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,*], Trevor A Graham[1,*]

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 Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK.

5 Departments of Medicine and Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

6 Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305, USA

7 UCL Genetics Institute, University College London, London, UK

* Correspondence should be addressed to:
christopher.barnes@ucl.ac.uk
andrea.sottoriva@icr.ac.uk; +44 208 722 4072
t.graham@qmul.ac.uk; +44 207 882 6231

Abstract

Subclonal architectures are prevalent across cancer types. However, the temporal evolutionary dynamics that produce tumour subclones remain unknown. Here we measure clone dynamics in human cancers using computational modelling of subclonal selection and theoretical population genetics applied to high throughput sequencing data. Our method determines the detectable subclonal architecture of tumour samples, and simultaneously measures 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 cancers, as well as metastatic deposits, showed that detectable subclones under selection, when present, consistently emerged early during tumour growth and had a large fitness advantage (>20%). Our quantitative framework provides new insight into the evolutionary trajectories of human cancers, facilitating predictive measurements in individual tumours from widely available sequencing data.

Introduction

Carcinogenesis is the result of Darwinian selection for malignant phenotypes, 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 tumours across all cancer types are highly heterogeneous[2,3] with complex clonal architectures[4]. However, because longitudinal observation of solid tumour 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 utilised to predict future tumour 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 intra-tumour genetic heterogeneity (ITH) – is a direct consequence of the unobservable evolutionary dynamics of tumour growth. Therefore, given a realistically constrained model of subclonal expansion, the pattern of ITH in a tumour can be used to infer its most probable evolutionary trajectory. ITH represented within the distribution of variant allele frequencies (VAF), as measured by high coverage sequencing, is particularly amenable to such an approach.

In this study, 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 tumours. This type of approach is established in the field of molecular evolution, where evolutionary processes are also difficult to measure directly[7,8], and examples of applications of these approaches to human cancers date back to the previous century[9,10].

Recently, we have shown that under a neutral “null” evolutionary model (i.e. when all selected driver alterations are truncal and present in all cancer cells), the VAF follows a characteristic power law distribution[11]. Subsequent simulations that modelled space and subclonal selection demonstrated that genetic divergence in multi-region sequencing data could be used to categorize tumours based on 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, 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 identify the characteristic patterns of subclonal selection in the cancer genome and measure fundamental evolutionary parameters in non-neutrally evolving human tumours.

Results

Theoretical framework of subclonal selection

We developed a stochastic computational model of tumour growth applicable to cancer genomic data that accounts for subclonal selection (see Methods). The model is based on a classical stochastic branching process approach from population genetics[13] that has been often used to model malignant populations[5,14] and is here extended 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 rate μ mutations per cell per division (Figure 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)

$$1 + s = \frac{\lambda_m}{\lambda_b}. \quad [1]$$

This definition[13] provides an intuitive interpretation for the fitness coefficient *s*: for example, *s*=1 implies that the mutant cell population grows twice as fast as the host tumour population, and *s*=0 implies $\lambda_m=\lambda_b$ such that the subclone evolves neutrally with respect to the background population. Within the model, neutral evolution (s=0) leads to a VAF distribution characterised by a power-law distributed subclonal tail of mutations[11,15-17] (Figure 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 Figure 1b, this shows as ~$1/f^2$). Alternatively, clonal selection (s>0) produces characteristic 'subclonal clusters' within the VAF distribution that have been observed in cancer genomes[18] (Figure 1c). Importantly, as neutral mutations continue to accumulate within each subclone, the 1/f tail is also present in tumours with selected subclones (Figure 1c).

A mathematical analysis of the model indicates how subclonal clusters encode the underlying evolutionary dynamics of a subclone: the mean VAF of the cluster is a measure of the relative size of the subclone within the tumour, 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[19]. We provide a summary derivation below and 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 tumour 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 tumour 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 higher frequencies than expected under the neutral null model. The number of hitchhiking mutations,

M_{sub} acquired by the founder cell of the fitter subclone which has experienced Γ successful divisions between t_0 and t_1 is therefore

$$M_{sub} = \mu\Gamma. \quad [2]$$

The relationship between the mean number of divisions of a lineage, Γ and time measured in population doublings is $\Gamma = 2log(2)t_1$ (see 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 f_{sub} in a diploid genome, and assuming a sample with 100% tumour purity), and given the limited accuracy of VAF measurement inherent to next generation sequencing this will appear as a cluster of mutations with a mean $f_{sub}/2$ in the VAF distribution. Therefore, Equation [2] provides an estimate of t_1, the time when the subclone appeared.

Assuming exponential growth and well mixed populations, and considering that the subclone grows *1+s* times faster than the background tumour population as defined by Equation [1], the frequency of the subclone will grow in time according to:

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

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

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

Given an estimate of the age of the tumour 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 (Figure 1d). t_{end} can be derived from the final tumour size N_{end} by the relation $2^{t_{end}} = (1 - f_{sub}) \times N_{end}$. 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 been previously used to describe the dynamics of asexual haploid populations [13].

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 (see Supplementary Figures 1-3 & Methods). However, the frequentist approach has limitations: it requires to choose the interval of the VAF distribution to test, and importantly 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 Figure 4 & Methods) that fits our computational model incorporating both selection and neutrality to sequencing data, and simultaneously estimates 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 tumour and specifically calculate probabilities that a tumour contained 0 subclones (s=0, neutral evolution), 1 or more subclones (non-neutral evolution). The advantage of the Bayesian approach is that we can directly ask which model (neutral or non-neutral) is best supported by the data, using the whole VAF distribution.

Our framework models mutation, selection and neutral drift 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 (see Methods and Supplementary Note). This approach allows sample-based schemes designed such that the data-generating process can be mimicked to account for complex experimental biases. Despite these confounding factors, we found that the 1/f tail accurately measures the mutation rate even in the presence of subclonal clusters (Supplementary Figure 5), and our inferred value of 1+s is largely insensitive to the final tumour size (N_{end}) when this value is realistically large ($N_{end}>10^9$) (Supplementary Figure 6 and Supplementary Note).

We note that the theoretical framework is based upon the assumption of exponential growth, which is a growth pattern well supported by empirical data in many cancer types[23-25]. The impact 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 where only a subset of cells has unlimited proliferation potential and found that for the purposes of this study this has little impact on the expected VAF distribution, which in this scenario only measure events that occur in the stem cell compartment (Supplementary Figure 7).

Recovery of evolutionary dynamics in synthetic tumours

First, we assessed the degree to which subclonal selection is detectable within VAF distributions by performing a frequentist power analysis to examine the conditions under which we correctly reject the null when the alternative (selection present) is true. We performed simulations to measure the values of t_1 (time of subclone formation) and s (magnitude of selective advantage of subclone) that lead to observable deviations from the null neutral model (see Methods) in high depth sequencing data (100X). Only subclones that arise sufficiently early (small t_1) or that were very fit (large *s*) were able to produce detectable deviations in the clonal composition of the tumour (Figure 1e).

We then applied our Bayesian framework to estimate evolutionary parameters from synthetic data (VAF distributions derived from computational simulations of tumour growth with known parameters). Our framework identified the correct underlying model with high probability for representative examples of a neutrally growing tumour (Figure 2a), a tumour with a single subclone (Figure

2b) and a tumour with 2 subclones (Figure 2c), and also recovers the evolutionary parameters in each case (Figures 2d-g). Given that we modelled tumour growth as a stochastic process, variability in our estimates was expected (see Supplementary Note). In a cohort of 100 synthetic tumours (20 examples selected in Supplementary Figure 8), where the ground truth was known, the mean percentage error on parameter inference was below 10% (Figure 2h). The stochasticity also explains the width of the posterior distributions (Figures 2d-g). In particular, the rate of stochastic cell death has a large effect on the variability of lineage age and consequently can cause a slight over-estimation of the mutation rate and variability in the time taken for a lineage to clonally expand increases with increased cell death (see Supplementary Note).

Monte Carlo analysis indicated that accurate measurement of subclonal evolutionary dynamics required high depth (>100X) for both whole-exome and whole-genome sequencing (Supplementary Figure 9). This analysis demonstrates how the clonal structure becomes progressively obscured as the sequencing depth decreases. Depths of sequencing of less than 100X preclude a robust quantification of subclonal dynamics, and moreover the neutral model is preferred by our Bayesian model selection framework, even when it is false (Supplementary Figure 9). Importantly, this analysis showed that even in some cases when selection is present (particularly weak selection), neutral evolution is the most parsimonious description of the data. In other words, the observed dynamics are then 'effectively neutral'. In addition, we note that while the increased mutational information provided by WGS and higher sequencing depths makes quantification of subclonal structure more robust, this can also reveal (neutrally) drifting populations that may be falsely ascribed as a selected clone (Supplementary Figure 10). We also investigated the robustness of the inference method to tumour purity and cancer cell fraction of the subclone finding that at 100X sequencing depth a minimum purity of 50% is needed to confidently identify subclones with cancer cell fraction >30% (15% VAF in a diploid genome), see Supplementary Figure 11.

Detectable subclones have a large selective advantage

We first used our approach to quantify evolutionary dynamics in primary human cancers where high depth (>150X) and validated sequencing data were available. We considered whole-genome sequencing (WGS) of a single AML sample[26], WGS of a single breast cancer sample[18] and multi-region high-depth whole exome sequencing (WXS) 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 (see 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 2 subclonal populations, corroborating the initial studies but instead finding the lowest frequency cluster to be a consequence of all within-clone neutral

mutations[18,26] (Figure 3a,b,h). Measurement of the evolutionary dynamics showed that for both cancers the subclones had considerably large fitness advantages (>20%, Figure 3i) and emerged within the first 15 population doublings (Figure 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 note that the original analysis found that subclone 1 (highest frequency subclone) had lost one copy of chromosome 13. Interestingly, the breast cancer sample also exhibited a 100-fold higher mutation rate per tumour doubling compared to the AML sample (Figure 3k). We note that our mutation rate estimate corresponds to the number of mutations per base per population doubling. Due to the high cell death and possibly differentiation in cancers (both leading to lineage extinction), doubling in volume may require several rounds of cell division. To derive the mutation rates per base per division an independent measurement of the probability β of a cell division to give rise to two surviving lineages is required (see Methods, Equation [9] and Supplementary Note). Mutational signature analysis[28] of subclonal mutations provided support for the assumption of a constant mutation rate during subclone evolution (Methods and Supplementary Figure 12).

In the lung adenocarcinoma case, multiple tumour regions (n=5) had been sequenced to high depth. Amongst these regions, only one region (region 12) showed strong evidence of a new subclone (Figures 3c,h, BF = 1.49) with a measured selective advantage of 30% (Figure 3j), while for all other regions a neutral evolutionary model was most probable (Figures 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 Figure 13). Together these data show spatial heterogeneity of the evolutionary dynamics within a single tumour.

We then applied our analysis to 4 additional large cohorts of variable sequencing depth: WXS colon cancers from TCGA[29] (Supplementary Figure 14), WGS gastric cancers from Wang et al[30] (Supplementary Figure 15), WXS lung cancers from the TRACERx trial[31] (Supplementary Figure 16), and WXS metastasis samples (multiple sites) from the MET500 cohort[32] (Supplementary Figure 17). Based on our previous analysis of minimum data quality needed (see Supplementary Figure 11), we selected samples with purity >40% and number of subclonal mutations ≥25 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 (Figure 4a). Interestingly the MET500 (51%, 58/113) data had a higher proportion of tumours with selected subclones. The measured selective advantage of these subclones was large (>20%) and emerged during the first few tumour doublings across all cohorts (Figures 4b,c). We note that in the metastases case, time is measured relative to the founding of the metastatic lesion, and differential selection of the subclone is 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, where 97% of cases (36 out of the 37 cases suitable for our analysis) were characterised by non-neutral dynamics (Supplementary Figure 16 and 18).

Forecasting cancer evolution

Measuring the evolutionary dynamics of individual human tumours facilitates prediction on the future evolutionary trajectory of these malignancies[33]. Specifically, we can predict how the clonal architecture of a tumour 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 tumour when making treatment decisions. We note we can only predict the future subclonal structure of a tumour assuming that environmental conditions stay the same – e.g. that subclone selective advantages are constant and intervention such as treatment is likely to invalidate this assumption.

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

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

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

In the case of the examined AML sample (Figure 3a), the measured fitness advantages predict the future clonal structure of the malignancy (in the absence of treatment). Specifically, the larger of the two subclones present at the point when the tumour was sampled is predicted to take over the tumour, while the smaller clone is projected to become too rare to remain detectable (Figure 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 tumour subclones. We confirmed that subclonal selection causes an overrepresentation of mutations within the expanding clone, manifested as an additional 'peak' in the VAF distribution, as suggested by many recent studies[18,26,34]. However, irrespective of subclonal selection, the tumour will still show an abundance of low frequency variants (a

1/f-like tail) as the natural consequence tumour growth, wherein the number of new mutations is proportional to the population size.

Our quantitative measurement of the selective advantage (relative fitness) of an expanding subclone revealed that detectable subclones had experienced remarkably large fitness increases, in excess of 20% greater than the background tumour population. Large increases in subclone fitness were also observed in metastatic lesions, indicating that there can still be on-going adaption even in late-stage disease, perhaps as a consequence of treatment. Because selection is inferred using only SNVs that shift in frequency due 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 reports from 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 tumour initiation (at constant population size) showed that *KRAS* and *APC* mutant stem cells have a ~2-4 fold increased fixation probability in single crypts[37] and *TP53* mutant cells in mouse epidermis exhibited a 10% bias toward self-renewal[38]. Moreover, our inferred fitness advantages compare to large fitness advantages measured in bacteria[39]. Nevertheless, we acknowledge that experimental systems may differ significantly from *in vivo* human tumour growth and that new experimental systems are necessary to test these measurements. We also note that we are only able to measure large changes in fitness, and additional efforts will be needed to measure the complete distribution of fitness effects (DFE) 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 in normal lymphocytes of 5.5×10^{-8}-24.6×10^{-8} and a 10-100 fold increase in mutation rate in cancer cell lines such as B-cell lymphoma (5.2×10^{-7}-13.1×10^{-7}) and ALL (66.6×10^{-7}). A recent analysis of a mouse tumour model indicates somatic mutation rates in neoplastic cells are 11x higher than in normal tissue.

Our analysis highlights that even if cancer subclones experience pervasive weak selection, it is not sufficient to alter the clonal composition of the tumour 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 tumours makes them peculiar evolutionary systems, as tumour growth dilutes the effects of selection[41]. Thus, our analysis does not discount the possibility of a multitude of 'mini-drivers'[42] but shows that these must have a corresponding 'mini' effect on the subclonal composition of a tumour (and that the VAF distribution in mini-driver tumours is well described by a neutral

model). We note however, that the ratio of non-synonymous to synonymous variants (dN/dS), a classical test for selection, identified only a small subset of genes (<20 in a pan-cancer analysis) with extreme dN/dS values indicative of strong selection[21,43].

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 tumour changes over time as demonstrated in our *in silico* prediction (Figure 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', where the goal is to maintain the existence of competing subclones that mutually supress the growth of another[44,45]. Our measurements of relative clone fitness could potentially be used to optimize treatment regimes in order to maintain the coexistence of competing populations.

We acknowledge that features not described in our model, e.g. the spatial structure of the tumour, could affect the estimates of the evolutionary parameters[46]. Indeed, our analysis shows that there can be heterogeneity in the evolutionary process within a tumour (only 1/5 regions of a single lung tumour showed strong evidence of subclonal selection). Spatial models of tumour evolution can help elucidate other important biological parameters such as the degree of mixing within tumour cell populations, a purely spatial phenomenon which cannot be quantified using non-spatial models such as ours. We have recently shown how multiple samples per tumour 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 modelled within our framework.

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

Contributions

MW wrote all simulation code and performed mathematical and bioinformatics analysis. BW performed mathematical analysis. TH performed bioinformatics analysis. MW, BW, TH, CC, CB, AS and TG analysed the data. MW, BW, CB, AS and TG wrote the paper. CB, AS and TG jointly conceived, designed, supervised and funded the study.

Acknowledgements

We thank Weini Huang and Kate Chkhaidze for fruitful discussions. We are grateful to Arul Chinnaiyan and Marcin Cieslik for providing us with data from the MET500 cohort, and to Suet Leung from 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 (A22909). T.A.G. is supported by Cancer Research UK (A19771). C.P.B. is supported by the Wellcome Trust (097319/Z/11/Z). B.W. is supported by the Geoffrey W. Lewis Post-Doctoral Training fellowship. A.S. and T.A.G. are jointly supported by the Wellcome Trust (202778/B/16/Z and 202778/Z/16/Z respectively). C.C is supported by NIH R01CA182514. M.J.W is supported by a Medical Research Council student scholarship. This work was also supported by Wellcome Trust funding to the Centre for Evolution and Cancer (105104/Z/14/Z).

Figure Legends

Figure 1. Modelling patterns of subclonal selection in sequencing data. **(a)** In a stochastic branching process model of tumour growth cells have birth rate *b* and death rate *d*, mutations accumulate with rate *μ*. Cells with fitness advantage (orange) grow at a faster net rate (*b*-*d*) than the host population (blue). **(b)** The variant allele frequency (VAF) distribution contains clonal (truncal) mutations around f=0.5 (in this example of diploid tumour), and subclonal mutations (f<0.5) which encode how a tumour has grown. In the absence of subclonal selection, a neutral $1/f^2$ tail describes the accumulation of passenger mutations as the tumour expands. **(c)** A selected subclone produces an additional peak in the distribution while a $1/f^2$ tail is still present due to passenger mutations accumulating in both the original population and the new subclone. **(d)** In the presence of subclonal selection, the magnitude and average frequency of the subclonal cluster of mutations (red) encode the age and size of a subclone respectively, which in turn allows measuring the clone's selective advantage. **(e)** Frequentist power analysis of 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 tumour, resulting in the rejection of the neutral (null) model. Tumours 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.

Figure 2. Accurate recovery of evolutionary parameters from simulated data using Approximate Bayesian Computation. Our method recovered the correct clonal structure in simulated tumour data for representative examples of **(a)** a neutral case, **(b)** a 1 subclone case and **(c)** a two subclones case. Grey 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), shaded areas are 95% intervals. The inferred posterior distributions of the evolutionary parameters contained the true values (dashed lines) for **(d,f)** the time of emergence of the subclones and **(e,g)** the selection coefficient 1+s. **(h)** The mean percentage error in inferred parameter values across a virtual tumour cohort (n=100 tumours) was below 10%. 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.

Figure 3. Quantifying selection from high-depth bulk sequencing of human cancers. Both **(a)** an acute myeloid leukemia (AML) sample and **(b)** a breast cancer sample sequenced at whole-genome resolution showed evidence of two selected subclones. **(c)** In the case of a multi-region whole-exome sequenced case of lung cancer, one sample showed evidence of a single subclone whereas four other samples **(d-g)** from the same patient were consistent with the neutral model**.** Grey 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), shaded areas are the 95% intervals. **(h)** Bayesian model selection reports the expected clonal structure for each case (Bayes Factors reported above histograms). **(i)** Inferred subclone fitness advantages were 20% and 80% faster than the

original population. **(j)** Inferred times of subclone emergence indicated subclones arose within the first 15 tumour population doublings. **(k)** Inferred mutation rates were of the order of 10^{-7} mutations per base per tumour doubling in solid tumours but ~10^{-9} in AML, reflecting the respective differences in mutational burden between cancer types. All posterior distributions were generated from 500 samples.

Figure 4. Quantifying selection in large cohorts of primary tumours and metastatic lesions. (a) 21% of colon cancers (N=70) from TCGA (sequenced to sufficient depth and with high enough cellularity for statistical inference), 29% of WGS 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. When present, differentially selected subclones were found to have **(b)** large fitness advantages with respect to the host population and **(c)** emerge early during growth. Bayes Factors for subclonal structures for all data are reported in Supplementary Table 5. 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.

Figure 5. Predicting the future evolution of subclones. (a) VAF distribution of an *in silico* tumour sampled at 10^5 cells was used to measure the fitness and time of emergence of a subclone. Grey 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), shaded areas are the 95% intervals. Inset shows error from ground truth. 500 posterior samples were taken to perform the inference. **(b)** These values were then used to predict the spread of the subclone as the tumour grew to 10^7 cells, showing the predictions matched the ground truth. Predictions were made by extrapolating the posterior distribution of 1+s using equations in the main text. Solid line shows the median value from the posterior distribution, shaded area shows the 95% 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 tumour doublings while subclone 1 will become too small to be detected.

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. Tumour Cell Heterogeneity. *F1000Res* **5,** 238–14 (2016).
3. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512,** 155–160 (2014).
4. Burrell, R. A. & Swanton, C. Re-Evaluating Clonal Dominance in Cancer Evolution. *Trends in Cancer* (2016). doi:10.1016/j.trecan.2016.04.002
5. Durrett, R. *Branching Process Models of Cancer*. (Springer, 2015).
6. Marjoram, P. & Tavaré, S. Modern computational approaches for analysing 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. *The American Journal of Pathology* **154,** 1815–1824 (1999).
10. Tsao, J. L. *et al.* Genetic reconstruction of individual colorectal tumor histories. *PNAS* **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. *Nature Genetics* **48,** 238–244 (2016).
12. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nature Genetics* **49,** 1015–1024 (2017).
13. Hartl, D. L. & Clark, A. G. *Principles of population genetics*. (Sinauer, 1997).
14. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 18545–18550 (2010).
15. Cheek, D. & Antal, T. Mutation frequencies in a birth-death branching process. *arXiv*
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 (2014).
17. Durrett, R. POPULATION GENETICS OF NEUTRAL MUTATIONS IN EXPONENTIALLY GROWING CANCER CELL POPULATIONS. *The Annals of Applied Probability* **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* (2015). doi:10.1038/nature14279
20. Gillespie, J. H. Genetic Drift in an Infinite Population: The Pseudohitchhiking 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. *Clinical Cancer Research* **13,** 7357–7362 (2007).
26. Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems* **1,** 210–223 (2015).
27. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion 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. 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. *Nature Publishing Group* **46,** 573–582 (2014).
31. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *N Engl J Med* NEJMoa1616288–13 (2017). doi:10.1056/NEJMoa1616288
32. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature Publishing Group* **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* 1–11 (2017). doi:10.1038/nature22312
36. Rutledge, S. D. *et al.* Selective advantage of trisomic human cells cultured in non- standard conditions. *Sci. Rep.* 1–12 (2016). doi:10.1038/srep22828
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 preneoplasia. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 270–275 (2010).
39. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *PNAS* **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. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth.

Nature Genetics **47,** 209–216 (2015).
42. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nature Reviews Cancer* 1–6 (2015). doi:10.1038/nrc3999
43. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* 1–35 (2017). doi:10.1016/j.cell.2017.09.042
44. Enriquez-Navas, P. M. *et al.* Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. *Science Translational Medicine* **8,** 327ra24–327ra24 (2016).
45. Zhang, J., Cunningham, J. J., Brown, J. S. & Gatenby, R. A. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat Commun* 1–9 (2017). doi:10.1038/s41467-017-01968-5
46. 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).

Methods

Simulating tumour growth

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

Tumour growth is assumed to begin with a single transformed cancer cell that has 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 thus appear as clonal (present in all cells and thus will generate the cluster of clonal mutations at frequency ½ for a diploid tumour) within the cancer population.

To simulate tumour, and subclone evolution, we specify a birth rate *b* and death rate *d* (*b*>*d*, for a growing population), meaning that the average population size at time t is:

$$N(t) = e^{(b-d)t} \quad [6]$$

We set *b=log(2)* for all simulations, such that in the absence of cell death the population will double in size at every unit of time. The tumour grows until it has reached a specified size N_{end}, where the simulation stops. At each division, cells acquire υ new mutations, where υ is drawn from a Poisson distribution with mean μ, the mutation rate per cell division. We assume new mutations are unique (infinite sites approximation). Not all divisions result in new surviving lineages because of cell death and differentiation. The probability of a cell division producing a surviving lineage β expressed can be expressed in terms of the birth and death rates:

$$\beta = \frac{b-d}{b}. \quad [7]$$

Simulating subclonal selection

To include the effects of subclonal selection, a mutant is introduced into the population that has a higher net growth rate (birth minus death) than the host population. We only consider the cases of one or two subclonal populations under selection at any given time. We deem 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[44]). Additionally, we found that sequencing depth >100X is required to resolve more than 1 subclone (Supplementary Figure 9). Fitter mutants can have a higher birth rate, a lower death rate, or a combination of the two, all of which results in the mutant growing at a faster

rate than the host population. Given that the host/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 define the selective advantage s of the fitter population as:

$$1 + s = \frac{b_F - d_F}{b_H - d_H} \quad [8]$$

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[45]. Due to the small number of possible reactions (we consider at most 3 populations with different birth and death rates) this algorithm is 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. Simulation initialized with 1 cell and set all simulation parameters.
2. Choose a random cell, i from the population.
3. Draw a random number r~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} there is a probability that cell i will neither divide nor die. If $\beta = 1$, ie no cell death then in the above d_{max} = 0.
5. If cell divides, daughter cells acquire ν new mutations where ν ~Poisson(μ).
6. Time is increased by a small increment $\frac{1}{N(b_{max}+d_{max})}\tau$, where τ is an exponentially distributed random variable[47].
7. Go to step 2 and repeat until 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 computationally efficiency when used in our Bayesian inference method, a procedure that requires potentially many millions of individual simulations to be run in order to get accurate inferences. Our ultimate goal was to measure the time subclones emerge and their fitness. These parameters are measured in terms of tumour volume doublings, not in terms of cell division durations (as this is unknown in human tumours). Our approximations allow us to quantify

relative fitness of subclones, measured in units of population doubling, from the VAF distribution. The approximations are:

Approximation 1: We model differential subclone fitness by varying the birth rate only, and setting the deth rate to 0 (e.g. $\beta = 1$, all lineages survive). This increases simulation speed because a smaller number of time steps are required to reach the same population size and ensures that tumours never die out in our simulations.

Timing the emergence of subclones depends on the number of mutations that have 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 tumour doublings × the effective mutation rate ($n_{doublings} \times \frac{\mu}{\beta}$). Given we measure everything in terms of tumour doublings and the effective mutation rate (μ/β) is the only measure available to us from the VAF distribution (from the low frequency 1/f tail), we reduce our search space by fixing $\beta = 1$ and varying μ, recognizing that in reality the effective mutation rate is 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 $\frac{\mu}{\beta}$ 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 $\frac{\mu}{\beta}$. Even in cases where the death rate was dramatically different in the subclone compared to the host population ($\beta = 1.0$ vs $\beta = 0.5$) the mean error on the estimates of the mutation rate was 42% (Supplementary Figure 5), significantly less than the order of magnitude previously measured between cancer type[11] and so 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 tumours with high death rates.

Approximation 2: We simulate a smaller tumour population size compared to typical tumour sizes at diagnosis, and scale the inferred values *a posteriori*. We note that the VAF distribution holds no information on the population size (it measures only relative proportions) and furthermore simulating realistic population sizes (in the order of tens or hundreds of billions of cells in human malignancies) is computationally unfeasible. To circumvent this, we generate synthetic datasets that capture the characteristics relevant to measuring the fitness and time subclones emerge, namely the effective mutation rate ($\frac{\mu}{\beta}$) encoded by the low frequency part of the distribution, the number of mutations in any subclonal cluster and their frequency. Theoretical population genetics is then used to transform these measurements into values of fitness

and time (via Equations [2] and [4]), and values are scaled by the realistic population size $N_{end} = 10^{10}$.

Simulation length was required to allow the single cell that gives rise to the subclone sufficient time to accumulate the number of mutations 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 where we used the more conservative $N_{end}=10^4$. In general, $N_{end}=10^4$ is sufficient to be able to measure the range of parameters considered in Figure 1e.

To appropriately scale the estimates of *s* requires an estimate of the age of the tumour in terms of tumour doublings. Using Equation [4] with a final population size of N_{end}, we can calculate t_{end} as:

$$t_{end} = \frac{\log\left((1-f_{sub})\times N_{end}\right)}{\log\left(2\right)}, \qquad [10]$$

where f_{sub} is the frequency of the subclone. We assumed a realistic $N_{end} = 10^{10}$, for generating the posterior distributions in Figures 3 & 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) the exact choice has minimal effect on our inferred values of *s* (Supplementary Figure 6).

To confirm that these assumptions do not invalidate our approach, we generated synthetic datasets with cell death and 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 tumours. This demonstrated that we were able to accurately recover the input parameters when the simplifications were applied (Figure 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 suffers from multiple sources of noise, most importantly for this study is that mutation counts (VAFs) are sampled from the true underlying frequencies in the tumour population (both because of the initial limited physical sampling of cells from the tumour for DNA extraction, and then due to 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 not measured at all. Consequently only mutations above a frequency of around 5-10% with 100X sequencing are observable with certainty[48]. 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 VAF_{true}, the sequence depth D_i is Binomially distributed:

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

for a tumour of size N. The sampled read count with the mutant is Binomially distributed with the following parameters:

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

or if over-dispersed sequencing is modelled[49,50] we use the Beta-Binomial model, which introduces additional variance to the sampling:

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

where ρ is the overdispersion parameter, and $\rho = 0$ reverts to the Binomial model. Finally, the sequenced VAF for mutation i is given by:

$$VAF_i = \frac{f_i}{D_i}$$

Modelling stem cells

Stem cell architecture was modelled 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 cells accumulated mutations as described above. We used $\alpha = 0.1$ and n=5. If $\alpha = 1.0$ then the model is 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 using simulated sequencing data where the true underlying evolutionary dynamics was known. The simulation approach to generate synthetic data was taken instead of a purely statistical approach, as the simulation naturally accounts 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 account 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 $\pi(\theta|D) = \frac{p(D|\theta)\pi(\theta)}{p(D)}$, where $\pi(\theta)$ is the prior distribution on θ and $p(D|\theta)$ is the likelihood of the data given θ. In cases where calculating the likelihood is intractable, as was the case here where our model cannot 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 using a distance measure. Through multiple

comparisons of different input parameter values, we can produce a posterior distribution of parameter values that minimise the distance measure, and in so doing accurately approximate the true posterior. The simplest approach is called the ABC rejection method and the algorithm is as follows[51]:

1) Sample candidate parameters θ* from prior distribution π(θ)
2) Simulate tumour growth with parameters θ*
3) Evaluate distance, δ between simulated data and target data
4) If $\delta < \varepsilon$ reject parameters θ*
5) If $\delta \geq \varepsilon$ accept parameters θ*
6) Return to 1

We used an extension of the simple ABC rejection algorithm, called Approximate Bayesian Computation Sequential Monte-Carlo (ABC SMC)[22,52]. 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, using an approach known as sequential importance sampling[53]. The ABC SMC algorithm also allows for Bayesian model selection to be performed 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 use the full data distribution, thus avoiding issues of inconsistent Bayes factors due to loss of information[54,55]. For further details on the algorithm see references[22] and the Supplementary Note on the specific details of our implementation. Bayes factors 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 Figure 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 for Selection in the Frequentist paradigm

We also refined a simple analytical test in order to rapidly determine what evolutionary parameters of selection lead to an observable deviation of the VAF distribution from that expected under neutrality. Previously, we showed that under neutrality, the distribution of mutations with a frequency greater than f is given by[11]:

$$M(f) = \frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right) \quad [11]$$

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 is to use the shape of the curve described by Equation [5] and test whether our empirical data collapses onto this curve. To implement

this approach, here we defined the *universal neutrality curve*, $\bar{M}(f)$. Given an appropriate normalization of the data, the mutant allele frequency distribution governed by neutral growth will 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 $f=f_{min}$.

$$\max\left(M(f)\right) = \frac{\mu}{\beta}\left(\frac{1}{f_{min}} - \frac{1}{f_{max}}\right) \quad [12]$$

$$\bar{M}(f) = \frac{\frac{\mu}{\beta}\left(\frac{1}{f} - \frac{1}{f_{max}}\right)}{\max\left(M(f)\right)} \quad [13]$$

$$\bar{M}(f) = \frac{\left(\frac{1}{f} - \frac{1}{f_{max}}\right)}{\left(\frac{1}{f_{min}} - \frac{1}{f_{max}}\right)} \quad [14]$$

$\bar{M}(f)$ is independent of the mutation rate and the death rate and 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 $\bar{M}(f)$ and the empirical data and the area between $\bar{M}(f)$ 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 4 classifiers we ran 10^5 neutral and non-neutral simulations and compared the distribution of the test statistics for these two cases. Due to the stochastic nature of the model, not all 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 where the fitter subpopulation was within a certain range (20% and 70% fraction of the final tumour size). All 4 test statistics showed significantly different distributions between neutral and non-neutral cases (Supplementary Figure 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 Figure 3). This analysis showed that R^2 had the least discriminatory power, with the other 3 performing approximately equally well (see Supplementary Table 4 for AUC). 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.

Code Availability Statement

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

Bioinformatics analysis

Variant calls from the original studies were used for the AML data[26], TRACERx[31] data and 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>0.4 to perform the analysis. For the breast cancer data[18] and lung cancer data[27], bam files from the original study were obtained and variants were called using Mutect2[56] and filtered to require at least 5 reads reporting the variants in the tumour and 0 reads in the normal. To mitigate the effects of low frequency mutations arising from paralogous regions of the genome we filtered any mutations where 75bp regions either side of the mutations had multiple BLAST hits (minimum of 100bp hit length, maximum of 3% mismatching bases).

Copy number aberrations could also potentially result in the multi-peaked distribution we observe, hence we only used mutations that were found in regions identified as diploid (and without copy-neutral LOH). 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 falling in non-diploid regions. For the TRACERx data and MET500 data we used allele specific copy number calls provided in the original studies to filter the data. For all other datasets we applied the Sequenza algorithm to infer allele specific copy number states and estimate the cellularity[57]. As the original breast cancer study found evidence of subclonal copy number alterations in multiple chromosomes we only used mutations on chromosome 3 for our analysis, (Supplementary Figure 20). BAFs of regions called as copy neutral by Sequenza in the lung cancer sample were consistent with a diploid genome (Supplementary Figure 21).

We used cellularity estimated 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 $\bar{d}_i = \kappa \times d_i$. 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 can account for the over-dispersion of allele read counts. To measure the over-dispersion parameter ρ, we fitted a Beta-Binomial model to the clonal cluster where we know $VAF_{true} = 0.5$. We used Markov Chain Monte Carlo (MCMC) to fit the following model to the right hand side of the clonal cluster so as to minimize the effects of the 1/f distribution or subclonal clusters:

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

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 model, and supplementary table S7 reports the over-dispersion parameter for each dataset. We also used this analysis to further refine the cellularity estimate provided by sequenza, ensuring that the clonal cluster was centred at VAF = 0.5. We note that some of the over-dispersion is likely artificial and introduced by the cellularity correction.

Mutational signatures in the breast cancer sample and AML sample (Supplementary Figure 12) were identified 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 report the mean value and the 95% CI.

Data Availability Statement

Only publically available data was used in this study, and data sources and handling of these data are described above.

References

47. Waclaw, B. *et al.* A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* (2015). doi:10.1038/nature14971
48. 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. *Human Mutation* **34,** 1432–1438 (2013).
49. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11,** 396–398 (2014).
50. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun* **3,** 811 (2012).
51. 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).
52. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26,** 104–110 (2010).
53. Del Moral, P., Doucet, A. & Jasra, A. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **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. U.S.A.* **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 Biology* 1–11 (2016). doi:10.1186/s13059-016-0893-4