nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0128-6

Quantification of subclonal selection in cancer from bulk sequencing data

Marc J. Williams[1,2,3], Benjamin Werner[4], Timon Heide[4], Christina Curtis[5,6], Chris P. Barnes[2,7*], Andrea Sottoriva[4*] and Trevor A. Graham[1*]

Subclonal architectures are prevalent across cancer types. However, the temporal evolutionary dynamics that produce tumor subclones remain unknown. Here we measure clone dynamics in human cancers by using computational modeling of subclonal selection and theoretical population genetics applied to high-throughput sequencing data. Our method determined the detectable subclonal architecture of tumor samples and simultaneously measured the selective advantage and time of appearance of each subclone. We demonstrate the accuracy of our approach and the extent to which evolutionary dynamics are recorded in the genome. Application of our method to high-depth sequencing data from breast, gastric, blood, colon and lung cancer samples, as well as metastatic deposits, showed that detectable subclones under selection, when present, consistently emerged early during tumor growth and had a large fitness advantage (>20%). Our quantitative framework provides new insight into the evolutionary trajectories of human cancers and facilitates predictive measurements in individual tumors from widely available sequencing data.

Carcinogenesis is the result of Darwinian selection for malignant phenotypes and is driven by genetic and epigenetic alterations that allow cells to evade normal homeostatic regulation and prosper in changing microenvironments[1]. High-throughput genomics has shown that tumors across all cancer types are highly heterogeneous[2,3], with complex clonal architectures[4]. However, because longitudinal observation of solid tumor growth unperturbed by treatment remains impractical, the temporal evolutionary dynamics that produce subclones remain undetermined, and consequently, there is no mechanistic basis that can be used to predict future tumor evolution and modes of relapse. More specifically, the magnitude of the fitness advantage experienced by a new cancer subclone has remained unknown.

The subclonal architecture of a cancer—as measured by the pattern of intratumoral genetic heterogeneity (ITH)—is a direct consequence of the unobservable evolutionary dynamics of tumor growth. Therefore, given a realistically constrained model of subclonal expansion, the pattern of ITH in a tumor can be used to infer its most probable evolutionary trajectory. ITH represented within the distribution of variant-allele frequencies (VAFs), as measured by high-coverage sequencing, is particularly amenable to such an approach.

Here we build upon theoretical population-genetics models of asexual evolution[5] and Bayesian statistical inference on genetic data[6] to measure cancer evolution in human tumors. This type of approach is established in the field of molecular evolution, in which evolutionary processes are also difficult to measure directly[7,8], and examples of applications of these approaches to human cancers date back nearly twenty years[9,10].

We recently showed that, under a neutral 'null' evolutionary model (i.e., when all selected driver alterations are truncal and present in all cancer cells), the VAFs within a tumor follow a characteristic power-law distribution[11]. Subsequent simulations that modeled space and subclonal selection demonstrated that genetic divergence in multiregion sequencing data could be used to categorize tumors on the basis of the mode of their evolution[12] (effectively neutral or non-neutral), but the specific evolutionary dynamics that produce subclonal architectures, such as the fitness advantage of subclones, remained unmeasured. Here, by using a combination of a stochastic branching process model of subclonal selection in cancer, an explicit sequencing error model, and Bayesian model selection and parameter inference, we identified the characteristic patterns of subclonal selection in the cancer genome and measured fundamental evolutionary parameters in non-neutrally evolving human tumors.

Results

Theoretical framework of subclonal selection. We developed a stochastic computational model of tumor growth applicable to cancer genomic data that accounted for subclonal selection (Methods). The model is based on a classical stochastic branching process approach from population genetics[5,13] that has been often used to model malignant populations[13,14], which we extended here to be applicable to cancer sequencing data. Cells divide and die according to defined birth and death rates, and daughter cells acquire new mutations at a rate of μ mutations per cell per division (Fig. 1a). The fitness advantage of a mutant subclone is defined by the ratio of net growth rates between the fitter mutant (λ_m) and the background host population (λ_b), such that

$$1 + s = \lambda_m / \lambda_b \quad (1)$$

This definition[5] provides an intuitive interpretation for the fitness coefficient s: for example, $s = 1$ implies that the mutant cell population

[1]Evolution and Cancer Laboratory, Barts Cancer Institute, Queen Mary University of London, London, UK. [2]Department of Cell and Developmental Biology, University College London, London, UK. [3]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London, UK. [4]Evolutionary Genomics & Modelling Lab, Centre for Evolution and Cancer, Institute of Cancer Research, London, UK. [5]Departments of Medicine and Genetics, Stanford University School of Medicine, Stanford, CA, USA. [6]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [7]UCL Genetics Institute, University College London, London, UK. *e-mail: christopher.barnes@ucl.ac.uk; andrea.sottoriva@icr.ac.uk; t.graham@qmul.ac.uk

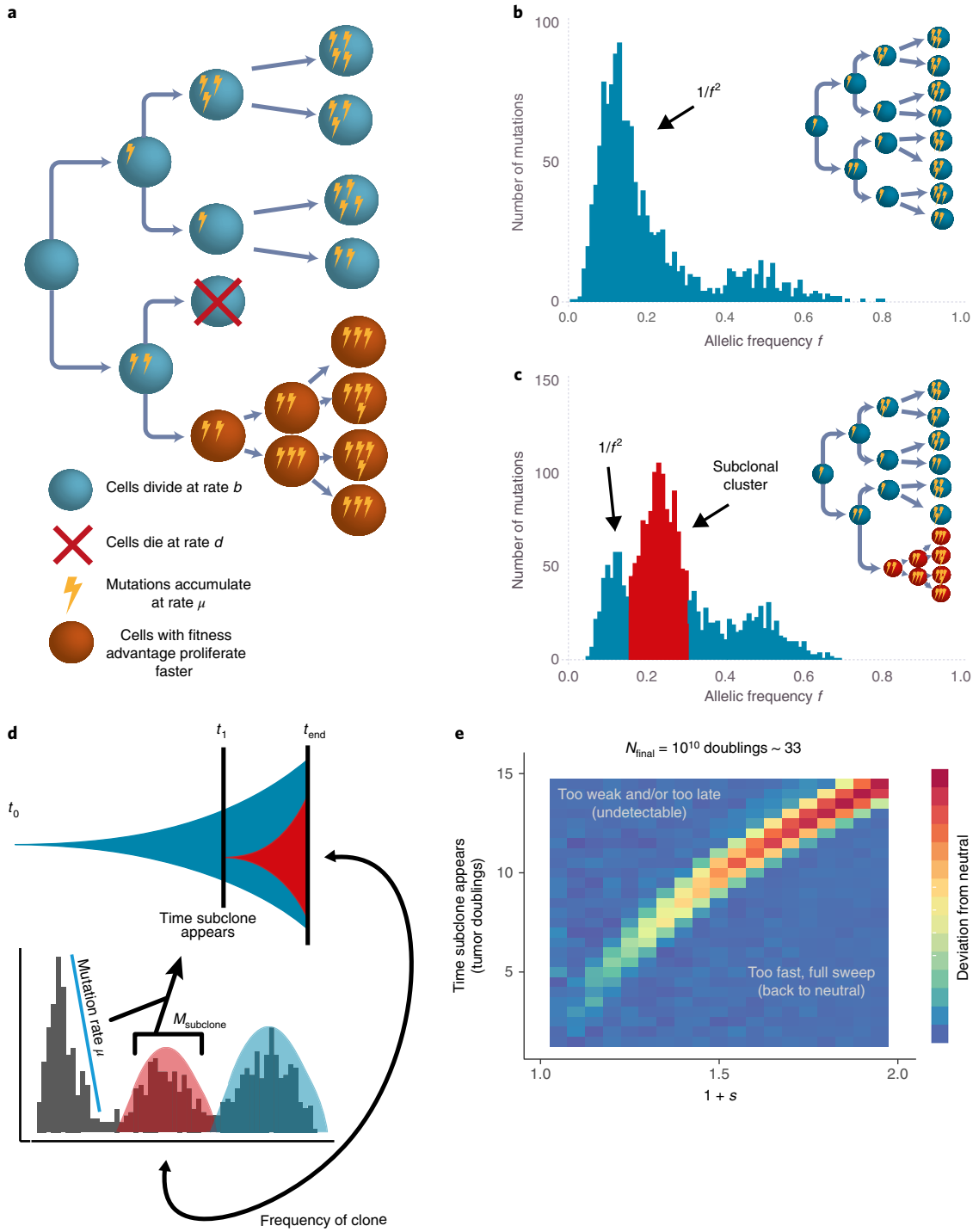

Fig. 1 | Modeling patterns of subclonal selection in sequencing data. a, In a stochastic branching process model of tumor growth, cells have birth rate b and death rate d, and mutations accumulate with rate μ. Cells with a fitness advantage (orange) grow at a faster net rate ($b - d$) than the host population (blue). **b**, The VAF distribution contains clonal (truncal) mutations at approximately $f = 0.5$ (in this example of a diploid tumor) and subclonal mutations ($f < 0.5$), which encode how a tumor has grown. In the absence of subclonal selection, a neutral $1/f^2$ tail describes the accumulation of passenger mutations as the tumor expands. **c**, A selected subclone produces an additional peak in the distribution, whereas a $1/f^2$-like tail is still present owing to passenger mutations that accumulate in both the original population and the new subclone. **d**, In the presence of subclonal selection, the magnitude and average frequency of a subclonal cluster of mutations (red) encode the age and size of the subclone, respectively, which in turn allow measurement of the clone's selective advantage. **e**, Frequentist power analysis of the detectability of an emerging selected subclone on simulated data. Only early and/or very fit subclones caused significant alterations of the clonal composition of a tumor, resulting in rejection of the neutral (null) model. Tumors were simulated to 10^6 cells and scaled to a final population size of 10^{10} with a mutation rate of 20 mutations per genome per division; each pixel represents the average value for the metric (area between curves) over 50 simulations.

grows twice as fast as the host tumor population, and $s=0$ implies that $\lambda_m=\lambda_b$, such that the subclone evolves neutrally with respect to the background population. Within the model, neutral evolution ($s=0$) leads to a VAF distribution characterized by a power-law-distributed subclonal tail of mutations[11,15–17] (Fig. 1b), where the cumulative number of mutations at a frequency f is proportional to the inverse of that frequency, $1/f$ (in the non-cumulative VAF distribution, such as in Fig. 1b, this shows as $\sim 1/f^2$). Alternatively, clonal selection ($s>0$) produces characteristic 'subclonal clusters' within the VAF distribution that have been observed in cancer genomes[18] (Fig. 1c). Notably, as neutral mutations continue to accumulate within each subclone, the $1/f$ tail is also present in tumors with selected subclones (Fig. 1c).

A mathematical analysis of the model indicates how subclonal clusters encode the underlying evolutionary dynamics of a subclone: the mean VAF of a cluster is a measure of the relative size of the subclone within the tumor, and the total number of mutations in the cluster (i.e., the area of the cluster) indicates the subclone's relative age (as later-arising subclones will have accumulated more mutations). Together, these two measures allow the fitness advantage s to be estimated[5,19]. We provide a summary derivation below; refer to the Supplementary Note for full details.

We define $t_0=0$ to be the time when the first transformed cancer cell begins to grow. At a later time t_1, a cell in the tumor acquires a subclonal 'driver' somatic alteration that confers a fitness advantage, giving rise to a new phenotypically distinct subclone that expands faster than the other tumor cells. We note that to measure selection dynamics it is not important what the actual driver event is—genetic (point mutation or copy number alteration), epigenetic or even microenvironmental drivers will all cause somatic mutations in the selected lineage to 'hitchhike'[20] to frequencies higher than expected under the neutral null model. The number of hitchhiking mutations, M_{sub}, acquired by the founder cell of the fitter subclone that has experienced Γ successful divisions between t_0 and t_1 is therefore

$$M_{sub}=\mu\Gamma \tag{2}$$

The relationship between the mean number of divisions of a lineage Γ and time measured in population doublings is $\Gamma=2\times\log(2)\times t_1$ (Supplementary Note). The mutation rate per population doubling can be estimated from the '$1/f$'-like tail[11]. For a subclone that emerges at time t_1, we would expect to observe M_{sub} mutations at some frequency $f_{sub}/2$ (for a subclone at a cancer cell fraction (CCF) f_{sub} in a diploid genome and assuming a sample with 100% tumor purity), and given the limited accuracy of VAF measurement inherent to next-generation sequencing, this will appear as a cluster of mutations with mean $f_{sub}/2$ in the VAF distribution. Therefore, equation (2) provides an estimate of t_1, the time at which the subclone appeared.

Assuming exponential growth and well-mixed populations, and considering that the subclone grows $1+s$ times faster than the background tumor population as defined by equation (1), the frequency of the subclone will grow in time according to

$$f_{sub}(t_{end})=\frac{e^{\lambda_b(1+s)(t_{end}-t_1)}}{e^{\lambda_b t_{end}}+e^{\lambda_b(1+s)(t_{end}-t_1)}} \tag{3}$$

This equation leads to an expression for the fitness advantage s given the frequency f_{sub} and the relative time of the subclone's appearance t_1

$$s=\frac{\lambda_b t_1+\ln\left(\frac{f_{sub}}{1-f_{sub}}\right)}{\lambda_b(t_{end}-t_1)} \tag{4}$$

Given an estimate of the age of the tumor expressed in population doublings t_{end}, equations (2) and (4) provide a means to measure the selective advantage of a subclone directly from the VAF distribution (Fig. 1d). t_{end} can be derived from the final tumor size N_{end} by the relationship $2^{t_{end}}=(1-f_{sub})\times N_{end}$. In the case of multiple subclones, equation (4) takes a slightly modified form (Supplementary Note). We note that equations (1)–(4) are known results in population genetics and have previously been used to describe the dynamics of asexual haploid populations[5].

Our previously presented frequentist approach to detect subclonal selection from bulk sequencing data involves an R^2 test statistic[19] to reject the hypothesis of neutral evolution ($s=0$), the null model in molecular evolution[21]. Here we extended our previous work to examine different test statistics for assessing deviations from the null neutral model (Supplementary Figs. 1–3 and Methods). However, the frequentist approach has limitations: it requires one to choose the interval of the VAF distribution to test and notably only allows for the rejection of the null hypothesis (which is not necessarily evidence for the null itself).

To address these shortcomings, we implemented a Bayesian statistical inference framework (Supplementary Fig. 4 and Methods) that fit our computational model incorporating both selection and neutrality to sequencing data and that simultaneously estimated the subclone fitness, time of occurrence and the mutation rate. This method allowed us to perform Bayesian model selection[22] for the number of subclones within the tumor and to specifically calculate the probabilities that a tumor contained zero subclones ($s=0$; neutral evolution) or one or more subclones (non-neutral evolution). The advantage of the Bayesian approach is that we could directly ask which model (neutral or non-neutral) was best supported by the data, using the whole VAF distribution.

Our framework can model mutation, selection and neutral drift by using a classical stochastic branching process[13] while integrating several confounding factors and sources of noise in bulk sequencing data, principally allele sampling and depth of sequencing (Supplementary Note and Methods). This approach allowed sample-based schemes to be designed such that the data-generating process could be mimicked to account for complex experimental biases. Despite these confounding factors, we found that the $1/f$ tail accurately measured the mutation rate, even in the presence of subclonal clusters (Supplementary Fig. 5), and our inferred value of $1+s$ was largely insensitive to the final tumor size (N_{end}) when this value was realistically large ($N_{end}>10^9$) (Supplementary Fig. 6 and Supplementary Note).

We note that the theoretical framework is based on the assumption of exponential growth, which is a growth pattern that is well supported by empirical data in many cancer types[23–25]. The effect of alternate models of growth, such as logistic and Gompertzian growth, is explored in the Supplementary Note. We also implemented a cancer stem cell model in which only a subset of cells had unlimited proliferation potential and found that, for the purposes of this study, this had little effect on the expected VAF distribution, which in this scenario only measured events that occurred in the stem cell compartment (Supplementary Fig. 7).

Recovery of evolutionary dynamics in synthetic tumors. First, we assessed the degree to which subclonal selection was detectable within VAF distributions by performing a frequentist power analysis to examine the conditions under which we correctly rejected the null when the alternative (selection present) was true. We performed simulations to measure the values of t_1 (time of subclone formation) and s (magnitude of the selective advantage of a subclone) that led to observable deviations from the null neutral model (Methods) in high-depth sequencing data (100×). Only subclones that arose sufficiently early (small t_1) or that were very fit (large s) were able to produce detectable deviations in the clonal composition of the tumor (Fig. 1e).

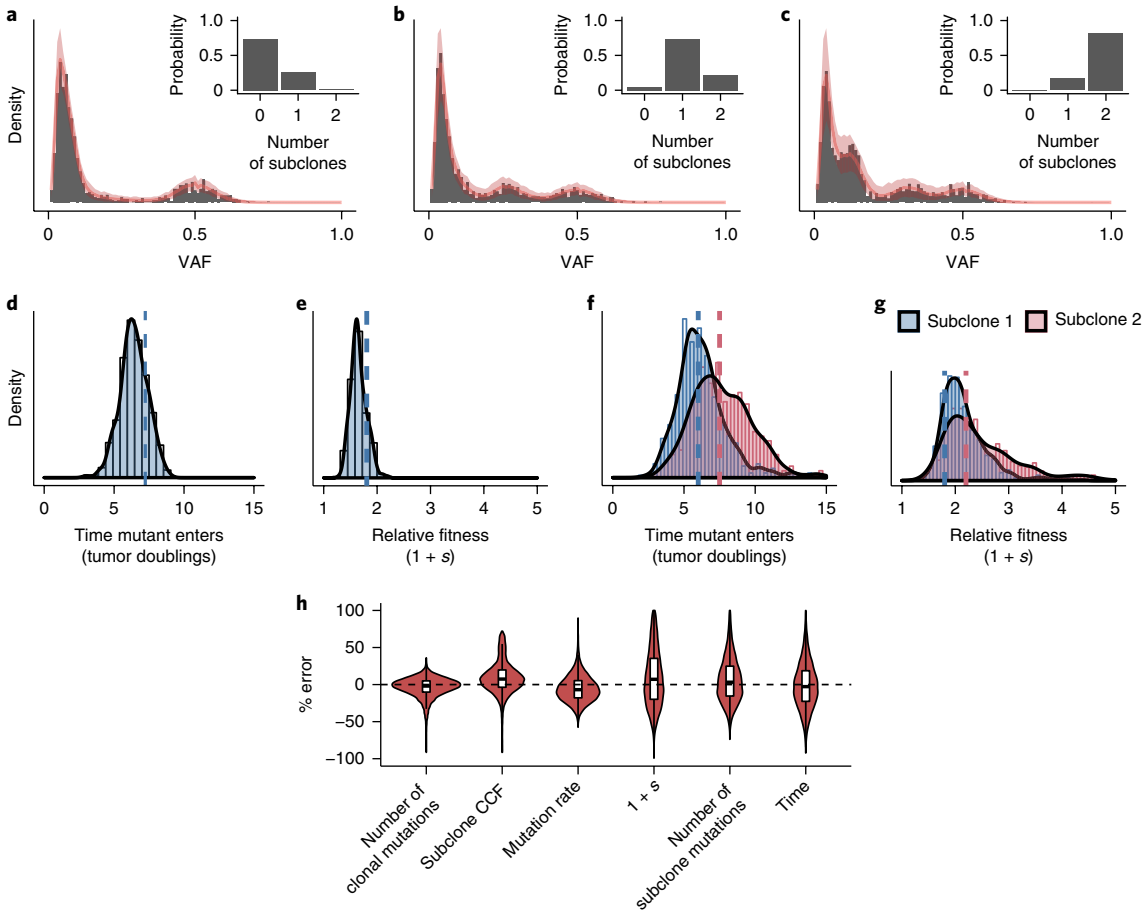

Fig. 2 | Accurate recovery of evolutionary parameters from simulated data using approximate Bayesian computation. a–c, Our method recovered the correct clonal structure in simulated tumor data (150× depth of sequencing, mutation rate of 16 mutations per tumor doubling) for representative examples of a neutral case (**a**), a one-subclone case (**b**) and a two-subclones case (**c**). Gray bars are simulated VAF data, solid red lines indicate the median histograms from the simulations that were selected by the statistical inference framework (500 posterior samples), and shaded areas are 95% confidence intervals. **d–g**, The inferred posterior distributions of the evolutionary parameters contained the true values (dashed lines) for the time of emergence of the subclones (**d**,**f**) and the selection coefficient '1+*s*' (**e**,**g**). **h**, The mean percentage error in inferred parameter values across a virtual tumor cohort ($n=100$ tumors) was below 10%. Box plots show the median and interquantile range (IQR); the upper whisker is the 3rd quantile + 1.5 × IQR; and the lower whisker is the 1st quantile − 1.5 × IQR.

We then applied our Bayesian framework to estimate evolutionary parameters from synthetic data (VAF distributions derived from computational simulations of tumor growth with known parameters). Our framework identified the correct underlying model with high probability for representative examples of a neutrally growing tumor (Fig. 2a), a tumor with a single subclone (Fig. 2b) and a tumor with two subclones (Fig. 2c), and it also recovered the evolutionary parameters in each case (Fig. 2d–g). Given that we modeled tumor growth as a stochastic process, variability in our estimates was expected (Supplementary Note). In a cohort of 100 synthetic tumors (20 examples selected in Supplementary Fig. 8) for which the ground truth was known, the mean percentage error on parameter inference was below 10% (Fig. 2h). The stochasticity also explained the width of the posterior distributions (Fig. 2d–g). In particular, the rate of stochastic cell death has a large effect on the variability of lineage age, and consequently it can cause a slight overestimation of the mutation rate and the variability in the time taken for a lineage to clonally expand increases with increased cell death (Supplementary Note).

Monte Carlo analysis indicated that accurate measurement of subclonal evolutionary dynamics required high depth (>100×) for both whole-exome sequencing (WES) and whole-genome sequencing (WGS) (Supplementary Fig. 9). This analysis demonstrates how clonal structure becomes progressively obscured as sequencing depth decreases. Depths of sequencing of less than 100× precluded a robust quantification of subclonal dynamics, and moreover the neutral model was preferred by our Bayesian model selection framework, even when it was false (Supplementary Fig. 9). Notably, this analysis showed that, even in some cases when selection was present (particularly weak selection), neutral evolution was the most parsimonious description of the data. In other words, the observed dynamics were then 'effectively neutral'. In addition, although the increased mutational information provided by WGS and higher sequencing depths made quantification of subclonal structure more robust, this could also identify (neutrally) drifting populations that might have been falsely ascribed as a selected clone (Supplementary Fig. 10). We also investigated the robustness of the inference method to tumor purity and the CCF of the subclone and found that at 100× sequencing depth a minimum purity of 50% was needed to confidently identify subclones with CCF >30% (15% VAF in a diploid genome) (Supplementary Fig. 11).

Detectable subclones have a large selective advantage. We first used our approach to quantify evolutionary dynamics in primary human cancers, for which high-depth (>150×) and validated sequencing data were available. We considered WGS of a single

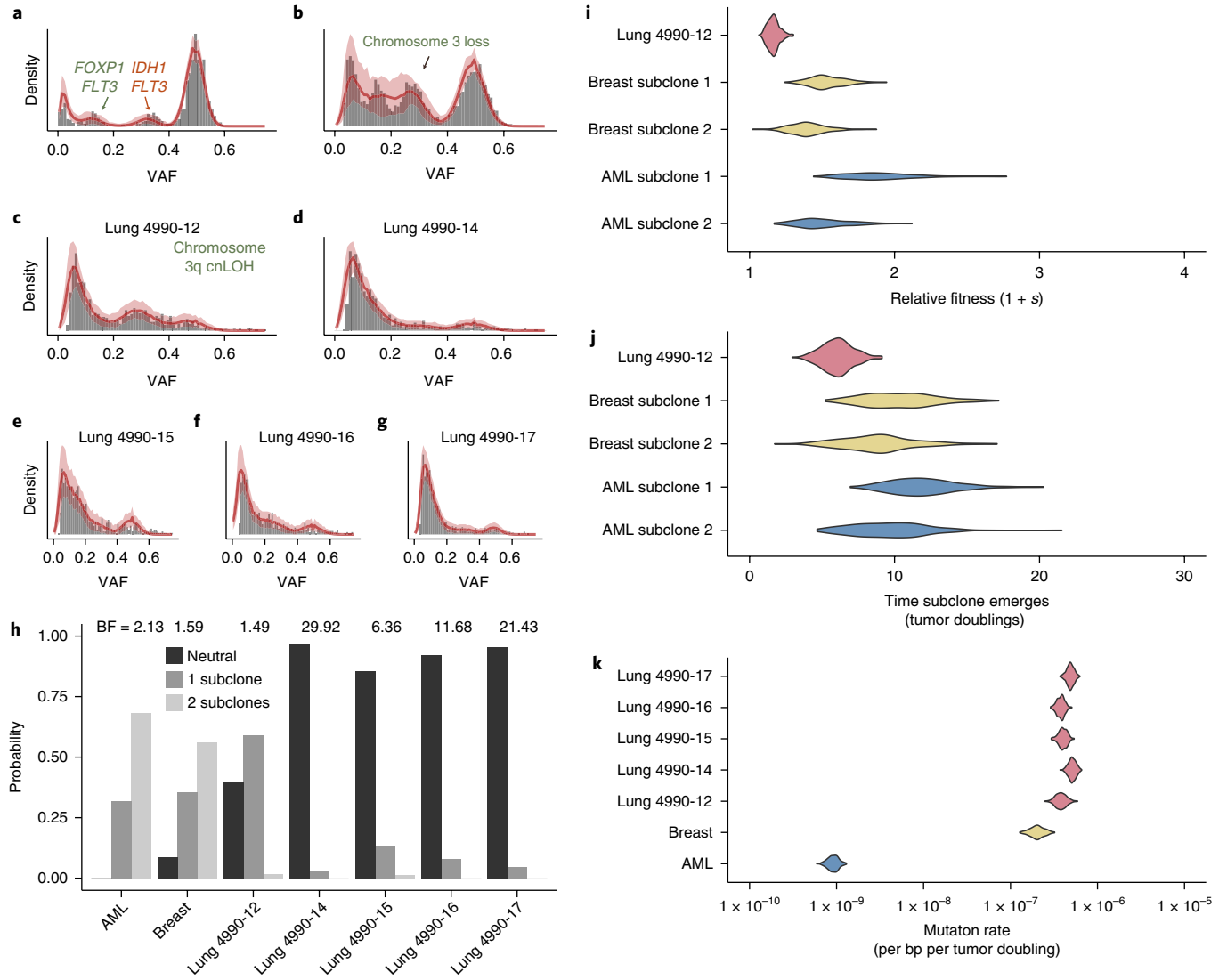

Fig. 3 | Quantifying selection from high-depth bulk sequencing of human cancers. a,**b**, Both an AML sample (**a**) and a breast cancer sample (**b**) sequenced at whole-genome resolution showed evidence of two selected subclones. **c**–**g**, In the case of a multiregion whole-exome-sequenced case of lung cancer, one sample showed evidence of a single subclone (**c**), whereas four other samples (**d**–**g**) from the same patient were consistent with the neutral model. Gray bars are the data, solid red lines indicate the median histograms from the simulations that were selected by the statistical inference framework (500 posterior samples), and shaded areas are the 95% credible intervals. **h**, Bayesian model selection reports the expected clonal structure for each case (Bayes factors (BFs) reported above histograms). **i**, Inferred subclone fitness advantages were 20% and 80% faster than that for the original population. **j**, Inferred times of subclone emergence indicate that subclones arise within the first 15 tumor-population doublings. **k**, Inferred mutation rates are of the order of 10^{-7} mutations per base per tumor doubling in solid tumors, but are ~10^{-9} in AML, reflecting the respective differences in mutational burden between cancer types. All posterior distributions were generated from 500 samples.

acute myeloid leukemia (AML) sample[26], WGS of a single breast cancer sample[18] and multiregion high-depth WES of a lung adenocarcinoma[27]. To avoid the confounding effects of copy number changes, we exploited the hitchhiking principle and restricted our analysis to consider only somatic single-nucleotide variants (SNVs) that were located within diploid regions (Methods). After correction for cellularity, the 'clonal cluster' at VAF = 0.5 and a potentially complex distribution of mutations with VAF <0.5 representing the subclonal architecture were clearly observable.

The AML and breast cancer cases both showed evidence of two subclonal populations, which corroborated the initial studies, but the lowest-frequency cluster was instead found to be a consequence of within-clone neutral mutations (Fig. 3a,b,h). Measurement of the evolutionary dynamics showed that for both cancers the subclones had considerably large fitness advantages (>20%; Fig. 3i) and emerged within the first 15 population doublings (Fig. 3j). In the AML sample, subclone 1 (highest-frequency subclone) had putative driver mutations in *IDH1* and *FLT3* and subclone 2 had a distinct *FLT3* mutation and a *FOXP1* mutation. In the breast cancer sample, no putative driver point mutations were found in the subclonal clusters, but we noted that the original analysis found that subclone 1 (highest-frequency subclone) had lost one copy of chromosome 13. Notably, the breast cancer sample also exhibited a 100-fold-higher mutation rate per tumor doubling than the AML sample (Fig. 3k). We note that our mutation rate estimate corresponded to the number of mutations per base per population doubling. Owing to the high amounts of cell death or possibly differentiation (both leading to lineage extinction), doubling cancer volume may require several rounds of cell division. To derive the mutation rates per base per division, an independent measurement of the probability

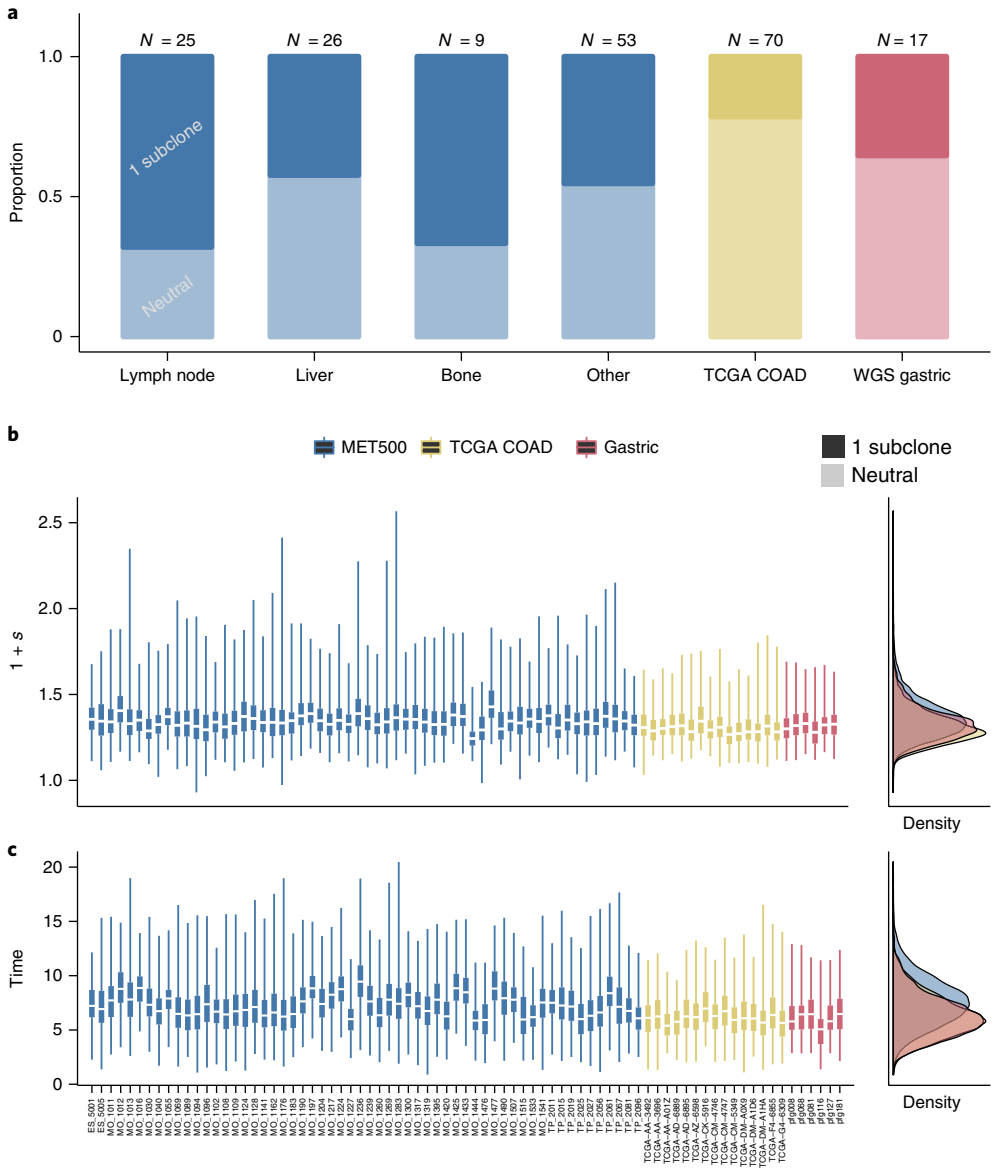

Fig. 4 | Quantifying selection in large cohorts of primary tumors and metastatic lesions. a, 21% of colon cancers ($N=70$) from TCGA samples (sequenced to sufficient depth and with sufficient cellularity for statistical inference), 29% of WGS of gastric cancers ($N=17$) (data from ref. [30], filtered for cellularity) and 53% of metastases ($N=113$) from sites had evidence of differentially selected subclones. **b**,**c**, When present, differentially selected subclones were found to have large fitness advantages with respect to the host population (**b**) and to emerge early during growth (**c**). BFs for subclonal structures for all data are reported in Supplementary Table 5. Posterior distributions were generated from 500 samples. Box plots show the median and IQR; the upper whisker is the 3rd quantile + 1.5 × IQR; and the lower whisker is the 1st quantile − 1.5 × IQR.

β of a cell division giving rise to two surviving lineages is required (equation (11); Methods and Supplementary Note). Mutational signature analysis[28] of subclonal mutations provided support for the assumption of a constant mutation rate during subclone evolution (Supplementary Fig. 12 and Methods).

In the lung adenocarcinoma case, multiple tumor regions ($n=5$) had been sequenced to high depth. Among these regions, only one (region 12) showed strong evidence of a new subclone (Fig. 3c,h; Bayes factor (BF) = 1.49), with a measured selective advantage of 30% (Fig. 3i), whereas for all of the other regions a neutral evolutionary model was most probable (Fig. 3d–g; BF = 6.36–29.92). Region 12 had unique copy number alterations on chromosome 3 that could plausibly have caused the subclonal expansion (Supplementary Fig. 13). Together, these data show spatial heterogeneity of the evolutionary dynamics within a single tumor.

We then applied our analysis to four additional large cohorts of variable sequencing depth: WES of colon cancers from The Cancer Genome Atlas (TCGA[29]; Supplementary Fig. 14), WGS of gastric cancers from Wang et al.[30] (Supplementary Fig. 15), WES of lung cancers from the TRACERx trial[31] (Supplementary Fig. 16) and WES of metastasis samples (multiple sites) from the MET500 cohort[32] (Supplementary Fig. 17). On the basis of our previous analysis for the minimum data quality needed (Supplementary Fig. 11), we selected samples with purity >40% and ≥25 subclonal mutations for further analysis. Differentially selected subclones were detected in 29% (5/17 cases) of the gastric cancers and 21% (15/70 cases) of the colon cancers (Fig. 4a). Of note, the MET500 data had a higher proportion of tumors with selected subclones (51%, 58/113). The measured selective advantage of these subclones was large (>20%) and emerged during the first few tumor doublings across all cohorts

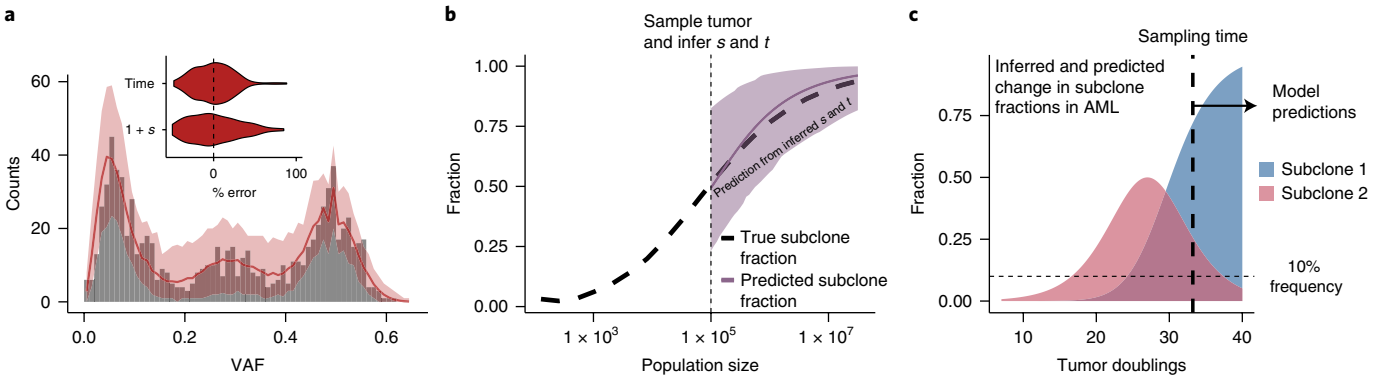

Fig. 5 | Predicting the future evolution of subclones. a, The VAF distribution of an in silico tumor sampled at 10^5 cells was used to measure the fitness and time of emergence of a subclone. Gray bars are the simulated data, solid red lines indicate the median histograms from the simulations that were selected by the statistical inference framework (500 posterior samples), and shaded areas are the 95% confidence intervals. The inset shows error from ground truth. 500 posterior samples were taken to perform the inference. **b**, The values from **a** were used to predict the spread of the subclone as the tumor grew to 10^7 cells, showing that the predictions matched the ground truth. Predictions were made by extrapolating from the posterior distribution of '1+*s*' using equations in the main text. The solid line shows the median value from the posterior distribution, and the shaded area shows the 95% confidence interval. **c**, Using the same approach in the AML sample, where we measured '1+*s*', t_1 and t_2, we would predict that subclone 2 would become dominant within 3–4 further tumor doublings, whereas subclone 1 would become too small to be detected.

(Fig. 4b,c). We note that, in the case of the metastases, time was measured relative to the founding of the metastatic lesion, and differential selection of the subclone was measured relative to the other cells in the metastasis. Eventual founder effects in the metastasis are, by definition, clonal events in the sample and so do not appear in the subclonal VAF spectrum. We also observed similarly large fitness advantages of subclones within the TRACERx cohort, in which 97% of cases (36/37 cases suitable for our analysis) were characterized by non-neutral dynamics (Supplementary Figs. 16 and 18).

Forecasting cancer evolution. Measuring the evolutionary dynamics of individual human tumors facilitates prediction of the future evolutionary trajectory of these malignancies[33]. Specifically, we can predict how the clonal architecture of a tumor is expected to change over time (in the absence of new drivers); such predictions could be useful, for instance, to decide how often to sample a tumor when making treatment decisions. We note that we can only predict the future subclonal structure of a tumor by assuming that environmental conditions stay the same—for example, that subclone selective advantages are constant—and we note that intervention, such as treatment, is likely to invalidate this assumption.

Suppose a biopsy is taken and the fitness of a subclone is measured at some time t, we can then ask how long it will take for the subclone to become dominant (>90% CCF) in the tumor. From our model, the time for a subclone to shift from a CCF of f_1 to a CCF of f_2 given a relative fitness advantage s would be

$$\Delta T = \frac{\log\left(\frac{f_2}{1-f_2}\right) - \log\left(\frac{f_1}{1-f_1}\right)}{\lambda s} \quad (5)$$

Figure 5 shows an in silico implementation of this method. The fitness advantage of a subclone was measured within a tumor at size $N = 10^5$ using the Bayesian inference framework (Fig. 5a), and the inferred values were then used to predict subsequent growth of the subclone. The prediction represented the ground truth quite accurately (Fig. 5b).

In the case of the examined AML sample (Fig. 3a), the measured fitness advantages predicted the future clonal structure of the malignancy (in the absence of treatment). Specifically, the larger of the two subclones that was present at the time point at which the tumor was sampled was predicted to take over the tumor, whereas the smaller clone was projected to become too rare to remain detectable (Fig. 5c). Despite the assumption of constant conditions, our framework could be extended in the future to simulate treatment effects when those mechanisms are known.

Discussion

Here we have demonstrated how the VAF distribution can be used to directly measure evolutionary dynamics of tumor subclones. We confirmed that subclonal selection causes an over-representation of mutations within the expanding clone, which is manifested as an additional 'peak' in the VAF distribution, as suggested by many recent studies[18,26,34]. However, irrespective of subclonal selection, the tumor will still show an abundance of low-frequency variants (a 1/*f*-like tail) as a natural consequence of tumor growth, in which the number of new mutations is proportional to the population size.

Our quantitative measurement of the selective advantage (relative fitness) of an expanding subclone showed that detectable subclones had experienced remarkably large fitness increases, in excess of 20% greater than that of the background tumor population. Large increases in subclone fitness were also observed in metastatic lesions, indicating that there can still be ongoing adaption, even in late-stage disease, perhaps as a consequence of treatment. Because selection is inferred using only SNVs that shift in frequency owing to hitchhiking, differential fitness can be measured by our analysis regardless of the underlying mechanism. Genetic driver mutations found within a subclone are one possible cause for the fitness increase.

The values of fitness advantage we infer in human malignancies are similar to those from reports of experimental systems. Evidence from growing human pluripotent stem cells indicates that *TP53* mutants may have a fitness advantage as high as 90% ($1+s=1.9$)[35] and that single chromosomal gains can provide a fitness advantage of up to 50%[36] (range 20–53%). A study of the competitive advantage of mutant stem cells in the mouse intestine during tumor initiation (at constant population size) showed that *KRAS*- and *APC*-mutated stem cells have a ~2- to 4-fold-increased fixation probability in single crypts[37] and that *TP53*-mutated cells in mouse epidermis exhibit a 10% bias toward self-renewal[38]. Moreover, our inferred fitness advantages are comparable to large fitness advantages measured in bacteria[39]. Nevertheless, we acknowledge that current experimental systems may differ significantly from in vivo human tumor growth and that new model systems are necessary to test these

measurements. We also note that we are only able to measure large changes in fitness and that additional efforts will be needed to measure the complete distribution of fitness effects (DFEs) within cancers. Furthermore, the inferred fitness value is sensitive to the underlying stochastic evolutionary model, and thus caution is warranted in directly comparing fitness values.

Our inferred in vivo mutation rates per population doubling are also in line with experimental evidence. Seshadri et al.[40] reported somatic mutation rates of 5.5×10^{-8} to 24.6×10^{-8} in normal lymphocytes and a 10- to 100-fold increase in mutation rate in cancer cell lines, such as B cell lymphoma (5.2×10^{-7} to 13.1×10^{-7}) and ALL (66.6×10^{-7}). A recent analysis of a mouse tumor model indicates that somatic mutation rates in neoplastic cells are 11× higher than those in normal tissue[41].

Our analysis highlights that, even if cancer subclones experience weak selection, this is not sufficient to markedly alter the clonal composition of the tumor and, therefore, to cause the VAF distribution to deviate detectably from the distribution expected under neutrality. It is important to note that the (initial) growth of tumors makes them peculiar evolutionary systems, as tumor growth dilutes the effects of selection[42]. Thus, our analysis does not discount the possibility of a multitude of 'mini-drivers'[43], but it does show that these must have a corresponding 'mini' effect on the subclonal composition of a tumor (and that the VAF distribution in mini-driver tumors is well described by a neutral model). We note, however, that the ratio of nonsynonymous to synonymous variants (dN/dS), a classical test for selection, identified only a small subset of genes with extreme dN/dS values, which are indicative of strong selection[21,44].

Our previous analysis[11] suggested that neutral dynamics were rejected in a higher percentage of colon cancers (approximately 65%) than the 21% reported here. The discrepancy is explained by the stochasticity in the evolutionary process, where chance events can lead to deviations from the neutral $1/f$ distribution. Unlike our previous analytic derivation, the Bayesian model selection framework presented here captures this stochasticity (and hence neutral evolution is preferred in a greater proportion of samples).

Our measurement of evolutionary trajectories facilitates mechanistic prediction of how a tumor changes over time, as demonstrated in our in silico prediction (Fig. 5a,b), with implications for anticipating the dynamics of treatment-resistant subclones. This may have particular value for novel evolutionary therapeutic approaches such as 'adaptive therapy', for which the goal is to maintain the existence of competing subclones that mutually suppress the growth of one another[45,46]. Our measurements of relative clone fitness could potentially be used to optimize treatment regimens to maintain the coexistence of competing populations.

We acknowledge that features not described in our model—for example, the spatial structure of the tumor—could affect the estimates of the evolutionary parameters[12,42,47]. Indeed, our analysis shows that there can be heterogeneity in the evolutionary process within a tumor (only one out of five regions of a single lung tumor showed strong evidence of subclonal selection). Spatial models of tumor evolution can help elucidate other important biological parameters, such as the degree of mixing within tumor cell populations, a purely spatial phenomenon that cannot be quantified using nonspatial models such as ours. We have recently shown how multiple samples per tumor increase the power to detect selection, in part because of the increased probability of sampling across a 'subclone boundary' where selection is evident[12]. We also acknowledge that complex, undetectable intermediate dynamics in the evolution of subclones, such as multiple small subclonal expansions before a subclone becomes detectable, are not modeled within our framework.

In summary, we have developed a quantitative framework to infer the timing and strength of subclonal selection in vivo in human malignancies. This is a step toward enabling mechanistic prediction of cancer evolution.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0128-6.

Received: 19 May 2017; Accepted: 23 March 2018;
Published online: 28 May 2018

References

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
2. Gay, L., Baker, A.-M. & Graham, T. A. Tumor cell heterogeneity. *F1000Res* **5**, 238 (2016).
3. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past. *Cell* **168**, 613–628 (2017).
4. Burrell, R. A. & Swanton, C. Re-evaluating clonal dominance in cancer evolution. *Trends Cancer* **2**, 263–276 (2016).
5. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics*. (Sinauer Associates, Inc.: Sunderland, MA, USA, 1997).
6. Marjoram, P. & Tavaré, S. Modern computational approaches for analyzing molecular-genetic-variation data. *Nat. Rev. Genet.* **7**, 759–770 (2006).
7. Fu, Y. X. & Li, W. H. Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14**, 195–199 (1997).
8. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
9. Tsao, J. L. et al. Colorectal adenoma and cancer divergence. Evidence of multilineage progression. *Am. J. Pathol.* **154**, 1815–1824 (1999).
10. Tsao, J. L. et al. Genetic reconstruction of individual colorectal tumor histories. *Proc. Natl Acad. Sci. USA* **97**, 1236–1241 (2000).
11. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
12. Sun, R. et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.* **49**, 1015–1024 (2017).
13. Durrett, R. *Branching Process Models of Cancer*. (Springer: New York, 2015).
14. Bozic, I. et al. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA* **107**, 18545–18550 (2010).
15. Cheek, D. & Antal, T. Mutation frequencies in a birth–death branching process. Preprint at https://arxiv.org/abs/1710.09783 (2017).
16. Kessler, D. A. & Levine, H. Scaling solution in the large population limit of the general asymmetric stochastic Luria–Delbrück evolution process. *J. Stat. Phys.* **158**, 783–805 (2015).
17. Durrett, R. Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probab.* **23**, 230–250 (2013).
18. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
19. Levy, S. F. et al. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
20. Gillespie, J. H. Genetic drift in an infinite population. The pseudo-hitchhiking model. *Genetics* **155**, 909–919 (2000).
21. Wu, C.-I., Wang, H.-Y., Ling, S. & Lu, X. The ecology and evolution of cancer: the ultra-microevolutionary process. *Annu. Rev. Genet.* **50**, 347–369 (2016).
22. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26**, 104–110 (2010).
23. Honda, O. et al. Doubling time of lung cancer determined using three-dimensional volumetric software: comparison of squamous cell carcinoma and adenocarcinoma. *Lung Cancer* **66**, 211–217 (2009).
24. Peer, P. G., van Dijck, J. A., Hendriks, J. H., Holland, R. & Verbeek, A. L. Age-dependent growth rate of primary breast cancer. *Cancer* **71**, 3547–3551 (1993).
25. Tilanus-Linthorst, M. M. A. et al. *BRCA1* mutation and young age predict fast breast cancer growth in the Dutch, United Kingdom and Canadian magnetic resonance imaging screening trials. *Clin. Cancer Res.* **13**, 7357–7362 (2007).
26. Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).
27. Zhang, J. et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multi-region sequencing. *Science* **346**, 256–259 (2014).
28. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
29. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
30. Wang, K. et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* **46**, 573–582 (2014).

31. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
32. Robinson, D. R. et al. Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
33. Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. *Nat. Ecol. Evol.* **1**, 77 (2017).
34. Shah, S. P. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
35. Merkle, F. T. et al. Human pluripotent stem cells recurrently acquire and expand dominant-negative P53 mutations. *Nature* **545**, 229–233 (2017).
36. Rutledge, S. D. et al. Selective advantage of trisomic human cells cultured in nonstandard conditions. *Sci. Rep.* **6**, 22828 (2016).
37. Vermeulen, L. et al. Defining stem cell dynamics in models of intestinal tumor initiation. *Science* **342**, 995–998 (2013).
38. Klein, A. M., Brash, D. E., Jones, P. H. & Simons, B. D. Stochastic fate of p53-mutant epidermal progenitor cells is tilted toward proliferation by UV B during pre-neoplasia. *Proc. Natl Acad. Sci. USA* **107**, 270–275 (2010).
39. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc. Natl Acad. Sci. USA* **91**, 6808–6814 (1994).
40. Seshadri, R., Kutlaca, R. J., Trainor, K., Matthews, C. & Morley, A. A. Mutation rate of normal and malignant human lymphocytes. *Cancer Res.* **47**, 407–409 (1987).
41. Lugli, N. et al. Enhanced rate of acquisition of point mutations in mouse intestinal adenomas compared to normal tissue. *Cell Rep* **19**, 2185–2192 (2017).
42. Sottoriva, A. et al. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
43. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer* **15**, 680–685 (2015).
44. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
45. Enriquez-Navas, P. M. et al. Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. *Science Transl. Med.* **8**, 327ra24 (2016).
46. Zhang, J., Cunningham, J. J., Brown, J. S. & Gatenby, R. A. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat. Commun.* **8**, 1816 (2017).
47. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. *Nat. Commun.* **7**, 12760 (2016).

Acknowledgements

We thank W. Huang and K. Chkhaidze for fruitful discussions. We are grateful to A. Chinnaiyan and M. Cieslik for providing us with data from the MET500 cohort and to S. Leung for providing access to the gastric cancer cohort. A.S. is supported by the Chris Rokos Fellowship in Evolution and Cancer and by Cancer Research UK (grant no. A22909). T.A.G. is supported by Cancer Research UK (grant no. A19771). C.P.B. is supported by the Wellcome Trust (grant no. 097319/Z/11/Z). B.W. is supported by the Geoffrey W. Lewis Postdoctoral Training fellowship. A.S. and T.A.G. are jointly supported by the Wellcome Trust (grant no. 202778/B/16/Z and 202778/Z/16/Z, respectively). C.C. is supported by awards from the NIH (R01CA182514), Susan G. Komen Foundation (IIR13260750) and the Breast Cancer Research Foundation (BCRF-16-032). M.J.W. is supported by a Medical Research Council student scholarship. This work was also supported by Wellcome Trust funding to the Center for Evolution and Cancer (grant no. 105104/Z/14/Z).

Author contributions

M.J.W. wrote all of the simulation code and performed mathematical and bioinformatics analysis; B.W. performed mathematical analysis; T.H. performed bioinformatics analysis; M.J.W., B.W., T.H., C.C., C.P.B., A.S. and T.A.G. analyzed the data; M.J.W., B.W., C.P.B., A.S. and T.A.G. wrote the manuscript; C.P.B., A.S. and T.G. jointly conceived, designed, supervised and funded the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0128-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.P.B. or A.S. or T.A.G.

Methods

Simulating tumor growth. We implemented a stochastic birth–death process simulation of tumor growth, followed by a sampling scheme that recapitulated the 'noise' of cancer sequencing data. The sampling scheme was required to ensure that the underlying evolutionary dynamics measured from the data were not confounded by such noise. We will first introduce the simulation framework for an exponentially expanding population where all cells have equal fitness, and then we will show how elements of the simulation can be modified to include differential fitness effects and non-exponential growth (see the Supplementary Note for details).

Tumor growth was assumed to begin with a single transformed cancer cell that had acquired the full set of alterations necessary for cancer expansion. In our model, this first cell will therefore be carrying a set of mutations (the number of these mutations can be modified) that will be present in all subsequent lineages and will thus appear to be clonal (present in all cells, and thus will generate the cluster of clonal mutations at frequency 0.5 for a diploid tumor) within the cancer population.

To simulate the tumor, and subclone evolution, we specified a birth rate b and a death rate d ($b > d$, for a growing population), meaning that the average population size at time t was

$$N(t) = e^{(b-d)\times t} \tag{6}$$

We set $b = \log(2)$ for all simulations, such that in the absence of cell death the population would double in size at every unit of time. The tumor would grow until it reached a specified size N_{end}, at which point the simulation would be stopped. At each division, cells acquire ν new mutations, where ν is drawn from a Poisson distribution with mean μ, the mutation rate per cell division. We assumed that new mutations are unique (infinite sites approximation). Not all divisions would result in new surviving lineages because of cell death and differentiation. The probability of a cell division producing a surviving lineage β in terms of the birth and death rates is given by

$$\beta = (b - d)/b \tag{7}$$

Simulating subclonal selection. To include the effects of subclonal selection, a mutant that has a higher net growth rate (birth minus death) than the host population was introduced into the population. We only considered the cases of one or two subclonal populations under selection at any given time. We deemed this simplification to be reasonable, as the number of large-effect driver mutations in a typical cancer is thought to be small (<10; see ref. [46]). Additionally, we found that sequencing depth >100× was required to resolve more than one subclone (Supplementary Fig. 9). Fitter mutants can have a higher birth rate, a lower death rate, or a combination of the two, all of which result in the mutant growing at a faster rate than the host population. Given that the host (or background) population has growth rate b_H and death rate d_H, and the fitter population has growth rate b_F and death rate d_F, we defined the selective advantage s of the fitter population as

$$1 + s = \frac{b_F - d_F}{b_H - d_H} \tag{8}$$

Fitter mutants can be introduced into the population with a specified selective advantage s and at a chosen time t_1, allowing us to explore the relationship between the strength of selection and the time the mutant enters the population.

Simulation method and parameters. We used a rejection-kinetic Monte Carlo algorithm to simulate the model[48]. Owing to the small number of possible reactions (we considered at most three populations with different birth and death rates), this algorithm was more computationally efficient than a rejection-free-kinetic Monte Carlo algorithm, such as the Gillespie algorithm. The input parameters of the simulation are given in Supplementary Table 1.

The simulation algorithm is as follows:

1. Initialize the simulation with one cell and set all simulation parameters;
2. Choose a random cell i from the population;
3. Draw a random number $r \sim \text{Uniform}(0, b_{max} + d_{max})$, where b_{max} and d_{max} are the maximum birth and death rates of all cells in the population;
4. Using r, cell i will divide with probability proportional to its birth rate b_i and die with probability proportional to its death rate d_i. If '$b_i + d_i$' < '$b_{max} + d_{max}$', then there is a probability that cell i will neither divide nor die. If $\beta = 1$ (i.e., no cell death), then in step 3 $d_{max} = 0$;
5. If a cell divides, then the daughter cells acquire ν new mutations, where $\nu \sim \text{Poisson}(\mu)$;
6. Time is increased by a small increment $\frac{1}{N(b_{max} + d_{max})}\tau$, where τ is an exponentially distributed random variable[49];
7. Go to step 2 and repeat until the population size is N_{end}.

The output of the simulation is a list of mutations for each cell in the final population.

Generating millions of simulations for parameter inference. A number of simplifications to our simulation scheme were made to improve computational efficiency when used in our Bayesian inference method, a procedure that requires potentially many millions of individual simulations to be run to get accurate inferences. Our ultimate goal was to measure the time at which the subclones emerged and their fitness. These parameters were measured in terms of tumor volume doublings, not in terms of cell division durations (as this is unknown in human tumors). Our approximations allowed us to quantify the relative fitness of subclones, which was measured in units of population doubling, from the VAF distribution. The approximations were as follows.

Approximation 1: we modeled differential subclone fitness by varying the birth rate only and setting the death rate to 0 (for example, for $\beta = 1$, all lineages survive). This increased the simulation speed because a smaller number of time steps were required to reach the same population size and ensured that the tumors never died out in our simulations.

Timing the emergence of subclones depended on the number of mutations that had accumulated in the first cell that gave rise to the subclone. This is the product of the number of divisions and the mutation rate ($n \times \mu$), or equivalently the number of tumor doublings times the effective mutation rate ($n_{doublings} \times \mu/\beta$). Given that we measured everything in terms of tumor doublings and the effective mutation rate, μ/β was the only measure available to us from the VAF distribution (from the low-frequency $1/f$ tail); we reduced our search space by fixing $\beta = 1$ and varying μ, recognizing that in reality the effective mutation rate was likely to have $\beta < 1$.

We do note, however, that cell death ($\beta < 1$) can affect our inferences in two ways. First of all, in the presence of one or more subclones, the low-frequency tail, which encodes μ/β, consists of a combination of two or more $1/f$ tails. If there are large differences in the β value between subclones, then the inference on the effective mutation rate from the gradient of the low-frequency tail may be incorrect. For example, a fitter subclone could arise due to decreased cell death rather than increased proliferation. To quantify this effect, we simulated subclones with differential fitness due to decreased cell death and measured the error on the inferred μ/β. Even in cases where the death rate was dramatically different in the subclone than in the host population ($\beta = 1.0$ versus $\beta = 0.5$), the mean error on the estimates of the mutation rate was 42% (Supplementary Fig. 5), which was significantly less than the order of magnitude previously measured between cancer types[11], and we conclude that the constant β assumption is therefore acceptable. We do acknowledge, however, that we may underestimate the effects of drift, which will be accentuated in tumors with high death rates.

Approximation 2: we simulated a smaller tumor population size, as compared to typical tumor sizes at diagnosis, and scaled the inferred values a posteriori. We note that the VAF distribution holds no information on population size (it measures only relative proportions), and furthermore, simulating realistic population sizes (on the order of tens or hundreds of billions of cells in human malignancies) is computationally unfeasible. To circumvent this, we generated synthetic datasets that captured the characteristics relevant to measuring fitness and the time that the subclones emerged, namely the effective mutation rate (μ/β) encoded by the low-frequency part of the distribution, the number of mutations in any subclonal cluster and their frequency. Theoretical population genetics was then used to transform these measurements into values of fitness and time (via equations (2) and (4)), and the values were scaled by the realistic population size $N_{end} = 10^{10}$.

Sufficient simulation length was required to allow the single cell that gives rise to the subclone time to accumulate the number of mutations that were ultimately observed in the empirical datum. In general, we found $N_{end} = 10^3$ to be sufficient, except for the breast cancer and AML samples, for which we used the more conservative $N_{end} = 10^4$. In general, $N_{end} = 10^4$ was sufficient to be able to measure the range of parameters considered in Fig. 1e.

Appropriately scaling the estimates of s requires an estimate of the age of the tumor in terms of tumor doublings. Using equation (4) with a final population size of N_{end}, we can calculate t_{end} as

$$t_{end} = \frac{\log((1 - f_{sub}) \times N_{end})}{\log(2)} \tag{9}$$

where f_{sub} is the frequency of the subclone. We assumed a realistic $N_{end} = 10^{10}$ to generate the posterior distributions in Figs. 3 and 4. We also generated posterior distributions for s as a function of N_{end} for the AML, breast and lung cancers. For realistically large N_{end} ($>10^9$) values, the exact choice had minimal effects on our inferred values of s (Supplementary Fig. 6).

To confirm that these assumptions did not invalidate our approach, we generated synthetic datasets with cell death and a large final population size (10^6). We then used our inference method (detailed below), with the simplifying assumptions, to infer the parameters used to generate these synthetic tumors. This demonstrated that we were able to accurately recover the input parameters when the simplifications were applied (Fig. 2).

Sampling. To mimic the process of data generation by high-throughput sequencing, we performed various rounds of empirically motivated sampling

of the simulation data. Sequencing data suffer from multiple sources of noise; most notably for this study, mutation counts (VAFs) were sampled from the true underlying frequencies in the tumor population (because of both the initial limited physical sampling of cells from the tumor for DNA extraction and the limited read depth of the sequencing). Additionally, it is challenging to discern mutations that are at low frequencies from sequencing errors, and the limited sampling of sequencing assays means that many low-frequency mutations are likely to not be measured at all. Consequently, only mutations with a frequency greater than ~5–10% with 100× sequencing depth are observable with certainty[49]. The ability to resolve subclonal structures is thus dependent on the depth of sequencing.

Our sampling scheme to generate synthetic datasets was as follows. For mutation i with true frequency $\mathrm{VAF}_{\mathrm{true}}$, the sequence depth D_i was binomially distributed

$$D_i \sim B_o\left(n=N, p=\frac{D}{N}\right)$$

for a tumor of size N. The sampled read count with the mutant was binomially distributed with the following parameters

$$f_i \sim B_o\left(n=D_i, p=\frac{\mathrm{VAF}_{\mathrm{true}}}{N}\right)$$

or, if overdispersed sequencing was modeled[50,51], we used the beta-binomial model, which introduces additional variance to the sampling

$$f_i \sim \mathrm{BetaBin}\left(n=D_i, p=\frac{\mathrm{VAF}_{\mathrm{true}}}{N}, \rho\right)$$

where ρ is the overdispersion parameter and $\rho=0$ reverts to the binomial model. Finally, the sequenced VAF for mutation i was given by $\mathrm{VAF}_i = f_i/D_i$.

Modeling stem cells. Stem cell architecture was modeled with two compartments: long-lived stem cells and short-lived non-stem cells. Stem cells divided symmetrically to produce two stem cells with probability α and asymmetrically to produce a single stem cell and a single differentiated cell with probability $1-\alpha$. Differentiated cells divided n further times before dying. At each division all of the cells accumulated mutations as described above. We used $\alpha=0.1$ and $n=5$. If $\alpha=1.0$, then the model was equivalent to the above exponential growth model.

Bayesian statistical inference. We used 'approximate Bayesian computation' (ABC) to infer the evolutionary parameters. We evaluated the accuracy of our inferences by using simulated sequencing data where the true underlying evolutionary dynamics were known. The simulation approach to generate synthetic data was used instead of a purely statistical approach, as the simulation naturally accounted for effects that would be difficult to represent in a pure statistical model (such as the convolution of multiple within-subclone mutations at lower frequency ranges). Furthermore, the posterior distribution reported from this method naturally accounts for uncertainties due to experimental noise and stochastic effects, such as Poisson-distributed mutation accumulation and stochastic birth–death processes. For in-depth discussion on these stochastic effects, see the Supplementary Note.

As in all Bayesian approaches, the goal of the ABC approach was to produce posterior distributions of parameters that give the degree of confidence that particular parameter values are true, given the data. Given a parameter vector of interest θ and data D, the aim was to compute the posterior distribution $\pi(\theta|D) = \frac{p(D\mid\theta)\pi(\theta)}{\beta(D)}$, where $\pi(\theta)$ is the prior distribution on θ and $p(\mathrm{D}|\theta)$ is the likelihood of the data given θ. In cases where calculating the likelihood is intractable, as was the case here where our model could not be expressed in terms of well-known and characterized probability distributions, approximate approaches must be sought. The basic idea of these 'likelihood-free' ABC methods is to compare simulated data, for a given set of parameter values, with observed data by using a distance measure. Through multiple comparisons of different input parameter values, we can produce a posterior distribution of parameter values that minimizes the distance measure, and in so doing accurately approximates the true posterior. The simplest approach is called the ABC rejection method, and the algorithm is as follows[52]:

1. Sample candidate parameters θ^* from prior distribution $\pi(\theta)$;
2. Simulate tumor growth with parameters θ^*;
3. Evaluate the distance δ between simulated data and target data;
4. If $\delta > \varepsilon$, then reject parameters θ^*;
5. If $\delta \leq \varepsilon$, then accept parameters θ^*;
6. Return to step 1.

We used an extension of the simple ABC rejection algorithm, called 'approximate Bayesian computation sequential Monte Carlo' (ABC SMC)[22]. This method achieves higher acceptance rates of candidate simulations and thus makes the algorithm more computationally efficient than the simple rejection ABC. It achieves this increased efficiency by propagating a set of 'particles' (sample parameter values) through a set of intermediate distributions with strictly decreasing ε until the target ε_T is reached, by using an approach known as sequential importance sampling[53]. The ABC SMC algorithm also allows for Bayesian model selection to be performed; that is, we can test various models on the data and see which fits best. This is done by placing a prior over models and performing inference on the joint space of models and model parameters, (m, θ_m). In contrast to many applications of ABC that use summary statistics, we used the full data distribution, thus avoiding issues of inconsistent BFs due to loss of information[54,55]. For further details on the algorithm, see ref. [22] and the Supplementary Note on the specific details of our implementation. BFs for all data are shown in Supplementary Tables 5 and 6. We found that the probability of neutrality was significantly correlated with our frequentist-based neutrality metrics and that the inferred mutation rates were highly similar (Supplementary Fig. 19).

The clonal structure of the cancer is encoded by the shape of the VAF distribution; we therefore used the Euclidean distance between the two cumulative distributions (simulated and target datasets) for our inference.

Testing selection in the frequentist paradigm. We also refined a simple analytical test to rapidly determine what evolutionary parameters of selection led to an observable deviation of the VAF distribution from that expected under neutrality. Previously[11], we showed that under neutrality the distribution of mutations with a frequency $>f$ is given by

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{\max}}\right) \tag{10}$$

We fit a linear model of $M(f)$ against $1/f$ and used the R^2 measure of the explained variance as our measure of the goodness of fit.

Another approach was to use the shape of the curve described by equation (10) and test whether our empirical data collapsed onto this curve. To implement this approach, here we defined the 'universal neutrality curve', $\overline{M}(f)$. Given an appropriate normalization of the data, the mutant allele frequency distribution governed by neutral growth would collapse onto this curve, although we recognize that deviations due to stochastic effects are possible. We can normalize the distribution described by equation (5) by considering the maximum value of $M(f)$ at $f=f_{\min}$.

$$\max(M(f)) = \frac{\mu}{\beta}\left(\frac{1}{f_{min}} - \frac{1}{f_{max}}\right) \tag{11}$$

$$\overline{M}(f) = \frac{\frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{\max}}\right)}{\max(M(f))} \tag{12}$$

$$\overline{M}(f) = \frac{\left(\frac{1}{f} - \frac{1}{f_{\max}}\right)}{\left(\frac{1}{f_{\min}} - \frac{1}{f_{\max}}\right)} \tag{13}$$

$\overline{M}(f)$ is independent of the mutation rate and death rate, and it therefore allows comparison with any dataset. To compare this theoretical distribution against empirical data, we used the Kolmogorov distance, D_k, the Euclidean distance between $\overline{M}(f)$ and the empirical data, and the area between $\overline{M}(f)$ and the empirical data. The Kolmogorov distance D_k is the maximum distance between two cumulative distribution functions. Supplementary Figure 1 provides a summary of the different metrics.

To assess the performance of the four classifiers, we ran 10^5 neutral and non-neutral simulations and compared the distribution of the test statistics for these two cases. Because of the stochastic nature of the model, not all of the simulations that include selection will result in subclones at a high enough frequency to be detected; therefore, to accurately assess the performance of our tests, we only included simulations for which the fitter subpopulation was within a certain range (20% and 70% fraction of the final tumor size). All four test statistics showed significantly different distributions between neutral and non-neutral cases (Supplementary Fig. 2). Under the null hypothesis of neutrality and a false-positive rate of 5%, the area between the curves was the test statistics with the highest power (67%) to detect selection, slightly outperforming the Kolmogorov distance and Euclidean distance, with the R^2 test statistics showing the poorest performance with a power of 61% (Supplementary Tables 2 and 3).

We also plotted receiver-operating characteristic (ROC) curves by varying the discrimination threshold of each of the tests of selection and calculating true-positive and false-positive rates (using a dataset derived from simulations with subclonal populations at a range of frequencies; Supplementary Fig. 3). This analysis showed that R^2 had the least discriminatory power, with the other three performing approximately equally well (see Supplementary Table 4 for area-under-the-curve values). Increasing the range of allowed subclone sizes decreased the

classifier performance, likely because the subclone could merge into the clonal cluster or $1/f$ tail when it took a more extreme size.

Bioinformatics analysis. Variant calls from the original studies were used for the AML data[26], the TRACERx[31] data and the MET500 data[32]. Our analysis of the TCGA colon cancer cohort and gastric cancers is explained in our previous publication[11]. For both these cohorts, we required the cellularity to be >0.4 to perform the analysis. For the breast cancer[18] and lung cancer[27] data, bam files from the original study were obtained, and variants were called using MuTect2[56] and filtered to require at least five reads reporting the variants in the tumor and zero reads in the normal sample. To mitigate the effects of low-frequency mutations arising from paralogous regions of the genome, we filtered any mutations for which 75-bp regions on either side of the mutations had multiple BLAST hits (minimum hit length of 100 bp, maximum of 3% mismatching bases).

Copy number aberrations could also potentially result in the multipeaked distribution we observed; hence, we used only mutations that were found in regions identified as diploid (and without copy-neutral loss of heterozygocity). The original AML study found no evidence of copy number alterations. For the TCGA colon cancer cohort, we used paired SNP array data to filter out mutations that fell in nondiploid regions. For the TRACERx data and the MET500 data, we used allele-specific copy number calls provided in the original studies to filter the data. For all of the other datasets, we applied the Sequenza algorithm to infer allele-specific copy number states and estimated the cellularity[57]. Because the original breast cancer study found evidence of subclonal copy number alterations in multiple chromosomes, we used only mutations on chromosome 3 for our analysis (Supplementary Fig. 20). B-allele frequencies (BAFs) of regions that were called as copy neutral by Sequenza in the lung cancer sample were consistent with a diploid genome (Supplementary Fig. 21).

We used the cellularity estimate provided by the Sequenza algorithm to correct the VAFs for each individual sample. For a cellularity estimate κ, the corrected depth for variant i will be $\overline{d_i} = \kappa \times d_i$. When cellularity estimates from Sequenza were unavailable (MET500 and TRACERx), we fitted the cellularity using our ABC method by including it as an additional parameter.

As noted, our simulation could account for the overdispersion of allele read counts. To measure the overdispersion parameter ρ, we fitted a beta-binomial model to the clonal cluster where we know $\mathrm{VAF}_{\mathrm{true}} = 0.5$. We used Markov chain Monte Carlo (MCMC) to fit the following model to the righthand side of the clonal cluster, to minimize the effects of the $1/f$ distribution or subclonal clusters

$$f_i \sim \mathrm{BetaBin}(n = D_i, p = \mathrm{VAF}_{\mathrm{true}}, \rho)$$

where D_i is the sequencing depth, f_i is the allele read count, and ρ is the overdispersion parameter. We then used this estimate for ρ in the simulation sampling scheme. Supplementary Figure 22 shows the fits to the clonal cluster for the AML data, using both the beta-binomial and binomial models, and Supplementary Table 7 reports the overdispersion parameter for each dataset. We also used this analysis to further refine the cellularity estimate provided by Sequenza, ensuring that the clonal cluster was centered at VAF = 0.5. We note that some of the overdispersion is likely artificial and was introduced by the cellularity correction.

Mutational signatures in the breast cancer sample and AML sample (Supplementary Fig. 12) were identified by using the deconstructSigs R package[58], using the latest mutational signature probability file from COSMIC. Signature assignment was restricted to signatures known to be active in the respective cancer types. All other parameters were set to default values. To generate confidence intervals, we bootstrapped the assignment by generating 50 datasets by sampling 90% of the mutations and running the regression on each dataset; we then reported the mean value and 95% confidence intervals.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. Code for the simulation and inference method, frequentist-based neutrality statistics and bioinformatic scripts is available at https://marcjwilliams1.github.io/quantifying-selection.

Data availability. Only publicly available data were used in this study, and data sources and handling of these data are described above.

References

48. Waclaw, B. et al. A spatial model predicts that dispersal and cell turnover limit intratumor heterogeneity. *Nature* **525**, 261–264 (2015).
49. Stead, L. F., Sutton, K. M., Taylor, G. R., Quirke, P. & Rabbitts, P. Accurately identifying low–allelic fraction variants in single samples with next-generation sequencing: applications in tumor subclone resolution. *Hum. Mutat.* **34**, 1432–1438 (2013).
50. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
51. Gerstung, M. et al. Reliable detection of subclonal single-nucleotide variants in tumor cell populations. *Nat. Commun.* **3**, 811 (2012).
52. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791–1798 (1999).
53. Del Moral, P., Doucet, A. & Jasra, A. Sequential Monte Carlo samplers. *J. R. Stat. Soc. Series B Stat. Methodol.* **68**, 411–436 (2006).
54. Robert, C. P., Cornuet, J.-M., Marin, J.-M. & Pillai, N. S. Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl Acad. Sci. USA* **108**, 15112–15117 (2011).
55. Barnes, C. P., Filippi, S., Stumpf, M. P. H. & Thorne, T. Considerate approaches to constructing summary statistics for ABC model selection. *Stat. Comput.* **22**, 1181–1197 (2012).
56. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
57. Favero, F. et al. Sequenza: allele-specific copy-number and mutation profiles from tumor-sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
58. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).

nature research

Corresponding author(s): Prof. Trevor Graham

☒ Initial submission ☐ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

Sample size was not guided by a pre-specified power analysis. Analysis is a sample-by-sample basis, so there were no issues arising of power to compare groups. We analysed all samples where we were able to obtain sufficiently high-quality sequence data.

2. Data exclusions

Describe any data exclusions.

Samples were excluded for having insufficient sequencing depth or cellularity. Samples with celluarlity>40% and number of subclonal mutations ≥25 were analysed.

3. Replication

Describe whether the experimental findings were reliably reproduced.

No experiments were performed.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Our analysis was carried out on a sample-by-sample basis, and hence randomisation was not relevant.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Our analysis was carried out on a sample-by-sample basis, and hence blinding was not relevant.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Simulation and inference algorithms are available at: https://marcjwilliams1.github.io/quantifying-selection We used the following software for the bioinformatic analysis: Mutect2 GATK v3.6 Sequenza v2.1.2 deconstructSigs v1.8

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No cell lines were used
b. Describe the method of cell line authentication used.	No cell lines were used
c. Report whether the cell lines were tested for mycoplasma contamination.	No cell lines were used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Only publicly available data was used. The original manuscripts describe these covariants.