

1
2 **Intratumor Heterogeneity and**
3 **Circulating Tumor Cell Clusters**
4 Zafarali Ahmed¹, Simon Gravel^{2*}

5 **1** Department of Biology, McGill University,
6 Montreal, Quebec, Canada
7 **2** Department of Human Genetics, McGill
8 University, Montreal, Quebec, Canada
9 * simon.gravel@mcgill.ca

10 **Summary**

11 Genetic diversity plays a central role in tumor
12 progression, metastasis, and resistance to treat-
13 ment. Experiments are shedding light on this
14 diversity at ever finer scales, but interpretation
15 is challenging. Using recent progress in numer-
16 ical models, we simulate macroscopic tumors to
17 investigate the interplay between global growth
18 dynamics, microscopic composition, and circ-
19 ulating tumor cell cluster diversity. We find that
20 modest differences in growth parameters can
21 profoundly change microscopic diversity. Simple
22 outwards expansion leads to spatially segregated
23 clones, as expected, but a modest cell turnover
24 can result in mixing at the microscopic scale,
25 consistent with experimental observations. Us-
26 ing multi-region sequencing data from a Hepato-
27 cellular Carcinoma patient to validate our mod-
28 els, we propose that deep multi-region sequenc-
29 ing is well-powered to distinguish between lead-
30 ing models of cancer evolution. The genetic com-
31 position of circulating tumor cell clusters, which
32 can be obtained from non-invasive blood draws,
33 is therefore informative about tumor evolution
34 and its metastatic potential.

35 **Highlights**

- 36 1. Numerical and theoretical models show in-
37 teraction of front expansion, selection, and
38 mixing in shaping tumor heterogeneity.
39 2. Cell turnover increases intratumor hetero-
40 geneity
41 3. Simulated circulating tumor cell clusters
42 and microbiopsies exhibit substantial diver-
43 sity.

- 44 4. Simulations suggest attainable sampling
45 schemes able to distinguish between preva-
46 lent tumor growth models.

Introduction

47
48 Most cancer deaths are due to metastasis of
49 the primary tumor, which complicates treatment
50 and promotes relapse (Nguyen, Bos, and Mas-
51 sagué 2009; Eccles and Welch 2007; Holohan et
52 al. 2013). Circulating tumor cells (CTC) are
53 bloodborne enablers of metastasis that can be
54 isolated and genetically characterized (Massagué
55 and Obenauf 2016; Aceto et al. 2014). Counts
56 of single CTCs have been used to predict tu-
57 mor progression (Cristofanilli, Budd, et al. 2004;
58 Cristofanilli, Hayes, et al. 2005) and monitor cu-
59 rative and palliative therapies in breast (Rack
60 et al. 2014; Hayes et al. 2006) and lung cancers
61 (Maheswaran et al. 2008). CTCs have also been
62 isolated in clusters of 2-30 cells (Marrinucci et al.
63 2012). These CTC clusters, though rare, are as-
64 sociated with more aggressive metastatic cancer
65 and poorer survival rates in mice and breast and
66 prostate cancer patients (Aceto et al. 2014).

67 Cellular growth within tumors follows Dar-
68 winian evolution with sequential accumulation
69 of mutations and selection resulting in subclones
70 of different fitness (Nowell 1976; Greaves and
71 Maley 2012). Certain classes of mutations are
72 known to give cancer cells advantages beyond
73 local growth rates. For example, acquiring mu-
74 tations in *ANGPTL4* in breast tumors does not
75 appear to provide a growth advantage to cells
76 in the primary, however it enhances metastatic
77 potential to the lungs (Padua et al. 2008). Sim-
78 ilarly, breast tumors are more likely to metasta-
79 size into the lung or brain if they acquire mu-
80 tations in *TGF β* or *ST6GALNAC5*, respectively
81 (Bos et al. 2009; Padua et al. 2008). These genes
82 are referred to as metastasis progression genes
83 or metastasis virulence genes (Nguyen, Bos, and
84 Massagué 2009; Nguyen and Massagué 2007).

85 Mutations, including those in metastasis pro-
86 gression and virulence genes, are not uniformly
87 distributed in the tumor. Tumors show substan-
88 tial intratumoral heterogeneity (ITH) (Navin et

89 al. 2010; Sottoriva et al. 2015; McGranahan and
90 Swanton 2015) where subclones have private mu-
91 tations that can lead to subclonal phenotypes
92 (Yates et al. 2015; J. Zhang et al. 2014; Ger-
93 linger, Horswell, et al. 2014). A high degree of
94 ITH can allow tumors to explore a wide range
95 of phenotypes: this might result in a few can-
96 cer cells that have a metastatic phenotype in
97 early tumor growth. Additionally, ITH can con-
98 tribute to therapy resistance and relapse (Hil-
99 ley et al. 2014; Holohan et al. 2013). Study-
100 ing ITH is therefore important for understand-
101 ing cancer progression and improving therapeutic
102 and prognostic decisions (Hiley et al. 2014;
103 Jamal-Hanjani, Hackshaw, et al. 2014; Alizadeh
104 et al. 2015). To capture the complete mutational
105 spectrum of a primary tumor, multiple study de-
106 signs have been proposed that divide the tumor
107 into regionally representative samples, known as
108 multiregion sequencing (Gerlinger, Rowan, et al.
109 2012; Gerlinger, Horswell, et al. 2014; J. Zhang
110 et al. 2014; Yates et al. 2015).

111 Next-generation sequencing (NGS) of single
112 CTCs has shown that they have similar genetic
113 composition to both the primary and metastatic
114 lesions (Heitzer et al. 2013), and can therefore be
115 used as a non-invasive liquid biopsy to study tu-
116 mors and tumor heterogeneity, monitor response
117 to therapy, and determine patient-specific course
118 of treatment (Powell et al. 2012; Heitzer et al.
119 2013; Krebs et al. 2014; Hodgkinson et al. 2014).

120 Here we ask whether genetic heterogeneity
121 within individual circulating tumor cell clusters
122 can be informative about solid tumor progres-
123 sion. Because CTC clusters are thought to origi-
124 nate from neighboring cells in the tumor (Aceto
125 et al. 2014), heterogeneity within CTC clusters
126 is closely related to cellular-scale genetic hetero-
127 geneity within tumors. We therefore interpret
128 our simulation results as informative about both
129 micro-biopsies and circulating tumor cell clus-
130 ters.

131 We used an extension¹ of the simulator de-
132 scribed in Waclaw *et al.* (Waclaw et al. 2015)
133 to study the interplay of tumor dynamics, CTC
134 cluster diversity, and metastatic outlook. We

show that fine-scale tumor heterogeneity, and
therefore CTC cluster composition, depend sen-
sitively on the tumor growth dynamics and sam-
pling location. Simulated data is consistent with
recent sequencing experiments, but slightly finer
sampling will provide stringent tests that dis-
tinguish between state-of-the-art models. These
findings further reinforce the utility of fine-scale
tumor profiling and CTC clusters as clinical tools
to elucidate tumor information and clinical out-
look (Mateo et al. 2014; Ignatiadis, Lee, and Jef-
frey 2015).

Simulation Model

To simulate the growth of solid tumors, we use
TumorSimulator² (Waclaw et al. 2015). The
software is able to simulate a tumor containing
 $10^8 - 10^9$ cells, or roughly 2 cubic centimeters,
in 24 core-hours. The tumor consists of cells
that occupy points in a 3D lattice. Cells do not
move in this model: The tumor evolves through
cell division and death. Empty lattice sites are
assumed to contain normal cells which are not
modelled in TumorSimulator.

Each cell has an associated list of genetic al-
terations which represent single nucleotide poly-
morphisms (SNPs) that can be either passenger
or driver. Driver mutations increase the growth
rate by a factor $1 + s$, where $s \geq 0$ is the average
selective advantage of a driver mutation.

The simulation begins with a single cell that
already has an unlimited growth potential. Tu-
mor growth then proceeds by selecting a mother
cell randomly. It then divides with a probabili-
ty $b_0(1 + s)^k$ where b_0 is the initial birth rate
and k is the number of driver mutations. New
cells are given new passenger and driver muta-
tions according to two independent Poisson dis-
tributions parameterized by two mutation rates.
The mother cell dies with a probability propor-
tional to the death rate, d . Further details of the
algorithm are described in Supplemental Meth-
ods. Values of b_0 , s are selected as in Waclaw et
al. 2015. The mutation rates are selected as in
Waclaw et al. 2015 to facilitate comparisons be-

¹<https://github.com/zafarali/tumorheterogeneity>

²<http://www2.ph.ed.ac.uk/~bwaclaw/cancer-code/>

179 tween simulations and Ling et al. 2015 to match
180 empirical observations.

181 We consider three turnover scenarios corre-
182 sponding to three models for the death rate d :
183 (i) No turnover ($d = 0$), corresponding to simple
184 clonal growth (Hallatschek et al. 2007); (ii) Sur-
185 face Turnover ($d(x, y, z) > 0$ only if x, y, z is on
186 the surface), corresponding to a quiescent core
187 model (Shweiki et al. 1995) (iii) Turnover ($d > 0$
188 everywhere), a model favored in Waclaw et al.
189 2015 to explore ITH.

190 Results

191 Global composition

192 To determine the effect of the growth dynam-
193 ics on global intratumor heterogeneity, we first
194 consider the allele frequency spectra for different
195 turnover models (Fig 1, S1). In all cases, a ma-
196 jority of driver and passenger genetic variants are
197 at frequency less than 1%, as expected from the-
198 oretical and empirical observations (Wang et al.
199 2014). Passenger mutations represent the bulk of
200 ITH, consistent with the theoretical and exper-
201 imental evidence that neutral evolution drives
202 most ITH (Williams et al. 2016). For simu-
203 lations with low to moderate death rate, $d \in$
204 $\{0.05, 0.1, 0.2\}$, we find that the frequency spec-
205 tra are very similar across the three turnover
206 models (Fig 1, S1): A low death rate has little
207 impact on the global composition of a tumor.

208 When the death rate is increased to $d = 0.65$,
209 as in Waclaw et al. 2015, the different models
210 produce distinct frequency spectra (Fig 1b). As
211 in Waclaw et al. 2015, we find that the number of
212 high-frequency drivers is higher in the turnover
213 model than in the no turnover model. Whereas
214 Waclaw *et al.* interpreted this observation as
215 an indication that turnover reduces diversity, we
216 find that diversity is in fact increased for all types
217 of variants and at all frequencies. The number of
218 somatic mutations in the turnover model is 3.4
219 times higher than in the surface turnover model
220 and 6.2 times higher than in the no turnover
221 model. This is primarily due to a higher number
222 of cell divisions required to reach a given tumor
223 size when cell death occurs throughout the tu-

224 mor (Table S1). The Waclaw *et al.* model uses
225 a death rate of $d = 0.65$, which is a staggering
226 95% of the birth rate. The turnover model there-
227 fore has 8.3 times more cell divisions to reach a
228 given size, and the surface turnover has 4 times
229 more cell divisions than the no turnover model
230 (Table S1).

231 We find a large excess of rare variants com-
232 pared to most previous analytical models of tu-
233 mor evolution. The Wright-Fisher model for a
234 constant-sized population (the “standard neutral
235 model”) predicts that the distribution $\phi(f)$ of
236 mutations with frequency f decays as f^{-1} . Re-
237 cently published tumor models that account for
238 exponential population growth in a coalescent or
239 branching process framework (Ohtsuki and In-
240 nan 2017) predict $\phi(f) \sim f^{-1}$ to $\phi(f) \sim f^{-2}$,
241 depending on model parameters

242 Here we observe that, for variants above 1%
243 in frequency, $\phi(f) \simeq f^{-2.5}$. The tumor model
244 studied here departs from these previous mod-
245 els in three ways: the rate of population growth,
246 the presence of selection, and differential growth
247 in the core and edge of the tumor. Selection
248 itself has a weak effect on the scaling behavior
249 (Fig S2), and the different turnover models ex-
250 hibit similar scaling, suggesting that the overall
251 growth rate is not the culprit. We find that dif-
252 ferential growth across the tumor explains most
253 of the discrepancy. In fact, a simple determinis-
254 tic and neutral geometric model with differential
255 growth accurately predicts the observed decay
256 $\phi(f) \sim f^{-2.5}$ (Figs 1 and S2).

257 A geometric model

Here we model the tumor as a continuously grow-
ing sphere where only surface cells divide. If a
mutation appears in a cell at the surface of the
tumor at a time when the tumor has radius r ,
we suppose that this mutation occupies a cross-
section area a^2 of the tumor surface. It therefore
occupies a fraction $\frac{a^2}{4\pi r^2}$ of the surface of the tu-
mor at that point. If the tumor grows radially
outwards, the descendants of this cell occupy a
fraction $\frac{a^2}{4\pi r^2}$ of the space yet to be occupied, and

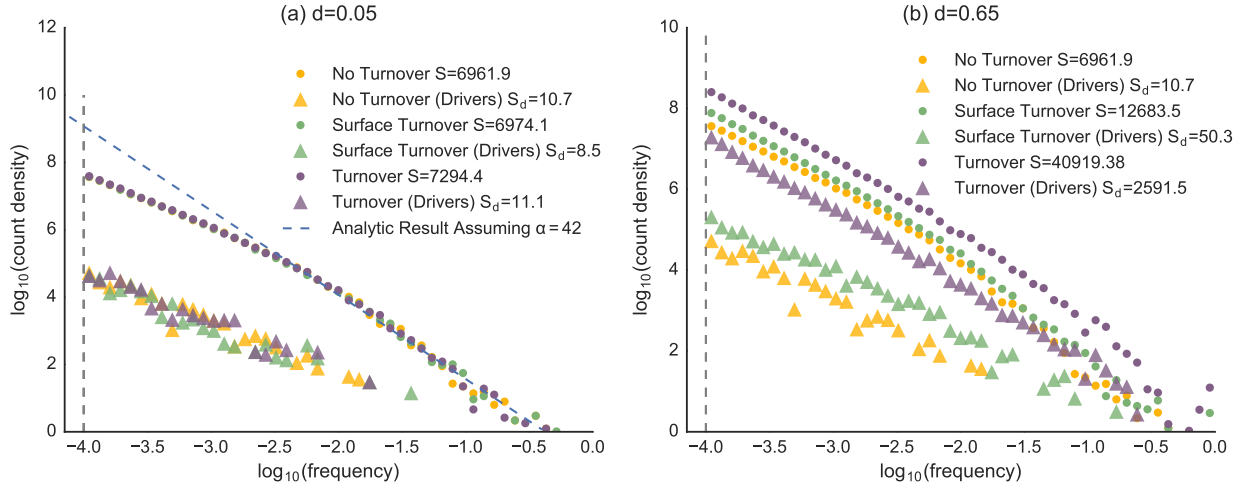


Figure 1: **Frequency Spectra for the Primary Tumor at (a) low death rate and (b) high death rate.** A histogram of the allele frequencies of all mutations (circles) and driver mutations (triangles) in the tumor. (a) At low death rate, the frequency spectra are indistinguishable, whereas for (b) higher death rate, the turnover model produces elevated diversity across the frequency spectrum for both driver and neutral mutations. (a) At low death rate, the frequency spectra are indistinguishable, whereas for (b) higher death rate, the turnover model produces elevated diversity across the frequency spectrum for both driver and neutral mutations. The total number of somatic mutations, S , and the total number of driver mutations, S_d , in the tumor is shown in the legend (average of 11 simulations). The gray dotted line shows the minimum frequency mutations returned by the tumor simulator. The blue dotted line shows the asymptotic result of a geometric model with a scaling of $\alpha = 42$. Fig S1 and S2 show simulations with intermediate values of d and $s = 0$ respectively.

the mutation itself will occupy a fraction

$$f(r) = \frac{a^2}{4\pi r^2} \left(1 - \frac{r^3}{R^3}\right)$$

258 of the final tumor, which is the volume of a spher-
 259 ical cone with its tip removed. We can then inte-
 260 grate over all possible radii r where mutations
 261 occur. The density $\rho(r)$ of mutations occurring
 262 at radius r is proportional to the density of cells
 263 at that locus

$$\rho(r) \simeq \mu \frac{4\pi r^2}{a^3},$$

with μ the mutation rate per cell. The frequency spectrum is therefore

$$\phi(f) = \int_0^R dr \rho(r) \delta(f - f(r)).$$

If we focus on common mutations, which occurred at $r \ll R$, we can approximate $f(r) \simeq \frac{a^2}{4\pi r^2}$, leading to

$$\phi(f) \simeq \frac{\mu}{4\sqrt{\pi} f^{\frac{5}{2}}}.$$

We show in Supplemental Methods that a model accounting for stochastic fluctuations in the early reproductive success of a mutation preserves this scaling behavior, but with an overall scale factor α that depends on details of the growth model, i.e.

$$\phi(f) \simeq \frac{\alpha\mu}{4\sqrt{\pi} f^{\frac{5}{2}}}.$$

Fig 1 shows the agreement of simulation results 264
 to the geometric model with $\alpha = 42$. Variants 265
 at less than 1% frequency follow a distinct power 266
 law that is closer to the $\phi(f) = f^{-2}$ described in 267
 (Ohtsuki and Innan 2017). 268

Cluster diversity depends on sampling position and turnover rate 269

To study the effect of cluster size, position of 271
 origin, and evolutionary model on CTC cluster 272
 composition, we sampled groups of cells across 273
 tumors (More details in CTC cluster synthesis). 274
 To assess genetic heterogeneity within clusters, 275

276 we consider the number of distinct somatic mu-
277 tations, $S(n)$, among cells in clusters of size n .

278 As expected, we find that larger CTC clus-
279 ters have more somatic mutations (Fig 2, S3).
280 By contrast with global diversity patterns, we
281 find that moderate turnover has a profound im-
282 pact: Clusters from models with low turnover
283 have many more somatic mutations than in the
284 no turnover model (Fig 2a,b). Surface turnover
285 has little effect on cluster diversity (Fig S3).

286 Fig 2 also shows the relationship between a
287 CTC cluster's shedding location (i.e. its distance
288 to the tumor center-of-mass when it was sam-
289 pled) and its genetic content. No turnover and
290 surface turnover models show similar trends of
291 increasing diversity with distance (Fig S3). Full
292 turnover models show an opposite trend of de-
293 creasing diversity with distance in clusters of in-
294 termediate size (Fig 2b-d and S4 for $d = 0.1, 0.2,$
295 and 0.65 , respectively). However, these trends
296 revert again when considering large clusters with
297 thousands of cells (Fig 3).

298 **Comparison with multi-region sequenc-** 299 **ing data**

300 We did not have access to large-scale sequencing
301 data for micro-biopsies. To validate predictions
302 of our model, we therefore used multi-region se-
303 quencing data from a Hepatocellular Carcinoma
304 (HCC) patient presented in (Ling et al. 2015)
305 (Fig 3a). The HCC data contained 23 sequenced
306 samples from a single tumor each with $\approx 20,000$
307 cells, therefore we used our sampling scheme to
308 produce 23 biopsies of comparable sizes (20,000
309 cells). The distance measurements were made
310 using ImageJ (Schneider, Rasband, and Eliceiri
311 2012) and Fig S1 from Ling et al. 2015. Since
312 (Ling et al. 2015) could only reliably call vari-
313 ants at more than 10% frequency, we used a
314 similar frequency cutoff in our simulations. The
315 HCC data does not show a clear spatial trend,
316 (Fig 3a) similarly to the model without turnover,
317 (Fig 3c), whereas the model with turnover pre-
318 dicta a detectable trend at comparable sample
319 size (Fig 3d). However, we have little statistical
320 power to distinguish between the models.

321 We therefore investigated the study design

that would be needed to effectively distinguish
between the different models proposed here.
Based on our simulations, power depends on
cluster size, number of clusters sampled, and the
choice of frequency cutoff. Interestingly, even
though the spatial trends in diversity are unde-
tectable in large clusters across all frequencies
(Fig S5), they are restored if we impose a fre-
quency cutoff (Fig 3c, d): The large number of
rare, recent variants overwhelms the signal about
early tumor evolution that can be gathered from
older, common variants.

Overall, the trends observed in Fig 2 are
barely detectable with the current sample size
but could be detected with modest increases
(Fig 3b). For biopsies containing tens of thou-
sands of cells, the number of spatially distributed
samples needed is ≈ 40 , roughly twice the size
of the HCC dataset. Alternatively, ≈ 30 small
cluster (23-30 cells) samples are necessary to
reliably detect spatial patterns. Furthermore,
intermediate-sized clusters show opposite trends
to both small and large clusters in the different
models (Fig 3b and S6). Thus small cluster se-
quencing may increase our power to discriminate
between leading models.

348 **CTC clusters derived from turnover** 349 **models are more likely to contain viru-** 350 **lent mutations**

Metastasis is an inefficient process (Massagué
and Obenauf 2016) in that most CTCs are elim-
inated from the circulatory system or fail to sur-
vive in the new microenvironment. We hypoth-
esize that the genetic composition of CTC clus-
ters influences the likelihood of implantation into
a new microenvironment. More specifically, ge-
netic heterogeneity within a cluster may con-
tribute to implantation by increasing the like-
lihood that a metastasis progression mutation is
present. If a cluster has S somatic mutations,
and each mutation has a small probability $p \ll 1$
of being a metastasis progression or virulence
gene, the probability of having at least one such
metastasis virulence gene is $1 - (1 - p)^S \approx Sp$.

Diverse CTC clusters do not carry more vir-
ulent mutations, on average, than homogeneous

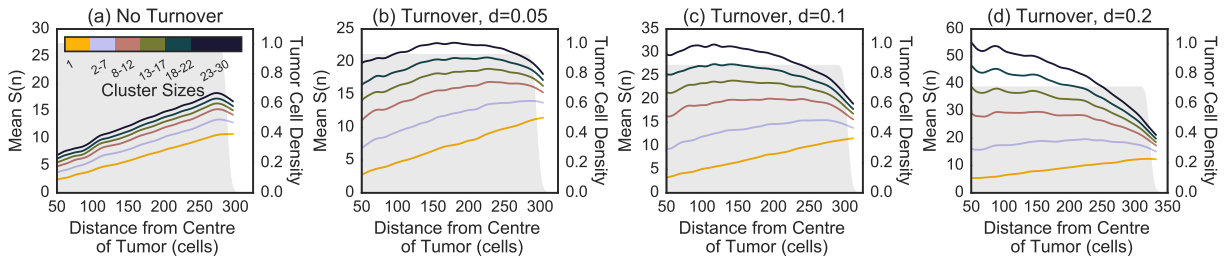


Figure 2: **Number of somatic mutations per cluster** as a function of cluster size and position for a model with (a) no turnover, (b) turnover with $d = 0.05$, (c) turnover with $d = 0.1$ and (d) turnover with $d = 0.2$. A higher number of somatic mutations increases the likelihood that a metastatic progression mutation is present. The number of mutations in single CTCs increases at the edge, reflecting the larger number of cell divisions. The trend is reversed for larger clusters with at higher death rate. The shaded gray area represents the density of tumor cells at each position. The smoothed curves were obtained by a Gaussian weighted average using weight $w_i(x) = \exp(-(x - x_i)^2)$, with x_i is the distance from the centre of the tumor. See Fig S3 and S4 for the surface turnover model and turnover model with $d = 0.65$ respectively.

368 ones, but they are more likely to carry *some* vir-
 369 ulent mutations because of the increased diver-
 370 sity. Unless implantation probability is exactly
 371 proportional to the number of cells carrying viru-
 372 lent mutations in a cluster, which seems unlikely,
 373 diversity will impact implantation rate.

374 To compare the increased likelihood that CTC
 375 clusters possess metastatic progression genes
 376 compared to single CTCs, we determine the rela-
 377 tive increase in the number of distinct somatic
 378 mutations in a CTC cluster versus a single CTC
 379 termed *cluster advantage*, $A(n)$. To disentangle
 380 the contributions from the microscopic and
 381 macroscopic diversity, as well as cluster size ef-
 382 fects, we compute the cluster advantage for clus-
 383 ters composed of neighboring cells, as well as for
 384 random sets of cells sampled across the tumor
 385 (Fig 4).

386 Whereas randomly sampled sets of cells show
 387 similar and almost linear increase of the cluster
 388 advantage with sample size, cell clusters show
 389 more variability. Turnover models have the
 390 highest cluster advantage, followed by the sur-
 391 face turnover model, and the no turnover model
 392 (Fig 4). Higher turnover increases the cluster
 393 advantage (Fig S7). Even low turnover with a
 394 death rate of $d = 0.05$ doubles the cluster ad-
 395 vantage compared to the no turnover and surface
 396 turnover model (Fig S7).

Discussion

397 Even though the results of our simulations are
 398 consistent with Waclaw *et al.* at the tumor-
 399 wide level (Waclaw et al. 2015), we reach oppo-
 400 site conclusions about the effect of cell turnover
 401 on genetic diversity. Waclaw *et al.* argued that
 402 turnover reduces diversity based on the obser-
 403 vation that more high-frequency variants were
 404 observed in the tumor with turnover: A small
 405 number of clones make up a larger proportion of
 406 the tumor. Even though we can reproduce the
 407 observation, we find that turnover models in fact
 408 vastly *increase* diversity according to more con-
 409 ventional metrics, for example by increasing the
 410 number of somatic mutations (by $\approx 5.9\times$) across
 411 the frequency spectrum. Both the increase in
 412 dominant clone frequency and increased overall
 413 diversity have the same simple origin: A tumor
 414 model with turnover requires more cell divisions
 415 to reach a given size. An early driver mutation
 416 has more time to realize a selective advantage
 417 and occupy a high fraction of the tumor, but car-
 418 rier cells are also more likely to accumulate new
 419 mutations along the way leading to increased di-
 420 versity (Fig 1 and Table S1).
 421

422 The impact of turnover on cellular heterogene-
 423 ity is particularly pronounced when considering
 424 small cell clusters. These fine-scale patterns,
 425 observed in Figs 2 and S3, can be interpreted
 426 by considering the expansion dynamics of each

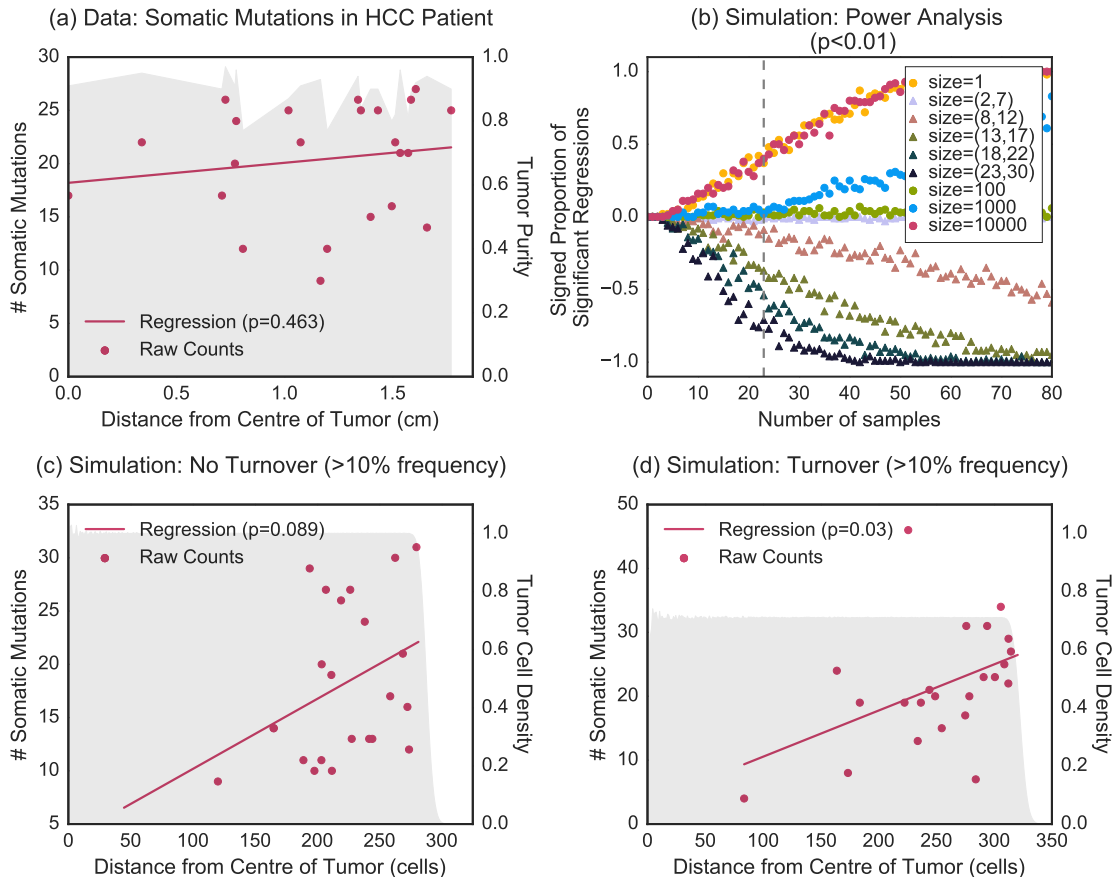


Figure 3: **Comparison of simulated multi-region NGS with empirical hepatocellular carcinoma.** (a) Spatial distribution and regression of the number of somatic mutations of 23 samples (20,000 cells each) in hepatocellular carcinoma patient. (b) Power to identify spatial trends in diversity as a function of cluster size and sample size. The signed proportion of significant regressions counts the number of regressions that were significant ($p < 0.01$) for positive and negative slopes (See Power Analysis). (c) and (d) Spatial trends in simulated tumors with sampling schemes as in (a), without turnover (c) and with turnover (d). The shaded gray area of (a) represents the tumor purity of the samples at each position. The shaded gray area of (c) and (d) represents the density of tumor cells at each position. See also Fig S5 and S6 for power analyses with different frequency cutoff and turnover model.

427 model and their impact on cell division and mix- 440
 428 ing. In all turnover models, the number of somatic 441
 429 mutations in a given cell is $\approx 2.75 \times$ higher 442
 430 at the edges than at the center of the tumor, re- 443
 431 flecting the higher number of divisions to reach 444
 432 the edge: The center of the tumor is occupied 445
 433 early, which slows down cell division. 446

434 In the no turnover and surface turnover mod- 447
 435 els, cell clusters show the same overall pattern 448
 436 of additional diversity at tumor edge. In the 449
 437 turnover model, however, we observe the oppo- 450
 438 site pattern: Even though edge *cells* still carry 451
 439 the most mutations, core *clusters* are now more 452

diverse than edge clusters.

Turnover increases the number of somatic mu- 441
 442 tations by increasing the number of cell divisions 443
 444 required to reach a given size, especially in the 445
 446 core. For example, core cells in the model with 447
 448 $d = 0.2$ have ≈ 3.99 somatic mutations, com- 449
 450 pared to ≈ 1.83 for the no turnover model. This 451
 452 effect is somewhat weaker for edge cells, leading 453
 to a modest spatial trend: Without turnover, the 454
 number of somatic mutations per cell is 3.5 times 455
 higher at the edge than in the core, and the ratio 456
 is reduced to 2.2 when turnover is present 457
 ($d = 0.2$). 458

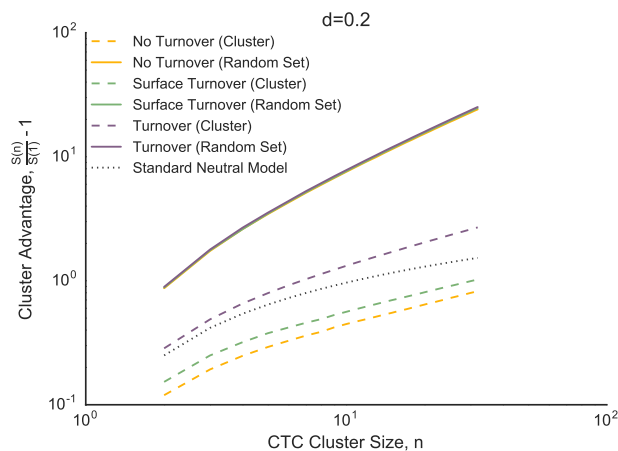


Figure 4: ‘Cluster advantage’ $A(n)$, or the increase in number of distinct somatic mutations in a CTC cluster relative to single CTC, as a function of cluster size for a random subset of 500 clusters drawn uniformly across the tumor. A law of diminishing returns applies to all models because of redundancy of mutations. The turnover model shows a 2-fold increase in the cluster advantage over the no turnover model. See also Fig S7 for death rates ≤ 0.1 .

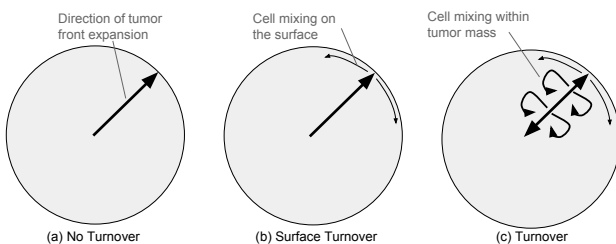


Figure 5: **Migration and Quiescent Core Explains Spatial Patterns** (a) In the no turnover model, the tumor front expands in the outward direction with no cell dying. There is little to no mixing and no divisions in the core: The number of somatic mutations increases with distance from the tumor center. (b) In the surface turnover model, the cells dying on the surface permit a small amount of mixing. This accounts for the higher number of somatic mutations per cluster. We still find increased diversity at the edge of the tumor because of the quiescent core. (c) In the turnover model, cells that die within the tumor can be replaced by cells from the surface as well as cells from the center.

453 More importantly for diversity, turnover al-
 454 lows for mixing of cells from nearby clones
 455 (Fig 5c). This mixing has a smaller effect at the
 456 edge of the tumor, where the range expansion
 457 produces serial bottlenecks which reduce the ef-
 458 fective population size relative to the tumor core.

459 For moderate cluster sizes, this differential mix-
 460 ing effect overwhelms the “number of divisions”
 461 effect, and core clusters are much more diverse
 462 than edge clusters, producing distinctive gradi-
 463 ents of diversity. Fig

464 The difference in somatic diversity between
 465 single CTCs and CTC clusters, measured
 466 through the cluster advantage, follows the ex-
 467 pected law of diminishing returns: the more cells
 468 in the cluster, the fewer the number of unique
 469 mutations per cell. However, the trends vary by
 470 growth model and cluster origin. Cell mixing af-
 471 forded by turnover reduces neighboring cell sim-
 472 ilarity and increases cluster advantage.

473 Under the assumption that the presence or ab-
 474 sence of a metastatic progression allele modu-
 475 lates metastatic potential of tumor cell clusters,
 476 the proportion of metastatic lesions that derive
 477 from circulating tumor cell clusters is highest in
 478 the turnover model. We can think of this as in-
 479 terference occurring between cells within a clus-
 480 ter. Alternately, this is an illustration of the ad-
 481 vantage of not putting all one’s egg in the same
 482 basket, applied to tumor metastasis: Assuming
 483 that there is a chance component to cluster im-
 484 plantation, mixing increases the likelihood that
 485 at least one virulence cell makes it to a hospitable
 486 site. Such an effect should be robust to details
 487 of the growth model.

488 In experiments, CTC clusters derived from
 489 primary breast and prostate tumors produced
 490 more aggressive metastatic tumors (Aceto et al.
 491 2014) compared to single CTCs. This is likely
 492 due to differences in mechanical properties of the
 493 cluster or the creation of a locally favorable en-
 494 vironment by the cluster, rather than by genetic
 495 differences. However, the present analysis sug-
 496 gests that this advantage can be enhanced by
 497 diversity within the cluster.

498 Both fine-scale mixtures of cell phenotypes
 499 and clonally constrained mutations have been
 500 observed experimentally in tumors (Yates et al.
 501 2015; Navin et al. 2010). Similarly, multi-region
 502 sequencing revealed high tumor heterogeneity in
 503 clear cell renal carcinoma (ccRCC) (Gerlinger,
 504 Horswell, et al. 2014), but low levels in lung
 505 adenocarcinomas (J. Zhang et al. 2014). This

506 strongly suggests that the amount of migration
507 and mixing varies substantially across tumors,
508 with ccRCC data being better described by a
509 model with turnover, whereas lung adenocarci-
510 noma data more closely resembles a model with
511 low or no turnover.

512 Distinguishing between migration effects,
513 turnover effects, and tumor growth idiosyn-
514 crasies is extremely challenging. Among lim-
515 itations of our model, we note the assump-
516 tion of spherical tumor shape and the absence
517 of complex physical constraints (which HCC tu-
518 mors may experience). Another limitation of the
519 present model is the rigid computational grid
520 which prevents cells from pushing each other out
521 of the way, which constrains growth in the cen-
522 ter of the tumor. This constraint plays a role
523 in reducing diversity at the center of the tumor,
524 but it may not be realistic in the earlier stages
525 of tumor growth.

526 The importance of such effects is largely un-
527 known, and it is likely to vary between tumors
528 and tumor types. Fortunately, we have shown
529 that we are at the cusp of being able to test
530 such models quantitatively. A sampling experi-
531 ment with twice as many samples than were col-
532 lected in the HCC patient studied above would
533 enable us to either validate or reject the current
534 state-of-the-art models (Fig 3b), and sequencing
535 of small clusters would further allow us to dis-
536 criminate between the different models studied
537 here.

538 Data collection schemes including the lung
539 TRACERx study (Jamal-Hanjani, Hackshaw, et
540 al. 2014; Jamal-Hanjani, Wilson, et al. 2017)
541 will help us put the state-of-the-art models to
542 the test and identify such important parameters
543 of tumor growth. Given our power analysis, we
544 find that sequencing small contiguous cell clus-
545 ters provides a richer picture of tumor dynamics
546 compared to larger biopsies, with little to no loss
547 in power, assuming that few-cell sequencing can
548 be performed accurately.

549 This work set out to answer two simple ques-
550 tions: First, should we expect substantial hetero-
551 geneity at the cellular scale within tumors and
552 within circulating tumor cell clusters? The an-

553 swer to the first question is most likely yes, as
554 even the models with no turnover exhibit mea-
555 surable cluster heterogeneity.

556 The second question was whether this het-
557 erogeneity, sampled through liquid biopsies or
558 multi-region sequencing, is informative about tu-
559 mor dynamics. Given that state-of-the-art mod-
560 els produce very different predictions about the
561 level of cluster heterogeneity, the answer is also
562 positive. This work identified some of the key
563 factors that determine cluster diversity, espe-
564 cially the interaction between range expansion,
565 cell turnover, and mixing. Even if no diversity
566 were observed at all in CTC clusters, it would
567 enable us to reject the present models in favor
568 of models including additional biological factors
569 that favor the clustering of genetically similar
570 cells. Measuring diversity, or the lack of di-
571 versity, within circulating tumor cell clusters or
572 fine-scale multi-region sequencing is therefore a
573 promising tool for both fundamental and medi-
574 cal oncology.

575 Author Contributions

576 Conceptualization, S.G.; Methodology, S.G.;
577 Software, Z.A.; Investigation, Z.A. and S.G.;
578 Writing Original Draft, Z.A.; Data Curation
579 Z.A.; Writing Review & Editing, Z.A & S.G.;
580 Visualization, Z.A.; Funding Acquisition, Z.A.
581 and S.G.; Resources, S.G.; Supervision, S.G.

582 Acknowledgments

We thank Julien Jouganous, Hamid Nikbakht,
Yasser Riazalhosseini, and Robert Sladek for
useful discussions. This research was made pos-
sible thanks to a Canadian Institutes of Health
Undergraduate Research Award in computa-
tional biology, funding reference numbers 139962
and 145987 and Frederick Banting and Charles
Best Canada Graduate Scholarship. This re-
search was undertaken, in part, thanks to fund-
ing from the Canada Research Chairs program
and a Sloan research fellowship.

References

- Nguyen, D. X., P. D. Bos, and J. Massagué (2009). “Metastasis: from dissemination to organ-specific colonization”. *Nature Reviews Cancer* 9.4, 274–284.
- Eccles, S. A. and D. R. Welch (2007). “Metastasis: recent discoveries and novel treatment strategies”. *The Lancet* 369.9574, 1742–1757.
- Holohan, C. et al. (2013). “Cancer drug resistance: an evolving paradigm”. *Nature Reviews Cancer* 13.10, 714–726.
- Massagué, J. and A. C. Obenauf (2016). “Metastatic colonization by circulating tumour cells”. *Nature* 529.7586, 298–306.
- Aceto, N. et al. (2014). “Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis”. *Cell* 158.5, 1110–1122.
- Cristofanilli, M., G. T. Budd, et al. (2004). “Circulating tumor cells, disease progression, and survival in metastatic breast cancer”. *New England Journal of Medicine* 351.8, 781–791.
- Cristofanilli, M., D. F. Hayes, et al. (2005). “Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer”. *Journal of Clinical Oncology* 23.7, 1420–1430.
- Rack, B. et al. (2014). “Circulating tumor cells predict survival in early average-to-high risk breast cancer patients”. *Journal of the National Cancer Institute* 106.5, dju066.
- Hayes, D. F. et al. (2006). “Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival”. *Clinical Cancer Research* 12.14, 4218–4224.
- Maheswaran, S. et al. (2008). “Detection of mutations in EGFR in circulating lung-cancer cells”. *New England Journal of Medicine* 359.4, 366–377.
- Marrinucci, D. et al. (2012). “Fluid biopsy in patients with metastatic prostate, pancreatic and breast cancers”. *Physical biology* 9.1, 016003.
- Nowell, P. C. (1976). “The clonal evolution of tumor cell populations”. *Science* 194.4260, 23–28.
- Greaves, M. and C. C. Maley (2012). “Clonal evolution in cancer”. *Nature* 481.7381, 306–313.
- Padua, D. et al. (2008). “TGF β primes breast tumors for lung metastasis seeding through angiopoietin-like 4”. *Cell* 133.1, 66–77.
- Bos, P. D. et al. (2009). “Genes that mediate breast cancer metastasis to the brain”. *Nature* 459.7249, 1005–1009.
- Nguyen, D. X. and J. Massagué (2007). “Genetic determinants of cancer metastasis”. *Nature Reviews Genetics* 8.5, 341–352.
- Navin, N. et al. (2010). “Inferring tumor progression from genomic heterogeneity”. *Genome Research* 20.1, 68–80.
- Sottoriva, A. et al. (2015). “A Big Bang model of human colorectal tumor growth”. *Nature genetics* 47.3, 209–216.
- McGranahan, N. and C. Swanton (2015). “Biological and therapeutic impact of intratumor heterogeneity in cancer evolution”. *Cancer Cell* 27.1, 15–26.
- Yates, L. R. et al. (2015). “Subclonal diversification of primary breast cancer revealed by multiregion sequencing”. *Nature medicine* 21.7, 751–759.
- Zhang, J. et al. (2014). “Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing”. *Science* 346.6206, 256–259.
- Gerlinger, M., S. Horswell, et al. (2014). “Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing”. *Nature genetics* 46.3, 225–233.
- Hiley, C. et al. (2014). “Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine”. *Genome Biology* 15.8, 453.
- Jamal-Hanjani, M., A. Hackshaw, et al. (2014). “Tracking genomic cancer evolution for precision medicine: the lung TRACERx study”. *PLoS Biology* 12.7, e1001906.
- Alizadeh, A. A. et al. (2015). “Toward understanding and exploiting tumor heterogeneity”. *Nature Medicine* 21.8, 846–853.
- Gerlinger, M., A. J. Rowan, et al. (2012). “Intratumor heterogeneity and branched evolu-

- tion revealed by multiregion sequencing”. *New England Journal of Medicine* 2012.366, 883–892.
- Heitzer, E. et al. (2013). “Complex tumor genomes inferred from single circulating tumor cells by array-CGH and next-generation sequencing”. *Cancer research* 73.10, 2965–2975.
- Powell, A. A. et al. (2012). “Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines”. *PloS one* 7.5, e33788.
- Krebs, M. G. et al. (2014). “Molecular analysis of circulating tumour cells-biology and biomarkers.” *Nature Reviews Clinical Oncology* 11.3, 129–44.
- Hodgkinson, C. L. et al. (2014). “Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer”. *Nature medicine* 20.8, 897–903.
- Waclaw, B. et al. (2015). “A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity”. *Nature* 525.7568, 261–264.
- Mateo, J. et al. (2014). “The promise of circulating tumor cell analysis in cancer management”. *Genome biology* 15.8, 448.
- Ignatiadis, M., M. Lee, and S. S. Jeffrey (2015). “Circulating tumor cells and circulating tumor DNA: challenges and opportunities on the path to clinical utility”. *Clinical Cancer Research* 21.21, 4786–4800.
- Ling, S. et al. (2015). “Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution”. *Proceedings of the National Academy of Sciences* 112.47.
- Hallatschek, O. et al. (2007). “Genetic drift at expanding frontiers promotes gene segregation”. *Proceedings of the National Academy of Sciences* 104.50, 19926–19930.
- Shweiki, D. et al. (1995). “Induction of vascular endothelial growth factor expression by hypoxia and by glucose deficiency in multicell spheroids: implications for tumor angiogenesis”. *Proceedings of the National Academy of Sciences* 92.3, 768–772.
- Wang, Y. et al. (2014). “Clonal evolution in breast cancer revealed by single nucleus genome sequencing”. *Nature* 512.7513, 155–160.
- Williams, M. J. et al. (2016). “Identification of neutral tumor evolution across cancer types”. *Nature Genetics* 48, 238–244.
- Ohtsuki, H. and H. Innan (2017). “Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population”. *Theoretical Population Biology* 117, 43–50.
- Schneider, C. A., W. S. Rasband, and K. W. Eliceiri (2012). “NIH Image to ImageJ: 25 years of image analysis”. *Nature Methods* 9.7, 671.
- Jamal-Hanjani, M., G. A. Wilson, et al. (2017). “Tracking the evolution of non-small-cell lung cancer”. *New England Journal of Medicine* 376.22, 2109–2121.
- Hou, J. M. et al. (2012). “Clinical significance and molecular characteristics of circulating tumor cells and circulating tumor microemboli in patients with small-cell lung cancer”. *Journal of Clinical Oncology* 30.5, 525–532.
- Durrett, R. (2008). *Probability models for DNA sequence evolution*. Springer Science & Business Media.

Supplemental Information

Supplemental Methods

Tumor growth model

The tumor consists of cells that occupy points in a 3D lattice. Empty lattice sites are assumed to contain normal cells which are not modelled explicitly in TumorSimulator.

Recall that each cell has an associated list of genetic alterations which represent single nucleotide polymorphisms (SNPs) that can be either passenger or driver. Driver mutations increase the growth rate by a factor $1 + s$, where $s \geq 0$ is the average selective advantage of a driver mutation.

At $t = 0$, the simulation begins with a single cell that already has an unlimited growth potential. The TumorSimulator algorithm then proceeds to grow the tumor through the following steps:

1. Select a random cell to be the mother cell.
2. Set the cell birth rate to $b' = b(1+s)^k$, where b is the initial tumor birth rate, s is the average selective advantage of a driver mutation, and k is the number of driver mutations present in the mother cell.
3. Randomly select a lattice point adjacent to the mother cell. If empty, create a genetically identical daughter cell at that position with a probability proportional to the birth rate, b' . If no cell created, or no empty sites are found proceed to 5.
4. Independently give mother and daughter cells additional passenger and driver mutation. The number of passenger and driver mutations are drawn according to Poisson distributions with mean λ_p and λ_d , respectively, and are drawn independently for the mother and daughter cell. Each mutation is unique and there is no back-mutations or recurrent mutations.
5. Kill (i.e., remove) the mother cell with probability proportional to the death rate d .
6. Update time by a small increment $dt = 1/(b_{max}N)$, where N is the total number of cancer cells in the tumor and b_{max} is the maximum birth rate in the population of cells.

In our analysis, we consider three turnover scenarios corresponding to three values of the death rate d : (i) No turnover ($d = 0$), corresponding to simple clonal growth (Hallatschek et al. 2007); (ii) Surface Turnover ($d(x, y, z) > 0$ only if x, y, z is on the surface), corresponding to a quiescent core model (Shweiki et al. 1995) (iii) Turnover ($d > 0$ everywhere), a model favored in Waclaw et al. 2015 to explore ITH.

The birth rate ($b = \ln(2)$), and selective advantage ($s = 1\%$) were kept consistent with Waclaw et al. 2015. In addition to varying the turnover model (full, surface, or none), we vary its intensity by controlling the death rate, $d \in \{0.05, 0.1, 0.2, 0.65\}$. TumorSimulator also has a parameter that controls migration of cells to form new independent cancer lesions. We did not allow such local migrations, as they would have little effect on the very fine-scale diversity in the primary tumor. We tried two values for the passenger mutation rate: $\lambda_p = 0.02$ to facilitate comparison with simulations from Waclaw et al. 2015, and $\lambda_p = 0.0375$ to match effective experimental observations from Ling et al. 2015.

TumorSimulator (Waclaw et al. 2015) is available at <http://www2.ph.ed.ac.uk/~bwaclaw/cancer-code/>.

CTC cluster synthesis

Experimental evidence suggests that CTC clusters are formed from neighboring cells in the primary tumor and not by agglomeration or proliferation of single CTCs in the blood (Hou et al. 2012; Aceto et al. 2014). To represent circulating tumor cell clusters, we therefore sampled spherical clusters (with a large radius) of cells in different areas of the tumor produced by the Waclaw *et al.* model. To get a fixed number of cells in the cluster, n , we picked the n closest cells to the center-of-mass of this sphere. We varied

the number of cells in the cluster from $n = 2$ to $n = 30$ to allow comparison to empirical findings Marrinucci et al. 2012.

Power Analysis

To establish the effectiveness of sequencing CTC clusters versus larger biopsies at detecting a trend and distinguishing between models, we conduct a power analysis. We do a linear regression on the number of somatic mutations per cluster (or biopsy) of size n as a function of distance from the center-of-mass (i.e, $S(n, r) = mr + c$ where m and c are discovered by the inference technique). We count the number of regressions that were significant ($p < 0.01$): This is denoted as the proportion of significant regressions (out of 100). To capture the direction of the slope, we calculate the sign of the coefficient m and report the *signed* proportion of significant regressions.

Standard Neutral Model for Cluster Advantage

The relative increase in the number of distinct somatic mutations in a CTC cluster versus a single CTC is given by the *cluster advantage*, i.e., $A(n) = \frac{S(n)-S(1)}{S(1)} = \frac{S(n)}{S(1)} - 1$, where $S(n)$ is the number of somatic mutations in a cluster of size n and $S(1)$ is the number of somatic mutations in the cell closest to the center-of-mass of the cluster (as described in Section). A higher cluster advantage indicates that a CTC cluster is more potent relative to a single CTC from the same tumor. In other words, a higher cluster advantage means less genetic redundancy within a cluster. To compare how clusters would behave under a model with no selection, we consider the *Standard Neutral Model*. We make the infinite sites assumptions, and therefore the expected number of somatic mutations in a sample of size n , $S(n)$, is proportional to the expected number of segregating sites, $S'(n)$. This is given by $E(S'(n)) = \mu H(n - 1)$ (Durrett 2008), where $H(n)$ is the n -th harmonic number, $\sum_{i=1}^n \frac{1}{i}$.

Allele frequency distribution under a stochastic spherical growth model

The deterministic model presented in the main text for the distribution of allele frequencies does not take into account the stochastic variation in the fate of cells, which is especially important in the first few generations after a mutation appears. To account for this, we can imagine that the initial frequency of each new mutation gets multiplied by a random factor i to account for the random differences in success in the original cells over the first few generations. In other words, i is the number of descendants produced by the original cell divided by the expected number of descendants for other cells at the same radius. If we only consider mutations with given i , we find

$$f_i(r) = \frac{ia^2}{4\pi r^2}$$

and

$$\phi_i(f) \simeq \frac{\mu i^{\frac{3}{2}}}{4\sqrt{\pi} f^{\frac{5}{2}}}.$$

If we assume that multipliers are drawn from a probability distribution $P(i)$ that is independent of r , we get an expected frequency spectrum

$$\phi(f) \simeq \sum_i P(i) \phi_i(f) = \frac{\mu \mathbb{E} \left[i^{\frac{3}{2}} \right]}{4\sqrt{\pi} f^{\frac{5}{2}}}.$$

Even though the $5/2$ scaling behavior is maintained, the expectation $\mathbb{E} \left[i^{\frac{3}{2}} \right]$ can be much larger than 1, as there is an early settler advantage in this model. However, the value of this scaling factor depends on the details of the growth model (Fig 1 and S2).

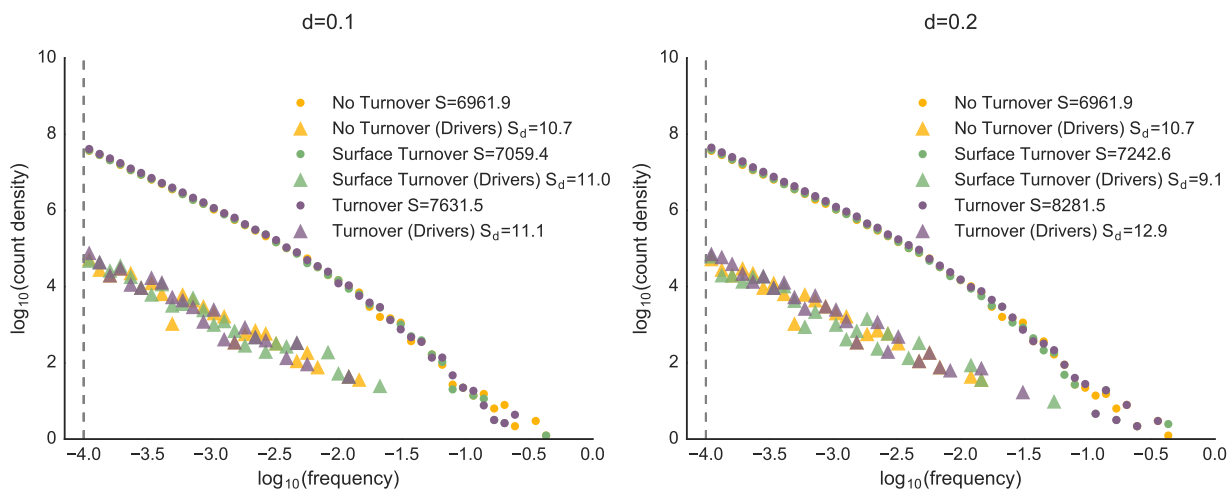
More generally, the $f^{-\frac{5}{2}}$ asymptotic result is derived under an extremely simple model: it does not take into account selection, turnover, and the fact that $P(i)$ likely varies with r . By focusing on high-frequency variants, the model also effectively ignores the contribution of variants that are ultimately unsuccessful and remain buried under the surface. Obtaining a general analytical approximation to the general allele frequency distribution appears extremely challenging.

Code Availability

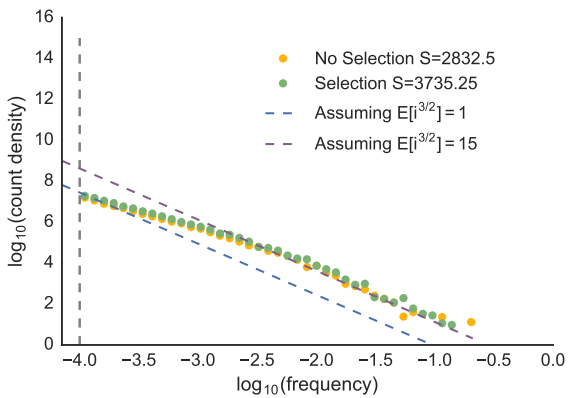
The code to reproduce simulations, analysis and figures can be found at <https://github.com/zafarali/tumorheterogeneity>.

Table 1: Average number of generations for a cell in each model (estimated from the number of somatic mutations per cell divided by the mutation rate).

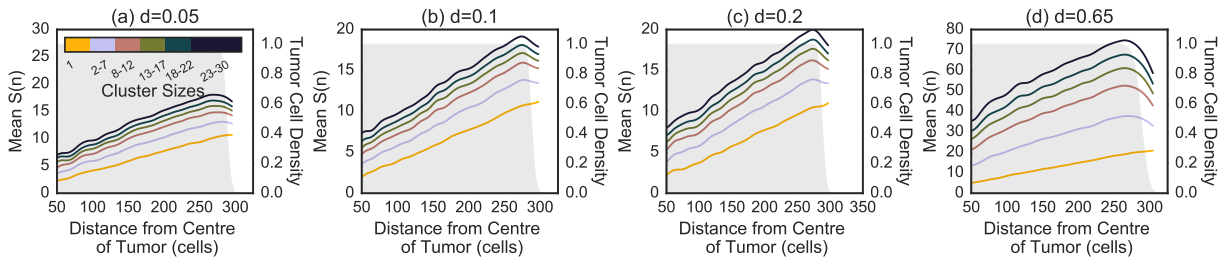
Average Number of Divisions in Model (mutation rate = 0.02, birth rate = 0.69)			
Death Rate (d)	No Turnover	Surface Turnover	Turnover
0.05	218.23 ± 13.99	216.51 ± 13.99	224 ± 11.00
0.1	218.23 ± 13.99	219.73 ± 7.11	239.38 ± 8.06
0.2	218.23 ± 13.99	227.27 ± 6.24	279.80 ± 13.00
0.65	218.23 ± 13.99	439.90 ± 18.21	1799.05 ± 55.81



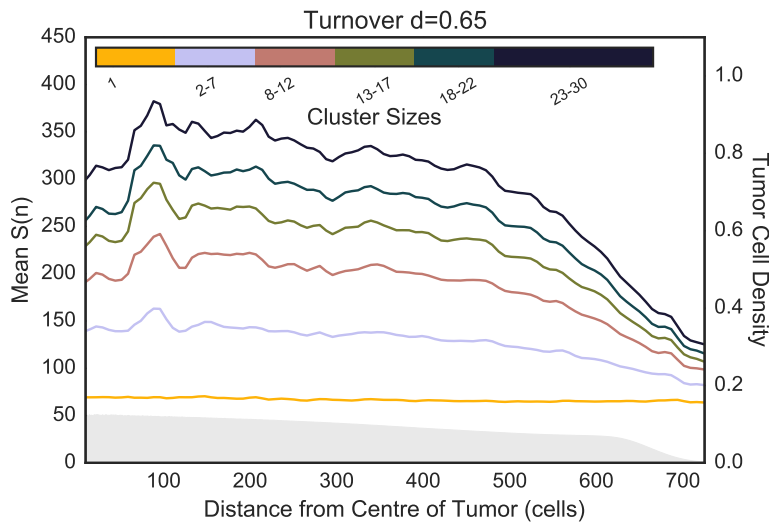
Supplemental Figure 1: Allele frequency spectra for low death rates, $d \in \{0.1, 0.2\}$ are indistinguishable.



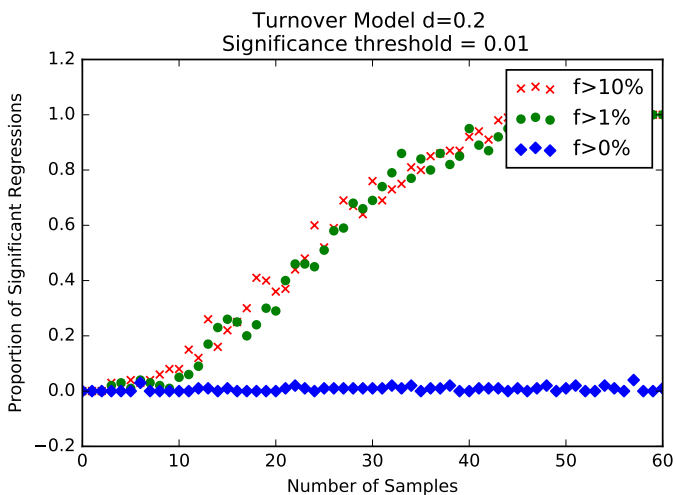
Supplemental Figure 2: Comparison of the allele frequency spectrum for simulations with and without selection, and analytic solutions of a tumor (size 10^8) with no death.



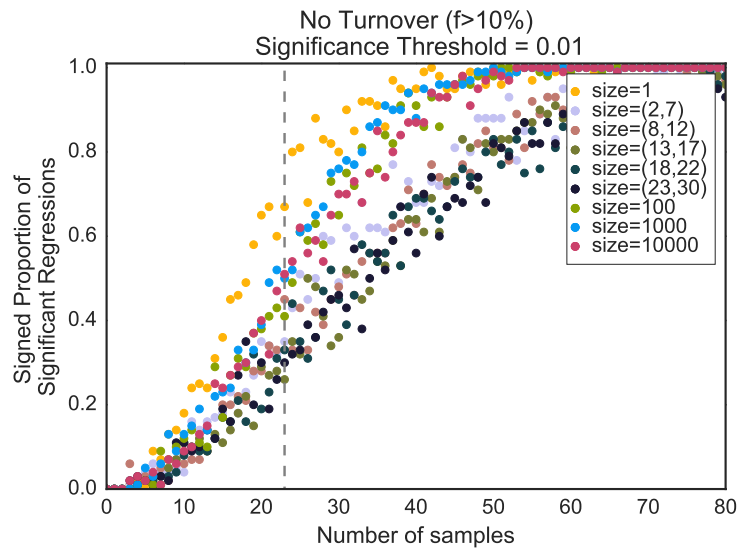
Supplemental Figure 3: The spatial distribution of the number of somatic mutations per cluster in the surface turnover model with death rates (a) $d = 0.05$, (b) $d = 0.1$, (c) $d = 0.2$ and (d) $d = 0.65$.



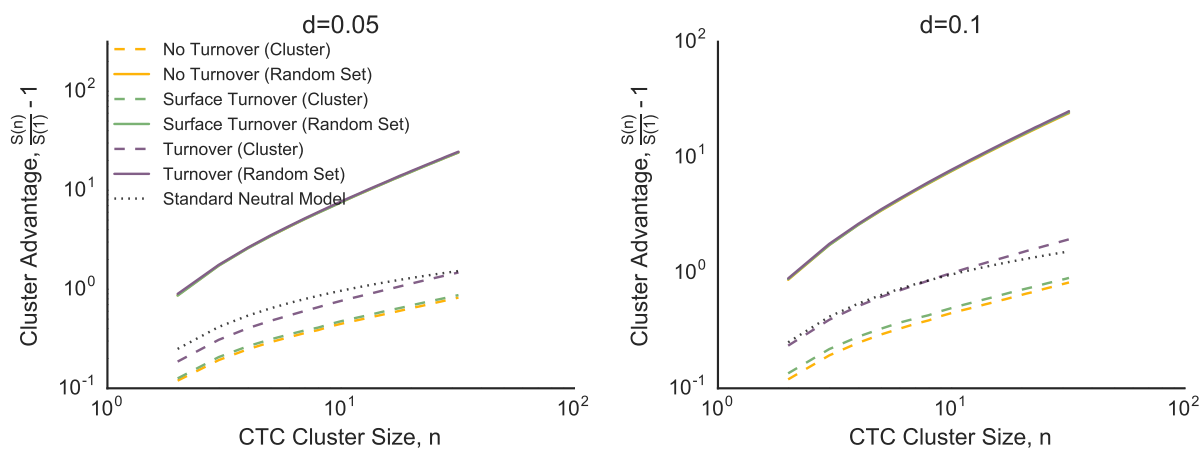
Supplemental Figure 4: The spatial distribution of the number of somatic mutation per cluster in a turnover model with $d = 0.65$.



Supplemental Figure 5: The power to detect spatial trends in diversity as a function of the frequency cutoff. With no frequency cutoff, the number of rare variants in a large biopsy ($n = 20,000$ cells) overwhelms the detectable spatial pattern contributed by common variants.



Supplemental Figure 6: The number of samples necessary to detect spatial trends from a regression analysis for CTCs and biopsies in the no turnover model.



Supplemental Figure 7: Cluster advantage for weak turnover models: even weak mixing (turnover model with $d = 0.05$) can lead to substantial differences in the cluster advantage.