Received: 27 December 2021 | Revised: 18 March 2022 | Accepted: 3 April 2022

DOI: 10.1002/cso2.1033

COMPUTATIONAL AND SYSTEMS ONCOLOGY Open Access

WILEY

DATA ARTICLE

# *SITH*: An R package for visualizing and analyzing a spatial model of intratumor heterogeneity

**Phillip B. Nicol**[1] | **Dániel L. Barabási**[2] | **Kevin R. Coombes**[3] | **Amir Asiaee**[4]

[1]Department of Biostatistics, Harvard University, Boston, Massachusetts, USA

[2]Biophysics Program, Harvard University, Boston, Massachusetts, USA

[3]Department of Biomedical Informatics, Ohio State University, Columbus, Ohio, USA

[4]Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

**Correspondence**
Phillip B. Nicol, Department of Biostatistics, Harvard University, Boston, MA, USA.
Email: philnicol740@gmail.com

**Funding information**
National Institutes of Health, Grant/Award Numbers: NIH NIGMS T32GM008313, NIH T32CA009337

**Abstract**

Cancer progression, including the development of intratumor heterogeneity, is inherently a spatial process. Mathematical models of tumor evolution may be a useful starting point for understanding the patterns of heterogeneity that can emerge in the presence of spatial growth. A commonly studied spatial growth model assumes that tumor cells occupy sites on a lattice and replicate into neighboring sites. Our R package *SITH* provides a convenient interface for exploring this model. Our efficient simulation algorithm allows for users to generate 3D tumors with millions of cells in under a minute. For the distribution of mutations throughout the tumor, *SITH* provides interactive graphics and summary plots. Additionally, *SITH* can produce synthetic bulk and single-cell DNA-seq datasets by sampling from the simulated tumor. A streamlined application programming interface (API) makes *SITH* a useful tool for investigating the relationship between spatial growth and intratumor heterogeneity. *SITH* is a part of CRAN and can be installed by running `install.packages(‘‘SITH’’)` from the R console. See https://CRAN.R-project.org/package=SITH for the user manual and package vignette.

KEYWORDS

sequencing, simulation, tumor evolution

## 1 | INTRODUCTION

A comprehensive understanding of how intratumor heterogeneity (ITH) develops is critical for effective cancer diagnosis and treatment [11]. Mathematical models of cancer evolution are a promising approach for studying ITH and are free of the ethical and logistical questions associated with collecting clinical data [2]. Although the general evolutionary dynamics of cancer growth are well-characterized [8], little is known about the effect of spatial growth on ITH. Developing an in silico model that captures the evolution of a spatially embedded tumor would be a starting point for investigating this relationship. Such a model may also be useful for developing novel statistical methods which can account for samples collected from a spatially heterogeneous tumor.

A simple model of spatial tumor growth assumes that cells occupy sites on a lattice and replicate into unoccupied adjacent sites. Waclaw et al. [12] studied the dynamics of this model with selective mutations and local migration and similar models were used by [9] and [3] to study the effects of spatial heterogeneity on sequencing data.

Existing software for simulating this model is either too slow to simulate large tumors, does not allow for 3D simulation, does not simulate synthetic sequencing datasets, or is written in a low-level programming language (Supplement A). Our package 'A Spatial model of Intra-Tumor Heterogeneity (*SITH*)' implements an efficient simulation algorithm which allows the user to generate tumors with millions of cells in under a minute, entirely within R. Additionally, *SITH* can produce synthetic bulk and single-cell DNA-seq datasets from the simulated tumor. In this paper, we describe the core functionality of *SITH* and provide two examples to demonstrate its utility.

## 2 | METHODS

### 2.1 | Mathematical model

We model tumor cells as occupying sites on the three-dimensional integer lattice $\mathbb{Z}^3$ with (attempted) replication and death events occurring at times given by a Poisson process with intensity $b$ and $d$, respectively. We refer to $b$ and $d$ as the *birth rate* and *death rate*. A cell can replicate if at least one of the six adjacent sites are unoccupied. During replication, both daughter cells can acquire new genetic alterations. For each cell, the number of new alterations is drawn from a Poisson distribution with *mutation rate* $u$. With probability $u_d$, a genetic alteration is a "driver" and confers a selective advantage of $s$ to the cell. A driver mutation increases the birth rate by a factor of $s$, so that a cell with $k$ driver mutations has a birth rate of $bs^k$. Since tumors often begin with a single mutated cell [4], the initial state of our model is a single cell at the origin with $b > d$.

### 2.2 | Simulation algorithm

We simulate our model using a Gillespie algorithm [5]. Given a population of $N$ cells at time $t$, cell $i$ is chosen to replicate with probability $b_i / \sum_{j=1}^{N}(b_j + d_j)$ and die with probability $d_i / \sum_{j=1}^{N}(b_j + d_j)$. After an event is selected, the time is updated to be $t + X$, where $X$ follows an exponential distribution:

$$X \sim \text{Expo}\left(\sum_{j=1}^{N} b_j + d_j\right). \qquad (1)$$

For faster simulation, we approximate the parameter of (1) with $Np_{\max}$, where $p_{\max} = \max_j(b_j + d_j)$. See Section B of the Supporting information for a full description of the simulation algorithm.

## 3 | RESULTS

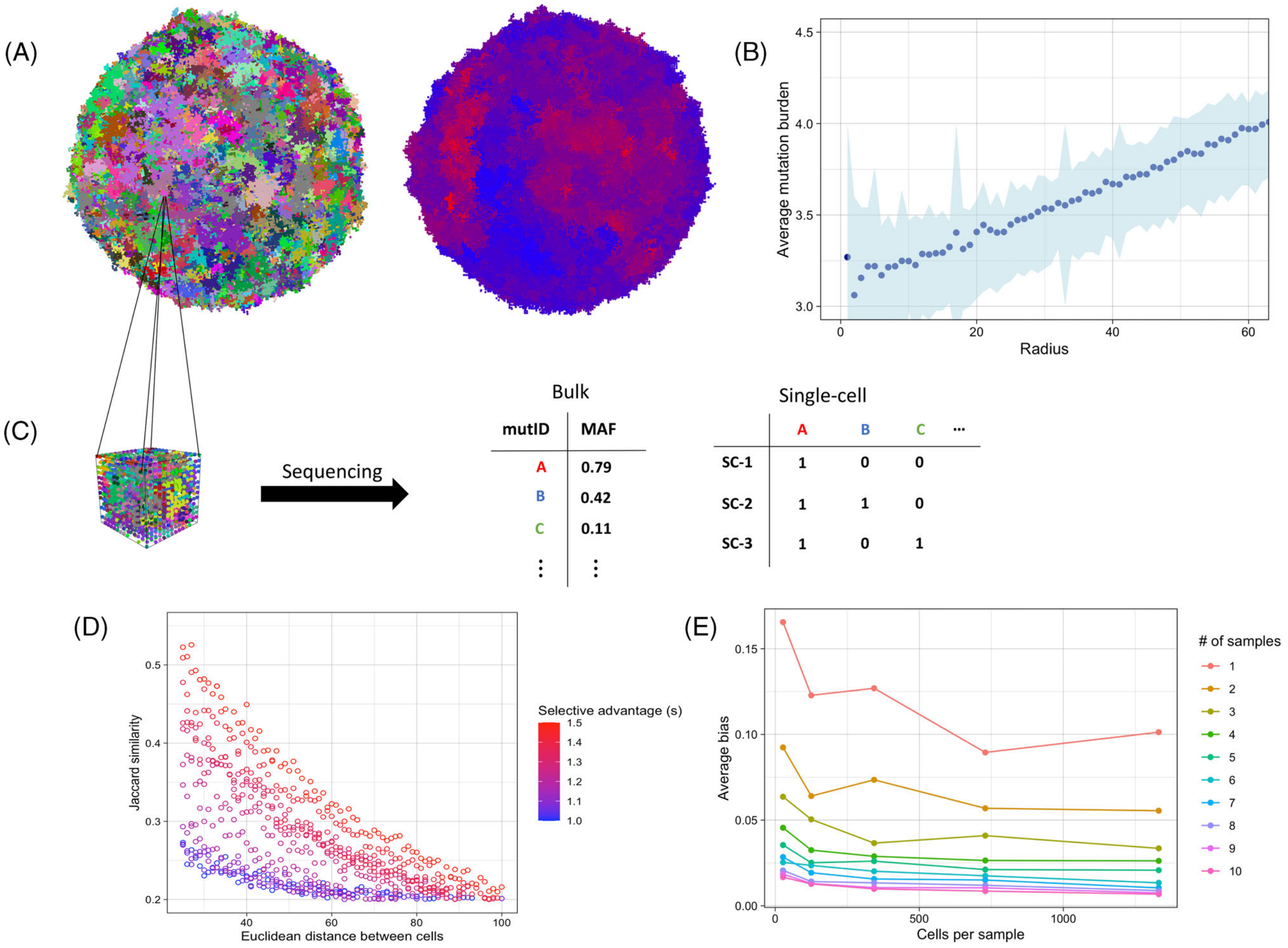
We implement the above model and simulation algorithm in the R package *SITH*. The core function of the package is `simulateTumor()`, which runs the simulation. The user can specify cell replication rate, death rate, mutation rate, and selective advantage conferred to cells with driver mutations. By default, the infinite sites model is assumed (so that each model occurs only once), but custom models of mutation can also be chosen (Supporting information C).

In silico tumors produced by *SITH* can be rendered in an interactive 3D environment through the *RGL* package [1]. As shown in Figure 1A, we have implemented two modes to visualize the tumor. On the left, each unique genotype is assigned a distinct color. On the right, cells are colored by their mutational burden, with blue corresponding to few and red corresponding to many mutations. Two-dimensional cross-sections can be visualized with the function `plotSlice()`.

### 3.1 | Driver mutations are associated with increased spatial clustering

A crucial unknown is how spatial growth biases the distribution of genetic diversity within the tumor. *SITH* was designed to provide a sandbox for asking questions about the spatial distribution of mutants within a tumor. `spatialDistribution()` produces relevant summaries of spatial heterogeneity, which can be either plotted through *SITH* (Figure 1B) or output as data for further study. For example, the function plots the average number of mutations per cell as a function of radial distance from the origin. An increasing trend suggests that the mutation burden is higher in the tumor periphery.

`spatialDistribution()` also plots a measure of genetic similarity (Jaccard index [7]) as a function of the Euclidean distance between cells. In general, we expect nearby cells to be more similar than cells on opposite sides of the tumor. *SITH* allows us to quantify how the amount of spatial clustering depends on the model parameters. To demonstrate this, we vary the selective advantage $s$ as the other model parameters are fixed at $b = 0.25$, $d = 0.18$, $u = 0.01$, $d_u = 0.003$ (standard values, see [12] or [6]). For each value of $s$, we simulate 100 tumors of size $10^6$ cells and report the average Jaccard similarity at each distance in Figure 1D. The average similarity increases as selective advantage increases, with the most significant difference at smaller distances. This suggests that increasing the strength of driver mutations can lead to increased spatial clustering and homogeneous subclones.



Bulk

| mutID | MAF |
|---|---|
| A | 0.79 |
| B | 0.42 |
| C | 0.11 |
| ⋮ | ⋮ |

Single-cell

| | A | B | C | ··· |
|---|---|---|---|---|
| SC-1 | 1 | 0 | 0 | |
| SC-2 | 1 | 1 | 0 | |
| SC-3 | 1 | 0 | 1 | |

**FIGURE 1** The main features of *SITH*. **(**A) 3D snapshots of a simulated tumor ($10^6$ cells). On the left, each unique genotype is assigned a color. On the right, regions with high mutation are colored red, while regions with low mutation are colored blue. **(**B) A plot of average mutations per cell as a function of Euclidean distance from the origin. **(**C) A cube is selected from the tumor and sequenced, returning bulk or single-cell data. **(**D) Jaccard similarity [7] as a function of Euclidean distance between cells. The genetic similarity between nearby cells increases as the selective advantage *s* increases. (E) Average bias of simulated bulk sequencing data as the number of cells and number of samples varies

## 3.2 | Multi-region sequencing leads to a reduction in bias

*SITH* is capable of generating synthetic sequencing datasets from the simulated tumor. Bulk sampling is performed by selecting all cells within an $n \times n \times n$ cube (with empty sites corresponding to normal "un-mutated" cells) and reporting the resulting mutation allele frequency (MAF). The function `bulkSample()` allows the user to select the number and location of samples to draw from the tumor. Additionally, technological noise can be introduced by decreasing the coverage.

We hypothesized that taking a local sample from a spatially structured tumor population could lead to significant bias in the estimated MAFs. To test this, we used *SITH* to obtain synthetic sequencing data from a tumor under a variety of sampling strategies. Specifically, we varied the number of cells per sample (a function of $n$) and the number of samples (cubes centered at spatially distinct locations). We quantify the accuracy of estimated MAFs by taking the mean difference between the estimated and ground truth MAFs. Figure 1E shows that the bulk sequencing procedure consistently overestimates MAFs. Importantly, the results suggest that the greatest reduction in bias is achieved by including more spatially distinct samples as opposed to increasing the number of cells per sample. Our results are consistent with a similar study from [9].

## 4 | DISCUSSION

With a straightforward API that can be used entirely within R, *SITH* provides a biologically motivated simulation of spatial tumor growth, coupled with methods for measuring ITH. Synthetic data generated from *SITH* can serve as the ground truth for benchmarking various computational methods. For example, the single-cell data could be used as input to various phylogenetic tree reconstruction algorithms, such as those presented by [10]. Additionally, *SITH* can be used to test the accuracy of algorithms designed to estimate subclonal composition, since the true MAF for each mutation is provided.

Additional features of *SITH* include simulations of metastatic seeding and treatment. For example, *SITH* can be used to simulate targeted therapy with the evolution of resistant subclones (Supporting information D). Incorporating simulations of treatment can allow for comparisons of the cancer recurrence time under a variety of surgical and therapeutic procedures. By analyzing cells near the tumor periphery, *SITH* can also provide insights into the likely genetic compositions of metastases.

## ACKNOWLEDGMENTS

Phillip B. Nicol is supported by NIH T32CA009337. Dániel L. Barabási is supported by NIH NIGMS T32 GM008313. Amir Asiaee is partially supported by NIH NHGRI R00 HG011367.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

*Phillip B. Nicol* https://orcid.org/0000-0002-8526-5889

## REFERENCES

1. D. Adler, and D. Murdoch, *rgl: 3D Visualization Using OpenGL* (2020), https://CRAN.R-project.org/package=rgl. R package version 0.100.50.
2. N. Beerenwinkel, R. Schwarz, M. Gerstrung, and F. Markowetz, *Cancer evolution: Mathematical models and computational inference*, Syst. Biol. **64** (2015), e1–e25.
3. K. Chkhaidze, T. Heide, B. Werner, M. Williams, W. Huang, G. Caravagna, T. Graham, and A. Sottoriva, *Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data*. PLoS Comput. Biol. **15** (2019), e007243.
4. G. Cooper, *The cell: A molecular approach*, 2nd ed., Sinauer Associates, 2000.
5. D. Gillespie, *Exact stochastic simulation of coupled chemical reactions*, J. Phys. Chem. **81** (1977), 2340–2361.
6. A. Heyde, J. Reiter, K. Naxerova, and M. Nowak, *Consecutive seeding and transfer of genetic diversity in metastasis*, PNAS **116** (2019), 14129–14137.
7. P. Jaccard, *The distribution of the flora in the alpine zone*, New Phytol. **11** (1912), 37–50.
8. F. Michor, Y. Iwasa, and M. Nowak, *Dynamics of cancer progression*, Nat. Rev. Cancer **4** (2004), 197–205. https://doi.org/10.1038/nrc1295.
9. L. Opasic, D. Zhou, B. Wener, D. Dingli, and A. Traulsen, *How many samples are needed to infer truly clonal mutations from heterogenous tumours?* BMC Cancer **403** (2019). https://doi.org/10.1186/s12885-019-5597-1.
10. R. Schwartz, and A. Schäffer, *The evolution of tumour phylogenetics: Principles and practice*, Nat. Rev. Gen. **18** (2017), 213–229. https://doi.org/10.1038/nrg.2016.170.
11. G. Stanta, and S. Bonin, *Overview on clinical relevance of intra-tumor heterogeneity*, Front. Med. **5** (2018), 85. ISSN 2296-858X. https://doi.org/10.3389/fmed.2018.00085. URL https://www.frontiersin.org/article/10.3389/fmed.2018.00085.
12. B. Waclaw, I. Bozic, M. Pittman, R. Hruban, B. Vogelstein, and M. Nowak, *A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity*, Nature **525** (2015), 261–264.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** P. B. Nicol, D. L. Barabási, K. R. Coombes, and A. Asiaee, SITH*: An R package for visualizing and analyzing a spatial model of intratumor heterogeneity*. Comp. Sys. Onco. (2022), **2,** e1033. https://doi.org/10.1002/cso2.1033