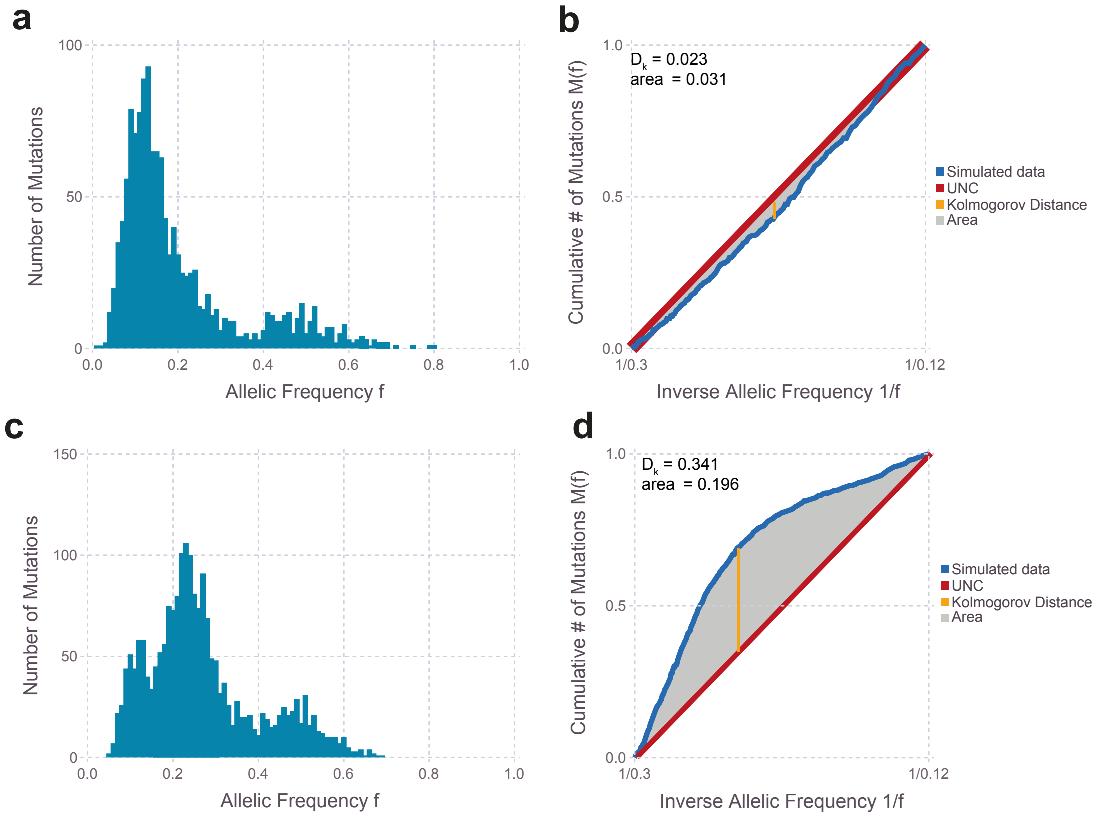
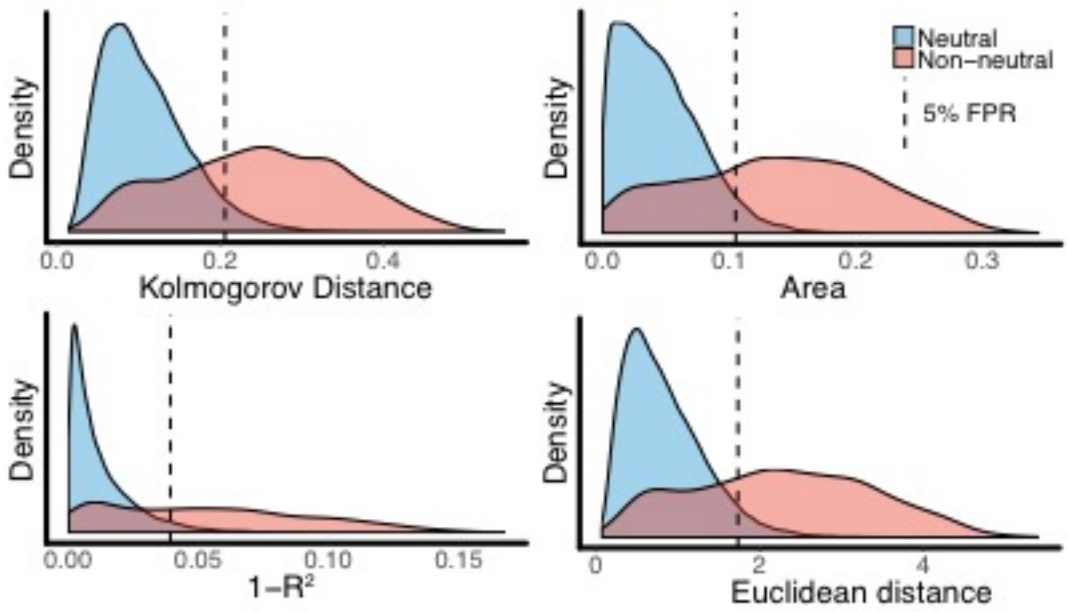
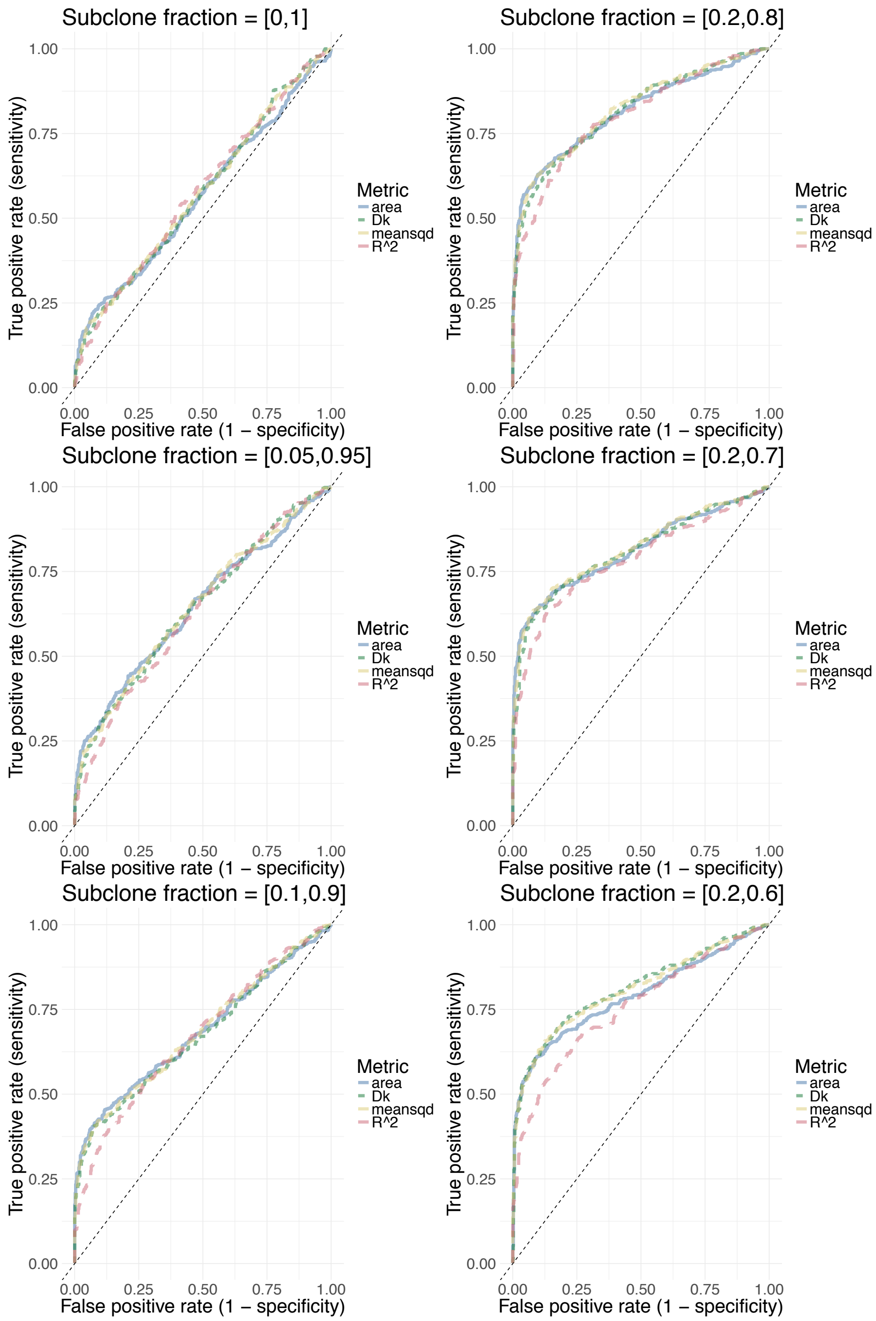


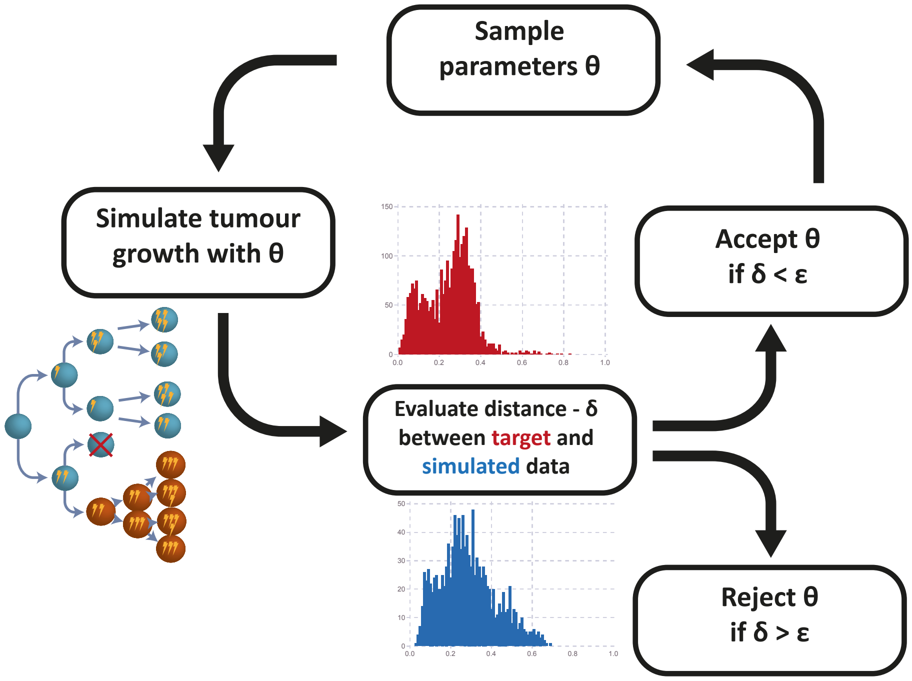
Supplementary Figure 1. New test statistics for neutrality. Representative example of VAF distribution in a neutral tumour **(a)** and a tumour with 1 selected subclone **(c)**. To optimise acceptance and rejection of the neutral ‘null’ model using the frequentist test we examined a number of test statistics where we compared the data (blue line) to the normalised distribution expected under neutrality (universal neutrality curve – UNC, red line), **(b,d).** We tested the area between the curves (AUC, shaded grey area), the Kolmogorov distance (orange line) and the Euclidean distance between all points (180 data points) on the two curves. These measurements improved the discrimination of non-neutral evolution over-and-above the R^2 method we previously proposed for neutrality testing.



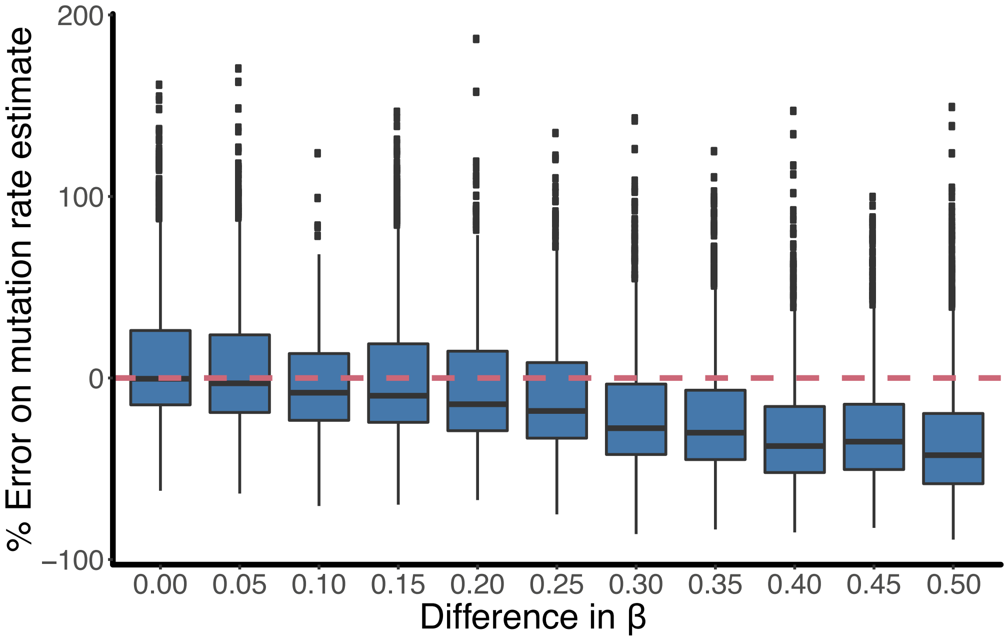
Supplementary Figure 2. Test statistics significantly differentiate neutral from non-neutral simulated cancers. Taking 10^5 neutral and 10^5 non-neutral simulations (100X simulated ‘sequencing’ depth) with a subclone with cancer cell fraction greater than 20% (VAF=10% in a diploid case) and smaller than 70% we confirmed that all metrics had significantly different distributions between the neutral and non-neutral cases, thus correctly distinguishing neutral from non-neutral dynamics (5% False Positive Rate reported as dash line).



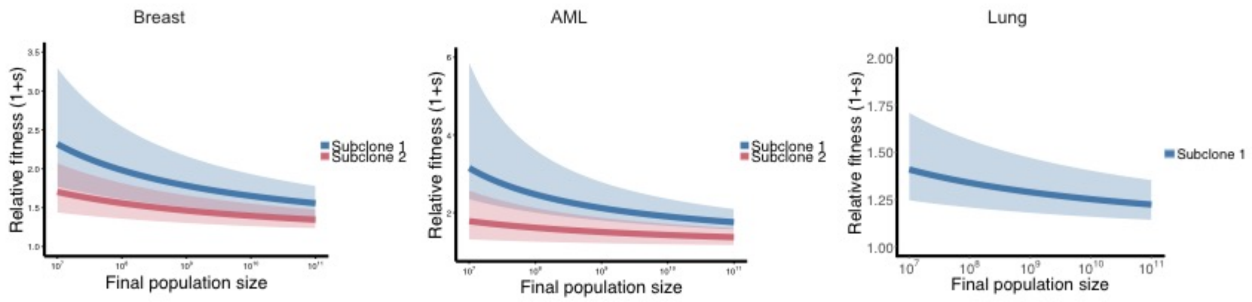
Supplementary Figure 3. ROC analysis for neutrality test. Receiver Operator Curve (ROC) analysis of the ability to correctly detect deviations from the neutral model for different VAF intervals considered. Accuracy depends on the fraction of the subclone over the total cancer cell population. If the subclone is very small (<20%) or very large (>80%), a frequentist test like this struggles to identify the correct model. This is because small subclones are hard to detect (VAF dominated by neutral tail of background clone) whereas large subclones which have almost reached fixation, an event that reverts the dynamics back to neutral (VAF dominated by neutral tail of the new subclone). The area test statistics performs best amongst all measures tested (largest area under the curve; see Supplementary Table 3). A cohort of 5000 synthetically generated tumors were used for this analysis. In the main text we present a Bayesian model selection method to overcome the problem of selecting a range in the VAF distribution.



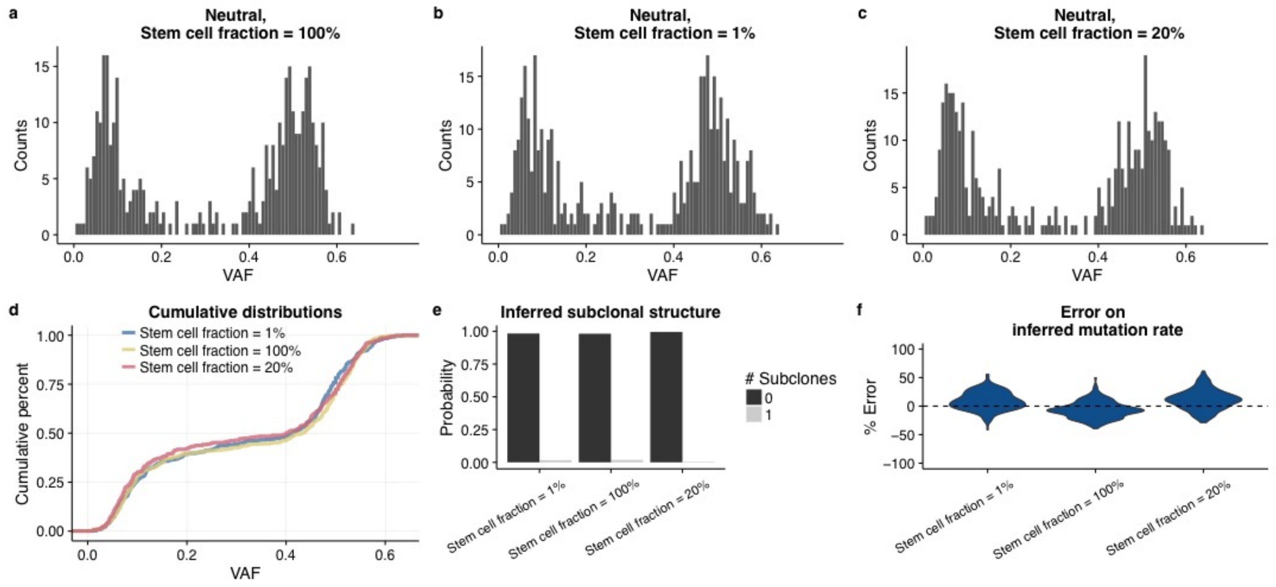
Supplementary Figure 4. Approximate Bayesian Computation (ABC) framework. We used ABC to fit our model to the data. Parameters are drawn from a prior distribution and then a tumour is simulated using those parameters. We then measured the distance between the simulated data and the target data using the Euclidean distance between the cumulative distributions from the two datasets. When the discrepancy, δ between the target data and simulated data is lower than a given ε, we accept these parameters. We used an extension of this basic ABC-rejection algorithm called ABC sequential Monte Carlo (ABC-SMC, ref[51,52] in the main text) where rather than repeatedly sampling from the prior, we sample from the prior once and accept a set of N particles (parameter sets) which produce simulated datasets with rather large discrepancies. We then sample from this set of particles and perturb the parameter values until we achieve a lower discrepancy δ, we continue through multiple rounds of this procedure lowering the ε value at each iteration, and thereby gradually evolving toward the posterior.



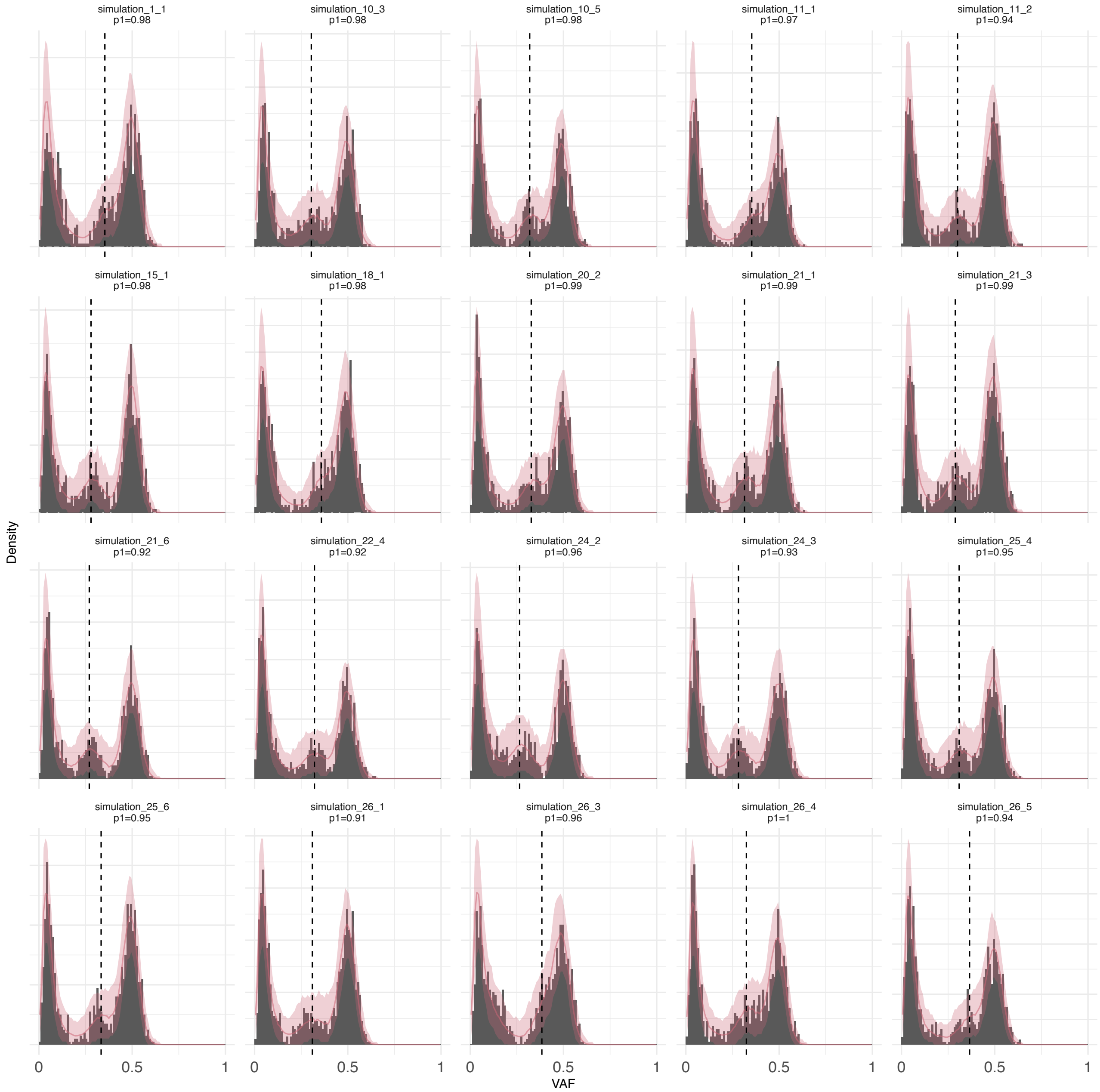
Supplementary Figure 5. Accurate estimation of the mutation rate in the presence of subclonal selection. Evaluation of the accuracy of mutation rate inference for differing levels of subclone selection. Differential subclonal selection was model by altering the probability β that a new lineage survives, and expressed in terms of $\Delta\beta$, the difference in lineage survival between the subclone and rest of the tumour population. Mutation rate inference was performed by fitting the linear 1/f cumulative model to the left hand neutral tail. Because in the presence of selection, the 1/f tail is the combination of the neutral tail of the background clone and the new neutral tail of the selected clone, one expects an error in the estimation of the mutation rate. The % error on the inferred mutation rate increased as the strength of subclonal selection increased (larger $\Delta\beta$), but the mean error was less than 50% even when selection was very strong ($\Delta\beta = 0.5$). This is reasonable when the aim is to measure mutation rates in humans with a level of precision of an order of magnitude. 100 simulated tumours were used for each $\Delta\beta$. Boxplots show the median and inter quantile range (IQR), upper whisker is 3rd quantile + 1.5*IQR and lower whisker is 1st quantile - 1.5*IQR.



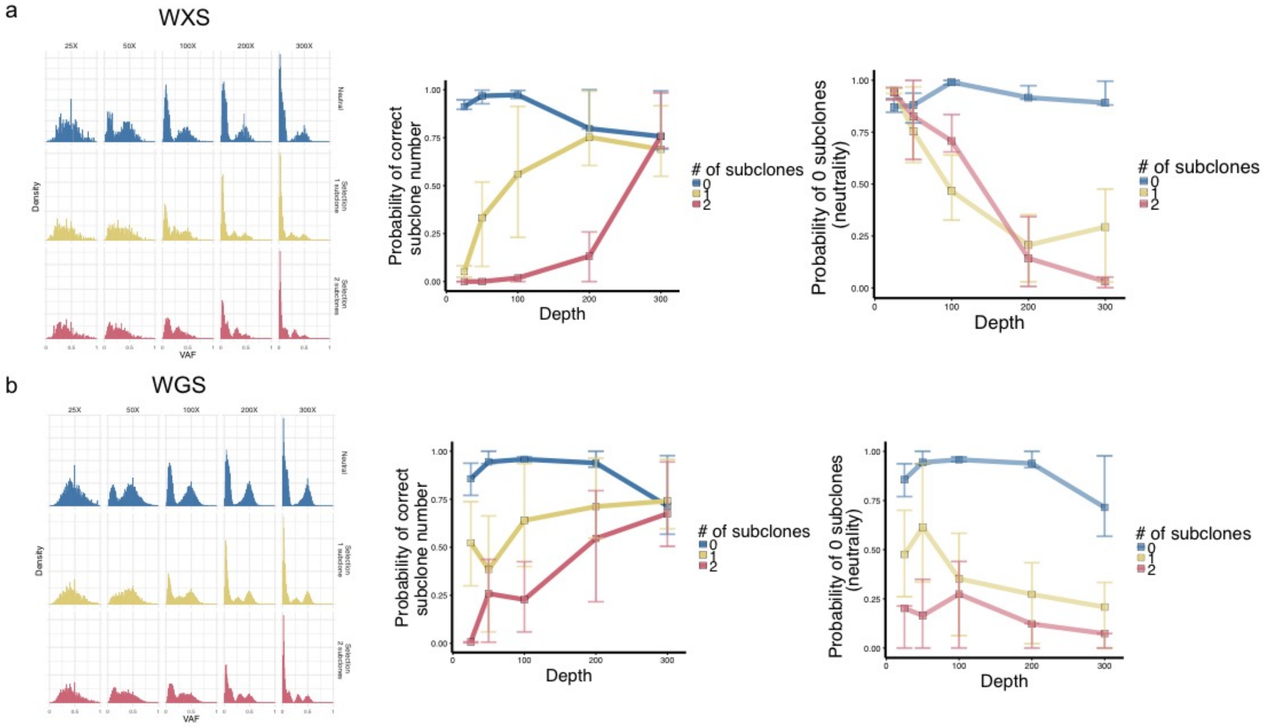
Supplementary Figure 6. Sensitivity of selective advantage estimates in relation to final population size. The estimates of the relative fitness advantage parameter *s* depend on the assumption of the final population size N_{end}, which is often not known with precision for any given tumour. Posterior distribution for the relative fitness as a function of N_{end} for AML, breast and lung cancer cases demonstrate that within the realistic range of N_{end}, *s* is not sensitive to the precise value of N_{end} (due to the properties of exponential growth). This confirms the robustness of our estimated values. 500 posterior samples were used for the inference. Solid lines indicate median values, shaded area is 95% intervals.



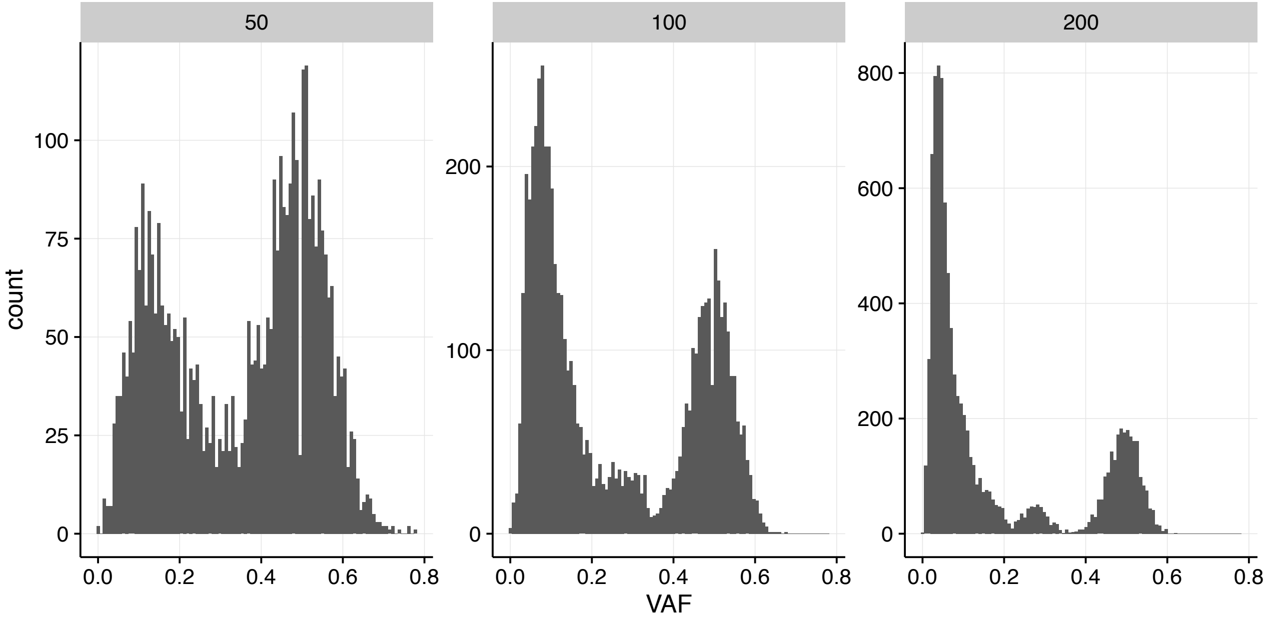
Supplementary Figure 7. Stem cell model is equivalent to exponential growth model in terms of VAF distributions. Using a two compartment stem cell model (long lived stem cell lineages that generate non-stem cell progeny, which undergo a restricted number of divisions) we simulated neutral tumour growth and generated synthetic datasets where the stem cell fraction was 100% (372 mutations) **(a)**, 1% (397 mutations) **(b)** and 20% (405 mutations) **(c)**. Plotting the cumulative distributions shows that the VAF distributions generated from these models are indistinguishable (p=0.41, p=0.20, p=0.65 by Kolmogorov-Smirnoff test), **(d)**. Using the ABC (500 posterior samples) inference a neutral exponential pattern of growth with 0 subclones captures the data well **(e)** and accurately measures the mutation rate **(f)**.



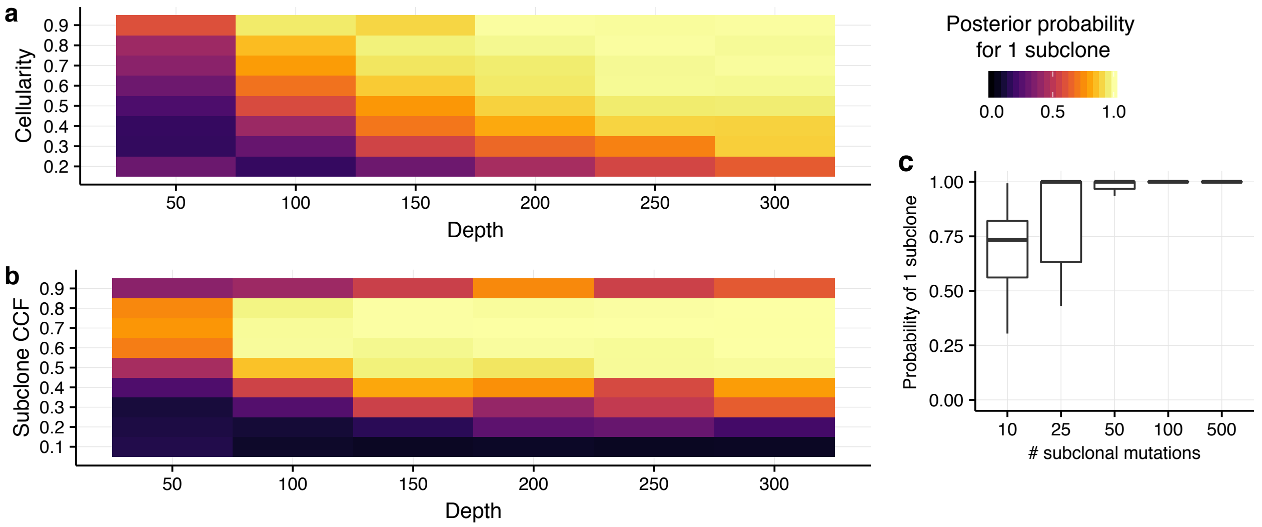
Supplementary Figure 8. Simulated VAF distributions containing one subclone with ABC fits. Random sample of 20 simulated VAF distributions with selected subclone (from cohort of 100 simulated tumours) and ABC fits, data used in Figure 2h. Grey histograms are empirical VAF distributions from simulations, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Dashed line shows the inferred VAF of the subclonal cluster which overlaps with the centre of the subclonal cluster.



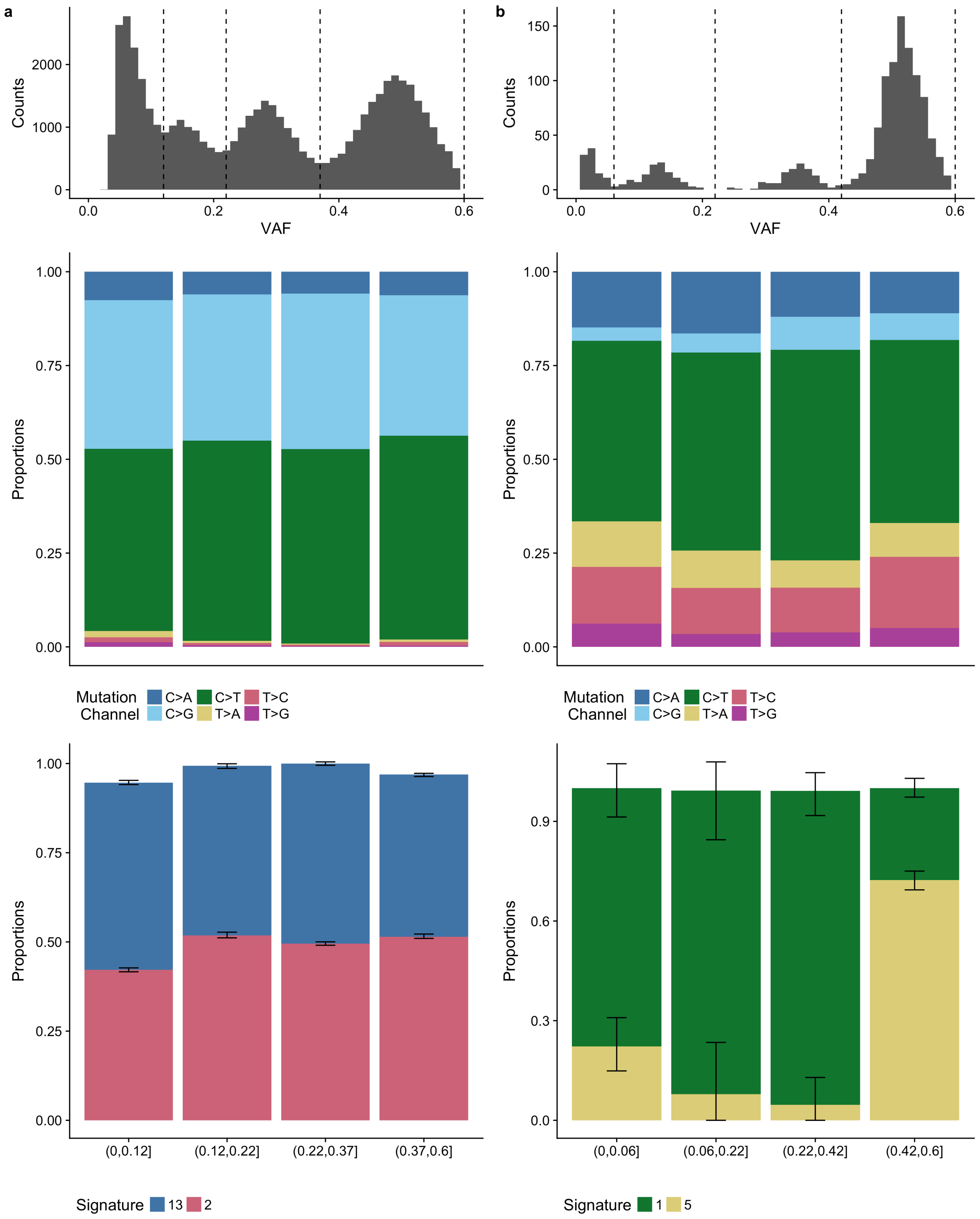
Supplementary Figure 9. Limited sequencing depth (<100X) obscures the correct clonal structure. Synthetic data were generated using different simulated sequencing depths for **(a)** exome sequencing (WXS) and **(b)** whole genome sequencing (WGS). Input parameters were chosen so that the WGS data had a mutation load of 10,000 mutations and WXS had a mutation load of 400 mutations at 100X depth, in line with results from genomic profiling studies. Applying our inference method to this data showed that a depth >100X was required to confidently identify more than 1 subclone for both WGS and WXS. WGS provides a denser VAF plot with respect to WXS, but the precision of the VAF mutational clusters is not improved (as it is dependent mainly on the sequencing depth). This indicates that for subclonal architecture deconvolution, the depth of sequencing is a critical factor. We note that the problem of low sequencing depth can be further exacerbated by low tumour purity in the sample. WGS may still be necessary for low mutational burden tumours in order to detect enough mutations to delineate subclone clusters. Each data point is the average probability from applying the method to 10 different simulations with the same parameters. Error bars show the minimum and maximum values from the 10 simulations.



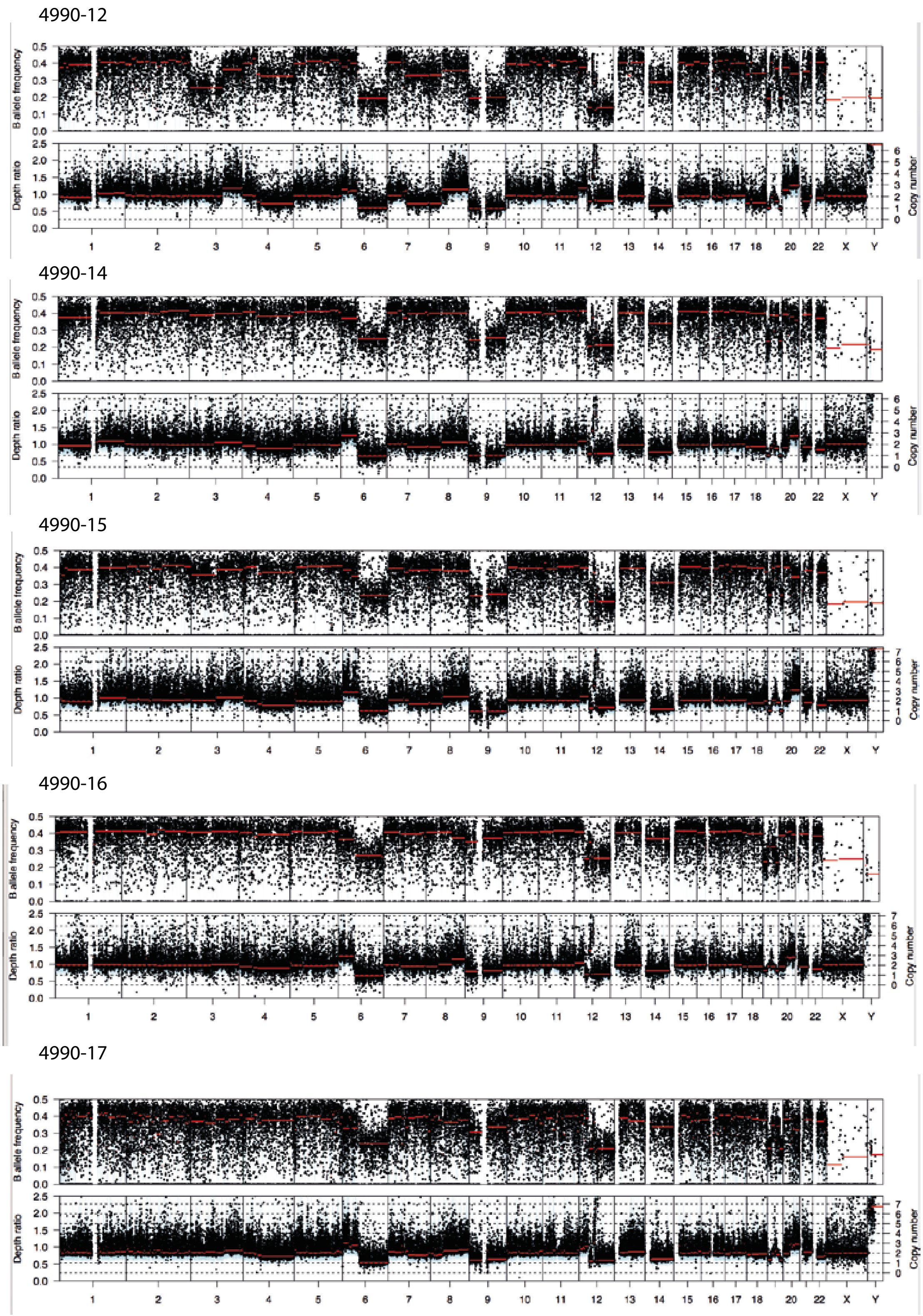
Supplementary Figure 10. High sequencing depth can reveal clusters that are due to genetic drift and not selection. Sequencing to very high depth can reveal apparent clusters arising due to stochastic neutral drift rather than selection. This occurs when a lineage increases in frequency just due to stochastic birth/death effects generating a non-selected subclonal cluster. The same neutral simulation sampled to different sequencing depths shows that a small cluster at VAF~0.25 becomes evident as the depth of sequencing increases. We note that, because genetic drift affects prevalently small populations, this is relevant only during the very early phase of tumour growth when the neoplasm is very small. This implies that too little time has passed for the accumulation of many mutations in a drifting subclonal cluster, and hence, contrary to selected clusters, drifting clusters are expected to be very small in terms of number of mutations.



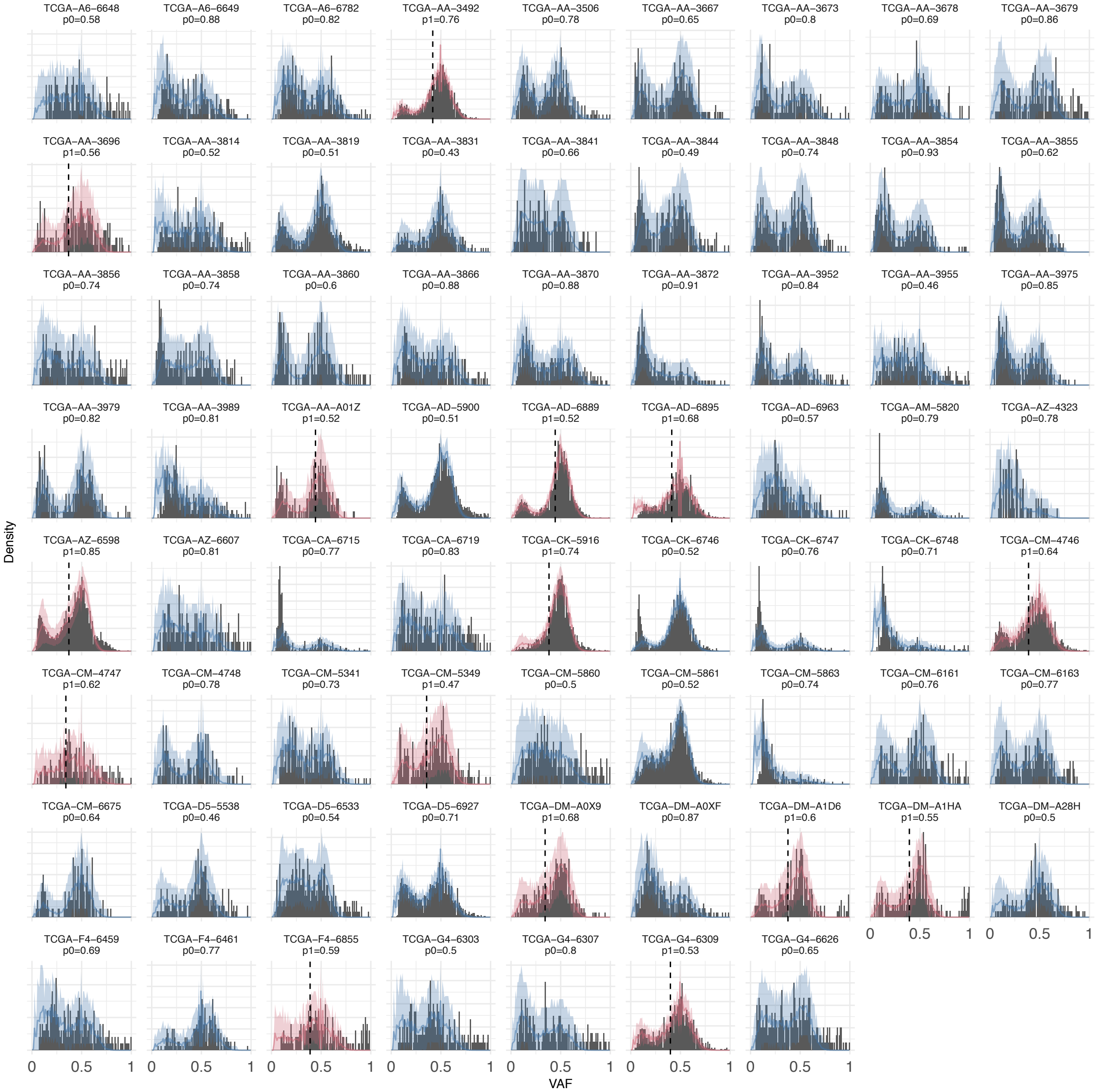
Supplementary Figure 11. Ability to detect subclones depend on cellularity and subclone fraction. We simulated tumours with a single subclone at various cancer cell fraction and with a range of cellularity, then the ABC inference was applied and posterior probability of a single subclone recorded. As expected, the ability to resolve a subclone in the VAF distribution increases with increasing depth for low cellularity samples **(a)** and low frequency subclones **(b)**. We note that, as shown in panel **(b)**, if a subclone has cancer cell fraction >0.9, it becomes difficult to distinguish it from the clonal cluster. Each coloured square is the average posterior probability from 25 simulations with the same input parameters. Subsampling the number of mutations per sample showed that a minimum of 25 subclonal mutations was needed to confidently identify a subclone **(c)**. Boxplots show the median and inter quantile range (IQR), upper whisker is 3rd quantile + 1.5*IQR and lower whisker is 1st quantile - 1.5*IQR.



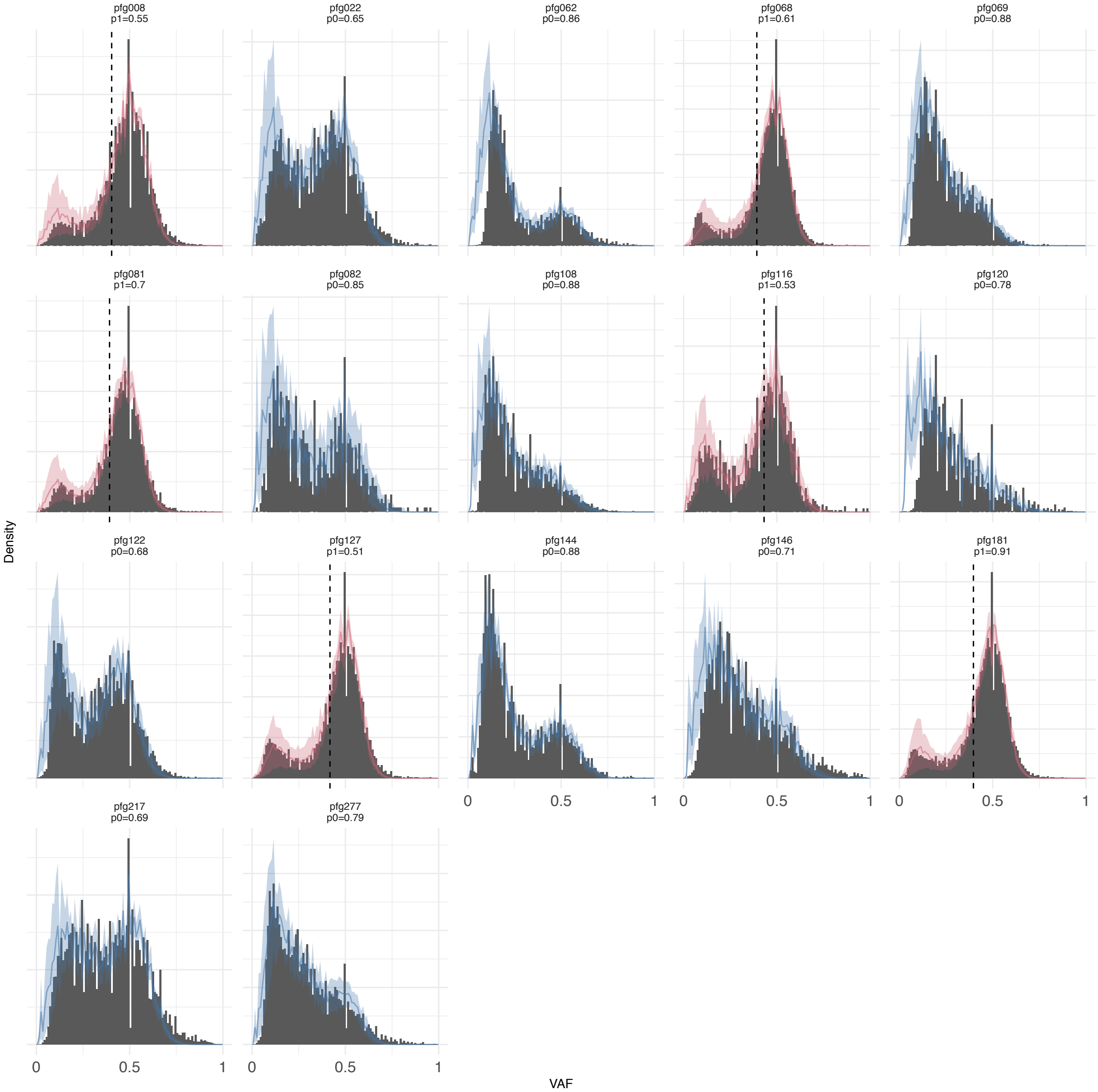
Supplementary Figure 12. Mutational processes are consistent throughout the subclonal frequency spectrum. All subclonal mutations in the breast cancer sample (64,677 mutations, ref[13]) **(a)** and AML sample (1302 mutations, ref[20]) **(b)** showed consistent mutational processes for both subclonal clusters and 1/f tail mutations. This supports our assumption of approximate constant mutation rate during the final growth of the tumour. Signatures for the clonal AML mutations appeared different but this has no impact on our inferences as we only consider subclonal mutations in our analysis. Error bars on the signature assignment were obtained via bootstrapping and show the 95% interval and coloured proportions show the mean value.



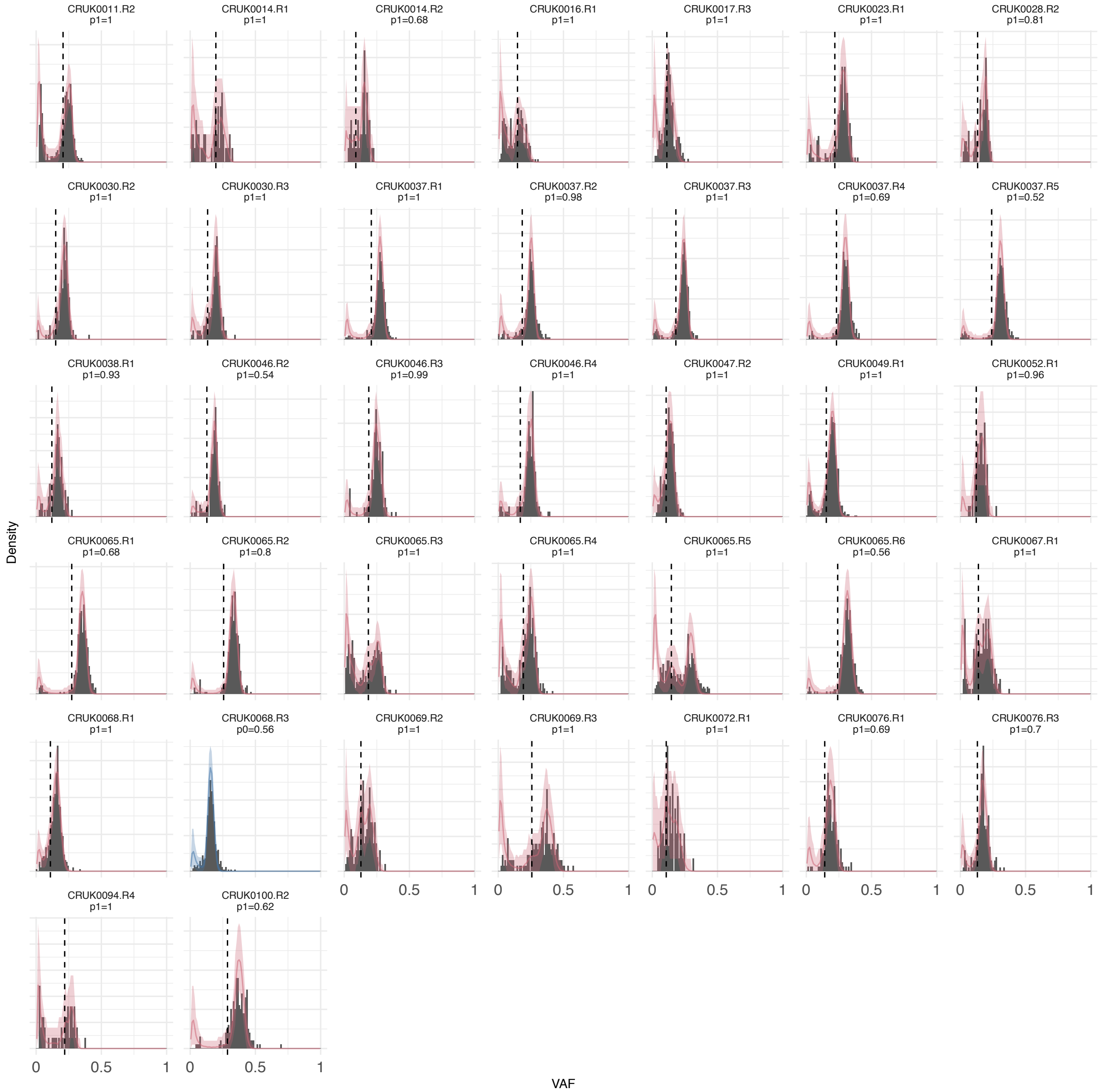
Supplementary Figure 13. Lung cancer copy number profiles from Zhang et al. 2014. Copy number profiles for the 5 lung adenocarcinoma samples (ref[21] in the main manuscript). Sample 4990-12, that was measured to contain a differentially selected subclone, has a CNA not present in the other samples (chromosome 3), suggesting that such alteration may confer a fitness advantage.



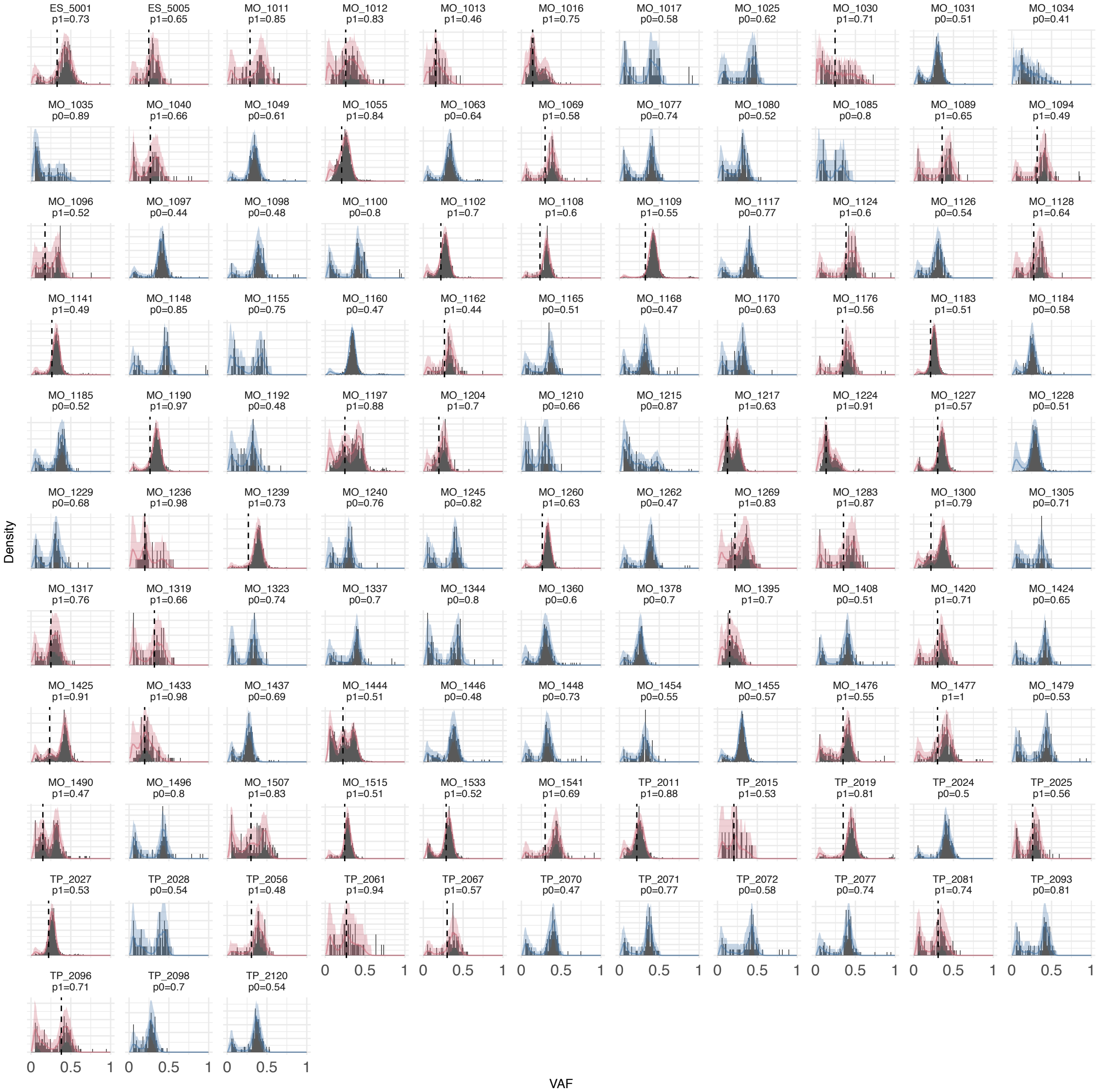
Supplementary Figure 14. TCGA colon cancer model fits. Grey histograms are empirical VAF distributions, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Blue tumours are those identified as neutral and red are those with subclonal selection. For those with a subclone, dashed line shows the mean of subclonal fraction. Title of each panel shows sample name and probability of the assigned subclonal structure. Only samples suitable for our analysis (purity >40%, number of subclonal mutations ≥25) are considered.



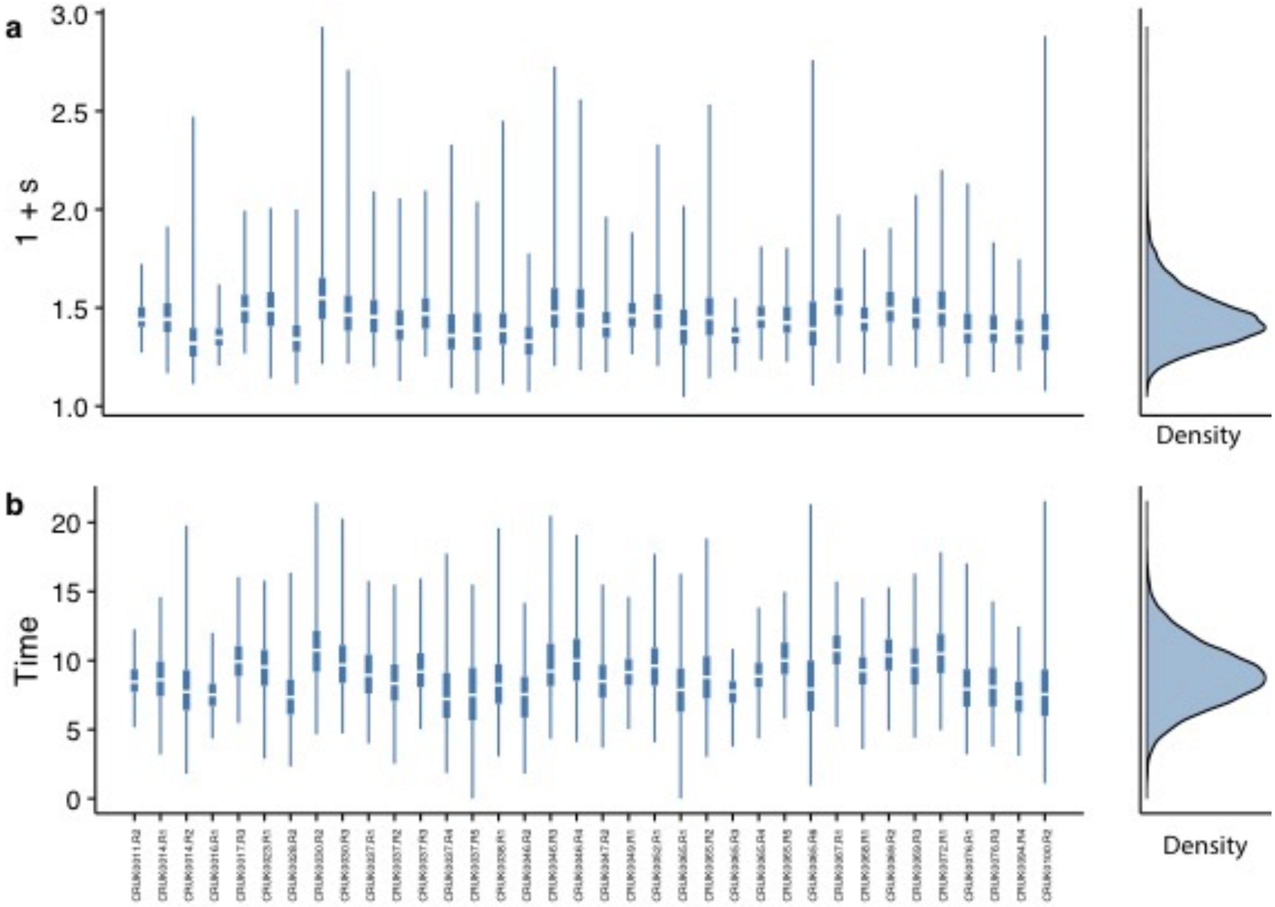
Supplementary Figure 15. Gastric cancer model fits. Model fits to WGS gastric cancer data summarised in Figure 4 from Wang et al. 2014 (ref[23] in the main manuscript). Grey histograms are empirical VAF distributions, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Blue tumours are those identified as neutral and red are those with subclonal selection. For those with a subclone, dashed line shows the mean of subclonal fraction. Title of each panel shows sample name and probability of the assigned subclonal structure. Only gastric samples suitable for our analysis (see main text) are presented. Only samples suitable for our analysis (purity >40%, number of subclonal mutations ≥25) are considered.



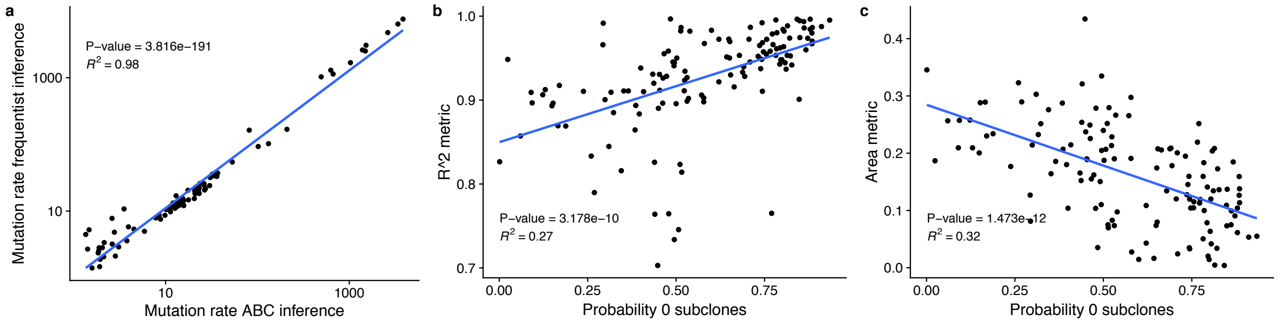
Supplementary Figure 16. TRACERx cancer model fits. Model fits to lung cancer TRACERx data from Jamal-Hanjani et al. 2017 (ref[24] in the main manuscript). Grey histograms are empirical VAF distributions, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Blue tumours are those identified as neutral and red are those with subclonal selection. For those with a subclone, dashed line shows the mean of subclonal fraction. Title of each panel shows sample name and probability of the assigned subclonal structure. Only samples suitable for our analysis (purity >40%, number of subclonal mutations ≥25) are considered.



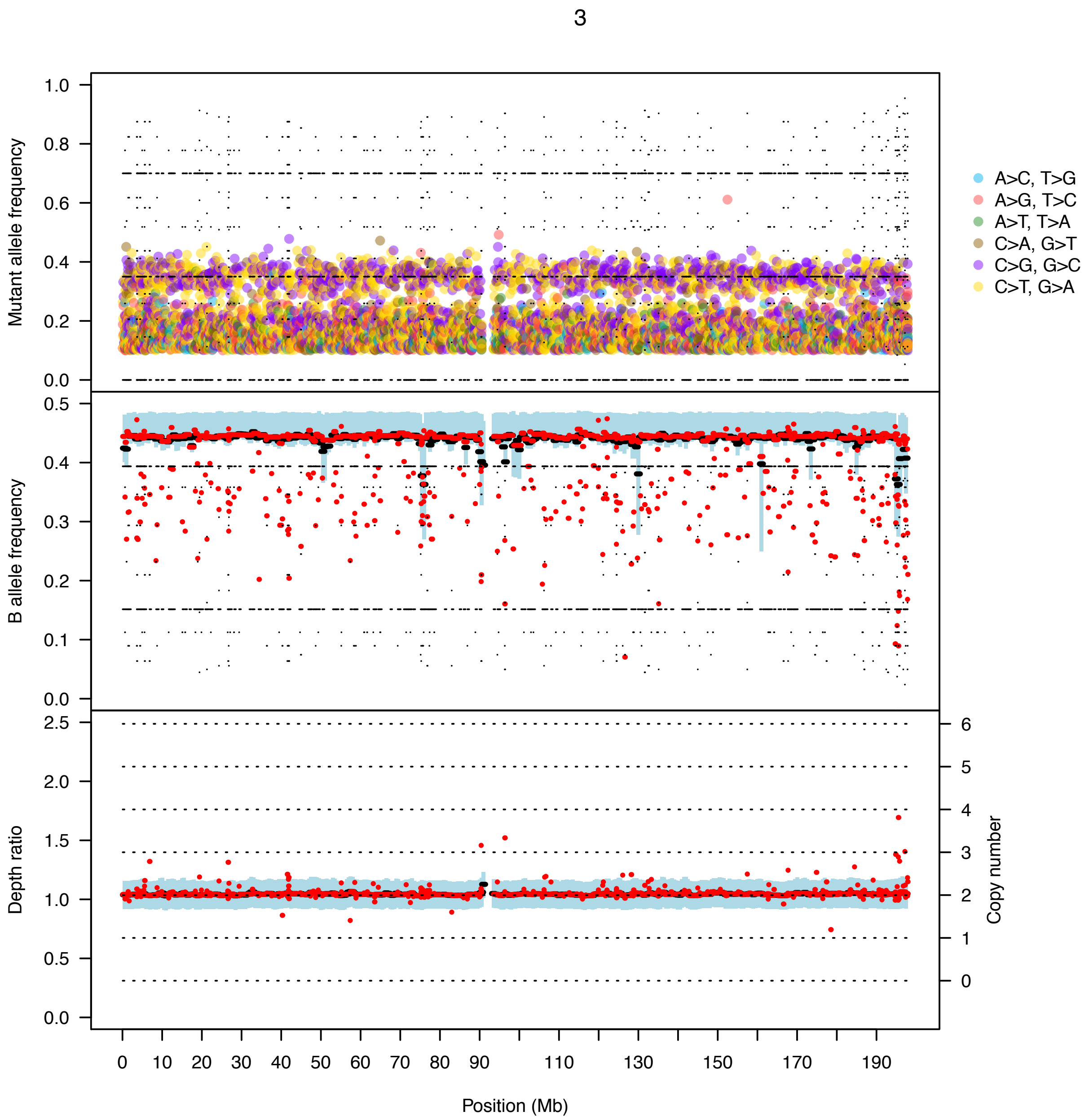
Supplementary Figure 17. MET500 metastatic deposits model fits. Model fits to MET500 data summarised in Figure 4 from Robinson et al. 2017 (ref[25] in the main manuscript). Grey histograms are empirical VAF distributions, line is mean value from 500 simulations that fitted the data and shaded area is 95% interval. Blue tumours are those identified as neutral and red are those with subclonal selection. For those with a subclone, dashed line shows the mean of subclonal fraction. Title of each panel shows sample name and probability of the assigned subclonal structure. Only samples suitable for our analysis (purity >40%, number of subclonal mutations ≥25) are considered.



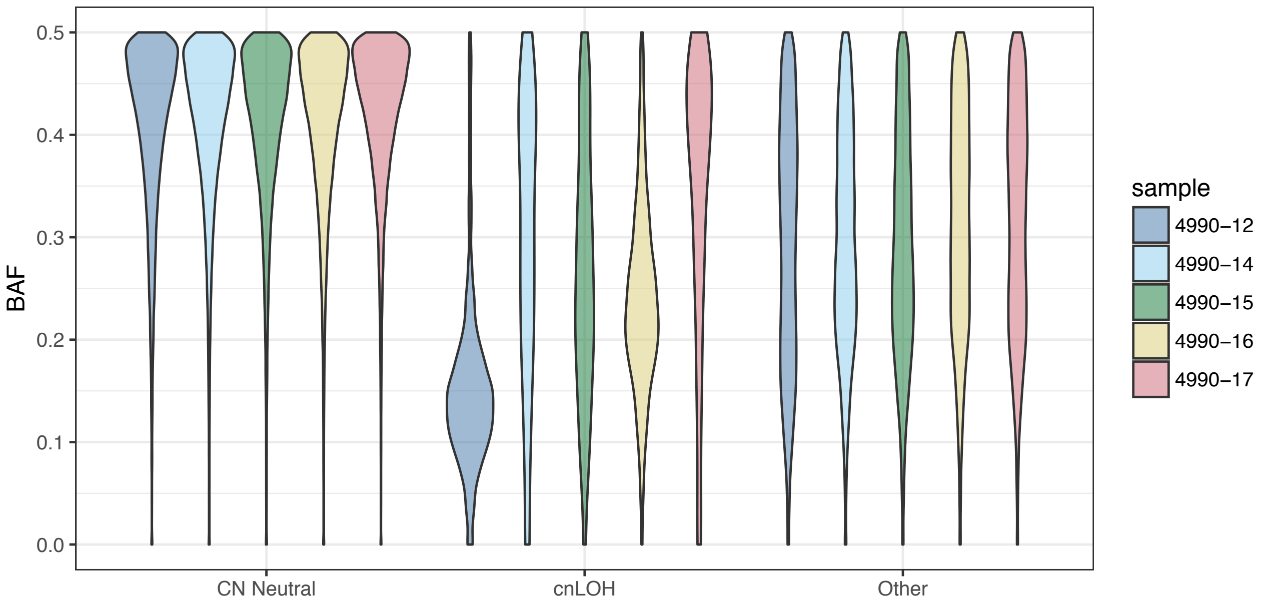
Supplementary Figure 18. Subclone parameter estimation for the TRACERx cohort. Time of emergence and fitness of subclones estimated for subclones present in the TRACERx data. Posterior distributions were generated from 500 samples. Boxplots show the median and inter quantile range (IQR), upper whisker is 3rd quantile + 1.5*IQR and lower whisker is 1st quantile - 1.5*IQR. Only samples suitable for our analysis (purity >40%, number of subclonal mutations ≥25) are considered.



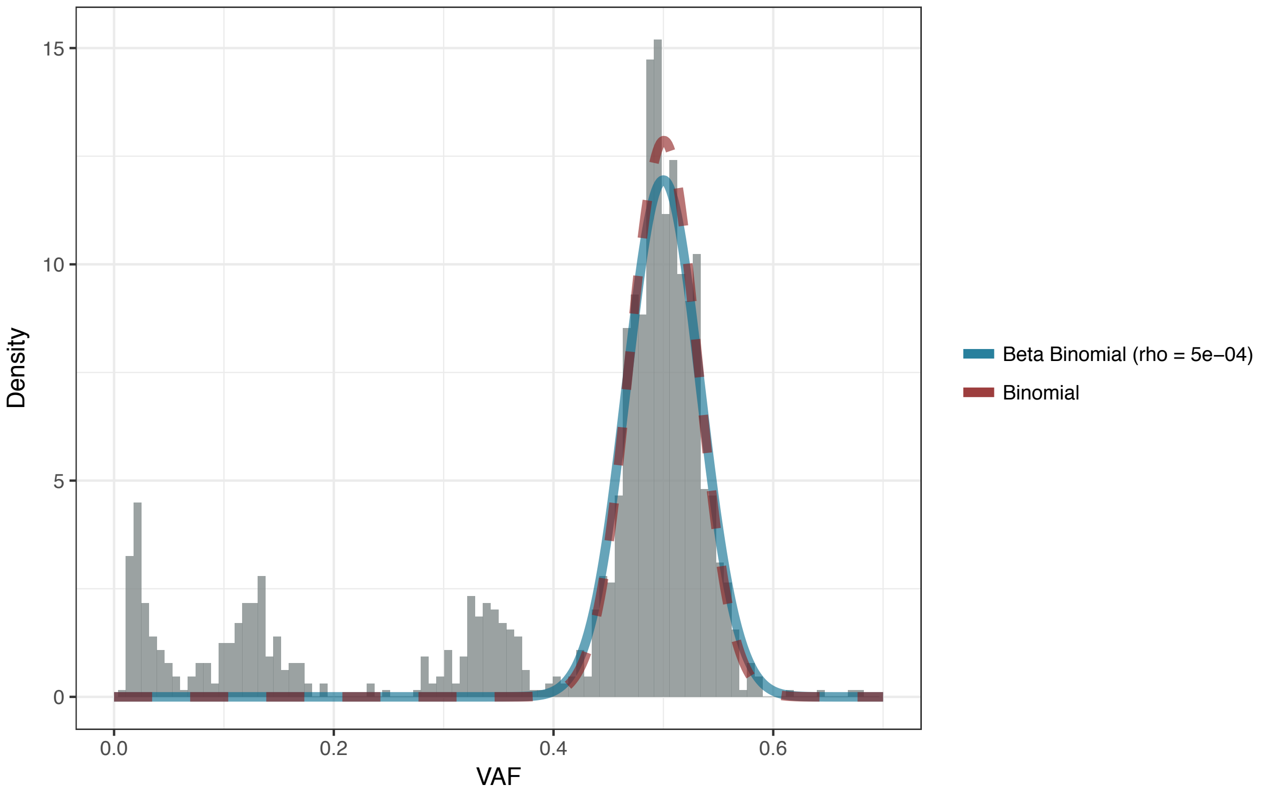
Supplementary Figure 19. Frequentist vs Bayesian neutrality classification. Comparison of measurement of neutrality between our previously reported analytical 1/f model and the posterior distributions from fitting our computational Bayesian approach. We used samples that had cellularity >0.8 so that the integration limits (0.1, 0.3) for the 1/f test was not influenced by the clonal peak. **(a)** Inferred mutation rates were highly similar in both cases for neutral tumours (N=87). **(b,c)** The probability of 0 subclones (neutrality) were significantly correlated (linear regression, p-values and R^2 values shown in figure insets) between methods using both the R^2 and the area between curves metric respectively, outliers are due to the ABC method detecting high frequency clones which we are unable to resolve using frequentist neutrality test statistics. A total of 130 samples were used for this analysis.



Supplementary Figure 20. Mutations used in the breast cancer high-depth sample from Nik-Zainal et al. 2012 were from truly diploid regions. B-Allele frequency plot and depth ratio plot for chromosome 3 for breast cancer sample (ref[13] in the main manuscript), demonstrating that this chromosome was consistent with a diploid genome. Only mutation from this chromosome were used in the analysis.



Supplementary Figure 21. Mutations used in the lung cancer samples from Zhang et al. 2014 were from true diploid regions. B-Allele Frequency (BAF) distributions stratified by copy number calls for lung cancer samples (ref[21] in the main manuscript) showing BAF was consistent with segments called as diploid, regions called as copy number LOH (cnLOH) and other copy number aberrations had different BAF distributions. For each sample the sequenza algorithm (ref[53] in methods) split the genome into the following number of segments 61,323 (4990-12) 58,289 (4990-14) 57,982 (4990-15) 58,828 (4990-16) 60,764 (4990-17).



Supplementary Figure 22. A binomial model well describes mutational clusters. We used Monte Carlo Markov Chain to fit beta-binomial and binomial models to the clonal cluster of the AML whole-genome sample (to avoid potential confounding factors from near-clonal CNAs in the left-hand tail of the distribution). We found that the binomial model was sufficient to describe the VAF dispersion in mutational clusters. Overdispersion parameters ρ, for these fits are reported in Supplementary Table 6.

Supplementary Table 1. Power to detect non-neutral VAF distribution. Power to detect non-neutral VAF distribution for 4 different test statistics at a FPR rate of 0.05 and at variable read depths and clone size.

Supplementary Table 2. False Positive Rate cutoff. Cut off values for the four different test statistics evaluated that give false positive rate (FPR) of 0.05 to discriminate a VAF distribution from the null neutral distribution.

Supplementary Table 3. Receiver Operator Curve area under the curve analysis. Area under curves from receiver operator curves (ROC) analysis of the four different test statistics at variable read depths and clone size.

Supplementary Table 4. Bayes Factors for Figure 3. Bayes factors and probabilities of 0, 1 and 2 clones for data in Figure 3.

Supplementary Table 5. Bayes Factors for Figure 4. Bayes factors and probabilities of 0, 1 and 2 clones for data in Figure 4.

Supplementary Table 6. Beta-Binomial model fits for clonal cluster. Number of clonal mutations and over-dispersion parameter ρ (MAP estimate) from fitting Beta-Binomial model to clonal cluster using MCMC.

Supplementary note for quantification of subclonal selection in cancer from bulk sequencing data

Contents

1 Detecting and quantifying selection in subclones 2
1.1 Number of mutations in a subclone 3
1.2 Fitness advantage of a subclone 5
1.3 Multiple subclones 6
1.3.1 Nested subclones 7

2 Expected variation in parameter inference 8
2.1 Stochastic birth death model 8
2.2 Stochastic birth-death model conditioned on N 9

3 Driver mutations 10

4 Statistical inference 11
4.1 Bayesian inference 12
4.1.1 ABC rejection 13
4.1.2 ABC SMC algorithm 14
4.2 Prior distributions 15
4.3 Algorithm convergence 16

5 Alternative growth models 20
5.1 Logistic growth 20
5.1.1 Modeling logistic growth 20
5.1.2 Moran model 21
5.2 Power law growth 22

1 Detecting and quantifying selection in subclones

Previously we showed that under a neutral evolutionary model, the variant allele frequency (VAF) distribution of passenger mutations that accumulate as the tumour grows collapses to a predictable form [1], where the cumulative number of mutations, $M(f)$ with a frequency greater than f is given by

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right), \tag{1}$$

where μ is the mutation rate per division, β is the probability at cell division of generating 2 surviving offspring (to reflect cell death and differentiation), and f_{max} is the maximum VAF possible in the spectrum given the tumour ploidy (e.g. for a diploid genomic region, $f_{max} = 0.5$). With the knowledge of what we would predict the frequency of mutations to be in a neutral population, in this paper we wished to identify the presence and strength of selection in non-neutral tumours, again using the VAF distribution.

We wish to quantify the strength of selection of one (or more) subclones growing in a tumour, using information we can extract from a single sample at a single time point. We also aim at measuring the mutation rate, the number of mutations in a subclone and the fraction of the subclone in the population at the same time. Our approach is summarised in Figure 1D (main text), where the number of mutations in the subclone tells us about the age of the subclone - i.e. the number of mutations accumulated between the most recent common ancestor (MRCA) of all cells in the tumour and the the single cell that gave rise to the subclone (MRCA of the cancer cells in the subclone) is proportional to the number of generations between these two MRCAs, and the fraction of the subclone in the population tells us how much the subclone expanded from that single cell.

First we will introduce some of the assumptions and the notation. We will assume exponential growth for all our populations, with the fitness advantage of a population over the host population defined as the ratio of net growth rates between the fitter population (λ_1) and the host tumour population (λ_H).

$$1 + s = \frac{\lambda_1}{\lambda_H} \tag{2}$$

Given birth and death rates b and d we can express this as:

$$1 + s = \frac{b_1 - d_1}{b_H - d_H} \tag{3}$$

The effective mutation rate $\hat{\mu}$ is given by the mutation rate per generation (e.g. number of mutations per genome per division) divided by the probability of having 2 surviving offspring (β), $\hat{\mu} = \mu/\beta$, ie the prefactor in equation (1). β in terms of the birth and death rates can be written as $\beta = \frac{b-d}{b}$. The effective mutation rate $\hat{\mu}$ is important because it is the only measurable value from the type of data we will consider, and represents the mutation rate per tumour population doubling. The mutation rate per cell division remains obscured in the data because μ and β cannot be estimated independently (see below).

1.1 Number of mutations in a subclone

First of all, we wish to know how the number of mutations in a subclone relates to the time the subclone arose. The subclonal cluster is primarily composed by mutations present in the single cell that gave rise to the subclone, i.e. the most recent common ancestor of the subclone. This number depends on the number of generations that passed between the single cell that founded the whole tumour population (MRCA of the whole tumour) and the appearance of the single cell that will then give rise to the subclone later. The number of mutations in a cell at some time t_1 is given by the product of the mutation rate and the number of divisions. Assuming the tumour starts growing at time $t = 0$ the number of mutations, M_c in a cell after Γ_1 successful divisions (at time t_1) is given by:

$$M_c = \hat{\mu}\Gamma_1 \tag{4}$$

Note that this Γ_1 is in effect the number of divisions assuming no cell death, as cell death is incorporated into $\hat{\mu}$, which is the effective mutation rate, or the true mutation rate scaled by the death rate. What remains is to relate the number of divisions Γ_1 to the time t_1 in terms of tumour doublings. We will consider a birth death model with birth rate b and death rate d. If $D_i(t)$ is the total number of cells that have completed i divisions at time t, then we can write a set of differential equations where $D_i(t)$ increases or decreases based on the rates b and d, similar to the approach taken in [2], which is a classical birth-death process.

$$\begin{aligned}
\frac{dD_0(t)}{dt} &= -(b+d)D_0(t) \\
\frac{dD_i(t)}{dt} &= -(b+d)D_i(t) + 2bD_{i-1}(t)
\end{aligned} \tag{5}$$

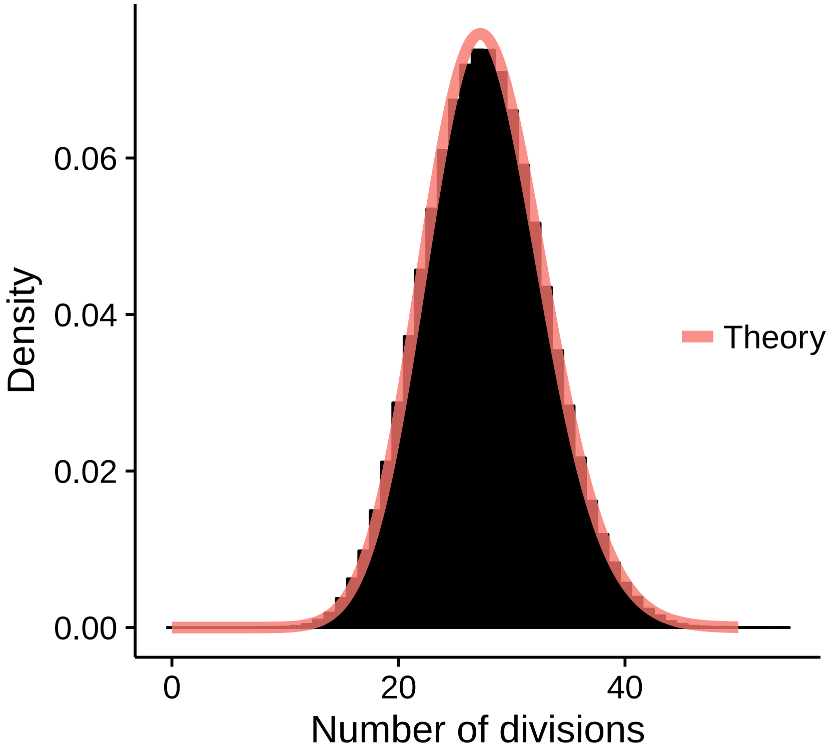


Figure 1: *Distribution of number of divisions for a tumour growing to size* 2^{20} *from simulation. Red line is the theoretical distribution predicted by* (7)

Intuitively, we can think of each D_i being a compartment, and b and d being the rates with which cells move from one compartment to another or are lost from the system. Cells can be lost from a compartment via death or via a birth event (hence the factor $(b+d)$), where for each birth event in compartment $i-1$, 2 cells are gained in compartment i. Equations (5) have the following solution, given the boundary condition $N_0(0) = 1$.

$$\begin{aligned} D_0(t) =& e^{-(b+d)t} \\ D_i(t) =& \frac{(2bt)^i}{i!} D_0(t) \end{aligned} \tag{6}$$

Given that the total population grows as $e^{(b-d)t}$, the pdf can be derived by dividing equation (6) by this expression.

$$P_i(t) = \frac{(2bt)^i}{i!} e^{-2bt} \tag{7}$$

Equation 7 is a poisson distribution with mean $2bt$. Supplementary Note Figure 1 shows the distribution of the number of divisions from a stochastic simulation where a tumour is grown to size 2^{20} along with the theoretical predictions from equation (7).

If t is in units of tumour doublings then $b = log(2)$, then the mean number of divisions, Γ_1 experienced by a cell after time t_1 is:

$$\Gamma_1 = 2\log(2)t_1 \tag{8}$$

Therefore in our framework where the number of divisions experienced by the founder cell of a subclone is measured, we can estimate also the most probable time (in tumour doublings) when the subclone emerged. We note that as Supplementary Note Figure 1 shows the distribution of the number of divisions is wide and by chance the single cell that gives rise to the subclonal population could be in the extremities of this distribution. In these cases we would over or underestimate the time the subclone emerges in terms of population doublings given we have a single measurement and not a distribution (number of mutations carried by the subclone). However given that this distribution is symmetric, on average the error should be 0. Figure 2E shows that on average we do manage to capture the correct time, however the large deviation from the true value in some cases will be partly due to this effect.

1.2 Fitness advantage of a subclone

To calculate the fitness advantage of a subclone we utilize information on the frequency of the subclone and assume the subclone starts from a single cell. Given we observe the tumour at time t_{end} the frequency of the subclone will increase as a function of this time:

$$f(t_{end}) = \frac{N_1(t_{end} - t_1)}{N_1(t_{end} - t_1) + N_H(t_{end})} \tag{9}$$

where N_1 and N_H are the population sizes of of the subclone and host tumour population respectively. Assuming exponential growth and using the definition of fitness provided by equation (1), we can write this as:

$$f(t_{end}) = \frac{e^{\lambda(1+s)(t_{end}-t_1)}}{e^{\lambda(1+s)(t_{end}-t_1)} + e^{\lambda t_{end}} - e^{\lambda(t_{end}-t_1)}} \tag{10}$$

Where the last term in the denominator is a correction that takes into account that a single cell from the host population mutates into the fitter type. By extracting a factor $e^{\lambda t_{end}}$ from each term we arrive at the following.

$$f(t_{end}) = \frac{e^{\lambda s(t_{end}-t_1)}e^{-\lambda t_1}}{e^{\lambda s(t_{end}-t_1)}e^{-\lambda t_1} + 1 - e^{-\lambda t_1}} \tag{11}$$

For even moderate t_1, $e^{-\lambda t_1} << 1$ hence we can neglect this term (although it is possible to solve for s with this term the correction is minimal). Solving for s (which amounts to solving an equation of the form $y = \frac{x}{1+x}$) we get:

$$s = \frac{\log(\frac{f}{1-f}) + \lambda t_1}{\lambda(t_{end} - t_1)} \tag{12}$$

Thus given an assumption on t_{end} measuring f and t_1 from the distribution (via equation (2)) we can estimate the relative fitness advantage s. For example, assuming a reasonable population size of the final tumour to be 10^{10} (100 billion cells is typical in, for example, colorectal cancers), would amount to solving $2^{t_{end}} = (1-f)10^{10}$, given we've measured the fraction of the clone f. As we are using the doubling time as our unit of time we can simply set $\lambda = \log(2)$. Turning to a fully stochastic model, the mean time taken for a population to reach a size N_{end} has an additional correction factor which has been calculated using a branching process [3]: $t_{end} = \frac{1}{\lambda}\log(N_{end}) + \gamma/\lambda$, where γ is Euler's constant. Thus the deterministic approach is a slight underestimate when compared to the stochastic case. See section 2.1.

1.3 Multiple subclones

When we have more than one subclone the picture is of course more complicated. Each subclone can have their own relative fitness advantage and time of emergence. Subclones can also be nested within one another, further complicating the underlying dynamics.

First of all we consider the case where clones are not nested, i.e. two clones arise independently within the host population. We then have 2 equations describing how the fraction (and hence the allele frequency) of the 2 clones increases in time.

$$f_1(t_{end}) = \frac{N_1(t_{end} - t_1)}{N_1(t_{end} - t_1) + N_2(t_{end} - t_2) + N_H(t_{end})} \tag{13}$$

$$f_2(t_{end}) = \frac{N_2(t_{end} - t_2)}{N_1(t_{end} - t_1) + N_2(t_{end} - t_2) + N_H(t_{end})} \tag{14}$$

Again substituting in for a exponentially growing populations and removing a factor $e^{\lambda t_{end}}$ from each term we arrive at the following:

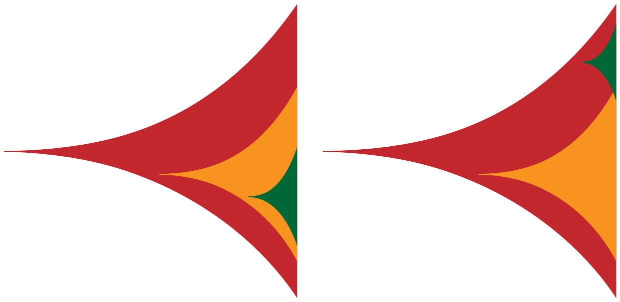


Figure 2: *When we have more than one subclone, they may be nested (left) or unnested (right).*

$$f_1(t_{end}) = \frac{e^{\lambda s_1(t_{end}-t_1)}e^{-\lambda t_1}}{e^{\lambda s_1(t_{end}-t_1)}e^{-\lambda t_1} + e^{\lambda s_2(t_{end}-t_2)}e^{-\lambda t_2} + 1} \tag{15}$$

$$f_2(t_{end}) = \frac{e^{\lambda s_2(t_{end}-t_2)}e^{-\lambda t_2}}{e^{\lambda s_1(t_{end}-t_1)}e^{-\lambda t_1} + e^{\lambda s_2(t_{end}-t_2)}e^{-\lambda t_2} + 1} \tag{16}$$

We then arrive at the following for s_1 and s_2:

$$s_1 = \frac{\log(\frac{f_1}{1-f_1-f_2}) + \lambda t_1}{\lambda(t_{end} - t_1)} \tag{17}$$

$$s_2 = \frac{\log(\frac{f_2}{1-f_1-f_2}) + \lambda t_2}{\lambda(t_{end} - t_2)} \tag{18}$$

1.3.1 Nested subclones

In the case of nested subclones, one subclone will grow inside the other thereby increasing the frequency of the major subclone. We define subclone 1 as the major subclone ($t_1 < t_2$) with subclone 2 growing inside, we also require ($s_1 < s_2$) . The frequency of suclone 1 will therefore be given by:

$$f_1(t_{end}) = \frac{N_1(t_{end} - t_1) + N_2(t_{end} - t_2)}{N_1(t_{end} - t_1) + N_2(t_{end} - t_2) + N_H(t_{end})} \tag{19}$$

Proceeding as before we get:

$$s_1 = \frac{\log(\frac{f_1-f_2}{1-f_1}) + \lambda t_1}{\lambda(t_{end} - t_1)} \tag{20}$$

$$s_2 = \frac{\log(\frac{f_2}{1-f_1}) + \lambda t_2}{\lambda(t_{end} - t_2)} \tag{21}$$

We also have a modified equation for the time the subclones emerged.

$$M_1 = \hat{\mu}\Gamma_1 \tag{22}$$

$$M_2 = \hat{\mu}\Gamma_2 \tag{23}$$

Here Γ_2 is the number of divisions between t_1 (time subclone 1 appears) and t_2 (time subclone 2 appears). Meanwhile as subclone 1 is growing faster by a factor $1 + s_1$, converting the number of divisions requires including this factor for the second subclone.

$$\Gamma_2 = t_1 + (1 + s_1) \times 2\log(2)t_2 \tag{24}$$

While for t_1, we have as before:

$$\Gamma_1 = 2\log(2)t_1 \tag{25}$$

For our reported values of $1 + s_n$ we assumed the non-nested case, for the AML sample this was experimentally validated in the original study with single cell sequencing [4].

2 Expected variation in parameter inference

The model discussed so far assumes deterministic growth of the tumour and subclones. A more realistic model would consider the tumour to grow stochastically according to some birth and death rates. This is what is implemented in our computational model of tumour growth, and therefore our posterior distributions will give some weight to all possible stochastic trajectories for any given parameter set. The relatively wide posterior distributions we obtain in some cases is therefore not entirely unexpected given the stochastic nature of the underlying processes of mutation and division. To gain some intuition on the expected variability we can consider a stochastic treatment of a birth-death process and derive the mean and variance of the population size at some time t.

2.1 Stochastic birth death model

Given that the probability of any cell giving birth in the interval $(t, t + \delta t)$ is $b\delta t$, and the probability of dying is given by $d\delta t$ we can write down the differential difference equation for the birth-death process as follows:

$$\frac{dp_0(t)}{dt} = dp_1(t)$$
$$\frac{dp_n(t)}{dt} = b(n-1)p_{n-1}(t) - (b+d)np_n(t) + d(n+1)p_{n+1}(t) \ \ , \ \ (n \geq 1) \tag{26}$$

The solution to this equation can be found in Bailey (1964)[5]:

$$\begin{gathered} p_0 = \alpha \\ p_n = (1-\alpha)(1-\beta)\beta^{n-1} \ \ , \ \ (n \geq 1) \end{gathered} \tag{27}$$

where

$$\begin{aligned} \alpha &= \frac{d(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \\ \beta &= \frac{b(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \end{aligned} \tag{28}$$

The mean and variance of the population size at a time can then be easily calculated from this probability distribution:

$$\bar{N} = e^{(b-d)t} \tag{29}$$

$$\sigma_N^2 = \frac{b+d}{b-d}e^{(b-d)t}(e^{(b-d)t}-1) \tag{30}$$

Therefore the standard deviation $\sigma_N \propto e^{(b-d)t}$. In our stochastic model of tumour growth, a subclone emerges that will eventually reach a fraction f_{sub}, given the large expected variance in population size it is then not unexpected to see our posterior estimates for the selection coefficients s to be relatively wide as in Figures 2 and 3 (main text). Note that the variance increases with increasing death rate.

2.2 Stochastic birth-death model conditioned on N

Another way to explore the expected variability due to stochastic effects is to ask the reverse question: what is the distribution of times, T conditioned on the population reaching a certain population size, N. This is particularly relevant to our inferences on the time a subclone emerges. We can use this reasoning to quantify how much of the variability in the posterior distribution is due to stochastic effects. Durrett has derived the probability density function for this reverse problem and it is given by [3]

$$p_{T_N}(t) = \frac{\lambda^2 N}{b}\exp(-\lambda t) \times \exp(-\frac{\lambda}{b}N \times \exp(-\lambda t)) \tag{31}$$

where $\lambda = b - d$. The mean and standard deviation can then be found to be:

$$\begin{aligned} \mathrm{E}(T_N) =& \frac{\gamma + \log(\frac{N\lambda}{b})}{\lambda} \\ \mathrm{SD}(T_N) =& \frac{\pi^2}{6 \times \lambda^2} \end{aligned} \tag{32}$$

where γ is Euler's constant. Expressing time in units of population doublings, $\lambda = \log(2)$, therefore the expected standard deviation due to stochastic effects is 1.85. We observed an average standard deviation of 1.56 across all data and simulations where we applied our ABC method, consistent with the variability being primarily due to stochastic effects.

3 Driver mutations

We note that, for simplicity, we assume that mutations effecting fitness occur at any given time t_1 (a parameter of our model on which we perform inference), but we do not model this as a stochastic variable dependent on the mutation rate of driver alterations, as it is the case in other models of cancer progression [6]. We can however, address this *a posteriori* and ask the question of what would the driver mutation rate need to be to observe the distribution of t_1 times we measure? Bozic et al. [6] and Durrett [3] both use branching process to calculate the expected waiting time for a cell with k driver mutations, we are particularly interested in the case $k = 1$, i.e. in a growing population what is the waiting time for the first subclone with a fitness advantage. We follow the formulation in Durrett as this approach uses a continuous time formulation more relevant to our model. Continuing with our notation where b is the birth rate and d is the death rate the mean time (in population doublings) until the first driver mutation t_1 is given by

$$t_1 = \frac{1}{b} \log\left(\frac{1}{\mu_d}\frac{b-d}{b}\right) \tag{33}$$

where μ_d is the driver mutation rate. This can be rearranged to express μ_d as a function of time:

$$\mu_d = \frac{b-d}{b} e^{-bt_1} = \beta e^{-bt_1} \tag{34}$$

Following our previous notation we can then write the effective driver mutation rate, $\hat{\mu}_d$, i.e. the driver mutation rate per effective population doubling representing an upper limit on the per cell division mutation rate.

$$\hat{\mu}_d = \frac{\mu_d}{\beta} = e^{-bt_1} \tag{35}$$

We will use $b = \log(2)$ so that time is in units of population doublings as before. The distribution of emergence times t_1 we observe in our cohorts peaks around $t = 6 - 9$ depending on the cohort (see Figure 4C). This gives $\hat{\mu}_d = \{0.002 - 0.015\}$ per cell division, corresponding to $\hat{\mu}_d = \{10^{-11} - 10^{-10}\}$ per bp per effective division in the exome. The average $\hat{\mu}_d$ across the cohort of tumours may well be less than this given the abundance of neutrally evolving tumours. Given that the mutation rates we measure are on the order 10-100's per effective cell division, this represents a driver mutation rate that is at least 10^3 smaller than the overall mutation rate, consistent with accumulating evidence of a limited number of cancer driver mutations [7]. We also note that the fitness advantage of subclones in our model needs not necessarily to be due to point mutations and could very well be due to any mechanism that provides a selective advantage. Our estimate of the driver mutation rate per effective cell division ($\hat{\mu}_d = \{0.002 - 0.015\}$) in reality combines all possible mechanisms such as copy number alterations and epigenetic modifications. We also note that mutation rate per division will in reality be $< \hat{\mu}_d$ depending on the value of β.

4 Statistical inference

Equations (12), (17), (18), (20) & (21) provide a means to estimate the selective advantage of a subclone and can be summarised by the following equation for subclone n:

$$s_n = \frac{\log(F(f_{1..n})) + \lambda t_n}{\lambda(t_{end} - t_n)} \tag{36}$$

where $F(f_{1..n})$ is some function of the subclone fractions. t_n, the time the subclone appears is a function of the effective mutation rate and possibly the time the parent subclone arises if it is nested:

$$t_n = F(\hat{\mu}, b, t_{1..n-1}) \tag{37}$$

Therefore if we can accurately measure the fraction of the subclone, the number of mutations in the subclone and the effective mutation rate we can estimate the selective advantage of a subclone s_n. For the inference, we used our cancer simulation scheme together with approximate Bayesian computation (ABC). The advantage of this approach over say a clustering based approach is that as our simulation simulates the stochastic process of tumour growth so any stochastic effects that may be important will be captured in the inference scheme. It also naturally accounts for aspects that would be difficult to accurately model in a clustering based approach such as accounting for within (sub)clone neutral mutations that accumulate as the tumour grows.

4.1 Bayesian inference

The quantity of interest in Bayesian methods is the posterior distribution, which tells us the probability of a particular parameter in our model being the correct parameter given the data. The posterior distribution is obtained by combining any prior beliefs we may have of a particular parameter with how well a particular parameter value explains our observed data. This is formally expressed via Bayes rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{38}$$

where θ is a vector of parameters for our model, $p(\theta)$ is the prior distribution, ie our prior beliefs on θ, and $p(\theta|D)$ is the likelihood function. Classical Bayesian inference would entail writing down the likelihood, choosing a prior distribution on θ and then calculating the posterior distribution, in most cases the posterior distribution is analytically intractable so it is necessary to use computational approaches such as Markov chain Monte Carlo (MCMC).

The likelihood is typically defined by a known probability distribution, where we would consider the probability of θ given the data D is fixed. In our case, and in many applications it is impossible to write down the likelihood but it is possible to simulate from a model. In such cases when the likelihood is unknown we can turn to approximate approaches. In these likelihood free or approximate Bayesian computation (ABC) approaches we obtain samples from a distribution $p(\theta|d(D^*, D) \leq \epsilon)$, where D^* is simulated data from our model, D is the observed data, $d(D^*, D)$ is some distance measure between the real data and ϵ is a tolerance level. If ϵ is sufficiently small then $p(\theta|d(D^*, D) \leq \epsilon)$ will be a good approximation of the posterior distribution $p(\theta|D)$.

4.1.1 ABC rejection

The simplest ABC algorithm is the rejection algorithm [8, 9] which proceeds as follows:

S1 Sample θ^* from $p(\theta)$

S2 Simulate a dataset D^* from model $M(D|\theta^*)$

S3 If $d(D^*, D) \leq \epsilon$, accept θ^*, otherwise reject

S4 Return to S1

In many applications we may have a number of competing models (including this one) and would like to infer the most probable model. For this we can turn to Bayesian model selection. In our case models are specified by the number of subclones under selection, where 0 subclones under selection would be the neutral case. If m_0 and m_1 are two models, we would like to choose which model provides the best support for the data, for this we can turn to Bayes factors, which is the ratio of posterior odds to prior odds of the two models. The Bayes factor in favour of m_0 over m_1 is defined as:

$$B_{01} = \frac{P(m_1|D)/P(m_2|D)}{P(m_1)/P(m_2)} \tag{39}$$

Where $P(m_n)$ is the prior probability of model n and $P(m_n|D)$ is the posterior probability.

Incorporating model selection into the ABC framework is relatively straightforward as we can effectively treat the model as an additional parameter in our inference scheme, where each model m_n will have a corresponding model specific parameter vector θ_n. The ABC rejection with model selection then becomes [10]:

S1 Sample m^* from $p(m)$

S2 Sample θ^* from $p(\theta|m^*)$

S3 Simulate a dataset D^* from model $M(D|\theta^*, m^*)$

S4 If $d(D^*, D) \leq \epsilon$, accept (m^*, θ^*), otherwise reject

S5 Return to S1

The downside of the ABC rejection algorithm is that the acceptance rate is generally low, requiring a large amount of datasets to be simulated, we therefore turned to the approximate Bayesian computation sequential monte carlo (ABC SMC) algorithm which increases the acceptance rate [11, 12] and thus the efficiency of the algorithm.

4.1.2 ABC SMC algorithm

The ABC SMC algorithm uses sequential importance sampling to increase the acceptance rate of simulated datasets. In ABC SMC, parameter vectors, particles (m_n, θ_n) are sampled from the prior distribution and then propagated through a series of distributions with decreasing tolerances, ϵ_i, until $\epsilon_i = \epsilon_T$ the target tolerance. We therefore gradually evolve toward the target posterior distribution $p(\theta | d(D^*, D) \leq \epsilon_T)$ as ϵ_i decreases. The ABC SMC model selection algorithm is as follows [12]:

S1 Set the population indiciator to $t = 1$

S2 Set the particle indiciator $i = 1$

S3 If $t = 1$, sample (m^{**}, θ^{**}) from the prior distribution $P(m, \theta)$
if $t > 1$, sample m^* from $P_{t-1}(m^*)$ and then perturb according to $m^{**} \sim KM_t(m|m^*)$. Sample θ^* from previous populations with weights w_{t-1} and perturb parameter vector according to $\theta^{**} \sim KP_{t,m^{**}}(\theta|\theta^*)$

S4 If $P(m^{**}, \theta^{**}) = 0$, return to **S3**

S5 Simulate data D^* for model m^{**} and parameters θ^{**}, then calculate $d(D^*, D)$, if $d(D^*, D) > \epsilon_t$ go to **S3**

S6 Set $(m_t^i, \theta_t^i) = (m^{**}, \theta^{**})$ and calculate the weight of the particle w_t. If $i < N$ set $i = i + 1$ and go to **S3**

S7 Normalize the particle weights and calculate the marginal model probabilities, $P_t(m_t = m) = \sum_{i, m_t^i = m} w_t^i(m_t^i, \theta_t^i)$

S8 Calculate the perturbation kernels and next tolerance value ϵ_t, if $\epsilon_t > \epsilon_T$, set $t = t + 1$ and go to **S3**.

The particle weights are calculated as follows:

$$w_t^i(m_t^i, \theta_t^i) = \begin{cases} 1, & \text{if } t = 1 \\ \frac{P(m_t^i, \theta_t^i)}{S}, & \text{if } t > 1 \end{cases} \tag{40}$$

where S is:

$$S = \sum_{j=1}^{M} P_{t-1}(m_{t-1}^j) KM_t(m_t^i, m_{t-1}^j) \times \sum_{k, m_{t-1} - m_t^i} \frac{w_{t-1}^k KP_{t,m_t}^i(\theta_t^i | \theta_{t-1}^k)}{P_{t-1}(m_{t-1} = m_t^i)} \tag{41}$$

Here KM is the model perturbation kernel and KP is the parameter perturbation kernel. Particles that have been sampled from the previous distribution are denoted by a single asterisk, the perturbed particles are denoted with a double asterisk. For the model perturbation kernel we used the following:

$$KM_t(m|m^*) = \begin{cases} 0.6, & \text{if } m = m^* \\ 0.4, & \text{if } m \neq m^* \end{cases} \tag{42}$$

For the particle perturbation kernel we used the uniform distribution with limits determined from the range of parameter values from the previous population [13], for parameter k, $KP_t(k|k^*) = U(k_i - \sigma, k_i + \sigma)$, where σ is given by:

$$\sigma = \frac{1}{2}(max(k)_{t-1} - min(k)_{t-1}) \tag{43}$$

We updated the tolerance at each population step and used the 0.3 quantile of the previous populations ϵ values. We set the number of particles N to be 500. The ABC SMC algorithm continues until one of the following conditions is met.

1. $\frac{\epsilon_t - \epsilon_{t-1}}{\epsilon_t} < 0.05$
2. Completed 5×10^6 simulations
3. 200 hours of computation time

4.2 Prior distributions

We used uniform prior distributions for all the parameters, limits on these parameters are given in Table 1. We also used a uniform distribution for the model prior. Often, when we simulate a model with selection, the size of the clones may be very small or very large. Given that we are constrained by the data in observing only relatively large subclones we only consider simulations that give clones greater than the detection limit of the data (the point where the 1/f peak decreases) and less than 0.95 (above this we would be measuring the neutral dynamics of the selected clone). Therefore when performing the ABC after sampling a model we continue to simulate until we return a simulation where clones are > detection limit and < 0.95.

4.3 Algorithm convergence

To asses the ability of the algorithm to converge to a stable posterior distribution we looked at the posterior distributions evolution over the population

1+s	μ	clonal mutations	t_1
[1, 26.0]	[0.1, 500]	[1, 5000]	[3, 14]

Table 1: *Limits on prior distributions for all parameters*

and reassuringly find that as ϵ decreased the posterior distributions becomes tighter and less variable. See Supplementary Note Figures 4-9.

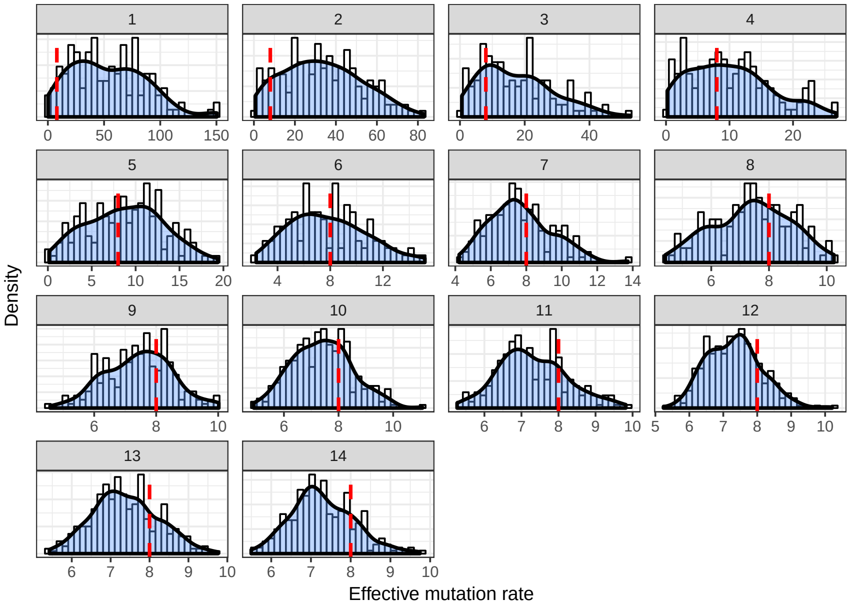


Figure 3: *Posterior distribution of mutation rate as ϵ decreases, red dashed line is true value.*

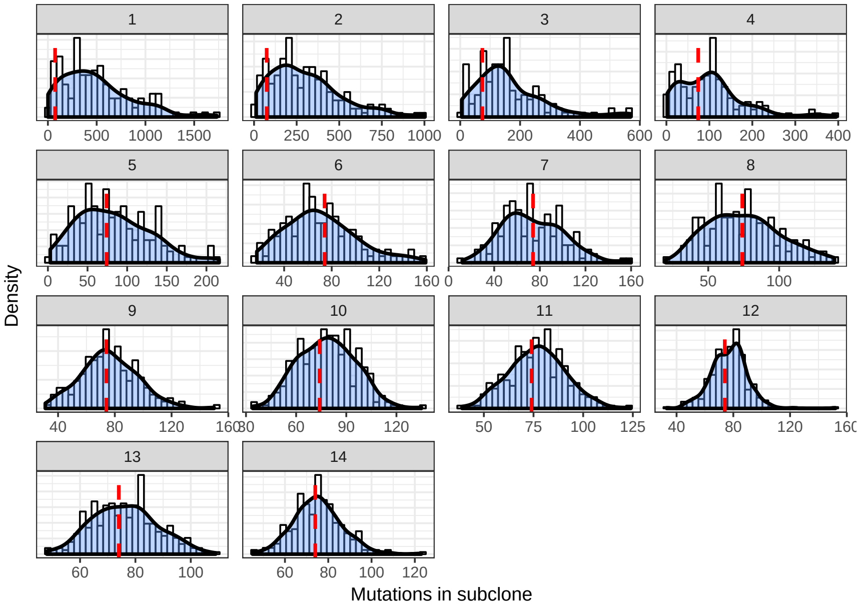


Figure 4: *Posterior distribution of mutations in founder cell of subclone as ϵ decreases, red dashed line is true value.*

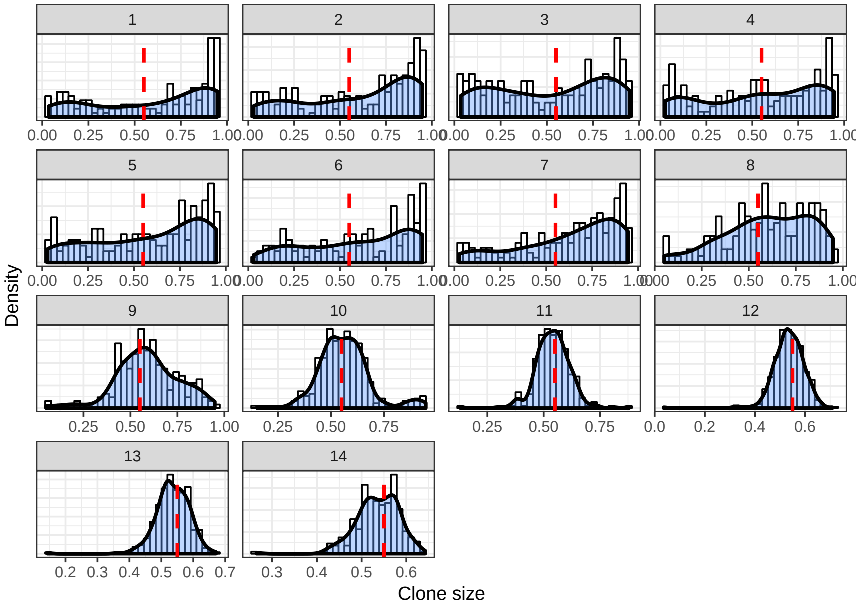


Figure 5: *Posterior distribution of the clone size of subclone as fraction of total tumour size as ϵ decreases, red dashed line is true value.*

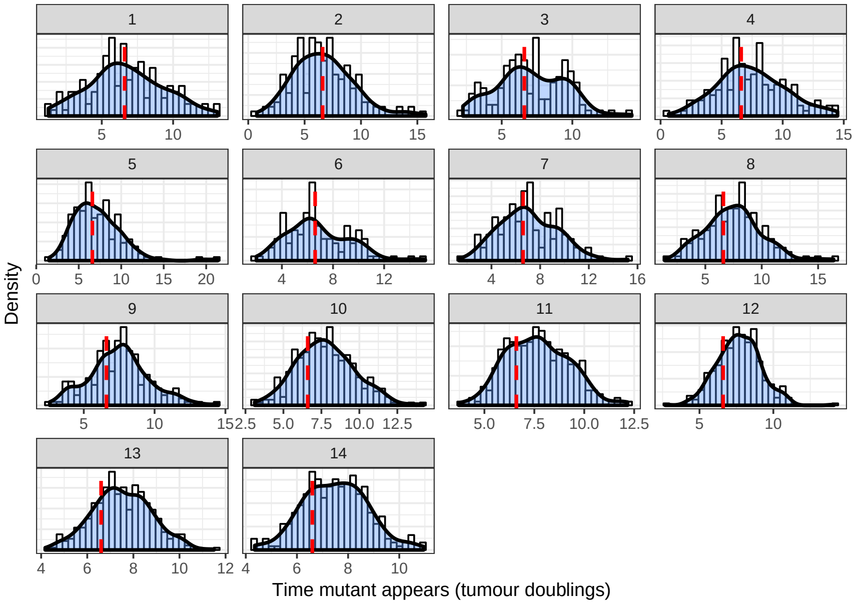


Figure 6: *Posterior distribution of time subclone appears as ϵ decreases. This is the inferred time calculated using equation* (4)*. Red dashed line is true value.*

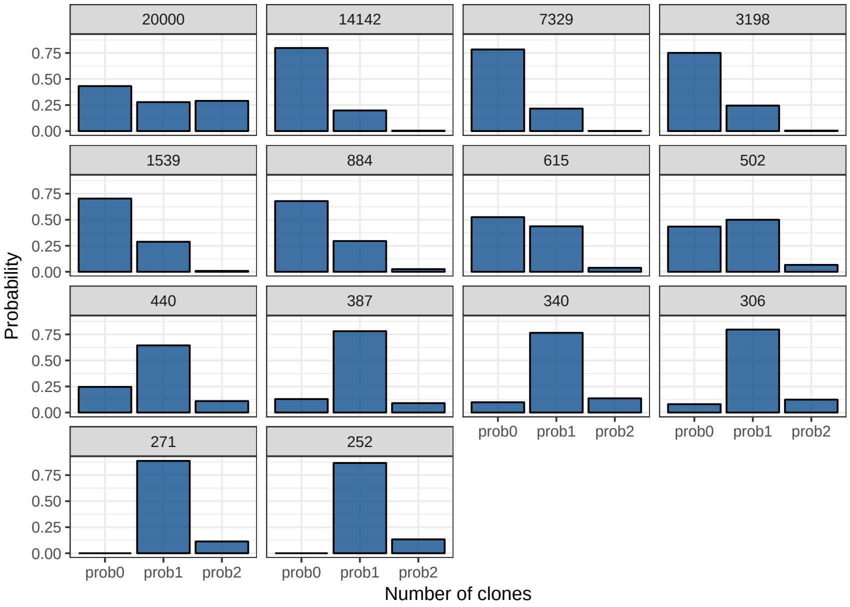


Figure 7: *Model probabilities as ϵ decreases. True model is the 1 subclone model.*

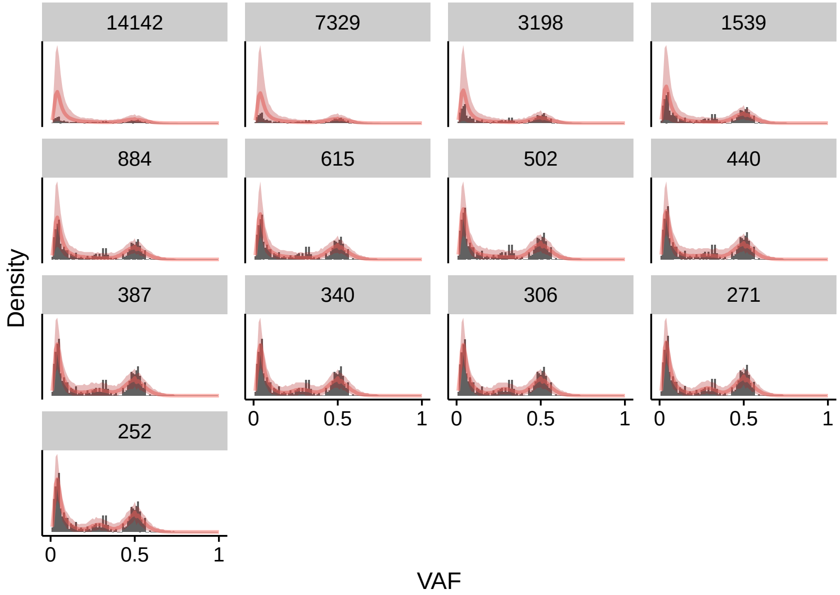


Figure 8: *Median histogram (red line) and 95% intervals (shaded red area) for simulations that pass ABC SMC as ϵ decreases. Results are from 500 posterior samples.*

5 Alternative growth models

5.1 Logistic growth

For our main analysis we have assumed an exponential growth model. Some tumours however, especially benign lesions, may reach a plateau in their growth, and consequently are better represented by sigmoidal-type growth models [14, 15]. In a sigmoidal model of tumour growth, the tumour population at late times can be approximated as a population of constant size with continual turnover of cells. Interestingly, in a fixed size population, it has been shown that the fixation time of beneficial mutations is proportional to the logarithm of the population size [16], which suggests that clonal expansions can be relatively rapid when the population is no longer growing. To explore this further we implemented a logistic growth model as well as a moran model for a fixed size population.

5.1.1 Modeling logistic growth

In the logistic growth model, growth is density-dependent and the environment has a maximum number of individuals it can support, commonly referred to as the carrying capacity, K of the population. The differential equation for logistic population growth is

$$\frac{dN}{dt} = \lambda \left(1 - \frac{N}{K}\right) N \tag{44}$$

In the logistic population growth model, the birth and death rates of individuals in the population are proportional to the population size

$$b(N) = b_1 - b_2 N \tag{45}$$

$$d(N) = d_1 + d_2 N \tag{46}$$

where b_1 and d_1 are the intrinsic birth and death rates, and b_2 and d_2 can be calculated given a carrying capacity K from:

$$K = \frac{b_1 - d_1}{b_2 + d_2} \tag{47}$$

When $b_2 = d_2 = 0$ we recover exponential growth.

5.1.2 Moran model

The Moran model is a classic model from population genetics and describes a stochastic birth death process where at each time step one individual is chosen to die and one is chosen to replicate [17] (see Supplementary Note Figure 9). Individuals that have fitness advantages are more likely to be chosen to replicate, the selection coefficient is often defined as relative increase in the average number of offspring per generation: a fitter individual will on average have $1+s$ more offspring. It has been shown that the average fixation time (in generations) of a neutral mutation is $\propto N$. In the case of a beneficial mutation the time to fixation τ_{fix} is given by [18],

$$\tau_{fix} = \frac{2}{s}\log(N). \tag{48}$$

Therefore for a fixed size neutral population, the timescales over which mutations may rise to observable frequencies is likely longer than the age of the tumour. Results consistent with our simulations demonstrated that if a tumour follows a logistic growth model, the dominant signal in the VAF distribution is that of the early exponential growth (Supplementary Note Figure 9). Selection however can result in mutations reaching observable frequencies rapidly.

To examine the effect of the population growth profile on subclone evolution, we simulated a model of fixed population size using a Moran process, and compared the speed with which subclones expand versus the exponential growth model described in the main text (Supplementary Note Figure 9A&B). The fitness advantage of a mutant in both fixed and growing populations was defined as the average offspring per generation (of the background host population). We introduced a fitter mutant in the growing population when the population was of size N, and simulated the Moran model for fixed size N; thus a new mutant starts out at a frequency $1/N$ in both cases. We followed the average frequency of the mutant over time. In the fixed population model the fitter mutant spreads through the population at a significantly faster rate (Supplementary Note Figure 9C; $p < 0.001$), and we noted that subclonal expansions can also lead to subclonal clusters in the VAF distribution in a fixed population (Supplementary Note Figure 9E). We note that a constant population of cells that acquires new passenger mutations and undergoes neutral drift results in a neutral tail in the VAF distribution that however, does not directly encode the mutation rate[1] as in the case of exponential growth.

Under a logistic regime, initial cancer growth is exponential, slowing to a constant population size (with turnover) once a carrying capacity is reached. We investigated how this pattern of population growth influenced the measurement of evolutionary dynamics. We simulated logistic growth where the population first grows exponentially and then transitions into a Moran model (Supplementary Note Figure 9B). We found that assuming for example a small carrying capacity of 10^4 cells, even if the fixed population size phase is 20 times longer than the growth phase, the dominant signature is that of the initial (neutral) growth, not the neutral drift within the fixed size population (Supplementary Note Figure 9F). Consequently, the mutation rate estimates match those measured in a purely exponential neutral model (figure 9F&G). We note that this is because tumours are very large populations, and effects of neutral drift during the constant phase are unlikely to be significant since the time it takes for variants to rise to a detectable frequency under these conditions is proportional to the population size N. Hence, even for barely-detectable tumours of 10^6 cells, it would take approximately a million generations before seeing those drift tails: much longer than a human lifetime. Hence in cancer data, irrespective of whether or not the population has become constant, the VAF distribution encodes initial tumour growth, and neutral tails do accurately inform on the mutation rate.

5.2 Power law growth

Power law type growth is another growth law used to model tumours. The expected allelic frequency distribution for neutral power law growth is derived below. Specifically, we wish to solve equation (1) for a growth law described by:

$$N(t) = 1 + bt^n \tag{49}$$

where b is the growth rate and n is the dimension. Solving equation (1) for $M(t)$ give,

$$M(t) = \pi\mu b \int_{t_{min}}^{t} (1 + bt^n)\, dt \tag{50}$$

$$M(t) = \pi\mu b \left(t - t_{min} + \frac{b}{n+1} \left(t^{1+n} - t_{min}^{1+n} \right) \right) \tag{51}$$

Mutation frequencies are inversely proportional to the population size at the time they arise, so f is given by:

$$f = \frac{1}{\pi N(t)} = \frac{1}{\pi(1 + bt^n)} \tag{52}$$

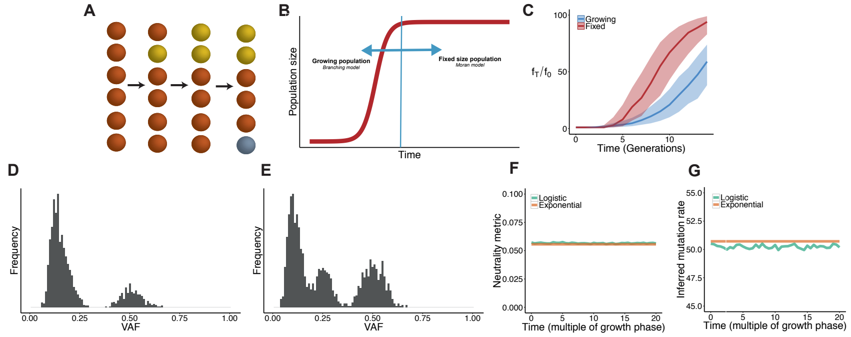


Figure 9: *We used a Moran model to compare the dynamics between fixed size populations and growing population, A, and found that in fixed size population selection can be more rapid C. We simulated a Moran model with $N = 100$, and introduced a mutation at $N = 100$ in the growing population so in both models the initial frequency of the mutation $f_0 = 1/100$, the subclone has fitness advantage $1 + s = 0.5$. We then measured the frequency at a later time T, in the fixed population size the ratio f_T/f_0 increases quicker than in the growing population. Shown here are the results from 1000 simulations, dark line shows median value of these 1000 simulations, shaded ares shows 95% interval. The Moran model can also produce VAF histograms similar to the neutral case, D (no selection, 300 generations) and the non neutral case E ($1 + s = 2$, number of generations = 10). However simulating a tumour that grows logistically and transitions into a Moran model B shows that the VAF distribution is a consequence of the early exponential dynamics. When the population was in fixed population size Moran model phase for 20 times longer (in generations) than it was in the growth phase, the main signature of the VAF distribution is that of exponential growth this can be seen as we observe no difference in our neutrality metric F, or the inferred mutation rate G, over what would be expected from the exponentially growing model.*

Under neutrality, frequency and time are interchangeable, so solving for the frequency $f(t)$ gives:

$$t = \left(\frac{1/f\pi - 1}{b}\right)^{1/n} \tag{53}$$

Using this in $M(t)$ we arrive at the following for the cumulative number of mutations at a frequency f.

$$\begin{aligned} M(f) = \pi\mu b \Bigg(\left(\frac{1/f\pi - 1}{b}\right)^{1/n} - \left(\frac{1/f_{max}\pi - 1}{b}\right)^{1/n} + \\ \frac{b}{n+1}\left(\left(\frac{1/f\pi - 1}{b}\right)^{n+1/n} + \left(\frac{1/f_{max}\pi - 1}{b}\right)^{n+1/n} \right)\Bigg) \end{aligned} \tag{54}$$

Removing all constants we get the following:

$$M(f) \sim \left(\frac{1}{\pi f} - 1\right)^{\frac{1}{n}} + \frac{1}{n+1}\left(\frac{1}{\pi f} - 1\right)^{\frac{n+1}{n}} \tag{55}$$

References

1. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. English. *Nature Genetics* **48,** 238–244 (Mar. 2016).
2. Werner, B. *et al.* Reconstructing the in vivodynamics of hematopoietic stem cells from telomere length distributions. English. *eLife* **4,** e08687 (Oct. 2015).
3. Durrett, R. Branching Process Models of Cancer, 1–61 (Dec. 2014).
4. Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. English. *Cell Systems* **1,** 210–223 (Sept. 2015).
5. Bailey, N. *The elements of stochastic processes with applications to the natural sciences* 1964. <`http://books.google.com/books?hl=en&lr=&id=yHPnwl4QOfIC&oi=fnd&pg=PA1&dq=The+Elements+of+Stochastic+Processes+with+Applications+to+the+Natural+Sciences&ots=DzjbVSVn1F&sig=UIcrNuxp1USRTHRQevh_UqtlW-w`>.
6. Bozic, I., Gerold, J. M. & Nowak, M. A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. English. *PLOS Computational Biology* **12,** e1004731 (Feb. 2016).
7. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell,* 1–35 (Oct. 2017).
8. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. English. *Molecular biology and evolution* **16,** 1791–1798 (Dec. 1999).
9. Tavaré, S, Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. English. *Genetics* **145,** 505–518 (Feb. 1997).
10. Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F. & Taly, J.-F. ABC likelihood-free methods for model choice in Gibbs random fields. English. *Bayesian Analysis* **4,** 317–335 (June 2009).
11. Toni, T, Welch, D, Strelkowa, N, Ipsen, A & Stumpf, M. P. H. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. English. *Journal of The Royal Society Interface* **6,** 187–202 (Feb. 2009).

12. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical systems in systems and population biology. English. *Bioinformatics (Oxford, England)* **26,** 104–110 (Jan. 2010).

13. Filippi, S., Barnes, C. P., Cornebise, J. & Stumpf, M. P. H. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. English. *Statistical applications in genetics and molecular biology* **12,** 87–107 (Mar. 2013).

14. Rodriguez-Brenes, I. A., Komarova, N. L. & Wodarz, D. Tumor growth dynamics: insights into evolutionary processes. English. *Trends in Ecology & Evolution* **28,** 597–604 (Oct. 2013).

15. Spratt, J. A., Von Fournier, D, Spratt, J. S. & Weber, E. E. Decelerating growth and human breast cancer. English. *Cancer* **71,** 2013–2019 (Mar. 1993).

16. Otto, S. P. & Whitlock, M. C. *Fixation Probabilities and Times* English. ISBN: 9780470015902. doi:10.1002/9780470015902.a0005464.pub3. <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0005464.pub3/full> (John Wiley & Sons, Ltd, 2001).

17. Nowak, M. A. *Evolutionary Dynamics: Exploring the Equations of Life* ISBN: 9780674023383. <https://books.google.co.uk/books?id=YXrIRDuAbEOC> (Belknap Press of Harvard University Press, 2006).

18. Durrett, R. *Probability Models for DNA Sequence Evolution* ISBN: 9780387781693. <https://books.google.co.uk/books?id=o4_bMHy7jFoC> (Springer New York, 2008).