

1 Quantification of subclonal selection in cancer from bulk 2 sequencing data

3
4 Marc J. Williams^{1,2,3}, Benjamin Werner⁴, Timon Heide⁴, Christina Curtis^{5,6},
5 Chris P Barnes^{2,7,*}, Andrea Sottoriva^{4,*}, Trevor A Graham^{1,*}

6
7 1 Evolution and Cancer Laboratory, Barts Cancer Institute, Queen Mary University of London, London, UK.

8 2 Department of Cell and Developmental Biology, University College London, London, UK.

9 3 Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University
10 College London, London, UK.

11 4 Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK.

12 5 Departments of Medicine and Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

13 6 Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305, USA

14 7 UCL Genetics Institute, University College London, London, UK

15
16 * Correspondence should be addressed to:

17 christopher.barnes@ucl.ac.uk

18 andrea.sottoriva@icr.ac.uk; +44 208 722 4072

19 t.graham@qmul.ac.uk; +44 207 882 6231

20 21 22 23 **Abstract**

24 Subclonal architectures are prevalent across cancer types. However, the
25 temporal evolutionary dynamics that produce tumour subclones remain
26 unknown. Here we measure clone dynamics in human cancers using
27 computational modelling of subclonal selection and theoretical population
28 genetics applied to high throughput sequencing data. Our method determines
29 the detectable subclonal architecture of tumour samples, and simultaneously
30 measures the selective advantage and time of appearance of each subclone.
31 We demonstrate the accuracy of our approach and the extent to which
32 evolutionary dynamics are recorded in the genome. Application of our method
33 to high-depth sequencing data from breast, gastric, blood, colon and lung
34 cancers, as well as metastatic deposits, showed that detectable subclones
35 under selection, when present, consistently emerged early during tumour
36 growth and had a large fitness advantage (>20%). Our quantitative framework
37 provides new insight into the evolutionary trajectories of human cancers,
38 facilitating predictive measurements in individual tumours from widely
39 available sequencing data.
40
41

42 **Introduction**

43 Carcinogenesis is the result of Darwinian selection for malignant phenotypes,
44 driven by genetic and epigenetic alterations that allow cells to evade normal
45 homeostatic regulation and prosper in changing microenvironments¹. High
46 throughput genomics has shown that tumours across all cancer types are
47 highly heterogeneous^{2,3} with complex clonal architectures⁴. However,
48 because longitudinal observation of solid tumour growth unperturbed by
49 treatment remains impractical, the temporal evolutionary dynamics that
50 produce subclones remain undetermined, and consequently, there is no
51 mechanistic basis that can be utilised to predict future tumour evolution and
52 modes of relapse. More specifically, the magnitude of the fitness advantage
53 experienced by a new cancer subclone has remained unknown.

54
55 The subclonal architecture of a cancer – as measured by the pattern of intra-
56 tumour genetic heterogeneity (ITH) – is a direct consequence of the
57 unobservable evolutionary dynamics of tumour growth. Therefore, given a
58 realistically constrained model of subclonal expansion, the pattern of ITH in a
59 tumour can be used to infer its most probable evolutionary trajectory. ITH
60 represented within the distribution of variant allele frequencies (VAF), as
61 measured by high coverage sequencing, is particularly amenable to such an
62 approach.

63
64 In this study, we build upon theoretical population genetics models of asexual
65 evolution⁵ and Bayesian statistical inference on genetic data⁶ to measure
66 cancer evolution in human tumours. This type of approach is established in
67 the field of molecular evolution, where evolutionary processes are also difficult
68 to measure directly^{7,8}, and examples of applications of these approaches to
69 human cancers date back to the previous century^{9,10}.

70
71 Recently, we have shown that under a neutral “null” evolutionary model (i.e.
72 when all selected driver alterations are truncal and present in all cancer cells),
73 the VAF follows a characteristic power law distribution¹¹. Subsequent
74 simulations that modelled space and subclonal selection demonstrated that
75 genetic divergence in multi-region sequencing data could be used to
76 categorize tumours based on the mode of their evolution¹² (effectively-neutral
77 or non-neutral), but the specific evolutionary dynamics that produce subclonal
78 architectures, such as the fitness advantage of subclones, remained
79 unmeasured. Here, using a combination of a stochastic branching process
80 model of subclonal selection in cancer, an explicit sequencing error model,
81 and Bayesian model selection and parameter inference, we identify the
82 characteristic patterns of subclonal selection in the cancer genome and
83 measure fundamental evolutionary parameters in non-neutrally evolving
84 human tumours.

85

86

87 Results

88

89 Theoretical framework of subclonal selection

90 We developed a stochastic computational model of tumour growth applicable
91 to cancer genomic data that accounts for subclonal selection (see Methods).

92 The model is based on a classical stochastic branching process approach
93 from population genetics¹³ that has been often used to model malignant

94 populations^{5,14} and is here extended to be applicable to cancer sequencing
95 data. Cells divide and die according to defined birth and death rates and

96 daughter cells acquire new mutations at rate μ mutations per cell per division
97 (Figure 1a). The fitness advantage of a mutant subclone is defined by the

98 ratio of net growth rates between the fitter mutant (λ_m) and the background
99 host population (λ_b)

100

$$101 \quad 1 + s = \frac{\lambda_m}{\lambda_b}. \quad [1]$$

102

103 This definition¹³ provides an intuitive interpretation for the fitness coefficient s :

104 for example, $s=1$ implies that the mutant cell population grows twice as fast as
105 the host tumour population, and $s=0$ implies $\lambda_m=\lambda_b$ such that the subclone

106 evolves neutrally with respect to the background population. Within the model,
107 neutral evolution ($s=0$) leads to a VAF distribution characterised by a power-

108 law distributed subclonal tail of mutations^{11,15-17} (Figure 1b), where the
109 cumulative number of mutations at a frequency f is proportional to the inverse

110 of that frequency, $1/f$ (in the non-cumulative VAF distribution such as Figure
111 1b, this shows as $\sim 1/f^2$). Alternatively, clonal selection ($s>0$) produces

112 characteristic 'subclonal clusters' within the VAF distribution that have been
113 observed in cancer genomes¹⁸ (Figure 1c). Importantly, as neutral mutations

114 continue to accumulate within each subclone, the $1/f$ tail is also present in
115 tumours with selected subclones (Figure 1c).

116

117 A mathematical analysis of the model indicates how subclonal clusters
118 encode the underlying evolutionary dynamics of a subclone: the mean VAF of

119 the cluster is a measure of the relative size of the subclone within the tumour,
120 and the total number of mutations in the cluster (i.e. the area of the cluster)

121 indicates the subclone's relative age (as later-arising subclones will have
122 accumulated more mutations). Together, these two measures allow the

123 fitness advantage s to be estimated¹⁹. We provide a summary derivation
124 below and refer to the Supplementary Note for full details.

125

126 We define $t_0=0$ to be the time when the first transformed cancer cell begins to
127 grow. At a later time t_1 , a cell in the tumour acquires a subclonal 'driver'

128 somatic alteration that confers a fitness advantage, giving rise to a new
129 phenotypically distinct subclone that expands faster than the other tumour

130 cells. We note that to measure selection dynamics it is not important what the
131 actual driver event is: genetic (point mutation or copy number alteration),

132 epigenetic, or even microenvironmental drivers will all cause somatic
133 mutations in the selected lineage to 'hitchhike'²⁰ to higher frequencies than

134 expected under the neutral null model. The number of hitchhiking mutations,

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

$$137$$

$$138 \quad M_{sub} = \mu\Gamma. \quad [2]$$

$$139$$

140 The relationship between the mean number of divisions of a lineage, Γ and
 141 time measured in population doublings is $\Gamma = 2\log(2)t_1$ (see Supplementary
 142 Note). The mutation rate per population doubling can be estimated from the
 143 $1/f$ -like tail¹¹. For a subclone that emerges at time t_1 , we would expect to
 144 observe M_{sub} mutations at some frequency $f_{sub}/2$ (for a subclone at a cancer
 145 cell fraction f_{sub} in a diploid genome, and assuming a sample with 100%
 146 tumour purity), and given the limited accuracy of VAF measurement inherent
 147 to next generation sequencing this will appear as a cluster of mutations with a
 148 mean $f_{sub}/2$ in the VAF distribution. Therefore, Equation [2] provides an
 149 estimate of t_1 , the time when the subclone appeared.

150
 151 Assuming exponential growth and well mixed populations, and considering
 152 that the subclone grows $1+s$ times faster than the background tumour
 153 population as defined by Equation [1], the frequency of the subclone will grow
 154 in time according to:

$$155$$

$$156 \quad f_{sub}(t_{end}) = \frac{e^{\lambda_b(1+s)(t_{end}-t_1)}}{e^{\lambda_b t_{end}} + e^{\lambda_b(1+s)(t_{end}-t_1)}}. \quad [3]$$

$$157$$

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

$$160$$

$$161 \quad s = \frac{\lambda_b t_1 + \ln\left(\frac{f_{sub}}{1-f_{sub}}\right)}{\lambda_b(t_{end}-t_1)}. \quad [4]$$

$$162$$

163 Given an estimate of the age of the tumour expressed in population doublings
 164 t_{end} , equations [2] and [4] provide a means to measure the selective
 165 advantage of a subclone directly from the VAF distribution (Figure 1d). t_{end}
 166 can be derived from the final tumour size N_{end} by the relation $2^{t_{end}} =$
 167 $(1 - f_{sub}) \times N_{end}$. In the case of multiple subclones, Equation [4] takes a
 168 slightly modified form (Supplementary Note). We note that Equations [1-4] are
 169 known results in population genetics and have been previously used to
 170 describe the dynamics of asexual haploid populations¹³.

171
 172 Our previously presented frequentist approach to detect subclonal selection
 173 from bulk sequencing data involves an R^2 test statistic¹⁹ to reject the
 174 hypothesis of neutral evolution ($s=0$), the null model in molecular evolution²¹.
 175 Here we extended our previous work to examine different test statistics for
 176 assessing deviations from the null neutral model (see Supplementary Figures
 177 1-3 & Methods). However, the frequentist approach has limitations: it requires
 178 to choose the interval of the VAF distribution to test, and importantly only
 179 allows for the rejection of the null hypothesis (which is not necessarily
 180 evidence for the null itself).

181

182 To address these shortcomings, we implemented a Bayesian statistical
183 inference framework (Supplementary Figure 4 & Methods) that fits our
184 computational model incorporating both selection and neutrality to sequencing
185 data, and simultaneously estimates the subclone fitness, time of occurrence,
186 and the mutation rate. This method allowed us to perform Bayesian model
187 selection²² for the number of subclones within the tumour and specifically
188 calculate probabilities that a tumour contained 0 subclones ($s=0$, neutral
189 evolution), 1 or more subclones (non-neutral evolution). The advantage of the
190 Bayesian approach is that we can directly ask which model (neutral or non-
191 neutral) is best supported by the data, using the whole VAF distribution.

192
193 Our framework models mutation, selection and neutral drift using a classical
194 stochastic branching process¹³, while integrating several confounding factors
195 and sources of noise in bulk sequencing data, principally allele sampling and
196 depth of sequencing (see Methods and Supplementary Note). This approach
197 allows sample-based schemes designed such that the data-generating
198 process can be mimicked to account for complex experimental biases.
199 Despite these confounding factors, we found that the $1/f$ tail accurately
200 measures the mutation rate even in the presence of subclonal clusters
201 (Supplementary Figure 5), and our inferred value of $1+s$ is largely insensitive
202 to the final tumour size (N_{end}) when this value is realistically large ($N_{\text{end}} > 10^9$)
203 (Supplementary Figure 6 and Supplementary Note).

204
205 We note that the theoretical framework is based upon the assumption of
206 exponential growth, which is a growth pattern well supported by empirical data
207 in many cancer types²³⁻²⁵. The impact of alternate models of growth, such as
208 logistic and Gompertzian growth, is explored in the Supplementary Note. We
209 also implemented a cancer stem cell model where only a subset of cells has
210 unlimited proliferation potential and found that for the purposes of this study
211 this has little impact on the expected VAF distribution, which in this scenario
212 only measure events that occur in the stem cell compartment (Supplementary
213 Figure 7).

214 215 **Recovery of evolutionary dynamics in synthetic tumours**

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

226
227 We then applied our Bayesian framework to estimate evolutionary parameters
228 from synthetic data (VAF distributions derived from computational simulations
229 of tumour growth with known parameters). Our framework identified the
230 correct underlying model with high probability for representative examples of a
231 neutrally growing tumour (Figure 2a), a tumour with a single subclone (Figure

232 2b) and a tumour with 2 subclones (Figure 2c), and also recovers the
233 evolutionary parameters in each case (Figures 2d-g). Given that we modelled
234 tumour growth as a stochastic process, variability in our estimates was
235 expected (see Supplementary Note). In a cohort of 100 synthetic tumours (20
236 examples selected in Supplementary Figure 8), where the ground truth was
237 known, the mean percentage error on parameter inference was below 10%
238 (Figure 2h). The stochasticity also explains the width of the posterior
239 distributions (Figures 2d-g). In particular, the rate of stochastic cell death has
240 a large effect on the variability of lineage age and consequently can cause a
241 slight over-estimation of the mutation rate and variability in the time taken for
242 a lineage to clonally expand increases with increased cell death (see
243 Supplementary Note).

244
245 Monte Carlo analysis indicated that accurate measurement of subclonal
246 evolutionary dynamics required high depth (>100X) for both whole-exome and
247 whole-genome sequencing (Supplementary Figure 9). This analysis
248 demonstrates how the clonal structure becomes progressively obscured as
249 the sequencing depth decreases. Depths of sequencing of less than 100X
250 preclude a robust quantification of subclonal dynamics, and moreover the
251 neutral model is preferred by our Bayesian model selection framework, even
252 when it is false (Supplementary Figure 9). Importantly, this analysis showed
253 that even in some cases when selection is present (particularly weak
254 selection), neutral evolution is the most parsimonious description of the data.
255 In other words, the observed dynamics are then 'effectively neutral'. In
256 addition, we note that while the increased mutational information provided by
257 WGS and higher sequencing depths makes quantification of subclonal
258 structure more robust, this can also reveal (neutrally) drifting populations that
259 may be falsely ascribed as a selected clone (Supplementary Figure 10). We
260 also investigated the robustness of the inference method to tumour purity and
261 cancer cell fraction of the subclone finding that at 100X sequencing depth a
262 minimum purity of 50% is needed to confidently identify subclones with cancer
263 cell fraction >30% (15% VAF in a diploid genome), see Supplementary Figure
264 11.

265 **Detectable subclones have a large selective advantage**

266 We first used our approach to quantify evolutionary dynamics in primary
267 human cancers where high depth (>150X) and validated sequencing data
268 were available. We considered whole-genome sequencing (WGS) of a single
269 AML sample²⁶, WGS of a single breast cancer sample¹⁸ and multi-region
270 high-depth whole exome sequencing (WXS) of a lung adenocarcinoma²⁷. To
271 avoid the confounding effects of copy number changes, we exploited the
272 hitchhiking principle and restricted our analysis to consider only somatic single
273 nucleotide variants (SNVs) that were located within diploid regions (see
274 Methods). After correction for cellularity the 'clonal cluster' at VAF=0.5, and a
275 potentially complex distribution of mutations with VAF<0.5 representing the
276 subclonal architecture were clearly observable.
277

278
279 The AML and breast cancer cases both showed evidence of 2 subclonal
280 populations, corroborating the initial studies but instead finding the lowest
281 frequency cluster to be a consequence of all within-clone neutral

282 mutations^{18,26} (Figure 3a,b,h). Measurement of the evolutionary dynamics
283 showed that for both cancers the subclones had considerably large fitness
284 advantages (>20%, Figure 3i) and emerged within the first 15 population
285 doublings (Figure 3j). In the AML sample, subclone 1 (highest frequency
286 subclone) had putative driver mutations in *IDH1* and *FLT3* and subclone 2
287 had a distinct *FLT3* mutation and a *FOXP1* mutation. In the breast cancer
288 sample, no putative driver point mutations were found in the subclonal
289 clusters but we note that the original analysis found that subclone 1 (highest
290 frequency subclone) had lost one copy of chromosome 13. Interestingly, the
291 breast cancer sample also exhibited a 100-fold higher mutation rate per
292 tumour doubling compared to the AML sample (Figure 3k). We note that our
293 mutation rate estimate corresponds to the number of mutations per base per
294 population doubling. Due to the high cell death and possibly differentiation in
295 cancers (both leading to lineage extinction), doubling in volume may require
296 several rounds of cell division. To derive the mutation rates per base per
297 division an independent measurement of the probability β of a cell division to
298 give rise to two surviving lineages is required (see Methods, Equation [9] and
299 Supplementary Note). Mutational signature analysis²⁸ of subclonal mutations
300 provided support for the assumption of a constant mutation rate during
301 subclone evolution (Methods and Supplementary Figure 12).

302
303 In the lung adenocarcinoma case, multiple tumour regions (n=5) had been
304 sequenced to high depth. Amongst these regions, only one region (region 12)
305 showed strong evidence of a new subclone (Figures 3c,h, BF = 1.49) with a
306 measured selective advantage of 30% (Figure 3j), while for all other regions a
307 neutral evolutionary model was most probable (Figures 3d-g, BF = 6.36-
308 29.92). Region 12 had unique copy number alterations on chromosome 3 that
309 could plausibly have caused the subclonal expansion (Supplementary Figure
310 13). Together these data show spatial heterogeneity of the evolutionary
311 dynamics within a single tumour.

312
313 We then applied our analysis to 4 additional large cohorts of variable
314 sequencing depth: WXS colon cancers from TCGA²⁹ (Supplementary Figure
315 14), WGS gastric cancers from Wang et al³⁰ (Supplementary Figure 15), WXS
316 lung cancers from the TRACERx trial³¹ (Supplementary Figure 16), and WXS
317 metastasis samples (multiple sites) from the MET500 cohort³²
318 (Supplementary Figure 17). Based on our previous analysis of minimum data
319 quality needed (see Supplementary Figure 11), we selected samples with
320 purity >40% and number of subclonal mutations ≥ 25 for further analysis.
321 Differentially selected subclones were detected in 29% (5/17 cases) of the
322 gastric cancers and 21% (15/70 cases) of the colon cancers (Figure 4a).
323 Interestingly the MET500 (51%, 58/113) data had a higher proportion of
324 tumours with selected subclones. The measured selective advantage of these
325 subclones was large (>20%) and emerged during the first few tumour
326 doublings across all cohorts (Figures 4b,c). We note that in the metastases
327 case, time is measured relative to the founding of the metastatic lesion, and
328 differential selection of the subclone is measured relative to the other cells in
329 the metastasis. Eventual founder effects in the metastasis are, by definition,
330 clonal events in the sample, and so do not appear in the subclonal VAF
331 spectrum. We also observed similarly large fitness advantages of subclones

332 within the TRACERx cohort, where 97% of cases (36 out of the 37 cases
333 suitable for our analysis) were characterised by non-neutral dynamics
334 (Supplementary Figure 16 and 18).

335

336 **Forecasting cancer evolution**

337 Measuring the evolutionary dynamics of individual human tumours facilitates
338 prediction on the future evolutionary trajectory of these malignancies³³.

339 Specifically, we can predict how the clonal architecture of a tumour is
340 expected to change over time (in the absence of new drivers): such
341 predictions could be useful, for instance, to decide how often to sample a
342 tumour when making treatment decisions. We note we can only predict the
343 future subclonal structure of a tumour assuming that environmental conditions
344 stay the same – e.g. that subclone selective advantages are constant and
345 intervention such as treatment is likely to invalidate this assumption.

346

347 Suppose a biopsy is taken and fitness of a subclone measured at some time t ,
348 we can then ask how long it will take for the subclone to become dominant
349 (>90% frequency) in the tumour. From our model, the time for a subclone to
350 shift from a frequency f_1 to a frequency of f_2 given a relative fitness advantage
351 s is:

352

$$353 \quad \Delta T = \frac{\log\left(\frac{f_2}{1-f_2}\right) - \log\left(\frac{f_1}{1-f_1}\right)}{\lambda s} \quad [5]$$

354

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

360

361 In the case of the examined AML sample (Figure 3a), the measured fitness
362 advantages predict the future clonal structure of the malignancy (in the
363 absence of treatment). Specifically, the larger of the two subclones present at
364 the point when the tumour was sampled is predicted to take over the tumour,
365 while the smaller clone is projected to become too rare to remain detectable
366 (Figure 5c). Despite the assumption of constant conditions, our framework
367 could be extended in the future to simulate treatment effects when those
368 mechanisms are known.

369

370

371

372

373 **Discussion**

374

375 Here we have demonstrated how the VAF distribution can be used to directly
376 measure evolutionary dynamics of tumour subclones. We confirmed that
377 subclonal selection causes an overrepresentation of mutations within the
378 expanding clone, manifested as an additional ‘peak’ in the VAF distribution, as
379 suggested by many recent studies^{18,26,34}. However, irrespective of subclonal
380 selection, the tumour will still show an abundance of low frequency variants (a

381 1/f-like tail) as the natural consequence tumour growth, wherein the number of
382 new mutations is proportional to the population size.

383

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

394

395 The values of fitness advantage we infer in human malignancies are similar to
396 reports from experimental systems. Evidence from growing human pluripotent
397 stem cells indicates that *TP53* mutants may have a fitness advantage as high
398 as 90% ($1+s=1.9$)³⁵ and that single chromosomal gains can provide a fitness
399 advantage of up to 50%³⁶ (range 20%-53%). A study of the competitive
400 advantage of mutant stem cells in the mouse intestine during tumour initiation
401 (at constant population size) showed that *KRAS* and *APC* mutant stem cells
402 have a ~2-4 fold increased fixation probability in single crypts³⁷ and *TP53*
403 mutant cells in mouse epidermis exhibited a 10% bias toward self-renewal³⁸.
404 Moreover, our inferred fitness advantages compare to large fitness
405 advantages measured in bacteria³⁹. Nevertheless, we acknowledge that
406 experimental systems may differ significantly from *in vivo* human tumour
407 growth and that new experimental systems are necessary to test these
408 measurements. We also note that we are only able to measure large changes
409 in fitness, and additional efforts will be needed to measure the complete
410 distribution of fitness effects (DFE) within cancers. Furthermore, the inferred
411 fitness value is sensitive to the underlying stochastic evolutionary model and
412 thus caution is warranted in directly comparing fitness values.

413

414 Our inferred *in vivo* mutation rates per population doubling are also in line with
415 experimental evidence. Seshadri et al.⁴⁰ reported somatic mutation rates in
416 normal lymphocytes of 5.5×10^{-8} - 24.6×10^{-8} and a 10-100 fold increase in
417 mutation rate in cancer cell lines such as B-cell lymphoma (5.2×10^{-7} - 13.1×10^{-7})
418 and ALL (66.6×10^{-7}). A recent analysis of a mouse tumour model indicates
419 somatic mutation rates in neoplastic cells are 11x higher than in normal
420 tissue.

421

422 Our analysis highlights that even if cancer subclones experience pervasive
423 weak selection, it is not sufficient to alter the clonal composition of the tumour
424 and therefore to cause the VAF distribution to deviate detectably from the
425 distribution expected under neutrality. It is important to note that the (initial)
426 growth of tumours makes them peculiar evolutionary systems, as tumour
427 growth dilutes the effects of selection⁴¹. Thus, our analysis does not discount
428 the possibility of a multitude of 'mini-drivers'⁴² but shows that these must have
429 a corresponding 'mini' effect on the subclonal composition of a tumour (and
430 that the VAF distribution in mini-driver tumours is well described by a neutral

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

435

436 Our previous analysis¹¹ suggested that neutral dynamics were rejected in a
437 higher percentage of colon cancers (approximately 65%) than the 21%
438 reported here. The discrepancy is explained by the stochasticity in the
439 evolutionary process where chance events can lead to deviations from the
440 neutral 1/f distribution. Unlike our previous analytic derivation, the Bayesian
441 model selection framework presented here captures this stochasticity (and
442 hence neutral evolution is preferred in a greater proportion of samples).

443

444 Our measurement of evolutionary trajectories facilitates mechanistic
445 prediction of how a tumour changes over time as demonstrated in our *in silico*
446 prediction (Figure 5a,b), with implications for anticipating the dynamics of
447 treatment resistant subclones. This may have particular value for novel
448 evolutionary therapeutic approaches such as ‘adaptive therapy’, where the
449 goal is to maintain the existence of competing subclones that mutually
450 suppress the growth of another^{44,45}. Our measurements of relative clone fitness
451 could potentially be used to optimize treatment regimes in order to maintain
452 the coexistence of competing populations.

453

454 We acknowledge that features not described in our model, e.g. the spatial
455 structure of the tumour, could affect the estimates of the evolutionary
456 parameters⁴⁶. Indeed, our analysis shows that there can be heterogeneity in
457 the evolutionary process within a tumour (only 1/5 regions of a single lung
458 tumour showed strong evidence of subclonal selection). Spatial models of
459 tumour evolution can help elucidate other important biological parameters
460 such as the degree of mixing within tumour cell populations, a purely spatial
461 phenomenon which cannot be quantified using non-spatial models such as
462 ours. We have recently shown how multiple samples per tumour increase the
463 power to detect selection, in part because of the increased probability of
464 sampling across a ‘subclone boundary’ where selection is evident¹². We also
465 acknowledge that complex, undetectable intermediate dynamics in the
466 evolution of subclones, such as multiple small subclonal expansions before a
467 subclone becomes detectable, are not modelled within our framework.

468

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

472

473 **Contributions**

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

479

480 **Acknowledgements**

481 We thank Weini Huang and Kate Chkhaidze for fruitful discussions. We are
482 grateful to Arul Chinnaiyan and Marcin Cieslik for providing us with data from
483 the MET500 cohort, and to Suet Leung from providing access to the gastric
484 cancer cohort. A.S. is supported by The Chris Rokos Fellowship in Evolution
485 and Cancer and by Cancer Research UK (A22909). T.A.G. is supported by
486 Cancer Research UK (A19771). C.P.B. is supported by the Wellcome Trust
487 (097319/Z/11/Z). B.W. is supported by the Geoffrey W. Lewis Post-Doctoral
488 Training fellowship. A.S. and T.A.G. are jointly supported by the Wellcome
489 Trust (202778/B/16/Z and 202778/Z/16/Z respectively). C.C is supported
490 by NIH R01CA182514. M.J.W is supported by a Medical Research Council
491 student scholarship. This work was also supported by Wellcome Trust funding
492 to the Centre for Evolution and Cancer (105104/Z/14/Z).
493

494 Figure Legends

495 **Figure 1. Modelling patterns of subclonal selection in sequencing data.**

496 **(a)** In a stochastic branching process model of tumour growth cells have birth
497 rate b and death rate d , mutations accumulate with rate μ . Cells with fitness
498 advantage (orange) grow at a faster net rate ($b-d$) than the host population
499 (blue). **(b)** The variant allele frequency (VAF) distribution contains clonal
500 (truncal) mutations around $f=0.5$ (in this example of diploid tumour), and
501 subclonal mutations ($f<0.5$) which encode how a tumour has grown. In the
502 absence of subclonal selection, a neutral $1/f^2$ tail describes the accumulation
503 of passenger mutations as the tumour expands. **(c)** A selected subclone
504 produces an additional peak in the distribution while a $1/f^2$ tail is still present
505 due to passenger mutations accumulating in both the original population and
506 the new subclone. **(d)** In the presence of subclonal selection, the magnitude
507 and average frequency of the subclonal cluster of mutations (red) encode the
508 age and size of a subclone respectively, which in turn allows measuring the
509 clone's selective advantage. **(e)** Frequentist power analysis of detectability of
510 an emerging selected subclone on simulated data. Only early and/or very fit
511 subclones caused significant alterations of the clonal composition of a tumour,
512 resulting in the rejection of the neutral (null) model. Tumours were simulated
513 to 10^6 cells and scaled to a final population size of 10^{10} with a mutation rate of
514 20 mutations per genome per division, each pixel represents the average
515 value for the metric (area between curves) over 50 simulations.

516 **Figure 2. Accurate recovery of evolutionary parameters from simulated**

517 **data using Approximate Bayesian Computation.** Our method recovered
518 the correct clonal structure in simulated tumour data for representative
519 examples of **(a)** a neutral case, **(b)** a 1 subclone case and **(c)** a two subclones
520 case. Grey bars are simulated VAF data, solid red lines indicate the median
521 histograms from the simulations that were selected by the statistical inference
522 framework (500 posterior samples), shaded areas are 95% intervals. The
523 inferred posterior distributions of the evolutionary parameters contained the
524 true values (dashed lines) for **(d,f)** the time of emergence of the subclones
525 and **(e,g)** the selection coefficient $1+s$. **(h)** The mean percentage error in
526 inferred parameter values across a virtual tumour cohort ($n=100$ tumours) was
527 below 10%. Boxplots show the median and inter quantile range (IQR), upper
528 whisker is 3rd quantile + 1.5*IQR and lower whisker is 1st quantile - 1.5*IQR.

530 **Figure 3. Quantifying selection from high-depth bulk sequencing of**

531 **human cancers.** Both **(a)** an acute myeloid leukemia (AML) sample and **(b)** a
532 breast cancer sample sequenced at whole-genome resolution showed
533 evidence of two selected subclones. **(c)** In the case of a multi-region whole-
534 exome sequenced case of lung cancer, one sample showed evidence of a
535 single subclone whereas four other samples **(d-g)** from the same patient were
536 consistent with the neutral model. Grey bars are the data, solid red lines
537 indicate the median histograms from the simulations that were selected by the
538 statistical inference framework (500 posterior samples), shaded areas are the
539 95% intervals. **(h)** Bayesian model selection reports the expected clonal
540 structure for each case (Bayes Factors reported above histograms). **(i)**
541 Inferred subclone fitness advantages were 20% and 80% faster than the

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

549
550 **Figure 4. Quantifying selection in large cohorts of primary tumours and**
551 **metastatic lesions. (a)** 21% of colon cancers (N=70) from TCGA (sequenced
552 to sufficient depth and with high enough cellularity for statistical inference),
553 29% of WGS gastric cancers (N=17) (data from ref.³⁰, filtered for cellularity)
554 and 53% of metastases (N=113) from sites had evidence of differentially
555 selected subclones. When present, differentially selected subclones were
556 found to have **(b)** large fitness advantages with respect to the host population
557 and **(c)** emerge early during growth. Bayes Factors for subclonal structures
558 for all data are reported in Supplementary Table 4. Posterior distributions
559 were generated from 500 samples. Boxplots show the median and inter
560 quantile range (IQR), upper whisker is 3rd quantile + 1.5*IQR and lower
561 whisker is 1st quantile - 1.5*IQR.

562
563 **Figure 5. Predicting the future evolution of subclones. (a)** VAF distribution
564 of an *in silico* tumour sampled at 10^5 cells was used to measure the fitness
565 and time of emergence of a subclone. Grey bars are the simulated data, solid
566 red lines indicate the median histograms from the simulations that were
567 selected by the statistical inference framework (500 posterior samples),
568 shaded areas are the 95% intervals. Inset shows error from ground truth. 500
569 posterior samples were taken to perform the inference. **(b)** These values were
570 then used to predict the spread of the subclone as the tumour grew to 10^7
571 cells, showing the predictions matched the ground truth. Predictions were
572 made by extrapolating the posterior distribution of 1+s using equations in the
573 main text. Solid line shows the median value from the posterior distribution,
574 shaded area shows the 95% interval. **(c)** Using the same approach in the
575 AML sample, where we measured 1+s, t_1 and t_2 , we would predict that
576 subclone 2 would become dominant within 3-4 further tumour doublings while
577 subclone 1 will become too small to be detected.

578
579
580

581 **References**

- 582 1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313
583 (2012).
- 584 2. Gay, L., Baker, A.-M. & Graham, T. A. Tumour Cell Heterogeneity. *F1000Res* **5**,
585 238–14 (2016).
- 586 3. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus
587 genome sequencing. *Nature* **512**, 155–160 (2014).
- 588 4. Burrell, R. A. & Swanton, C. Re-Evaluating Clonal Dominance in Cancer
589 Evolution. *Trends in Cancer* (2016). doi:10.1016/j.trecan.2016.04.002
- 590 5. Durrett, R. *Branching Process Models of Cancer*. (Springer, 2015).
- 591 6. Marjoram, P. & Tavaré, S. Modern computational approaches for analysing
592 molecular genetic variation data. *Nat Rev Genet* **7**, 759–770 (2006).
- 593 7. Fu, Y. X. & Li, W. H. Estimating the age of the common ancestor of a sample of
594 DNA sequences. *Mol Biol Evol* **14**, 195–199 (1997).
- 595 8. Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. Inferring coalescence
596 times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
- 597 9. Tsao, J. L. *et al.* Colorectal adenoma and cancer divergence. Evidence of
598 multilineage progression. *The American Journal of Pathology* **154**, 1815–1824
599 (1999).
- 600 10. Tsao, J. L. *et al.* Genetic reconstruction of individual colorectal tumor histories.
601 *PNAS* **97**, 1236–1241 (2000).
- 602 11. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A.
603 Identification of neutral tumor evolution across cancer types. *Nature Genetics*
604 **48**, 238–244 (2016).
- 605 12. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo
606 of tumor evolution. *Nature Genetics* **49**, 1015–1024 (2017).
- 607 13. Hartl, D. L. & Clark, A. G. *Principles of population genetics*. (Sinauer, 1997).
- 608 14. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor
609 progression. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18545–18550 (2010).
- 610 15. Cheek, D. & Antal, T. Mutation frequencies in a birth-death branching process.
611 *arXiv*
- 612 16. Kessler, D. A. & Levine, H. Scaling Solution in the Large Population Limit of the
613 General Asymmetric Stochastic Luria–Delbrück Evolution Process. *J Stat Phys*
614 **158**, 783–805 (2014).
- 615 17. Durrett, R. POPULATION GENETICS OF NEUTRAL MUTATIONS IN
616 EXPONENTIALLY GROWING CANCER CELL POPULATIONS. *The Annals of*
617 *Applied Probability* **23**, 230–250 (2013).
- 618 18. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007
619 (2012).
- 620 19. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution
621 lineage tracking. *Nature* (2015). doi:10.1038/nature14279
- 622 20. Gillespie, J. H. Genetic Drift in an Infinite Population: The Pseudohitchhiking
623 Model. *Genetics* **155**, 909–919 (2000).
- 624 21. Wu, C.-I., Wang, H.-Y., Ling, S. & Lu, X. The Ecology and Evolution of Cancer:
625 The Ultra-Microevolutionary Process. *Annu. Rev. Genet.* **50**, 347–369 (2016).
- 626 22. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical
627 systems in systems and population biology. *Bioinformatics* **26**, 104–110

- 628 (2010).
- 629 23. Honda, O. *et al.* Doubling time of lung cancer determined using three-
630 dimensional volumetric software: comparison of squamous cell carcinoma and
631 adenocarcinoma. *Lung Cancer* **66**, 211–217 (2009).
- 632 24. Peer, P. G., van Dijck, J. A., Hendriks, J. H., Holland, R. & Verbeek, A. L. Age-
633 dependent growth rate of primary breast cancer. *Cancer* **71**, 3547–3551
634 (1993).
- 635 25. Tilanus-Linthorst, M. M. A. *et al.* BRCA1 mutation and young age predict fast
636 breast cancer growth in the Dutch, United Kingdom, and Canadian magnetic
637 resonance imaging screening trials. *Clinical Cancer Research* **13**, 7357–7362
638 (2007).
- 639 26. Griffith, M. *et al.* Optimizing Cancer Genome Sequencing and Analysis. *Cell*
640 *Systems* **1**, 210–223 (2015).
- 641 27. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas
642 delineated by multiregion sequencing. *Science* **346**, 256–259 (2014).
- 643 28. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer.
644 *Nature* **500**, 415–421 (2013).
- 645 29. Cancer Genome Atlas Network. Comprehensive molecular characterization of
646 human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
- 647 30. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular
648 profiling identify new driver mutations in gastric cancer. *Nature Publishing*
649 *Group* **46**, 573–582 (2014).
- 650 31. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer.
651 *N Engl J Med* NEJMoa1616288–13 (2017). doi:10.1056/NEJMoa1616288
- 652 32. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer.
653 *Nature Publishing Group* **548**, 297–303 (2017).
- 654 33. Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. *Nat. ecol. evol.*
655 **1**, 77 (2017).
- 656 34. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary
657 triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
- 658 35. Merkle, F. T. *et al.* Human pluripotent stem cells recurrently acquire and
659 expand dominant negative P53 mutations. *Nature* 1–11 (2017).
660 doi:10.1038/nature22312
- 661 36. Rutledge, S. D. *et al.* Selective advantage of trisomic human cells cultured in
662 non- standard conditions. *Sci. Rep.* 1–12 (2016). doi:10.1038/srep22828
- 663 37. Vermeulen, L. *et al.* Defining stem cell dynamics in models of intestinal tumor
664 initiation. *Science* **342**, 995–998 (2013).
- 665 38. Klein, A. M., Brash, D. E., Jones, P. H. & Simons, B. D. Stochastic fate of p53-
666 mutant epidermal progenitor cells is tilted toward proliferation by UV B during
667 preneoplasia. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 270–275 (2010).
- 668 39. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a
669 10,000-generation experiment with bacterial populations. *PNAS* **91**, 6808–
670 6814 (1994).
- 671 40. Seshadri, R., Kutlaca, R. J., Trainor, K., Matthews, C. & Morley, A. A. Mutation
672 rate of normal and malignant human lymphocytes. *Cancer Res* **47**, 407–409
673 (1987).
- 674 41. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth.

675 *Nature Genetics* **47**, 209–216 (2015).
676 42. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of
677 polygenic cancer evolution. *Nature Reviews Cancer* 1–6 (2015).
678 doi:10.1038/nrc3999
679 43. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic
680 Tissues. *Cell* 1–35 (2017). doi:10.1016/j.cell.2017.09.042
681 44. Enriquez-Navas, P. M. *et al.* Exploiting evolutionary principles to prolong
682 tumor control in preclinical models of breast cancer. *Science Translational*
683 *Medicine* **8**, 327ra24–327ra24 (2016).
684 45. Zhang, J., Cunningham, J. J., Brown, J. S. & Gatenby, R. A. Integrating
685 evolutionary dynamics into treatment of metastatic castrate-resistant
686 prostate cancer. *Nat Commun* 1–9 (2017). doi:10.1038/s41467-017-01968-5
687 46. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of
688 mutational jackpot events in expanding populations revealed by spatial Luria-
689 Delbrück experiments. *Nat Commun* **7**, 12760 (2016).
690
691
692
693
694

695 **Methods**

696

697 **Simulating tumour growth**

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

707

708 Tumour growth is assumed to begin with a single transformed cancer cell that
709 has acquired the full set of alterations necessary for cancer expansion. In our
710 model, this first cell will therefore be carrying a set of mutations (the number
711 of these mutations can be modified) that will be present in all subsequent
712 lineages, and thus appear as clonal (present in all cells and thus will generate
713 the cluster of clonal mutations at frequency $\frac{1}{2}$ for a diploid tumour) within the
714 cancer population.

715

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

719

$$720 N(t) = e^{(b-d)t} \quad [6]$$

721

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

731

$$732 \beta = \frac{b-d}{b}. \quad [7]$$

733

734 **Simulating subclonal selection**

735 To include the effects of subclonal selection, a mutant is introduced into the
736 population that has a higher net growth rate (birth minus death) than the host
737 population. We only consider the cases of one or two subclonal populations
738 under selection at any given time. We deem this simplification to be
739 reasonable as the number of large-effect driver mutations in a typical cancer
740 is thought to be small (< 10 see ref⁴⁴). Additionally, we found that sequencing
741 depth $> 100X$ is required to resolve more than 1 subclone (Supplementary
742 Figure 9). Fitter mutants can have a higher birth rate, a lower death rate, or a
743 combination of the two, all of which results in the mutant growing at a faster

744 rate than the host population. Given that the host/background population has
 745 growth rate b_H and death rate d_H , and the fitter population has growth rate b_F
 746 and death rate d_F , we define the selective advantage s of the fitter population
 747 as:

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

748
 749
 750
 751 Fitter mutants can be introduced into the population with a specified selective
 752 advantage s and at a chosen time t_1 , allowing us to explore the relationship
 753 between the strength of selection and the time the mutant enters the
 754 population.

755 **Simulation method and parameters**

756 We used a rejection kinetic Monte Carlo algorithm to simulate the model⁴⁵.
 757 Due to the small number of possible reactions (we consider at most 3
 758 populations with different birth and death rates) this algorithm is more
 759 computationally efficient than a rejection-free kinetic Monte Carlo algorithm
 760 such as the Gillespie algorithm. The input parameters of the simulation are
 761 given in table 1.

762
 763
 764

b	Birth rate within host population
d	Death rate within host population
b_F	Birth rate of fitter populations, each new population will have a unique b_F
d_F	Death rate of fitter populations, each new population will have a unique d_F
s	Selective advantage of fitter populations (calculated from b_F and d_F)
μ	Mutation rate
t_{event}	Time when fitter mutant is introduced
N_{end}	Maximum population size, simulation stops once this maximum is reached

765 **Table 1:** Input parameters for simulation

766

767 The simulation algorithm is as follows:

768

- 769 1. Simulation initialized with 1 cell and set all simulation parameters.
- 770 2. Choose a random cell, i from the population.
- 771 3. Draw a random number $r \sim \text{Uniform}(0, b_{\max} + d_{\max})$, where b_{\max} and d_{\max}
 772 are the maximum birth and death rates of all cells in the population.
- 773 4. Using r , cell i will divide with probability proportional to its birth rate b_i
 774 and die with probability proportional to its death rate d_i . If $b_i + d_i$
 775 $< b_{\max} + d_{\max}$ there is a probability that cell i will neither divide nor die. If
 776 $\beta = 1$, ie no cell death then in the above $d_{\max} = 0$.
- 777 5. If cell divides, daughter cells acquire ν new mutations where
 778 $\nu \sim \text{Poisson}(\mu)$.
- 779 6. Time is increased by a small increment $\frac{1}{N(b_{\max} + d_{\max})} \tau$, where τ is an
 780 exponentially distributed random variable⁴⁷.

781 7. Go to step 2 and repeat until population size is N_{end} .

782

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

785

786 **Generating millions of simulations for parameter inference**

787

788 A number of simplifications to our simulation scheme were made to improve
789 computationally efficiency when used in our Bayesian inference method, a
790 procedure that requires potentially many millions of individual simulations to
791 be run in order to get accurate inferences. Our ultimate goal was to measure
792 the time subclones emerge and their fitness. These parameters are measured
793 in terms of tumour volume doublings, not in terms of cell division durations (as
794 this is unknown in human tumours). Our approximations allow us to quantify
795 relative fitness of subclones, measured in units of population doubling, from
796 the VAF distribution. The approximations are:

797

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

803

804 Timing the emergence of subclones depends on the number of mutations that
805 have accumulated in the first cell that gave rise to the subclone. This is the
806 product of the number of divisions and the mutation rate ($n \times \mu$), or
807 equivalently the number of tumour doublings \times the effective mutation rate
808 ($n_{doublings} \times \frac{\mu}{\beta}$). Given we measure everything in terms of tumour doublings
809 and the effective mutation rate (μ/β) is the only measure available to us from
810 the VAF distribution (from the low frequency $1/f$ tail), we reduce our search
811 space by fixing $\beta = 1$ and varying μ , recognizing that in reality the effective
812 mutation rate is likely to have $\beta < 1$.

813

814 We do note however that cell death ($\beta < 1$) can affect our inferences in two
815 ways. First of all, in the presence of one or more subclones, the low-frequency
816 tail which encodes $\frac{\mu}{\beta}$ consists of a combination of two or more $1/f$ tails. If there
817 are large differences in the β value between subclones, then the inference on
818 the effective mutation rate from the gradient of the low-frequency tail may be
819 incorrect. For example, a fitter subclone could arise due to decreased cell
820 death rather than increased proliferation. To quantify this effect, we simulated
821 subclones with differential fitness due to decreased cell death and measured
822 the error on the inferred $\frac{\mu}{\beta}$. Even in cases where the death rate was

823 dramatically different in the subclone compared to the host population

824 ($\beta = 1.0$ vs $\beta = 0.5$) the mean error on the estimates of the mutation rate was
825 42% (Supplementary Figure 5), significantly less than the order of magnitude
826 previously measured between cancer type¹¹ and so we conclude that the
827 constant β assumption is therefore acceptable. We do acknowledge however

828 that we may underestimate the effects of drift, which will be accentuated in
829 tumours with high death rates.

830

831 Approximation 2: We simulate a smaller tumour population size compared to
832 typical tumour sizes at diagnosis, and scale the inferred values *a posteriori*.
833 We note that the VAF distribution holds no information on the population size
834 (it measures only relative proportions) and furthermore simulating realistic
835 population sizes (in the order of tens or hundreds of billions of cells in human
836 malignancies) is computationally unfeasible. To circumvent this, we generate
837 synthetic datasets that capture the characteristics relevant to measuring the
838 fitness and time subclones emerge, namely the effective mutation rate
839 ($\frac{\mu}{\beta}$) encoded by the low frequency part of the distribution, the number of
840 mutations in any subclonal cluster and their frequency. Theoretical population
841 genetics is then used to transform these measurements into values of fitness
842 and time (via Equations [2] and [4]), and values are scaled by the realistic
843 population size $N_{end} = 10^{10}$.

844

845 Simulation length was required to allow the single cell that gives rise to the
846 subclone sufficient time to accumulate the number of mutations ultimately
847 observed in the empirical datum. In general, we found $N_{end}=10^3$ to be
848 sufficient, except for the breast cancer and AML samples where we used the
849 more conservative $N_{end}=10^4$. In general, $N_{end}=10^4$ is sufficient to be able to
850 measure the range of parameters considered in Figure 1e.

851

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

855

$$856 \quad t_{end} = \frac{\log((1-f_{sub}) \times N_{end})}{\log(2)}, \quad [10]$$

857

858 where f_{sub} is the frequency of the subclone. We assumed a realistic $N_{end} =$
859 10^{10} , for generating the posterior distributions in Figures 3 & 4. We also
860 generated posterior distributions for s as a function of N_{end} , for the AML,
861 breast and lung cancers. For realistically large $N_{end} (>10^9)$ the exact choice
862 has minimal effect on our inferred values of s (Supplementary Figure 6).

863

864 To confirm that these assumptions do not invalidate our approach, we
865 generated synthetic datasets with cell death and large final population size
866 (10^6). We then used our inference method (detailed below) with the
867 simplifying assumptions to infer the parameters used to generate these
868 synthetic tumours. This demonstrated that we were able to accurately recover
869 the input parameters when the simplifications were applied (Figure 2).

870

871

872 **Sampling**

873 To mimic the process of data generation by high-throughput sequencing we
874 performed various rounds of empirically-motivated sampling of the simulation
875 data. Sequencing data suffers from multiple sources of noise, most
876 importantly for this study is that mutation counts (VAFs) are sampled from the

877 true underlying frequencies in the tumour population (both because of the
 878 initial limited physical sampling of cells from the tumour for DNA extraction,
 879 and then due to the limited read depth of the sequencing). Additionally, it is
 880 challenging to discern mutations that are at low frequencies from sequencing
 881 errors, and the limited sampling of sequencing assays means that many low
 882 frequency mutations are likely not measured at all. Consequently only
 883 mutations above a frequency of around 5-10% with 100X sequencing are
 884 observable with certainty⁴⁸. The ability to resolve subclonal structures is thus
 885 dependent on the depth of sequencing.

886

887 Our sampling scheme to generate synthetic datasets was as follows. For
 888 mutation i with true frequency VAF_{true} , the sequence depth D_i is Binomially
 889 distributed:

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

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

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

892 or if over-dispersed sequencing is modelled^{49,50} we use the Beta-Binomial
 893 model, which introduces additional variance to the sampling:

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

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

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

896

897 **Modelling stem cells**

898 Stem cell architecture was modelled with two-compartments: long lived stem
 899 cells and short lived non-stem cells. Stem cells divided symmetrically to
 900 produce two stem cells with probability α and asymmetrically to produce a
 901 single stem cell and a single differentiated cell with probability $1 - \alpha$.
 902 Differentiated cells divided n further times before dying. At each division all
 903 cells accumulated mutations as described above. We used $\alpha = 0.1$ and $n=5$. If
 904 $\alpha = 1.0$ then the model is equivalent to the above exponential growth model.

905

906 **Bayesian Statistical Inference**

907 We used Approximate Bayesian Computation (ABC) to infer the evolutionary
 908 parameters. We evaluated the accuracy of our inferences using simulated
 909 sequencing data where the true underlying evolutionary dynamics was known.
 910 The simulation approach to generate synthetic data was taken instead of a
 911 purely statistical approach, as the simulation naturally accounts for effects that
 912 would be difficult to represent in a pure statistical model (such as the
 913 convolution of multiple within subclone mutations at lower frequency ranges).
 914 Furthermore, the posterior distribution reported from this method naturally
 915 account for uncertainties due to experimental noise and stochastic effects
 916 such as Poisson-distributed mutation accumulation and stochastic birth-death
 917 processes. For in-depth discussion on these stochastic effects, see the
 918 Supplementary Note.

919
920 As in all Bayesian approaches, the goal of the ABC approach was to produce
921 posterior distributions of parameters that give the degree of confidence that
922 particular parameter values are true, given the data. Given a parameter
923 vector of interest θ and data D , the aim was to compute the posterior
924 distribution $\pi(\theta|D) = \frac{p(D|\theta)\pi(\theta)}{p(D)}$, where $\pi(\theta)$ is the prior distribution on θ and
925 $p(D|\theta)$ is the likelihood of the data given θ . In cases where calculating the
926 likelihood is intractable, as was the case here where our model cannot be
927 expressed in terms of well-known and characterized probability distributions,
928 approximate approaches must be sought. The basic idea of these ‘likelihood
929 free’ ABC methods is to compare simulated data, for a given set of parameter
930 values, with observed data using a distance measure. Through multiple
931 comparisons of different input parameter values, we can produce a posterior
932 distribution of parameter values that minimise the distance measure, and in so
933 doing accurately approximate the true posterior. The simplest approach is
934 called the ABC rejection method and the algorithm is as follows⁵¹:

- 935
936 1) Sample candidate parameters θ^* from prior distribution $\pi(\theta)$
937 2) Simulate tumour growth with parameters θ^*
938 3) Evaluate distance, δ between simulated data and target data
939 4) If $\delta < \epsilon$ reject parameters θ^*
940 5) If $\delta \geq \epsilon$ accept parameters θ^*
941 6) Return to 1

942
943 We used an extension of the simple ABC rejection algorithm, called
944 Approximate Bayesian Computation Sequential Monte-Carlo (ABC SMC)^{22,52}.
945 This method achieves higher acceptance rates of candidate simulations and
946 thus makes the algorithm more computationally efficient than the simple
947 rejection ABC. It achieves this increased efficiency by propagating a set of
948 ‘particles’ (sample parameter values) through a set of intermediate
949 distributions with strictly decreasing ϵ until the target ϵ_T is reached, using an
950 approach known as sequential importance sampling⁵³. The ABC SMC
951 algorithm also allows for Bayesian model selection to be performed by placing
952 a prior over models and performing inference on the joint space of models
953 and model parameters, (m, θ_m) . In contrast to many applications of ABC that
954 use summary statistics, we use the full data distribution, thus avoiding issues
955 of inconsistent Bayes factors due to loss of information^{54,55}. For further details
956 on the algorithm see references²² and the Supplementary Note on the specific
957 details of our implementation. Bayes factors for all data are shown in
958 Supplementary Tables 4 and 5. We found that the probability of neutrality was
959 significantly correlated with our frequentist based neutrality metrics and that
960 the inferred mutation rates were highly similar (Supplementary Figure 19).

961
962 The clonal structure of the cancer is encoded by the shape of the VAF
963 distribution, we therefore used the Euclidean distance between the two
964 cumulative distributions (simulated and target datasets) for our inference.

965
966 **Testing for Selection in the Frequentist paradigm**

967 We also refined a simple analytical test in order to rapidly determine what
 968 evolutionary parameters of selection lead to an observable deviation of the
 969 VAF distribution from that expected under neutrality. Previously, we showed
 970 that under neutrality, the distribution of mutations with a frequency greater
 971 than f is given by¹¹:

$$972 \quad M(f) = \frac{\mu}{\beta} \left(\frac{1}{f} - \frac{1}{f_{max}} \right) \quad [11]$$

974
 975 We fit a linear model of $M(f)$ against $1/f$ and used the R^2 measure of the
 976 explained variance as our measure of the goodness of fit.

977
 978 Another approach is to use the shape of the curve described by Equation [5]
 979 and test whether our empirical data collapses onto this curve. To implement
 980 this approach, here we defined the *universal neutrality curve*, $\bar{M}(f)$. Given an
 981 appropriate normalization of the data, the mutant allele frequency distribution
 982 governed by neutral growth will collapse onto this curve, although we
 983 recognize that deviations due to stochastic effects are possible. We can
 984 normalize the distribution described by Equation [5] by considering the
 985 maximum value of $M(f)$ at $f=f_{min}$.

$$986 \quad \max(M(f)) = \frac{\mu}{\beta} \left(\frac{1}{f_{min}} - \frac{1}{f_{max}} \right) \quad [12]$$

$$987 \quad \bar{M}(f) = \frac{\frac{\mu}{\beta} \left(\frac{1}{f} - \frac{1}{f_{max}} \right)}{\max(M(f))} \quad [13]$$

$$988 \quad \bar{M}(f) = \frac{\left(\frac{1}{f} - \frac{1}{f_{max}} \right)}{\left(\frac{1}{f_{min}} - \frac{1}{f_{max}} \right)} \quad [14]$$

992
 993 $\bar{M}(f)$ is independent of the mutation rate and the death rate and therefore
 994 allows comparison with any dataset. To compare this theoretical distribution
 995 against empirical data we used the Kolmogorov distance, D_k , the Euclidean
 996 distance between $\bar{M}(f)$ and the empirical data and the area between $\bar{M}(f)$
 997 and the empirical data. The Kolmogorov distance D_k is the maximum distance
 998 between two cumulative distribution functions. Supplementary Figure 1
 999 provides a summary of the different metrics.

1000
 1001 To assess the performance of the 4 classifiers we ran 10^5 neutral and non-
 1002 neutral simulations and compared the distribution of the test statistics for
 1003 these two cases. Due to the stochastic nature of the model, not all simulations
 1004 that include selection will result in subclones at a high enough frequency to be
 1005 detected, therefore to accurately assess the performance of our tests we only
 1006 included simulations where the fitter subpopulation was within a certain range
 1007 (20% and 70% fraction of the final tumour size). All 4 test statistics showed
 1008 significantly different distributions between neutral and non-neutral cases
 1009 (Supplementary Figure 2). Under the null hypothesis of neutrality and a false
 1010 positive rate of 5%, the area between the curves was the test statistics with
 1011 the highest power (67%) to detect selection, slightly outperforming the
 1012 Kolmogorov distance and Euclidean distance, with the R^2 test statistics

1013 showing the poorest performance with a power of 61% (Supplementary
1014 Tables 1 and 2).

1015

1016 We also plotted receiver operating characteristic (ROC) curves by varying the
1017 discrimination threshold of each of the tests of selection and calculating true
1018 positive and false positive rates (using a dataset derived from simulations with
1019 subclonal populations at a range of frequencies, Supplementary Figure 3).

1020 This analysis showed that R^2 had the least discriminatory power, with the
1021 other 3 performing approximately equally well (see Supplementary Table 3 for
1022 AUC). Increasing the range of allowed subclone sizes decreased the classifier
1023 performance, likely because the subclone could merge into the clonal cluster
1024 or 1/f tail when it took a more extreme size.

1025

1026 **Code Availability Statement**

1027 Code for the simulation and inference method, frequentist based neutrality
1028 statistics and bioinformatic scripts are available at:

1029 <https://marcjwilliams1.github.io/quantifying-selection>

1030

1031 **Bioinformatics analysis**

1032 Variant calls from the original studies were used for the AML data²⁶,
1033 TRACERx³¹ data and MET500 data³². Our analysis of the TCGA colon cancer
1034 cohort and gastric cancers is explained in our previous publication¹¹. For both
1035 these cohorts, we required the cellularity > 0.4 to perform the analysis. For the
1036 breast cancer data¹⁸ and lung cancer data²⁷, bam files from the original study
1037 were obtained and variants were called using Mutect2⁵⁶ and filtered to require
1038 at least 5 reads reporting the variants in the tumour and 0 reads in the normal.
1039 To mitigate the effects of low frequency mutations arising from paralogous
1040 regions of the genome we filtered any mutations where 75bp regions either
1041 side of the mutations had multiple BLAST hits (minimum of 100bp hit length,
1042 maximum of 3% mismatching bases).

1043

1044 Copy number aberrations could also potentially result in the multi-peaked
1045 distribution we observe, hence we only used mutations that were found in
1046 regions identified as diploid (and without copy-neutral LOH). The original AML
1047 study found no evidence of copy number alterations. For the TCGA colon
1048 cancer cohort we used paired SNP array data to filter out mutations falling in
1049 non-diploid regions. For the TRACERx data and MET500 data we used allele
1050 specific copy number calls provided in the original studies to filter the data.
1051 For all other datasets we applied the Sequenza algorithm to infer allele
1052 specific copy number states and estimate the cellularity⁵⁷. As the original
1053 breast cancer study found evidence of subclonal copy number alterations in
1054 multiple chromosomes we only used mutations on chromosome 3 for our
1055 analysis, (Supplementary Figure 20). BAFs of regions called as copy neutral
1056 by Sequenza in the lung cancer sample were consistent with a diploid
1057 genome (Supplementary Figure 21).

1058

1059 We used cellularity estimated provided by the Sequenza algorithm to correct
1060 the VAFs for each individual sample. For a cellularity estimate κ , the corrected
1061 depth for variant i will be $\bar{d}_i = \kappa \times d_i$. When cellularity estimates from

1062 Sequenza were unavailable (MET500 and TRACERx) we fitted the cellularity
1063 using our ABC method by including it as an additional parameter.

1064

1065 As noted our simulation can account for the over-dispersion of allele read
1066 counts. To measure the over-dispersion parameter ρ , we fitted a Beta-
1067 Binomial model to the clonal cluster where we know $VAF_{true} = 0.5$. We used
1068 Markov Chain Monte Carlo (MCMC) to fit the following model to the right hand
1069 side of the clonal cluster so as to minimize the effects of the 1/f distribution or
1070 subclonal clusters:

1071

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

1072

1073 where D_i is the sequencing depth, f_i is the allele read count and ρ is the
1074 overdispersion parameter. We then used this estimate for ρ in the simulation
1075 sampling scheme. Supplementary figure 22 shows the fits to the clonal cluster
1076 for the AML data using both the Beta-Binomial and Binomial model, and
1077 supplementary table S6 reports the over-dispersion parameter for each
1078 dataset. We also used this analysis to further refine the cellularity estimate
1079 provided by sequenza, ensuring that the clonal cluster was centred at VAF =
1080 0.5. We note that some of the over-dispersion is likely artificial and introduced
1081 by the cellularity correction.

1082

1083 Mutational signatures in the breast cancer sample and AML sample
1084 (Supplementary Figure 12) were identified using the deconstructSigs R
1085 package⁵⁸ using the latest mutational signature probability file from COSMIC.
1086 Signature assignment was restricted to signatures known to be active in the
1087 respective cancer types. All other parameters were set to default values. To
1088 generate confidence intervals, we bootstrapped the assignment by generating
1089 50 datasets by sampling 90% of the mutations and running the regression on
1090 each dataset, we then report the mean value and the 95% CI.

1091

1092 **Data Availability Statement**

1093 Only publically available data was used in this study, and data sources and
1094 handling of these data are described above.

1095

1096

1097 **References**

1098

1099

- 1100 47. Waclaw, B. *et al.* A spatial model predicts that dispersal and cell turnover limit
1101 intratumour heterogeneity. *Nature* (2015). doi:10.1038/nature14971
- 1102 48. Stead, L. F., Sutton, K. M., Taylor, G. R., Quirke, P. & Rabbitts, P. Accurately
1103 identifying low-allelic fraction variants in single samples with next-generation
1104 sequencing: applications in tumor subclone resolution. *Human Mutation* **34**,
1105 1432–1438 (2013).
- 1106 49. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in
1107 cancer. *Nat Methods* **11**, 396–398 (2014).
- 1108 50. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in
1109 tumour cell populations. *Nat Commun* **3**, 811 (2012).

- 1110 51. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population
1111 growth of human Y chromosomes: a study of Y chromosome microsatellites.
1112 *Mol Biol Evol* **16**, 1791–1798 (1999).
- 1113 52. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical
1114 systems in systems and population biology. *Bioinformatics* **26**, 104–110
1115 (2010).
- 1116 53. Del Moral, P., Doucet, A. & Jasra, A. Sequential Monte Carlo samplers. *Journal*
1117 *of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 411–436
1118 (2006).
- 1119 54. Robert, C. P., Cornuet, J.-M., Marin, J.-M. & Pillai, N. S. Lack of confidence in
1120 approximate Bayesian computation model choice. *Proc. Natl. Acad. Sci. U.S.A.*
1121 **108**, 15112–15117 (2011).
- 1122 55. Barnes, C. P., Filippi, S., Stumpf, M. P. H. & Thorne, T. Considerate approaches
1123 to constructing summary statistics for ABC model selection. *Stat Comput* **22**,
1124 1181–1197 (2012).
- 1125 56. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure
1126 and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213–219 (2013).
- 1127 57. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles
1128 from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
- 1129 58. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C.
1130 deconstructSigs: delineating mutational processes in single tumors
1131 distinguishes DNA repair deficiencies and patterns of carcinoma evolution.
1132 *Genome Biology* 1–11 (2016). doi:10.1186/s13059-016-0893-4
1133
1134







