Chapter 5

The subclonal architecture and life history of a single cancer

5.1 Introduction

In the previous chapters I have described methods to call copy number, infer the subclonal architecture of a tumour and quality control the results. In this chapter I explore what the inferred architecture reveals about the events that took place during a tumour's development and what can be learned about a single cancer from these data. The aim of this chapter is to visibly show what the previously introduced methods do, without a deep understanding of their internal workings.

To this end I have selected a single breast cancer sample of which the copy number profile and subclonal architecture are very clear. The sample however originates from a project that is not yet complete and therefore not described in this thesis. This project comprises the whole genome sequencing of breast cancers from patients in Nigeria with the aim to explore the tumours' subclonal architecture and life history, the background of this project is briefly described in the next section. For the purposes of this chapter I present the selected tumour as 'a cancer' and therefore do not consider its additional unique features, such as the germline context, occurrence rates of breast cancer subtypes and differences in (access to) healthcare when compared to tumours from patients in the Western world.

The sequencing read alignment and variant calling work described in section 5.3 was performed by Jason Pitt.

5.2 Background

Over the last few decades, enormous progress has been made in treatment of breast cancer. In the UK, between 1988 and 2013 mortality has dropped from 60 per 100,000 to below 40, even though incidence rate has gone up from 120 to 170 per 100,000 [1]. However, a big discrepancy remains between women of different ethnic backgrounds. American women of African American ancestry consistently showing lower treatment success, even though survival rates show a similar improvement obtained for American women of European descent (Servick, 2014).

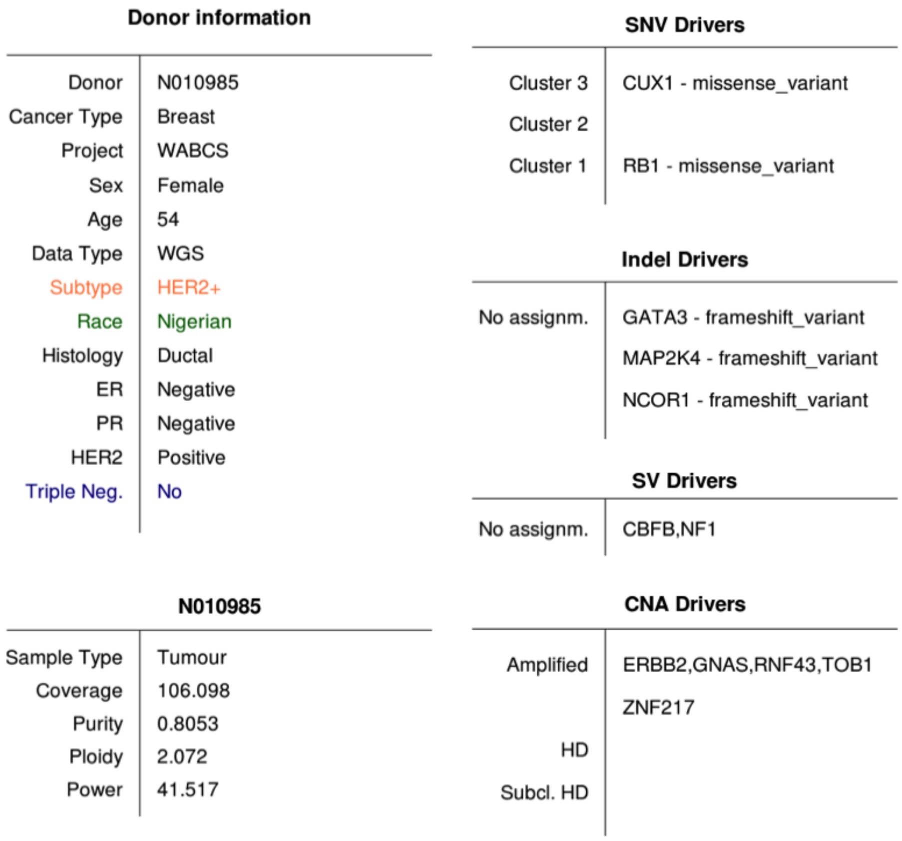
The reasons behind this disparity are thought to be a complex interplay between socio-economic and tumour biology differences (Daly and Olopade, 2015). American women of African ancestry are less likely to be diagnosed with breast cancer, however tumours are diagnosed at an earlier age and at higher tumour stage compared to American women of European descent (Iqbal et al., 2015). Breast tumours are more often of the triple negative subtype (Ray and Polite, Feb) and there is a higher prevalence of *BRCA1* and *BRCA2* germline carriers among African women (Fackenthal et al., 2012). Meanwhile, several social boundaries (Jones et al., 2014) and differences in patterns of referral have been described (Daly and Olopade, 2015), including that African American women with a family history of breast cancer are less likely to undergo genetic counselling (Armstrong et al., 2005).

The West African Breast Cancer Study (WABCS) was set up to further investigate the tumour biology and genetics of breast cancers from western Africa and is aimed to provide a comprehensive overview by sourcing and sequencing tumours (WXS or WGS of DNA and RNA-seq) from West Africa. The study consists of various projects focussing on predisposing germline loci, a landscape of somatic alterations and unveiling patterns tumour evolution. I am part of the tumour evolution project where we aim to describe the life history of breast cancers from Africa and investigate whether there are different patterns of evolution, when compared to those obtained from women in North America. At the point of writing, the study is in progress, with no finalised results. The sample described in this chapter is one of 98 whole genome sequenced tumours that are part of the study and was specifically picked for its clear copy number profile and subclonal architecture to aid the purpose of this chapter.

5.3 Methods

Sample N010985 was resected from a 54 year old patient in Nigeria (Fig. 5.1). Six needle biopsy samples were taken, of which one was prepared for whole genome sequencing. The

[1] https://visual.ons.gov.uk/40-years-of-cancer



Donor information

Donor	N010985
Cancer Type	Breast
Project	WABCS
Sex	Female
Age	54
Data Type	WGS
Subtype	HER2+
Race	Nigerian
Histology	Ductal
ER	Negative
PR	Negative
HER2	Positive
Triple Neg.	No

SNV Drivers

Cluster 3	CUX1 - missense_variant
Cluster 2	
Cluster 1	RB1 - missense_variant

Indel Drivers

No assignm.	GATA3 - frameshift_variant MAP2K4 - frameshift_variant NCOR1 - frameshift_variant

SV Drivers

No assignm.	CBFB,NF1

N010985

Sample Type	Tumour
Coverage	106.098
Purity	0.8053
Ploidy	2.072
Power	41.517

CNA Drivers

Amplified	ERBB2,GNAS,RNF43,TOB1 ZNF217
HD	
Subcl. HD	

Fig. 5.1 General annotations of breast cancer case N010985 (left column) and identified potential drivers (right column). The top left table contains information about the donor, including age, ethnicity, project (WABCS stands for West African Breast Cancer Study), the type of sequencing and the inferred ER/PR/HER2 status. The bottom left table shows statistics about the tumour sample: coverage, purity and ploidy. It also shows the number of reads per chromosome copy (labelled as power), which determines the power to detect subclones (see Chapter 6 for a description of the metric).

tumour biopsy and a blood sample from the same patient were sequenced on an Illumina X10 machine to a coverage of 100x and 30x respectively. Histology of the tumour was examined by pathologists in Nigeria and at the University of Chicago, after which it was classified as a ductal carcinoma. Accompanying RNA-seq data was used to infer that the tumour ER-negative and HER2-positive.

After passing initial sequencing quality control metrics the obtained reads were aligned to the GRCh37 reference genome using BWA (Li and Durbin, 2009), after which SNV calling was performed using Strelka (Saunders et al., 2012) and Mutect (Cibulskis et al., 2013), indel calling using Strelka and SVs were obtained by applying Delly (Rausch et al., 2012) and Lumpy (Layer et al., 2014). To obtain reliable SNV and SV calls the results from the two methods were intersected and filtered by an unmatched normal panel. For indels only the filtering by panel was applied.

5.4 Subclonal architecture

The sequencing yielded 18,813 SNVs, 382 indels and 335 SVs. A copy number profile was fit using the Battenberg algorithm, which yields a relatively quiet profile with a ploidy just over 2 and a purity of 81% (Fig. 5.2a). The tumour consists of a clone with an estimated 9,794 SNVs and two subclones with 2,792 and 5,530 SNVs (Fig. 5.2b). At the time of writing the VAF adjustment pipeline for indels is not ready, hence indels are not assigned to mutation clusters.

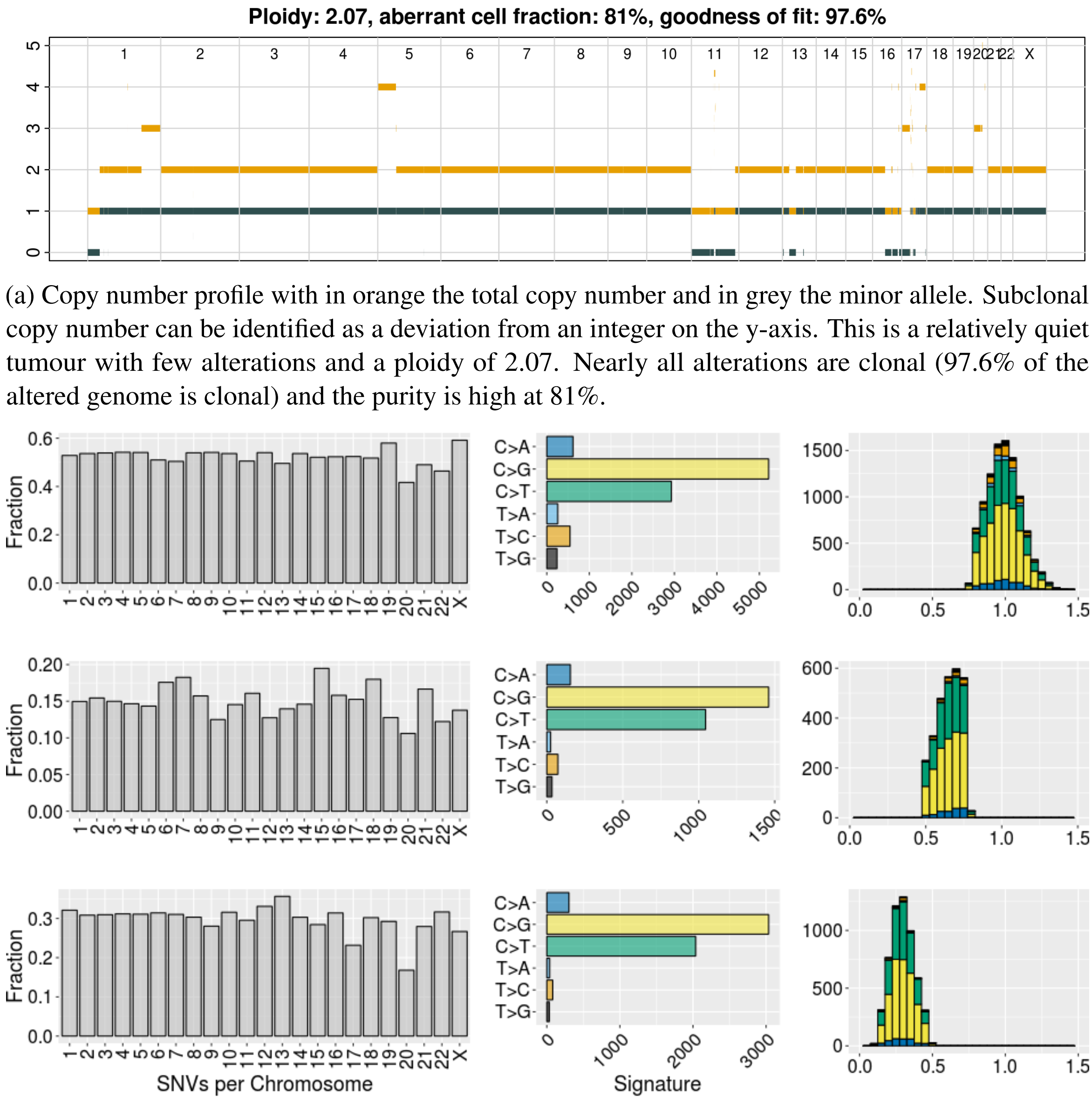
The mutations (SNVs, indels, SVs, amplifications and homozygous deletions) were intersected with a putative list of 149 genes thought to be involved in breast cancer development (Fig. 5.1). This list consists of genes taken as the top hits reported in (Nik-Zainal et al., 2016) and all genes in which a driver was found in a breast cancer in the ICGC pan-cancer dataset (Sabarinathan et al., 2017).

This analysis yields a clonal missense variant in *CUX1*, which is carried by 1 chromosome copy in a balanced copy number region and a subclonal missense variant in *RB1*, which falls in a region of clonal LOH where only a single copy of the locus is available. It's unclear whether the RB1 mutation deactivates the remaining copy of *RB1*.

Analysis of indels yields deletions of 11 bases in *MAP2K4* and 14 in *NCOR1* in a region of LOH where there are 3 copies of one allele and an insertion of a single base into *GATA3* in a region of balanced copy number. Raw CCF values for these variants are 1.00, 1.07 and 1.05 respectively, all three are therefore most likely clonal. The *MAP2K4* deletion is reported by 104 out of 124 reads and the *NCOR1* deletion by 132 out of 159, which suggests the mutations are clonal and carried by multiple chromosome copies.

Amplifications and homozygous deletions were obtained from the copy number data by selecting focal segments (< 1Mb). A segment is classified as an amplification when the total copy number exceeds 2*ploidy+1 and as a homozygous deletion when both alleles have been lost (clonally or subclonally). This results in three genes on chromosome 17 (*ERBB2*, *TOB1* and *RNF43*) and two on chromosome 20 (*ZNF217* and *GNAS*) being classified as amplified. The *ERBB2* is also known as *HER2* and is a primary driver of this tumour.

A copy number or SV breakpoint is found within the *CBFB* and *NF1* genes (Fig. 5.3). *CBFB* contains a copy number breakpoint where the first 3 exons are deleted. No other disrupting event is found, which suggests one copy of *CBFB* remains intact. *NF1* contains multiple breakpoints, which results in gaps in the local copy number as segments in Bat-

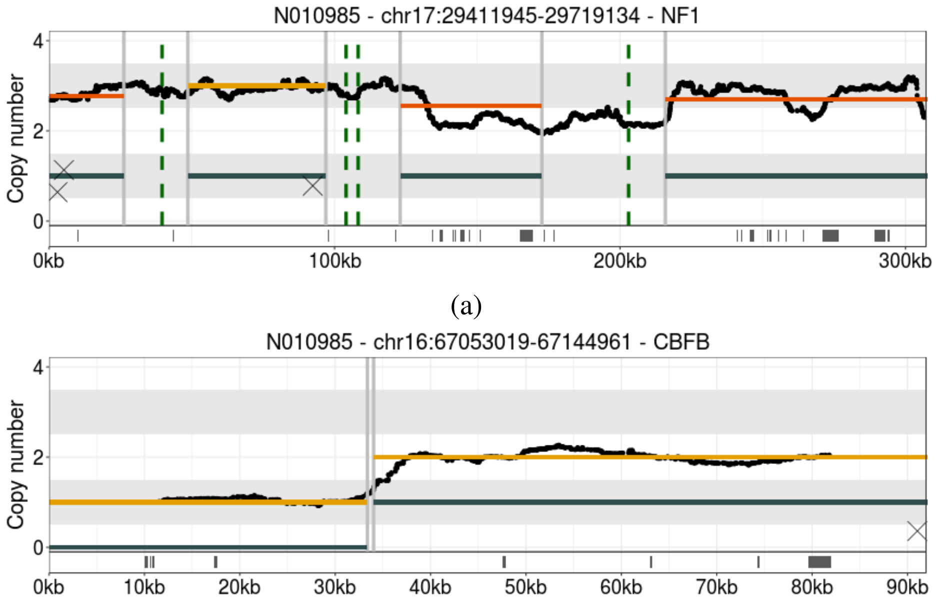


(a) Copy number profile with in orange the total copy number and in grey the minor allele. Subclonal copy number can be identified as a deviation from an integer on the y-axis. This is a relatively quiet tumour with few alterations and a ploidy of 2.07. Nearly all alterations are clonal (97.6% of the altered genome is clonal) and the purity is high at 81%.

(b) Summary of the subclonal architecture with a row for each mutation cluster identified. The left column shows the number of SNVs per chromosome, the middle column counts for each of the six possible base substitutions and the right column the raw CCF values of the SNVs assigned to the cluster. This tumour consists of three mutation clusters, a clone and two subclones. All three contain a high number of C>G and C>T mutation types.

Fig. 5.2 Subclonal architecture and copy number profile of N010985.

tenberg start and end at a germline heterozygous SNP. The copy number fit suggests there are three copies of *NF1*, of which one contains a deletion of exons 6-16. The SVs suggest the whole gene up to 200kb was duplicated, and both regions marked with subclonal copy number are supported by deletion calls. Regardless, given the copy number, there is at least



(a)

(b)

Fig. 5.3 Detailed figures showing the mutations measured in *NF1* and *CBFB*. The figures contain the copy number profile in orange and grey (total copy number and minor allele), raw total copy number calculated from the coverage in the background in black, SNVs as X-es (grey when non-coding, black when coding, there are no coding mutations found in either gene), copy number breakpoints as grey vertical lines and SV breakpoints as green dashed lines. Below the mutations is a track that shows the exons of the default transcript from Ensembl.

one working copy of *NF1* remaining as no other disrupting events have been found. *NF1* is unlikely to be a driver of this tumour as it is a tumour suppressor gene (Cichowski and Jacks, 2001). One copy of *CBFB* remains in tact, and without evidence of a fusion with *MYH11* or deactivation of *RUNX1* this gene is also unlikely to be a driver (Banerji et al., 2012).

5.5 Mutational Signatures

Mutational signature analysis was restricted to nine selected signatures that have been called de novo by Jason Pitt on a large set of breast cancer exomes from Nigerian patients, also part of the WABCS project (Pitt et al., 2018). The signatures found have been matched against the COSMIC signatures to determine the labels (Forbes et al., 2017). I subsequently quantified the activity of each of the nine signatures using the MutationalPatterns R package (Blokzijl et al., 2018).

The signatures reveal strong APOBEC activity in the clone and both subclones (Fig. 5.4). There is a larger relative contribution of the C>T APOBEC signature in both subclones when compared to the clone, which may be an indication that the C>T signature has a later onset in this tumour or that the activity rate of the two APOBEC signatures varies. The other seven signatures do not contribute substantially and their detected presence in low proportions could be noise.

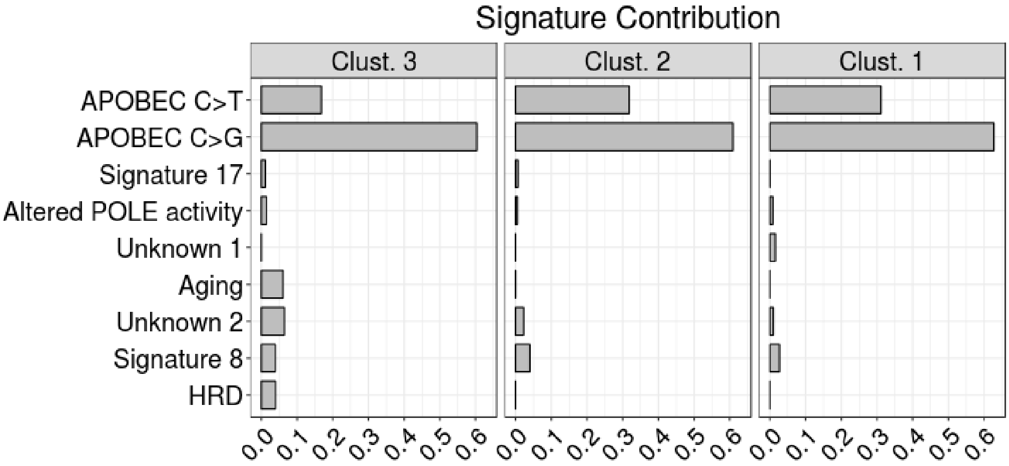


Fig. 5.4 Mutational signature analysis of SNVs assigned to the three mutation clusters reveals steady APOBEC C>G and C>T signature activity.

Kataegis events in N010985

Region	Chromosome	Size (bp)	Num SNVs	C>A	C>G	C>T	T>G
1	2	12356	23	3	10	10	0
2	17	1063	11	2	0	7	2
3	19	6079	13	1	6	6	0
4	20	10666	35	3	22	10	0

Table 5.1 Four regions containing kataegis have been identified in N010985. All four regions contain a large proportion of C>G and C>T mutations associated with APOBEC activity. Regions 1 and 3 contain an equal number of C>Gs and C>Ts, while regions 2 and 4 show an imbalance between the two types of substitutions. No T>A and T>C substitutions have been identified in any of the regions.

5.6 Kataegis

APOBEC activity is associated with local hypermutation, known as kataegis (Nik-Zainal et al., 2012b). Regions containing kataegis are obtained by first segmenting the intermutational

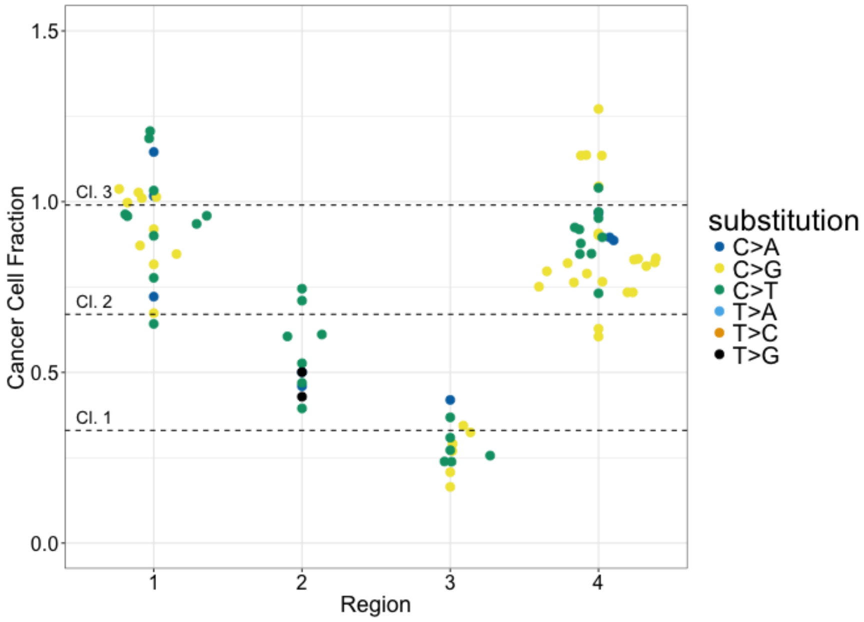


Fig. 5.5 Overview of four kataegis regions found in N010985. Regions 2 and 3 appear to be subclonal, while regions 1 and 4 may consist of multiple kataegis events. All four regions predominantly contain SNVs in APOBEC C>G and C>T contexts.

distance using (i.e. grouping mutations in stretches of similar distance) and subsequently selecting regions with a consistent short distance. The distance threshold is set depending on the mutation rate of the tumour: it must be below an average of 100 base-pair in tumours with over 90,000 SNVs, 250 in tumours with between 50,000 and 90,000 SNVs, 500 when between 10,000 and 50,000 and the threshold is set to 1,000 base-pair below 10,000 SNVs.

I identify four regions are with local hypermutation in this tumour (Table 5.1). All four regions contain SNVs that can be explained as the result of APOBEC signature activity. Two regions show an imbalance between the number of C>G and C>T substitutions, with region 2 containing no C>Gs and region 4 containing more than double the number of C>Gs. This suggest that both APOBEC signatures can independently generate kataegis events and that some of the regions identified may be a combination of multiple events.

SNVs in kataegis regions are routinely excluded from subclonal architecture inference. The localised hypermutation causes reads to contain multiple variants, which may impact the read alignment quality and result in more variable VAFs. However, analysis of the raw CCF estimates of the SNVs in the four regions suggests regions 2 and 3 contain subclonal kataegis events (Fig. 5.5). There appears to be separation between C>G and C>T SNVs in region 4, with the C>G SNVs possibly belonging to cluster 2.

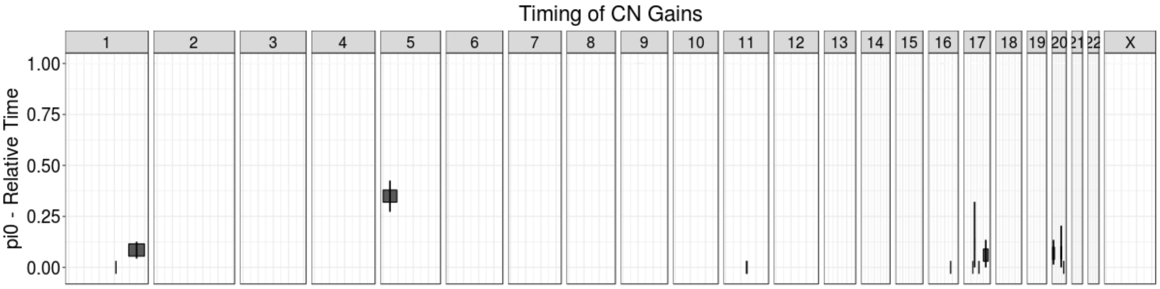


Fig. 5.6 Timing of gains analysis showing all gains with at least 10 SNVs. The y-axis represents the relative ordering, with a low value meaning the gain occurred relatively early according to the mutation data on the segment. The thin vertical lines are confidence intervals.

5.7 The life history of N010985

Analysis of the timing of gains (Fig. 5.6) reveals that gains on chromosome 17, 20 and 1q are early, followed by a gain of chromsome arm 5p. This analysis utilises the ratio of SNVs that are on multiple copies and on a single chromosome copy, where a low ratio indicates the gain is early. Timing of gains was performed using cancerTiming (Purdom et al., 2013), with segments restricted to those with 10 or more SNVs.

Previously I have split events measured in this tumour into clonal and subclonal. The timing of gains analysis allows for splitting clonal events into clonal early, late or undefined. Meanwhile, potential driver mutations can also be classified by taking into account the multiplicity.

The tumour's life history can now be compiled from the accumulated evidence (Fig. 5.7). It starts with deletions in *NCOR1* and *MAP2K4*, which are subsequently gained. The loss of 17p is most likely also early as it deletes the remaining intact copy of both genes, but the loss cannot be timed. Gains of chromosomes 1q and 20 are also early. These events appear in the first 150 measured SNVs. Mutational signature analysis suggests both APOBEC mutational signatures are already active.

Then follows a range of events that cannot be accurately timed. This phase contains an SNV in *CUX1* and an insertion into *GATA3*, in both cases on one of two available copies, and a loss of one copy of *RB1*. It also contains amplifications of *ERBB2* and other genes on chromsome 17p and 20q and losses of 1p, 11 and 16q. This period represents a large proportion of the tumour's life history consisting of 9,647 SNVs. APOBEC signatures remain constantly active. Multiple kataegis events with an APOBEC context are observed.

Finally, subclonally there is the second deactivating event for *RB1*, which suggests *RB1* is the driver of a subclonal expansion. Also observed are further gains of segments on chromosomes 11 and 17 and multiple kataegis events associated with APOBEC activity.

The subclonal architecture leaves two possible tree representations, a branching and a linear tree (Fig. 5.8). Phasing of SNVs did not yield a mutually exclusive pair that could have ruled out the linear tree. It is therefore not possible to resolve the tree topology.

This life history can be put in perspective by comparing it to the combined life history of all breast cancers (which is available as Appendix B, as it is part of the supplementary figures of Gerstung et al. (2017)). N010985 does not have a whole genome duplication. PCAWG tumours without a genome doubling event typically contain early gains, which is also observed in N010985 (Appendix B Fig. A).

The overall breast cancer life history (Appendix B Fig. B) shows that loss of 17p and 13q (*RB1*) are indeed most likely early. The gain of 1q and driver mutations in *GATA3* are typically early, but later than losses of 17p and 13q. Mutational signature analysis of the PCAWG breast cancers suggests that APOBEC activity is highly variable between cancers, with some tumours showing high early or late activity, while in others APOBEC activity remains constant (signatures 2+13 in Appendix B Fig. D/E).

These findings highlight what a subclonal reconstruction can tell about the life history of a single cancer.

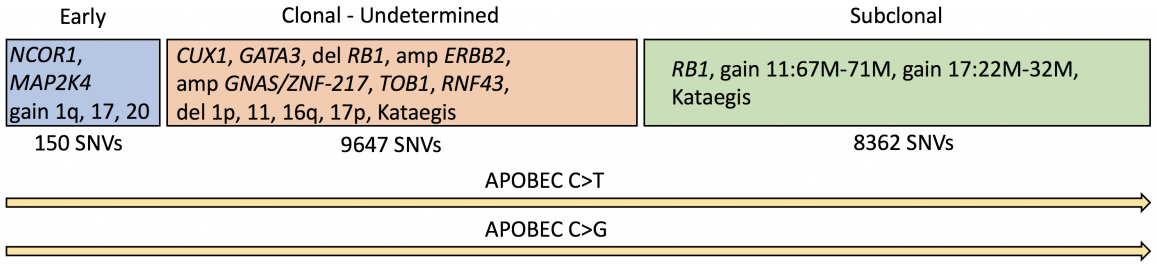

Early
Clonal - Undetermined
Subclonal
NCOR1, MAP2K4 gain 1q, 17, 20
CUX1, GATA3, del RB1, amp ERBB2, amp GNAS/ZNF-217, TOB1, RNF43, del 1p, 11, 16q, 17p, Kataegis
RB1, gain 11:67M-71M, gain 17:22M-32M, Kataegis
150 SNVs
9647 SNVs
8362 SNVs
APOBEC C>T
APOBEC C>G

Fig. 5.7 The compiled life history for N010985. Early drivers are *NCOR1* and *MAP2K4* (blue square). A range of events cannot be timed and could be early or late (red square). An *RB1* mutation and two gains are subclonal (green square). APOBEC mutational signatures are active throughout the life history of this tumour.

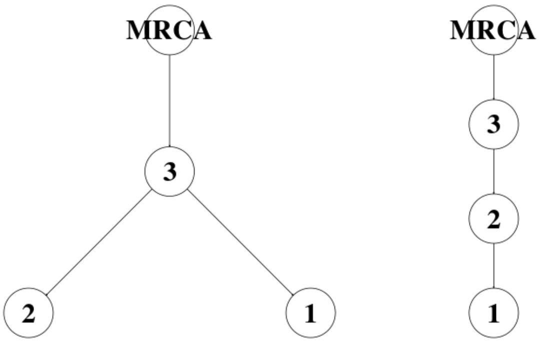


Fig. 5.8 Possible trees reconstructed from the subclonal architecture. The numbers on each node refer to the cluster number, cluster 3 is the clone. The cluster locations provide the option of either a linear or a branching tree. Mutation phasing information did not provide evidence to rule out one of the scenarios.