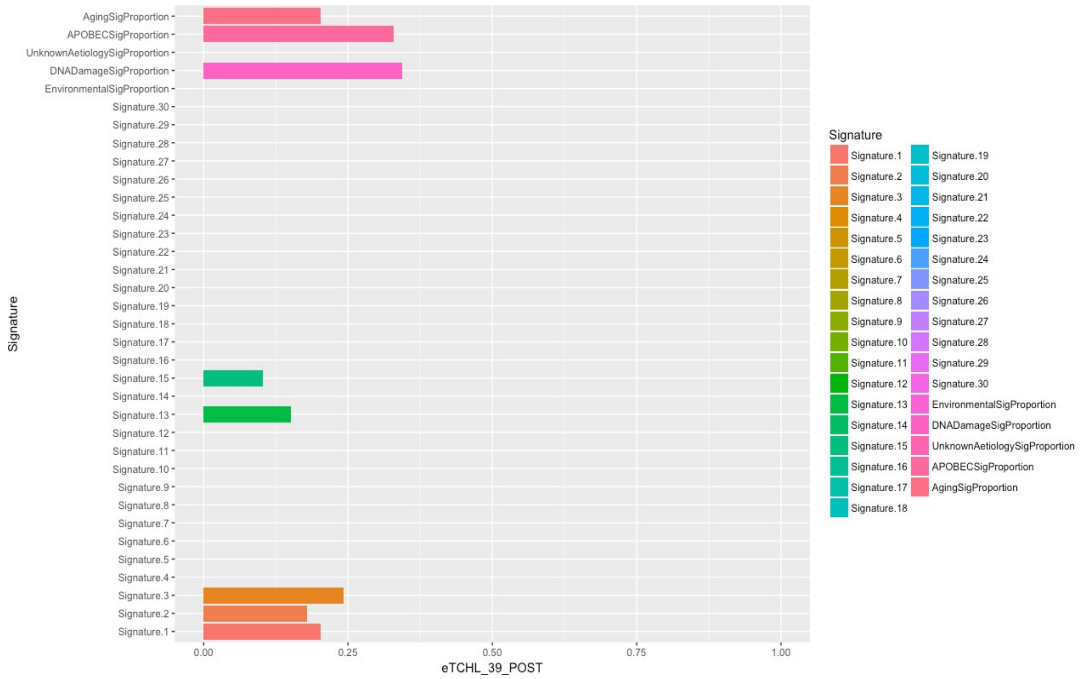


Figure 3.4.6.3 - Mutational signatures in the TCHL 39 Post-treatment sample

The TCHL 39 Post Treatment sample shows a mutational spectrum composed mostly of DNA damage and APOBEC related signatures. As with the corresponding Pre-treatment sample, the DNA damage related signatures include Signature 3 (DSB repair deficiency) and Signature 15 (DNA mismatch repair deficiency), though Signature 6 is now absent. The APOBEC signatures 2 and 13 are present in fairly similar proportions to the Pre-treatment sample. The aging signature is present to a slightly larger degree than the Pre-treatment sample. Signature 5 (aetiology unknown) is absent entirely from this sample, despite making a substantial contribution to the Pre-treatment sample.

SciClone Data

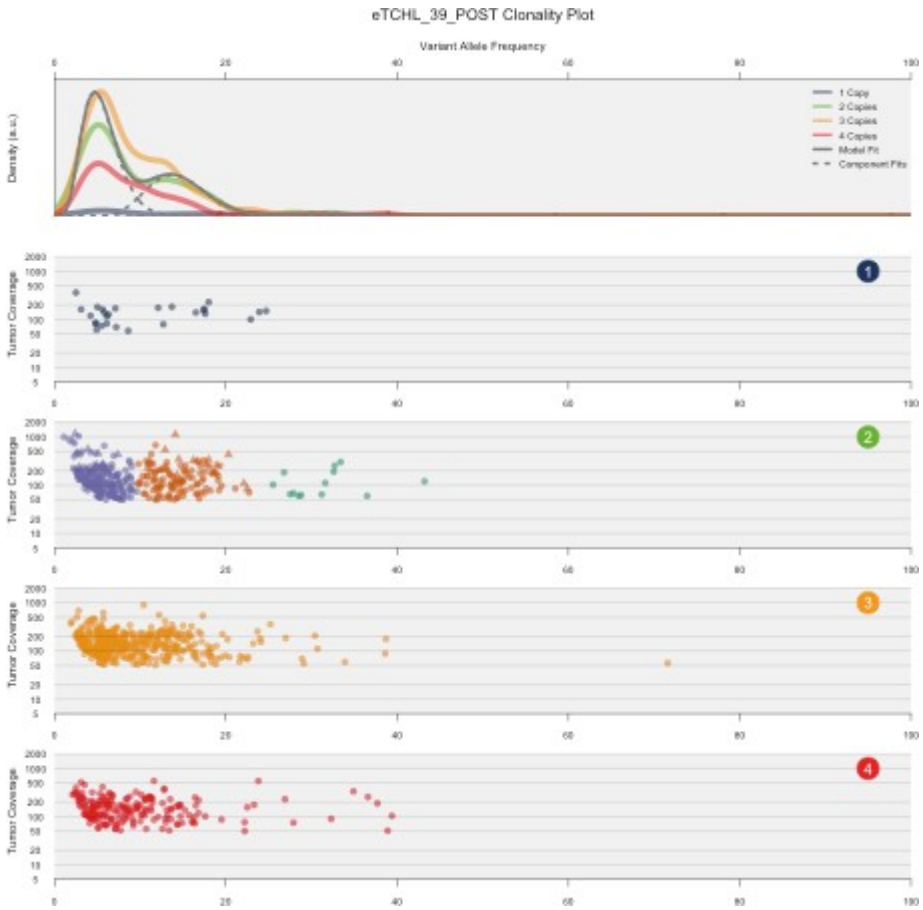


Figure 3.4.6.4- Subclonal architecture in the TCHL 39 Post-treatment sample

The TCHL 39 Post treatment sample shows a subclonal architecture with 3 clusters. The highest VAF cluster, corresponding to the least mutated subclonal population, is centred at a VAF of 35. The next highest VAF cluster, corresponding to the second most mutated subclonal population, is centred at a VAF of 15. Finally, the most mutated subclonal population is centred at a VAF of 5.

Driver analysis

Table 3.4.6.2 – Known or predicted driver SNVs and indels for TCHL 39 Post treatment. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
6|157522410|G|A	***ARID1B***	**c.4682G> A**	**chr6:g.157522410G >A**	**p.R1561H**	**LoF**	**predicted driver: tier 1**
18|51013310|G|A	***DCC***	**c.3880G> A**	**chr18:g.51013310G >A**	**p.G1294R**	**LoF**	**predicted driver: tier 2**
18|20529651|G|A	*RBBP8*	c.223G>A	chr18:g.20529651G >A	p.E75K	LoF	predicted driver: tier 1
11|93526897|C|G	*MED17*	c.641C>G	chr11:g.93526897C >G	p.S214C	Act	predicted driver: tier 2

The TCHL 39 Post treatment sample shows 4 predicted driver mutations, one of which is shared with the Pre-treatment sample (the *RBBP8* E75K mutation). The other mutations are new to the Post treatment sample. The fact that the Post treatment sample shows a greater number of SNVs and Indels than the Pre- treatment sample (see the table in section 3.5.1), that the Post treatment sample shows a different mutational signature spectrum to the Pre-treatment sample (with the absence of Signature 5) and that the Post treatment samples shows a very different subclonal architecture to the Pre-treatment sample suggests that these new predicted driver mutations are the result of the cells that survived therapy up to the point the Post treatment sample was

The predicted driver mutations present in the Pre-treatment sample but absent in this sample (the *SEC31A* and *PBRM1* mutations) may be associated with vulnerability to therapy, while the mutations seen only in this sample and not in the Pre-treatment sample (the *DCC, MED17* and *ARID1B* mutations) may be associated with enhanced survival of therapy.

Mutations in the following genes are likely oncogenic: *MED17*

Mutations in the following genes are likely tumour suppressor inactivating: *DCC, RBBP8, ARID1B*

3.4.6.3 - TCHL 39C - Relapse sample

Mutational Signatures

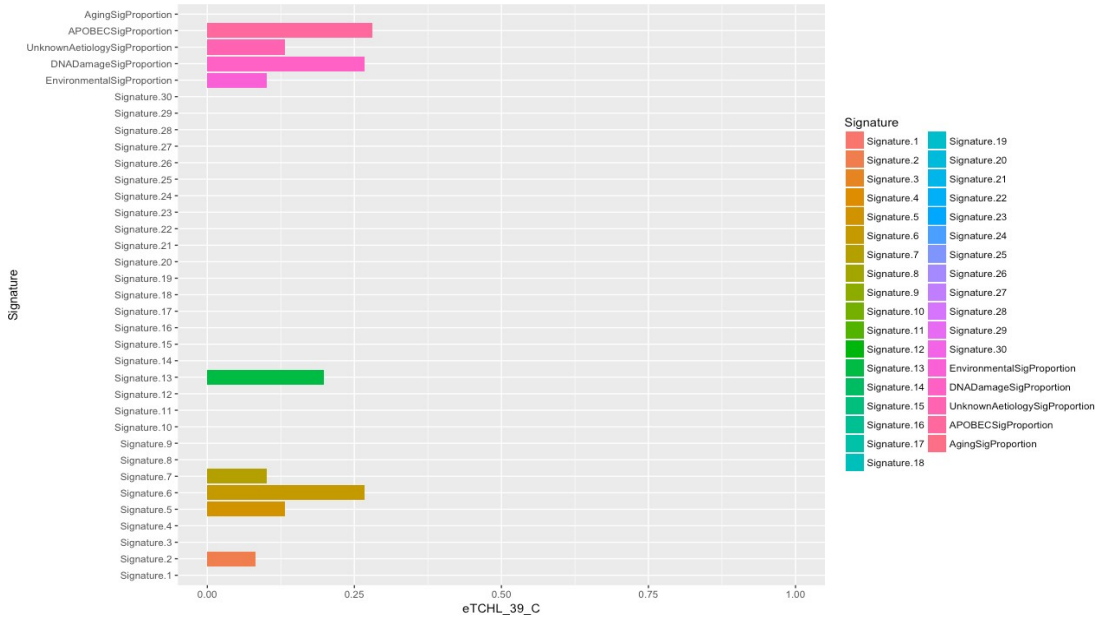


Figure 3.4.6.5 - Mutational signatures in the TCHL 39 Relapse sample

The TCHL 39 C sample (the relapse sample) shows a striking absence of Signature 1 compared to the Pre or Post treatment samples, as well as the presence of an environmental signature, signature 7, which is not present in the pre or post treatment samples. Since, as explained above, Signature 1 is necessarily present in all somatic cells, the absence of it on this spectrum is likely the result of an error of the deconstructSigs program rather than a genuine absence of that signature in this sample. The presence of Signatures 2, 5, 6 and 13 is shared with the Pre-treatment sample. The relapse sample has fewer SNVS and indels than the other samples from the TCHL 39 patient (see section 3.5.1) and deconstructSigs tends to be less accurate the fewer mutations are present, possibly explaining the odd mutational spectrum assigned to this sample.

SciClone Data

SciClone was unable to find any clusters in the data provided from this sample.

Driver analysis

Table 3.4.6.3 – Known or predicted driver SNVs and indels for TCHL 39 Relapse. Predicted or known driver genes that are not shared with other samples from the same patient are highlighted in bold.

Input	gene	cdna	gdna	protein	gene_role	Driver statement
7|6017386|G|T	***PMS2***	**c.2278C>A**	**chr7:g.6017386 G>T**	**p.P760T**	**LoF**	**predicted driver tier 2**
2|25458646|C|T	***DNMT3A***	**c.2527G>A**	**chr2:g.25458646 C>T**	**p.G843S**	**LoF**	**predicted driver tier 2**
12|57501959|C|G	***STAT6***	**c.103G>C**	**chr12:g.57501959C>G**	**p.E35Q**	**Act**	**predicted driver tier 2**
11|93526897|C|G	*MED17*	c.641C>G	chr11:g.93526897C>G	p.S214C	Act	predicted driver: tier 2

The TCHL 39 relapse treatment sample shows 4 predicted driver mutations. The *MED17* S214C mutation is shared by the 39 Post treatment sample. The other predicted driver genes in this sample are in genes that do not show predicted driver mutations in any other sample in this cohort. The relapse sample shows far fewer SNVs and Indels overall than the Pre or Post treatment samples.

Overall these facts suggest that heavily mutated cells in this tumour were killed by the treatment, but surviving cells with a lower mutational burden underwent mutagenesis during the treatment process, generating the new predicted driver mutations seen here, which may be responsible for the relapse.

Mutations in the following genes are likely oncogenic: *STAT6, MED17*
Mutations in the following genes are likely tumour suppressor inactivating: *PMS2, DNMT3A*

3.4.7 TCHL 45

Mutational Signatures

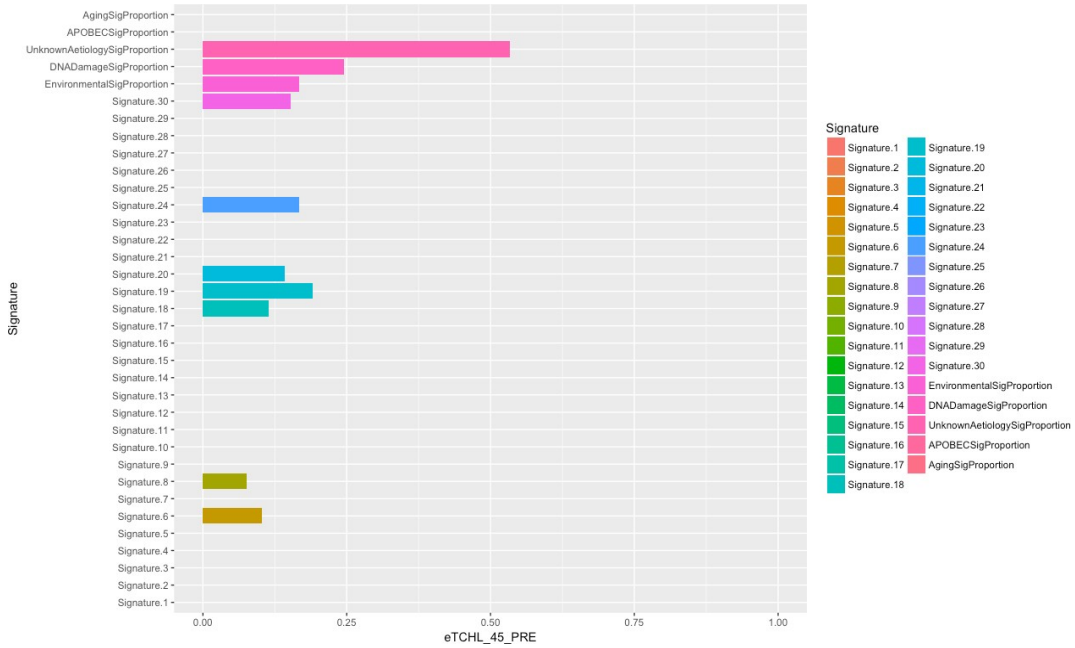


Figure 3.4.7.1 – TCHL 45 Mutational Signatures

The TCHL 45 Pre-treatment sample shows a mutational spectrum dominated by signatures of unknown origin. It is therefore hard to say what caused the majority of mutations in the tumour and what caused the driver mutations that lead to tumourigenesis. We can see some evidence of DNA mismatch repair errors playing a role, due to the presence of Signatures 6 and 20. There is also a relatively minor environmental influence in the form of signature 24, a signature associated with aflatoxin exposure.

SciClone Data

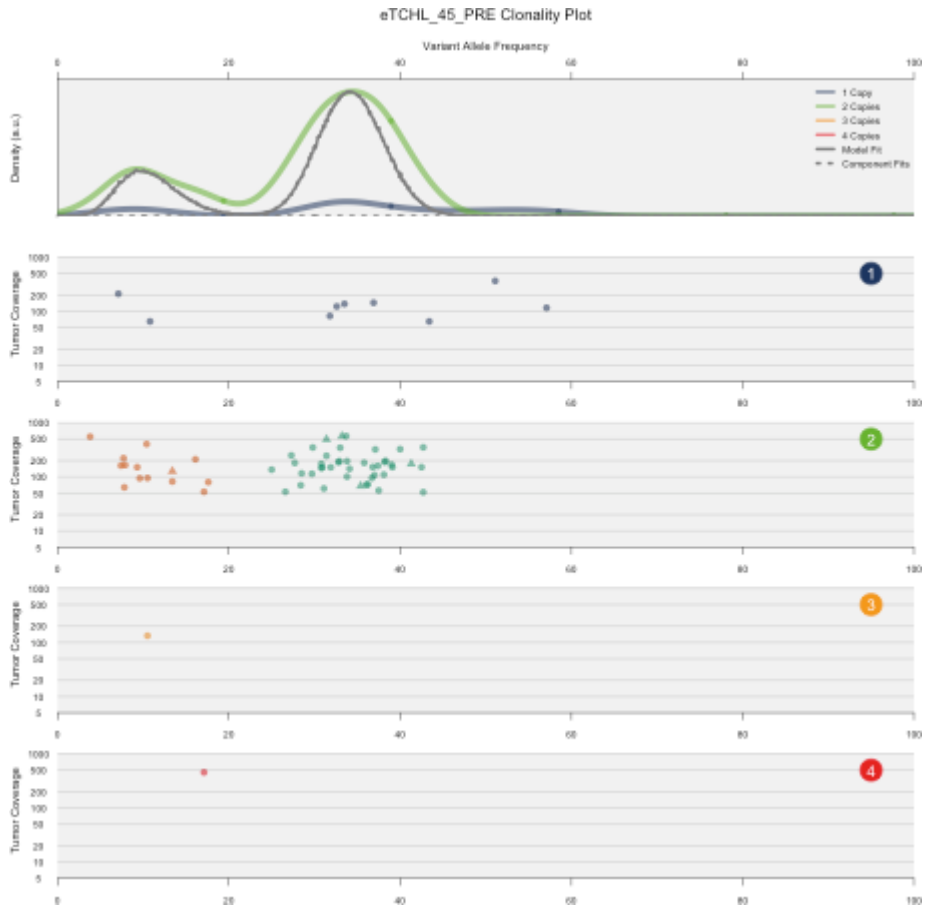


Figure 3.4.7.2 – TCHL 45 SciClone clonality plot.

Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
13|25671267|C|T	PABPC3	c.931C>T	chr13:g.25671267C>T	p.R311W	ambiguous	predicted driver: tier 2
12|6692411|C|A	CHD4	c.4013G>T	chr12:g.6692411C>A	p.R1338I	Act	predicted driver: tier 1

The TCHL 45 Pre-treatment sample shows only 2 predicted driver mutations, both of which are shared with few other samples in the cohort. As discussed for

other samples with a similar number of predicted driver mutations, this seems a low number of driver mutations for a cancer to have, suggesting that there are further driver mutations in the sample that were either filtered out during variant calling or were not recognised by the CGI algorithm.

Mutations in the following genes are likely oncogenic: CHD4

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: PABPC3

3.5 SciClone time point comparisons

3.5.1 TCHL 3 Sample comparisons

TCHL 3 Pre vs Post

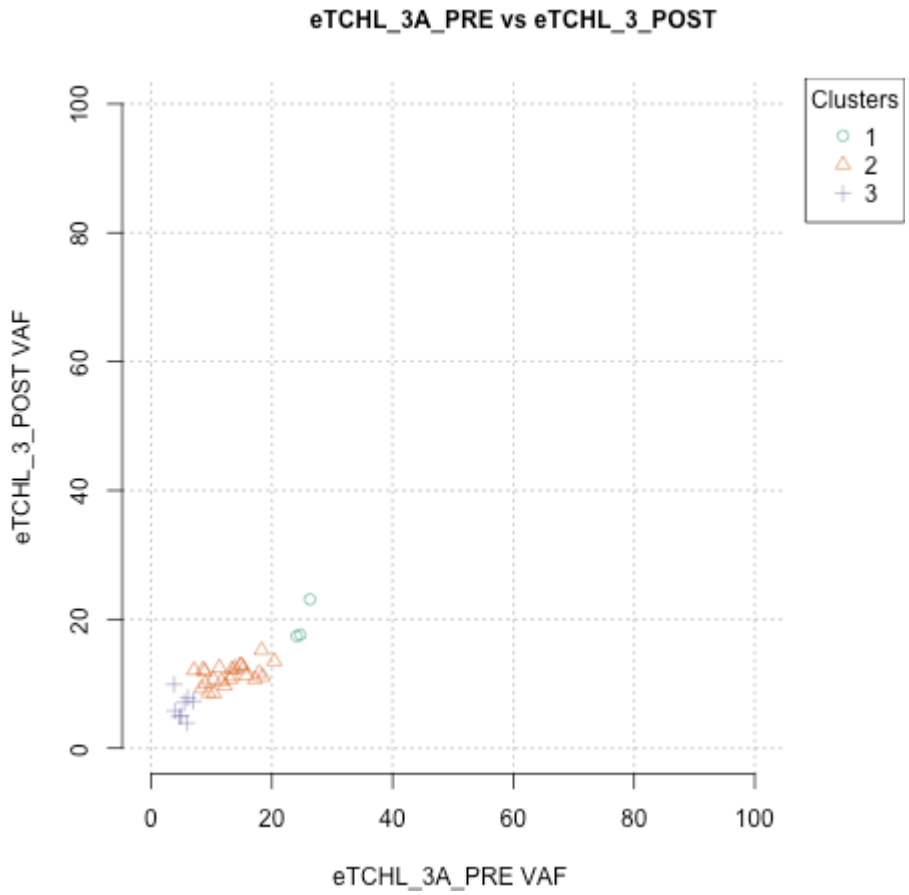


Figure 3.5.1 – Subclonal architecture in the TCHL 3 Pre and Post treatment samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The comparison of the clonal architecture of the TCHL 3 pre and post treatment samples shows 3 clusters, compared to the 2 clusters discovered when analysing the Pre-treatment sample by itself. The first and third cluster (purple crosses and green circles) show similar VAFs in both samples, suggesting that therapy has little effect on these subclones. The middle cluster (orange

triangles) shows a reduced VAF in the Post treatment sample, suggesting that therapy is successfully killing some cells in this subclone.

TCHL 3 Post vs Surgery

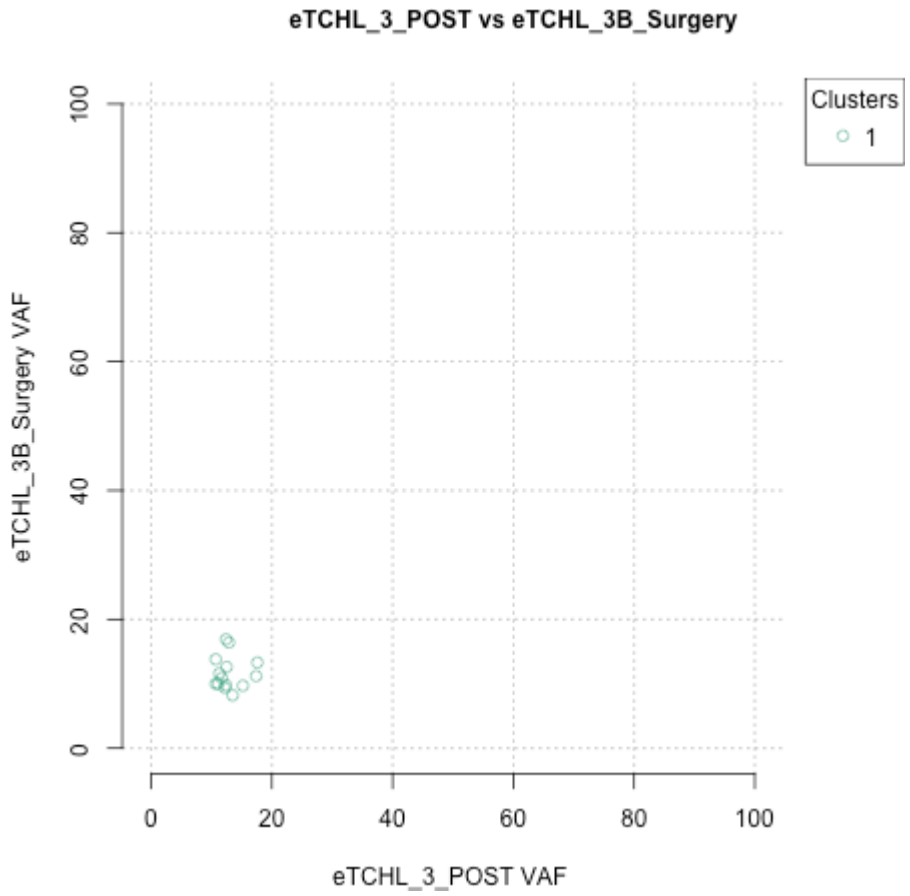


Figure 3.5.2 – Subclonal architecture in the TCHL 3 Post treatment and surgery samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The comparison of the clonal architecture of the TCHL 3 post treatment and surgery samples shows one cluster, with similar VAFs in both samples. This suggests that only one subclonal population survives the full therapy process, but that said subclone does not reduce in population between the time the post treatment sample was taken and the time the surgery sample was taken.

TCHL 3 Pre vs Surgery

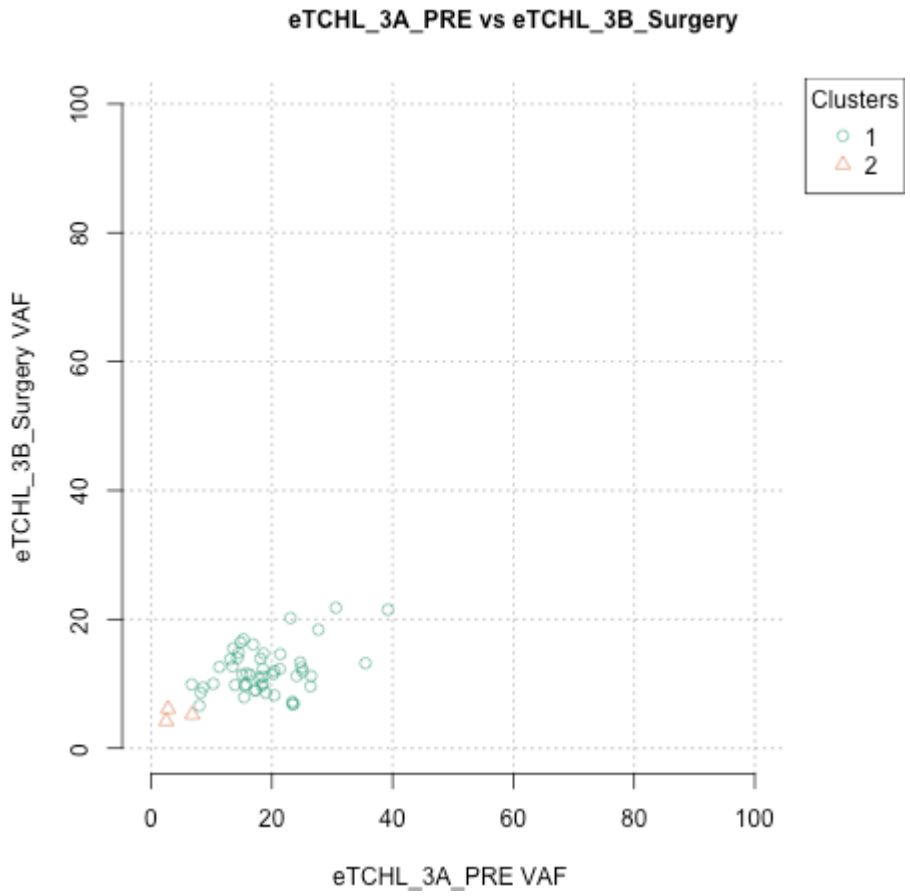


Figure 3.5.3 – Subclonal architecture in the TCHL 3 Pre-treatment and surgery samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The Pre-treatment sample vs post treatment sample comparison shows two clusters, suggesting two subclone in common between the two samples. Cluster 2, the orange triangles, is comprised of a small number of mutations at a very low VAF, suggesting a small-sized subclone with a low mutational burden that may have survived therapy due to its small size. The other cluster shows consistently lower VAFs in the surgery sample than the Pre-treatment sample, showing how the treatment process reduces the VAFs of the mutations in the cluster by killing cells with those mutations.

3.5.3 TCHL 6 sample comparisons

TCHL 6 Pre vs Post

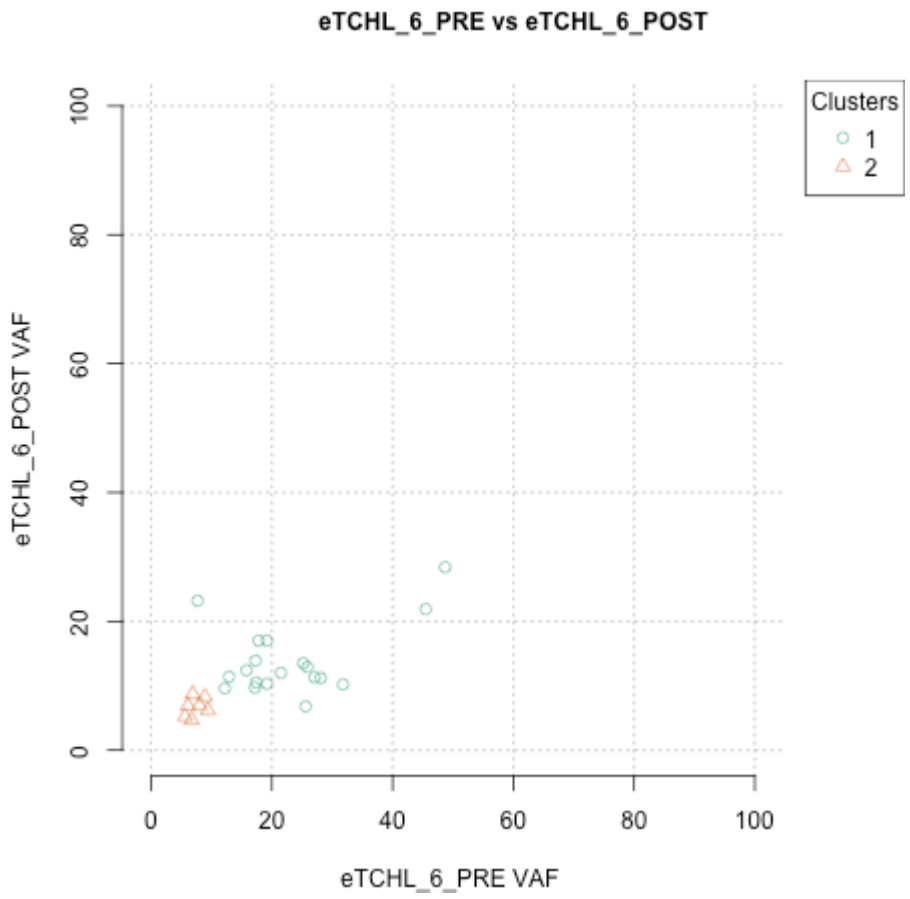


Figure 3.5.4 – Subclonal architecture in the TCHL 6 Pre and Post treatment samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The TCHL 6 pre vs post treatment comparison shows 2 clusters, suggesting two subclonal populations shared by the samples, consistent with the results from the solo sample Sciclone analyses above. Cluster 2 (orange triangles) is comprised of a small number of mutations at a very low VAF, suggesting a small-sized subclone with a low mutational burden that may have survived therapy due to its small size. Cluster 1 shows lower VAFs overall in the Post treatment sample, showing that the treatment process has diminished the size of the subclone and its associated mutations.

TCHL 6 Pre vs Relapse

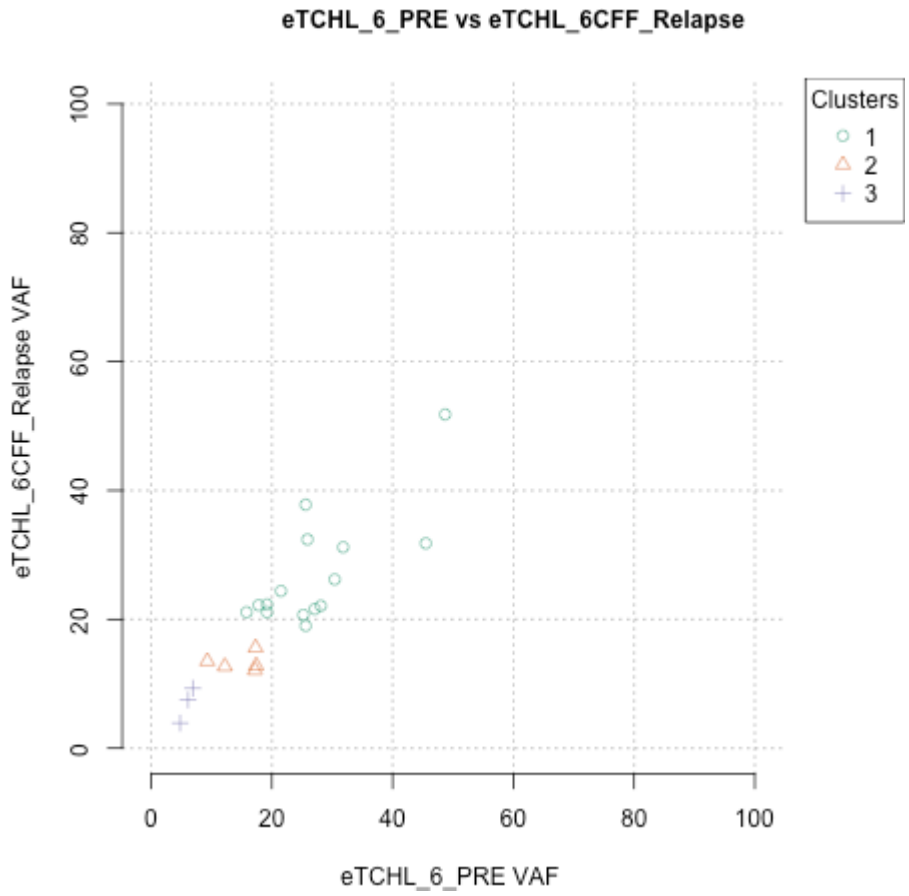


Figure 3.5.5 – Subclonal architecture in the TCHL 6 Pre-treatment and relapse samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The TCHL 6 Pre-treatment vs Relapse comparison shows 3 clusters shared by the samples, suggesting 3 subclonal populations in common. All clusters show similar VAFs in both samples. In the context of some subclones showing lower VAFs in the Post treatment sample above, this suggests that some subclones in the tumour were diminished during treatment, but then grew back to their original frequency after treatment during relapse.

TCHL 6 Post vs Relapse

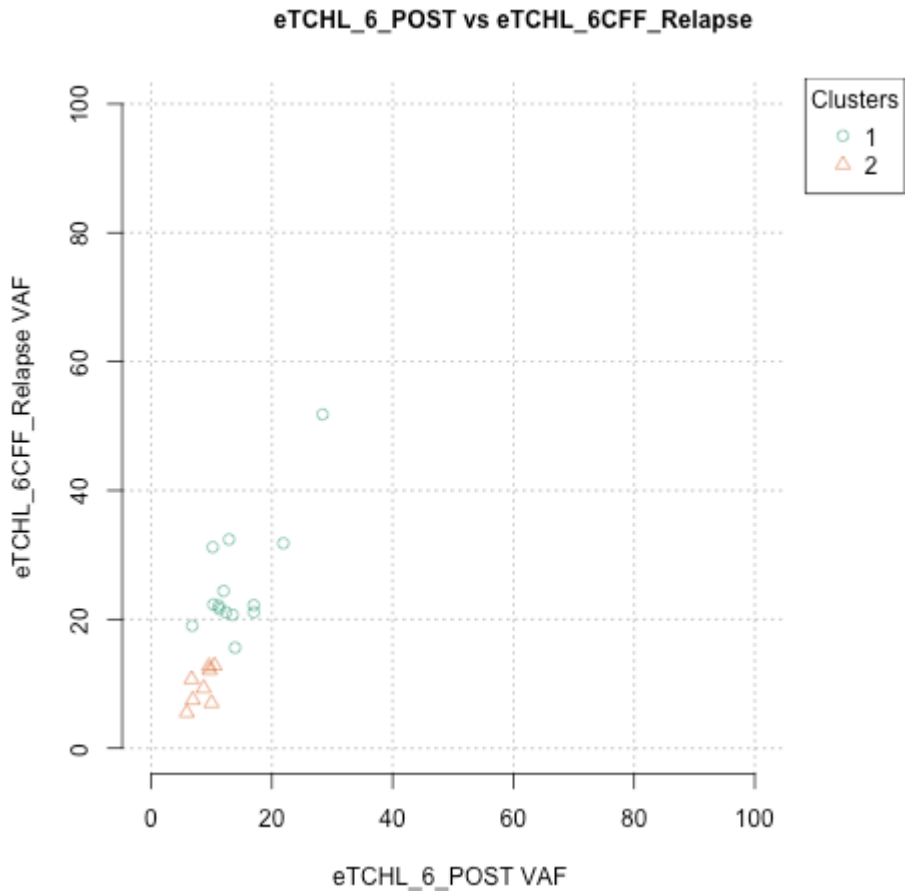


Figure 3.5.6 – Subclonal architecture in the TCHL 6 Post-treatment and Relapse samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The TCHL 6 Post treatment vs Relapse comparison shows 2 clusters shared between the samples, suggesting two subclonal populations shared by the samples. Both clusters, especially cluster 1 (orange triangles), show higher VAFs in the Relapse sample than the post treatment sample. This gives further credence to the hypothesis above, that some subclones in the tumour were diminished during treatment, but then grew back to their original frequency after treatment during relapse. Though TCHL 6 did not show pCR during therapy, these results suggest that therapy did initially succeed in reducing some subclonal populations, but that these populations grew back during the relapse.

3.5.3 TCHL 12 sample comparisons

SciClone was unable to find any clusters in the data provided when comparing these samples.

3.5.4 TCHL 29 sample comparisons

TCHL 29 Pre vs Post

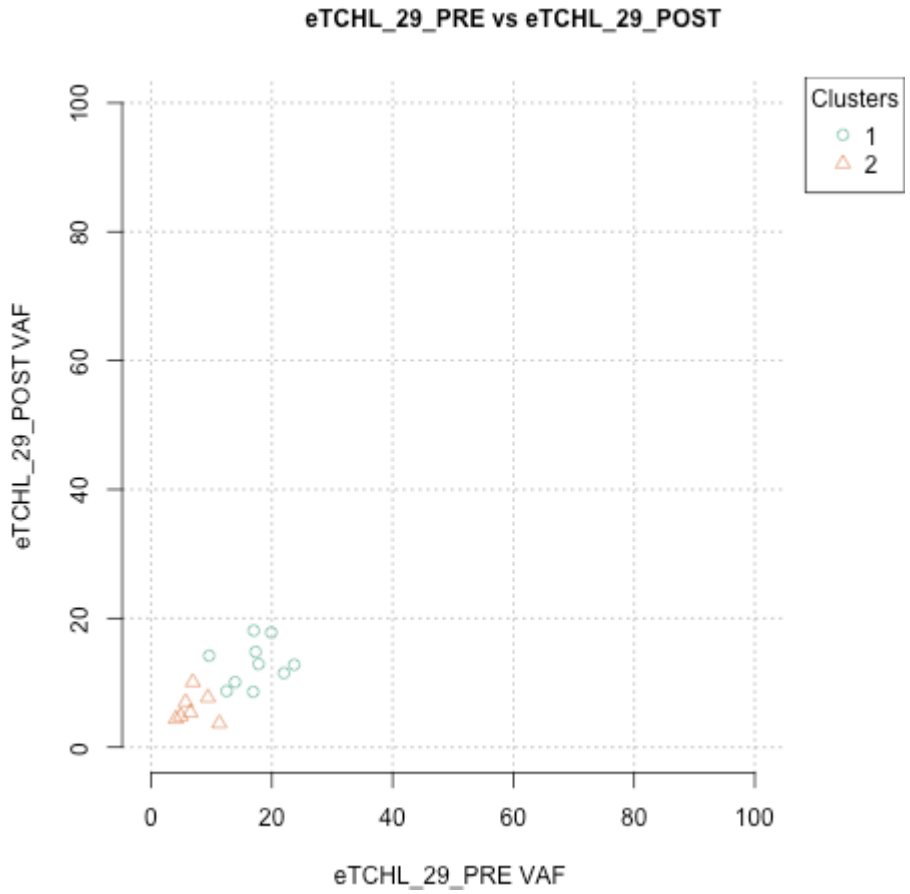


Figure 3.5.7 – Subclonal architecture in the TCHL 29 Pre and Post treatment samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The TCHL 29 Pre-treatment vs Post treatment sample comparison shows two clusters, suggesting two subclonal populations shared between the cells. This is consistent with the results of the solo SciCLone analysis the TCHL 29 Pre-treatment sample. Both clusters show similar VAFs in both samples, suggesting that therapy was not very effective in this sample. This is potentially linked to the fact that TCHL 29 was a non-responder sample (i.e. did not show pCR during therapy).

3.5.5 TCHL 32 sample comparisons

TCHL 32 Pre vs Relapse

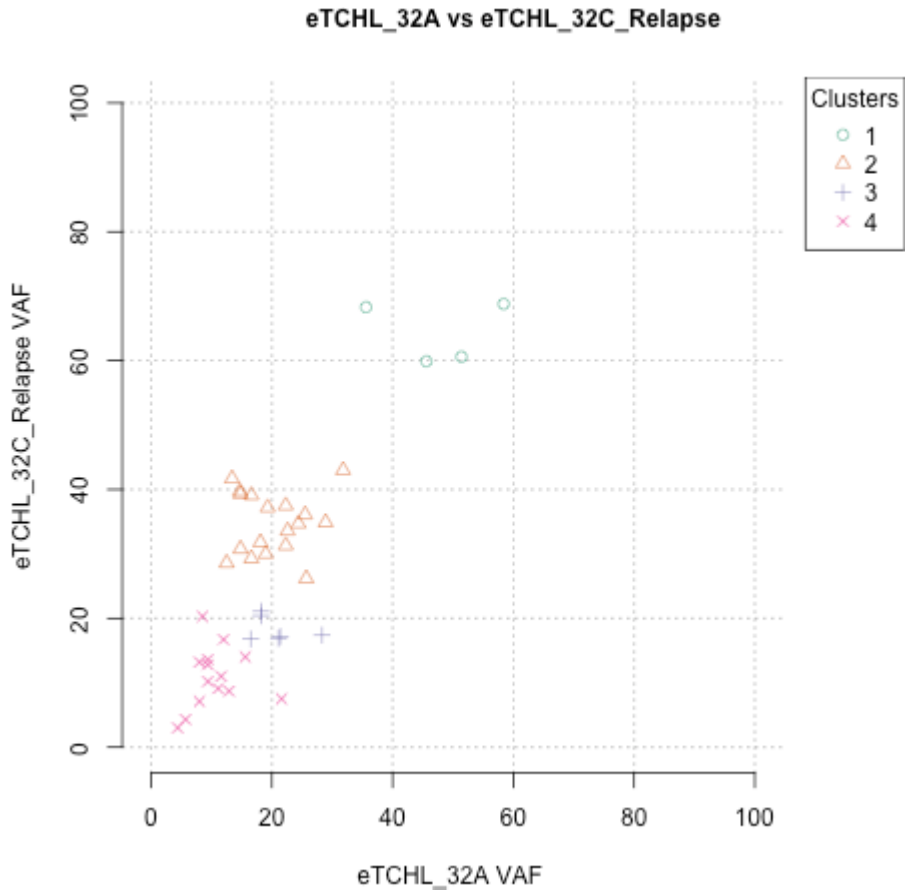


Figure 3.5.8 – Subclonal architecture in the TCHL 32 Pre-treatment and Relapse samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The TCHL 32a Pre-treatment vs Relapse comparison shows 4 clusters, suggesting 4 subclonal populations shared between the samples. Clusters 3 and 4 (purple crosses, red X's)show similar VAFs in both samples, suggesting that they did not change in size during the therapy or relapse process. Clusters 1 and 2 (green circles and orange triangles), however, show higher VAFs in the relapse sample, suggesting that these subclones grew in size over the course of the therapy and relapse process. Mutations in these subclones may have been triggered by the relapse.

3.5.6 TCHL 39 sample comparisons

TCHL 39 Pre vs Post

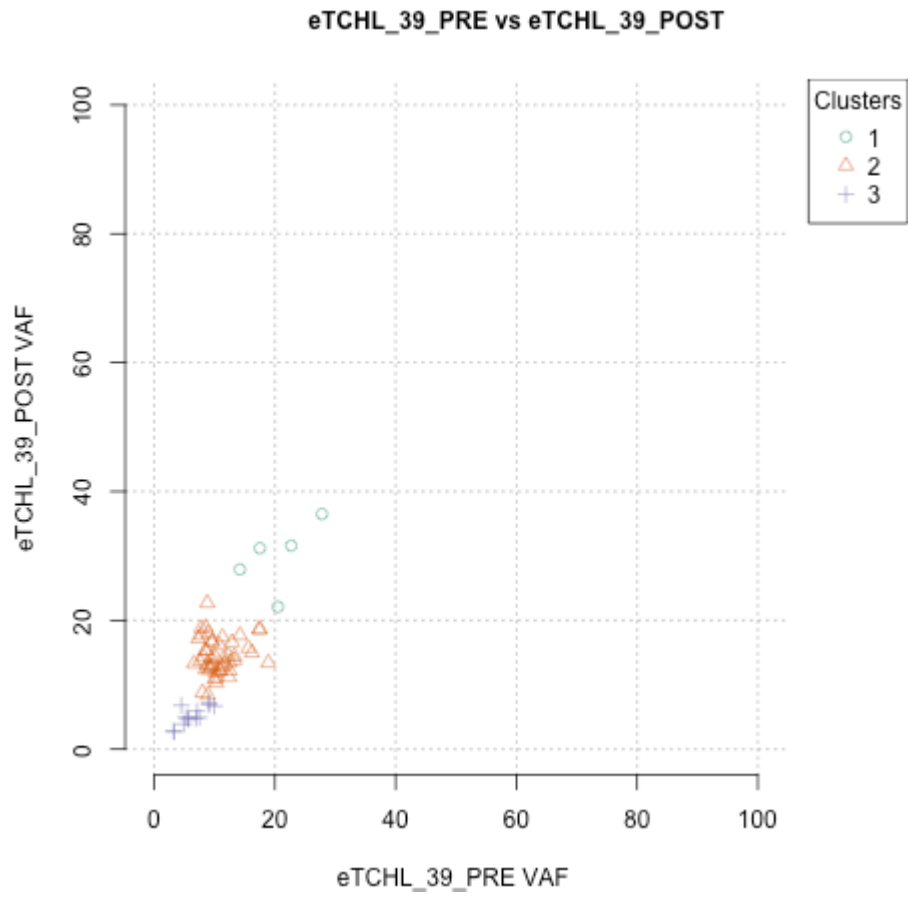


Figure 3.5.9 – Subclonal architecture in the TCHL 39 Pre and Post treatment samples. The graph shows the Variant allele frequencies (VAF) of mutations shared by the two samples, grouped into clusters of mutations predicted to originate from the same subclone.

The TCHL 39 Pre-treatment vs Post Treatment comparison shows 3 clusters. Cluster 1 and 2 (purple crosses and orange triangles) show similar VAFs in both samples, suggesting that they are not very affected by the therapy process.

Cluster 1 (green circles) shows higher average VAFs in the Post treatment sample. Taken with the fact that the Post treatment sample shows a higher number of SNVs and indels than the Pre-treatment sample, this suggests that the subclone that the green circles correspond to was expanding during the therapy process, and acquiring new mutations as it grew, leaving the Post

treatment sample with a higher mutational burden than the Pre-treatment sample.

SciClone was unable to find any clusters in the remaining pairs of TCHL 39 samples (Pre vs Relapse, Post vs Relapse)

3.6 Relapse sample analysis

The only notable difference between the Relapse samples and the other samples in the cohort is the increased presence of *RAD50* mutations in the relapse samples compared to the non-Relapse samples. For this reason, a fisher exact test is performed below to see whether the association of *RAD50* with the relapse samples is statistically significant. This test is done for *RAD50* only because no other gene has a notably increased/decreased frequency of being mutated in the relapse samples compared to the non-relapse samples.

Table 3.6.1 – Contingency table, on which a Fisher Exact Test is performed to test the statistical significance of the association between RAD50 and Relapse samples

RAD50	Samples with gene mutated	Samples without gene mutated
Relapse samples	2 (66.6%)	1 (33.3%)
Other samples	3	27

P-value of Fisher's exact test is 0.05315. The result is not significant under a condition of $p < 0.05$ for significance. The p-value is very close to 0.05 however, suggesting that a larger scale study might uncover a genuinely statistically significant association between RAD50 mutation status and the relapse status of a tumour after therapy.

4. Discussion

4.1 General comment/Overall landscape

As mentioned in the Introduction section, targeted therapy for *HER2* positive breast cancer has seen some great successes but frustratingly, tumours do not always respond to therapy. By analysing which driver mutations and mutational signature patterns are associated with response to therapy, we can hopefully build a model of the genetic landscape of the type of tumour most likely to show a clear response to known therapies. This will aid in selecting which therapies to give to a patient once the genotype of their tumour is identified, and will provide a framework around which to design new therapies for tumours that are unresponsive to the current selection of available therapies.

Patient	Treatment	Response category
TCHL 3	TCHL	Non-responder
TCHL 6	TCH	Non-responder
TCHL 12	TCHL	Non-responder
TCHL 29	TCHL	Non-responder
TCHL 32	TCH	Responded to therapy initially. Later showed relapse in the brain
TCHL 39	TCH	Non-responder
TCHL 45	TCH	Responded to initial therapy

Table 4.1 – Treatment category and responder status for all of the deep sequencing samples.

In terms of small scale somatic mutations, the samples show a mean average of 784 SNVs, 413 indels, and 40 sequence alterations. In terms of CNVs, they show a mean average of 98 amplifications and 57 deletions. Table 4.1, above, shows the Treatment category and responder status for the deep sequencing

samples in order to put the following results in context: The TCHL 3, 6, and 29 pre-treatment samples show more mutations than their post treatment counterparts, while the TCHL 39 pre-treatment sample shows fewer mutations than the post treatment counterpart. The TCHL 12 post-treatment sample shows slightly more SNVs and indels than the Pre-treatment sample but far fewer CNVs. As discussed in the Results section, the cases where the number of mutations decreases during treatment likely reflects therapy successfully killing highly mutated cells in the tumour, while the cases where the number of mutations increases over the course of therapy likely represents highly mutated subclone(s) carrying mutation(s) conferring resistance to the therapy proliferating during the course of therapy.

The Non-Responder samples showed a higher average number of mutations (both median and mean average) than the Responder samples. However, as noted in the results section, this may be influenced by there being more Higher depth sequencing samples in the Non-Responder cohort than the Responder cohort.

4.2 Mutational signatures

Based on the current literature, we would expect signatures 1, 5, 8, 3,2 and 13 to feature most prominently in a breast cancer cohort (145). All of these signatures feature prominently in this cohort, implying that the mutational signatures found in a cohort of breast cancer samples taken from Irish women match those found in studies in other countries. This implies that roughly the same mutational processes drive breast cancer in the Irish population as in other populations. This in turn implies that any alleles more prevalent or less prevalent in the Irish population than in the other populations have little impact on the processes that are likely to drive breast cancer in an Irish population compared to other populations, and that the Irish population of breast cancer patients responds to therapy in the same way as other patients around the world. We cannot find a study in the literature that specifically compares the Irish population of breast cancer patients to the world population in terms of the mutational processes that drive HER2 positive breast cancer, or in terms of how these mutational processes evolve during treatment in breast cancer patients. Therefore, we consider this a novel finding of this project. We consider it reasonable to compare the Irish population under this treatment regime to other breast cancer patients around the world due to the fact that most breast cancer patients around the world will have a similar treatment regime. This is evidenced by the fact that carboplatin, docetaxel, and trastuzumab are all on the World Health Organization (WHO) list of essential medicines for a country to have (list is available at this URL: https://www.who.int/medicines/publications/essentialmedicines/en/).

The signatures observed in this cohort rarely show an environmental influence, with the exception of a small number of samples carrying an aflatoxin associated signature (e.g. TCHL 11, 61, and 45). Most of the signatures observed are aging based, APOBEC based, or DNA damage based. This suggests that the mutational processes underlying breast cancer development in this cohort were a combination of mutations caused naturally by the aging process in the cells of the patients in the lead up to tumourigenesis, and a

dysregulation of enzymes that interact with DNA in those same cells. The DNA damage based signatures (mostly signatures 3 and 6) are more prevalent in the non-responder cohort than in the responder cohort, suggesting that dysfunction in the DNA damage repair enzymes in a cell could affect patient response to therapy. This suggests that checking patient genomes for signs of DNA damage repair dysfunction (e.g. the presence of signatures 3 and 6) could be a useful tool for predicting ahead of time which patients are likely to be non-responders and avoiding damaging the health of a patient with a treatment that is unlikely to help them.

The other signature type that is more prevalent in the non-responders than the responders is the Aging based signature, suggesting that being older is associated with a worse chance of responding to therapy. Ageing is a known risk factor for breast cancer – see Figure 4.1, which shows a clear increase in both breast cancer incidence and mortality the older a woman is (146).

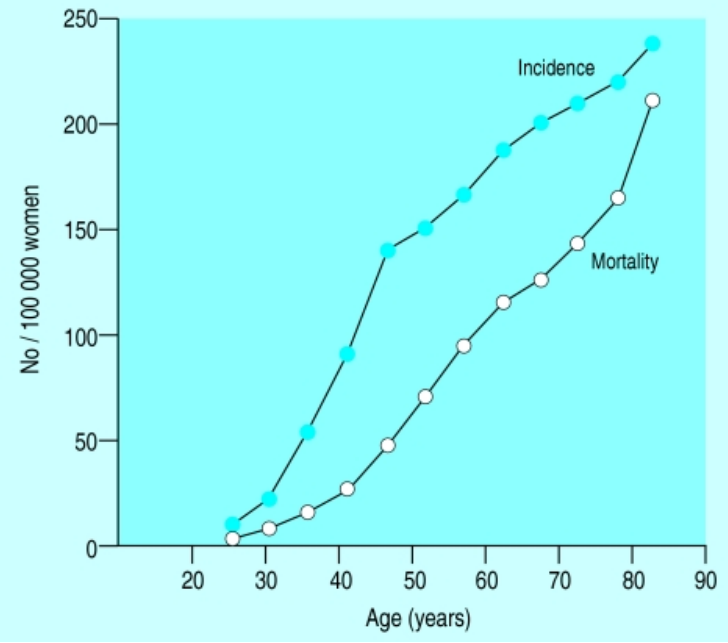


Figure 4.1 – Risk of incidence of breast cancer and mortality from breast cancer in women. Taken from (146)

Age is also known to be a prognostic factor in breast cancer – a prognostic factor is a factor that can be used to estimate the probability of disease recovery, or relapse. In the case of breast cancer overall, some studies have shown that younger age is actually a negative prognostic factor. This means

that in those studies, younger patients were shown to be more likely to die from the disease than older patients (younger being defined as being <40 years old at diagnosis) (147) (148) (149), though it is important to note that some other studies have found that, despite the distinct histopathological features of tumours in younger woman, there was no statistically significant difference in Overall Survival (OS) between younger and older patients in those studies (150). A potential resolution to this apparent contradiction is found in a 2015 paper entitled “The prognostic impact of age in different molecular subtypes of breast cancer”, published in Breast Cancer Treatment and Research, in which the authors stratified by breast cancer subtype before testing for an association between OS and age. This study showed that the association between younger age and poorer OS differs by breast cancer subtype – the association is strongest for triple negative breast cancer, and is absent for the HER2 positive subtype (151). Since the cohort under examination is all of the HER2 positive subtype, the discovery of an association between the ageing signature and non-response to therapy is not in contradiction with the existing literature. The final note to make about the association between age and breast cancer is that later age of menopause starting is associated with an increased risk of breast cancer (152).

Age has been shown to have an effect on probability of responding to therapy in thyroid cancer (153), so these results, taken together with the facts explained above about the relationship between age and HER2 positive breast cancer, suggests that the same effects may apply in HER2 positive breast cancer as well. This result also suggests a mechanism by which older age would increase the chances of non-response to therapy: older cells will have more mutations caused by the ageing signature. The more mutations present, the higher the probability that one of these mutations will be a mutation that confers protection to a tumour that causes it to not respond to therapy.

One other major signature shows up across the cohort - signature 5, a signature of unknown aetiology. As noted in the results section, this signature appears in many cancer types and may be responsible for the relapse observed

in patient 32. The prevalence of signature 5 in this cohort emphasises the importance of learning what causes this signature, so that in future whatever process is causing this signature can be avoided and therapy can be designed to reduce or reverse the impact that this process has on the genomes of somatic cells.

4.3 Driver gene analysis - SNVs and indels

Apart from *TP53* and *PIK3CA*, there are few driver genes that show predicted or known driver SNVs or indels in multiple different samples in the cohort. This shows that different tumours within the cohort, despite sharing a tissue type, utilise different biological pathways to achieve the uncontrollable cell division characteristic of cancer. Many of the samples show fewer than 6 known or predicted driver SNVs or indels, and some show none at all, suggesting that either driver SNVs and indels have been missed by the CGI algorithm, or that they were filtered in the variant calling stage, or that these tumours were driven by CNV mutations rather than SNVs or indels. We note at this stage, as was noted in the Results section, that even state-of-the-art software and pipelines show a false negative rate around 2-3% (144), and logically this false negative rate should increase at least by at least some amount as more filtration is applied to remove false positives from the dataset (unless the filtration is so accurate that it only removes false positives and does not remove any false negatives, which is highly improbable).

In this project, we filtered stringently by having several filtration steps to ensure that our final callset had as few false positives as possible. However, in doing so, we may have removed a number of genuine mutations, and this could explain the low number of known or predicted driver mutations observed in some of the samples. An alternative possibility is that these tumours may have been driven by epigenetic changes rather than by genetic changes (see Intro section 1.1 for a description of what “epigenetic changes are”). It is known from the literature that epigenetic modifications that modify the rate of gene expression can drive cancer just as genetic changes can (154). Unfortunately there was no epigenomic data available for these patients, as this study was only designed to examine genomic changes. This suggests that analyzing epigenomic changes as genomic changes may be worthwhile in future studies. However obtaining this epigenomic data will of course make the study more expensive – the ideal of using more data to make our conclusions as scientifically accurate as possible must always be balanced with feasibility.

In contrast to the samples showing no driver mutations, some samples, such as the TCHL 29 Pre-treatment sample and the TCHL 39 relapse sample, show a very high number of predicted or known driver mutations. This emphasises the diversity in biology seen across the cohort.

Since most of the predicted driver genes show predicted or known driver mutations in only a single sample across the cohort, it is not surprising that many predicted driver genes show predicted driver mutations in only one of the heatmaps in section 3.4, where the samples are divided into responders and non responders. It is also notable when looking at the Pre-treatment sample driver SNV and indel heatmap in section 3.3 that several of the genes that show predicted or known driver mutations in some part of the cohort do not show any predicted or known driver mutations in any other samples in the cohort. This demonstrates the distinct biology of the different samples in the cohort.

The Fisher's exact tests performed in the Results section show that none of the genes show a statistically significant association with the Responder or non-Responder section. However, as noted in the results section, this this may be due to the small sample size, and the results do suggest some genes as targets for more in depth study.

4.4 SciClone info

Comparing the SciClone data for the samples across the cohort shows a wide diversity of subclonal architectures. For the patients with samples from multiple timepoints, the results section explains how the subclonal architecture revealed by the SciClone data, when combined with the mutational signature data and driver gene data for that sample, can be used to hypothesise the evolutionary path that that tumour took through treatment and why, in cases where the tumour ultimately relapsed, the relapse occurred.

Based on the table in section 3.3, the responder cohort shows a mean average of 1.75 subclones per sample, while the non-responder cohort shows a mean average of 2.2105 subclones per sample. As noted in that section, the association does not reach statistical significance in this cohort, but this result does suggest this as a target for further research in future. After all, a tested potential association not reaching statistical signifigance does not mean that no association is non-existent: it simply means that it is plausible that the difference seen between the two populations is caused by chance rather than by a genuine biological phenomenon.

4.5 Final conclusions

The overall picture painted by examining the cohort as a whole is that the tumours in the cohort often show a biology distinct from each other - many samples show predicted driver mutations in genes that do not show predicted driver mutations in any other sample in the cohort, and there are few consistent trends in the mutational signatures or subclonal architectures of the tumours in the cohort overall. The mutational signature patterns observed are consistent with the predictions in the literature for what the mutational signatures of a breast cancer cohort would be expected to look like.

The results section of this thesis shows how knowing the subclonal architecture of the samples gives us an insight into the evolutionary history of a tumour and allows us to confirm or reject hypotheses formed about why the tumour developed the way it did. The results also suggest that more complex subclonal architecture may be associated with non-response to therapy, though a larger scale study would be needed to prove this in a statistically significant manner.

The predicted and known driver SNVs and indels in each sample tend to be in genes that do not show predicted driver SNVs and indels in many other samples in the cohort. Often, a mutation in a predicted driver gene appears only once across the entire cohort. However, TP53 and PIK3CA mutations do appear frequently. The presence of a small number of very commonly mutated genes and a much larger number of infrequently mutated genes across a cohort shows that the genomic “landscape” of this cohort, and breast cancer cohorts in general is similar to cohorts of other cancer types (155).The small group of frequently mutated genes are called “mountains” and the larger group of infrequently mutated genes are referred to as “hills”. TP53 and PIK3CA are known in the literature to be “mountain” genes in breast cancer (156).

None of the individual driver genes examined showed a statistically significant association with either the responder or non-responder cohorts. However, SNVs/indels in *RAD50, ARID1B, DHX9, IZF3, TNPO2, UBR5* and *TAOK1* did show an altered frequency between the two cohorts, suggesting that the

mutation status of these genes may have a genuine association with response status of a tumour, which could be uncovered by a larger scale study with more samples.

The results did hint at a connection between RAD50 and relapse, though the association did not reach statistical signifigance (see Results section 3.6). *RAD50* mutations are known to be associated with breast cancer, and also with genomic instability (157). Therefore, further study is required to uncover firstly, whether *RAD50* mutations are genuinely more common in relapse samples than non-relapse samples at a statistically significant level, and secondly whether this association is actually caused by the *RAD50* mutations or whether it is an artefact of the fact that greater genomic instability is associated with higher propensity for a tumour to relapse in general (158).

Overall, this project has highlighted several promising areas for future research. However, to reach statistically significant findings that can confidently be used for the design of future therapies, larger scale cohorts with a greater number of samples will be needed

5 Bibliography

1. GBD 2015 Mortality and Causes of Death Collaborators G 2015 M and C of D. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet (London, England) [Internet]. 2016 Oct 8 [cited 2019 Jul 30];388(10053):1459–544. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27733281

2. Wishart DS. Is Cancer a Genetic Disease or a Metabolic Disease? EBioMedicine [Internet]. 2015 Jun [cited 2019 Jul 30];2(6):478–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26288805

3. You JS, Jones PA. Cancer genetics and epigenetics: two sides of the same coin? Cancer Cell [Internet]. 2012 Jul 10 [cited 2019 Jul 30];22(1):9–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22789535

4. Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM. Introduction to genetic analysis, 7th edition. W. H. Freeman; 2000.

5. Ahmad A. Introduction to Cancer Metastasis [Internet]. [cited 2019 Jul 31]. 416 p. Available from: https://www.sciencedirect.com/book/9780128040034/introduction-to-cancer-metastasis#book-description

6. Seyfried TN, Huysentruyt LC. On the origin of cancer metastasis. Crit Rev Oncog [Internet]. 2013 [cited 2017 Aug 11];18(1–2):43–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23237552

7. Milanese J-S, Wang E. Germline mutations and their clinical applications in cancer. Breast Cancer Manag [Internet]. 2019 Mar 20 [cited 2019 Jul 31];8(1):BMT23. Available from: https://www.futuremedicine.com/doi/10.2217/bmt-2019-0005

8. Tomlinson IPM, Novelli MR, Bodmer WF. The mutation rate and cancer (tumorigenesisevolutionselectionmutator phenotypegenomic instability) [Internet]. Vol. 93, Medical Sciences. 1996 [cited 2019 Jul 28]. Available from: https://www.pnas.org/content/pnas/93/25/14800.full.pdf

9. Komarova NL, Wodarz D. Evolutionary Dynamics of Mutator Phenotypes in Cancer: Implications for Chemotherapy. CANCER Res [Internet]. 2003 [cited 2017 Aug 13];63:6635–42. Available from: http://cancerres.aacrjournals.org/content/canres/63/20/6635.full.pdf

10. Tomasetti C, Voeglstein B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science (80-) [Internet]. 2015;347(6217):78–81. Available from: http://science.sciencemag.org/content/347/6217/78

11. Costamagna D, Berardi E, Ceccarelli G, Sampaolesi M. Adult Stem Cells and Skeletal Muscle Regeneration. Curr Gene Ther [Internet]. 2015 [cited 2019 Aug 1];15(4):348–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26122100

12. Clevers H. STEM CELLS. What is an adult stem cell? Science. 2015 Dec 11;350(6266):1319–20.

13. McFarland CD, Mirny LA, Korolev KS. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. Proc Natl Acad Sci U S A [Internet]. 2014 Oct 21 [cited 2017 Aug 31];111(42):15138–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25277973

14. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature [Internet]. 2009 Apr 9 [cited 2017 Aug 31];458(7239):719–24. Available from: http://www.nature.com/doifinder/10.1038/nature07943

15. Mills GB. An emerging toolkit for targeted cancer therapies. [cited 2019 Jul 31]; Available from: www.genome.org

16. Almendro V, Polyak K. INtra-tumour heterogeneity: a looking glass for cancer? 2012 [cited 2019 Aug 4]; Available from:

https://www.researchgate.net/publication/224708018

17. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. bioRxiv [Internet]. 2018 May 5 [cited 2019 Aug 4];312041. Available from: https://www.biorxiv.org/content/10.1101/312041v1

18. Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acad Sci U S A [Internet]. 2010 Oct 26 [cited 2018 Aug 23];107(43):18545–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20876136

19. Swanton C. Intratumor heterogeneity: evolution through space and time. Cancer Res [Internet]. 2012 Oct 1 [cited 2019 Aug 1];72(19):4875–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23002210

20. Gay L, Baker A-M, Graham TA. Tumour Cell Heterogeneity. F1000Research [Internet]. 2016 [cited 2019 Aug 5];5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26973786

21. Robertson-Tessi M, Gillies RJ, Gatenby RA, Anderson ARA. Impact of metabolic heterogeneity on tumor growth, invasion, and treatment outcomes. Cancer Res [Internet]. 2015 Apr 15 [cited 2019 Aug 5];75(8):1567–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25878146

22. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. Nature [Internet]. 2013 Sep 19;501(7467):338–45. Available from: http://dx.doi.org/10.1038/nature12625

23. Thapar A, Cooper M. Copy number variation: what is it and what has it told us about child psychiatric disorders? J Am Acad Child Adolesc Psychiatry [Internet]. 2013 Aug [cited 2019 Aug 5];52(8):772–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23880486

24. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V, et al. Signatures of mutational processes in human cancer. Nature

[Internet]. 2013 Aug 22;500(7463):415–21. Available from: http://dx.doi.org/10.1038/nature12477

25. McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci Transl Med [Internet]. 2015 [cited 2017 Aug 13];7(283). Available from: http://stm.sciencemag.org/content/7/283/283ra54

26. Bashashati A, Ha G, Tone A, Ding J, Prentice LM, Roth A, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. J Pathol [Internet]. 2013 Sep 1 [cited 2017 Aug 13];231(1):21–34. Available from: http://doi.wiley.com/10.1002/path.4230

27. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. Cell [Internet]. 2013 Feb [cited 2017 Aug 13];152(4):714–26. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867413000718

28. Caswell DR, Swanton C. The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. BMC Med [Internet]. 2017 [cited 2019 Aug 5];15(1):133. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28716075

29. Stanta G, Bonin S. Overview on Clinical Relevance of Intra-Tumor Heterogeneity. Front Med [Internet]. 2018 Apr 6 [cited 2019 Aug 5];5:85. Available from: http://journal.frontiersin.org/article/10.3389/fmed.2018.00085/full

30. Seyfried TN, Huysentruyt LC. On the origin of cancer metastasis. Crit Rev Oncog [Internet]. 2013 [cited 2018 Aug 23];18(1–2):43–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23237552

31. Pennisi E. Is cancer a breakdown of multicellularity? Science [Internet]. 2018 Jun 29 [cited 2019 Aug 11];360(6396):1391. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29954963

32. Nowell PC. The clonal evolution of tumor cell populations. Science [Internet]. 1976 Oct 1 [cited 2018 Aug 24];194(4260):23–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/959840

33. Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. Trends in Cancer [Internet]. 2016 Jan [cited 2018 Aug 23];2(1):49–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26949746

34. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell [Internet]. 2011 Mar 4 [cited 2019 Aug 11];144(5):646–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21376230

35. Yates LR, Campbell PJ. Evolution of the cancer genome. Nat Rev Genet [Internet]. 2012 Nov [cited 2018 Aug 23];13(11):795–806. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23044827

36. Valdes-Mora F, Handler K, Law AMK, Salomon R, Oakes SR, Ormandy CJ, et al. Single-Cell Transcriptomics in Cancer Immunobiology: The Future of Precision Oncology. Front Immunol [Internet]. 2018 [cited 2019 Aug 11];9:2582. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30483257

37. Jones PA, Baylin SB. The epigenomics of cancer. Cell [Internet]. 2007 Feb 23 [cited 2019 Aug 11];128(4):683–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17320506

38. Martincorena I, Raine KM, Davies H, Stratton MR, Campbell PJ. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell [Internet]. 2017 [cited 2018 Aug 23];171:1029-1041.e21. Available from: https://doi.org/10.1016/j.cell.2017.09.042

39. Lacina L, Čoma M, Dvořánková B, Kodet O, Melegová N, Gál P, et al. Evolution of Cancer Progression in the Context of Darwinism. Anticancer Res [Internet]. 2019 Jan 1 [cited 2019 Aug 11];39(1):1–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30591435

40. Klein CA. Parallel progression of primary tumours and metastases. Nat Rev

Cancer [Internet]. 2009 Apr 1 [cited 2018 Aug 23];9(4):302–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19308069

41. Lathia JD, Heddleston JM, Venere M, Rich JN. Deadly Teamwork: Neural Cancer Stem Cells and the Tumor Microenvironment. Cell Stem Cell [Internet]. 2011 May 6 [cited 2018 Aug 23];8(5):482–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21549324

42. Turajlic S, Swanton C. Metastasis as an evolutionary process. Science (80-) [Internet]. 2016 Apr 8 [cited 2018 Aug 23];352(6282):169–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27124450

43. Powell E, Piwnica-Worms D, Piwnica-Worms H. Contribution of p53 to Metastasis. Cancer Discov [Internet]. 2014 Apr 1 [cited 2018 Aug 23];4(4):405–14. Available from: http://cancerdiscovery.aacrjournals.org/cgi/doi/10.1158/2159-8290.CD-13-0136

44. Merino D, Malkin D. p53 and Hereditary Cancer. In: Sub-cellular biochemistry [Internet]. 2014 [cited 2019 Aug 11]. p. 1–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25201186

45. Hanel W, Moll UM. Links Between Mutant p53 and Genomic Instability. J Cell Biochem [Internet]. 2012 Feb [cited 2018 Aug 23];113(2):433. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22006292

46. Rausch T, Jones DTW, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, et al. Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. Cell [Internet]. 2012 Jan [cited 2018 Aug 23];148(1–2):59–71. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867411015169

47. Luijten MNH, Lee JXT, Crasta KC. Mutational game changer: Chromothripsis and its emerging relevance to cancer. Mutat Res Mutat Res [Internet]. 2018 Jul 1 [cited 2019 Aug 11];777:29–51. Available from: https://www.sciencedirect.com/science/article/pii/S1383574218300322

48. Bloomfield M, Duesberg P. Inherent variability of cancer-specific aneuploidy generates metastases. Mol Cytogenet [Internet]. 2016 [cited 2019 Aug 11];9:90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28018487

49. Nouhaud F-X, Blanchard F, Sesboue R, Flaman J-M, Sabourin J-C, Pfister C, et al. Clinical Relevance of Gene Copy Number Variation in Metastatic Clear Cell Renal Cell Carcinoma. Clin Genitourin Cancer [Internet]. 2018 Aug [cited 2019 Aug 11];16(4):e795–805. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29548613

50. Bakhoum SF, Ngo B, Laughney AM, Cavallo J-A, Murphy CJ, Ly P, et al. Chromosomal instability drives metastasis through a cytosolic DNA response. Nature [Internet]. 2018 Jan 17 [cited 2019 Aug 11];553(7689):467–72. Available from: http://www.nature.com/articles/nature25432

51. Tijhuis AE, Johnson SC, McClelland SE. The emerging links between chromosomal instability (CIN), metastasis, inflammation and tumour immunity. Mol Cytogenet [Internet]. 2019 Dec 14 [cited 2019 Aug 11];12(1):17. Available from: https://molecularcytogenetics.biomedcentral.com/articles/10.1186/s13039-019-0429-1

52. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated Evolution of Prostate Cancer Genomes. Cell [Internet]. 2013 Apr 25 [cited 2019 Aug 11];153(3):666–77. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867413003437

53. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell [Internet]. 2011 Jan 7 [cited 2019 Aug 11];144(1):27–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21215367

54. Krøigård AB, Larsen MJ, Lænkholm A-V, Knoop AS, Jensen JD, Bak M, et al. Clonal expansion and linear genome evolution through breast cancer progression from pre-invasive stages to asynchronous metastasis. Oncotarget

[Internet]. 2015 Mar 20 [cited 2019 Aug 5];6(8):5634–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25730902

55. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature [Internet]. 2009 Oct 8 [cited 2018 Aug 23];461(7265):809–13. Available from: http://www.nature.com/doifinder/10.1038/nature08489

56. 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 [Internet]. 2010 Apr 15 [cited 2018 Aug 23];464(7291):999–1005. Available from: http://www.nature.com/doifinder/10.1038/nature08989

57. 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 [Internet]. 2015 Jul;21(7):751–9. Available from: http://dx.doi.org/10.1038/nm.3886

58. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature [Internet]. 2012 Jan 11 [cited 2018 Aug 23];481(7382):506–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22237025

59. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. Cell [Internet]. 2013 Feb [cited 2017 Aug 12];152(4):714–26. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867413000718

60. Schwarz RF, Ng CKY, Cooke SL, Newman S, Temple J, Piskorz AM, et al. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. Kemp C, editor. PLOS Med [Internet]. 2015 Feb 24 [cited 2018 Aug 23];12(2):e1001789. Available from: http://dx.plos.org/10.1371/journal.pmed.1001789

61. Mroz EA, Rocco JW. MATH, a novel measure of intratumor genetic

heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. Oral Oncol [Internet]. 2013 Mar [cited 2018 Aug 23];49(3):211–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23079694

62. Burrell RA, Swanton C. Tumour heterogeneity and the evolution of polyclonal drug resistance. Mol Oncol [Internet]. 2014 Sep 12 [cited 2018 Aug 23];8(6):1095–111. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25087573

63. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer [Internet]. 2018 Dec 6 [cited 2019 Aug 5];ijc.31937. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.31937

64. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol [Internet]. 2011 Feb [cited 2019 Aug 11];5(1):5–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21147047

65. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A [Internet]. 2001 Sep 11 [cited 2019 Aug 18];98(19):10869–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11553815

66. Prat A, Cheang MCU, Galván P, Nuciforo P, Paré L, Adamo B, et al. Prognostic Value of Intrinsic Subtypes in Hormone Receptor–Positive Metastatic Breast Cancer Treated With Letrozole With or Without Lapatinib. JAMA Oncol [Internet]. 2016 Oct 1 [cited 2019 Aug 18];2(10):1287. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27281556

67. Prat A, Pineda E, Adamo B, Galván P, Fernández A, Gaba L, et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. The Breast [Internet]. 2015 Nov [cited 2019 Aug 18];24:S26–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26253814

68. Sandhu R, Parker JS, Jones WD, Livasy CA, Coleman WB. Microarray-Based Gene Expression Profiling for Molecular Classification of Breast Cancer and Identification of New Targets for Therapy. Lab Med [Internet]. 2010 Jun 1 [cited 2018 Aug 24];41(6):364–72. Available from: https://academic.oup.com/labmed/article-lookup/doi/10.1309/LMLIK0VIE3CJK0WD

69. Hashmi AA, Aijaz S, Khan SM, Mahboob R, Irfan M, Zafar NI, et al. Prognostic parameters of luminal A and luminal B intrinsic breast cancer subtypes of Pakistani patients. World J Surg Oncol [Internet]. 2018 Dec 2 [cited 2019 Aug 18];16(1):1. Available from: https://wjso.biomedcentral.com/articles/10.1186/s12957-017-1299-9

70. Toft DJ, Cryns VL. Minireview: Basal-like breast cancer: from molecular profiles to targeted therapies. Mol Endocrinol [Internet]. 2011 Feb [cited 2019 Aug 18];25(2):199–211. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20861225

71. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res [Internet]. 2015 [cited 2019 Aug 18];5(10):2929–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26693050

72. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. Clin Cancer Res [Internet]. 2007 Aug 1 [cited 2019 Aug 18];13(15):4429–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17671126

73. Fragomeni SM, Sciallis A, Jeruss JS. Molecular Subtypes and Local-Regional Control of Breast Cancer. Surg Oncol Clin N Am [Internet]. 2018 [cited 2019 Aug 18];27(1):95–120. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29132568

74. Vallejos CS, Gómez HL, Cruz WR, Pinto JA, Dyer RR, Velarde R, et al. Breast Cancer Classification According to Immunohistochemistry Markers: Subtypes and Association With Clinicopathologic Variables in a Peruvian Hospital

Database. Clin Breast Cancer [Internet]. 2010 Aug 1 [cited 2018 Aug 24];10(4):294–300. Available from: https://www.sciencedirect.com/science/article/pii/S1526820911700416?via%3Dihub

75. Shabbir A, Qureshi MA, Mirza T, Khalid A Bin. Human epidermal growth factor (Her-2) in gastric and colorectal adenocarcinoma. J Pak Med Assoc [Internet]. 2017 Jul [cited 2019 Aug 12];67(7):1085–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28770892

76. Gutierrez C, Schiff R. HER2: biology, detection, and clinical implications. Arch Pathol Lab Med [Internet]. 2011 Jan [cited 2019 Aug 12];135(1):55–62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21204711

77. Roskoski R. The ErbB/HER family of protein-tyrosine kinases and cancer. Pharmacol Res [Internet]. 2014 Jan [cited 2019 Aug 12];79:34–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24269963

78. Olayioye MA. Update on HER-2 as a target for cancer therapy: intracellular signaling pathways of ErbB2/HER-2 and family members. Breast Cancer Res [Internet]. 2001 [cited 2018 Aug 24];3(6):385–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11737890

79. Mitri Z, Constantine T, O'Regan R. The HER2 Receptor in Breast Cancer: Pathophysiology, Clinical Use, and New Advances in Therapy. Chemother Res Pract [Internet]. 2012 [cited 2019 Aug 18];2012:743193. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23320171

80. Cronin KA, Harlan LC, Dodd KW, Abrams JS, Ballard-Barbash R. Population-based estimate of the prevalence of HER-2 positive breast cancer tumors for early stage patients in the US. Cancer Invest [Internet]. 2010 Nov [cited 2018 Aug 24];28(9):963–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20690807

81. Zhang L, Huang Y, Feng Z, Wang X, Li H, Song F, et al. Comparison of breast cancer risk factors among molecular subtypes: A case-only study. Cancer Med

[Internet]. 2019 Apr 14 [cited 2019 Aug 18];8(4):1882–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30761775

82. Daniels B, Lord SJ, Kiely BE, Houssami N, Haywood P, Lu CY, et al. Use and outcomes of targeted therapies in early and metastatic HER2-positive breast cancer in Australia: protocol detailing observations in a whole of population cohort. BMJ Open [Internet]. 2017 Jan 24 [cited 2018 Aug 24];7(1):e014439. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28119394

83. Loibl S, von Minckwitz G, Schneeweiss A, Paepke S, Lehmann A, Rezai M, et al. *PIK3CA* Mutations Are Associated With Lower Rates of Pathologic Complete Response to Anti–Human Epidermal Growth Factor Receptor 2 (HER2) Therapy in Primary HER2-Overexpressing Breast Cancer. J Clin Oncol [Internet]. 2014 Oct 10 [cited 2018 Aug 24];32(29):3212–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25199759

84. Huszno J, Nowara E. Risk factors for disease progression in HER2-positive breast cancer patients based on the location of metastases. Prz menopauzalny = Menopause Rev [Internet]. 2015 Sep [cited 2019 Aug 18];14(3):173–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26528105

85. Joensuu H, Kellokumpu-Lehtinen P-L, Bono P, Alanko T, Kataja V, Asola R, et al. Adjuvant Docetaxel or Vinorelbine with or without Trastuzumab for Breast Cancer. N Engl J Med [Internet]. 2006 Feb 23 [cited 2019 Aug 18];354(8):809–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16495393

86. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE, Davidson NE, et al. Trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer. N Engl J Med [Internet]. 2005 Oct 20 [cited 2019 Aug 18];353(16):1673–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16236738

87. Boekhout AH, Beijnen JH, Schellens JHM. Trastuzumab. Oncologist [Internet]. 2011 [cited 2019 Aug 12];16(6):800–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21632460

88. Vu T, Claret FX. Trastuzumab: updated mechanisms of action and resistance in breast cancer. Front Oncol [Internet]. 2012 [cited 2019 Aug 12];2:62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22720269

89. Triulzi T, Regondi V, De Cecco L, Cappelletti MR, Di Modica M, Paolini B, et al. Early immune modulation by single-agent trastuzumab as a marker of trastuzumab benefit. Br J Cancer [Internet]. 2018 Dec 27 [cited 2019 Aug 12];119(12):1487–94. Available from: http://www.nature.com/articles/s41416-018-0318-0

90. Tsimberidou A-M. TARGETED THERAPY IN CANCER HHS Public Access. Cancer Chemother Pharmacol [Internet]. 2015 [cited 2018 Aug 24];76(6):1113–32. Available from: http://www.sanger.ac.uk/genetics/CGP/cosmic/

91. Montero A, Fossella F, Hortobagyi G, Valero V. Docetaxel for treatment of solid tumours: a systematic review of clinical data. Lancet Oncol [Internet]. 2005 Apr [cited 2019 Aug 12];6(4):229–39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15811618

92. Pienta KJ. Preclinical mechanisms of action of docetaxel and docetaxel combinations in prostate cancer. Semin Oncol [Internet]. 2001 Aug [cited 2019 Aug 12];28(4 Suppl 15):3–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11685722

93. Herbst RS, Khuri FR. Mode of action of docetaxel - a basis for combination with novel anticancer agents. Cancer Treat Rev [Internet]. 2003 Oct 1 [cited 2019 Aug 12];29(5):407–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12972359

94. Cortes JE, Pazdur R. Docetaxel. J Clin Oncol [Internet]. 1995 Oct 21 [cited 2018 Aug 24];13(10):2643–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7595719

95. Pokhriyal R, Hariprasad R, Kumar L, Hariprasad G. Chemotherapy Resistance in Advanced Ovarian Cancer Patients. Biomark Cancer [Internet]. 2019 Jan 5

[cited 2019 Aug 12];11:1179299X1986081. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31308780

96. Wang S, Zimmermann S, Parikh K, Mansfield AS, Adjei AA. Current Diagnosis and Management of Small-Cell Lung Cancer. Mayo Clin Proc [Internet]. 2019 Aug [cited 2019 Aug 12];94(8):1599–622. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31378235

97. Sousa GF de, Wlodarczyk SR, Monteiro G, Sousa GF de, Wlodarczyk SR, Monteiro G. Carboplatin: molecular mechanisms of action associated with chemoresistance. Brazilian J Pharm Sci [Internet]. 2014 Dec [cited 2018 Aug 24];50(4):693–701. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-82502014000400693&lng=en&tlng=en

98. Bukowska B, Gajek A, Marczak A. Two drugs are better than one. A short history of combined therapy of ovarian cancer. Contemp Oncol (Poznan, Poland) [Internet]. 2015 [cited 2019 Aug 12];19(5):350–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26793017

99. MEDINA P, GOODIN S. Lapatinib: A dual inhibitor of human epidermal growth factor receptor tyrosine kinases. Clin Ther [Internet]. 2008 Aug [cited 2019 Aug 13];30(8):1426–47. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18803986

100. Keegan NM, Toomey S, Fay J, Madden SF, Moran B, Milewska M, et al. Effect of TCHL-based therapy on immune cell content in on-treatment, neoadjuvant-treated HER2-positive breast cancer patients. J Clin Oncol [Internet]. 2017 May 20 [cited 2019 Aug 18];35(15_suppl):583–583. Available from: http://ascopubs.org/doi/10.1200/JCO.2017.35.15_suppl.583

101. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed [Internet]. 2013 Dec [cited 2018 Aug 23];98(6):236–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23986538

102. Mardis ER. Next-Generation DNA Sequencing Methods. 2008 [cited 2019 Aug

4]; Available from: www.helicosbio.com

103. Pienaar E, Theron M, Nelson M, Viljoen H. A QUANTITATIVE MODEL OF ERROR ACCUMULATION DURING PCR AMPLIFICATION. Comput Biol Chem [Internet]. 2006 Apr [cited 2019 Aug 4];30(2):102. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16412692

104. Walsh PS, Erlich HA, Higuchi R. Preferential PCR amplification of alleles: mechanisms and solutions. PCR Methods Appl [Internet]. 1992 May [cited 2018 Aug 24];1(4):241–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1477658

105. Shen R, Fan J-B, Campbell D, Chang W, Chen J, Doucet D, et al. High-throughput SNP genotyping on universal bead arrays. Mutat Res Mol Mech Mutagen [Internet]. 2005 Jun 3 [cited 2018 Aug 24];573(1–2):70–82. Available from: https://www.sciencedirect.com/science/article/pii/S0027510705000278?via%3Dihub

106. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature [Internet]. 2008 Nov 6 [cited 2018 Aug 24];456(7218):53–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18987734

107. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res [Internet]. 2010 Sep [cited 2018 Aug 23];20(9):1297–303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20644199

108. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet [Internet]. 2011 May 10 [cited 2018 Aug 24];43(5):491–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21478889

109. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res [Internet]. 2010 Apr [cited 2018 Aug 23];38(6):1767–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20015970

110. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2010;

111. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics [Internet]. 2014 Aug 1 [cited 2018 Aug 23];30(15):2114–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24695404

112. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. 2009 Jul 15 [cited 2018 Aug 23];25(14):1754–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19451168

113. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics [Internet]. 2009 Aug 15 [cited 2018 Aug 23];25(16):2078–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19505943

114. 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. In: Current Protocols in Bioinformatics [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013 [cited 2018 Aug 24]. p. 11.10.1-11.10.33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25431634

115. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res [Internet]. 2009 Jun 1 [cited 2019 Aug 4];19(6):1124–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19420381

116. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics [Internet]. 2011 Aug 1 [cited 2018 Aug 24];27(15):2156–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21653522

117. Shen R, Seshan VE, P.-Y. K, N. R, L. W, C.S. S, et al. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res [Internet]. 2016 Sep 19 [cited 2017 Aug 13];44(16):e131–e131. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw520

118. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016 Dec 6 [cited 2018 Aug 24];17(1):122. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27268795

119. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome Med [Internet]. 2018 Dec 28 [cited 2018 Aug 23];10(1):25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29592813

120. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res [Internet]. 2010 Sep 1 [cited 2018 Aug 23];38(16):e164–e164. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20601685

121. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. Beerenwinkel N, editor. PLoS Comput Biol [Internet]. 2014 Aug 7 [cited 2018 Aug 23];10(8):e1003665. Available from: http://dx.plos.org/10.1371/journal.pcbi.1003665

122. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res [Internet]. 2001 Jan 1 [cited 2018 Aug 23];29(1):308–11. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/11125122

123. Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, Kovar C, et al. A global reference for human genetic variation. Nature [Internet]. 2015 Oct 1 [cited 2018 Aug 23];526(7571):68–74. Available from: http://www.nature.com/articles/nature15393

124. Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, et al. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. In: Current Protocols in Human Genetics [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2016 [cited 2017 Aug 31]. p. 10.11.1-10.11.37. Available from: http://doi.wiley.com/10.1002/cphg.21

125. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell [Internet]. 2012 May 25 [cited 2019 Aug 11];149(5):979–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22608084

126. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep [Internet]. 2013 Jan 31 [cited 2019 Aug 11];3(1):246–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23318258

127. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet [Internet]. 2015 Dec [cited 2018 Aug 23];47(12):1402–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26551669

128. 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 [Internet]. 2016 Dec 22 [cited 2018 Aug 23];17(1):31. Available from: http://genomebiology.com/2016/17/1/31

129. Teng BB, Ochsner S, Zhang Q, Soman K V, Lau PP, Chan L. Mutational analysis of apolipoprotein B mRNA editing enzyme (APOBEC1). structure-

function relationships of RNA editing and dimerization. J Lipid Res [Internet]. 1999 Apr [cited 2018 Aug 23];40(4):623–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10191286

130. Harris RS, Dudley JP. APOBECs and virus restriction. Virology [Internet]. 2015 May [cited 2018 Aug 23];479–480:131–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25818029

131. Olson ME, Harris RS, Harki DA. APOBEC Enzymes as Targets for Virus and Cancer Therapy. Cell Chem Biol [Internet]. 2018 Jan 18 [cited 2018 Aug 23];25(1):36–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29153851

132. Xiao X, Melton DW, Gourley C. Mismatch repair deficiency in ovarian cancer — Molecular characteristics and clinical implications. Gynecol Oncol [Internet]. 2014 Feb [cited 2018 Aug 29];132(2):506–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24333356

133. Fisher RA. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. J R Stat Soc [Internet]. 1922 Jan 1 [cited 2019 Aug 19];85(1):87. Available from: https://www.jstor.org/stable/2340521?origin=crossref

134. Kim TK. T test as a parametric statistic. Korean J Anesthesiol [Internet]. 2015 Dec [cited 2019 Aug 19];68(6):540–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26634076

135. Dahiru T. P - value, a true test of statistical significance? A cautionary note. Ann Ibadan Postgrad Med [Internet]. 2008 Jun [cited 2019 Aug 19];6(1):21–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25161440

136. Meléndez B, Van Campenhout C, Rorive S, Remmelink M, Salmon I, D'Haene N. Methods of measurement for tumor mutational burden in tumor tissue. Transl lung cancer Res [Internet]. 2018 Dec [cited 2019 Aug 19];7(6):661–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30505710

137. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS,

et al. Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. Cell [Internet]. 2013 Feb 14 [cited 2019 Aug 19];152(4):714–26. Available from: https://www.sciencedirect.com/science/article/pii/S0092867413000718

138. Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nat Methods [Internet]. 2016 Oct 29 [cited 2019 Aug 18];13(10):806–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27684579

139. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res [Internet]. 2016 Jan 4 [cited 2019 Aug 18];44(D1):D862–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26582918

140. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol [Internet]. 2017 Nov [cited 2019 Aug 18];1(1):1–16. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28890946

141. Sobhani N, Roviello G, Corona SP, Scaltriti M, Ianza A, Bortul M, et al. The prognostic value of PI3K mutational status in breast cancer: A meta-analysis. J Cell Biochem [Internet]. 2018 [cited 2019 Aug 18];119(6):4287–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29345357

142. Madsen RR, Vanhaesebroeck B, Semple RK. Cancer-Associated PIK3CA Mutations in Overgrowth Disorders. Trends Mol Med [Internet]. 2018 Oct 1 [cited 2019 Aug 18];24(10):856–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30197175

143. Harley AS, Lau CK, Shinbrot E. Abstract 2474: Automated somatic variant classifier to reduce false positives identified by tumor normal variant callers. In: Bioinformatics, Convergence Science, and Systems Biology [Internet]. American Association for Cancer Research; 2019 [cited 2019 Aug 19]. p. 2474–2474. Available from: http://cancerres.aacrjournals.org/lookup/doi/10.1158/1538-7445.AM2019-2474

144. Field MA, Cho V, Andrews TD, Goodnow CC. Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. PLoS One [Internet]. 2015 [cited 2019 Aug 19];10(11):e0143199. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26600436

145. Nik-Zainal S, Morganella S. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. Clin Cancer Res [Internet]. 2017 Jun 1 [cited 2018 Sep 7];23(11):2617–29. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28572256

146. McPherson K, Steel CM, Dixon JM. ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. BMJ [Internet]. 2000 Sep 9 [cited 2019 Aug 17];321(7261):624–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10977847

147. Kheirelseid EA, Boggs JM, Curran C, Glynn RW, Dooley C, Sweeney KJ, et al. Younger age as a prognostic indicator in breast cancer: A cohort study. BMC Cancer [Internet]. 2011 Dec 28 [cited 2019 Aug 17];11(1):383. Available from: http://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-11-383

148. Høst H, Lund E. Age as a prognostic factor in breast cancer. Cancer [Internet]. 1986 Jun 1 [cited 2019 Aug 17];57(11):2217–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3697919

149. Brandt J, Garne J, Tengrup I, Manjer J. Age at diagnosis in relation to survival following breast cancer: a cohort study. World J Surg Oncol [Internet]. 2015 Feb 7 [cited 2019 Aug 17];13(1):33. Available from: http://www.wjso.com/content/13/1/33

150. Clagnan WS, Andrade JM de, Carrara HHA, Tiezzi DG, Reis FJC dos, Marana HRC, et al. [Age as an independent prognostic factor in breast cancer]. Rev Bras Ginecol Obstet [Internet]. 2008 Feb [cited 2019 Aug 17];30(2):67–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19142478

151. Liedtke C, Rody A, Gluz O, Baumann K, Beyer D, Kohls E-B, et al. The

prognostic impact of age in different molecular subtypes of breast cancer. Breast Cancer Res Treat [Internet]. 2015 Aug 21 [cited 2019 Aug 17];152(3):667–73. Available from: http://link.springer.com/10.1007/s10549-015-3491-3

152. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. Lancet Oncol [Internet]. 2012 Nov [cited 2019 Aug 18];13(11):1141–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23084519

153. Shah S, Boucai L. Effect of Age on Response to Therapy and Mortality in Patients With Thyroid Cancer at High Risk of Recurrence. J Clin Endocrinol Metab [Internet]. 2018 Feb 1 [cited 2018 Sep 7];103(2):689–97. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29220531

154. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis [Internet]. 2010 Jan [cited 2019 Aug 17];31(1):27–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19752007

155. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science (80-) [Internet]. 2013 Mar 29 [cited 2017 Aug 30];339(6127):1546–58. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1235122

156. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science [Internet]. 2007 Nov 16 [cited 2018 Aug 23];318(5853):1108–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17932254

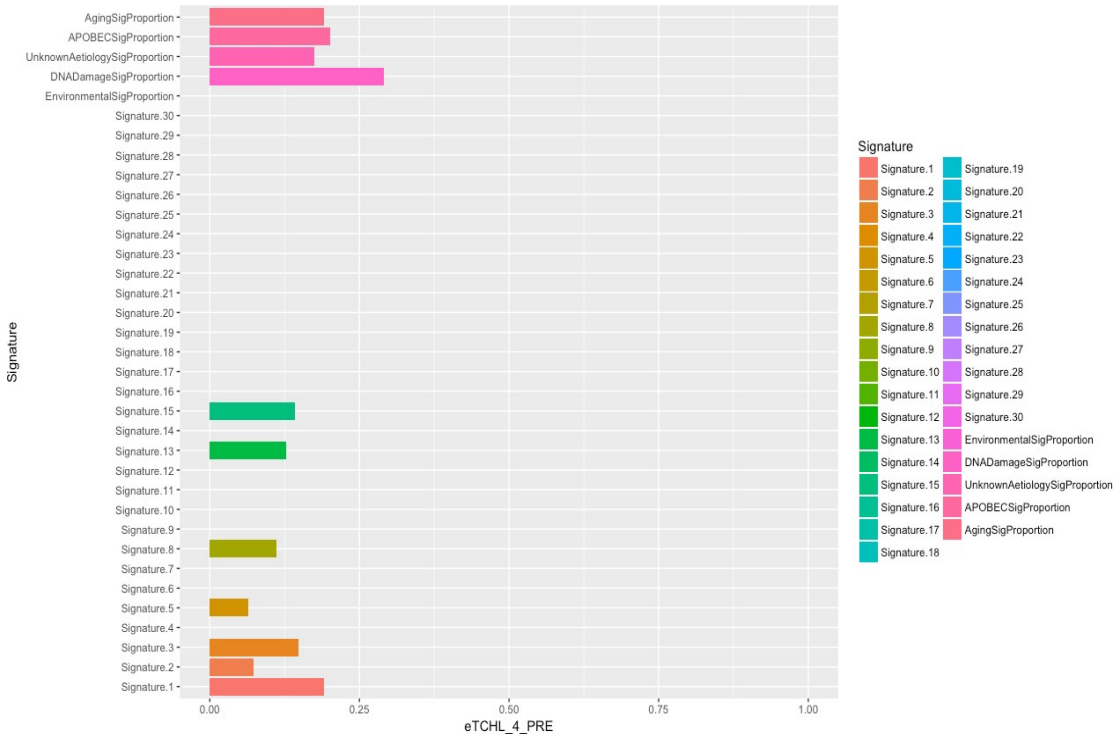
157. Heikkinen K, Rapakko K, Karppinen S-M, Erkko H, Knuutila S, Lundán T, et al. RAD50 and NBS1 are breast cancer susceptibility genes associated with genomic instability. Carcinogenesis [Internet]. 2006 Aug [cited 2018 Sep 7];27(8):1593–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16474176

158. Habermann JK, Doering J, Hautaniemi S, Roblick UJ, Bündgen NK, Nicorici D, et al. The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. Int J Cancer [Internet]. 2009 Apr 1 [cited 2018 Sep 7];124(7):1552–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19101988

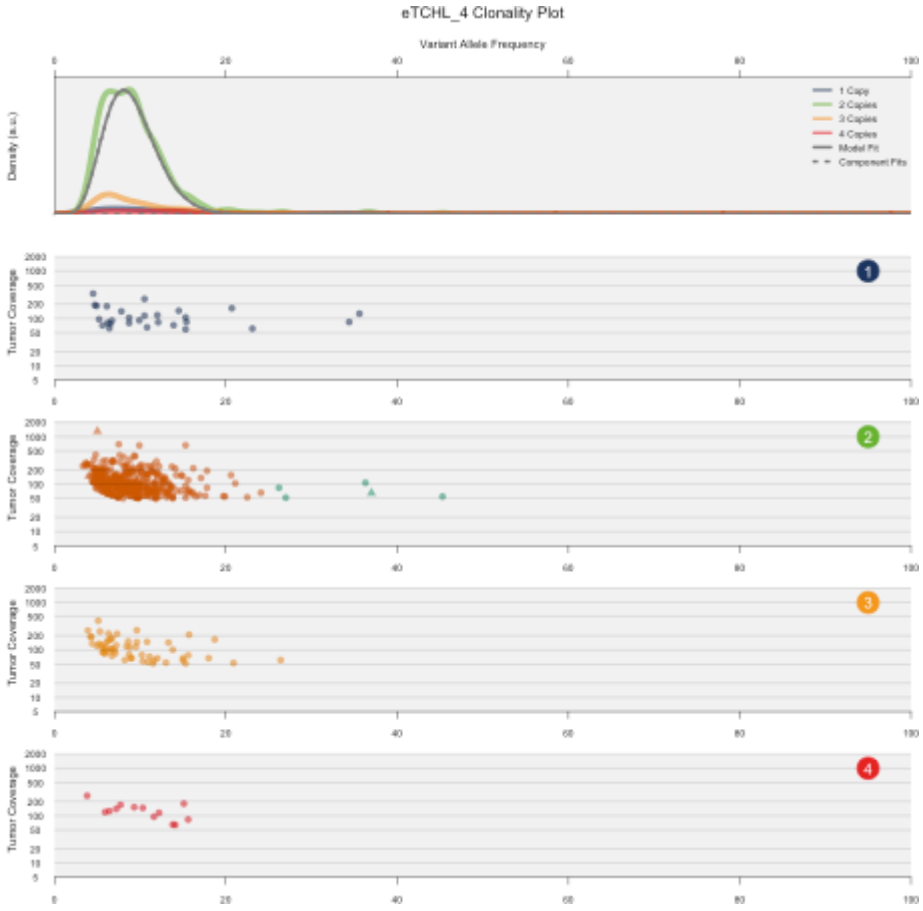
6 Supplementary materials

TCHL 4

Mutational Signatures



The TCHL 4 Pre-treatment sample shows a mutational spectrum with a variety of influences. As with all somatic cells, the Aging signature, signature 1 shows an influence. There is also an influence of DNA damage based signatures in the form of Signature 3 (DSB repair deficiency) and Signature 15 (DNA mismatch repair deficiency). APOBEC dysregulation also shows an influence in the form of Signatures 2 and 13. The final contribution is from signature 5, the signature of unknown aetiology found in all cancer types and seen in several samples in this cohort.



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
5|149441295|G|T	CSF1R	c.1744C>A	chr5:g.149441295G>T	p.L582M	Act	predicted driver: tier 2
1|120612018|C|T	NOTCH2	c.3G>A	chr1:g.120612018C>T	.	ambiguous	predicted driver: tier 2
17|7573987|G|GCCTC		c.1032_1039dupGAAT	chr17:g.7573988_757	p.A347Gfs*		predicted

ATTC	TP53	GAGG	3995dupCCTCATTC	26	LoF	driver: tier 1
17|11984705|C|G	MAP2K4	c.251C>G	chr17:g.11984705C> G	p.S84*	LoF	predicted driver: tier 1
12|6680190|G|C	CHD4	c.5566C>G	chr12:g.6680190G>C	p.Q1856E	Act	predicted driver: tier 1
11|108173675|G|C	ATM	c.5415G>C	chr11:g.108173675G> C	p.W1805C	LoF	predicted driver: tier 1

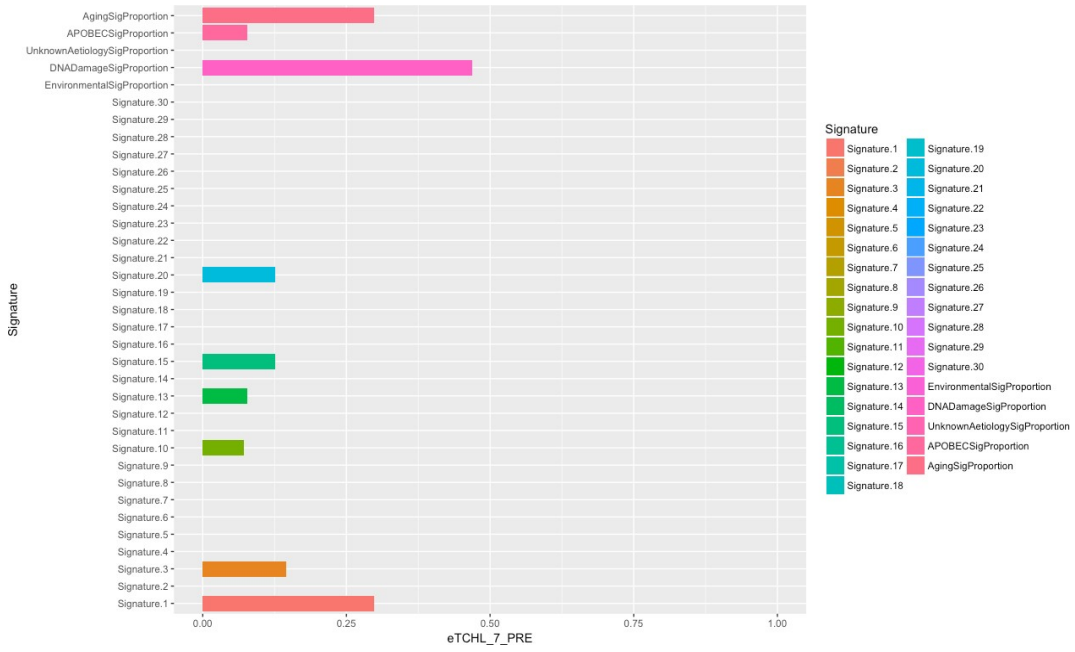
The TCHL 4 Pre-treatment sample shows 6 predicted driver mutations. As with many other samples in the cohort, one of these predicted driver mutations is a loss of function event in TP53. This is the only sample with predicted driver mutations in MAP2K4 and ATM. The remaining predicted driver mutations are in genes that also show predicted driver mutations in a small number of other samples in the cohort. Mutations in the following genes are likely oncogenic: CSF1R, CHD4

Mutations in the following genes are likely tumour suppressor inactivating: TP53, MAP2K4, ATM

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: NOTCH2

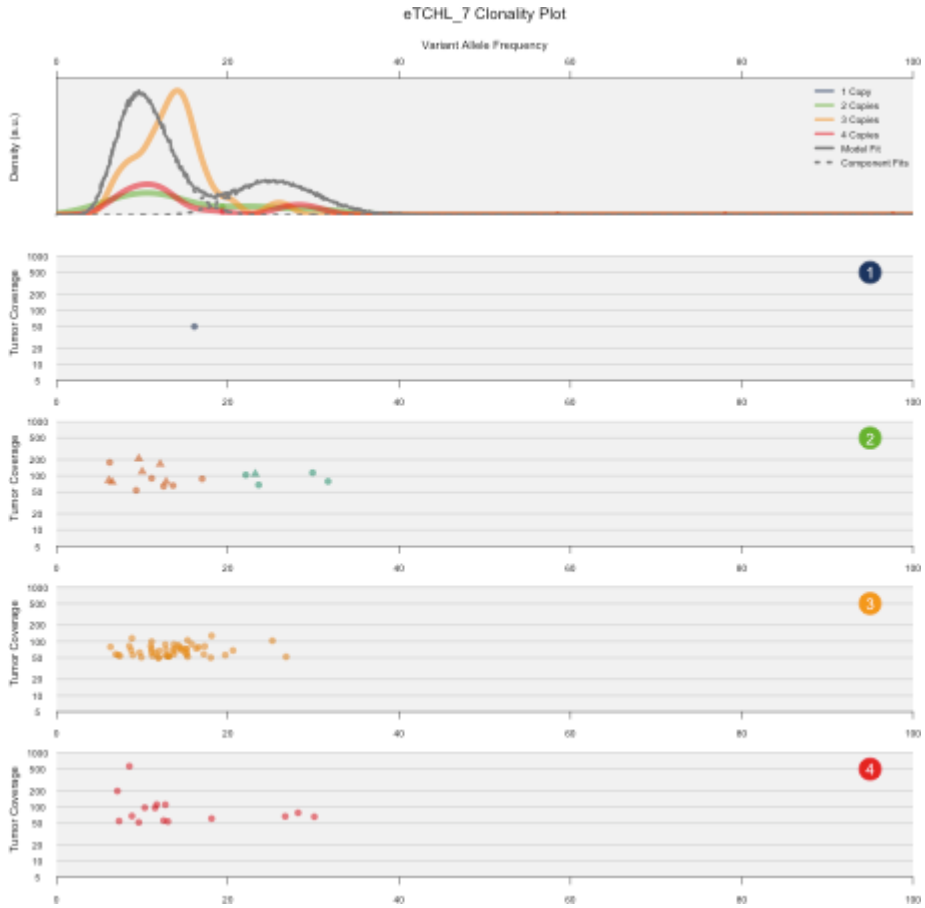
TCHL 7

Mutational Signatures



The TCHL 7 Pre-treatment sample shows a mutational spectrum dominated by DNA damaged related signatures, specifically Signature 3 (DSB repair deficiency), Signature 10 (altered activity of the error DNA polymerase POLE, aka DNA polymerase epsilon) and Signature 20 (defective DNA mismatch repair). As with all somatic cells, it is also influenced by the Aging signature, Signature 1. Finally, an APOBEC related signature (Signature 13) makes a small contribution to the mutational landscape of this sample.

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
5|131925489|C|T	RAD50	c.1412C>T	chr5:g.131925489C>T	p.S471L	ambiguous	predicted driver: tier 1
2|225371598|T|		c.1005_1006ins	chr2:g.225371598_225			predicted driver: tier

TC	CUL3	G	371599insC	p.K336Efs*5	LoF	1
1|179078400|G|A	ABL2	c.2002C>T	chr1:g.179078400G>A	p.R668C	Act	predicted driver: tier 2

The TCHL 7 Pre-treatment sample shows 3 predicted driver mutations. The RAD50 gene also shows predicted driver mutations in several other samples in the cohort (specifically 3 Post treatment, 6 relapse, 29 Pre-treatment and 32 relapse). The

remaining predicted driver mutations are in genes that do not show predicted driver mutations in any other samples in the cohort.

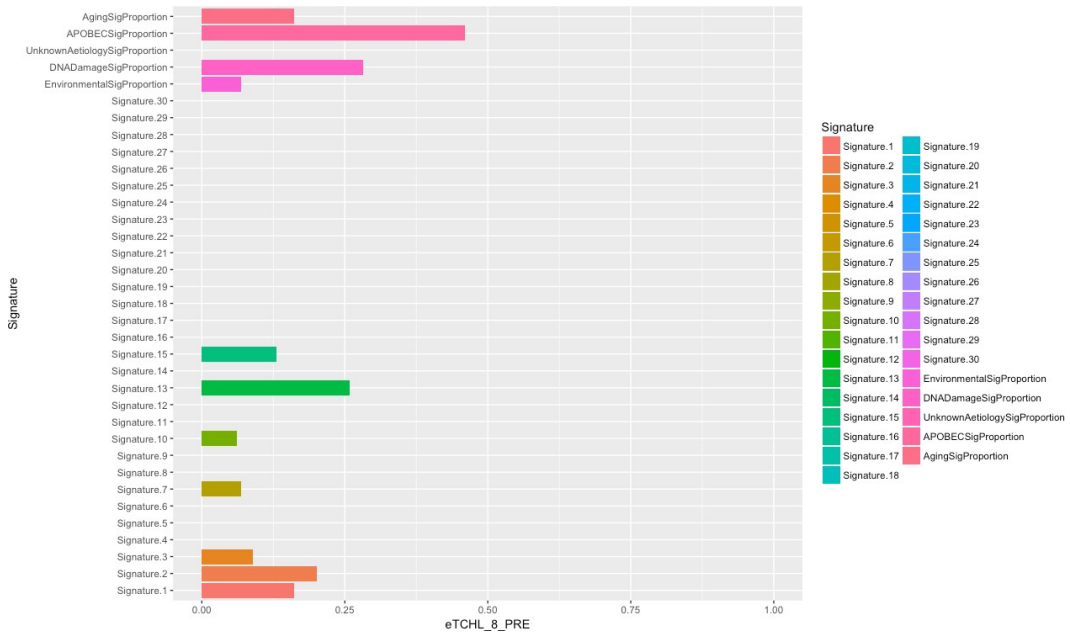
Mutations in the following genes are likely oncogenic: ABL2

Mutations in the following genes are likely tumour suppressor inactivating: CUL3

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: RAD50

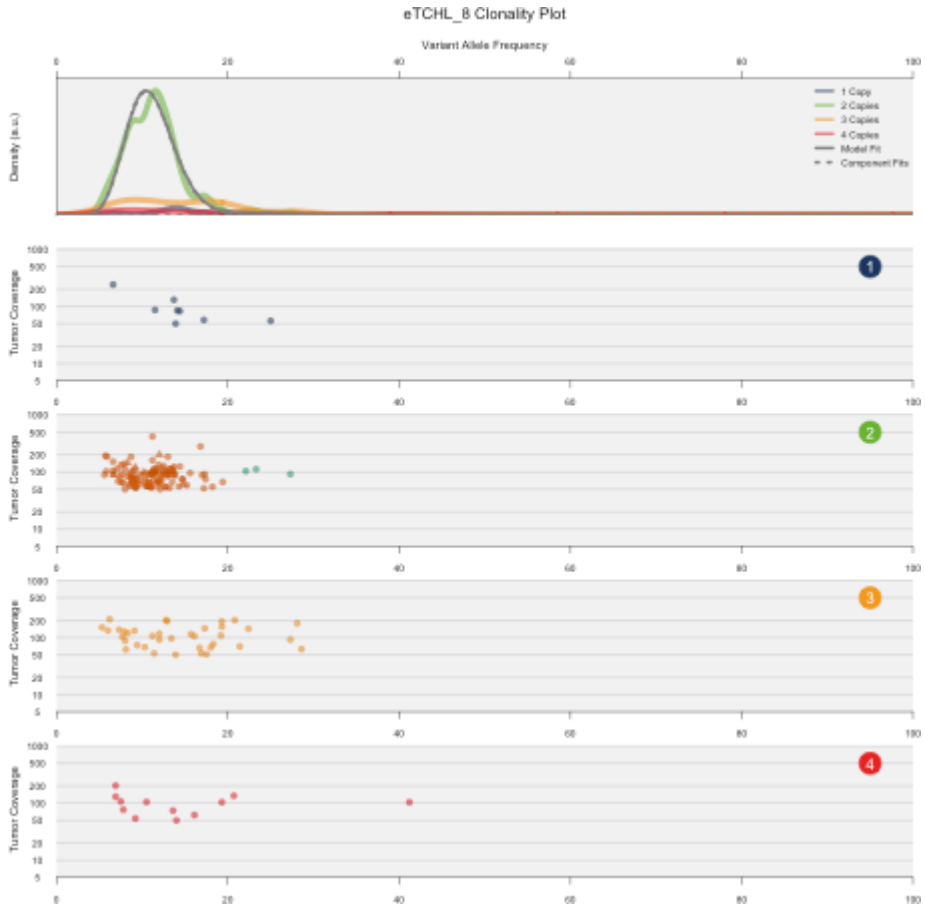
TCHL 8

Mutational Signatures



The TCHL 8 Pre-treatment sample shows a mutational landscape most heavily influenced by APOBEC related signatures (Signature 2 and 13). The next most influential contribution is DNA damage related signatures, specifically Signature 3 (DSB repair deficiency), Signature 10 (altered activity of error-prone polymerase POLE), and Signature 15 (defective DNA mismatch repair). The remaining contribution to the mutational landscape is from the Aging signature, Signature 1, and there is also a small Environmental signature contribution in the form of Signature 7. Signature 7 is the UV light associated signature. Signature 7 being present in this sample may indicate the patient was exposed to UV light prior to tumourigenesis (e.g. from sunlight) or it may be a mis-assigned signature due to an error of the deconstructSigs algorithm (see section 1.7 for an explanation of how this occurs).

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	**Driver statement**
X|47428122|C|T	ARAF	c.1082C>T	chrX:g.47428122C>T	p.T361M	Act	predicted driver: tier 1
17|7576889|C						predicted driver:

|A	TP53	c.957G>T	chr17:g.7576889C>A	p.K319N	LoF	tier 1
17|7216565|T C|T	GPS2	c.769delG	chr17:g.7216567delC	p.E257Nfs*86	LoF	predicted driver: tier 1
17|27829680| C|T	TAOK1	c.1277C>T	chr17:g.27829680C>T	p.S426F	ambiguous	predicted driver: tier 2
15|43713344| TC|T	TP53BP 1	c.4128del G	chr15:g.43713347del C	p.T1377Rfs*51	LoF	predicted driver: tier 1

The TCHL 8 Pre-treatment sample shows 5 predicted driver mutations. As with many other samples in the cohort, one of these predicted driver mutations is in the TP53 gene. TAOK1 shows predicted driver mutations in 2 other samples in the cohort (37 Pre-treatment and 87 Pre-treatment). The remaining predicted driver mutations are in

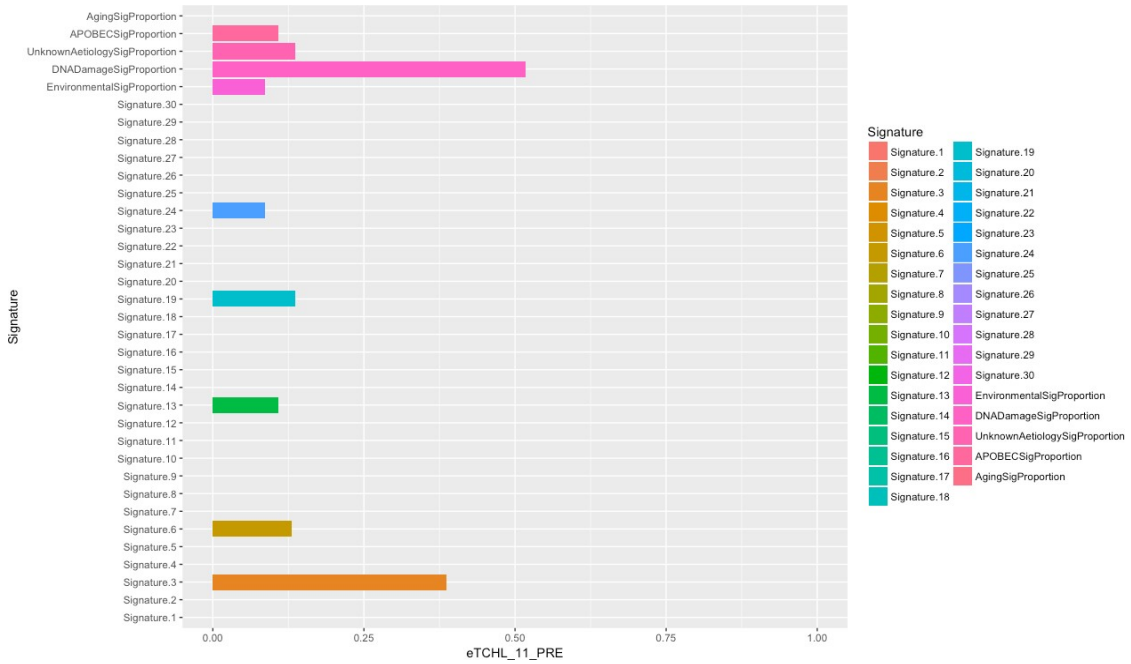
genes that do not show predicted driver mutations in many other samples in the cohort Mutations in the following genes are likely oncogenic: ARAF

Mutations in the following genes are likely tumour suppressor inactivating: TP53, GPS2, TP53BP1

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: TAOK1

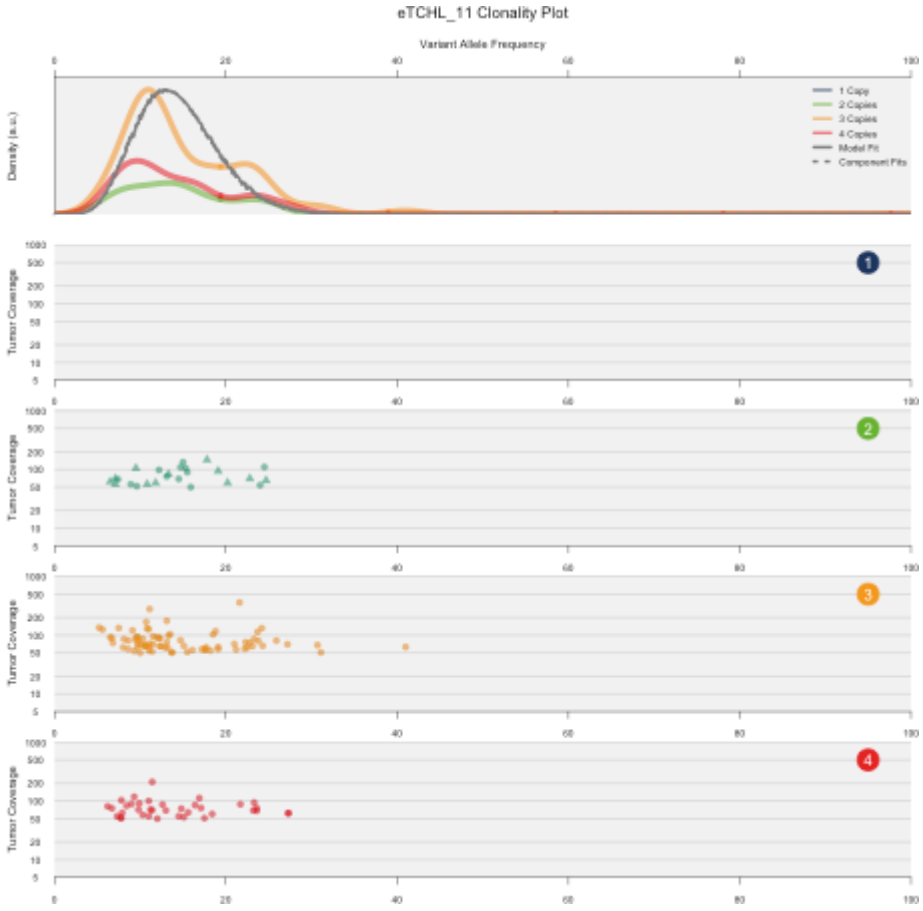
TCHL 11

Mutational Signatures



The TCHL 11 Pre-treatment sample shows a mutational landscape dominated by DNA damage related signatures, specifically Signature 3 (DSB repair deficiency) and Signature 6 (DNA mismatch repair deficiency). The next most important influence is Signature 19, which is of unknown aetiology. The APOBEC associated Signature 13 shows a similar level of influence to Signature 19. There is also an environmental influence in the form of Signature 24, which is associated with exposure to aflatoxin. No Aging Signature (Signature 1) is present. As discussed under other samples, this is presumably due to an error in how deconstructSigs assigns signatures rather than due to the signature genuinely not having affected the sample (see section 1.7 for an explanation)

SciClone Data



Driver analysis

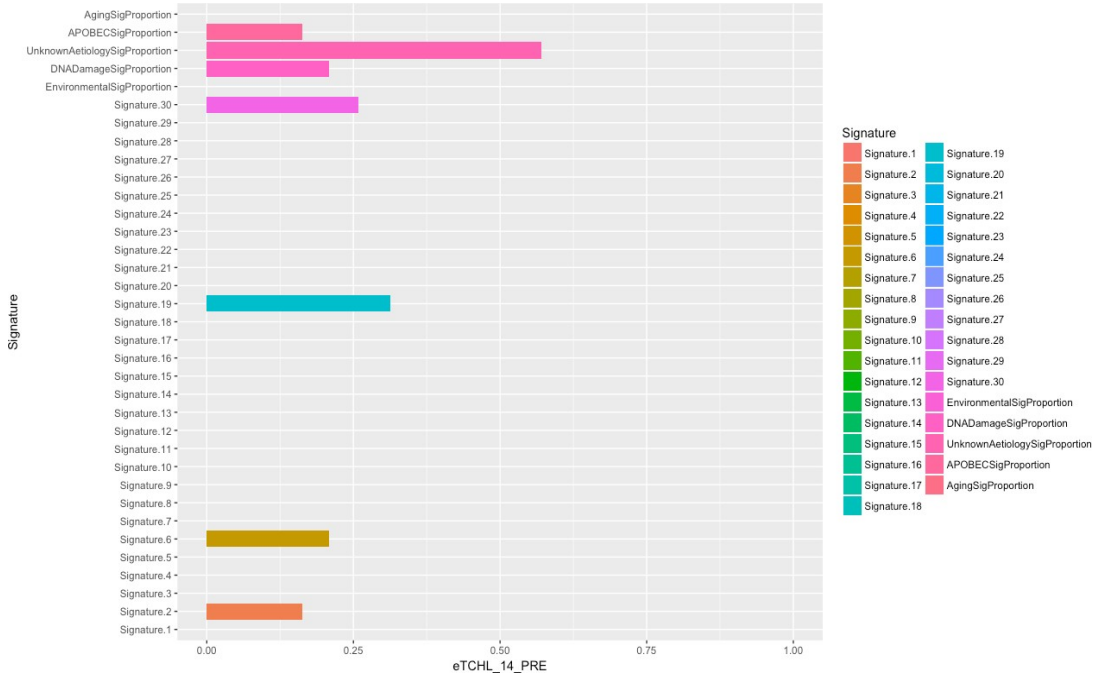
input	gene	cdna	gdna	protein	gene_role	Driver statement
17|7578442|T|C	TP53	c.488A>G	chr17:g.7578442T>C	p.Y163C	LoF	known in: CANCER;CANCER-PR
9|131339463|G|A	SPTAN1	c.841G>A	chr9:g.131339463G>A	p.D281N	LoF	predicted driver: tier 1
2|230683100|TG|T	TRIP12	c.1434delC	chr2:g.230683101delG	p.K479Rfs*7	LoF	predicted driver: tier 1

The TCHL 11 Pre-treatment sample shows 3 predicted driver mutations. As with many other samples in the cohort, one of these predicted driver mutations is in the TP53 gene. The other predicted driver mutations are in genes that do not show predicted driver mutations in any other samples in the cohort. 3 driver mutations seems like a relatively low number of driver mutations for a tumour to have, suggesting that some of the driver mutations in this sample were either filtered during variant calling or not recognised by the CGI algorithm.

Mutations in the following genes are likely tumour suppressor inactivating: TP53, SPTAN1, TRIP12

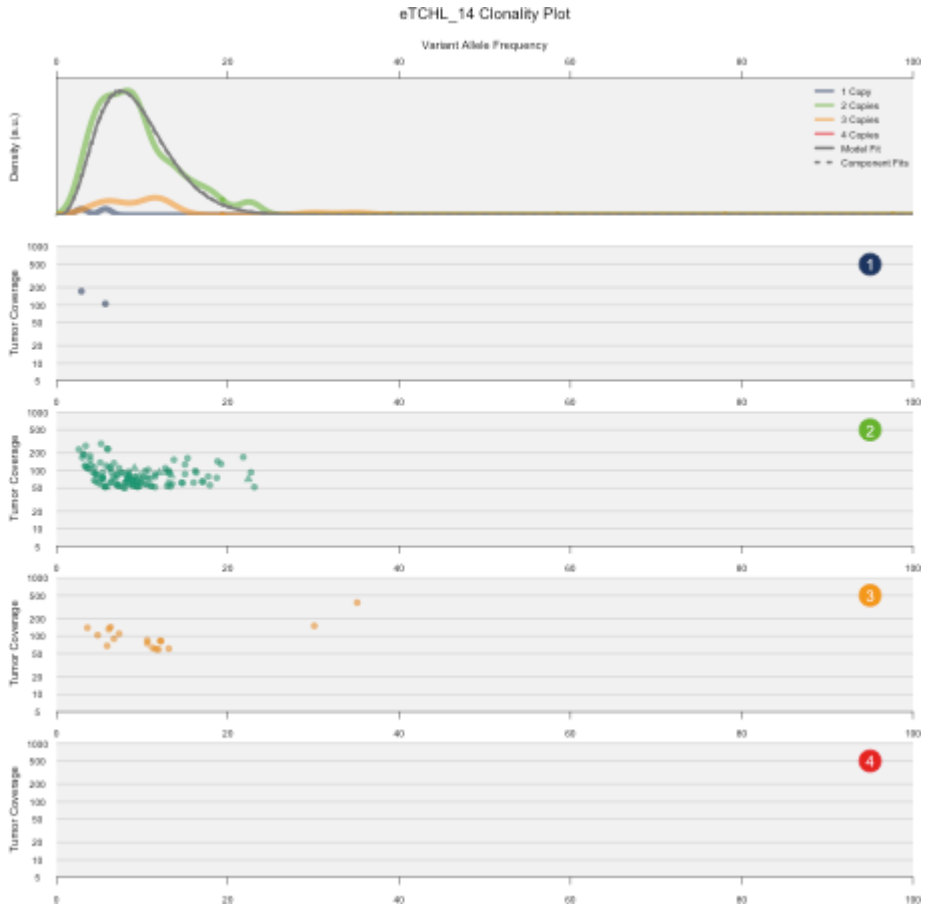
TCHL 14

Mutational Signatures



The TCHL 14 Pre-treatment sample shows a mutational landscape predominantly caused by signatures of unknown origin, specifically Signature 19 and Signature 30 (known to be observed in some breast cancers). The DNA damage associated Signature 6 (DNA mismatch repair deficiency) and the APOBEC associated Signature 6 also contribute. No Aging Signature (Signature 1) is present. As discussed under other samples, this is presumably due to an error in how deconstructSigs assigns signatures rather than due to the signature genuinely not having affected the sample (see section 1.7 for an explanation)

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178952085|A|G	PIK3CA	c.3140A>G	chr3:g.178952085A>G	p.H1047R	Act	known in: BRCA;OV;COREAD;NSCLC

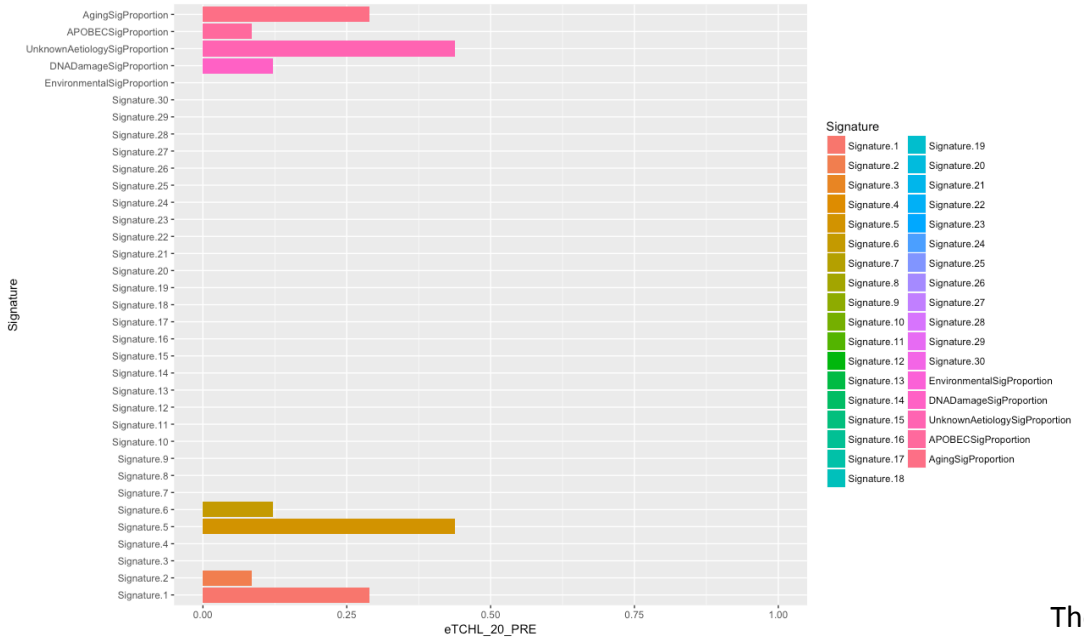
The TCHL 14 Pre-treatment sample shows 1 validated driver mutation, the PIK3CA H1047R mutation. PIK3CA is also mutated in several other samples in the cohort. Since

the sample must have more than 1 active driver mutation to be an active tumour, this suggests that some of the driver mutations in the sample may have been filtered during variant calling or not recognised by the CGI algorithm.

Mutations in the following genes are likely oncogenic: PIK3CA

TCHL 20

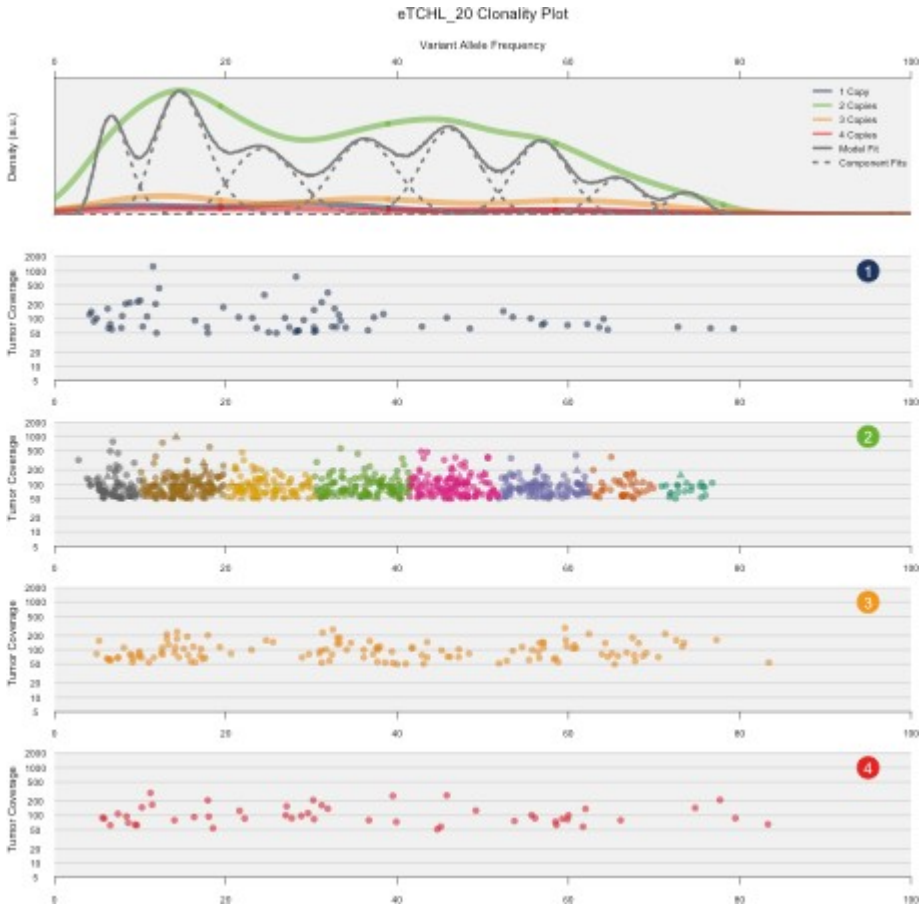
Mutational Signatures



The TCHL 20 Pre-treatment sample shows a mutational landscape predominated by Signature 5, a signature of unknown aetiology that has been observed in all cancer types. The next most influential contribution is from signature 1, the Aging based signature that influences all somatic cells. Finally, there is a contribution observed from the DNA damage associated Signature 6 (DNA mismatch repair associated) and

the APOBEC associated Signature 2.

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
7|151932949|C|A	MLL3	c.2722G>T	chr7:g.151932949 C>A	p.G908C	LoF	predicted driver: tier 1
5|14394217|C|T	TRIO	c.4289C>T	chr5:g.14394217C >T	p.T1430M	Act	predicted driver: tier 1

19|6743224|G|A	TRIP10	c.365G>A	chr19:g.6743224 G>A	p.R122Q	Act	predicted driver: tier 1
18|48604751|A|G	SMAD4	c.1573A>G	chr18:g.48604751 A>G	p.I525V	LoF	predicted driver: tier 1
18|20529651|G|A	RBBP8	c.223G>A	chr18:g.20529651 G>A	p.E75K	LoF	predicted driver: tier 1
12|88487680|AT|A	CEP290	c.3175delA	chr12:g.88487688 delT	p.I1059*fs*1	LoF	predicted driver: tier 1

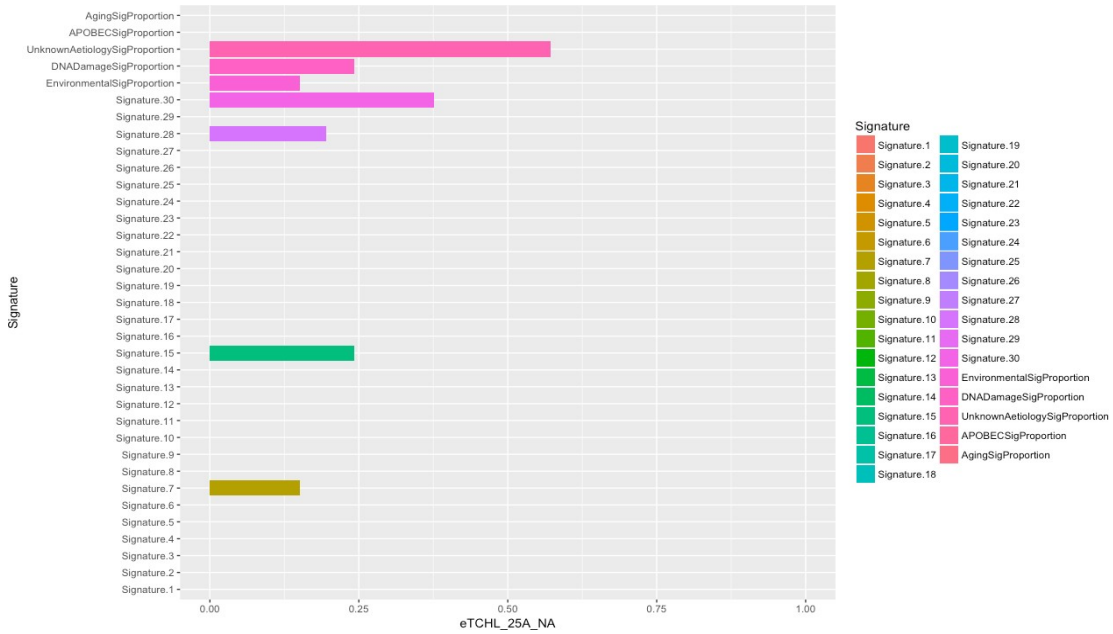
The TCHL 20 Pre-treatment sample shows 6 predicted driver mutations. TRIP10, CEP290 and MLL3 do not show predicted driver mutations in any other samples in the cohort. The remaining predicted driver mutations are in genes that show predicted driver mutations in a small number of other samples in the cohort.

Mutations in the following genes are likely oncogenic: TRIP10, TRIO

Mutations in the following genes are likely tumour suppressor inactivating: MLL3, SMAD4, RBBP8, CEP290

TCHL 25A – Pre-treatment sample

Mutational Signatures



The TCHL 25 Pre-treatment sample shows a mutational landscape dominated by signatures of unknown aetiology, specifically signatures 28 and 30. There is also a signature of DNA damage repair deficiency in the form of Signature 15 (DNA mismatch repair deficiency). Finally, there is an environmentally based signature in the form of Signature 7, the UV light based signature. This may reflect extended exposure to UV light (e.g. from sunlight) in the patient prior to tumourigenesis.

SciClone Data

SciClone was unable to generate data from the TCHL 25 Pre-treatment sample because there were no copy number 2 regions for SciClone to operate on.

Driver analysis

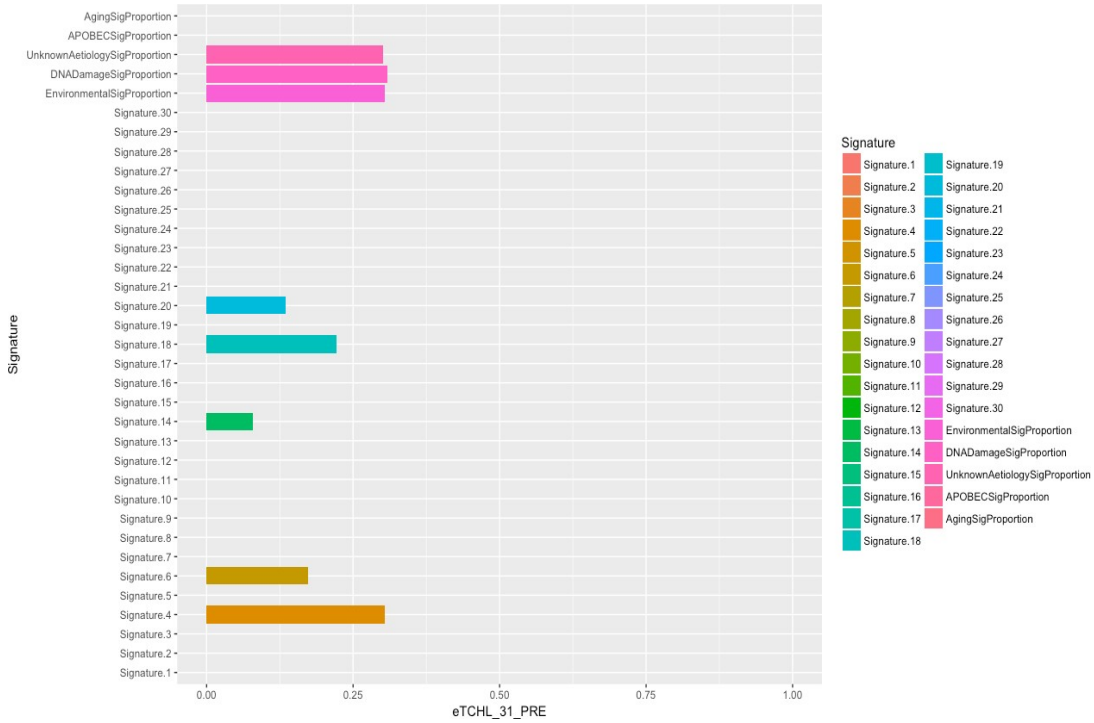
input	gene	cdna	gdna	protein	gene_role	Driver statement
15|42057114|C|T	MGA	c.7775C>T	chr15:g.42057114C>T	p.T2592I	LoF	predicted driver: tier 1

The TCHL 25 Pre-treatment sample shows 1 predicted driver mutation. The MGA gene shows a predicted driver mutation in one other sample in the cohort, the TCHL 12 Pre-treatment sample. Since the sample must have more than 1 active driver mutation to be an active tumour, this suggests that some of the driver mutations in the sample may have been filtered during variant calling or not recognised by the CGI algorithm.

Mutations in the following genes are likely tumour suppressor inactivating: MGA

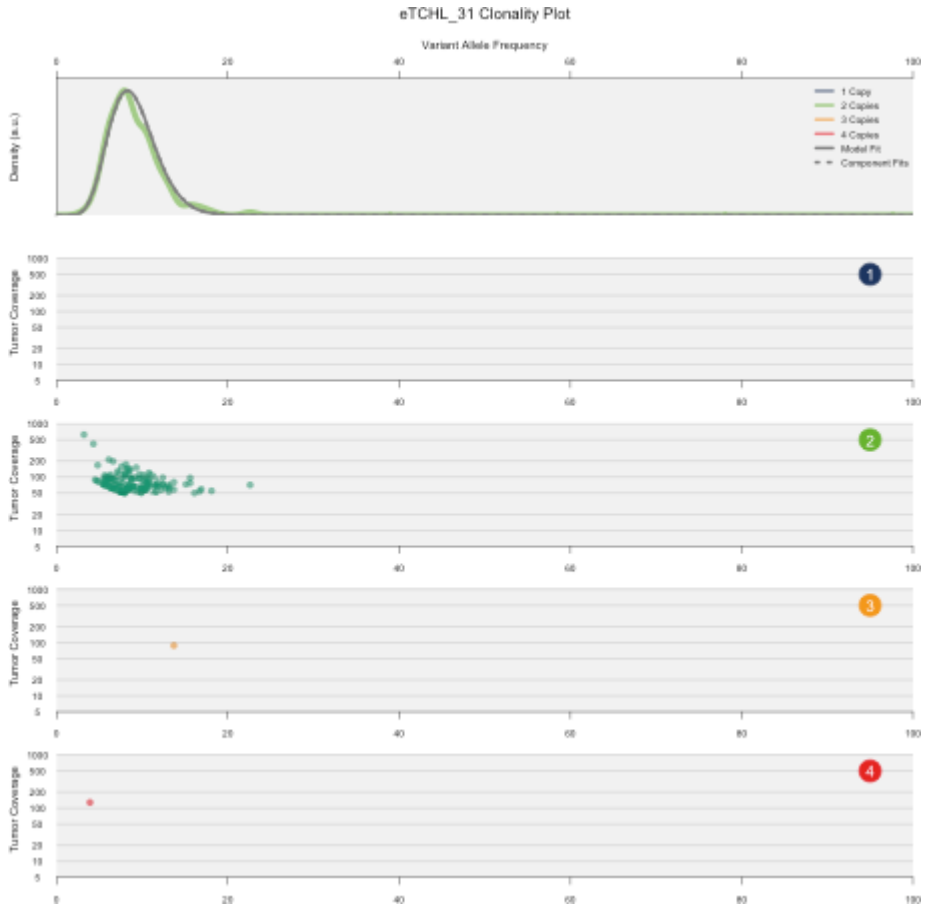
TCHL 31

Mutational Signatures



The TCHL 31 Pre-treatment sample shows a mutational landscape with equal contributions of DNA damage based signatures, an environmental signature and signatures of unknown aetiology. The DNA damage based Signatures are Signature 6 and Signature 20 (both (DNA mismatch repair deficiency based). The Environmental signature is signature 4, the tobacco smoke associated signature - this may indicate that the patient smoked tobacco and that the effect spread to the breast tissue, ormay be a signature mis-assignment on the part of deconstructSigs. The signatures of unknown aetiology are Signature 14 and Signature 19. As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7)

SciClone Data

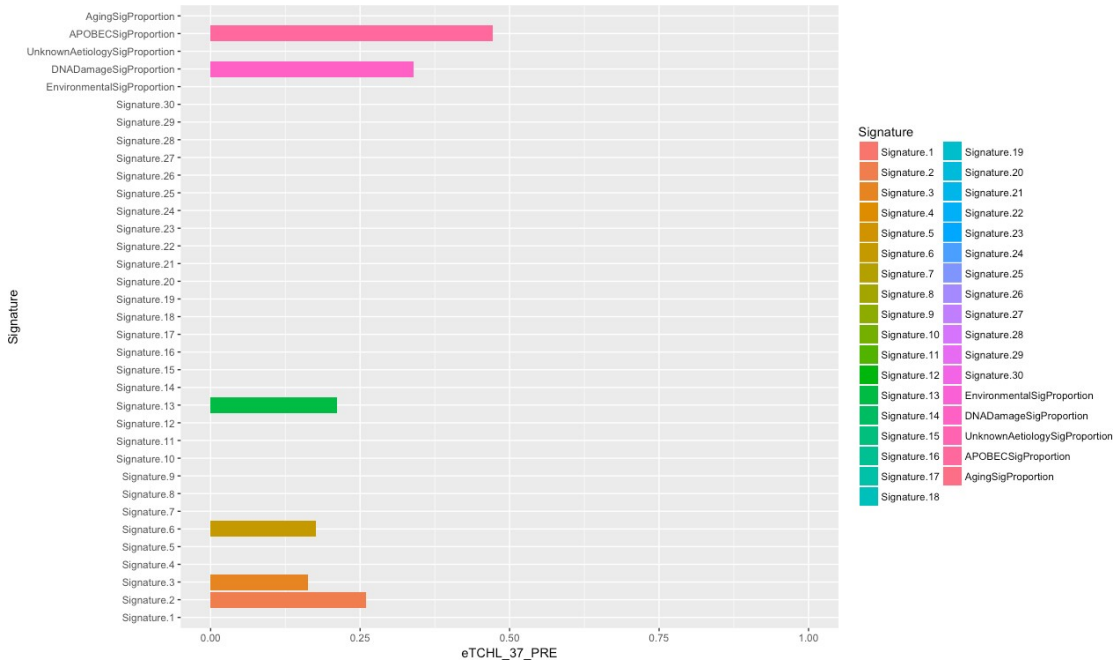


Driver analysis

None of the SNVs or indels present in this sample are known or predicted drivers, according to CGI. This suggests that some of the driver mutations in the sample may have been filtered during variant calling or not recognised by the CGI algorithm.

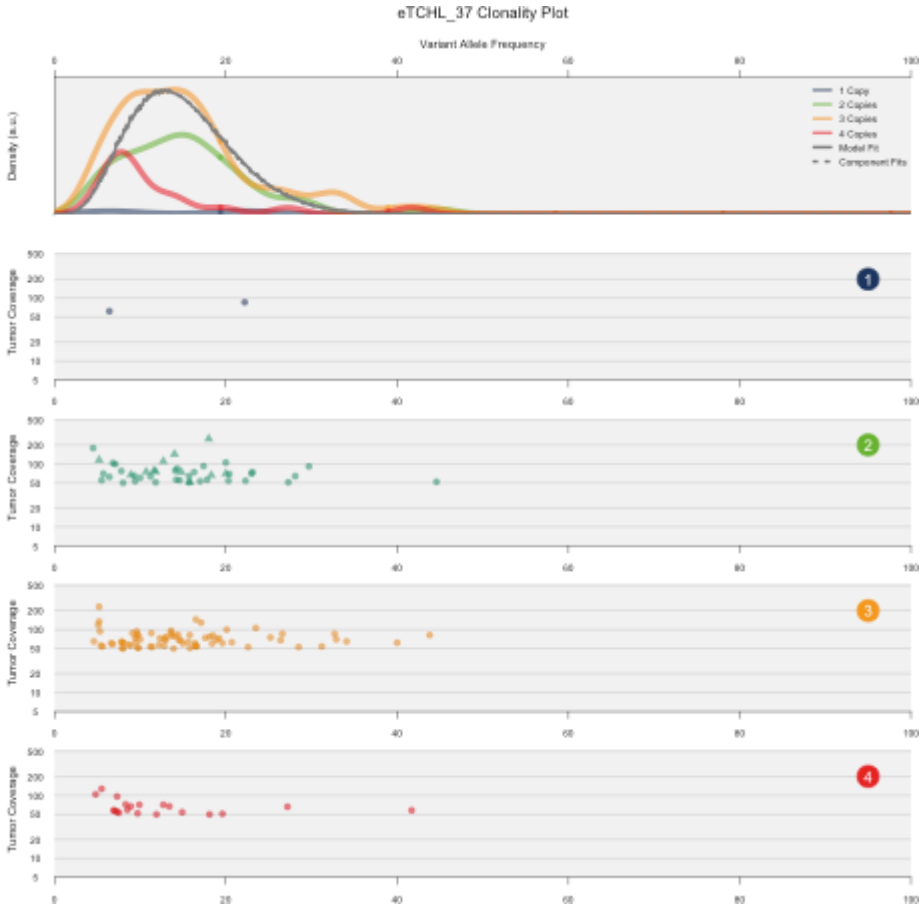
TCHL 37

Mutational Signatures



The TCHL 37 Pre-treatment sample shows a mutational spectrum comprised entirely of DNA damage associated signatures and APOBEC associated signatures. The DNA damage associated signatures are Signature 3 (DSB repair deficiency associated). The APOBEC associated signatures are Signature 2 and Signature 13. As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7)

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
17|7578265|A|G	TP53	c.584T>C	chr17:g.7578265A>G	p.I195T	LoF	known in: CANCER;CANCER- PR
8|68138293|G|A	ARFGEF1	c.4042C>T	chr8:g.68138293G>A	p.R1348C	ambiguous	predicted driver: tier 1

7|2977543|G|T	CARD11	c.1141C>A	chr7:g.29775 43G>T	p.Q381K	Act	predicted driver: tier 2
3|195507004|G|C	MUC4	c.11447C>G	chr3:g.19550 7004G>C	p.S3816*	ambiguous	predicted driver: tier 2
17|27804710|C|T	TAOK1	c.338C>T	chr17:g.2780 4710C>T	p.S113L	ambiguous	predicted driver: tier 1

The TCHL 37 Pre-treatment sample shows 4 predicted driver mutations and 1 validated driver mutation. As with many other samples in the cohort, the validated driver

mutation in this sample is a mutation in TP53. TAOK1 shows predicted driver mutations in two other samples in the cohort (8 Pre-treatment and 87 Pre-treatment). The remaining predicted driver mutations in this sample are in genes that do not shows predicted driver mutations in any other samples in this cohort.

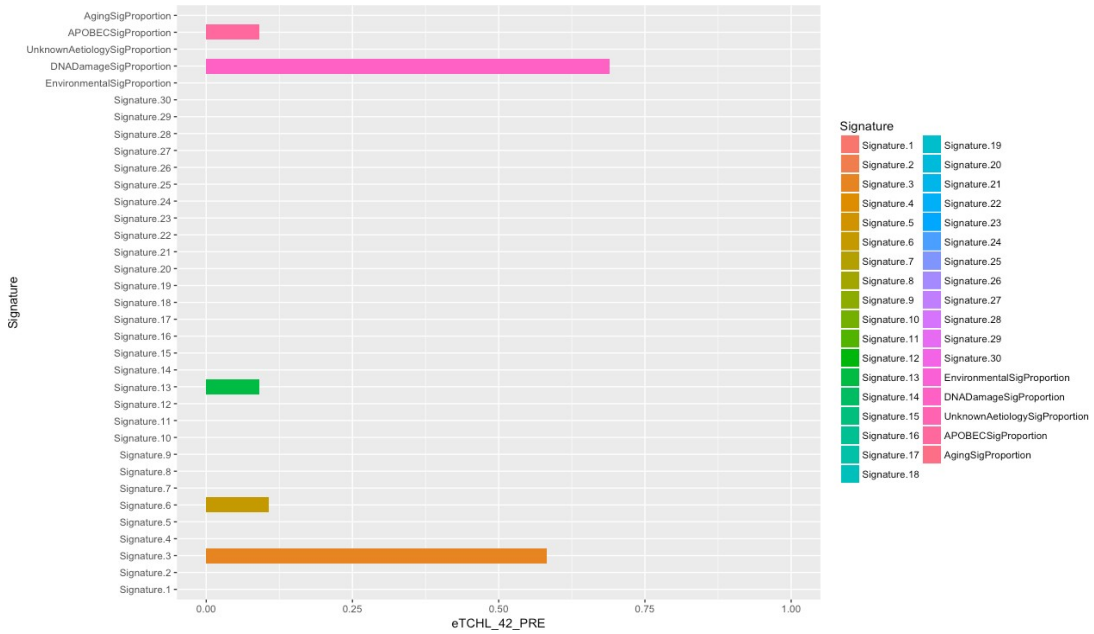
Mutations in the following genes are likely oncogenic: CARD11

Mutations in the following genes are likely tumour suppressor inactivating: TP53

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: ARFGEF1, MUC4, TAOK1

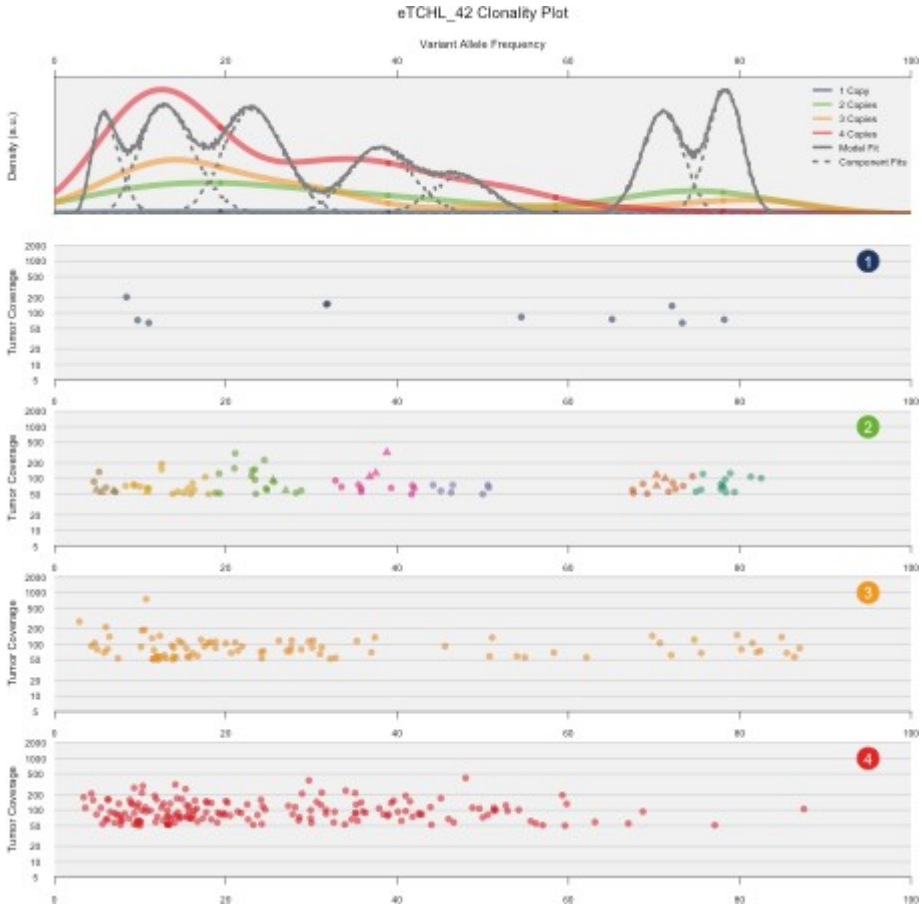
TCHL 42

Mutational Signatures



The TCHL 42 Pre-treatment sample shows a mutational landscape almost entirely composed of DNA damage associated signatures. This is almost all due to the heavy presence of Signature 3 (DSB repair deficiency associated), with a small contribution by Signature 6 (DNA repair deficiency associated). There is also a small contribution by the APOBEC associated Signature 13. As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7)

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
17|7577120|C|T	TP53	c.818G>A	chr17:g.7577120C>T	p.R273H	LoF	known in: AML;THCA;CANCER-PR
5|80074565|C|G	MSH3	c.2345C>G	chr5:g.80074565C>G	p.S782C	LoF	predicted driver: tier 2
4|153245503|ATTGA		c.1678_1687del	chr4:g.153245504_15324	p.D560Sfs		predicted

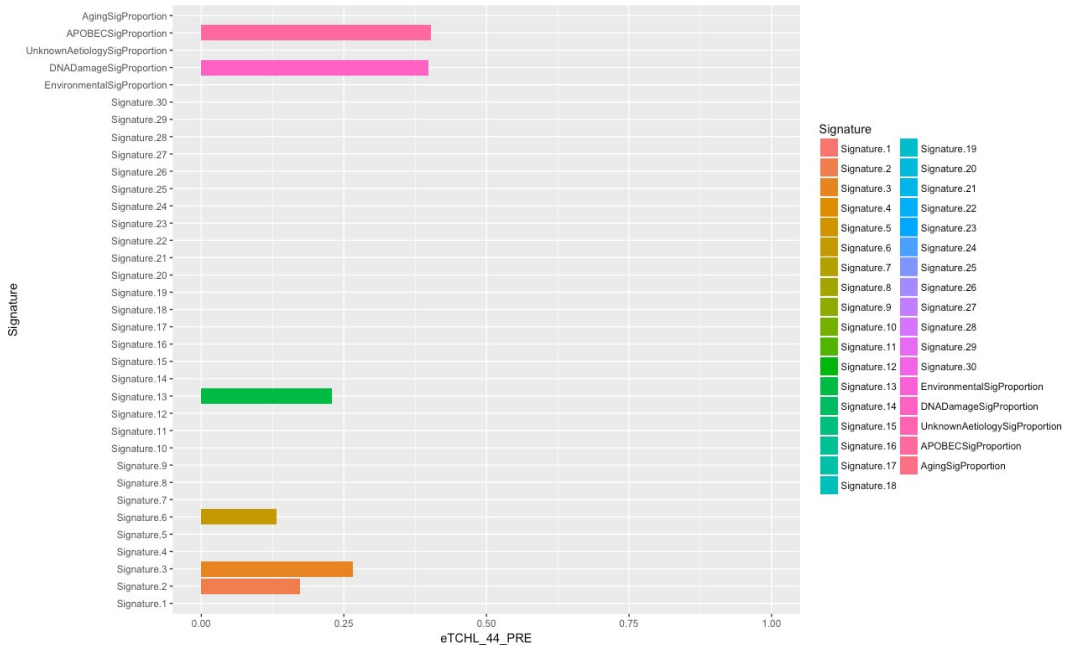
TGTATC|A	FBXW7	GATACATCAA	5513delTTGATGTATC	*15	LoF	driver: tier 1
18|48573663|CAGGT T|C	SMAD4	c.249+1_249+5 delGTTAG	chr18:g.48573666_48573670delGTTAG	.	LoF	predicted driver: tier 1

The TCHL 42 Pre-treatment sample shows 1 validated driver mutation and 3 predicted driver mutations. As with many other samples in this cohort, the validated driver mutation is a mutation in TP53. SMAD4 shows a predicted driver mutation in one other

sample in the cohort (TCHL 42 PRE). The remaining predicted driver mutations were in genes that do not show predicted driver mutations in any other sample in the cohort. Mutations in the following genes are likely tumour suppressor inactivating: TP53, MSH3, FBXW7, SMAD4

TCHL 44

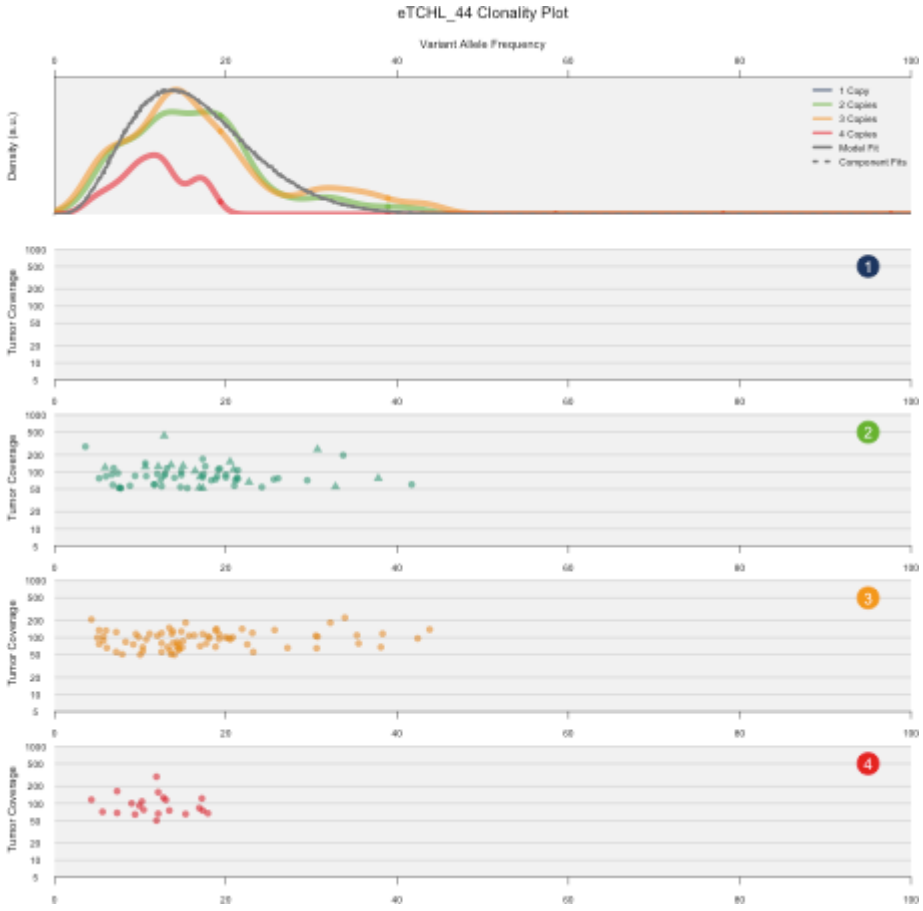
Mutational Signatures



The TCHL 44 Pre-treatment sample shows a mutational landscape with equal contributions by DNA damage associated signatures and APOBEC associated signatures. The DNA damage signature contribution is mostly Signature 3 (DSB repair deficiency), with Signature 6 (DNA mismatch repair deficiency) making a smaller contribution.

Signature 2 and Signature 13 make a fairly even contribution to the APOBEC associated part of the mutational landscape of this sample. As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7)

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178936091|G|A	PIK3CA	c.1633G>A	chr3:g.178936091G>A	p.E545K	Act	known in: COREAD;NSCLC;OV;BRCA
17|7577085|C|T	TP53	c.853G>A	chr17:g.7577085C>T	p.E285K	LoF	known in: CANCER-PR;CANCER

X|38146372|G|C	RPGR	c.1880C>G	chrX:g.38146372G>C	p.S627*	LoF	predicted driver: tier 1
8|38133960|C|A	WHSC1L1	c.3926G>T	chr8:g.38133960C>A	p.R1309I	Act	predicted driver: tier 1
5|112179771|G|A	APC	c.8480G>A	chr5:g.112179771G>A	p.G2827E	LoF	predicted driver: tier 1
2|43452437|T|G	ZFP36L2	c.506A>C	chr2:g.43452437T>G	p.K169T	LoF	predicted driver: tier 1

19|13050425|CAG|C	CALR	c.379_380delGA	chr19:g.13050427_13050428delGA	p.E127Ifs*39	ambiguous	predicted driver: tier 2
13|33741754|C|T	STARD13	c.175G>A	chr13:g.33741754C>T	p.E59K	Act	predicted driver: tier 1
10|43596088|G|C	RET	c.255G>C	chr10:g.43596088G>C	p.W85C	Act	predicted driver: tier 1

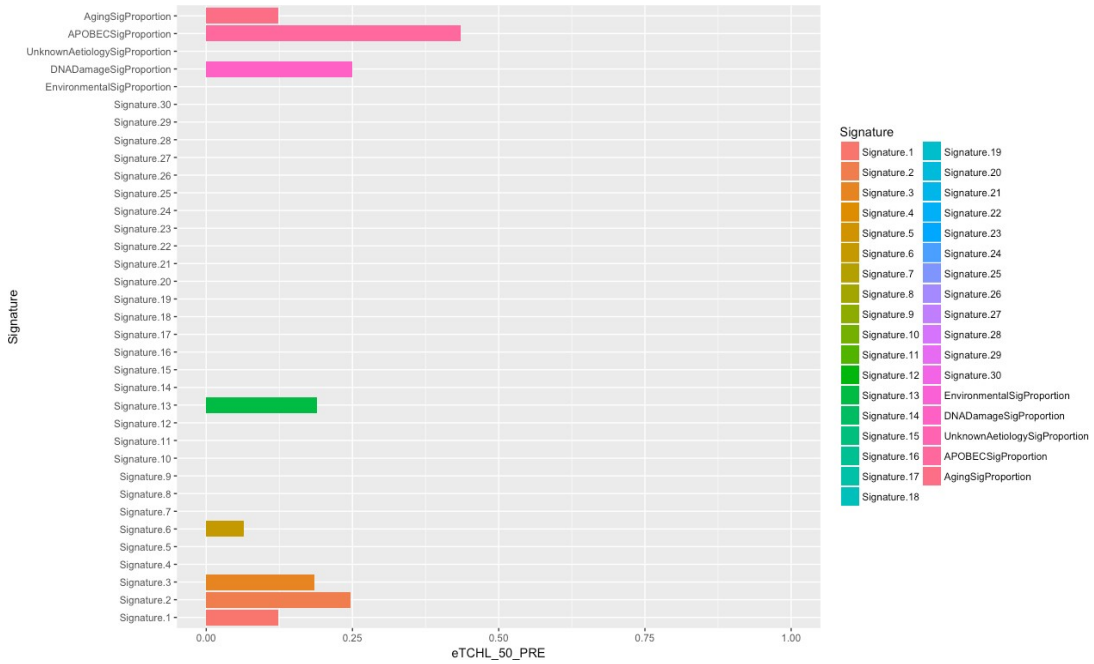
The TCHL 44 Pre-treatment sample shows 2 validated driver mutations and 7 predicted driver mutations. As with many other samples in the cohort, the predicted driver mutations are in TP53 and PIK3CA. APC shows a predicted driver mutation in one other sample in the cohort (TCHL 32 relapse). RPGR also shows a predicted driver mutation in one other sample in the cohort (TCHL 76 Pre-treatment). The remaining predicted driver mutations in this sample are in genes that do not show predicted driver mutations in any other sample in this cohort.

Mutations in the following genes are likely oncogenic: PIK3CA, WHSC1L1, STARD13, RET Mutations in the following genes are likely tumour suppressor inactivating: TP53, RPGR, APC, ZFP36L2

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: CALR

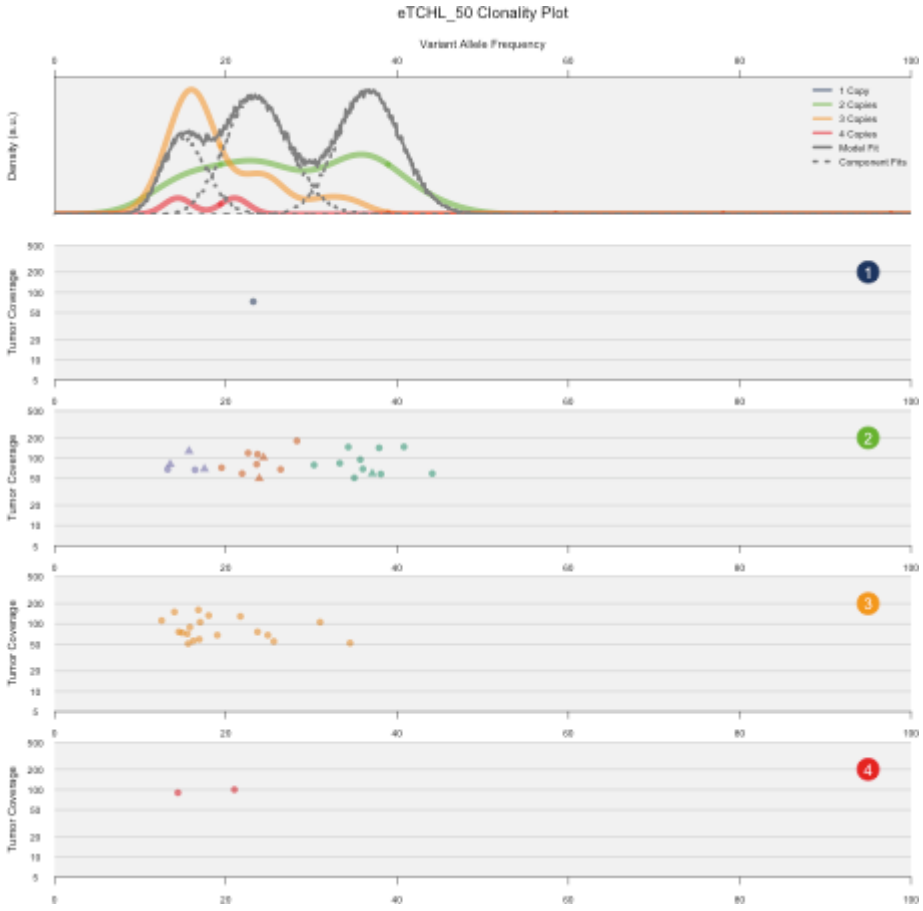
TCHL 50

Mutational Signatures



The TCHL 50 Pre-treatment sample shows a mutational landscape dominated by APOBEC associated signatures, specifically Signature 2 and 13. There is also a smaller contribution from DNA damage associated signatures, mostly Signature 3 (DSB repair deficiency), with a minor contribution from Signature 6 (DNA mismatch repair deficiency). Finally, the Aging-based Signature 1, present in all somatic cells, makes a small contribution.

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
3|178936091|G|A	PIK3CA	c.1633G>A	chr3:g.178936091G> A	p.E545K	Act	known in: COREAD;BRCA;NSCLC ;OV
17|7576873|C|A	TP53	c.973G>T	chr17:g.7576873C>A	p.G325*	LoF	known in: CANCER-PR
18|52921912|C|T	TCF4	c.1166G>A	chr18:g.52921912C>T	p.R389H	Act	predicted driver: tier 1

17|37680929|T|C	CDK12	c.3098T>C	chr17:g.37680929T>C	p.L1033P	LoF	predicted driver: tier 1
17|29592291|T|G	NF1	c.4769T>G	chr17:g.29592291T>G	p.L1590*	LoF	predicted driver: tier 1

The TCHL 50 Pre-treatment sample shows 2 validated driver mutations and 3 predicted driver mutations. As with many other samples in the cohort, the validated driver mutations are mutations in TP53 and PIK3CA. CDK12 shows a predicted driver mutation

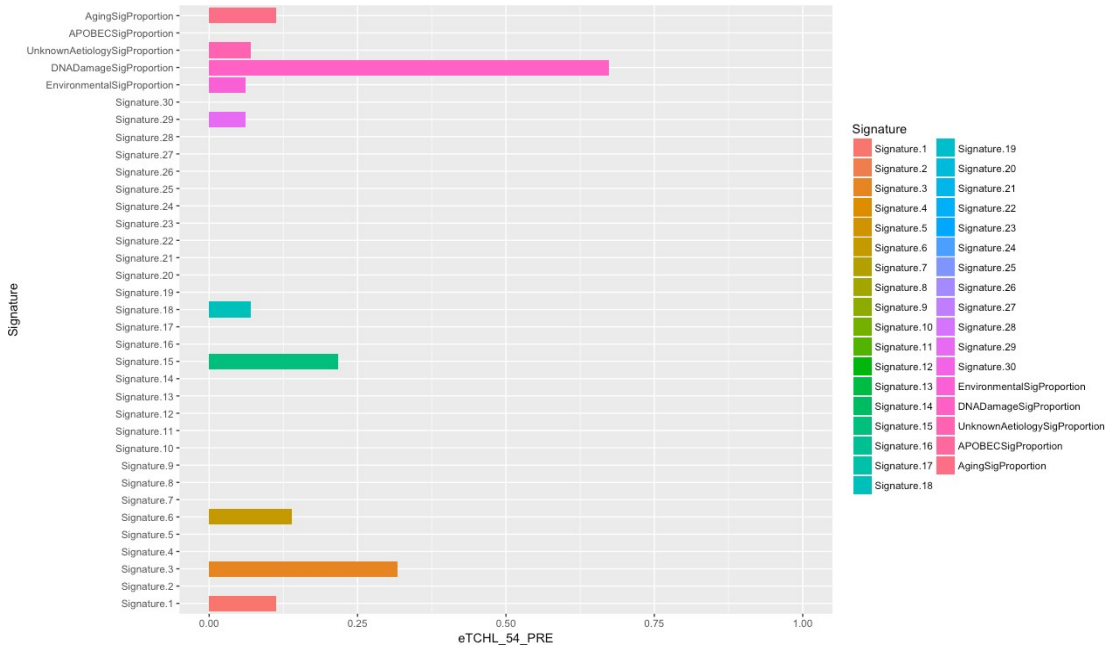
in one other sample in the cohort (TCHL 87 Pre-treatment). The remaining predicted driver mutations are in genes that do not show predicted driver mutations in any other samples in the cohort.

Mutations in the following genes are likely oncogenic: PIK3CA, TCF4

Mutations in the following genes are likely tumour suppressor inactivating: TP53, CDK12, NF1

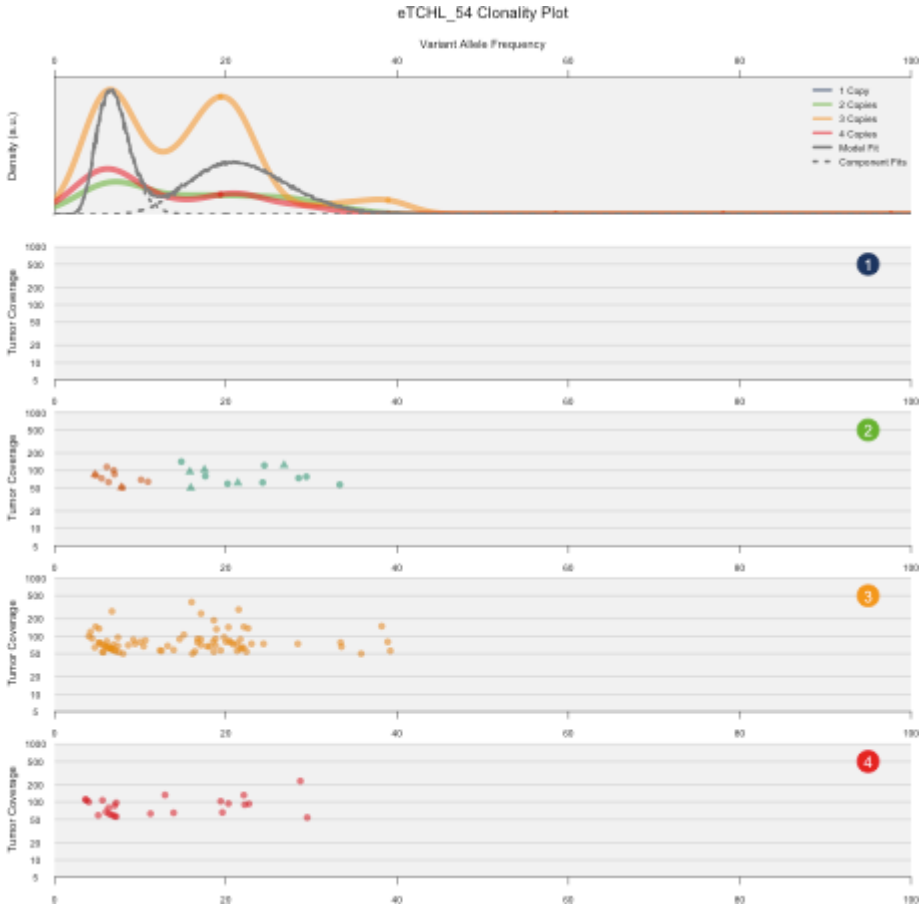
TCHL 54

Mutational Signatures



The TCHL 54 Pre-treatment sample shows a mutational spectrum completely dominated by DNA damage related signatures. The most prevalent DNA damage based signature present is Signature 3 (DSB repair deficiency), followed by Signature 15 (DNA mismatch repair deficiency) and Signature 6 (DNA mismatch repair deficiency. There is also a minor contribution by Signature 1, the Aging signature, as well as Signature 18 (unknown aetiology) and the environment-based Signature 29. Signature 29 is associated with tobacco chewing in mouth cancer. It is unlikely that the process that causes Signature 29 is actually active in a breast cancer cell, so this is probably an error of deconstructSigs.

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	**Driver statement**
1|183536344|C|T	NCF2	c.850G>A	chr1:g.183536344C>T	p.G284R	ambiguous	predicted driver: tier 1
17|7577145|GTAGAT|G	TP53	c.788_792delATCTA	chr17:g.7577148_7577152delGATTA	p.N263Tfs*7	LoF	predicted driver: tier 1

16|3830776|C|T	CREBBP	c.1780G>A	chr16:g.3830776C>T	p.E594K	LoF	predicted driver: tier 1
10|17271455|C|T	VIM	c.34C>T	chr10:g.17271455C>T	p.R12C	ambiguous	predicted driver: tier 2

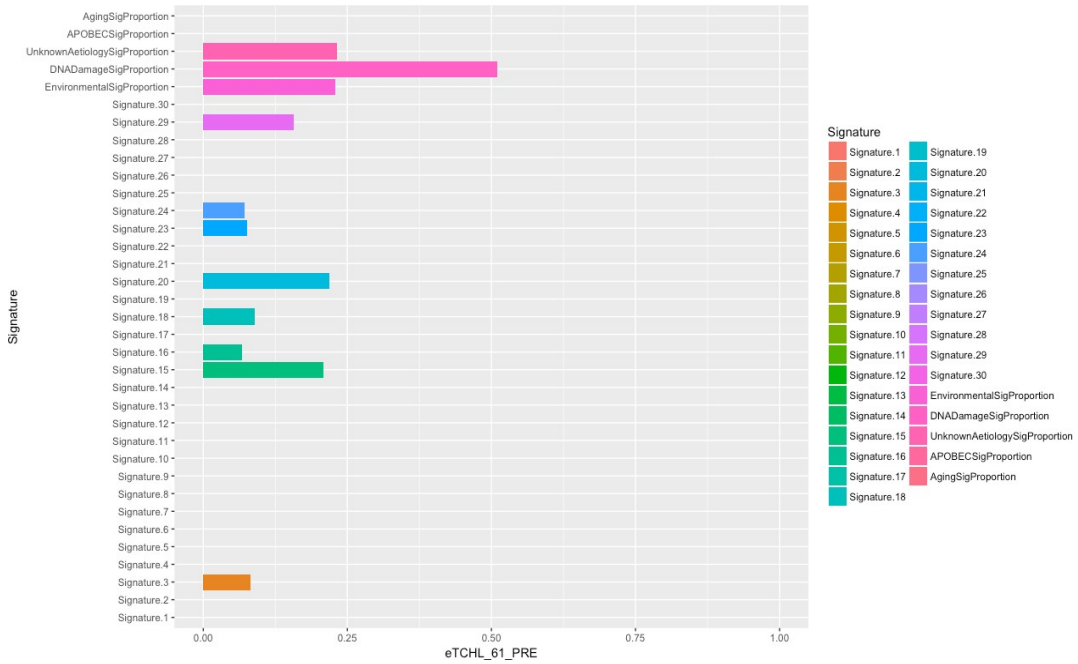
The TCHL 54 Pre-treatment sample shows 4 predicted driver mutations. As with many oth3r samples in the cohort, one of the predicted driver mutations is a mutation in TP53. The remaining predicted driver mutations are in genes that do not show predicted driver mutations in any other samples in the cohort.

Mutations in the following genes are likely tumour suppressor inactivating: TP53, CREBBP

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: VIM, NCF2

TCHL 61

Mutational Signatures



The TCHL 61 Pre-treatment sample shows a mutational landscape dominated by DNA damage related signatures, specifically Signature 3 (DSB repair deficiency), Signature 15 (DNA mismatch repair deficiency) and Signature 20 (DNA mismatch repair deficiency). The remainder of the mutational landscape is split evenly between signatures of unknown aetiology (specifically Signatures 16, 18 and 23) and Environmentally based signatures. The environmentally based signatures are Signature 24 (aflatoxin exposure associated) and Signature 29 (associated with tobacco chewing). As discussed previously for other samples showing these signatures, Signature 24 may reflect actual exposure to aflatoxin, whereas Signature 29 is unlikely to genuinely reflect exposure to the known cause of the signature and has likely been assigned in error. As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7)

SciClone Data

SciClone was unable to find any clusters in the data provided from this sample.

Driver analysis

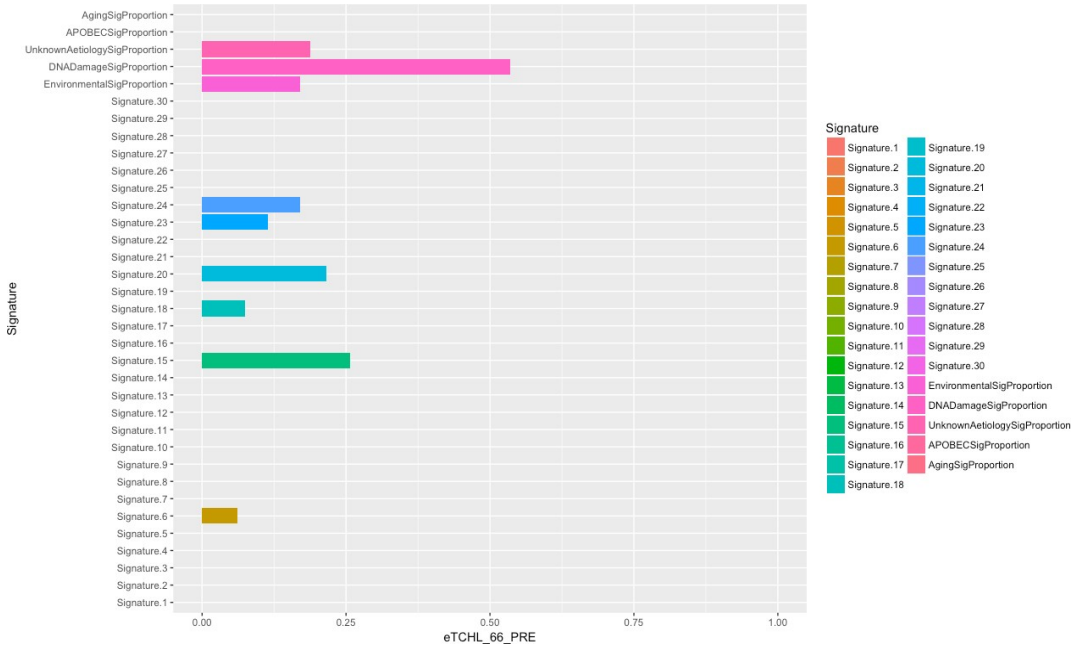
input	gene	cdna	gdna	protein	gene_role	Driver statement
17|7578263|G|A	TP53	c.586C>T	chr17:g.7578263G>A	p.R196*	LoF	known in: CANCER-PR
9|32427317|C|G	ACO1	c.1367C>G	chr9:g.32427317C>G	p.A456G	LoF	predicted driver: tier 1
7|50450397|C|T	IKZF1	c.581C>T	chr7:g.50450397C>T	p.T194M	ambiguous	predicted driver: tier 1
19|3110189|G|A	GNA11	c.179G>A	chr19:g.3110189G>A	p.R60H	Act	predicted driver: tier 2

The TCHL 61 Pre-treatment sample shows 1 validated driver mutation and 3 predicted driver mutations. As with many other samples in the cohort, the validated driver mutation is a mutation in the TP53 gene. GNA11 also shows predicted driver mutations in the TCHL 6 Pre, Post and Relapse samples. The remaining predicted driver mutations are in genes that do not show predicted driver mutations in any other samples in the cohort.

Mutations in the following genes are likely oncogenic: GNA11

Mutations in the following genes are likely tumour suppressor inactivating: TP53, ACO1 It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: IKZF1

TCHL 66



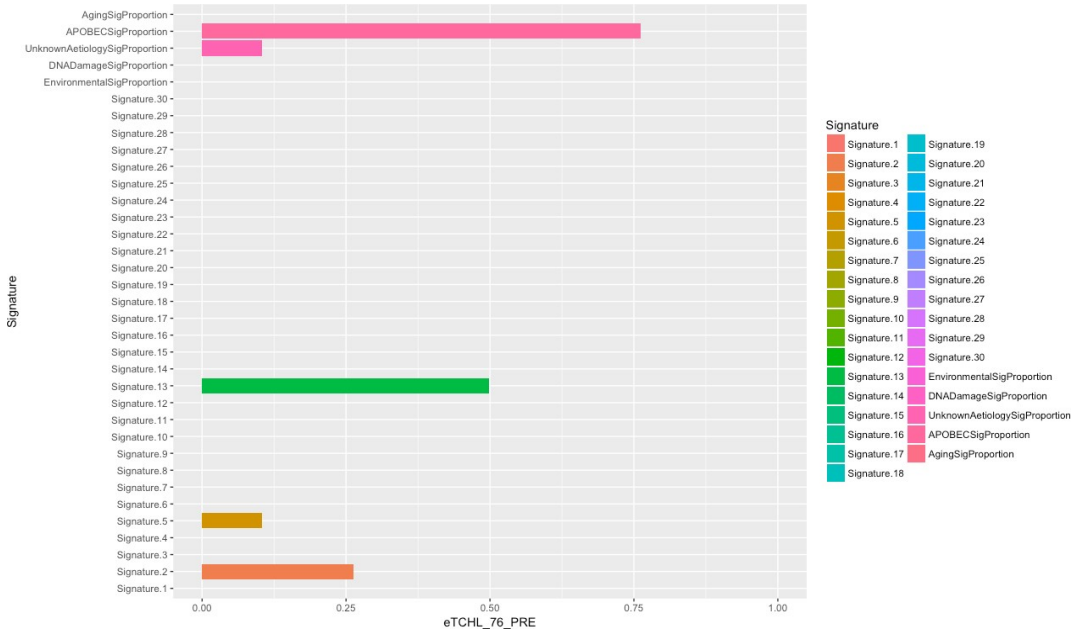
The TCHL 66 Pre-treatment sample shows a mutational landscape dominated by DNA damage related signatures, specifically Signature 6, Signature 15 and Signature 20 (all DNA mismatch repair related). There is also an environmental signature contribution, specifically Signature 24, the aflatoxin exposure associated signature. This may reflect the patient having had exposure to aflatoxin. The remainder of the mutational landscape is made up of signatures of unknown aetiology (Signatures 18 and 23). As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7)

SciClone Data

SciClone was unable to find any clusters in the data provided from this sample. Driver analysis

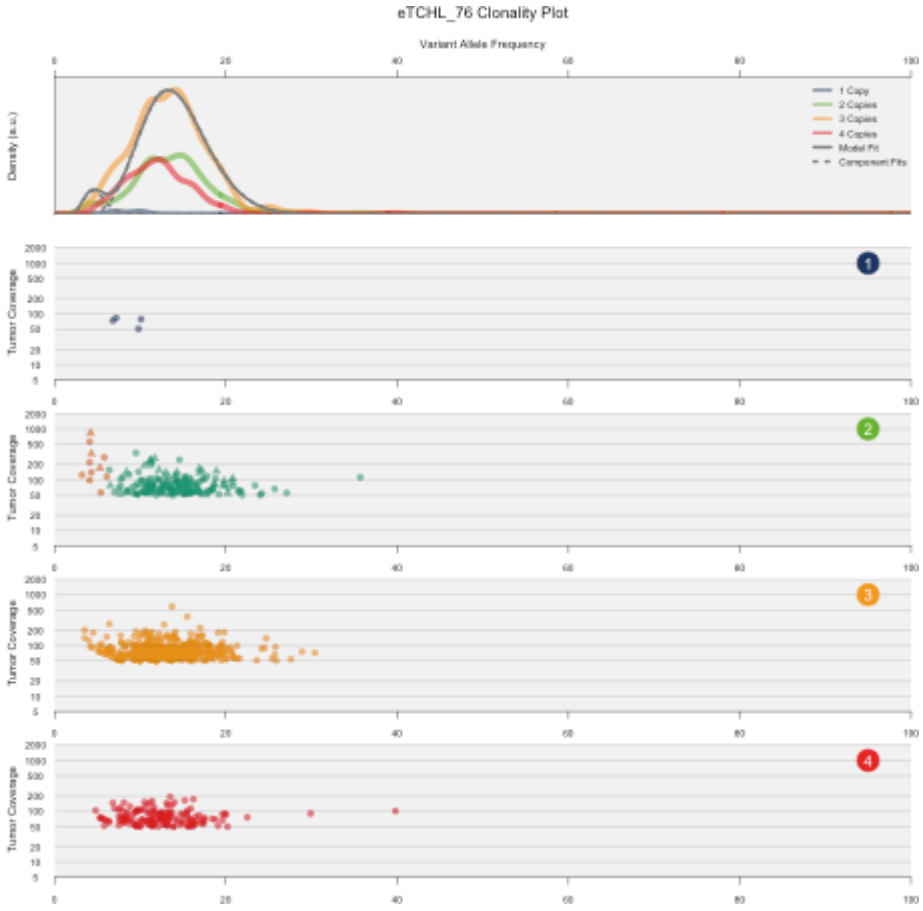
None of the SNVs or indels present in this sample are known or predicted drivers, according to CGI. This suggests that some of the driver mutations in the sample may have been filtered during variant calling or not recognised by the CGI algorithm.

TCHL 76



The TCHL 76 Pre-treatment samples shows a mutational landscape completely dominated by the APOBEC associated Signature 2 and 13. The only other contribution is a minor contribution by Signature 5, a signature of unknown aetiology found in all cancer types and seen in some other samples in this cohort. As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7)

SciClone Data



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
12|112888166|A|G	PTPN11	c.182A>G	chr12:g.112888166A>G	p.D61G	Act	known in: MML
X|38146499|C|T	RPGR	c.1754- 1G>A	chrX:g.38146499C> T	.	LoF	predicted driver: tier 1

X|16870688|C|G	RBBP7	c.1081G>C	chrX:g.16870688C>G	p.D361H	Act	predicted driver: tier 1
9|113259100|C|T	SVEP1	c.1795G>A	chr9:g.113259100C>T	p.E599K	ambiguous	predicted driver: tier 1
7|152346219|CA| (	XRCC2	c.350delT	chr7:g.152346227delA	p.L117Wfs*17	ambiguous	predicted driver: tier 2
7|151874347|C|G	MLL3	c.8191G>C	chr7:g.151874347C>G	p.E2731Q	LoF	predicted driver: tier 1

6|168348622|G|A	MLLT4	c.3622G>A	chr6:g.168348622G>A	p.E1208K	Act	predicted driver: tier 1
2|39505604|C|T	MAP4K3	c.1738G>A	chr2:g.39505604C>T	p.E580K	LoF	predicted driver: tier 1
20|40111993|C|A	CHD6	c.2424G>T	chr20:g.40111993C>A	p.M808I	Act	predicted driver: tier 1
1|27105513|G|C	ARID1A	c.5125- 1G>C	chr1:g.27105513G>C	.	LoF	predicted driver: tier 1
1|197057559|C|G	ASPM	c.9988G>C	chr1:g.197057559C>G	p.E3330Q	Act	predicted driver: tier 1
17|8052807|G|A	PER1	c.826C>T	chr17:g.8052807G>A	p.P276S	Act	predicted driver: tier 2
17|62552026|A|T	SMURF2	c.1522T>A	chr17:g.62552026A>T	p.F508I	ambiguous	predicted driver: tier 2
17|48268243|C|G	COL1A1	c.2278G>C	chr17:g.48268243C>G	p.D760H	Act	predicted driver: tier 1
17|40991365|C|T	PSME3	c.691C>T	chr17:g.40991365C>T	p.R231W	LoF	predicted driver: tier 2
16|2225359|G|T	TRAF7	c.1444G>T	chr16:g.2225359G>T	p.V482L	LoF	predicted driver: tier 2
15|28443879|C|T	HERC2	c.7753G>A	chr15:g.28443879C>T	p.E2585K	ambiguous	predicted driver: tier 1
14|95571531|C|T	DICER1	c.3146G>A	chr14:g.95571531C>T	p.R1049K	LoF	predicted driver: tier 1

13|48881426|G|T	RB1	c.148G>T	chr13:g.48881426G>T	p.E50*	LoF	predicted driver: tier 1

The TCHL 76 Pre-treatment sample shows 1 validated driver mutation and 17 predicted driver mutations. As mentioned in the intro, this is a large number of driver mutations for one tumour to have, suggesting that potentially not all of the predicted driver mutations in this sample are contributing to the cancer. RPGR also shows a predicted driver mutation in the TCHL 44 Pre-treatment sample, while MLL3 also shows a predicted driver mutation in the TCHL 20 Pre-treatment sample. The remaining driver mutations assigned by CGI, including the validated driver mutation in PTPN11, are mutations in genes that do not show predicted driver mutations in any other samples in the cohort.

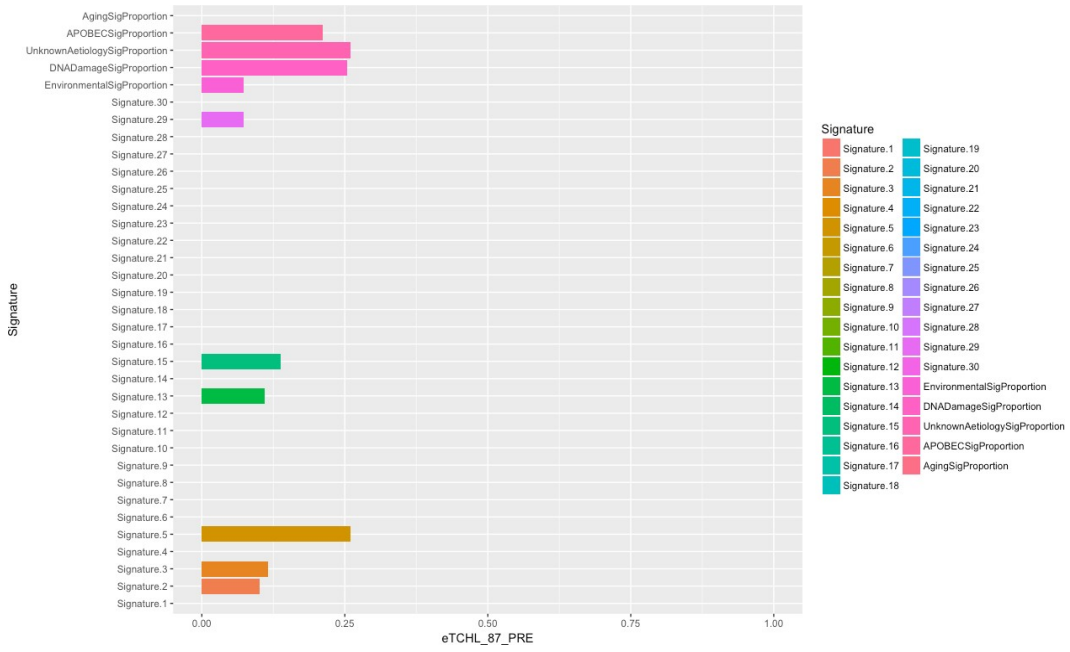
Mutations in the following genes are likely oncogenic: PTPN11, RBBP7, MLLT4, ASPM, PER1, CHD6, COL1A1

Mutations in the following genes are likely tumour suppressor inactivating: RGPR, MLL3, MAP4K3, ARID1A, PSME3, TRAF7, DICER1, RB1

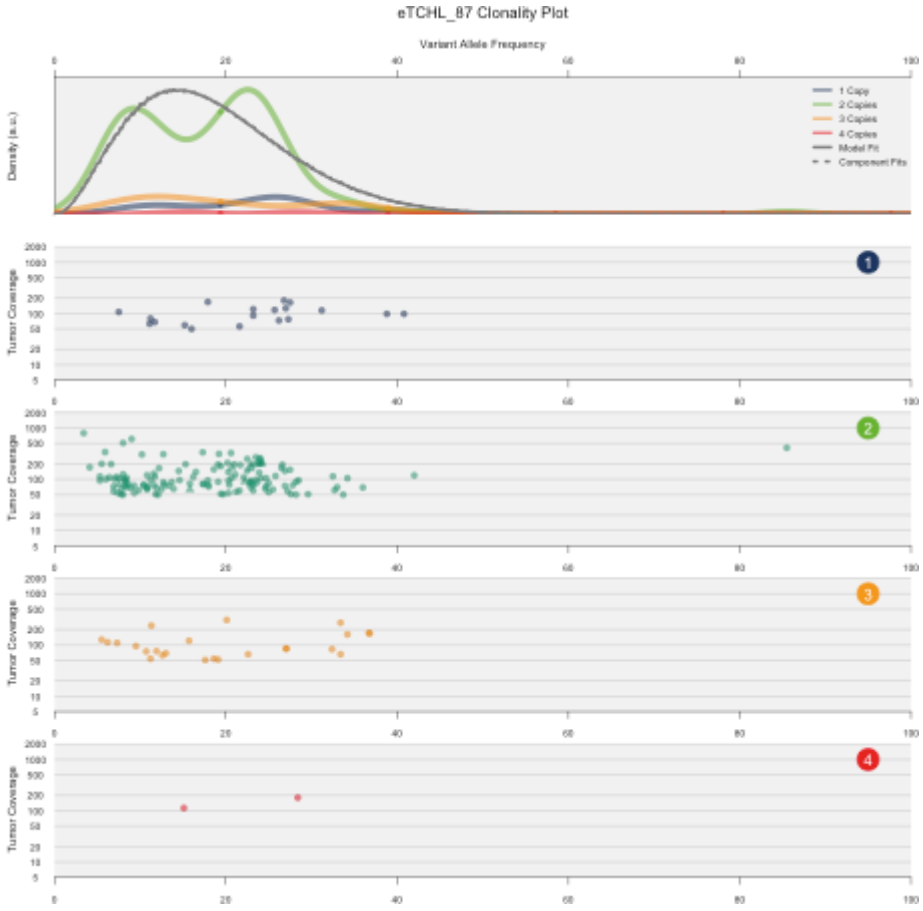
It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: SVEP1, XRCC2, SMURF2, HERC2

TCHL 87

Mutational Signatures



The TCHL 87 Pre-treatment sample shows a mutational spectrum influenced by all signature groups other than the aging signature. As discussed for other samples, the absence of Signature 1 is most likely an error of deconstructSigs rather than a genuine absence of the signature in the sample (see section 1.7). The most prevalent individual signature is Signature 5, a signature of unknown aetiology found in all cancer types and seen in several other samples in this cohort. The next most prevalent signature groupis DNA damage based signatures, comprised of Signature 3 (DSB repair deficiency) and Signature 15 (DNA mismatch repair deficiency), followed by APOBEC related signature (Signatures 2 and 13). Finally, there is a minor contribution by an environmentally associated signature, Signature 29, associated with tobacco chewing in oral cancer. As discussed under other samples with this signature, the sample in question was probably not actually exposed to the process known to cause this signature and the signature being assigned to this sample is probably a quirk of the deconstructSigs algorithm.



Driver analysis

input	gene	cdna	gdna	protein	gene_role	Driver statement
4|119660367|G|A	SEC24D	c.2314C>T	chr4:g.119660367G> A	p.Q772*	LoF	predicted driver: tier 1
17|37627538|G|T	CDK12	c.1453G> T	chr17:g.37627538G>T	p.E485*	LoF	predicted driver: tier 1
17|37627298|G|T	CDK12	c.1213G> T	chr17:g.37627298G>T	p.E405*	LoF	predicted driver: tier 1

17|37627277|G|A	CDK12	c.1192G> A	chr17:g.37627277G> A	p.E398K	LoF	predicted driver: tier 1
17|27807493|C|G	TAOK1	c.557C>G	chr17:g.27807493C> G	p.P186R	ambiguous	predicted driver: tier 1

The TCHL 87 Pre-treatment sample shows 5 predicted driver mutations. TAOK1 shows a predicted driver mutation in a few other samples in the cohort (TCHL 8 Pre-treatment and TCHL 37 Pre-treatment). CDK12 shows a predicted driver mutation in one other

sample in the cohort (TCHL 50 Pre-treatment). SEC24D does not show predicted driver mutations in any other sample in the cohort.

Mutations in the following genes are likely tumour suppressor inactivating: CDK12, SEC24D

It is ambiguous whether mutations in the following genes are oncogenic or tumour suppressor mutations: TAOK1