# Genetic Diversity in Circulating Tumor Cell Clusters

Zafarali Ahmed[1], Simon Gravel[2*]

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

## Abstract

Genetic diversity plays a central role in tumor progression, metastasis, and resistance to treatment. Experiments are shedding light on this diversity at ever finer scales, but interpretation is challenging. Using recent progress in numerical models, we simulate macroscopic tumors to investigate the interplay between global growth dynamics, microscopic composition, and circulating tumor cell cluster diversity. We find that modest differences in growth parameters can profoundly change microscopic diversity. Simple outwards expansion leads to spatially segregated clones, as expected, but a modest cell turnover can result in mixing at the microscopic scale, consistent with experimental observations. Whereas simple range expansion models predict maximum diversity at the tumor edge, turnover models predict maximum diversity near the core of the tumor and a higher potency of CTCs for metastasis. Using multi-region sequencing data from a Hepatocellular Carcinoma patient to validate our models, we propose that deep multi-region sequencing is well-powered to distinguish between some of the leading models of cancer evolution. The genetic composition of circulating tumor cell clusters, which can be obtained from non-invasive blood draws, is therefore informative about tumor evolution, the position of origin of the cluster within the tumor, and its metastatic potential. It is therefore a promising tool for both fundamental and medical research.

## Introduction

Most cancer deaths are due to metastasis of the primary tumor, which complicates treatment and promotes relapse [1–3]. Circulating tumor cells (CTC) are bloodborne enablers of metastasis that can be isolated and genetically characterized [4, 5]. Counts of single CTCs have been used to predict tumor progression [6, 7] and monitor curative and palliative therapies in breast [8, 9] and lung cancers [10]. CTCs have also been isolated in clusters of 2-30 cells [11]. These CTC clusters, though rare, are associated with more aggressive metastatic cancer and poorer survival rates in mice and breast and prostate cancer patients [5].

Cellular growth within tumors follows Darwinian evolution with sequential accumulation of mutations and selection resulting in subclones of different fitness [12, 13]. Certain classes of mutations are known to give cancer cells advantages beyond local growth rates. For example, acquiring mutations in *ANGPTL4* in breast tumors does not appear to provide a growth advantage to cells in the primary, however it enhances metastatic potential to the lungs [14]. Similarly, breast tumors are more likely to metastasize into the lung or brain if they acquire mutations in *TGFβ* or *ST6GALNAC5*, respectively [14, 15]. These genes are referred to as metastasis progression genes or metastasis virulence genes [1, 16].

Mutations, including those in metastasis progression and virulence genes, are not uniformly distributed in the tumor. Tumors show substantial intratumoral heterogeneity (ITH) [17–19] where subclones have private mutations that can lead to subclonal phenotypes [20–22]. A high degree of ITH can allow tumors to explore a wide range of phenotypes: this might result in a few cancer cells that have a metastatic phenotype in early tumor growth. Additionally, ITH can contribute to therapy resistance and relapse when a treatment is applied during the late tumor stages [3, 23]. Studying ITH is therefore important for cancer progression and will be a key consideration during therapeutic and prognostic decisions [23–25]. To capture the complete mutational spectrum of a primary tumor, multiple samples across a tumor are required: Indeed, this has been the highlight of multiple empirical

studies [20–22, 26].

Next-generation sequencing (NGS) of single CTCs has shown that they have similar genetic composition to both the primary and metastatic lesions [27]. This opens the way for using CTC and CTC clusters as a non-invasive liquid biopsy to study tumors, monitor response to therapy, and determine patient-specific course of treatment [27–30].

Here we ask whether genetic heterogeneity within individual circulating tumor cell clusters can be informative about solid tumor progression. Because CTC clusters are thought to originate from neighboring cells in the tumor [5], heterogeneity within CTC clusters is closely related to cellular-scale genetic heterogeneity within tumors. We therefore suppose that CTC cluster diversity is a direct function of diversity in small cell clusters within the tumor.

We used an extension[1] of the simulator described in Waclaw *et al.* [31] to study the interplay of tumor dynamics, CTC cluster diversity, and metastatic outlook. We show that fine-scale tumor heterogeneity, and therefore CTC cluster composition, depend sensitively on the tumor growth dynamics and sampling location. Simulated data is consistent with recent sequencing experiments, but slightly finer sampling will provide stringent tests that distinguish 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 outlook [32, 33].

# Results

## Global composition

To determine the effect of the growth dynamics on global intra-tumor heterogeneity, we first consider the allele frequency spectra for different turnover models (Fig 1, S1). In all cases, a majority of driver and passenger genetic variants are at frequency less than 1%, as expected from theoretical and empirical observations [34]. Passenger mutations represent the bulk of ITH, consistent with the theoretical and experimental evidence that neutral evolution drives most ITH [35]. For simulations with low to moderate death rate, $d = \{0.05, 0.1, 0.2\}$, we find that the frequency spectra are indistinguishable between the three turnover models (Fig 1, S1). This suggests that a low death rate does not affect the global composition of a tumor.

When the death rate is increased to $d = 0.65$, as in [31], the different models produce distinct frequency spectra (Fig 1b). As in [31], we find that the number of high-frequency drivers is higher in the turnover model than in the no turnover model. Whereas [31] interpreted this observation as an indication that turnover reduces diversity, we find that diversity is in fact increased for all types of variants and at all frequencies. The number of somatic mutations in the turnover model is 3.4 times higher than in the surface turnover model and 6.2 times higher than in the no turnover model. This is primarily due to a higher number of cell divisions required to reach a given tumor size when cell death occurs throughout the tumor (Table S1). The Waclaw *et al.* model uses a death rate of $d = 0.65$, which is a staggering 95% of the birth rate. The turnover model therefore has 8.3 times more cell divisions to reach a given size, and the surface turnover has 4 times more cell divisions than the no turnover model (Table S1).

## Cluster diversity depends on sampling position and turnover rate

To study the effect of cluster size, position of origin, and evolutionary model on CTC cluster composition, we sampled groups of cells across tumors. To assess genetic heterogeneity within clusters, we consider the number of distinct somatic mutations, $S(n)$, among cells in clusters of size $n$.

As expected, we find that larger CTC clusters have more somatic mutations (Fig 2, S2). By contrast with global diversity patterns, we find that moderate turnover has a profound impact: Clusters from models with low turnover have many more somatic mutations than in the no turnover model (Fig 2a,b). Surface turnover

[1] https://github.com/zafarali/tumorheterogeneity

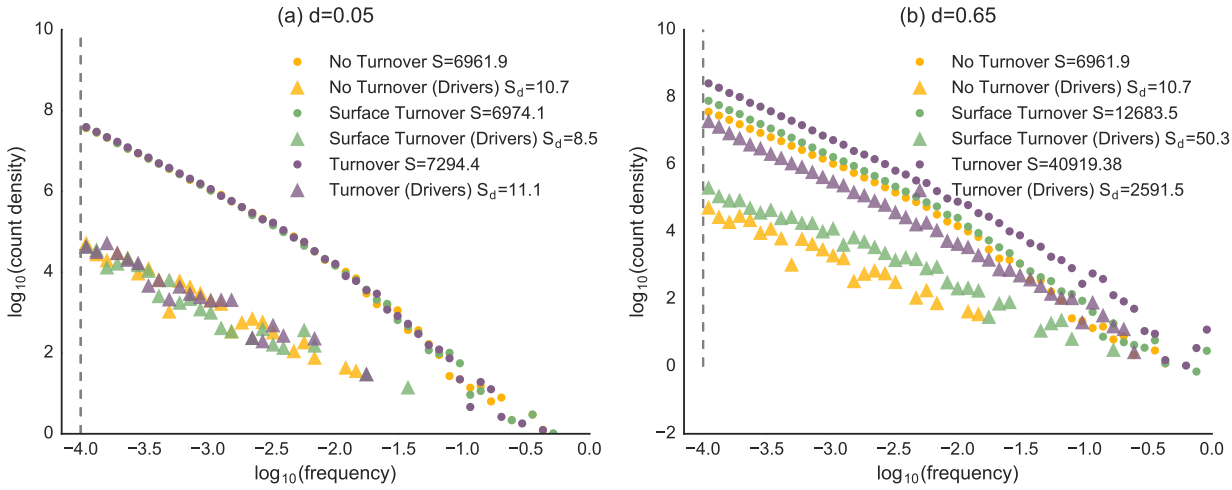



Figure 1: **Frequency Spectra for the Primary Tumor at (a) low death rate and (b) high death rate.** A histogram of the allele frequencies of all mutations (circles) and driver mutations (triangles) in the tumor. (a) At low death rate, the frequency spectra are indistinguishable, whereas for (b) higher 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.

has little effect on cluster diversity (Fig S2).

Fig 2 also shows the relationship between a CTC cluster's shedding location (i.e. its distance to the tumor center-of-mass when it was sampled) and its genetic content. No turnover and surface turnover models show similar trends of increasing diversity with distance (Fig S2). Full turnover models show an opposite trend of decreasing diversity with distance in clusters of intermediate size (Fig 2b-d and S3 for $d \in \{0.1, 0.2\}$ and $\{0.65\}$, respectively). However, these trends revert again when considering large clusters with thousands of cells (Fig 3).

## Comparison with multi-region sequencing data

To validate predictions of our model, we used multi-region sequencing data from a Hepatocellular Carcinoma (HCC) patient presented in [36] (Fig 3a). The HCC data contained 23 sequenced samples each with $\approx 20,000$ cells, therefore we used our sampling scheme to produce 23 biopsies of comparable sizes (20,000 cells). The distance measurements were made using ImageJ [37] and Fig S1 from [36]. Since [36] could only reliably call variants at more than 10% frequency, we used a similar frequency cutoff in our simulations. Interestingly, even though the spatial trends in diversity are undetectable in large clusters (Fig S4), they are restored if we impose a frequency cutoff (Fig 3c, d). Therefore, the spatial trends strongly depend on our choice of sample size and frequency cutoff (Fig S4), with low cutoff showing weaker spatial patterns. For large samples and low cutoffs, the large number of rare, recent variants overwhelms the signal for older common variants. Such trends are similar across turnover models (Fig 2c, d) and are barely detectable with the current sample size (Fig 3b). The trends observed in the HCC data (Fig 3a) are consistent with these but not sigificant.

Fig 3b shows the number of different samples necessary to reliably identify spatial trends. For biopsies containing tens of thousands of cells, the number of spatially distributed samples needed is $\approx$ 40, roughly twice the size of the HCC dataset. Furthermore, these show similar qualitative trends for both models, with an increase

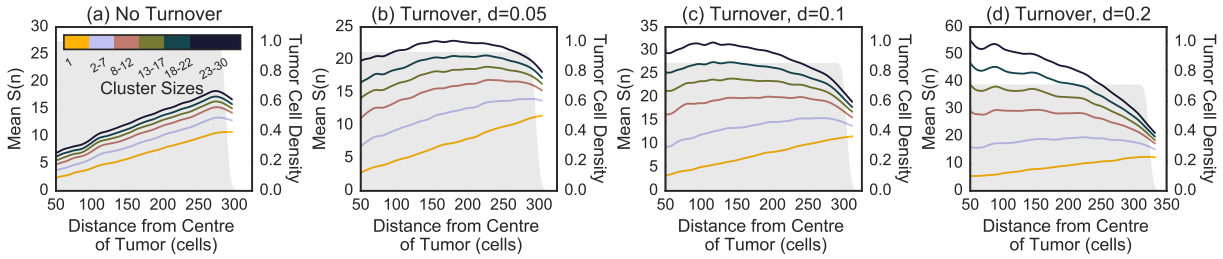



Figure 2: **Number of somatic mutations per cluster** as a function of cluster size and position for a model with (a) no turnover, (b) turnover with $d = 0.05$, (c) turnover with $d = 0.1$ and (d) turnover with $d = 0.2$. A higher number of somatic mutations increases the likelihood that a metastatic 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.

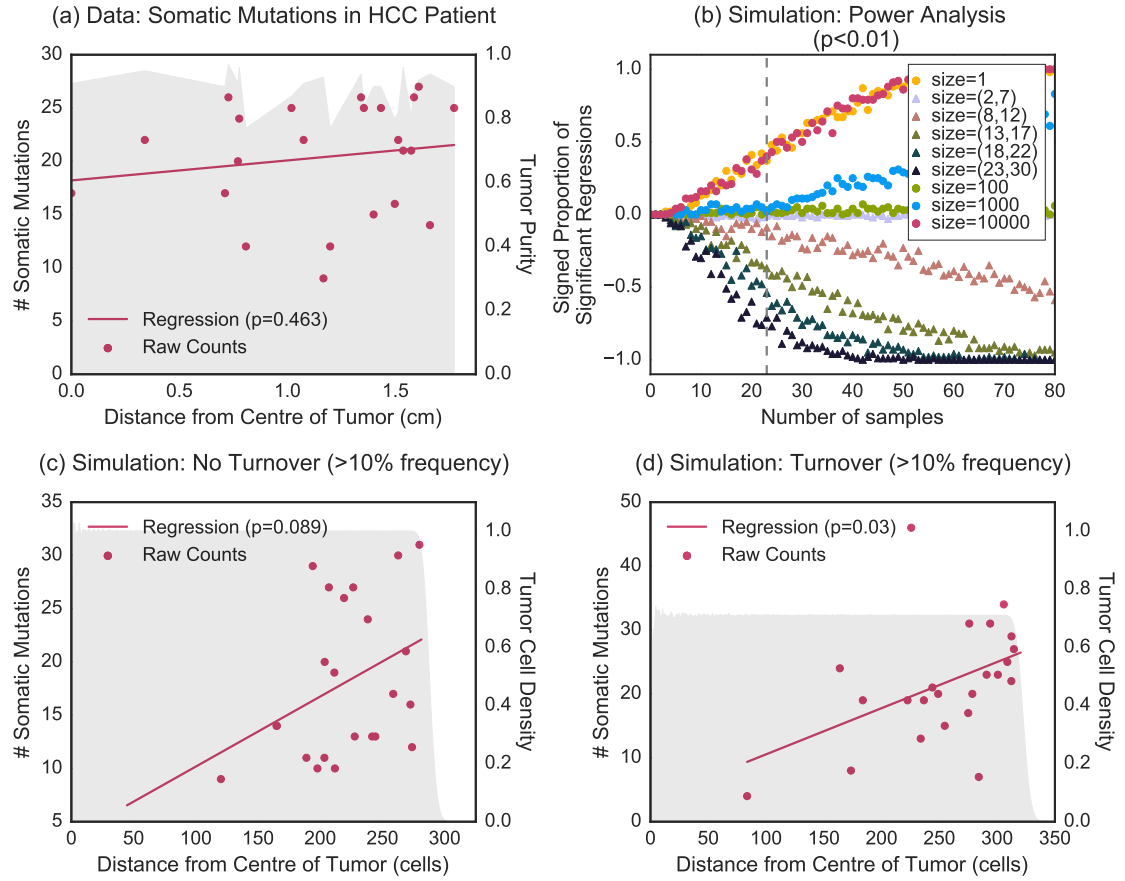



Figure 3: **Comparison of simulated multi-region NGS with empirical hepatocellular carcinoma.** Spatial distribution of the number of somatic mutations of 23 samples (20,000 cells each) in the (a) hepatocellular carcinoma patient, and (c) no turnover and (d) turnover simulated models. (b) shows the number of samples necessary to detect spatial trends from a regression analysis for CTCs and biopsies in the turnover model ($d = 0.2$). 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.

in diversity at the edge (Fig S5). Alternatively, $\approx$ 30 small cluster (23-30 cells) samples are necessary to detect spatial patterns. Furthermore, intermediate-sized clusters show qualitatively opposite trends in the different models (Fig 3b and S6). Thus small cluster sequencing may increase our power in discriminating between leading models.

## CTC clusters derived from turnover models are more likely to contain virulent mutations

Metastasis is an inefficient process [4] in that most CTCs are eliminated from the circulatory system or fail to survive in the new microenvironment. We hypothesize that the genetic composition of CTC clusters influences the likelihood of implantation into a new microenvironment. More specifically, genetic heterogeneity within a cluster may contribute to implantation by increasing the likelihood 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 virulent mutations, on average, than homogeneous ones, but they are more likely to carry *some* virulent mutations because of the increased diversity. Unless implantation probability is exactly proportional to the number of cells carrying virulent mutations in a cluster, which seems unlikely, diversity will impact implantation rate.

To compare the increased likelihood that CTC clusters possess metastatic progression genes compared to single CTCs, we determine the relative increase in the number of distinct somatic mutations in a CTC cluster versus a single CTC termed *cluster advantage*, $A(n)$. To disentangle the contributions from the microscopic and macroscopic diversity, as well as cluster size effects, we compute the cluster advantage for clusters composed of neighboring cells, as well as for random sets of cells sampled across the tumor (Fig 4).

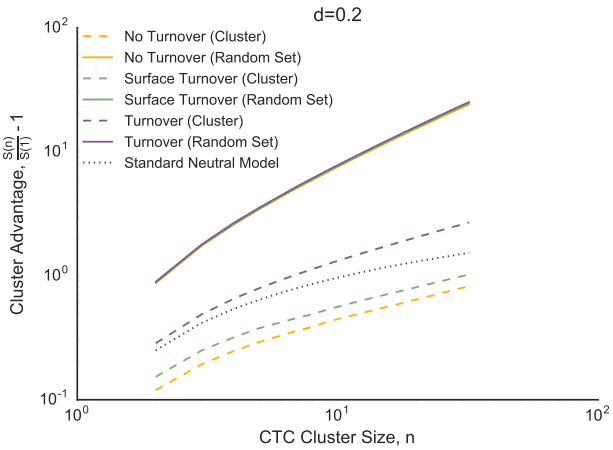



Figure 4: **'Cluster advantage' $A(n)$, or the increase in number of distinct somatic mutations in a CTC cluster relative to single CTC**, as a function of cluster size for a random subset of 500 clusters drawn uniformly across the tumor. A law of diminishing returns applies to all models because of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

Whereas randomly sampled sets of cells show similar and almost linear increase of the cluster advantage with sample size, cell clusters show more variability. Turnover models have the highest cluster advantage, followed by the surface turnover model, and the no turnover model (Fig 4). Higher turnover increases the cluster advantage (Fig S7). Even low turnover with a death rate of $d = 0.05$ doubles the cluster advantage compared to the no turnover and surface turnover model (Fig S7).

# Discussion

Even though the results of our simulations are consistent with Waclaw *et al.* at the tumor-wide level [31], we reach opposite conclusions about the effect of cell turnover on genetic diversity. Waclaw *et al.* argued that turnover reduces diversity based on the observation that more high-frequency variants were observed in the tumor with turnover: A small number of clones make up a larger proportion of the tumor. Even though we can reproduce the observation, we find that turnover models in fact vastly *in-*

*crease* diversity according to more conventional metrics, for example by increasing the number of somatic mutations (by $\approx 5.9\times$) across the frequency spectrum. Both the increase in dominant clone frequency and increased overall diversity have the same simple origin: A tumor model with turnover requires more cell divisions to reach a given size. An early driver mutation has more time to realize a selective advantage and occupy a high fraction of the tumor, but carrier cells are also more likely to accumulate new mutations along the way leading to increased diversity (Fig 1 and Table S1).

The impact of turnover on cellular heterogeneity is particularly pronounced when considering small cell clusters. These fine-scale patterns, observed in Figs 2 and S2, can be interpreted by considering the expansion dynamics of each model and their impact on cell division and mixing. In all turnover models, the number of somatic mutations in a given cell is $\approx 2.75\times$ higher at the edges than at the center of the tumor, reflecting the higher number of divisions to reach the edge: The center of the tumor is occupied early, which slows down cell division.

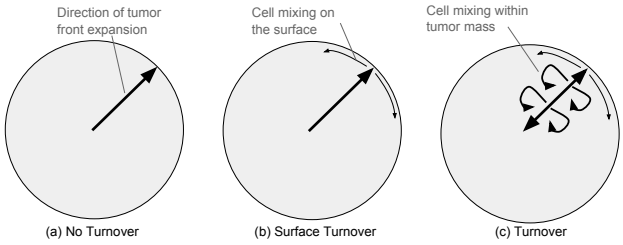



Figure 5: **Migration and Quiescent Core Explains Spatial Patterns** (a) In the no turnover model, the tumor front expands in the outward direction with no cell dying. There is little to no mixing and no divisions in the core: The number of somatic mutations increases with distance from the tumor center. (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.

In the no turnover and surface turnover models, cell clusters show the same overall pattern of additional diversity at tumor edge. In the turnover model, however, we observe the opposite pattern: Even though edge *cells* still carry the most mutations, core *clusters* are now more diverse than edge clusters.

Turnover increases diversity by increasing the number of cell divisions required to reach a given size, especially in the core. More cell divisions lead to more somatic mutations in single cells: core cells in the model with $d = 0.2$ have $\approx 3.99$ somatic mutations, compared to $\approx 1.83$ for the no turnover model. However, this has only a modest effect on the spatial patterns of diversity: without turnover, the number of somatic mutations per cell is 3.5 times higher at the edge than in the core, and the ratio is reduced to 2.2 when turnover is present ($d = 0.2$).

More importantly for diversity, turnover allows for mixing of cells from nearby clones (Fig 5c). This mixing has a smaller effect at the edge of the tumor, where the range expansion produces serial bottlenecks which reduce the effective population size relative to the tumor core. For moderate cluster sizes, this differential mixing effect overwhelms the "number of divisions" effect, and core clusters are much more diverse than edge clusters, producing distinctive gradients of diversity.

The difference in somatic diversity between single CTCs and CTC clusters, measured through the cluster advantage, follows the expected law of diminishing returns: the more cells in the cluster, the fewer the number of unique mutations per cell. However, the trends vary by growth model and cluster origin. Cell mixing afforded by turnover reduces neighboring cell similarity and increases cluster advantage.

Under the assumption that the presence or absence of a metastatic progression allele modulates metastatic potential of tumor cell clusters, the proportion of metastatic lesions that derive from circulating tumor cell clusters is highest in the turnover model. We can think of this as interference occurring between cells within a cluster. Alternately, this is an illustration of the advantage of not putting all one's egg in the same basket, applied to tumor metastasis: Assuming that there is a chance component to cluster im-

plantation, mixing increases the likelihood that at least one virulence cell makes it to a hospitable site. Such an effect should be robust to details of the growth model.

In experiments, CTC clusters derived from primary breast and prostate tumors produced more aggressive metastatic tumors [5] compared to single CTCs. This is likely due to differences in mechanical properties of the cluster or the creation of a locally favorable environment by the cluster, rather than by genetic differences. However, the present analysis suggests that this advantage can be enhanced by diversity within the cluster.

Both fine-scale mixtures of cell phenotypes and clonally constrained mutations have been observed experimentally in tumors [17, 20]. Similarly, multi-region sequencing revealed high tumor heterogeneity in clear cell renal carcinoma (ccRCC) [22], but low levels in lung adenocarcinomas [21]. This strongly suggests that the amount of migration and mixing varies substantially across tumors, with ccRCC data being better described by a model with turnover, whereas lung adenocarcinoma data more closely resembles a model with low or no turnover.

Distinguishing between migration effects, turnover effects, and tumor growth idiosyncrasies is obviously challenging. Among limitations of our model, we note the assumption of spherical tumor shape and the absence of complex physical contraints (which HCC tumors may experience). Another limitation of the present model is the rigid computational grid which prevents cells from pushing each other out of the way, which constrains growth rate in the center of the tumor. This constraint plays a role in reducing diversity at the center of the tumor, but it may not be realistic in the earlier stages of tumor growth.

The importance of such effects is largely unknown, and it is likely to vary between tumors and tumor types. Fortunately, we have shown that we are at the cusp of being able to test such models quantitatively. A sampling experiment with twice as many samples than were collected in the HCC patient studied above would enable us to either validate or reject the current state-of-the-art models (Fig 3b), and sequencing of small clusters would further allow us to discriminate between the different models studied here. The HCC data is from whole exome sequencing, as are most deep tumor sequencing datasets. We expect that power would be further increased in a whole-genome sequencing experiment, however, we were unable to perform whole-genome simulations due to memory constraints.

Future data collection schemes including the lung TRACERx study [24] will help us put the state-of-the-art models to the test and identify such important parameters of tumor growth. Given our power analysis, we find that sequencing small contiguous cell clusters provides a richer picture of tumor dynamics compared to larger biopsies, with little to no loss in power, assuming that few-cell sequencing can be performed accurately.

This work set out to answer two simple questions: First, should we expect substantial heterogeneity at the cellular scale within tumors and within circulating tumor cell clusters? The answer to the first question is most likely yes, as even the models with no turnover exhibit measurable cluster heterogeneity.

The second question was whether this heterogeneity, sampled through liquid biopsies or multi-region sequencing, is informative about tumor dynamics. Given that state-of-the-art models produce very different predictions about the level of cluster heterogeneity, the answer is also positive. This work identified some of the key factors that determine cluster diversity, especially the interaction between range expansion, cell turnover, and mixing. Even if no diversity were observed at all in CTC clusters, it would enable us to reject the present models in favor of models including additional biological factors that favor the clustering of genetically similar cells. Measuring diversity, or the lack of diversity, within circulating tumor cell clusters or fine-scale multi-region sequencing is therefore a promising tool for both fundamental and medical oncology.

## Acknowledgments

We thank Julien Jouganous, Hamid Nikbakht, Yasser Riazalhosseini, and Robert Sladek for useful discussions. This research was made possible thanks to a Canadian Institutes of Health Undergraduate Research Award in computational biology, funding reference numbers 139962 and 145987. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program and a Sloan research fellowship.

## References

[1] D. X. Nguyen, P. D. Bos, and J. Massagué. "Metastasis: from dissemination to organ-specific colonization". In: *Nature Reviews Cancer* 9.4 (2009), pp. 274–284.

[2] S. A. Eccles and D. R. Welch. "Metastasis: recent discoveries and novel treatment strategies". In: *The Lancet* 369.9574 (2007), pp. 1742–1757.

[3] C. Holohan et al. "Cancer drug resistance: an evolving paradigm". In: *Nature Reviews Cancer* 13.10 (2013), pp. 714–726.

[4] J. Massagué and A. C. Obenauf. "Metastatic colonization by circulating tumour cells". In: *Nature* 529.7586 (2016), pp. 298–306.

[5] N. Aceto et al. "Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis". In: *Cell* 158.5 (2014), pp. 1110–1122.

[6] M. Cristofanilli et al. "Circulating tumor cells, disease progression, and survival in metastatic breast cancer". In: *New England Journal of Medicine* 351.8 (2004), pp. 781–791.

[7] M. Cristofanilli et al. "Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer". In: *Journal of Clinical Oncology* 23.7 (2005), pp. 1420–1430.

[8] B. Rack et al. "Circulating tumor cells predict survival in early average-to-high risk breast cancer patients". In: *Journal of the National Cancer Institute* 106.5 (2014), dju066.

[9] D. F. Hayes et al. "Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival". In: *Clinical Cancer Research* 12.14 (2006), pp. 4218–4224.

[10] S. Maheswaran et al. "Detection of mutations in EGFR in circulating lung-cancer cells". In: *New England Journal of Medicine* 359.4 (2008), pp. 366–377.

[11] D. Marrinucci et al. "Fluid biopsy in patients with metastatic prostate, pancreatic and breast cancers". In: *Physical biology* 9.1 (2012), p. 016003.

[12] P. C. Nowell. "The clonal evolution of tumor cell populations". In: *Science* 194.4260 (1976), pp. 23–28.

[13] M. Greaves and C. C. Maley. "Clonal evolution in cancer". In: *Nature* 481.7381 (2012), pp. 306–313.

[14] D. Padua et al. "TGF$\beta$ primes breast tumors for lung metastasis seeding through angiopoietin-like 4". In: *Cell* 133.1 (2008), pp. 66–77.

[15] P. D. Bos et al. "Genes that mediate breast cancer metastasis to the brain". In: *Nature* 459.7249 (2009), pp. 1005–1009.

[16] D. X. Nguyen and J. Massagué. "Genetic determinants of cancer metastasis". In: *Nature Reviews Genetics* 8.5 (2007), pp. 341–352.

[17] N. Navin et al. "Inferring tumor progression from genomic heterogeneity". In: *Genome Research* 20.1 (2010), pp. 68–80.

[18] A. Sottoriva et al. "A Big Bang model of human colorectal tumor growth". In: *Nature genetics* 47.3 (2015), pp. 209–216.

[19] N. McGranahan and C. Swanton. "Biological and therapeutic impact of intratumor heterogeneity in cancer evolution". In: *Cancer Cell* 27.1 (2015), pp. 15–26.

[20] L. R. Yates et al. "Subclonal diversification of primary breast cancer revealed by multiregion sequencing". In: *Nature medicine* 21.7 (2015), pp. 751–759.

[21] J. Zhang et al. "Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing". In: *Science* 346.6206 (2014), pp. 256–259.

[22] M. Gerlinger et al. "Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing". In: *Nature genetics* 46.3 (2014), pp. 225–233.

[23] C. Hiley et al. "Deciphering intratumor heterogeneity and temporal acquisition of driver events to refine precision medicine". In: *Genome Biology* 15.8 (2014), p. 453.

[24] M. Jamal-Hanjani et al. "Tracking genomic cancer evolution for precision medicine: the lung TRACERx study". In: *PLoS Biology* 12.7 (2014), e1001906.

[25] A. A. Alizadeh et al. "Toward understanding and exploiting tumor heterogeneity". In: *Nature Medicine* 21.8 (2015), pp. 846–853.

[26] M. Gerlinger et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing". In: *New England Journal of Medicine* 2012.366 (2012), pp. 883–892.

[27] E. Heitzer et al. "Complex tumor genomes inferred from single circulating tumor cells by array-CGH and next-generation sequencing". In: *Cancer research* 73.10 (2013), pp. 2965–2975.

[28] A. A. Powell et al. "Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines". In: *PloS one* 7.5 (2012), e33788.

[29] M. G. Krebs et al. "Molecular analysis of circulating tumour cells-biology and biomarkers." In: *Nature Reviews Clinical Oncology* 11.3 (2014), pp. 129–44.

[30] C. L. Hodgkinson et al. "Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer". In: *Nature medicine* 20.8 (2014), pp. 897–903.

[31] B. Waclaw et al. "A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity". In: *Nature* 525.7568 (2015), pp. 261–264.

[32] J. Mateo et al. "The promise of circulating tumor cell analysis in cancer management". In: *Genome biology* 15.8 (2014), p. 448.

[33] M. Ignatiadis, M. Lee, and S. S. Jeffrey. "Circulating tumor cells and circulating tumor DNA: challenges and opportunities on the path to clinical utility". In: *Clinical Cancer Research* 21.21 (2015), pp. 4786–4800.

[34] Y. Wang et al. "Clonal evolution in breast cancer revealed by single nucleus genome sequencing". In: *Nature* 512.7513 (2014), pp. 155–160.

[35] M. J. Williams et al. "Identification of neutral tumor evolution across cancer types". In: *Nature Genetics* 48 (2016), pp. 238–244.

[36] S. Ling et al. "Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution". In: *Proceedings of the National Academy of Sciences* 112.47 (2015).

[37] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri. "NIH Image to ImageJ: 25 years of image analysis". In: *Nature Methods* 9.7 (2012), p. 671.

[38] O. Hallatschek et al. "Genetic drift at expanding frontiers promotes gene segregation". In: *Proceedings of the National Academy of Sciences* 104.50 (2007), pp. 19926–19930.

[39] D. Shweiki et al. "Induction of vascular endothelial growth factor expression by hypoxia and by glucose deficiency in multicell spheroids: implications for tumor angiogenesis". In: *Proceedings of the National Academy of Sciences* 92.3 (1995), pp. 768–772.

[40] J. M. Hou et al. "Clinical significance and molecular characteristics of circulating tumor cells and circulating tumor microemboli in patients with small-cell lung cancer". In: *Journal of Clinical Oncology* 30.5 (2012), pp. 525–532.

[41] R. Durrett. *Probability models for DNA sequence evolution*. Springer Science & Business Media, 2008.

# Supporting Information

## Methods

### Tumor growth model

To simulate the growth of solid tumors, we use TumorSimulator [31]. 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. Empty lattice sites are assumed to contain normal cells which are not modelled in TumorSimulator.

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 $\lambda_p$ and $\lambda_d$, 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.

We consider three turnover scenarios corresponding to three values of the death rate $d$: (i) No turnover ($d = 0$), corresponding to simple clonal growth [38]; (ii) Surface Turnover ($d(x, y, z) > 0$ only if $x, y, z$ is on the surface), corresponding to a quiescent core model [39] (iii) Turnover ($d > 0$ everywhere), a model favored in [31] to explore ITH.

The birth rate ($b = \ln(2)$), and selective advantage ($s = 1\%$) were kept consistent with [31]. 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 [31], and $\lambda_p = 0.0375$ to match effective experimental observations from [36].

### 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 [5, 40]. 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 [11].

### 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

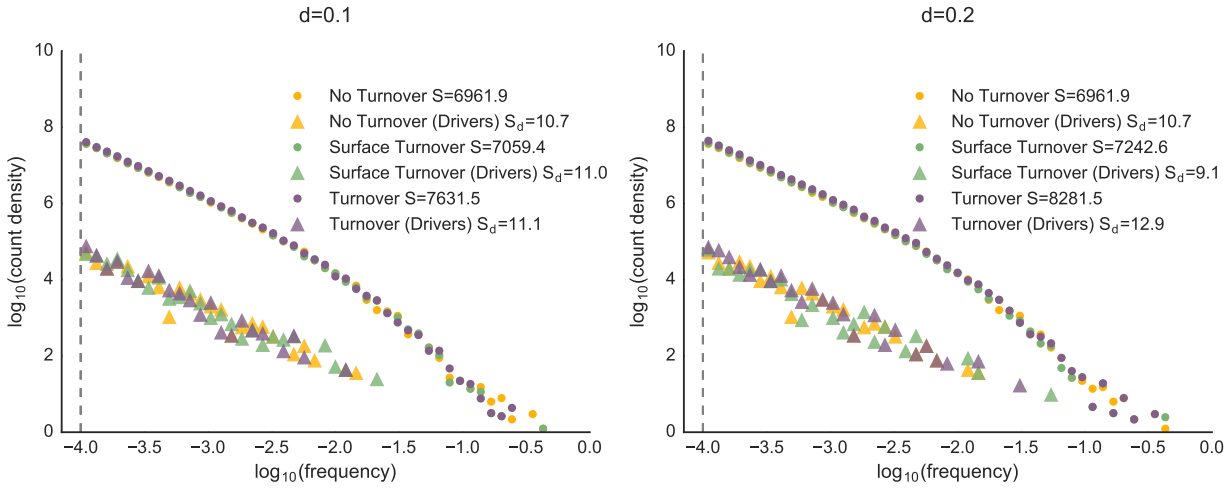
The relative increase in the number of distinct somatic mutations in a CTC cluster versus a single CTC is given by the *cluster advantage*, i.e., $A(n) = \frac{S(n)-S(1)}{S(1)} = \frac{S(n)}{S(1)} - 1$, where $S(n)$ is the number of somatic mutations in a cluster of size $n$ and $S(1)$ is the number 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 comapre 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 $E(S'(n)) = \mu H(n-1)$ [41], where $H(n)$ is the $n$-th harmonic number, $\sum_{i=1}^{n} \frac{1}{i}$.

### Code Availability

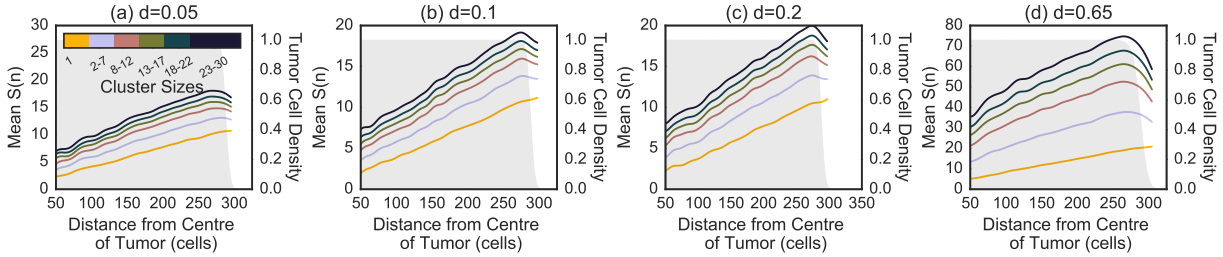
The code to reproduce simulations, analysis and figures can be found at https://github.com/zafarali/tumorheterogeneity.

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).

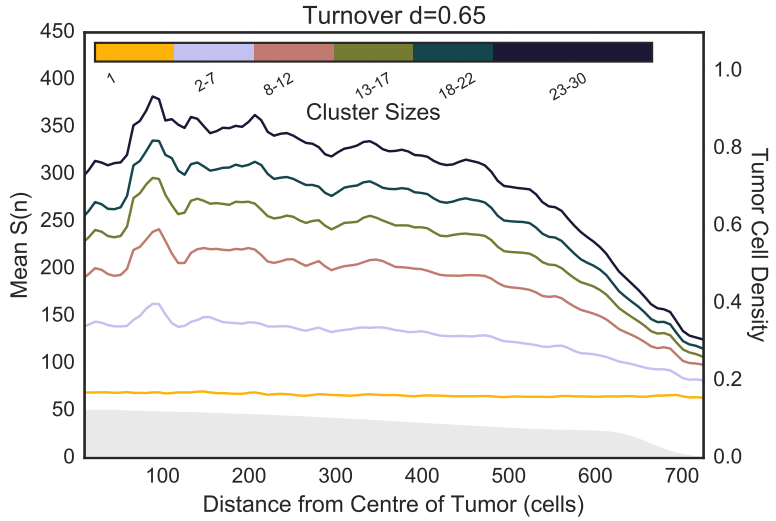
| | Average Number of Divisions in Model (mutation rate = 0.02, birth rate = 0.69) | | |
|---|---|---|---|
| Death Rate (d) | No Turnover | Surface Turnover | Turnover |
| 0.05 | $218.23 \pm 13.99$ | $216.51 \pm 13.99$ | $224 \pm 11.00$ |
| 0.1 | $218.23 \pm 13.99$ | $219.73 \pm 7.11$ | $239.38 \pm 8.06$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



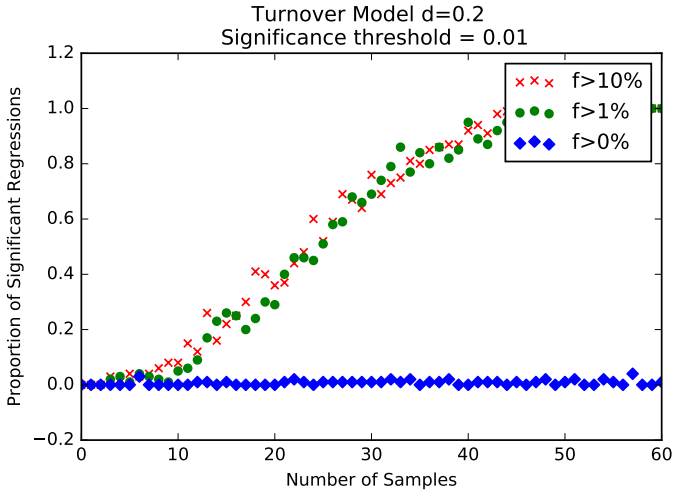

Supplementary Figure 1: Allele frequency spectra for low death rates, $d \in \{0.1, 0.2\}$ are indistinguishable.



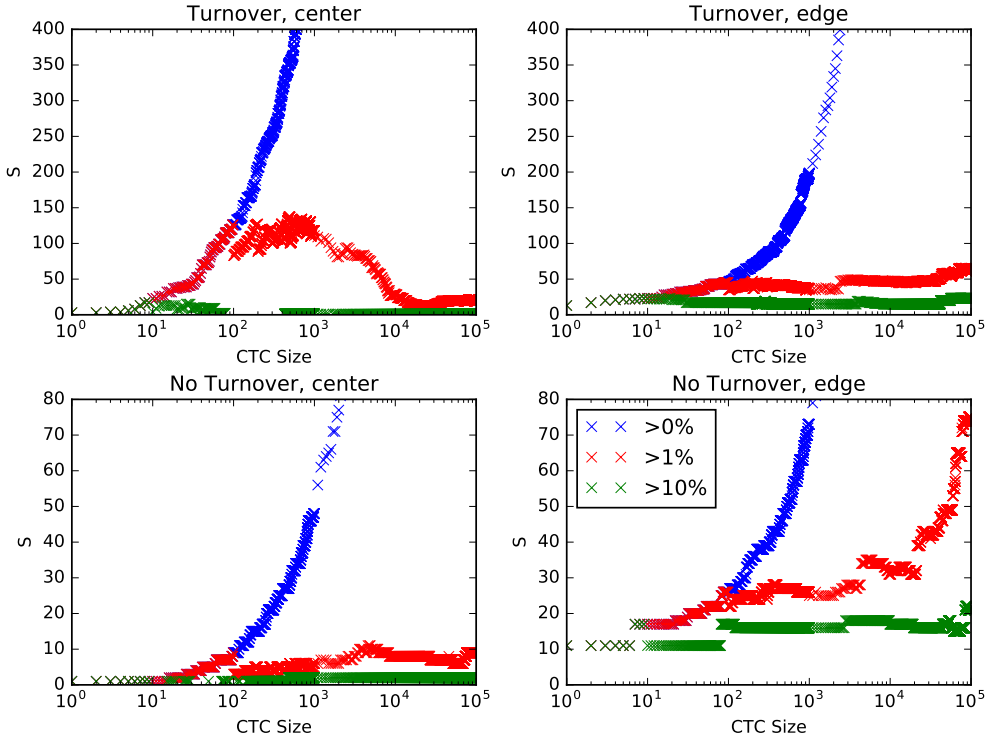

Supplementary Figure 2: The spatial distribution of the number of somatic mutations per cluster in the surface turnover model with death rates (a) $d = 0.05$, (b) $d = 0.1$, (c) $d = 0.2$ and (d) $d = 0.65$.



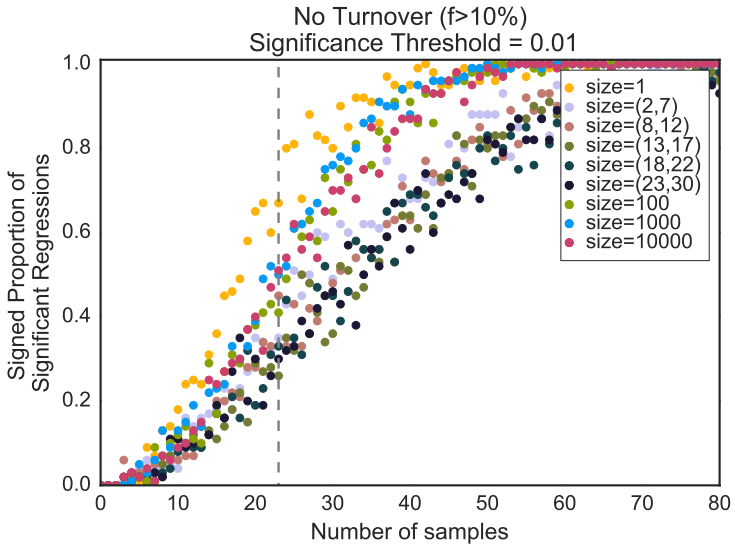

Supplementary Figure 3: The spatial distribution of the number of somatic mutation per cluster in a turnover model with $d = 0.65$.



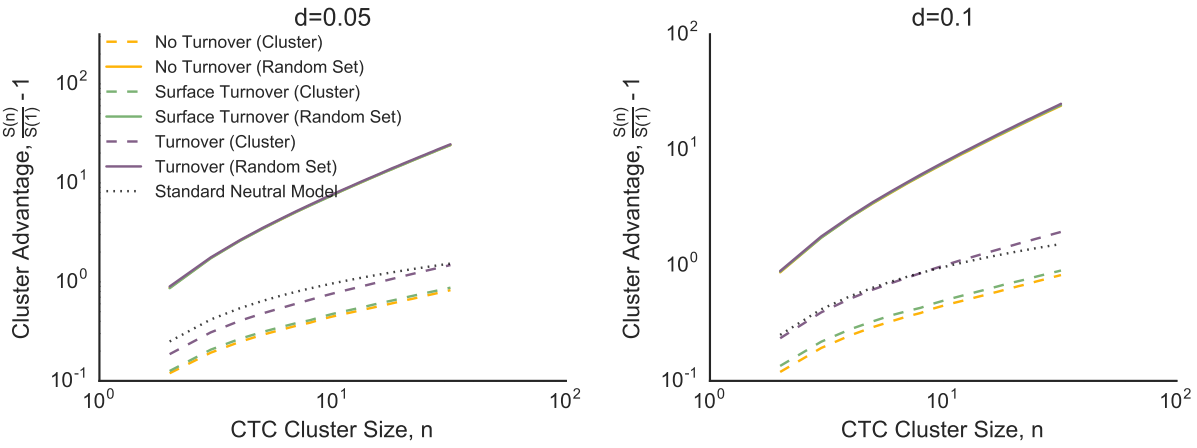

Supplementary Figure 4: The power to detect spatial trends in diversity as a function of the frequency cutoff. With no frequency cutoff, the number of rare variants in a large biopsy ($n = 20,000$ cells) overwhelms the detectable spatial pattern contributed by common variants.




Supplementary Figure 5: Number of somatic mutations observed in a sample as a function of the CTC size compared between the tumor center and edge.




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




Supplementary 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.