Breast Tumors Maintain a Reservoir of Subclonal Diversity During Primary Expansion

Supplementary Information

Conterno Minussi, Nicholson, Ye et al. 2021

Table of Contents

SUPPLEMENTARY INFORMATION GUIDE

Supplementary Table 1 - Relevant clinical information for the eight TNBC tumors and four TNBC cell lines that were analyzed in this study.

Supplementary Table 2 – Exome sequencing mutational information from bulk DNA exome of the 8 TNBC tumors.

Supplementary Table 3 - List of dual barcodes from ACT protocol.

Supplementary Methods – Extensive description of the methodology used for the mathematical modeling section.

TNBC Tumors

ID	age	ER	PR	HER2	grade	pathology	specimen size (cm)	lymph	treatment	ploidy	cells	reads per cell	%dup	mean bin count
TN1	60	<1%	<1%	neg	3	DCIS	1.4x1.2x1.0	neg	AC	3.45	1100	954367	9.77	62.20
TN2	79	<1%	<1%	neg	3	IDC	1.1x1.0x0.6	neg	untreated	3.03	1024	1446404	9.72	90.10
TN3	71	<1%	<1%	neg	3	IDC/DCIS	2.0x1.8x1.1	neg	untreated	3.44	1101	938037	7.68	63.40
TN4	53	<1%	<1%	neg	3	IDC/DCIS	0.9x0.9x0.7	neg	untreated	3.76	1307	894304	7.92	55.70
TN5	37	<1%	<1%	neg	3	IDC/DCIS	1.2x0.4x0.8	neg	untreated	2.65	1238	919209	8.12	60.90
TN6	50	<1%	<1%	neg	3	IDC	1.4x0.6x0.8	neg	untreated	3.17	1378	1117188	12.40	69.10
TN7	47	<1%	<1%	neg	3	IDC	0.8x1.2x0.5	neg	untreated	3.15	1393	1321931	14.00	74.10
TN8	74	<1%	<1%	neg	3	IDC/DCIS	0.6x0.6x0.9	neg	untreated	3.95	1224	894949	10.4	54.4

Cell Lines

ID	age	ER	PR	HER2	ploidy	cells	reads per cell	%dup	bin count
MDA-MB-231 EX-2	NA	<1%	<1%	neg	2.41	897	915415	6.98	66.99
MDA-MB-231 EX-1	NA	<1%	<1%	neg	2.41	995	810439	6.73	61.63
MDA-MB-231 PARENTAL	NA	<1%	<1%	neg	2.41	820	1172793	9.68	89.57
MDA-MB-157	NA	<1%	<1%	neg	2.55	1210	94240	8.83	64.4
BT-20	NA	<1%	<1%	neg	2.7	1231	857024	9.82	54.9
MDA-MB-453	NA	<1%	<1%	neg	4.17	1260	912792	7.66	62.1

Supplementary Table 1 – Clinical Tumor Information and Sequencing Metrics
This table lists relevant clinical information for the eight TNBC tumors and four TNBC cell lines that were analyzed in this study. Clinical information listed includes patient identifier, patient age, estrogen receptor and progesterone receptor IHC status, Her2 cytogenetic FISH status, tumor grade, pathological classification, tumor specimen size, lymph node status, prior treatment status and drug type, FACS mean ploidy of the aneuploid cell population, number of single cells sequenced, mean number of reads per cell, mean number of duplicate reads per cell and mean number of median bin counts in 220kb genomic intervals.

Patient	somatic mutations	nonsynonymous	clonal	subclonal	TP53 mutation class	TP53 mutation
TN1	61	43	61	0	SNV	c.C346T,p.R116W
TN2	39	28	34	4	SNV	c.C437G,p.P146R
TN3	105	73	85	17	SNV	c.A319G,p.N107D
TN4	28	17	25	2	SNV	c.G460A,p.E154K
TN5	106	80	14	91	SNV	c.G649T,p.E217X
TN6	131	94	126	4	Indel	del.7579356:GACGGA
TN7	100	69	33	65	SNV	c.C241T,p.R81X
TN8	1728	1173	1294	387	SNV	c.T188C:p.I63T

Supplementary Table 2 – Bulk DNA Exome Sequencing Mutations
This tables lists information about the exome sequencing data of the eight TNBC tumors including the number of somatic mutations, number of nonsynonymous mutations, number of clonal and subclonal mutations identified and information on the *TP53* mutations detected in each tumor.

i7 primer		i5 primer	
N703	AGGCAGAA	S504	TCTACTCT
N704	TCCTGAGC	S505	CTCCTTAC
N705	GGACTCCT	S506	TATGCAGT
N706	TAGGCATG	S507	TACTCCTT
N707	CTCTCTAC	S508	AGGCTTAG
N708	CAGAGAGG	S510	ATTAGACG
N709	GCTACGCT	S511	CGGAGAGA
N710	CGAGGCTG	S513	CTAGTCGA
N711	AAGAGGCA	S515	AGCTAGAA
N712	GTAGAGGA	S516	ACTCTAGG
N714	GCTCATGA	S517	TCTTACGC
N715	ATCTCAGG	S518	CTTAATAG
N716	ACTCGCTA	S520	ATAGCCTT
N718	GGAGCTAC	S521	TAAGGCTC
N719	GCGTAGTA	S522	TCGCATAA
N720	CGGAGCCT	S525	AACGCTTA
N722	ATGCGCAG	S526	AAGACGGA
N723	TAGCGCTC	S527	AAGGTACA
N724	ACTGAGCG	S528	ACACAGAA
N726	CCTAAGAC	S529	ACAGCAGA
N727	CGATCAGT	S530	ACCTCCAA
N728	TGCAGCTA	S531	ACGCTCGA
N731	AACGTGAT	S532	ACGTATCA
N732	AAACATCG	S533	ACTATGCA
N733	ATGCCTAA	S534	AGAGTCAA
N734	AGTGGTCA	S535	AGATCGCA
N735	ACCACTGT	S536	AGCAGGAA
N736	ACATTGGC	S537	AGTCACTA
N737	CAGATCTG	S538	ATCCTGTA
N738	CATCAAGT	S539	ATTGAGGA
N739	CGCTGATC	S540	CAACCACA
N740	ACAAGCTA	S541	GACTAGTA
		S542	CAATGGAA
		S543	CACTTCGA
		S544	CAGCGTTA
		S545	CATACCAA
		S546	CCAGTTCA
		S547	CCGAAGTA
		S548	CCGTGAGA

Supplementary Table 3 – ACT Dual Barcoding Primer Sequences
This table lists the dual barcodes sequences for the sixteen unique N7XX barcodes and 24 unique S5XX barcodes that are used for the ACT protocol (see Acoustic Cell Tagmentation Procedure, Methods).

Breast Tumors Maintain a Reservoir of Subclonal Diversity During Primary Expansion: Mathematical modeling

This document provides the mathematical details of our study. The workflow is summarized in Fig. 1. In brief, we used a mathematical modeling approach to investigate whether the patient data can more likely be explained by a stochastic model in which individual copy number alterations arise gradually during tumor evolution at a constant rate ('the gradual model') or at a higher rate during an early phase of transient genomic instability followed by a phase in which alterations emerge at the baseline rate ('the transient instability model'). The computational pipeline starts with the single cell copy number sequencing data, which is first segmented to obtain shared breakpoints. This approach enables us to determine the breakpoint frequencies, i.e. the fraction of cells harboring a specific breakpoint, and in turn to obtain the number of breakpoints at a given frequency – the breakpoint frequency spectrum. The frequency spectrum is then examined using a maximum likelihood framework, incorporating breakpoint detection error, to determine which of the two stochastic models are more likely to explain the data for each individual patient. In the sections below we discuss each step of the pipeline in detail.

1 Model and approximations

We developed a computational framework capable of detecting a transient period of elevated genomic instability in the patient single cell copy number data. The accumulation of structural variants is modeled as a branching process [1] – a stochastic process model in which individual cells can divide with or without accumulating a new set of breakpoints, or die.

As copy number alterations may be subject to natural selection [2], in the context of this model, cells that acquire alterations also obtain heritable changes to their fitness (i.e. reproductive ability), with the changes drawn from a flexible fitness distribution. However, simulation-based inference using such a model when the final number of simulated cells matches those detected in patients ($\approx 3 \times 10^9$ cells) is computationally infeasible. To circumvent this issue, we developed the following strategy (Fig. 1): we considered a reduced fitness distribution, according to which mutations are either lethal or selectively neutral, which is analytically tractable so that analytical expressions can be derived. Using this approximate model, we then developed an inference scheme, which was applied to simulations including a more complex fitness distribution in order to evaluate the difference in model predictions using the reduced and full fitness distribution assumptions. Since the simulations using the full distribution are computationally expensive, they were only simulated to 10^5 cells. This approach demonstrated that the biological conclusions obtained via the analytic model are robust in the presence of a complex fitness distribution (Fig. 2). We proceed to discuss the 'full model' before providing details on the approximations.

In the full model, a tumor starts from a single cell with division rate b and death rate d. This originating cell possesses k_0 copy number alterations (CNAs). The tumor grows until a final observation size N. Upon a cell dividing, one of the daughter cells acquires a CNA with probability μ per cell division. To account for the potential selective effect of CNAs [2], each CNA heritably alters the division rate such that the birth rate of the cell acquiring the CNA changes to $b \mapsto b + \Delta b$, where Δb follows a double exponential distribution (a.k.a. the Laplace distribution) with parameter α ($\Delta b > 0$ with probability 1/2, and $\mathbb{E}[|\Delta b|] = \alpha^{-1}$). The double exponential fitness distribution has previously been used in cancer modeling [3, 4] and further justification for using this distribution can be found in ref. [5]. To explore transient instability, we allow for

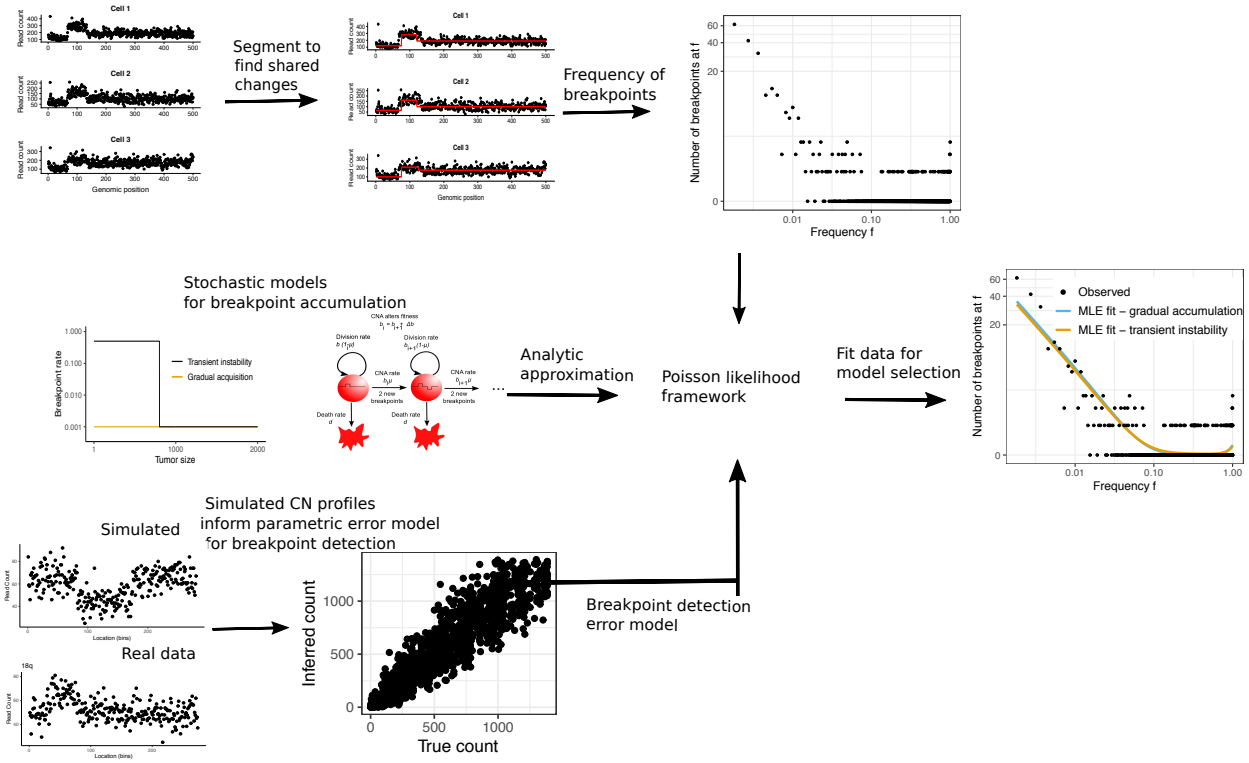


Figure 1: **Schematic of workflow**. Here we aim to determine whether a stochastic model of breakpoint accumulation with an early phase of genomic instability provides a superior explanation of the patient data than a model assuming a constant rate of breakpoint accumulation. To this end, we first segment the single cell copy number data to find shared breakpoints, thus determining the frequency of breakpoints in the population. We then determine the frequency spectrum (i.e. number of breakpoints present at greater than a given frequency), which is analyzed using a likelihood framework, incorporating breakpoint detection errors, to examine whether the gradual evolution or transient instability models have a higher likelihood of explaining the data.

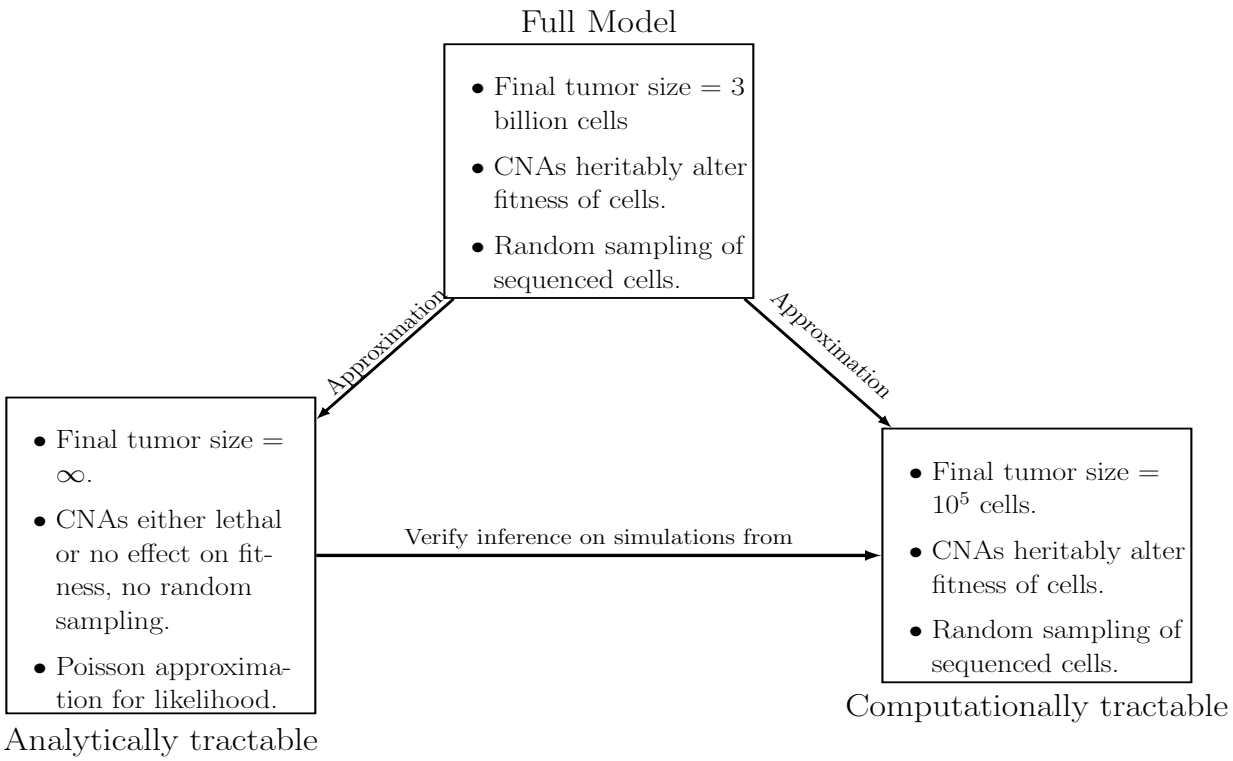


Figure 2: **Model approximations.** Due to computational intractability, we approximate the 'full model' in two ways. Biological conclusions remain robust under both approximations.

the possibility that μ is a function of the tumor size, and in particular we assume that there exists N_c such that

$$\mu = \begin{cases} \mu_1 & \text{until population size exceeds } N_c \text{ for the first time} \\ \mu_2 & \text{thereafter.} \end{cases}$$

The process is depicted schematically in Fig. 3. We make the infinite sites assumption (each new CNA results in two new unique breakpoints in the copy number profiles) and provide justification for doing so in Section 8. Our aim is to detect whether there is a period of transient instability $\mu_1 > \mu_2$, which we contrast with gradual accumulation $\mu_1 = \mu_2$. Sample realizations for both cases are displayed as Muller plots in Fig. 4, which were generated using the package ggmuller [6].

Our stochastic model starts from a single cell initiating an exponentially expanding population. This cell is defined by possessing the clonal structural variants, which may have arisen during an event previously termed 'the punctuated burst' [4]. From then on, two different scenarios can occur: 1) Immediately after the punctuated burst, population expansion occurs and genomic instability remains resulting in subclonal structural variants. 2) After the punctuated burst, the clonal lineage persists for many generations. At a later point, population expansion begins in tandem with an increase in genomic instability. If scenario 1) holds true, then the stochastic model starts at the punctuated burst. If scenario 2) holds true, the model starts at the population expansion.

2 Analytical models and expected frequency spectra

To derive analytical results for the model outlined above, we now consider situations in which CNAs do not affect the division or death rates of cells ($\Pr(\Delta b = 0) = 1$). Subsequent results also hold in the scenario in

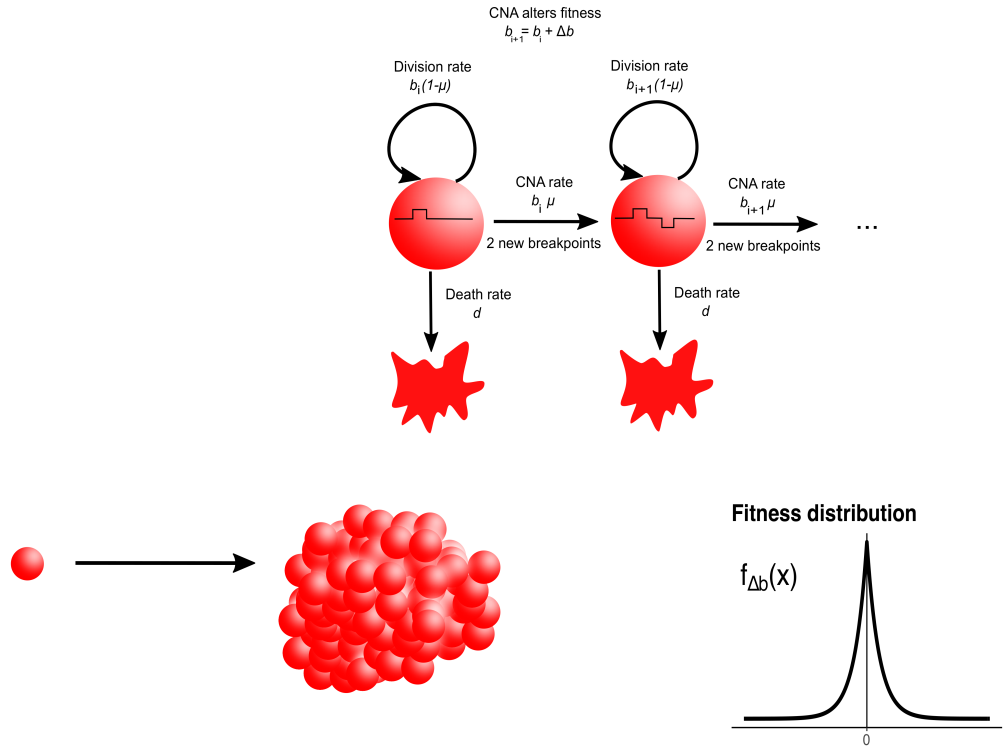


Figure 3: **Model illustration**. We model the acquisition of CNAs during tumor growth. To assess whether a period of transient instability exists early in tumorigenesis we suppose that the probability of acquiring a CNA, μ, is dependent on the tumor size.

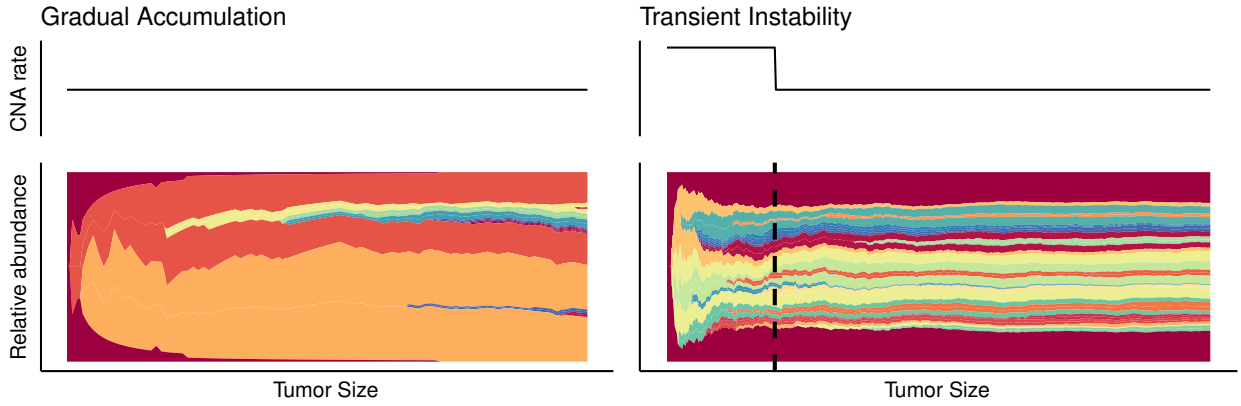


Figure 4: **CNAs are acquired in a period of transient instability or gradually during tumor growth**. The panels display sample realizations of the model with gradual CNA accumulation (left), or during a period of transient instability. Realizations were obtained by exact simulation of the model with the following parameters: gradual accumulation model: $N = 100$, $b = 1.5$, $d = 0.3$, $\mu_1 = 0.08$, $\alpha = 10^4$; transient instability model: $N = 500$, $\mu_1 = 0.2$, $\mu_2 = 0.01$, $N_c = 100$, and all other parameters were the same as for the gradual accumulation model. Simulations were displayed as Muller plots using the package ggmuller. The relative abundance displayed corresponds to the last time the population achieved the x-axis size, which is required as the population size can be non-monotone in time.

which a CNA is instantly lethal with probability u, with the mapping $\mu \mapsto \mu(1-u)$. Currently we focus on the scenario in which there is one change in the rate of CNA accumulation, μ. The more general setting with multiple CNA rate changes was not found to give increased explanatory power (see Section 6).

Any new CNA emerges in a daughter cell resulting from a division event. The lineage initiated by this daughter cell may eventually become extinct with probability $\delta = d/b$ [1]. CNAs emerging in daughter cells whose lineage goes extinct are unobservable and we therefore only consider surviving CNAs; the term CNA henceforth only refers to these. Let the total number of unique CNAs when the tumor is diagnosed be K_N. We consider the site frequency spectrum:

$$F_N(x) = \sum_{i=1}^{K_N} 1(\text{CNA } i \text{ is at a frequency} \geq x\ ) + k_0$$

and in particular its expectation for large N,

$$f(x) \approx \mathbb{E}[F_N(x)] \text{ for } N \gg 1.$$

The appearance of k_0 in Eq. 2 is because those CNAs which are present in the founding cell will always be at frequency 1. At places, our derivation uses formal arguments, and thus we term our statements 'results'.

In Section 9 we provide a derivation which shows that the expected site frequency spectrum is well approximated by

Result 1.

$$f(x) = \mu_1 \frac{r(x) - r(x)^{N_c+1}}{x(1-\delta)} + \mu_2 \frac{r(x)^{N_c+1}}{x(1-\delta)} + k_0, \tag{1}$$

where $r(x) = 1 - x(1-\delta)$. *The expected number of clonal CNAs is thus* $f(1) = \frac{\mu_1\delta + \delta^{N_c+1}(\mu_2-\mu_1)}{1-\delta} + k_0$.

The number of CNAs present at frequencies in (a, b) is $f(a) - f(b)$. While $f(x)$ concerns the expected frequency spectrum, it well describes the shapes of the frequency spectra obtained by individual realizations as shown in Fig. 7. For these simulations, we used the package SIApopr [7]. To obtain an intuitive understanding of the system, we also provide the following cruder approximation to $f(x)$ (with derivation again in Section 9). Let $x_1 = \frac{\mu_2}{\mu_1(1-\delta)N_c}$ and $x_2 = \frac{1}{(1-\delta)N_c}$, then $f(x)$ can be crudely approximated by

$$f(x) \approx f_{\text{crude}}(x) = \begin{cases} N_c(\mu_1 - \mu_2) + \frac{\mu_2}{x(1-\delta)} + k_0 & x < x_1, \\ \mu_1 N_c + k_0 & x_1 \leq x < x_2, \\ \frac{\mu_1}{x(1-\delta)} + k_0 & x_2 \leq x \leq 1. \end{cases} \tag{2}$$

Comparisons of $f(x)$ and $f_{\text{crude}}(x)$ are shown in Fig. 5. Thus $f(x)$ has three main regimes and to be able to distinguish between the cases of $\mu_1 > \mu_2$ and $\mu_1 = \mu_2$, we require the detection window of $[x_1, x_2)$ to be sufficiently large. The frequency spectrum is often plotted on a doubly logarithmic scale; therefore, note that with a logarithmic base-10 x-axis, the detection window is of size $\log_{10}(x_2/x_1) = \log_{10}(\mu_1/\mu_2)$. If either $N_c \to \infty$ (causing $x_1, x_2 \to 0$), or $\mu_1 = \mu_2$, we collapse to the case of no change in the rate of CNA accumulation.

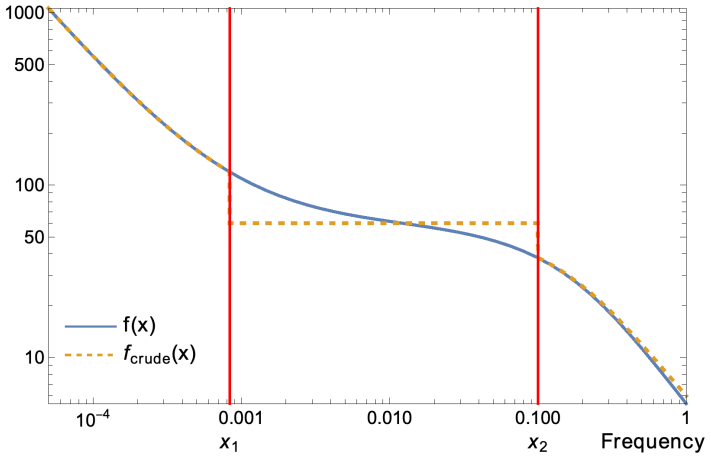


Figure 5: **CNA accumulation rate change has a window of detection**. Comparison of $f(x)$ with $f_{\text{crude}}(x)$. Observe that the general form of $f_{\text{crude}}(x)$ holds. Red vertical lines indicate the region where the lack of intermediate frequency CNAs is apparent; this window is of size $\log_{10}(\mu_1/\mu_2)$. Parameters are $\mu_1 = .6$, $\mu_2 = 5 \times 10^{-3}$, $\delta = 0.9$, $N_c = 100$, $k_0 = 0$.

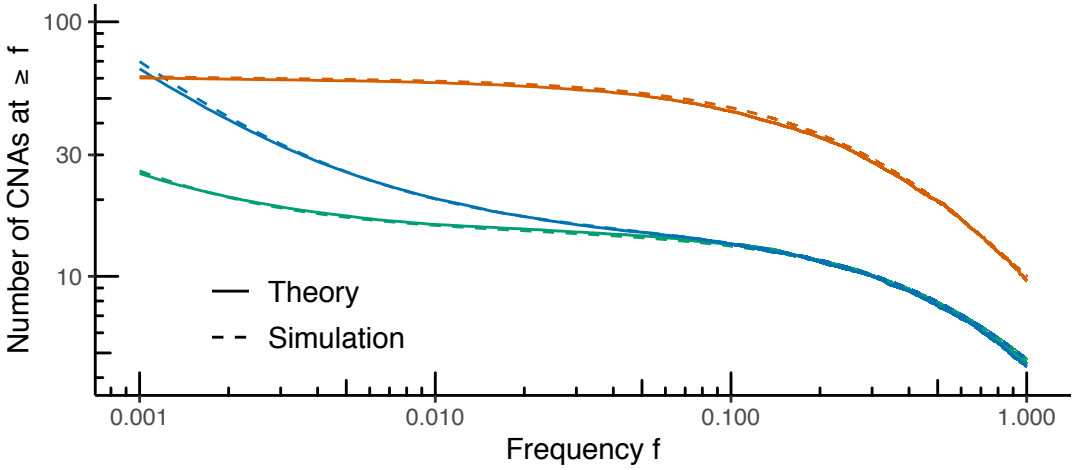


Figure 6: **Expected frequency spectrum: simulations vs analytic formula**. Comparison of the expected site frequency spectrum obtained from simulation and theory, Eq. (1). Parameters (common) $N = 10^5$, $b = 1.1$, $d = 1$ (green) $\mu_1 = 0.5$, $\mu_2 = 0.001$, $N_c = 30$ (blue) $\mu_1 = 0.5$, $\mu_2 = 0.005$, $N_c = 30$ (red) $\mu_1 = 1$, $\mu_2 = 10^{-4}$, $N_c = 60$. For simulations the fitness distribution parameter was $\alpha = 10^3$. Simulated results are averaged over 500 realizations.

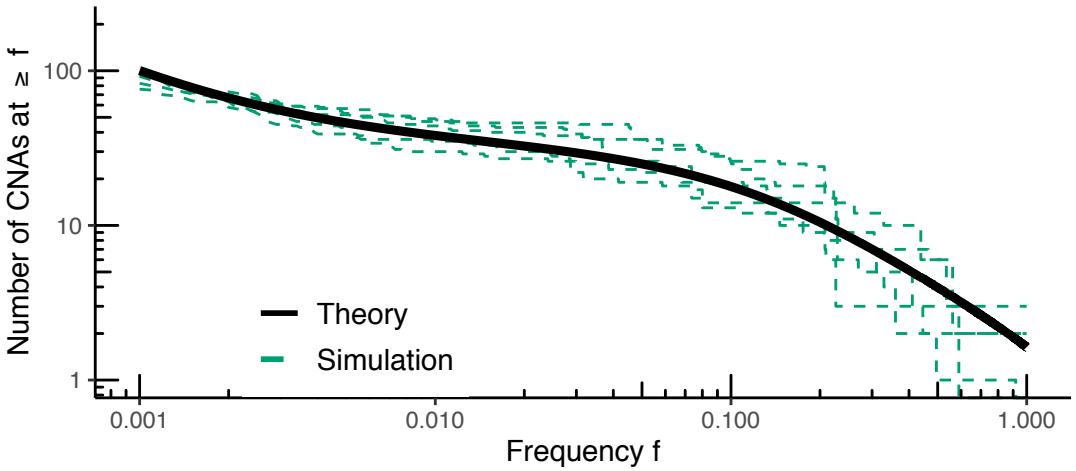


Figure 7: **The shape of expected frequency spectrum matches individual realizations**. Comparison of the expected site frequency spectrum and individual realizations of the stochastic simulations. Using $f(x)$ to interpret individual datasets is only valid if individual realizations follow the general form it indicates; observe that this broadly holds. Parameters are $N = 10^4$, $b = 1$, $d = 0.7$, $\mu_1 = 0.7$, $\mu_2 = 0.02$, $N_c = 50$. 5 realizations. For simulations the fitness parameter was $\alpha = 10^3$.

3 Breakpoint detection analysis

While our interest is in the accumulation of CNAs, breakpoints in the obtained copy number profiles are more easily detected and are less likely to be obscured by further alterations. Thus we compare our model to the frequency of breakpoints detected in the segmented copy number profiles. To obtain the frequency of a breakpoint we jointly segment the copy number profiles obtained for each sample using the package Piet as described in [8]. In this section we aim to: (i) determine reasonable parameters of the segmentation algorithm used, (ii) determine a plausible error model which may be used in our inference to account for errors in ascertaining the frequencies of breakpoints.

To do so we simulated copy number profiles and then segmented them using the same process used on the patient and cell line samples. Note that below in Section 5 we discuss simulations from a branching process to obtain the frequency of breakpoints. Here we will assume known frequencies of breakpoints, simulate copy number profiles and assess our ability to detect breakpoints at the correct frequency. Our aim here is to broadly assess the robustness of our conclusions, which is reflected in our simulation methodology.

3.1 Simulating copy number profiles

In this section we outline the simulation process. We simulate a population of 1393 cells (which matches the number of cells sequenced for TN7) with stochastically generated copy number profiles. We generated 500 simulated 'populations' of cells. The input to the segmentation process is the bin counts: the number of reads mapped to a specific genomic region. As each chromosome arm is segmented independently, it is sufficient to simulate the bin counts of a single arm. The simulated arm length was chosen as 279 bins, as, the average chromosome arm length is 279 bins, in units of the variable bins with average genomic length 200kb. The simulation comprises three steps.

First, for each of the 500 'populations' the total number of CNAs in the population, and how many cells possessed each CNA, was determined. Of the 500 populations, each received 2-5 CNAs, in equal proportions. Each population had the same evolutionary tree - specifying the order and relation of CNAs - $\mathcal{T} = \{(0), (0,1), (0,2), (0,1,3), (0,1,4)\}$ (so any cell can be labelled by $x \in \mathcal{T}$, e.g. a cell is $(0,1)$ if it possesses the first two CNAs). For those populations with only two CNAs, only (0) and $(0,1)$ are simulated - the analogous statement holds for populations with 3 or 4 CNAs. The CNA represented in the tree as (0) was always present in all cells in a population. For the remaining CNAs, let l_x be the number of cells labelled as $x \in \mathcal{T}$. Then for each x, we let the number of cells labelled as x be uniformly distributed between 1 and 'the number of cells which possessed the parent of x' -1, while keeping the tree structure intact. In detail,

$$\begin{aligned} l_{(0,1)} &\sim \text{Unif}(1, X-1) \\ l_{(0,2)} &\sim \text{Unif}(1, X - l_{(0,1)}) \\ l_{(0,1,3)} &\sim \text{Unif}(1, l_{(0,1)} - 1) \\ l_{(0,1,4)} &\sim \text{Unif}(1, l_{(0,1)} - l_{(0,1,3)}). \end{aligned}$$

If the population of cells has only four CNAs in total, then $l_{(0,1,4)}$ is set to zero, and similarly when having two or three total CNAs. If k CNAs are specified but $\sum_x 1(l_x > 0) \neq k$ (which could happen if e.g. $l_{(0,2)} = 1$ so $l_{(0,1,3)} = 0$ by the above designation), we resample all l_x.

Second, we build the true ploidy profile for each cell, based on which CNAs each cell acquired and in which order (the order being specified by the tree). For each alteration, a starting position S is uniformly chosen between 1 and 279 - the simulated chromosome arm length. The alteration length is then sampled according to $L \sim \text{Geom}(3/279)$. The end of the segment is given by $E = S + L$, if $S < 279/2$ (alteration starts in left-half of segment), else $E = S - L$. If $E < 1$, $E > 279$, or either S or E has been sampled for a CNA higher up the tree (which would contradict our infinite sites assumption for breakpoints), we repeat the process starting with sampling S. Each CNA is assigned to be either an amplification or a deletion with equal probability. The true ploidy profiles are then created for all cells, starting with a 'neutral' ploidy value of 3 since the mean ploidy for TN7 obtained by DAPI staining is 3.15 (and roughly 3 for most other tumor

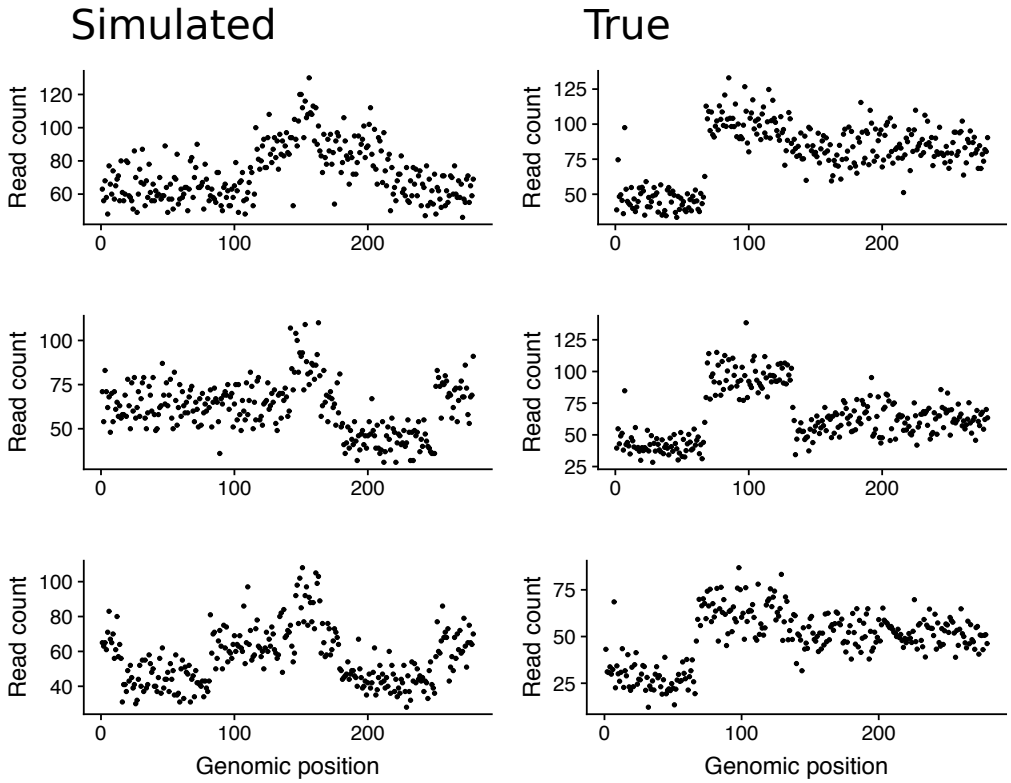


Figure 8: **Simulated vs true copy number profiles.** Example copy number profiles from sequenced cells, right, and from the simulation scheme described in Section 3.1, left.

samples), i.e. each cell is initiated with a vector $(3, 3, \ldots, 3)$ of length 279. Then appropriate vectors are added for each cell depending on their final genotype.

Finally, the 'observed' copy number profiles are generated. For each bin in each cell, we sample a negative binomial distribution. As the mean ploidy for TN7 is roughly 3 and the median bin count is 63, we expect a bin with true ploidy of 1 to have a count of 63/3. Therefore we specified that the mean of the negative binomial distribution is 'the true ploidy of that cell at that bin' $\times 63/3$ (where 'true ploidy' of a simulated cell is determined by step 2 of the simulation described in the preceding paragraph). The index of dispersion (iod) of the negative binomial distribution was chosen to be 1.33, which matches the median iod for TN7.

3.2 Segmentation error and parameters

After obtaining the simulated copy number profiles, they are segmented using a group fused lasso approach [8]. The output of this segmentation approach provides a list containing the observed breakpoints, the location of those breakpoints, and the number of cells each observed breakpoint is detected in. As our profiles are simulated we also have a list containing the true number of breakpoints, their true location, and how many cells possessed each breakpoint. Each observed breakpoint was matched to a true breakpoint by Euclidean distance, which was applied to the location and $\sqrt{\text{count}}$ (the square root was applied due to the larger range of cell counts, 0-1393, relative to location, 1-279; this was empirically found to more faithfully match breakpoints). True breakpoints which where not matched to an observed breakpoint are termed dropouts, while observed breakpoints which are not the closest to their matched true breakpoint are false positives. Thus we obtained the number of false positives, dropouts, and for those true breakpoints detected, the observed vs true cell count.

Next we used this information to select reasonable parameters for breakpoint detection. While we sim-

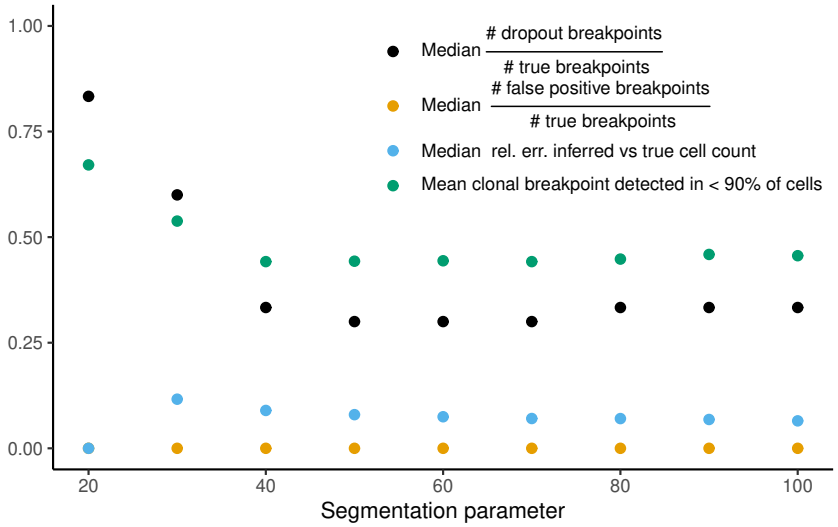


Figure 9: **Selecting optimal breakpoint detection parameters.** We simulated 500 populations of 1,393 cells, each with either 4, 6, 8, or 10 total breakpoints. Using this data, we investigated the dropouts (i.e. undetected breakpoints), false positives, errors on the number of cells possessing a true detected breakpoint (rel.err = |inferred count/ true count -1|), and the misclassification of clone breakpoints. The median/mean was taken over the 500 populations. We proceeded with the segmentation parameter equal to 70. Details for simulating the copy number profiles are provided in the main text.

ulated populations of differing cell numbers corresponding to the sampled number of cells per patient, for the parameter selection we focused on those populations containing 1,393 cells, corresponding to TN7. The errors obtained as a function of the segmentation parameter is displayed in Fig. 9. Based on these results, we proceeded with the segmentation parameter as 70.

At the selected segmentation parameter, we still observed significant errors in terms of the number of true breakpoints not detected (dropouts), clonal breakpoints being detected at lower frequencies, and generally in the inferred cell count of detected breakpoints. All 3 errors may bias our results, and so to account for these errors we used our simulations to construct a parametric error model.

3.3 Parametric error model

There are 3 error types we include in our error model: breakpoints not detected (dropouts), clonal breakpoints being detected at lower frequencies, and random noise in the frequency of detected non-clonal beakpoints. Note that the formulas in the error model are neither derived from the branching process model, nor the model used to generate the simulated copy number profiles; instead these expressions are used for capturing the key features of the observed breakpoint detection errors based on the simulated copy number profiles. To incorporate the errors due to dropouts we note that - based on the simulated copy numbers - the probability of not observing a breakpoint as a function of the number cells which possess the breakpoint, x, is well described by $p(x, 0) = \beta e^{-\gamma x}$ for $x \in \{1, \ldots, n\}$, see Fig. 10. Second, from our simulated copy numbers, the error on clonal breakpoints that are detected is well described by a trunctated Geometric distribution. That is for $y \in \{1, \ldots, n\}$, $p(n, y) = (1 - p(n, 0))\frac{q(1-q)^{n-y}}{1-(1-q)^n}$. Finally for non-clonal detected breakpoints in x cells we desired $p(x, y)$ to be integer valued and have variance increasing with x. Hence we model the number of cells we detect the breakpoint in, y, as having a negative binomial distribution with mean x and variance κx, truncated between $\{1, \ldots, n\}$.

The error model has 4 parameters β, γ, q, κ, and we sought reasonable estimates for these based on the simulated copy number profiles. However sensitivity analysis was carried out to explore the dependency of

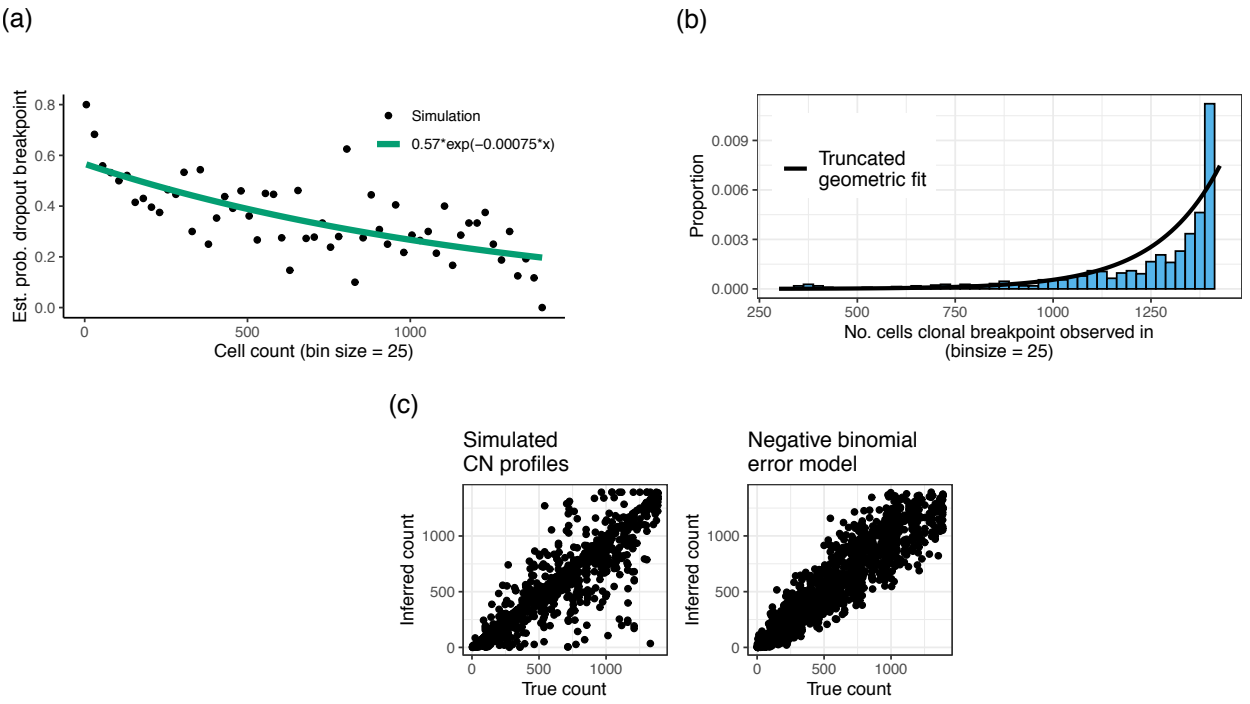


Figure 10: **Incorporating errors in breakpoint detection into likelihood.** There are 3 sources of error we take into account. In each case we select an appropriate parametric form to model the error, and select the parameters of the parametric function based on the simulated copy number profiles. (a) Probability of not observing a breakpoint as a function of the number of cells the breakpoint is in. (b) Probability of observing a clonal breakpoint at a reduced cell count. (c) For non-clonal detected breakpoints, the detection error is modeled by a negative binomial distribution centered at the true count and trunctated to be in the range $[1, n]$. The variance of the negative binomial was set by comparison with the simulated copy number profiles.

our conclusions based on these parameter estimates. Least square fits were used to determine β, γ, q and resulted in $\beta = 0.57$, $\gamma = 7.5 \times 10^{-4}$, $q = 6.1 \times 10^{-3}$ (for $n = 1393$ this q corresponds to an average error on clonal breakpoints of ≈ 163). For non-clonal detected breakpoints ($x < n$, $y > 0$), we binned the true breakpoint count into bins of size 100, and then estimated the variance on the detected number of cells with the breakpoint. The least squares fit resulted in $\kappa = 45.6$.

4 Statistical model to detect transient instability

Here we present a statistical scheme to detect the existence of transient instability based on the results presented in Section 2. We first outline the statistical model assuming no measurement error, and then describe the modifications used to incorporate the uncertainty in breakpoint detection.

4.1 Statistical model without breakpoint detection error

If we neglect stochastic drift – i.e. assume that the number of cells carrying a given CNA increases with deterministic exponential growth, and suppose that CNAs occur at a rate proportional to the population size (so that CNA arrival times are drawn from a nonhomogeneous Poisson process with rate $(b\mu e^{(b-d)s})_{s\geq 0}$) - then (i) the number of CNAs present at frequencies in (a, b) is Poisson distributed, and (ii) the number of CNAs present at frequencies (a_1, b_1) and (a_2, b_2) are independent, assuming the two intervals are disjoint. Motivated by this observation, that our data takes integer values, and that each CNA event ideally contributes two new breakpoints, we assume that the breakpoint frequency spectrum follows

$$(g_N(x))_{x=2/n}^{1} \sim (\text{Pois}(2(f(x) - f(x+1))_{x=2/n}^{1}, \tag{3}$$

where all $g_N(x)$ are independent and we set $f(x) = 0$ for $x > 1$. We neglect those breakpoints present in only one cell (frequency $1/n$). Our statistical model does not distinguish between the sampled frequency spectrum (obtained by sampling n cells) and that acquired via sequencing the whole tumor. Mathematically this extension is possible, but one obtains hypergeometric functions which are computationally expensive (relevant as we numerically maximize likelihoods). However the simulations presented below in Section 5 do include sampling, and the results presented there demonstrate the adequacy of the statistical model. Note that for the model as currently stated, the number of breakpoints present at frequencies (a, b) should simply be twice the number of CNAs present in the same frequency range - hence the number of breakpoints present in y cells should always be a multiple of 2. However, due to experimental artifacts, the data do not follow such a rule. The modification of doubling the Poisson parameter in Eq. 3 allows us to roughly account for each CNA introducing two new breakpoints, while ensuring a computationally tractable likelihood.

To assess whether a change in the CNA accumulation rate has occurred, we adopt a model selection approach: we numerically maximize the likelihood of our observed breakpoint frequency spectrum under gradual accumulation, $\mu_1 = \mu_2$, and transient instability, $\mu_1 > \mu_2$. The package bbmle [9] was used for likelihood maximization for both models. We chose the underlying optimization algorithm to be simulated annealing ("SANN" in R). Simulated annealing uses an adaptation of the Metropolis-Hastings algorithm to find an approximate optimum over the parameter space; hence there is no guarantee that a global optimum has been found. However, when applied to simulations, this approach has a reasonable performance, see Section 5. Having found the approximate maximum likelihood for both models, the AIC provides a means to assess superior model fit while penalizing the extra two parameters that have been introduced for the transient instability model (μ_2 and N_c). We uniformly sampled 50 initial values used as starting parameters for the maximization from the following grids:

$\log_{10}\mu_1 \in [-4, -1]$, $\delta \in [.01, .99]$, $k_0 \in \{0, \ldots, 50\}$	Gradual Accumulation
$\log_{10}\mu_1 \in [-1.5, 0]$, $\log_{10}\mu_2 \in [-4, -1.5]$, $N_c \in \{1, \ldots, 1500\}$ $\delta \in [.01, .99]$, $k_0 \in \{0, \ldots, 50\}$	Transient Instability

The optimization scheme will depart from these initial parameters but some brief justification for these parameter ranges are as follows. Regarding k_0, the median number of clonal CNAs that separated diploid cells from aneuploid (cancer) cells in [4] was approximately 25, while the range for δ is discussed in [10] which takes into account in vitro, epidemiological and sequencing data. For CNA rates, firstly we have our cell line data which resulted in rates in the range $[0.08, 0.3]$, which agrees with the values obtained by ref. [11]. Furthermore Refs. [12, 13] find an amplification rates for specific genes of $\approx 10^{-4}$. The CNA rate for a specific gene = CNA

rate $\times$ Pr(CNA affects specific gene), hence the CNA rate $\approx 10^{-4}(\Pr(\text{CNA affects specific gene}))^{-1} \geq 10^{-4}$ (as $\Pr(\text{CNA affects specific gene})^{-1} \geq 1$). For the transient instability scenario we keep roughly the same parameter space however we are also partially motivated by identifiability considerations. If $\mu_1 \approx \mu_2$, or N_c is too large or too small, then no differences between the models exist - we require the detection window displayed in Fig. 5 to be sufficiently large. While we do not explicitly model the biological mechanism behind an elevated CNA rate, taking the scenario of breakage-fusion-bridge (BFB) cycles as an example, the logarithm of the corresponding N_c would be proportional to the number of generations the BFB phenomenon persists for.

Of the 50 searches, the search leading to the highest likelihood was selected for further analysis, from which AICs were obtained. Control parameters used for the stochastic optimization algorithm SANN were: parscale - $N_c = 1$, $\delta = 0.02$, $\mu_1 = 0.02$, $\mu_2 = 0.001$, $k_0 = 1$) and maxiter = 2000, which were found to have reasonable performance on simulated data (see Section 5).

4.2 Incorporating breakpoint detection errors

To incorporate the breakpoint detection errors into our statistical model we include $p(w, z)$ representing the probability a breakpoint truly in w cells is detected in z cells, as discussed in Section 3.3 (with $w \in \{1, \ldots, n\}$, $z \in \{0, \ldots, n\}$). Then our updated likelihood is

$$(g_N(y))_{y=2/n}^{1} \sim \left(\text{Pois} \left[\sum_{x=1/n}^{1} p(nx, ny) 2(f(x) - f(x+1)) \right] \right)_{y=2/n}^{1}, \tag{4}$$

where, as before, n is the number of cells sampled and $f(x)$ is defined in Eq. 1. Note that in Eq. 4 $x \in \{1/n, \ldots, 1\}$ and therefore $nx \in \{1, \ldots, n\}$, and similar for y.

5 Inference on simulated data

We applied the inference scheme presented in Section 4 to simulated data in order to (i) examine its validity, and (ii) provide context to the point estimates obtained from the true data. Simulations were performed using the R-package SIApopr [7] with a modification to allow for a CNA accumulation rate change. Note that in the analytical model of Section 2, non-lethal CNAs have no effect on division and death rates, while in the stochastic simulation framework, CNAs alter the birth rate of cells as described in Section 1. For each realization of the stochastic process, we continued the simulation until the population reached 10^5 cells, a value which was chosen for computational efficiency. We then uniformly sampled 1,393 cells – the number of cells sequenced for TN7. In the simulation, any CNA contributed two breakpoints, from which we constructed the sampled breakpoint frequency spectrum. To account for measurement error for any breakpoint present in x cells we sampled from the distribution $p(x, y)$, where $p(x, y)$ is detailed in Section 3.3. This gives the 'observed' breakpoint frequency spectrum

Overall, we performed 60,000 simulations for each model. Parameters for the simulations were as follows. For all simulations $d = 1$, and equal proportions of simulations were carried out for $b = 1.01, 1.1, 1.5$, which correspond to slow, medium, and fast growing tumors, and cover the hypothesized range for $\delta = d/b$ based on sequencing data, observed division/death rates, and incidence data [10]. The remaining parameters were uniformly sampled from

$\log_{10} \mu_1 \in [-4, -1]$, $k_0 \in \{0, \ldots, 50\}$, $\log_{10}(\alpha) \in [2, 5]$ — Gradual Accumulation

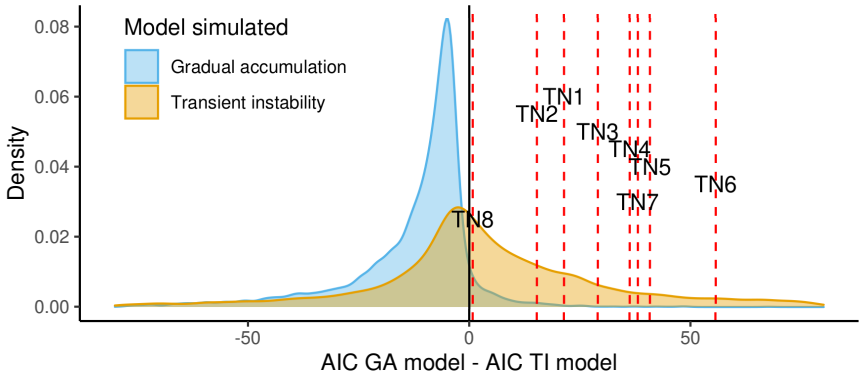
$\log_{10} \mu_1 \in [-1.5, 0]$, $\log_{10} \mu_2 \in [-4, -1]$, $N_c \in \{1, \ldots, 1000\}$
$k_0 \in \{0, \ldots, 40\}$, $\log_{10}(\alpha) \in [2, 5]$ — Transient Instability

Across all patients we see a range of 280-631 observed breakpoints present in greater than 1 cell. To make the signal strength of the data comparable with the patient data, and to exclude simulations with minimal

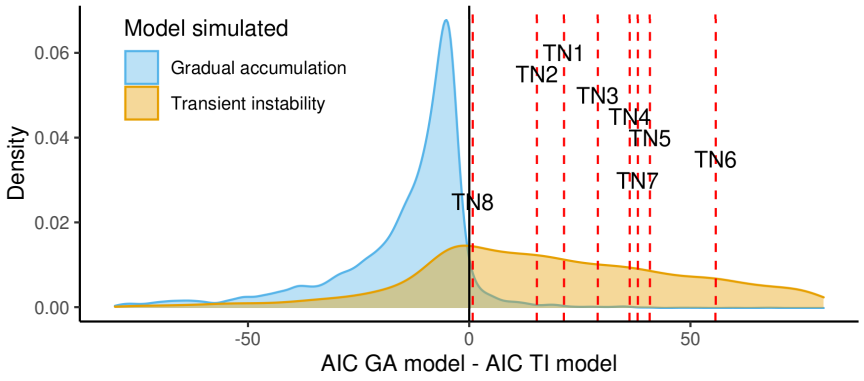
numbers of CNAs, only simulations yielding greater than 40 breakpoints present in greater than 1 cell were considered. As we performed our simulations until the total population size reached 10^5 cells due to computational cost, we cannot be sure that these parameter ranges carry biological significance since an unknown size-dependent rescaling of the parameter ranges might have to be applied. However, for reference, the selective advantage of driver mutations in a similar model [14] was estimated to be of the order of 10^{-3}, which corresponds approximately (due to model differences) to $\log_{10}\alpha = 3$. A discussion of the parameter regime for the other parameters (not pertaining to selection) was given in Section 4.1.

The initial birth and death rates control much of the stochasticity and number of CNAs observed since a high δ - the ratio of the death rate to the initial birth rate - increases the variance in the tumor size as a function of time and thus the expected number of CNAs. Due to this, we present the results separately for each initial birth rate b. Fig. 11 displays the difference of the AIC assuming the gradual accumulation and the AIC assuming transient instability, obtained from applying the likelihood model to the simulated data from either model (AIC GA model - AIC TI model > 0 implies a better fit for the transient instability scenario). The vertical red bars in Fig. 11 display the difference in AICs obtained for the patient data. Given that all patient samples are better explained by the transient instability scenario, our main concern is the false positive rate (misclassification of gradual accumulation simulations as transient instability). The false positive rate for the three different birth rates was 0.06, 0.05, and 0.05 (for $b = 1.5$, 1.1, 1.01). We note that if the simulated data is not perturbed to account for breakpoint-detection errors, greater sensitivity and specificity can be obtained with this method.

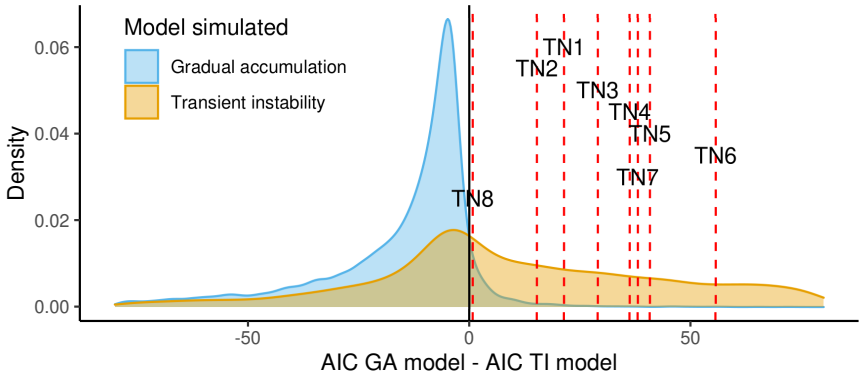
We next examined how accurately we could estimate the extent of the elevated instability, μ_1/μ_2. From the discussion surrounding Fig. 5, we expect that the detection window is of size $\log_{10}(\mu_1/\mu_2)$ hence we focus on estimating the order of magnitude of the CNA accumulation rate change. From the set of 6×10^5 simulations of the model with transient instability, we again restricted our analysis to only those that acquired greater than 40 breakpoints present in more than 1 cell. The histogram of the estimation error, i.e. the true $\log_{10}(\mu_1/\mu_2)$ - inferred $\log_{10}(\mu_1/\mu_2)$, obtained over all simulations is shown in Fig. 12(a), (b). From this data we observed that the CNA accumulation rate change is likely to be underestimated. This underestimation is due to the lower limit of the 'detection window' (x_1 - defined in the paragraph immediately preceding Eq. (2)) falling below the lower limit of frequencies observed, 2/1393). Restricting to only those simulations with x_1, $x_2 \in [2/1393, 1]$ extinguishes the underestimation bias and increases the accuracy. While this remark explains the bias in the simulated data it offers no information on the error of the inferred fold change from the patient data, as x_1, x_2 would be unknown for the patient data. Similarly, as for $x \in [x_1, x_2]$, $f(x)$ is approximately $\mu_1 N_c + k_0$, we looked to see whether the composite parameter $\mu_1 N_c$ could be inferred for simulations from the transient instability model that were correctly classified. The results are shown in Fig. 12 (c), demonstrating that the order of magnitude of $\mu_1 N_c$ can be inferred using our likelihood method.



(a)

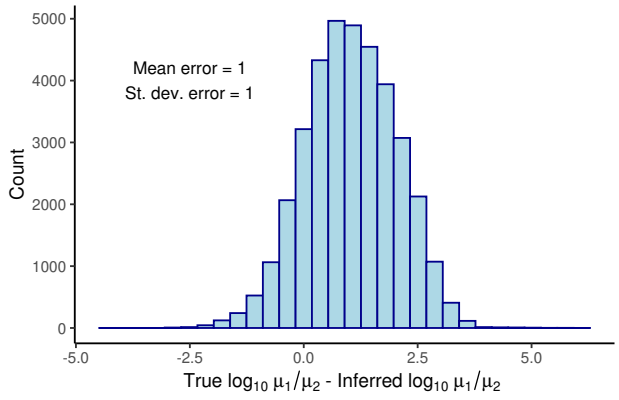


(b)

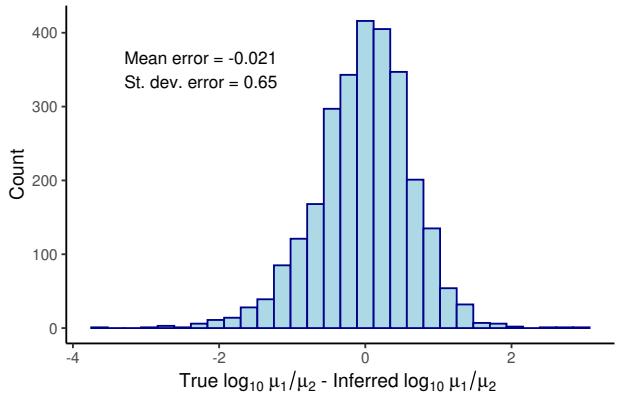


(c)

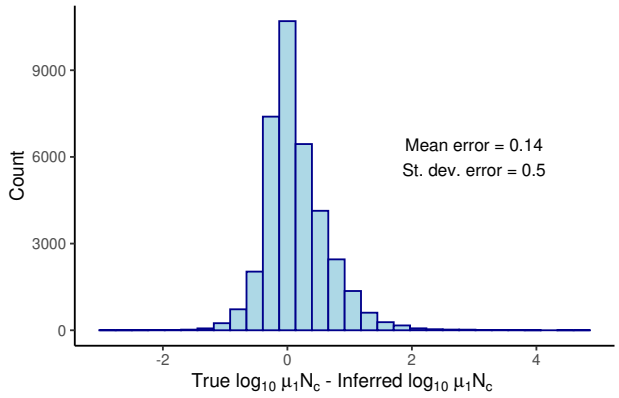
Figure 11: **Model selection scheme applied to simulated data**. We applied our likelihood framework described in Section 4 to simulated data. Simulations incorporate CNAs which heritably alter cells' division rate, subsampling of cells to mimic the experimental process, and breakpoint detection error as described in Section 5. Panels correspond to different initial division rates for fast, medium and slowly growing tumors ($b = 1.5, 1.1, 1.01$ (a), (b), (c)), and the red vertical lines denote the difference in AIC obtained for the patient samples. Our aim is to minimize the false positive rate (misclassification of gradual accumulations simulations). For the three difference replication rates the false positive rate was 0.06, 0.05, 0.05. Parameters were sampled from specified prior distributions for each simulated realization. 2×10^4 simulations were performed for each model for each initial division rate. Simulations that yielded too few CNAs were discarded in order to maintain a comparable signal to the patient data.



(a)



(b)



(c)

Figure 12: **Estimation of parameters on simulated data**. (a) We applied our likelihood framework described in Section 4 to simulations of the transient instability model which were correctly classified to estimate the increase in the CNA accumulation rate. From Fig. 5 we expect to be able to estimate the order of magnitude of the CNA accumulation rate change ($\log_{10}(\mu_1/\mu_2)$ – the size of the detection window in Fig. 5). The error in the estimates when all simulations are considered is displayed. Note that $\log_{10}(\mu_1\mu_2)$ is likely to be underestimated in this case. (b) We restrict the analysis to only simulations where the detection window is within the frequencies at which we detect CNAs (both $x_1, x_2 \in [2/1393, 1]$, where x_1, x_2 are defined immediately before Eq. (2)). The underestimation of the CNA accumulation rate change in the panel (a) is therefore due to the detection window extending below the lowest CNA frequency observed. (c) Errors on the estimate of the product of the elevated CNA rate and the population size at which the instability subsides.

6 Patient and cell expansions analysis

We then applied the statistical model of Section 4 to the breakpoint frequency distributions of the 8 sequenced patient samples and the two cell expansions. Breakpoint frequency spectra for each sample were obtained using the package Piet as described in [8], with parameters $\rho_1 = 0$, $\rho_2 = 0$, $\rho_3 = 70$. Brief justification for these parameter choices and an analysis of the errors the segmentation process introduces was given in Section 3.

Figs. 13 and 14 demonstrate the fit obtained by maximizing the likelihood of the patient data, using the likelihood of Eq. 4 for the cases of transient instability, $\mu_1 > \mu_2$, and gradual accumulation, $\mu_1 = \mu_2$. The difference in AIC under each model provides a measure of whether including a period of transient instability in the model results in a superior fit. As outlined in the figures, we found a superior fit of the transient instability model for all patient samples, and that the gradual accumulation model provided a better explanation for the cell expansion data. However we do note the far smaller difference in AICs for TN8. Of further note is that the lower frequency breakpoints appear to follow a power-law decay (notice the double-logarithmic scales in Fig. 14) as theoretically predicted for exponentially growing tumors [15], and more general tumor growth patterns [16].

Assuming a period of transient instability, we can also obtain point estimates of the extent of the increased CNA accumulation rate. For the reasons given in Section 5 we focus on inferring the composite variable μ_1/μ_2. Point estimates for the patients were 16, 18, 15, 8, 22, 18, 14, and 11, for TN1, TN2, TN3, TN4, TN4, TN5, TN6, TN7, and TN8 respectively. Under a $\log_{10}$ transform, these values are also the distance between the red vertical lines in Fig. 15 for reasons given in the discussion surrounding Fig. 5. We recall that when we estimated the CNA fold change on simulated data, discussed in Section 5, we typically underestimated the change: therefore the values obtained for the patients are likely to be lower bounds. The point estimates are those obtained for the simulated-annealing search which resulted in the maximum likelihood (among the 50 initiated simulated-annealing searches). The log-fold change for all 50 approximate maximum likelihoods is given in Fig. 17. Additionally we estimate the composite parameter $\mu_1 N_c$, that is the product of the elevated CNA rate and the population size at which the instability subsides. The estimates for $\mu_1 N_c$ were 35.7, 42, 30.4, 74.9, 63.8, 55.5, 55, 13.

To assess the dependency of our conclusions on the parameters of the error model we performed sensitivity analysis. We again maximized the likelihood of the patient data under both models for $\beta \in \{0.47, 0.57, 0.67\}$, $\gamma \in \{\frac{7.5}{2} \times 10^{-4}, 7.5 \times 10^{-4}, 15 \times 10^{-4}\}$, $q \in \{\frac{6.1}{2} \times 10^{-3}, 6.1 \times 10^{-3}, 12.2 \times 10^{-3}\}$, $\kappa \in \{35.6, 45.6, 55.6\}$. The resulting differences in AICs are shown in Fig. 18, where we have stratified the results by the clonal loss parameter q (controlling the extent of underestimation in the frequency of truly clonal breakpoints), which we found to have the dominant effect. We see that our conclusions are robust unless clonal breakpoints are detected at a substantially lower frequency (note that $q = \frac{6.1}{2} \times 10^{-3}$ corresponds to an average error of 307 breakpoints for $n = 1393$, that is clonal breakpoints are detected on average in 1393- 307 = 1086 cells). Indeed for all but one of the samples (TN8) we do not see a qualitative difference in conclusion without $q = \frac{6.1}{2} \times 10^{-3}$.

For reasons of biological relevance, and statistical identifiability, we have focused on the scenario where the CNA rate alters at a critical size threshold. However, a more general scenario may be treated similarly. For the setting where the CNA rate changes multiple times, the modifications to $f(x)$ (the expected number of CNAs present at frequency $\geq x$) can be found in Section 9. The likelihood method discussed in Section 4.1 may then be applied with the modified $f(x)$. To assess whether a further CNA rate change improves our fit, we maximized the likelihood of the breakpoint frequencies using Eq. 4 with 2 CNA rate changes. The resulting AICs are given in Fig. 19, and provide support that further alterations to the CNA rate do not improve explanatory power of the model given the patient data.

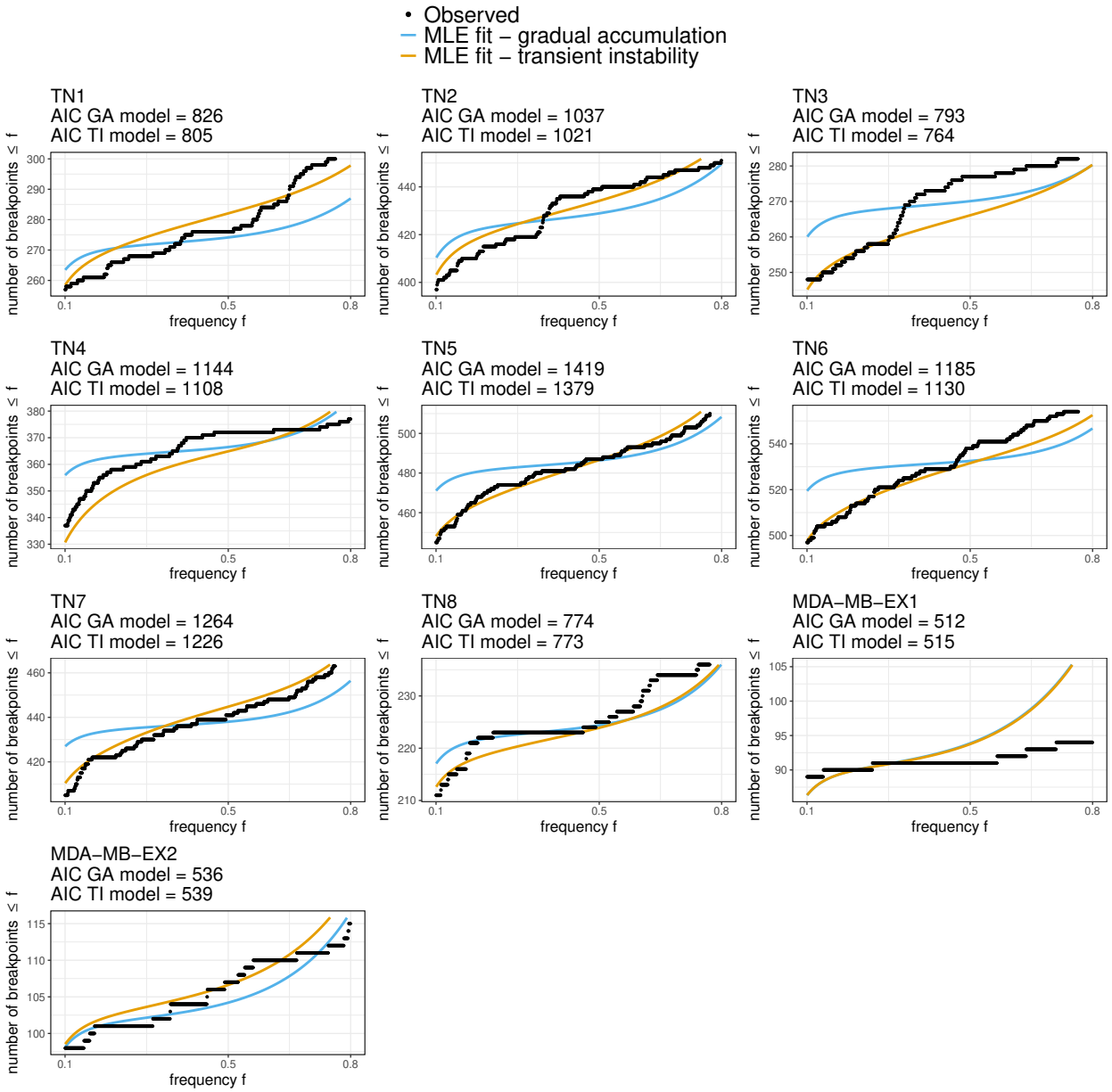


Figure 13: **Support for transient instability in patient samples but not in cell line data**. The likelihood of the observed frequency spectrum was maximized in the settings of transient instability (TI, $\mu_1 > \mu_2$) and gradual accumulation (GA, $\mu_1 = \mu_2$) using (3). For each sample, the plots show the best fit under each model together with the corresponding AIC values. Any difference between the models would be expected to exist at intermediate frequencies, and so we focus on this region and display the data as the cumulative number of breakpoints present at less than or equal to a frequency f.

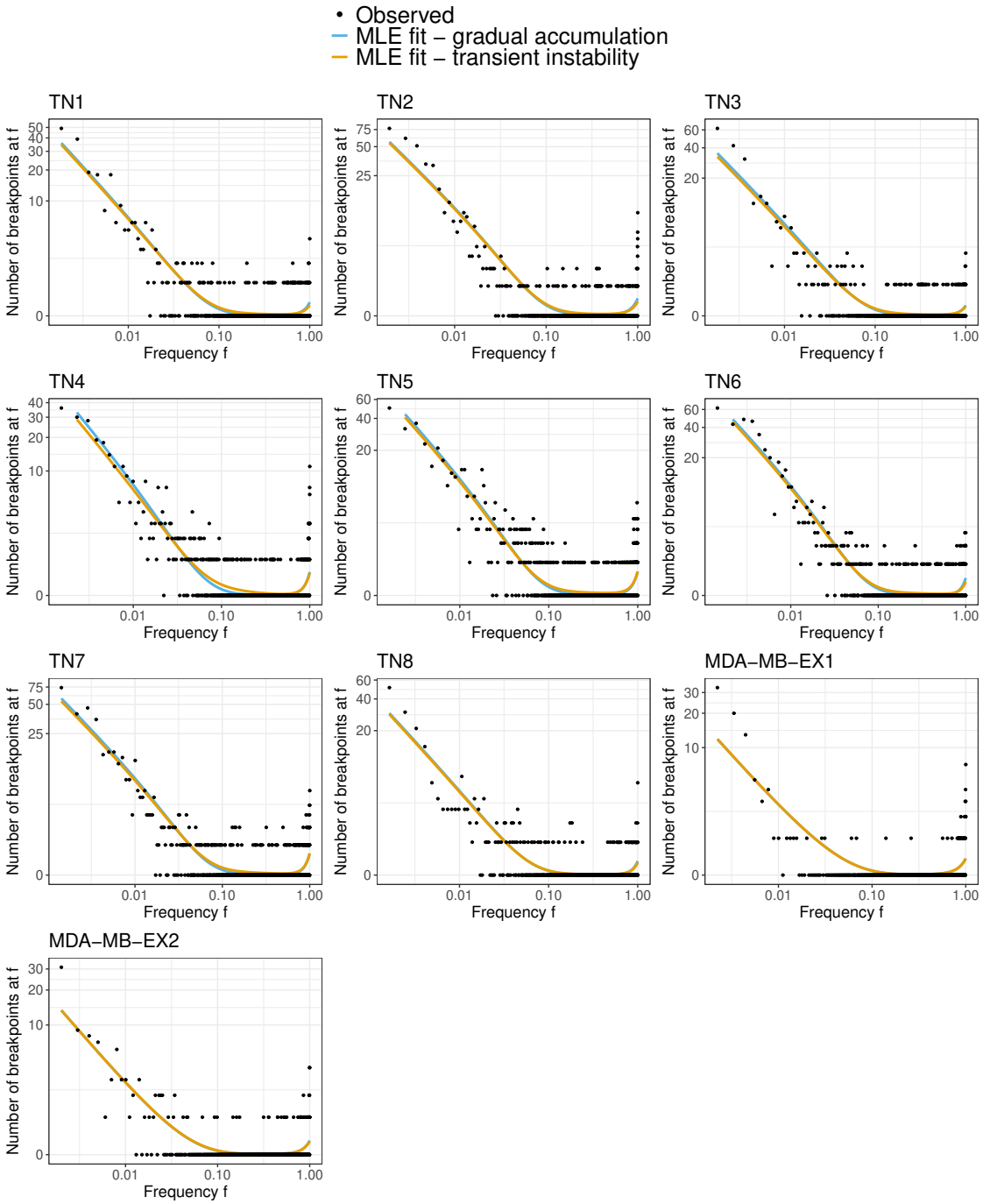


Figure 14: **Support for transient instability in patient samples but not in cell line data**. The likelihood of the observed frequency spectrum was maximized in the settings of transient instability (TI, $\mu_1 > \mu_2$) and gradual accumulation (GA, $\mu_1 = \mu_2$) using (3). For each sample, the plots show the best fit under each model together with the corresponding AIC values. Here the data is displayed as the number of breakpoints present at a frequency f. The y axis is displayed on a $\log(1 + y)$ scale. The low frequency breakpoints follow a power-law trend as expected by theory.

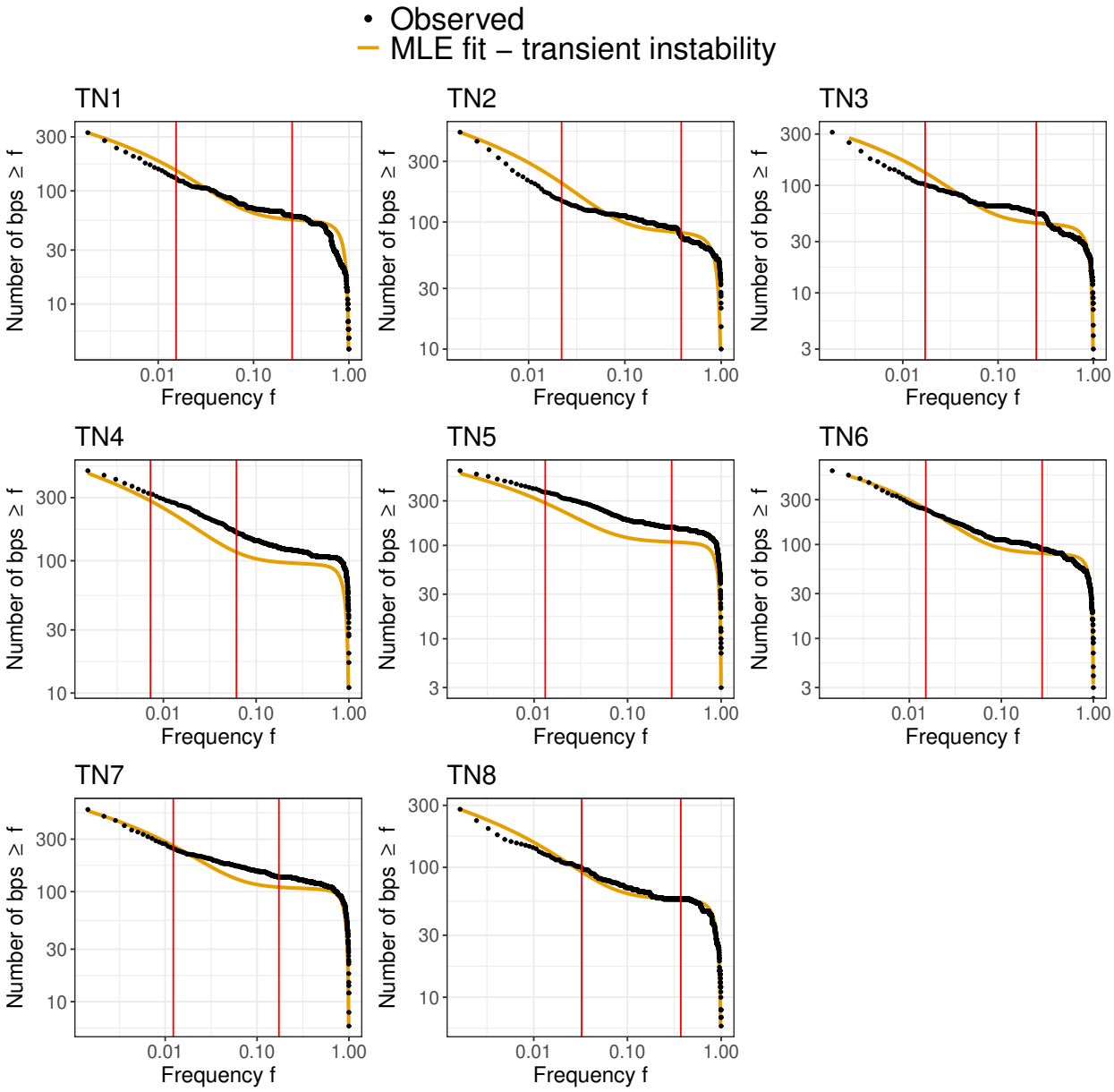


Figure 15: **Comparison of patient samples with expected breakpoint frequency**. For the patient samples, we compare the fits obtained with the general shape of the breakpoint frequency spectrum, see the discussion around in Eqn 2. Here the data is displayed as the cumulative number of breakpoints present at greater than or equal to a frequency f. Vertical red lines correspond to point estimates of $x_1 = \frac{\mu_2}{\mu_1(1-\delta)N_c}$ (left line) and $x_2 = \frac{1}{(1-\delta)N_c}$ (right line) and these should be compared to Fig. 5.

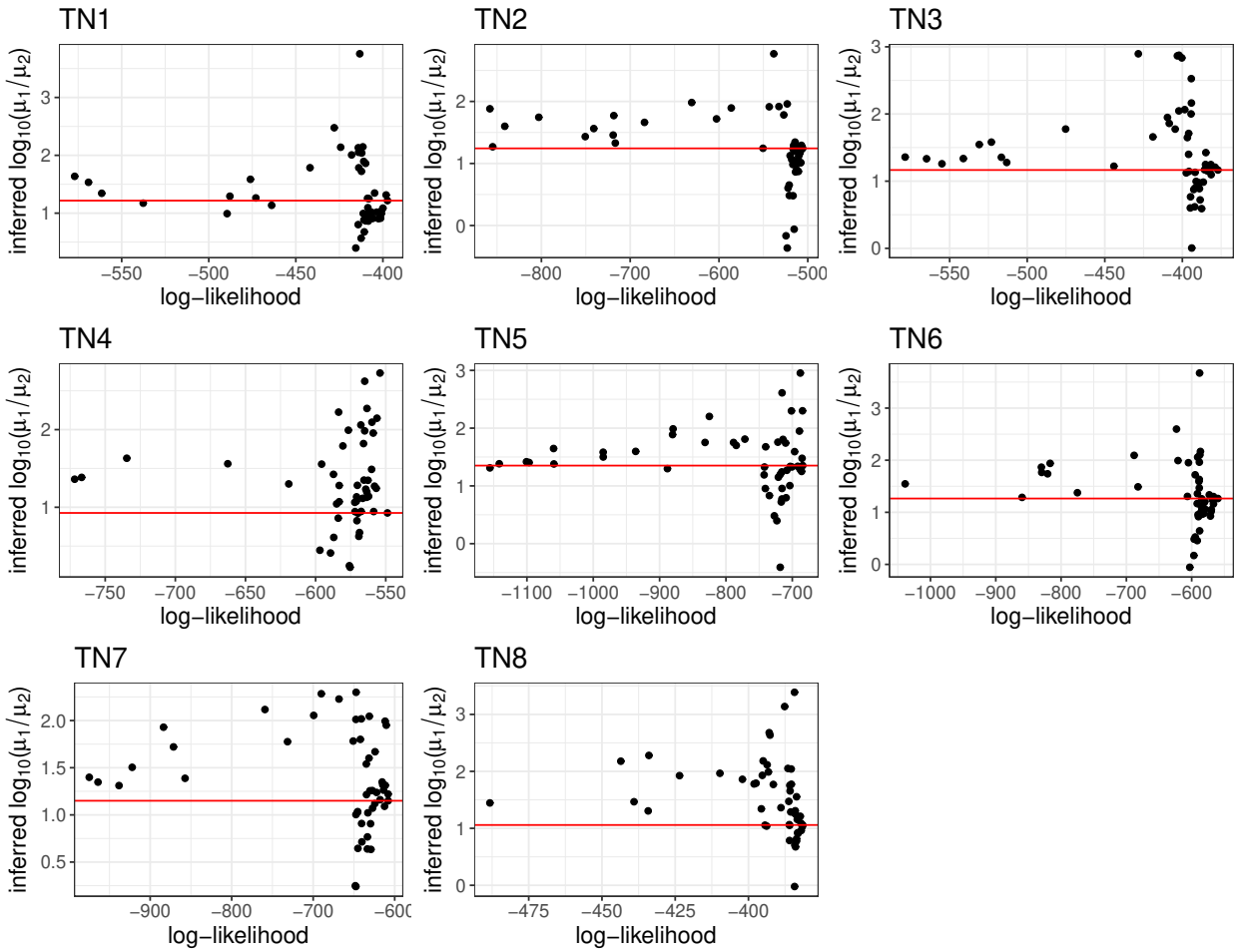


Figure 16: **Estimates of the fold change of the CNA accumulation rate**. We used a stochastic optimization scheme to obtain approximate maximum likelihoods, starting at 50 random initial parameter vectors. The panels show the log-likelihoods obtained at the end of the optimization scheme versus the inferred CNA accumulation rate fold change resulting from the approximate maximum likelihood fits. The right-most point on each panel corresponds to the CNA accumulation rate change reported in the text, and its value is indicated by the red horizontal line..

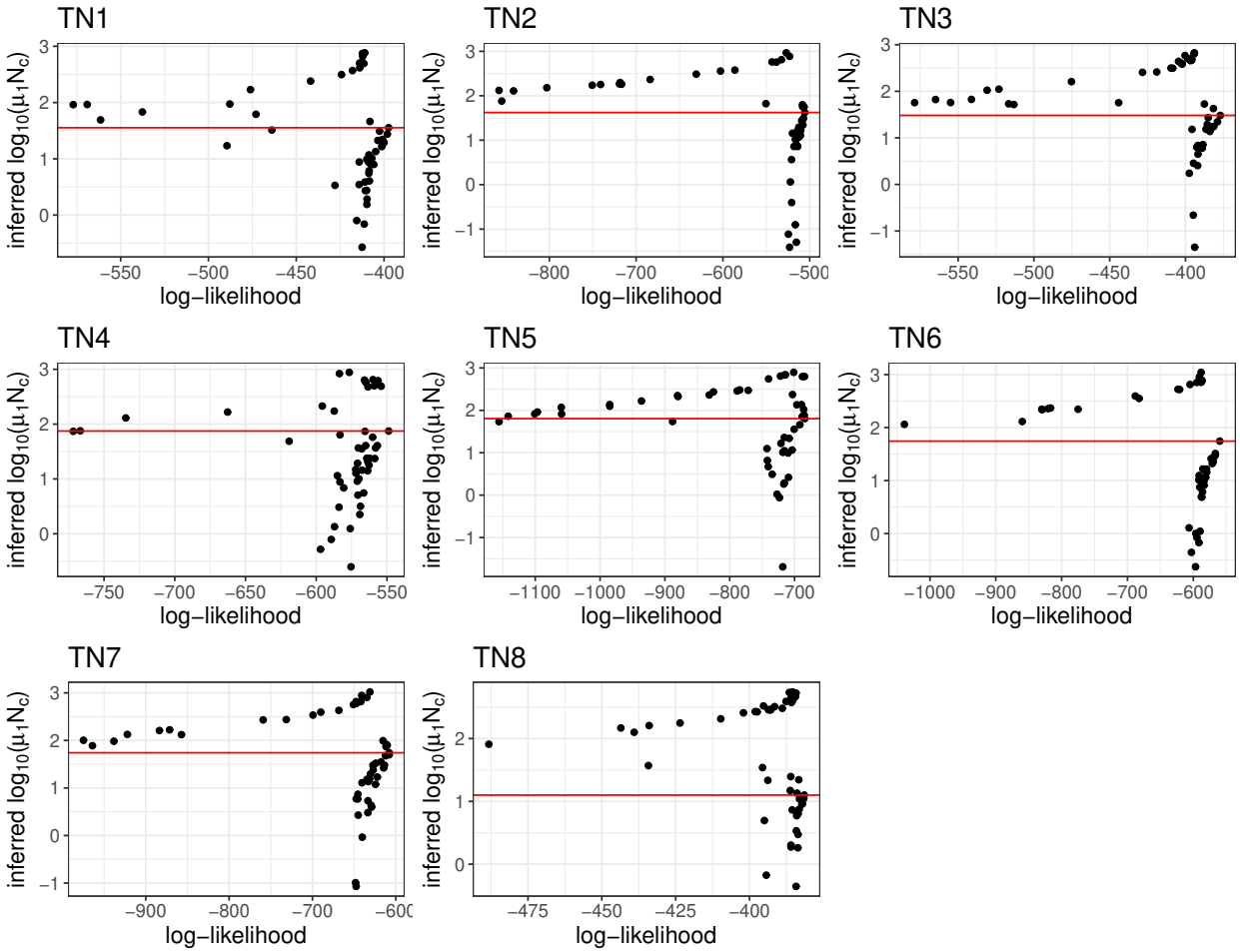


Figure 17: **Estimates of $\mu_1 N_c$.** We used a stochastic optimization scheme to obtain approximate maximum likelihoods, starting at 50 random initial parameter vectors. The panels show the log-likelihoods obtained at the end of the optimization scheme versus the inferred $\mu_1 N_c$ change resulting from the approximate maximum likelihood fits. The right-most point on each panel corresponds to the estimated $\mu_1 N_c$ reported in the text, and its value is indicated by the red horizontal line.

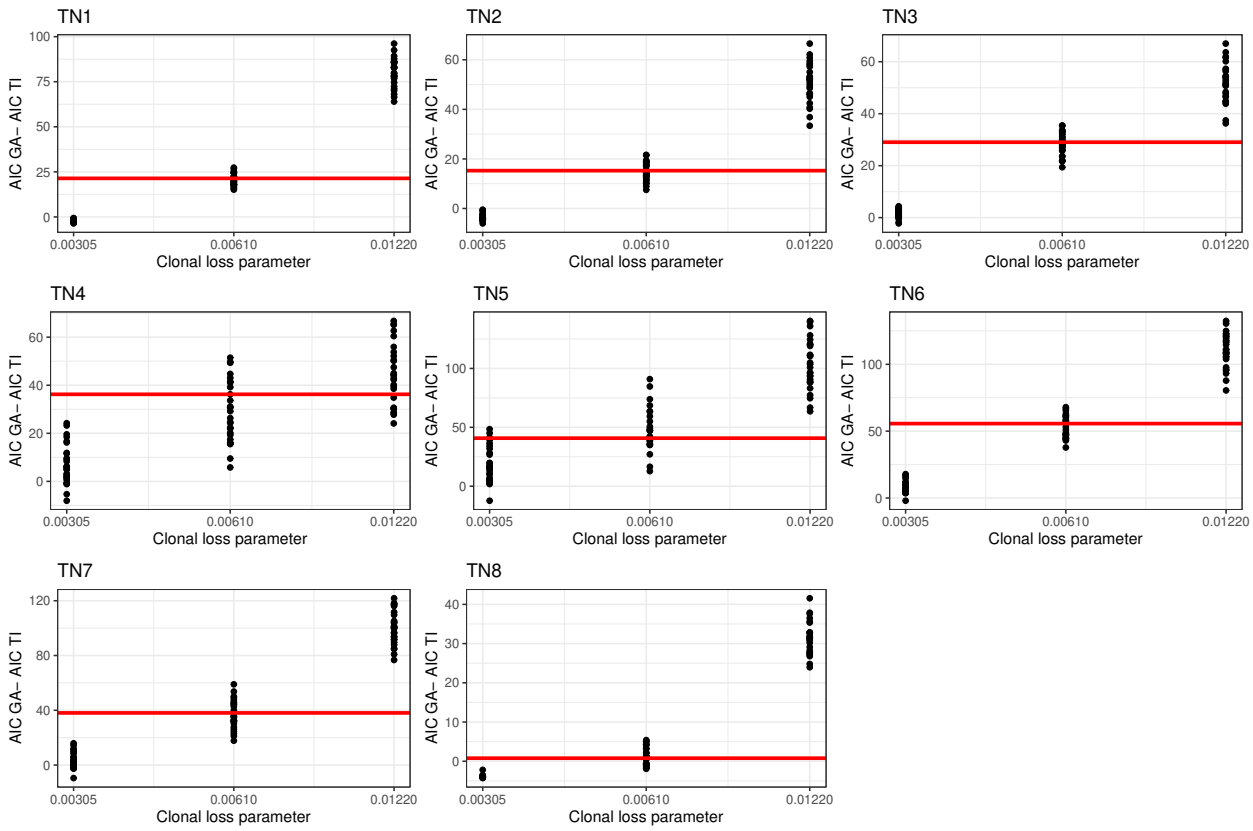


Figure 18: **Conclusions are robust without severe errors in detecting clonal breakpoints.** Here we fit each dataset with the model incorporating errors in breakpoint detection Eq. (4), across different parameters for the error model. For each set of parameters, we display the difference in the AIC. Due to the strong influence of the clonal loss parameter (the error on clonal breakpoints) we stratify the difference in AICs. Horizontal red bar corresponds to the difference in AIC reported in the text.

	AIC 0–change	AIC 1–change	AIC 2–change
TN1	826.1	**804.7**	808.2
TN2	1036.7	**1021.4**	1026.6
TN3	793.1	**764.1**	764.2
TN4	1143.7	**1107.5**	1128.7
TN5	1419.4	1378.6	**1377.9**
TN6	1185.2	**1129.6**	1133.8
TN7	1264.1	**1226**	1237.4
TN8	773.7	**773**	776.9
MDA–MB–EX1	**511.9**	515.5	516.4
MDA–MB–EX2	**536.1**	539.2	543

Figure 19: **Comparison of AIC values with multiple CNA rate changes**

7 CNA rate inference

To estimate the focal CNA rate (focal as we only consider intra-arm breakpoints) we adopt the statistical model described in Eq. (4) assuming gradual accumulation of breakpoints; however, we neglect the erroneous detection of clonal breakpoints at subclonal frequencies and the random noise term previously modeled by a truncated negative binomial distribution. To justify using this reduced model we restrict ourselves to breakpoints at frequencies ≤ 0.5 so that truly clonal breakpoints do not influence the estimates. Note that the cell expansions are grown to $\approx 10^7$ cells, hence we are in the large N regime and can use the analytical expressions.

We estimate the focal CNA rate in the two cell expansions, MDA-MB-EX1 and MDA-MB-EX2. Suppose n_i cells are sequenced in sample i ($i = 1, 2$) and let X_i be the number of breakpoints observed in the ith sample in the frequency range $[2/n_i, 0.5]$. For the cell lines we can assume $\delta = 0$. Then our assumed model is

$$X_i \sim \text{Pois}\left[2\mu_i \sum_{y=2}^{\lfloor n_i/2 \rfloor} \left(\frac{n_i}{y} - \frac{n_i}{y+1}\right)(1 - \beta e^{-\gamma y})\right] = \text{Pois}\left[2\mu_i \mathcal{A}\right],$$

where we introduce $\mathcal{A}$ for notational convenience. Then given the number of observed breakpoints in the ith colony, x_i, our point estimate for μ_i is $\frac{x_i}{2\mathcal{A}}$. Using the exact method within the survival package [17], we obtain z_1, z_2 as lower and upper bonds for a 95% CI for $2\mu_i\mathcal{A}$ and hence the relevant lower and upper bounds for the 95% CI of μ_i is $\frac{z_1}{2\mathcal{A}}, \frac{z_2}{2\mathcal{A}}$.

Applying this to the samples MDA-MB-EX1 and MDA-MB-EX2 with $\beta = 0.57$, $\gamma = 7.5 \times 10^{-4}$ (for the reasons given in Section 4.2) we obtain the estimates 0.235 (0.189, 0.288); 0.249 (0.204, 0.3). In the absence of dropout errors ($\beta = 0$) these estimates instead are 0.102 (0.082, 0.130); 0.108 (0.088, 0.13).

8 Discussion

Recent work has suggested the existence of punctuated evolution at the level of structural variants, both en route to tumor initiation and during cancer evolution [4, 18, 19]. Most relevant to our current study is ref. [19] who, using SNVs as a molecular clock, provided evidence that arm level changes occur early in colorectal carcinoma. A clock like approach was unavailable due to the sequencing depth in this study (0.01X), and so instead we sought to identify an enrichment of breakpoints at high frequency. This approach has some caveats which we now discuss.

8.1 Alternative models

We have adopted a branching process model of breakpoint accumulation. Such models have been widely used in cancer modeling, particularly to describe stochastically acquired heritable mutations [20, 21, 4, 16, 14, 22, 23]. This model is simplistic in that is assumes independence between cells, exponential growth, and a common mutation CNA rate for all cells. However, we believe that such a model is able to provide suggestive evidence for the specific question of detecting elevated instability. Broadly, early elevated genomic turbulence - here termed transient instability - will give rise to a relative enrichment of high frequency breakpoints. In the patient data, due to the existence of these high frequency breakpoints, a period of transient instability was deemed plausible when compared to a model of gradual accumulation. However, alternative explanations for an abundance of high frequency breakpoints also exist, which we now discuss.

Selection Within the same modeling framework, it has been recently demonstrated that mutations hitchhiking in a selectively advantageous subclone can be present at high frequencies [21, 24]. While we cannot rule out such a scenario, it is expected that these subclones will result in 'peaks' in the density of the frequency spectrum, resulting from the mutations present in the founding cell of the subclone (here the subclone is defined by all cells possessing the advantageous mutation). Such peaks are not seen in our patient data. While our analytical model was derived with mutations being either lethal, or not altering the fitness of cells, we showed in Section 5 that our results are robust under a model with a double-exponential

fitness distribution. One instead might consider a mutation arising which significantly increases the selective advantage (for example a driver mutation), so much so that the resulting subclone sweeps to near fixation. In order for this to occur this mutation must occur early or have an extremely strong selective advantage [21]. Considering the former case, if the mutation arises τ_s time units after the tumor initiated, and we assume deterministic growth of the tumor - so has size $e^{(b-d)t}$ at time t - then a period of transient instability might still be detected so long as the instability persists after τ_s. In the language of our model this amounts to the threshold size being sufficiently large so that $\log(N_c)/(b-d) > \tau_s$. In such a scenario, the above arguments would still be expected to hold but with the critical size N_c decreased. If a selective sweep occurs outside this regime we would not expect to detect elevated instability. We note also that ref. [25] has recently given a convincing heuristic argument that, within a finite sites version of our gradual accumulation model, if only a small number of sites provide a selective advantage then this will be unlikely to impact high frequency clones.

Finite sites effects The infinite sites assumption was used for computational tractability. Our conclusions are based on an enrichment of high frequency clones (group of cells sharing a breakpoint), and it could be the case that this is due to breakpoints occurring in the same genomic regions in different lineages - inflating the clone sizes. To account for this we could use a finite sites model, where sites here would denote the number of possible locations that breakpoints could occur in - if copy number profiles have $S+1$ bins then we have S sites. In this finite sites version of the model, our $f(x)$ would be an analogue of the generalized Luria-Delbrück distribution, which is difficult to evaluate numerically (relevant for numerical maximization of likelihoods). Adopting a finite sites approach with gradual accumulation, the issue of the inflation of clone sizes due to the same breakpoints occurring in different lineages corresponds to the mode of the size normalized Luria-Delbrück distribution (i.e. the distribution of the number of cells with an altered site divided by N) becoming present in the frequency range we consider $\approx [x_{\min}, 1]$ (the right tail of the Luria-Delbrück distribution, past the mode, agrees with the infinite sites case [25]). We now give a crude heuristic argument that if this were true we would observe more breakpoints.

An approximate mapping between the models is that when a CNA occurs during a division, two of the S sites are chosen uniformly at random to contain breakpoints. If we let μ_* be the per site breakpoint rate, then μ_* is related to per cell division CNA rate by $\mu_* = \mu 2/S$ (as $\mu \leq 1$ we henceforth assume $\mu_* \leq S/2$). When $N\mu_* \gg 1$, the Luria-Delbrück distribution is conjectured to have a mode at $\frac{N\mu_*(\log N\mu_* - 0.23)}{1-\delta}$, see [26] Section 5 (the conjectured mode given in ref. [26] is when death is neglected, i.e. $\delta = 0$, but the alteration required to the mode when death is included is clear from the scaling presented in Section 5.2 in ref. [27]). Hence the size-normalized distribution has mode at $\frac{\mu_*(\log N\mu_* - 0.23)}{1-\delta}$. For our application, $S \approx 13000$, while $N \approx 3 \times 10^9$. For the mode to be greater than $x_{\min}$, we then require $\frac{\mu_*(\log(\mu_*)+21.59)}{1-\delta} > x_{\min}$ (†). Assume finite sites effects are visible and so the inequality (†) holds, and let B_{N,μ_*} follow the Luria-Delbrück distribution with final population size N and per division mutation rate μ_*. The expected number of observed breakpoints would be

$$S\Pr(B_{N,\mu_*}/N \in (x_{\min}, 1)) \gtrapprox \frac{S\mu_*}{1-\delta}(x_{\min}^{-1} - 1) > \frac{S}{(\log(\mu_*) + 21.59)}(1 - x_{\min}). \tag{5}$$

In (5), the expression immediately following $\gtrapprox$ is an approximation of the size-normalized Luria-Delbrück distribution which holds when N is large and $\mu_* \propto 1/N$ (see Theorem 4.4 in [25]). Our interest is in the parameter regime $\mu_* \gg 1/N$, in which case we expect more sites containing breakpoints at large frequencies, hence the approximate lower bound. The second inequality in (5) uses inequality (†). With $S = 2/\mu_* = 13000$, $x_{\min} = 1/500$ the right-most expression of (5) is approximately 1013. So if we assume finite sites effects are visible then we expect at least 1013 non-unique breakpoints. However, the number of non-unique breakpoints observed in the patients are in the range 280-631 and hence we do not expect the assumption of infinite sites to bias our conclusions.

Hotspots of instability A further alternative explanation for the enrichment of high frequency breakpoints is that certain genomic regions have a greatly increased chance of acquiring structural variants per cell division. Within the finite-sites framework, using similar heuristic arguments as above, one can argue that the frequency spectrum can attain a form similar to that given in $f(x)$. To examine this alternative

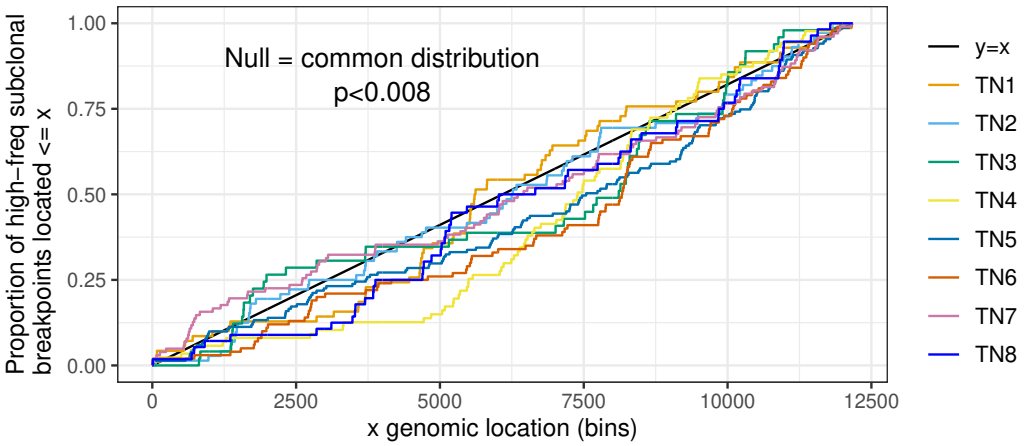


Figure 20: **Locations of high frequency subclonal breakpoints** For each patient we extract the genomic locations of those breakpoints present at frequencies in $(0.05, 0.9]$ and plot the cumulative distribution of these locations. A k-sample Anderson-Darling test rejects the null hypothesis that these locations are sampled from a common distribution. This provides evidence against global (shared amongst patients) 'hotspots' of genomic instability.

explanation, for each patient we looked at the genomic location of breakpoints present at frequencies between 0.05 and 0.9 - high frequency but subclonal. The cumulative distribution of the genomic location of the high frequency subclonal breakpoints is shown in Fig 20. Patient to patient differences exist, and the null hypothesis that there is a common distribution specifying the locations of the high frequency subclonal breakpoints is rejected by a k-sample Anderson-Darling test (the package [28] with "Version 1" of the test and the "asymptotic" method was used). This provides evidence against a common set of 'hotspot' genomic locations. However, we cannot rule out the possibility that each patient has a unique set of sites with an elevated rate of acquiring breakpoints.

Spatial growth The branching process model used neglects spatial considerations; however, it is technically possible that mutations are observed at high frequencies due to spatial effects. Several recent studies [29, 30, 31] have considered this issue. While it has been argued that spatial growth leads to a depletion of high frequency mutations (see the supplementary material of ref.[21] Eq. (55)), it has also been observed that boundary driven growth can induce a form of 'gene surfing' [31] leading to high frequency mutations being enriched. Such an effect could in principle be detected by combining genotypic information of single cells with their spatial location. As spatial data was not collected as part of the experimental protocol of our study, and due to the challenges of parameterizing a spatially explicit mathematical model, we did not consider such models in detail

Non-uniform sampling When comparing against simulations, we sampled cells uniformly at random from the full simulated tumor. However if cells are sampled in a non-uniform manner, then mutations may appear to be at higher frequencies in the sample than they actually are in the tumor (in an extreme case a mutation could appear in all cells sampled, but not be present in any other cell in the tumor). While due to the experimental design under which the cells were sampled in this study, assuming uniform at random sampling seems valid, we cannot rule out elevated frequencies due to sampling errors (see Fig. 4 in ref. [32] for further discussion).

9 Derivations

9.1 Frequency spectrum for tumor

Here 'mutation' refers to any heritable alteration to the genome - our primary example being CNAs. The number of mutations present in the founder cell k_0 will always be at frequency 1, hence we ignore this term in the derivation.

Our interest is in

$$f_N(x) = \mathbb{E}[F_N(x)] = \mathbb{E}[K_N]\Pr(\text{randomly selected mutation is at a frequency} \geq x\ ) \tag{6}$$

Random here refers to uniform at random sampling assuming $K_N > 0$, and our decomposition of $f_N(x)$ follows from Wald's equation. We turn to approximating $f_N(x)$, and use similar approximations to those outlined in [33, 10]. In ref. [34], in the setting of $\mu_1 = \mu_2$, it was shown that the mutation accumulation process may be approximated as a Poisson process on $[1, N]$ with rate μ. The derivation in the SI of [34] needs no alteration with a change in the mutation rate and so henceforth we assume that mutations arrive as a Poisson process with rate μ on $[1, N]$. Under the Poisson approximation, $\mathbb{E}[K_N] = \mu_1 N_c + \mu_2(N - N_c)$. We now turn to the second term in the right-hand side of (6).

Let us uniformly at random select one of our K_N mutations (assuming $K_N > 0$). Suppose this mutation arrived on a division event taking the population size from X to $X + 1$ cells. The lineage of the cell which originally received our mutation of interest must survive. Of the remaining X cells, let the number whose lineage survives be Y. Finally let the frequency of our chosen mutation at the observation size be ϕ_N. To complete our approximation of $f_N(x)$, we desire $\Pr(\phi_N \geq x)$. We collect some useful facts. First, given Y

$$\lim_{N\to\infty} \Pr(\phi_N \geq x|Y) = (1 - x)^Y$$

(for a derivation see the proof of Theorem 1 in [15]). Second for fixed X, as $N \to \infty$ the distribution of $Y|X$ converges to that of Binomial$(X, 1 - \delta)$. Under the Poisson approximation, the arrival size X will be a continuous random variable on $[1, N]$. However the statement $Y|X \sim$ Binomial$(X, 1 - \delta)$ is true only for discrete X, hence we discretize X in the following way; X is equal in distribution to $BU_1 + (1 - B)U_2$, with B Bernoulli with success parameter $p = \frac{\mu_1 N_c}{\mu_1 N_c + \mu_2(N - N_c)}$, U_1 discrete Uniform on $\{1, \ldots, N_c\}$ and U_2 discrete Uniform on $\{N_c + 1, \ldots, N\}$. The choice of discrete versus continuous X makes only a small difference, however we've found the discrete X to be more accurate relative to simulations. Thus our interest is in,

$$\Pr(\phi_N \geq x) = p\mathbb{E}[\Pr(\phi_N \geq x|1 \leq X \leq N_c)] + (1 - p)\mathbb{E}[\Pr(\phi_N \geq x|N_c < X \leq N)]$$

Returning to $f_N(x)$ we have

$$\begin{aligned} f_N(x) &= (\mu_1 N_c + \mu_2(N - N_c))\left(p\mathbb{E}[\Pr(\phi_N \geq x|1 \leq X \leq N_c)] + (1 - p)\mathbb{E}[\Pr(\phi_N \geq x|N_c < X \leq N)]\right) \\ &= \mu_1 N_c \mathbb{E}[\Pr(\phi_N \geq x|U_1)] + \mu_2(N - N_c)\mathbb{E}[\Pr(\phi_N \geq x|U_2)]. \end{aligned}$$

In turn let's consider each of the above summands:

$$\begin{aligned} \mu_1 N_c \mathbb{E}[\Pr(\phi_N \geq x|U_1)] \xrightarrow{N\to\infty} &\mu_1 N_c \mathbb{E}\left[\sum_{y=0}^{U_1} \binom{U_1}{y}(1 - x)^y(1 - \delta)^y \delta^{U_1 - y}\right] \\ = &\mu_1 N_c \mathbb{E}[(1 - x(1 - \delta))^{U_1}] \end{aligned}$$

Note that for U discrete uniform on $\{a, \ldots, b\}$, its moment generating function is $\mathbb{E}[e^{sU}] = \frac{e^{sa} - e^{s(b+1)}}{(b-a+1)(1-e^s)}$, implying $\mathbb{E}[s^U] = \mathbb{E}[e^{U\log(s)}] = \frac{s^a - s^{b+1}}{(b-a+1)(1-s)}$. Hence

$$\lim_{N\to\infty} \mu_1 N_c \mathbb{E}[\Pr(\phi_N \geq x|U_1)] = \mu_1 \frac{(1 - x(1 - \delta)) - (1 - x(1 - \delta))^{N_c+1}}{x(1 - \delta)}$$

Similarly, formally substituting in the asymptotic form of $\Pr(\phi_N \geq x|U_2)$

$$\begin{aligned}\mu_2(N-N_c)\mathbb{E}[\Pr(\phi_N \geq x|U_2)] &\approx \mu_2(N-N_c)\mathbb{E}[(1-x(1-\delta))^{U_2}] \\ &= \mu_2 \frac{(1-x(1-\delta))^{N_c+1} - (1-x(1-\delta))^{N+1}}{x(1-\delta)} \\ &\xrightarrow{N\to\infty} \mu_2 \frac{(1-x(1-\delta))^{N_c+1}}{x(1-\delta)}\end{aligned}$$

where the last limit is due to $1-x+\delta x < 1$ for $\delta < 1$ and $x > 0$.

Collecting our results we have (formally), with $r(x) = 1-x(1-\delta)$

$$f(x) = \lim_{N\to\infty} f_N(x) = \mu_1 \frac{r(x) - r(x)^{N_c+1}}{x(1-\delta)} + \mu_2 \frac{r(x)^{N_c+1}}{x(1-\delta)}.$$

Immediately the number of clonal mutations is apparent as $r(1) = \delta$ leading to $f(1) = \frac{\mu_1\delta + \delta^{N_c+1}(\mu_2-\mu_1)}{1-\delta}$. Note as $N_c \to \infty$ or for $\mu_1 = \mu_2$ we have $f(1) = \mu_1\delta/(1-\delta)$ which matches Eq. 6 in [10].

We now look at differing regimes to understand the behaviour of the limiting expected site frequency spectrum. It is convenient to rewrite $f(x)$ as

$$\begin{aligned}f(x) &= r(x)\left(\mu_1 \frac{1-r(x)^{N_c}}{x(1-\delta)} + \mu_2 \frac{r(x)^{N_c}}{x(1-\delta)}\right) \\ &\approx \mu_1 N_c \frac{1-e^{-x(1-\delta)N_c}}{x(1-\delta)N_c} + \mu_2 \frac{e^{-x(1-\delta)N_c}}{x(1-\delta)}\end{aligned}$$

where terms of order $x(1-\delta)N_c$ have been dropped. As x moves from 0 to 1, this function has broadly 3 regimes, dictated by which of the two summands above dominate. For $x(1-\delta)N_c \ll 1$ the first summand is constant at $\mu_1 N_c$, and so the decay is controlled completely by the second summand. The first summand starts to dominate when $x \approx x_1 = \frac{\mu_2}{\mu_1(1-\delta)N_c}$. This corresponds to the tumor having exceeded N_c, but still being too small to pick up any mutations with the lower mutation rate μ_2. Once x exceeds x_1, as $\frac{1-e^{-x(1-\delta)N_c}}{x(1-\delta)N_c} \approx 1$ for small $x(1-\delta)N_c$ we see $f(x)$ remains constant at $\mu_1 N_c$ until $x \approx x_2 = \frac{1}{(1-\delta)N_c}$. In summary, as a very crude approximation, we broadly see

$$f(x) \approx f_{\text{crude}}(x) = \begin{cases} N_c(\mu_1-\mu_2) + \frac{\mu_2}{x(1-\delta)} & x < x_1, \\ \mu_1 N_c & x_1 \leq x < x_2, \\ \frac{\mu_1}{x(1-\delta)} & x_2 \leq x < 1. \end{cases}$$

The agreement between the crude approximation and $f(x)$ is shown in Fig. 5. The differing behaviour from the scenario with no change in mutation is the constant, middle regime in $f_{\text{crude}}(x)$. We see this will exist for $x \in [x_1, x_2)$. With a logarithmic base 10 x-axis the window for this different behaviour will be of size $\log_{10}(x_2/x_1) = \log_{10}(\mu_1/\mu_2)$. For the parameters of Fig. 5, $\log_{10}(\mu_1/\mu_2) \approx 2.08$, which is why the window enclosed by the red vertical lines spans 2 orders of magnitude.

9.2 Multiple mutation rate changes

Suppose the mutation rate alters at l thresholds $(N_c^{(i)})_{i=1}^l$ (positive integer valued and strictly increasing). For notational convenience we let $N_c^{(0)} = 0$, $N_c^{(l+1)} = N$. Once the population size surpasses $N_c^{(i)}$ for the first time, the mutation rate is set to μ_i, with $i = 1, \ldots, l+1$. In such a scenario, with an identical arguments to that of Section 9.1 (and under identical caveats), we obtain

$$f(x) = \sum_{i=1}^{l} \mu_i \frac{r(x)^{N_c^{(i-1)}+1} - r(x)^{N_c^{(i)}+1}}{x(1-\delta)} + \mu_{l+1} \frac{r(x)^{N_c^{(l)}+1}}{x(1-\delta)}.$$

References

[1] Athreya KB, Ney PE. Branching Processes. Dover Publications; 2004.

[2] Davoli T, Xu A, Mengwasser K, Sack L, Yoon J, Park P, et al. Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. Cell. 2013;155(4):948 – 962. Available from: http://www.sciencedirect.com/science/article/pii/S0092867413012877.

[3] McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of deleterious passenger mutations on cancer progression. Proceedings of the National Academy of Sciences. 2013;110(8):2910–2915. Available from: https://www.pnas.org/content/110/8/2910.

[4] Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nature Genetics. 2016 Aug;48:1119. Available from: http://dx.doi.org/10.1038/ng.364110.1038/ng.3641https://www.nature.com/articles/ng.3641#supplementary-information.

[5] Durrett R, Foo J, Leder K, Mayberry J, Michor F. Evolutionary dynamics of tumor progression with random fitness values. Theoretical Population Biology. 2010;78(1):54 – 66. Available from: http://www.sciencedirect.com/science/article/pii/S0040580910000444.

[6] Noble R, Team RDC. ggmuller: Create Muller Plots of Evolutionary Dynamics; 2019. R package version 0.5.4. Available from: https://CRAN.R-project.org/package=ggmuller.

[7] McDonald TO, Michor F. SIApopr: a computational method to simulate evolutionary branching trees for analysis of tumor clonal evolution. Bioinformatics. 2017 03;33(14):2221–2223. Available from: https://doi.org/10.1093/bioinformatics/btx146.

[8] Zhang Z, Lange K, Sabatti C. Reconstructing DNA copy number by joint segmentation of multiple sequences. BMC Bioinformatics. 2012;13(1):205. Available from: https://doi.org/10.1186/1471-2105-13-205.

[9] Bolker B, Team RDC. bbmle: Tools for General Maximum Likelihood Estimation; 2017. R package version 1.0.20. Available from: https://CRAN.R-project.org/package=bbmle.

[10] Bozic I, Gerold JM, Nowak MA. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. PLOS Computational Biology. 2016 02;12(2):1–19. Available from: https://doi.org/10.1371/journal.pcbi.1004731.

[11] Zhang Y, Li Y, Li T, Shen X, Zhu T, Tao Y, et al. Genetic Load and Potential Mutational Meltdown in Cancer Cell Populations. Molecular Biology and Evolution. 2019 01;36(3):541–552. Available from: https://doi.org/10.1093/molbev/msy231.

[12] Tlsty TD, Margolin BH, Lum K. Differences in the rates of gene amplification in nontumorigenic and tumorigenic cell lines as measured by Luria-Delbruck fluctuation analysis. Proc Natl Acad Sci U S A. 1989;86(23):9441–9445.

[13] Singer MJ, Mesner LD, Friedman CL, Trask BJ, Hamlin JL. Amplification of the human dihydrofolate reductase gene via double minutes is initiated by chromosome breaks. Proceedings of the National Academy of Sciences. 2000;97(14):7921–7926. Available from: https://www.pnas.org/content/97/14/7921.

[14] Bozic I, Antal T, Ohtsuki H, Carter H, Kim D, Chen S, et al. Accumulation of driver and passenger mutations during tumor progression. Proceedings of the National Academy of Sciences. 2010;107(43):18545–18550. Available from: https://www.pnas.org/content/107/43/18545.

[15] Durrett R. Population genetics of neutral mutations in exponentially growing cancer cell populations. Ann Appl Probab. 2013 02;23(1):230–250. Available from: https://doi.org/10.1214/11-AAP824.

[16] Nicholson MD, Antal T. Universal asymptotic clone size distribution for general population growth. Bull Math Biol. 2016;78:2243–2276.

[17] Therneau TM. A Package for Survival Analysis in R; 2020. R package version 3.1-11. Available from: https://CRAN.R-project.org/package=survival.

[18] Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578(7793):122–128. Available from: https://doi.org/10.1038/s41586-019-1907-7.

[19] Cross W, Kovac M, Mustonen V, Temko D, Davis H, Baker AM, et al. The evolutionary landscape of colorectal tumorigenesis. Nature Ecology & Evolution. 2018;2(10):1661–1672. Available from: https://doi.org/10.1038/s41559-018-0642-z.

[20] Haeno H, Iwasa Y, Michor F. The Evolution of Two Mutations During Clonal Expansion. Genetics. 2007;177(4):2209–2221. Available from: http://www.genetics.org/content/177/4/2209.

[21] Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics. 2018;50(6):895–903. Available from: https://doi.org/10.1038/s41588-018-0128-6.

[22] Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 2017;543(7643):122–125. Available from: https://doi.org/10.1038/nature21356.

[23] Durrett R, Moseley S. Evolution of resistance and progression to disease during clonal expansion of cancer. Theoretical Population Biology. 2010;77(1):42–48. Available from: http://dx.doi.org/10.1016/j.tpb.2009.10.008.

[24] Dinh KN, Jaksik R, Kimmel M, Lambert A, Tavare S. Statistical Inference for the Evolutionary History of Cancer Genomes. Statist Sci. 2020;35(1):129–144. Available from: https://projecteuclid.org:443/euclid.ss/1583226033.

[25] Cheek D, Antal T. Genetic composition of an exponentially growing cell population. Stochastic Processes and their Applications. 2020;130(11):6580 – 6624. Available from: http://www.sciencedirect.com/science/article/pii/S0304414920303033.

[26] Möhle M. Convergence Results for Compound Poisson Distributions and Applications to the Standard Luria-Delbrück Distribution. Journal of Applied Probability. 2005;42(3):pp. 620–631. Available from: http://www.jstor.org/stable/30040845.

[27] Keller P, Antal T. Mutant number distribution in an exponentially growing population. J Stat Mech P01011. 2015;(1). Available from: http://stacks.iop.org/1742-5468/2015/i=1/a=P01011?key=crossref.5a2f85cd93ef7e088940e9cbe2209bee.

[28] Scholz F, Zhu A. kSamples: K-Sample Rank Tests and their Combinations; 2019. R package version 1.2-9. Available from: https://CRAN.R-project.org/package=kSamples.

[29] Chkhaidze K, Heide T, Werner B, Williams MJ, Huang W, Caravagna G, et al. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. PLOS Computational Biology. 2019 07;15(7):1–26. Available from: https://doi.org/10.1371/journal.pcbi.1007243.

[30] Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature Genetics. 2016;48:238–244.

[31] Fusco D, Gralka M, Kayser J, Anderson A, Hallatschek O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. Nature Communications. 2016;7(1):12760. Available from: https://doi.org/10.1038/ncomms12760.

[32] Caravagna G, Heide T, Williams M, Zapata L, Nichol D, Chkhaidze K, et al. Model-based tumor subclonal reconstruction. bioRxiv. 2019. Available from: https://www.biorxiv.org/content/early/2019/03/26/586560.

[33] Iwasa Y, Nowak MA, Michor F. Evolution of resistance during clonal expansion. Genetics. 2006;172(4):2557–2566.

[34] Bozic I, Nowak MA. Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. Proceedings of the National Academy of Sciences. 2014;111(45):15964–15968. Available from: https://www.pnas.org/content/111/45/15964.